STATISTICAL MODELLING AND INFERENCE FOR MULTIVARIATE AND LONGITUDINAL DISCRETE RESPONSE DATA

by

James Jianmeng Xu

B.Sc., Sichuan University

M.Sc., Université Laval

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF

THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

THE FACULTY OF GRADUATE STUDIES

Department of Statistics

We accept this thesis as conforming

to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA

September 1996

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Statistics

The University of British Columbia
Vancouver, Canada

Date Sept. 30, 1996

Abstract

This thesis presents research on modelling, statistical inference and computation for multivariate discrete data. I address the problem of how to systematically model multivariate discrete response data including binary, ordinal categorical and count data, and how to carry out statistical inference and computations. To this end, I relate the multivariate models to similar univariate models already widely used in applications and to some multivariate models that hitherto were available but scattered in the literature, and I introduce new classes of models.

The main contributions in this thesis to multivariate discrete data analysis are in several distinct directions. In *modelling of multivariate discrete data* , we propose two new classification of multivariate parametric discrete models: *multivariate copula discrete (MCD) models* and *multivariate mixture discrete (MMD) models.* Numerous new multivariate discrete models are introduced through these two classes and several multivariate discrete models which have appeared in the literature are unified by these two classes. With appropriate choices of copulas, these two classes of models allow the marginal parameters and dependence parameters to vary with covariates in a natural way. By using special dependence structures, the models can be used for longitudinal data with short time series or repeated measures data. As a result, the scope of multivariate discrete data analysis is substantially broadened. In *statistical inference and computation* for multivariate models, we propose the inference function of margins (IFM) approach in which each inference function is a likelihood equation for some marginal distribution of a multivariate distribution. Examples where the approach applies are the multivariate logit model with the copulas having certain closure properties and the multivariate probit model for binary data. This general approach makes the estimation of parameters for the multivariate models computationally feasible. The corresponding asymptotic theory, the estimation of standard errors by the Godambe information matrix as well as the jackknife method, and the efficiency of the IFM approach relative to full multivariate likelihood function approach are studied. Particular attention has been given to the models with special dependence structure

(e.g. the copula dependence structure is exchangeable or AR(1) type if applicable), and efficient parameter estimation schemes based on IFM (weighting approach and pool-marginal-likelihood approach) are developed. We also give detailed assessments of the efficiency of the GEE approach for estimating regression parameters in multivariate models; this is lacking in the literature. Detailed data analyses of existing data sets are provided to give concrete application of multivariate models and the statistical inference procedures in this thesis.

Contents

List of Tables

List of Figures

Basic Notation and Definitions

The following notation and definitions are used throughout the thesis.

1. cdf stands for *cumulative distribution function*; pdf stands for *probability density function*, and pmf stands for *probability mass function*; Pr stands for *probability of.*

2. rv stands for *random variable* or *random vector* depending on the context; iid stands for *independent and identically distributed.*

3. BVN and MVN are the abbreviations for *bivariate normal* and *multivariate normal* respectively.

4. CUOM is the abbreviation for *closure under taking of margins.* MUBE is the abbreviation for *model univariate and bivariate expressible.* PUBE is the abbreviation for *parameter univariate and bivariate expressible.* MPME is the abbreviation for *model parameters marginally expressible.* (Definitions given in Section 2.1.)

5. ML and MLE are the abbreviations for *maximum likelihood* and *maximum likelihood estimates or estimation.* An MLE of $\boldsymbol{\theta}$ will usually be denoted by $\hat{\boldsymbol{\theta}}$.

6. IFM and IFME are the abbreviations for *inference functions of margins* and *inference functions of margins estimates or estimation.* An IFME of $\boldsymbol{\theta}$ will usually be denoted by $\tilde{\boldsymbol{\theta}}$. IFS is the abbreviation for *inference functions of scores.*

7. MCD and MMD are the abbreviations for *multivariate copula discrete* and *multivariate mixture discrete.*

8. The symbol "□" indicates the end of a definition, the statement of assumptions, a proof, a result, or an example.

9. For a vector or matrix, the transpose is indicated with a superscript of T or $'$, depending on convenience in the context.

10. All vectors are column vectors; hence transposed vectors such as X', x' (or X^T, x^T) are row vectors.

11. $\mathbb{R}^k = \{\mathbf{x} : \mathbf{x} = (x_1, \ldots, x_k)', -\infty < x_j < \infty \text{ for } j = 1, \ldots, k\}$ denotes the k-dimensional Euclidean space.

12. d is used for *dimension* of the multivariate response vector of multivariate distribution.

13. Boldfaced Roman upper case letter $\mathbf{Y} = (Y_1, \ldots, Y_d)'$, usually with subscripts, is used to denote a response random vector and $\mathbf{y}$ is used for the observed value of this response vector. A vector of explanatory variables or covariates is usually denoted by $\mathbf{x}$ or $\mathbf{w}$.

14. Boldfaced Roman upper case letters $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$ and so on, usually with subscripts, are used for (random) vectors, boldfaced Roman lower case letters $\mathbf{x}, \mathbf{y}, \mathbf{z}$ and so on are used for the observed vector values.

15. Roman upper case letters X, Y, Z and so on, usually with subscripts, are used for random variables, roman lower case letters x, y, z and so on are used for the observed values.

16. Greek boldfaced lower case letters, often with subscripts, are used for a collection of parameters of families of distributions, *e.g.* $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{\delta}$. They are in vector format. Greek lower case letters, often with subscripts, are used for parameters of families of distributions, *e.g.* $\alpha, \beta, \theta, \delta$.

17. Greek upper case letters Θ, Σ are used for a set of parameters (often dependence parameters) in multivariate family, they are mostly in matrix format.

18. $\Re$ is the symbol for parameter space, usually $\Re \subseteq \mathbb{R}^k$ for some k.

19. Script Roman upper case letters $\mathcal{F}$ and $\mathcal{G}$ are used for classes of functions or distribution families.

20. F, G, H are the symbols for a (multivariate) cdf.

21. For a d-variate cdf F, the set of its marginal distributions is denoted as $\{F_S : S \in S_d\}$, where S_d is the set of non-empty subsets of $\{1, \ldots, d\}$. For a specific S, the subscript is written without braces, *e.g.*, $F_1, \ldots, F_d, F_{12}, F_{123}$, etc..

22. We define the pdf or pmf of $\mathbf{Y}$ at $\mathbf{y} = (y_1, \ldots, y_d)$ as $P_{12\cdots d}(y_1 \cdots y_d)$ or simply $P(y_1 \cdots y_d)$, with the corresponding jth marginal $P_j(y_j)$, the bivariate (j,k) marginal $P_{jk}(y_j y_k)$, and so on. We also write $P(y_1 \cdots y_d; \boldsymbol{\theta})$ to denote that the pdf or pmf of $\mathbf{Y}$ depends on a parameter (or parameter vector) $\boldsymbol{\theta}$.

23. The frequency of observing a particular outcome $(y_1, \ldots, y_d)'$ in a data set is denoted by $n_{12\cdots d}(y_1 \cdots y_d)$ or simply $n(y_1 \cdots y_d)$. The frequency corresponding to the jth marginal outcome y_j is $n_j(y_j)$, and that corresponding to the (j,k) bivariate marginal outcome y_j and y_k is $n_{jk}(y_j y_k)$, and so on.

24. $\sum_{\{y_j\}}$ means the summation over all possible different values of y_j. $\sum_{\{x_1 \le y_1, \ldots, x_d \le y_d\}}$ means the summation over all possible different vector values of $\mathbf{x} = (x_1, \ldots, x_d)'$ which satisfy $\{x_1 \le y_1, \ldots, x_d \le y_d\}$. $\sum_{\{y_1 \cdots y_d\} \setminus \{y_j\}}$ means the summation over all possible different vector value $\mathbf{y} = (y_1, \ldots, y_d)'$ with the jth component absent, and so on.

25. $U(a,b)$ denotes the uniform distribution on the interval $[a,b]$. $N(\mu, \sigma^2)$ denotes the univariate normal with mean μ and variance σ^2. $N_d(\boldsymbol{\mu}, \Sigma)$ denotes the d-variate normal with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ; $\Phi_d(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ (or $\Phi_d(\mathbf{x})$) denotes the corresponding cdf and $\phi_d(\mathbf{x}, \boldsymbol{\mu}, \Sigma)$ (or $\phi_d(\mathbf{x})$) the pdf.

26. The partial derivative of a scalar function $\psi(\boldsymbol{\theta})$, $\partial\psi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$, is the $q \times 1$ vector

$$\left(\frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \psi(\boldsymbol{\theta})}{\partial \theta_q} \right)',$$

where $\theta_1, \ldots, \theta_q$ are the components of the vector $\boldsymbol{\theta}$.

27. The partial derivative of a vector function $\Psi = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_r(\boldsymbol{\theta}))'$, $\partial\Psi/\partial\boldsymbol{\theta}'$, is the $r \times q$ matrix

$$\begin{pmatrix} \frac{\partial}{\partial \theta_1}\psi_1(\boldsymbol{\theta}) & \cdots & \frac{\partial}{\partial \theta_q}\psi_1(\boldsymbol{\theta}) \\ \vdots & \ddots & \vdots \\ \frac{\partial}{\partial \theta_1}\psi_r(\boldsymbol{\theta}) & \cdots & \frac{\partial}{\partial \theta_q}\psi_r(\boldsymbol{\theta}) \end{pmatrix},$$

where $\theta_1, \ldots, \theta_q$ are the components of the vector $\boldsymbol{\theta}$.

Acknowledgements

I wish to record my warmest thanks to my thesis supervisor, Professor Harry Joe, for his continuous encouragement and for numerous suggestions and discussions during the development of this thesis. I acknowledge, with gratitude, the work of Harry Joe on multvariate copulas and dependence, and especially his invaluable book draft on "Multivariate Models and Dependence Concepts"; this has had a significant impact on the development of this thesis, and served as a foundation for this thesis. There are also many computer programming techniques and computations involved in this work where Harry Joe's tremendous experience was a crucial help.

I would like to thank Professors John Petkau and Michael Schulzer for serving on my supervisory committee. In addition, I thank Professor John Petkau for helpful discussions, his valuable comments on the thesis, as well as his encouragement and support. Also, many thanks go to Professors Nancy Heckman, James Zidek and Ruben Zamar for their encouragement and support.

Special thanks to Rinaldo Artes, Billy Ching and John Smith for their encouragement and support.

I would also like to thank my fellow students, friends, professors and staff members in the Department of Statistics at UBC for providing a pleasant and stimulating research environment.

Finally, I would like to thank Harry Joe for his financial support through an NSERC grant. The financial support from the Department of Statistics is acknowledged with great appreciation.

Chapter 1

Introduction

This chapter starts by discussing the structure of the multivariate data for which we are going to build appropriate multivariate models. We motivate our thesis research through reviewing and criticizing the relevant literature on the modelling of the multivariate discrete data.

This chapter is organized in the following way. In section 1.1, we introduce the multivariate data structure, for which we are going to develop multivariate models. In this section, we discuss in detail multivariate binary, multivariate ordinal categorical and multivariate count data. The models developed in this thesis are general in nature, but the illustrative examples will be mainly based on the forementioned three types of multivariate discrete data. In section 1.2, we briefly summarize and criticize the relevant statistical literature on the modelling of data of the types described in section 1.1, point out the inadequacies thereof, and thus motivate our thesis research. In section 1.3, we outline some desirable features of multivariate models and briefly discuss some of my understandings about statistical modelling. Section 1.4 provides an overview of the thesis.

1.1 Multivariate discrete response data

The data structure

Many data sets consist of discrete variables. Familiar examples of such variables are religion, nationality, level of education, degree of disability, attitude to a social issue, and the number of job changes for an individual during a certain period of time. These variables are categorical or count, they may be unordered (religion, nationality) or ordered (degree of disability, attitude to a social issue). In real life, what is more complicated is that often the discrete data are multivariate and

Table 1.1: The structure of general multivariate discrete data

d-variate resp.	margin-indep. cova.	margin-dep. cova.
$y_{11} \cdots y_{1d}$	$x_{11} \cdots x_{1p}$	$z_{111} \cdots z_{11p_1}, \ \cdots, \ z_{1d1} \cdots z_{1dp_d}$
$\vdots$	$\vdots$	$\vdots$
$y_{i1} \cdots y_{id}$	$x_{i1} \cdots x_{ip}$	$z_{i11} \cdots z_{i1p_1}, \ \cdots, \ z_{id1} \cdots z_{idp_d}$
$\vdots$	$\vdots$	$\vdots$
$y_{n1} \cdots y_{nd}$	$x_{n1} \cdots x_{np}$	$z_{n11} \cdots z_{n1p_1}, \ \cdots, \ z_{nd1} \cdots z_{ndp_d}$

the multiple measurements may be interdependent in some way. The dependence may be general or special. The multivariate data structure can be further complicated by having missing data, random covariates and so on.

In this thesis, we shall concentrate mainly on multivariate discrete response data, with or without covariates. The general multivariate discrete data set of interest is given in Table 1.1. The data structure in Table 1.1 consists of basically three parts: d-dimensional discrete response observations $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})'$, a *margin-independent* covariate vector of p components $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})'$, that is, a covariate vector which is constant across margins, and d *margin-dependent* (or marginal specific) covariate vectors $\mathbf{z}_{i1}, \ldots, \mathbf{z}_{id}$, where $\mathbf{z}_{ij} = (z_{ij1}, \ldots, z_{ijp_j})'$ is a vector of p_j components for the jth margin, $j = 1, \ldots, d$, $i = 1, \ldots, n$. In the longitudinal or repeated measures settings, the marginals might be defined by successive points in time. In these situations, we can call the margin-independent covariates *time-independent*, that is, constant across times, and the margin-dependent covariates *time-dependent*. The response vector $\mathbf{y}_i$ can be measures on d variates with general or special dependence structure, such as multiple measures from a human, a litter of animals, a piece of equipment, a geographical location, or any other unit for which the observations are a collection of related measures. The measures can be spatial or temporal.

One way to make inferences from such a data structure is through a multivariate parametric model. (Nonparametric multivariate inference requires much more data than parametric multivariate inference.) The development and analysis of suitable models for the multivariate data in Table 1.1 are the main objectives of this thesis.

Some typical multivariate discrete data

Binary data. Binary data arises when measurements can have only one of two values. Conventionally these are represented by 0 and 1, with 1 usually representing the occurrence of an event and 0 representing non-occurrence. For example, the reaction of a living organism to some material, often

observed as presence or absence of the reaction (usually called quantal response), is binary. Alive throughout a specified period or died during the period, won or lost, success or failure in a specified task, gender, agree or disagree, are all examples of sources of binary data. Multivariate binary data are frequent in statistical applications. Whenever multivariate data are coded, for each dimension, as one of two mutually exclusive categories, the data are multivariate binary. In the more complicated situation, covariates can be included when one is considering binary response data. An example of multivariate binary data is the Six Cities Study case analyzed in subsection 5.2.1.

Ordinal categorical data. An ordinal variable is one that has a natural ordering of its possible values, but for which the distances between the values are undefined, such as a four-category Likert scale. Ordinal categorical (or ordered categorical) response data, often accompanied with a set of covariates, arise frequently in a wide variety of experimental studies, such as in bioassay, epidemiology, econometrics and medicine. For example, in medicine it may be possible to classify a patient as, say, severely, moderately or mildly ill, when a more exact measurement of the severity of the disease is not possible; the covariates may be age, gender and so on. With ordinal variables, the categories are known to have an order but knowledge of the scale is insufficient to consider them as forming a metric. We may treat the ordinal categories simply as nominal categories – which is unordered categorical measures, but by doing so the valuable information of order is lost. So the consideration of the order is important for optimal information extraction. For an ordinal variable, it is often reasonable to assume that the ordered categories correspond to non-overlapping and exhaustive intervals of the real line. Multivariate ordinal data are frequent in applications. Whenever multivariate response variables are each ordinal categorical, the data are multivariate ordinal categorical. More complicated situations include covariates for each of the response variables. A case of multivariate ordinal data from the Three Mile Island (TMI) nuclear power plant accident study can be found in subsection 5.2.2.

Count data. Data in the form of counts appear regularly in life. In the simplest case, the number of occurrences of some phenomena on each unit are counted. Because no explanatory variable (e.g. time, treatment) distinguishes among these observed events, they can be aggregated as single numbers, the counts. Examples of count data are the counts of pest eggs on plant leaves, the counts of bacteria in different kinds of bacteria colonies, the number of organic cells with fixed number of chromosome interchanges produced by X-ray irradiation, etc.. Consul (1989) discussed many count data examples in a variety of situations, including home injuries, and strikes in industries. Other examples include the number of units of different commodities purchased by consumers over a

period of time, the number of times authors cited over a number of years, spatial patterns of plants, the number of television commercials, or the number of speakers in a meeting. Multivariate count data are also frequent in applications. Whenever multivariate response variables are each count in nature, the data are multivariate count. The more complicated situations also include covariates to the response variables. An example of multivariate count data can be found in subsection 5.2.3.

1.2 Review of literature and research motivation

For the data types we have seen in section 1.1, one of the questions is how to build a model or a probability distribution as an approximation to the stochastic phenomenon of multivariate nature, and based on the available data, to estimate the distribution, and make some inference or predictions. For this purpose, the construction of an appropriate probability distribution or statistical model in accordance with the available data generated by the stochastic phenomenon is essential.

Models for univariate discrete data have been studied extensively. The well-known *generalized linear models* for a univariate variable are such examples (McCullagh and Nelder 1989, Nelder and Wedderburn 1972). However, general studies on multivariate models for the type of data outlined in Table 1.1 are lacking in the statistical literature. One difficulty with the analysis of nonnormal multivariate data (including continuous and discrete data) has been the lack of rich classes of models such as the multivariate Gaussian. Some isolated studies on the modelling of a particular data set or under a particular multivariate setting of the type of data in Table 1.1 have appeared in the literature. These studies can be classified in general as being based either on a *completely specified probability model* or on a *partially specified probability model.* We overview some of them here, and point out their drawbacks or weaknesses.

Completely specified probability models

Exponential family: Following Cox (1972), the probability distribution for a binary random vector $\mathbf{Y}$ can be represented as a saturated log-linear model

$$P(\mathbf{y}) = \exp(u_0 + \sum_{j=1}^{d} u_j y_j + \sum_{j<k} u_{jk} y_j y_k + \cdots + u_{12\cdots d} y_1 \cdots y_d) \tag{1.1}$$

where u_0 is a normalizing constant. The $2^d - 1$ parameters $u_1, \ldots, u_d, \ldots, u_{12}, u_{13}, \ldots, u_{(d-1)d}, \ldots, u_{12\cdots d}$ vary independently from $-\infty$ to ∞. Expressions similar to (1.1) can also be found in Zhao and Prentice (1990), Liang *et al.* (1992) and Fitzmaurice and Laird (1993). The representation

(1.1) is not closed under taking of margins (see Section 2.1 for a definition). In fact, if we write $P(y_1y_2) = \exp(u_0^* + \sum_{j=1}^2 u_j^* y_j + u_{12}^* y_1 y_2)$, then u_0^*, u_j^* and u_{12}^* must depend on all the parameters u_0, u_j, u_{jk}, ..., $u_{12\cdots d}$. This fact makes the interpretation of the parameters u_j, u_{jk}, ..., $u_{12\cdots d}$ very difficult, and it is not clear how covariates could be included. For the general form, there are too many parameters.

Bahadur representation: Bahadur (1961) gave a representation of the distribution for a binary random vector $\mathbf{Y}$, in terms of the moments:

$$P(\mathbf{y}) = \prod_{j=1}^{d} P_j(1)^{y_j} P_j(0)^{1-y_j}[1 + \sum_{j<k} \rho_{jk} e_j e_k + \sum_{j<k<l} \rho_{jkl} e_j e_k e_l + \cdots + \rho_{12\cdots d} e_1 e_2 \cdots e_d] \quad (1.2)$$

where $e_j = (y_j - P_j(1))/\sqrt{P_j(1)P_j(0)}$, and $\rho_{jk} = E(e_j e_k)$, ..., $\rho_{12\cdots d} = E(e_1 e_2 \cdots e_d)$. This representation has the closure property under taking of margins, but the parameterization may not be desirable since the ρ_{jk}'s and the parameters of higher order are constrained by the marginal probabilities (see Prentice 1988), and the extension to include covariates may be difficult. For the general form, there are also too many parameters for the model to be useful.

Multivariate Poisson convolution-closed model: Teicher (1954) and Mahamunulu (1967) discussed a class of multivariate Poisson convolution-closed models. For example, a trivariate Poisson convolution-closed model has the stochastic representation

$$(Y_1, Y_2, Y_3) \stackrel{\text{def}}{=} (X_1 + X_{12} + X_{13} + X_{123}, X_2 + X_{12} + X_{23} + X_{123}, X_3 + X_{13} + X_{23} + X_{123}) \quad (1.3)$$

where X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X_{123} are independent Poisson rv's with parameters λ_1, λ_2, λ_3, λ_{12}, λ_{13}, λ_{23}, λ_{123} respectively. These models may be suitable for counts in overlapping regions or time periods if the Poisson process is a reasonable model of the underlying count process. The model has a closure property under taking of margins, but it is not "model univariate-bivariate expressible" (see Chapter 2 and Example 2.5 for further explanation of this expression), and it can only accommodate multivariate count data with a limited type of dependence range (positive dependence).

Exchangeable mixture models: Prentice (1986) gives an expression for a joint distribution of a binary random vector $\mathbf{Y}$, with

$$P(\mathbf{y}) = \frac{\prod_{j=0}^{m}(p+\gamma j)\prod_{j=0}^{d-m-1}(1-p+\gamma j)}{\prod_{j=0}^{d-1}(1+\gamma j)} \quad (1.4)$$

where $0 < p < 1$, $m = Y_1 + \cdots + Y_d$ and $\gamma \geq -(d-1)^{-1}\min\{p, 1-p\}$. The model (1.4) is an extension of a beta-Bernoulli model derived from the mixture model $P(\mathbf{y}) = \int_0^1 p^{y_+}(1-p)^{d-y_+} g(p)\, dp$, where $y_+ = \sum_{j=1}^d y_j$ and $g(p)$ is the density of a Beta(α, β) distribution. This model implies equicorrelation

of $\mathbf{Y}$ with correlation parameter of $(1+\gamma^{-1})^{-1}$. The representation (1.4) has the closure property under taking of margins, but it is limited to equicorrelation of response variables. Joe (1996) has discussions on the range of negative dependence on this family. Discussions of extensions to incorporate covariates appeared in Prentice (1986) and Connolly and Liang (1988).

Multivariate probit model: A d-variate probit model for binary data is

$$Y_j = I(Z_j \le z_j), \quad j = 1, \ldots, d, \tag{1.5}$$

where $I(\cdot)$ is indicator function, $\mathbf{Z} = (Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \Theta)$, $\Theta = (\theta_{jk})$ is a correlation matrix. The z_j's are often referred to as cut-off points. Ashford and Sowden (1970) used the bivariate probit model for binary data to describe a coal miner's status of development of breathlessness (present or absent) and wheeze (present or absent) as a function of the miner's age. Anderson and Pemberton (1985) used a trivariate probit model for the analysis of an ornithological data set on the three aspects of colouring of blackbirds. A general introduction to the multivariate probit model is Lesaffre and Molenberghs (1991). The multivariate probit model has many nice properties, such as closure under taking of margins, model univariate-bivariate expressibility, and a wide range of dependence. MLE is considered but is computationally more difficult as d increases. New approaches to estimation and inference are explored in this thesis.

Multivariate Poisson-lognormal model: Aitchison and Ho (1989) studied a model for count random vector $\mathbf{Y}$, with

$$P(\mathbf{y}) = \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{d} f_j(y_j; \lambda_j) g(\boldsymbol{\lambda})\, d\lambda_1 \cdots d\lambda_d. \tag{1.6}$$

where $f_j(y_j; \lambda_j)$ is a Poisson pmf with parameter λ_j and $g(\boldsymbol{\lambda})$ is the density of a multivariate lognormal distribution. This model also has many nice properties, such as closure under taking of margins, model univariate-bivariate expressibility, and a wide range of dependence. Again the MLE is computationally difficult.

Molenberghs-Lesaffre model: A model that may be suitable for binary and ordinal data is studied in Molenberghs and Lesaffre (1994). This model can accommodate general dependence structure from the Molenberghs-Lesaffre construction (Joe 1996) with bivariate copulas, such as in Joe (1993). The multivariate objects in the Molenberghs-Lesaffre construction have not been proved to be proper multivariate copulas, but they can be used for the parameters that lead to positive orthant probabilities for the resulting probabilities for the multivariate binary vector.

Other miscellaneous models (some for time series or longitudinal data):

- Kocherlakota and Kocherlakota (1992) provide a good summary of bivariate discrete distributions (including bivariate Poisson, bivariate negative binomial, etc.).

- Markov chain of first order for binary data with $\Pr(Y_{j+1} = 1|Y_j = 0) = P_{j,j+1}(01)$ and $\Pr(Y_{j+1} = 1|Y_j = 1) = P_{j,j+1}(11)$. It can be generalized to higher order Markov chains. Some combinations of $P_{j,j+1}(01)$, $P_{j,j+1}(11)$ and $\Pr(Y_{j+1} = 1)$ could be replaced by logistic functions (but not all three) to incorporate covariates. Examples are in Darlington and Farewell (1992), Muenz and Rubinstein (1985), Zeger, Liang and Self (1985) and Gardner (1990).

- Poisson AR(1) time series, as in Al-Osh and Alzaad (1987) and McKenzie (1988). The bivariate Poisson margin (for consecutive Y_t's) from this Poisson AR(1) time series is the same as a bivariate margin of (1.3).

- Negative binomial AR(1) time series, as in McKenzie (1986), Al-Osh and Aly (1992) and Joe (1996b). The model of Al-Osh and Aly has range of serial correlation depending on the parameters of the negative binomial distribution (and hence is not very flexible).

- When the binary or count variables are observed sequentially in time, one could use a model consisting of a product of a sequence of logit models for binary data (logit of Y_t given $Y_1, \ldots, Y_{t-1}, \mathbf{x}$) and of Poisson models for counts (Poisson of Y_t given $Y_1, \ldots, Y_{t-1}, \mathbf{x}$). This is proposed in Bonney (1987) and Fahrmeir and Kaufmann (1987). The advantage of such models is that one can use widely available software for univariate logit and Poisson models. One disadvantage of such models is that it would be difficult to predict Y_t based on $\mathbf{x}$ alone.

- Meester and MacKay (1994) studied a class of multivariate exchangeable models with the multivariate Frank copula. The models have limited application since only exchangeable dependence structures are considered.

- Glonek and McCullagh (1995) have a similar bivariate model to the Molenberghs-Lesaffre model in that the dependence parameter is linear in covariates and the related bivariate copula is the Plackett copula. Their multivariate extension appears to overlap with that of Molenberghs and Lesaffre (1994), but with a different model construction approach.

Partially specified probability models – generalized estimating equations approach

General application of many of the preceding models was impeded, however, by their mathematical complications and by the computational difficulty usually encountered in multivariate analysis.

A different body of methodology, called the generalized estimating equations (GEE) approach, has been developed based on moment-type methods which do not require explicit distributional assumptions. References for this methodology are Liang and Zeger (1986) and Zeger and Liang (1986), Zhao and Prentice (1990), Fitzmaurice and Laird (1993), among others. However the GEE approach has several disadvantages mainly related to the modelling, inference, diagnostics checking and interpretations. Furthermore, the GEE approach does not apply directly to multivariate ordinal data. A detailed study of the GEE approach, including a discussion of some of its shortcomings, can be found in Chapter 6.

Research motivation

In summary, although some approaches have appeared in the literature to model specific instances/examples for the data in Table 1.1, there are at least two major features lacking in the statistical literature in terms of modelling multivariate discrete data:

1. A unified, systematic approach to multivariate discrete modelling, with classes of models for multivariate discrete data where some models in the class have nice properties (see section 1.3 for some desirable features of multivariate models).

2. A model-fitting strategy with computationally feasible parameter estimation and inference procedures, with good asymptotic properties and efficiency.

This thesis makes contributions to these two lackings in multivariate discrete (more generally, nonnormal) data modelling. We study systematic approaches to the modelling of multivariate discrete response data with covariates. The response types include binary, ordinal categorical and count. Statistical inference and computational aspects of the multivariate nonnormal models are studied.

1.3 Statistical modelling

We discuss here two issues in statistical modelling. One is what we mean by statistical modelling in general. The other is the construction of multivariate models with desirable properties. Other aspects of statistical modelling as part of data analysis will be discussed in Chapter 5.

In practice, with a finite sample of data, to capture exactly the possibly complex multivariate system which generated the data is impossible. The problem can even be more complicated than modelling a system; it might be that the system itself does not exist and it is forever a hypothetical

one. In statistical modelling, the specification of a particular model for the data is always somehow arbitrary; what we hope is that the stochastic models we use may reflect relatively well the randomness or uncertainty in the system, as well as the significant features of the systematic relationships. The statistical models should be considered as a means of providing statistical inference; they should be viewed as tentative approximations to the truth. The most important consideration in using any statistical method (or model) is whether the method (or model) can give insight into important practical problems. All models are subjective in some degree. Often the modeller chooses those elements of the system under investigation that should be included in the model as well as the mode of representation. Modelling should not be a substitute for thinking and will only be effective if combined with an interest in and knowledge of the system being modelled.

The construction of multivariate nonnormal models is not easy. For modelling purposes, we would like to have parametric families of models that (i) cover the different types of dependence, (ii) have interpretable parameters, and (iii) apply to multivariate discrete data. Some desirable properties of a multivariate model are the following:

1. The model is natural. That is, the model is interpretable in terms of mixture, stochastic or latent variable representations, etc..

2. Parameters in the model are interpretable. A parametric family has extra interpretability if some of the parameters can be identified as dependence or multivariate parameters, such that some range of the parameters corresponds to positive dependence and some corresponds to negative dependence, and it is desirable to have the amount of dependence to be increasing as parameters increase.

3. The model allows wide and flexible range of dependence, with interpretable dependence parameters which are flexible to the needs for different applications.

4. The model extends naturally to include covariates for the univariate marginal parameters as well as dependence parameters, in the sense that after the extension, we still have probabilistic model and proper interpretations.

5. The model has marginally expressible properties, such as model parameters expressible by parameters in univariate and bivariate distributions property and closure property with extension of univariate to bivariate and to higher order margins.

6. The model has a simple form, preferably with closed form representations of the cdf and density, or at least is easy to use for computation, estimation and inference.

Generally, it is not possible to achieve all of these desirable properties simultaneously, in which case one must decide the relative importance of the properties and sacrifice one or more of them. There is no known multivariate family having all of these properties but the family of multivariate normal distributions may be the closest. Multinormal distributions satisfy (1), (2), (3), (4) and (5) but not (6) since the multinormal has no closed form cdf. The mixture of max-id copulas (Joe and Hu 1996) satisfy (1), (2), (3), (4) and (6) but only partially (5). In Chapter 3, these desirable properties of a multivariate model will be used as criteria to compare different models.

1.4 Overview of thesis

This thesis consists of seven chapters. In Chapter 2, we develop the theoretical background for the multivariate discrete models, statistical inference and computation procedures. Two general classes of multivariate discrete models are introduced; their common feature is that both rely on the copula concept. Several new concepts related to multivariate models are proposed. The asymptotic theory for parameter estimation based on the inference functions of margins (IFM) is also given in this chapter. In Chapter 3, we study and compare many specific models in the two general classes of multivariate models proposed in Chapter 2. Mathematical details for parameter estimation for some of the models are provided. In Chapter 4, the efficiency of IFM approach relative to the classical maximum likelihood approach is investigated. The major advantage of IFM is its computational feasibility and its good asymptotic properties. We demonstrate that IFM is an efficient parameter estimation approach when it is applicable. We also study the efficiency of the jackknife method of variance estimation proposed in Chapter 2. In Chapter 5, some important issues such as a proper data analysis cycle, model selection and diagnostic checking are discussed. Data analysis examples illustrating modelling and inference procedures developed in this thesis are also carried out. In Chapter 6, we study the usefulness and efficiency of the GEE approach which has been the focus of many recent statistical applications dealing with multivariate and longitudinal data with univariate margins covered by the theory of generalized linear models. In Chapter 7, the final chapter, we discuss some further important research topics closely related to this thesis work. Finally, the Appendix contains a Maple symbolic manipulation program example used in Chapter 4.

Chapter 2

Foundation: models, statistical inference and computation

In this chapter, we propose two classes of multivariate discrete models: *multivariate copula discrete (MCD) models* and *multivariate mixture discrete (MMD) models.* These two classes of models provide a new classification of multivariate discrete models, and allow a general approach to modelling multivariate discrete data. The two classes unify a number of multivariate discrete models appearing in the literature, such as the multivariate probit model, multivariate Poisson-lognormal model, etc. At the same time, numerous new models are proposed under these two classes. We also propose an *inference functions of margins (IFM)* approach to parameter estimation for MCD and MMD models. This estimation approach is built on the general theory of inference functions (or estimating equations). Asymptotic theory for IFM is developed and applied to the specific models in Chapter 3. While similar ideas about the same kind of estimating functions for a specific model have appeared in the literature, the general development of the procedure as an approach for the parameter estimation for a class of multivariate discrete models, and the related asymptotic results, are new. We also show that a jackknife estimate of the covariance matrix of the estimates from the IFM approach is asymptotically equivalent to the asymptotic covariance matrix from the Godambe information matrix. The jackknife procedure has the advantage of general computational feasibility. These results are used extensively in the applications in Chapter 5. The efficiency of IFM versus the optimal estimation procedure based on maximum likelihood estimation and the numerical assessment of the efficiency of jackknife covariance matrix estimates compared with Godambe information

matrix are studied in detail in Chapter 4.

The present chapter is organized as follows. Section 2.1 introduces the multivariate copula models, some multivariate dependence concepts and a number of new concepts regarding the properties of a multivariate model. In section 2.2, we introduce two classes of multivariate discrete models: the *multivariate copula discrete models* and the *multivariate mixture discrete models.* These two classes of models are the focus of this thesis, and specific models in these two classes will be extensively studied in Chapter 3. In section 2.3, we propose an *inference functions of margins (IFM)* approach for the parameter estimation of MCD and MMD models; the theoretical foundation is built on the theory of inference functions for the multi-parameter situation. Section 2.4 is devoted to the study of the asymptotic properties of parameter estimates based on the IFM approach. Under regularity conditions, the IFM estimators (IFME) for parameters are shown to be consistent and asymptotically normal with a Godambe information matrix as the variance-covariance matrix. These are done for the models with no covariates as well as models with covariates. The extension of models with no covariates to models with covariates will be made clear in this section. In section 2.5, we propose a jackknife approach to the asymptotic variance estimation of IFME, and show theoretically that the jackknife estimate of variance is asymptotically equivalent to the Godambe information matrix. The importance of the jackknife estimate of variance will be demonstrated in Chapter 5 for real data analysis.

2.1 Multivariate copulas and dependence measures

2.1.1 Multivariate distribution functions

We begin by recalling the definition of a distribution on $I\!R^d$.

Definition 2.1 *A d-dimensional distribution function is a function* $F : I\!R^d \to I\!R$, *which is right continuous, with*

$$\text{(i)} \lim_{y_j \to -\infty} F(y_1, \ldots, y_d) = 0, \quad j = 1, \ldots, d, \qquad \text{(ii)} \lim_{y_j \to \infty \forall j} F(y_1, \ldots, y_d) = 1$$

and which satisfies the following rectangle inequality: For all $(a_1, \ldots, a_d), (b_1, \ldots, b_d)$ *with* $a_j < b_j$, $j = 1, \ldots, d$,

$$\sum_{k_1=1}^{2} \cdots \sum_{k_d=1}^{2} (-1)^{k_1 + \cdots + k_d} F(x_{1k_1}, \ldots, x_{dk_d}) \geq 0, \tag{2.1}$$

where $x_{j1} = a_j, x_{j2} = b_j$. □

The following are several remarks related to Definition 2.1:

i. If F has dth order derivatives, then (2.1) is equivalent to $\partial^d F/\partial y_1 \cdots \partial y_d \geq 0$.

ii. Letting $a_2, \ldots, a_d \to -\infty$, then (2.1) reduces to $F(b_1, b_2, \ldots, b_d) - F(a_1, b_2, \ldots, b_d) \geq 0$, so F is increasing in the first variable. Similarly, by symmetry, F is increasing in the remaining variables.

iii. Let S be a subset of $\{1, \ldots, d\}$. The margins F_S of $F(y_1, \ldots, y_d)$ are obtained by letting $y_i \to \infty$ for $i \notin S$.

There are two important types of cdf generated from a random vector $\mathbf{Y}$: discrete and continuous. In the case of an absolutely continuous random vector $\mathbf{Y}$, there is a corresponding density function $f(y_1, \ldots, y_d)$ which satisfies $f(y_1, \ldots, y_d) \geq 0$ and $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(y_1 \cdots y_d) dy_1 \cdots dy_d = 1$. The cdf can be written by

$$F(y_1, \ldots, y_d) = \int_{-\infty}^{y_d} \cdots \int_{-\infty}^{y_1} f(x_1 \cdots x_d) dx_1 \cdots dx_d.$$

In the case of a discrete random vector $\mathbf{Y}$, the probability that $\mathbf{Y}$ takes on a value $\mathbf{y} = (y_1, \ldots, y_d)'$ is defined by the pmf

$$P(y_1 \cdots y_d) = \Pr(Y_1 = y_1, \ldots, Y_d = y_d),$$

which satisfies $P(y_1 \cdots y_d) \geq 0$ and $\sum_{\{y_1\}} \cdots \sum_{\{y_d\}} P(y_1 \cdots y_d) = 1$. The cdf can be written as

$$F(y_1, \ldots, y_d) = \sum_{\{x_1 \leq y_1, \ldots, x_d \leq y_d\}} P(x_1 \cdots x_d).$$

For a discrete random vector, the jth marginal distribution is defined by

$$P_j(y_j) = \sum_{\{y_1 \cdots y_d\} \setminus \{y_j\}} P(y_1 \cdots y_d).$$

The (j, k) marginal distribution is defined by

$$P_{jk}(y_j y_k) = \sum_{\{y_1 \cdots y_d\} \setminus \{y_j y_k\}} P(y_1 \cdots y_d).$$

In general, the marginal distributions can be obtained from the previous remark (iii).

2.1.2 Multivariate copulas and Fréchet bounds

The multivariate normal distribution is used extensively in multivariate analysis because of its many nice properties (see for example, Seber 1984). The wide range of successful application of the

multivariate normal distribution is because of its flexibility in representing different dependence structures rather than for physical reasons or as an approximation from the Central Limit Theorem. The dependence structure plays a crucial role in multivariate modelling. But the multivariate normal model is not sufficient for every multivariate modelling situation. To be able to model multivariate data in general, a good understanding of the general parametric families of multivariate distribution functions – the constructs which describe the characteristic of the random phenomena, is necessary. One useful and well-known approach to understanding a multivariate distribution function F is to express F in terms of its marginals and its associated dependence function $C(\cdot)$. This $C(\cdot)$ (or simply C) is commonly called the *copula.*

Definition 2.2 (Copula) *C is a copula if*

$$G(y_1, \ldots, y_d) = C(G_1(y_1), \ldots, G_d(y_d)).$$

is a distribution function, whenever G_1, ..., G_d are all arbitrary univariate distribution functions. □

Let $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ be a d-variate continuous random vector with cdf $G(y_1, \ldots, y_d)$ and with continuous univariate marginal distribution functions $G_1(y_1)$, ..., $G_d(y_d)$ respectively. Then $U_1 = G_1(Y_1)$, ..., $U_d = G_d(Y_d)$ are uniformly distributed on $[0, 1]$. Let $G_1^{-1}, \ldots, G_d^{-1}$ be the univariate quantile functions, where G_j^{-1} is defined by $G_j^{-1}(t) = \inf\{y : G_j(y) \geq t\}$, $j = 1, \ldots, d$. The copula, C, of $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ is constructed by making marginal probability integral transforms on $Y_1, \ldots, Y_d$ to $U_1, \ldots, U_d$. That is, the copula is the joint distribution function of $U_1, \ldots, U_d$:

$$C(u_1, \ldots, u_d) = G(G_1^{-1}(u_1), \ldots, G_d^{-1}(u_d)). \tag{2.2}$$

C is non-unique if the G_j's are not all continuous. This point will be made clear in section 2.2. Suppose $\mathbf{Y}$ is a continuous random vector with distribution function $G(y_1, \ldots, y_d)$ and the corresponding copula is $C(u_1, \ldots, u_d)$ with density function $c(u_1, \ldots, u_d)$. The density function of $G(y_1, \ldots, y_d)$ in terms of copula density function is $g(y_1, \ldots, y_d) = c(G_1(y_1), \ldots, G_d(y_d)) \prod_{j=1}^{d} g_j(y_j)$.

The copula captures the dependence among the components of the random vector $\mathbf{Y}$; it contains all of the information that couples the d marginal distributions together to yield the joint distribution of $\mathbf{Y}$. This understanding is essential for the subsequent development of the multivariate discrete models. The copula was first introduced by Sklar (1959). For parametric families of copulas with good properties, see Joe (1993, 1996). Through the copula, a distribution function is decomposed into two parts: a set of marginal distribution functions and the dependence structure which is

specified in terms of the copula. This suggests that one natural way to model multivariate data is to find the dependence structure in terms of copula and the univariate marginals separately. This important feature will be extended to form multivariate discrete models by using the copula concept in section 2.2 and in Chapter 3.

Next we define the Fréchet bounds.

Definition 2.3 (Fréchet bounds) *Let $F(\mathbf{x})$ be a d-variate cdf with univariate margins $F_1, \ldots, F_d$. Then for $\forall x_1, \ldots, x_d$,*

$$\max\{0, F_1(x_1) + \cdots + F_d(x_d) - (d-1)\} \leq F(x_1, \ldots, x_d) \leq \min\{F_1(x_1), \ldots, F_d(x_d)\}, \qquad (2.3)$$

where $\min\{F_1(x_1), \ldots, F_d(x_d)\}$ is the Fréchet upper bound, and $\max\{0, F_1(x_1)+\cdots+F_d(x_d)-(d-1)\}$ is the Fréchet lower bound. □

We state here some important properties of the Fréchet bounds.

Properties

1. The Fréchet upper bound is a cdf.
2. The Fréchet lower bound is a cdf for $d = 2$.
3. The Fréchet upper bound copula is $C_U(\mathbf{u}) = \max\{u_1, \ldots, u_d\}$. For $d = 2$, the Fréchet lower bound copula is $C_L(\mathbf{u}) = \min\{0, u_1 + u_2 - 1\}$.

For a proof of the properties 1,2,3 and other properties of Fréchet bounds, see Joe (1996).

Under independence, the copula is

$$C_I(u_1, \ldots, u_d) = \prod_{j=1}^{d} u_j,$$

and any copula must pointwise fall between $\max\{0, u_1 + \cdots + u_d - (d-1)\}$ and $\min\{u_1, \ldots, u_d\}$.

2.1.3 Dependence measures

It is desirable for a parametric family of multivariate distributions to have a flexible and wide range of dependence. For non-normal random variables, correlation is not the best measure of dependence, and concepts based on linearity are not necessarily the best to work with. More general concepts of positive and negative dependence and measures of monotone dependence are needed. These are necessary for analyzing the type of dependence and range of dependence in a parametric family of

multivariate models. For a thorough treatment of dependence concepts and dependence orderings, see Joe (1996, Chapter 2).

In multivariate analysis, one of the most important activities is to model the dependence structure among the random variables. The complexity of the dependence structure and its range often determines the practical usefulness of the model. The dependence structure of a multivariate model can be considered somehow equivalent to the copula; for example, Schweizer and Wolff (1981) used copulas to define several natural nonparametric measures of dependence for pairs of random variables. The parameters in a multivariate copula reflect the degree of dependence among variables; for example, the multivariate normal copula can be adequately expressed in terms of a correlation matrix of which the elements consist of the pairwise correlation coefficients of a multinormal random vector, with a large correlation coefficients indicating strong dependence among variables. However, it is not always possible to express a copula in term of correlation coefficients of a set of random variables. There is also a mathematical reason, *e.g.* mathematical simplicity, not to express a copula in terms of correlation coefficients.

A measure of dependence for two random variables indicates how closely these two random variables are related. The extreme situations would be mutual independence and complete mutual dependence. Some very useful dependence concepts, such as positive and negative dependence concepts, are based on the refinement of some intuitive understanding of dependence among random variables. For example, for two random variables X and Y, the positive dependence concept means roughly that large (small) values of X tend to accompany large (small) values of Y. Often in practice, this knowledge of the amount of dependence is good enough for some modelling purposes.

Some well-known measures of dependence for two random variables are Pearson's correlation coefficient r, Spearman's rho and Kendall's tau. These measures are defined as follows: Let X, Y be random variables with continuous distribution function $F(x)$ and $G(y)$ and copula C. We further assume that (X_1, Y_1), (X_2, Y_2) and (X, Y) are independent with the same joint distribution. Then Pearson's correlation coefficient is $r = \mathrm{Cov}(X, Y)/\sigma_X \sigma_Y$ or

$$r = \frac{1}{\sigma_X \sigma_Y} \int_0^1 \int_0^1 [C(u, v) - uv]\, dF^{-1}(u)\, dG^{-1}(v),$$

Kendall's tau is $\tau = \mathrm{Corr}(sgn(X_1 - X), sgn(Y_1 - Y)) = 2\mathrm{Pr}((X_1 - X)(Y_1 - Y) > 0) - 1$, or

$$\tau = 4 \int_0^1 \int_0^1 C(u, v)\, dC(u, v) - 1,$$

and Spearman's rho is $\rho = \mathrm{Corr}(sgn(X_1 - X), sgn(Y_2 - Y))$ or

$$\rho = 12 \int_0^1 \int_0^1 [C(u, v) - uv]\, du\, dv,$$

where σ_X and σ_Y stand for the standard deviation of random variables X and Y, $sgn(\cdot)$ denotes the sign function. Both Kendall's tau and Spearman's rho are invariant to strictly increasing transformations. They are equal to 1 for the Fréchet upper bound and -1 for the Fréchet lower bound. These properties do not hold for Pearson's correlation. Essentially, Pearson's r measures the strength of the linear relationship between two random variables X and Y, whereas the Kendall's tau and Spearman's rho are measures of monotone correlation (strength of monotone relationship). For bivariate quantitative data, Spearman's rho corresponds to the rank correlation (Pearson's correlation applied to the ranks of the 2 variables). That the copula captures the basic dependence structure among the components of $\mathbf{Y}$ can be seen by the fact that all nonparametric measures of association, such as Kendall's tau, Spearman's rho, are normed distances of the copula from the independence copula. In general, it is difficult to judge the intensity of dependence for a given multivariate model solely based on one dependence measure; the three common dependence measures can be used as a reference for the attainable intensity of the dependences of a given multivariate model.

For ease of interpretation of the dependence structure, we would like to see the dependence structure expressed in easily interpretable parameters. For example, for arbitrary marginals, a question is how to express a copula in terms of the most common measures of association, such as Pearson's r (from some specific marginals), Spearman's rho, or Kendall's tau, in a natural way. For some well-defined classes of distribution, such as the multivariate normal, Pearson's correlation coefficient is the measure of choice. In other classes, other measures may be more appropriate. (For example, the Morgenstern copula in subsection 2.1.4 is expressed in terms of Kendall's tau in a natural way.) A parametric family has extra interpretability if some of the parameters can be identified as dependence parameters. More specifically, one would like to be able to say that some range of the parameters corresponds to positive dependence and some corresponds to negative dependence. Furthermore, it would be desirable to have the amount of dependence to be increasing (decreasing) as parameters increase (decrease).

2.1.4 Examples of multivariate copulas

Some well-known examples of copula families are: the multivariate normal copula, Morgenstern copula, Plackett copula, Frank copula, etc. Joe (1993, 1996) provides a detailed list of families of copulas with good properties. In Genest and Mackay (1986), a class of copulas, called Archimedean copulas, is studied extensively. Most existing parametric families of copulas represent monotone dependence structures where the intensity of the dependence is determined by the value of the

dependence parameter. Some families, such as the normal family, possess a complete range of dependence intensities whereas others, such as the Morgenstern family, possess only a limited range. In fact, the Morgenstern copula never attains the Fréchet bounds; Spearman's rho lies between $-1/3$ and $1/3$. For general modelling purposes, we would naturally seek families with a wide range of dependence intensities.

Here we give some examples of multivariate copulas. More examples of multivariate copulas will be given in Chapter 3 for constructing multivariate discrete models.

Example 2.1 (Multivariate normal copula) Let Φ be the standard normal distribution function and let Φ_d be the d-variate normal distribution function with mean vector $\mathbf{0}$, variance vector $\mathbf{1}$ and correlation matrix Θ. Then the corresponding d-variate copula is

$$C(u_1, \ldots, u_d; \Theta) = \Phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); \Theta), \tag{2.4}$$

where every bivariate copula $C_{jk}(u_j, u_k; \theta_{jk})$, $1 \leq j < k \leq d$, attains the lower Fréchet bound, the independence case, or the upper Fréchet bound according to $\theta_{jk} = -1, 0$, or 1. Pearson's correlation coefficient for the corresponding bivariate normal distribution is $r = \theta_{jk}$. For Spearman's rho and Kendall's tau, we can also establish the following relationships: $\tau = (2/\pi)\sin^{-1} r$ and $\rho = (6/\pi)\sin^{-1}(r/2)$. With this copula, we have

$$G(z_1, \ldots, z_d) = C(G_1(z_1), \ldots, G_d(z_d); \Theta) = \Phi_d(\Phi^{-1}(G_1(z_1)), \ldots, \Phi^{-1}(G_d(z_d)); \Theta),$$

where the G_j's are arbitrary cdfs. For example, we can have $G_j(z_j) = \exp(z_j)/(1 + \exp(z_j))$, $G_j(z_j) = \Phi(z_j)$, $G_j(z_j) = 1 - \exp(-\exp(z_j))$ or $G_j(z_j) = \exp(-\exp(-z_j))$. $G_j(z_j)$ can even be taken as a mixture distribution function, for example, $G_j(z_j) = \pi_j \Phi(z_j) + (1 - \pi_j) \int_{-\infty}^{z_j} \exp(-|x|)/2 \, dx$, where $0 \leq \pi_j \leq 1$. These flexible choices of univariate marginal distributions combined with the complete range of the dependence parameter matrix Θ make the multivariate normal copula a powerful copula for general modelling purposes. In Chapter 3, we will use this copula extensively. □

Example 2.2 (Morgenstern copula) In the literature, sometimes the names of several people are put together in naming this copula; Farlie-Gumbel-Morgenstern copula is one of them. In this thesis, we simply call this copula the *Morgenstern copula* (Morgenstern 1956). One simpler version of a d-dimensional Morgenstern copula, which does not include higher order terms, is

$$C(u_1, u_2, \ldots, u_d) = \left[1 + \sum_{j<k}^{d} \theta_{jk}(1 - u_j)(1 - u_k)\right] \prod_{h=1}^{d} u_h. \tag{2.5}$$

It has density function

$$c(u_1, u_2, \ldots, u_d) = 1 + \sum_{j<k}^{d} \theta_{jk}(1-2u_j)(1-2u_k).$$

The conditions on the parameters θ_{jk} so that (2.5) is indeed a copula are straightforward. For $d = 3$, the conditions can be conveniently summarized as follows: $1+\theta_{12}+\theta_{13}+\theta_{23} \geq 0$, $1+\theta_{13} \geq \theta_{23}+\theta_{23}$, $1+\theta_{12} \geq \theta_{13}+\theta_{23}$, $1+\theta_{23} \geq \theta_{12}+\theta_{13}$, or more succinctly $-1+|\theta_{12}+\theta_{23}| \leq \theta_{13} \leq 1-|\theta_{12}-\theta_{23}|$, $-1 \leq \theta_{12}, \theta_{13}, \theta_{23} \leq 1$. Similar conditions for higher dimension $d = 4, 5, \ldots$, can also be obtained by considering the 2^d cases for $u_j = 0$ or 1, $i = 1, \ldots, d$, and verifying that $c(u_1, \ldots, u_d) \geq 0$.

It is easy to see that for any $j, k = 1, \ldots, d; j \neq k$,

$$C_{jk}(u_j, u_k; \theta_{jk}) = [1 + \theta_{jk}(1-u_j)(1-u_k)]\, u_j u_k, \quad -1 \leq \theta_{jk} \leq 1,$$

with density function

$$c_{jk}(u_j, u_k) = 1 + \theta_{jk}(1-2u_j)(1-2u_k).$$

The dependence structure between U_j and U_k is controlled by the parameter θ_{jk}. Spearman's rho is $\rho = \theta_{jk}/3$. The maximum Pearson's correlation coefficient over all choices of G_j and G_k is $1/3$ (when $\theta_{jk} = 1$) which occurs for uniform marginals. For normal marginals, the maximum of the Pearson's correlation coefficient is $1/\pi$; for exponential marginals it is $1/4$; for double exponential marginals, the limit is 0.281. Kendall's tau is $2\theta_{jk}/9$, with the maximum range of $-2/9$ to $2/9$. Because of the dependence range limitation, the Morgenstern copula is not very useful for general modelling. Nevertheless, because the Morgenstern copula has such a simple form, it can be used as an investigation tool in, for example, simulation studies to check properties of some general modelling procedures. An example of its use is provided in section 4.3. If a new procedure breaks down with a distribution based on the Morgenstern copula, then it will probably have difficulties with other models that admit a wider range of dependence.

A version of the d-dimensional Morgenstern copula with higher order terms has the following density function

$$\begin{aligned} c(u_1, u_2, \ldots, u_d) = 1 &+ \sum_{j_1<j_2}^{d} \beta_{j_1 j_2}[1-2u_{j_1}][1-2u_{j_2}] \\ &+ \sum_{j_1<j_2<j_3}^{d} \beta_{j_1 j_2 j_3}[1-2u_{j_1}][1-2u_{j_2}][1-2u_{j_3}] + \cdots + \beta_{12\cdots d}\prod_{j=1}^{d}(1-2u_j). \end{aligned} \tag{2.6}$$

This form expands the correlation structure of the Morgenstern distribution (2.5). For more details, see Johnson and Kotz (1975, 1977). □

2.1.5 CUOM, CUOM(k), MUBE, PUBE and MPME concepts

In this thesis, we are mainly interested in (a) parametric models (or copulas) with wide range of dependence intensities, and (b) parametric models (or copulas) with certain types of marginal distribution closure properties. In this subsection, we introduce several concepts for the marginal behaviour of a distribution.

Definition 2.4 (Closure under taking of margins, or CUOM) *A parametric model (copula) is said to have the property of closure under taking of margins, if the bivariate margins and higher-order margins belong to the same parametric family.* □

Definition 2.5 (Closure under taking of margins of order k, or CUOM(k)) *A parametric model (copula) is said to have the property of closure under taking of margins of order k, if the k-variate margins belong to the same parametric family.* □

Definition 2.6 (Model univariate-bivariate expressible, or MUBE) *A parametric model (copula) is called model univariate-bivariate expressible, or MUBE, if all the parameters in the model can be expressed by parameters in the model's univariate and bivariate marginal distributions.* □

Definition 2.7 (Parameter univariate-bivariate expressible, or PUBE) *If a parameter in a model can be expressed by the model's univariate and bivariate marginal distributions, then this parameter is called parameter univariate-bivariate expressible, or PUBE, under the model.* □

Definition 2.8 (Model parameters marginally expressible, or MPME) *If all the parameters in a model can be expressed by the model's lower dimensional (lower than full) marginal distributions, then the model is said to have the property of model parameters being marginally expressible or the MPME property.* □

If we are thinking about parameter estimation, then the expressions such as "expressible" and "be expressed" in the above definitions should be understood as "estimable" and "be estimated" respectively from lower-dimensional margins.

A model with CUOM is also said to have reproducibility or upward compatibility under taking of margins. Basically, the marginal distributions "reproduce" themselves under taking of margins. This property is desirable in many applications in multivariate because initial data analysis often starts with lower dimensional margins.

A model with the MUBE property means that all the parameters appearing in the multivariate distribution appear in univariate and bivariate marginal distributions. A model with the PUBE property may have multivariate parameters of order higher than 3, but part of its parameters of interest can be univariate or bivariate expressed without the involvement of other multivariate parameters (e.g. trivariate parameters). A model with the MPME property means that all of its parameters may be expressed marginally. These are important properties of a multivariate model that allows for a simplification in the parameter estimation through the IFM approach (defined in section 2.3).

Based on the above definitions, the following implications hold:

i. If a model has the CUOM property, then it also has the CUOM(k) property. If a model is not CUOM(k), then it is not CUOM.

ii. CUOM(r_1) implies CUOM(r_0) if $r_1 > r_0$. That is, there exists a parameterization of the lower dimensional margins so that the lower order closure property hold.

iii. If a model has the MUBE property, then all the parameters in this model are PUBE. Furthermore, this model is also MPME.

iv. If every parameter is PUBE, then the model is MUBE.

No other implications hold in general.

In the following, a few examples are used to illustrate the above concepts and some of their relationships.

Example 2.3 (Models with CUOM and MUBE properties) A familiar example of a model with the CUOM and PUBE properties is the multivariate normal model. The closure under taking of margins for the multinormal distribution is somewhat stronger than the CUOM property defined here, since it is also closed under taking of univariate margins, which is not required in our definition. □

Example 2.4 (Models with MUBE property) For some copulas, such as (2.4) and (2.5), the dependence structure can be expressed by a $d \times d$ matrix parameter $\Theta = (\theta_{jk})$ with $\theta_{jj} = 1$. For such a d-dimensional copula $C(\cdot; \Theta)$, the 2-dimensional margins can be expressed by a bivariate copula $C_{jk}(\cdot; \theta_{jk})$ with one dependence parameter θ_{jk}, for $j, k = 1, \ldots, d; j \neq k$. Thus each element in the dependence structure described by the parametric matrix $\Theta = (\theta_{jk})$ can be equivalently expressed

by a set of bivariate copulas $C_{jk}(\cdot;\theta_{jk})$. The distribution with this copula is thus MUBE. Some copulas such as (2.4) have a wide range of dependence; some such as (2.5) do not. □

Example 2.5 (Models with CUOM but not MUBE property) We give two examples here:
a. Consider the generalized Morgenstern copula (2.6). This copula has the CUOM property, since for any $\{j_1,\ldots,j_m\} \in \{1,\ldots,d\}$ where $m < d$, it is straightforward to verify that $c(u_{j_1}, u_{j_2},\ldots,u_{j_m})$ has the form (2.6). But this generalized Morgenstern copula is not MUBE.
b. Another example is the multivariate Poisson distribution. Let us examine the trivariate Poisson distribution. Let the random variables X_1, X_2, X_3, X_{12}, X_{13}, X_{23}, X_{123} have independent Poisson distributions with the mean parameters $\lambda_1, \lambda_2, \lambda_3, \lambda_{12}, \lambda_{13}, \lambda_{23}, \lambda_{123}$ respectively. We now construct new random variables as follows:

$$Y_1 = X_1 + X_{12} + X_{13} + X_{123},$$
$$Y_2 = X_2 + X_{12} + X_{23} + X_{123},$$
$$Y_3 = X_3 + X_{13} + X_{23} + X_{123}.$$

Using the convolution property of the Poisson, we derive that $Y_1 \sim Po(\lambda_1 + \lambda_{12} + \lambda_{13} + \lambda_{123})$, $Y_2 \sim Po(\lambda_2 + \lambda_{12} + \lambda_{23} + \lambda_{123})$, $Y_3 \sim Po(\lambda_3 + \lambda_{13} + \lambda_{23} + \lambda_{123})$, that (Y_1, Y_2), (Y_1, Y_3), (Y_2, Y_3) have bivariate Poisson distributions, and that (Y_1, Y_2, Y_3) has a trivariate Poisson distribution. This 3-dimensional Poisson model has the CUOM property because the bivariate margins have a similar stochastic representation. But it is not MUBE nor PUBE. In fact, with univariate and bivariate margins, we can only estimate $\lambda_1 + \lambda_{13}$, $\lambda_2 + \lambda_{23}$ and $\lambda_{12} + \lambda_{123}$ from the $(1,2)$ margins, $\lambda_1 + \lambda_{12}$, $\lambda_3 + \lambda_{23}$ and $\lambda_{13} + \lambda_{123}$ from the $(1,3)$ margins, and $\lambda_2 + \lambda_{12}$, $\lambda_3 + \lambda_{13}$ and $\lambda_{23} + \lambda_{123}$ from the $(2,3)$ margins. These nine linear expressions form only six independent linear expressions. Since we have seven parameters in the model, thus the model is not MUBE. Furthermore, it can be easily verified that no any single parameter can be univariate-bivariate expressed. □

Example 2.6 (Models with MUBE but not CUOM(2) property) Consider a trivariate copula constructed in the following way:

$$C_{123}(u_1, u_2, u_3) = \int_0^{u_2} C_{1|2}(u_1|x;\delta_{12}) C_{3|2}(u_3|x;\delta_{23})\, dx, \tag{2.7}$$

where $C_{1|2}$ and $C_{3|2}$ are conditional cdfs obtained from two arbitrary bivariate copulas families $C_{12}(u_1, u_2;\delta_{12})$ and $C_{23}(u_2, u_3;\delta_{23})$. This trivariate copula has $(1,2)$ bivariate margin $C_{12}(u_1, u_2;\delta_{12})$, $(1,3)$ bivariate margin $C_{13}(u_1, u_3) = \int_0^1 C_{1|2}(u_1|x;\delta_{12}) C_{3|2}(u_3|x;\delta_{23})\, dx$, and $(2,3)$ bivariate margin $C_{23}(u_2, u_3;\delta_{23})$. Suppose we can let C_{12} be the Plackett copula

$$C(u, v;\delta) = 0.5\eta^{-1}\{1 + \eta(u+v) - [(1+\eta(u+v))^2 - 4\delta\eta uv]^{1/2}\}, \quad 0 \le \delta < \infty, \tag{2.8}$$

where $\eta = \delta - 1$, and we can let C_{23} be the Frank copula

$$C(u, v; \delta) = -\delta^{-1} \log([\xi - (1 - e^{-\delta u})(1 - e^{-\delta v})]/\xi), \quad 0 \leq \delta < \infty, \tag{2.9}$$

where $\xi = 1 - e^{-\delta}$. Then the model (2.7) is well-defined, and is obviously MUBE with 2 bivariate dependence parameters δ_{12} and δ_{23}. But the model (2.7) is not CUOM(2), since the Plackett copula and the Frank copula are not in the same parametric family.

Generally speaking, given bivariate distributions F_{12}, F_{23} with univariate margins F_1, F_2, F_3, it can be shown that

$$F_{123}(y_1, y_2, y_3; \theta_{12}, \theta_{13}, \theta_{23}) = \int_{-\infty}^{y_2} C_{13}(F_{1|2}(y_1|z_2; \theta_{12}), F_{3|2}(y_3|z_2; \theta_{23}))\, F_2(dz_2) \tag{2.10}$$

is a proper trivariate distribution with univariate margins F_1, F_2, F_3, (1,2) bivariate margin F_{12}, and (2,3) bivariate margin F_{23}. In (2.10), $F_{1|2}, F_{3|2}$ are conditional cdfs obtained from F_{12}, F_{23}, and C_{13} is a bivariate copula associated with the $(1, 3)$ margin (it can be interpreted as a copula representing the amount of conditional dependence in the first and third univariate margin given the second). Specifically, $C_{13}(u_1, u_3) = u_1 u_3$ corresponds to conditional independence and $C_{13}(u_1, u_3) = \min\{u_1, u_3\}$ corresponds to perfect conditional dependence. The model (2.10) is MUBE, but it may not be CUOM(2) – it is enough to see this fact by choosing F_{12} and F_{23} from different parametric family. The model (2.7) is a special case of (2.10) obtained by letting F_{12}, F_{23} be the Plackett and Frank copulas respectively, and $C_{13}(u_1, u_3) = u_1 u_3$. The construction (2.10) is a special case of Joe (1996a). □

Example 2.7 (Models with CUOM(2) but not CUOM property) Let $F(u, v; \theta) = uv(1 + \theta(1 - u)(1 - v))$, $-1 \leq \theta \leq 1$, be the bivariate Morgenstern family (2.5). Let F_{12} and F_{23} are in this family with parameters θ_{12} and θ_{23} respectively. Let $C_{13}(u_1, u_3) = u_1 u_3$. The conditional distributions are $F_{j|2}(u_j|u_2) = u_j + \theta_{j2} u_j (1 - u_j)(1 - 2u_2)$, $j = 1, 3$. Hence by (2.10), we have

$$F_{13}(u_1, u_3) = \int_0^1 F_{1|2}(u_1|z_2) F_{3|2}(u_3|z_2)\, dz_2 = u_1 u_3 [1 + 3^{-1} \theta_{12} \theta_{23} (1 - u_1)(1 - u_3)],$$

which is in the bivariate Morgenstern family (2.5) with parameter $\theta_{12}\theta_{23}/3$. Hence the model

$$F_{123}(u_1, u_2, u_3) = \int_0^{u_2} F_{1|2}(u_1|z_2) F_{3|2}(u_3|z_2)\, dz_2 \tag{2.11}$$

is CUOM(2). But (2.11) is not CUOM. In fact, we find

$$\begin{aligned} F_{123}(u_1, u_2, u_3) =& u_1 u_2 u_3 [1 + \theta_{12}(1 - u_1)(1 - u_2) + 3^{-1} \theta_{12} \theta_{23} (1 - u_1)(1 - u_3) + \\ & \theta_{23}(1 - u_2)(1 - u_3) + 2\theta_{12}\theta_{23}(1 - u_1)(1 - u_2)(1 - u_3)(1 - 2u_2)/3], \end{aligned}$$

which is not in the trivariate Morgenstern family (2.5). □

Example 2.8 (Models with CUOM(r_0) but not CUOM(r_1) property, when $r_0 < r_1$) Consider a 4-variate copula model:

$$F_{1234}(u_1,u_2,u_3,u_4) =$$
$$u_1u_2u_3u_4[1+\theta_{12}(1-u_1)(1-u_2)+3^{-1}\theta_{12}\theta_{23}(1-u_1)(1-u_3)+\theta_{14}(1-u_1)(1-u_4)+$$
$$\theta_{23}(1-u_2)(1-u_3)+\theta_{24}(1-u_2)(1-u_4)+\theta_{34}(1-u_3)(1-u_4)+$$
$$2\theta_{12}\theta_{23}(1-u_1)(1-u_2)(1-u_3)(1-2u_2)/3],$$

where $|\theta_{14}+\theta_{24}+\theta_{34}|-\theta_{12}-1<\theta_{23}(1+\theta_{12})<1+\theta_{12}-|\theta_{14}+\theta_{24}-\theta_{34}|$, $|\theta_{14}-\theta_{24}-\theta_{34}|+\theta_{12}-1<\theta_{23}(1-\theta_{12})<1-\theta_{12}-|\theta_{14}-\theta_{24}+\theta_{34}|$ and $|\theta_{jk}|\le 1$, $1\le j<k\le 4$. It can be shown that F_{12}, F_{13}, F_{14}, F_{23}, F_{24}, and F_{34} are in the bivariate Morgenstern family (2.5), but F_{123}, F_{124}, F_{134} and F_{234} are not in the same parametric family. In fact, F_{124}, F_{134} and F_{234} are in the trivariate Morgenstern family (2.5), but F_{123} is not. □

Example 2.9 (Models with PUBE but not MPME property) We give two examples here:
a. In the generalized Morgenstern copula (2.6), the parameters $\beta_{j_1j_2}$ $(1\le j_1<j_2\le d)$ are PUBE, but the model is not MPME, as the parameter $\beta_{12\cdots d}$ cannot be expressed by any marginal copula.
b. Another example is the Molenberghs-Lesaffre model in Example 2.17. The parameters η_j $(1\le j\le d)$ and η_{jk} $(1\le j<k\le d)$ are PUBE, but the model is not MPME, as the parameter $\eta_{12\cdots d}$ cannot be expressed by any marginal pmf. □

2.2 Multivariate discrete models

Assume $\mathcal{F}$ is a parametric family defined on a common measurable space $(\mathcal{Y},\mathcal{A})$, where $\mathcal{Y}$ is a discrete sample space and $\mathcal{A}$ the corresponding σ-field. We further assume

$$\mathcal{F}=\{P(\mathbf{y};\boldsymbol{\theta}):\boldsymbol{\theta}\in\Re\},\quad \Re\subseteq I\!R^q, \tag{2.12}$$

where $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_q)'$ is a q-component vector, and $\Re$ is the parameter space. The parameter space is usually a subset of q-dimensional Euclidean space. We presume the existence of a measure μ on $\mathcal{Y}$ such that for each fixed value of the parameter $\boldsymbol{\theta}$, the function $P(\mathbf{y};\boldsymbol{\theta})$ is the density with respect to μ of a probability measure $\mathcal{P}$ on $\mathcal{Y}$. For a d-dimensional random discrete vector $\mathbf{Y}=(Y_1,\ldots,Y_d)'$, its pmf $P(y_1\cdots y_d;\boldsymbol{\theta})$ (or simply $P(y_1\cdots y_d)$) is assumed to be in $\mathcal{F}$.

2.2.1 Multivariate copula discrete models

We define a cdf of a discrete random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ as

$$G(y_1, \ldots, y_d) = C(G_1(y_1), \ldots, G_d(y_d)), \tag{2.13}$$

where C is a d-dimensional copula and G_j $(j = 1, \ldots, d)$ is the cdf of the discrete rv Y_j. Thus $G(y_1, \ldots, y_d)$ is a well-defined cdf for a discrete random vector $\mathbf{Y}$. The pmf of $\mathbf{Y} = \mathbf{y} = (y_1, \ldots, y_d)'$ is

$$P(y_1 \cdots y_d) = \sum_{k_1=1}^{2} \cdots \sum_{k_d=1}^{2} (-1)^{k_1+\cdots+k_d} C(x_{1k_1}, \ldots, x_{dk_d}), \tag{2.14}$$

where $x_{j1} = G_j(y_j), x_{j2} = G_j(y_j^*)$ with $G_j(y_j) < G_j(y_j^*)$ and for any x such that $y_j < x < y_j^*$, we have $\Pr(Y_j = x) = 0$. We call the model (2.13) for a discrete random vector $\mathbf{Y}$ a *multivariate copula discrete (MCD) model.*

The family of MCD models is a big family. With MCD models, we have flexible choices of marginal cdfs, including standard distributions such as Bernoulli, binomial, negative binomial, Poisson and generalized Poisson, etc., and these allow the models to accommodate a wide range of data. We may also have flexible choices of copulas; examples are multinormal copula, Hüsler-Reiss copula, Morgenstern copula, etc.. For a summary of properties of MCD models, see subsection 2.2.4.

For a given d-variate discrete distribution F, we can often find multiple copulas which match F into a MCD model. For example, suppose we have a bivariate binary random vector $\mathbf{Y} = (Y_1, Y_2)'$, where Y_j $(j = 1, 2)$ takes values 0 and 1. The probability of observing $(1, 1)$, $(1, 0)$, $(0, 1)$ and $(0, 0)$ are $P(11)$, $P(10)$, $P(01)$ and $P(00)$ respectively. Then for any given one-parameter family of bivariate copulas $C(u_1, u_2; \theta)$ that ranges from the Fréchet lower bound to upper bound, we can find a θ to express the four probability masses in the following way

$$\begin{cases} C(u_1, u_2; \theta) = P(11), \\ u_1 = P(11) + P(10), \\ u_2 = P(11) + P(01). \end{cases} \tag{2.15}$$

(2.15) may not hold if $C(\cdot; \theta)$ cannot attain the Fréchet bounds. The above observation suggests that to model multivariate discrete data, different copulas could do the modelling job equally well. To make the modelling successful in the general sense, it is important that the copula has a wide dependence range. Evidently, with different copulas, we will not be estimating the same dependence parameters, but nevertheless the fitted model should lead to the similar inference or interpretations.

2.2.2 Multivariate mixture discrete models

Multivariate discrete models can be constructed in different ways than the derivation of MCD models. We can envisage circumstances that the multivariate discrete random vector $\mathbf{Y}$ at $\mathbf{y} = (y_1, \ldots, y_d)'$ has pmf $f(y_1 \cdots y_d; \boldsymbol{\lambda})$ for a given $\boldsymbol{\lambda}$. Suppose further that $\boldsymbol{\lambda}$ is a random outcome which we assume to be a p-component vector (p may be different from d) subject to chance variation described by a certain (continuous) multivariate distribution $G(\lambda_1, \ldots, \lambda_p)$, which in turn can be expressed in terms of a copula function $C(u_1, \ldots, u_p)$ with (continuous) univariate marginal distribution G_j $j = 1, \ldots, p$. This is similar to imagining a group of outcomes, with random traits or effects for the individuals in the group, and having a common constant trait or element through the distribution of the random effects. Then the probability of $\mathbf{Y} = \mathbf{y}$, or the pmf of $\mathbf{Y}$ at $\mathbf{y}$ is

$$P(y_1 \cdots y_d) = \int \cdots \int f(y_1 \cdots y_d; \boldsymbol{\lambda}) c(G_1(\lambda_1), \ldots, G_p(\lambda_p)) \prod_{j=1}^{p} g_j(\lambda_j) d\lambda_1 \cdots d\lambda_p. \tag{2.16}$$

We call (2.16) a *multivariate mixture discrete (MMD) model.* We use the word mixture since the distribution function is constructed as a mixture of $\{f(y_1 \cdots y_d; \boldsymbol{\lambda})\}$ over $\boldsymbol{\lambda}$. A special case of (2.16) obtains by assuming that the outcome of each univariate marginal probability mass corresponding to the outcome of Y_j, which is $P_j(y_j)$, depends on a parameter γ_j, $j = 1, \ldots, d$, (or a vector of parameters), and given γ_j, the variables Y_j are independent. If $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_p)'$ is the p-component vector formed by the non-singular components of γ_j, $j = 1, \ldots, d$, then the model (2.16) becomes

$$P(y_1 \cdots y_d) = \int \cdots \int \prod_{j=1}^{d} f_j(y_j; \gamma_j) c(G_1(\lambda_1), \ldots, G_p(\lambda_p)) \prod_{j=1}^{p} g_j(\lambda_j)\, d\lambda_1 \cdots d\lambda_p, \tag{2.17}$$

where $f_j(y_j; \gamma_j) = \Pr(Y_j = y_j | \Gamma_j = \gamma_j)$. The dependence among the response variables is induced through the mixing distribution of $\boldsymbol{\lambda}$. Usually $\lambda_j = \gamma_j$, $j = 1, \ldots, d$. A special case is $\gamma_j = \lambda_j = \lambda$ for all j.

2.2.3 Examples of MCD and MMD models

From their definitions, we see that the above two classes are rather general. We can choose any appropriate multivariate copula as the copula in the construction of the distribution. The sets of MCD and MMD models are not disjoint, as we can see from Example 2.13.

From practical viewpoint, we need to find some specific multivariate copulas C which offer good modelling properties and have a simple analytic form. One such choice is the multivariate normal copula (2.4). With this copula, we have $C(G_1(z_1), \ldots, G_d(z_d)) = \Phi_d(\Phi^{-1}(G_1(z_1)), \ldots, \Phi^{-1}(G_d(z_d)); \Theta)$,

where G_j's are arbitrary cdfs. The multivariate normal copula allows us to fully or almost fully exploit the dependence structure among the response variables. Its primary disadvantage may be computational difficulties when d is large (e.g. $d \geq 7$, see Schervish 1984).

This subsection consists of examples of MCD and MMD models. Discussion concerning the inclusion of covariates is given in some cases. More extensive studies of specific MCD and MMD models are given in Chapter 3.

Example 2.10 (MCD binary model)

1. *General models.* Let Y_j $(j = 1, \ldots, d)$ be a binary random variable taking values 0 or 1, and suppose the probability of outcome 1 is p_j. The cdf for Y_j is

$$G_j(y_j) = \begin{cases} 0, & y_j < 0, \\ 1 - p_j, & 0 \leq y_j < 1, \\ 1, & y_j \geq 1. \end{cases} \tag{2.18}$$

For a given d-dimensional copula $C(u_1, \ldots, u_d; \boldsymbol{\theta})$, $C(G_1(y_1), \ldots, G_d(y_d); \boldsymbol{\theta})$ is a well-defined distribution for the binary random vector $\mathbf{Y} = (Y_1, \ldots, Y_d)'$. When $d = 2$, with a one-parameter copula $C(u_1, u_2; \theta_{12})$, we can write down the pmf of $\mathbf{Y}$ as

$$P(y_1 y_2) = C(b_1, b_2; \theta_{12}) - C(b_1, a_2; \theta_{12}) - C(a_1, b_2; \theta_{12}) + C(a_1, a_2; \theta_{12}),$$

where $a_1 = G_1(y_1 - 1)$, $b_1 = G_1(y_1)$, $a_2 = G_2(y_2 - 1)$ and $b_2 = G_2(y_2)$. The pmf of $\mathbf{Y} = \mathbf{y}$ for a general d is expressed by (2.14).

One simple way to reparameterize p_j in (2.18), so that the new parameter associated to the univariate margin has the range in $(-\infty, \infty)$, is by letting $p_j = F_j(z_j)$, where F_j is a proper cdf. This is equivalent to writing $Y_j = I(Z_j < z_j)$, where Z_j is a rv with cdf F_j, and the random vector $\mathbf{Z} = (Z_1, \ldots, Z_d)'$ has a multivariate cdf $F_{12\cdots d}$. In the literature, this approach is referred to as a latent variable model or a *multivariate latent model*, since $\mathbf{Z}$ is an unobserved (latent) vector. There is also the option of including covariates to the parameter z_j, as well as to the dependence parameters $\boldsymbol{\theta}$ in the copula $C(u_1, \ldots, u_d; \boldsymbol{\theta})$. We will show these by examples.

2. *Multivariate probit model with no covariates.* The classical multivariate probit model for the multivariate binary response vector $\mathbf{Y}$ is (2.14) with the multinormal copula (2.4), where p_j is reparameterized as $p_j = \Phi(z_j)$ and G_j has form (2.18). This model has the CUOM and MUBE properties. Through its latent variable representation, the model can also be written as $Y_j = I(Z_j \leq z_j)$, $j = 1, \ldots, d$, where $\mathbf{Z} = (Z_1, \ldots, Z_d)' \sim N(\mathbf{0}, \Theta)$, $\Theta = (\theta_{jk})$; z_j is often referred to as the cut-off point. Θ is a correlation matrix, which (a) has elements bounded by 1 in absolute value

and (b) is nonnegative definite. To avoid the constraint of the bounds, we can reparameterize θ_{jk} through the hyperbolic tangent transform as

$$\theta_{jk} = \frac{\exp(\gamma_{jk}) - 1}{\exp(\gamma_{jk}) + 1}, \tag{2.19}$$

so that the new parameter γ_{jk} is in the range $(-\infty, \infty)$. The right hand side of (2.19) is an increasing function in γ_{jk}. Condition (a) is not sufficient to guarantee Θ be nonnegative definite except when $d = 2$. For $d = 2$, Θ is always nonnegative definite since the determinant of Θ, $1 - \theta_{12}^2$, is always nonnegative. For $d = 3$, Θ is nonnegative definite matrix provided

$$\det(\Theta) = 1 + 2\theta_{12}\theta_{13}\theta_{23} - \theta_{12}^2 - \theta_{13}^2 - \theta_{23}^2 \geq 0; \tag{2.20}$$

this constraint is satisfied for about 61.7% of the cube $[-1, 1]^3$ for $(\theta_{12}, \theta_{13}, \theta_{23})$. For $d = 4$, only about 18.3% of the hyper cube $[-1, 1]^6$ leads to a nonnegative definite matrix Θ; see Rousseeuw and Molenberghs (1994). Theoretically, the constraint (b) causes no trouble for the usefulness of the model. But numerically, this constraint may be a problem, since the space where the numerical computation can be carried out is quite limited. For the numerical computation to be successful, we have to guarantee that the current values are not out of the space of constraint, which, in some situations (e.g. the real parameters are close to the space boundaries), may render the computation time consuming or even not possible. In some situations, these problems with the constraint (b) can be avoided by limiting consideration to a simple correlation structure, so that the nonnegative definite condition is always satisfied. Examples include an exchangeable correlation matrix with all correlations equal to the same θ, and an AR(1) correlation matrix with the (j, k) component equal to $\theta^{|j-k|}$ for some θ.

3. *Multivariate probit model with covariates.* The classical multivariate probit model for a binary response vector $\mathbf{Y}_i$, $i = 1, \ldots, n$, with covariate vector $\mathbf{x}_{ij}$ for the jth univariate marginal parameter, if we use the latent variable representation, is that $Y_{ij} = I(Z_{ij} \leq \alpha_j + \boldsymbol{\beta}_j \mathbf{x}_{ij})$, $j = 1, \ldots, d$, $i = 1, \ldots, n$, where $\mathbf{Z}_i \sim N(\mathbf{0}, \Theta_i)$, $\Theta_i = (\theta_{i,jk})$. A modelling question may be whether dependence parameters should also be functions of covariates. If so, what are natural function to choose, so that Θ_i are all correlation matrices? If Θ_i does not depend on any covariates, then $\mathbf{Z}_i$ are iid $N(\mathbf{0}, \Theta)$, with $\Theta_i = \Theta = (\theta_{jk})$. If Θ_i depends on some covariate vectors, say $\theta_{i,jk}$ depends on $\mathbf{w}_{i,jk}$, then to satisfy $|\theta_{i,jk}| < 1$, we can let

$$\theta_{i,jk} = \frac{\exp(\gamma_{jk,0} + \boldsymbol{\gamma}_{jk}\mathbf{w}_{i,jk}) - 1}{\exp(\gamma_{jk,0} + \boldsymbol{\gamma}_{jk}\mathbf{w}_{i,jk}) + 1}. \tag{2.21}$$

Since all Θ_i, $i = 1, \ldots, n$, must be nonnegative definite, this may be a very strong restriction on the regression parameters $(\gamma_{jk,0}, \boldsymbol{\gamma}_{jk})$. In some situations, choices of the parameters $(\gamma_{jk,0}, \boldsymbol{\gamma}_{jk})$ in (2.21)

making all Θ_i nonnegative definite may not exist. The inclusion of covariates to the dependence parameters $\theta_{i,jk}$ as expressed in (2.21) is a mathematical construction. In the Example 2.13, we will give a more "natural" way to include covariates to dependence parameters. □

Example 2.11 (MCD count model)

1. *General models.* Consider a d-variate random count vector $\mathbf{Y} = (Y_1, \ldots, Y_d)'$. Let Y_j be a random variable taking the integer values $0, 1, 2, \ldots, \infty$, $j = 1, 2, \ldots, d$. Let $\Pr(Y_j = m) = p_j^{(m)}$. Then we have $\sum_{m=0}^{\infty} p_j^{(m)} = 1$ and the cdf of Y_j is

$$G_j(y_j) = \sum_{m=0}^{[y_j]} p_j^{(m)}, \tag{2.22}$$

where $[y_j]$ means the largest integer less or equal than y_j. Thus for a given d-dimensional copula $C(u_1, \ldots, u_d; \boldsymbol{\theta})$, $C(G_1(y_1), \ldots, G_d(y_d); \boldsymbol{\theta})$ is a well-defined distribution for the count random vector $\mathbf{Y}$. The pmf of $\mathbf{Y} = \mathbf{y}$ for a general d is expressed by (2.14). If we further assume that Y_j has a Poisson distribution with parameter λ_j, that is

$$p_j^{(m)} = \frac{\lambda_j^m \exp(-\lambda_j)}{m!}, \quad \lambda_j > 0, \tag{2.23}$$

then we will say we have a *MCD Poisson model.*

For the MCD Poisson model, the univariate parameter λ_j can be reparameterized by $\eta_j = \log(\lambda_j)$, so that the new parameter η_j has the range $(-\infty, \infty)$. Covariates can be included to η_j in an appropriate way. The comments on modelling of the dependence structure in the copula C for the MCD binary model are also relevant here.

To represent the MCD Poisson model by latent variables, let $Y_j = m$ if $z_{m-1} < Z_j \leq z_m$, $-\infty = z_{-1} \leq z_0 \leq \cdots \leq z_\infty = \infty$, where Z_j is a rv with cdf F_j, and the random vector $\mathbf{Z} = (Z_1, \ldots, Z_d)'$ has a multivariate cdf $F_{12\ldots d}$. The form of F_j does not have much importance since for count data, we are seldom interested in the cut-off points $z_0, z_1, \ldots, z_\infty$. But the copula related to $F_{12\ldots d}$ has essential importance for the modelling of count data, since it determines the multivariate structure of the random count vector $\mathbf{Y}$. Thus we may say that for count data, the MCD representation (2.14) is more relevant than the latent variable representation.

2. *Multivariate Poisson model with multinormal copula.* The multivariate Poisson model with multinormal copula for a count response vector $\mathbf{Y}$ is that in (2.14), where the copula is the multinormal copula (2.4) and $p_j^{(m)}$ has the form (2.23). This model has the CUOM and MUBE properties. The univariate marginal parameters λ_j can be transformed to $\eta_j = \log(\lambda_j)$ so that η_j has range $(-\infty, \infty)$. For a random vector $\mathbf{Y}_i$, $i = 1, \ldots, n$, if there is a covariate vector $\mathbf{x}_{ij}$ for λ_{ij}, a possible way to

include $\mathbf{x}_{ij}$ is by letting $\eta_{ij} = \alpha_j + \boldsymbol{\beta}_j \mathbf{x}_{ij}$, where $\eta_{ij} = \log(\lambda_{ij})$. Similarly, if $\Theta_i = (\theta_{i,jk})$ with $\theta_{i,jk}$ depending on a covariate vector $\mathbf{w}_{i,jk}$, a possible way to include $\mathbf{w}_{i,jk}$ is by letting $\theta_{i,jk}$ have the form (2.21). The difficulties with adding covariates to Θ_i remain, as in the previous example. □

Example 2.12 (MMD Poisson model)

1. *General models.* Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be a random vector of count data, where Y_j, $j = 1, \ldots, d$, has a Poisson distribution. The *MMD Poisson model* for the random vector $\mathbf{Y}$ is

$$P(y_1 \cdots y_d) = \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^d f_j(y_j; \lambda_j) c(G_1(\eta_1), \ldots, G_d(\eta_p)) \prod_{j=1}^p g_j(\eta_j)\, d\eta_1 \cdots d\eta_p, \tag{2.24}$$

where

$$f_j(y_j; \lambda_j) = \exp^{-\lambda_j} \lambda_j^{y_j} / y_j! \tag{2.25}$$

is the probability mass function of a Poisson distribution for Y_j given the parameter λ_j. In (2.24), $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)'$ is a $p \times 1$ vector of the collection of functions of $\lambda_1, \ldots, \lambda_d$; it is assumed to be random with a density function $c(G_1(\eta_1), \ldots, G_p(\eta_d)) \prod_{j=1}^p g_j(\eta_j)$, where $c(\cdot)$ is the density function of a copula C and $g_j(\cdot)$ the marginal density of η_j. The model can cover a wide range of dependence through appropriate parametric families of the copula C. Through conditional expectations one can study the covariances and correlations of $\mathbf{Y}$. If $\lambda_j = \eta_j$, $j = 1, \ldots, d$, we have

$$\begin{cases} \mathrm{E}(Y_j) = \mathrm{E}(\mathrm{E}(Y_j|\lambda_j)) = \mathrm{E}(\lambda_j), \\ \mathrm{Var}(Y_j) = \mathrm{E}(\mathrm{Var}(Y_j|\lambda_j)) + \mathrm{Var}(\mathrm{E}(Y_j|\lambda_j)) = \mathrm{Var}(\lambda_j) + \mathrm{E}(\lambda_j), \\ \mathrm{Cov}(Y_j, Y_k) = \mathrm{E}(\mathrm{Cov}(Y_j, Y_k|\lambda_j, \lambda_k)) + \mathrm{Cov}(\mathrm{E}(Y_j|\lambda_j), \mathrm{E}(Y_2|\lambda_k)) = \mathrm{Cov}(\lambda_j, \lambda_k). \end{cases} \tag{2.26}$$

Therefore the correlation of Y_j and Y_k is

$$\mathrm{Corr}(Y_j, Y_k) = \frac{\mathrm{Cov}(\lambda_j, \lambda_k)}{\{[\mathrm{Var}(\lambda_j) + \mathrm{E}(\lambda_j)][\mathrm{Var}(\lambda_k) + \mathrm{E}(\lambda_k)]\}^{1/2}}, \tag{2.27}$$

which has the same sign as the correlation of λ_j and λ_k. $\mathrm{Corr}(Y_j, Y_k)$ is smaller than $\mathrm{Corr}(\lambda_j, \lambda_k)$ and tends to $\mathrm{Corr}(\lambda_j, \lambda_k)$ when $\mathrm{E}(\lambda_j)/\mathrm{Var}(\lambda_j)$ and $\mathrm{E}(\lambda_k)/\mathrm{Var}(\lambda_k)$ tend to zero. When $\lambda_j = \eta$, $j = 1, \ldots, d$, $\mathbf{Y}$ is equicorrelated with $\mathrm{Corr}(Y_j, Y_k) = \mathrm{Var}(\eta)/[\mathrm{Var}(\eta) + \mathrm{E}(\eta)]$. The range of dependence for this special situation is quite restricted. For the general model (2.24), the parameters are introduced by the marginal distribution of η_j and the copula C. Letting the parameters depend on covariates is possible, as we can see from the next example with a specific copula.

2. *Multivariate Poisson-lognormal model.* The Multivariate Poisson-lognormal model for a random Poisson vector $\mathbf{Y}$ is that in (2.24), where the copula is the multinormal copula (2.4), and η_j has a

lognormal distribution with parameters μ_j and σ_j. The pmf for $\mathbf{Y} = \mathbf{y}$ is

$$P(y_1 \cdots y_d) = \int_0^\infty \cdots \int_0^\infty \prod_{j=1}^{d} f_j(y_j; \lambda_j) g(\boldsymbol{\eta}; \boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)\, d\eta_1 \cdots d\eta_p, \tag{2.28}$$

where $f_j(y_j; \lambda_j)$ is of the form (2.25), and

$$g_d(\boldsymbol{\eta}; \boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta) = \frac{1}{(2\pi)^{d/2}(\eta_1 \cdots \eta_p)|\boldsymbol{\sigma}'\Theta\boldsymbol{\sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\log \boldsymbol{\eta} - \boldsymbol{\mu})'(\boldsymbol{\sigma}'\Theta\boldsymbol{\sigma})^{-1}(\log \boldsymbol{\eta} - \boldsymbol{\mu})\right\}, \tag{2.29}$$

with $\eta_j > 0$, $j = 1, \ldots, p$, is a multivariate lognormal density function. The model (2.28) has the CUOM and MUBE properties. The parameters in the model are $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$, $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)'$ and Θ. By (2.26) and (2.27), we have

$$\begin{aligned}
&\mathrm{E}(Y_j) = \exp\{\mu_j + \frac{1}{2}\sigma_j^2\} \stackrel{\text{def}}{=} a_j, \\
&\mathrm{Var}(Y_j) = a_j + a_j^2[\exp(\sigma_j^2) - 1], \\
&\mathrm{Cov}(Y_j, Y_k) = a_j a_k[\exp(\theta_{jk}\sigma_j\sigma_k) - 1], \quad j \neq k, \\
&\mathrm{Corr}(Y_j, Y_k) = \frac{\exp(\theta_{jk}\sigma_j\sigma_k) - 1}{\{[\exp(\sigma_j^2) - 1 + a_j^{-1}][\exp(\sigma_k^2) - 1 + a_k^{-1}]\}^{1/2}}.
\end{aligned} \tag{2.30}$$

The margins are overdispersed Poisson since $\mathrm{Var}(Y_j)/\mathrm{E}(Y_j) > 1$. $|\mathrm{Corr}(Y_j, Y_k)|$ is less than $|\mathrm{Corr}(\eta_j, \eta_k)|$ and $\mathrm{Corr}(Y_j, Y_k)$ approaches $\mathrm{Corr}(\eta_j, \eta_k)$ when $a_j, a_k \to \infty$. A covariate vector $\mathbf{x}$ can be included in the model, say by letting the components of $\boldsymbol{\mu}$ be linear functions of $\mathbf{x}$. $\boldsymbol{\sigma}$ can be assumed to have some special pattern, for example $\sigma_1 = \cdots = \sigma_p = \sigma$. It is harder to naturally let the correlation matrix Θ depend on covariates, as already discussed for the multivariate probit model for binary data. □

Example 2.13 (MMD model for binary data)

1. *General models.* Let $\mathbf{Y} = (Y_1, \ldots, Y_d)'$ be a binary random vector. Assume that $\mathbf{Y}$ has the MCD binary model in Example 2.10 for a given cut-off point vector $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)'$. $\boldsymbol{\alpha}$ in turn is assumed to be a random vector. Let $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)$ be the collection of functions of $\boldsymbol{\alpha}$. With the latent variable representation, we have that for given $\boldsymbol{\eta}$

$$\mathbf{Y} = (Y_1, \ldots, Y_d)' = (I(A_1 \leq \alpha_1), \ldots, I(A_d \leq \alpha_d))', \tag{2.31}$$

where $\mathbf{A} = (A_1, \ldots, A_d)'$ has a multivariate cdf F, and $\boldsymbol{\eta}$ has a multivariate cdf G. Thus

$$P(y_1 \cdots y_d) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} P(y_1 \cdots y_d|\boldsymbol{\eta}) c(G_1(\eta_1), \ldots, G_p(\eta_p)) \prod_{j=1}^{p} g_j(\eta_j)\, d\eta_1 \cdots d\eta_p,$$

where $c(G_1(\eta_1), \ldots, G_p(\eta_p)) \prod_{j=1}^{p} g_j(\eta_j)$ is the density function of $\boldsymbol{\eta}$, with $c(\cdot)$ the density function of a copula C and $g_j(\cdot)$ the marginal density of η_j. A more general case is when there is a covariate vector $\mathbf{x}$. In this situation, we may let $\alpha_j = \beta_{j,0} + \boldsymbol{\beta}_j \mathbf{x}$, $j = 1, \ldots, d$, where the $\beta_{j,0}$'s and $\boldsymbol{\beta}_j$'s are random, and $\boldsymbol{\eta}$ is now assumed to be the collection of functions of the random components $\beta_{j,0}$'s and $\boldsymbol{\beta}_j$'s.

2. *Multivariate probit-normal model.* The MMD probit model is obtained by assuming that in (2.31), $\mathbf{A} = (A_1, \ldots, A_d)' \sim N_d(\mathbf{0}, \Theta)$ and $\boldsymbol{\eta} \sim N_p(\boldsymbol{\mu}, \Sigma)$, where $\Theta = (\theta_{jk})$ is a correlation matrix and $\Sigma = (\sigma_{jk})$ is a variance-covariance matrix. Without loss of generality, let us assume $\boldsymbol{\eta} = \boldsymbol{\alpha}$. Then the MMD probit model of the form (2.31) becomes

$$\mathbf{Y} = (Y_1, \ldots, Y_d)' = (I(Z_1 \leq z_1^*), \ldots, I(Z_d \leq z_d^*))', \tag{2.32}$$

where $Z_j = (A_j - \alpha_j - \mu_j)/\sqrt{1 + \sigma_{jj}}$, $z_j^* = \mu_j/\sqrt{1 + \sigma_{jj}}$, $j = 1, \ldots, d$, and $\mathbf{Z} = (Z_1, \ldots, Z_d)' \sim N_d(\mathbf{0}, R)$, where $R = (r_{jk})$ is a correlation matrix with $r_{jk} = (\theta_{jk} + \sigma_{jk})/\{(1 + \sigma_{jj})(1 + \sigma_{kk})\}^{1/2}$, $j \neq k$. This is a special class of multivariate probit model in Example 2.10. When $\sigma_{jj} = 0$, it is the multivariate probit model discussed in Example 2.10. This example demonstrates that the intersection of the sets of MCD and MMD models is not empty. It is straightforward to extend such a construction to the more general situation with a covariate vector $\mathbf{x}$, such that $\alpha_j = \beta_{j,0} + \boldsymbol{\beta}_j \mathbf{x}$ with the $\beta_{j,0}$'s and $\boldsymbol{\beta}_j$'s random. With one covariate x_j, for example, one might take $\alpha_j = \beta_{j,0} + \beta_j x_j$ with $\boldsymbol{\beta}_0 = (\beta_{1,0}, \ldots, \beta_{d,0})' \sim N_d(\boldsymbol{\mu}_0, \Sigma_0)$ independent of $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)' \sim N_d(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu}_0 = (\mu_{1,0}, \ldots, \mu_{d,0})'$, $\Sigma_0 = (\sigma_{jk,0})$, $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ and $\Sigma = (\sigma_{jk})$. Now in (2.32), we have $Z_j = (A_j - \beta_{j,0} - \beta_j x_j - \mu_{j,0} - \mu_j x_j)/\sqrt{1 + \sigma_{jj,0} + \sigma_{jj} x_j^2}$, with $\mathbf{Z} = (Z_1, \ldots, Z_d)' \sim N_d(\mathbf{0}, R)$, $R = (r_{jk})$, such that

$$\begin{cases} z_j^* = \dfrac{\mu_{j,0} + \mu_j x_j}{\{1 + \sigma_{jj,0} + \sigma_{jj} x_j^2\}^{1/2}}, & j = 1, \ldots, d, \\ r_{jk} = \dfrac{\theta_{jk} + \sigma_{jk,0} + \sigma_{jk} x_j x_k}{\{(1 + \sigma_{jj,0} + \sigma_{jj} x_j^2)(1 + \sigma_{kk,0} + \sigma_{kk} x_k^2)\}^{1/2}}, & j \neq k. \end{cases} \tag{2.33}$$

The function of r_{jk} in (2.33) can be considered as a "natural" form for the correlation parameters as functions of the covariates, since this function representation is derived directly from the linear regression for marginal parameters. As long as the conditions for linear regression for marginal parameter hold, r_{jk} will always satisfy the constraints for forming a correlation matrix. For R to be nonnegative definite, it suffice that Θ, Σ_0 and Σ be nonnegative definite. These three matrices do not depend on covariates, which is very attractive numerically compared with the nonnegative definite requirement on Θ_i in (2.21). A special case is $\theta_{jk} = 0$ and $\sigma_{jk,0} = 0$ $(j \neq k)$, in which

case the only constraint is that Σ be nonnegative definite. Finally, we notice that in contrast to the conventional univariate probit analysis, the regression function in (2.33) for the cut-off points are not linear functions of covariates. Nevertheless, (2.33) can be used in lieu of the multivariate probit model with covariates in Example 2.10, since the parameters in (2.33) are also interpretable. To use the model (2.33), it is necessary to reparameterize the parameters $\sigma_{jj,0}, \sigma_{kk,0}, \sigma_{jk,0}, \sigma_{jj}, \sigma_{kk}, \sigma_{jk}$ and θ_{jk} such that the new parameters have $(-\infty, \infty)$ as their domain. □

2.2.4 Some properties of MCD and MMD models

We summarize some of the properties of MCD and MMD models:

1. MCD and MMD models, constructed through stochastic or latent variable representation, provide a clear probabilistic description of multivariate discrete random phenomenon. In some situations, the pmf and cdf have closed forms; in other situations, the pmf or cdf can be numerically computed in a reasonably short time. Likelihood inference can be used, with the help of the theory in section 2.3 and section 2.4.

2. MCD and MMD models allow flexible choices of multivariate copulas (Multinormal copula, Hüsler-Reiss copula, Morgenstern copula, etc.) as well as flexible choices of all the univariate marginal distributions (any discrete distributions: Bernoulli, binomial, negative binomial, Poisson and generalized Poisson, etc.), and they allow relevant covariates to be included in the appropriate parameters in the models. In this way, these two classes of models are able to capture the nature of discrete data in an individual or grouped observation basis, thus they allow the drawing of appropriate inferences from the data.

3. With appropriate copulas, many MCD and MMD models have the CUOM and MUBE properties. The CUOM property, sometimes referred to as "reproducibility" or "upward compatibility" in the literature, is also sought for modelling longitudinal and repeated measures. With appropriate families of parametric copulas, a wide range of dependence, including negative dependence, is possible.

4. With appropriate copulas, the parameters related to the univariate margins structure and the parameters related to dependence structure can be allowed to vary independently in separate parameter spaces. This is a good property that the multivariate Gaussian model also enjoys.

5. By choosing appropriate marginal distributions, the MCD and MMD models can naturally account for a variety of situations occuring with discrete data, such as over-dispersion which is independent of covariates, skewed distributions, multimodality, etc.

6. For a given d-variate discrete distribution F, there may be many copulas which match F into MCD model class; MCD models are robust in terms of data modelling with copulas of similar structure.

   Some of the points above will be made clear in Chapter 3 as well as in Chapter 5.

2.3 Inference functions of margins

For a general multivariate model, parameter estimation is often a difficult computational issue. Without readily available parameter estimation methods, any model, even though interpretable, will not have practical usefulness. For situations involving univariate models, many methods have been devised for parameter estimation, ranging from the method of moments through formal maximum likelihood to informal graphical techniques. The maximum likelihood approach is used in general because it has a number of desirable statistical properties. For example, under general regularity conditions, ML estimators are consistent, and asymptotically normal. With some weak additional assumptions, the MLE is also asymptotically efficient. However, the method has not been successfully applied for estimating the parameters of multivariate models, except for the multivariate normal and a few cases with low dimension (*e.g.* $d = 2$). A primary cause of this unsatisfactory situation is the computational difficulty involved with multivariate models, even with modern powerful computers. The ML approach for parameter estimation in multivariate situations is still not routine. The question is: can we have a general effective estimation procedure to estimate parameters for a model in the MCD and MMD classes?

In this section, we first discuss model fitting strategies for multivariate models in subsection 2.3.1. One strategy leads to the inference functions of margins approach, that we propose as the parameter estimation approach for MCD and MMD models with the MUBE, PUBE or MPME properties. In subsection 2.3.2, we introduce some important results in inference function theory for multiple parameters needed for developing the inference basis for MCD and MMD models. In subsection 2.3.3, we introduce the inference functions of margins (IFM) approach and give some examples.

2.3.1 Approaches for fitting multivariate models

There are at least three possible likelihood-based approaches to estimate parameters in a multivariate model:

Approach 1. All univariate and multivariate parameters are estimated simultaneously by maximizing the full-dimensional likelihood function. This is the MLE approach.

Approach 2. For a model where all multivariate parameters are in a copula, univariate parameters are estimated from the separate univariate likelihoods. The multivariate parameters are then estimated from multivariate likelihoods with the univariate parameters fixed as estimated from separate univariate likelihoods.

Approach 3. For a model with the MUBE, PUBE or MPME property, univariate parameters are estimated from separate univariate likelihoods. Bivariate, trivariate and multivariate parameters are then estimated from bivariate, trivariate and multivariate likelihoods, with lower order parameters fixed as estimated from lower order likelihoods.

The first approach is general and direct. While this strategy sounds most natural from the likelihood point of view, it could be computationally very difficult for most of the multivariate models, even in relatively low dimensional situations. The multivariate normal distribution, which can be easily handled by this approach, is an exception. The second approach makes the computational task easier, but it still has the difficulties of dealing with a multivariate object in general. These difficulties are mainly two: the high-dimensional maximization problem and the multivariate probability calculation. The third approach reduces these difficulties by working with lower dimensional maximizations or lower dimensional probability calculations. This is a valuable approach if the parametric family of interest has the MUBE, PUBE or MPME properties. It is important because it makes statistical inference for multivariate data easier. Computational tractability is an important factor for the popularity of certain statistical tools, as we observe in many areas of statistics. The third approach to stochastic modeling is often convenient, since many tractable models are readily available for the marginal distributions. It is also invaluable as a general strategy for data analysis in that it allows one to investigate the dependence structure independently of marginals effects (through copula) and computationally only dealing with lower dimensional (often two-dimensional) models.

Example 2.14 Consider the multivariate probit model for a d-dimensional binary vector $\mathbf{Y}$ with pmf

$$P(y_1 \cdots y_d) = \sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} \Phi_d(\Phi^{-1}(a_{1i_1}), \ldots, \Phi^{-1}(a_{di_d}); \Theta), \tag{2.34}$$

where $\Theta = (\theta_{jk})$, $a_{j1} = G_j(y_j - 1)$ and $a_{j2} = G_j(y_j)$, with $G_j(1) = 1$ and $G_j(0) = 1 - \Phi(z_j)$. This model has the CUOM and MUBE properties. For estimation from a random sample of iid $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$, the three approaches for fitting multivariate models could be used here:

Approach 1. Estimate the parameters $\mathbf{z} = (z_1, \ldots, z_d)'$ and Θ by maximizing the multivariate likelihood $L = \prod_{i=1}^{n} P(y_{i1} \cdots y_{id})$. Let the resulting estimates be $\hat{\mathbf{z}}$ and $\hat{\Theta}$.

Approach 2. (a) Obtain the estimates $\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_d)'$ by maximizing separately d univariate marginal likelihoods. (b) Estimate the parameters Θ from the multivariate likelihood $L = \prod_{i=1}^{n} P(y_{i1} \cdots y_{id})$ with the parameters $\mathbf{z}$ fixed at the estimated values $\tilde{\mathbf{z}}$ from (a).

Approach 3. (a) Obtain the estimates $\tilde{\mathbf{z}} = (\tilde{z}_1, \ldots, \tilde{z}_d)'$ by maximizing separately d univariate marginal likelihoods. (b) Estimate the parameters θ_{jk}, $1 \leq j < k \leq d$, by maximizing separately $d(d-1)/2$ bivariate likelihoods $L_{jk} = \prod_{i=1}^{n} P_{jk}(y_{ij} y_{ik})$ with the parameters z_j, z_k fixed at the estimated values $\tilde{z}_j, \tilde{z}_k$ from (a). Let the resulting estimate be $\tilde{\Theta}$.

Approach 1 is computationally demanding, since it requires the calculation of high-dimensional multinormal probabilities and a numerical optimization on many parameters. Approach 2 reduces the numerical optimization problem to fewer parameters, but the high-dimensional multinormal probability calculation is still required. Approach 3 reduces the numerical optimization in Approach 2 into several numerical optimizations, each involving fewer parameters. Further, the high-dimensional multinormal probability calculation is no longer required; all that is needed are the binormal probability calculations, which are readily feasible with modern computers. Multi-dimensional calculation are needed for predicted or expected frequencies, but this is much less effort compared with multi-dimensional numerical integrations within a numerical optimization procedure.

Since it is computationally easier to obtain $\tilde{\mathbf{z}}$ and $\tilde{\Theta}$, a natural question is what is the asymptotic efficiency of $\tilde{\mathbf{z}}$ and $\tilde{\Theta}$ compared with $\hat{\mathbf{z}}$ and $\hat{\Theta}$. In Chapter 4, we will deal with this problem in a general context. □

2.3.2 Inference functions for multiple parameters

Introduction

The common approach to the problem of estimation is to propose an estimator $T(\mathbf{x})$ and then study its properties. For estimators with specific properties such as unbiasedness, minimum variance or minimum mean squared error, or asymptotic normality, theories for ordering these estimators are developed. Standard methods for obtaining the estimator $T(\mathbf{x})$ include least squares (LS), maximum likelihood (ML), best linear unbiased, method of moments, uniform minimum variance (UMV), and so on. However, many point estimation procedures may be viewed as the solution of an (or some) appropriate estimating equation(s). Indeed, any estimator may be regarded as a solution to an equation or a set of equations of the form $\Psi(\mathbf{x}, \boldsymbol{\theta}) = 0$, where Ψ is a vector of functions (or a single function in the one-parameter case) of the data $\mathbf{x}$ and the parameter $\boldsymbol{\theta}$. $\Psi(\mathbf{x}, \boldsymbol{\theta})$ is commonly called a vector of *inference functions* or *estimating functions*. In this thesis, we use mainly the term "inference functions". But when we focus more on the use of the inference functions for estimation, we also employ the term "estimating functions".

The theory of inference functions is studied in, for example, Godambe (1960, 1976, 1991), McLeish and Small (1988) and Jϕrgensen and Labouriau (1995). The theory of inference functions imposes optimality criteria on the function Ψ rather than the estimators obtained from it. The approach of considering a class of inference functions and finding the optimal inference function has the advantage of retaining the strengths of the estimation method (e.g LS, ML, UMV) and at the same time eliminates some of their weaknesses. For example, in point estimation, the Cramér-Rao lower bound is attained only in rare occasions whereas the optimality of the score function among inference functions holds merely under regularity conditions (see below). Inference functions may be used either as estimating equations to determine a point estimate or as the basis for constructing tests or confidence intervals for the parameters. An example is the maximum likelihood estimators, which are obtained as the solutions of estimating equations from the score functions. Thus the inference functions for MLE are the score functions. Other examples of the application of inference functions are the theory of M-estimators for obtaining robust estimators and the quasi-likelihood methods used in generalized linear models. Inference functions have also found application in a wide variety of applied fields; examples in biostatistics, stochastic processes, and survey sampling can be found in Godambe (1991).

In the following, we introduce the notion of regular inference functions and study the asymptotic properties of resulting estimates in the iid situation.

Inference functions for a vector parameter

In the following, we will give a series of definitions for the inference functions for a vector of parameters and a general asymptotic result for the parameter estimates from the defined inference functions.

Let us consider a parametric family $\mathcal{F}$ defined on a common measurable space $(\mathcal{Y}, \mathcal{A})$, where $\mathcal{A}$ is the σ-field associated with $\mathcal{Y}$. We further assume

$$\mathcal{F} = \{P(\mathbf{y};\boldsymbol{\theta}) : \boldsymbol{\theta} \in \Re\},\ \Re \subseteq I\!R^q, \tag{2.35}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is q-component vector, and $\Re$ the parameter space. The parameter space is usually a subset of q-dimensional Euclidean space. We presume the existence of a measure μ on $\mathcal{Y}$ such that for each fixed value of the parameter $\boldsymbol{\theta}$ the function $P(\mathbf{y};\boldsymbol{\theta})$ is the density with respect to μ of a probability measure $\mathcal{P}$ on $\mathcal{Y}$.

Definition 2.9 (Inference functions) *A $I\!R^q$-valued vector of functions*

$$\Psi(\mathbf{y};\boldsymbol{\theta}) = (\psi_1(\mathbf{y};\boldsymbol{\theta}), \ldots, \psi_q(\mathbf{y};\boldsymbol{\theta}))^T :\ \mathcal{Y} \times \Re \to I\!R^q$$

is called a vector of inference functions, if the component functions of $\Psi(\mathbf{y};\boldsymbol{\theta})$ are measurable for each fixed $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q) \in \Re$. □

Definition 2.10 (Unbiased inference functions) *Ψ is said to be unbiased if for each $\boldsymbol{\theta} \in \Re$ and $j = 1, \ldots, q$, $E_{\boldsymbol{\theta}}\{\psi_j\} = 0$, where $E_{\boldsymbol{\theta}}$ means expectation relative to $P(\cdot;\boldsymbol{\theta})$.* □

Unbiasedness is a natural requirement which ensures that the roots of the equations are close to the true values when little random variation is present. Whereas $\boldsymbol{\theta}$ may not have an unbiased estimator, unbiased inference functions exist under fairly general circumstances. For any given inference function vector Ψ and any $\mathbf{y} \in \mathcal{Y}$, an estimator of $\boldsymbol{\theta}$, say $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y})$, can be obtained as the solution to $\Psi = \mathbf{0}$.

In order for the estimate $\tilde{\boldsymbol{\theta}}$ to be well-defined and well-behaved, the inference function vector Ψ must satisfy some regularity conditions, that is, Ψ must consist of regular inference functions.

Definition 2.11 (Regular inference functions) *The vector of inference functions Ψ is said to be a vector of regular inference functions if, for all $\boldsymbol{\theta} \in \Re$, the following assumptions are satisfied:*

1. *The support of $\mathbf{y}$ does not depend on any $\boldsymbol{\theta} \in \Re$.*
2. *$E\{\psi_j\} = 0$, $j = 1, \ldots, q$.*

3. *The partial derivative* $\partial\psi_j/\partial\theta_k$ *exists for almost every* $\mathbf{y} \in \mathcal{Y}$, $j, k = 1, \ldots, q$.

4. *The order of integration and differentiation may be interchanged as follows:*

$$\frac{\partial}{\partial\theta_k}\int_{\mathcal{Y}} \psi_j P(\mathbf{y};\boldsymbol{\theta}) d\mu(\mathbf{y}) = \int_{\mathcal{Y}} \frac{\partial}{\partial\theta_k}\left[\psi_j P(\mathbf{y};\boldsymbol{\theta})\right] d\mu(\mathbf{y}),$$

   $j, k = 1, \ldots, q$.

5. $E\{\psi_j\psi_k\}$ *exists,* $j, k = 1, \ldots, q$, *and the* $q \times q$ *matrix*

$$M_\Psi(\boldsymbol{\theta}) = E\{\Psi\Psi^T\}$$

   is positive-definite.

6. *The* $q \times q$ *matrix*

$$D_\Psi(\boldsymbol{\theta}) = E\left\{\frac{\partial\Psi}{\partial\boldsymbol{\theta}'}\right\}$$

   is non-singular.

□

A model $P(\mathbf{y};\boldsymbol{\theta})$ in (2.35) is said to be regular, if the score functions are regular inference functions and $\Re$ is an open region of $\mathbb{R}^q$. We are only interested in regular models, such that the asymptotic theory concerning MLEs is readily available for use. This is not a strong assumption for applications. (The main limitation may be the exclusion of models in 1 of Definition 2.11.)

Definition 2.12 (Fisher information matrix) *The Fisher information matrix is the matrix-valued function* $I : \Re \to \mathbb{R}^{q\times q}$ *defined by*

$$I(\boldsymbol{\theta}) = E\{U(\boldsymbol{\theta})U^T(\boldsymbol{\theta})\},$$

where $U(\boldsymbol{\theta})$ *is the vector of score functions,* $U(\boldsymbol{\theta}) \stackrel{\text{def}}{=} \partial/\partial\boldsymbol{\theta} \log P(\mathbf{y};\boldsymbol{\theta})$. □

Definition 2.13 (Godambe information matrix) *For a regular inference function vector* Ψ, *the Godambe information matrix is the matrix-valued function* $J_\Psi : \Re \to \mathbb{R}^{q\times q}$ *defined by*

$$J_\Psi(\boldsymbol{\theta}) = D_\Psi^T(\boldsymbol{\theta})M_\Psi^{-1}(\boldsymbol{\theta})D_\Psi(\boldsymbol{\theta}),$$

where $M_\Psi(\boldsymbol{\theta}) = E\{\Psi\Psi^T\}$ *and* $D_\Psi(\boldsymbol{\theta}) = E\{\partial\Psi/\partial\boldsymbol{\theta}'\}$. □

Consider n iid observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$ from a model $P(\mathbf{y}; \boldsymbol{\theta})$ in (2.35). Let $\Psi(\mathbf{y}_i; \boldsymbol{\theta}) = (\psi_{i1}, \ldots, \psi_{iq})'$. The inference function vector based on the n observations is $\Psi_n : \mathcal{Y}^n \times \Re \to I\!R^q$ given by

$$\Psi_n = \sum_{i=1}^{n} \Psi(\mathbf{y}_i; \boldsymbol{\theta}).$$

We define the estimator $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ as the solution of $\Psi_n = \mathbf{0}$.

The following theorem establishes the asymptotic normality of the solution $\tilde{\boldsymbol{\theta}}$ based on regular inference functions and gives an asymptotic interpretation of the Godambe information matrix.

Theorem 2.1 *Assume that the estimator $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ associated with the regular inference function vector $\Psi_n : \mathcal{Y}^n \times \Re \to I\!R^q$ is a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}$, that is, $\sqrt{n}(\tilde{\theta}_j - \theta_j)$, $j = 1, \ldots, q$, is bounded in probability so that $\tilde{\theta}_j$ tends to θ_j at least at the rate of $1/\sqrt{n}$. We further assume that there exist functions $M_{jkl}(\mathbf{y})$ such that $|\partial^2 \psi_j / \partial\theta_k \partial\theta_l| \leq M_{jkl}(\mathbf{y})$ for all $\boldsymbol{\theta} \in \Re$, where $E\{M_{jkl}(\mathbf{y})\} < \infty$ for all j, k, l. Then as $n \to \infty$, we have asymptotically*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_{\Psi}^{-1}(\boldsymbol{\theta}))$$

under $P(\cdot; \boldsymbol{\theta})$.

Proof: The proof is similar to the corresponding theorem for the asymptotic normality of the MLE. We therefore only sketch it.

Ψ_n has the following expansion around $\boldsymbol{\theta}$

$$\mathbf{0} = \Psi_n(\tilde{\boldsymbol{\theta}}) = \Psi_n(\boldsymbol{\theta}) + H_n(\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathrm{R}_n,$$

where H_n is a $q \times q$ matrix $\partial\Psi_n / \partial\boldsymbol{\theta}$ and $\mathrm{R_n} = O_p(||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2) = o_p(n^{-1})$ by assumptions.

Thus

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \left[\frac{1}{n} H_n(\boldsymbol{\theta})\right]^{-1} \frac{1}{\sqrt{n}} [-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n]. \tag{2.36}$$

By the Law of Large Numbers

$$\frac{1}{n} H_n(\boldsymbol{\theta}) \xrightarrow{p} D_{\Psi}(\boldsymbol{\theta}).$$

Now for any fixed vector $\mathbf{u} = (u_1, \ldots, u_q)'$, consider the sequence of one-dimensional rv's

$$\begin{aligned} \mathbf{u}' \frac{\Psi_n}{\sqrt{n}} &= u_1 \sum_{i=1}^{n} \frac{\psi_{i1}(\mathbf{y}_i; \boldsymbol{\theta})}{\sqrt{n}} + \cdots + u_q \sum_{i=1}^{n} \frac{\psi_{iq}(\mathbf{y}_i; \boldsymbol{\theta})}{\sqrt{n}} \\ &= \sum_{i=1}^{n} \frac{u_1 \psi_{i1}(\mathbf{y}_i; \boldsymbol{\theta}) + \cdots + u_q \psi_{iq}(\mathbf{y}_i; \boldsymbol{\theta})}{\sqrt{n}}. \end{aligned}$$

By the central limit theorem (Lindberg-Lévy), $\mathbf{u}'\Psi_n/\sqrt{n}$ is $N_1(0, \mathbf{u}'M_\Psi\mathbf{u})$. This result leads to

$$\frac{1}{\sqrt{n}}\Psi_n \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, M_\Psi).$$

Applying Slutsky's Theorem to (2.36), we obtain

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, D_\Psi^{-1} M_\Psi (D_\Psi^{-1})^T)$$

or

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_\Psi^{-1}(\boldsymbol{\theta})).$$

□

Optimality criteria for inference functions

In this subsection, we will summarize optimality results for inference functions in the multi-parameter situation. These results will be referred to later for comparing two sets of regular inference functions.

Consider a scalar inference function Ψ. It is natural to seek an unbiased estimating function Ψ for which the variance $\mathrm{E}\{\Psi^2\}$ is as small as possible. This is analogous to the theory of minimum variance unbiased (MVU) estimation. Since the variance may be changed by multiplying Ψ with an arbitrary constant, some further standardization is necessary for the purpose of comparing variances. Godambe (1960) suggested considering the variance of the standardized estimating function $\Psi_s = \Psi/\mathrm{E}\{\partial\Psi/\partial\theta\}$, and defined an optimal estimating function to be one which minimizes $\mathrm{Var}(\Psi_s) = \mathrm{E}\{\Psi^2\}/\{\mathrm{E}(\partial\Psi/\partial\theta)\}^2$, or maximizes $\mathrm{Var}^{-1}(\Psi_s)$, the *Godambe information* for Ψ. Godambe showed that in the one-parameter case the usual maximum likelihood estimating equation has this optimal property within a wide class of regular unbiased inference functions. Thus Godambe information can be used to compare two regular inference functions, and the function with the larger Godambe information is generally preferred.

Given two vectors of inference functions, Ψ and Ω, several different optimality criteria can be used to say that Ω is preferred (or optimal) to Ψ.

Definition 2.14 (M-optimality) *A vector of inference functions Ω is said to have matrix optimality or M-optimality versus a vector of inference functions Ψ if the difference of the inverses of the Godambe information matrices*

$$J_\Psi^{-1}(\boldsymbol{\theta}) - J_\Omega^{-1}(\boldsymbol{\theta})$$

is non-negative definite. □

Definition 2.15 (T-optimality) *A vector of inference functions Ω is said to have trace optimality or T-optimality versus a vector of inference functions Ψ if the difference of the trace of the inverse of Godambe information matrices*

$$Tr(J_{\Psi}^{-1}(\boldsymbol{\theta})) - Tr(J_{\Omega}^{-1}(\boldsymbol{\theta}))$$

is positive. □

Definition 2.16 (D-optimality) *A vector of inference functions Ω is said to have determinant optimality or D-optimality versus a vector of inference functions Ψ if the difference of determinant of the inverse of Godambe information matrices*

$$|J_{\Psi}^{-1}(\boldsymbol{\theta})| - |J_{\Omega}^{-1}(\boldsymbol{\theta})|$$

is positive.

Chandrasekar and Kale (1984) proved that M-optimality implies T-optimality and D-optimality. Joseph and Durairajan (1991) further proved that the above three criteria are equivalent in the sense that if Ψ is optimal with respect to any one of the three criteria then it is also optimal with respect to the remaining two.

When comparing two sets of regular inference functions, we could examine a slightly different version of T-optimality and D-optimality. For example, for T-optimality, we may examine

$$\frac{Tr(J_{\Psi}^{-1}(\boldsymbol{\theta}))}{Tr(J_{\Omega}^{-1}(\boldsymbol{\theta}))},$$

and for the D-optimality

$$\frac{\sqrt[q]{|J_{\Psi}^{-1}(\boldsymbol{\theta})|}}{\sqrt[q]{|J_{\Omega}^{-1}(\boldsymbol{\theta})|}}.$$

In practice and often in simulation studies, only the estimated values of $J_{\Psi}^{-1}(\boldsymbol{\theta})$ and $J_{\Omega}^{-1}(\boldsymbol{\theta})$ are available, M-optimality or T-optimality or D-optimality may be violated slightly numerically based on only one set of observations.

We end this subsection by stating an extended Cramér-Rao inequality for inference functions:

Theorem 2.2 *For any given vector of regular inference functions Ψ, and for all $\boldsymbol{\theta} \in \Re$, $J_{\Psi}^{-1}(\boldsymbol{\theta}) - I^{-1}(\boldsymbol{\theta})$ is non-negative definite.*

For a proof of this result, see Jϕrgensen and Labouriau (1995). Related references include Ferreira (1982) and Chandrasekar (1988), among others. □

This theorem states that, for a regular model $P(\mathbf{y};\boldsymbol{\theta})$, the vector score functions

$$U(\boldsymbol{\theta}) = \frac{\partial \log P(\mathbf{y};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \log P(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \log P(\mathbf{y};\boldsymbol{\theta})}{\partial \theta_k} \right)$$

are M-optimal within the class of all regular unbiased estimating functions.

2.3.3 Inference function of margins

We have seen from previous subsection that, under fairly general regularity conditions, the score functions are asymptotically optimal among regular inference functions. However, with multivariate models, except in a few special cases (e.g multivariate normal), the estimating equations based on the score functions are computationally very cumbersome or intractable. It would be an invaluable alternative to have inference functions which are computationally feasible in general and also efficient compared to the score functions.

In the ensuing subsection, we introduce a set of inference functions, we call *the inference functions of margins (IFM)*. In Chapter 4, we show that IFM shares the asymptotic optimality properties of the score functions, and this is particularly true for the multivariate models with MUBE and PUBE properties. One major advantage of IFM is that it is computationally feasible in general and more flexible for handlinge different types of data. This leads us to develop a new inference theory and computationally feasible procedures for many MCD and MMD models.

Inference function of scores

We consider the family (2.35) and assume it is a regular parametric family. The likelihood function of $\boldsymbol{\theta}$, given $\mathbf{y}$, is $L(\boldsymbol{\theta};\mathbf{y}) = P(\mathbf{y};\boldsymbol{\theta})$, the corresponding loglikelihood function is $\ell(\boldsymbol{\theta};\mathbf{y}) = \log P(\mathbf{y};\boldsymbol{\theta})$. Let

$$L_n(\boldsymbol{\theta}) = \prod_{i=1}^{n} L(\boldsymbol{\theta};\mathbf{y}_i)$$

denote the likelihood of $\boldsymbol{\theta}$ based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$, a sample from $\mathcal{Y}$. The loglikelihood function of $\boldsymbol{\theta}$ based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$ is

$$\ell_n(\boldsymbol{\theta}) = \log L_n(\boldsymbol{\theta}) = \sum_{i=1}^{n} \ell(\boldsymbol{\theta};\mathbf{y}_i).$$

Definition 2.17 (Inference functions of scores, or IFS) *The vector of score functions*

$$\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \left(\frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial \ell_n(\boldsymbol{\theta})}{\partial \theta_q} \right)$$

is called inference function vector of scores, or IFS. □

The maximum likelihood estimate (MLE) is generally determined as the solution to the likelihood equations $\partial \ell_n(\boldsymbol{\theta})/\partial \boldsymbol{\theta} = \mathbf{0}$. The Hessian matrix of the function $-\ell_n(\boldsymbol{\theta})/n$ is $J(\boldsymbol{\theta})$, where $(J(\boldsymbol{\theta}))_{jk} = -(1/n)(\partial^2 \ell_n(\boldsymbol{\theta})/\partial\theta_j \partial\theta_k)$. The expected value of $J(\boldsymbol{\theta})$, $I(\boldsymbol{\theta}) = \mathrm{E}\{J(\boldsymbol{\theta})\}$, is the Fisher information matrix. The value $J(\hat{\boldsymbol{\theta}})$ of $J(\cdot)$ at the maximum likelihood estimate $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}(\mathbf{y}_1, \ldots, \mathbf{y}_n)$ is referred to as the observed information. $J(\hat{\boldsymbol{\theta}})$ will generally be positive definite since $\hat{\boldsymbol{\theta}}$ is the point of maximum likelihood. A consistent estimate of $I(\boldsymbol{\theta})$ is $\hat{I}(\boldsymbol{\theta}) = J(\hat{\boldsymbol{\theta}})$.

Under very general regularity conditions, it is known that the MLEs are asymptotically normal, in the sense that as $n \to \infty$,

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, I(\boldsymbol{\theta})^{-1}).$$

See Sen and Singer (1993, p.209) for a proof.

Inference function of margins

We now introduce the loglikelihood function of a margin, the inference function of a margin for one parameter, and then define the inference functions of margins (IFM) for a parameter vector $\boldsymbol{\theta}$. The asymptotic results for the estimates from IFM will be established in the next section.

Consider the parametric family (2.35) and assume $P(\mathbf{y}; \boldsymbol{\theta})$ is a d-dimensional density function with respect to a probability measure μ on $\mathcal{Y}$. Let S_d denote the set of non-empty subsets of $\{1, \ldots, d\}$. For any $S \in S_d$, we use $|S|$ to denote the *cardinality* of S. Let $P_S(\mathbf{y}_S)$ be the S-margin of $P(\mathbf{y}; \boldsymbol{\theta})$, where $\mathbf{y}_S = \{y_j : j \in S\}$. Assume $P_S(\mathbf{y}_S)$ depends on $\boldsymbol{\theta}_S$, where $\boldsymbol{\theta}_S$ is a subvector of $\boldsymbol{\theta}$.

Definition 2.18 *Let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$. Suppose the parameter θ_k appears in S-margin $P_S(\mathbf{y}_S)$. The loglikelihood function of the S-margin is*

$$\ell_S(\boldsymbol{\theta}_S) = \log P_S(\mathbf{y}_S).$$

An inference function for θ_k is

$$\frac{\partial \ell_S(\boldsymbol{\theta}_S)}{\partial \theta_k}.$$

□

The inference function of a margin for a parameter θ is not necessarily uniquely defined by the above definition. In this thesis, unless specified otherwise, we always work with the inference function from a margin with the smallest $|S|$. For a specific model, it is often evident when $|S|$ is the smallest for a parameter, so we will not be concerned with the proof of this feature in most applied situations. If there are two or more inference functions for θ with the same smallest $|S|$, than there

is a question of how to combine these inference functions to optimally extract information. We will discuss this issue in section 2.6.

Note that with the assumption of MPME (or MUBE), one can use S with $|S| < q$ ($|S| \leq 2$ for MUBE) for every parameter θ_k. In the case where MPME does not hold, then one has $S = \{1, \ldots, d\}$ for some θ_k in the model. For the new theory below, we assume MPME or MUBE or PUBE in the remainder of this chapter.

Assume for the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$, the corresponding smallest cardinality subsets associated with the parameters are $S_1, \ldots, S_q$ (as q is usually greater than d, there are duplicates among the S_k's).

Definition 2.19 (Inference functions of margins, or IFM) *The vector of inference functions*

$$\Psi_{\text{IFM}} = \left(\frac{\partial \ell_{n,S_1}(\boldsymbol{\theta}_{S_1})}{\partial \theta_1}, \ldots, \frac{\partial \ell_{n,S_q}(\boldsymbol{\theta}_{S_q})}{\partial \theta_q} \right)'$$

is called the inference functions of margins, or IFM, for $\boldsymbol{\theta}$. □

For a regular model, the inference functions derived from the likelihood functions of margins also satisfy the regularity conditions. Thus asymptotic properties related to the regular inference functions should apply to IFM. Detailed development of this aspect will be given in section 2.4.

Definition 2.20 (Inference functions of margins estimates, or IFME) *Any* $\tilde{\boldsymbol{\theta}} \in \Re$ *which is the solution of*

$$\Psi_{\text{IFM}} = \left(\frac{\partial \ell_{n,S_1}(\boldsymbol{\theta}_{S_1})}{\partial \theta_1}, \ldots, \frac{\partial \ell_{n,S_q}(\boldsymbol{\theta}_{S_q})}{\partial \theta_q} \right)' = \mathbf{0}$$

is called the inference functions of margins estimate, or IFME, of the unknown true parameter vector $\boldsymbol{\theta}$. □

In a few cases, $\tilde{\boldsymbol{\theta}}$ has an analytic expression (e.g. Example 4.3). In general, $\tilde{\boldsymbol{\theta}}$ has to be obtained by means of numerical methods.

Examples of inference functions of margins

Example 2.15 Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be n iid rv's from the distribution $C(G_1(x_1), \ldots, G_d(x_d); \Theta)$ with $G_j(x_j) = \Phi(x_j; \mu_j, \sigma_j^2)$, where C is the multinormal copula (2.4). Let $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)$ and $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_d)$. The loglikelihood function is

$$\ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta) = \sum_{i=1}^{n} \log \left(c(\Phi(x_{i1}), \ldots, \Phi(x_{id}); \Theta) \prod_{j=1}^{d} \phi(x_{ij}; \mu_j, \sigma_j) \right).$$

Thus the IFS is

$$\Psi_{\text{IFS}} = \left(\frac{\partial \ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)}{\partial \mu_1}, \ldots, \frac{\partial \ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)}{\partial \mu_d}, \frac{\partial \ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)}{\partial \sigma_1}, \ldots, \right.$$
$$\left. \frac{\partial \ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)}{\partial \sigma_d}, \frac{\partial \ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)}{\partial \theta_{12}}, \ldots, \frac{\partial \ell_n(\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta)}{\partial \theta_{d-1,d}} \right).$$

The loglikelihood functions of 1 and 2-dimensional margins for the parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}, \Theta$ are

$$\begin{cases} \ell_{nj}(\mu_j, \sigma_j) = \sum_{i=1}^{n} \log \phi(x_{ij}; \mu_j, \sigma_j), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_{jk}, \mu_j, \mu_k, \sigma_j, \sigma_k) = \sum_{i=1}^{n} \log \left(c(\Phi(x_{ij}), \Phi(x_{ik}); \theta_{jk}) \phi(x_{ij}; \mu_j, \sigma_j) \phi(x_{ik}; \mu_k, \sigma_k) \right), \ 1 \le j < k \le d. \end{cases}$$

Thus the IFM is

$$\Psi_{\text{IFM}} = \left(\frac{\partial \ell_{n1}(\mu_1, \sigma_1)}{\partial \mu_1}, \ldots, \frac{\partial \ell_{nd}(\mu_d, \sigma_d)}{\partial \mu_d}, \frac{\partial \ell_{n1}(\mu_1, \sigma_d)}{\partial \sigma_1}, \ldots, \frac{\partial \ell_{nd}(\mu_d, \sigma_d)}{\partial \sigma_d}, \right.$$
$$\left. \frac{\partial \ell_{n12}(\theta_{12}, \mu_1, \mu_2, \sigma_1, \sigma_2)}{\partial \theta_{12}}, \ldots, \frac{\partial \ell_{nd-1,d}(\theta_{d-1,d}, \mu_{d-1}, \mu_d, \sigma_{d-1}, \sigma_d)}{\partial \theta_{d-1,d}} \right).$$

It is known that Ψ_{IFS} and Ψ_{IFM} lead to the same estimates; see for example Seber (1984). □

Example 2.16 Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be n iid rv from the multivariate Poisson model with multinormal copula in Example 2.11. The loglikelihood functions of margins for the parameters $\boldsymbol{\lambda}$ and Θ are

$$\begin{cases} \ell_{nj}(\lambda_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_{jk}, \lambda_j, \lambda_k) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d, \end{cases}$$

where $P_j(y_{ij}) = \lambda_j^{y_{ij}} \exp(-\lambda_j)/y_{ij}!$ and $P_{jk}(y_{ij}y_{ik}) = \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) - \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk}) - \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) + \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk})$, where $a_{ij} = G_{ij}(y_{ij} - 1)$, $b_{ij} = G_{ij}(y_{ij})$, $a_{ik} = G_{ik}(y_{ik} - 1)$ and $b_{ik} = G_{ik}(y_{ik})$, with $G_{ij}(y_{ij}) = \sum_{x=0}^{y_{ij}} P_j(x)$ and $G_{ik}(y_{ik}) = \sum_{x=0}^{y_{ik}} P_k(x)$.

Let $\eta_j = \log(\lambda_j)$. The IFM for η_j, $j = 1, \ldots, d$, and θ_{jk}, $1 \le j < k \le d$ are

$$\Psi_{\text{IFM}} = \left(\sum_{i=1}^{n} \frac{1}{P_1(y_{i1})} \frac{\partial P_1(y_{i1})}{\partial \eta_1}, \ldots, \sum_{i=1}^{n} \frac{1}{P_d(y_{id})} \frac{\partial P_d(y_{id})}{\partial \eta_d}, \right.$$
$$\left. \sum_{i=1}^{n} \frac{1}{P_{12}(y_{i1}y_{i2})} \frac{\partial P_{12}(y_{i1}y_{i2})}{\partial \theta_{12}}, \ldots, \sum_{i=1}^{n} \frac{1}{P_{d-1,d}(y_{id-1}y_{id})} \frac{\partial P_{d-1,d}(y_{id-1}y_{id})}{\partial \theta_{d-1,d}} \right).$$

For a similar random vector $\mathbf{Y}_i$, $i = 1, \ldots, n$ with a covariate vector $\mathbf{x}_{ij}$ for λ_{ij}, a possible way to include $\mathbf{x}_{ij}$ is by letting $\eta_{ij} = \alpha_j + \boldsymbol{\beta}_j \mathbf{x}_{ij}$, where $\eta_{ij} = \log(\lambda_{ij})$. We can similarly write down the IFM for parameters α_j, $\boldsymbol{\beta}_j$ and θ_{jk}. □

Example 2.17 (Multivariate binary Molenberghs-Lesaffre model) We first define the multivariate binary Molenberghs-Lesaffre model (Molenberghs and Lesaffre, 1994), or M-L model. Let $\mathbf{Y} = (Y_1, \ldots, Y_d)$ be a d-variate binary random vector taking values 0 or 1 for each component. A model for $\mathbf{Y}$ is defined in the following way. Consider a set of $2^d - 1$ generalized cross-ratios with values in $(0, \infty)$: η_j, $1 \le j \le d$, η_{jk}, $1 \le j < k \le d$, $\ldots$, and $\eta_{12\cdots d}$ such that:

$$\eta_{j_1 \cdots j_q} = \frac{\prod_{(y_{j_1}, \ldots, y_{j_q}) \in A_q^+} P_{j_1 \cdots j_q}(y_{j_1} \cdots y_{j_q})}{\prod_{(y_{j_1}, \ldots, y_{j_q}) \in A_q^-} P_{j_1 \cdots j_q}(y_{j_1} \cdots y_{j_q})}, \tag{2.37}$$

where $A_q^+ = \{(y_{j_1}, \ldots, y_{j_q}) \in \{1,0\}^q \mid (q - \sum_{t=1}^q y_{j_t}) \equiv 0 (\text{mod } 2)\}$ and $A_q^- = \{1,0\}^q \backslash A_q^+$, and $\{j_1, \ldots, j_q\}$ is a subset of $\{1, 2, \ldots, d\}$ with cardinality q. We can verify for example when $q = 1, 2, 3, 4$ that

$$\begin{aligned}
\eta_j &= \frac{P(1)}{P(0)}, \quad 1 \le j \le d, \\
\eta_{jk} &= \frac{P(11)P(00)}{P(10)P(01)}, \quad 1 \le j < k \le d, \\
\eta_{jkl} &= \frac{P(111)P(100)P(010)P(001)}{P(110)P(101)P(011)P(000)}, \quad 1 \le j < k < l \le d, \\
\eta_{jklm} &= \frac{P(1111)P(1100)P(1010)P(1001)P(0110)P(0101)P(0011)P(0000)}{P(1110)P(1101)P(1011)P(1000)P(0111)P(0100)P(0010)P(0001)}, \quad 1 \le j < k < l < m \le d,
\end{aligned} \tag{2.38}$$

where subscripts on P are suppressed to simplify the notation.

Molenberghs and Lesaffre (1994) show that the $2^d - 1$ equations in (2.37) together with $\sum P(y_1 \cdots y_d) = 1$ leads to unique nonnegative solutions for $P(y_1 \cdots y_d)$, $(y_1, \ldots, y_d) \in \{1,0\}^d$, under some compatibility conditions on the $d-1$ and lower-dimensional probabilities. If all these conditions in the Molenberghs-Lesaffre construction are satisfied, we have a well-defined multivariate Bernoulli model. We call this model multivariate M-L binary model. The multivariate M-L binary model is not MUBE, but the parameters η_j and η_{jk} are PUBE. The special case where $\eta_S = 1$ for $|S| \ge 3$ is MUBE.

Related to the MCD model, it is not clear if there exists a MCD model such that (2.37) is true and under what conditions a MCD model is equivalent to (2.37). The difficulty is to prove there exists a copula, such that (2.37) is a equivalent expression to a MCD model. The existence of a copula is needed in order to properly define this model with covariates (e.g. logistic regression univariate margins). For a discussion of whether the Molenberghs-Lesaffre construction leads to a copula model, see Joe (1996). Nevertheless, (2.37) in terms of $P_{j_1 \cdots j_q}(y_{j_1} \cdots y_{j_q})$ certainly defines a multivariate model for binary data for some ranges of the parameters.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be n iid binary rv's from-a proper multivariate M-L binary model. Assume the parameters of interest are $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_d, \eta_{12}, \ldots, \eta_{d-1,d})'$ and let η_S be arbitrary for $|S| \geq 3$. The loglikelihood functions of margins for $\boldsymbol{\eta}$ are

$$\begin{cases} \ell_{nj}(\eta_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\eta_{jk}, \eta_j, \eta_k) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \leq j < k \leq d. \end{cases}$$

Thus the IFM is

$$\Psi_{\text{IFM}} = \left(\frac{\partial \ell_{n1}(\eta_1)}{\partial \eta_1}, \ldots, \frac{\partial \ell_{nd}(\eta_d)}{\partial \eta_d}, \frac{\partial \ell_{n12}(\eta_{12}, \eta_1, \eta_2)}{\partial \eta_{12}}, \ldots, \frac{\partial \ell_{nd-1,d}(\eta_{d-1,d}, \eta_{d-1}, \eta_d)}{\partial \eta_{d-1,d}} \right).$$

For an interpretation of the parameters (2.37), see Joe (1996). □

Some advantages of IFM approach

The IFM approach has many advantages for parameter estimation and statistical inference:

1. The IFM approach for parameter estimation is computationally simpler than estimating all the parameters from the IFS approach. A numerical optimization with many parameters is much more time-consuming (sometimes beyond the capacity of current computers) compared with several numerical optimizations, each with fewer parameters. In some cases, optimization is done with parameters from lower-dimensional margins already estimated (that is, there is some order to the sequence of numerical optimizations). IFM leads to estimates of the parameters of many multivariate nonnormal models efficiently and quickly.

2. A potential problem with the IFS approach is the lack of stability of the solution when there are outliers or perturbations of the data in one or few dimensions. With the IFM approach, we suggest that only the contaminated margins will have such nonrobustness problems. In other words, IFM has some robustness properties in multivariate analysis. It would be interesting to study theoretically and numerically how outliers perturb the IFS and IFM estimates.

3. A large sample size is often needed for a large dimension of the responses. This may not be easily satisfied in most applied problems. Rather, sparse data are commonplace when there are multiple responses; these often create problems for ML estimation. By working with the lower dimensional likelihoods, the IFM approach avoids the sparseness problem in multivariate situations to a certain degree; this could be a major advantage in small sample situations.

4. The IFM approach should be robust against some misspecification in the multivariate model. Also some assessment of the goodness-of-fit of the copula can be made after solving part of the estimation equations from IFM, corresponding to parameters of univariate margins.

5. Finally, IFM leads to separate modelling of the relationship of the response with marginal covariates, and the association among the response variables in some situations. This feature can be exploited to shorten the modelling cycle when some quick answer on the marginal behaviour of the covariates is the scientific focus.

In the above, we listed some advantages of IFM approach. In the next section, we study the asymptotic properties of IFM approach. The remaining question of efficiency of IFM will be studied in Chapter 4.

2.4 Parameter estimation with IFM and asymptotic results

In this section we will be concerned with the asymptotic properties of the parameter estimates from the IFM approach. We will develop in detail the parameter estimation procedure with the IFM approach for a MCD or MMD model with MUBE or with some parameters of the models having PUBE properties. The situations we consider include models with covariates. Sufficient conditions for the consistency and asymptotic normality of IFME are given. Some theory concerning the asymptotic variance matrix (Godambe information matrix) for the estimates from the IFM approach is also developed. Detailed direct calculations of the Godambe information matrix for the estimates based on the data and fitted models are given. An alternative computational approach, namely the jackknife method, for the estimation of the Godambe information matrix is given in section 2.5. This technique has considerable importance because of its practical usefulness (See Chapter 5). Later in section 2.6, we will propose computational algorithms, which are based on IFM, for the parameter estimation where common parameters appear in different inference functions of margins.

2.4.1 Models with no covariates

In this subsection, we confine our discussion to the case of samples of n independent observations from the same distributions. The case of samples of n independent observations from different distributions will be studied in the next subsection. We consider a regular MCD or MMD model in (2.12)

$$P(y_1 \cdots y_d; \boldsymbol{\theta}), \quad \boldsymbol{\theta} \in \Re, \tag{2.39}$$

where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d, \theta_{12}, \ldots, \theta_{d-1,d})'$. The model (2.39) is assumed to have MUBE or to have some of its parameters having PUBE properties. In general, we assume that θ_j $(j = 1, \ldots, d)$ is a parameter vector for the jth univariate margin of (2.39) such that $P_j(y_j) = P_j(y_j; \theta_j)$, and θ_{jk} $(1 \le j < k \le d)$ is a parameter vector for the (j,k) bivariate margin of (2.39) such that $P_{jk}(y_j y_k) = P_{jk}(y_j, y_k; \theta_j, \theta_k, \theta_{jk})$. The situation for models with higher order (> 2) parameters are similar; the extension of the results here should be straightforward. For the purpose of illustration, and without loss of generality, we assume in the following that θ_j and θ_{jk} are scalar parameters.

Let $\mathbf{Y}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be iid rv with model (2.39). The loglikelihood functions of margins of $\boldsymbol{\theta}$ are

$$\begin{cases} \ell_{nj}(\theta_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_j, \theta_k, \theta_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d. \end{cases} \tag{2.40}$$

These expressions can also be rewritten as

$$\begin{cases} \ell_{nj}(\theta_j) = \sum_{\{y_j\}} n_j(y_j) \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_j, \theta_k, \theta_{jk}) = \sum_{\{y_j y_k\}} n_{jk}(y_j y_k) \log P_{jk}(y_j y_k), \quad 1 \le j < k \le d, \end{cases} \tag{2.41}$$

based on the summary data $n_j(y_j)$ and $n_{jk}(y_j y_k)$. In the following we continue to use the expression (2.40) for technical development, for consistency with the case where covariates are present. But (2.41) is a more economic form for computation, and should be used for model fitting whenever possible.

Let

$$\begin{cases} \psi_j = \psi_j(\theta_j) \stackrel{\text{def}}{=} \dfrac{1}{P_j(y_j)} \dfrac{\partial P_j(y_j)}{\partial \theta_j}, \quad j = 1, \ldots, d, \\ \psi_{jk} = \psi_{jk}(\theta_j, \theta_k, \theta_{jk}) \stackrel{\text{def}}{=} \dfrac{1}{P_{jk}(y_j y_k)} \dfrac{\partial P_{jk}(y_j y_k)}{\partial \theta_{jk}}, \quad 1 \le j < k \le d, \end{cases}$$

and

$$\begin{cases} \psi_{i;j} = \psi_{i;j}(\theta_j) \stackrel{\text{def}}{=} \dfrac{1}{P_j(y_{ij})} \dfrac{\partial P_j(y_{ij})}{\partial \theta_j}, \quad j = 1, \ldots, d, \\ \psi_{i;jk} = \psi_{i;jk}(\theta_j, \theta_k, \theta_{jk}) \stackrel{\text{def}}{=} \dfrac{1}{P_{jk}(y_{ij} y_{ik})} \dfrac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \theta_{jk}}, \quad 1 \le j < k \le d, \end{cases}$$

for $i = 1, \ldots, n$. Let $\boldsymbol{\Psi} = \boldsymbol{\Psi}(\boldsymbol{\theta}) = (\psi_1, \ldots, \psi_d, \psi_{12}, \ldots, \psi_{d-1,d})'$, and $\boldsymbol{\Psi}_n = \boldsymbol{\Psi}_n(\boldsymbol{\theta}) = (\Psi_{n1}, \ldots, \Psi_{nd}, \Psi_{n12}, \ldots, \Psi_{nd-1,d})'$, where $\Psi_{nj} = \sum_{i=1}^{n} \psi_{i;j}$ $(j = 1, \ldots, d)$ and $\Psi_{njk} = \sum_{i=1}^{n} \psi_{i;jk}$ $(1 \le j < k \le d)$.

From (2.40), we derive the IFM for $\boldsymbol{\theta}$

$$\begin{cases} \Psi_{nj} = \sum_{i=1}^{n} \psi_{i;j}, \quad j = 1, \ldots, d, \\ \Psi_{njk} = \sum_{i=1}^{n} \psi_{i;jk}, \quad 1 \leq j < k \leq d. \end{cases} \tag{2.42}$$

Since (2.39) is a regular model, the regularity conditions of Definition 2.11 are also true for the inference functions (2.42). With the IFM approach, an estimate of $\boldsymbol{\theta}$ that we denote by $\tilde{\boldsymbol{\theta}} = \tilde{\boldsymbol{\theta}}(\mathbf{y}_1, \ldots, \mathbf{y}_n) = (\tilde{\theta}_1, \ldots, \tilde{\theta}_d, \tilde{\theta}_{12}, \ldots, \tilde{\theta}_{d-1,d})'$ is obtained by solving the following system of nonlinear equations

$$\begin{cases} \Psi_{nj} = 0, \quad j = 1, \ldots, d, \\ \Psi_{njk} = 0, \quad 1 \leq j < k \leq d. \end{cases}$$

Properties of estimators

We start by examining the simple case (particularly, for notation) when $d = 2$ to illustrate how the consistency and asymptotic normality of $\tilde{\boldsymbol{\theta}}$ can be established. The model (2.39) is now

$$P(y_1, y_2; \theta_1, \theta_2, \theta_{12}) \tag{2.43}$$

with $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{12})' \in \Re$. Without loss of generality, $\theta_1, \theta_2, \theta_{12}$ are assumed to be scalar parameters. Let the random vector $\mathbf{Y} = (Y_1, Y_2)'$ and $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$ $(i = 1, \ldots, n)$ be iid with (2.43), and $\mathbf{y}$, $\mathbf{y}_i$ be the observed values of $\mathbf{Y}$ and $\mathbf{Y}_i$ respectively. Thus the IFM for $\theta_1, \theta_2, \theta_{12}$ are

$$\begin{cases} \Psi_{nj} = \sum_{i=1}^{n} \psi_{i;j}, \quad j = 1, 2, \\ \Psi_{n12} = \sum_{i=1}^{n} \psi_{i;12}. \end{cases} \tag{2.44}$$

We denote the true value of $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_{12})'$ by $\boldsymbol{\theta}_0 = (\theta_{1,0}, \theta_{2,0}, \theta_{12,0})'$. Using Taylor's theorem on (2.44) at $\boldsymbol{\theta}_0$ to the first order, we have

$$\begin{cases} 0 = \Psi_{n1}(\tilde{\theta}_1) = \Psi_{n1}(\theta_{1,0}) + (\tilde{\theta}_1 - \theta_{1,0}) \left.\dfrac{\partial \Psi_{n1}}{\partial \theta_1}\right|_{\theta_1^*}, \\ 0 = \Psi_{n2}(\tilde{\theta}_2) = \Psi_{n2}(\theta_{2,0}) + (\tilde{\theta}_2 - \theta_{2,0}) \left.\dfrac{\partial \Psi_{n2}}{\partial \theta_2}\right|_{\theta_2^*}, \\ 0 = \Psi_{n12}(\tilde{\theta}_{12}, \tilde{\theta}_1, \tilde{\theta}_2) = \Psi_{n12}(\theta_{12,0}) + (\tilde{\theta}_{12} - \theta_{12,0}) \left.\dfrac{\partial \Psi_{n12}}{\partial \theta_{12}}\right|_{\boldsymbol{\theta}^{**}} \\ \qquad\qquad + (\tilde{\theta}_1 - \theta_{1,0}) \left.\dfrac{\partial \Psi_{n12}}{\partial \theta_1}\right|_{\boldsymbol{\theta}^{**}} + (\tilde{\theta}_2 - \theta_{2,0}) \left.\dfrac{\partial \Psi_{n12}}{\partial \theta_2}\right|_{\boldsymbol{\theta}^{**}}, \end{cases} \tag{2.45}$$

where θ_1^* is some value between $\tilde{\theta}_1$ and $\theta_{1,0}$, θ_2^* is some value between $\tilde{\theta}_2$ and $\theta_{2,0}$, and $\boldsymbol{\theta}^{**}$ is some vector value between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Note that Ψ_{n12} also depends on θ_1 and θ_2.

Let

$$H_n = H_n(\boldsymbol{\theta}) = \begin{pmatrix} \frac{\partial \Psi_{n1}}{\partial \theta_1} & 0 & 0 \\ 0 & \frac{\partial \Psi_{n2}}{\partial \theta_2} & 0 \\ \frac{\partial \Psi_{n12}}{\partial \theta_1} & \frac{\partial \Psi_{n12}}{\partial \theta_2} & \frac{\partial \Psi_{n12}}{\partial \theta_{12}} \end{pmatrix}$$

and $D_\Psi = D_\Psi(\boldsymbol{\theta}) = \mathrm{E}\{n^{-1}H_n\}$. Since (2.43) is assumed to be a regular model, we have that $\mathrm{E}(\Psi_n) = \mathbf{0}$ and D_Ψ non-singular.

On the right-hand side of (2.45), Ψ_{n1}, Ψ_{n2}, Ψ_{n12}, $\partial\Psi_{n1}/\partial\theta_1$, $\partial\Psi_{n2}/\partial\theta_2$, $\partial\Psi_{n12}/\partial\theta_{12}$, $\partial\Psi_{n12}/\partial\theta_1$ and $\partial\Psi_{n12}/\partial\theta_2$ are all sums of independent identical variates, and as $n \to \infty$ each therefore converges to its expectation by the strong law of large numbers. The expectations of $\Psi_{n1}(\theta_{1,0})$, $\Psi_{n2}(\theta_{2,0})$, and $\Psi_{n12}(\theta_{12,0})$ are zero and the expectations of $\partial\Psi_{n1}/\partial\theta_1$, $\partial\Psi_{n2}/\partial\theta_2$, $\partial\Psi_{n12}/\partial\theta_{12}$, are non-zero by regularity assumptions. Since all terms on the right-hand side must converge to zero to remain equal to the left-hand sides, we see that we must have $(\tilde{\theta}_1 - \theta_{1,0})$, $(\tilde{\theta}_2 - \theta_{2,0})$ and $(\tilde{\theta}_{12} - \theta_{12,0})$ converging to zero as $n \to \infty$, so that $\tilde{\theta}_1$, $\tilde{\theta}_2$ and $\tilde{\theta}_{12}$ are consistent estimators under our assumptions (for a more rigorous proof along these lines, see Cramér 1946, page 500-504).

Now let

$$H_n^* = \begin{pmatrix} \left.\frac{\partial \Psi_{n1}}{\partial \theta_1}\right|_{\theta_1^*} & 0 & 0 \\ 0 & \left.\frac{\partial \Psi_{n2}}{\partial \theta_2}\right|_{\theta_2^*} & 0 \\ \left.\frac{\partial \Psi_{n12}}{\partial \theta_1}\right|_{\boldsymbol{\theta}^{**}} & \left.\frac{\partial \Psi_{n12}}{\partial \theta_2}\right|_{\boldsymbol{\theta}^{**}} & \left.\frac{\partial \Psi_{n12}}{\partial \theta_{12}}\right|_{\boldsymbol{\theta}^{**}} \end{pmatrix}.$$

It follows from the convergence in probability of $\tilde{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_0$ that

$$\frac{1}{n}\left[H_n(\tilde{\boldsymbol{\theta}}) - H_n(\boldsymbol{\theta}_0)\right] \xrightarrow{\mathrm{p}} \mathbf{0}.$$

Since each element of $n^{-1}H_n(\boldsymbol{\theta})$ is the mean of n independent identically-distributed variates, by the strong law of large numbers, it converges with probability 1 to its expectation. This implies that $n^{-1}H_n(\boldsymbol{\theta}_0)$ converges with probability 1 to $D_\Psi(\boldsymbol{\theta}_0)$. Thus

$$\frac{1}{n}H_n(\tilde{\boldsymbol{\theta}}) \xrightarrow{\mathrm{p}} D_\Psi(\boldsymbol{\theta}_0).$$

Now we rewrite (2.45) in the following form

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) = \left[\frac{1}{n}H_n^*\right]^{-1} \frac{1}{\sqrt{n}}[-\Psi_n(\boldsymbol{\theta}_0)]. \tag{2.46}$$

Since θ_1^* lies between $\tilde{\theta}_1$ and $\theta_{1,0}$, θ_2^* lies between $\tilde{\theta}_2$ and $\theta_{2,0}$, and $\boldsymbol{\theta}^{**}$ lies between $\tilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$, thus we also have

$$\frac{1}{n}H_n^* \xrightarrow{\mathrm{p}} D_\Psi(\boldsymbol{\theta}_0).$$

Along the same lines as the proof in Theorem 2.1, we see that

$$\frac{1}{\sqrt{n}}\Psi_n(\boldsymbol{\theta}_0)\xrightarrow{\mathrm{D}}N_3(\mathbf{0}, M_\Psi(\boldsymbol{\theta}_0)),$$

where $M_\Psi(\boldsymbol{\theta}_0) = \mathrm{E}(\Psi\Psi')$. Applying Slutsky's Theorem to (2.46), we obtain

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\xrightarrow{\mathrm{D}}N_3(\mathbf{0}, D_\Psi^{-1}(\boldsymbol{\theta}_0)M_\Psi(\boldsymbol{\theta}_0)(D_\Psi^{-1}(\boldsymbol{\theta}_0))^T),$$

or equivalently

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\xrightarrow{\mathrm{D}}N_3(\mathbf{0}, J_\Psi^{-1}(\boldsymbol{\theta}_0)).$$

Thus we can extend to the following theorem for the IFME of model (2.43):

Theorem 2.3 *Consider the model (2.43) and let the dimension of θ_j $(j = 1, 2)$ be p_j and that of θ_{12} be p_{12}. Let $\tilde{\boldsymbol{\theta}}$ denote the IFME of $\boldsymbol{\theta}$ under the IFM corresponding to (2.44). Then $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. Furthermore, as $n \to \infty$,*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta})\xrightarrow{\mathrm{D}}N_{p_1+p_2+p_{12}}(\mathbf{0}, J_\Psi^{-1}),$$

where $J_\Psi = J_\Psi(\boldsymbol{\theta}) = D_\Psi^T M_\Psi^{-1} D_\Psi$, with $M_\Psi = E\{\Psi\Psi'\}$ and $D_\Psi = E\{\partial\Psi/\partial\boldsymbol{\theta}'\}$. □

Inverting the Godambe information matrix J_Ψ yields the asymptotic variances and covariances of $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \tilde{\theta}_2, \tilde{\theta}_{12})$. We provide the calculations for the situation where $\theta_1, \theta_2, \theta_{12}$ are scalars. The asymptotic variance of $\tilde{\theta}_j$, $j = 1, 2$, is $n^{-1}[\mathrm{E}\psi_j^2][\mathrm{E}\partial\psi_j/\partial\theta_j]^{-2}$ and the asymptotic covariance of $\tilde{\theta}_1, \tilde{\theta}_2$ is $n^{-1}[\mathrm{E}\psi_1\psi_2][\mathrm{E}\partial\psi_1/\partial\theta_1]^{-1}[\mathrm{E}\partial\psi_2/\partial\theta_2]^{-1}$. The asymptotic variance of $\tilde{\theta}_{12}$ is

$$n^{-1}\left[\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_{12}}\right]^{-2}\left\{\mathrm{E}\psi_{12}^2 + \sum_{j=1}^{2}\left[\mathrm{E}\frac{\partial\psi_j}{\partial\theta_j}\right]^{-2}\mathrm{E}\psi_j^2\left[\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_j}\right]^2\right.$$
$$\left.-2\sum_{j=1}^{2}\left[\mathrm{E}\frac{\partial\psi_j}{\partial\theta_j}\right]^{-1}\left[\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_j}\right][\mathrm{E}\psi_{12}\psi_j] + 2\prod_{j=1}^{2}\left[\mathrm{E}\frac{\partial\psi_j}{\partial\theta_j}\right]^{-1}\left[\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_j}\right][\mathrm{E}\psi_1\psi_2]\right\},$$

and the asymptotic covariance of $\tilde{\theta}_{12}$, $\tilde{\theta}_j$ is

$$n^{-1}\left[\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_{12}}\right]^{-1}\left[\mathrm{E}\frac{\partial\psi_j}{\partial\theta_j}\right]^{-1}\left\{\mathrm{E}\psi_{12}\psi_j - \left[\mathrm{E}\frac{\partial\psi_j}{\partial\theta_j}\right]^{-1}\left[\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_j}\right]\mathrm{E}\psi_j^2\right.$$
$$\left.-\prod_{k=1}^{2}\left(\left[\mathrm{E}\frac{\partial\psi_k}{\partial\theta_k}\right]^{-1}\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_k}\right)\left(\left[\mathrm{E}\frac{\partial\psi_j}{\partial\theta_j}\right]^{-1}\mathrm{E}\frac{\partial\psi_{12}}{\partial\theta_j}\right)^{-1}\mathrm{E}\psi_1\psi_2\right\}.$$

Furthermore, from the calculation steps leading to the asymptotic variance expression, we can see that $\tilde{\theta}_1$, $\tilde{\theta}_2$ and $\tilde{\theta}_{12}$ are $\sqrt{n}$-consistent.

Now we turn to the general model (2.39) where d is arbitrary. As we can see from the detailed development for $d = 2$, it is straightforward to generalize Theorem 2.3 to the case where $d \geq 3$, since in the general situation of the model (2.39), the corresponding IFM are

$$\begin{cases} \Psi_{nj} = \sum_{i=1}^{n} \psi_{i;j}, \quad j = 1, \ldots, d, \\ \Psi_{njk} = \sum_{i=1}^{n} \psi_{i;jk}, \quad 1 \leq j < k \leq d. \end{cases}$$

In (2.39), θ_j $(j = 1, \ldots, d)$ and θ_{jk} $(1 \leq j < k \leq d)$ can be scalars or vectors, and in the latter case, $\psi_j(\theta_j)$ and $\psi_{jk}(\theta_{jk})$ are function vectors.

The asymptotic properties of $\tilde{\boldsymbol{\theta}}$ for a general d is given by the following theorem:

Theorem 2.4 *Consider the model (2.39) and let the dimension of θ_j $(j = 1, \ldots, d)$ be p_j and that of θ_{jk} $(1 \leq j < k \leq d)$ be p_{jk}. Let $\tilde{\boldsymbol{\theta}}$ denote the IFME of $\boldsymbol{\theta}$ under the IFM corresponding to (2.42). Then $\tilde{\boldsymbol{\theta}}$ is a consistent estimator of $\boldsymbol{\theta}$. Furthermore, as $n \to \infty$, we have asymptotically*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_\Psi^{-1}),$$

where $q = \sum_{j=1}^{d} p_j + \sum_{1 \leq j < k \leq d} p_{jk}$, $J_\Psi = J_\Psi(\boldsymbol{\theta}) = D_\Psi' M_\Psi^{-1} D_\Psi$, with $M_\Psi = M_\Psi(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\Psi\Psi'\}$ and $D_\Psi = E\{\partial\Psi/\partial\boldsymbol{\theta}\}$. □

For many models, we have $p_{jk} = 1$; that is θ_{jk} is a scalar (see Chapter 3).

The asymptotic variance-covariance of $\tilde{\boldsymbol{\theta}}$ is expressed in terms of a Godambe information matrix $J_\Psi(\boldsymbol{\theta})$. Assume θ_j $(j = 1, \ldots, d)$ and θ_{jk} $(1 \leq j < k \leq d)$ are scalars. Let us see how we can calculate the Godambe information matrix J_Ψ. We first calculate the matrix M_Ψ and then D_Ψ. Since $M_\Psi = \mathrm{E}(\Psi\Psi')$, we only need to calculate its typical components $\mathrm{E}(\psi_j\psi_k)$ $(1 \leq j, k \leq d)$, $\mathrm{E}(\psi_j\psi_{km})$ $(k < m)$ where j may be equal to k or m, and $\mathrm{E}(\psi_{jk}\psi_{lm})$ $(j < k, l < m)$, where j may be equal to l and k may be equal to m.

For $\mathrm{E}(\psi_j\psi_k)$, we have

$$\begin{aligned} \mathrm{E}(\psi_j\psi_k) &= \mathrm{E}\left(\frac{1}{P_j(y_j)}\frac{1}{P_j(y_k)}\frac{\partial P_j(y_j)}{\partial\theta_j}\frac{\partial P_k(y_k)}{\partial\theta_k}\right) \\ &= \sum_{\{y_j y_k\}} \frac{P_{jk}(y_j y_k)}{P_j(y_j)P_k(y_k)}\frac{\partial P_j(y_j)}{\partial\theta_j}\frac{\partial P_k(y_k)}{\partial\theta_k}. \end{aligned}$$

It can be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})P_k(y_{ik})}\frac{\partial P_j(y_{ij})}{\partial\theta_{ij}}\frac{\partial P_k(y_{ik})}{\partial\theta_{i;k}}\bigg|_{\tilde{\theta}_j\tilde{\theta}_k}, \tag{2.47}$$

or equivalently by

$$\frac{1}{n}\sum_{\{y_j y_k\}} \frac{n_{jk}(y_j y_k)}{P_j(y_j)P_k(y_k)} \frac{\partial P_j(y_j)}{\partial \theta_j} \frac{\partial P_k(y_k)}{\partial \theta_k}\bigg|_{\tilde{\theta}_j \tilde{\theta}_k},$$

based on the summary data. For the case $j = k$, we need to replace $P_{jk}(y_j y_k)$ by $P_j(y_j)$, $\{y_j y_k\}$ by $\{y_j\}$ and $n_{jk}(y_j y_k)$ by $n_j(y_j)$ in the above expressions.

For $\mathrm{E}(\psi_j \psi_{km})$ $(k < m)$, we have

$$\begin{aligned}\mathrm{E}(\psi_j \psi_{km}) &= \mathrm{E}\left(\frac{1}{P_j(y_j)} \frac{1}{P_{km}(y_k y_m)} \frac{\partial P_j(y_j)}{\partial \theta_j} \frac{\partial P_{km}(y_k y_m)}{\partial \theta_{km}}\right) \\ &= \sum_{\{y_j y_k y_m\}} \frac{P_{jkm}(y_j y_k y_m)}{P_j(y_j)P_{km}(y_k y_m)} \frac{\partial P_j(y_j)}{\partial \theta_j} \frac{\partial P_{km}(y_k y_m)}{\partial \theta_{km}}.\end{aligned}$$

It can be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})P_{km}(y_{ik}y_{im})} \frac{\partial P_j(y_{ij})}{\partial \theta_j} \frac{\partial P_{km}(y_{ik}y_{im})}{\partial \theta_{km}}\bigg|_{\tilde{\theta}_j \tilde{\theta}_k \tilde{\theta}_{km}}. \tag{2.48}$$

For the case $j = k$ or $j = m$, a slight modification on the above expression is necessary. For example for $j = k$, we need to replace $P_{jkm}(y_j y_k y_m)$ by $P_{jm}(y_j y_m)$, $\{y_j y_k y_m\}$ by $\{y_j y_m\}$ and $n_{jkm}(y_j y_k y_m)$ by $n_{jm}(y_j y_m)$ in the above expressions.

For $\mathrm{E}(\psi_{jk}\psi_{lm})$ $(j < k, l < m)$, we have

$$\begin{aligned}\mathrm{E}(\psi_{jk}\psi_{lm}) &= \mathrm{E}\left(\frac{1}{P_{jk}(y_j y_k)} \frac{1}{P_{lm}(y_l y_m)} \frac{\partial P_{jk}(y_j y_k)}{\partial \theta_{jk}} \frac{\partial P_{lm}(y_l y_m)}{\partial \theta_{lm}}\right) \\ &= \sum_{\{y_j y_k y_l y_m\}} \frac{P_{jklm}(y_j y_k y_l y_m)}{P_{jk}(y_j y_k)P_{lm}(y_l y_m)} \frac{\partial P_{jk}(y_j y_k)}{\partial \theta_{jk}} \frac{\partial P_{lm}(y_l y_m)}{\partial \theta_{lm}}.\end{aligned}$$

It can be estimated consistently by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_{jk}(y_{ij}y_{ik})P_{lm}(y_{il}y_{im})} \frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial \theta_{jk}} \frac{\partial P_{lm}(y_{il}y_{im})}{\partial \theta_{lm}}\bigg|_{\tilde{\theta}_j \tilde{\theta}_k \tilde{\theta}_i \tilde{\theta}_m \tilde{\theta}_{jk} \tilde{\theta}_{lm}}. \tag{2.49}$$

For the particular case $j = l, k \neq m$ (or similarly $j = m$ or $k = l$ or $j \neq l, k = m$ cases), we need to replace $P_{jklm}(y_j y_k y_l y_m)$ by $P_{jkm}(y_j y_k y_m)$, $\{y_j y_k y_l y_m\}$ by $\{y_j y_k y_m\}$ and $n_{jklm}(y_j y_k y_l y_m)$ by $n_{jkm}(y_j y_k y_m)$ in the above expressions. For the particular case $j = l, k = m$, we need to replace $P_{jklm}(y_j y_k y_l y_m)$ by $P_{jk}(y_j y_k)$, $\{y_j y_k y_l y_m\}$ by $\{y_j y_k\}$ and $n_{jklm}(y_j y_k y_l y_m)$ by $n_{jk}(y_j y_k)$ in the above expressions.

Now let us calculate D_Ψ. Since $D_\Psi = D_\Psi(\boldsymbol{\theta})$ is a matrix with (p, q) element $\mathrm{E}(\partial \psi_p / \partial \theta_q)$ $(1 \leq j, k \leq q)$, where ψ_p is the pth components of Ψ and θ_q is the qth component of $\boldsymbol{\theta}$, we only need to calculate its typical component $\mathrm{E}(\partial \psi_j / \partial \theta_m)$ $(1 \leq j, m \leq d)$, $\mathrm{E}(\partial \psi_j / \partial \theta_{lm})$ $(1 \leq j \leq d; 1 \leq l < m \leq$

d), $\mathrm{E}(\partial\psi_{jk}/\partial\theta_m)$ $(1 \le j < k \le d; 1 \le m \le d)$, and $\mathrm{E}(\partial\psi_{jk}/\partial\theta_{lm})$ $(1 \le j < k \le d; 1 \le l < m \le d)$. Since

$$\frac{\partial\psi_j}{\partial\theta_m} = -\frac{1}{P_j^2(y_j)}\frac{\partial P_j(y_j)}{\partial\theta_j}\frac{\partial P_j(y_j)}{\partial\theta_m} + \frac{1}{P_j(y_j)}\frac{\partial^2 P_j(y_j)}{\partial\theta_j\partial\theta_m},$$

we have

$$\begin{aligned}\mathrm{E}\left(\frac{\partial\psi_j}{\partial\theta_m}\right) &= \sum_{\{y_j\}} P_j(y_j)\left(-\frac{1}{P_j^2(y_j)}\frac{\partial P_j(y_j)}{\partial\theta_j}\frac{\partial P_j(y_j)}{\partial\theta_m} + \frac{1}{P_j(y_j)}\frac{\partial^2 P_j(y_j)}{\partial\theta_j\partial\theta_m}\right)\\ &= \sum_{\{y_j\}} -\frac{1}{P_j(y_j)}\frac{\partial P_j(y_j)}{\partial\theta_j}\frac{\partial P_j(y_j)}{\partial\theta_m}.\end{aligned}$$

It can be estimated consistently by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_j^2(y_{ij})}\frac{\partial P_j(y_{ij})}{\partial\theta_j}\frac{\partial P_j(y_{ij})}{\partial\theta_m}\bigg|_{\tilde\theta_j\tilde\theta_m}. \tag{2.50}$$

Because P_j depends only on univariate parameter θ_j, thus ψ_j does not depend on the parameter θ_{lm}. So $\partial\psi_j(\theta_j)/\partial\theta_{lm} = 0$; this also leads to

$$\mathrm{E}\left(\frac{\partial\psi_j(\theta_j)}{\partial\theta_{lm}}\right) = 0.$$

Since

$$\frac{\partial\psi_{jk}}{\partial\theta_m} = -\frac{1}{P_{jk}^2(y_j y_k)}\frac{\partial P_{jk}(y_j y_k)}{\partial\theta_{jk}}\frac{\partial P_{jk}(y_j y_k)}{\partial\theta_m} + \frac{1}{P_{jk}(y_j y_k)}\frac{\partial^2 P_{jk}(y_j y_k)}{\partial\theta_{jk}\partial\theta_m},$$

we have

$$\mathrm{E}\left(\frac{\partial\psi_{jk}}{\partial\theta_m}\right) = \sum_{\{y_j y_k\}} -\frac{1}{P_{jk}(y_j y_k)}\frac{\partial P_{jk}(y_j y_k)}{\partial\theta_{jk}}\frac{\partial P_{jk}(y_j y_k)}{\partial\theta_m}.$$

It can be estimated consistently by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_{jk}^2(y_{ij}y_{ik})}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\theta_{jk}}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\theta_m}\bigg|_{\tilde\theta_j\tilde\theta_k\tilde\theta_m\tilde\theta_{jk}}. \tag{2.51}$$

Similarly, we find

$$\mathrm{E}\left(\frac{\partial\psi_{jk}}{\partial\theta_{lm}}\right) = \begin{cases}\displaystyle\sum_{\{y_j y_k\}} -\frac{1}{P_{jk}(y_j y_k)}\left(\frac{\partial P_{jk}(y_j y_k)}{\partial\theta_{jk}}\right)^2, & j = l,\ k = m,\\ 0, \quad \text{otherwise},\end{cases}$$

where $\mathrm{E}\left(\partial\psi_{jk}/\partial\theta_{jk}\right)$ can be estimated consistently by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_{jk}^2(y_{ij}y_{ik})}\left(\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\theta_{jk}}\right)^2\bigg|_{\tilde\theta_j\tilde\theta_k\tilde\theta_{jk}}. \tag{2.52}$$

2.4.2 Models with covariates

In this subsection, we consider extensions of models to include covariates. Under regularity conditions, the IFME for parameters are shown to be consistent and asymptotically normal and the form of the asymptotic variance-covariance matrix is derived. One approach for asymptotic results is given in this subsection, a second approach is given in the next subsection. Our objective is to set forth results as simply as possible and in useful forms; more general theorems for multivariate models could be obtained.

Let $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be a sequence of independent random vectors of dimension d, which are defined on the probability measure space $(\mathcal{Y}, \mathcal{A}, P(\mathbf{Y}; \boldsymbol{\theta}))$, $\boldsymbol{\theta} \in \Re \subset I\!R^q$. The marginal probability measure spaces are defined as $(\mathcal{Y}_j, \mathcal{A}_j, P(Y_j; \boldsymbol{\theta}))$ $(j = 1, \ldots, d)$ for jth margin, and $(\mathcal{Y}_{jk}, \mathcal{A}_{jk}, P(Y_j, Y_k; \boldsymbol{\theta}))$ $(1 \leq j < k \leq d)$ for the (j, k) bivariate margin and so on. Particularly, the random vectors $\mathbf{Y}_i$ $(i = 1, \ldots, n)$ are assumed to have the distribution

$$P(y_{i1} \cdots y_{id}; \boldsymbol{\theta}_i), \ \boldsymbol{\theta}_i \in \Re, \tag{2.53}$$

where $P(y_{i1} \cdots y_{id}; \boldsymbol{\theta}_i)$ is a MCD or MMD model with univariate and bivariate expressible (PUBE) parameters $\boldsymbol{\theta}_i = (\theta_{i;1}, \ldots, \theta_{i;d}, \theta_{i;12}, \ldots, \theta_{i;d-1,d})$. We also assume that $\theta_{i;j}$ $(j = 1, \ldots, d)$ is the parameter for the jth univariate margin of $\mathbf{Y}_i$ such that $P_j(y_{ij})$ depends only on the parameter $\theta_{i;j}$, and $\theta_{i;jk}$ $(1 \leq j < k \leq d)$ is the parameter for the (j, k) bivariate margin of $\mathbf{Y}_i$ such that $\theta_{i;jk}$ is the only bivariate parameter appearing in $P_{jk}(y_{ij} y_{ik})$. $\theta_{i;j}$ and $\theta_{i;jk}$ can both be vectors, but for the purpose of illustration, and without loss of generality, we assume they are scalar parameters. Furthermore, we assume for $i = 1, \ldots, n$,

$$\begin{cases} \theta_{i;j} = g_j(\boldsymbol{\alpha}_j' \mathbf{x}_{ij}), & j = 1, 2, \ldots, d, \\ \theta_{i;jk} = h_{jk}(\boldsymbol{\beta}_{jk}' \mathbf{w}_{ijk}), & 1 \leq j < k \leq d, \end{cases} \tag{2.54}$$

where the functions $g_j(\cdot)$ and $h_{jk}(\cdot)$ are usually monotonic increasing (or decreasing) and differentiable functions (for examples of the choice of $g_j(\cdot)$ and $h_{jk}(\cdot)$ in a specific model, see Chapter 3), called *link functions*. They link the model parameters to a set of covariates through a functional relationship. In (2.54), $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp_j})'$ is a $p_j \times 1$ parameter vector, $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp_j})'$ is a $p_j \times 1$ vector of observable covariates corresponding to the response $\mathbf{y}_i$. $\boldsymbol{\beta}_{jk} = (\beta_{jk1}, \ldots, \beta_{jkq_{jk}})'$ is a $q_{jk} \times 1$ parameter vector, and $\mathbf{w}_{ijk} = (w_{ijk1}, \ldots, w_{ijkq_{jk}})'$ is a $q_{jk} \times 1$ vector of observable covariates corresponding to the response $\mathbf{y}_i$. Usually $\sum_j p_j + \sum_{j<k} q_{jk}$ is much smaller than the total number of observations n. $\mathbf{x}_{ij}$ and $\mathbf{w}_{ijk}$ may or may not include the same covariate components. In terms of margins, the covariates $\mathbf{x}_{ij}$ and $\mathbf{w}_{ijk}$ may be margin-dependent, or margin-independent,

or partly margin-dependent and partly margin-independent. The marginal parameters may also be margin-dependent or margin-independent.

We consider the problem of how to estimate the regression parameters in the statistical model defined by (2.53) and (2.54). We assume we have a well-structured data set as in Table 1.1. Problems such as the misspecification of the link function, the omission of one or several important covariates, the random nature of some covariates, missing values, and so on will not be dealt with here.

From (2.54) and the PUBE assumption, P_j can be considered as a function of $\boldsymbol{\alpha}_j$ and P_{jk} can be considered as a function of $\boldsymbol{\alpha}_j$, $\boldsymbol{\alpha}_k$, and $\boldsymbol{\beta}_{jk}$. Let $\mathbf{y}_1, \ldots, \mathbf{y}_n$ be the observed values of $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$. The loglikelihood functions of margins of the parameter vectors $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$ based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$ are

$$\begin{cases} \ell_{nj}(\boldsymbol{\alpha}_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\boldsymbol{\alpha}_j, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d. \end{cases} \tag{2.55}$$

Let

$$\begin{cases} \varphi_{i;j} = \varphi_{i;j}(\theta_{i;j}) \stackrel{\text{def}}{=} \dfrac{1}{P_j(y_{ij})} \dfrac{\partial P_j(y_{ij})}{\partial \theta_{i;j}}, \quad j = 1, \ldots, d, \\ \varphi_{i;jk} = \varphi_{i;jk}(\theta_{i;jk}) \stackrel{\text{def}}{=} \dfrac{1}{P_{jk}(y_{ij} y_{ik})} \dfrac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \theta_{i;jk}}, \quad 1 \le j < k \le d, \end{cases}$$

and

$$\begin{cases} \dot{\varphi}_{i;j,j} \stackrel{\text{def}}{=} \dfrac{\partial \varphi_{i;j}(\theta_{i;j})}{\partial \theta_{i;j}} \quad j = 1, \ldots, d, \\ \dot{\varphi}_{i;jk,j} \stackrel{\text{def}}{=} \dfrac{\partial \varphi_{i;jk}(\theta_{i;jk})}{\partial \theta_{i;j}}, \quad \dot{\varphi}_{i;jk,k} \stackrel{\text{def}}{=} \dfrac{\partial \varphi_{i;jk}(\theta_{i;jk})}{\partial \theta_{i;k}}, \quad \dot{\varphi}_{i;jk,jk} \stackrel{\text{def}}{=} \dfrac{\partial \varphi_{i;jk}(\theta_{i;jk})}{\partial \theta_{i;jk}}, \quad 1 \le j < k \le d, \end{cases}$$

and

$$\begin{cases} \psi_{i;js} = \psi_{i;js}(\alpha_{js}) \stackrel{\text{def}}{=} \dfrac{1}{P_j(y_{ij})} \dfrac{\partial P_j(y_{ij})}{\partial \alpha_{js}}, \quad j = 1, \ldots, d;\ s = 1, \ldots, p_j, \\ \psi_{i;jkt} = \psi_{i;jkt}(\beta_{jkt}) \stackrel{\text{def}}{=} \dfrac{1}{P_{jk}(y_{ij} y_{ik})} \dfrac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \beta_{jkt}}, \quad 1 \le j < k \le d;\ t = 1, \ldots, q_{jk}. \end{cases}$$

Let $\boldsymbol{\gamma} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_d', \boldsymbol{\beta}_{12}', \ldots, \boldsymbol{\beta}_{d-1,d}')'$, and $\boldsymbol{\gamma}_0 = (\boldsymbol{\alpha}_{1,0}', \ldots, \boldsymbol{\alpha}_{2,0}', \boldsymbol{\beta}_{12,0}', \ldots, \boldsymbol{\beta}_{d-1,d,0}')'$, where $\boldsymbol{\gamma}_0$ is the true vector value of $\boldsymbol{\gamma}$. Let $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\alpha}}_1', \ldots, \tilde{\boldsymbol{\alpha}}_d', \tilde{\boldsymbol{\beta}}_{12}', \ldots, \tilde{\boldsymbol{\beta}}_{d-1,d}')'$ be the IFME. Assume $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}_0$ and $\tilde{\boldsymbol{\gamma}}$ are all $q \times 1$ vectors. Let

$$\begin{aligned} \boldsymbol{\Psi}_{i;j} &= \boldsymbol{\Psi}_{i;j}(\boldsymbol{\alpha}_j) = (\psi_{i;j1}, \ldots, \psi_{i;jp_j})', \\ \boldsymbol{\Psi}_{i;jk} &= \boldsymbol{\Psi}_{i;jk}(\boldsymbol{\beta}_{jk}) = (\psi_{i;jk1}, \ldots, \psi_{i;jkq_{jk}})', \\ \boldsymbol{\Psi}_{nj} &= \boldsymbol{\Psi}_{nj}(\boldsymbol{\alpha}_j) = (\Psi_{nj1}, \ldots, \Psi_{nj,p_j})', \\ \boldsymbol{\Psi}_{njk} &= \boldsymbol{\Psi}_{njk}(\boldsymbol{\beta}_{jk}) = (\Psi_{njk1}, \ldots, \Psi_{njk,q_{jk}})', \end{aligned}$$

where $\Psi_{njs} = \sum_{i=1}^{n} \psi_{i;js}$ $(s = 1, \ldots, p_j)$ and $\Psi_{njkt} = \sum_{i=1}^{n} \psi_{i;jkt}$ $(t = 1, \ldots, q_{jk})$. Let $\Psi_{i;}(\boldsymbol{\gamma}) = (\Psi_{i;1}(\boldsymbol{\alpha}_1)', \ldots, \Psi_{i;d}(\boldsymbol{\alpha}_d)', \Psi_{i;12}(\boldsymbol{\beta}_{12})', \ldots, \Psi_{i;d-1,d}(\boldsymbol{\beta}_{d-1,d})')'$, $\Psi_n(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \Psi_{i;}(\boldsymbol{\gamma})$, and we define

$$M_n(\boldsymbol{\gamma}) = n^{-1} \sum_{i=1}^{n} \mathrm{E}(\Psi_{i;}(\boldsymbol{\gamma})\Psi_{i;}(\boldsymbol{\gamma})') \quad \text{and} \quad D_n(\boldsymbol{\gamma}) = n^{-1}\mathrm{E}\left\{\frac{\partial \Psi_n(\boldsymbol{\gamma})}{\partial \boldsymbol{\gamma}}\right\}. \tag{2.56}$$

From (2.55), the IFM for $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$ are

$$\begin{cases} \Psi_{nj} = \sum_{i=1}^{n} \varphi_{i;j} \dfrac{\partial g_j(\cdot)}{\partial \boldsymbol{\alpha}_j}, \quad j = 1, \ldots, d, \\ \Psi_{njk} = \sum_{i=1}^{n} \varphi_{i;jk} \dfrac{\partial h_{jk}(\cdot)}{\partial \boldsymbol{\beta}_{jk}}, \quad 1 \le j < k \le d, \end{cases} \tag{2.57}$$

and the estimating equations based on IFM are

$$\begin{cases} \Psi_{nj} = \mathbf{0}, \quad j = 1, \ldots, d, \\ \Psi_{njk} = \mathbf{0}, \quad 1 \le j < k \le d. \end{cases} \tag{2.58}$$

With the IFM approach, estimates of $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$, denoted by $\tilde{\boldsymbol{\alpha}}_j = (\tilde{\alpha}_{j1}, \ldots, \tilde{\alpha}_{j,p_j})'$ and $\tilde{\boldsymbol{\beta}}_{jk} = (\tilde{\beta}_{jk1}, \ldots, \tilde{\beta}_{jk,q_{jk}})'$, are obtained by solving the nonlinear system of equations (2.58).

We next give several necessary assumptions for establishing asymptotic properties of $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_{jk}$.

Assumptions 2.1 *For $1 \le j < k \le d$, $1 \le l < m \le d$ and $i = 1, \ldots, n$, we make the following assumptions:*

1. *(a) For almost all $\mathbf{y}_i \in \mathcal{Y}$ and for all $\boldsymbol{\theta}_i \in \Re$,*

$$\varphi_{i;j},\ \varphi_{i;jk},\ \dot{\varphi}_{i;j,j},\ \dot{\varphi}_{i;jk,j},\ \dot{\varphi}_{i;jk,k},\ \dot{\varphi}_{i;jk,jk}$$

exist.

(b) For almost all $\mathbf{y}_i \in \mathcal{Y}$ and for every $\boldsymbol{\theta}_i \in \Re$,

$$\left|\frac{\partial P_j(y_{ij})}{\partial \theta_{i;j}}\right| < K_{i;j};\ \left|\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;j}}\right| < L_{i;j};\ \left|\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;k}}\right| < L_{i;k};\ \left|\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;jk}}\right| < L_{i;jk};$$
$$\left|\frac{\partial^2 P_j(y_{ij})}{\partial \theta_{i;j}^2}\right| < S_{i;j};\ \left|\frac{\partial^2 P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;j}^2}\right| < T_{i;j};\ \left|\frac{\partial^2 P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;k}^2}\right| < T_{i;k};$$
$$\left|\frac{\partial^2 P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;jk}^2}\right| < T_{i;jk};\ \left|\frac{\partial^2 P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;jk}\partial \theta_{i;j}}\right| < T_{i;jk,j};\ \left|\frac{\partial^2 P_{jk}(y_{ij}y_{ik})}{\partial \theta_{i;jk}\partial \theta_{i;k}}\right| < T_{i;jk,k},$$

where $K_{i;j}$, $L_{i;j}$, $L_{i;k}$, $L_{i;jk}$, $S_{i;j}$, $T_{i;j}$, $T_{i;k}$, $T_{i;jk}$, $T_{i;jk,j}$ and $T_{i;jk,k}$ are integrable (summable) over $\mathcal{Y}$, and are all bounded by some positive constants.

2. (a)

$$\begin{cases} \sum_{i=1}^{n} \sum_{\{y_{ij} \,:\, |\varphi_{i;j}|>n\}} P_j(y_{ij}) = o_p(1), \\ \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\varphi_{i;jk}|>n\}} P_{jk}(y_{ij}y_{ik}) = o_p(1), \end{cases}$$

and

$$\begin{cases} \sum_{i=1}^{n} \sum_{\{y_{ij} \,:\, |\varphi_{i;j}|<n\}} \varphi_{i;j}^2 P_j(y_{ij}) = o_p(n^2), \\ \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, \sup(|\varphi_{i;j}|,|\varphi_{i;k}|)<n\}} \varphi_{i;j}\varphi_{i;k} P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\ \sum_{i=1}^{n} \sum_{\{y_{ij},y_{il},y_{im} \,:\, \sup(|\varphi_{i;j}|,|\varphi_{i;lm}|)<n\}} \varphi_{i;j}\varphi_{i;lm} P_{jlm}(y_{ij}y_{il}y_{im}) = o_p(n^2), \\ \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\varphi_{i;jk}|<n\}} \varphi_{i;jk}^2 P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\ \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik},y_{il},y_{im} \,:\, \sup(|\varphi_{i;jk}|,|\varphi_{i;lm}|)<n\}} \varphi_{i;jk}\varphi_{i;lm} P_{jklm}(y_{ij}y_{ik}y_{il}y_{im}) = o_p(n^2). \end{cases}$$

(b)

$$\begin{cases} \sum_{i=1}^{n} \sum_{\{y_{ij} \,:\, |\dot{\varphi}_{i;j,j}|>n\}} P_j(y_{ij}) = o_p(1), \quad \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\dot{\varphi}_{i;jk,j}|>n\}} P_{jk}(y_{ij}y_{ik}) = o_p(1), \\ \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\dot{\varphi}_{i;jk,k}|>n\}} P_{jk}(y_{ij}y_{ik}) = o_p(1), \quad \sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\dot{\varphi}_{i;jk,jk}|>n\}} P_{jk}(y_{ij}y_{ik}) = o_p(1), \end{cases}$$

and

$$\begin{cases}
\sum_{i=1}^{n} \sum_{\{y_{ij} \,:\, |\dot{\varphi}_{i;j,j}|<n\}} \dot{\varphi}^2_{i;j,j} P_j(y_{ij}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\dot{\varphi}_{i;jk,j}|<n\}} \dot{\varphi}^2_{i;jk,j} P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\dot{\varphi}_{i;jk,k}|<n\}} \dot{\varphi}^2_{i;jk,k} P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, |\dot{\varphi}_{i;jk,jk}|<n\}} \dot{\varphi}^2_{i;jk,jk} P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij}y_{ik} \,:\, \sup(|\dot{\varphi}_{i;j,j}|,|\dot{\varphi}_{i;k,k}|)<n\}} \dot{\varphi}_{i;j,j}\dot{\varphi}_{i;k,k} P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{il},y_{im} \,:\, \sup(|\dot{\varphi}_{i;j,j}|,|\dot{\varphi}_{i;lm,l}|)<n\}} \dot{\varphi}_{i;j,j}\dot{\varphi}_{i;lm,l} P_{jlm}(y_{ij}y_{il}y_{im}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{il},y_{im} \,:\, \sup(|\dot{\varphi}_{i;j,j}|,|\dot{\varphi}_{i;lm,lm}|)<n\}} \dot{\varphi}_{i;j,j}\dot{\varphi}_{i;lm,lm} P_{jlm}(y_{ij}y_{il}y_{im}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik} \,:\, \sup(|\dot{\varphi}_{i;jk,j}|,|\dot{\varphi}_{i;jk,k}|)<n\}} \dot{\varphi}_{i;jk,j}\dot{\varphi}_{i;jk,k} P_{jk}(y_{ij}y_{ik}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik},y_{il},y_{im} \,:\, \sup(|\dot{\varphi}_{i;jk,j}|,|\dot{\varphi}_{i;lm,lm}|)<n\}} \dot{\varphi}_{i;jk,j}\dot{\varphi}_{i;lm,lm} P_{jklm}(y_{ij}y_{ik}y_{il}y_{im}) = o_p(n^2), \\
\sum_{i=1}^{n} \sum_{\{y_{ij},y_{ik},y_{il},y_{im} \,:\, \sup(|\dot{\varphi}_{ijk,jk}|,|\dot{\varphi}_{ilm,lm}|)<n\}} \dot{\varphi}_{ijk,jk}\dot{\varphi}_{ilm,lm} P_{jklm}(y_{ij}y_{ik}y_{il}y_{im}) = o_p(n^2).
\end{cases}$$

□

Assumptions 2.2 *Let $\mathcal{G} \in \mathbb{R}^k$ is the parameter space of $\boldsymbol{\gamma}$, where $\boldsymbol{\gamma}$ is assumed to be vector of length q. Suppose $\mathcal{G}$ is an open convex set.*

1. *Each $g_j(\cdot)$ and $h_{jk}(\cdot)$ is twice continuously differentiable.*

2. *(a) The covariates are uniformly bounded, that is, there exist an M_0, such that $||\mathbf{x}_{ij}|| \leq M_0$, $||\mathbf{w}_{ijk}|| \leq M_0$. Furthermore, $\sum_i \mathbf{x}_{ij}\mathbf{x}_{ij}^T$, $\sum_i \mathbf{w}_{ijk}\mathbf{w}_{ijk}^T$ have full rank for $n \geq n_0$, where n_0 is some positive integer value.*

   (b) In the neighborhood of the true $\boldsymbol{\gamma}$ defined by

$$B(\delta) = \{\boldsymbol{\gamma}^* \in \mathcal{G} \,:\, ||\boldsymbol{\gamma}^* - \boldsymbol{\gamma}|| < \delta\}, \quad \delta \downarrow 0,$$

$g_j(\cdot)$, $g_j'(\cdot)$,$g_j''(\cdot)$, $h_{jk}(\cdot)$, $h_{jk}'(\cdot)$ and $h_{jk}''(\cdot)$ are bounded away from zero.

□

Assumption 2.3 *For all $\epsilon > 0$ and for any fixed vector $||\mathbf{u}|| \neq 0$, the following condition is satisfied*

$$\lim_{n\to\infty} \frac{1}{(\mathbf{u}' M_n(\boldsymbol{\gamma}_0)\mathbf{u})^{1/2}} \sum_{i=1}^{n} \sum_{\{|\mathbf{u}'\xi_i(\boldsymbol{\gamma}_0)| \geq \epsilon(\mathbf{u}' M_n(\boldsymbol{\gamma}_0)\mathbf{u})^{1/2}\}} [\mathbf{u}'\xi_i(\boldsymbol{\gamma}_0)]^2 P(\mathbf{Y};\boldsymbol{\gamma}_0) = 0.$$

□

Assumptions 2.1 and 2.2 are needed so that we may apply some weak law of large numbers theorem to prove the consistency of the estimates. Assumption 2.3 is needed for applying the central limit theorem for deriving the asymptotic normality of the estimates. These conditions appear to be complicated, but for special cases they are often not difficult to verify. For instance, for the models we will use in Chapter 5, if the covariates are bounded and have full rank in the sense of Assumptions 2.2, with appropriate choice of the link functions, the conditions are almost empty.

Related to statistical literature, Bradley and Gart (1962) studied the asymptotic properties of maximum likelihood estimates (MLE's) where the observations are independent but not identically distributed (i.n.i.d.). Hoadley (1971) established general conditions (for cases where the observations are i.n.i.d. and there is a lack of densities with respect to Lebesque or counting measure) under which maximum likelihood estimators are consistent and asymptotically normal. Assumptions 2.1, 2.2 and 2.3 are closely related to the conditions in Bradley and Gart (1962) and Hoadley (1971), but adapted to multivariate discrete models with MPME property. In particular, the Assumptions 2.1 reflect the *uniform integrability* concept (for discrete models) employed in Hoadley (1971).

Properties of estimators

As with the model (2.39) with no covariates, we first develop the asymptotic results for the simplest situation when $d = 2$ such that $\mathbf{Y}_i = (Y_{i1}, Y_{i2})'$ $(i = 1, \ldots, n)$ has the distribution

$$P(y_{i1}y_{i2}; \boldsymbol{\theta}_i), \tag{2.59}$$

where $\boldsymbol{\theta}_i = (\theta_{i;1}, \theta_{i;2}, \theta_{i;12})$. Without loss of generality, $\theta_{i;1}, \theta_{i;2}, \theta_{i;12}$ are assumed to be scalar parameters. We further assume

$$\begin{cases} \theta_{i;j} = g_j(\boldsymbol{\alpha}_j'\mathbf{x}_{ij}), \quad j = 1, 2, \\ \theta_{i;12} = h_{12}(\boldsymbol{\beta}_{12}'\mathbf{w}_{i12}), \end{cases} \tag{2.60}$$

where the functions $g_j(\cdot)$ and $h_{12}(\cdot)$ are monotonic increasing (or decreasing) and differentiable functions. In (2.60), $\boldsymbol{\alpha}_j = (\alpha_{j1}, \ldots, \alpha_{jp_j})'$ is a $p_j \times 1$ parameter vector and $\mathbf{x}_{ij} = (x_{ij1}, \ldots, x_{ijp_j})'$

is a $p_j \times 1$ vector of covariates corresponding to the response variables Y_{ij} $(j = 1, 2)$. Similarly, $\boldsymbol{\beta}_{12} = (\beta_{121}, \ldots, \beta_{12q_{12}})'$ is a $q_{12} \times 1$ parameter vector and $\mathbf{w}_{i12} = (w_{i121}, \ldots, w_{i12q_{12}})'$ is a $q_{12} \times 1$ vector of covariates corresponding to the response vector $\mathbf{y}_i$, where $\mathbf{y}_i = (y_{i1}, y_{i2})$ is the observed value of $\mathbf{Y}_i$.

Theorem 2.5 *Consider the model (2.59) together with (2.60) and let* $\tilde{\boldsymbol{\gamma}} = (\tilde{\boldsymbol{\alpha}}_1', \tilde{\boldsymbol{\alpha}}_2', \tilde{\boldsymbol{\beta}}_{12}')'$ *denote the IFME of* $\boldsymbol{\gamma}$ *corresponding to IFM (2.57). Assume* $\boldsymbol{\gamma}$ *is a* $q \times 1$ *vector. Under Assumptions 2.1 and 2.2,* $\tilde{\boldsymbol{\gamma}}$ *is a consistent estimator of* $\boldsymbol{\gamma}_0$*. Furthermore, under Assumptions 2.1, 2.2 and 2.3, as* $n \to \infty$*, we have*

$$\sqrt{n} A_n^{-1} (\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{D} N_q(\mathbf{0}, I),$$

where $A_n = D_n^{-1/2}(\boldsymbol{\gamma}_0) M_n^{1/2}(\boldsymbol{\gamma}_0) (D_n^{-1/2}(\boldsymbol{\gamma}_0))^T$*, with* $D_n(\boldsymbol{\gamma}_0)$ *and* $M_n(\boldsymbol{\gamma}_0)$ *defined in (2.56).*

Proof: Using Taylor's theorem to the first order, we have

$$\begin{cases} \Psi_{n1}(\tilde{\boldsymbol{\alpha}}_1) = \Psi_{n1}(\boldsymbol{\alpha}_{1,0}) + (\tilde{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1,0}) \left. \dfrac{\partial \Psi_{n1}}{\partial \boldsymbol{\alpha}_1} \right|_{\boldsymbol{\alpha}_1^*}, \\ \Psi_{n2}(\tilde{\boldsymbol{\alpha}}_2) = \Psi_{n2}(\boldsymbol{\alpha}_{2,0}) + (\tilde{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_{2,0}) \left. \dfrac{\partial \Psi_{n2}}{\partial \boldsymbol{\alpha}_2} \right|_{\boldsymbol{\alpha}_2^*}, \\ \Psi_{n12}(\tilde{\boldsymbol{\beta}}_{12}) = \Psi_{n12}(\boldsymbol{\beta}_{12,0}) + (\tilde{\boldsymbol{\beta}}_{12} - \boldsymbol{\beta}_{12,0}) \left. \dfrac{\partial \Psi_{n12}}{\partial \boldsymbol{\beta}_{12}} \right|_{\boldsymbol{\gamma}^{**}} \\ \qquad + (\tilde{\boldsymbol{\alpha}}_1 - \boldsymbol{\alpha}_{1,0}) \left. \dfrac{\partial \Psi_{n12}}{\partial \boldsymbol{\alpha}_1} \right|_{\boldsymbol{\gamma}^{**}} + (\tilde{\boldsymbol{\alpha}}_2 - \boldsymbol{\alpha}_{2,0}) \left. \dfrac{\partial \Psi_{n12}}{\partial \boldsymbol{\alpha}_2} \right|_{\boldsymbol{\gamma}^{**}}, \end{cases} \tag{2.61}$$

where $\boldsymbol{\alpha}_1^*$ is some vector value between $\tilde{\boldsymbol{\alpha}}_1$ and $\boldsymbol{\alpha}_{1,0}$, $\boldsymbol{\alpha}_2^*$ is some vector value between $\tilde{\boldsymbol{\alpha}}_2$ and $\boldsymbol{\alpha}_{2,0}$, and $\boldsymbol{\gamma}^{**}$ is some vector value between $\tilde{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}_0$.

Note that in (2.61), $\Psi_{n12}(\tilde{\boldsymbol{\beta}}_{12})$ also depends on $\tilde{\boldsymbol{\alpha}}_1$ and $\tilde{\boldsymbol{\alpha}}_2$, and $\Psi_{n12}(\boldsymbol{\beta}_{12,0})$ depends on $\boldsymbol{\alpha}_{1,0}$ and $\boldsymbol{\alpha}_{2,0}$. Furthermore, in (2.61), we have

$$\begin{cases} \dfrac{\partial \Psi_{n1}(\boldsymbol{\alpha}_1)}{\partial \boldsymbol{\alpha}_1} = \displaystyle\sum_{i=1}^{n} [\varphi_{i;1}'(\theta_{i;1})(g_j'(\boldsymbol{\alpha}_1^T \mathbf{x}_{i1}))^2 + \varphi_{i;1}(\theta_{i;1}) g_j''(\boldsymbol{\alpha}_1^T \mathbf{x}_{i1})] \mathbf{x}_{i1} \mathbf{x}_{i1}^T, \\ \dfrac{\partial \Psi_{n2}(\boldsymbol{\alpha}_2)}{\partial \boldsymbol{\alpha}_2} = \displaystyle\sum_{i=1}^{n} [\varphi_{i;2}'(\theta_{i;2})(g_j'(\boldsymbol{\alpha}_2^T \mathbf{x}_{i2}))^2 + \varphi_{i;2}(\theta_{i;2}) g_j''(\boldsymbol{\alpha}_2^T \mathbf{x}_{i2})] \mathbf{x}_{i2} \mathbf{x}_{i2}^T, \\ \dfrac{\partial \Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial \boldsymbol{\beta}_{12}} = \displaystyle\sum_{i=1}^{n} [\varphi_{i;12}'(\theta_{i;12})(h_{jk}'(\boldsymbol{\beta}_{12}^T \mathbf{w}_{i12}))^2 + \varphi_{i;12}(\theta_{i;12}) h_{jk}''(\boldsymbol{\beta}_{12}^T \mathbf{w}_{i12})] \mathbf{w}_{i12} \mathbf{w}_{i12}^T, \\ \dfrac{\partial \Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial \boldsymbol{\alpha}_1} = \displaystyle\sum_{i=1}^{n} \left[\dfrac{\partial \varphi_{i;12}(\theta_{i;12})}{\partial \theta_{i;1}} g_j'(\boldsymbol{\alpha}_1' \mathbf{x}_{i1}) h_{jk}'(\boldsymbol{\beta}_{12}^T \mathbf{w}_{i12}) \right] \mathbf{w}_{i12} \mathbf{x}_{i1}^T, \\ \dfrac{\partial \Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial \boldsymbol{\alpha}_2} = \displaystyle\sum_{i=1}^{n} \left[\dfrac{\partial \varphi_{i;12}(\theta_{i;12})}{\partial \theta_{i;2}} g_j'(\boldsymbol{\alpha}_2' \mathbf{x}_{i2}) h_{jk}'(\boldsymbol{\beta}_{12}^T \mathbf{w}_{i12}) \right] \mathbf{w}_{i12} \mathbf{x}_{i2}^T, \end{cases} \tag{2.62}$$

and $\partial\Psi_{n1}(\boldsymbol{\alpha}_1)/\partial\boldsymbol{\alpha}_2 = 0$, $\partial\Psi_{n1}(\boldsymbol{\alpha}_1)/\partial\boldsymbol{\beta}_{12} = 0$, $\partial\Psi_{n2}(\boldsymbol{\alpha}_2)/\partial\boldsymbol{\alpha}_1 = 0$ and $\partial\Psi_{n2}(\boldsymbol{\alpha}_2)/\partial\boldsymbol{\beta}_{12} = 0$.

To establish the consistency and asymptotic normality of $\tilde{\boldsymbol{\gamma}}$, note that with Assumptions 2.1, we have that $n^{-2}\mathrm{E}(\Psi_{nj}(\boldsymbol{\alpha}_{j,0})\Psi_{nj}(\boldsymbol{\alpha}_{j,0})^T) \to \mathbf{0}$ $(j = 1, 2)$, and $n^{-2}\mathrm{E}(\Psi_{n12}(\boldsymbol{\beta}_{12,0})\Psi_{n12}(\boldsymbol{\beta}_{12,0})^T) \to \mathbf{0}$. By the assumed monotonicity (*e.g.* $\nearrow$) and differentiability of $g_j(\cdot)$ and $h_{12}(\cdot)$, $g_j'(\cdot)$ is a non-negative function of the unknown $\boldsymbol{\alpha}_j$ and the given $\mathbf{x}_{ij}$, and $h_{12}'(\cdot)$ is a non-negative function of the unknown $\boldsymbol{\beta}_{12}$ and the given $\mathbf{w}_{i12}$. From Assumptions 2.1 and 2.2, $\partial\Psi_{nj}(\boldsymbol{\alpha}_j)/\partial\boldsymbol{\alpha}_j$ $(j = 1, 2)$ and $\partial\Psi_{n12}(\boldsymbol{\beta}_{12})/\partial\boldsymbol{\beta}_{12}$ have full rank. By Markov's weak law of large numbers (see for instance Petrov 1995, p.134), we obtain that $n^{-1}\Psi_{n1}(\boldsymbol{\alpha}_{1,0}) \to^p \mathbf{0}$, $n^{-1}\Psi_{n2}(\boldsymbol{\alpha}_{2,0}) \to^p \mathbf{0}$ and $n^{-1}\Psi_{n12}(\boldsymbol{\beta}_{12,0}) \to^p \mathbf{0}$. Since $\Psi_{n1}(\tilde{\boldsymbol{\alpha}}_1) = \mathbf{0}$, $\Psi_{n2}(\tilde{\boldsymbol{\alpha}}_2) = \mathbf{0}$ and $\Psi_{n12}(\tilde{\boldsymbol{\beta}}_{12}) = \mathbf{0}$, by following similar arguments as for the consistency of $\tilde{\boldsymbol{\theta}}$ in the model (2.39), we establish the consistency of $\tilde{\boldsymbol{\gamma}}$.

Now let

$$\Psi_n(\boldsymbol{\gamma}) = \begin{pmatrix} \Psi_{n1}(\boldsymbol{\alpha}_1) \\ \Psi_{n2}(\boldsymbol{\alpha}_2) \\ \Psi_{n12}(\boldsymbol{\beta}_{12}) \end{pmatrix}, \quad H_n(\boldsymbol{\gamma}) = \begin{pmatrix} \frac{\partial\Psi_{n1}(\boldsymbol{\alpha}_1)}{\partial\boldsymbol{\alpha}_1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\partial\Psi_{n2}(\boldsymbol{\alpha}_2)}{\partial\boldsymbol{\alpha}_2} & \mathbf{0} \\ \frac{\partial\Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial\boldsymbol{\alpha}_1} & \frac{\partial\Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial\boldsymbol{\alpha}_2} & \frac{\partial\Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial\boldsymbol{\beta}_{12}} \end{pmatrix},$$

and

$$H_n^* = \begin{pmatrix} \frac{\partial\Psi_{n1}(\boldsymbol{\alpha}_1)}{\partial\boldsymbol{\alpha}_1}\Big|_{\boldsymbol{\alpha}_1^*} & 0 & 0 \\ 0 & \frac{\partial\Psi_{n2}(\boldsymbol{\alpha}_2)}{\partial\boldsymbol{\alpha}_2}\Big|_{\boldsymbol{\alpha}_2^*} & 0 \\ \frac{\partial\Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial\boldsymbol{\alpha}_1}\Big|_{\boldsymbol{\gamma}^{**}} & \frac{\partial\Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial\boldsymbol{\alpha}_2}\Big|_{\boldsymbol{\gamma}^{**}} & \frac{\partial\Psi_{n12}(\boldsymbol{\beta}_{12})}{\partial\boldsymbol{\beta}_{12}}\Big|_{\boldsymbol{\gamma}^{**}} \end{pmatrix}.$$

(2.61) can be rewritten in the following matrix form

$$\sqrt{n}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) = \left[\frac{1}{n}H_n^*\right]^{-1} \frac{1}{\sqrt{n}}[-\Psi_n(\boldsymbol{\gamma}_0)]. \tag{2.63}$$

It follows from the convergence in probability of $\tilde{\boldsymbol{\gamma}}$ to $\boldsymbol{\gamma}_0$ that

$$\frac{1}{n}\left[H_n(\tilde{\boldsymbol{\gamma}}) - H_n(\boldsymbol{\gamma}_0)\right] \xrightarrow{\mathrm{P}} \mathbf{0}.$$

Since each element of $n^{-1}H_n(\boldsymbol{\gamma}_0)$ is the mean of n independent variates, under some conditions for the law of large numbers, we have

$$\frac{1}{n}H_n(\boldsymbol{\gamma}_0) - D_n(\boldsymbol{\gamma}_0) \xrightarrow{p} \mathbf{0}, \tag{2.64}$$

where $D_n(\boldsymbol{\gamma}_0) = n^{-1}\mathrm{E}\{H_n(\boldsymbol{\gamma}_0)\}$.

Assumptions 2.1 and 2.2 imply that

$$\lim_{n\to\infty} n^{-2}\mathrm{Var}(H_n(\boldsymbol{\gamma}_0)) = \mathbf{0}. \tag{2.65}$$

Thus by Markov's weak law of large numbers, we derive that

$$\frac{1}{n}H_n^* - D_n(\boldsymbol{\gamma}_0) \xrightarrow{p} \mathbf{0}. \tag{2.66}$$

Next, we note that $\Psi_n(\boldsymbol{\gamma}_0)$ involves independent centered summands. Therefore, we may directly appeal to the Lindberg-Feller central limit theorem to establish its asymptotic normality. From Assumption 2.3, by the Lindberg-Feller central limit theorem, we have

$$\frac{\mathbf{u}'\Psi_n(\boldsymbol{\gamma}_0)}{(\mathbf{u}'M_n(\boldsymbol{\gamma}_0)\mathbf{u})^{1/2}} \xrightarrow{D} N(0,1).$$

Applying Slutsky's Theorem to (2.63), we obtain

$$\sqrt{n}A_n^{-1}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{D} N_q(\mathbf{0}, I),$$

where $A_n = D_n^{-1/2}(\boldsymbol{\gamma}_0)M_n^{1/2}(\boldsymbol{\gamma}_0)(D_n^{-1/2}(\boldsymbol{\gamma}_0))^T$, and I is a $q \times q$ identity matrix. □

Next we turn to the general model (2.53) where d is arbitrary. With the Assumptions 2.1, 2.2 and 2.3, Theorem 2.5 can be generalized to the case $d > 2$. Compared with the $d = 2$ situation, the IFM for the general model (2.53) is a system of equations (2.58), which do not introduce any complication in terms of the asymptotic properties for IFME. Thus we have the following:

Theorem 2.6 *Consider the general model (2.53) with arbitrary d. Let $\tilde{\boldsymbol{\gamma}}$ denote the IFME of $\boldsymbol{\gamma}$ under the IFM (2.58). Under Assumptions 2.1 and 2.2, $\tilde{\boldsymbol{\gamma}}$ is a consistent estimator of $\boldsymbol{\gamma}$. Furthermore, under Assumptions 2.1, 2.2 and 2.3, as $n \to \infty$, we have*

$$\sqrt{n}A_n^{-1}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{D} N_q(\mathbf{0}, I),$$

where $A_n = D_n^{-1/2}(\boldsymbol{\gamma}_0)M_n^{1/2}(\boldsymbol{\gamma}_0)(D_n^{-1/2}(\boldsymbol{\gamma}_0))^T$, with $D_n(\boldsymbol{\gamma}_0)$ and $M_n(\boldsymbol{\gamma}_0)$ are defined in (2.56) □

We now calculate the matrix $M_n(\boldsymbol{\gamma})$ and $D_n(\boldsymbol{\gamma})$ (or just part of the matrices, depending on the need) corresponding to the IFM for $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$. For example, suppose α_{jl} is a parameter appearing in $P_j(y_{ij})$ and α_{km} is a parameter appearing in $P_k(y_{ik})$. Then the element of the matrix $M_n(\boldsymbol{\gamma})$ corresponding to the parameters α_{jl} and α_{km} can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})P_k(y_{ik})} \frac{\partial P_j(y_{ij})}{\partial \alpha_{jl}} \frac{\partial P_k(y_{ik})}{\partial \alpha_{km}}\bigg|_{\tilde{\boldsymbol{\alpha}}_j \tilde{\boldsymbol{\alpha}}_k}, \tag{2.67}$$

where j may equal to k.

If α_{jl} is a parameter appearing in $P_j(y_{ij})$, and β_{kms} is a parameter appearing in $P_{km}(y_{ik}y_{im})$, then the element of the matrix $M_n(\boldsymbol{\gamma})$ corresponding to the parameters α_{jl} and β_{kms} can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})P_{km}(y_{ik}y_{im})} \frac{\partial P_j(y_{ij})}{\partial \alpha_{jl}} \frac{\partial P_{km}(y_{ik}y_{im})}{\partial \beta_{kms}}\bigg|_{\tilde{\boldsymbol{\alpha}}_j \tilde{\boldsymbol{\alpha}}_k \tilde{\boldsymbol{\alpha}}_m \tilde{\boldsymbol{\beta}}_{km}}, \tag{2.68}$$

where $k < m$ and j may equal to k or m. Furthermore, if β_{jks} is a parameter appearing in $P_{jk}(y_{ij}y_{ik})$ and β_{lmt} a parameter appearing in $P_{lm}(y_{il}y_{im})$, then the element of the matrix $M_n(\boldsymbol{\gamma})$ corresponding to the parameters β_{jks} and β_{lmt} can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_{jk}(y_{ij}y_{ik})P_{lm}(y_{il}y_{im})}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\beta_{jks}}\frac{\partial P_{lm}(y_{il}y_{im})}{\partial\beta_{lmt}}\Bigg|_{\tilde{\boldsymbol{\alpha}}_j\tilde{\boldsymbol{\alpha}}_k\tilde{\boldsymbol{\alpha}}_l\tilde{\boldsymbol{\alpha}}_m\tilde{\boldsymbol{\beta}}_{jk}\tilde{\boldsymbol{\beta}}_{lm}}, \tag{2.69}$$

where $j < k$, $l < m$ and $(j,k) = (l,m)$ is allowed.

For the elements of the matrix $D_n(\boldsymbol{\gamma})$, suppose α_{jl} and α_{jm} are parameters appearing in $P_j(y_{ij})$. Then the element of the matrix $D_n(\boldsymbol{\gamma})$ corresponding to the expression $\partial\Psi_{njl}(\alpha_{jl})/\partial\alpha_{jm}$ can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_j^2(y_{ij})}\frac{\partial P_j(y_{ij})}{\partial\alpha_{jl}}\frac{\partial P_j(y_{ij})}{\partial\alpha_{jm}}\Bigg|_{\tilde{\boldsymbol{\alpha}}_j}. \tag{2.70}$$

If α_{jl} is a parameter appearing in $P_j(y_{ij})$ and β_{jks} is a parameter appearing in $P_{jk}(y_{ij}y_{ik})$, then the element of the matrix $D_n(\boldsymbol{\gamma})$ corresponding to the expression $\partial\Psi_{nj}(\alpha_{jl})/\partial\beta_{jks}$ is 0.

If β_{jks} is a parameter appearing in $P_{jk}(y_{ij}y_{ik})$ and α_l is a parameter corresponding to a univariate margin, then the element of the matrix $D_n(\boldsymbol{\gamma})$ corresponding to the expression $\partial\Psi_{njks}(\beta_{jks})/\partial\alpha_l$ can be estimated by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_{jk}^2(y_{ij}y_{ik})}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\beta_{jks}}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\alpha_l}\Bigg|_{\tilde{\boldsymbol{\alpha}}_j\tilde{\boldsymbol{\alpha}}_k\tilde{\boldsymbol{\beta}}_{jk}}. \tag{2.71}$$

If β_{jks} is a parameter appearing in $P_{jk}(y_{ij}y_{ik})$ and β_t is a parameter corresponding to a bivariate margin, then the element of the matrix $D_n(\boldsymbol{\gamma})$ corresponding to the expression $\partial\Psi_{njks}(\beta_{jks})/\partial\beta_t$ can be estimated by

$$-\frac{1}{n}\sum_{i=1}^{n}\frac{1}{P_{jk}^2(y_{ij}y_{ik})}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\beta_{jks}}\frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial\beta_t}\Bigg|_{\tilde{\boldsymbol{\alpha}}_j\tilde{\boldsymbol{\alpha}}_k\tilde{\boldsymbol{\beta}}_{jk}}. \tag{2.72}$$

However, as in section 2.5 and the data analysis examples in Chapter 5, it is easier to use the jackknife technique to estimate $M_n(\boldsymbol{\gamma})$ and $D_n(\boldsymbol{\gamma})$.

2.4.3 Asymptotic results for the models assuming a joint distribution for response vector and covariates

The asymptotic developments in subsection 2.4.2 treat $\mathbf{x}_{i1},\ldots,\mathbf{x}_{id}$ and $\mathbf{w}_{i12},\ldots,\mathbf{w}_{id-1,d}$ as known constant vectors. An alternative would be to consider the covariates as marginal stochastic outcomes of a vector $\mathbf{V}_i$, and to consider the distribution of the random vector formed by the response vector

together with the covariate vector. Then the model (2.53) can be interpreted as a conditional model, *i.e.*, (2.53) is the conditional mass function given the covariate vectors, where the $\mathbf{Y}_i$ are conditionally independent. Specifically, let

$$\mathbf{Z}_i = \begin{pmatrix} \mathbf{Y}_i \\ \mathbf{V}_i \end{pmatrix}, \quad i = 1, \ldots, n,$$

be iid with distribution $F_{\boldsymbol{\delta}}$ belonging to a family $\mathcal{F} = \{F_{\boldsymbol{\delta}}, \boldsymbol{\delta} \in \Delta\}$. Suppose that the distributions $F_{\boldsymbol{\delta}}$ possess densities or mass functions (or the mixture of density functions and mass functions) with support $\mathcal{Z}$; denote these functions by $f(\mathbf{z}; \boldsymbol{\delta})$. Let $\boldsymbol{\delta} = (\boldsymbol{\gamma}', \boldsymbol{\eta}')'$, where $\boldsymbol{\gamma} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_d', \boldsymbol{\beta}_{12}', \ldots, \boldsymbol{\beta}_{d-1,d}')'$. Assume that the conditional distribution of $\mathbf{Y}_i$ given $\mathbf{V}_i = \mathbf{v}_i$,

$$P(\mathbf{y}_i; \boldsymbol{\theta}_i | \mathbf{V}_i = \mathbf{v}_i), \tag{2.73}$$

is as in (2.53), that is $P(\mathbf{y}_i; \boldsymbol{\theta}_i | \mathbf{V}_i = \mathbf{v}_i)$ is a MCD or MMD model with univariate and bivariate expressible parameters (PUBE)

$$\boldsymbol{\theta}_i = \left(\theta_1(\boldsymbol{\alpha}_1', \mathbf{v}_i), \ldots, \theta_d(\boldsymbol{\alpha}_d', \mathbf{v}_i), \theta_{12}(\boldsymbol{\beta}_{12}', \mathbf{v}_i), \ldots, \theta_{d-1,d}(\boldsymbol{\beta}_{d-1,d}', \mathbf{v}_i)\right),$$

where $\boldsymbol{\theta}_i$ is assumed to be completely independent of $\boldsymbol{\eta}$ (defined below), and is a function of $\boldsymbol{\gamma}$. We also assume that θ_j $(j = 1, \ldots, d)$ is the parameter for the jth univariate margin $P_j(y_{ij})$ of $\mathbf{Y}_i$ from the conditional distribution (2.73), such that $P_j(y_{ij})$ depends only on the parameter θ_j, and θ_{jk} $(1 \leq j < k \leq d)$ is the parameter for the (j,k) bivariate margin $P_{jk}(y_{ij}y_{ik})$ of $\mathbf{Y}_i$ from the conditional distribution (2.73) such that θ_{jk} is the only bivariate parameter. In contrast to (2.54), here we only need to assume that θ_j is a function of $\boldsymbol{\alpha}_j$ and $\mathbf{v}$, $j = 1, 2, \ldots, d$, and θ_{jk} is a function of $\boldsymbol{\beta}_{jk}$ and $\mathbf{v}$, $1 \leq j < k \leq d$, without explicitly imposing the type of link functions in (2.54).

The marginal distribution of $\mathbf{V}_i$ is assumed to depend only on the parameter vector $\boldsymbol{\eta}$, which is treated as a nuisance parameter vector. Its density or mass function (or the mixture of the two) is denoted by $g_j(\mathbf{v}_i; \boldsymbol{\eta})$. Thus

$$f(\mathbf{z}_i; \boldsymbol{\delta}) = P(\mathbf{y}_i; \boldsymbol{\theta}_i | \mathbf{V}_i = \mathbf{v}_i) g_j(\mathbf{v}_i; \boldsymbol{\eta}). \tag{2.74}$$

Under this framework, weak assumptions on the support of $\mathbf{V}_i$, in particular not requiring any special feature of the link function, are sufficient for consistency and asymptotic normality of the IFME in the conditional sense.

Let the density of $\mathbf{Z} = (\mathbf{Y}', \mathbf{V}')'$ be $f(\mathbf{z}; \boldsymbol{\delta})$, and that of $\mathbf{V}$ be $g_j(\mathbf{v}; \boldsymbol{\eta})$. Based on our assumptions on $P(\mathbf{y}_i; \boldsymbol{\theta}_i | \mathbf{V}_i = \mathbf{v}_i)$, we see that the joint marginal distribution of Y_j and $\mathbf{V}$ is

$$P_{j,\mathbf{V}} = P_j(y_j | \mathbf{V} = \mathbf{v}) g_j(\mathbf{v}; \boldsymbol{\eta})$$

and the joint marginal distribution of Y_j, Y_k and $\mathbf{V}$ is

$$P_{jk,\mathbf{v}} = P_{jk}(y_j y_k|\mathbf{V}=\mathbf{v})g_j(\mathbf{v};\boldsymbol{\eta}).$$

For notational simplicity, in the following, we simply write $P_j(y_j)$ and $P_{jk}(y_j y_k)$ in lieu of $P_j(y_j|\mathbf{V}=\mathbf{v})$ and $P_{jk}(y_j y_k|\mathbf{V}=\mathbf{v})$. The corresponding marginal distributions for $\mathbf{Z}_i$ are

$$\begin{cases} P_{j,\mathbf{v}_i} = P_j(y_{ij})g_j(\mathbf{v}_i;\boldsymbol{\eta}), \\ P_{jk,\mathbf{v}_i} = P_{jk}(y_{ij}y_{ik})g_j(\mathbf{v}_i;\boldsymbol{\eta}). \end{cases}$$

Thus we obtain a set of loglikelihood functions of margins for $\boldsymbol{\gamma}$

$$\begin{cases} \ell_{nj} = \sum_{i=1}^{n} \log P_{j,\mathbf{v}_i} = \sum_{i=1}^{n} \log P_j(y_{ij}) + \sum_{i=1}^{n} g_j(\mathbf{v}_i;\boldsymbol{\eta}), \quad j=1,\ldots,d, \\ \ell_{njk} = \sum_{i=1}^{n} \log P_{jk,\mathbf{v}_i} = \sum_{i=1}^{n} \log P_{jk}(y_{ij}y_{ik}) + \sum_{i=1}^{n} g_j(\mathbf{v}_i;\boldsymbol{\eta}), \quad 1\le j<k\le d. \end{cases} \tag{2.75}$$

Let

$$\begin{cases} \omega_{j,s} = \omega_{j,s}(\alpha_{js}) \stackrel{\text{def}}{=} \dfrac{1}{P_{j,\mathbf{v}}}\dfrac{\partial P_{j,\mathbf{v}}}{\partial \alpha_{js}}, \quad j=1,\ldots,d;\ s=1,\ldots,p_j, \\ \omega_{jk,t} = \omega_{jk,t}(\beta_{jk,t}) \stackrel{\text{def}}{=} \dfrac{1}{P_{jk,\mathbf{v}}}\dfrac{\partial P_{jk,\mathbf{v}}}{\partial \beta_{jk,t}}, \quad 1\le j<k\le d;\ t=1,\ldots,q_{jk}, \end{cases}$$

and

$$\begin{cases} \omega_{i;j,s} = \omega_{i;j,s}(\alpha_{js}) \stackrel{\text{def}}{=} \dfrac{1}{P_{j,\mathbf{v}_i}}\dfrac{\partial P_{j,\mathbf{v}_i}}{\partial \alpha_{js}}, \quad j=1,\ldots,d;\ s=1,\ldots,p_j, \\ \omega_{i;jk,t} = \omega_{i;jk,t}(\beta_{jkt}) \stackrel{\text{def}}{=} \dfrac{1}{P_{jk,\mathbf{v}_i}}\dfrac{\partial P_{jk,\mathbf{v}_i}}{\partial \beta_{jkt}}, \quad 1\le j<k\le d;\ t=1,\ldots,q_{jk}, \end{cases}$$

for $i=1,\ldots,n$. From (2.75), we derive the IFM for $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$

$$\begin{cases} \Omega_{nj,s} = \sum_{i=1}^{n}\omega_{i;j,s} = \sum_{i=1}^{n}\dfrac{1}{P_{j,\mathbf{v}_i}}\dfrac{\partial P_{j,\mathbf{v}_i}}{\partial \alpha_{js}} = \sum_{i=1}^{n}\dfrac{1}{P_j(y_{ij})}\dfrac{\partial P_j(y_{ij})}{\partial \alpha_{js}} = \sum_{i=1}^{n}\psi_{i;js},\ j=1,\ldots,d;\ s=1,\ldots,p_j; \\ \Omega_{njk,t} = \sum_{i=1}^{n}\omega_{i;jk,t} = \sum_{i=1}^{n}\dfrac{1}{P_{jk,\mathbf{v}_i}}\dfrac{\partial P_{jk,\mathbf{v}_i}}{\partial \beta_{jkt}} = \sum_{i=1}^{n}\dfrac{1}{P_{jk}(y_{ij}y_{ik})}\dfrac{\partial P_{jk}(y_{ij}y_{ik})}{\partial \beta_{jkt}} = \sum_{i=1}^{n}\psi_{i;jkt}, \\ \qquad\qquad 1\le j<k\le d;\ t=1,\ldots,q_{jk}, \end{cases} \tag{2.76}$$

and the estimating equations for $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$ based on IFM are

$$\begin{cases} \Omega_{nj,s} = 0, \quad j=1,\ldots,d;\ s=1,\ldots,p_j, \\ \Omega_{njk,t} = 0, \quad 1\le j<k\le d;\ t=1,\ldots,q_{jk}. \end{cases} \tag{2.77}$$

With the IFM approach, estimates of $\boldsymbol{\alpha}_j, j=1,\ldots,d$, and $\boldsymbol{\beta}_{jk}$, $1\le j<k\le d$, denoted by $\tilde{\boldsymbol{\alpha}}_j = (\tilde{\alpha}_{j1},\ldots,\tilde{\alpha}_{j,p_j})'$ and $\tilde{\boldsymbol{\beta}}_{jk} = (\tilde{\beta}_{jk1},\ldots,\tilde{\beta}_{jk,q_{jk}})'$, are obtained by solving the nonlinear system

of equations (2.77). Note that (2.77) is computationally equivalent to (2.57); they both lead to the same numerical solutions (assuming the link functions given in (2.54)).

Let $\Omega_{nj} = (\Omega_{nj,1}, \ldots, \Omega_{nj,p_j})'$ and $\Omega_{njk} = (\Omega_{njk,1}, \ldots, \Omega_{njk,q_{jk}})'$. Then (2.76) can be rewritten in function vector form

$$\begin{cases} \Omega_{nj} = \mathbf{0}, & j = 1, \ldots, d, \\ \Omega_{njk} = \mathbf{0}, & 1 \le j < k \le d. \end{cases}$$

Let $\Omega_j = (\omega_{j,1}, \ldots, \omega_{j,p_j})'$ and $\Omega_{jk} = (\omega_{jk,1}, \ldots, \omega_{jk,q_{jk}})'$. Let $\Omega = (\Omega_1', \ldots, \Omega_d', \Omega_{12}', \ldots, \Omega_{d-1,d}')'$. Let $M_\Omega = \mathrm{E}(\Omega\Omega')$, $D_\Omega = \partial\Omega/\partial\boldsymbol{\gamma}$. Under some regularity conditions similar to subsection 2.4.1, consistency and asymptotic normality for $\tilde{\boldsymbol{\gamma}}$ can be established.

Basically, the assumptions that we need are those making Ω a regular inference function vector.

Assumptions 2.3

1. *The support of* $\mathbf{Z}$, $\mathcal{Z}$ *does not depend on any* $\boldsymbol{\delta} \in \Delta$.

2. $E_{\boldsymbol{\delta}}\{\Omega\} = \mathbf{0}$;

3. *The partial derivative* $\partial\Omega/\partial\boldsymbol{\gamma}$ *exists for almost every* $\mathbf{z} \in \mathcal{Z}$;

4. *Assume* Ω *and* $\boldsymbol{\gamma}$ *are* $q \times 1$ *vectors respectively, and their components have subindex* $j = 1, \ldots, q$. *The order of integration and differentiation may be interchanged as follows:*

$$\frac{\partial}{\partial\beta_j}\int_{\mathcal{Z}} \omega_j f(\mathbf{z};\boldsymbol{\delta}) d\mu(\mathbf{z}) = \int_{\mathcal{Z}} \frac{\partial}{\partial\beta_j}\left[\omega_j f(\mathbf{z};\boldsymbol{\delta})\right] d\mu(\mathbf{z});$$

5. $E_{\boldsymbol{\delta}}\{\Omega\Omega'\} \in I\!R^{q\times q}$ *and the* $q \times q$ *matrix*

$$M_\Omega = E_{\boldsymbol{\delta}}\{\Omega\Omega'\}$$

is positive-definite.

6. *The* $q \times q$ *matrix* $D_\Omega = \partial\Omega(\boldsymbol{\theta})/\partial\boldsymbol{\gamma}$ *is non-singular.*

□

Assumptions 2.3 are equivalent to assuming that $M_n(\boldsymbol{\gamma})$ in (2.56) is positive-definite and $D_n(\boldsymbol{\gamma})$ in (2.56) is non-singular for certain $n \ge n_0$, where n_0 is a fixed integer.

We have the following theorem

Theorem 2.7 *Consider the model (2.73) and let $\tilde{\boldsymbol{\gamma}}$ denote the IFME of $\boldsymbol{\gamma}$ under the IFM (2.77). Under Assumptions 2.3, $\tilde{\boldsymbol{\gamma}}$ is a consistent estimator of $\boldsymbol{\gamma}$. Furthermore, as $n \to \infty$, we have asymptotically*

$$\sqrt{n}(\tilde{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_\Omega^{-1}),$$

where $J_\Omega = D_\Omega' M_\Omega^{-1} D_\Omega$.

Proof: Under the model (2.74) and the Assumptions 2.3, the proof is similar to that of Theorem 2.3 and 2.4. □

We believe this approach for deriving asymptotic properties of an estimate has not appeared in the statistical literature. The assumptions are suitable for an observational study but not an experimental study in which one has control of the **v**'s.

Theorem 2.7 is different from Theorem 2.6 in that M_Ω and D_Ω both depend on the distribution function of **V**. Nevertheless, because $\omega_{i;j,s} = \psi_{i;j,s}$ and $\omega_{i;jk,t} = \psi_{i;jk,t}$, the numerical evaluation of M_Ω and $M_n(\boldsymbol{\gamma})$ in (2.59) based on data are the same because only the empirical distribution for Ψ is needed. For example, suppose α_l is a parameter of $P_j(y_{ij})$ and α_m a parameter of $P_k(y_{ik})$, then the element of the matrix M_Ω corresponding to the parameters α_l and α_m can be estimated by

$$\frac{1}{n}\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})P_k(y_{ik})} \frac{\partial P_j(y_{ij})}{\partial \alpha_l} \frac{\partial P_k(y_{ik})}{\partial \alpha_m}\bigg|_{\tilde{\boldsymbol{\alpha}}_j \tilde{\boldsymbol{\alpha}}_k},$$

which is the same as (2.67). We can similarly obtain (2.68) and (2.69). The same result is true for D_Ω versus $D_n(\boldsymbol{\gamma})$ in (2.56); they both lead to the same numerical results based on the data. We thus derive the same formulas (2.70)-(2.72) for numerical evaluation of D_Ω.

2.5 The Jackknife approach for the variance of IFME

The calculation of the Godambe information matrix based on regular IFM for the models (2.39) and (2.53) is straightforward in terms of symbolic representation. However, the actual computation of the Godambe information matrix requires many derivatives of first and second order, and in terms of computer implementation, considerable programming effort would be required. With this consideration, an alternative jackknife procedure for the calculation of the IFME asymptotic variance is developed. The jackknife idea is simple, but very useful, especially in cases where the analytical answer is very difficult to obtain or computationally very complex. This procedure has the advantage of general computational feasibility.

In this section, we show that our jackknife method for calculating the corresponding asymptotic variance matrix of $\tilde{\boldsymbol{\theta}}$ is asymptotically equivalent to the Godambe information matrix. We examine the situation for models with covariates and with no covariates. Our main interest in using the jackknife is to obtain the SE of an estimate and not for bias correction (because for multivariate statistical inference the data set cannot be small), though several results about jackknife parameter estimation are also given.

The jackknife estimate of variance may be preferred when the appropriate computer code is not available to compute the Godambe information matrix or there are other complications such as the need to calculate the asymptotic variance of a function of an estimator. Some numerical comparisons of the Godambe information and the jackknife variance estimate based on simulations are reported in Chapter 4. The jackknife procedure is demonstrated to be satisfactory. Some general references about jackknife methods are Quenouille (1956), Miller (1974), Efron and Stein (1981), and Efron (1982), among others. A recent reference on the use of jackknife estimators of variance for parameter estimates from estimating equations is Lipsitz *et al.* (1994), though their one-step jackknife estimator is not as general as what we are studying here, and their main application is to clustered survival data.

In the following, we assume that we can partition the n observations into g groups with m observations each so that $n = m \times g$, m is an integer. We discuss two situations for applying jackknife idea: leaving out one observation at a time and leaving out more than one observation at a time.

2.5.1 Jackknife approach for models with no covariates

Let $\mathbf{Y}, \mathbf{Y}_1, \ldots, \mathbf{Y}_n$ be iid rv's from a regular discrete model

$$P(y_1 \cdots y_d; \boldsymbol{\theta}),\ \boldsymbol{\theta} \in \Re$$

in (2.12), and $\mathbf{y}, \mathbf{y}_1, \ldots, \mathbf{y}_n$ be their observed values respectively. Let $\Psi(\boldsymbol{\theta}) = (\psi_1(\boldsymbol{\theta}), \ldots, \psi_q(\boldsymbol{\theta}))$ be the IFM based on $\mathbf{y}$, $\Psi_i(\boldsymbol{\theta}) = (\psi_{i;1}(\boldsymbol{\theta}), \ldots, \psi_{i;q}(\boldsymbol{\theta}))$ be the IFM based on $\mathbf{y}_i$, and $\Psi_n(\boldsymbol{\theta}) = (\Psi_{n1}(\boldsymbol{\theta}), \ldots, \Psi_{nq}(\boldsymbol{\theta}))$ be the IFM based on $\mathbf{y}_1, \ldots, \mathbf{y}_n$, where $\Psi_{nj}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \psi_{i;j}(\boldsymbol{\theta})$ $(j = 1, \ldots, q)$. Let $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}_1, \ldots, \tilde{\theta}_q)$ be the estimate of $\boldsymbol{\theta}$ from $\Psi_n(\boldsymbol{\theta}) = \mathbf{0}$.

Leave out one observation at a time

Let $\tilde{\boldsymbol{\theta}}_{(i)}$ be an estimate of $\boldsymbol{\theta}$ based on the same set of inference functions Ψ_n but with the ith observation $\mathbf{y}_i$ from the data set $\mathbf{y}_1, \ldots, \mathbf{y}_n$ deleted, $i = 1, \ldots, n$. In this situation, we have $m = 1$

and $g = n$. That is, we delete one group of size 1 each time and calculate the same estimate of $\boldsymbol{\theta}$ based on the remaining $n-1$ observations. Let $\tilde{\boldsymbol{\theta}}_i = n\tilde{\boldsymbol{\theta}} - (n-1)\tilde{\boldsymbol{\theta}}_{(i)}$, and $\tilde{\boldsymbol{\theta}}_{(\cdot)} = \sum_{i=1}^{n} \tilde{\boldsymbol{\theta}}_{(i)}/n$. $\tilde{\boldsymbol{\theta}}_i$ are called "pseudo-values" in the literature. The *jackknife estimate* of $\boldsymbol{\theta}$ is defined as

$$\tilde{\boldsymbol{\theta}}_J \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} \tilde{\boldsymbol{\theta}}_i = n\tilde{\boldsymbol{\theta}} - (n-1)\tilde{\boldsymbol{\theta}}_{(\cdot)}. \tag{2.78}$$

The jackknife estimator has the property that it eliminates the order $1/n$ term from a bias of the form $E(\tilde{\boldsymbol{\theta}}) = \boldsymbol{\theta} + \mathbf{u}_1/n + \mathbf{u}_2/n^2 + \cdots$, where the functions $\mathbf{u}_1, \mathbf{u}_2, \ldots$ do not depend upon n (see Miller 1974). In fact, the jackknife statistic $\tilde{\boldsymbol{\theta}}_J$ often has less bias than the original statistic $\tilde{\boldsymbol{\theta}}$.

The early version of the *jackknife estimate of variance* for the jackknife estimator $\tilde{\boldsymbol{\theta}}_J$ was suggested by Tukey (1958). It is defined as

$$V_J = \frac{1}{n(n-1)}\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}_J)(\tilde{\boldsymbol{\theta}}_i - \tilde{\boldsymbol{\theta}}_J)^T = \frac{n-1}{n}\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}}_{(\cdot)})(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}}_{(\cdot)})^T. \tag{2.79}$$

In (2.79), the pseudo-values $\tilde{\boldsymbol{\theta}}_i$ are treated as if they are independent and identically distributed, disregarding the fact that the pseudo-values $\tilde{\boldsymbol{\theta}}_i$ actually depend on the common observations. To justify (2.79), Thorburn (1976) proved that under rather natural conditions on the original statistic $\tilde{\boldsymbol{\theta}}$, the pseudo-values are asymptotically uncorrelated.

In practice, the pseudo-values have been used to estimate the variance not only of the jackknife estimate $\tilde{\boldsymbol{\theta}}_J$, but also $\tilde{\boldsymbol{\theta}}$. But if the bias correction in jackknife estimate is not needed, we propose to use a simpler estimate of the asymptotic variance (matrix) of $\tilde{\boldsymbol{\theta}}$:

$$V_J = \sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})^T. \tag{2.80}$$

In our context, unless stated otherwise, we always call V_J defined by (2.80) the *jackknife estimate of variance*. In the following, we first prove that the asymptotic distribution of a jackknife statistic $\tilde{\boldsymbol{\theta}}_J$ is the same as the original statistic $\tilde{\boldsymbol{\theta}}$ under the same model assumptions; subsequently, we prove that V_J defined by (2.80) is a consistent estimator of inverse of the Godambe information matrix $J_\Psi(\boldsymbol{\theta})$.

Theorem 2.8 *Under the same assumptions as in Theorem 2.4, the jackknife estimate $\tilde{\boldsymbol{\theta}}_J$ in (2.78) has the same asymptotic distribution as $\tilde{\boldsymbol{\theta}}$. That is, as $n \to \infty$,*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) \xrightarrow{\text{D}} N_q(\mathbf{0}, J_\Psi^{-1}(\boldsymbol{\theta})),$$

where $J_\Psi(\boldsymbol{\theta}) = D_\Psi^T(\boldsymbol{\theta})M_\Psi^{-1}(\boldsymbol{\theta})D_\Psi(\boldsymbol{\theta})$, with $M_\Psi(\boldsymbol{\theta}) = E\{\Psi(\boldsymbol{\theta})\Psi^T(\boldsymbol{\theta})\}$ and $D_\Psi(\boldsymbol{\theta}) = \partial\Psi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

Proof: We sketch the proof. For $\Psi_n(\boldsymbol{\theta}) = (\Psi_{n1}(\boldsymbol{\theta}), \ldots, \Psi_{nq}(\boldsymbol{\theta})) : \mathcal{Y}^n \times \Re \to I\!R^q$ has the following expansion around $\boldsymbol{\theta}$

$$\mathbf{0} = \Psi_n(\tilde{\boldsymbol{\theta}}) = \Psi_n(\boldsymbol{\theta}) + H_n(\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) + \mathrm{R}_n,$$

where $H_n(\boldsymbol{\theta}) = \partial\Psi_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is a $q \times q$ matrix and $\mathrm{R_n} = O_p(||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2) = O_p(n^{-1})$ by assumptions. Thus

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) = \left(\frac{1}{n}H_n(\boldsymbol{\theta})\right)^{-1} \frac{1}{\sqrt{n}}\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right).$$

Let $\Psi_{(i)}(\boldsymbol{\theta})$ be $\Psi_n(\boldsymbol{\theta})$ calculated without the ith observation, and $H_{(i)}(\boldsymbol{\theta})$ be the $q \times q$ matrix $\partial\Psi_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ calculated without the ith observation. Similarly, we have

$$\sqrt{n-1}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}) = \left(\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta})\right)^{-1} \frac{1}{\sqrt{n-1}}\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right),$$

where $\mathrm{R}_{n-1,i} = O_p(||\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}||^2) = o_p(n^{-1})$.

Since

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}) = n\left(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta})\right) + \sqrt{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}),$$

we have

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) &= n\left(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n}\sqrt{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta})\right) + \frac{1}{n}\sum_{i=1}^{n}\sqrt{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}) \\ &= n\left(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}\sqrt{n-1}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta})\right) + \sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}\sqrt{n-1}(\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}).\end{aligned}$$

Thus

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) = n\Bigg[&\left(\frac{1}{n}H_n(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n}}\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right) - \\ &\sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-1}}\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right)\right\}\Bigg] + \\ &\sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-1}}\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right)\right\}.\end{aligned}$$

By the Law of Large Numbers, we have

$$\frac{1}{n}H_n(\boldsymbol{\theta}) \xrightarrow{p} D_\Psi(\boldsymbol{\theta}),$$

and

$$\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta}) \xrightarrow{p} D_\Psi(\boldsymbol{\theta}).$$

From the central limit theorem,

$$\frac{1}{\sqrt{n}}\Psi_n(\boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, M_\Psi).$$

We further have

$$\sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}-\frac{1}{\sqrt{n-1}}\Psi_{(i)}(\boldsymbol{\theta})=-\frac{1}{\sqrt{n}}\Psi_n(\boldsymbol{\theta}).$$

This, together with the $\sqrt{n}$-consistency of $\tilde{\boldsymbol{\theta}}$ (Theorem 2.4), lead to

$$n\left[\left(\frac{1}{n}H_n(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n}}\left(-\Psi_n(\boldsymbol{\theta})-\mathrm{R}_n\right)\right.$$
$$\left.-\sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-1}}\left(-\Psi_{(i)}(\boldsymbol{\theta})-\mathrm{R}_{n-1}\right)\right\}\right]\xrightarrow{\mathrm{p}}0$$

and

$$\sqrt{\frac{n}{n-1}}\frac{1}{n}\sum_{i=1}^{n}\left\{\left(\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-1}}\left(-\Psi_{(i)}(\boldsymbol{\theta})-\mathrm{R}_{n-1}\right)\right\}\xrightarrow{\mathrm{D}}N_q(\mathbf{0},D_{\Psi}^{-1}M_{\Psi}(D_{\Psi}^{-1})^T).$$

Thus by applying Slutsky's Theorem, we obtain

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_J-\boldsymbol{\theta})\xrightarrow{\mathrm{D}}N_q(\mathbf{0},J_{\Psi}^{-1}(\boldsymbol{\theta})).$$

□

Theorem 2.9 *Under the same assumptions as in Theorem 2.4, the sample size n times the jackknife estimate of variance V_J defined by (2.80) is a consistent estimator of $J_{\Psi}^{-1}(\boldsymbol{\theta})$.*

Proof. We have

$$\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)}-\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(i)}-\tilde{\boldsymbol{\theta}})^T=\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)}-\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}_{(i)}-\boldsymbol{\theta})^T$$
$$-\left[\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)}-\boldsymbol{\theta})\right](\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})^T-(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\left[\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)}-\boldsymbol{\theta})\right]^T+n(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})^T.$$

Recall that

$$\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta}=H_n^{-1}(\boldsymbol{\theta})\left(-\Psi_n(\boldsymbol{\theta})-\mathrm{R}_n\right),$$

and

$$\tilde{\boldsymbol{\theta}}_{(i)}-\boldsymbol{\theta}=H_{(i)}^{-1}(\boldsymbol{\theta})\left(-\Psi_{(i)}(\boldsymbol{\theta})-\mathrm{R}_{n-1,i}\right),$$

where $\mathrm{R}_n = O_p(||\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}||^2) = O_p(n^{-1})$ and $\mathrm{R}_{n-1,i} = O_p(||\tilde{\boldsymbol{\theta}}_{(i)} - \boldsymbol{\theta}||^2) = O_p(n^{-1})$. Thus

$$\begin{aligned}\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})^T = &\sum_{i=1}^{n} H_{(i)}^{-1}(\boldsymbol{\theta})\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right)\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right)^T \left(H_{(i)}^{-1}(\boldsymbol{\theta})\right)^T \\ &- \left[\sum_{i=1}^{n} H_{(i)}^{-1}(\boldsymbol{\theta})\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right)\right]\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right)^T \left(H_n^{-1}(\boldsymbol{\theta})\right)^T \\ &- H_n^{-1}(\boldsymbol{\theta})\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right)\left[\sum_{i=1}^{n} H_{(i)}^{-1}(\boldsymbol{\theta})\left(-\Psi_{(i)}(\boldsymbol{\theta}) - \mathrm{R}_{n-1,i}\right)\right]^T \\ &+ nH_n^{-1}(\boldsymbol{\theta})\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right)\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right)^T \left(H_n^{-1}(\boldsymbol{\theta})\right)^T.\end{aligned} \tag{2.81}$$

As $\Psi_{(i)}(\boldsymbol{\theta}) = \Psi_n(\boldsymbol{\theta}) - \Psi_{i;}(\boldsymbol{\theta})$, thus $\sum_{i=1}^n \Psi_{(i)}(\boldsymbol{\theta}) = (n-1)\Psi_n(\boldsymbol{\theta})$ and $\sum_{i=1}^n \Psi_{(i)}(\boldsymbol{\theta})\Psi_{(i)}^T(\boldsymbol{\theta}) = (n-2)\Psi_n(\boldsymbol{\theta})\Psi_n^T(\boldsymbol{\theta}) + \sum_{i=1}^n \Psi_{i;}(\boldsymbol{\theta})\Psi_{i;}^T(\boldsymbol{\theta})$. By the Law of Large Numbers, we have

$$\frac{1}{n}H_n(\boldsymbol{\theta}) \xrightarrow{p} D_\Psi(\boldsymbol{\theta}),$$
$$\frac{1}{n-1}H_{(i)}(\boldsymbol{\theta}) \xrightarrow{p} D_\Psi(\boldsymbol{\theta})$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\Psi_{i;}(\boldsymbol{\theta})\Psi_{i;}^T(\boldsymbol{\theta}) \xrightarrow{p} E(\Psi(\boldsymbol{\theta})\Psi^T(\boldsymbol{\theta})).$$

From (2.81), we have

$$\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})^T - \sum_{i=1}^{n} H_{(i)}^{-1}(\boldsymbol{\theta})\Psi_{i;}(\boldsymbol{\theta})\Psi_{i;}(\boldsymbol{\theta})^T \left(H_{(i)}^{-1}(\boldsymbol{\theta})\right)^T \xrightarrow{p} \mathbf{0}$$

and this implies that

$$n\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(i)} - \tilde{\boldsymbol{\theta}})^T \xrightarrow{p} D_\Psi^{-1}(\boldsymbol{\theta})M_\Psi(\boldsymbol{\theta})(D_\Psi(\boldsymbol{\theta})^{-1})^T.$$

In other words, we proved that nV_J is a consistent estimator of $J_\Psi^{-1}(\boldsymbol{\theta})$. □

Leave out more than one observation at a time

Now for general $g > 1$, we assume a random subdivision of $\mathbf{y}_1, \ldots, \mathbf{y}_n$ into g groups ($n = gm$). Let $\tilde{\boldsymbol{\theta}}_{(\nu)}$ ($\nu = 1, \ldots, g$) be an estimate of $\boldsymbol{\theta}$ based on the same set of inference functions Ψ from the data set $\mathbf{y}_1, \ldots, \mathbf{y}_n$ but deleting the ν-th group of size m (m is fixed), thus $\tilde{\boldsymbol{\theta}}_{(\nu)}$ is calculated based on a subsample of size $m(g-1)$. The *jackknife estimate* of $\boldsymbol{\theta}$ in this general setting is the mean of $\tilde{\boldsymbol{\theta}}_\nu$, which is

$$\tilde{\boldsymbol{\theta}}_J = \frac{1}{g}\sum_{\nu=1}^{g}\tilde{\boldsymbol{\theta}}_\nu = g\tilde{\boldsymbol{\theta}} - (g-1)\tilde{\boldsymbol{\theta}}_{(\cdot)}, \tag{2.82}$$

where $\tilde{\boldsymbol{\theta}}_{(\cdot)} = g^{-1}\sum_{\nu=1}^{g}\tilde{\boldsymbol{\theta}}_{(\nu)}$, and $\tilde{\boldsymbol{\theta}}_{\nu} = g\tilde{\boldsymbol{\theta}} - (g-1)\tilde{\boldsymbol{\theta}}_{(\nu)}$ $(\nu = 1, \ldots, g)$ are the pseudo-values.

In this situation, the *jackknife estimate of variance* for $\tilde{\boldsymbol{\theta}}$, V_J, is defined as

$$V_J = \sum_{\nu=1}^{g}\left(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}}\right)\left(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}}\right)^T. \tag{2.83}$$

Theorem 2.8 and 2.9 can be easily generalized to the situation with fixed $m > 1$.

Theorem 2.10 *Under the same assumptions as in Theorem 2.4, the jackknife estimate $\tilde{\boldsymbol{\theta}}_J$ defined by (2.82) with m fixed has the same asymptotic distribution as $\tilde{\boldsymbol{\theta}}$. That is, as $n \to \infty$ (thus $g \to \infty$),*

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_{\Psi}^{-1}(\boldsymbol{\theta})),$$

where $J_{\Psi}(\boldsymbol{\theta}) = D_{\Psi}^T(\boldsymbol{\theta})M_{\Psi}^{-1}(\boldsymbol{\theta})D_{\Psi}(\boldsymbol{\theta})$, with $M_{\Psi}(\boldsymbol{\theta}) = E_{\boldsymbol{\theta}}\{\Psi(\boldsymbol{\theta})\Psi^T(\boldsymbol{\theta})\}$ and $D_{\Psi}(\boldsymbol{\theta}) = \partial\Psi(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$.

Proof: We sketch the proof. Let $\Psi_{(\nu)}(\boldsymbol{\theta})$ be $\Psi_n(\boldsymbol{\theta})$ calculated without the νth group, and $H_{(\nu)}(\boldsymbol{\theta})$ be the $q \times q$ matrix $\partial\Psi_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ calculated without the νth group. We have

$$\sqrt{n-m}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta}) = \left(\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-m}}\left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right),$$

where $\mathrm{R}_{n-m,\nu} = O_p(||\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta}||^2) = O_p(n^{-1})$. Recall that

$$\sqrt{g}(\tilde{\boldsymbol{\theta}}_{\nu} - \boldsymbol{\theta}) = g\left(\sqrt{g}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta})\right) + \sqrt{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta}),$$

we thus have

$$\begin{aligned}\sqrt{g}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) &= g\left(\sqrt{g}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \frac{1}{g}\sum_{\nu=1}^{g}\sqrt{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta})\right) + \frac{1}{g}\sum_{\nu=1}^{g}\sqrt{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta}) \\ &= g\left(\sqrt{g}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\sqrt{g-1}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta})\right) + \sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\sqrt{g-1}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta}),\end{aligned}$$

this implies that

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) = g\left(\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) - \sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\sqrt{n-m}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta})\right) + \sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\sqrt{n-m}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta}).$$

Thus

$$\begin{aligned}\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) = g\Bigg[&\left(\frac{1}{n}H_n(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n}}\left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right) - \\ &\sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\left\{\left(\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-m}}\left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right)\right\}\Bigg] + \\ &\sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\left\{\left(\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-m}}\left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right)\right\}.\end{aligned} \tag{2.84}$$

By the Law of Large Numbers, we have

$$\frac{1}{n}H_n(\boldsymbol{\theta})\xrightarrow{p}D_\Psi(\boldsymbol{\theta}),$$

and

$$\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta})\xrightarrow{p}D_\Psi(\boldsymbol{\theta}).$$

From the central limit theorem,

$$\frac{1}{\sqrt{n}}\Psi(\boldsymbol{\theta})\xrightarrow{\mathrm{D}}N_q(\mathbf{0},M_\Psi).$$

We also have

$$\sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}-\frac{1}{\sqrt{n-m}}\Psi_{(\nu)}(\boldsymbol{\theta})=-\frac{1}{\sqrt{n}}\Psi_n(\boldsymbol{\theta}).$$

This, together with the $\sqrt{n}$-consistency of $\tilde{\boldsymbol{\theta}}$ (Theorem 2.4), lead to

$$g\left[\left(\frac{1}{n}H_n(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n}}\left(-\Psi_n(\boldsymbol{\theta})-\mathrm{R}_n\right)\right.$$
$$\left.-\sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\left\{\left(\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-m}}\left(-\Psi_{(\nu)}(\boldsymbol{\theta})-\mathrm{R}_{n-m,\nu}\right)\right\}\right]\xrightarrow{\mathrm{p}}0$$

and

$$\sqrt{\frac{g}{g-1}}\frac{1}{g}\sum_{\nu=1}^{g}\left\{\left(\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta})\right)^{-1}\frac{1}{\sqrt{n-m}}\left(-\Psi_{(\nu)}(\boldsymbol{\theta})-\mathrm{R}_{n-m,\nu}\right)\right\}\xrightarrow{\mathrm{D}}N_q(\mathbf{0},D_\Psi^{-1}M_\Psi(D_\Psi^{-1})^T).$$

Applying Slutsky's Theorem to (2.84), we obtain

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_J-\boldsymbol{\theta})\xrightarrow{\mathrm{D}}N_q(\mathbf{0},J_\Psi^{-1}(\boldsymbol{\theta})).$$

□

Theorem 2.11 *Under the same assumptions as in Theorem 2.4, the sample size n times the jackknife estimate of variance V_J defined by (2.83) is a consistent estimator of $J_\Psi^{-1}(\boldsymbol{\theta})$, when m is fixed and $g\to\infty$.*

Proof: We use the same notation as in the proof of Theorem 2.10. We have

$$\sum_{\nu=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)}-\tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(\nu)}-\tilde{\boldsymbol{\theta}})^T=\sum_{\nu=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)}-\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}_{(\nu)}-\boldsymbol{\theta})^T$$
$$-\left[\sum_{\nu=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)}-\boldsymbol{\theta})\right](\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})^T-(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})\left[\sum_{\nu=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)}-\boldsymbol{\theta})\right]^T+g(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})(\tilde{\boldsymbol{\theta}}-\boldsymbol{\theta})^T,$$

and

$$\tilde{\boldsymbol{\theta}}_{(\nu)} - \boldsymbol{\theta} = H_{(\nu)}^{-1}(\boldsymbol{\theta}) \left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right).$$

Thus

$$\begin{aligned}\sum_{\nu=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})^T = &\sum_{\nu=1}^{g} H_{(\nu)}^{-1}(\boldsymbol{\theta}) \left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right) \left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right)^T \left(H_{(\nu)}^{-1}(\boldsymbol{\theta})\right)^T \\ &- \left[\sum_{\nu=1}^{g} H_{(\nu)}^{-1}(\boldsymbol{\theta}) \left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right)\right] \left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right)^T \left(H_n^{-1}(\boldsymbol{\theta})\right)^T \\ &- H_n^{-1}(\boldsymbol{\theta}) \left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right) \left[\sum_{\nu=1}^{g} H_{(\nu)}^{-1}(\boldsymbol{\theta}) \left(-\Psi_{(\nu)}(\boldsymbol{\theta}) - \mathrm{R}_{n-m,\nu}\right)\right]^T \\ &+ gH_n^{-1}(\boldsymbol{\theta}) \left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right) \left(-\Psi_n(\boldsymbol{\theta}) - \mathrm{R}_n\right)^T \left(H_n^{-1}(\boldsymbol{\theta})\right)^T. \end{aligned} \tag{2.85}$$

Let $\Psi_\nu^*(\boldsymbol{\theta}) = \Psi_n(\boldsymbol{\theta}) - \Psi_{(\nu)}(\boldsymbol{\theta})$. Then

$$\sum_{\nu=1}^{g} \Psi_{(\nu)}(\boldsymbol{\theta}) = (g-1)\Psi_n(\boldsymbol{\theta})$$

and

$$\sum_{\nu=1}^{g} \Psi_{(\nu)}(\boldsymbol{\theta})\Psi_{(\nu)}^T(\boldsymbol{\theta}) = (g-2)\Psi_n(\boldsymbol{\theta})\Psi_n^T(\boldsymbol{\theta}) + \sum_{\nu=1}^{g} \Psi_\nu^*(\boldsymbol{\theta})(\Psi_\nu^*(\boldsymbol{\theta}))^T.$$

By the Law of Large Numbers, we have

$$\begin{aligned}&\frac{1}{n}H_n(\boldsymbol{\theta}) \xrightarrow{p} D_\Psi(\boldsymbol{\theta}), \\ &\frac{1}{n-m}H_{(\nu)}(\boldsymbol{\theta}) \xrightarrow{p} D_\Psi(\boldsymbol{\theta}).\end{aligned}$$

We also have $E\{\Psi_\nu^*(\boldsymbol{\theta})(\Psi_\nu^*(\boldsymbol{\theta}))^T/m\} = E\{\Psi(\boldsymbol{\theta})\Psi^T(\boldsymbol{\theta})\}$, thus by the Law of Large Numbers,

$$\frac{1}{n}\sum_{\nu=1}^{g} \Psi_\nu^*(\boldsymbol{\theta})(\Psi_\nu^*(\boldsymbol{\theta}))^T = \frac{1}{g}\sum_{\nu=1}^{g} \{\Psi_\nu^*(\boldsymbol{\theta})(\Psi_\nu^*(\boldsymbol{\theta}))^T/m\} \xrightarrow{p} E(\Psi(\boldsymbol{\theta})\Psi^T(\boldsymbol{\theta})).$$

From (2.85), we have

$$\sum_{\nu=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})^T - \sum_{\nu=1}^{g} H_{(\nu)}^{-1}(\boldsymbol{\theta})\Psi_\nu^*(\boldsymbol{\theta})(\Psi_\nu^*(\boldsymbol{\theta}))^T \left(H_{(\nu)}^{-1}(\boldsymbol{\theta})\right)^T \xrightarrow{p} 0,$$

which implies that

$$n\sum_{i=1}^{n}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})^T \xrightarrow{p} D_\Psi(\boldsymbol{\theta})^{-1} M_\Psi(\boldsymbol{\theta})(D_\Psi(\boldsymbol{\theta})^{-1})^T.$$

In other words, we proved that $n\sum_{i=1}^{g}(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})(\tilde{\boldsymbol{\theta}}_{(\nu)} - \tilde{\boldsymbol{\theta}})^T$ is a consistent estimator of $J_\Psi^{-1}(\boldsymbol{\theta})$, when m is fixed and $g \to \infty$. □

The main motive for the leave-out-more-than-one-observation-at-a-time approach is to reduce the amount of computation required for the jackknife method. For large samples, it may be helpful to make an initial random grouping of the data by randomly deleting a few observations, if necessary, into g groups of size m. The choice of the number of groups g may be based on the computation costs and the precision or accuracy of the resulting estimators. As regards of computation costs, the choice $(m, g) = (1, n)$ is most expensive. For large samples, $g = n$ may not be computationally feasible, thus some values of g less than n may be preferred. The grouping, however, introduces a degree of arbitrariness, a problem not encountered when $g = n$. This results in an analysis that is not uniquely defined. This is generally not a problem for SE estimation for application purposes, as usually then a rough assessment is sufficient. As regards the precision of the estimators, when the sample size n is small to moderate, the choice $(m, g) = (1, n)$ is preferred. See Chapter 5 for examples.

2.5.2 Jackknife for a function of $\tilde{\boldsymbol{\theta}}$

The jackknife method can also be used for estimates of functions of parameters, such as the asymptotic variance of $P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}})$ for a MCD or MMD model. The usual delta method requires partial derivatives of the function with respect to the parameters, and these may be very difficult to obtain. The jackknife method eliminates the need for these partial derivatives. In the following, we present some results on the jackknife method for functions of $\tilde{\boldsymbol{\theta}}$.

Suppose $\mathbf{b}(\boldsymbol{\theta}) = (b_1(\boldsymbol{\theta}), \ldots, b_s(\boldsymbol{\theta}))'$ is a vector-valued function defined on $\Re$ and taking values in s-dimensional space. We assume that each component function of $\mathbf{b}$, $b_j(\cdot)$ $(j = 1, \ldots, s)$, is real valued and has a differential at $\boldsymbol{\theta}_0$, thus $\mathbf{b}$ has the following expansion as $\boldsymbol{\theta} \to \boldsymbol{\theta}_0$:

$$\mathbf{b}(\boldsymbol{\theta}) = \mathbf{b}(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)\left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'}\right)^T + o(||\boldsymbol{\theta} - \boldsymbol{\theta}_0||), \tag{2.86}$$

where $\partial \mathbf{b}/\partial \boldsymbol{\theta}_0' = (\partial \mathbf{b}/\partial \boldsymbol{\theta}')|_{\boldsymbol{\theta}=\boldsymbol{\theta}_0}$ is of rank $t = \min(s, q)$.

By Theorem 2.4, $\tilde{\boldsymbol{\theta}}$ has an asymptotic normal distribution

$$\sqrt{n}(\tilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_\Psi^{-1}).$$

Similarly by Theorem 2.8 and 2.10, $\tilde{\boldsymbol{\theta}}_J$ has an asymptotic normal distribution in the sense that

$$\sqrt{n}(\tilde{\boldsymbol{\theta}}_J - \boldsymbol{\theta}) \xrightarrow{\mathrm{D}} N_q(\mathbf{0}, J_\Psi^{-1}).$$

We have the following results for $\mathbf{b}(\tilde{\boldsymbol{\theta}})$ and $\mathbf{b}(\tilde{\boldsymbol{\theta}}_J)$:

Theorem 2.12 *Let* $\mathbf{b}$ *be as described above and suppose (2.86) holds. Under the same assumptions as in Theorem 2.4,* $\mathbf{b}(\tilde{\boldsymbol{\theta}})$ *has the asymptotic distribution given by*

$$\sqrt{n}(\mathbf{b}(\tilde{\boldsymbol{\theta}}) - \mathbf{b}(\boldsymbol{\theta})) \xrightarrow{\mathrm{D}} N_t \left(\mathbf{0}, \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'} \right) J_{\Psi}^{-1} \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'} \right)^T \right).$$

Proof: See Serfling (1980, Ch.3). □

Theorem 2.13 *Let* $\mathbf{b}$ *be as described above and suppose (2.86) hold. Under the same assumptions as in Theorem 2.4,* $\mathbf{b}(\tilde{\boldsymbol{\theta}}_J)$ *has the asymptotic distribution given by*

$$\sqrt{n}(\mathbf{b}(\tilde{\boldsymbol{\theta}}_J) - \mathbf{b}(\boldsymbol{\theta})) \xrightarrow{\mathrm{D}} N_t \left(\mathbf{0}, \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'} \right) J_{\Psi}^{-1} \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'} \right)^T \right).$$

Proof: See Serfling (1980, Ch.3). □

As in the previous subsection, let $\tilde{\boldsymbol{\theta}}_{(\nu)}$ be the estimator of $\boldsymbol{\theta}$ with the ν-th group of size m deleted, $\nu = 1, \ldots, g$. We define the *jackknife estimate of variance* of $\mathbf{b}(\tilde{\boldsymbol{\theta}})$, which we denote by $V_{J\mathbf{b}}$, as

$$V_{J\mathbf{b}} = \sum_{\nu=1}^{g} \left(\mathbf{b}(\tilde{\boldsymbol{\theta}}_{(\nu)}) - \mathbf{b}(\tilde{\boldsymbol{\theta}}) \right) \left(\mathbf{b}(\tilde{\boldsymbol{\theta}}_{(\nu)}) - \mathbf{b}(\tilde{\boldsymbol{\theta}}) \right)^T. \tag{2.87}$$

We have the following theorem.

Theorem 2.14 *Let* $\mathbf{b}$ *be as described above and suppose (2.86) holds. Under the same assumptions as in Theorem 2.4, the sample size* n *times the jackknife estimate of variance* $V_{J\mathbf{b}}$ *defined by (2.87) is a consistent estimator of*

$$\left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'} \right) J_{\Psi}^{-1} \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\theta}_0'} \right)^T.$$

Proof: The proof is similar to that of Theorem 2.11, and thus omitted here. □

To carry out the above computational results related to the estimates of functions of parameters, it would be desirable to maintain a table of the parameter estimates for the full sample and each jackknife subsample. Then one can use this table for computing estimates of one or more functions of the parameters, and their corresponding SEs.

The results in Theorems 2.12, 2.13 and 2.14 have immediate applications. One example is given next.

Example 2.18 For a MCD or MMD model in (2.12), say $P(y_1 \cdots y_d; \boldsymbol{\theta})$, we could apply the above results to say something about the asymptotic behaviour of $P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}})$ and $P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}}_J)$. From Theorems 2.12 and 2.13, we derive that as $n \to \infty$

$$\sqrt{n}(P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}}) - P(y_1 \cdots y_d; \boldsymbol{\theta})) \xrightarrow{\mathrm{D}} N \left(0, \left(\frac{\partial P}{\partial \boldsymbol{\theta}'} \right) J_{\Psi}^{-1} \left(\frac{\partial P}{\partial \boldsymbol{\theta}'} \right)^T \right)$$

and

$$\sqrt{n}(P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}}_J) - P(y_1 \cdots y_d; \boldsymbol{\theta})) \xrightarrow{\mathrm{D}} N\left(0, \left(\frac{\partial P}{\partial \boldsymbol{\theta}'}\right) J_{\Psi}^{-1} \left(\frac{\partial P}{\partial \boldsymbol{\theta}'}\right)^T\right).$$

Furthermore, by Theorem 2.14, we obtain a consistent estimator of $(\partial P/\partial \boldsymbol{\theta}')J_{\Psi}^{-1}(\partial P/\partial \boldsymbol{\theta}')^T$, *i.e.*

$$n\sum_{\nu=1}^{g}\left(P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}}_{(\nu)}) - P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}})\right)^2.$$

Also see Chapter 5 for direct application in data analysis. □

2.5.3 Jackknife approach for models with covariates

Suppose we have the model defined by (2.53) and (2.54). Let $\tilde{\boldsymbol{\gamma}}_{(\nu)}$ $(\nu = 1, \ldots, g)$ be an estimate of $\boldsymbol{\gamma}$ based on the same set of inference functions $\Psi_n(\boldsymbol{\gamma})$ from the data set $\mathbf{y}_1, \ldots, \mathbf{y}_n$ but deleting the ν-th group of size m (m is fixed). The *jackknife estimate* of $\boldsymbol{\gamma}$ is

$$\tilde{\boldsymbol{\gamma}}_J = \frac{1}{g}\sum_{\nu=1}^{g}\tilde{\boldsymbol{\gamma}}_\nu = g\tilde{\boldsymbol{\gamma}} - (g-1)\tilde{\boldsymbol{\gamma}}_{(\cdot)}, \tag{2.88}$$

where $\tilde{\boldsymbol{\gamma}}_{(\cdot)} = 1/g\sum_{\nu=1}^{g}\tilde{\boldsymbol{\gamma}}_{(\nu)}$, and $\tilde{\boldsymbol{\gamma}}_\nu = g\tilde{\boldsymbol{\gamma}} - (g-1)\tilde{\boldsymbol{\gamma}}_{(\nu)}$ $(\nu = 1, \ldots, g)$.

We define the *jackknife estimate of variance* $V_{J,\boldsymbol{\gamma}}$ for $\tilde{\boldsymbol{\gamma}}$ as follows:

$$V_{J,\boldsymbol{\gamma}} = \sum_{\nu=1}^{g}\left(\tilde{\boldsymbol{\gamma}}_{(\nu)} - \tilde{\boldsymbol{\gamma}}\right)\left(\tilde{\boldsymbol{\gamma}}_{(\nu)} - \tilde{\boldsymbol{\gamma}}\right)^T. \tag{2.89}$$

Under the assumptions for the establishment of Theorem 2.6, in parallel to Theorems 2.10 and 2.11, we have the following theorems for the models with covariates. The proofs are generalizations of the proofs for Theorem 2.5, 2.6, 2.10 and 2.11. We omit the proofs and only state the results here.

Theorem 2.15 *Consider the general model (2.53) with d arbitrary. Let $\tilde{\boldsymbol{\gamma}}$ denote the IFME of $\boldsymbol{\gamma}$ under the IFM corresponding to (2.57). Under Assumptions 2.1, 2.2 and 2.3, the jackknife estimator $\tilde{\boldsymbol{\gamma}}_J$ defined by (2.88) is asymptotically normal in the sense that, as $n \to \infty$,*

$$\sqrt{n}A_n^{-1}(\tilde{\boldsymbol{\gamma}}_J - \boldsymbol{\gamma}_0)) \xrightarrow{D} N_q(\mathbf{0}, I),$$

where $A_n = D_n^{-1/2}(\boldsymbol{\gamma}_0)M_n^{1/2}(\boldsymbol{\gamma}_0)(D_n^{-1/2}(\boldsymbol{\gamma}_0))^T$, $D_n(\boldsymbol{\gamma}_0)$ and $M_n(\boldsymbol{\gamma}_0)$ are defined by (2.56). □

Theorem 2.16 *Under the same assumptions as in Theorem 2.7, we have*

$$nV_{J,\boldsymbol{\gamma}} - D_n^{-1}(\boldsymbol{\gamma}_0)M_n(\boldsymbol{\gamma}_0)(D_n^{-1}(\boldsymbol{\gamma}_0))^T \xrightarrow{\mathrm{P}} \mathbf{0},$$

where $V_{J,\boldsymbol{\gamma}}$ is the jackknife estimate of variance defined by (2.89), and $D_n(\boldsymbol{\gamma}_0)$ and $M_n(\boldsymbol{\gamma}_0)$ are defined by (2.56). □

Theorem 2.17 *Let* $\mathbf{b}$ *be as described in subsection 2.5.2 and suppose (2.86) hold. Under the same assumptions as in Theorem 2.7,* $\mathbf{b}(\tilde{\boldsymbol{\gamma}})$*, a function of IFME* $\tilde{\boldsymbol{\gamma}}$*, has the asymptotic distribution given by*

$$\sqrt{n}B_n^{-1}(\mathbf{b}(\tilde{\boldsymbol{\gamma}}) - \mathbf{b}(\boldsymbol{\gamma})) \xrightarrow{\mathrm{D}} N_t(\mathbf{0}, I),$$

where $B_n = \left[(\partial \mathbf{b}/\partial \boldsymbol{\gamma}_0')\, D_n^{-1}(\boldsymbol{\gamma}_0) M_n(\boldsymbol{\gamma}_0)(D_n^{-1}(\boldsymbol{\gamma}_0))^T (\partial \mathbf{b}/\partial \boldsymbol{\gamma}_0')^T\right]^{1/2}$*, and* $D_n(\boldsymbol{\gamma}_0)$ *and* $M_n(\boldsymbol{\gamma}_0)$ *are defined by (2.56).* □

Theorem 2.18 *Let* $\mathbf{b}$ *be as described in subsection 2.5.2 and suppose (2.86) hold. Under the same assumptions as in Theorem 2.7,* $\mathbf{b}(\tilde{\boldsymbol{\gamma}}_J)$*, of the jackknife estimate* $\tilde{\boldsymbol{\gamma}}_J$ *derived from* $\tilde{\boldsymbol{\gamma}}$*, has the asymptotic distribution given by*

$$\sqrt{n}B_n^{-1}(\mathbf{b}(\tilde{\boldsymbol{\gamma}}_J) - \mathbf{b}(\boldsymbol{\gamma})) \xrightarrow{\mathrm{D}} N_t(\mathbf{0}, I),$$

where $B_n = \left[(\partial \mathbf{b}/\partial \boldsymbol{\gamma}_0')\, D_n^{-1}(\boldsymbol{\gamma}_0) M_n(\boldsymbol{\gamma}_0)(D_n^{-1}(\boldsymbol{\gamma}_0))^T (\partial \mathbf{b}/\partial \boldsymbol{\gamma}_0')^T\right]^{1/2}$*, and* $D_n(\boldsymbol{\gamma}_0)$ *and* $M_n(\boldsymbol{\gamma}_0)$ *are defined by (2.56).* □

We define the *jackknife estimate of variance* of $\mathbf{b}(\tilde{\boldsymbol{\gamma}})$, $V_{J\mathbf{b}}$, as follows:

$$V_{J\mathbf{b}} = \sum_{\nu=1}^{g} \left(\mathbf{b}(\tilde{\boldsymbol{\gamma}}_{(\nu)}) - \mathbf{b}(\tilde{\boldsymbol{\gamma}})\right)\left(\mathbf{b}(\tilde{\boldsymbol{\gamma}}_{(\nu)}) - \mathbf{b}(\tilde{\boldsymbol{\gamma}})\right)^T. \tag{2.90}$$

Theorem 2.19 *Let* $\mathbf{b}$ *be as described in subsection 2.5.2 and suppose (2.86) hold. Under the same assumptions as in Theorem 2.6, we have*

$$nV_{J\mathbf{b}} - \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\gamma}_0'}\right) D_n^{-1}(\boldsymbol{\gamma}_0) M_n(\boldsymbol{\gamma}_0)(D_n^{-1}(\boldsymbol{\gamma}_0))^T \left(\frac{\partial \mathbf{b}}{\partial \boldsymbol{\gamma}_0'}\right)^T \xrightarrow{\mathrm{p}} \mathbf{0},$$

where the jackknife estimate of variance $V_{J\mathbf{b}}$ *defined by (2.90) and* $D_n(\boldsymbol{\gamma}_0)$ *and* $M_n(\boldsymbol{\gamma}_0)$ *are defined by (2.56).* □

2.6 Estimation for models with parameters common to more than one margin

One potential application of the MCD and MMD models is for longitudinal or repeated measures studies with short time series, in which the interest may be on how the distribution of the response changes over time. Some common characteristics, which carry over time, may appear in the form of common regression parameters or common dependence parameters. There are also general situations in which the same parameters appear in more than one margin. This happens with the MCD and

MMD models, for example, when there is a special dependence structure in the copula C, such as in the multinormal copula (2.4), where $\Theta = (\theta_{jk})$ is an exchangeable correlation matrix with all correlations equal to θ, or Θ is an AR(1) correlation matrix with the (j,k) component equal to $\theta^{|j-k|}$ for some θ.

Example 2.19 Suppose for the d-variate binary vector $\mathbf{Y_i}$ with a covariate vector for the jth univariate margin can be represented as $Y_{ij} = I(Z_{ij} \leq \alpha_j + \boldsymbol{\beta}_j \mathbf{x}_{ij})$, $i = 1, \ldots, n$, where $\mathbf{Z_i} \sim N(\mathbf{0}, \Theta)$. This is a multivariate probit model. Assume $\boldsymbol{\beta}_j = \boldsymbol{\beta}$, then the common regression coefficients appear in more than one margin. We could estimate $\boldsymbol{\beta}$ from the d univariate margins based on the IFM approach, but then we have d estimates of $\boldsymbol{\beta}$. Taking any one of them as the estimate of $\boldsymbol{\beta}$ evidently results in some loss of information. Can we pool the information together to get a better estimate of $\boldsymbol{\beta}$? The same question arises for the correlation matrix Θ. Assume there is no covariate for Θ. When Θ has certain special forms, for example exchangeable or AR(1), the same parameter appears in $d(d-1)/2$ bivariate margins. Can we get a more efficient estimate from the IFM approach? There are also situations where a parameter is common some margins, such as $\theta_{12} = \theta_{23} = \cdots = \theta_{d,d-1}$ in Θ. The same question about getting a more efficient estimate arises. □

A direct approach for common parameter estimation is to use the likelihood of a higher-order margin, if this is computationally feasible. Otherwise, the IFM approach for model fitting can be applied. With the IFM approach, appropriately taking the information about common parameters into account can improve the efficiency of the parameter estimates. Analytical and numerical evidence supporting this claim are given in Chapter 4 for these two approaches of information pooling for IFM that we propose here. The first approach, called the *weighting approach* (WA), is to form a new estimate based on some weighting of the estimates for the same parameter from different margins. A special case is the simple average. The second approach, called the *pool-marginal-likelihoods approach* (PMLA), is to rewrite the inference function of margins under the assumption that the same parameter appears in several margins. In the following, we outline the two estimating approaches in general terms.

2.6.1 Weighting approach

WA is a method to get an efficient estimate based on a weighted average of different estimates for the same parameter. We state this approach in general terms. Assume $\tilde{\gamma}_1, \ldots, \tilde{\gamma}_q$ are estimates of the same parameter γ, but from different inference functions. Let $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}_1, \ldots, \tilde{\gamma}_q)'$, and let $\Sigma_{\tilde{\boldsymbol{\gamma}}}$ be

the asymptotic variance-covariance matrix based on Godambe information matrix. One relatively efficient estimate of the parameter γ is based on the following result, which easily obtains from the method of Lagrange multipliers.

Result 2.1 *Suppose* $\mathbf{X}$ *is a q-variate random vector with mean vector* $\mu_{\mathbf{X}} = (\mu, \ldots, \mu)' = \mu\mathbf{1}$ *and* $Var(\mathbf{X}) = \Sigma_{\mathbf{X}}$, *where* μ *is a scalar and* $\mathbf{1} = (1, \ldots, 1)'$. *A linear unbiased estimate of* μ, $\mathbf{u}'\mathbf{X}$, *has the smallest variance when*

$$\mathbf{u} = \frac{\Sigma_{\mathbf{X}}^{-1}\mathbf{1}}{\mathbf{1}'\Sigma_{\mathbf{X}}^{-1}\mathbf{1}}.$$

□

Applying the above result to our problem, the resulting estimate of γ is

$$\tilde{\gamma} = \frac{\mathbf{1}'\Sigma_{\tilde{\boldsymbol{\gamma}}}^{-1}\tilde{\boldsymbol{\gamma}}}{\mathbf{1}'\Sigma_{\tilde{\boldsymbol{\gamma}}}^{-1}\mathbf{1}}. \tag{2.91}$$

If $\tilde{\gamma}_j$s are consistent estimates of γ, then $\tilde{\gamma}$ is also a consistent estimate of γ and it has smaller asymptotic variance than any of the individual estimates of $\tilde{\gamma}$ from one particular inference function. The asymptotic variance of $\tilde{\gamma}$ is

$$\sigma_{\tilde{\gamma}}^2 = \mathbf{u}'\Sigma_{\tilde{\boldsymbol{\gamma}}}\mathbf{u} = \frac{1}{\mathbf{1}'\Sigma_{\tilde{\boldsymbol{\gamma}}}^{-1}\mathbf{1}}. \tag{2.92}$$

A computationally simpler but slightly less efficient estimate of γ is

$$\tilde{\gamma} = \frac{\mathbf{1}'\mathrm{diag}\{\Sigma_{\tilde{\boldsymbol{\gamma}}}^{-1}\}\tilde{\boldsymbol{\gamma}}}{\mathbf{1}'\mathrm{diag}\{\Sigma_{\tilde{\boldsymbol{\gamma}}}^{-1}\}\mathbf{1}}, \tag{2.93}$$

and an approximation of the asymptotic variance of $\tilde{\gamma}$ from (2.93) is $\sigma_{\tilde{\gamma}}^2 = 1/(\mathbf{1}'\mathrm{diag}\{\Sigma_{\tilde{\boldsymbol{\gamma}}}^{-1}\}\mathbf{1})$. A naive estimate of γ is $\tilde{\gamma} = \mathbf{1}'\tilde{\boldsymbol{\gamma}}/\mathbf{1}'\mathbf{1}$, which is just the simple average. In some special situations, such as in the Example 4.4, the estimate of (2.91) reduces to a simple average.

In practice, $\Sigma_{\tilde{\boldsymbol{\gamma}}}$ may not be available. We may have to estimate it based on the data. The following algorithm may be useful for getting a final estimate of γ. Assume we already have $\tilde{\boldsymbol{\gamma}}$ and $\widetilde{\Sigma}_{\tilde{\boldsymbol{\gamma}}}$.

Computation algorithm:

1. Let $\mathbf{u} = \widetilde{\Sigma}_{\tilde{\boldsymbol{\gamma}}}^{-1}\mathbf{1}/\mathbf{1}'\widetilde{\Sigma}_{\tilde{\boldsymbol{\gamma}}}^{-1}\mathbf{1}$.
2. Find $\tilde{\gamma} = \mathbf{u}'\tilde{\boldsymbol{\gamma}}$.
3. Set $\tilde{\boldsymbol{\gamma}} = (\tilde{\gamma}, \ldots, \tilde{\gamma})$, and update $\widetilde{\Sigma}_{\tilde{\boldsymbol{\gamma}}}$ with this new $\tilde{\boldsymbol{\gamma}}$.

4. Go back to step 1 until the final estimate of $\tilde{\gamma}$ satisfies a prescribed convergence criterion.

□

The asymptotic variance of $\tilde{\gamma}$ at the final iteration can be calculated by (2.92).

2.6.2 The pool-marginal-likelihoods approach

In this approach, different inference functions for the same parameter are pooled together as a sum, where each inference function is the loglikelihood of some margin. We use the model (2.54) to illustrate the approach. Suppose that in (2.54), $\boldsymbol{\alpha}_j = \boldsymbol{\alpha}$, $j = 1, \ldots, d$ and $b_{jk} = \boldsymbol{\beta}$, $1 \le j < k \le d$, then more efficient estimators of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ may be obtained. For example, we can sum up the the loglikelihood functions of margins corresponding to the parameter vectors $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in the following way:

$$\begin{cases} \ell_n^*(\boldsymbol{\alpha}) = \sum_{i=1}^{n}\sum_{j=1}^{d} \log P_j(y_{ij}), \\ \ell_n^*(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j<k} \log P_{jk}(y_{ij} y_{ik}). \end{cases} \tag{2.94}$$

(2.94) is an example of PMLA. The inference functions of margins from (2.94) corresponding to $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\begin{aligned} \Psi_{\mathrm{IFM}} = \Bigg(& \sum_{i=1}^{n}\sum_{j=1}^{d} \frac{1}{P_j(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \alpha_1}, \ldots, \sum_{i=1}^{n}\sum_{j=1}^{d} \frac{1}{P_j(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \alpha_p}, \\ & \sum_{i=1}^{n}\sum_{j<k}^{d} \frac{1}{P_{jk}(y_{ij} y_{ik})} \frac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \beta_1}, \ldots, \sum_{i=1}^{n}\sum_{j<k}^{d} \frac{1}{P_{jk}(y_{ij} y_{ik})} \frac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \beta_q} \Bigg), \end{aligned} \tag{2.95}$$

and the estimating equations based on IFM is

$$\begin{cases} \sum_{i=1}^{n}\sum_{j=1}^{d} \frac{1}{P_j(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \alpha_s} = 0, \quad s = 1, \ldots, p, \\ \sum_{i=1}^{n}\sum_{j<k}^{d} \frac{1}{P_{jk}(y_{ij} y_{ik})} \frac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \beta_t} = 0, \quad t = 1, \ldots, q. \end{cases}$$

If we consider (2.95) as inference functions, we see that asymptotic theory on the estimates from the PMLA can be established by applying the general inference function theory.

For completeness, we give here algebraic expressions on IFM with the PMLA for the Godambe information for the model (2.39) with $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d, \theta_{12}, \ldots, \theta_{d-1,d})'$, where we assume $\theta_1, \ldots, \theta_d$

are each a function of one parameter λ, and $\theta_{12}, \ldots, \theta_{d-1,d}$ are each a function of one parameter ρ. The loglikelihood functions of margin corresponding to the parameter vectors λ and ρ are

$$\begin{cases} \ell_n^*(\lambda) = \sum_{i=1}^{n} \sum_{j=1}^{d} \log P_j(y_{ij}), \\ \ell_n^*(\lambda, \rho) = \sum_{i=1}^{n} \sum_{j,k=1; j<k}^{d} \log P_{jk}(y_{ij} y_{ik}), \end{cases} \tag{2.96}$$

and the estimating equations based on IFM is

$$\begin{cases} \sum_{i=1}^{n} \sum_{j=1}^{d} \frac{1}{P_j(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \lambda} = 0, \\ \sum_{i=1}^{n} \sum_{j<k}^{d} \frac{1}{P_{jk}(y_{ij} y_{ik})} \frac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \rho} = 0. \end{cases}$$

The IFM corresponding to one observation is

$$\Psi = (\psi_1, \psi_2) = \left(\sum_{j=1}^{d} \frac{1}{P_j(y_j)} \frac{\partial P_j(y_j)}{\partial \lambda}, \sum_{j<k}^{d} \frac{1}{P_{jk}(y_j y_k)} \frac{\partial P_{jk}(y_j y_k)}{\partial \rho} \right).$$

The Godambe information matrix is a 2×2 matrix. To calculate the Godambe information matrix $J_\Psi = D_\Psi^T M_\Psi^{-1} D_\Psi$, we need the matrices

$$M_\Psi = \begin{pmatrix} \mathrm{E}(\psi_1^2) & \mathrm{E}(\psi_1 \psi_2) \\ \mathrm{E}(\psi_1 \psi_2) & \mathrm{E}(\psi_2^2) \end{pmatrix} \quad \text{and} \quad D_\Psi = \begin{pmatrix} \mathrm{E}(\partial \psi_1 / \partial \lambda) & 0 \\ 0 & \mathrm{E}(\partial \psi_2 / \partial \rho) \end{pmatrix}.$$

For $\mathrm{E}(\psi_1^2)$, we have

$$\begin{aligned} \mathrm{E}(\psi_1^2) &= \mathrm{E} \left(\sum_{j=1}^{d} \frac{1}{P_j(y_j)} \frac{\partial P_j(y_j)}{\partial \lambda} \right)^2 \\ &= \sum_{\{y_1 \cdots y_d\}} P(y_1 \cdots y_d) \left(\sum_{j=1}^{d} \frac{1}{P_j(y_j)} \frac{\partial P_j(y_j)}{\partial \lambda} \right)^2. \end{aligned}$$

It can be estimated consistently by

$$\frac{1}{n} \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \frac{1}{P_j(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \lambda} \right)^2 \Bigg|_{\tilde{\lambda}}. \tag{2.97}$$

Similarly, we have

$$\begin{aligned} \mathrm{E}(\psi_2^2) &= \mathrm{E} \left(\sum_{j<k}^{d} \frac{1}{P_{jk}(y_j y_k)} \frac{\partial P_{jk}(y_j y_k)}{\partial \rho} \right)^2 \\ &= \sum_{\{y_1 \cdots y_d\}} P_{jk}(y_1 \cdots y_d) \left(\sum_{j<k}^{d} \frac{1}{P_{jk}(y_j y_k)} \frac{\partial P_{jk}(y_j y_k)}{\partial \rho} \right)^2 \end{aligned}$$

and

$$\mathrm{E}(\psi_1\psi_2) = \mathrm{E}\left\{\left(\sum_{j=1}^{d}\frac{1}{P_j(y_j)}\frac{\partial P_j(y_j)}{\partial\lambda}\right)\left(\sum_{j<k}^{d}\frac{1}{P_{jk}(y_jy_k)}\frac{\partial P_{jk}(y_jy_k)}{\partial\rho}\right)\right\}$$
$$= \sum_{\{y_1\cdots y_d\}} P(y_1\cdots y_d)\left[\left(\sum_{j=1}^{d}\frac{1}{P_j(y_j)}\frac{\partial P_j(y_j)}{\partial\lambda}\right)\left(\sum_{j<k}^{d}\frac{1}{P_{jk}(y_jy_k)}\frac{\partial P_{jk}(y_jy_k)}{\partial\rho}\right)\right].$$

For $\mathrm{E}(\partial\psi_1/\partial\lambda)$,

$$\frac{\partial\psi_1}{\partial\lambda} = \sum_{j=1}^{d}\left[-\frac{1}{P_j^2(y_j)}\left(\frac{\partial P_j(y_j)}{\partial\lambda}\right)^2 + \frac{1}{P_j(y_j)}\frac{\partial^2 P_j(y_j)}{\partial\lambda^2}\right],$$

so

$$\mathrm{E}(\partial\psi_1/\partial\lambda) = \sum_{\{y_1\cdots y_d\}} P(y_1\cdots y_d)\sum_{j=1}^{d}\left[-\frac{1}{P_j^2(y_j)}\left(\frac{\partial P_j(y_j)}{\partial\lambda}\right)^2 + \frac{1}{P_j(y_j)}\frac{\partial^2 P_j(y_j)}{\partial\lambda^2}\right].$$

Similarly, we find

$$\mathrm{E}(\partial\psi_2/\partial\rho) = \sum_{\{y_1\cdots y_d\}} P(y_1\cdots y_d)\sum_{j<k}^{d}\left[-\frac{1}{P_{jk}^2(y_jy_k)}\left(\frac{\partial P_{jk}(y_jy_k)}{\partial\rho}\right)^2 + \frac{1}{P_{jk}(y_jy_k)}\frac{\partial^2 P_{jk}(y_jy_k)}{\partial\rho^2}\right].$$

Consistent estimates for $\mathrm{E}(\psi_2^2)$, $\mathrm{E}(\psi_1\psi_2)$, $\mathrm{E}(\partial\psi_1/\partial\lambda)$ and $\mathrm{E}(\partial\psi_2/\partial\rho)$ can be similarly written as in (2.97).

2.6.3 Examples

We give two examples of WA and PMLA.

Example 2.20 *1. A trivariate probit model with exchangeable dependence structure*: Suppose we have a trivariate probit model with known cut-off points and $P(111) = \Phi_3(0,0,0,\rho,\rho,\rho)$. It can be shown (see Example 4.4) that the asymptotic variance of $\tilde{\rho}$ from one bivariate margin is $[(\pi^2 - 4(\sin^{-1}\rho)^2)(1-\rho^2)]/4n$, and the asymptotic variance of $\tilde{\rho}$ from WA or PMLA is $[(1-\rho^2)(\pi + 6\sin^{-1}\rho)(\pi - 2\sin^{-1}\rho)]/12n$. The ratio of the former to the latter is $[3(\pi+2\sin^{-1}\rho)]/[\pi+6\sin^{-1}\rho]$, which decreases from ∞ to 1.5 as ρ increases from -0.5 to 1. In this example, the optimal weighting is equivalent to a simple average (see Example 4.4 for details).

2. A trivariate probit model with AR(1) dependence structure: Suppose we have a trivariate probit model with cut-off points known, such that $P(111) = \Phi_3(0,0,0,\rho,\rho^2,\rho)$. Let σ_w^2 be the asymptotic variance of $\tilde{\rho}$ from WA, σ_p^2 be the asymptotic variance of $\tilde{\rho}$ from PMLA, σ_{12}^2 be the asymptotic variance of $\tilde{\rho}$ from the $(1,2)$ margin, and σ_{13}^2 be the asymptotic variance of $\tilde{\rho}$ from the $(1,3)$ margin. In Example 4.5, we show that $\sigma_p^2/\sigma_w^2 \geq 1$, with a maximum value of 1.0391, which is attained at

$\rho = 0.3842$ and $\rho = -0.3842$; σ_{12}^2/σ_w^2 increases from 1.707 to 2 as ρ goes from -1 to 0 and from 2 to 1.707 as ρ goes from 0 to 1; and σ_{13}^2/σ_w^2 increases from 1.207 to ∞ as ρ goes from -1 to 0, and decreases from ∞ to 1.207 as ρ goes from 0 to 1. □

PMLA in the form presented in this section can also be considered as a simple weighted likelihood approach. More complicated weighting schemes for the PMLA can be sought. In general, as long as a reasonable efficiency is preserved, we prefer the simple weighting schemes.

2.7 Numerical methods for the model fitting

From previous sections, we see that the IFM approach for parameter estimation leads to the problem of optimization of a set of loglikelihood functions of margins. The typical system of functions for the models with no covariates in the form of loglikelihood functions of margins are

$$\begin{cases} \ell_{nj}(\lambda_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d, \end{cases} \tag{2.98}$$

and the estimating equations (derived from the loglikelihood functions of margins) are

$$\begin{cases} \Psi_{nj}(\lambda_j) = \sum_{i=1}^{n} \dfrac{1}{P_j(y_{ij})} \dfrac{\partial P_j(y_{ij})}{\partial \lambda_j} = 0, \quad j = 1, \ldots, d, \\ \Psi_{njk}(\theta_{jk}) = \sum_{i=1}^{n} \dfrac{1}{P_{jk}(y_{ij} y_{ik})} \dfrac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \theta_{jk}} = 0, \quad 1 \le j < k \le d. \end{cases} \tag{2.99}$$

For the models with covariates, the typical system of functions in the form of loglikelihood functions of margins are

$$\begin{cases} \ell_{nj}(\boldsymbol{\alpha}_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\boldsymbol{\beta}_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d \end{cases} \tag{2.100}$$

and the estimating equations are

$$\begin{cases} \Psi_{nj}(\boldsymbol{\alpha}_j) = \sum_{i=1}^{n} \dfrac{1}{P_j(y_{ij})} \dfrac{\partial P_j(y_{ij})}{\partial \boldsymbol{\alpha}_j} = 0, \quad j = 1, \ldots, d, \\ \Psi_{njk}(\boldsymbol{\beta}_{jk}) = \sum_{i=1}^{n} \dfrac{1}{P_{jk}(y_{ij} y_{ik})} \dfrac{\partial P_{jk}(y_{ij} y_{ik})}{\partial \boldsymbol{\beta}_{jk}} = 0, \quad 1 \le j < k \le d. \end{cases} \tag{2.101}$$

Newton-Raphson method

The traditional approach for numerical optimization or root-finding is the *Newton-Raphson method.* This method requires the evaluation of both the first and the second derivations of the objective functions in (2.98) and (2.100). This method, with its good rate of convergence, is the preferred method if the derivatives can be easily obtained analytically and coded in a program. But in many cases, for example with $\ell_{njk}(\theta_{jk})$ or $\ell_{njk}(\boldsymbol{\beta}_{jk})$, where bivariate objects involve non-closed form two-dimensional integrals, application of the Newton-Raphson method is difficult since analytical derivatives in such situations are very hard to obtain. The Newton-Raphson method can only be easily applied for a few cases with $\ell_{nj}(\lambda_j)$ or $\ell_{nj}(\boldsymbol{\alpha}_j)$, where only univariate objects are involved. For example, the Newton-Raphson method may be used to solve $\Psi_{nj}(\lambda_j) = 0$ to find $\tilde{\lambda}_j$, $j = 1, \ldots, d$. In this case, based on Newton-Raphson method, for a given initial value $\lambda_{j,0}$, an updated value of $\tilde{\lambda}_j$ is

$$\tilde{\lambda}_{j,\text{new}} = \lambda_{j,0} - \left\{ \left[\frac{\partial \Psi_{nj}(\lambda_j)}{\partial \lambda_j} \right]^{-1} \Psi_{nj}(\lambda_j) \right\} \Bigg|_{\lambda_{j,0}}. \tag{2.102}$$

This is repeated until successive $\tilde{\lambda}_{j,\text{new}}$ agree to a specified precision. In (2.102), we need to be able to code

$$\Psi_{nj}(\lambda_j) = \sum_{i=1}^{n} [1/P_j(y_{ij})][\partial P_j(y_{ij})/\partial \lambda_j] \tag{2.103}$$

and

$$\frac{\partial \Psi_{nj}(\lambda_j)}{\partial \lambda_j} = \sum_{i=1}^{n} \frac{1}{P_j(y_{ij})} \frac{\partial^2 P_j(y_{ij})}{\partial \lambda_j^2} - \frac{1}{P_j^2(y_{ij})} \left(\frac{\partial P_j(y_{ij})}{\partial \lambda_j} \right)^2. \tag{2.104}$$

This is for the case with no covariates. For the case with covariates, similar iteration equations to (2.102) can be written down. We need to calculate

$$\Psi_{nj}(\boldsymbol{\alpha}_j) = \sum_{i=1}^{n} [1/P_j(y_{ij})][\partial P_j(y_{ij})/\partial \boldsymbol{\alpha}_j], \tag{2.105}$$

which is a $p_j \times 1$ vector and

$$\frac{\partial \Psi_{nj}(\boldsymbol{\alpha}_j)}{\partial \boldsymbol{\alpha}_j} = \sum_{i=1}^{n} \left[\frac{1}{P_j(y_{ij})} \frac{\partial^2 P_j(y_{ij})}{\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j'} - \frac{1}{P_j^2(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \boldsymbol{\alpha}_j} \left(\frac{\partial P_j(y_{ij})}{\partial \boldsymbol{\alpha}_j} \right)' \right], \tag{2.106}$$

which is a $p_j \times p_j$ matrix. It is equivalent to calculate the gradient of $P_j(y_{ij})$ at the point $\boldsymbol{\alpha}_j$, which is the p_j-vector of (first order) partial derivatives:

$$\frac{\partial}{\partial \boldsymbol{\alpha}_j} P_j(y_{ij}) = \left[\frac{\partial}{\partial \alpha_{j1}} P_j(y_{ij}), \ldots, \frac{\partial}{\partial \alpha_{jp_j}} P_j(y_{ij}) \right]^T$$

and the Hessian of $P_j(y_{ij})$ at the point $\boldsymbol{\alpha}_j$, which is a $p_j \times p_j$ matrix of second order partial derivatives with (s,t) $(s,t = 1,\ldots,p_j)$ component $(\partial^2/\partial\alpha_s\partial\alpha_t)P_j(y_{ij})$.

To avoid the often tedious algebraic derivatives in (2.103) – (2.106), modern symbolic computation software, such as Maple (Char *et al.*, 1992), may be used. This software is also convenient in that it outputs the results in the form of C or Fortran code.

Quasi-Newton method

For many multivariate models, it is inconvenient to supply both first and second partial derivatives of the objective functions as required by the Newton-Raphson method. For example, to get the partial derivatives of the forms (2.104) – (2.106) may be tedious, particularly with function objects such as $\ell_{njk}(\theta_{jk})$ or $\ell_{njk}(\boldsymbol{\beta}_{jk})$, where 2-dimensional integrations are often involved. A numerical method for optimization that is useful for many multivariate models in this thesis is the *quasi-Newton method* (or *variable-metric method*). This method uses the numerical approximation to the derivatives (gradients and Hessian matrix) in the Newton-Raphson iteration; thus it can be considered as a derivative-free method. In many situations, a crude approximation to the derivatives can lead to convergence in the Newton-Raphson iteration as well. Application of this method requires only the objective functions, such as those in (2.98) and (2.100), to be coded. The gradients are computed numerically and the inverse Hessian matrix of second order derivatives is updated after each iteration. This method has the advantage of not requiring the analytic derivatives of the objective functions with respect to the parameters. Its disadvantage is that convergence could be slow compared with the Newton-Raphson approach. An example of a quasi-Newton routine, which is used in the programs written for this thesis work, is a quasi-Newton minimization routine in Nash (1990) (Algorithm 21, p192). This is a modified Fletcher variable-metric method; the original method is due to Fletcher (1970).

With the quasi-Newton methods, all we need to do is to write down the optimization (minimization or maximization) objective function (such as $\ell_{njk}(\boldsymbol{\beta}_{jk})$), and then let a quasi-Newton routine take care of the rest. A quasi-Newton routine works fine if the objective function can be computed to arbitrary precision, say ϵ_0. The numerical gradients are then based on a step size (or step length) $\epsilon < \epsilon_0$. The calculation of the optimization objective function with multivariate model often involves the evaluation of multiple integration at some arbitrary points. One-dimensional and two-dimensional numerical integrals can usually be computed quite quickly to around six digits of precision, but there is a problem of computational time in trying to achieve many digits of precision for numerical integrals of dimension three or more. When the objective function is not computed

sufficiently accurately, the numerical gradients are poor approximations to the true gradients and this will lead to poor performance of the quasi-Newton method. On the other hand, for statistical problems, great accuracy is seldom required; it is often suffice to obtain two or three significant digits, and we expect that in most of situations, we are not dealing with the worst cases.

Starting points for numerical optimization

In general, an objective function may have many local optima in addition to possibly a single global optimum. There is no numerical method which will always locate an existing global optimum, and the computational complexity in general increases either linearly or quadratically in the number of parameters. The best scenario is that we have a dependable method which converges to a local optimum based on initial guesses of the values which optimize the objective function. Thus good starting points for the numerical optimization methods are important. It is desirable to locate a good starting point based on a simple method, rather than trying many random starting points. An example based on method of moments estimation for deciding the starting points is for the multivariate Poisson-lognormal model (see Example 2.12), where the initial values for an estimate of μ_j and σ_j based on the the sample mean ($\bar{y}_j$), sample variance (s_j^2) and sample correlations (r_{jk}) are $\tilde{\sigma}_j^0 = \{\log[(s_j^2 - \bar{y}_j)/\bar{y}_j^2 + 1]\}^{1/2}$, $\tilde{\mu}_j^0 = \log \bar{y}_j - 0.5(\tilde{\sigma}_j^0)^2$ and $\tilde{\theta}_{jk}^0 = \log[r_{jk} s_j s_k/(\bar{y}_j \bar{y}_k) + 1]/(\tilde{\sigma}_j^0 \tilde{\sigma}_k^0)$ respectively. If the problem involves covariates, one can solve some linear equation systems with appropriately chosen covariate values to obtain initial values for the regression parameters. Initial values may also be obtained from prior knowledge of the study or by trial and error. Generally speaking, it is easier to have a good starting point for a model with interpretable parameters or parameters which are easily related to interpretable characteristics of the model. In the situations where closed-form moment characteristics of the model are not available, we may numerically compute the model moments.

Numerical integration

There are several methods for obtaining numerical integrals, among them are Romberg integration, adaptive integration and Monte-Carlo integration. The latter isg especially useful for high dimensional integration provided the accuracy requirements are modest. With the IFM approach, as often only one or two-dimensional integrations are needed, the necessary numerical integrations are not a problem in most cases. (Thus IFM can be considered as a tractable computational method, which alleviates the often extremely heavy computational burdens in fitting multivariate models.) For

this thesis work, an integration routine based on the Romberg integration method in Davis and Rabinowitz (1984) is used; this routine is good for two to about four dimensional integrations. A routine in Fortran code for computing the multinormal probability (or multinormal copula) can be found in Schervish (1984). Joe (1995) provides some good approximation methods for computing the multinormal cdf and rectangle probabilities.

2.8 Summary

In this chapter, two classes of models, MCD and MMD, are proposed and studied. The IFM is proposed as a parameter estimation and inference procedure, and its asymptotic properties are studied. Most of the results are for the the models with MUBE or PUBE properties. But the results of this chapter should apply to a very wide range of inference problems for numerous popular models in MCD and MMD classes. The IFME has the advantage of computational feasibility; this makes numerous models in the MCD and MMD classes practically useful. We also proposed a jackknife procedure for computing the asymptotic covariance matrix of the IFME, and demonstrated the asymptotic equivalence of this estimate to the Godambe information matrix. Based on the IFM approach, we also proposed estimation procedures for models with parameters common to more than one margin. One problem of great interest is that of determining the efficiency of the IFME relative to the conventional MLE. Clearly, a general comparison would be very difficult and next to impossible. Analytic and simulation studies on IFM efficiency will be given in Chapter 4. Another problem of interest is to see how the jackknife procedure compares with the Godambe information calculation numerically. A study of this is also given in Chapter 4. Our results may have several interesting extensions; some of these will be discussed in Chapter 7 as possible future research topics.

Chapter 3

Modelling of multivariate discrete data

In this chapter, we study some specific models in the class of MCD and MMD models and develop the corresponding IFM approach for fitting the model based on data. The models are first discussed for the case with no covariates and then for the case with covariates. For the dependence structure in the models, we distinguish the models with general dependence structure and the models with special dependence structure. Different ways to extend the models, especially to include covariates for the dependence parameters, are discussed.

This chapter is organized as the following. In section 3.1, we study MCD models for binary data, with emphasis on multivariate logit models and probit models for binary data. In section 3.2, we make some comparison of the models discussed in section 3.1. The general ideas in this section should also extend to other MCD and MMD models. In section 3.3, we study MCD models for count data, and in section 3.4, we study MCD models for ordinal data. MMD models for binary data are studied in section 3.5, and MMD models for count data are studied in section 3.6. Finally in section 3.7, we discuss the use of MCD and MMD models for longitudinal and repeated measures data. In each section, only a few parametric models are given, but many others can be derived. For data analysis examples with different models presented in this chapter, see Chapter 5.

3.1 Multivariate copula discrete models for binary data

3.1.1 Multivariate logit model

A multivariate logit model should be based on a multivariate logistic distribution. As there is no natural multivariate logistic distribution, we construct multivariate logit models by putting univariate logistic margins into a family of copulas with a wide range of dependence, and simple computational form if possible. As a result, multivariate logit models for binary data are obtained by letting $G_j(0) = 1/[1+\exp(z_j)]$ and $G_j(1) = 1$ in the model (2.13), with arbitrary copula C. Some choices of the copula C are:

- Multinormal copula

$$C(u_1, \ldots, u_d; \Theta) = \Phi_d(\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d); \Theta), \tag{3.1}$$

  where $\Theta = (\theta_{jk})$ is a correlation matrix (of normal margins). The bivariate margins of (3.1) are $C_{jk}(u_j, u_k) = \Phi_2(\Phi^{-1}(u_j), \Phi^{-1}(u_k); \theta_{jk})$

- Mixture of max-id copula (id for infinitely divisible, see Joe and Hu 1996)

$$C(\mathbf{u}) = \psi\left(-\sum_{j<k} \log K_{jk}(e^{-p_j\psi^{-1}(u_j)}, e^{-p_k\psi^{-1}(u_k)}) + \sum_{j=1}^{d} \nu_j p_j \psi^{-1}(u_j)\right), \tag{3.2}$$

  where K_{jk} are max-id copulas, $p_j = (\nu_j + d - 1)^{-1}$, $\nu_j > 0$. For interpretation, we may say that ψ can be considered as providing the minimal level of (pairwise) dependence, the copula K_{jk} adds some pairwise dependence to the global dependence, and ν_j's can be used for bivariate and multivariate asymmetry (the asymmetries are represented through $\nu_j/(\nu_j + \nu_k)$, $j \neq k$). (3.2) has the MUBE and partially CUOM properties. With $\psi(s) = -\theta^{-1}\log(1-[1-e^{-\theta}]e^{-s})$, $\theta > 0$,

$$C(\mathbf{u}) = -\theta^{-1}\log\left[1 - (1-e^{-\theta})\prod_{j<k} K_{jk}(x_j, x_k)\prod_{j=1}^{d} x_j^{\nu_j}\right], \tag{3.3}$$

  where $x_j = [(1-e^{-\theta u_j})/(1-e^{-\theta})]^{p_j}$, $p_j = (\nu_j + d - 1)^{-1}$. The bivariate margins of (3.3) are $C_{jk}(u_j, u_k) = -\theta^{-1}\log[1-(1-e^{-\theta})x_j^{\nu_j+d-2}x_k^{\nu_k+d-2}K_{jk}(x_j, x_k)]$. One good choice of K_{jk} would be $K_{jk}(x_j, x_k) = (x_j^{-\delta_{jk}} + x_k^{-\delta_{jk}} - 1)^{-1/\delta_{jk}}$, $0 \leq \delta_{jk} < \infty$; because the resulting copula is simple in form. (See Kimeldorf and Sampson (1975) and Joe (1993) for a discussion of this bivariate copula.)

- Molenberghs-Lesaffre (M-L) construction (Joe 1996). The construction generates a multivariate "copula" with general dependence structure. An example for a 4-dimensional copula is the following:

$$\eta_{1234} = \frac{x(x-a_0)\prod_{1\le j<k\le 4}(x-a_{jk})}{(C_{123}-x)(C_{124}-x)(C_{134}-x)(C_{234}-x)\prod_{j=1}^{4}(a_j-x)}, \tag{3.4}$$

where $x = C_{1234}$ is the copula, $a_0 = C_{123} + C_{124} + C_{134} + C_{234} - C_{12} - C_{13} - C_{14} - C_{23} - C_{24} - C_{34} + u_1 + u_2 + u_3 + u_4 - 1$, $a_1 = u_1 - C_{12} - C_{13} - C_{14} + C_{123} + C_{124} + C_{134}$, $a_2 = u_2 - C_{12} - C_{23} - C_{24} + C_{123} + C_{124} + C_{234}$, $a_3 = u_3 - C_{13} - C_{23} - C_{34} + C_{123} + C_{134} + C_{234}$, $a_4 = u_4 - C_{14} - C_{24} - C_{34} + C_{124} + C_{134} + C_{234}$, and $a_{jk} = C_{jkl} + C_{jkm} - C_{jk}$, for $1 \le j < k \le 4$ and $1 \le l, m \ne j, k \le 4$. In (3.4), C_{123}, C_{124}, C_{134} and C_{234} are compatible trivariate copulas such that

$$\eta_{jkl} = \frac{z b_1 b_2 b_3}{b_4 b_5 b_6 b_7}, \tag{3.5}$$

where $z = C_{jkl}$, $b_1 = u_j - C_{jk} - C_{jl} + z$, $b_2 = u_k - C_{jk} - C_{kl} + z$, $b_3 = u_l - C_{jl} - C_{kl} + z$, $b_4 = C_{jk} - z$, $b_5 = C_{jl} - z$, $b_6 = C_{kl} - z$ and $b_7 = 1 - u_j - u_k - u_l + C_{jk} + C_{jl} + C_{kl} - z$, for $1 \le j < k < l \le 4$. The bivariate copulas C_{12}, C_{13}, C_{14}, C_{23}, C_{24}, C_{34} are arbitrary compatible bivariate copulas. Examples are the bivariate normal, Plackett (2.8) or Frank copula (2.9); see Joe (1993, 1996) for a list of bivariate copula families with good properties. The parameters in C_{1234} are the parameters from the bivariate copulas, plus η_{jkl}, $1 \le j < k < l \le 4$, and η_{1234}. The generalization of this construction to arbitrary dimension can be found in Joe (1996). Notice that we have quotation marks on the word copula, because the multivariate object obtained from (3.4) and (3.5) or the corresponding form for general dimension has not been proven to be a proper multivariate copula. But they can be used for the parameter range that leads to positive orthant probabilities for the resulting probabilities for the multivariate binary vector.

- Morgenstern copula

$$C(u_1, \ldots, u_d) = \left[1 + \sum_{j<k}^{d} \theta_{jk}(1-u_j)(1-u_k)\right] \prod_{h=1}^{d} u_h, \tag{3.6}$$

where θ_{jk} must satisfy some constraints such that (3.6) is a proper copula. The bivariate margins of (3.6) are $C_{jk}(u_j, u_k) = [1 + \theta_{jk}(1-u_j)(1-u_k)]u_j u_k$, $|\theta_{jk}| \le 1$.

- The permutation symmetric copulas of the form

$$C(u_1, \ldots, u_d) = \phi\left(\sum_{i=1}^{d} \phi^{-1}(u_i)\right), \tag{3.7}$$

where $\phi : [0,\infty) \to [0,1]$ is strictly decreasing and continuously differentiable (of all orders), and $\phi(0) = 1$, $\phi(\infty) = 0$, $(-1)^j \phi^{(j)} \geq 0$. With $\psi(s) = -\theta^{-1}\log(1-[1-e^{-\theta}]e^{-s})$, $\theta > 0$, (3.7) is

$$C(u_1,\ldots,u_d) = \phi\left(\sum_{i=1}^{d}\phi^{-1}(u_i)\right) = -\frac{1}{\theta}\log\left(1-\frac{\prod_j(1-e^{-\theta u_j})}{(1-e^{-\theta})^{d-1}}\right). \tag{3.8}$$

This choice of $\psi(s)$ leads to 3.8 with bivariate marginal copulas that are reflection symmetric.

It is straightforward to extend the univariate marginal parameters to include covariates. For example, for z_{ij} corresponding to the random vector $\mathbf{Y}_i$, we can let $z_{ij} = \boldsymbol{\alpha}_j \mathbf{x}_{ij}$, where $\boldsymbol{\alpha}_j$ is a parameter vector and $\mathbf{x}_{ij}$ is a covariate vector corresponding to the jth margin. But what about the dependence parameters in the copula? Should the dependence parameters be functions of covariates? If so, what are the functions? These questions have no obvious answers. It is evident that if the dependence parameters are functions of covariates, the form of the functions will depend on the particular copula associated with the model. A simple way to deal with the dependence parameters is to let the dependence parameters in the copula be independent of covariates, and sometimes this may be good enough for the modelling purposes. If we need to include covariates for the dependence parameters, careful consideration should be given. In the following, in referring to specific copulas, we give some discussion on different ways of letting the dependence parameters depend on covariates:

- With the multinormal copula, the dependence structure in the copula for the ith response vector $\mathbf{Y}_i$ is $\Theta_i = (\theta_{i,jk})$. It is well-known that (i) Θ_i has to be nonnegative definite, and (ii) the component $\theta_{i,jk}$ of Θ_i has to be bounded by 1 in absolute value. Under these constraints, different ways of letting Θ_i depend on covariates are possible: (a) let $\theta_{i,jk} = [\exp(\boldsymbol{\beta}_{jk}\mathbf{w}_{i,jk}) - 1]/[\exp(\boldsymbol{\beta}_{jk}\mathbf{w}_{i,jk}) + 1]$; (b) let Θ_i have a simple correlation structure such as exchangeable and AR(1); (c) use a representation such as $z_{ij} = [\boldsymbol{\alpha}_j'\mathbf{x}_{ij}]/[1+\mathbf{x}_{ij}'\mathbf{x}_{ij}]^{1/2}$, $\theta_{i,jk} = r_{jk}/[(1+\mathbf{x}_{ij}'\mathbf{x}_{ij})(1+\mathbf{x}_{ik}'\mathbf{x}_{ik})]^{1/2}$; (d) use a more general representation such as $\theta_{i,jk} = r_{jk}\mathbf{w}_{i,jk}'\mathbf{w}_{i,jk}$; or (e) reparameterize Θ_i into the form of $d-1$ correlations and $(d-1)(d-2)/2$ partial correlations.

  The extension (a) satisfies condition (ii), but not necessarily (i). The extension (b) satisfies conditions (i) and (ii), but is only suitable for data with a special dependence structure. The extension (c) is more natural, as it is derived from a mixture representation (see section 3.5 for a more general form) and it satisfies condition (ii) and also condition (i) as long as the correlation matrix (r_{jk}) is nonnegative definite. This is an advantage in comparison with (a), as in (a), for (i) to be satisfied, all n correlation matrices must be nonnegative definite. The disadvantage of (c) is that the dependence range may be limited once a particular formal

representation is chosen. The extension (d) is similar to the the extension (c), except that now the $\theta_{i,jk}$s are not required to depend on the same covariate vectors as z_{ij}. The extension (e) completely avoids the matrix constraint (i), thus relieving a big burden on the constraint for the appropriate inclusion of covariate to the dependence parameters. But this extension implicitly implies an indexing of variables, which may not be rational in general, although this is not a problem in many applications as the indexing of variables is often evident, such as with time series; see Joe (1996).

- With the mixture of max-id copula (3.3), extensions to parameters θ, ν_j, δ_{jk} as functions of the covariates are straightforward. For example, for θ_i, ν_{ij}, $\delta_{i,jk}$ corresponding to the random vector $\mathbf{Y}_i$, we may have θ_i, ν_{ij} constant, and $\delta_{i,jk} = \exp(\boldsymbol{\beta}'_{jk}\mathbf{w}_{i,jk})$.

- With the Molenberghs-Lesaffre construction, the extension to include covariates is possible. In applications, it is often good enough to let the bivariate parameters $\delta_{i,jk}$ be function of covariates, such as $\delta_{i,jk} = \exp(\boldsymbol{\beta}'_{jk}\mathbf{w}_{i,jk})$ for bivariate Plackett copula, or Frank copula, and to let the higher order parameters be constant values, such as 1. This is a simple and useful approach, but there is no guarantee that this leads to compatible parameters. (See Joe 1996 for a maximum entropy interpretation in this case.)

- With the Morgenstern copula, the extension to let the parameters $\theta_{i,jk}$ be functions of some covariates is not easy, since the $\theta_{i,jk}$ must satisfy some constraints. This is rather complicated and difficult to manage when the dimension is high. The situation is very similar to the multinormal copula where Θ_i should be nonnegative definite.

- With the permutation symmetric copula (3.8), the extension to include covariates is to let θ_i be function of covariates, such as to let $\theta_i = \exp(\boldsymbol{\beta}'\mathbf{w}_i)$.

We see that for different copulas, there are many ways to extend the model to include covariates. Some are obvious and thus appear to be natural, others are not easy or obvious to specify. Note also that an exchangeable structure within the copula does not imply an exchangeable structure for the response variables. For MCD models for binary data, an exchangeable structure within the copula plus constant cut-off points across all the margins implies an exchangeable structure with the response variables. The AR dependence structure for the discrete response variables should be understood as latent Markov dependence structure (see section 3.7). When we mention an AR(1) (or AR) dependence structure, we are referring to the latent dependence structure within the multinormal copula.

In summary, under "multivariate logit models", many slightly different models are available. For example, we have multivariate logit model with

i. multinormal copula (3.1),

ii. multivariate Molenberghs-Lesaffre construction

   a. with bivariate normal copula,

   b. with Plackett copula (2.8),

   c. with Frank copula (2.9),

iii. mixture of max-id copula (3.3),

iv. Morgenstern copula (3.6),

v. the permutation symmetric copula (3.8).

Indeed, such multiple choices of models are available for any kind of MCD model. For a discrete random vector $\mathbf{Y}$, different copulas in (2.13) lead to different probabilistic models. The question is when is one model preferred to another? We will discuss this question in section 3.2. In the following, as an example, we examine estimation aspects of the multivariate logit model with multinormal copula.

The multivariate logit model with multinormal copula can also be called *multivariate normal-copula logit model* to highlight the fact that multinormal copula is used. For the case with no covariates, the estimating equations for multivariate normal-copula logit model based on IFM can be written as the following

$$\begin{cases} \Psi_{nj}(z_j) = [n_j(1)(1+\exp(-z_j)) - n_j(0)(1+\exp(-z_j))\exp(z_j)]\dfrac{\exp(-z_j)}{(1+\exp(-z_j))^2} = 0,\ j = 1,\ldots,d, \\ \Psi_{njk}(\theta_{jk}) = \left[\dfrac{n_{jk}(11)}{P_{jk}(11)} - \dfrac{n_{jk}(10)}{P_{jk}(10)} - \dfrac{n_{jk}(01)}{P_{jk}(01)} + \dfrac{n_{jk}(00)}{P_{jk}(00)}\right]\phi_2(\Phi^{-1}(u_j),\Phi^{-1}(u_k),\theta_{jk}) = 0, \\ \qquad\qquad 1 \le j < k \le d, \end{cases}$$

where $P_{jk}(11) = C_{jk}(u_j,u_k;\theta_{jk})$, $P_{jk}(10) = u_j - C_{jk}(u_j,u_k;\theta_{jk})$, $P_{jk}(11) = u_k - C_{jk}(u_j,u_k;\theta_{jk})$, $P_{jk}(11) = 1 - u_j - u_k + C_{jk}(u_j,u_k;\theta_{jk})$ with $C_{jk}(u_j,u_k;\theta_{jk}) = \Phi_2(\Phi^{-1}(u_j),\Phi^{-1}(u_k),\theta_{jk})$, $u_j = 1/[1+\exp(-z_j)]$, $u_k = 1/[1+\exp(-z_k)]$, and ϕ_2 is the BVN density. We obtain the estimates $\tilde{z}_j = \log(n_j(1)/n_j(0))$, $j = 1,\ldots,d$, and $\tilde{\theta}_{jk}$ is the root of the equation $\Phi_2(\Phi^{-1}(\tilde{u}_j),\Phi^{-1}(\tilde{u}_k),\tilde{\theta}_{jk}) = n_{jk}(11)/n$, $1 \le j < k \le d$, where $\tilde{u}_j = 1/(1+\exp(-\tilde{z}_j))$ and $\tilde{u}_k = 1/(1+\exp(-\tilde{z}_k))$.

For the situation with covariates, we may have the cut-off points depending on some covariate vectors, such that

$$z_{ij} = \alpha_{j0}x_{ij0} + \alpha_{j1}x_{ij1} + \cdots + \alpha_{jp_j}x_{ijp_j} = \boldsymbol{\alpha}_j'\mathbf{x}_{ij}, \tag{3.9}$$

where $x_{ij0} = 1$, and possibly

$$\theta_{i,jk} = \frac{\exp(\boldsymbol{\beta}_{jk}\mathbf{w}_{i,jk}) - 1}{\exp(\boldsymbol{\beta}_{jk}\mathbf{w}_{i,jk}) + 1}, \tag{3.10}$$

where $\boldsymbol{\beta}_{jk} = (b_{jk,0}, b_{jk,1}, \ldots, b_{jk,p_{jk}})$. We recognize that (3.10) is one of the form of functions of dependence parameters (in multinormal copula) on covariates that we discussed previously. We use this form of functions for the purpose of illustration. Other function forms can also be used instead. Because of the linearity of (3.9) and (3.10), the regression parameter vectors $\boldsymbol{\alpha}_j$, $\boldsymbol{\beta}_{jk}$ have marginal interpretations. The loglikelihood functions of margins are

$$\begin{cases} \ell_{nj}(\boldsymbol{\alpha}_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\boldsymbol{\alpha}_j, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij}y_{ik}), \quad 1 \le j < k \le d, \end{cases}$$

where

$$P_{i,j}(y_{ij}) = y_{ij}\frac{\exp(z_{ij})}{1+\exp(z_{ij})} + (1-y_{ij})\frac{1}{1+\exp(z_{ij})},$$

$$P_{i,jk}(y_{ij}y_{ik}) = \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) - \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk}) -$$
$$\Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) + \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk}),$$

where $a_{ij} = G_{ij}(y_{ij} - 1)$, $b_{ij} = G_{ij}(y_{ij})$, $a_{ik} = G_{ik}(y_{ik} - 1)$ and $b_{ik} = G_{ik}(y_{ik})$, with $G_{ij}(1) = 1$ and $G_{ij}(0) = 1/1 + \exp(z_{ij})$. We can apply quasi-Newton minimization to the loglikelihood functions of margins for getting the estimates of the parameters $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$. The Newton-Raphson method can also be used for getting the estimates of $\boldsymbol{\alpha}_j$ (what we used in our computer programs). In this case, we have to solve the estimating equations

$$\Psi_{nj} = \sum_{i=1}^{n} \frac{1}{P_j(y_{ij})}\frac{\partial P_j(y_{ij})}{\partial \boldsymbol{\alpha}_j} = \mathbf{0}.$$

For applying Newton-Raphson method, we need to calculate $\partial P(y_{ij})/\partial\alpha_{js}$ and $\partial^2 P(y_{ij})/\partial\alpha_{js}\partial\alpha_{jt}$. We have $\partial P(y_{ij})/\partial\alpha_{js} = (2y_{ij}-1)\{\exp(z_{ij})/(1+\exp(z_{ij}))^2\}x_{ijs}$, $s = 0, 1, \ldots, p_j$, and $\partial^2 P(y_{ij})/\partial\alpha_{js}\partial\alpha_{jt}$ $= (2y_{ij} - 1)\{\exp(z_{ij})(1 - \exp(z_{ij}))/(1 + \exp(z_{ij}))^3\}x_{ijs}x_{ijt}$, $s, t = 0, 1, \ldots, p_j$. For details about Newton-Raphson and quasi-Newton methods, see section 2.7.

M_Ψ and D_Ψ can be calculated and estimated by following the results in section 2.4. In applications, to avoid the tedious coding of M_Ψ and D_Ψ, we may use the jackknife technique to obtain the

asymptotic variance of $\tilde{z}_j$ and $\tilde{\theta}_{jk}$ in case there are no covariates, or that of $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_{jk}$ in case there are covariates.

3.1.2 Multivariate probit model

The general *multivariate probit model*, similar to that of multivariate logit model, is obtained by letting $G_j(0) = 1 - \Phi(z_j)$ and $G_j(1) = 1$ in the model (2.13). The *multivariate probit model* in Example 2.10 is the classical multivariate probit model discussed in the literature, where the copula in (2.13) is the multinormal copula. All the discussion of the multivariate logit model is relevant and can be directly applied to the multivariate probit model. For completeness, we give in the following some detailed discussion about the multivariate probit model when the copula is the multinormal copula, as a continuation of Example 2.10.

For the multivariate probit model in Example 2.10, it is easy to see that $\mathrm{E}(Y_j) = \Phi(z_j)$, $\mathrm{Var}(Y_j) = \Phi(z_j)(1-\Phi(z_j))$, $\mathrm{Cov}(Y_j, Y_k) = \Phi_2(z_j, z_k, \theta_{jk}) - \Phi(z_j)\Phi(z_k)$, $j \neq k$. The correlation of the response variable Y_j and Y_k is

$$\mathrm{Corr}(Y_j, Y_k) = \frac{\mathrm{Cov}(Y_j, Y_k)}{\{\mathrm{Var}(Y_j)\mathrm{Var}(Y_k)\}^{1/2}} = \frac{\Phi_2(z_j, z_k, \theta_{jk}) - \Phi(z_j)\Phi(z_k)}{\{\Phi(z_j)(1-\Phi(z_j))\Phi(z_k)(1-\Phi(z_k))\}^{1/2}}.$$

The variance of Y_j achieves its maximum when $z_j = 0$. In this case $\mathrm{E}(Y_j) = 1/2$, $\mathrm{Var}(Y_j) = 1/4$. If $z_j = 0, z_k = 0$, we have $\mathrm{Cov}(Y_j, Y_k) = (\sin^{-1}\theta_{jk})/(2\pi)$, and $\mathrm{Corr}(Y_j, Y_k) = (2\sin^{-1}\theta_{jk})/\pi$. Without loss of generality, assume $z_j \leq z_k$, then when θ_{jk} is at its boundary values,

$$\mathrm{Corr}(Y_j, Y_k) = \begin{cases} \{\Phi(z_j)(1-\Phi(z_k))/(1-\Phi(z_j))\Phi(z_k)\}^{1/2}, & \theta_{jk} = 1\ , \\ -\{(1-\Phi(z_j))(1-\Phi(z_k))/\Phi(z_j)\Phi(z_k)\}^{1/2}, & \theta_{jk} = -1,\ -z_k \leq z_j\ , \\ -\{\Phi(z_j)\Phi(z_k)/(1-\Phi(z_j))(1-\Phi(z_k))\}^{1/2}, & \theta_{jk} = -1,\ z_j \leq -z_k\ , \\ 0, & \theta_{jk} = 0\ . \end{cases}$$

From Fréchet bound inequalities,

$$-\min\{a^{1/2}, a^{-1/2}\} \leq \mathrm{Corr}(Y_j, Y_k) \leq \min\{b^{1/2}, b^{-1/2}\},$$

where $a = [P_j(1)P_k(1)]/[P_j(0)P_k(0)]$, $b = [P_j(1)P_k(0)]/[P_j(0)P_k(1)]$, we see that $\mathrm{Corr}(Y_j, Y_k)$ attains its upper and lower bound when $\theta_{jk} = 1$ and $\theta_{jk} = -1$ respectively. $\mathrm{Corr}(Y_j, Y_k)$ is an increasing function in θ_{jk}, it varies over its full range as θ_{jk} varies over its full range. Thus in a general situation a multivariate probit model of dimension d consists of d univariate probit models describing some marginal characteristics and $d(d-1)/2$ latent correlations θ_{jk}, $1 \leq j < k \leq d$, expressing the strength of the association among the response variables. $\theta_{jk} = 0$ corresponds to independence among the

response variables. The response variables are exchangeable if Θ has an exchangeable structure and the cut-off points are constant across all the margins. Note that when Θ has exchangeable structure, we must have $\theta_{jk} = \theta \geq -1/(d-1)$.

The estimating equations for the multivariate probit model with multinormal copula, based on n response vectors $\mathbf{y}_i$, $i = 1, \ldots, n$, are

$$\begin{cases} \Psi_{nj}(z_j) = \left(\dfrac{n_j(1)}{\Phi(z_j)} - \dfrac{n_j(0)}{1-\Phi(z_j)} \right) \phi(z_j) = 0, \quad j = 1, \ldots, d, \\ \Psi_{njk}(\theta_{jk}) = \left(\dfrac{n_{jk}(11)}{\Phi_2(z_j, z_k, \theta_{jk})} - \dfrac{n_{jk}(10)}{\Phi(z_j) - \Phi_2(z_j, z_k, \theta_{jk})} - \dfrac{n_{jk}(01)}{\Phi(z_k) - \Phi_2(z_j, z_k, \theta_{jk})} + \right. \\ \qquad \left. \dfrac{n_{jk}(00)}{1 - \Phi(z_j) - \Phi(z_k) + \Phi_2(z_j, z_k, \theta_{jk})} \right) \phi_2(z_j, z_k, \theta_{jk}) = 0, \quad 1 \leq j < k \leq d. \end{cases}$$

These lead to the solutions $\tilde{z}_j = \Phi^{-1}\left(n_j(1)/n\right)$, $j = 1, \ldots, d$, and $\tilde{\theta}_{jk}$ is the root of the equation $\Phi_2(\tilde{z}_j, \tilde{z}_k, \tilde{\theta}_{jk}) = n_{jk}(11)/n$, $1 \leq j < k \leq d$.

For the situation with covariates, the details are similar to the multivariate logit model with multinormal copula in the preceding subsection, except now we have

$$\begin{aligned} P_{i,j}(y_{ij}) &= y_{ij}\Phi(z_{ij}) + (1 - y_{ij})(1 - \Phi(z_{ij})), \\ P_{i,jk}(y_{ij}\, y_{ik}) &= \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) - \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk}) - \\ &\qquad \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) + \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk}), \end{aligned}$$

with $a_{ij} = G_{ij}(y_{ij} - 1)$, $b_{ij} = G_{ij}(y_{ij})$, $a_{ik} = G_{ik}(y_{ik} - 1)$ and $b_{ik} = G_{ik}(y_{ik})$, with $G_{ij}(1) = 1$ and $G_{ij}(0) = 1 - \Phi(z_{ij})$. We also have $\partial P(y_{ij})/\partial \alpha_{js} = (2y_{ij} - 1)\phi(z_{ij})x_{ijs}$, $s = 0, 1, \ldots, p_j$ and $\partial^2 P(y_{ij})/\partial \alpha_{js} \partial \alpha_{jt} = (1 - 2y_{ij})\phi(z_{ij}) z_{ij} x_{ijs} x_{ijt}$, $s, t = 0, 1, \ldots, p_j$; these expressions are needed for applying the Newton-Raphson method to get estimates of $\boldsymbol{\alpha}_j$.

M_Ψ and D_Ψ can be calculated and estimated by following the results in section 2.4. For example, for the case with no covariates, we have $\mathrm{E}(\psi^2(z_j)) = \phi^2(z_j)/\{\Phi(z_j)(1 - \Phi(z_j))\}$ and $\mathrm{E}(\psi^2(\theta_{jk})) = [1/P_{jk}(11) + 1/P_{jk}(10) + 1/P_{jk}(01) + 1/P_{jk}(00)][\partial P_{jk}(11)/\partial \theta_{jk}]^2$, where $\partial P_{jk}(11)/\partial \theta_{jk} = \partial \Phi_2(z_j, z_k, \theta_{jk})/\partial \theta_{jk} = \phi_2(z_j, z_k, \theta_{jk})$, a result due to Plackett (1954). In applications, to avoid the tedious computer coding of M_Ψ and D_Ψ, we may use the jackknife technique to obtain the asymptotic variance of $\tilde{z}_j$ and $\tilde{\theta}_{jk}$ in case there are no covariates, or that of $\tilde{\boldsymbol{\alpha}}_j$ and $\tilde{\boldsymbol{\beta}}_{jk}$ in case there are covariates.

3.2 Comparison of models

We obtain many models under the name of multivariate logit model (also multivariate probit model) for binary data. An immediate question is when is one model preferred to another?

In section 1.3, we outlined some desirable features of multivariate models; among them (2) and (3) may be the most important. But in applications, the importance of a particular desirable feature of multivariate model may well depend on the practical need and constraints. As an example, we briefly compare the multivariate logit models and the multivariate probit models with different copulas studied in the section 3.1.

The multivariate logit model with multinormal copula satisfies the desirable properties (1), (2), (3) and (4) of a multivariate model outlined in section 1.3, but not (5). The multivariate probit model with multinormal copula is similar, except that one has logit univariate margins and the other has probit univariate margins. In applications, the multivariate logit model with multinormal copula may be preferred to the multivariate probit model with multinormal copula, as the multivariate logit model with multinormal copula has the advantage of having a closed form univariate marginal cdf. This consideration also leads to the general preference of multivariate logit model to multivariate probit model when both have the same associated multivariate copula. For this reason, in the following, we concentrate on discussion of multivariate logit models.

The multivariate logit model with the mixture of max-id copula (3.3) satisfies the desirable properties (1), (3) and (5) of a multivariate model outlined in section 1.3, but only partially (2) and (4). The model only admits positive dependence (otherwise, it is flexible and wide in terms of dependence range) and it is CUOM(k) ($k \geq 2$) but not CUOM. The closed form cdf of this model is a very attractive feature. If the data exhibit only positive dependence (or prior knowledge tells us so), then the multivariate logit model with mixture of max-id copula (3.3) may be preferred to the multivariate logit model with multinormal copula.

The multivariate logit model with the M-L construction satisfies the desirable properties (1), (2), (3) and (4) of a multivariate model outlined in section 1.3, but not (5). The computation of the cdf may be easier numerically than that of multivariate logit model with multinormal copula since the former only requires solving a set of polynomial equations, but the latter requires multiple integration. The disadvantage with this model, as stated earlier, is that the object from the construction has not been proven to be a proper multivariate copula. What has been verified numerically (see Joe 1996) is that (3.4) and its extensions do not yield proper distributions if η_{1234} and η_{jkl} ($1 \leq j < k < l \leq 4$) are either too small or too large. In any case, the useful thing about this

model is that it leads to multivariate objects with given proper univariate and bivariate margins.

The multivariate logit model with the Morgenstern copula satisfies the desirable properties (1), (4) and (5) of a multivariate model outlined in section 1.3, but not (2) and (3). This is a major drawback. Thus this model is not very useful.

The multivariate logit models with the permutation symmetric copulas (3.7) are only suitable for the modelling of data with special exchangeable dependence patterns. They cannot be considered as widely applicable models, because the desirable property (2) of multivariate models is not satisfied. Nevertheless, this model may be one of the interesting considerations in some applications, such as when the data to be modelled are repeated measures over different treatments, or familial data.

In summary, for general applications, the multivariate logit model with the multinormal copula or the mixture of max-id copula (3.3) may be preferred. If the condition of positive dependence holds in a study, then the multivariate logit model with the mixture of max-id copula (3.3) may be preferred to the multivariate logit model with multinormal copula because the former has a closed form multivariate cdf; this is particularly attractive for moderate to large dimension of response, d. The multivariate logit model with Molenberghs-Lesaffre construction may be another interesting option. When several models fit the data about equally well, a preference for one should be based on which desirable feature is considered important to the successful application of the models. In many situations, several equally good models may be possible; see Chapter 5 for discussion and data analysis examples.

In the statistical literature, the multivariate probit model with multinormal copula has been studied and applied. An early reference on an application to binary data is Ashford and Sowden (1970). An explanation of the popularity of multivariate probit model with multinormal copula is that the model is related to the multivariate normal distribution, which allows the multivariate probit model to accommodate the dependence in its full range for the response variables. Furthermore, marginal models follow the simple univariate probit models.

3.3 Multivariate copula discrete models for count data

Univariate count data may be modelled by binomial, negative binomial, logarithmic, Poisson, or generalized Poisson distributions, depending on the amount of dispersion. In this section, we study some MCD models for multivariate count data.

3.3.1 Multivariate Poisson model

The *multivariate Poisson model* for Poisson count data is obtained by letting $G_j(y_j) = \sum_{m=0}^{[y_j]} p_j^{(m)}$, $y_j = 0, 1, 2, \ldots, \infty$, $j = 1, 2, \ldots, d$, in the model (2.13), where $p_j^{(m)} = [\lambda_j^m \exp(-\lambda_j)]/m!$, $\lambda_j > 0$. The copula C in (2.13) is arbitrary. Copulas (3.1)–(3.8) are some interesting choices here.

The multivariate Poisson model has univariate Poisson marginals. We have $\mathrm{E}(Y_j) = \mathrm{Var}(Y_j) = \lambda_j$, which is a characterizing feature of the Poisson distribution called *equidispersion*. There are situations where the variance of count data is greater than the mean, or the variance is smaller than the mean. The former case is called *overdispersion* and the latter case is called *underdispersion*. We will see models dealing with overdispersion and underdispersion in the subsequent sections. Although the multivariate Poisson model has Poisson univariate marginal distribution, the conditional distributions are not Poisson.

The univariate parameter λ_j in the multivariate Poisson model can be reparameterized by taking $\eta_j = \log(\lambda_j)$ so that the new parameter η_j has the range $(-\infty, \infty)$. It is also straightforward to extend the univariate marginal parameters to include covariates. For example, for λ_{ij} corresponding to random vector $\mathbf{Y}_i$, we can let $\lambda_{ij} = \exp(\boldsymbol{\alpha}_j' \mathbf{x}_{ij})$, where $\boldsymbol{\alpha}_j$ is a parameter vector and $\mathbf{x}_{ij}$ is a covariate vector corresponding to the jth margin. The discussion on modelling the dependence parameters in the copulas in section 3.1 are also relevant here. Most of the discussion in section 3.2 about the comparisons of models is also relevant here as the comparison is essentially the comparison of the associated multivariate copulas.

In summary, under the name "multivariate Poisson models", we may have multivariate Poisson model with

i. multinormal copula (3.1),

ii. multivariate Molenberghs-Lesaffre construction

   a. with bivariate normal copula,

   b. with Plackett copula (2.8),

   c. with Frank copula (2.9),

iii. mixture of max-id copula (3.3),

iv. Morgenstern copula (3.6),

v. the permutation symmetric copula (3.8).

These are similar to the multivariate logit models for binary data.

For illustration purposes, in the following, we provide some details on the multivariate Poisson model with the multinormal copula. The multivariate Poisson model with the multinormal copula can also be called the *multivariate normal-copula Poisson model.* This model was already introduced in Example 2.11. For the multivariate normal-copula Poisson model, the Fréchet upper bound is reached in the limit if $\Theta = J$, where J is matrix of 1's. In fact, when $\theta_{jk} = 1$ and $\lambda_j = \lambda_k$, the correlation of the response variable Y_j and Y_k is

$$\text{Corr}(Y_j, Y_k) = \frac{\sum_{\{y_j\}} y_j^2 P(y_j) - \lambda_j^2}{\lambda_j} = \frac{\lambda_j + \lambda_j^2 - \lambda_j^2}{\lambda_j} = 1.$$

When Θ has an exchangeable structure and λ_j does not depend on j, then there is also an exchangeable correlation structure in the response vector $\mathbf{Y}$.

The loglikelihood functions of margins for the parameters $\boldsymbol{\lambda}$ and Θ based on n observed response vectors $\mathbf{y}_i$, $i = 1, \ldots, n$, are:

$$\begin{cases} \ell_{nj}(\lambda_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d, \end{cases} \tag{3.11}$$

where $P_j(y_{ij}) = \lambda_j^{y_{ij}} \exp(-\lambda_j)/y_{ij}!$ and $P_{jk}(y_{ij}y_{ik}) = \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) - \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk}) - \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(b_{ik}); \theta_{jk}) + \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(a_{ik}); \theta_{jk})$, where $a_{ij} = G_{ij}(y_{ij} - 1)$, $b_{ij} = G_{ij}(y_{ij})$, $a_{ik} = G_{ik}(y_{ik} - 1)$ and $b_{ik} = G_{ik}(y_{ik})$, with $G_{ij}(y_{ij}) = \sum_{x=0}^{y_{ij}} P_j(x)$ and $G_{ik}(y_{ik}) = \sum_{x=0}^{y_{ik}} P_k(x)$. The estimating equations based on IFM are

$$\begin{cases} \Psi_{nj}(\lambda_j) = \sum_{i=1}^{n} \frac{1}{P_j(y_{ij})} \frac{\partial P_j(y_{ij})}{\partial \lambda_j} = 0, \quad j = 1, \ldots, d, \\ \Psi_{njk}(\theta_{jk}) = \sum_{i=1}^{n} \frac{1}{P_{jk}(y_{ij}y_{ik})} \frac{\partial P_{jk}(y_{ij}y_{ik})}{\partial \theta_{jk}} = 0, \quad 1 \le j < k \le d, \end{cases} \tag{3.12}$$

which lead to $\tilde{\lambda}_j = \sum_{i=1}^{n} y_{ij}/n$, and $\tilde{\theta}_{jk}$ can be found through numerical computation. An extension of the multivariate normal-copula Poisson model with covariate $\mathbf{x}_{ij}$ for the response observation $\mathbf{y}_{ij}$ is to let $\lambda_{ij} = h_j(\boldsymbol{\gamma}_j, \mathbf{x}_{ij})$ for some function h_j in the range $[0, \infty)$. An example of the function h_j is $\lambda_{ij} = \exp(\boldsymbol{\gamma}_j' \mathbf{x}_{ij})$ (or $\log(\lambda_{ij}) = \boldsymbol{\gamma}_j' \mathbf{x}_{ij}$). The ways to let the dependence parameters θ_{jk} be functions of covariates follow the discussion in section 3.1 for the multivariate logit model with multinormal copula. We can apply quasi-Newton minimization to the loglikelihood functions of margins (3.11) to obtain the estimates of the parameters $\boldsymbol{\gamma}_j$ and the dependence parameters θ_{jk} (or the regression

parameters for the dependence parameters if applicable). The Newton-Raphson method can also be used to obtain the estimates of $\boldsymbol{\gamma}_j$ from $\Psi_{nj}(\lambda_j) = 0$. Let $\log(\lambda_{ij}) = \gamma_{j0} + \gamma_{j1}x_{ij1} + \cdots + \gamma_{jp_j}x_{ijp_j}$ For applying the Newton-Raphson method, we need to calculate $\partial P(y_{ij})/\partial\gamma_{js}$ and $\partial^2 P(y_{ij})/\partial\gamma_{js}\partial\gamma_{jt}$. If we let $x_{ij0} = 1$, we have $\partial P(y_{ij})/\partial\gamma_{js} = \{\lambda_{ij}^{y_{ij}}\exp(-\lambda_{ij})/y_{ij}!\}[y_{ij} - \lambda_{ij}]x_{ijs}$, $s = 0, 1, \ldots, p_j$, and $\partial^2 P(y_{ij})/\partial\gamma_{js}\partial\gamma_{jt} = \{\lambda_{ij}^{y_{ij}}\exp(-\lambda_{ij})/y_{ij}!\}[(y_{ij} - \lambda_{ij})^2 - \lambda_{ij}]x_{ijs}x_{ijt}$, $s, t = 0, 1, \ldots, p_j$. For details about numerical methods, see section 2.7.

3.3.2 Multivariate generalized Poisson model

The *multivariate generalized Poisson model* for count data is obtained by letting $G_j(y_j) = \sum_{s=0}^{[y_j]} p_j^{(s)}$, $y_j = 0, 1, 2, \ldots, \infty$, $j = 1, 2, \ldots, d$, in the model (2.13), where

$$p_j^{(s)} = \begin{cases} \dfrac{\lambda_j(\lambda_j + s\alpha_j)^{s-1}\exp(-\lambda_j - s\alpha_j)}{s!}, & s = 0, 1, 2, \ldots, \\ 0 \quad \text{for } s > m, \text{ when } \alpha_j < 0, \end{cases} \tag{3.14}$$

where $\lambda_j > 0$, $\max(-1, -\lambda_j/m) < \alpha_j \leq 1$ and m (≥ 4) is the largest positive integer for which $\lambda_j + m\alpha_j > 0$ when α_j is negative. The copula C in (2.13) is arbitrary. The copulas (3.1)–(3.8) are some choices here.

The multivariate generalized Poisson model has as its jth ($j = 1, \ldots, d$) margin the generalized Poisson distribution with pmf (3.14). This generalized Poisson distribution is extensively studied in a monograph by Consul (1989). Its main characteristic is that it allows for both overdispersion and underdispersion by introducing one additional parameter α_j. The generalized Poisson distribution has the Poisson distribution as a special case when $\alpha_j = 0$. The mean and variance for Y_j are $\mathrm{E}(Y_j) = \lambda_j(1 - \alpha_j)^{-1}$ and $\mathrm{Var}(Y_j) = \lambda_j(1 - \alpha_j)^{-3}$, respectively. Thus, the generalized Poisson distribution displays overdispersion for $0 < \alpha_j < 1$, equidispersion for $\alpha_j = 0$ and underdispersion for $\max(-1, \lambda_j/m) < \alpha_j \leq 0$. The restrictions leading to underdispersion are rather complicated, as the parameters α_j are restricted by the sample space. It is easier to work with the overdispersion situation where the restrictions are simply $\lambda_j > 0$, $0 < \alpha_j \leq 1$.

The details of applying the IFM procedure to the generalized Poisson model are similar to that of the multivariate Poisson model. For the situation with no covariates, the univariate estimating equations for the parameters λ_j and α_j, $j = 1, \ldots, d$, are

$$\begin{cases} \Psi_{nj}(\lambda_j) = \displaystyle\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})}\frac{\partial P_j(y_{ij})}{\partial\lambda_j} = 0, \\ \Psi_{nj}(\alpha_j) = \displaystyle\sum_{i=1}^{n} \frac{1}{P_j(y_{ij})}\frac{\partial P_j(y_{ij})}{\partial\alpha_j} = 0. \end{cases} \tag{3.15}$$

They lead to

$$\begin{cases} \sum_{i=1}^{n} \dfrac{n y_{ij}(y_{ij}-1)}{y_{+j}+(n y_{ij}-y_{+j})\alpha_j} - y_{+j} = 0, \\ y_{+j}(1-\alpha_j) - n\lambda_j = 0, \end{cases}$$

where $y_{+j} = \sum_{i=1}^{n} y_{ij}$. When there is a covariate vector $\mathbf{x}_{ij}$ for the response observation $\mathbf{y}_{ij}$, we may let $\lambda_{ij} = a_j(\boldsymbol{\gamma}_j, \mathbf{x}_{ij})$ for some function a_j in the range $[0,\infty)$, and let $\alpha_{ij} = b_j(\boldsymbol{\eta}_j, \mathbf{x}_{ij})$ for some function b_j in the range $[0,1]$. An example is $\lambda_{ij} = \exp(\boldsymbol{\gamma}_j'\mathbf{x}_{ij})$ (or $\log(\lambda_{ij}) = \boldsymbol{\gamma}_j'\mathbf{x}_{ij}$) and $\alpha_{ij} = 1/[1+\exp(-\boldsymbol{\eta}_j'\mathbf{x}_{ij})]$. The discussion on modelling the dependence parameters in the copulas in section 3.1 is also appropriate here. Furthermore, most of the discussion in section 3.2 about the comparisons of models is also relevant here since the comparison is essentially the comparison of the associated multivariate copulas.

3.3.3 Multivariate negative binomial model

The *multivariate negative binomial model* for count data is obtained by letting $G_j(y_j) = \sum_{s=0}^{[y_j]} p_j^{(s)}$, $y_j = 0,1,2,\ldots,\infty$, $j = 1,2,\ldots,d$, in the model (2.13), where

$$p_j^{(s)} = \frac{\Gamma(\alpha_j+s)}{\Gamma(\alpha_j)\Gamma(s+1)} p_j^{\alpha_j}(1-p_j)^s, \quad s = 0,1,2,\ldots, \tag{3.16}$$

with $\alpha_j > 0$ and $0 < p_j < 1$. The mean and variance for Y_j are $\mathrm{E}(Y_j) = \alpha_j(1-p_j)/p_j$ and $\mathrm{Var}(Y_j) = \alpha_j(1-p_j)/p_j^2$, respectively. Since $\alpha_j > 0$, we see that this model allows for overdispersion. When there is a covariate vector $\mathbf{x}_{ij}$ for the response observation $\mathbf{y}_{ij}$, we may let $\alpha_{ij} = a_j(\boldsymbol{\gamma}_j, \mathbf{x}_{ij})$ for some function a_j in the range $[0,\infty)$, and let $p_{ij} = b_j(\boldsymbol{\eta}_j, \mathbf{x}_{ij})$ for some function b_j in the range $[0,1]$. See Lawless (1987) for another way to deal with covariates. Other details are similar to that of the multivariate generalized Poisson model.

3.3.4 Multivariate logarithmic series model

The *multivariate logarithmic series model* for count data is obtained by letting $G_j(y_j) = \sum_{s=1}^{[y_j]} p_j^{(s)}$, $y_j = 0,1,2,\ldots,\infty$, $j = 1,2,\ldots,d$, in the model (2.13), where

$$p_j^{(s)} = \alpha_j p_j^s/s, \quad s = 1,2,\ldots, \tag{3.17}$$

with $\alpha_j = -[\log(1-p_j)]^{-1}$ and $0 < p_j < 1$. The mean and variance for Y_j are $\mathrm{E}(Y_j) = \alpha_j p_j/(1-\alpha_j)$ and $\mathrm{Var}(Y_j) = \alpha_j p_j(1-\alpha_j p_j)/(1-p_j)^2$, respectively. This model allows for overdispersion when $p_j > 1-e^{-1}$ and underdispersion when $p_j < 1-e^{-1}$. Note that for this model to allow a zero count, we need a shift of one such that $p_j^{(t)} = \alpha_j p_j^{t+1}/(t+1)$ for $t = 0,1,2,\ldots$.

For the situation where there is a covariate vector $\mathbf{x}_{ij}$ for the response observation $\mathbf{y}_{ij}$, we may let $p_{ij} = F_j(\boldsymbol{\gamma}_j, \mathbf{x}_{ij})$ where F_j is a univariate cdf.

An unattractive feature of this model is that $p_j^{(s)}$ is a decreasing function of s, which may not be suitable in many applications.

3.4 Multivariate copula discrete models for ordinal data

In this section, we shall discuss the modelling of multivariate ordinal categorical data with multivariate copula discrete (MCD) models. We first briefly discuss some special features of ordinal categorical data before we introduce the general MCD model for ordinal data and some specific models.

When a polytomous variable has an ordered structure, we may assume the existence of a latent continuous random variable that measures the level of the ordered polytomous variable. For a binary variable, models for ordered data and unordered data are equivalent, but for categories variables with more than 2 categories, ordered data and unordered data are quite different. The modelling of unordered data is not as straightforward as the modelling of ordered data. This is especially so in the multivariate situation, where it is not obvious how to model the dependence structure of unordered data. We will discuss briefly the modelling of multivariate polytomous unordered data in Chapter 7. One aspect of ordinal data worth noticing is that it is possible to combine one category with an adjacent category for data analysis. But this practice may not be as meaningful for unordered categorical data, since the notion of adjacent category is not meaningful, and arbitrary clumping of categories may be unsatisfactory.

We next introduce the MCD model for ordinal data. Consider d-dimension ordinal categorical random vectors $\mathbf{Y}$ with m_j categories for the jth margin ($j = 1, 2, \ldots, d$) and with the categories coded as $1, 2, \ldots, m_j$. For the jth margin, the outcome y_j can take values $1, 2, \ldots, m_j$, where m_j can differ with the index j. For Y_j, suppose the probability of outcome s, $s = 1, 2, \ldots, m_j$, is $p_j^{(s)}$. We define

$$G_j(y_j) = \begin{cases} 0, & y_j < 1, \\ \sum_{s=1}^{[y_j]} p_j^{(s)}, & 1 \le y_j < m_j, \\ 1, & y_j \ge m_j \ , \end{cases} \tag{3.18}$$

where $[y_j]$ means the largest integer less or equal than y_j. For a given d-dimensional copula $C(u_1, \ldots, u_d; \boldsymbol{\theta})$, $C(G_1(y_1), \ldots, G_d(y_d); \boldsymbol{\theta})$ is a well-defined distribution for the ordinal random vector

$\mathbf{Y}$. The pmf of $\mathbf{Y}$ is

$$P(y_1 \cdots y_d) = \sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(a_{1i_1}, \ldots, a_{di_d}; \boldsymbol{\theta}), \tag{3.19}$$

where $a_{j1} = G_j(y_j - 1)$, $a_{j2} = G_j(y_j)$. (3.19) is called the *multivariate copula discrete models for ordinal data.*

Since Y_j is an ordered categorical variable, one simple way to reparameterize $p_j^{(s)}$, so that the new parameter associated to the univariate margin has the range in the entire space, is to let $G_j(y_j) = F_j(z_j(y_j))$ where F_j is a cdf of a continuous random variable Z_j. Thus $p_j^{(s)} = F_j(z_j(y_j^{(s)})) - F_j(z_j(y_j^{(s-1)}))$. This is equivalent to

$$\begin{cases} Y_j = 1 & \text{iff } z_j(0) < Z_j \leq z_j(1), \\ Y_j = 2 & \text{iff } z_j(1) < Z_j \leq z_j(2), \\ \cdots & \cdots, \\ Y_j = m_j & \text{iff } z_j(m_j - 1) < Z_j \leq z_j(m_j), \end{cases} \tag{3.20}$$

where $-\infty = z_j(0) < z_j(1) < \cdots < z_j(m_j - 1) < z_j(m_j) = \infty$ are constants, $j = 1, 2, \ldots, d$, and the random vector $\mathbf{Z} = (Z_1, \ldots, Z_d)'$ has a multivariate cdf $F_{12\cdots d}$. In the literature, the representation in (3.20) is referred to as modelling $\mathbf{Y}$ through the latent random vector $\mathbf{Z}$, and the parameter $z_j(y_j)$ is called the y_jth cut-off point for the random variable Z_j.

As for the MCD model for binary data, the choices of F_j are abundant. It can be standard logistic, normal, extreme value, gamma, lognormal cdf, and so on. Furthermore, $F_j(z_j)$ need not to be in the same distribution family for different j. Similarly, in terms of the form of copula $C(u_1, \ldots, u_d; \Theta)$, it can be the multivariate normal copula, a mixture of max-id copula, the Molenberghs-Lesaffre construct, the Morgenstern copula and a permutation symmetric copula.

It is also possible to express the parameters $z_j(y_j)$ as functions of covariates, as we will see through examples. For the dependence parameters $\boldsymbol{\theta}$ in the copula $C(u_1, \ldots, u_d; \boldsymbol{\theta})$, there is also the option of including covariates. The discussion in section 3.1 on the extension for letting the dependence parameters in the copulas be functions of covariates is also relevant here, since this only depends on the associated copulas. In the following, in parallel with the multivariate models for binary data, we will see some examples of multivariate models for ordinal data.

3.4.1 Multivariate logit model

The multivariate logit model for ordinal data is obtained by letting $G_j(y_j) = \exp(z_j(y_j))/[1 + \exp(z_j(y_j)]$ in (3.19), where $-\infty = z_j(0) < z_j(1) < \cdots < z_j(m_j - 1) < z_j(m_j) = \infty$ are constants,

$j = 1, 2, \ldots, d$. It is equivalent letting $F_j(z) = \exp(z)/[1+\exp(z)]$, or choosing F_j to be the standard logistic cdf. The copula C in (3.19) is arbitrary. The copulas (3.1)–(3.8) are some choices here.

It is relatively straightforward to extend the univariate marginal parameters to include covariates. For example, for z_{ij} corresponding to random vector $\mathbf{Y}_i$, we can let $\mathbf{z}_{ij}(y_{ij}) = \gamma_j(y_{ij}) + g_j(\boldsymbol{\alpha}_j, \mathbf{x}_{ij})$, for some constants $-\infty = \gamma_j(1) < \gamma_j(2) < \cdots < \gamma_j(m_j - 1) < \gamma_j(m_j) = \infty$, and some function g_j in the range of $(-\infty, \infty)$. An example of the function g_j is $g_j(x) = x$. As we have discussed for the multivariate copula discrete models for binary data, a simple way to deal with the dependence parameters is to let the dependence parameters in the copula be independent of covariates. To extend the model to let the dependence parameters be functions of covariates requires specific knowledge of the associated copula C. The discussion on the extension of letting the dependence parameters in the copulas be functions of covariates for the multivariate logit models for binary data in section 3.1 are also relevant here. As with the multivariate logit models for binary data, we may also have multivariate logit model for ordinal data with

i. multinormal copula (3.1),

ii. multivariate Molenberghs-Lesaffre construction

   a. with bivariate normal copula,

   b. with Plackett copula (2.8),

   c. with Frank copula (2.9),

iii. mixture of max-id copula (3.3),

iv. Morgenstern copula (3.6),

v. the permutation symmetric copula (3.8).

For illustrative purposes, we give some details on the multivariate logit model with multinormal copula for ordinal data. The multivariate logit model with multinormal copula for ordinal data is also called the *multivariate normal-copula logit model* for ordinal data. Let the data be $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$. For the situation with no covariates, there are $\sum_{j=1}^{d}(m_j - 1)$ univariate parameters and $d(d-1)/2$ dependence parameters. The estimating equations from (2.42) are

$$\begin{cases} \Psi_{nj}(z_j(y_j)) = \left(\dfrac{n_j(y_j)}{F(z_j(y_j)) - F(z_j(y_j - 1))} - \dfrac{n_j(y_j + 1)}{F(z_j(y_j + 1)) - F(z_j(y_j))} \right) \dfrac{\exp(-z_j(y_j))}{(1 + \exp(-z_j(y_j)))^2} = 0, \\ \Psi_{njk}(\theta_{jk}) = \displaystyle\sum_{y_j=1}^{m_j} \sum_{y_k=1}^{m_k} \frac{n(y_j y_k)}{P_{jk}(y_j y_k)} \frac{\partial P_{jk}(y_j y_k)}{\partial \theta_{jk}} = 0, \end{cases}$$

where $F(z) = 1/[1 + \exp(-z)]$, and

$$P_{jk}(y_j y_k) = \Phi_2(\Phi^{-1}(u_j), \Phi^{-1}(u_k), \theta_{jk}) - \Phi_2(\Phi^{-1}(u_j^*), \Phi^{-1}(u_k), \theta_{jk}) - \Phi_2(\Phi^{-1}(u_j), \Phi^{-1}(u_k^*), \theta_{jk}) + \Phi_2(\Phi^{-1}(u_j^*), \Phi^{-1}(u_k^*), \theta_{jk})$$

with $u_j = 1/[1 + \exp(-z_j(y_j))]$, $u_k = 1/[1 + \exp(-z_k(y_k))]$, $u_j^* = 1/[1 + \exp(-z_j(y_j - 1))]$, $u_k^* = 1/[1 + \exp(-z_k(y_k - 1))]$. From $\Psi_{nj}(z_j(y_j)) = 0$, we obtain

$$F(z_j(y_j + 1)) - F(z_j(y_j)) = \frac{n_j(y_j + 1)}{n_j(y_j)} (F(z_j(y_j)) - F(z_j(y_j - 1))) = \frac{n_j(y_j + 1)}{n_j(1)} F(z_j(1)).$$

This implies that

$$\sum_{y_j=0}^{m_j-1} (F(z_j(y_j + 1)) - F(z_j(y_j))) = \sum_{y_j=0}^{m_j-1} n_j(y_j + 1) \frac{F(z_j(1))}{n_j(1)},$$

which leads to

$$F(z_j(1)) = \frac{n_j(1)}{n},$$

where $n = \sum_{y_j=1}^{m_j} n_j(y_j)$. It is thus easy to see that

$$F(z_j(y_j)) = \frac{\sum_{r=1}^{y_j} n_j(r)}{n},$$

which means that the estimate of $z_j(y_j)$ from IFM is

$$\tilde{z}_j(y_j) = \log \left(\frac{\sum_{r=1}^{y_j} n_j(r)}{n - \sum_{r=1}^{y_j} n_j(r)} \right).$$

The closed form of $\tilde{\theta}_{jk}$ is not available. We need to numerically solve $\Psi_{njk}(\theta_{jk}) = 0$ to find $\tilde{\theta}_{jk}$.

For the situation with covariate vector $\mathbf{x}_{ij}$ for the marginal parameters $z_j(y_{ij})$ for Y_{ij}, and a covariate vector $\mathbf{w}_{i,jk}$ for the dependence parameter $\theta_{i,jk}$, $i = 1, \ldots, n$, one way to extend the model to include the covariates is as follows:

$$\begin{cases} \mathbf{z}_{ij}(y_{ij}) = \gamma_j(y_{ij}) + \boldsymbol{\alpha}_j' \mathbf{x}_{ij}, & j = 1, 2, \ldots, d, \\ \theta_{i,jk} = \dfrac{\exp(\boldsymbol{\beta}_{jk}' \mathbf{w}_{i,jk}) - 1}{\exp(\boldsymbol{\beta}_{jk}' \mathbf{w}_{i,jk}) + 1}, & 1 \le j < k \le d. \end{cases} \tag{3.21}$$

The loglikelihood functions of margin for the parameter vectors $\boldsymbol{\gamma}_j = (\gamma_j(2), \ldots, \gamma_j(m_1 - 1))'$, $\boldsymbol{\alpha}_j$ $(j = 1, \ldots, d)$ and $\boldsymbol{\beta}_{jk}$ $(1 \le j < k \le d)$ are

$$\begin{cases} \ell_{nj}(\boldsymbol{\gamma}_j, \boldsymbol{\alpha}_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \\ \ell_{njk}(\boldsymbol{\gamma}_j, \boldsymbol{\gamma}_k, \boldsymbol{\alpha}_j, \boldsymbol{\alpha}_k, \boldsymbol{\beta}_{jk}) = \sum_{i=1}^{n} \log P_{ijk}(y_{ij} y_{ik}), \end{cases} \tag{3.22}$$

where $P_{i,j}(y_{ij}) = F(z_{ij}(y_{ij})) - F(z_{ij}(y_{ij}-1))$ and

$$P_{i,jk}(y_{ij}y_{ik}) = \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(b_{ik}); \theta_{ijk}) - \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(a_{ik}); \theta_{ijk}) - \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(b_{ik}); \theta_{ijk}) + \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(a_{ik}); \theta_{ijk}),$$

with $a_{ij} = F(z_j(y_{ij}-1))$, $b_{ij} = F(z_j(y_{ij}))$, $a_{ik} = F(z_k(y_{ik}-1))$ and $b_{ik} = F(z_k(y_{ik}))$. We can apply quasi-Newton minimization to the loglikelihood functions of margins (3.22) to obtain the estimates of the parameters $\boldsymbol{\gamma}_j$, $\boldsymbol{\alpha}_j$ and the dependence parameters θ_{jk} (or the regression parameters for the dependence parameters, $\boldsymbol{\beta}_{jk}$, if applicable). The Newton-Raphson method can also be used to obtain the estimates of $\boldsymbol{\gamma}_j$ from $\Psi_{nj}(\boldsymbol{\gamma}_j) = \mathbf{0}$, and the estimates of $\boldsymbol{\alpha}_j$ from $\Psi_{nj}(\boldsymbol{\alpha}_j) = \mathbf{0}$. For applying the Newton-Raphson method, we need to calculate $\partial P_j(y_{ij})/\partial \boldsymbol{\gamma}_j$, $\partial P_j(y_{ij})/\partial \boldsymbol{\alpha}_j$, $\partial^2 P_j(y_{ij})/\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\gamma}_j^T$, $\partial^2 P_j(y_{ij})/\partial \boldsymbol{\alpha}_j \partial \boldsymbol{\alpha}_j^T$ and $\partial^2 P_j(y_{ij})/\partial \boldsymbol{\gamma}_j \partial \boldsymbol{\alpha}_j^T$. The mathematical details for applying the Newton-Raphson method are the following. Let $z_{ij}(y_{ij}) = \gamma_j(y_{ij}) + \alpha_{j1}x_{ij1} + \cdots + \alpha_{jp_j}x_{ijp_j}$. For $y_{ij} \neq 1, m_j$, we have $P_j(y_{ij}) = \exp(z_{ij}(y_{ij}))/[1+\exp(z_{ij}(y_{ij}))] - \exp(z_{ij}(y_{ij}-1))/[1+\exp(z_{ij}(y_{ij}-1))]$, thus $\partial P_j(y_{ij})/\partial \alpha_{js} = \{[\exp(z_{ij}(y_{ij}))/(1+\exp(z_{ij}(y_{ij})))^2] - [\exp(z_{ij}(y_{ij}-1))/(1+\exp(z_{ij}(y_{ij}-1)))^2]\}x_{ijs}$, $s = 1, \ldots, p_j$, and $\partial^2 P_j(y_{ij})/\partial \alpha_{js}\partial \alpha_{jt} = \{[\exp(z_{ij}(y_{ij}))(1-\exp(z_{ij}(y_{ij})))/(1+\exp(z_{ij}(y_{ij})))^3] - [\exp(z_{ij}(y_{ij}-1))(1-\exp(z_{ij}(y_{ij}-1)))/(1+\exp(z_{ij}(y_{ij}-1)))^3]\}x_{ijs}x_{ijt}$, $s,t = 1, \ldots, p_j$. For $r = 1, 2, \ldots, m_j - 1$, we have

$$\frac{\partial P_j(y_{ij})}{\partial \gamma_j(r)} = \begin{cases} \frac{\exp(z_{ij}(y_{ij}))}{(1+\exp(z_{ij}(y_{ij})))^2} & \text{if } r = y_{ij}\ , \\ -\frac{\exp(z_{ij}(y_{ij}-1))}{(1+\exp(z_{ij}(y_{ij}-1)))^2} & \text{if } r = y_{ij}-1\ , \\ 0 & \text{otherwise}\ , \end{cases}$$

and for $r_1, r_2 = 1, 2, \ldots, m_j - 1$, we have

$$\frac{\partial^2 P_j(y_{ij})}{\partial \gamma_j(r_1)\partial \gamma_j(r_2)} = \begin{cases} \frac{\exp(z_{ij}(y_{ij}))(1-\exp(z_{ij}(y_{ij})))}{(1+\exp(z_{ij}(y_{ij})))^3} & \text{if } r_1 = r_2 = y_{ij}\ , \\ -\frac{\exp(z_{ij}(y_{ij}-1))(1-\exp(z_{ij}(y_{ij}-1)))}{(1+\exp(z_{ij}(y_{ij}-1)))^3} & \text{if } r_1 = r_2 = y_{ij}-1\ , \\ 0 & \text{otherwise}\ , \end{cases}$$

and

$$\frac{\partial^2 P_j(y_{ij})}{\partial \gamma_j(r)\partial \alpha_{js}} = \begin{cases} \frac{\exp(z_{ij}(y_{ij}))(1-\exp(z_{ij}(y_{ij})))}{(1+\exp(z_{ij}(y_{ij})))^3}x_{ijs} & \text{if } r = y_{ij}\ , \\ -\frac{\exp(z_{ij}(y_{ij}-1))(1-\exp(z_{ij}(y_{ij}-1)))}{(1+\exp(z_{ij}(y_{ij}-1)))^3}x_{ijs} & \text{if } r = y_{ij}-1\ , \\ 0 & \text{otherwise}\ . \end{cases}$$

For $y_{ij} = 1$, $P_j(y_{ij}) = \exp(z_{ij}(y_{ij}))/[1+\exp(z_{ij}(y_{ij}))]$ and for $y_{ij} = m_j$, $P_j(y_{ij}) = 1 - \exp(z_{ij}(y_{ij}-1))/[1+\exp(z_{ij}(y_{ij}-1))]$, thus corresponding slight modification on the above formulas should be made. For details about numerical methods, see section 2.7.

M_Ψ and D_Ψ can be calculated and estimated by following the results in section 2.4. In applications, to avoid the tedious coding of M_Ψ and D_Ψ, we may use the jackknife technique to obtain the

asymptotic variance of $\tilde{z}_j(y_j)$ and $\tilde{\theta}_{jk}$ when there is no covariates, or that of $\boldsymbol{\gamma}_j$, $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$ when there are covariates.

3.4.2 Multivariate probit model

Similar to the multivariate probit model for binary data, the general *multivariate probit model* is obtained by letting $G_j(y_j) = \Phi(z_j(y_j))$ in (3.19), where $-\infty = z_j(0) \leq z_j(1) \leq \cdots \leq z_j(m_j - 1) \leq z_j(m_j) = \infty$ are constants, $j = 1, 2, \ldots, d$. It is equivalent to letting $F_j(z) = \Phi(z)$, or choosing F_j to be the standard normal cdf. The copula C in (3.19) is arbitrary. The copulas (3.1)–(3.8) are some choices here. The multivariate probit model with multinormal copula for ordinal data is discussed in the literature (see for example Anderson and Pemberton 1985). The discussion of the multivariate logit model for ordinal data in the previous subsection is relevant and can be directly applied to the multivariate probit model for ordinal data. For completeness, we provide some detailed discussion for the multivariate probit model for ordinal data when the copula is the multinormal copula.

Let the data be $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$. For the situation with no covariates, there are $\sum_{j=1}^{d}(m_j - 1)$ univariate parameters and $d(d-1)/2$ dependence parameters. As for the multivariate logit model, with the IFM approach, we find that $\tilde{z}_j(y_j) = \Phi^{-1}(\sum_{r=1}^{y_j} n_j(r)/n)$, and $\tilde{\theta}_{jk}$ must be obtained numerically.

For the situation with covariate vector $\mathbf{x}_{ij}$ for the marginal parameters $z_j(y_{ij})$ for Y_{ij}, and a covariate vector $\mathbf{w}_{i,jk}$ for the dependence parameter $\theta_{i,jk}$, $i = 1, \ldots, n$, the details on IFM for parameter estimation are similar to the multivariate logit model for ordinal data in the preceding subsection. We here provide some mathematical details for this model. We have $P_{i,j}(y_{ij}) = \Phi(z_{ij}(y_{ij})) - \Phi(z_{ij}(y_{ij} - 1))$ and

$$\begin{aligned} P_{i,jk}(y_{ij}y_{ik}) =& \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(b_{ik}); \theta_{ijk}) - \Phi_2(\Phi^{-1}(b_{ij}), \Phi^{-1}(a_{ik}); \theta_{ijk}) - \\ & \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(b_{ik}); \theta_{ijk}) + \Phi_2(\Phi^{-1}(a_{ij}), \Phi^{-1}(a_{ik}); \theta_{ijk}), \end{aligned}$$

where $a_{ij} = \Phi(z_j(y_{ij} - 1))$, $b_{ij} = \Phi(z_j(y_{ij}))$, $a_{ik} = \Phi(z_k(y_{ik} - 1))$ and $b_{ik} = \Phi(z_k(y_{ik}))$. The mathematical details for applying the Newton-Raphson method are the following. For $y_{ij} \neq 1, m_j$, we have $P_j(y_{ij}) = \Phi(z_{ij}(y_{ij})) - \Phi(z_{ij}(y_{ij} - 1))$, thus $\partial P_j(y_{ij})/\partial \alpha_{js} = [\phi(z_{ij}(y_{ij})) - \phi(z_{ij}(y_{ij} - 1))]x_{ijs}$, $s = 1, \ldots, p_j$, and $\partial^2 P_j(y_{ij})/\partial \alpha_{js} \partial \alpha_{jt} = [-\phi(z_{ij}(y_{ij}))z_{ij}(y_{ij}) + \phi(z_{ij}(y_{ij} - 1))z_{ij}(y_{ij} - 1)]x_{ijs}x_{ijt}$, $s, t = 1, \ldots, p_j$. For $r = 1, 2, \ldots, m_j - 1$, we have

$$\frac{\partial P_j(y_{ij})}{\partial \gamma_j(r)} = \begin{cases} \phi(z_{ij}(y_{ij})) & \text{if } r = y_{ij}\ , \\ -\phi(z_{ij}(y_{ij} - 1)) & \text{if } r = y_{ij} - 1, \\ 0 & \text{otherwise}\ , \end{cases}$$

and for $r_1, r_2 = 1, 2, \ldots, m_j - 1$, we have

$$\frac{\partial^2 P_j(y_{ij})}{\partial\gamma_j(r_1)\partial\gamma_j(r_2)} = \begin{cases} -\phi(z_{ij}(y_{ij}))z_{ij}(y_{ij}) & \text{if } r_1 = r_2 = y_{ij}\ , \\ \phi(z_{ij}(y_{ij}-1))z_{ij}(y_{ij}-1) & \text{if } r_1 = r_2 = y_{ij} - 1, \\ 0 & \text{otherwise}\ , \end{cases}$$

and

$$\frac{\partial^2 P_j(y_{ij})}{\partial\gamma_j(r)\partial\alpha_{js}} = \begin{cases} -\phi(z_{ij}(y_{ij}))z_{ij}(y_{ij})x_{ijs} & \text{if } r = y_{ij}\ , \\ \phi(z_{ij}(y_{ij}-1))z_{ij}(y_{ij}-1)x_{ijs} & \text{if } r = y_{ij} - 1, \\ 0 & \text{otherwise}\ . \end{cases}$$

For $y_{ij} = 1$, $P_j(y_{ij}) = \Phi(z_{ij}(y_{ij}))$ and for $y_{ij} = m_j$, $P_j(y_{ij}) = 1 - \Phi(z_{ij}(y_{ij} - 1))$, thus the corresponding slight modification on the above formulas should be made. For details on numerical methods, see section 2.7.

M_Ψ and D_Ψ can be calculated and estimated by following the results in section 2.4. For example, for the case with no covariate, we have $E(\psi^2(z_j(y_j))) = \{[P_j(y_j+1) + P_j(y_j)]\phi^2(z_j(y_j))\}/\{P_j(y_j + 1)P_j(y_j)\}$, where $P_j(y_j) = \Phi(z_j(y_j)) - \Phi(z_j(y_j - 1))$, and so on. In applications, to avoid the tedious coding of M_Ψ and D_Ψ, we may use the jackknife technique to obtain the asymptotic variances of $\tilde{z}_j(y_j)$ and $\tilde{\theta}_{jk}$ in case there is no covariates, or those of $\boldsymbol{\gamma}_j$, $\boldsymbol{\alpha}_j$ and $\boldsymbol{\beta}_{jk}$ in case there are covariates.

The multivariate probit model with multinormal copula for ordinal data has been studied and applied in the literature. For example, Anderson and Pemberton (1985) used a trivariate probit model for the analysis of an ornithological data set on the three aspects of colouring of blackbirds.

3.4.3 Multivariate binomial model

In the previous subsections, we supposed that for Y_j, the probability of outcome s is $p_j^{(s)}$, $s = 1, 2, \ldots, m_j$, $j = 1, \ldots, d$, and we linked the m_j probabilities $p_j^{(s)}$ to $m_j - 1$ cut-off points $z_j(1), z_j(2), \ldots, z_j(m_j - 1)$. We keep as many independent parameters within the margins and between the margins as possible. In some situations, it is worthwhile to reduce the number of free parameters and obtain a more parsimonious model which may still capture the major features of the data and serve the inference purpose. One way to reduce the number of free parameters for the ordinal variable is to reparameterize the marginal distribution. Because $\sum_{s=1}^{m_j} p_j^{(s)} = 1$ and $p_j^{(s)} \geq 0$, we may let

$$p_j^{(s)} = \binom{m_j - 1}{s - 1} p_j^{s-1}(1 - p_j)^{m_j - s} \tag{3.23}$$

for some $0 < p_j < 1$. In other words, we assume that Y_j follows a binomial distribution $\mathrm{Bi}(m_j - 1, p_j)$. This reparameterization of the distribution of Y_j reduces the number of free parameter to one, namely p_j. The model constructed in this way is called the *multivariate binomial model* for ordinal data.

By treating $P_j(y_j)$ as a binomial probabilities, we need only deal with one parameter p_j for the jth margin. (3.23) is artificial for the ordinal data as s in (3.23) is based on letting Y_j take the integer values in $\{0, 1, \ldots, m_j\}$ as its category indicator. But s is a qualitative quantity; it should reflect the ordinal nature of the variable Y_j, not necessarily take on the integer values in $\{0, 1, \ldots, m_j\}$. In applications, if one feels justified in assuming the binomial behaviour in the sense of (3.23) for the univariate margin, then this model may be considered. (3.23) is a more natural assumption if the categorical outcome of each univariate response can be considered as the number of realizations of an event in a fixed number of random trials. In this situation, it is a MCD model for binomial count data. When there is a covariate vector $\mathbf{x}_{ij}$ for the response observation $\mathbf{y}_{ij}$, we may let $p_{ij} = b_j(\boldsymbol{\eta}_j, \mathbf{x}_{ij})$ for some function b_j in the range $[0, 1]$. Other details are similar to the multivariate logit model for binary data.

3.5 Multivariate mixture discrete models for binary data

The multivariate mixture discrete models (2.16) or (2.17) are flexible for the type of discrete data and for the multivariate structure by allowing different choices of copulas. However, they generally do not have closed form pmf or cdf. The choice of models should be based on the desirable features for multivariate models outlined in section 1.3, among them, (2) and (3) are considered to be essential.

In this and the next section, we study some specific MMD models. The mathematical development for other MMD models with different choices of copulas should be similar.

3.5.1 Multivariate probit-normal model

The *multivariate probit-normal model* for binary data is introduced in (2.32) in Example 2.13. Following the notation in Example 2.13, the corresponding cut-off points α_{ij} is $\alpha_{ij} = \boldsymbol{\beta}_j' \mathbf{x}_{ij}$ for a more general situation, where $\mathbf{x}_{ij}$ is a covariate vector, $j = 1, \ldots, d$ and $i = 1, \ldots, n$. Assume $\boldsymbol{\beta}_j \sim N_{p_j}(\boldsymbol{\mu}_j, \Sigma_j)$, $j = 1, \ldots, d$. Let $\boldsymbol{\gamma} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_d)' \sim N_q(\boldsymbol{\mu}, \Sigma)$, where $q = \sum p_j$, and $\mathrm{Cov}(\boldsymbol{\beta}_j, \boldsymbol{\beta}_k) = \Sigma_j$. From the stochastic representation in Example 2.13, we have

$$\begin{cases} z_{ij}^* = \dfrac{\boldsymbol{\mu}_j' \mathbf{x}_{ij}}{\{1 + \mathbf{x}_{ij}' \Sigma_j \mathbf{x}_{ij}\}^{1/2}}, \quad j = 1, \ldots, d, \\ r_{i,jk} = \dfrac{\theta_{jk} + \mathbf{x}_{ij}' \Sigma_{jk} \mathbf{x}_{ik}}{\{(1 + \mathbf{x}_{ij}' \Sigma_j \mathbf{x}_{ij})(1 + \mathbf{x}_{ik}' \Sigma_k \mathbf{x}_{ik})\}^{1/2}}, \quad j \neq k. \end{cases} \tag{3.24}$$

The jth and (j,k) marginal pmf are

$$\begin{aligned}
P_{i,j}(y_{ij}) &= y_{ij}\Phi(z_{ij}^*) + (1-y_{ij})(1-\Phi(z_{ij}^*)),\\
P_{i,jk}(y_{ij}y_{ik}) &= \Phi_2(\Phi^{-1}(b_{ij}),\Phi^{-1}(b_{ik});r_{i,jk}) - \Phi_2(\Phi^{-1}(b_{ij}),\Phi^{-1}(a_{ik});r_{i,jk}) -\\
&\qquad \Phi_2(\Phi^{-1}(a_{ij}),\Phi^{-1}(b_{ik});r_{i,jk}) + \Phi_2(\Phi^{-1}(a_{ij}),\Phi^{-1}(a_{ik});r_{i,jk}),
\end{aligned}$$

where $a_{ij} = G_{ij}(y_{ij}-1)$, $b_{ij} = G_{ij}(y_{ij})$, $a_{ik} = G_{ik}(y_{ik}-1)$ and $b_{ik} = G_{ik}(y_{ik})$, with $G_{ij}(1) = 1$ and $G_{ij}(0) = 1-\Phi(z_{ij}^*)$. We can thus apply quasi-Newton minimization to the log-likelihood functions of margins

$$\begin{cases}
\ell_{nj}(\boldsymbol{\mu}_j,\Sigma_j) = \sum_{i=1}^n \log P_j(y_{ij}), \quad j=1,\ldots,d,\\
\ell_{njk}(\boldsymbol{\mu}_j,\Sigma_j,\boldsymbol{\mu}_k,\Sigma_k,\theta_{jk},\Sigma_{jk}) = \sum_{i=1}^n \log P_{jk}(y_{ij}y_{ik}), \quad 1\le j<k\le d,
\end{cases}$$

to obtain the estimates of the parameters $\boldsymbol{\mu}_j$, Σ_j, $\boldsymbol{\mu}_k$, Σ_k, θ_{jk} and Σ_{jk}.

From appropriate assumptions, many simplifications of (3.24) are possible. For example, if $\Sigma_j = I$ and $\Sigma_{jk} = \mathbf{0}$, $j \neq k$, then (3.24) simplifies to

$$\begin{cases}
z_{ij}^* = \dfrac{\boldsymbol{\mu}_j'\mathbf{x}_{ij}}{\{1+\mathbf{x}_{ij}'\mathbf{x}_{ij}\}^{1/2}}, \quad j=1,\ldots,d,\\
r_{i,jk} = \dfrac{\theta_{jk}}{\{(1+\mathbf{x}_{ij}'\mathbf{x}_{ij})(1+\mathbf{x}_{ik}'\mathbf{x}_{ik})\}^{1/2}}, \quad j\neq k,
\end{cases} \tag{3.25}$$

which is a simple example of having the dependence parameters be functions of covariates in a natural way, as they are derived. The numerical advantage is that as long as $\Theta = (\theta_{jk})$ is positive-definite, then all $R_i = (r_{i,jk})$, $i=1,\ldots,n$, are positive-definite.

An extension of (3.24) is to let

$$\begin{cases}
z_{ij}^* = \boldsymbol{\mu}_j'\mathbf{x}_{ij}, \quad j=1,\ldots,d,\\
r_{i,jk} = \dfrac{\theta_{jk}+\mathbf{w}_{ij}'\Sigma_{jk}\mathbf{w}_{ik}}{\{(1+\mathbf{w}_{ij}'\Sigma_j\mathbf{w}_{ij})(1+\mathbf{w}_{ik}'\Sigma_k\mathbf{w}_{ik})\}^{1/2}}, \quad j\neq k,
\end{cases} \tag{3.26}$$

where $\mathbf{x}_{ij}$ and $\mathbf{w}_{ij}$ may differ. However this does not obtain from a mixture model.

3.5.2 Multivariate Bernoulli-Beta model

For a d-variate binary random vector $\mathbf{Y}$ taking value 0 or 1 for each component, assume we have the MMD model (2.17), such that

$$P(y_1\cdots y_d) = \int_0^1\cdots\int_0^1 \prod_{j=1}^d f(y_j;p_j)c(G_1(p_1),\ldots,G_q(p_q))\prod_{j=1}^q g_j(p_j)\,dp_1\cdots dp_q, \tag{3.27}$$

where $f(y_j;p_j) = p_j^{y_j}(1-p_j)^{1-y_j}$. If in (3.27), G_j is a Beta(α_j,β_j) distribution, with density $g_j(p_j) = [B(\alpha_j,\beta_j)]^{-1}p_j^{\alpha_j-1}(1-p_j)^{\beta_j-1}$, $0<p_j<1$, then (3.27) is called a *multivariate Bernoulli-Beta model.* The copula C in (3.27) is arbitrary; a good choice may be the normal copula (2.4). With the normal copula, the model (3.27) is MUBE, thus the IFM approach can be applied to fit the model. We can then write the jth and (j,k) marginal pmf as

$$P_j(y_j) = \int_0^1 p_j^{y_j}(1-p_j)^{1-y_j}g_j(p_j)\,dp_j = B(\alpha_j+y_j,\beta_j+1-y_j)/B(\alpha_j,\beta_j),$$
$$P_{jk}(y_j\,y_k) = \int_0^1\int_0^1 p_j^{y_j}(1-p_j)^{1-y_j}p_k^{y_k}(1-p_k)^{1-y_k}\phi_2(x,y;\theta_{jk})g_j(p_j)g_k(p_k)\,dp_j\,dp_k,$$

where $x=\Phi^{-1}(G_j(p_j))$, $y=\Phi^{-1}(G_k(p_k))$, and $\phi_2(x,y;\theta)$ is the density of the bivariate normal, and g_j the density of Beta(α_j,β_j). Given data $\mathbf{y}_i=(y_{i1},\ldots,y_{id})$ with no covariates, we may obtain $\tilde{\alpha}_j$, $\tilde{\beta}_j$ and $\tilde{\theta}_{jk}$ with the IFM approach. For the case of an individual covariate $\mathbf{x}_{ij}$ for Y_{ij}, an interpretable extension of (3.27) is

$$P(\mathbf{y}_i) = \int_0^1\cdots\int_0^1\prod_{j=1}^d[h_j(\mathbf{x}_{ij},p_j)]^{y_{ij}}[1-h_j(\mathbf{x}_{ij},p_j)]^{1-y_{ij}}c(G_1(p_1),\ldots,G_q(p_q))\prod_{j=1}^q g_j(p_j)\,dp_1\cdots dp_q, \tag{3.28}$$

for some function h_j with range in $[0,1]$. A large family of such functions is $h_j(\mathbf{x}_{ij},p_j) = F_j(F_j^{-1}(p_j)+\boldsymbol{\beta}_j'\mathbf{x}_{ij})$ where F_j is a univariate cdf. $P_j(y_{ij})$ and $P_{jk}(y_{ij}\,y_{ik})$ can be written accordingly. For example, if $F_j(z)=\exp(-e^{-z})$, then $h_j(\mathbf{x}_{ij},p_j) = p_j^{\exp(-\boldsymbol{\beta}_j'\mathbf{x}_{ij})}$, and we have that when $y_{ij}=1$, $P_j(y_{ij}) = B(\alpha_j+\exp(-\boldsymbol{\beta}_j'\mathbf{x}_{ij}),\beta_j)/B(\alpha_j,\beta_j)$. If covariates are not subject dependent, but only margin-dependent, an alternative extension is to let α_{ij} and β_{ij} depend on the covariates for some functions a_j and b_j with range in $[0,\infty]$, such that $\alpha_{ij}=a_j(\boldsymbol{\gamma}_j,\mathbf{x}_j)$ and $\beta_{ij}=b_j(\boldsymbol{\eta}_j,\mathbf{x}_j)$. In this situation, we have, for example, $P_j(y_{ij}) = B(a_j(\boldsymbol{\gamma}_j'\mathbf{x}_j)+y_{ij},\ b_j(\boldsymbol{\eta}_j'\mathbf{x}_j)+1-y_{ij})/B(a_j(\boldsymbol{\gamma}_j'\mathbf{x}_j),\ b_j(\boldsymbol{\eta}_j'\mathbf{x}_j))$. An example of the functions a_j and b_j is $\alpha_{ij}=\exp(\boldsymbol{\gamma}_j'\mathbf{x}_j)$ and $\beta_{ij}=\exp(\boldsymbol{\eta}_j'\mathbf{x}_j)$. When applying the IFM approach to parameter estimation, the numerical computation involves 2-dimensional integration which would be feasible in most cases.

A special case of the model (3.27), where $p_j=p$, $j=1,\ldots,d$, is the model (1.1), studied in Prentice (1986). The pmf of the model is

$$P(y_1\cdots y_d) = \int_0^1 p^{y_+}(1-p)^{d-y_+}g(p)\,dp, \tag{3.29}$$

where $y_+=\sum_{j=1}^d y_j$ and $g(p)$ is the density of a Beta(α,β) distribution. The model (3.29) has exchangeable dependence and admits only positive dependence. A discussion of this special model, with extensions to include covariates and to admit negative dependence, can be found in Joe (1996).

3.5.3 Multivariate logit-normal model

For a d-variate binary random vector $\mathbf{Y}$ taking value 0 or 1 for each component, suppose we have the MMD model (2.17), such that

$$P(y_1 \cdots y_d) = \int_0^1 \cdots \int_0^1 \prod_{j=1}^{d} f(y_j\,;p_j) g(p_1,\ldots,p_d)\, dp_1 \cdots dp_d, \tag{3.30}$$

where $f(y_j\,;p_j) = p_j^{y_j}(1-p_j)^{1-y_j}$, and $g(\cdot)$ is the density function of a normal copula, with univariate marginal cdf G_j

$$G_j(p_j) = \int_0^{p_j} \phi\left(\frac{\log(x/(1-x)) - \mu_j}{\sigma_j}\right) \frac{1}{\sigma_j x(1-x)} dx, \quad 0 < p_j < 1,\ j = 1,\ldots,d.$$

In other words, if p_j is the outcome of a rv P_j, and $Z_j = logit(P_j) = \log(P_j/1-P_j)$, $j = 1,\ldots,d$, then $(Z_1,\ldots,Z_d)'$ has a joint d-dimensional normal distribution with mean vector $\boldsymbol{\mu}$, variance vector $\boldsymbol{\sigma}^2$ and a correlation matrix $\Theta = (\theta_{jk})$. We have

$$g(p_1,\ldots,p_d) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\sigma}'\Theta\boldsymbol{\sigma}|^{1/2}\prod_{j=1}^{d} p_j(1-p_j)} \exp\left\{-\frac{1}{2}(\mathbf{z}-\boldsymbol{\mu})'(\boldsymbol{\sigma}'\Theta\boldsymbol{\sigma})^{-1}(\mathbf{z}-\boldsymbol{\mu})\right\},$$

where $\mathbf{z} = (z_1,\ldots,z_d)'$, with $z_j = \log(p_j/1-p_j)$. We call this model the *multivariate logit-normal model.* The Fréchet upper bound is reached in the limit if $\Theta = J$, where J is matrix of 1's and $\sigma_j^2 \to \infty$. The multivariate probit model obtains in the limit as σ goes to ∞, by assuming Θ be a fixed correlation matrix and the mean parameters $\mu_j = \sigma_j z_j$ where z_j is constant.

The jth and the (j,k) marginal pmf are

$$P_j(y_j) = \int_0^1 \sigma_j^{-1} p_j^{y_j-1}(1-p_j)^{-y_j}\phi(x_j)\, dp_j, \quad j = 1,\ldots,d,$$

$$P_{jk}(y_j y_k) = \int_0^1 \int_0^1 (\sigma_j\sigma_k)^{-1} p_j^{y_j-1}(1-p_j)^{-y_j} p_k^{y_k-1}(1-p_k)^{-y_k}\phi_2(x_j,x_k;\theta_{jk})\, dp_j dp_k,\ 1 \le j < k \le d,$$

where $x_j = \{[\log(p_j/(1-p_j)) - \mu_j]/\sigma_j\}$, $j = 1,\ldots,d$, ϕ is the standard univariate normal density function and ϕ_2 the standard bivariate normal density function. Given data $\mathbf{y}_i = (y_{i1},\ldots,y_{id})$ with no covariates, we may obtain $\tilde{\mu}_j$, $\tilde{\sigma}_j$ and $\tilde{\theta}_{jk}$ by the IFM approach. For the case of different covariates for different margins, similar to the multivariate Bernoulli-Beta model, an interpretable extension of (3.30) is obtained by letting

$$P(y_{i1}\cdots y_{id}) = \int_0^1 \cdots \int_0^1 \prod_{j=1}^{d} [h_j(\mathbf{x}_{ij},p_j)]^{y_{ij}}[1-h_j(\mathbf{x}_{ij},p_j)]^{1-y_{ij}} g(p_1,\ldots,p_d)\, dp_1 \cdots dp_d, \tag{3.31}$$

for some function h_j with range in $[0,1]$. $P_j(y_{ij})$ and $P_{jk}(y_{ij}y_{ik})$ can be written accordingly. If covariates are not different, an interpretable extension to include covariates to the parameters in

(3.30) obtains by letting $\mu_{ij} = a_j(\boldsymbol{\gamma}_j, \mathbf{x}_j)$ and $\sigma_{ij} = b_j(\boldsymbol{\eta}_j, \mathbf{x}_{ij})$ for some functions a_j and b_j. The loglikelihood functions of margins for the parameters are now

$$\begin{cases} \ell_{nj}(\mu_j, \sigma_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_{jk}) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d, \end{cases}$$

where

$$P_j(y_{ij}) = \int_{-\infty}^{\infty} \frac{\exp\{y_{ij}(\mu_{ij} + \sigma_{ij}x)\}}{1 + \exp(\mu_{ij} + \sigma_{ij}x)} \phi(x)\, dx,$$
$$P_{jk}(y_{ij} y_{ik}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \frac{\exp\{y_{ij}(\mu_{ij} + \sigma_{ij}x)\}}{1 + \exp(\mu_{ij} + \sigma_{ij}x)} \frac{\exp\{y_{ik}(\mu_{ik} + \sigma_{ik}y)\}}{1 + \exp(\mu_{ik} + \sigma_{ik}y)} \phi_2(x, y; \theta_{jk})\, dxdy.$$

An example of the functions a_j and b_j is $\mu_{ij} = \boldsymbol{\gamma}_j' \mathbf{x}_j$ and $\sigma_{ij} = \exp(\boldsymbol{\eta}_j' \mathbf{x}_j)$. It is also possible to include covariates to the dependence parameters θ_{jk}; a discussion of this can be found in section 3.1. Again, when applying the IFM approach to parameter estimation, the numerical computation involves 2-dimensional integration, which should be feasible in most cases.

3.6 Multivariate mixture discrete models for count data

3.6.1 Multivariate Poisson-lognormal model

The *multivariate Poisson-lognormal model* for count data is introduced in Example 2.12. The pmf of $\mathbf{y}_i = (y_{i1}, \ldots, y_{id})$, $i = 1, \ldots, n$, is

$$P(y_{i1} \cdots y_{id}) = \int_0^{\infty} \cdots \int_0^{\infty} \prod_{j=1}^{d} f(y_{ij}; \lambda_{ij}) g(\boldsymbol{\eta}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i, \Theta_i)\, d\eta_1 \cdots d\eta_p, \tag{3.32}$$

where $f(y_{ij}; \lambda_{ij}) = \exp(-\lambda_{ij}) \lambda_{ij}^{y_{ij}} / y_{ij}!$, and

$$g(\boldsymbol{\eta}; \boldsymbol{\mu}_i, \boldsymbol{\sigma}_i, \Theta_i) = \frac{1}{(2\pi)^{d/2} (\eta_1 \cdots \eta_p) |\boldsymbol{\sigma}_i' \Theta_i \boldsymbol{\sigma}_i|^{1/2}} \exp\left\{ -\frac{1}{2} (\log \boldsymbol{\eta} - \boldsymbol{\mu}_i)' (\boldsymbol{\sigma}_i \Theta_i \boldsymbol{\sigma}_i)^{-1} (\log \boldsymbol{\eta} - \boldsymbol{\mu}_i) \right\}, \tag{3.33}$$

with $\eta_j > 0$, $j = 1, \ldots, p$, is the multivariate lognormal density. For simple situation with no covariates, $\boldsymbol{\mu}_i = \boldsymbol{\mu}$, $\boldsymbol{\sigma}_i = \boldsymbol{\sigma}$ and $\Theta_i = \Theta$. This model is studied in Aitchison and Ho (1989). The model (3.32) can accommodate a wide range of dependence, as we have seen in Example 2.12. $\mathrm{Corr}(Y_j, Y_k)$ is an increasing function of θ_{jk}, and varies over its full range when θ_{jk} varies over its full range. Thus in a general situation a multivariate Poisson-lognormal model of dimension d consists of d univariate Poisson-lognormal models describing some marginal characteristics and $d(d-1)/2$

dependence parameters θ_{jk}, $1 \le j < k \le d$, expressing the strength of the associations among the response variables. $\theta_{jk} = 0$ for all $j \ne k$ correspond to independence among the response variables. The response variables are exchangeable if Θ has an exchangeable structure and μ_j and σ_j are constant across the margins. We will see another special case later which also leads to exchangeable response variables. The loglikelihood functions of margins for the parameters are now

$$\begin{cases} \ell_{nj}(\mu_j, \sigma_j) = \sum_{i=1}^{n} \log P_j(y_{ij}), \quad j = 1, \ldots, d, \\ \ell_{njk}(\theta_{jk}, \mu_j, \mu_k, \sigma_j, \sigma_k) = \sum_{i=1}^{n} \log P_{jk}(y_{ij} y_{ik}), \quad 1 \le j < k \le d, \end{cases}$$

where

$$\begin{cases} P_j(y_{ij}) = \dfrac{\exp(y_{ij}\mu_j)}{\sqrt{2\pi} y_{ij}!} \displaystyle\int_{-\infty}^{\infty} \exp(y_{ij}\sigma_j z_j - e^{\mu_j + z_j \sigma_j}) \exp(-z_j^2/2) d\, z_j, \\ P_{jk}(y_{ij} y_{ik}) = \displaystyle\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \dfrac{\exp\left[y_{ij}(\mu_j + \sigma_j z_j) + y_{ik}(\mu_k + \sigma_k z_k)\right]}{y_{ij}! y_{ik}! \exp(e^{\mu_j + \sigma_j z_j} + e^{\mu_k + \sigma_k z_k})} \phi_2(z_j, z_k; \theta_{jk}) dz_j dz_k, \end{cases}$$

where ϕ_2 is the standard binormal density. To get the IFME of $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$ and Θ, quasi-Newton minimization method can be used. Good starting points can be obtained from the method of moments estimates. Let $\bar{y}_j$, s_j^2 and r_{jk} be the sample mean, sample variance and sample correlations respectively. The method of moments estimates based on the expected values given in (2.30) are $\tilde{\sigma}_j^0 = \{\log[(s_j^2 - \bar{y}_j)/\bar{y}_j^2 + 1]\}^{1/2}$, $\tilde{\mu}_j^0 = \log \bar{y}_j - 0.5(\tilde{\sigma}_j^0)^2$ and $\tilde{\theta}_{jk}^0 = \log[r_{jk} s_j s_k/(\bar{y}_j \bar{y}_k) + 1]/(\tilde{\sigma}_j^0 \tilde{\sigma}_k^0)$.

When there is a covariate vector $\mathbf{x}_{ij}$ for the response observation $\mathbf{y}_{ij}$, we may let $\mu_{ij} = a_j(\boldsymbol{\gamma}_j, \mathbf{x}_{ij})$ for some function a_j in the range $(-\infty, \infty)$, and let $\sigma_{ij} = b_j(\boldsymbol{\eta}_j, \mathbf{x}_{ij})$ for some function b_j in the range $[0, \infty)$. An example of the functions a_j and b_j is $\mu_{ij} = \boldsymbol{\gamma}_j' \mathbf{x}_{ij}$ and $\sigma_{ij} = \exp(\boldsymbol{\eta}_j' \mathbf{x}_{ij})$. It is also possible to let the dependence parameters θ_{jk} be functions of covariates; a discussion of this can be found section 3.1. For details on numerical methods for obtaining the parameter estimates, see section 2.7.

A special situation of the multivariate Poisson-lognormal model is to assume that $f(y_j; \lambda_j) = e^{-\lambda_j} \lambda_j^{y_j}/y_j!$, where $\lambda_j = \lambda \beta_j$. $\beta_j > 0$ is considered as a scale factor (known or unknown) and the common parameter λ has the lognormal distribution $LN(\mu, \sigma^2)$. In this situation we have

$$P(y_1 \cdots y_d) = \frac{\prod_{j=1}^{d} \beta_j^{y_j} \exp(\mu \sum_{j=1}^{d} y_j)}{\sqrt{2\pi} \prod_{j=1}^{d} y_j!} \int_{-\infty}^{\infty} \frac{\exp(\sigma z \sum_{j=1}^{d} y_j)}{\exp(e^{\mu + z\sigma} \sum_{j=1}^{d} \beta_j)} \exp(-z^2/2) d\, z, \tag{3.34}$$

and the parameters μ and σ are common across all the margins. To calculate $P(y_1 \cdots y_d)$, we need only calculate a one-dimensional integral; thus full maximum likelihood estimation can be used to get the estimates of μ, σ and β_j (if it is unknown). By the formulas in (2.26), it can be shown that there

is an exchangeable correlation structure in the response vector $\mathbf{Y}$, with the pairwise correlations tending to 1 when μ or σ tend to infinity. Independence is achieved when $\sigma \to 0$. The model (3.34) does not admit negative dependence.

3.6.2 Multivariate Poisson-gamma model

The *multivariate Poisson-gamma model* is obtained by letting $G_j(\eta_j)$ in (2.24) be the cdf of a univariate gamma distribution with shape parameter α_j and scale parameter β_j, with the density function $g_j(x;\alpha_j,\beta_j) = \beta_j^{-\alpha_j} x^{\alpha_j-1} e^{-x/\beta_j}/\Gamma(\alpha_j)$, $x > 0$, $\alpha_j > 0$ and $\beta_j > 0$. The Gamma family is closed under convolution for fixed β. The copula C in (2.24) is arbitrary; (3.1)–(3.8) are some choices here. For example, with the multinormal copula, the multivariate Poisson-gamma model is MUBE. Thus the IFM approach can be applied to fit the model. The jth marginal distribution of a multivariate Poisson-gamma distribution is

$$\begin{aligned} P_j(y_j) &= \int_0^\infty f(y_j;z_j) g_j(z_j)\, dz_j = \frac{\int e^{-z_j} z_j^{y_j} z_j^{\alpha_j-1} e^{-z_j/\beta_j}\, dz_j}{y_j!\beta_j^{\alpha_j}\Gamma(\alpha_j)} \\ &= \frac{\Gamma(y_j+\alpha_j)}{y_j!\Gamma(\alpha_j)} \left(\frac{1}{1+\beta_j}\right)^{\alpha_j} \left(\frac{\beta_j}{1+\beta_j}\right)^{y_j}, \end{aligned} \tag{3.35}$$

which implies that Y_j has a negative binomial distribution (in the generalized sense). We have $E(Y_j) = \alpha_j\beta_j$ and $Var(Y_j) = \alpha_j\beta_j(1+\beta_j)$. The margins are overdispersed since $Var(Y_j)/E(Y_j) > 1$. Based on (3.35), if α_j is an integer, y_j can be interpreted as the number of observed failures in $y_j+\alpha_j$ trials, with α_j a previously fixed number of successes.

The parameter estimation procedure based on IFM is similar to that for the multivariate Poisson-lognormal model. Some simplifications are possible. One simplification for the Poisson-gamma model is to hold the shape parameter α_j constant across j. In this situation, we have $E(Y_j) = \mu_j = \alpha\beta_j$ and $Var(Y_j) = \mu_j(1+\mu_j/\alpha)$. Similarly, we can also require β_j be constant across j and obtain the same functional relationship between the mean and the variance across j. By doing so, we reduce the total number of parameters. With this simplification in the number of parameters, the same parameter appears in different margins. The IFM approach for estimating parameters common to more than one margin discussed in section 2.6 can be applied. Another special case is to let $\lambda_j = \lambda\beta_j$, where $\beta_j > 0$ is considered to be a scale factor (known or unknown) and the common parameter λ has a Gamma distribution. This is similar to the multivariate Poisson-lognormal model (3.34). Negative dependence cannot be admitted into this special situation, which is similar to the multivariate Poisson-lognormal model (3.34).

3.6.3 Multivariate negative-binomial mixture model

Consider d-dimensional count data with $y_j = r_j, r_j + 1, \ldots, \; r_j \geq 1, \; j = 1, 2 \ldots, d$. For example, with given integer value r_j, y_j might be the total number of Bernoulli trials until the r_jth success, where the probability of success in each trial is p_j; that is

$$P_j(y_j|p_j) = \binom{y_j - 1}{r_j - 1} p_j^{r_j}(1 - p_j)^{y_j - r_j}.$$

If p_j is itself the outcome of a random variate X_j, $j = 1, \ldots, d$, which have the joint distribution $G(p_1, \ldots, p_d)$, then the distribution for $\mathbf{Y} = (Y_1, \ldots, Y_d)$ is called the *multivariate negative-binomial mixture model.* If the inverse of $1/X_j$ has a distribution with mean μ_j and variance σ_j^2, then simple calculations lead to $\mathrm{E}(Y_j) = r_j\mu_j$ and $\mathrm{Var}(Y_j) = r_j\mu_j(mu_j - 1) + r_j(r_j + 1)\sigma_j^2$. This multivariate negative-binomial mixture model for count data is similar to the multivariate Bernoulli-Beta model for binary data in section 3.5. Thus the comments on the extensions to include covariates apply here as well.

A more general form of negative binomial distribution is (3.16), such that

$$P_j(y_j|p_j) = \frac{\Gamma(\beta_j + y_j)}{\Gamma(\beta_j)\Gamma(y_j + 1)} p_j^{\beta_j}(1 - p_j)^{y_j}, \quad \beta_j > 0, \; y_j = 0, 1, 2, \ldots.$$

Using the recursive relation $\Gamma(x) = (x - 1)\Gamma(x - 1)$, $P_j(y_j|p_j)$ can be written as

$$P_j(y_j|p_j) = p_j^{\beta_j} \begin{cases} \displaystyle\prod_{k=1}^{y_j} \frac{(\beta_j + k - 1)(1 - p_j)}{k}, & y_j = 1, 2, \ldots, \\ 1, & y_j = 0. \end{cases}$$

The multivariate negative-binomial mixture model can be defined with this general negative binomial distribution as the discrete mixing part. β_j, $j = 1, \ldots, d$, can be considered as parameters in the model.

3.6.4 Multivariate Poisson-inverse Gaussian model

The *multivariate Poisson-inverse Gaussian model* is obtained by letting $G_j(\eta_j)$ in (2.24) be the cdf of a three-parameter univariate inverse Gaussian distribution with density function

$$g_j(\lambda_j) = \frac{\xi_j^{-\gamma_j}}{K_{\gamma_j}(\omega_j)} \lambda_j^{\gamma_j - 1} \exp[(\omega_j/2)(\xi_j/\lambda_j + \lambda_j/\xi_j)], \quad \lambda_j > 0, \tag{3.36}$$

where $\omega_j = \sqrt{\xi_j^2 + \alpha_j^2} - \xi_j$, $\alpha_j > 0$, $\xi_j > 0$ and $-\infty < \gamma_j < \infty$. In the density expression, $K_v(z)$ denotes the modified Bessel function of the second kind of order v and argument z. It satisfies the

relationship

$$K_{v+1}(z) = \frac{2v}{z} K_v(z) + K_{v-1}(z),$$

with $K_{-1/2}(z) = K_{1/2}(z) = \sqrt{\pi/2z}\exp(-z)$. The copula C in (2.24) is arbitrary; interesting choices are copulas (3.1)–(3.8). With the multinormal copula, the multivariate Poisson-inverse Gaussian model is MUBE; thus the IFM approach can be applied to fit the model.

A special case of the multivariate Poisson-inverse Gaussian model results when $f(y_j; z_j) = e^{-z_j} z_j^{y_j}/y_j!$, where $z_j = \lambda t_j$, with $t_j > 0$ considered as a scale factor ($j = 1, \ldots, d$). Then the pmf for $\mathbf{Y}$ is

$$P(y_1 \cdots y_d) = \frac{K_{k+\gamma}\left(\sqrt{\omega(\omega + 2\xi \sum t_j}\right)}{K_\gamma(\omega)} \left(\frac{\omega}{\omega + 2\xi \sum t_j}\right)^{(k+\gamma)/2} \prod_{j=1}^{d} \frac{(\xi t_j)^{y_j}}{y_j!},$$

where $k = \sum_{j=1}^{d} y_j$. An extensive study of this special model can be found in Stein *et al.* (1987).

3.7 Application to longitudinal and repeated measures data

Multivariate copula discrete (MCD) and multivariate mixture discrete (MMD) models can be used for longitudinal and repeated measures (over time) data when the response variables are discrete (binary, ordinal and count), and the number of measures is small and constant over subjects. The multivariate dependence structure has the form of time series dependence or of dependence decreasing with lag. Examples include MCD and MMD models with special copula dependence structure and special patterns of marginal parameters. These models include stationary time series models that allow arbitrary univariate margins and non-stationary cases, in which there are time-dependent or time-independent covariates or time trends.

In classical time series analysis, the standard models are autoregressive (AR) and moving average (MA) models. The generalization of these concepts to MCD and MMD models for discrete time series is that "autoregressive" is replaced by Markov and "moving average" is replaced by k-dependent (only rv's that are separated by a lag of k or less are dependent). A particularly interesting model is the Markov model of order one, which can be considered as a replacement for AR(1) model in classical time series analysis; and these types of Markov models can be constructed from families of bivariate copulas. For a more detailed discussion of related topics, such as the extension of models to include covariates and models for different individuals observed at different times, see Joe (1996, Chapter 8).

If the copula is the multinormal copula (3.1), the correlation matrix in the multinormal copula may have patterns of correlations depending on lags, such as exchangeable or AR type. For example, for exchangeable pattern, $\theta_{jk} = \theta$ for all $1 \leq j < k \leq d$. For AR(1), $\theta_{jk} = \theta^{|j-k|}$ for some $|\theta| < 1$. For AR(2), $\theta_{jk} = \rho_s$, with $s = |j-k|$. ρ_s is the autocorrelations of lag s; the autocorrelation satisfy $\rho_s = \phi_1\rho_{s-1} + \phi_2\rho_{s-2}$, $s \geq 3$, $\phi_1 = \rho_1(1-\rho_2)/(1-\rho_1^2)$, $\phi_2 = (\rho_2 - \rho 1)/(1-\rho_1^2)$, and are determined from ρ_1 and ρ_2.

Some examples of models suitable for modelling longitudinal data and repeated measures (over time) are the multivariate Poisson-lognormal model, multivariate logit-normal model, multivariate logit model with multinormal copula or with M-L construction, multivariate probit model with multinormal copula, and so on. In fact, the multivariate probit model with multinormal copula is equivalent to the discretization of ARMA normal time series for binary and ordinal response. For the discrete time series and $d \geq 4$, approximations can be used for the probabilities $\Pr(Y_j = y_j, j = 1, \ldots, d)$ which in general are multidimensional integrals.

3.8 Summary

In this chapter, we studied specific MCD models for binary, ordinal and count data, and MMD models for binary and count data. (MMD models for ordinal data are not presented, since there is no natural simple way to represent such models, however MMD models for binary data can be extended to MMD models for nominal categorical data.) Extension to let the marginal parameters as well as the dependence parameter be functions of covariates are discussed. We also outlined the potential application of MCD and MMD models for longitudinal data, repeated measures and time series data. However, this chapter does not contain an exhaustive list of models in the family of MCD and MMD classes. Many additional interesting models in MCD and MMD classes could be introduced and studied. Our purpose in this chapter is to demonstrate the richness of the classes of MCD and MMD models, and to make several specific models available for applications. Some examples of the application of models introduced in this chapter can be found in Chapter 5.

Chapter 4

The efficiency of IFM approach and the efficiency of jackknife variance estimate

It is well known that under regularity conditions, the (full-dimensional) maximum likelihood estimator (MLE) is asymptotically efficient and optimal. But in multivariate situations, except the multinormal model, the computation of the MLE is often complicated or impossible. The IFM approach is proposed in Chapter 2 as an alternative estimation approach. We have shown that the IFM approach provides consistent estimators with some good asymptotic properties (such as asymptotic normality of the estimators). This approach has many advantages; computational feasibility is main one. It can be applied to many MCD and MMD models (models with MUBE, PUBE properties) with appropriate choices of the copulas; examples of such copulas are multinormal copula, M-L construction, copulas from mixture of max-id distributions, copulas from mixture of conditional distributions, and so on. The IFM theory is a new statistical inference theory for the analysis of multivariate non-normal models. However, the efficiency of estimators obtained from IFM in comparison with ML estimators is not clear.

In this chapter, we investigate the efficiency of the IFM approach relative to maximum likelihood. Our studies suggest that the IFM approach is a viable alternative to ML for models with MUBE, PUBE or MPME properties. This chapter is organized as follows. In section 4.1, we discuss how to assess the efficiency of the IFM approach. In section 4.2, we carry out some analytical comparisons

of the IFM approach to ML for some models. These studies show that the IFM approach is quite efficient. A general analytical investigation is not possible, as closed form expressions for estimators and the corresponding asymptotic variance-covariance matrices from ML and IFM are not possible for the majority of multivariate non-normal models. Most often numerical assessment of their performance must be used. In section 4.3, we carry out extensive numerical studies of the efficiency of IFM approach relative to ML approach. These studies are done mainly for MCD and MMD models with MUBE or PUBE properties. The situations include models without and with covariates. In section 4.4, we numerically study the efficiency of IFM approach relative to ML approach for models with special dependence structure. The IFM approach extends easily to the models with parameters common to more than one margin. Section 4.5 is devoted to the numerical assessment of the efficiency of the jackknife approach for variance estimation of IFME. The numerical results show that the jackknife variance estimates are quite satisfactory.

4.1 The assessment of the efficiency of IFM approach

In section 2.3, we have given some optimality criteria for inference functions. We concluded that in the class of all regular unbiased estimating functions, the inference functions of scores (IFS) is M-optimal (so T-optimal or D-optimal as well). For the regular model (2.12), the inference function in the IFM approach are in the class of regular unbiased inference functions; thus all the (asymptotic) properties of regular inference functions apply to IFM.

To assess the efficiency of IFM relative to IFS, at least three approaches are possible:

A1. Examine the M-optimality (or T-optimality or D-optimality) of IFM relative to IFS.

A2. Compare the MSE of the estimates from IFM and IFS based on simulation.

A3. Examine the asymptotic behaviour of $2\ell(\tilde{\boldsymbol{\theta}}) - 2\ell(\boldsymbol{\theta})$ based on the knowledge that $2\ell(\hat{\boldsymbol{\theta}}) - 2\ell(\boldsymbol{\theta})$ has an asymptotic χ^2_q distribution when $\boldsymbol{\theta}$ is the true parameter vector (of length q).

A1 is along the lines of inference function theory. As an estimator may be regarded as a solution to an equation of the form $\Psi(\mathbf{y};\boldsymbol{\theta}) = \mathbf{0}$, we study the inference functions instead of the estimators. This approach can be carried out analytically in a few cases when both the Godambe information matrix of IFM and the Fisher information matrix for IFS are available in closed form, or otherwise numerically by computing (or estimating) the Godambe information matrix and the Fisher information matrix (based on simulation). With this approach, we do not need to actually find the parameter estimates

for the purposes of comparison. The disadvantage is that the Godambe information matrix or Fisher information matrix may be difficult to calculate, because partial derivatives are needed for the computation and they are difficult to calculate for most multivariate non-normal models. Also this is an asymptotic comparison. A2 is a conventional approach, it provides a way to investigate the small sample properties of the estimates. This possibility is especially interesting in comparison with A1, since although MLEs are asymptotically optimal, this may not generally be the case for finite samples. The disadvantage with A2 is that it may computationally demanding with multivariate non-normal models, because for each simulation, parameters estimation based on IFM and IFS are carried out. A3 is based on the understanding that if the estimates from IFM are efficient, we would envisage that the full-dimensional likelihood function evaluated at these estimates should have the similar asymptotic behaviour as when the full-dimensional likelihood function is evaluated at the MLE. More specifically, suppose the loglikelihood function is $\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \log f(\mathbf{y}_i|\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a vector of length q. Under regularity conditions, $2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}))$ has an asymptotic χ^2_q distribution (see for example, Sen and Singer 1993, p236). Thus a rough method of assessing the efficiency of $\tilde{\boldsymbol{\theta}}$ is to see if $2(\ell(\tilde{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}))$ is in the likelihood-based confidence interval for $2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}))$; this interval of $1-\alpha$ confidence is $(\chi^2_{q;\alpha/2}, \chi^2_{q;1-\alpha/2})$, where $\chi^2_{q,\beta}$ is the lower β quantile of a chi-square distribution with q degrees of freedom. The assessment can be carried out by comparing the frequency of (empirical confidence level of) $2(\ell(\tilde{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta}))$ in the $(\chi^2_{q;\alpha/2}, \chi^2_{q;1-\alpha/2})$ with $1-\alpha$. In other words, we check the frequency of

$$\tilde{\boldsymbol{\theta}} \in \{\hat{\boldsymbol{\theta}} : \chi^2_{q;\alpha/2} < 2(\ell(\hat{\boldsymbol{\theta}}) - \ell(\boldsymbol{\theta})) < \chi^2_{q;1-\alpha/2}\},$$

and $\tilde{\boldsymbol{\theta}}$ is considered to be efficient if the empirical frequency is close to $1-\alpha$. The advantage of this approach is that only $\tilde{\boldsymbol{\theta}}$, $\ell(\tilde{\boldsymbol{\theta}})$ and $\ell(\boldsymbol{\theta})$ need to be calculated; this leads to much less computation compare with finding $\hat{\boldsymbol{\theta}}$. The disadvantage is that this approach may not be very informative about efficiency of $\tilde{\boldsymbol{\theta}}$ in comparison with $\hat{\boldsymbol{\theta}}$ in relatively small sample situations. In our studies, A3 will not be used. We mention this approach merely for further potential investigations.

To compare IFM with IFS by A1, we need to calculate the Fisher information (matrix) and the Godambe information (matrix). Suppose $P(y_1 \cdots y_d; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Re$, is a regular MCD or MMD model in (2.12), where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_q)'$ is q-component vector, and $\Re$ is the parameter space. The Fisher information matrix from one observation for the parameter vector $\boldsymbol{\theta}$, I, has the following expression

$$I = \begin{pmatrix} I_{11} & \cdots & I_{1q} \\ \vdots & \ddots & \vdots \\ I_{1q} & \cdots & I_{qq} \end{pmatrix},$$

where

$$I_{jj} = \sum_{\{y_1 \cdots y_d\}} \frac{1}{P_{1\cdots q}(y_1 \cdots y_d)} \left(\frac{\partial P_{1\cdots q}(y_1 \cdots y_d)}{\partial \theta_j} \right)^2, \quad j = 1, \ldots, q,$$

$$I_{jk} = \sum_{\{y_1 \cdots y_d\}} \frac{1}{P_{1\cdots q}(y_1 \cdots y_d)} \frac{\partial P_{1\cdots q}(y_1 \cdots y_d)}{\partial \theta_j} \frac{\partial P_{1\cdots q}(y_1 \cdots y_d)}{\partial \theta_k}, \quad 1 \le j < k \le q.$$

Assume the IFM for one observation is $\Psi = (\psi_1, \ldots, \psi_q)$. The Godambe information matrix J_Ψ based on IFM for one observation is $J_\Psi = D_\Psi M_\Psi^{-1} D_\Psi^T$, where

$$M_\Psi = \begin{pmatrix} M_{11} & \cdots & M_{1q} \\ \vdots & \ddots & \vdots \\ M_{1q} & \cdots & M_{qq} \end{pmatrix} \quad D_\Psi = \begin{pmatrix} D_{11} & \cdots & D_{1q} \\ \vdots & \ddots & \vdots \\ D_{q1} & \cdots & D_{qq} \end{pmatrix}$$

with $M_{jj} = E(\psi_j^2)$ $(j = 1, \ldots, q)$, $M_{jk} = E(\psi_j \psi_k)$ $(j, k = 1, \ldots, q, j \neq k)$, $D_{jj} = E(\partial \psi_j / \partial \theta_j)$ $(j = 1, \ldots, q)$, and $D_{jk} = E(\partial \psi_j / \partial \theta_k)$ $(j, k = 1, \ldots, q, j \neq k)$. The detailed calculation of the elements of M_Ψ and D_Ψ can be found in section 2.4 for the models without covariates. The M-optimality assessment examines the positive-definiteness of $J_\Psi^{-1} - I^{-1}$. It is equivalent to T-optimality which examines the ratio of the trace of the two information matrices, $\mathrm{Tr}(J_\Psi^{-1})/\mathrm{Tr}(I^{-1})$, and D-optimality which examines the ratio of the determinant of the two information matrices, $\det(J_\Psi^{-1})/\det(I^{-1})$. T-optimality is a suitable index for the efficiency investigation as it is easier to compute. An equivalent index to D-optimality is $\sqrt[q]{\det(J_\Psi^{-1})/\det(I^{-1})}$. In our efficiency assessment studies, we will use M-optimality, T-optimality or D-optimality interchangeably depending on which is most convenient.

In most multivariate settings, A1 is not feasible analytically and extremely difficult computationally (involving tedious programming of partial derivatives). A2 is an approach which eliminates the above problems as long as MLEs are available. As MLEs and IFMEs are both unbiased only asymptotically, the actual bias related to the sample size is an important issue. For an investigation related to sample size, it is more sensible to examine the measures of closeness of an estimator to its true value. Such a measure is the mean squared error (MSE) of an estimator. For an estimator $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a random sample of size n from a distribution indexed by θ, the MSE of $\tilde{\theta}$ about the true value θ is

$$\mathrm{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 = Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2.$$

Suppose that $\tilde{\theta}$ has a sampling distribution F, and suppose $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$ are iid of F, then an obvious estimator of $\mathrm{MSE}(\tilde{\theta})$ is

$$\widehat{\mathrm{MSE}}(\tilde{\theta}) = \frac{\sum_{i=1}^m (\tilde{\theta}_i - \theta)^2}{m}. \tag{4.1}$$

If $\tilde{\theta}$ is from the IFM approach and $\hat{\theta}$ from the IFS approach, A2 suggests that we compare $\widehat{\mathrm{MSE}}(\tilde{\theta})$ and $\widehat{\mathrm{MSE}}(\hat{\theta})$. For a fixed sample size, $\hat{\theta}$ need not be the optimal estimate of θ in terms of MSE, since now the bias of the estimate is also taken into consideration. The measure $\widehat{\mathrm{MSE}}(\tilde{\theta})/\widehat{\mathrm{MSE}}(\hat{\theta})$ thus gives us an idea how IFME performs relative to MLE. The approach is mainly computational, based on the computer implementation of specific models and subsequent intensive simulation and parameter estimation. A2 can be easily used to study models with no covariates as well as with covariates.

4.2 Analytical assessment of the efficiency

In this section, we study the efficiency of the IFM approach analytically for some special models where the Godambe information matrix and the Fisher information matrix are available in closed form, or computable.

Example 4.1 (Multinormal, general) Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$. Given n independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathbf{X}$, the MLEs are

$$\hat{\boldsymbol{\mu}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i, \quad \hat{\Sigma} = n^{-1} \sum_{i=1}^{n} (x_i - \hat{\boldsymbol{\mu}})(x_i - \hat{\boldsymbol{\mu}})^T.$$

It can be easily shown that the IFME $\tilde{\boldsymbol{\mu}}$ and $\tilde{\Sigma}$ are equal to $\hat{\boldsymbol{\mu}}$ and $\hat{\Sigma}$ respectively, so MLE and IFME are equivalent for the general multinormal distribution. □

Example 4.2 (Multinormal, common marginal mean) Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)' = \mu \mathbf{1}$ for a scalar parameter μ and Σ is known. Given n independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with same distributions as $\mathbf{X}$, the MLE of μ is

$$\hat{\mu} = \frac{\mathbf{1}' \Sigma^{-1} \sum_{i=1}^{n} \mathbf{x}_i}{n \mathbf{1}' \Sigma^{-1} \mathbf{1}}.$$

The IFM of $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ are equivalent to

$$\Psi_{nj} = \sum_{i=1}^{n} \frac{x_{ij} - \mu_j}{\sigma_j^2}, \quad j = 1, \ldots, d,$$

which leads to $\tilde{\mu}_j = n^{-1} \sum_{i=1}^{n} x_{ij}$, or $\tilde{\boldsymbol{\mu}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$. A simple calculation leads to $J_{\Psi}^{-1}(\boldsymbol{\mu}) = n^{-1}\Sigma$. Thus if we incorporate the knowledge that $\mu_1, \ldots, \mu_d$ are the same value, with WA and PMLA, we have

i. WA: the final IFME of μ is

$$\tilde{\mu}_w = \frac{\mathbf{1}'\Sigma^{-1}\tilde{\boldsymbol{\mu}}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}},$$

which is exactly the same as $\hat{\mu}$ since $\tilde{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^{n}\mathbf{x}_i$. So in this situation, the IFME is equivalent to the MLE.

ii. PMLA: the final IFME of μ is

$$\tilde{\mu}_p = \frac{\mathbf{1}'\{\mathrm{diag}(\Sigma)\}^{-1}\tilde{\boldsymbol{\mu}}}{\mathbf{1}'\{\mathrm{diag}(\Sigma)\}^{-1}\mathbf{1}} = \frac{\sum\sigma_{jj}^{-1}\tilde{\mu}_j}{\sum\sigma_{jj}^{-1}}.$$

With this approach, IFME is not equivalent to MLE. The ratio of $\mathrm{Var}(\tilde{\mu}_p)$ to $\mathrm{Var}(\hat{\mu})$ is

$$\frac{\mathbf{1}'\{\mathrm{diag}(\Sigma)\}^{-1}\Sigma\{\mathrm{diag}(\Sigma)\}^{-1}\mathbf{1}\mathbf{1}'\Sigma^{-1}\mathbf{1}}{(\mathbf{1}'\{\mathrm{diag}(\Sigma)\}^{-1}\mathbf{1})^2}.$$

There is some loss of efficiency with simple PMLA.

□

Example 4.3 (Trivariate probit, general) Suppose we have a trivariate probit model with known cut-off points, such that $P(111) = \Phi_3(0,0,0,\rho_{12},\rho_{13},\rho_{23})$. We have the following (Tong 1990):

$$\begin{cases} P_1(1) = P_2(1) = P_3(1) = \Phi(0) = \dfrac{1}{2}, \\ P_{jk}(11) = \Phi_2(0,0,\rho_{jk}) = \dfrac{1}{4} + \dfrac{1}{2\pi}\sin^{-1}\rho_{jk}, \quad 1 \le j < k \le 3, \\ P(111) = \Phi_3(0,0,0,\rho_{12},\rho_{13},\rho_{23}) = \dfrac{1}{8} + \dfrac{1}{4\pi}(\sin^{-1}\rho_{12} + \sin^{-1}\rho_{13} + \sin^{-1}\rho_{23}). \end{cases} \tag{4.2}$$

The full loglikelihood function is

$$\begin{aligned} \ell_n =& n(111)\log P(111) + n(110)\log P(110) + n(101)\log P(101) + n(100)\log P(100) \\ & n(011)\log P(011) + n(010)\log P(010) + n(001)\log P(001) + n(000)\log P(000). \end{aligned} \tag{4.3}$$

Even in this simple situation, the MLE of ρ_{jk} is not available in closed form. The information matrix for ρ_{12}, ρ_{13} and ρ_{23} from one observation is

$$I = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{12} & I_{22} & I_{23} \\ I_{13} & I_{23} & I_{33} \end{pmatrix},$$

where, for example,

$$\begin{aligned} I_{11} = & \frac{1}{P(111)}\left(\frac{\partial P(111)}{\partial\rho_{12}}\right)^2 + \frac{1}{P(110)}\left(\frac{\partial P(110)}{\partial\rho_{12}}\right)^2 + \frac{1}{P(101)}\left(\frac{\partial P(101)}{\partial\rho_{12}}\right)^2 + \frac{1}{P(100)}\left(\frac{\partial P(100)}{\partial\rho_{12}}\right)^2 \\ & + \frac{1}{P(011)}\left(\frac{\partial P(011)}{\partial\rho_{12}}\right)^2 + \frac{1}{P(010)}\left(\frac{\partial P(010)}{\partial\rho_{12}}\right)^2 + \frac{1}{P(001)}\left(\frac{\partial P(001)}{\partial\rho_{12}}\right)^2 + \frac{1}{P(000)}\left(\frac{\partial P(000)}{\partial\rho_{12}}\right)^2. \end{aligned}$$

Simple calculation gives us $\partial P(111)/\partial\rho_{12} = 1/(4\pi\sqrt{1-\rho_{12}^2})$; and other terms also have similar expressions. After simplification, we get

$$I_{11} = \frac{\pi^3 + 64a - 16\pi b}{\pi(1-\rho_{12}^2)cdef},$$

where

$$\begin{aligned}
a &= \sin^{-1}\rho_{12}\sin^{-1}\rho_{13}\sin^{-1}\rho_{23},\\
b &= (\sin^{-1}\rho_{12})^2 + (\sin^{-1}\rho_{13})^2 + (\sin^{-1}\rho_{23})^2,\\
c &= \pi + 2\sin^{-1}\rho_{12} + 2\sin^{-1}\rho_{13} + 2\sin^{-1}\rho_{23},\\
d &= \pi + 2\sin^{-1}\rho_{12} - 2\sin^{-1}\rho_{13} - 2\sin^{-1}\rho_{23},\\
e &= \pi - 2\sin^{-1}\rho_{12} + 2\sin^{-1}\rho_{13} - 2\sin^{-1}\rho_{23},\\
f &= \pi - 2\sin^{-1}\rho_{12} - 2\sin^{-1}\rho_{13} + 2\sin^{-1}\rho_{23}.
\end{aligned}$$

Other components in the matrix I can be computed similarly. The inverse of I, after simplification, is found to be

$$I^{-1} = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{12} & a_{22} & a_{23} \\ a_{13} & a_{23} & a_{33} \end{pmatrix},$$

where

$$\begin{aligned}
a_{11} &= \frac{(\pi^2 - 4(\sin^{-1}\rho_{12})^2)(1-\rho_{12}^2)}{4},\\
a_{22} &= \frac{(\pi^2 - 4(\sin^{-1}\rho_{13})^2)(1-\rho_{13}^2)}{4},\\
a_{33} &= \frac{(\pi^2 - 4(\sin^{-1}\rho_{23})^2)(1-\rho_{23}^2)}{4},\\
a_{12} &= \frac{(2\sin^{-1}\rho_{12}\sin^{-1}\rho_{13} - \pi\sin^{-1}\rho_{23})(1-\rho_{12}^2)^{1/2}(1-\rho_{13}^2)^{1/2}}{2},\\
a_{13} &= \frac{(2\sin^{-1}\rho_{12}\sin^{-1}\rho_{23} - \pi\sin^{-1}\rho_{13})(1-\rho_{12}^2)^{1/2}(1-\rho_{23}^2)^{1/2}}{2},\\
a_{23} &= \frac{(2\sin^{-1}\rho_{13}\sin^{-1}\rho_{23} - \pi\sin^{-1}\rho_{12})(1-\rho_{13}^2)^{1/2}(1-\rho_{23}^2)^{1/2}}{2}.
\end{aligned}$$

For the IFM approach, we have

$$\Psi_{njk} = \left(\frac{n_{jk}(11)+n_{jk}(00)}{P_{jk}(11)} - \frac{n_{jk}(10)+n_{jk}(01)}{1/2 - P_{jk}(11)}\right)\frac{\partial P_{jk}(11)}{\partial\rho}.$$

$\Psi_{njk} = 0$ leads to

$$\tilde{\rho}_{jk} = \sin\left(\frac{\pi}{2}\frac{n_{jk}(11)+n_{jk}(00)-n_{jk}(10)-n_{jk}(01)}{n}\right), \quad 1 \le j < k \le 3.$$

If the IFM for one observation is $\Psi = (\psi_{12}, \psi_{13}, \psi_{23})$, then from section 2.4, we have

$$E(\psi_{jk}\psi_{lm}) = \sum_{\{y_j y_k y_l y_m\}} \frac{P_{jklm}(y_j y_k y_l y_m)}{P_{jk}(y_j y_k)P_{lm}(y_l y_m)}\frac{\partial P_{jk}(y_j y_k)}{\partial\rho_{jk}}\frac{\partial P_{lm}(y_l y_m)}{\partial\rho_{lm}},$$

where $1 \le j < k \le 3$, $1 \le l < m \le 3$, and

$$E\left(\frac{\partial \psi_{jk}}{\partial \rho_{jk}}\right) = \sum_{\{y_j y_k\}} -\frac{1}{P_{jk}(y_j y_k)} \left(\frac{\partial P_{jk}(y_j y_k)}{\partial \rho_{jk}}\right)^2, \quad 1 \le j < k \le 3.$$

We thus find that

$$M_\Psi = \begin{pmatrix} b_{11} & b_{12} & b_{13} \\ b_{12} & b_{22} & b_{23} \\ b_{13} & b_{23} & b_{33} \end{pmatrix} \quad \text{and} \quad D_\Psi = \begin{pmatrix} -b_{11} & 0 & 0 \\ 0 & -b_{22} & 0 \\ 0 & 0 & -b_{33} \end{pmatrix},$$

where

$$\begin{aligned}
b_{11} &= \frac{4}{(\pi^2 - 4(\sin^{-1}\rho_{12})^2)(1-\rho_{12}^2)}, \\
b_{22} &= \frac{4}{(\pi^2 - 4(\sin^{-1}\rho_{13})^2)(1-\rho_{13}^2)}, \\
b_{33} &= \frac{4}{(\pi^2 - 4(\sin^{-1}\rho_{23})^2)(1-\rho_{23}^2)}, \\
b_{12} &= \frac{16\sin^{-1}\rho_{12}\sin^{-1}\rho_{13} - 8\pi\sin^{-1}\rho_{23}}{(\pi^2 - 4(\sin^{-1}\rho_{12})^2)(\pi^2 - 4(\sin^{-1}\rho_{13})^2)(1-\rho_{12}^2)^{1/2}(1-\rho_{13}^2)^{1/2}}, \\
b_{13} &= \frac{16\sin^{-1}\rho_{12}\sin^{-1}\rho_{23} - 8\pi\sin^{-1}\rho_{13}}{(\pi^2 - 4(\sin^{-1}\rho_{12})^2)(\pi^2 - 4(\sin^{-1}\rho_{23})^2)(1-\rho_{12}^2)^{1/2}(1-\rho_{23}^2)^{1/2}}, \\
b_{23} &= \frac{16\sin^{-1}\rho_{13}\sin^{-1}\rho_{23} - 8\pi\sin^{-1}\rho_{12}}{(\pi^2 - 4(\sin^{-1}\rho_{13})^2)(\pi^2 - 4(\sin^{-1}\rho_{23})^2)(1-\rho_{13}^2)^{1/2}(1-\rho_{23}^2)^{1/2}}.
\end{aligned}$$

After simplification, $J_\Psi^{-1} = D_\Psi^{-1} M_\Psi (D_\Psi^{-1})^T$ turns out to be equal to I^{-1}. Therefore by M-optimality, the IFM approach is as efficient as the IFS approach.

The algebraic computation in this example was carried out with the help of the symbolic manipulation software Maple (Char *et al.* 1992). Maple is also used for other analytical examples in this section. For completeness, the Maple program for this example is listed in Appendix A. The Maple programs for other examples in this section are similar. □

Example 4.4 (Trivariate probit, exchangeable) Suppose now we have a trivariate probit model with known cut-off points, such that $P(111) = \Phi_3(0,0,0,\rho,\rho,\rho)$. That is, the latent variables are permutation-symmetric or exchangeable. With (4.2), we obtain

$$\begin{cases}
P_1(1) = P_2(1) = P_3(1) = \Phi(0) = \dfrac{1}{2}, \\
P_{12}(11) = P_{13}(11) = P_{23}(11) = \Phi_2(0,0,\rho) = \dfrac{1}{4} + \dfrac{1}{2\pi}\sin^{-1}\rho, \\
P(111) = \Phi_3(0,0,0,\rho,\rho,\rho) = \dfrac{1}{8} + \dfrac{3}{4\pi}\sin^{-1}\rho.
\end{cases}$$

The MLE of ρ is

$$\hat{\rho} = \sin\left(\frac{\pi(4n_{111} + 4n_{000} - n)}{6n}\right).$$

Based on the full loglikelihood function (4.3), we calculate the Fisher information for ρ (using Maple). The asymptotic variance of $\hat{\rho}$ is found to be

$$\mathrm{Var}(\hat{\rho}) = \frac{(1-\rho^2)(\pi + 6\sin^{-1}\rho)(\pi - 2\sin^{-1}\rho)}{12n}.$$

Let the IFM for one observation be $\Psi = (\psi_{12}, \psi_{13}, \psi_{23})$. We use WA and PMLA to estimate the common parameter ρ:

i. WA: We have

$$M_\Psi = \begin{pmatrix} a & b & b \\ b & a & b \\ b & b & a \end{pmatrix} \quad \text{and} \quad D_\Psi = \begin{pmatrix} -a & 0 & 0 \\ 0 & -a & 0 \\ 0 & 0 & -a \end{pmatrix},$$

where

$$a = \frac{4}{(\pi^2 - 4(\sin^{-1}\rho)^2)(1-\rho^2)},$$
$$b = \frac{8\sin^{-1}\rho}{(\pi - 2\sin^{-1}\rho)(\pi + 2\sin^{-1}\rho)^2(1-\rho^2)}.$$

Thus

$$J_\Psi^{-1} = D_\Psi^{-1} M_\Psi (D_\Psi^{-1})^T = \begin{pmatrix} a^{-1} & a^{-2}b & a^{-2}b \\ a^{-2}b & a^{-1} & a^{-2}b \\ a^{-2}b & a^{-2}b & a^{-1} \end{pmatrix}.$$

Assume the IFME of $\rho_{12}, \rho_{13}, \rho_{23}$ are $\tilde{\rho}_{12}, \tilde{\rho}_{13}, \tilde{\rho}_{23}$ respectively. With WA, we find the weighting vector $\mathbf{u} = (1/3, 1/3, 1/3)'$. So the IFME of ρ, $\tilde{\rho}_w$, is

$$\tilde{\rho}_w = \frac{1}{3}(\tilde{\rho}_{12} + \tilde{\rho}_{13} + \tilde{\rho}_{23}),$$

and the asymptotic variance of $\tilde{\rho}_w$ is

$$\begin{aligned} \mathrm{Var}(\tilde{\rho}) &= \frac{1}{n}\mathbf{u}' J_\Psi^{-1}\mathbf{u} \\ &= \frac{1}{9}\frac{(1-\rho^2)(\pi^2 - 4(\sin^{-1}\rho)^2)}{4n} + \frac{2}{3}\frac{(1-\rho^2)(\pi - 2\sin^{-1}\rho)\sin^{-1}\rho}{2n} \\ &= \frac{(1-\rho^2)(\pi + 6\sin^{-1}\rho)(\pi - 2\sin^{-1}\rho)}{12n}. \end{aligned}$$

ii. PMLA: The IFM is $\Psi = \psi_{12} + \psi_{13} + \psi_{23}$. Thus

$$\begin{aligned} M_\Psi &= E(\psi_{12} + \psi_{13} + \psi_{23}) \\ &= \sum_{\{y_1 y_2 y_3\}} P(y_1 y_2 y_3) \left(\sum_{j,k=1; j<k}^{3} \frac{1}{P(y_j y_k)} \frac{\partial P(y_j y_k)}{\partial \rho} \right)^2, \end{aligned} \tag{4.4}$$

and

$$D_\Psi = E(\partial(\psi_{12}+\psi_{13}+\psi_{23})/\partial\rho)$$
$$= \sum_{\{y_1y_2y_3\}} P(y_1y_2y_3) \sum_{j,k=1;j<k}^{3} \left[-\frac{1}{P_{jk}^2(y_jy_k)} \left(\frac{\partial P_{jk}(y_jy_k)}{\partial\rho} \right)^2 + \frac{1}{P_{jk}(y_jy_k)} \frac{\partial^2 P_{jk}(y_jy_k)}{\partial\rho^2} \right]. \tag{4.5}$$

We find (using Maple)

$$M_\Psi = \frac{12(\pi + 6\sin^{-1}\rho)}{(1-\rho^2)(\pi - 2\sin^{-1}\rho)(\pi + 2\sin^{-1}\rho)^2},$$
$$D_\Psi = -\frac{12}{(\pi^2 - 4(\sin^{-1}\rho)^2)(1-\rho^2)}.$$

The evaluation of $J_\Psi^{-1} = D_\Psi^{-1} M_\Psi (D_\Psi^{-1})^T$ leads to the asymptotic variance of $\tilde{\rho}_p$

$$\mathrm{Var}(\tilde{\rho}_p) = \frac{(1-\rho^2)(\pi + 6\sin^{-1}\rho)(\pi - 2\sin^{-1}\rho)}{12n}.$$

We have so far shown that $\mathrm{Var}(\tilde{\rho}) = \mathrm{Var}(\tilde{\rho}_w) = \mathrm{Var}(\tilde{\rho}_p)$, which means that the IFM with WA and PMLA lead to an estimate as efficient as that from IFS approach.

Any single estimating equation from IFM also gives an asymptotically unbiased estimate of ρ, and the $\tilde{\rho}$ from each of these estimating equations has the same asymptotic properties because of the exchangeability of the latent variables. The ratio of the asymptotic variance of the IFME of ρ from one estimating equation to the asymptotic variance of the MLE ρ is found to be $[3(\pi + 2\sin^{-1}\rho)]/[\pi + 6\sin^{-1}\rho]$. Figure 4.1 is a plot of the ratio versus $\rho \in [-0.5, 1]$. The ratio decreases from ∞ to 1.5 as ρ increases from -0.5 to 1. When $\rho = 0$, the ratio value is 3. These imply that the estimate from a single estimating equation has relatively high efficiency relative to the MLE when there is high positive correlation, but performs poorly when there is high negative correlation.

□

Example 4.5 (Trivariate probit, AR(1)) Suppose we have a trivariate probit model with known cut-off points, such that $P(111) = \Phi_3(0,0,0,\rho,\rho^2,\rho)$. That is, the latent variables have AR(1) correlation structure. With (4.2), we obtain

$$\begin{cases} P_1(1) = P_2(1) = P_3(1) = \Phi(0) = \frac{1}{2}, \\ P_{12}(11) = P_{23}(11) = \Phi_2(0,0,\rho) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho, \\ P_{13}(11) = \Phi_2(0,0,\rho^2) = \frac{1}{4} + \frac{1}{2\pi}\sin^{-1}\rho^2, \\ P(111) = \Phi_3(0,0,0,\rho,\rho^2,\rho) = \frac{1}{8} + \frac{1}{2\pi}\sin^{-1}\rho + \frac{1}{4\pi}\sin^{-1}\rho^2. \end{cases}$$

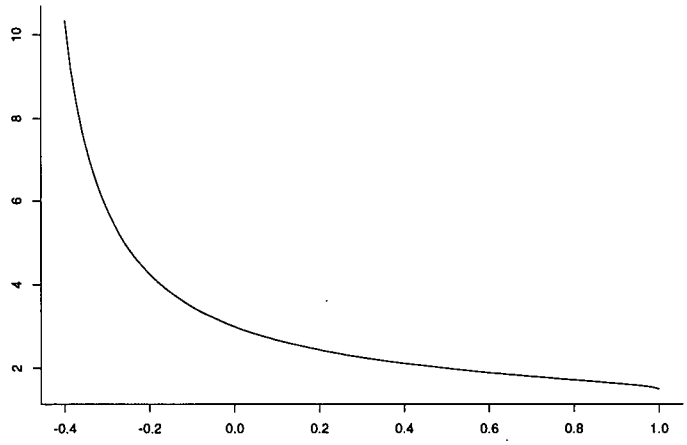

Figure 4.1: Trivariate probit, exchangeable: The efficiency of $\tilde{\rho}$ from the margin $(1,2)$ (or $(1,3)$, $(2,3)$)

Based on the full loglikelihood function (4.3), the asymptotic variance of $\hat{\rho}$ is found (using Maple) to be

$$\mathrm{Var}(\hat{\rho}) = \frac{\pi(1-\rho^4)a_1a_2a_3}{8\left[\rho^2a_4 + (1+\rho^2)a_5 + \rho(1+\rho^2)^{1/2}a_6\right]}, \tag{4.6}$$

where

$$\begin{aligned}
a_1 &= \pi - 2\sin^{-1}\rho^2,\\
a_2 &= \pi - 4\sin^{-1}\rho + 2\sin^{-1}\rho^2,\\
a_3 &= \pi + 4\sin^{-1}\rho + 2\sin^{-1}\rho^2,\\
a_4 &= 2\pi^2 - 16(\sin^{-1}\rho)^2 + 4\pi\sin^{-1}\rho^2,\\
a_5 &= \pi^2 - 4(\sin^{-1}\rho^2)^2,\\
a_6 &= 16\sin^{-1}\rho\sin^{-1}\rho^2 - 8\pi\sin^{-1}\rho.
\end{aligned}$$

Let the IFM for one observation be $\Psi = (\psi_{12}, \psi_{13}, \psi_{23})$. We use WA and PMLA to estimate the common parameter ρ:

i. WA: We have

$$M_\Psi = \begin{pmatrix} a & c & d \\ c & b & c \\ d & c & a \end{pmatrix} \quad \text{and} \quad D_\Psi = \begin{pmatrix} -a & 0 & 0 \\ 0 & -b & 0 \\ 0 & 0 & -a \end{pmatrix},$$

where

$$a = \frac{4}{(\pi^2 - 4(\sin^{-1}\rho)^2)(1-\rho^2)},$$
$$b = \frac{16\rho^2}{(\pi^2 - 4(\sin^{-1}\rho^2)^2)(1-\rho^4)},$$
$$c = \frac{16\rho\sin^{-1}\rho}{(\pi^2 - 4(\sin^{-1}\rho)^2)(\pi + 2\sin^{-1}\rho^2)(1-\rho^2)(1+\rho^2)^{1/2}},$$
$$d = \frac{8\pi\sin^{-1}\rho^2 - 16(\sin^{-1}\rho)^2}{(\pi^2 - 4(\sin^{-1}\rho)^2)^2(1-\rho^2)}.$$

Thus

$$J_\Psi^{-1} = D_\Psi^{-1} M_\Psi (D_\Psi^{-1})^T = \begin{pmatrix} a^{-1} & c(ab)^{-1} & da^{-2} \\ c(ab)^{-1} & b^{-1} & c(ab)^{-1} \\ da^{-2} & c(ab)^{-1} & a^{-1} \end{pmatrix}.$$

Assume the IFME of $\rho_{12}, \rho_{13}, \rho_{23}$ are $\tilde{\rho}_{12}, \tilde{\rho}_{13}, \tilde{\rho}_{23}$ respectively, and let $\tilde{\boldsymbol{\rho}} = (\tilde{\rho}_{12}, \tilde{\rho}_{13}, \tilde{\rho}_{23})'$. With WA, the IFME of ρ, $\tilde{\rho}_w$, is

$$\tilde{\rho}_w = \mathbf{u}'\tilde{\boldsymbol{\rho}},$$

where the weighting vector $\mathbf{u} = (u_1, u_2, u_3)' = J_\Psi \mathbf{1}/(\mathbf{1}' J_\Psi \mathbf{1})$. We find that

$$u_1 = u_3 = \frac{a_1 a_2 a_3 a_7}{2a_8\left[\rho^2 a_4 + (1+\rho^2)a_5 + \rho(1+\rho^2)^{1/2}a_6\right]},$$
$$u_2 = \frac{(2\rho^2 a_4 + \rho(1+\rho^2)^{1/2}a_6)a_2 a_3}{2a_8\left[\rho^2 a_4 + (1+\rho^2)a_5 + \rho(1+\rho^2)^{1/2}a_6\right]},$$

where a_1, a_2, a_3, a_4, a_5, and a_6 as above, and

$$a_7 = \pi + \pi\rho^2 + 2\sin^{-1}\rho^2 + 2\rho^2\sin^{-1}\rho^2 - 4\rho(\rho^2+1)^{1/2}\sin^{-1}\rho,$$
$$a_8 = \pi^2 + 4\pi\sin^{-1}\rho^2 + 4(\sin^{-1}\rho^2)^2 - 16(\sin^{-1}\rho)^2.$$

Figure 4.2 is a plot of the weights versus $\rho \in [-1, 1]$. The asymptotic variance of $\tilde{\rho}$ is

$$\mathrm{Var}(\tilde{\rho}) = \frac{1}{n}\mathbf{u}' J_\Psi^{-1}\mathbf{u},$$

which turns out to be the same as (4.6).

ii. PMLA: The IFM is $\Psi = \psi_{12} + \psi_{13} + \psi_{23}$. Following (4.4) and (4.5), we calculate (using Maple) the corresponding M_Ψ and D_Ψ, and then $\mathrm{Var}(\tilde{\rho}_p) = J_\Psi^{-1}$. The algebraic expression for $\mathrm{Var}(\tilde{\rho}_p)$ is complicated, so we do not display it. Instead, we plot the ratio $\mathrm{Var}(\tilde{\rho}_p)/\mathrm{Var}(\hat{\rho})$ versus $\rho \in [-1, 1]$ in the Figure 4.3. The maximum of the ratio is 1.0391, which is attained at $\rho = 0.3842$ and $\rho = -0.3842$.

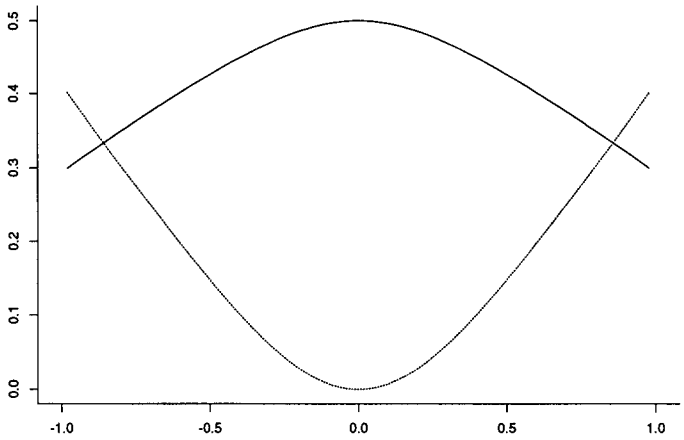

Figure 4.2: Trivariate probit, AR(1): The weight u_1 (or u_3) versus ρ (solid line) and the weight u_2 versus ρ (dash line).

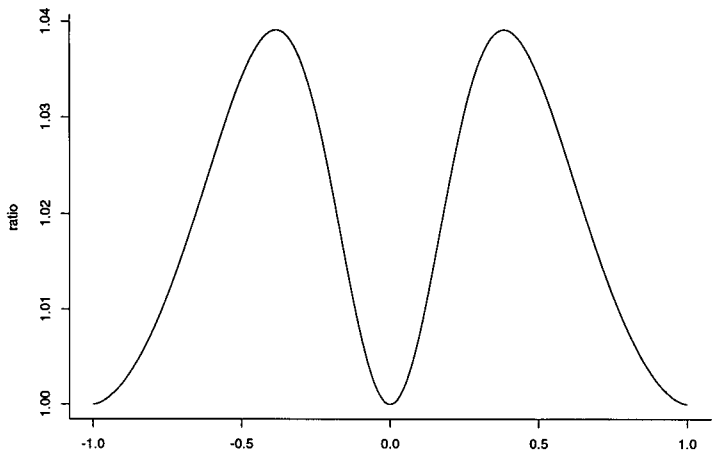

Figure 4.3: Trivariate probit, AR(1): The efficiency of $\tilde{\rho}_p$

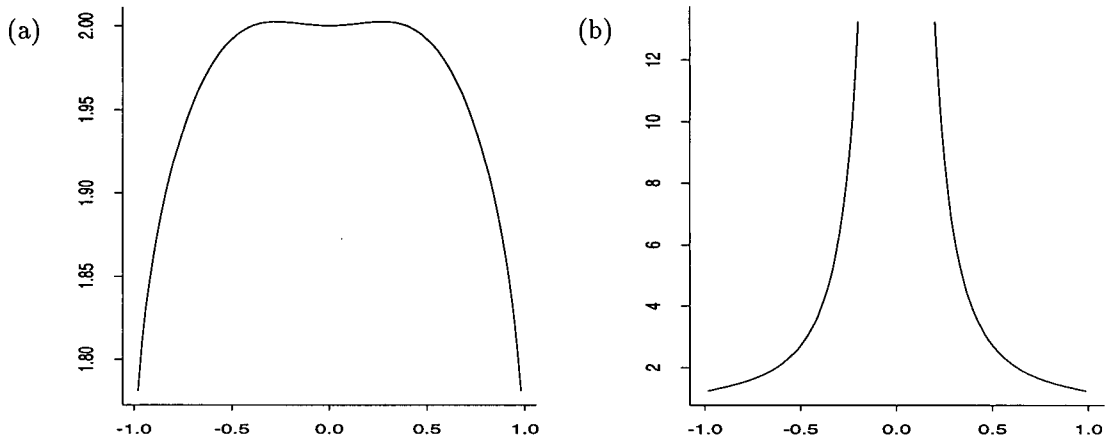

Figure 4.4: Trivariate probit, AR(1): (a) The efficiency of $\tilde{\rho}$ from the margins $(1,2)$ or $(2,3)$. (b) The efficiency of $\tilde{\rho}$ from the margin $(1,3)$.

The above results show that IFM with WA leads to an estimate as efficient as the IFS approach in the AR(1) situation, and IFM with simple PMLA leads to a slightly less efficient estimator (ratio≤ 1.04).

The $\tilde{\rho}$ from the estimating equations based on margin $(1,2)$ (or $(2,3)$) is different from the $\tilde{\rho}$ based on margin $(1,3)$. For $\tilde{\rho}$ from IFM with the $(1,2)$ (or $(2,3)$) margin, the ratio of the asymptotic variance of the IFME of ρ to the asymptotic variance $\hat{\rho}$ is

$$r_1 = \frac{2(\pi^2 - 4(\sin^{-1}\rho)^2)\left[\rho^2 a_4 + (1+\rho^2)a_5 + \rho(1+\rho^2)^{1/2}a_6\right]}{\pi(1+\rho^2)a_1 a_2 a_3}.$$

For $\tilde{\rho}$ from IFM with $(1,3)$ margin, the corresponding ratio is

$$r_2 = \frac{(\pi^2 - 4(\sin^{-1}\rho^2)^2)\left[\rho^2 a_4 + (1+\rho^2)a_5 + \rho(1+\rho^2)^{1/2}a_6\right]}{2\pi\rho^2 a_1 a_2 a_3}.$$

We plot r_1 and r_2 versus $\rho \in [-1,1]$ in Figure 4.4. We see that when ρ goes from -1 to 0, r_1 increases from 1.707 to values around 2. When ρ goes from 0 to 1, r_1 decreases from values around 2 to 1.707. Similarly, r_2 increases from 1.207 to ∞ as ρ goes from -1 to 0, and decreases from ∞ to 1.207 as ρ goes from 0 to 1. We conclude that the $(1,3)$ margin by itself leads to an inefficient estimator in a wide range of the values of ρ. We notice that $r_2 > r_1$ when $\rho < 0.6357$, $r_2 < r_1$ when $\rho > 0.6357$, and $r_1 = r_2 = 1.97$ when $\rho = 0.6357$.

□

Example 4.6 (Trivariate MCD model for binary data with Morgenstern copula) Suppose we have a trivariate MCD model for binary data with Morgenstern copula, such that

$$P(111) = u_1u_2u_3[1 + \theta_{12}(1-u_1)(1-u_2) + \theta_{13}(1-u_1)(1-u_3) + \theta_{23}(1-u_2)(1-u_3)],\ |\theta_{ij}| \leq 1,$$

where the dependence parameters θ_{12}, θ_{13} and θ_{23} obey several constraints:

$$\begin{aligned} &1 + \theta_{12} + \theta_{13} + \theta_{23} \geq 0,\ \ 1 + \theta_{13} \geq \theta_{23} + \theta_{23}, \\ &1 + \theta_{12} \geq \theta_{13} + \theta_{23},\ \ 1 + \theta_{23} \geq \theta_{12} + \theta_{13}. \end{aligned} \tag{4.7}$$

We have $P_j(1) = u_j$, $j = 1, 2, 3$, and $P_{jk}(11) = [1 + \theta_{jk}(1-u_j)(1-u_k)]\, u_j u_k$, $1 \leq j < k \leq 3$. Assume u_j are given, and the parameters of interest are θ_{12}, θ_{13} and θ_{23}. The full loglikelihood function is (4.3). The Fisher information matrix for the parameters θ_{12}, θ_{13} and θ_{23} is I. Assume we have IFM for one observation $\Psi = (\psi_{12}, \psi_{13}, \psi_{23})$. The Godambe information for Ψ is $J_\Psi = D_\Psi M_\Psi^{-1}(D_\Psi)^T$. We proceed to calculate M_Ψ and D_Ψ. The algebraic expression of I and J_Ψ are extremely complicated. We used Maple to output algebraic results in the form of C code and then numerically evaluated the ratio

$$r_g = \frac{\mathrm{Tr}(J_\Psi^{-1})}{\mathrm{Tr}(I^{-1})},$$

where r_g means the general efficiency ratio. For this purpose, we first generate n_1 uniform points $(\theta_{12}, \theta_{13}, \theta_{23})$ from the cube $[-1, 1]^3$ in three dimensional space under the constraints (4.7), and then order these n_1 points based on the value of $|\theta_{12}| + |\theta_{13}| + |\theta_{23}|$ from the smallest to the largest. For each one of the n_1 points $(\theta_{12}, \theta_{13}, \theta_{23})$, we generate n_2 points of (u_1, u_2, u_3) with $(\theta_{12}, \theta_{13}, \theta_{23})$ as given dependence parameters in a trivariate Morgenstern copula in (2.5) (see section 4.3 for how to generate multivariate Morgenstern variate), and then order these n_2 points based on the value of $u_1+u_2+u_3$ from the smallest to the largest. Each generated set of $(u_1, u_2, u_3, \theta_{12}, \theta_{13}, \theta_{23})$ determines a trivariate MCD model with Morgenstern copula for binary data. We calculate r_g corresponding to each particular model. Figure 4.5 presents the values of r_g at $n_1 \times n_2 = 300 \times 300$ "grid" points.

We can see from Figure 4.5 that the IFM approach is reasonably efficient in most situations. It is also clear that the magnitude of $|\theta_{12}| + |\theta_{13}| + |\theta_{23}|$ has an effect on the efficiency of the IFM approach, with generally speaking higher efficiency (r_g's value close to 1) when $|\theta_{12}| + |\theta_{13}| + |\theta_{23}|$ is relatively smaller. The magnitude of $u_1 + u_2 + u_3$ has some effect such that the efficiency of the IFM approach is lower at the area close to the boundary of $u_1 + u_2 + u_3$ (that is close to 0 or 3). The following facts show that the general efficiency of IFM approach is quite good: in these 90,000 efficiency (r_g) evaluations, 50% of the r_g values are less than 1.0196, 90% of the r_g values are less than 1.0722, 99% of the r_g values are less than 1.1803 and 99.99% of the r_g values are less

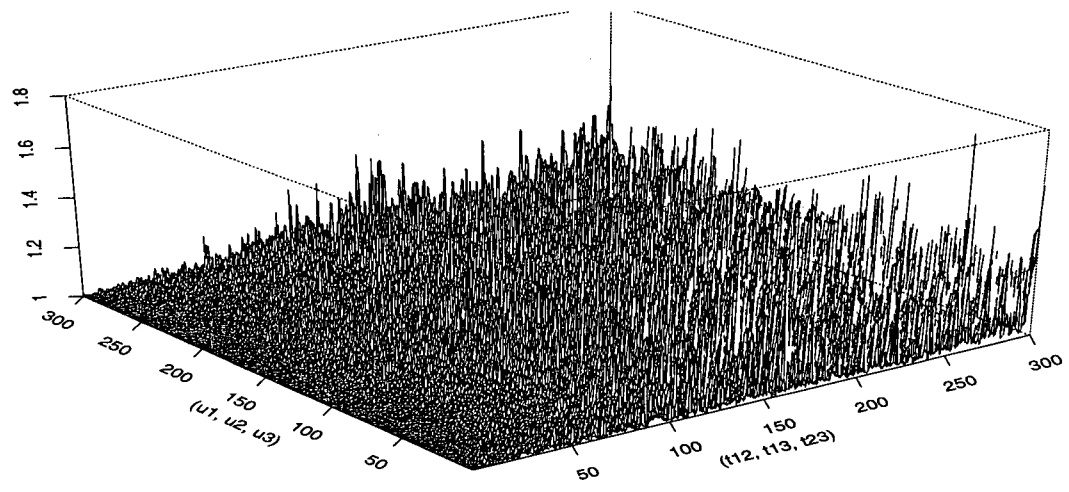

Figure 4.5: Trivariate Morgenstern-binary model: Relative efficiency of IFM approach versus IFS approach.

than 1.4654. The maximum is 1.7084. The minimum is 1. The two plots in Figure 4.6 are used to clarify the above observations. Plot (a) consists of the 90,000 ordered r_g values versus their ordered positions (from 1 to 90,000) in the data set. Plot (b) is a histogram of the r_g values. Overall, we consider the IFM approach to be efficient.

It is also possible to examine the efficiency ratio in some special situations. We study two of them here. The first one is the situation where $u_1 = u_2 = u_3 = u$ and $\theta_{12} = \theta_{13} = \theta_{23} = \theta$ $(-1/3 \leq \theta \leq 1)$. The ratio of the asymptotic variance of $\tilde{\theta}$ (based on WA) versus the asymptotic variance of $\hat{\theta}$ is found to be

$$r_1(u,\theta) = \frac{a_1 a_2 a_3}{b_1 b_2},$$

where

$$a_1 = 27\theta^2u^4 - 54\theta^2u^3 + 33\theta^2u^2 - 10\theta u^2 - 6\theta^2u + 10\theta u - 3\theta - 1,$$

$$a_2 = 3\theta^3u^6 - 9\theta^3u^5 + 9\theta^3u^4 - 11\theta^2u^4 - 3\theta^3u^3 + 22\theta^2u^3 - 12\theta^2u^2 + \theta u^2 + \theta^2u - \theta u - \theta - 1,$$

$$a_3 = \theta u^2 - \theta u + 1,$$

$$b_1 = (3\theta u^2 - 6\theta u + 3\theta + 1)(3\theta u^2 - 4\theta u + \theta + 1),$$

$$b_2 = (3\theta u^2 - 2\theta u + 1)(3\theta u^2 + 1)(\theta u^2 - \theta u - 1)^2.$$

Figure 4.7 is a plot of $r_1(u,\theta)$ versus u $(0 \leq u \leq 1)$ and θ $(-1/3 \leq \theta \leq 1)$. We observe that at the boundaries, when $\theta = 1$, $r_1(u,\theta)$ is in the interval $(1, 1.232)$, and the maximum 1.232 is attained at

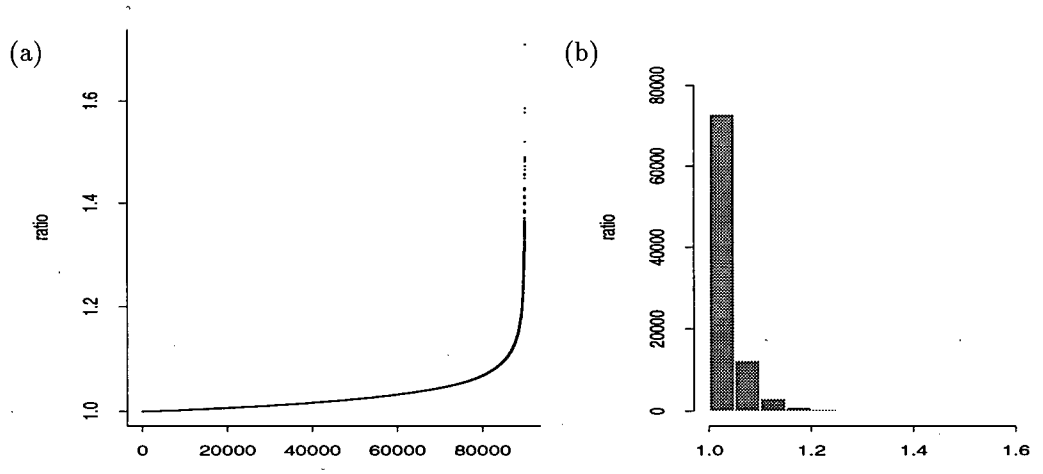

Figure 4.6: Trivariate Morgenstern-binary model: (a). Ordered relative efficiency values of IFM approach versus IFS approach; (b) A histogram of the efficiency value r_g.

$u = 0.2175$ or $u = 0.7825$. When $\theta = -1/3$, $r_1(u, \theta)$ is in the interval $(1, 2)$, and the maximum 2 is attained at $u = 0$ or $u = 1$. Since the maximum ratio is 2 at some extreme points in the parameter space and for the most part the ratio is less than 1.1, we consider the IFM approach to be efficient.

The second special situation is where $u_1 = u_2 = u_3 = u$, $\theta_{12} = \theta_{23} = \theta$ and $\theta_{13} = \theta^2$. The algebraic expression of the ratio $r_2(u, \theta)$ of the asymptotic variance of $\tilde{\theta}$ (based on WA) versus the asymptotic variance of $\hat{\theta}$ extends to several pages. We thus only present a plot of $r_2(u, \theta)$ versus u $(0 \leq u \leq 1)$ and θ $(-1 \leq \theta \leq 1)$ in Figure 4.8. We observe that at the boundaries when $\theta = 1$, the ratio $r_2(u, \theta)$ is in the interval $(1, 1.200097)$, and the maximum is attained at $u = 0.2139$ or $u = 0.7861$. When $\theta = -1$, the ratio $r_2(u, \theta)$ is in the interval $(1, 1.148333)$, and the maximum is attained at $u = 0.154$ or 0.846. Overall, the IFM approach is demonstrated again to be efficient. □

Example 4.7 (Trivariate normal-copula model for binary data) In Examples 4.3, 4.4 and 4.5, we studied the efficiency of the IFM approach versus the IFS approach in the special situations of $P(111) = \Phi_3(0, 0, 0, \rho_{12}, \rho_{13}, \rho_{23})$, $P(111) = \Phi_3(0, 0, 0, \rho, \rho, \rho)$ and $P(111) = \Phi_3(0, 0, 0, \rho, \rho^2, \rho)$. We found that the IFM approach was fully efficient in these situations. For a general trivariate normal-binary model

$$P(111) = \Phi_3(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3), \rho_{12}, \rho_{13}, \rho_{23}), \tag{4.8}$$

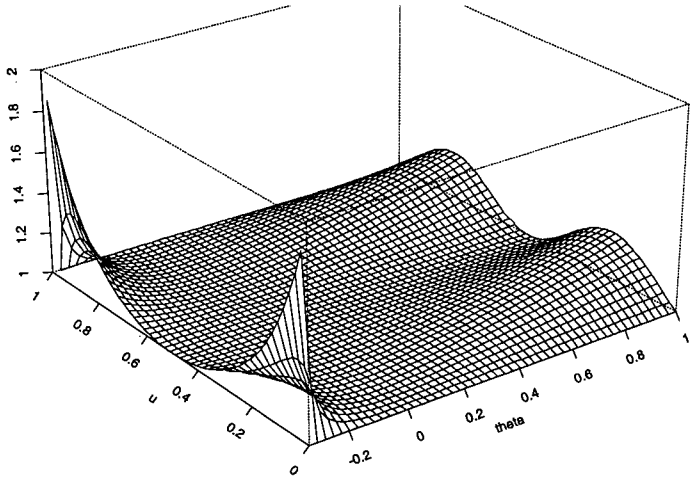

Figure 4.7: Trivariate Morgenstern-binary model: Relative efficiency of IFM approach versus IFS approach when $u_1 = u_2 = u_3$ and $\theta_{12} = \theta_{13} = \theta_{23}$.

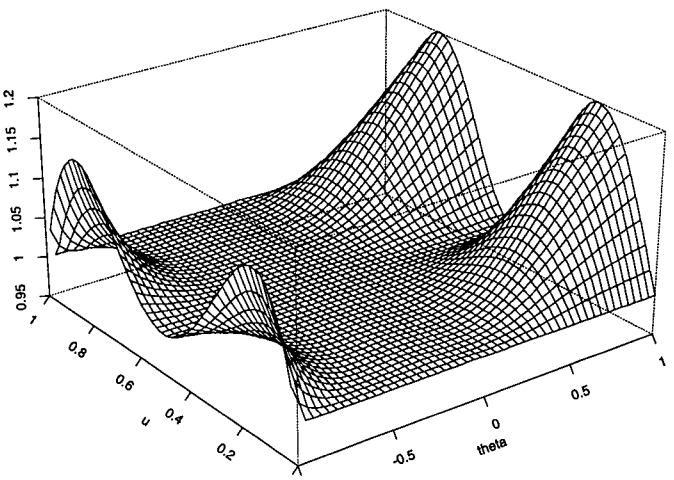

Figure 4.8: Trivariate Morgenstern-binary model: Relative efficiency of IFM approach versus IFS approach when $u_1 = u_2 = u_3$, $\theta_{12} = \theta_{23} = \theta$ and $\theta_{13} = \theta^2$.

the closed form efficiency evaluation, as provided for the trivariate Morgenstern-binary model in Example 4.6, is not possible because $\Phi_3(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \Phi^{-1}(u_3), \rho_{12}, \rho_{13}, \rho_{23})$ does not have closed form. Nevertheless, since a high precision multinormal probability calculation subroutine (Schervish 1984) is available, we can evaluate the efficiency numerically.

With the model (4.8), we have $P_j(1) = u_j$, $j = 1, 2, 3$ and $P_{jk}(11) = \Phi_2(\Phi^{-1}(u_j), \Phi^{-1}(u_k); \rho_{jk})$, $1 \le j < k \le 3$. Assume u_j are given, and the parameters of interest are ρ_{12}, ρ_{13} and ρ_{23}. Let $\theta_1 = \rho_{12}$, $\theta_2 = \rho_{13}$ and $\theta_3 = \rho_{23}$. The Fisher information matrix from one observation for the parameters θ_1, θ_2 and θ_3, I, has the following expression

$$I = \begin{pmatrix} I_{11} & I_{12} & I_{13} \\ I_{12} & I_{22} & I_{23} \\ I_{13} & I_{23} & I_{33} \end{pmatrix},$$

where

$$I_{jj} = \sum_{\{y_1y_2y_3\}} \frac{1}{P_{123}(y_1y_2y_3)} \left(\frac{\partial P_{123}(y_1y_2y_3)}{\partial \theta_j} \right)^2 \quad j = 1, 2, 3,$$

$$I_{jk} = \sum_{\{y_1y_2y_3\}} \frac{1}{P_{123}(y_1y_2y_3)} \frac{\partial P_{123}(y_1y_2y_3)}{\partial \theta_j} \frac{\partial P_{123}(y_1y_2y_3)}{\partial \theta_k} \quad 1 \le j < k \le 3.$$

We can similarly calculate the Godambe information matrix J_Ψ based on the IFM approach for one observation. We then numerically evaluate the ratio (T-optimality)

$$r_g = \frac{\mathrm{Tr}(J_\Psi^{-1})}{\mathrm{Tr}(I^{-1})}$$

in the joint trinormal copula sample space and its parameter space. Similar to Example 4.6 for the trivariate Morgenstern-binary model, we first generate n_1 uniform points of $(\rho_{12}, \rho_{13}, \rho_{23})$ from the cube $[-1, 1]^3$ in three dimensional space under the constraints $1 + 2\rho_{12}\rho_{13}\rho_{23} - \rho_{12}^2 - \rho_{13}^2 - \rho_{23}^2 > 0$ (which guarantees that the determinant of a trinormal correlation matrix is positive) and order these n_1 points based on the value of $|\rho_{12}| + |\rho_{13}| + |\rho_{23}|$ from the smallest to the largest. Then for each one of the n_1 points $(\rho_{12}, \rho_{13}, \rho_{23})$, we generate n_2 points (u_1, u_2, u_3) with $(\rho_{12}, \rho_{13}, \rho_{23})$ as given dependence parameters in a trinormal copula, and order these n_2 points based on the value of $u_1 + u_2 + u_3$ from the smallest to the largest. Each generated set of $(u_1, u_2, u_3, \rho_{12}, \rho_{13}, \rho_{23})$ determines a trivariate normal-binary model. We evaluate r_g corresponding to each particular model. The plot in Figure 4.9 presents the values of r_g at $n_1 \times n_2 = 300 \times 300$ "grid" points for the trivariate normal-copula model for binary data. We observe from the plot that the IFM approach is reasonably efficient in most situations. It is also clear that the magnitude of $|\rho_{12}| + |\rho_{13}| + |\rho_{23}|$ has an effect on the efficiency of the IFM approach, with generally higher efficiency (r_g's value close to 1) when

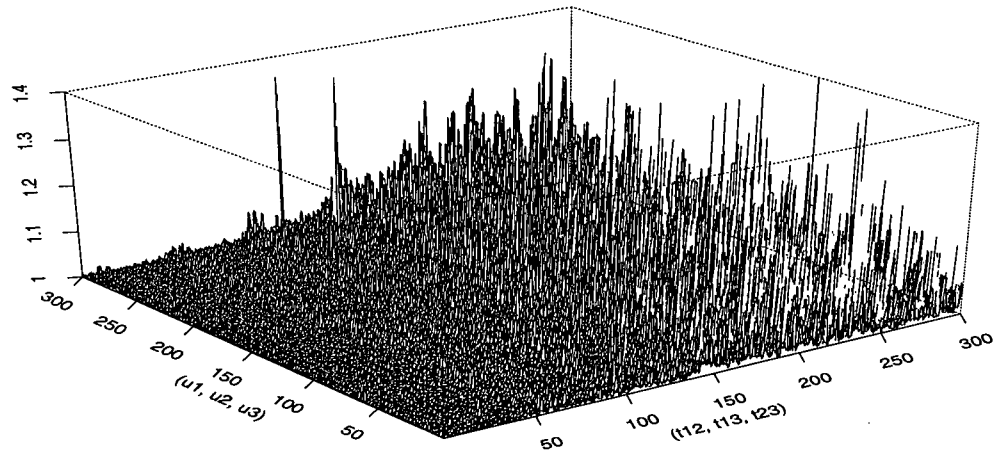

Figure 4.9: Trivariate normal-binary model: Relative efficiency of IFM approach versus IFS approach.

$|\rho_{12}|+|\rho_{13}|+|\rho_{23}|$ is smaller. The magnitude of $u_1+u_2+u_3$ has some effect such that the efficiency of IFM approach is lower at the area close to the boundaries of $u_1+u_2+u_3$ (that is close to 0 or 3). In general the IFM approach is quite efficient: in these 90,000 efficiency (r_g) evaluation, 50% of the r_g values are less than 1.0128, 90% of the r_g values are less than 1.0589, 99% of the r_g values are less than 1.1479 and 99.99% of the r_g values are less than 1.3672. The maximum is 1.8097. The minimum is 1. The two plots in Figure 4.10 are used to clarify the above observations. Plot (a) consists of the 90,000 ordered r_g values versus ordered positions (from 1 to 90,000) in the data set. Plot (b) is a histogram of the r_g values. Overall, we draw the conclusion that the IFM approach is efficient.

In the situation where $u_1=u_2=u_3=u$ and $\rho_{12}=\rho_{13}=\rho_{23}=\rho$ $(-1/2\leq\theta\leq 1)$, let us denote $r_1(u,\rho)$ the ratio of the asymptotic variance of $\tilde{\rho}$ (based on WA) versus the asymptotic variance of $\hat{\theta}$. $r_1(u,\rho)$ has to be evaluated numerically. Figure 4.11 shows a plot of $r_1(u,\rho)$ versus u $(0\leq u\leq 1)$ and ρ $(-1/2\leq\rho\leq 1)$. It is difficult to evaluate $r_1(u,\rho)$ numerically when the values of u and ρ are near the boundaries of the sample space and the parameter space, but generally speaking, the efficiency is lower when the values of u and ρ are close to the boundaries.

In the situation where $u_1=u_2=u_3=u$, $\rho_{12}=\rho_{23}=\rho$ and $\rho_{13}=\rho^2$ $(\rho\in[-1,1])$, we observed similar efficiency behaviour. These results are not presented here. □

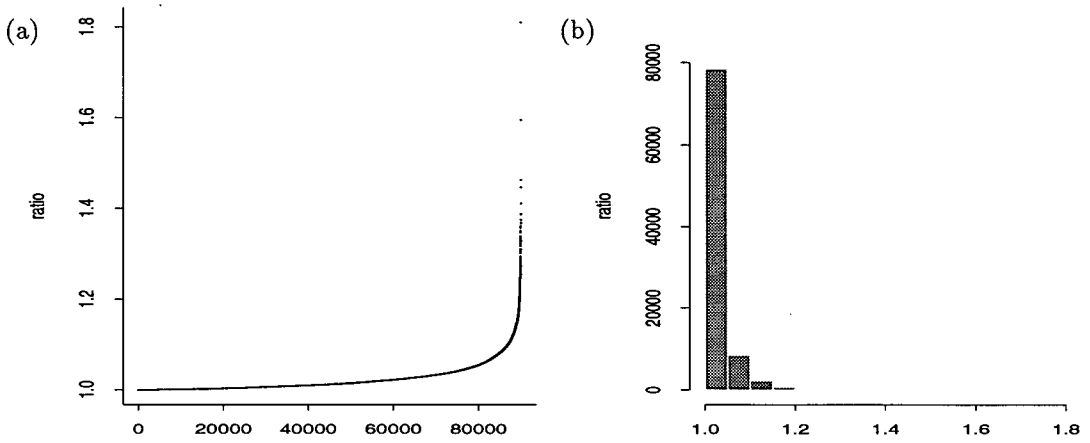

Figure 4.10: Trivariate normal-binary model: (a). Ordered relative efficiency of IFM approach versus IFS approach; (b) A histogram of the efficiency r_g.

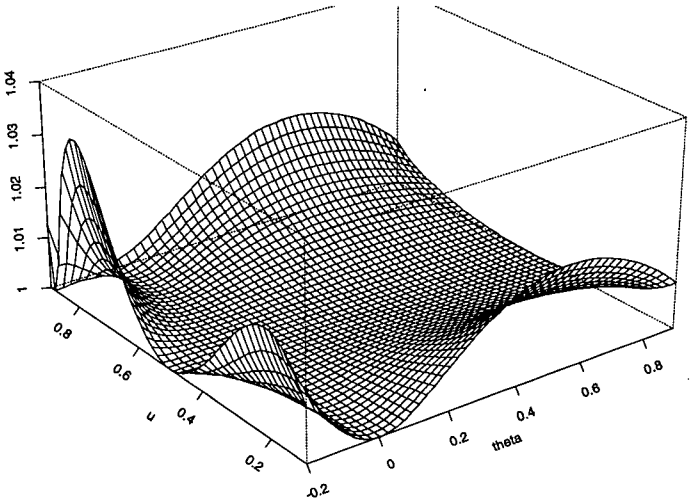

Figure 4.11: Trivariate normal-binary model: Relative efficiency of IFM approach versus IFS approach when $u_1 = u_2 = u_3$ and $\rho_{12} = \rho_{13} = \rho_{23}$.

We have seen from the trivariate normal-copula model for binary data and the trivariate Morgenstern-copula model for binary data that, in some situations, IFM is as efficient as IFS (*e.g.* when $u = 0.5$ for normal-binary model and $u = 0$, 0.5 or 1 for Morgenstern-binary model). In other situations, the efficiency of IFM relative to IFS varies from 1 to a value very close to 1. It is hoped the above results may help to develop intuition for the efficiency of IFM. We would guess that the relative efficiency of IFM to IFS for a model with the MUBE property should be good, as we have seen with the trivariate normal-copula model for binary data and the trivariate Morgenstern-copula model for binary data. However, a general exhaustive analytical investigation such as above is not possible; we have to rely on numerical investigation based on simulation for most of the complicated (higher dimensions or models with covariates) situations.

4.3 Efficiency assessment through simulation

In this section, we give efficiency assessment results through simulation studies with various models. The following are the steps in the simulation and computation: (1) a MCD or MMD model (with MUBE property) is chosen; (2) different sets of model parameters are specified; (3) with a given set of parameters, a sample of size n is generated from the model, and IFM and IFS approaches are used on the same generated data set to estimate the model parameters; (4) with the same set of parameters, step (3) is repeated m times; (5) for any single parameter in the model, say θ, if the estimates of θ with the IFM approach from step (3) and (4) are $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$, and the estimates of θ with IFS approach from step (3) and (4) are $\hat{\theta}_1, \ldots, \hat{\theta}_m$, then we compute

$$\bar{\tilde{\theta}} = \frac{\sum_{i=1}^{m} \tilde{\theta}_i}{m}, \quad \widehat{\mathrm{MSE}}(\tilde{\theta}) = \frac{\sum_{i=1}^{m} (\tilde{\theta}_i - \theta)^2}{m}, \tag{4.9}$$

and

$$\bar{\hat{\theta}} = \frac{\sum_{i=1}^{m} \hat{\theta}_i}{m}, \quad \widehat{\mathrm{MSE}}(\hat{\theta}) = \frac{\sum_{i=1}^{m} (\hat{\theta}_i - \theta)^2}{m}. \tag{4.10}$$

The relative efficiency of IFME to MLE is defined as the ratio r where $r^2 = \widehat{\mathrm{MSE}}(\tilde{\theta})/\widehat{\mathrm{MSE}}(\hat{\theta})$. The values of $\bar{\tilde{\theta}}$, $\sqrt{\widehat{\mathrm{MSE}}(\tilde{\theta})}$, $\bar{\hat{\theta}}$, $\sqrt{\widehat{\mathrm{MSE}}(\hat{\theta})}$ and r are tabulated, with $\sqrt{\widehat{\mathrm{MSE}}(\tilde{\theta})}$ and $\sqrt{\widehat{\mathrm{MSE}}(\hat{\theta})}$ presented in parentheses.

For a fixed sample size, a parameter estimation approach is said to be good if $\bar{\tilde{\theta}}$ (or $\bar{\hat{\theta}}$) is close to θ, and if $\sqrt{\widehat{\mathrm{MSE}}(\tilde{\theta})}$ (or $\sqrt{\widehat{\mathrm{MSE}}(\hat{\theta})}$) is small. There is no "good" in the strict sense, it should be understood in terms of inference, interpretation (*i.e.* no misleading interpretation or false inference would be derived, assuming the model is correct) and in comparison with conventional, well-established

approach. The main objective of this section is to show that with fairly complex models, the IFM approach still has high efficiency.

Multivariate copula discrete models for binary data

In this subsection, we study the MCD models for binary data. The parameters are assumed to be margin-dependent. In our simulation, we use the MVN copula, and simulate d-dimensional binary observations $\mathbf{y}_i$ $(i = 1, \ldots, n)$ from a multivariate probit model

$$Y_{ij} = I(Z_{ij} < z_{ij}), \quad j = 1, \ldots, d, \ i = 1, \ldots, n,$$

where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})' \sim MVN_d(\mathbf{0}, \Theta_i)$ with $z_{ij} = \boldsymbol{\beta}_j' \mathbf{x}_{ij}$, and $\Theta_i = (\theta_{i,jk})$ assumed to be free of covariates, that is $\Theta_i = \Theta$ or $\theta_{i,jk} = \theta_{jk}$, $\forall\ i$. We transform the dependence parameter θ_{jk} with $\theta_{jk} = (\exp(\alpha_{jk}) - 1)/(\exp(\alpha_{jk}) + 1)$, and estimate α_{jk} instead of θ_{jk}. We use the following simulation scheme:

1. The sample size is n, the number of simulations is N; both are reported in the tables.

2. For $d = 3$, we study the two situations: $Y_{ij} = I(Z_{ij} < z_j)$ and $Y_{ij} = I(Z_{ij} < \beta_{j0} + \beta_{j1} x_{ij})$. For each situation, two general dependence structures are chosen: $\theta_{12} = \theta_{13} = \theta_{23} = 0.6$ (or $\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$) and $\theta_{12} = \theta_{23} = 0.8$ (or $\alpha_{12} = \alpha_{23} = 2.1972$), $\theta_{13} = 0.64$ (or $\alpha_{13} = 1.5163$). Other parameters are:

   (a) With no covariates, with $\mathbf{z} = (0, 0, 0)'$.

   (b) With covariates, with $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \beta_{30})' = (0.7, 0.5, 0.3)'$ and $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{21}, \beta_{31})' = (0.5, 0.5, 0.5)'$. Situations where x_{ij} is discrete and continuous are considered. For the discrete situation, $x_{ij} = I(U \leq 0)$ where $U \sim \mathrm{U}(-1, 1)$; for the continuous situation, x_{ij}s are margin-independent, that is $x_{ij} = x_i$, with $x_i \sim N(0, 1/4)$.

3. For $d = 4$, we only study $Y_{ij} = I(Z_{ij} < z_j)$. Two dependence structures in the study are $\theta_{12} = \theta_{13} = \theta_{14} = \theta_{23} = \theta_{24} = \theta_{34} = 0.6$ (or $\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$) and $\theta_{12} = \theta_{23} = \theta_{34} = 0.8$ (or $\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$), $\theta_{13} = \theta_{24} = 0.64$ (or $\alpha_{13} = \alpha_{24} = 1.5163$) and $\theta_{14} = 0.512$ (or $\alpha_{14} = 1.1309$). The cut-off points are (a) $\mathbf{z} = (0, 0, 0, 0)'$, (b) $\mathbf{z} = (0.7, 0.7, 0.7, 0.7)'$, (c) $\mathbf{z} = (0.7, 0, 0.7, 0)'$.

The numerical results from MCD models for binary data are presented in Table 4.1 to Table 4.5. These tables lead to two clear conclusions: i) The IFM approach is efficient relative to the

Table 4.1: Efficiency assessment with MCD model for binary data: $d = 3$, $\mathbf{z} = (0, 0, 0)'$, $N = 1000$

n	margin parameters	1 z_1	2 z_2	3 z_3	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
		$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$					
100	IFM	0.003	-0.002	0.005	1.442	1.426	1.420
		(0.131)	(0.121)	(0.128)	(0.376)	(0.380)	(0.378)
	MLE	0.002	-0.003	0.004	1.441	1.426	1.420
		(0.131)	(0.121)	(0.128)	(0.376)	(0.380)	(0.378)
	r	0.998	0.999	0.999	0.999	0.999	0.999
1000	IFM	-0.0006	-0.0016	-0.0008	1.3924	1.3897	1.3906
		(0.040)	(0.038)	(0.039)	(0.114)	(0.114)	(0.113)
	MLE	-0.0018	-0.0028	-0.0019	1.3919	1.3893	1.3902
		(0.040)	(0.038)	(0.039)	(0.114)	(0.114)	(0.113)
	r	0.997	0.997	0.997	1.000	1.001	1.000
		$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$					
100	IFM	0.0027	-0.0006	0.0003	2.2664	1.5571	2.2586
		(0.131)	(0.123)	(0.130)	(0.454)	(0.377)	(0.453)
	MLE	0.0015	-0.0020	-0.0012	2.2646	1.5552	2.2579
		(0.131)	(0.123)	(0.131)	(0.453)	(0.377)	(0.452)
	r	0.999	1.000	0.999	1.001	1.001	1.002
1000	IFM	-0.0006	-0.0001	-0.0005	2.2009	1.5174	2.2043
		(0.040)	(0.038)	(0.039)	(0.135)	(0.118)	(0.136)
	MLE	-0.0023	-0.0020	-0.0022	2.2003	1.5166	2.2036
		(0.040)	(0.038)	(0.039)	(0.135)	(0.118)	(0.137)
	r	0.996	1.000	0.996	0.999	1.000	1.000

ML approach, for small to large sample sizes. The ratio values r are very close to 1 in almost all the situations studied. These results are consistent with the results from the analytical studies reported in the previous section. ii) The MLE may be slightly more efficient than the IFME, but this observation is not conclusive. We would say that IFME and MLE are comparable.

Multivariate copula discrete models for ordinal data

In this subsection, we study the MCD models for ordinal data. The parameters are assumed to be margin-dependent. In our simulation, we use the MVN copula. We simulate d-dimensional ordinal observations $\mathbf{y}_i$ $(i = 1, \ldots, n)$ from a multivariate probit model for ordinal data, such that

$$\begin{cases} Y_j = 1 \text{ iff } z_j(0) < Z_j \leq z_j(1), \\ Y_j = 2 \text{ iff } z_j(1) < Z_j \leq z_j(2), \\ \cdots, \\ Y_j = m_j \text{ iff } z_j(m_j - 1) < Z_j \leq z_j(m_j), \end{cases}$$

where $-\infty = z_j(0) \leq z_j(1) \leq \cdots \leq z_j(m_j - 1) \leq z_j(m_j) = \infty$ are constants, $j = 1, 2, \ldots, d$, and $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})' \sim MVN_d(\mathbf{0}, \Theta_i)$ with $z_{ij}(y_{ij}) = \gamma_j(y_{ij}) + \boldsymbol{\beta}_j' \mathbf{x}_{ij}$. $\Theta_i = (\theta_{i,jk})$ is assumed to be free of covariates, that is, $\Theta_i = \Theta$ or $\theta_{i,jk} = \theta_{jk}$, $\forall\, i$. We transform the dependence parameters

Table 4.2: Efficiency assessment with MCD model for binary data: $d = 3$, $\boldsymbol{\beta}_0 = (0.7, 0.5, 0.3)'$, $\boldsymbol{\beta}_1 = (0.5, 0.5, 0.5)'$, x_{ij} discrete, $N = 1000$

n	margin parameters	1 β_{10}	β_{11}	2 β_{20}	β_{21}	3 β_{30}	β_{31}	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
		$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$								
100	IFM	0.694	0.559	0.496	0.547	0.294	0.526	1.446	1.447	1.435
		(0.199)	(0.385)	(0.192)	(0.335)	(0.186)	(0.282)	(0.530)	(0.492)	(0.455)
	MLE	0.692	0.557	0.495	0.545	0.293	0.526	1.447	1.450	1.438
		(0.198)	(0.347)	(0.192)	(0.319)	(0.185)	(0.282)	(0.498)	(0.490)	(0.458)
	r	1.005	1.108	1.001	1.050	1.001	1.001	1.064	1.004	0.993
1000	IFM	0.700	0.501	0.498	0.509	0.298	0.503	1.395	1.386	1.385
		(0.063)	(0.100)	(0.058)	(0.089)	(0.058)	(0.085)	(0.145)	(0.136)	(0.131)
	MLE	0.699	0.500	0.497	0.508	0.298	0.503	1.395	1.387	1.387
		(0.063)	(0.099)	(0.058)	(0.089)	(0.058)	(0.085)	(0.145)	(0.136)	(0.131)
	r	1.002	1.013	1.000	1.006	0.998	1.001	1.001	1.001	1.001
		$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$								
100	IFM	0.694	0.559	0.496	0.540	0.293	0.534	2.392	1.599	2.314
		(0.199)	(0.385)	(0.198)	(0.312)	(0.186)	(0.289)	(0.790)	(0.592)	(0.669)
	MLE	0.692	0.558	0.494	0.542	0.292	0.534	2.352	1.591	2.314
		(0.199)	(0.344)	(0.198)	(0.309)	(0.187)	(0.290)	(0.675)	(0.548)	(0.597)
	r	1.001	1.117	0.999	1.007	0.995	0.996	1.171	1.081	1.119
1000	IFM	0.700	0.501	0.499	0.503	0.299	0.502	2.205	1.521	2.201
		(0.064)	(0.100)	(0.058)	(0.092)	(0.057)	(0.086)	(0.167)	(0.141)	(0.155)
	MLE	0.699	0.501	0.498	0.503	0.297	0.502	2.207	1.523	2.204
		(0.063)	(0.098)	(0.058)	(0.092)	(0.057)	(0.086)	(0.167)	(0.141)	(0.155)
	r	1.005	1.019	1.000	1.008	0.998	1.001	0.999	1.001	0.999

Table 4.3: Efficiency assessment with MCD model for binary data: $d = 3$, $\boldsymbol{\beta}_0 = (0.7, 0.5, 0.3)'$, $\boldsymbol{\beta}_1 = (0.5, 0.5, 0.5)'$, $x_{ij} = x_i$ continuous, $N = 100$

n	margin parameters	1 β_{10}	β_{11}	2 β_{20}	β_{21}	3 β_{30}	β_{31}	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
		$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$								
100	IFM	0.722	0.529	0.488	0.520	0.312	0.524	1.453	1.403	1.473
		(0.136)	(0.326)	(0.144)	(0.278)	(0.137)	(0.310)	(0.398)	(0.402)	(0.401)
	MLE	0.722	0.532	0.486	0.519	0.311	0.522	1.458	1.407	1.476
		(0.137)	(0.320)	(0.144)	(0.278)	(0.138)	(0.308)	(0.402)	(0.412)	(0.406)
	r	0.999	1.019	0.999	1.002	0.993	1.005	0.990	0.976	0.989
1000	IFM	0.704	0.495	0.501	0.504	0.306	0.504	1.413	1.380	1.391
		(0.042)	(0.089)	(0.046)	(0.084)	(0.041)	(0.093)	(0.140)	(0.109)	(0.124)
	MLE	0.703	0.494	0.500	0.503	0.305	0.503	1.415	1.381	1.393
		(0.042)	(0.090)	(0.045)	(0.084)	(0.040)	(0.093)	(0.139)	(0.109)	(0.124)
	r	1.004	0.988	1.004	0.993	1.007	1.001	1.000	1.006	0.999
		$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$								
100	IFM	0.722	0.529	0.490	0.543	0.300	0.544	2.303	1.556	2.303
		(0.136)	(0.326)	(0.133)	(0.272)	(0.131)	(0.309)	(0.494)	(0.362)	(0.525)
	MLE	0.721	0.532	0.488	0.542	0.298	0.538	2.318	1.550	2.310
		(0.136)	(0.317)	(0.134)	(0.279)	(0.131)	(0.306)	(0.504)	(0.371)	(0.533)
	r	1.000	1.028	0.993	0.976	1.002	1.010	0.981	0.978	0.985
1000	IFM	0.704	0.495	0.502	0.500	0.303	0.506	2.220	1.541	2.213
		(0.042)	(0.089)	(0.045)	(0.076)	(0.041)	(0.091)	(0.155)	(0.123)	(0.142)
	MLE	0.703	0.494	0.499	0.498	0.301	0.505	2.222	1.541	2.215
		(0.042)	(0.089)	(0.045)	(0.075)	(0.041)	(0.091)	(0.156)	(0.124)	(0.142)
	r	1.011	1.002	1.008	1.015	1.010	1.010	0.991	0.997	0.997

Table 4.4: Efficiency assessment with MCD model for binary data: $d = 4$, $\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$, $N = 1000$

n	margin parameters	1 z_1	2 z_2	3 z_3	4 z_4	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
					$\mathbf{z} = (0,0,0,0)'$						
100	IFM	0.002	-0.008	0.002	-0.002	1.417	1.421	1.419	1.420	1.406	1.397
		(0.121)	(0.122)	(0.128)	(0.125)	(0.369)	(0.374)	(0.374)	(0.370)	(0.374)	(0.3 66)
	MLE	0.001	-0.009	0.000	-0.003	1.417	1.419	1.418	1.419	1.405	1.395
		(0.121)	(0.123)	(0.128)	(0.125)	(0.369)	(0.374)	(0.374)	(0.370)	(0.373)	(0.3 65)
	r	0.997	0.996	0.999	0.999	1.001	0.999	1.000	1.000	1.002	1.003
1000	IFM	-0.001	-0.001	-0.001	-0.003	1.388	1.386	1.392	1.385	1.391	1.387
		(0.040)	(0.037)	(0.039)	(0.039)	(0.108)	(0.112)	(0.112)	(0.115)	(0.114)	(0.1 18)
	MLE	-0.002	-0.003	-0.002	-0.004	1.388	1.386	1.392	1.385	1.390	1.387
		(0.040)	(0.037)	(0.039)	(0.039)	(0.108)	(0.112)	(0.112)	(0.115)	(0.114)	(0.1 17)
	r	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
					$\mathbf{z} = (0.7,0.7,0.7,0.7)'$						
100	IFM	0.709	0.703	0.710	0.708	1.447	1.403	1.444	1.405	1.401	1.404
		(0.139)	(0.139)	(0.140)	(0.139)	(0.444)	(0.443)	(0.466)	(0.430)	(0.431)	(0.4 44)
	MLE	0.707	0.702	0.709	0.706	1.447	1.402	1.442	1.403	1.401	1.403
		(0.138)	(0.139)	(0.140)	(0.139)	(0.442)	(0.441)	(0.465)	(0.430)	(0.429)	(0.4 45)
	r	1.002	1.000	0.998	1.000	1.005	1.004	1.001	1.000	1.005	0.997
1000	IFM	0.700	0.703	0.699	0.700	1.388	1.384	1.390	1.380	1.383	1.389
		(0.043)	(0.044)	(0.043)	(0.045)	(0.134)	(0.130)	(0.131)	(0.131)	(0.136)	(0.1 32)
	MLE	0.699	0.701	0.698	0.699	1.389	1.385	1.390	1.382	1.384	1.390
		(0.043)	(0.044)	(0.043)	(0.044)	(0.133)	(0.130)	(0.131)	(0.130)	(0.136)	(0.1 32)
	r	1.000	1.001	1.000	1.004	1.001	0.999	1.001	1.004	1.000	1.000
					$\mathbf{z} = (0.7,0,0.7,0)'$						
100	IFM	0.709	-0.007	0.710	-0.002	1.464	1.403	1.480	1.463	1.406	1.454
		(0.139)	(0.122)	(0.140)	(0.124)	(0.567)	(0.443)	(0.596)	(0.533)	(0.374)	(0.5 74)
	MLE	0.708	-0.009	0.709	-0.003	1.458	1.400	1.472	1.459	1.405	1.447
		(0.138)	(0.122)	(0.140)	(0.124)	(0.512)	(0.445)	(0.541)	(0.493)	(0.373)	(0.5 18)
	r	1.002	0.999	0.998	1.000	1.107	0.996	1.102	1.081	1.003	1.108
1000	IFM	0.700	-0.001	0.699	-0.002	1.392	1.384	1.398	1.386	1.391	1.392
		(0.043)	(0.037)	(0.043)	(0.039)	(0.128)	(0.130)	(0.132)	(0.133)	(0.114)	(0.1 31)
	MLE	0.699	-0.002	0.698	-0.004	1.394	1.385	1.399	1.387	1.390	1.393
		(0.043)	(0.038)	(0.043)	(0.039)	(0.128)	(0.129)	(0.132)	(0.133)	(0.114)	(0.1 31)
	r	1.002	0.999	1.001	0.996	1.000	1.008	1.001	0.999	0.999	1.001

Table 4.5: Efficiency assessment with MCD model for binary data: $d = 4$, $\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$, $\alpha_{13} = \alpha_{24} = 1.5163$, $\alpha_{14} = 1.1309$, $N = 1000$

n	margin parameters	1 z_1	2 z_2	3 z_3	4 z_4	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
					$\mathbf{z} = (0,0,0,0)'$						
1000	IFM	0.002	0.001	0.001	0.001	2.210	1.518	1.137	2.198	1.516	2.204
		(0.039)	(0.038)	(0.039)	(0.041)	(0.135)	(0.116)	(0.106)	(0.131)	(0.115)	(0.128)
	MLE	-0.000	-0.002	-0.001	-0.001	2.209	1.517	1.136	2.197	1.515	2.204
		(0.039)	(0.039)	(0.039)	(0.041)	(0.135)	(0.115)	(0.106)	(0.131)	(0.115)	(0.127)
	r	1.000	0.996	1.004	0.997	1.000	1.002	1.001	1.000	1.000	1.000
					$\mathbf{z} = (0.7,0.7,0.7,0.7)'$						
1000	IFM	0.700	0.701	0.700	0.701	2.206	1.514	1.136	2.205	1.521	2.206
		(0.042)	(0.043)	(0.043)	(0.044)	(0.154)	(0.132)	(0.125)	(0.153)	(0.134)	(0.159)
	MLE	0.699	0.699	0.699	0.700	2.207	1.514	1.135	2.206	1.521	2.208
		(0.042)	(0.044)	(0.043)	(0.044)	(0.154)	(0.130)	(0.124)	(0.153)	(0.132)	(0.159)
	r	1.000	0.999	1.003	0.999	1.000	1.010	1.008	1.000	1.011	1.001
					$\mathbf{z} = (0.7,0,0.7,0)'$						
1000	IFM	0.700	0.001	0.700	0.001	2.212	1.514	1.139	2.209	1.516	2.214
		(0.042)	(0.038)	(0.043)	(0.041)	(0.159)	(0.132)	(0.122)	(0.162)	(0.115)	(0.162)
	MLE	0.699	-0.001	0.699	-0.001	2.215	1.513	1.140	2.214	1.515	2.218
		(0.042)	(0.039)	(0.043)	(0.041)	(0.159)	(0.130)	(0.121)	(0.163)	(0.115)	(0.162)
	r	0.996	0.994	0.998	0.997	0.999	1.014	1.009	0.996	1.008	1.001

θ_{jk} with $\theta_{jk} = (\exp(\alpha_{jk}) - 1)/(\exp(\alpha_{jk}) + 1)$, and estimate α_{jk} instead of θ_{jk}. In our simulation study, we only examine the situation where no covariates are involved in the marginal parameters, and further assume that $m_j = 3$. In these situations, for each margin, we need to estimate two parameters: $z_j(1)$ and $z_j(2)$. We use the following simulation scheme:

1. The sample size is n, the number of simulations is N; both are reported in the tables.

2. For $d = 3$, we study two situations of marginal parameters:

   (a) $\mathbf{z}(1) = (-0.5, -0.5, -0.5)'$, $\mathbf{z}(2) = (0.5, 0.5, 0.5)'$

   (b) $\mathbf{z}(1) = (-0.5, 0, -0.5)'$, $\mathbf{z}(2) = (0.5, 1, 0.5)'$

   and for each situation, two dependence structures are used in the simulation study: $\theta_{12} = \theta_{13} = \theta_{23} = 0.6$ (or $\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$) and $\theta_{12} = \theta_{23} = 0.8$ (or $\alpha_{12} = \alpha_{23} = 2.1972$), $\theta_{13} = 0.64$ (or $\alpha_{13} = 1.5163$).

3. For $d = 4$, we similarly study two situations of marginal parameters:

   (a) $\mathbf{z}(1) = (-0.5, -0.5, -0.5, -0.5)'$, $\mathbf{z}(2) = (0.5, 0.5, 0.5, 0.5)'$

   (b) $\mathbf{z}(1) = (-0.5, 0, -0.5, 0)'$, $\mathbf{z}(2) = (0.5, 1, 0.5, 1)'$

Table 4.6: Efficiency assessment with MCD model for ordinal data: $d = 3$, $\mathbf{z}(1) = (-0.5, -0.5, -0.5)'$, $\mathbf{z}(2) = (0.5, 0.5, 0.5)'$, $N = 1000$

n	margin parameters	1 $z_1(1)$	1 $z_1(2)$	2 $z_2(1)$	2 $z_2(2)$	3 $z_3(1)$	3 $z_3(2)$	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$										
100	IFM	-0.500	0.508	-0.508	0.500	-0.507	0.508	1.413	1.407	1.414
		(0.135)	(0.135)	(0.130)	(0.133)	(0.134)	(0.137)	(0.275)	(0.284)	(0.287)
	MLE	-0.503	0.507	-0.511	0.498	-0.510	0.507	1.413	1.408	1.415
		(0.134)	(0.135)	(0.130)	(0.133)	(0.134)	(0.136)	(0.275)	(0.284)	(0.287)
	r	1.004	1.003	1.000	1.003	1.006	1.003	1.000	0.999	0.998
1000	IFM	-0.501	0.498	-0.501	0.499	-0.502	0.500	1.390	1.386	1.387
		(0.043)	(0.041)	(0.041)	(0.041)	(0.042)	(0.042)	(0.086)	(0.089)	(0.088)
	MLE	-0.504	0.497	-0.504	0.498	-0.504	0.498	1.390	1.386	1.387
		(0.043)	(0.041)	(0.041)	(0.041)	(0.042)	(0.042)	(0.085)	(0.089)	(0.088)
	r	0.998	1.002	0.998	1.002	0.997	1.006	1.005	1.004	1.005
$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$										
100	IFM	-0.500	0.508	-0.506	0.502	-0.509	0.508	2.251	1.542	2.242
		(0.135)	(0.135)	(0.132)	(0.136)	(0.136)	(0.134)	(0.323)	(0.282)	(0.324)
	MLE	-0.504	0.506	-0.512	0.500	-0.513	0.506	2.252	1.539	2.243
		(0.136)	(0.136)	(0.132)	(0.136)	(0.138)	(0.135)	(0.321)	(0.285)	(0.322)
	r	0.990	0.999	0.997	1.005	0.991	0.999	1.005	0.992	1.005
1000	IFM	-0.501	0.498	-0.500	0.498	-0.502	0.500	2.202	1.516	2.199
		(0.043)	(0.041)	(0.041)	(0.040)	(0.042)	(0.041)	(0.093)	(0.088)	(0.097)
	MLE	-0.505	0.496	-0.504	0.496	-0.506	0.498	2.203	1.516	2.200
		(0.043)	(0.041)	(0.041)	(0.040)	(0.042)	(0.041)	(0.093)	(0.088)	(0.097)
	r	0.997	0.999	1.005	1.008	0.998	1.001	0.999	1.004	1.000

and for each situation, two dependence structures are used in the simulation study: $\theta_{12} = \theta_{13} = \theta_{14} = \theta_{23} = \theta_{24} = \theta_{34} = 0.6$ (or $\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$) and $\theta_{12} = \theta_{23} = \theta_{34} = 0.8$ (or $\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$), $\theta_{13} = \theta_{24} = 0.64$ (or $\alpha_{13} = \alpha_{24} = 1.5163$) and $\theta_{14} = 0.512$ (or $\alpha_{14} = 1.1309$).

The numerical results from MCD models for ordinal data are presented in Table 4.6 to Table 4.11. Again, from these tables, we have the following two clear conclusions: i) The IFM approach is efficient relative to the ML approach, for small to large sample sizes. The ratio values r are very close to 1 in almost all the studied situations. ii) MLE may be slightly more efficient than IFME, but this observation is not conclusive. We would say that IFME and MLE are comparable.

Multivariate copula discrete models for count data

In this subsection, we study the MCD models for count data. The parameters are assumed to be margin-dependent. In our simulation, we use the MVN copula. We simulate d-dimensional Poisson observations $\mathbf{y}_i$ $(i = 1, \ldots, n)$ from a multivariate normal-copula Poisson model

$$P(y_1 \cdots y_d) = \sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots+i_d} C(a_{1i_1}, \ldots, a_{di_d}; \Theta),$$

Table 4.7: Efficiency assessment with MCD model for ordinal data: $d = 3$, $\mathbf{z}(1) = (-0.5, 0, -0.5)'$, $\mathbf{z}(2) = (0.5, 1, 0.5)'$, $N = 1000$

n	margin parameters	1 $z_1(1)$	1 $z_1(2)$	2 $z_2(1)$	2 $z_2(2)$	3 $z_3(1)$	3 $z_3(2)$	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
		$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$								
100	IFM	-0.500	0.508	-0.002	1.01	-0.507	0.508	1.429	1.407	1.416
		(0.135)	(0.135)	(0.121)	(0.16)	(0.134)	(0.137)	(0.294)	(0.284)	(0.298)
	MLE	-0.503	0.507	-0.004	1.00	-0.510	0.507	1.430	1.408	1.417
		(0.134)	(0.135)	(0.120)	(0.16)	(0.134)	(0.136)	(0.293)	(0.284)	(0.297)
	r	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
1000	IFM	-0.501	0.498	-0.002	0.998	-0.502	0.500	1.393	1.386	1.389
		(0.043)	(0.041)	(0.038)	(0.049)	(0.042)	(0.042)	(0.089)	(0.089)	(0.090)
	MLE	-0.503	0.497	-0.003	0.996	-0.504	0.498	1.393	1.385	1.389
		(0.043)	(0.041)	(0.038)	(0.049)	(0.042)	(0.042)	(0.089)	(0.089)	(0.090)
	r	0.998	1.000	0.996	1.000	0.997	1.005	1.006	1.004	1.004
		$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$								
100	IFM	-0.500	0.508	-0.001	1.012	-0.509	0.508	2.261	1.542	2.247
		(0.135)	(0.135)	(0.123)	(0.164)	(0.136)	(0.134)	(0.343)	(0.282)	(0.344)
	MLE	-0.504	0.505	-0.004	1.011	-0.513	0.505	2.268	1.538	2.254
		(0.136)	(0.136)	(0.121)	(0.166)	(0.139)	(0.135)	(0.345)	(0.290)	(0.347)
	r	0.994	0.995	1.016	0.989	0.984	0.997	0.993	0.975	0.991
1000	IFM	-0.501	0.498	-0.000	0.999	-0.502	0.500	2.204	1.516	2.199
		(0.043)	(0.041)	(0.038)	(0.049)	(0.042)	(0.041)	(0.099)	(0.088)	(0.101)
	MLE	-0.504	0.496	-0.003	0.996	-0.505	0.498	2.204	1.516	2.199
		(0.043)	(0.041)	(0.038)	(0.049)	(0.042)	(0.041)	(0.099)	(0.087)	(0.100)
	r	0.998	0.999	1.018	1.002	0.998	1.002	1.002	1.006	1.011

Table 4.8: Efficiency assessment with MCD model for ordinal data: $d = 4$, $\mathbf{z}(1) = (-0.5, -0.5, -0.5, -0.5)'$, $\mathbf{z}(2) = (0.5, 0.5, 0.5, 0.5)'$, $\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$, $N = 100$

n	margin parameters	1 $z_1(1)$	1 $z_1(2)$	2 $z_2(1)$	2 $z_2(2)$	3 $z_3(1)$	3 $z_3(2)$	4 $z_4(1)$	4 $z_4(2)$
100	IFM	-0.4928	0.5025	-0.5037	0.4820	-0.4997	0.4986	-0.5088	0.4890
		(0.1238)	(0.1345)	(0.1139)	(0.1377)	(0.1145)	(0.1293)	(0.1365)	(0.1349)
	MLE	-0.4930	0.5023	-0.5043	0.4819	-0.5007	0.4983	-0.5101	0.4877
		(0.1225)	(0.1325)	(0.1145)	(0.1368)	(0.1149)	(0.1292)	(0.1373)	(0.1361)
	r	1.011	1.015	0.995	1.006	0.996	1.001	0.994	0.991
1000	IFM	-0.4954	0.5044	-0.5068	0.4981	-0.4987	0.5015	-0.4966	0.5016
		(0.0433)	(0.0415)	(0.0457)	(0.0436)	(0.0413)	(0.0413)	(0.0459)	(0.0414)
	MLE	-0.4973	0.5026	-0.5094	0.4963	-0.5008	0.4998	-0.4988	0.4998
		(0.0431)	(0.0416)	(0.0463)	(0.0439)	(0.0413)	(0.0413)	(0.0460)	(0.0419)
	r	1.003	0.999	0.987	0.993	1.000	1.001	0.997	0.988

n	margin parameters	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
100	IFM	1.4489	1.4332	1.4174	1.4539	1.4050	1.4149
		(0.2741)	(0.2858)	(0.2906)	(0.2903)	(0.3005)	(0.3088)
	MLE	1.4498	1.4351	1.4203	1.4562	1.4070	1.4175
		(0.2732)	(0.2842)	(0.2903)	(0.2928)	(0.3019)	(0.3042)
	r	1.003	1.006	1.001	0.991	0.995	1.015
1000	IFM	1.3939	1.3937	1.3942	1.3828	1.3739	1.3785
		(0.0786)	(0.0815)	(0.0869)	(0.0833)	(0.0775)	(0.0822)
	MLE	1.3949	1.3950	1.3958	1.3827	1.3745	1.3800
		(0.0795)	(0.0817)	(0.0886)	(0.0834)	(0.0790)	(0.0823)
	r	0.988	0.997	0.980	0.999	0.981	0.998

Table 4.9: Efficiency assessment with MCD model for ordinal data: $d = 4$, $\mathbf{z}(1) = (-0.5, -0.5, -0.5, -0.5)'$, $\mathbf{z}(2) = (0.5, 0.5, 0.5, 0.5)'$, $\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$, $\alpha_{13} = \alpha_{24} = 1.5163$, $\alpha_{14} = 1.1309$, $N = 100$

n	margin parameters	1		2		3		4	
		$z_1(1)$	$z_1(2)$	$z_2(1)$	$z_2(2)$	$z_3(1)$	$z_3(2)$	$z_4(1)$	$z_4(2)$
100	IFM	-0.4928	0.5025	-0.5032	0.4924	-0.5038	0.4917	-0.5142	0.4909
		(0.1238)	(0.1345)	(0.1156)	(0.1357)	(0.1155)	(0.1342)	(0.1349)	(0.1424)
	MLE	-0.4965	0.5050	-0.5049	0.4945	-0.5122	0.4936	-0.5203	0.4936
		(0.1265)	(0.1339)	(0.1162)	(0.1322)	(0.1238)	(0.1328)	(0.1388)	(0.1457)
	r	0.979	1.005	0.994	1.026	0.933	1.011	0.972	0.978
1000	IFM	-0.4954	0.5044	-0.5047	0.5018	-0.5011	0.5019	-0.5035	0.5012
		(0.0433)	(0.0415)	(0.0454)	(0.0435)	(0.0417)	(0.0401)	(0.0433)	(0.0391)
	MLE	-0.4986	0.5017	-0.5079	0.4988	-0.5047	0.4989	-0.5069	0.4984
		(0.0433)	(0.0416)	(0.0460)	(0.0434)	(0.0412)	(0.0408)	(0.0436)	(0.0394)
	r	0.999	0.999	0.988	1.004	1.014	0.984	0.992	0.992

n	margin parameters	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
100	IFM	2.2807	1.5610	1.1782	2.2764	1.5603	2.2516
		(0.3091)	(0.2732)	(0.2897)	(0.3453)	(0.3270)	(0.3331)
	MLE	2.2754	1.5503	1.1795	2.2750	1.5491	2.2477
		(0.2933)	(0.2621)	(0.2802)	(0.3354)	(0.3263)	(0.3291)
	r	1.050	1.040	1.030	1.030	1.000	1.010
1000	IFM	2.2190	1.5263	1.1396	2.1868	1.5055	2.1851
		(0.0915)	(0.0803)	(0.0790)	(0.0884)	(0.0874)	(0.0957)
	MLE	2.2217	1.5267	1.1394	2.1887	1.5055	2.1865
		(0.0916)	(0.0791)	(0.0789)	(0.0871)	(0.0865)	(0.0951)
	r	0.999	1.015	1.002	1.014	1.011	1.007

Table 4.10: Efficiency assessment with MCD model for ordinal data: $d = 4$, $\mathbf{z}(1) = (-0.5, 0, -0.5, 0)'$, $\mathbf{z}(2) = (0.5, 1, 0.5, 1)'$, $\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$, $N = 100$

n	margin parameters	1		2		3		4	
		$z_1(1)$	$z_1(2)$	$z_2(1)$	$z_2(2)$	$z_3(1)$	$z_3(2)$	$z_4(1)$	$z_4(2)$
100	IFM	-0.4928	0.5025	-0.01817	0.9924	-0.4997	0.4986	-0.0108	0.9994
		(0.1238)	(0.1345)	(0.12289)	(0.1507)	(0.1145)	(0.1293)	(0.1159)	(0.1528)
	MLE	-0.4939	0.5020	-0.01733	0.9886	-0.5013	0.4979	-0.0115	0.9964
		(0.1217)	(0.1318)	(0.12165)	(0.1505)	(0.1141)	(0.1296)	(0.1146)	(0.1532)
	r	1.018	1.020	1.010	1.002	1.003	0.998	1.011	0.997
1000	IFM	-0.4954	0.5044	-0.0076	1.0021	-0.4987	0.5015	-0.0017	1.0044
		(0.0433)	(0.0415)	(0.0437)	(0.0450)	(0.0413)	(0.0413)	(0.0405)	(0.0474)
	MLE	-0.4975	0.5026	-0.0095	0.9996	-0.5009	0.5000	-0.0037	1.0018
		(0.0431)	(0.0413)	(0.0436)	(0.0448)	(0.0414)	(0.0413)	(0.0404)	(0.0473)
	r	1.003	1.005	1.001	1.005	0.998	1.000	1.003	1.002

n	margin parameters	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
100	IFM	1.4398	1.4332	1.4428	1.4452	1.4228	1.4345
		(0.2874)	(0.2858)	(0.2801)	(0.2866)	(0.2966)	(0.3429)
	MLE	1.4468	1.4363	1.4467	1.4509	1.4313	1.4341
		(0.2888)	(0.2838)	(0.2781)	(0.2882)	(0.2971)	(0.3409)
	r	0.995	1.007	1.007	0.995	0.998	1.006
1000	IFM	1.4060	1.3937	1.3907	1.3866	1.3798	1.3756
		(0.0822)	(0.0815)	(0.0891)	(0.0806)	(0.0940)	(0.0919)
	MLE	1.4067	1.3947	1.3924	1.3877	1.3813	1.3769
		(0.0820)	(0.0811)	(0.0911)	(0.0808)	(0.0941)	(0.0908)
	r	1.003	1.005	0.978	0.997	0.999	1.012

Table 4.11: Efficiency assessment with MCD model for ordinal data: $d = 4$, $\mathbf{z}(1) = (-0.5, 0, -0.5, 0)'$, $\mathbf{z}(2) = (0.5, 1, 0.5, 1)'$, $\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$, $\alpha_{13} = \alpha_{24} = 1.5163$, $\alpha_{14} = 1.1309$, $N = 100$

n	margin parameters	1		2		3		4	
		$z_1(1)$	$z_1(2)$	$z_2(1)$	$z_2(2)$	$z_3(1)$	$z_3(2)$	$z_4(1)$	$z_4(2)$
100	IFM	-0.4928	0.5025	-0.0217	0.9877	-0.5038	0.4917	-0.01462	0.9791
		(0.1238)	(0.1345)	(0.1241)	(0.1516)	(0.1155)	(0.1342)	(0.10744)	(0.1439)
	MLE	-0.4944	0.5010	-0.0189	0.9796	-0.5090	0.4892	-0.01577	0.9780
		(0.1274)	(0.1342)	(0.1244)	(0.1548)	(0.1176)	(0.1319)	(0.11106)	(0.1449)
	r	0.972	1.002	0.998	0.979	0.981	1.018	0.967	0.994
1000	IFM	-0.4954	0.5044	-0.0006	0.9995	-0.5011	0.5019	-0.0013	1.0018
		(0.0433)	(0.0415)	(0.0406)	(0.0476)	(0.0417)	(0.0401)	(0.0382)	(0.0489)
	MLE	-0.4985	0.5010	-0.0034	0.9956	-0.5046	0.4988	-0.0044	0.9988
		(0.0433)	(0.0410)	(0.0396)	(0.0476)	(0.0416)	(0.0405)	(0.0392)	(0.0482)
	r	0.999	1.013	1.025	0.999	1.003	0.991	0.974	1.014

n	margin parameters	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
100	IFM	2.2370	1.5610	1.1717	2.2543	1.5524	2.2616
		(0.2871)	(0.2732)	(0.2737)	(0.3444)	(0.3290)	(0.3438)
	MLE	2.2484	1.5582	1.1819	2.2640	1.5540	2.2616
		(0.2867)	(0.2743)	(0.2832)	(0.3330)	(0.3279)	(0.3406)
	r	1.001	0.996	0.966	1.034	1.003	1.010
1000	IFM	2.2143	1.5263	1.1385	2.1965	1.5095	2.1807
		(0.0957)	(0.0803)	(0.0811)	(0.0980)	(0.0842)	(0.0951)
	MLE	2.2180	1.5268	1.1377	2.1988	1.5102	2.1812
		(0.0941)	(0.0799)	(0.0814)	(0.0953)	(0.0841)	(0.0975)
	r	1.018	1.005	0.996	1.028	1.001	0.975

where $a_{j1} = G_j(y_j - 1)$, $a_{j2} = G_j(y_j)$. C is d-dimensional normal copula. $G_j(\cdot)$ is defined as

$$G_j(y_j) = \begin{cases} 0, & \text{if} y_j < 0, \\ \sum_{s=0}^{[y_j]} p_j^{(s)}, & \text{if} 0 \le y_j, \end{cases}$$

where $p_j^{(s)} = [\lambda_j^s \exp(-\lambda_j)]/s!$, $s = 0, 1, 2, \ldots, \infty$. In general, we assume $\lambda_{ij} = \exp(\boldsymbol{\beta}_j' \mathbf{x}_{ij})$, and $\Theta_i = (\theta_{i,jk})$ to be free of covariates. We further transform the dependence parameters θ_{jk} with $\theta_{jk} = (\exp(\alpha_{jk}) - 1)/(\exp(\alpha_{jk}) + 1)$, and estimate α_{jk} instead of θ_{jk}. We use the following simulation scheme:

1. The sample size is n, the number of simulations is N; both are reported in the tables.

2. For $d = 3$, we study the two situations: $\log(\lambda_{ij}) = \beta_j$ and $\log(\lambda_{ij}) = \beta_{j0} + \beta_{j1} x_{ij}$. For each situation, we chose two dependence structures: $\theta_{12} = \theta_{13} = \theta_{23} = 0.6$ (or $\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$) and $\theta_{12} = \theta_{23} = 0.8$ (or $\alpha_{12} = \alpha_{23} = 2.1972$), $\theta_{13} = 0.64$ (or $\alpha_{13} = 1.5163$). Other parameters are

   (a) $\boldsymbol{\beta}_0 = (\beta_{10}, \beta_{20}, \beta_{30})' = (1, 1, 1)'$ and $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{21}, \beta_{31})' = (0.5, 0.5, 0.5)'$. Situations where x_{ij} is discrete is considered. For the discrete situation, $x_{ij} = I(U \le 0)$ where

Table 4.12: Efficiency assessment with MCD model for count data: $d = 3$, $\boldsymbol{\beta}_0 = (1,1,1)'$ and $\boldsymbol{\beta}_1 = (0.5,0.5,0.5)'$, x_{ij} discrete, $N = 1000$

n	margin parameters	1 β_{10} β_{11}	2 β_{20} β_{21}	3 β_{30} β_{31}	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
		$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$					
100	IFM	1.0054 0.490	1.0017 0.493	1.0029 0.493	1.423	1.412	1.420
		(0.0867) (0.109)	(0.0883) (0.110)	(0.0870) (0.106)	(0.198)	(0.191)	(0.188)
	MLE	1.0018 0.488	0.9985 0.491	0.9998 0.491	1.422	1.409	1.417
		(0.0876) (0.111)	(0.0892) (0.112)	(0.0872) (0.107)	(0.194)	(0.183)	(0.187)
	r	0.990 0.985	0.990 0.979	0.997 0.993	1.019	1.041	1.009
1000	IFM	1.0007 0.4991	1.0013 0.4975	1.0014 0.4984	1.3953	1.3864	1.3896
		(0.0287) (0.0358)	(0.0271) (0.0343)	(0.0278) (0.0352)	(0.0625)	(0.0564)	(0.0595)
	MLE	0.9976 0.4993	0.9984 0.4977	0.9983 0.4987	1.3918	1.3849	1.3875
		(0.0292) (0.0362)	(0.0274) (0.0343)	(0.0283) (0.0354)	(0.0578)	(0.0563)	(0.0574)
	r	0.983 0.988	0.989 0.998	0.983 0.995	1.081	1.003	1.035
		$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$					
100	IFM	1.0054 0.490	1.0044 0.490	1.0046 0.491	2.242	1.551	2.236
		(0.0867) (0.109)	(0.0884) (0.110)	(0.0870) (0.107)	(0.190)	(0.196)	(0.186)
	MLE	1.0006 0.489	0.9993 0.488	0.9998 0.489	2.239	1.545	2.233
		(0.0878) (0.111)	(0.0895) (0.112)	(0.0881) (0.109)	(0.187)	(0.180)	(0.185)
	r	0.987 0.986	0.987 0.981	0.987 0.980	1.015	1.087	1.007
1000	IFM	1.0007 0.4991	1.0013 0.4979	1.0014 0.4982	2.2055	1.5185	2.1975
		(0.0287) (0.0358)	(0.0272) (0.0350)	(0.0277) (0.0351)	(0.0555)	(0.0620)	(0.0572)
	MLE	0.9962 0.4992	0.9963 0.4981	0.9971 0.4984	2.2037	1.5156	2.1952
		(0.0291) (0.0364)	(0.0279) (0.0355)	(0.0279) (0.0354)	(0.0553)	(0.0553)	(0.0569)
	r	0.985 0.983	0.977 0.986	0.993 0.990	1.003	1.121	1.006

$U \sim$ uniform$(-1,1)$.

(b) $(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}) = (5,3,5)$ (or $(\beta_1, \beta_2, \beta_3) = (1.6094, 1.0986, 1.6094)$).

3. For $d = 4$, we only study $\log(\lambda_{ij}) = \beta_j$. Two dependence structures are considered: $\theta_{12} = \theta_{13} = \theta_{14} = \theta_{23} = \theta_{24} = \theta_{34} = 0.6$ (or $\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$) and $\theta_{12} = \theta_{23} = \theta_{34} = 0.8$ (or $\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$), $\theta_{13} = \theta_{24} = 0.64$ (or $\alpha_{13} = \alpha_{24} = 1.5163$) and $\theta_{14} = 0.512$ (or $\alpha_{14} = 1.1309$). Other parameters are

   (a) $(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}, \lambda_{i4}) = (5,5,5,5)$ (or equivalently $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.6094, 1.6094, 1.6094, 1.6094)$).

   (b) $(\lambda_{i1}, \lambda_{i2}, \lambda_{i3}, \lambda_{i4}) = (4,2,5,8)$ (or equivalently $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.3863, 0.6931, 1.6094, 2.0794)$).

The numerical results from MCD models for count data are presented in Table 4.12 to Table 4.15. We obtain the similar conclusions to those for the MCD models for binary data and ordinal data.

Table 4.13: Efficiency assessment with MCD model for count data: $d = 3$, $(\beta_1, \beta_2, \beta_3) = (1.6094, 1.0986, 1.6094)$, $N = 1000$

n	margin parameters	1 β_1	2 β_2	3 β_3	(1,2) α_{12}	(1,3) α_{13}	(2,3) α_{23}
		$\alpha_{12} = \alpha_{13} = \alpha_{23} = 1.3863$					
100	IFM	1.6075	1.1000	1.6076	1.415	1.403	1.408
		(0.0465)	(0.0593)	(0.0456)	(0.194)	(0.190)	(0.195)
	MLE	1.6024	1.0943	1.6032	1.413	1.398	1.403
		(0.0519)	(0.0648)	(0.0490)	(0.191)	(0.189)	(0.192)
	r	0.896	0.915	0.929	1.018	1.003	1.018
1000	IFM	1.6098	1.0988	1.6095	1.3885	1.3885	1.3880
		(0.0141)	(0.0185)	(0.0140)	(0.0597)	(0.0575)	(0.0586)
	MLE	1.6077	1.0963	1.6076	1.3877	1.3855	1.3869
		(0.0146)	(0.0191)	(0.0143)	(0.0588)	(0.0577)	(0.0574)
	r	0.966	0.967	0.975	1.015	0.996	1.021
		$\alpha_{12} = \alpha_{23} = 2.1972$, $\alpha_{13} = 1.5163$					
100	IFM	1.6075	1.0991	1.6089	2.234	1.539	2.219
		(0.0465)	(0.0599)	(0.0455)	(0.187)	(0.187)	(0.188)
	MLE	1.6017	1.0912	1.6032	2.231	1.533	2.217
		(0.0509)	(0.0667)	(0.0490)	(0.187)	(0.181)	(0.188)
	r	0.913	0.897	0.929	0.999	1.033	0.995
1000	IFM	1.6098	1.0992	1.6097	2.2027	1.5176	2.2002
		(0.0141)	(0.0182)	(0.0140)	(0.0565)	(0.0579)	(0.0547)
	MLE	1.6063	1.0944	1.6063	2.1992	1.5149	2.1968
		(0.0155)	(0.0197)	(0.0152)	(0.0563)	(0.0567)	(0.0551)
	r	0.915	0.924	0.923	1.003	1.021	0.993

Table 4.14: Efficiency assessment with MCD model for count data: $d = 4$, $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.6094, 1.0986, 1.6094, 1.6094)$, $N = 1000$

n	margin parameters	1 β_1	2 β_2	3 β_3	4 β_4	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
		$\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$									
100	IFM	1.6072	1.6099	1.6076	1.6085	1.417	1.410	1.413	1.402	1.407	1.398
		(0.0434)	(0.0443)	(0.0459)	(0.0451)	(0.179)	(0.188)	(0.190)	(0.185)	(0.183)	(0.186)
	MLE	1.6016	1.6044	1.6023	1.6031	1.415	1.410	1.413	1.401	1.404	1.401
		(0.0477)	(0.0494)	(0.0495)	(0.0490)	(0.179)	(0.189)	(0.193)	(0.184)	(0.183)	(0.184)
	r	0.910	0.895	0.927	0.920	1.001	0.991	0.986	1.009	1.003	1.008
		$\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$, $\alpha_{13} = \alpha_{24} = 1.5163$, $\alpha_{14} = 1.1309$									
100	IFM	1.6072	1.6090	1.6084	1.6084	2.228	1.546	1.159	2.218	1.536	2.215
		(0.0434)	(0.0445)	(0.0454)	(0.0456)	(0.176)	(0.184)	(0.195)	(0.183)	(0.179)	(0.177)
	MLE	1.5996	1.6003	1.5993	1.5999	2.228	1.538	1.153	2.215	1.527	2.217
		(0.0499)	(0.0513)	(0.0544)	(0.0525)	(0.176)	(0.187)	(0.193)	(0.184)	(0.177)	(0.177)
	r	0.871	0.867	0.835	0.867	0.997	0.982	1.015	0.995	1.009	1.001

Table 4.15: Efficiency assessment with MCD model for count data: $d = 4$, $(\beta_1, \beta_2, \beta_3, \beta_4) = (1.3863, 0.6931, 1.6094, 2.0794)$, $N = 1000$

n	margin parameters	1 β_1	2 β_2	3 β_3	4 β_4	(1,2) α_{12}	(1,3) α_{13}	(1,4) α_{14}	(2,3) α_{23}	(2,4) α_{24}	(3,4) α_{34}
$\alpha_{12} = \alpha_{13} = \alpha_{14} = \alpha_{23} = \alpha_{24} = \alpha_{34} = 1.3863$											
100	IFM	1.3835	0.6918	1.6076	2.0787	1.418	1.411	1.414	1.405	1.406	1.398
		(0.0489)	(0.0705)	(0.0459)	(0.0356)	(0.185)	(0.189)	(0.185)	(0.194)	(0.189)	(0.186)
	MLE	1.3772	0.6840	1.6024	2.0744	1.415	1.411	1.411	1.405	1.407	1.402
		(0.0549)	(0.0756)	(0.0496)	(0.0384)	(0.185)	(0.188)	(0.185)	(0.193)	(0.187)	(0.183)
	r	0.891	0.932	0.927	0.927	1.003	1.005	0.998	1.002	1.011	1.017
$\alpha_{12} = \alpha_{23} = \alpha_{34} = 2.1972$, $\alpha_{13} = \alpha_{24} = 1.5163$, $\alpha_{14} = 1.1309$											
100	IFM	1.3835	0.6906	1.6084	2.0790	2.239	1.546	1.155	2.219	1.535	2.216
		(0.0489)	(0.0693)	(0.0454)	(0.0361)	(0.191)	(0.184)	(0.194)	(0.198)	(0.191)	(0.173)
	MLE	1.3758	0.6792	1.6006	2.0735	2.236	1.537	1.152	2.221	1.529	2.216
		(0.0550)	(0.0764)	(0.0524)	(0.0412)	(0.191)	(0.184)	(0.188)	(0.202)	(0.186)	(0.174)
	r	0.890	0.907	0.866	0.876	1.001	0.997	1.032	0.981	1.023	0.991

Multivariate mixture discrete models for count data

We now consider a MMD model for count data with the Morgenstern copula

$$P(y_1 \cdots y_d) = \int \cdots \int \prod_{j=1}^{d} f(y_j; \lambda_j) \prod_{j=1}^{d} g(\lambda_j) \left[1 + \sum_{j<k}^{d} \theta_{jk} (1 - 2G(\lambda_j))(1 - 2G(\lambda_k)) \right] d\lambda_1 \cdots d\lambda_d, \tag{4.11}$$

where $f(y_j; \lambda_j) = e^{-\lambda_j} \lambda_j^{y_j} / y_j!$ is the Poisson frequency function with parameter λ_j ($\lambda_j > 0$), $g(\lambda)$ a Gamma density function, having the form $g(\lambda_j) = \{1/[\beta_j^{\alpha_j} \Gamma(\alpha_j)]\} \lambda_j^{\alpha_j - 1} e^{-\lambda_j/\beta_j}$, $\lambda_j \geq 0$, with β_j being a scale parameter, and $G(\lambda_j)$ is a Gamma cdf. We have $E(\lambda_j) = \alpha_j \beta_j$ and $Var(\lambda_j) = \alpha_j \beta_j^2$. (4.11) is a *multivariate Poisson-Morgenstern-gamma model.* The multiple integral in (4.11) over the joint space of $\lambda_1, \ldots, \lambda_d$ can be decomposed into a product of integrals of a single variable. The calculation of $P(y_1 \cdots y_d)$ can thus be accomplished by calculating $2d$ univariate integrals. In fact, we have

$$P(y_1 \cdots y_d) = \prod_{j=1}^{d} P(y_j) + \sum_{j<k}^{d} \theta_{jk} \left\{ \prod_{\substack{m=1 \\ m \neq j,k}}^{d} P(y_m)[P(y_j) - 2P_w(y_j)][P(y_k) - 2P_w(y_k)] \right\},$$

and

$$P_{jk}(y_j y_k) = P(y_j)P(y_k) + \theta_{jk}[P(y_j) - 2P_w(y_j)][P(y_k) - 2P_w(y_k)],$$

where $P(y_j) = \int f(y_j; \lambda_j) g(\lambda_j)\, d\lambda_j$ and $P_w(y_j) = \int f(y_j; \lambda_j) g(\lambda_j) G(\lambda_j)\, d\lambda_j$. Now

$$P(y_j) = \int \frac{e^{-\lambda_j} \lambda_j^{y_j}}{y_j!} \frac{1}{\beta_j^{\alpha_j} \Gamma(\alpha_j)} \lambda_j^{\alpha_j - 1} e^{-\lambda_j/\beta_j}\, d\lambda_j = \frac{\beta_j^{y_j} \Gamma(\alpha_j + y_j)}{(\beta_j + 1)^{y_j + \alpha_j} \Gamma(\alpha_j) y_j!}. \tag{4.12}$$

Further, as $f(y_j;\lambda_j)g(\lambda_j)$ is proportional to the density of a Gamma$(y+\alpha, \beta/(\beta+1))$ random variable, if we let $\mu_j = \beta_j(y_j+\alpha_j)/(\beta_j+1)$, $\sigma_j^2 = (y_j+\alpha_j)\beta_j^2/(\beta_j+1)^2$, then the upper and lower integration ranges of $P_w(y_j)$ can be set up as $L_j = \mu_j - 5\sigma_j$ and $U_j = \mu_j + 5\sigma_j$ for numerical evaluation of the integrals.

To carry out the efficiency assessment through simulation, we need to simulate the multivariate Poisson-Morgenstern-gamma distribution. Let C be the Morgenstern copula, and $G(x)$ be the cdf of a univariate Gamma distribution. The following simulation algorithm is used:

1. Generate $U_1, \ldots, U_d$ from $C(U_1, \ldots, U_d)$.
2. Let $\lambda_j = G^{-1}(U_j),\ j = 1, \ldots, d$.
3. Generate Y_j from Poisson(λ_j), $j = 1, \ldots, d$.

In the above algorithm, the difficult part is the generation of $U_1, \ldots, U_d$ from $C(U_1, \ldots, U_d)$. The conditional distribution approach to generate multivariate random variates can be used here. The conditional distribution approach is to obtain $\mathbf{x} = (x_1, \ldots, x_d)'$ with $F(x_1) = V_1$, $F(x_2|x_1) = V_2$, $\ldots$, $F(x_d|x_1, \ldots, x_{d-1}) = V_d$, where $V_1, \ldots, V_d$ are independent uniform$(0,1)$. With the Morgenstern copula, for $m \le d$,

$$C(u_m|U_1 = u_1, \ldots, U_{m-1} = u_{m-1}) = \int_0^{u_m} \frac{f(u_1, \ldots, u_{m-1}, u)}{f(u_1, \ldots, u_{m-1})}\, du,$$

where $f(u_1, \ldots, u_m) = 1 + \sum_{j<k}^{m} \theta_{jk}(1-2u_j)(1-2u_k)$. Since $f(u_1, \ldots, u_m) = f(u_1, \ldots, u_{m-1}) + \sum_{j=1}^{m-1} \theta_{jm}(1-2u_j)(1-2u_m)$, it follows that

$$\frac{f(u_1, \ldots, u_m)}{f(u_1, \ldots, u_{m-1})} = 1 + \frac{\sum_{j=1}^{m-1} \theta_{jm}(1-2u_j)}{f(u_1, \ldots, u_{m-1})} - \frac{2(\sum_{j=1}^{m-1} \theta_{jm}(1-2u_j))}{f(u_1, \ldots, u_{m-1})} u_m.$$

Hence

$$C(u_m|U_1 = u_1, \ldots, U_{m-1} = u_{m-1}) = \left(1 + \frac{\sum_{j=1}^{m-1} \theta_{jm}(1-2u_j)}{f(u_1, \ldots, u_{m-1})}\right) u_m - \frac{\sum_{j=1}^{m-1} \theta_{jm}(1-2u_j)}{f(u_1, \ldots, u_{m-1})} u_m^2.$$

Let $A = f(u_1, \ldots, u_{m-1})$, $B = \sum_{j=1}^{m-1} \theta_{jm}(1-2u_j)$, and $D = B/A$. From $Du_m^2 - (D+1)u_m + V_m = 0$, we get

$$u_m = \frac{(D+1) \pm \sqrt{(D+1)^2 - 4DV_m}}{2D}.$$

Thus the algorithm for generating $U_1, \ldots, U_d$ from $C(u_1, \ldots, u_d)$ is as the following:

1. Generate $V_1, \ldots, V_d$ from Uniform$(0,1)$.

2. Let $U_1 = V_1$.

3. Let $A = 1$ if $m = 2$ and $A = 1 + \sum_{j<k}^{m-1} \theta_{jk}(1-2u_j)(1-2u_k)$ if $m > 2$. Let $B = \sum_{j=1}^{m-1} \theta_{jm}(1-2u_j)$, and $D = B/A$.

4. For $m \geq 2$, if $D = 0$, $U_m = V_m$. If $D \neq 0$, U_m takes one of the values of $[(D+1) \pm \sqrt{(D+1)^2 - 4DV_m}]/[2D]$ for which it is positive and less than 1.

The efficiency studies with the multivariate Poisson-Morgenstern-gamma model are carried out only for the dependence parameters θ_{jk}, in that univariate parameters are fixed. We use the following simulation scheme:

1. The sample size is $n = 3000$, the number of simulations is $N = 200$.

2. The dimension d is chosen to be 3, 4 and 5.

3. The marginal parameters α_j and β_j are fixed. They are $\alpha_j = \beta_j = 1$ for $j = 1, \ldots, d$.

4. For each dimension, two dependence structures are considered:

   (a) For $d = 3$, we have $(\theta_{12}, \theta_{13}, \theta_{23}) = (0.5, 0.5, 0.5)$ and $(\theta_{12}, \theta_{13}, \theta_{23}) = (0.6, 0.7, 0.8)$.

   (b) For $d = 4$, we have $(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34}) = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$ and $(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34}) = (0.6, 0.7, 0.8, 0.6, 0.7, 0.6)$.

   (c) For $d = 5$, we have
   $(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{23}, \theta_{24}, \theta_{25}, \theta_{34}, \theta_{35}, \theta_{45}) = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$ and
   $(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{23}, \theta_{24}, \theta_{25}, \theta_{34}, \theta_{35}, \theta_{45}) = (0.6, 0.7, 0.8, 0.8, 0.6, 0.7, 0.8, 0.6, 0.7, 0.8)$.

The numerical results from the MMD models for count data with the Morgenstern copula are presented in Table 4.16 to Table 4.18. We obtain similar conclusions to those for the MCD models for binary, ordinal and count data. Basically, they are: i) The IFM approach is efficient relative to the ML approach; the ratio values r are very close to 1 in almost all the situations studied. ii) MLE may be slightly more efficient than IFME, but this observation is not conclusive. IFME and MLE are comparable.

Table 4.16: Efficiency assessment with multivariate Poisson-Morgenstern-gamma model, $d = 3$

parameters	θ_{12}	θ_{13}	θ_{23}
$(\theta_{12}, \theta_{13}, \theta_{23}) = (0.5, 0.5, 0.5)$			
IFM	0.495	0.500	0.501
	(0.125)	(0.125)	(0.124)
MLE	0.494	0.499	0.500
	(0.122)	(0.124)	(0.123)
r	1.022	1.008	1.003
$(\theta_{12}, \theta_{13}, \theta_{23}) = (0.6, 0.7, 0.8)$			
IFM	0.603	0.699	0.792
	(0.127)	(0.118)	(0.119)
MLE	0.600	0.697	0.790
	(0.127)	(0.120)	(0.119)
r	1.000	0.985	0.995

Table 4.17: Efficiency assessment with multivariate Poisson-Morgenstern-gamma model, $d = 4$

parameters	θ_{12}	θ_{13}	θ_{14}	θ_{23}	θ_{24}	θ_{34}
$(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34}) = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$						
IFM	0.500	0.495	0.513	0.498	0.494	0.488
	(0.131)	(0.128)	(0.124)	(0.133)	(0.134)	(0.138)
MLE	0.501	0.493	0.512	0.497	0.495	0.485
	(0.130)	(0.124)	(0.122)	(0.131)	(0.132)	(0.135)
r	1.008	1.026	1.014	1.021	1.018	1.019
$(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{23}, \theta_{24}, \theta_{34}) = (0.6, 0.7, 0.8, 0.6, 0.7, 0.6)$						
IFM	0.593	0.680	0.794	0.589	0.692	0.599
	(0.130)	(0.127)	(0.120)	(0.121)	(0.133)	(0.124)
MLE	0.589	0.678	0.792	0.585	0.689	0.598
	(0.127)	(0.124)	(0.117)	(0.118)	(0.129)	(0.124)
r	1.017	1.026	1.024	1.025	1.037	1.006

Table 4.18: Efficiency assessment with multivariate Poisson-Morgenstern-gamma model, $d = 5$

parameters	θ_{12}	θ_{13}	θ_{14}	θ_{15}	θ_{23}	θ_{24}	θ_{25}	θ_{34}	θ_{35}	θ_{45}
$(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{23}, \theta_{24}, \theta_{25}, \theta_{34}, \theta_{35}, \theta_{45}) = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$										
IFM	0.501	0.496	0.477	0.486	0.511	0.467	0.478	0.504	0.493	0.495
	(0.122)	(0.131)	(0.137)	(0.132)	(0.123)	(0.130)	(0.123)	(0.128)	(0.131)	(0.116)
MLE	0.495	0.493	0.473	0.482	0.508	0.466	0.475	0.503	0.489	0.494
	(0.121)	(0.125)	(0.134)	(0.128)	(0.122)	(0.125)	(0.119)	(0.127)	(0.128)	(0.113)
r	1.012	1.046	1.023	1.026	1.002	1.043	1.037	1.011	1.026	1.018
$(\theta_{12}, \theta_{13}, \theta_{14}, \theta_{15}, \theta_{23}, \theta_{24}, \theta_{25}, \theta_{34}, \theta_{35}, \theta_{45}) = (0.6, 0.7, 0.8, 0.8, 0.6, 0.7, 0.8, 0.6, 0.7, 0.8)$										
IFM	0.595	0.667	0.775	0.767	0.597	0.693	0.778	0.590	0.693	0.602
	(0.140)	(0.137)	(0.132)	(0.130)	(0.127)	(0.139)	(0.118)	(0.136)	(0.126)	(0.125)
MLE	0.590	0.666	0.772	0.766	0.593	0.690	0.778	0.588	0.687	0.604
	(0.137)	(0.132)	(0.128)	(0.126)	(0.119)	(0.135)	(0.115)	(0.132)	(0.124)	(0.113)
r	1.023	1.036	1.029	1.032	1.067	1.029	1.028	1.028	1.018	1.103

4.4 IFM efficiency for models with special dependence structure

The IFM approach may have important applications for models with special dependence structure. Data with special dependence structure arise often in practice: longitudinal studies, repeated measures, Markov type dependence data, k-dependent data, and so on.

The analytical assessment of the efficiency of the IFM approach for several models with special dependence structure were studied in section 4.2. In the following, we give some numerical results for IFM efficiency for some more complex models with special dependence structure. The estimation approach that we used here is PMLA. We only present representative results from the MCD model for binary data, with the MVN copula of exchangeable and AR(1) dependence structures. Results with other models are quite similar, as we also observed in section 4.3 for various situations with a general model. We use the following simulation scheme:

1. The sample size is $n = 1000$, the number of simulations is $N = 200$.

2. The dimension d are chosen to be 3 and 4.

3. For $d = 3$, we considered two marginal models $Y_{ij} = I(Z_{ij} < z_j)$ and $Y_{ij} = I(Z_{ij} < \alpha_{j0} + \alpha_{j1}x_{ij})$, with $x_{ij} = I(U \leq 0)$ where $U \sim$ uniform$(-1, 1)$, and with the regression parameters

   (a) with no covariates: $\mathbf{z} = (0.5, 0.5, 0.5)'$ and $\mathbf{z} = (0.5, 1.0, 1.5)'$;

   (b) with covariates: $\boldsymbol{\alpha}_0 = (\alpha_{10}, \alpha_{20}, \alpha_{30})' = (0.5, 0.5, 0.5)'$, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{21}, \alpha_{31})' = (1, 1, 1)'$ and $\boldsymbol{\alpha}_0 = (\alpha_{10}, \alpha_{20}, \alpha_{30})' = (0.5, 0.5, 0.5)'$, $\boldsymbol{\alpha}_1 = (\alpha_{11}, \alpha_{21}, \alpha_{31})' = (1, 0.5, 1.5)'$.

   For each marginal model, exchangeable and AR(1) dependence structures in the MVN copula are considered, with the single dependence parameter in both cases being $\theta_i = [\exp(\beta_0 + \beta_1 w_i) - 1]/[\exp(\beta_0 + \beta_1 w_i) + 1]$, with $w_i = I(U \leq 0)$ where $U \sim$ uniform$(-1, 1)$, and parameters $\beta_0 = 1$ and $\beta_1 = 1.5$.

4. For $d = 4$, we only study $Y_{ij} = I(Z_{ij} < z_j)$, with the marginal parameters $\mathbf{z} = (0.5, 0.5, 0.5, 0.5)'$, and $\mathbf{z} = (0.5, 0.8, 1.2, 1.5)'$. For each marginal model, exchangeable and AR(1) dependence structures in MVN copula are considered. The single dependence parameter in both cases is $\theta_i = [\exp(\beta_0) - 1]/[\exp(\beta_0) + 1]$, with $\beta_0 = 1.386$ and $\beta_0 = 2.197$ for both situations.

The numerical results from these models with special dependence structure are presented in Table 4.19 to Table 4.26. We basically have the same conclusions as with all other general cases

Table 4.19: Efficiency assessment with special dependence structure: $d = 3$, $\mathbf{z} = (0.5, 0.5, 0.5)'$

parameters	z_1	z_2	z_3	β_0	β_1
exchangeable, $\beta_0 = 1$, $\beta_1 = 1.5$					
IFM	0.496	0.497	0.497	0.996	1.511
	(0.043)	(0.041)	(0.042)	(0.118)	(0.194)
MLE	0.494	0.496	0.496	0.996	1.520
	(0.041)	(0.040)	(0.041)	(0.118)	(0.195)
r	1.047	1.021	1.015	1.003	0.998
AR(1), $\beta_0 = 1$, $\beta_1 = 1.5$					
IFM	0.496	0.497	0.496	0.992	1.506
	(0.043)	(0.041)	(0.041)	(0.123)	(0.185)
MLE	0.494	0.497	0.495	0.994	1.512
	(0.042)	(0.040)	(0.041)	(0.119)	(0.183)
r	1.034	1.027	1.012	1.031	1.015

Table 4.20: Efficiency assessment with special dependence structure: $d = 3$, $\mathbf{z} = (0.5, 1.0, 1.5)'$

parameters	z_1	z_2	z_3	β_0	β_1
exchangeable, $\beta_0 = 1$, $\beta_1 = 1.5$					
IFM	0.496	0.997	1.499	0.998	1.531
	(0.043)	(0.047)	(0.064)	(0.154)	(0.249)
MLE	0.496	0.996	1.499	0.997	1.534
	(0.043)	(0.047)	(0.063)	(0.156)	(0.247)
r	1.009	0.999	1.010	0.986	1.008
AR(1), $\beta_0 = 1$, $\beta_1 = 1.5$					
IFM	0.496	0.997	1.500	0.994	1.509
	(0.043)	(0.047)	(0.063)	(0.158)	(0.250)
MLE	0.496	0.996	1.500	0.993	1.518
	(0.043)	(0.046)	(0.062)	(0.156)	(0.249)
r	1.017	1.013	1.018	1.011	1.003

studied previously. These conclusions are: i) The IFM approach (PMLA) is efficient relative to the ML approach; the ratio values r are very close to 1 in almost all the studied situations. ii) MLE may be slightly more efficient than IFME, but this observation is not conclusive. IFME and MLE are comparable.

4.5 Jackknife variance estimate compared with Godambe information matrix

Now we turn to numerical evaluation of the performance of jackknife variance estimates of IFME. We have shown, in Chapter 2, that the jackknife estimate of variance is asymptotically equivalent to the estimate of variance from the corresponding Godambe information matrix. The jackknife approach may be preferred when the appropriate computer packages are not available to compute

Table 4.21: Efficiency assessment with special dependence structure: $d = 3$, $\boldsymbol{\alpha}_0 = (0.5, 0.5, 0.5)'$, $\boldsymbol{\alpha}_1 = (1, 1, 1)'$

parameters	α_{10} α_{11}	α_{20} α_{21}	α_{30} α_{31}	β_0	β_1
	exchangeable, $\beta_0 = 1$, $\beta_1 = 1.5$				
IFM	0.500 1.020	0.499 1.010	0.500 1.002	0.980	1.536
	(0.055) (0.109)	(0.060) (0.108)	(0.059) (0.104)	(0.153)	(0.242)
MLE	0.500 1.018	0.498 1.010	0.500 0.999	0.978	1.556
	(0.052) (0.104)	(0.059) (0.107)	(0.058) (0.102)	(0.152)	(0.250)
r	1.052 1.048	1.011 1.007	1.018 1.028	1.002	0.968
	AR(1), $\beta_0 = 1$, $\beta_1 = 1.5$				
IFM	0.500 1.020	0.499 1.010	0.497 1.002	0.988	1.529
	(0.055) (0.109)	(0.060) (0.108)	(0.058) (0.101)	(0.158)	(0.233)
MLE	0.501 1.017	0.499 1.009	0.497 0.999	0.985	1.545
	(0.052) (0.104)	(0.059) (0.105)	(0.058) (0.100)	(0.157)	(0.235)
r	1.043 1.047	1.023 1.022	1.004 1.004	1.008	0.991

Table 4.22: Efficiency assessment with special dependence structure: $d = 3$, $\boldsymbol{\alpha}_0 = (0.5, 0.5, 0.5)'$, $\boldsymbol{\alpha}_1 = (1, 0.5, 1.5)'$

parameters	α_{10} α_{11}	α_{20} α_{21}	α_{30} α_{31}	β_0	β_1
	exchangeable, $\beta_0 = 1$, $\beta_1 = 1.5$				
IFM	0.500 1.020	0.499 0.510	0.500 1.512	0.985	1.528
	(0.055) (0.109)	(0.060) (0.089)	(0.059) (0.141)	(0.160)	(0.238)
MLE	0.500 1.017	0.498 0.510	0.500 1.506	0.983	1.539
	(0.052) (0.103)	(0.059) (0.089)	(0.058) (0.132)	(0.159)	(0.239)
r	1.047 1.050	1.011 1.002	1.017 1.070	1.004	0.996
	AR(1), $\beta_0 = 1$, $\beta_1 = 1.5$				
IFM	0.500 1.020	0.499 0.510	0.497 1.514	0.996	1.518
	(0.055) (0.109)	(0.060) (0.089)	(0.058) (0.140)	(0.159)	(0.225)
MLE	0.500 1.017	0.499 0.510	0.497 1.510	0.994	1.530
	(0.053) (0.104)	(0.059) (0.089)	(0.057) (0.133)	(0.158)	(0.223)
r	1.041 1.045	1.021 1.003	1.006 1.049	1.007	1.010

Table 4.23: Efficiency assessment with special dependence structure: $d = 4$, $\mathbf{z} = (0.5, 0.5, 0.5, 0.5)'$

parameters	z_1	z_2	z_3	z_4	β_0
	exchangeable, $\beta_0 = 1.386$				
IFM	0.502	0.499	0.501	0.501	1.387
	(0.041)	(0.043)	(0.042)	(0.042)	(0.071)
MLE	0.501	0.499	0.500	0.500	1.389
	(0.041)	(0.043)	(0.042)	(0.042)	(0.070)
r	1.000	1.002	1.005	1.003	1.013
	AR(1), $\beta_0 = 1.386$				
IFM	0.502	0.499	0.501	0.497	1.385
	(0.041)	(0.043)	(0.041)	(0.042)	(0.072)
MLE	0.502	0.499	0.500	0.496	1.387
	(0.041)	(0.043)	(0.041)	(0.042)	(0.069)
r	0.996	1.000	0.998	0.998	1.047

Table 4.24: Efficiency assessment with special dependence structure: $d = 4$, $\mathbf{z} = (0.5, 0.8, 1.2, 1.5)'$

parameters	z_1	z_2	z_3	z_4	β_0
exchangeable, $\beta_0 = 1.386$					
IFM	0.502	0.803	1.199	1.494	1.389
	(0.041)	(0.045)	(0.052)	(0.061)	(0.087)
MLE	0.502	0.802	1.198	1.492	1.391
	(0.041)	(0.045)	(0.052)	(0.061)	(0.087)
r	0.998	1.004	1.002	1.007	1.004
AR(1), $\beta_0 = 1.386$					
IFM	0.502	0.803	1.20	1.495	1.388
	(0.041)	(0.045)	(0.05)	(0.067)	(0.085)
MLE	0.502	0.802	1.20	1.494	1.389
	(0.041)	(0.045)	(0.05)	(0.065)	(0.083)
r	0.999	1.006	1.00	1.017	1.025

Table 4.25: Efficiency assessment with special dependence structure: $d = 4$, $\mathbf{z} = (0.5, 0.5, 0.5, 0.5)'$

parameters	z_1	z_2	z_3	z_4	β_0
exchangeable, $\beta_0 = 2.197$					
IFM	0.502	0.501	0.501	0.501	2.200
	(0.041)	(0.042)	(0.042)	(0.042)	(0.093)
MLE	0.500	0.499	0.499	0.499	2.202
	(0.041)	(0.042)	(0.042)	(0.042)	(0.092)
r	0.999	1.000	0.999	1.000	1.015
AR(1), $\beta_0 = 2.197$					
IFM	0.502	0.501	0.501	0.499	2.194
	(0.041)	(0.042)	(0.042)	(0.042)	(0.086)
MLE	0.501	0.499	0.499	0.498	2.199
	(0.041)	(0.043)	(0.042)	(0.042)	(0.084)
r	0.995	0.993	1.000	0.999	1.025

Table 4.26: Efficiency assessment with special dependence structure: $d = 4$, $\mathbf{z} = (0.5, 0.8, 1.2, 1.5)'$

parameters	z_1	z_2	z_3	z_4	β_0
exchangeable, $\beta_0 = 2.197$					
IFM	0.502	0.802	1.201	1.499	2.203
	(0.041)	(0.046)	(0.056)	(0.060)	(0.114)
MLE	0.501	0.801	1.199	1.496	2.204
	(0.041)	(0.046)	(0.055)	(0.059)	(0.111)
r	0.996	1.002	1.005	1.003	1.031
AR(1), $\beta_0 = 2.197$					
IFM	0.502	0.802	1.198	1.500	2.200
	(0.041)	(0.046)	(0.052)	(0.060)	(0.110)
MLE	0.501	0.801	1.196	1.500	2.200
	(0.041)	(0.046)	(0.052)	(0.060)	(0.100)
r	0.997	1.005	0.993	1.000	1.040

the Godambe information matrix or when the asymptotic variance in terms of Godambe information matrix is difficult to compute analytically or computationally. For example, to compute the asymptotic variance of $P(y_1 \cdots y_d; \tilde{\boldsymbol{\theta}})$ by means of Godambe information is not an easy task. To complement the theoretical results in Chapter 2, in this subsection, we give some analytical and numerical comparisons of the variance estimates from Godambe information and the jackknife method. The application of jackknife methods to modelling and inference of real data sets is demonstrated in Chapter 5.

Analytical comparison of the two approaches

Example 4.8 (Multinormal, general) Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, and suppose we are interested in estimating $\boldsymbol{\mu}$. Given n independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ from $\mathbf{X}$, the IFME of $\boldsymbol{\mu}$ is $\tilde{\boldsymbol{\mu}} = n^{-1} \sum_{i=1}^{n} \mathbf{x}_i$, and the corresponding inverse of the Godambe information matrix is $J_{\Psi}^{-1} = \Sigma$. A consistent estimate of J_{Ψ}^{-1} is

$$\tilde{J}_{\Psi}^{-1} = n^{-1} \sum_{i=1}^{n} (x_i - \tilde{\boldsymbol{\mu}})(x_i - \tilde{\boldsymbol{\mu}})^T.$$

The jackknife estimate of the Godambe information matrix is

$$nV_J = n \sum_{i=1}^{n} (\tilde{\boldsymbol{\mu}}_{(i)} - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_{(i)} - \tilde{\boldsymbol{\mu}})^T,$$

where $\tilde{\boldsymbol{\mu}}_{(i)} = (n-1)^{-1}(n\tilde{\boldsymbol{\mu}} - \mathbf{x}_i)$. Some algebraic manipulation leads to

$$nV_J = \frac{n^2}{(n-1)^2} \frac{1}{n} \sum_{i=1}^{n} (x_i - \tilde{\boldsymbol{\mu}})(x_i - \tilde{\boldsymbol{\mu}})^T,$$

which is a consistent estimate of Σ. Furthermore, we see that

$$nV_J = \frac{n^2}{(n-1)^2} \tilde{J}_{\Psi}^{-1},$$

which shows that the jackknife estimate of the Godambe information matrix is also good when the sample size is moderate to small. □

Example 4.9 (Multinormal, common marginal mean) Let $\mathbf{X} \sim N_d(\boldsymbol{\mu}, \Sigma)$, where $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)' = \mu \mathbf{1}$ and Σ is known. We are interested in estimating the common parameter μ. Given n independent observations $\mathbf{x}_1, \ldots, \mathbf{x}_n$ with same distributions as $\mathbf{X}$, the IFME of μ by the weighting approach is (see Example 4.2)

$$\tilde{\mu}_w = \frac{\mathbf{1}'\Sigma^{-1}\tilde{\boldsymbol{\mu}}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}.$$

The inverse of Godambe information of $\tilde{\mu}_w$ is

$$J_{\Psi}^{-1} = \frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}.$$

The jackknife estimate of the Godambe information is

$$nV_J = n\sum_{i=1}^{n}(\tilde{\boldsymbol{\mu}}_{w(i)} - \tilde{\boldsymbol{\mu}}_w)(\tilde{\boldsymbol{\mu}}_{w(i)} - \tilde{\boldsymbol{\mu}}_w)^T,$$

where $\tilde{\boldsymbol{\mu}}_{w(i)} = \mathbf{1}'\Sigma^{-1}\tilde{\boldsymbol{\mu}}_{(i)}/\mathbf{1}'\Sigma^{-1}\mathbf{1}$. Some algebraic manipulation leads to

$$nV_J = \frac{\mathbf{1}'\Sigma^{-1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}\left[n\sum_{i=1}^{n}(\tilde{\boldsymbol{\mu}}_{(i)} - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_{(i)} - \tilde{\boldsymbol{\mu}})^T\right]\frac{\Sigma^{-1}\mathbf{1}}{\mathbf{1}'\Sigma^{-1}\mathbf{1}}.$$

We replace $n\sum_{i=1}^{n}(\tilde{\boldsymbol{\mu}}_{(i)} - \tilde{\boldsymbol{\mu}})(\tilde{\boldsymbol{\mu}}_{(i)} - \tilde{\boldsymbol{\mu}})^T$ with $n^2/(n-1)^2\Sigma$. Thus

$$nV_J \approx \frac{n^2}{(n-1)^2}\frac{1}{\mathbf{1}'\Sigma^{-1}\mathbf{1}},$$

and

$$nV_J \approx \frac{n^2}{(n-1)^2}J_{\Psi}^{-1},$$

which shows that the jackknife estimate of the Godambe information is also good when the sample size is moderate to small. □

Numerical comparison of the two approaches

In this subsection, we numerically compare the variance estimates of IFME from the jackknife method and from the Godambe information. For this purpose, we use a 3-dimensional probit model with normal copula. The comparison studies are carried out only for the dependence parameters θ_{jk}. For the chosen model parameters, we carry out N simulations for each sample size n. For each simulation s $(s = 1, \ldots, N)$ of sample size n, we estimate model parameters $\theta_{12}, \theta_{13}, \theta_{23}$ with the IFM approach. Let us denote these estimates $\tilde{\theta}_{12}^{(s)}, \tilde{\theta}_{13}^{(s)}, \tilde{\theta}_{23}^{(s)}$. We then compute the jackknife estimate of variance (with g groups of size m such that $g \times m = n$) for $\tilde{\theta}_{12}^{(s)}, \tilde{\theta}_{13}^{(s)}, \tilde{\theta}_{23}^{(s)}$. We denote these variance estimates by $v_{12}^{(s)}, v_{13}^{(s)}, v_{23}^{(s)}$. Let the asymptotic variance estimate of $\tilde{\theta}_{12}, \tilde{\theta}_{13}, \tilde{\theta}_{23}$ based on the Godambe information matrix from a sample of size n be v_{12}, v_{13}, v_{23}. We compare the following three variance estimates:

(i). MSE: $\frac{1}{N}\sum_{s=1}^{N}(\tilde{\theta}_{12}^{(s)} - \theta_{12})^2$, $\frac{1}{N}\sum_{s=1}^{N}(\tilde{\theta}_{13}^{(s)} - \theta_{13})^2$, $\frac{1}{N}\sum_{s=1}^{N}(\tilde{\theta}_{23}^{(s)} - \theta_{23})^2$;

(ii). Godambe: v_{12}, v_{13}, v_{23};

(iii). Jackknife: $\frac{1}{N}\sum_{s=1}^{N} v_{12}^{(s)}$, $\frac{1}{N}\sum_{s=1}^{N} v_{13}^{(s)}$, $\frac{1}{N}\sum_{s=1}^{N} v_{23}^{(s)}$.

The MSE in (i) should be considered as the true variance of the parameter estimate assuming unbiasedness. (ii) and (iii) should be compared with each other and also with (i). Table 4.27 and Table 4.28 summarize the numerical computation of the variance estimates of $\tilde{\theta}_{12}, \tilde{\theta}_{13}, \tilde{\theta}_{23}$ based on approaches (i), (ii) and (iii). For the jackknife method, the results for different combinations of (g, m) are reported in the two tables. In total four models with different marginal parameters $\mathbf{z} = (z_1, z_2, z_3)$ and different dependence parameters $\boldsymbol{\theta} = (\theta_{12}, \theta_{13}, \theta_{23})$ are studied. The details about the parameter values are reported in the tables. We have studied two sample sizes: $n = 500$ and $n = 1000$. For both sample sizes, the number of simulations is $N = 500$. From examining the two tables, we see that the three measures are very close to each other. We conclude that the jackknife method is indeed consistent with the Godambe information computation approach. Both approaches yields variance estimates which are comparable to MSE.

In conclusion, we have shown theoretically and demonstrated numerically in several cases that the jackknife method for variance estimation compares very favorably with the Godambe information computation. We are willing to extrapolate to general situations. The jackknife approach is simple and computationally straightforward (computationally, it only requires the code for obtaining the parameter estimates); it also has the advantage of easily handling more complex situations where the Godambe information computation is not possible. One major concern with the jackknife approach is the computational time needed to carry out the whole process. If the computing time problem is due to an extremely large sample size, appropriate grouping of the sample for the sake of applying the jackknife approach may improve the situation. A discussion is given in Section 2.5. Overall, we recommend the general use of the jackknife approach in applications.

4.6 Summary

In this chapter, we demonstrated analytically and numerically that the IFM approach is an efficient parameter estimation procedure for MCD and MMD models with MUBE or PUBE properties. We have chosen a wide variety of cases so that we can extrapolate this conclusion to the general situation. Theoretically, we expect IFM to be quite efficient because it is closely tied to MLE in that each inference function is a likelihood score function of a margin. For comparison purposes, we carried out ML estimates for several multivariate models. Our experience was that finding the MLE is a difficult and very time consuming task for multivariate models, while the IFME is

Table 4.27: Comparison of estimates of standard error, (i) true, (ii) Godambe, (iii) jackknife with g groups; $N = 500$, $n = 1000$

	approach	θ_{12}	θ_{13}	θ_{23}
$\mathbf{z} = (0.0, 0.7, 0.0)'$, $\boldsymbol{\theta} = (-0.5, 0.5, -0.5)$				
	(i)	0.002079	0.001704	0.002085
	(ii)	0.002012	0.001645	0.002012
(g, m)	(iii)			
$(1000, 1)$		0.002030	0.001646	0.002038
$(500, 2)$		0.002028	0.001653	0.002043
$(250, 4)$		0.002025	0.001658	0.002047
$(125, 8)$		0.002058	0.001653	0.002046
$(100, 10)$		0.002046	0.001663	0.002046
$(50, 20)$		0.002089	0.001685	0.002089
$\mathbf{z} = (0.7, 0.0, 0.7)'$, $\boldsymbol{\theta} = (0.5, 0.9, 0.5)$				
	(i)	0.002090	0.000281	0.002200
	(ii)	0.002012	0.000295	0.002012
(g, m)	(iii)			
$(1000, 1)$		0.002026	0.000299	0.002023
$(500, 2)$		0.002027	0.000300	0.002021
$(250, 4)$		0.002036	0.000300	0.002035
$(125, 8)$		0.002056	0.000302	0.002049
$(100, 10)$		0.002063	0.000301	0.002054
$(50, 20)$		0.002088	0.000301	0.002067
$\mathbf{z} = (0.7, 0.7, 0.7)'$, $\boldsymbol{\theta} = (0.9, 0.7, 0.5)$				
	(i)	0.000333	0.001218	0.002319
	(ii)	0.000295	0.001239	0.002187
(g, m)	(iii)			
$(1000, 1)$		0.000302	0.001254	0.002208
$(500, 2)$		0.000303	0.001257	0.002210
$(250, 4)$		0.000302	0.001260	0.002212
$(125, 8)$		0.000303	0.001267	0.002216
$(100, 10)$		0.000305	0.001261	0.002214
$(50, 20)$		0.000310	0.001252	0.002220
$\mathbf{z} = (1.0, 0.5, 0.0)'$, $\boldsymbol{\theta} = (0.8, 0.6, 0.8)$				
	(i)	0.000821	0.002147	0.000766
	(ii)	0.000869	0.002089	0.000666
(g, m)	(iii)			
$(1000, 1)$		0.000873	0.002129	0.000683
$(500, 2)$		0.000874	0.002118	0.000683
$(250, 4)$		0.000877	0.002108	0.000681
$(125, 8)$		0.000884	0.002119	0.000688
$(100, 10)$		0.000887	0.002138	0.000687
$(50, 20)$		0.000899	0.002151	0.000690

Table 4.28: Comparison of estimates of standard error, (i) true, (ii) Godambe, (iii) jackknife with g groups; $N = 500$, $n = 500$

	approach	θ_{12}	θ_{13}	θ_{23}
$\mathbf{z} = (0.0, 0.7, 0.0)'$, $\boldsymbol{\theta} = (-0.5, 0.5, -0.5)$				
	(i)	0.004158	0.003135	0.004262
	(ii)	0.004024	0.003290	0.004024
(g, m)	(iii)			
$(500, 1)$		0.004085	0.003315	0.004104
$(250, 2)$		0.004071	0.003333	0.004122
$(125, 4)$		0.004053	0.003331	0.004119
$(50, 10)$		0.004115	0.003396	0.004176
$\mathbf{z} = (0.7, 0.0, 0.7)'$, $\boldsymbol{\theta} = (0.5, 0.9, 0.5)$				
	(i)	0.003998	0.000602	0.003768
	(ii)	0.004024	0.000591	0.004024
(g, m)	(iii)			
$(500, 1)$		0.004062	0.000604	0.004049
$(250, 2)$		0.004062	0.000601	0.004054
$(125, 4)$		0.004091	0.000607	0.004103
$(50, 10)$		0.004123	0.000617	0.004171
$\mathbf{z} = (0.7, 0.7, 0.7)'$, $\boldsymbol{\theta} = (0.9, 0.7, 0.5)$				
	(i)	0.000632	0.002688	0.004521
	(ii)	0.000591	0.002479	0.004374
(g, m)	(iii)			
$(500, 1)$		0.000607	0.002501	0.004410
$(250, 2)$		0.000611	0.002510	0.004425
$(125, 4)$		0.000616	0.002533	0.004467
$(50, 10)$		0.000622	0.002539	0.004501
$\mathbf{z} = (1.0, 0.5, 0.0)'$, $\boldsymbol{\theta} = (0.8, 0.6, 0.8)$				
	(i)	0.001634	0.003846	0.001413
	(ii)	0.001738	0.004179	0.001332
(g, m)	(iii)			
$(500, 1)$		0.001821	0.004397	0.001365
$(250, 2)$		0.001837	0.004407	0.001368
$(125, 4)$		0.001846	0.004433	0.001360
$(50, 10)$		0.001876	0.004476	0.001388

computationally simple and results in significant saving of computing time. We further demonstrated numerically that the jackknife method yields SEs for the IFME, which are comparable to the SEs obtained from the Godambe information matrix. The jackknife method for variance estimates has significant practical importance as it eliminates the need to calculate the partial derivatives which are required for calculating the Godambe information matrix. The jackknife method can also be used for estimates of functions of parameters (such as probabilities of being in some category or probabilities of exceedances).

The IFM approach together with the jackknife estimation of SE's make many more multivariate models computationally feasible for working with real data. The IFM theory as part of statistical inference theory for multivariate non-normal models is highly recommended because of its good asymptotic properties and its computational feasibility. This approach should have significant practical usefulness. We will demonstrate its application in Chapter 5.

Chapter 5

Modelling, data analysis and examples

Possessing a tool is one thing, but using it effectively is quite another. In this chapter, we explore the possibility of effectively using the tools developed in this thesis for multivariate statistical modelling (including IFM theory, jackknife variance estimation, etc.) and provide data analysis examples. In section 5.1, we first discuss out view of the proper data analysis cycle. This is an important issue since the interpretation of the results and maybe the possible indication of further studies are directly related to the way that the data analysis was carried out. We next discuss several other important issues in multivariate discrete modelling, such as how to make the the choice of models and how to deal with checking the adequacy of models. We also provide some discussion on the testing of dependence structure hypotheses, which is useful for identifying some specific multivariate models. In section 5.2, we carry out several data analysis examples with the models and inference procedure developed in the previous chapters. We show some applications of the models and inference procedures developed in this thesis and point out difficulties related to multivariate nonnormal analysis.

5.1 Some issues on modelling

5.1.1 Data analysis cycle

A proper data analysis cycle usually consists of initial data analysis, statistical modelling, diagnostic model assessment and inferences.

The *initial data analysis* may consist of computing various data summaries and examining various graphical representation of data. The type of summary statistics and graphical representations depend on the basic features of the data set. For example, for binary, ordinal and count data, we can compute the empirical frequencies (and percentages) of response variables as well as covariates, separately and jointly. If some covariates are continuous, then standard summaries such as the mean, median, standard deviation, quartiles, maximum, minimum, as well as graphical displays such as boxplots and histograms could be examined. To have a rough idea of the dependence among the response variables, for binary data, a check of the pairwise log odds ratios of the responses could be helpful. Another convenient empirical pairwise dependence measure for multivariate discrete data, which is particularly useful for ordinal and count data, is a measure called *gamma*. This measure, for ordinal and count data $\mathbf{y}_i = (y_{i1}, y_{i2})$, $i = 1, \ldots, n$, is defined as

$$\gamma = \frac{C - D}{C + D}, \tag{5.1}$$

where $C = \sum_{i=1}^{n} \sum_{i'=1}^{n} I(y_{i1} > y_{i'1}) * I(y_{i2} > y_{i'2})$ and $D = \sum_{i=1}^{n} \sum_{i'=1}^{n} I(y_{i1} < y_{i'1}) * I(y_{i2} > y_{i'2})$, and I is the indicator function. In (5.1), C can be interpreted as the number of concordant pairs and D the number of discordant pairs. The *gamma* measure is studied in Goodman and Kruskal (1954), and is considered as a discrete generalization of Kendall's tau for continuous variables. The properties of the *gamma* measure follow directly from its definition. Like the correlation coefficient, its range is $-1 \leq \gamma \leq 1$: $\gamma = 1$ when the number of discordant pairs $D = 0$, $\gamma = -1$ when the number of concordant pairs $C = 0$, and $\gamma = 0$ when the number of concordant pairs equals the number of discordant pairs. Other dependence measures as the discrete generalizations of Kendall's tau or Spearman's ρ can also be used for ordinal response and count response as well as binary response. Furthermore, summaries such as means, variances and correlations could also be meaningful and useful for count data. Initial data analysis is particularly important in multivariate analysis, since the structure of multivariate data is much more complicated than that of univariate data, and the initial data analysis results will shed light on identifying the suitable statistical models.

Statistical modelling usually consists of specification, estimation, and evaluation steps. The specification formulates a probabilistic model which is assumed to have generated the observed

data. At this stage, to choose appropriate models, relevant questions are: "What is the nature of the data?" and " How have the data been generated?" The chosen models should make sense for the data. The decision concerning which model to fit to a set of data should, if possible, be the result of a prior consideration of what might be a suitable model for the process under investigation, as well as the result of computation. In some situations, a data set may have several suitable alternative models. After obtaining estimation and computation results, model selections could be made based on certain criteria.

Diagnostics consist of assessments of the reliability of the estimates, the fit of the model and the overall performance of the model. Both the fitting error of the model and possibly prediction error should be studied. We should also bear in mind that often a small fitting error does not lead to a small prediction error. Sometimes, it is necessary to seek a balance between the two. Appropriate diagnostic checking is an important but not easy step in the whole modelling process.

At the *inference* stage, relevant statements about the population from which the sample was taken can be made based on the statistical modelling (mainly probabilistic models) results from the previous stages. These inferences may be the explanation of changes in responses over margin or time, the effects of covariates on the probabilities of occurrence, the marginal and conditional behaviour of response variables, the probability of exceedance, as well as of hypothesis testing as suggested by the theory in the application domain, and so on. Some relevant questions are: "How can valid inference be drawn?", "What interpretation can be given to the estimates?", "Is there a structural interpretation, relating to the underlying theory in the application?", and "Are the results pointing to further studies?"

5.1.2 Model selection

When modelling a data set, usually it is required only that the model provide accurate predictions or other aspects of data, without necessarily duplicating every detail of the real system. A valid model is any model that gives an adequate representation of the system that is of interest to the model user.

Often a large number of equally good models exist for a particular data set in terms of the specific inference aspect of interest to the practitioner. Model selection is carried out by comparing alternative models. If a model fits the data approximately as well as the other more complex models, we usually prefer the simple one. There are many criteria to distinguish between models. One suitable criterion for choosing a model is the associated maximum loglikelihood value. However, within the

same family, the maximum loglikelihood value usually depends on the number of parameters estimated in the model, with more parameters yielding a bigger value. Thus maximizing this statistic cannot be the sole criterion since we would inevitably choose models with more parameters and more complex structure. In application, parsimonious models which identify the essential relations between the variables and capture the major characteristic features of the problem under study are more useful. Such models often lead to clear and simple interpretation. The ideal situation is that we arrive at a simple model which is consistent with the observed data. In this vein, a balance between the size of the maximum loglikelihood value and the number of parameters is important. But it is often difficult to judge the appropriateness of the balance. One widely used criterion is the Akaike Information Criterion (AIC), which is defined as

$$\text{AIC} = -2\ell(\hat{\boldsymbol{\theta}}; \mathbf{y}) + 2s,$$

where $\ell(\hat{\boldsymbol{\theta}}; \mathbf{y})$ is the maximum loglikelihood of the model, and s is the number of estimated parameters of the model. (With IFM estimation, the AIC is modified to AIC $= -2\ell(\tilde{\boldsymbol{\theta}}; \mathbf{y}) + 2s$.) By definition, a model with a smaller AIC is preferable. The AIC considers the principles of maximum likelihood and the model dimensions (or number of parameters) simultaneously, and thus aims for a balance of maximum likelihood value and model complexity. The negative of AIC/2 is asymptotically an unbiased estimator of the mean expected loglikelihood (see Sakamoto *et al.* 1986); thus AIC can be interpreted as an unbiased estimator of the -2 times the expected loglikelihood of the maximum likelihood. The model having minimum AIC should have minimum prediction error, at least asymptotically. In the use of AIC, it is the difference of AIC values that matters and not the actual values themselves. This is because of the fact that AIC is an estimate of the mean expected loglikelihood of a model. If the difference is less than 1, the goodness-of-fit of these models are almost the same. For a detailed account of AIC, see Sakamoto *et al.* (1986). The AIC was introduced by Akaike (1973) for the purpose of selecting an optimal model from within a set of proposed models (hypotheses). The AIC procedure has been used successfully to identify models; see, for example, Akaike (1977).

The selection of models should also be based on the understanding that it is an essential part of modelling to direct the analysis to aspects which are relevant to the context and to omit other aspects of the real world situation which often lead to spurious results. This is also the reason that we have to be careful not to overparameterize the model, since, although this might improve the goodness-of-fit, it is likely to result in the model portraying spurious features of the sampled data, which may detract from the usefulness of the achieved fit and may lead to poor prediction. The

selection of models should also be based on the consideration of the practical importance of the models, which in turn is based on the nature and extent of the models and their contribution to our understanding to the problem.

Statistical modelling is often an iterative process. The general process is such that after a promising member from a family of models is tentatively chosen, parameters in the model are next efficiently estimated; and finally, the success of the resulting fit is assessed. The now precisely defined model is either accepted by this verification stage or the diagnostic checks carried out will find it lacking in certain respects and should then suggest a sensible modified identification. Further estimation and checking may take place, and the cycle of identification, estimation, and verification is repeated until some satisfactory fits obtain.

5.1.3 Diagnostic checking

A model should be judged by its predictive power as well as its goodness-of-fit. Diagnostic checking is a procedure for evaluating to what extent the data support the model. The AIC only compares models through their relative predictive power; it doesn't assess the goodness-of-fit of the model to the data. In multivariate nonnormal analysis, it is not obvious how the goodness-of-fit checking could be carried out. We discuss this issue in the following.

There are many conventional ways to check the goodness-of-fit of a model. One direct way to check the model is by means of residuals (mainly for continuous data). A diagnostic check based on residuals consists of making a residual plot of the (standardized) residuals. Another frequently applied approach is to calculate some goodness-of-fit statistics. When the checking of residuals is feasible, the goodness-of-fit statistics are often used as a supplement. In multivariate analysis, direct comparison of estimated probabilities with the corresponding empirical probabilities may also be considered as a good and efficient diagnostic checking method.

For multivariate binary or ordinal categorical data, a diagnostic check based on residuals of observed data is not meaningful. However statistics of goodness-of-fit are available in these situations. We illustrate the situation here by means of multivariate binary data. For a d-dimensional random binary vector $\mathbf{Y}$ with a model P, its sample space contains 2^d elements. We denote these by $k = 1, \ldots, 2^d$ with k representing the kth particular outcome pattern and P_k the corresponding probability, with $\sum_{k=1}^{2^d} P_k = 1$. Assume n is the number of observations and $n_1, \ldots, n_{2^d}$ are the empirical frequencies corresponding to $k = 1, \ldots, 2^d$. Let $\tilde{P}_k$ be the estimate of P_k for a specified model. Under the hypothesis that the specified model is the true model and with the assumption of

some regularity conditions (e.g. efficient estimates, see Read and Cressie 1988, §4.1), Fisher (1924) shows, in the case if $\tilde{P}_k$ depends on one estimated parameter, that the Pearson χ^2 type statistic

$$X^2 = \sum_{k=1}^{2^d} \frac{(n_k - n\tilde{P}_k)^2}{n\tilde{P}_k} \tag{5.2}$$

is asymptotically chi-squared with $2^d - 2$ degrees of freedom. If $\tilde{P}_k$ depends on s ($s > 1$) estimated parameters, then the generalization is that (5.2) is asymptotically chi-squared with $2^d - s - 1$ degrees of freedom. A more general situation is that $\mathbf{Y}$ depends on a covariate of g categories. For each category of the covariate, it has the situation of (5.2). If we assume independence between the categories of the covariate, we can form an overall Pearson χ^2 type test statistic for the goodness-of-fit of the model as

$$X^2 = \sum_{\nu=1}^{g} \sum_{k=1}^{2^d} \frac{(n_k^{(\nu)} - n^{(\nu)}\tilde{P}_k^{(\nu)})^2}{n^{(\nu)}\tilde{P}_k^{(\nu)}}, \tag{5.3}$$

where ν is the index of the categories in the covariate. Suppose we estimated s parameters in the model; thus $\tilde{P}_k^{(\nu)}$ depends on s parameters. Under the hypothesis that the specified model is the true model, the test statistic X^2 in (5.3), with some regularity conditions (e.g. efficient estimates), is asymptotically $\chi^2_{g(2^d-1)-s}$, where g is the number of categories of the covariate, and s is the total number of parameters estimated in the model. Similarly, an overall loglikelihood ratio type statistic

$$G^2 = 2\sum_{\nu=1}^{g} \sum_{k=1}^{2^d} n_k^{(\nu)} \log[n_k^{(\nu)}/(n^{(\nu)}\tilde{P}_k^{(\nu)})] \tag{5.4}$$

is also asymptotically $\chi^2_{g(2^d-1)-s}$. X^2 and G^2 are asymptotically equivalent, but there are not the same in finite sample case, so sometimes there is a question of which statistic to choose. Read and Cressie (1988) may shed some light on this matter. The computation of the test statistic X^2 or G^2 requires the calculation of $\tilde{P}_k^{(\nu)}$, which may not be easily obtained, depending on the copula associated with the model (for example, it is generally feasible with mixture of max-id copula but only feasible with relatively low dimension for multinormal copula, unless approximations are used).

One frequently encountered problem in applications with multivariate binary or ordinal categorical data (also count data) is that when the dimension of the response is relatively high, the empirical frequency for some particular outcomes of the response vector is relatively small or even zero. Thus $\tilde{P}_k$ or $\tilde{P}_k^{(\nu)}$ would usually be very small for any particular model, and (5.2) or (5.3) with its related statistical inferences are not suitable in these situations. What we may still do in terms of goodness-of-fit checking in these situations is to limit the comparison of $\tilde{P}_k(\mathbf{x}_i)$ with $n_k(\mathbf{x}_i)/n(\mathbf{x}_i)$ by tables and graphics to outcomes of non-zero frequency (where $\mathbf{x}_i$ is the covariate vector), or to

calculate

$$X^2_{(a)} = \sum_{\{n_k \geq a\}} \frac{(n_k - \tilde{n}_k)^2}{\tilde{n}_k} \quad \text{or} \quad G^2_{(a)} = 2 \sum_{\{n_k \geq a\}} n_k \log(n_k/\tilde{n}_k), \tag{5.5}$$

where $\tilde{n}_k = \sum_{i=1}^{n} \tilde{P}_k(\mathbf{x}_i)$, where k represent the kth patterns of the response variables, and plot $X^2_{(a)}$ (or $G^2_{(a)}$) versus $a = \{1, 2, 3, 4, 5\}$ to get a rough idea of how the model fits the non-zero frequency observations. The data obviously support the model if the observed values of $X^2_{(a)}$ (or $G^2_{(a)}$) go down quickly to zero, while large values indicate potential model departures.

Obviously, in any case, some partial assessments using (5.2) or (5.3) may be done for some lower-dimensional margins where frequencies are sufficiently large. Sometimes, these kinds of goodness-of-fit checking may be used to retain a model while (5.5) is not helpful.

The statistics in (5.5) and related analysis can be applied to multivariate count data as well. Furthermore, a diagnostic check based on the residuals of the observed counts is also meaningful. If there are no covariates, quick and overall residual checking cab be based on examining

$$\tilde{e}_{ij} = y_{ij} - \mathrm{E}[Y_{ij} | \mathbf{Y}_{i,-j}, \tilde{\boldsymbol{\theta}}] \tag{5.6}$$

for a particular fixed j, where $\mathbf{Y}_{i,-j}$ means the response vector $\mathbf{Y}_i$ with the jth margin omitted. The model is considered as adequate based on residual plot in terms of goodness-of-fit if the residuals are small and do not exhibit systematic patterns. Note that the computation of $\mathrm{E}[Y_{ij} | \mathbf{Y}_{i,-j}, \tilde{\boldsymbol{\theta}}]$ may not be a simple task when the dimension d is large (e.g $d > 3$). Another rough check of the goodness-of-fit of a model for multivariate count data is to compare the empirical marginal means, variances and pairwise correlation coefficients with the corresponding means, variances and pairwise correlation coefficients calculated from the fitted model.

In principle, a model can be forced to fit the data increasingly well by increasing its number of parameters. However, the fact that the fitting errors are small is no guarantee that the prediction errors will be. Many of the terms in a complex model may simply be accounting for noise in the data. The overfitted models may predict future values quite poorly. Thus to arrive at a model which represents only the main features of the data, selection and diagnostic criteria which balance model complexity and goodness-of-fit must be used simultaneously. As we have discussed, often there are many relevant models that provide an acceptable approximation to reality or data. The purpose of statistical modelling is not to get the "true" model, but rather to obtain one or several models which extract the most information and better serve the inference purposes.

5.1.4 Testing the dependence structure

We next discuss a topic related to model identification. Short series of longitudinal data or repeated measures with many subjects often exhibit highly structured pattern of dependence structure, with the dependence usually becoming weaker as the time separation (if the observation point is time) increases. Valid inferences can be made by borrowing strength across subjects. That is, the consistency of a pattern across subjects is the basis for substantive conclusions. For this reason, inferences from longitudinal or repeated measures studies can be made more robust to model assumptions than those from time series data, particularly to assumptions about the nature of the dependence.

There are many possible structures for longitudinal or repeated measures type dependence. The exchangeable or AR(1)-like dependence structures are the simplest. But in a particular situation, how to test to see if a particular dependence structure is more plausible? The AIC for model comparison may be a useful index. In the following, we provide an alternative approach for testing special dependence structures. For this purpose, we first give a definition and state two results that we are going to use in the later development. A reference for these materials is Rao (1973).

Definition 5.1 (Generalized inverse of a matrix) *A generalized inverse of an $n \times m$ matrix A of any rank is an $m \times n$ matrix denoted by A^- which satisfies the following equality:*

$$AA^-A = A.$$

□

Result 5.1 (Spectral decomposition theorem) *Let A be a real $n \times n$ symmetric matrix. Then there exists an orthogonal matrix Q such that $Q'AQ$ is a diagonal matrix whose diagonal elements $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ are the characteristic roots of A, that is*

$$Q'AQ = \begin{pmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{pmatrix}.$$

□

Result 5.2 *If $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma_{\mathbf{X}})$, and $\Sigma_{\mathbf{X}}$ is positive semidefinite, then a set of necessary and sufficient conditions for $\mathbf{X}'A\mathbf{X} \sim \chi^2_r(\delta^2)$ is (i) $tr(A\Sigma_{\mathbf{X}}) = r$ and $\boldsymbol{\mu}'A\boldsymbol{\mu} = \delta^2$, (ii) $\Sigma_{\mathbf{X}}A\Sigma_{\mathbf{X}}A\Sigma_{\mathbf{X}} = \Sigma_{\mathbf{X}}A\Sigma_{\mathbf{X}}$, (iii) $\boldsymbol{\mu}'A\Sigma_{\mathbf{X}}A\boldsymbol{\mu} = \boldsymbol{\mu}'A\boldsymbol{\mu}$, (iv) $\boldsymbol{\mu}'(A\Sigma_{\mathbf{X}})^2 = \boldsymbol{\mu}'A\Sigma$. $\chi^2_r(\delta^2)$ denotes the non-central chi-square distribution with noncentality parameter δ^2.* □

In the following, we are going to build up a general statistical test, which in turn can be used to test exchangeable or AR(1)-type dependence assumptions.

Suppose $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \Sigma_{\mathbf{X}})$ where $\Sigma_{\mathbf{X}}$ is known. We want to test if $\boldsymbol{\mu} = \mu\mathbf{1}$, where μ is a constant. Let $\boldsymbol{\alpha} = \Sigma_{\mathbf{X}}^{-1}\mathbf{1}/\mathbf{1}'\Sigma_{\mathbf{X}}^{-1}\mathbf{1}$, then

$$\mathbf{X} - \boldsymbol{\alpha}'\mathbf{X}\mathbf{1} = (X_1 - \boldsymbol{\alpha}'\mathbf{X}, \ldots, X_p - \boldsymbol{\alpha}'\mathbf{X})' = B\mathbf{X},$$

where $B = I - \mathbf{1}\mathbf{1}'\Sigma_{\mathbf{X}}^{-1}/\mathbf{1}'\Sigma_{\mathbf{X}}^{-1}\mathbf{1}$, and I is the identity matrix. Thus $B\mathbf{X} \sim N_p(B\boldsymbol{\mu}, B\Sigma_{\mathbf{X}}B')$. It is easy to see that $Rank(B) = p - 1$, it implies that $Rank(B\Sigma_{\mathbf{X}}B') = p - 1$.

By Result 5.1, there is an orthogonal matrix Q, such that

$$B\Sigma_{\mathbf{X}}B' = Q \begin{pmatrix} \lambda_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \lambda_{p-1} & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} Q',$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{p-1} > 0$. Let

$$A = Q \begin{pmatrix} \frac{1}{\lambda_1} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_{p-1}} & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} Q',$$

then A is a full rank matrix. It is also easy to show that A is a generalized inverse of $B\Sigma_{\mathbf{X}}B'$, and all the conditions in Result 5.2 are satisfied, we thus have

$$\mathbf{X}'B'AB\mathbf{X} \sim \chi^2_{p-1}(\delta^2),$$

where $\delta^2 = \boldsymbol{\mu}'B'AB\boldsymbol{\mu}$, $\delta^2 \geq 0$. $\delta^2 = 0$ is true if and only if $B\boldsymbol{\mu} = 0$, and this in turn is true if and only if $\boldsymbol{\mu} = \mu\mathbf{1}$, that is $\boldsymbol{\mu}$ should be an equal constant vector. Thus under the null hypothesis $\boldsymbol{\mu} = \mu\mathbf{1}$, we should have

$$\mathbf{X}'B'AB\mathbf{X} \sim \chi^2_{p-1},$$

where χ^2_{p-1} means central chi-square distribution with $p - 1$ degrees of freedom.

Now we use an example to illustrate the use of above results.

Example 5.1 Suppose we choose the multivariate logit model with multinormal copula (3.1) with correlation matrix $\Theta = (\theta_{jk})$ to model the d-dimensional binary observations $\mathbf{y}_1, \ldots, \mathbf{y}_n$. We want to know if an exchangeable (that is $\theta_{jk} = \theta$ for all $1 \leq j < k \leq d$ and for some $|\theta| < 1$) or an AR(1)

(that is $\theta_{jk} = \theta^{|j-k|}$ for all $1 \leq j < k \leq d$ and for some $|\theta| < 1$) correlation matrix in the multinormal copula is the suitable assumptions. The above results can be used to test these assumptions. Let $\tilde{\theta}^{(jk)}$ be the IFME of θ from the (j,k) bivariate margin, and $\tilde{\boldsymbol{\theta}} = (\tilde{\theta}^{(12)}, \tilde{\theta}^{(13)}, \ldots, \tilde{\theta}^{(d-1,d)})$. By Theorem 2.4, we have asymptotically

$$\tilde{\boldsymbol{\theta}} \sim N_{d(d-1)/2}(\theta \mathbf{1}, \Sigma_{\tilde{\boldsymbol{\theta}}}),$$

where $\Sigma_{\tilde{\boldsymbol{\theta}}}$ is the inverse of Godambe information matrix of $\tilde{\boldsymbol{\theta}}$. Thus under the exchangeable or AR(1) assumptions of Θ, we have asymptotically

$$\tilde{\boldsymbol{\theta}}' B' A B \tilde{\boldsymbol{\theta}} \sim \chi^2_{d(d-1)/2-1},$$

where

$$B = I - \mathbf{1}\frac{\mathbf{1}'\Sigma_{\tilde{\boldsymbol{\theta}}}^{-1}}{\mathbf{1}'\Sigma_{\tilde{\boldsymbol{\theta}}}^{-1}\mathbf{1}}, \qquad A = Q \begin{pmatrix} \frac{1}{\lambda_1} & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \frac{1}{\lambda_{d(d-1)/2-1}} & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix} Q',$$

and Q is an orthogonal matrix from the spectral decomposition

$$B\Sigma_{\tilde{\boldsymbol{\theta}}}B' = Q \begin{pmatrix} \lambda_1 & \cdots & 0 & 0 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & \lambda_{d(d-1)/2-1} & 0 \\ 0 & \cdots & 0 & 0 \end{pmatrix} Q',$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{d(d-1)/2-1} > 0$. □

The above results are valid for large samples, and can be used in the applications for a rough judgement about the special dependence structure assumptions, though $\Sigma_{\tilde{\boldsymbol{\theta}}}$ would typically have to be estimated from the data.

5.2 Data analysis examples

In this section, we apply and compare some models developed in Chapter 3 on some real data sets, and illustrate the estimation procedures of Chapter 2. Following the discussion in section 5.1, the examples show the stages of the data analysis cycle and the special features related to the specific type of data.

5.2.1 Example with multivariate/longitudinal binary response data

In this subsection, several models for multivariate binary response data with covariates are applied to a subset of a data set from the "Six Cities Study" discussed and analyzed by Ware *et al.* (1984) and Stram *et al.* (1988).

The Six Cities Study is a longitudinal investigation of the effects of indoor and outdoor air pollution on respiratory health. As in most longitudinal studies, there were missing data for some subjects. In this analysis we consider a subset of data with no missing values, gathered in the study on the occurrence of persistent wheeze (graded as wheeze 1 and none 0) of children (total number of 1020) followed from ages 9 to 12 yearly in two different cities: Kingston-Harriman, Tenessee (KHT), and Portage, Wisconsin (PW) in the US. The outdoor air pollution is measured by the children's residence location, that is, the two cities. These two cities have very different ambient air quality. KHT (coded as 1 in the data set) is influenced by air pollution from several metropolitan and industrial areas, and thus has relatively high average concentrations of fine particulate matter and acid aerosols. PW (coded as 0 in the data set) is located in a region that has relatively low concentrations of these polluants. Indoor pollution is measured by level of maternal smoking graded as 1 ($>$ 10 cigarettes) or 0 ($<$ 10 cigarettes). Let us call the outdoor air pollution variable "City", and the indoor pollution variable "Smoking". Smoking is a time-dependent covariate since level of maternal smoking may vary from year to year, and City is considered as time-independent covariate (for the four-year period) since no one in the study moved over the four years. More documentation of the study can be found in Ware *et al.* (1984) and Stram *et al.* (1988). Some of the potential scientific questions are: (1) Does the prevalence of wheeze differ between cities or smoking groups? If so, does the difference change over time. If the effects are constant over time, how should they be estimated? (2) How should the rate of respiratory disease for children whose mothers smoke be compared to the rate for children whose mothers do not smoke?

Tables 5.1 – 5.3 summarize the initial data analysis. Table 5.1 provides the univariate summaries of the data, with the percentages of 1's for the binary response and predictor variables (City and Smoking at 4 time points which we denote by Smoking9, Smoking10, Smoking11 and Smoking12). We see, from response variables Age 9 to Age 12, that the incidence of persistent wheeze for ages 9 to 12 decreases slightly across the ages. The same is true for the maternal smoking levels. Table 5.2 contains the frequencies of the response vector of the 4 time points when ignoring the effects of the covariates. Table 5.3 has the pairwise log odds ratio for the response variables, ignoring the covariates; it gives some indication of the amount of dependence in the response variables in addition

to Table 5.2. Table 5.3 indicates that the dependence for consecutive years is larger.

Multivariate binary response models that were used to model the data include

1. The multivariate logit model from section 3.1, with
   a. multinormal copula (3.1),
   b. multivariate Molenberghs-Lesaffre construction
      i. with bivariate normal copula,
      ii. with Plackett copula (2.8),
      iii. with Frank copula (2.9).
   c. mixture of max-id copula (3.3),
   d. the permutation symmetric copula (3.8).
2. The multivariate probit model with multinormal copula.

The Multivariate logit-normal model (a MMD model) is also used to model this data set, but since in this model fitting, the variance parameters estimates ($\tilde{\sigma}_j$, $j = 1, 2, 3, 4$) all go to 0, it reduces this model in fact to a MCD model, thus we will not pursue the MMD models fitting with this data set further. Only the results with MCD model fitting are reported here.

Since we have the covariates City and Smoking, there is a question of how to include these variables into the models. For subject i ($i = 1, \ldots, 1020$), the cut-off points are z_{ij} ($j = 1, 2, 3, 4$) for an univariate probit or logit model. A suitable approach for the cut-off points to be functions of covariates is to let $z_{ij} = \alpha_{j0} + \alpha_{j1} * \text{City}_i + \alpha_{j2} * \text{Smoking}_{ij}$. To let the dependence parameters be functions of covariates is more complicated. Many possibilities are open. A simple approach is to let the dependence parameters be independent of covariates. This may serve the general modelling purpose in many situation while keeping the model simple. Besides this simple approach, partly for illustrative purposes, we also examine the situation where the dependence parameters depend on the covariate City. For model (1a), the dependence parameters are $\theta_{i,jk}$ for the subject i, $1 \leq j < k \leq 4$. There are many ways to include covariates to the dependence parameters $\theta_{i,jk}$, as we have discussed in section 3.1 for model (1a). For a general dependence structure, we may simply let $\theta_{i,jk} = [\exp(\beta_{jk,0}+\beta_{jk,1}*\text{City}_i)-1]/[\exp(\beta_{jk,0}+\beta_{jk,1}*\text{City}_i)+1]$. Another two dependence structures appropriate (suggested by the nature of the study and the initial data analysis) for this data set are exchangeable and AR(1) type structure with $\Theta_i = (\theta_{i,jk})$ for the ith subject. The exchangeable situation is that $\theta_{i,jk} = \theta_i$ for some $|\theta_i| < 1$. The AR(1) situation is $\theta_{i,jk} = \theta_i^{|j-k|}$ for some $|\theta_i| < 1$. In

both situations, we let $\theta_i = [\exp(\beta_0+\beta_1*\text{City}_i)-1]/[\exp(\beta_0+\beta_1*\text{City}_i)+1]$. For models (1bi), (1bii), (1biii), we first let higher order (≥ 3) parameters $\eta_{i,jkl}$ and $\eta_{i,1234}$ be constant, say 1. (This is usually good enough for practical purposes, refer to section 3.1.) We next let the parameters appearing in the bivariate copulas be functions of covariates. Assume that for model (1bi), the dependence parameters in the bivariate copulas are $\theta_{i,jk}$. Since $\theta_{i,jk}$ are correlation coefficients in bivariate normal copulas, we let $\theta_{i,jk} = [\exp(\beta_{jk,0} + \beta_{jk,1} * \text{City}_i) - 1]/[\exp(\beta_{jk,0} + \beta_{jk,1} * \text{City}_i) + 1]$. For model (1bii), assume $\delta_{i,jk}$s are the parameters in the Plackett copulas; we let $\delta_{i,jk} = \exp(\beta_{jk,0} + \beta_{jk,1} * \text{City}_i)$. For model (1biii), assume $\delta_{i,jk}$s are dependence parameters in the bivariate Frank copulas; we let $\delta_{i,jk} = \exp(\beta_{jk,0} + \beta_{jk,1} * \text{City}_i)$. For model (1c), the dependence parameters are θ_i and $\delta_{i,jk}$ ($1 \leq j < k \leq 4$). (We let the parameter of asymmetry $\nu_{i,j} = 0$ for all i and j.) θ_i represent a general minimum level of dependence, and $\delta_{i,jk}$ represent bivariate dependence exceeding the minimum dependence. For the dependence parameters, we let $\delta_{i,jk} = \exp(\beta_{jk,0} + \beta_{jk,1} * \text{City}_i)$ and $\theta_i = \exp(\beta_0)$ be independent of covariates. For model (1d), the dependence parameters are θ_i. We let $\theta_i = \exp(\beta_0 + \beta_1 * \text{City}_i)$. For model (2), the dependence structure is the same as model (1a).

We use "l" to denote the logit model and "p" to denote the probit model. For the univariate marginal regressions, at least two situations could be considered: regression coefficients differ across margins (or times), denoted by "md"; and regression coefficients common across margins (or times), denoted by "mc". For the regression of the dependence parameters, for models (1a) and (2), we consider the general (denoted by "g"), exchangeable (denoted by "e") and AR(1) (denoted by "a") dependence structures. We also consider the situations with covariate (denoted by "wc") and with no covariate (denoted by "wn") for the dependence parameters. Thus a total of 12 submodels of model (1a) are considered; they are l.md.g.wc, l.md.g.wn, l.md.e.wc, l.md.e.wn, l.md.a.wc, l.md.a.wn, l.mc.g.wc, l.mc.g.wn, l.mc.e.wc, l.mc.e.wn, l.mc.a.wc and l.mc.a.wn, where for example "l.md.g.wc" stands for the multivariate logit model with marginal regression coefficients differ across margins and general dependence structure with covariates. There are also 12 submodels for the model (2): these are p.md.g.wc, p.md.g.wn, p.md.e.wc, p.md.e.wn, p.md.a.wc, p.md.a.wn, p.mc.g.wc, p.mc.g.wn, p.mc.e.wc, p.mc.e.wn, p.mc.a.wc and p.mc.a.wn. For models (1bi), (1bii), (1biii), (1c) and (1d), the AR(1) type latent dependence structure may not be well-defined. In any case, for not repeating similar analysis, we will only consider possible models within the models (1bi), (1bii), (1biii), (1c) and (1d) with similar structure of models retained by the analysis with models (1a) and (2).

For all the models except (1d), the IFM estimation theory is applied. That is, the univariate (regression) parameters are estimated from separate univariate likelihoods (using the Newton-Raphson

method), and bivariate and multivariate (regression) parameters are estimated from bivariate likelihoods, using a quasi-Newton optimization routine, with univariate parameters fixed as estimated from the separate univariate likelihoods. Furthermore, for the situation of "mc" for common marginal regression coefficients and exchangeable (or AR(1) if applicable) dependence structure, WA of (2.93) in section 2.6 for parameter estimation based on IFM is used. It is also used for estimating the parameter β_0 in $\theta = \exp(\beta_0)$ in the model (1c) since θ is an overall parameter and common across all margins. Notice that only one choice of parametric families for ψ and K_{jk}'s were used, but it is expected that other choices could lead to a better fit according to AIC. The model (1d) has a copula with closed form cdf and there is only one dependence (or regression parameters related to it) parameter in the model, thus MLE(s) are computed in this situation. Model (1d) here is used to compare a simple permutation symmetric MCD model with the other models which all allow a general dependence structure. Model (1c) and (1d) have the advantage of having a copula with closed form cdf; this is particularly convenient for dealing with multivariate discrete data of high dimension, as it leads to faster computation in computing probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y})$ or $\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{x})$.

For standard errors (SEs) of parameter estimates and prediction probabilities, the jackknife method from Chapter 2 is used with 255 random groups of 4. Furthermore, the weights used for WA for common parameter estimation are based on the jackknife SEs, and these weights in turn are used, based on (2.93), to each step of jackknife parameter estimation.

Summaries of the fits of the models are given in several tables. Table 5.4 contains the estimates and SEs of the regression parameters for the marginal parameters with the logit model when the regression parameters are considered to be differ and common across the margins. Table 5.5 contains the estimates and SEs of the regression parameters for the dependence parameters under various settings for the multivariate logit model with multinormal copula (model (1a)). Table 5.6 contains AIC values and X^2 (calculated based on (5.5) with $a = 0$) values for all the submodels of multivariate logit and probit models with multinormal copula (that is models (1a) and (2)). Care must be taken in the comparison since the AICs here are not calculated from the ML of all parameters simultaneously, the parameters estimates are IFME. The AIC values and X^2 values for the corresponding submodel of models (1a), (2) are comparable; this echoes the well-known fact that the univariate probit and logit models are comparable. We thus only compare the submodels within the multivariate logit model. From examining the AIC and X^2 values for the 12 models, the models l.md.g.wn, l.md.e.wc, l.md.e.wn and l.mc.g.wn seem to stand out as interesting choices.

Since l.md.e.wc and l.md.e.wn are about the same in terms of AIC and X^2 values, and l.md.e.wn is simpler than l.md.e.wc, we only consider l.md.e.wn. At this stage, three models are retained for further inspection: l.md.g.wn, l.md.e.wn and l.mc.g.wn. Table 5.7 contains AIC values and Table 5.8 contains X^2 values of submodels l.md.g.wn, l.md.e.wn and l.mc.g.wn of models (1a), (1bi), (1bii), (1biii), (1c) and (1d). These two tables suggest that the models are comparable in general, with models (1c) and (1d) performing relatively poorly; possibly other parametric families of mixture of max-id copulas would do better. The model (1bi) seems to be the best for this data set. Note that since there is only one dependence structure with model (1d), the submodel l.md.g.wn and l.md.e.wn are equivalent in this case. Table 5.9 contains estimates and SEs of the bivariate dependence parameters of the submodel l.md.g.wn of models (1bi), (1bii), (1biii) and (1c). This and Table 5.5 also suggest that the models are comparable; the conclusion about which bivariate margins are more or less dependent are the same from the models. They show that the dependence for consecutive years is slightly stronger; this is also observed in Table 5.3 for the initial data analysis. (Note also the closeness of the dependence parameter estimates with the model (1bii) to the empirical pairwise log odds ratio in Table 5.3.) For comparison, the estimate of the dependence parameter for model (1d), with an permutation symmetric copula, is 1.719 with SE equal 0.067. As we have pointed out, the model (1d) does not perform as well as other models with this data set. This may indicate that, even though a model with exchangeable dependence structure may be acceptable, a better model is to have a general dependence structure. All these models are quite similar in term of computer time for the parameter estimation (because of the IFM approach), but models (1a) and (2) used much more computer time than the other models to compute AIC, X^2 and the prediction probabilities since 4-dimensional integrations were involved.

For the predicted probabilities and inference, and also as a supplement of X^2 values for an assessment of goodness-of fit, Table 5.10 contains estimates of probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y})$ for all possible $\mathbf{y}$ from submodels l.md.g.wn, l.md.e.wn and l.mc.g.wn of the model (1a), Table 5.11 contains estimates of probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y})$ for all possible $\mathbf{y}$ from submodels l.md.g.wn of models (1a), (1bi), (1bii), (1biii), (1c) and (1d), and these $\Pr(\mathbf{Y} = \mathbf{y})$ are estimated with $\sum_{i=1}^{1020} \tilde{\Pr}(\mathbf{Y} = \mathbf{y}|\mathbf{x}_i)/1020$. Table 5.12 contains estimates of probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{x})$ for various $\mathbf{y}$ and $\mathbf{x}$ from submodels l.md.g.wn, l.md.e.wn and l.mc.g.wn of the model (1a), and Table 5.13 contains estimates of probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{x})$ for various $\mathbf{y}$ and $\mathbf{x}$ from submodels l.md.g.wn of models (1a), (1bi), (1bii), (1biii), (1c) and (1d). In Table 5.12, the n^* is the subset sizes for the specific value of $\mathbf{x}$ and "rel. freq" is the observed relative frequency for the given $\mathbf{y}$ under

that value of $\mathbf{x}$. In Table 5.13, to save space, for each line, only the maximum estimated SE over the different models is given; actually the SEs are quite close to each other. The selected $\mathbf{x}$ and $\mathbf{y}$ values in the Table 5.12 and Table 5.13 are common values in the data set. Tables 5.10 – 5.13 suggest that the submodels l.md.g.wn, l.md.e.wn and l.mc.g.wn are all adequate for predictive purposes, since the prediction probabilities are comparable when the SEs are taken into account. These three submodels can be used to complement each other for a slightly different inference purposes. The large divergences in estimated probabilities occur with the model (1d) and only in the case where the vector $\mathbf{x}$ is at the extreme of the covariate space, for example $x = (1, 0, 0, 0, 0)$ and $x = (0, 0, 0, 0, 0)$ for $\mathbf{y} = (1, 1, 1, 1)$. There is a simple exchangeable dependence model l.md.e.wn among the three submodels l.md.g.wn, l.md.e.wn and l.mc.g.wn. An explanation for this may be that the dependence in the bivariate margins are different but not different enough to make a difference in prediction probabilities. Another possibility may be due to the dominance of the response vector $(0, 0, 0, 0)$.

The analysis (e.g. from submodel l.md.g.wn) indicates a slight decline in the rate of wheeze over time (the intercepts in Table 5.4 for regression parameters differing across margins decrease gradually over time from -1.090 to -1.564) and a moderate increase in wheeze for children of mothers who smoke (the corresponding regression parameters increases over time from 0.144 to 0.444) and for the city with pollution (the corresponding regression parameters increase in time from 0.003 to 0.209). There is an indication that the excess of maternal smoking and the city with pollution both increase significantly the probability of the occurrence of wheeze (e.g. from submodel l.mc.g.wn). This is also consistent with the observation in Ware *et al.* (1984), where it is believed that maternal smoking is predictive of respiratory illness. If we study the model with covariate (city) for the dependence (e.g. the submodel l.md.e.wc), we see that high city pollution level has a negative effect on the correlation; it possibly means that the low level city pollution leads to a slightly higher correlation on the occurrence of persistent wheeze. We can interpret this as the wheeze occurrence situation not caused by pollution is more stable over time. The analysis indicates that the dependence for consecutive years is stronger and the dependence (pairwise) are all significant. The rate of respiratory disease for children whose mothers smoke heavily is higher than the rate for children whose mothers do not smoke or only smoke slightly, these can be seen from Table 5.12 (with l.md.g.wn submodel), where for example for $\mathbf{y} = (1, 1, 1, 1)$, $\tilde{P}(\mathbf{y}|\mathbf{x}^{(a)}) = 0.099 > 0.071 = \tilde{P}(\mathbf{y}|\mathbf{x}^{(b)})$ where $\mathbf{x}^{(a)} = (0, 1, 1, 1, 1)$ and $\mathbf{x}^{(b)} = (0, 0, 0, 0, 0)$, and $\tilde{P}(\mathbf{y}|\mathbf{x}^{(c)}) = 0.118 > 0.085 = \tilde{P}(\mathbf{y}|\mathbf{x}^{(d)})$ where $\mathbf{x}^{(c)} = (1, 1, 1, 1, 1)$ and $\mathbf{x}^{(d)} = (1, 0, 0, 0, 0)$. Similarly, we also observe that rate of persistent wheeze for children whose mothers smoke is lower than the rate for children whose mothers do not

Table 5.1: Six Cities Study: Percentages for binary variables

Variables	# 1's	Percentage
Age 9	266	26.07%
Age 10	256	25.09%
Age 11	241	23.62%
Age 12	217	21.27%
City	512	50.19%
Smoking9	325	31.86%
Smoking10	313	30.68%
Smoking11	311	30.49%
Smoking12	309	30.29%

Table 5.2: Six Cities Study: Frequencies of the response vector (Age 9, 10, 11, 12)

Response pattern	Observed numbers	Relative frequency
1 1 1 1	95	0.093
1 1 1 0	30	0.029
1 1 0 1	15	0.015
1 1 0 0	28	0.027
1 0 1 1	14	0.014
1 0 1 0	9	0.009
1 0 0 1	12	0.012
1 0 0 0	63	0.062
0 1 1 1	19	0.019
0 1 1 0	15	0.015
0 1 0 1	10	0.010
0 1 0 0	44	0.043
0 0 1 1	17	0.017
0 0 1 0	42	0.041
0 0 0 1	35	0.034
0 0 0 0	572	0.561

smoke (e.g. for $\mathbf{y} = (0,0,0,0)$, $\tilde{P}(\mathbf{y}|\mathbf{x}^{(e)}) = 0.606 > 0.541 = \tilde{P}(\mathbf{y}|\mathbf{x}^{(f)})$ where $\mathbf{x}^{(e)} = (0,0,0,0,0)$ and $\mathbf{x}^{(f)} = (0,1,1,1,1)$). Also similarly, the rate of persistent wheeze for children who reside in the city with more pollution is higher than the rate for children who reside in the city with less pollution, e.g., for $\mathbf{y} = (1,1,1,1)$, $\tilde{P}(\mathbf{y}|\mathbf{x}^{(g)}) = 0.118 > 0.099 = \tilde{P}(\mathbf{y}|\mathbf{x}^{(h)})$ where $\mathbf{x}^{(g)} = (1,1,1,1,1)$ and $\mathbf{x}^{(h)} = (0,1,1,1,1)$, or $\tilde{P}(\mathbf{y}|\mathbf{x}^{(i)}) = 0.085 > 0.071 = \tilde{P}(\mathbf{y}|\mathbf{x}^{(j)})$ where $\mathbf{x}^{(i)} = (1,0,0,0,0)$ and $\mathbf{x}^{(j)} = (0,0,0,0,0)$. More detailed comparisons for different situations can be made.

Similar results to the partial interpretations given above are also obtained in the literature on the analysis of a similar data set from the same study; see for example Fitzmaurice and Laird (1993), Zeger *et al.* (1988) and Stram *et al.* (1988).

Table 5.3: Six Cities Study: Pairwise log odds ratios for Age 9, 10, 11, 12

Pair	odds	log odds
Age 9,10	12.97	2.56
Age 9,11	8.91	2.19
Age 9,12	8.69	2.16
Age 10,11	13.63	2.61
Age 10,12	10.45	2.35
Age 11,12	14.83	2.69

Table 5.4: Six Cities Study: Estimates of marginal regression parameters for multivariate logit model

margin	intercept (SE)	city (SE)	smoking (SE)
differ across the margins			
1	-1.090 (0.113)	0.003 (0.150)	0.144 (0.144)
2	-1.229 (0.120)	0.080 (0.148)	0.293 (0.155)
3	-1.412 (0.136)	0.311 (0.161)	0.237 (0.166)
4	-1.564 (0.123)	0.209 (0.155)	0.444 (0.166)
common across the margins			
	-1.308 (0.061)	0.144 (0.077)	0.270 (0.078)

Table 5.5: Six Cities Study: Estimates of dependence regression parameters for multivariate logit model with multinormal copula

margin	intercept (SE)	city (SE)
general dependence, with covariate		
12	2.156 (0.217)	-0.373 (0.288)
13	1.740 (0.183)	-0.178 (0.249)
14	1.583 (0.209)	0.041 (0.288)
23	2.334 (0.210)	-0.628 (0.287)
24	1.891 (0.214)	-0.294 (0.281)
34	2.079 (0.224)	-0.109 (0.287)
general dependence, without covariate		
12	1.960 (0.143)	
13	1.645 (0.124)	
14	1.604 (0.143)	
23	1.987 (0.143)	
24	1.733 (0.139)	
34	2.020 (0.142)	
exchangeable dependence, with covariate		
	1.948 (0.085)	-0.254 (0.114)
exchangeable dependence, without covariate		
	1.815 (0.057)	
AR(1) dependence, with covariate		
	2.380 (0.086)	-0.258 (0.115)
AR(1) dependence, without covariate		
	2.236 (0.057)	

Table 5.6: Six Cities Study: Comparisons of AIC values and X^2 values from various submodels of models (1a) and (2)

Logit Models	AIC	X^2	Probit Models	AIC	X^2
l.md.g.wc	3642.882	7.862	p.md.g.wc	3642.810	7.863
l.md.g.wn	3637.415	7.992	p.md.g.wn	3637.347	7.993
l.md.a.wc	3659.683	42.472	p.md.a.wc	3659.615	42.480
l.md.a.wn	3660.126	42.329	p.md.a.wn	3660.063	42.331
l.md.e.wc	3641.662	19.941	p.md.e.wc	3641.585	19.950
l.md.e.wn	3641.637	19.926	p.md.e.wn	3641.561	19.939
l.mc.g.wc	3646.339	21.031	p.mc.g.wc	3646.396	21.086
l.mc.g.wn	3640.444	21.119	p.mc.g.wn	3640.499	21.175
l.mc.a.wc	3661.238	52.281	p.mc.a.wc	3661.308	52.386
l.mc.a.wn	3660.876	51.942	p.mc.a.wn	3660.938	52.044
l.mc.e.wc	3647.731	37.758	p.mc.e.wc	3647.784	37.839
l.mc.e.wn	3646.785	37.687	p.mc.e.wn	3646.833	37.771

Table 5.7: Six Cities Study: Comparisons of AIC values from various models

Models	(1a)	(1bi)	(1bii)	(1biii)	(1c)	(1d)
l.md.g.wn	3637.415	3631.991	3634.040	3636.773	3668.059	3703.763
l.md.e.wn	3641.637	3637.184	3638.884	3642.080	3663.932	–
l.mc.g.wn	3640.444	3633.903	3635.855	3638.423	3674.095	3708.404

Table 5.8: Six Cities Study: Comparisons of X^2 values from various models

Models	(1a)	(1bi)	(1bii)	(1biii)	(1c)	(1d)
l.md.g.wn	7.992	1.765	1.795	1.866	31.747	77.157
l.md.e.wn	19.926	15.832	15.485	16.131	36.315	–
l.mc.g.wn	21.119	13.959	14.249	14.436	50.613	96.223

Table 5.9: Six Cities Study: Estimates (SE) of dependence regression parameters from the submodel l.md.g.wn of various models

margin	(1bi)	(1bii)	(1biii)	(1c)
12	1.960 (0.143)	2.556 (0.173)	1.861 (0.094)	2.715 (0.202)
13	1.645 (0.124)	2.183 (0.153)	1.654 (0.090)	2.181 (0.183)
14	1.604 (0.143)	2.150 (0.178)	1.638 (0.107)	2.066 (0.179)
23	1.987 (0.143)	2.605 (0.173)	1.893 (0.095)	2.621 (0.188)
24	1.733 (0.139)	2.316 (0.173)	1.737 (0.101)	2.362 (0.215)
34	2.020 (0.142)	2.681 (0.173)	1.953 (0.095)	2.833 (2.996)
$\log(\theta)$				1.513 (0.051)

Table 5.10: Six Cities Study: Estimates of $\Pr(\mathbf{Y} = \mathbf{y})$ from various submodels of model (1a)

y	freq.	rel. freq.	l.md.g.wn pred. prob	l.md.e.wn pred. prob.	l.mc.g.wn pred. prob.
1 1 1 1	95	0.093	0.087	0.087	0.089
1 1 1 0	30	0.029	0.029	0.027	0.023
1 1 0 1	15	0.015	0.017	0.021	0.017
1 1 0 0	28	0.027	0.031	0.023	0.024
1 0 1 1	14	0.014	0.015	0.018	0.017
1 0 1 0	9	0.009	0.013	0.020	0.011
1 0 0 1	12	0.012	0.013	0.015	0.015
1 0 0 0	63	0.062	0.055	0.049	0.045
0 1 1 1	19	0.019	0.021	0.017	0.025
0 1 1 0	15	0.015	0.018	0.018	0.017
0 1 0 1	10	0.010	0.010	0.014	0.012
0 1 0 0	44	0.043	0.037	0.044	0.034
0 0 1 1	17	0.017	0.018	0.012	0.024
0 0 1 0	42	0.041	0.034	0.038	0.035
0 0 0 1	35	0.034	0.031	0.028	0.042
0 0 0 0	572	0.561	0.569	0.568	0.570

Table 5.11: Six Cities Study: Estimates of $\Pr(\mathbf{Y} = \mathbf{y})$ from the submodel l.md.g.wn of various models

y	rel. freq.	(1a)	(1bi)	prob (1bii)	(1biii)	(1c)	(1d)
1 1 1 1	0.093	0.087	0.092	0.092	0.091	0.087	0.064
1 1 1 0	0.029	0.029	0.028	0.028	0.028	0.026	0.036
1 1 0 1	0.015	0.017	0.016	0.016	0.016	0.018	0.029
1 1 0 0	0.027	0.031	0.029	0.029	0.029	0.030	0.027
1 0 1 1	0.014	0.015	0.014	0.014	0.014	0.014	0.026
1 0 1 0	0.009	0.013	0.011	0.012	0.012	0.022	0.023
1 0 0 1	0.012	0.013	0.011	0.011	0.011	0.019	0.018
1 0 0 0	0.062	0.055	0.060	0.060	0.060	0.045	0.039
0 1 1 1	0.019	0.021	0.020	0.020	0.020	0.016	0.024
0 1 1 0	0.015	0.018	0.016	0.016	0.016	0.020	0.021
0 1 0 1	0.010	0.010	0.008	0.008	0.008	0.016	0.016
0 1 0 0	0.043	0.037	0.042	0.042	0.043	0.038	0.034
0 0 1 1	0.017	0.018	0.016	0.017	0.017	0.018	0.015
0 0 1 0	0.041	0.034	0.039	0.039	0.039	0.032	0.028
0 0 0 1	0.034	0.031	0.036	0.036	0.036	0.025	0.021
0 0 0 0	0.561	0.569	0.561	0.561	0.560	0.574	0.580

Table 5.12: Six Cities Study: Observed frequencies in comparison with estimates of $\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{x})$ from various models, $\mathbf{x}$ =(City, Smoking9, Smoking10, Smoking11, Smoking12).

$\mathbf{x}$	n^*	$\mathbf{y}$	rel. freq.	l.md.g.wn pred. prob (SE)	l.md.e.wn pred. prob. (SE)	l.mc.g.wn pred. prob. (SE)
(0,1,1,1,1)	118	(1,1,1,1)	0.076	0.099 (0.015)	0.099 (0.015)	0.101 (0.015)
(0,1,1,1,1)	118	(1,1,1,0)	0.025	0.027 (0.007)	0.025 (0.006)	0.024 (0.004)
(0,1,1,1,1)	118	(1,1,0,0)	0.025	0.035 (0.008)	0.026 (0.006)	0.025 (0.003)
(0,1,1,1,1)	118	(1,0,0,0)	0.059	0.054 (0.011)	0.049 (0.009)	0.046 (0.005)
(0,1,1,1,1)	118	(0,0,0,0)	0.534	0.541 (0.030)	0.541 (0.030)	0.542 (0.030)
(0,0,0,0,0)	344	(1,1,1,1)	0.087	0.071 (0.010)	0.070 (0.010)	0.074 (0.011)
(0,0,0,0,0)	344	(1,1,1,0)	0.035	0.028 (0.005)	0.027 (0.004)	0.020 (0.003)
(0,0,0,0,0)	344	(1,1,0,0)	0.026	0.035 (0.006)	0.027 (0.004)	0.022 (0.003)
(0,0,0,0,0)	344	(1,0,0,0)	0.076	0.063 (0.009)	0.058 (0.008)	0.043 (0.004)
(0,0,0,0,0)	344	(0,0,0,0)	0.608	0.606 (0.023)	0.606 (0.023)	0.609 (0.023)
(1,1,1,1,1)	131	(1,1,1,1)	0.145	0.118 (0.015)	0.118 (0.015)	0.118 (0.015)
(1,1,1,1,1)	131	(1,1,1,0)	0.015	0.028 (0.007)	0.026 (0.006)	0.027 (0.004)
(1,1,1,1,1)	131	(1,1,0,0)	0.046	0.026 (0.006)	0.019 (0.005)	0.027 (0.003)
(1,1,1,1,1)	131	(1,0,0,0)	0.076	0.043 (0.009)	0.037 (0.008)	0.047 (0.005)
(1,1,1,1,1)	131	(0,0,0,0)	0.397	0.505 (0.028)	0.504 (0.028)	0.505 (0.028)
(1,0,0,0,0)	306	(1,1,1,1)	0.085	0.085 (0.011)	0.085 (0.011)	0.087 (0.012)
(1,0,0,0,0)	306	(1,1,1,0)	0.023	0.031 (0.006)	0.028 (0.004)	0.022 (0.003)
(1,0,0,0,0)	306	(1,1,0,0)	0.016	0.027 (0.005)	0.020 (0.004)	0.024 (0.003)
(1,0,0,0,0)	306	(1,0,0,0)	0.049	0.052 (0.009)	0.046 (0.007)	0.045 (0.004)
(1,0,0,0,0)	306	(0,0,0,0)	0.592	0.574 (0.023)	0.573 (0.023)	0.574 (0.023)

Table 5.13: Six Cities Study: Estimates of $\Pr(\mathbf{Y} = \mathbf{y}|\mathbf{x})$ from the submodel l.md.g.wn of various models, $\mathbf{x}$ =(City, Smoking9, Smoking10, Smoking11, Smoking12).

x	rel. freq.	y	(1a)	(1bi)	(1bii)	prob. (1biii)	(1c)	(1d)	maxSE
(0,1,1,1,1)	0.076	(1,1,1,1)	0.099	0.104	0.105	0.106	0.099	0.076	0.017
(0,1,1,1,1)	0.025	(1,1,1,0)	0.027	0.025	0.025	0.024	0.024	0.035	0.010
(0,1,1,1,1)	0.025	(1,1,0,0)	0.035	0.033	0.033	0.033	0.033	0.028	0.011
(0,1,1,1,1)	0.059	(1,0,0,0)	0.054	0.059	0.058	0.058	0.043	0.037	0.011
(0,1,1,1,1)	0.534	(0,0,0,0)	0.541	0.533	0.535	0.537	0.549	0.557	0.030
(0,0,0,0,0)	0.087	(1,1,1,1)	0.071	0.076	0.074	0.071	0.070	0.046	0.012
(0,0,0,0,0)	0.035	(1,1,1,0)	0.028	0.027	0.027	0.028	0.025	0.035	0.008
(0,0,0,0,0)	0.026	(1,1,0,0)	0.035	0.033	0.034	0.034	0.034	0.032	0.007
(0,0,0,0,0)	0.076	(1,0,0,0)	0.063	0.068	0.069	0.068	0.054	0.048	0.009
(0,0,0,0,0)	0.608	(0,0,0,0)	0.606	0.598	0.595	0.591	0.608	0.611	0.023
(1,1,1,1,1)	0.145	(1,1,1,1)	0.118	0.121	0.125	0.128	0.118	0.097	0.017
(1,1,1,1,1)	0.015	(1,1,1,0)	0.028	0.027	0.026	0.025	0.025	0.035	0.009
(1,1,1,1,1)	0.046	(1,1,0,0)	0.026	0.024	0.024	0.024	0.024	0.020	0.011
(1,1,1,1,1)	0.076	(1,0,0,0)	0.043	0.048	0.047	0.047	0.032	0.026	0.009
(1,1,1,1,1)	0.397	(0,0,0,0)	0.505	0.498	0.503	0.507	0.515	0.525	0.028
(1,0,0,0,0)	0.085	(1,1,1,1)	0.085	0.090	0.090	0.089	0.085	0.061	0.012
(1,0,0,0,0)	0.023	(1,1,1,0)	0.031	0.029	0.030	0.030	0.027	0.037	0.007
(1,0,0,0,0)	0.016	(1,1,0,0)	0.027	0.025	0.025	0.025	0.027	0.023	0.007
(1,0,0,0,0)	0.049	(1,0,0,0)	0.052	0.057	0.057	0.058	0.042	0.036	0.009
(1,0,0,0,0)	0.592	(0,0,0,0)	0.574	0.566	0.566	0.564	0.578	0.584	0.024

5.2.2 Example with multivariate/longitudinal ordinal response data

In this subsection, several models for multivariate ordinal response data are applied to a longitudinal data set from Fienberg *et al.* (1985) and Conaway (1989) from a study on the psychological effects of the accident at the Three Mile Island (TMI) nuclear power plant in 1979.

The study focuses on the changes in levels of stress of mothers of young children living within 10 miles of the plant. Four waves of interviews were conducted in 1979, 1980, 1981, 1982, and the levels of stress (categorized as low, medium, or high from a composite score of 90-item checklist) of 268 mothers at each time point are measured. Hence stress is treated as an ordinal response variables with three categories, now labelled as 1 for low, 2 for medium, and 3 for high (levels of stress). Respondents were stratified into two groups, those living within 5 miles (labelled as 0) of the plant and those living between 5 and 10 miles (labelled as 1) from the plant. Let we call this variable "Distance". Some of the potential issues of interest are: (1) to compare the groups with respect to the changes in stress levels over time; (2) to assess the degree of stress associated with the accident.

The response vectors are 4-dimensional ordinal categorical measures with three levels. There are 81 possible four-tuples of the form (1,1,1,1) to (3,3,3,3). Table 5.14 lists the frequencies of the four-tuples by group (based on distance); only the 35 four-tuples with non-zero frequency in at least one the groups are listed. The variable ID is used to identify the non-zero frequency response patterns in the table. There are 115 mothers in the group within 5 miles of the plant and 153 in the group exceeding 5 miles. The table shows that there is only one subject (ID= 11) with a big change in the stress level (3 to 1 and 1 to 3) from one year to another. 42% of the subjects are categorized into the same stress level (ID= $1, 19, 35$) in all four years; 80.5% of these subjects are in the medium stress category (ID= 19). The frequencies by univariate margin (year) are given in Table 5.15. The medium stress category predominates and there is a higher relative frequency of subjects in the high stress category for the group within 5 miles of the plant. From Table 5.15, there are not big changes over time, but there is slight trend towards lower stress level for the group exceeding 5 miles. Table 5.16 has the pairwise *gamma* measures (5.1) for the response variables, for the group within 5 miles, the group exceeding 5 miles and also ignoring the covariate "Distance"; this gives a more detailed indication of the dependence in the response variables than Table 5.14. Table 5.16 indicates the dependence for consecutive years are larger, and all the *gamma* measures are larger for the group within 5 miles of the plant.

Multivariate ordinal response models that were used to model the data are:

1. The multivariate logit model from section 3.5, with
   a. multinormal copula (3.1),
   b. multivariate Molenberghs-Lesaffre construction
      i. with bivariate normal copula,
      ii. with Plackett copula (2.8),
      iii. with Frank copula (2.9),
   c. mixture of max-id copula (3.3),
   d. the permutation symmetric copula (3.8).
2. The multivariate probit model with multinormal copula.

In this data set, the single dichotomous covariate "Distance" can be considered as margin-independent (or time-independent). For subject i ($i = 1, \ldots, 268$), the cut-off points are $-\infty = z_{ij}(0) \leq z_{ij}(1) \leq z_{ij}(2) \leq z_{ij}(3) = \infty$ ($j = 1, 2, 3, 4$), where $z_{ij}(1)$ and $z_{ij}(2)$ need to be modelled and estimated. A suitable approach to let the cut-off points $z_{ij}(1)$ and $z_{ij}(2)$ be functions of covariates (see section 3.5) is to let $z_{ij}(1) = \gamma_j(1) + \alpha_j * \text{Distance}$ and $z_{ij}(2) = \gamma_j(2) + \alpha_j * \text{Distance}$. The ways for the dependence parameters be functions of covariates are similar to that of multivariate binary response examples in subsection 5.2.1. The simplest approach is to let the dependence parameters be independent of covariates. The various situations for the dependence parameters depending on the covariate are the following. We here only list the models, for details please refer to subsection 5.2.1. For model (1a), let $\theta_{i,jk} = [\exp(\beta_{jk,0} + \beta_{jk,1} * \text{Distance}_i) - 1]/[\exp(\beta_{jk,0} + \beta_{jk,1} * \text{Distance}_i) + 1]$ for a general dependence structure and $\theta_i = [\exp(\beta_0 + \beta_1 * \text{Distance}_i) - 1]/[\exp(\beta_0 + \beta_1 * \text{Distance}_i) + 1]$ for exchangeable and AR(1) dependence structure. For models (1bi), (1bii), (1biii), we first let higher order (> 3) parameters $\delta_{i,jkl}$ and $\delta_{i,1234}$ be 1 (see explanation in the example in subsection 5.2.1). We then let $\theta_{i,jk} = [\exp(\beta_{jk,0} + \beta_{jk,1} * \text{Distance}_i) - 1]/[\exp(\beta_{jk,0} + \beta_{jk,1} * \text{Distance}_i) + 1]$ for model (1bi), $\delta_{i,jk} = \exp(\beta_{jk,0} + \beta_{jk,1} * \text{Distance}_i)$ for model (1bii), and for model (1biii). For model (1c), we let $\delta_{i,jk} = \exp(\beta_{jk,0} + \beta_{jk,1} * \text{Distance}_i)$ and $\theta_i = \exp(\beta_0)$ be independent of covariate. (The parameter of asymmetry $\nu_{i,j}$ is set to 0 for all i and j.) Again, notice that only one choice of parametric families for ψ and K_{jk}'s were used, but it is expected that other choices could lead to a better fit according to AIC. For model (1d), the dependence parameters are θ_i and let $\theta_i = \exp(\beta_0 + \beta_1 * \text{Distance}_i)$. For model (2), the dependence structure is the same as model (1a).

As for the example in subsection 5.2.1, we study 12 submodels for the model (1a). They are: l.md.g.wc, l.md.g.wn, l.md.e.wc, l.md.e.wn, l.md.a.wc, l.md.a.wn, l.mc.g.wc, l.mc.g.wn, l.mc.e.wc,

l.mc.e.wn, l.mc.a.wc and l.mc.a.wn. The 12 submodels for the model (2) are: p.md.g.wc, p.md.g.wn, p.md.e.wc, p.md.e.wn, p.md.a.wc, p.md.a.wn, p.mc.g.wc, p.mc.g.wn, p.mc.e.wc, p.mc.e.wn, p.mc.a.wc and p.mc.a.wn. For models (1bi), (1bii), (1biii), (1c) and (1d), the AR(1) type latent dependence structure may not be well-defined. In any case, to avoid repeating similar analysis, we will only consider possible models within the models (1bi), (1bii), (1biii), (1c) and (1d) with similar structure of models retained by the analysis with models (1a) and (2).

For all the models except (1d), the IFM estimation theory is applied. That is, the univariate (regression) parameters are estimated from separate univariate likelihoods, and bivariate and multivariate (regression) parameters are estimated from bivariate likelihoods, with univariate parameters fixed as estimated from the separate univariate likelihoods. For "mc" models involving common marginal regression coefficients and exchangeable (or AR(1) if applicable) dependence structure, WA of (2.93) in section 2.6 for parameter estimation based on IFM is used, and it is also used for estimating the parameter β_0 in $\theta = \exp(\beta_0)$ in the model (1c). MLEs are computed in the model (1d). For standard errors (SEs) of parameter estimates and prediction probabilities, the (delete-one) jackknife method from Chapter 2 is used. These are all similar to the use of these models in subsection 5.2.1.

Summaries of the model fits are given in several tables. Table 5.17 contains the estimates and SEs of the regression parameters for the univariate parameters with the logit model when the regression parameters are considered to be different and common across the univariate margins. Table 5.18 contains the estimates and SEs of the regression parameters for the dependence parameters under various settings for the multivariate logit model with multinormal copula (the model (1a)). Table 5.19 contains AIC values and $X^2_{(2)}$ (calculated based on (5.5) with $a = 2$) values for all the submodels of multivariate logit and probit models with multinormal copula (that is, models (1a) and (2)). The AIC values and $X^2_{(a)}$ (not only $a = 2$, but for all a) values for the corresponding submodel of models (1a), (2) are comparable, similar to what we have observed for the models examples in subsection 5.2.1. We thus only compare the submodels within the multivariate logit model. From examining the AIC values and $X^2_{(2)}$ values for the 12 models, the models l.md.g.wn, l.md.a.wc seem to stand out as interesting choices, with l.md.a.wc appearing to be the better one. Since there is no equivalent way to express the AR(1) structure with the models (1bii), (1biii), (1c) and (1d), for the comparison study, we focus on the submodel l.md.g.wn. Table 5.20 contains AIC values and $X^2_{(2)}$ values of submodel l.md.g.wn of models (1a), (1bii), (1biii), (1c) and (1d). The AIC value and $X^2_{(2)}$ value are not available for (1bi) model, since the dependence parameter estimates obtained from IFM deviate

slightly from forming a compatible set of dependence parameters for a proper Molenberghs-Lesaffre construction multivariate object evaluation. Based on the available AIC values and $X^2_{(2)}$ values, Table 5.20 suggests that the models (1a), (1bii), (1biii) are comparable in general, with models (1c) and (1d) fitting relatively poorly. Table 5.21 contains estimates and SEs of the bivariate dependence parameters of the submodel l.md.g.wn of models (1bi), (1bii), (1biii) and (1c). This and Table 5.20 also suggest that the models (1a), (1bi), (1bii), (1biii) are comparable. The conclusion about which bivariate margins are more dependent or less dependent are the same from models (1a), (1bi), (1bii), (1biii). They show that the dependence for consecutive years are slightly stronger; this is consistent with the *gamma* measures in Table 5.16 for the initial data analysis. The dependence parameter estimates for model (1c) reveal that this model leads to a domination of dependence by the overall dependence ($\log\tilde{\theta} = 1.808$ with SE=0.073), which is close to assume a permutation symmetric copula. For comparison, for the model (1d) with a permutation symmetric copula, the dependence parameter estimate is $\log\tilde{\theta} = 1.700$ with SE= 0.111. From the above comparisons, it seems that the model (1a) is an adequate and better model for this data set. Thus in the following, we will concentrate on comparing the two submodels l.md.g.wn, l.md.a.wc of model (1a).

Table 5.19 suggests that the submodel l.md.a.wc is a better model than the submodel l.md.g.wn; it also indicates that there is a justifiable AR(1) latent dependence structure, which describes the data set better than a general or exchangeable dependence structure. The exchangeable dependence structure assumption would be the least acceptable hypothesis. To compare the submodel l.md.a.wc and l.md.g.wn, Table 5.22 lists the values of $X^2_{(a)}$ for different values of a ($a = 1, 2, \ldots, 10$). This table reveals that the submodel l.md.a.wc fits the response vectors with higher frequency (≥ 5) better, while the submodel l.md.g.wn fits the response vectors with lower frequency (≤ 4) better. In other words, neither submodel is clearly better. Different models capture the data set equally well in certain way; and they may be together useful to reveal the features of the data set and lead to some useful interpretations.

As a complement to the X^2 values for assessment of goodness-of fit, Table 5.23 contains estimates of probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y})$ for all possible $\mathbf{y}$ and the corresponding frequencies from the submodels l.md.g.wn and l.md.a.wc of model (1a), and these $\Pr(\mathbf{Y} = \mathbf{y})$ are estimated with $\sum_{i=1}^{268} \tilde{\Pr}(\mathbf{Y} = \mathbf{y}|x_i)/268$. Table 5.24 contains estimates of frequencies and probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y}|x)$ for various $\mathbf{y}$ at $x = 1$ (distance bigger than 5 miles) and at $x = 0$ (distance less than 5 miles) from submodels l.md.g.wn and l.md.a.wc of the model (1a). In Table 5.24, to save space, for each line, only the maximum estimated SE over models is given; actually the SEs are

quite close to each other. Table 5.23 and Table 5.24 suggest that the submodels l.md.g.wn and l.md.a.wc are both adequate for predictive purposes, since the prediction probabilities are comparable when the SEs are considered. These two submodels can be used to complement each other for slightly different inference purposes. The largest divergence in estimated probabilities with observed frequency occurs when $x = 1$ for $\mathbf{y} = (1, 1, 1, 1)$. The submodel l.md.a.wc, with a AR(1) latent dependence structure with dependence parameters depend significantly on the covariate, indicates that not only the dependence for consecutive years are larger significantly, but the strength of dependences also differ for those who live within 5 miles from those who live between 5 and 10 miles of the plant.

The analysis (e.g. from submodels l.md.g.wn as well as l.md.a.wc) indicates that, comparing the stress levels of the mothers living less 5 miles from the plant, there is a slight trend over time towards lower stress level for mothers living between 5 and 10 miles from the plant. There are no large changes of stress levels over time, but the stress levels of mothers living between 5 and 10 miles from the plant are a bit higher in the first year following the accident; they decrease in the second year and remain stable over the subsequent years. If we study the model with covariate (distance) for the dependence (e.g. the submodel l.md.a.wc), we see that living far from the plant has a negative effect on the dependence; it indicates that the dependence parameters are larger for those who live within 5 miles from the plant. This means that the mothers living within 5 miles from the plant are in probability more consistent over time in the original 90-item checklist; there could be a number of reasons for this. We can interpret this as the stress symptoms caused by the accident being more persistent over time for the group living closer to the plant. The analysis indicates that the dependence for consecutive years are larger and the dependence (pairwise) are all significant. The rate of a persistent high stress level is higher for mothers living closer to the plant. This can be seen from Table 5.24 (e.g. with l.md.a.wc submodel), where for example for $\mathbf{y} = (3, 3, 3, 3)$, $\tilde{P}(\mathbf{y}|x = 0) = 0.076 > 0.037 = \tilde{P}(\mathbf{y}|x = 1)$. The rate for a persistent medium stress level ($\mathbf{y} = (2, 2, 2, 2)$) is slightly higher for the group living closer to the plant, while the rates of persistent low stress level ($\mathbf{y} = (1, 1, 1, 1)$) is comparable for the two groups.

Similar results to the partial interpretations given above are also obtained in Fienberg *et al.* (1985) and Conaway (1989). They conclude that mothers within the five mile radius were in fact experiencing greater stress symptom than mothers living between 5 to 10 miles away; this is consistent with our observations.

Table 5.14: TMI Accident Study: Stress levels for 4 years following accident at TMI. Responses with non zero frequencies.

ID	Response pattern	Distance < 5 mi.	> 5 mi.
1	3 3 3 3	12	7
2	3 3 3 2	5	2
3	3 3 2 3	0	2
4	3 3 2 2	2	7
5	3 3 2 1	1	1
6	3 2 3 3	4	0
7	3 2 3 2	1	0
8	3 2 2 3	3	0
9	3 2 2 2	4	13
10	3 2 2 1	0	1
11	3 1 1 3	0	1
12	2 3 3 3	1	1
13	2 3 3 2	1	3
14	2 3 2 3	0	1
15	2 3 2 2	2	1
16	2 2 3 3	3	1
17	2 2 3 2	2	5
18	2 2 2 3	4	6
19	2 2 2 2	38	53
20	2 2 2 1	2	6
21	2 2 1 2	2	2
22	2 2 1 1	3	2
23	2 1 2 3	0	1
24	2 1 2 2	4	15
25	2 1 2 1	1	5
26	2 1 1 2	1	4
27	2 1 1 1	5	4
28	1 2 2 2	4	3
29	1 2 2 1	2	0
30	1 2 1 2	1	0
31	1 2 1 1	0	1
32	1 1 2 2	3	0
33	1 1 2 1	2	2
34	1 1 1 2	0	2
35	1 1 1 1	2	1
Total		115	153

Table 5.15: TMI Accident Study: Univariate marginal (and relative) frequencies.

	margin			
Outcomes	1979	1980	1981	1982
		< 5 mi.		
3	32 (0.278)	24 (0.209)	29 (0.252)	27 (0.235)
2	69 (0.600)	73 (0.635)	72 (0.626)	70 (0.609)
1	14 (0.122)	18 (0.157)	14 (0.122)	18 (0.157)
		> 5 mi.		
3	34 (0.222)	25 (0.163)	19 (0.124)	20 (0.131)
2	110 (0.719)	93 (0.608)	117 (0.765)	110 (0.719)
1	9 (0.059)	35 (0.229)	17 (0.111)	23 (0.150)
		all		
3	66 (0.246)	49 (0.183)	48 (0.179)	47 (0.175)
2	179 (0.668)	166 (0.619)	189 (0.705)	180 (0.672)
1	23 (0.086)	53 (0.198)	31 (0.116)	41 (0.153)

Table 5.16: TMI Accident Study: Pairwise *gamma* measures for Year 1979, 1980, 1981, 1982

Pair	< 5 mi.	> 5 mi.	all
(1979, 1980)	0.894	0.829	0.852
(1979, 1981)	0.831	0.635	0.758
(1979, 1982)	0.782	0.595	0.702
(1980, 1981)	0.907	0.882	0.887
(1980, 1982)	0.756	0.638	0.700
(1981, 1982)	0.924	0.738	0.851

Table 5.17: TMI Accident Study: Estimates of univariate marginal regression parameters for multivariate logit models

margin	$\gamma_j(1)$ (SE)	$\gamma_j(2)$ (SE)	distance (SE)
	differ across the margins		
1	-2.376 (0.304)	1.109 (0.227)	0.017 (0.272)
2	-1.629 (0.215)	1.291 (0.203)	0.384 (0.250)
3	-2.349 (0.299)	1.250 (0.234)	0.497 (0.287)
4	-1.938 (0.263)	1.343 (0.232)	0.368 (0.273)
	common across the margins		
	-1.984 (0.131)	1.250 (0.112)	0.315 (0.135)

Table 5.18: TMI Accident Study: Estimates of dependence regression parameters for multivariate logit model with multinormal copula

margin	intercept (SE)	distance (SE)
general dependence, with covariate		
12	1.960 (0.289)	-0.240 (0.422)
13	1.594 (0.250)	-0.470 (0.391)
14	1.430 (0.304)	-0.386 (0.414)
23	2.079 (0.310)	-0.092 (0.424)
24	1.428 (0.321)	-0.271 (0.425)
34	2.358 (0.363)	-0.960 (0.505)
general dependence, without covariate		
12	1.824 (0.212)	
13	1.356 (0.192)	
14	1.243 (0.195)	
23	2.032 (0.219)	
24	1.277 (0.205)	
34	1.779 (0.273)	
exchangeable dependence, with covariate		
	1.772 (0.123)	-0.377 (0.174)
exchangeable dependence, without covariate		
	1.546 (0.086)	
AR(1) dependence, with covariate		
	2.208 (0.124)	-0.385 (0.178)
AR(1) dependence, without covariate		
	2.008 (0.088)	

Table 5.19: TMI Accident Study: Comparisons of AIC values and $X^2_{(2)}$ values from various submodels of models (1a) and (2)

Logit Models	AIC	$X^2_{(2)}$	Probit Models	AIC	$X^2_{(2)}$
l.md.g.wc	1542.443	28.018	p.md.g.wc	1542.788	27.975
l.md.g.wn	1537.235	29.786	p.md.g.wn	1537.499	29.600
l.md.e.wc	1549.740	98.795	p.md.e.wc	1549.977	97.138
l.md.e.wn	1550.219	117.144	p.md.e.wn	1550.403	115.501
l.md.a.wc	1534.116	26.547	p.md.a.wc	1534.388	26.594
l.md.a.wn	1535.942	28.551	p.md.a.wn	1536.168	28.479
l.mc.g.wc	1568.305	83.795	p.mc.g.wc	1568.351	85.011
l.mc.g.wn	1563.332	86.916	p.mc.g.wn	1563.416	88.284
l.mc.e.wc	1568.103	197.711	p.mc.e.wc	1568.324	204.103
l.mc.e.wn	1570.950	217.573	p.mc.e.wn	1571.111	226.992
l.mc.a.wc	1557.114	77.743	p.mc.a.wc	1557.288	78.831
l.mc.a.wn	1562.229	80.869	p.mc.a.wn	1562.378	81.990

Table 5.20: TMI Accident Study: Comparisons of AIC values and $X^2_{(2)}$ values from the submodel l.md.g.wn of various models

Models	AIC	$X^2_{(2)}$
(1a)	1537.235	29.786
(1bi)	–	–
(1bii)	1540.846	23.485
(1biii)	1542.312	33.303
(1c)	1566.475	970.756
(1d)	1553.640	422.000

Table 5.21: TMI Accident Study: Estimates (SE) of dependence regression parameters from the submodel l.md.g.wn of various models

margin	(1bi)	(1bii)	(1biii)	(1c)
12	1.824 (0.212)	2.697 (0.289)	1.960 (0.158)	-8.250 (0.172)
13	1.356 (0.192)	2.035 (0.262)	1.628 (0.171)	-8.996 (0.010)
14	1.243 (0.195)	1.928 (0.273)	1.485 (0.184)	-8.397 (0.001)
23	2.032 (0.219)	2.857 (0.289)	2.122 (0.150)	-7.495 (0.202)
24	1.277 (0.205)	2.014 (0.271)	1.495 (0.185)	0.927 (1.910)
34	1.779 (0.273)	2.710 (0.290)	1.978 (0.190)	1.084 (1.827)
$\log(\theta)$				1.808 (0.073)

Table 5.22: TMI Accident Study: Comparisons of $X^2_{(a)}$ values from the submodels l.md.g.wn and l.md.a.wc of model (1a)

a	l.md.g.wn $X^2_{(a)}$	l.md.a.wc $X^2_{(a)}$
1	623.507	4390.084
2	29.786	26.547
3	15.809	16.243
4	10.339	11.766
5	9.086	7.453
6	8.217	7.240
7	7.583	6.798
8	4.625	3.575
9	2.983	2.481
10	2.467	2.310

Table 5.23: TMI Accident Study: Estimates of $\Pr(\mathbf{Y} = \mathbf{y})$ and frequencies from the submodels l.md.g.wn and l.md.a.wc of model (1a)

Response pattern	Observed numbers	l.md.g.wn Expected numbers	l.md.a.wc Expected numbers	Observed prob.	l.md.g.wn Expected prob.	l.md.a.wc Expected prob.
3 3 3 3	19	14.1	14.4	0.071	0.053	0.054
3 3 3 2	7	7.3	7.5	0.026	0.027	0.028
3 3 2 3	2	3.0	2.5	0.007	0.011	0.009
3 3 2 2	9	8.3	9.8	0.034	0.031	0.037
3 3 2 1	2	0.2	0.4	0.007	0.001	0.002
3 2 3 3	4	3.7	2.9	0.015	0.014	0.011
3 2 3 2	1	2.6	2.7	0.004	0.010	0.010
3 2 2 3	3	5.3	3.0	0.011	0.020	0.011
3 2 2 2	17	19.2	19.5	0.063	0.072	0.073
3 2 2 1	1	1.0	2.0	0.004	0.004	0.007
3 1 1 3	1	0.0	0.0	0.004	0.000	0.000
2 3 3 3	2	3.7	4.5	0.007	0.014	0.017
2 3 3 2	4	4.3	2.9	0.015	0.016	0.011
2 3 2 3	1	1.1	1.3	0.004	0.004	0.005
2 3 2 2	3	6.3	5.4	0.011	0.023	0.020
2 2 3 3	4	5.1	6.7	0.015	0.019	0.025
2 2 3 2	7	7.0	6.3	0.026	0.026	0.024
2 2 2 3	10	9.9	10.4	0.037	0.037	0.039
2 2 2 2	91	86.0	87.4	0.340	0.321	0.326
2 2 2 1	8	12.5	11.6	0.030	0.047	0.043
2 2 1 2	4	3.4	3.2	0.015	0.013	0.012
2 2 1 1	5	3.4	4.1	0.019	0.013	0.015
2 1 2 3	1	0.9	0.8	0.004	0.003	0.003
2 1 2 2	19	17.1	16.6	0.071	0.064	0.062
2 1 2 1	6	4.3	4.6	0.022	0.016	0.017
2 1 1 2	5	6.0	4.7	0.019	0.022	0.018
2 1 1 1	9	7.2	8.1	0.034	0.027	0.030
1 2 2 2	7	3.7	3.6	0.026	0.014	0.014
1 2 2 1	2	1.4	0.6	0.007	0.005	0.002
1 2 1 2	1	0.2	0.2	0.004	0.001	0.001
1 2 1 1	1	0.5	0.3	0.004	0.002	0.001
1 1 2 2	3	4.8	6.3	0.011	0.018	0.024
1 1 2 1	4	2.5	1.9	0.015	0.009	0.007
1 1 1 2	2	2.5	2.9	0.007	0.009	0.011
1 1 1 1	3	6.7	6.3	0.011	0.025	0.023
others	0	2.8	2.6	0.000	0.009	0.008

Table 5.24: TMI Accident Study: Estimates of $\Pr(\mathbf{Y} = \mathbf{y}|x)$ and frequencies from the submodels l.md.g.wn and l.md.a.wc of model (1a)

Response pattern	Observed numbers	l.md.g.wn Expected numbers	l.md.a.wc Expected numbers	maxSE	Observed prob.	l.md.g.wn Expected prob.	l.md.a.wc Expected prob.	maxSE
< 5 miles								
3 3 3 3	12	7.5	8.7	2.4	0.104	0.065	0.076	0.021
3 3 3 2	5	3.6	3.8	1.2	0.043	0.031	0.033	0.010
3 3 2 2	2	3.4	4.0	1.0	0.017	0.030	0.035	0.009
3 3 2 1	1	0.1	0.1	0.1	0.009	0.001	0.001	0.001
3 2 3 3	4	1.7	1.4	0.7	0.035	0.015	0.012	0.006
3 2 3 2	1	1.2	1.0	0.5	0.009	0.010	0.009	0.004
3 2 2 3	3	2.1	1.0	0.8	0.026	0.018	0.009	0.007
3 2 2 2	4	6.9	6.8	1.5	0.035	0.060	0.059	0.013
2 3 3 3	1	2.3	2.6	0.7	0.009	0.020	0.023	0.006
2 3 3 2	1	2.5	1.5	0.9	0.009	0.022	0.013	0.008
2 3 2 2	2	3.2	2.4	0.9	0.017	0.028	0.021	0.008
2 2 3 3	3	2.9	3.4	0.9	0.026	0.025	0.030	0.008
2 2 3 2	2	3.8	3.0	1.2	0.017	0.033	0.026	0.010
2 2 2 3	4	4.7	4.5	1.3	0.035	0.041	0.039	0.011
2 2 2 2	38	37.3	40.6	3.8	0.330	0.324	0.353	0.033
2 2 2 1	2	4.8	4.4	1.3	0.017	0.042	0.038	0.011
2 2 1 2	2	1.2	0.8	0.6	0.017	0.010	0.007	0.005
2 2 1 1	3	1.2	1.3	0.6	0.026	0.010	0.011	0.005
2 1 2 2	4	6.3	5.8	1.5	0.035	0.055	0.050	0.013
2 1 2 1	1	1.5	1.6	0.6	0.009	0.013	0.014	0.005
2 1 1 2	1	1.8	1.3	0.7	0.009	0.016	0.011	0.006
2 1 1 1	5	2.1	2.5	0.8	0.043	0.018	0.022	0.007
1 2 2 2	4	2.0	1.6	0.8	0.035	0.017	0.014	0.007
1 2 2 1	2	0.7	0.2	0.3	0.017	0.006	0.002	0.003
1 2 1 2	1	0.1	0.1	0.1	0.009	0.001	0.001	0.001
1 1 2 2	3	2.2	2.8	0.8	0.026	0.019	0.024	0.007
1 1 2 1	2	1.0	0.8	0.5	0.017	0.009	0.007	0.004
1 1 1 1	2	2.3	2.8	0.9	0.017	0.020	0.024	0.008
others	0	4.5	3.9	–	0.000	0.039	0.034	–
> 5 miles								
3 3 3 3	7	6.6	5.7	1.7	0.046	0.043	0.037	0.011
3 3 3 2	2	3.7	3.7	0.9	0.013	0.024	0.024	0.006
3 3 2 3	2	1.7	1.4	0.6	0.013	0.011	0.009	0.004
3 3 2 2	7	4.7	5.8	1.4	0.046	0.031	0.038	0.009
3 3 2 1	1	0.2	0.3	0.2	0.007	0.001	0.002	0.001
3 2 2 2	13	12.4	12.7	2.1	0.085	0.081	0.083	0.014
3 2 2 1	1	0.8	1.5	0.5	0.007	0.005	0.010	0.003
3 1 1 3	1	0.0	0.0	0.0	0.007	0.000	0.000	0.000
2 3 3 3	1	1.4	1.8	0.5	0.007	0.009	0.012	0.003
2 3 3 2	3	1.8	1.4	0.8	0.020	0.012	0.009	0.005
2 3 2 3	1	0.5	0.8	0.3	0.007	0.003	0.005	0.002
2 3 2 2	1	3.1	3.1	0.9	0.007	0.020	0.020	0.006
2 2 3 3	1	2.3	3.2	0.8	0.007	0.015	0.021	0.005
2 2 3 2	5	3.4	3.4	1.1	0.033	0.022	0.022	0.007
2 2 2 3	6	5.0	5.8	1.4	0.039	0.033	0.038	0.009
2 2 2 2	53	48.7	46.8	4.9	0.346	0.318	0.306	0.032
2 2 2 1	6	7.7	7.2	1.7	0.039	0.050	0.047	0.011
2 2 1 2	2	2.1	2.3	0.9	0.013	0.014	0.015	0.006
2 2 1 1	2	2.3	2.9	0.9	0.013	0.015	0.019	0.006
2 1 2 3	1	0.6	0.6	0.3	0.007	0.004	0.004	0.002
2 1 2 2	15	10.9	10.9	2.3	0.098	0.071	0.071	0.015
2 1 2 1	5	2.9	3.1	0.9	0.033	0.019	0.020	0.006
2 1 1 2	4	4.1	3.5	1.4	0.026	0.027	0.023	0.009
2 1 1 1	4	5.2	5.5	1.2	0.026	0.034	0.036	0.008
1 2 2 2	3	1.7	2.0	0.8	0.020	0.011	0.013	0.005
1 2 1 1	1	0.3	0.2	0.2	0.007	0.002	0.001	0.001
1 1 2 1	2	1.5	1.1	0.6	0.013	0.010	0.007	0.004
1 1 1 2	2	1.5	1.8	0.6	0.013	0.010	0.012	0.004
1 1 1 1	1	4.4	3.5	1.1	0.007	0.029	0.023	0.007
others	0	11.6	11.2	–	0.000	0.076	0.073	–

5.2.3 Example with multivariate count response data

In this subsection, several models are applied to a data set of trivariate counts of pathogenic bacteria at 50 different sterile locations measured by three different air samplers. Aitchison and Ho (1989) studied this data set. One of the objectives of the study is to investigate the relative effectiveness of three different air samplers to detect pathogenic bacteria.

The response vectors are 3-dimensional count measures. Table 5.25 lists the count measures of the three samplers from the 50 locations. The table shows that there are no duplicate trivariate response observations. The frequencies by univariate margin (or by sampler) given in Table 5.26 indicate that sampler 3 is more variable than sampler 1 and 2. The pairwise *gamma* measures in Table 5.27 indicate that the samplers 1 and 3 and the samplers 2 and 3 are negatively associated. Summary statistics (means, variance, quartiles, maximum, minimum and pairwise correlations) are given in Table 5.28. This initial data analysis indicates that there is some extra-Poisson variation as the variance to mean ratio for the margins (or sampler) range from 2 to 5, with the sampler 3 more variable than the other two samplers.

This initial data analysis suggests that the MCD models with Poisson variation may not be suitable, but for illustrative purposes, we applied the multivariate count model with Poisson variation as well as multivariate count model with extra Poisson variation. The multivariate count response models that were used to model this trivariate count data are:

1. The multivariate Poisson model with multinormal copula (3.1).

2. The multivariate Poisson-lognormal model in section 3.6.

This data set has no covariates, we thus directly estimate the parameters in the multivariate Poisson model with multinormal copula (see section 3.3) and the multivariate Poisson-lognormal model (see section 3.6). The multivariate Poisson model has Poisson marginals. The univariate parameter λ_j $(j = 1, 2, 3)$ in the multivariate Poisson model is reparameterized by taking log-transformation, $\eta_j = \log(\lambda_j)$, such that the new parameter η_j has the range $(-\infty, \infty)$. For the dependence parameters θ_{jk} in the multinormal copula, we let $\theta_{jk} = [\exp(\beta_{jk}) - 1]/[\exp(\beta_{jk}) + 1]$ such that β_{jk} has the range $(-\infty, \infty)$. We proceed to estimate η_j and β_{jk}. For the multivariate Poisson-lognormal model, the marginal parameters are $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$. The dependence parameters are θ_{jk} which are similar to those of the multivariate Poisson model with multinormal copula.

For the univariate marginal parameters, at least two situations could be considered: parameters

differ across margins, denoted by "md", and parameters common across margins, denoted by "mc". For the dependence parameters, for both models (1) and (2), we consider the general (denoted by "g") and exchangeable (denoted by "e") dependence structures. Thus 4 submodels for both model (1) and (2) are considered: md.g, md.e, mc.g, mc.e.

For all the models, the IFM estimation theory is applied. That is, the univariate (regression) parameters are estimated from separate univariate likelihoods, and bivariate and multivariate (regression) parameters are estimated from bivariate likelihoods, with univariate parameters fixed as estimated from the separate univariate likelihoods. For the situation of "mc" for common marginal parameters and exchangeable dependence structure, WA of (2.93) in section 2.6 for parameter estimation based on IFM is used. For standard errors (SEs) of parameter estimates and prediction probabilities, the (delete-one) jackknife method from Chapter 2 is used. These are all similar to the use of these models in subsection 5.2.1 and 5.2.2.

Summaries of the modelling are given in several tables. Table 5.29 contains the estimates and SEs of the regression parameters for the marginal parameters with the multivariate Poisson model. Table 5.30 contains the estimates and SEs of the dependence parameters for multivariate Poisson model with multinormal copula. Table 5.31 contains AIC values for all the submodels of multivariate Poisson models multinormal copula. The $X^2_{(a)}$ for $a \geq 2$ are not available since all the 3-tuples in the data set have frequency 1. $X^2_{(1)}$ is very large since many estimated probabilities for the 3-tuples of frequency 1 are very close to zero. In this situation, because of the frequency 1 occurrence for all 3-tuples, the X^2 measure as well as the estimated probabilities of the form $\Pr(\mathbf{Y} = \mathbf{y})$ may not be suitable measures for a rough assessment of the goodness-of-fit. Instead, residual measures such as (5.6) should be considered if feasible. Other rough goodness-of-fit checks may consist of comparing some empirical statistics (means, variances, correlations, etc.) with the counterparts estimated from the fitted model. The latter approach would rule out all submodels of model (1) for the goodness-of-fit of this data set, since the extra-Poisson variation demonstrated by the empirical statistics are not matched by the model (1). For the residual checking based on (5.6), we give an illustration here with the submodel md.g. We first compute $\tilde{e}_{i3} = y_{i3} - \mathrm{E}[Y_{i3}|Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \tilde{\boldsymbol{\lambda}}]$, where $\mathrm{E}[Y_{i3}|Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \tilde{\boldsymbol{\lambda}}] = \sum_{y=1}^{\infty} y * P(y_{i1}y_{i2}y)/P(y_{i1}y_{i2})$. We then plot $\tilde{e}_{i3}$ versus y_{i1} and $\tilde{e}_{i3}$ versus y_{i2} for all $i = 1, \ldots, 50$. The model would be considered as adequate based on residual plots if the residuals are small and do not exhibit systematic patterns. The two plots in Figure 5.1 do not show evident systematic patterns (except for a few outliers), but almost all the residuals are quite large judging from the observed values of y_{i3}; it indicates that the models do not fit the data well

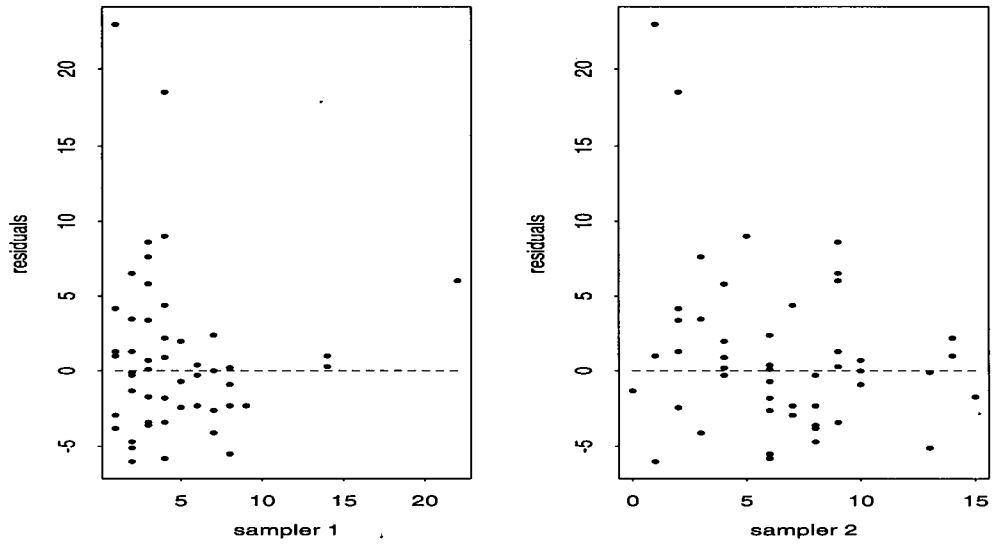

Figure 5.1: Bacteria Counts: Residuals from the submodel md.g of model (1).

enough. This is expected since the multivariate Poisson models only fit data with Poisson variation.

Next we consider fitting the multivariate Poisson-lognormal model to the data. From IFM estimation theory, we first estimate the univariate marginal parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$ from separate univariate likelihoods, and then the dependence parameters from bivariate likelihoods, with univariate parameters fixed as estimated from the separate univariate likelihoods. For the multivariate Poisson-lognormal model, several hypotheses are possible: $\mu_j = \mu$, $\sigma_j = \sigma$, $\theta_{jk} = \theta$; μ_j margin-dependent; or plus σ_j margin-dependent; or plus θ_{jk} margin-dependent, etc.. Similarly to the model (1), we consider the four submodels, md.g, md.e, mc.g and mc.e of the multivariate Poisson-lognormal model. For standard errors (SEs) of parameter estimates and prediction probabilities, the (delete-one) jackknife method from Chapter 2 is used.

Summaries of the model fits with the multivariate Poisson-lognormal model are given in several tables. Table 5.32 contains the estimates and SEs of the marginal parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \mu_3)$ and $\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \sigma_3)$. Table 5.33 contains the estimates and SEs of the dependence parameters θ_{12}, θ_{13} and θ_{23}. Table 5.34 contains AIC values for all the submodels of multivariate Poisson-lognormal model. From examining the AIC values for the 4 submodels, the models md.g and md.e seem to be better choices. For a rough assessment of the goodness-of-fit, Table 5.35 contains the estimated means, variances and correlations based on the fitted model md.g; they are quite close to the empirical means, variances, correlations in Table 5.28. This implies that the multivariate Poisson-lognormal model may be considered as a more appropriate model for this data set. To further check the goodness-of-fit

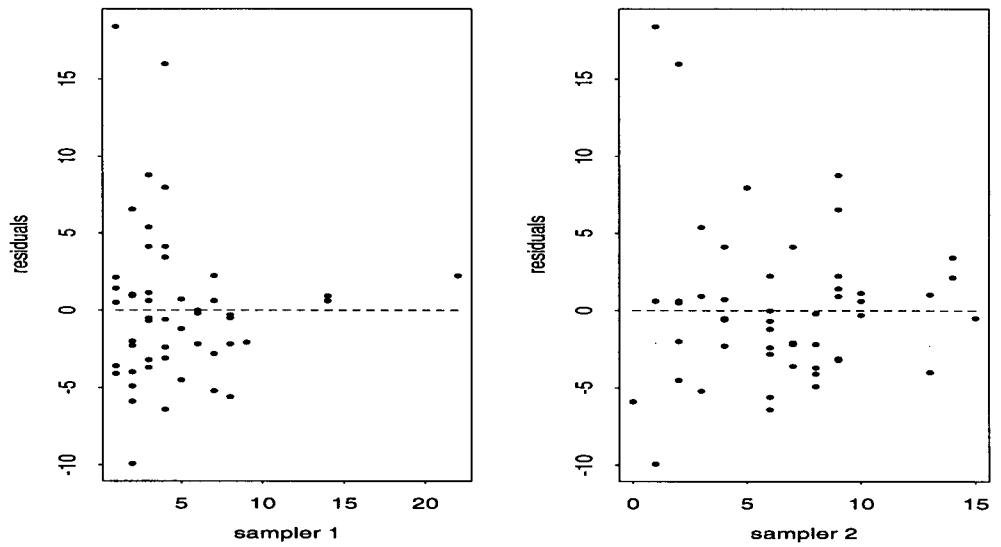

Figure 5.2: Bacteria Counts: Residuals from the submodel md.g of trivariate Poisson lognormal model.

of the fitted model, we compute residual measures of the form $\tilde{e}_{i3} = y_{i3} - \mathrm{E}[Y_{i3}|Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \tilde{\boldsymbol{\lambda}}]$, where $\mathrm{E}[Y_{i3}|Y_{i1} = y_{i1}, Y_{i2} = y_{i2}, \tilde{\boldsymbol{\lambda}}] = \sum_{y=1}^{\infty} y * P(y_{i1}y_{i2}y)/P(y_{i1}y_{i2})$. We then plot $\tilde{e}_{i3}$ versus y_{i1} and $\tilde{e}_{i3}$ versus y_{i2} for all $i = 1, \ldots, 50$. The two plots in Figure 5.2 are residual plots from the submodel md.g; they do not show evident systematic patterns except for a few outliers. The magnitude of the residuals are smaller than those obtained with the multivariate Poisson models. In these circumstances, we would further study the fitted trivariate Poisson-lognormal submodels md.g and md.e.

The analysis (e.g. from submodels md.g as well as md.e) indicates that sampler 3 tend to be negatively correlated with sampler 2 and sampler 1 (based on the submodel md.g), and with significantly negative correlation if we based our interpretation on the submodel md.e. The samplers appear to have been competing in some way for the capture of bacteria; this competing behaviour is particular evident when we compare sampler 3 with sampler 2 or sampler 1. It would be interesting to see if there is any practical reason explaining this observation. From the estimates, sampler 2 seems to be the most effective sampler, and sampler 1 the least. Similar results are also obtained in Aitchison and Ho (1989). For a particular model with the assumption of equal σ (that is $\sigma_1 = \sigma_2 = \sigma_3 = \sigma$) and exchangeable correlation, the estimates in Aitchison and Ho (1989) are $\hat{\mu}_1 = 1.39\ (0.11)$, $\hat{\mu}_2 = 1.75\ (0.10)$, $\hat{\mu}_3 = 1.70\ (0.10)$, $\hat{\sigma} = 0.56\ (0.05)$, and $\hat{\theta} = -0.28\ (0.10)$ which are quite close to our estimates (based on the IFM approach) $\tilde{\mu}_1 = 1.388\ (0.098)$, $\tilde{\mu}_2 = 1.784\ (0.098)$, $\tilde{\mu}_3 = 1.660\ (0.120)$,

Table 5.25: Bacteria Counts: Bacterial counts by 3 samplers in 50 sterile locations

y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3	y_1	y_2	y_3
1	2	11	3	6	6	3	8	2	7	10	5	22	9	6
8	6	0	3	9	14	1	1	30	2	2	8	5	2	4
2	13	5	4	2	25	4	5	15	3	15	3	2	0	6
2	8	1	9	7	3	7	6	3	1	8	2	2	1	1
5	6	5	5	4	8	8	10	4	4	6	0	4	6	4
14	1	7	4	4	7	3	2	10	8	7	3	4	9	2
3	9	2	7	3	2	6	8	5	6	6	6	8	4	6
7	6	8	1	14	6	2	3	10	4	14	7	3	10	6
3	4	12	2	13	0	1	7	3	3	3	14	4	7	10
1	9	7	14	9	5	2	9	12	6	8	3	2	4	6

Table 5.26: Bacteria Counts: Univariate marginal frequencies

margin																		
1	count	1	2	3	4	5	6	7	8	9	14	22						
	freq.	6	9	9	8	3	3	4	4	1	2	1						
2	count	0	1	2	3	4	5	6	7	8	9	10	13	14	15			
	freq.	1	3	5	3	5	1	8	4	5	7	3	2	2	1			
3	count	0	1	2	3	4	5	6	7	8	10	11	12	14	15	25	30	
	freq.	3	2	5	6	3	5	8	4	3	3	1	2	2	1	1	1	

$\tilde{\sigma} = 0.523$ (0.062), and $\tilde{\theta} = -0.347$ (0.133).

The IFM approach should be considered as a better approach than ML approach with this data set since the sample size is small and all 3-tuple responses have a frequency of 1. It is better not only in the sense of efficiency, but in the sense that the IFM approach may lead to more reliable estimates. The ML approach may be severely affected by the small sample size and thus lower frequencies of response outcome. In contrast, the inference from IFM approach should be more reliable, since IFM estimation method is robust marginally, thus the estimated parameters are still able to capture the general feature of the data even the full response frequencies are low.

Table 5.27: Bacteria Counts: Pairwise *gamma* measures for samplers 1, 2, 3

Pair	*gamma*
(1, 2)	0.038
(1, 3)	-0.125
(2, 3)	-0.261

Table 5.28: Bacteria Counts: Moment estimate of means, variances, correlations and other summary statistics of responses

margin	mean	variance	min	Q1	med	Q3	max	margin	correlation
1	4.7	15.07	1	2	4	6	22	1,2	0.0192
2	6.5	13.64	0	4	6	9	15	1,3	-0.1666
3	6.6	32.61	0	3	6	8	30	2,3	-0.3667

Table 5.29: Bacteria Counts: Estimates of marginal parameters for multivariate Poisson model

margin	η_j (SE)
differ across the margins	
1	1.469 (0.101)
2	1.872 (0.081)
3	1.746 (0.097)
common across the margins	
	1.723 (0.053)

Table 5.30: Bacteria Counts: Estimates of dependence regression parameters for multivariate Poisson model with multinormal copula

margin	β_{jk} (SE)
general dependence	
12	0.009 (1.114)
13	-0.153 (1.336)
23	-0.334 (0.207)
exchangeable dependence	
	-0.319 (0.201)

Table 5.31: Bacteria Counts: Comparisons of AIC values from various submodels of multivariate Poisson model with multinormal copula

Models	AIC
md.g	787.924
md.e	785.600
mc.g	806.351
mc.e	801.657

Table 5.32: Bacteria Counts: Estimates of marginal parameters from multivariate Poisson-lognormal model

margin	$\tilde{\mu}$ (SE)	$\tilde{\sigma}$ (SE)
differ across the margins		
1	1.388 (0.098)	0.551 (0.122)
2	1.784 (0.098)	0.425 (0.090)
3	1.660 (0.120)	0.672 (0.121)
common across the margins		
	1.604 (0.060)	0.523 (0.062)

Table 5.33: Bacteria Counts: Estimates of dependence parameters from multivariate Poisson-lognormal model

margin	θ_{jk} (SE)
general dependence	
12	0.059 (0.315)
13	-0.260 (0.208)
23	-0.605 (0.206)
exchangeable dependence	
	-0.347 (0.133)

Table 5.34: Bacteria Counts: Comparisons of AIC values from various submodels of multivariate Poisson-lognormal model

Models	AIC
md.g	813.898
md.e	813.546
mc.g	818.575
mc.e	815.568

Table 5.35: Bacteria Counts: Estimates of means, variances and correlations of responses from the submodel md.g of multivariate Poisson-lognormal model

margin	mean	variance	margin	correlation
1	4.66	12.38	12	0.031
2	6.52	14.92	13	-0.143
3	6.59	31.39	23	-0.315

5.3 Summary

In this chapter, we studied some issues on modelling, illustrating the data analysis cycle, model selection and diagnostic checking. We also provided three detailed data analysis examples with the models developed in Chapter 3. The fact that there are many possibilities with the multivariate modelling is highlighted by the examples. AIC is one way to choose any model. In addtion, models are compared by their predictibility. The sensitivity analysis from comparison of models is important. If inference and prediction are similar, then one does not have to worry so much about the validity of SEs, etc., after choosing a model from AIC or another criterion. One finding from our data analysis is that there is insensitivity to multivariate models that have similar qualitative dependence structure and similar form of univariate margins.

The models that we used to model the data are just some of the available models. Other examples of models include MMD binary models for Six Cities Study and MCD model with univariate Poisson-lognormal margins for the bacteria counts data. There is still a lot to be investigated in terms of the applications of different type of models to the real data. Some of them will be studied in the further research.

Chapter 6

GEE methodology and its comparison with ML and IFM approaches

Partly because of a lack of suitable multivariate non-normal distributions, and the problems of mathematical tractibility and computational difficulties in statistical inference with the multivariate models, Liang and Zeger (1986), Zeger and Liang (1986) and others have developed the generalized estimating equations (GEE) approach with a partly specified probability model (some moment characteristics of the distribution are specified but not the joint distribution), for the estimation of regression parameters. It is claimed, as one of the advantages of GEE, that the GEE approach requires no further distributional assumptions other than that the observations across primary units (subjects) are independent. Under the correct specification of the marginal expectation for the outcome, consistent and asymptotically normal estimators of regression coefficients can be obtained, with or without the correct specification of the dependence among the response variables. However the GEE approach has several disadvantages, including (i) limited types of inferences that can be made, (ii) incompleteness of the data analysis cycle of initial data analysis, statistical modelling, diagnostics and inference, (iii) lack of clear accompanying means of assessing the implicit assumptions, and (iv) possible interpretability problems. In fact, the main inferences provided by the GEE method are for the regression coefficients. Furthermore, GEE only deals with the case of univariate margins in the generalized linear model class, and extensions to quasi-likelihood models (and this

does not include ordinal response data). The extension of GEE by Zhao and Prentice (1990) allows for estimating equations for correlations. This then has some analogy to the IFM approach when the model is MUBE.

The models of the earlier chapters provide a framework for evaluation of the GEE and IFM approaches. Both methods can be sensitive to incorrect marginal model specifications but the IFM approach has much greater flexibility for a sensitivity analysis.

In this chapter, we discuss the drawbacks of GEE in the multivariate analysis framework and examine the efficiency of the GEE approach. In section 6.1, we briefly introduce the GEE approach. In section 6.2, we discuss problems with the GEE approach when it is considered within a multivariate analysis framework. In section 6.3, we carry out some comparison studies of GEE with the ML approach and the IFM approach to better understand how efficient, or inefficient, the GEE estimates can be. These are done analytically and computationally. Finally, in section 6.4, we propose a new likelihood-based computational strategy which combines the GEE and IFM approaches to achieve greater efficiency for marginal regression parameter estimation, when this is the only inference of interest.

6.1 Generalized estimating equations

In likelihood analysis, we must specify the actual form of the distribution. In quasi-likelihood analysis, one specifies only the relationships between the means of the outcomes and the covariates and the relationships between the mean and variance. A univariate quasi-likelihood function can be written as

$$Q(\mu; y) = \int_y^\mu (y - t)/\mathrm{Var}(Y)dt. \tag{6.1}$$

The estimating equations for some regression parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_q)'$ (these appear only in μ_i) based on the quasi-likelihood function (6.1) are

$$\sum_{i=1}^{n} \frac{\partial \mu_i}{\partial \boldsymbol{\beta}} \frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} = \mathbf{0},$$

where $\mu_i = \mathrm{E}(Y_i)$. The quasi-likelihood (6.1) can be identified with a loglikelihood corresponding to an exponential family distribution in the univariate situation. By adopting a quasi-likelihood approach and specifying only the mean-variance structure, such as $\mu = \mathrm{E}(Y)$ and $\mathrm{Var}(Y) = \phi V(\mu)$, where $V(\mu)$ is a specified function of μ (see McCullagh and Nelder 1989), the estimating equations are applicable to different types of variables (continuous and discrete), with no assumptions about

the distribution of the response. (Actually the form $\text{Var}(Y) = \phi V(\mu)$ is quite restrictive. We will discuss this in section 6.2.) For the d-dimensional multivariate response, if the responses are naively assumed to be independent and $\boldsymbol{\beta}$ is assumed to be common for different Y_{ij} or μ_{ij}, $j = 1, \ldots, d$, then the quasi-likelihood estimation equations become

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} \right)^T V^{-1}(\mathbf{Y}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})' = \text{E}(\mathbf{Y}_i)$, $V(\mathbf{Y}_i) = \text{diag}[\text{Var}(Y_{i1}), \ldots, \text{Var}(Y_{id})]$, and

$$\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'} = \begin{pmatrix} \frac{\partial \mu_1}{\partial \beta_1} & \cdots & \frac{\partial \mu_1}{\partial \beta_q} \\ \vdots & \ddots & \vdots \\ \frac{\partial \mu_d}{\partial \beta_1} & \cdots & \frac{\partial \mu_d}{\partial \beta_q} \end{pmatrix}.$$

To gain more efficiency in estimating these regression parameters of the univariate margins, Liang and Zeger (1986) and Zeger and Liang (1986) propose to estimate $\boldsymbol{\beta}$ from

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0}, \tag{6.2}$$

where $D_i = \partial \boldsymbol{\mu}_i / \partial \boldsymbol{\beta}'$. Here V_i is a "working" or approximate covariance matrix for $\mathbf{Y}_i$, chosen by the investigator. The "working" covariance can be expressed in the form:

$$V_i = A_i^{1/2} R_i(\boldsymbol{\alpha}) A_i^{1/2},$$

where $A_i = \text{diag}[\text{Var}(Y_{i1}), \ldots, \text{Var}(Y_{id})]$ and $R_i(\boldsymbol{\alpha}) = \text{Corr}(\mathbf{Y}_i)$. $R_i(\boldsymbol{\alpha})$ is termed a "working" correlation matrix, with $\boldsymbol{\alpha}$ representing a vector of parameters associated with a specified model for $\text{Corr}(\mathbf{Y}_i)$. Note that the correlation matrix can differ from subject to subject, but $R_i(\boldsymbol{\alpha})$ is fully specified by the vector of unknown parameters, $\boldsymbol{\alpha}$, which is usually the same for all subjects. $R_i(\boldsymbol{\alpha})$ is referred to as a "working" correlation matrix, as Liang and Zeger argued, because it need not to be correctly specified. The equations (6.2) are thus called generalized estimating equations, or GEE.

The extension in (6.2) made by Liang and Zeger is basically about the specification of the "working" correlation matrix. Liang and Zeger (1986) showed that as the sample size tends to infinity, the estimates of the regression coefficients obtained from the GEE approach are consistent and asymptotically normal, with an asymptotic variance-covariance matrix which can be consistently estimated even under misspecification of the dependence structure.

If $V_i = \text{Cov}(\mathbf{Y}_i)$ is correctly specified, then a consistent estimate of the asymptotic variance of $\hat{\boldsymbol{\beta}}$ is given by

$$\Sigma_1^{-1}(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \hat{D}_i^T \hat{V}_i^{-1} \hat{D}_i,$$

where $\hat{V}_i$ is V_i evaluated at $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}})$, and $\hat{D}_i$ is D_i evaluated at $\hat{\boldsymbol{\beta}}$, respectively. However, if the "working" correlation $R_i(\boldsymbol{\alpha})$ is misspecified, $\Sigma_1^{-1}(\hat{\boldsymbol{\beta}})$ can give inconsistent estimates. Liang and Zeger (1986) suggest using the following "robust" estimate:

$$\Sigma_1^{-1}(\hat{\boldsymbol{\beta}})\Sigma_2(\hat{\boldsymbol{\beta}})\Sigma_1^{-1}(\hat{\boldsymbol{\beta}}),$$

where

$$\Sigma_2(\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \hat{D}_i^T \hat{V}_i^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)(\mathbf{y}_i - \hat{\boldsymbol{\mu}}_i)' \hat{V}_i^{-1} \hat{D}_i.$$

This estimate is "robust" since it is consistent even if the "working" covariance V_i is not equal to $\mathrm{Cov}(\mathbf{Y}_i)$. An alternative approach, which we recommend, is to apply the jackknife method to (6.2) to obtain an estimate of $\Sigma(\hat{\boldsymbol{\beta}})$.

There are several choices for the "working" correlation matrix R_i. The simplest choice is to use $R_i = I$, where I is a identity matrix. This is equivalent to assume that the response variables are not linearly correlated. One can also assume $R_i(\boldsymbol{\alpha}) = R(\boldsymbol{\alpha})$ for all i, and let $R(\boldsymbol{\alpha})$ be fully unspecified. Two simple special cases of $R(\boldsymbol{\alpha})$ are an exchangeable correlation matrix where $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \alpha$ and an autoregressive correlation matrix where $\mathrm{Corr}(Y_{ij}, Y_{ik}) = \alpha^{|j-k|}$. If $\boldsymbol{\alpha}$ is known, (6.2) is sufficient alone for the estimation of $\boldsymbol{\beta}$. Otherwise, $\boldsymbol{\alpha}$ must be estimated. We discuss this next.

The "working" correlation matrix may be obtained through additional modelling. Prentice (1988) has considered extensions of the GEE in Zeger and Liang (1986) to explicitly estimate the covariances of the responses. He proposed to work with

$$\begin{cases} \displaystyle\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}\right)^T \mathrm{Cov}^{-1}(\mathbf{Y}_i)(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0 \\ \displaystyle\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\alpha}'}\right)^T \mathrm{Cov}^{-1}(\mathbf{W}_i)(\mathbf{w}_i - \boldsymbol{\eta}_i) = 0 \end{cases} \tag{6.3}$$

for finding $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, where $\mathbf{W}_i = (Y_{i1}Y_{i2}, Y_{i1}Y_{i13}, \ldots, Y_{i(d-1)}Y_{id})'$, $\boldsymbol{\mu}_i = \mathrm{E}(\mathbf{Y}_i)$, $\boldsymbol{\eta}_i = \mathrm{E}(\mathbf{W}_i)$. $\boldsymbol{\beta}$ characterizes the marginal means $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{id})^t$ and $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_r)'$ characterizes the marginal pairwise association $\boldsymbol{\eta}_i$. $\mathrm{Cov}^{-1}(\mathbf{Y}_i)$ and $\mathrm{Cov}^{-1}(\mathbf{W}_i)$ could be replaced by some special chosen matrices. Let $\boldsymbol{\delta} = (\boldsymbol{\beta}', \boldsymbol{\alpha}')'$. Zhao and Prentice(1990) proposed to work with

$$\sum_{i=1}^{n} \left(\frac{\partial(\boldsymbol{\mu}_i, \boldsymbol{\eta}_i)}{\partial \boldsymbol{\delta}'}\right)^T \mathrm{Cov}^{-1}\begin{pmatrix}\mathbf{Y}_i \\ \mathbf{W}_i\end{pmatrix}\begin{pmatrix}\mathbf{y}_i - \boldsymbol{\mu}_i \\ \mathbf{w}_i - \boldsymbol{\eta}_i\end{pmatrix} = 0 \tag{6.4}$$

for finding $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. $\mathrm{Cov}\begin{pmatrix}\mathbf{Y}_i \\ \mathbf{W}_i\end{pmatrix}$ could be replaced by a "working" covariance matrix for $(\mathbf{Y}_i', \mathbf{W}_i')'$. The main idea here was to add extra estimation equations for the dependence parameters, to improve the parameter estimation from (6.2).

6.2 GEE in multivariate analysis

In this section, the GEE method is illustrated with some examples. Drawbacks of the method are discussed.

Example 6.1 Suppose $\mathbf{Y}_i \sim N_d(\mathbf{X}_i\boldsymbol{\beta}, \Sigma_i)$, $i = 1, \ldots, n$, where $\mathbf{X}_i$ is a known $d \times q$ matrix and Σ_i is a $d \times d$ covariance matrix, and $\boldsymbol{\beta}$ is a $q \times 1$ parameter vector. The maximum likelihood method is to minimize $\sum_{i=1}^{n}(\mathbf{y}_i - \boldsymbol{\mu}_i)'\Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i)$ with $\boldsymbol{\mu}_i = (\mu_{i1}, \mu_{i2}, \ldots, \mu_{id})' = \mathbf{X}_i\boldsymbol{\beta}$. It leads to

$$\sum_{i=1}^{n}\left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}\right)^T \Sigma_i^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = 0,$$

where $\partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}' = \mathbf{X}_i$. In this particular situation, the ML estimating equations are exactly the same as GEE. □

Example 6.2 Suppose $\mathbf{Y}_i$ has the multivariate probit model with multinormal copula. The mean vector is $\boldsymbol{\mu}_i = (\mu_{i1}, \ldots, \mu_{id})'$, where $\mu_{ij} = \Phi(\boldsymbol{\beta}\mathbf{x}_{ij})$, $i = 1, \ldots, n$, $j = 1, \ldots, d$. The GEE for $\boldsymbol{\beta}$, with the correct specification of the response variance matrix are

$$U(\boldsymbol{\beta}) = \sum_{i=1}^{n} D_i^T V_i(\mathbf{Y}_i)^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_i) = \mathbf{0},$$

where $D_i = \partial\boldsymbol{\mu}_i/\partial\boldsymbol{\beta}'$, and

$$V(\mathbf{Y}_i) = \begin{pmatrix} \sigma_{i1}^2 & \cdots & \alpha_{i,1d} \\ \vdots & \ddots & \vdots \\ \alpha_{i,1d} & \cdots & \sigma_{id}^2 \end{pmatrix}, \tag{6.5}$$

with $\sigma_{ij}^2 = \Phi(\boldsymbol{\beta}\mathbf{x}_{ij})(1 - \Phi(\boldsymbol{\beta}\mathbf{x}_{ij}))$, and $\alpha_{i,jk} = \Phi_2(\boldsymbol{\beta}\mathbf{x}_{ij}, \boldsymbol{\beta}\mathbf{x}_{ik}; \rho_{jk}) - \Phi(\boldsymbol{\beta}\mathbf{x}_{ij})\Phi(\boldsymbol{\beta}\mathbf{x}_{ik})$, $1 \le j < k \le d$. The GEE for $\boldsymbol{\beta}$ is different from MLE for $\boldsymbol{\beta}$ in this case. We also notice that in this example, the actual correlation among the responses depend on the mean values, and hence $\boldsymbol{\beta}\mathbf{x}_{ij}$, $j = 1, \ldots, d$. This is not considered in the GEE assumptions for the "working" correlation matrix. In GEE for multivariate binary data, the "working" correlation is usually assumed to be independent of the mean parameters. In the next section, more studies to compare GEE with MLE under the multivariate probit model assumption will be given. □

Example 6.3 We examine the multivariate Poisson-lognormal model of Example 2.12. Suppose in (2.28) – (2.30), $\mu_j = \nu$ and $\sigma_j = \eta$, $j = 1, \ldots, d$. Let $\boldsymbol{\beta} = (\nu, \eta)'$. We can obtain the estimates of ν and η by ML or IFM. Now we apply the GEE approach. Since for $i = 1, \ldots, n$, $\mathrm{E}(Y_{ij}) =$

$\exp\{\nu+\eta^2/2\}\stackrel{\text{def}}{=}a$, $\text{Var}(Y_{ij})=a+a^2[\exp(\eta^2)-1]\stackrel{\text{def}}{=}b$ and $\text{Cov}(Y_{ij},Y_{ik})=a^2[\exp(\theta_{jk}\eta^2)-1]$, $j\neq k$. Thus

$$\left(\frac{\partial \text{E}(\mathbf{Y}_i)}{\partial \boldsymbol{\beta}}\right)^T=\begin{pmatrix} a & \cdots & a \\ a\eta & \cdots & a\eta \end{pmatrix}_{2\times d}.$$

With the correct specification of the response mean function and the variance-covariance matrix, the GEE for the parameters ν and η are

$$\sum_{i=1}^{n}\left(\frac{\partial \text{E}(\mathbf{Y}_i)}{\partial \boldsymbol{\beta}}\right)^T V(\mathbf{Y}_i)^{-1}(\mathbf{y}_i-\text{E}(\mathbf{Y}_i))=\mathbf{0}, \tag{6.6}$$

where $V(\mathbf{Y}_i)$ has the diagonal component $a+a^2[\exp(\eta^2)-1]$, and off-diagonal (j,k) component $a^2[\exp(\theta_{jk}\eta^2)-1]$. Since the two rows in $(\partial \text{E}(\mathbf{Y}_i)/\partial\boldsymbol{\beta})^T$ are proportional, (6.6) reduces to a single equation, thus estimates for ν and η cannot be obtained with the quasi-likelihood approach. We can see this more clearly by examining a special situation where $\theta_{jk}=0$. In this case (6.6) becomes

$$\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{ij}-a)=0,$$

which leads only to an estimate of $a=\exp\{\nu+\eta^2/2\}$. In this situation, we obtain a consistent estimator for a, but not for ν and η separately. This is a situation where consistent estimators of the parameters of interest are not obtainable with GEE approach even with the correct specification of the mean functions, the variance functions and the correlation structure. Following GEE, if the interpretation of the model or parameters are carried out based on a, we will have the same interpretation for $\nu+\eta^2/2$ being constant irrespective of whether ν is in fact a relatively big value or small value. For if $\nu=h(\boldsymbol{\beta}\mathbf{x})$ where $\boldsymbol{\beta}$ is a parameter vector and $\mathbf{x}$ is a covariate vector, then the correct interpretation of the effect of covariates is not possible from the GEE approach. The problem with this example is that ν and η are confounded in $\text{E}(Y_j)$, and GEE fails to use the information on ν and η in the second moment of Y_j. This is an example showing that the form $\text{Var}(Y)=\phi V(\mu)$ is restrictive. □

The above examples provide some flavor of the GEE approach. It is clear that to use GEE for a meaningful purpose, the method requires the correct specification of marginal mean (and possibly correct specification of variance). It expects to get useful marginal regression parameter estimates with the dependence structure of multivariate data treated as a nuisance. An attractive feature of GEE is the partial requirement on the model through the specification only of lower order moments. But the GEE approach has a number of drawbacks if it is considered under the multivariate data analysis framework; some of the drawbacks are direct results of its attractiveness:

i. The GEE approach is incomplete for the data analysis cycle of initial data analysis, statistical modelling, estimation, diagnostics and inference. In published work, GEE focuses mainly on the estimation stage with emphasis on some marginal regression parameters estimation, which can only be considered as a small part of the whole multivariate analysis process. In multivariate data analysis, the proper analysis cycle is important for the interpretation of the findings to be statistically meaningful.

ii. With the GEE approach, the type of inferences can be made from the estimation results are limited. GEE is mainly useful for marginal regression parameter estimation, regardless of the possible multivariate models for the data. If the objective of scientific investigation is to find the probability occurrences of some phenomena, such as in multivariate discriminant analysis, GEE is not helpful. GEE also treats the dependence as a nuisance, and then use a "working" correlation matrix in estimation. This may deviate from the purpose of multivariate analysis which is often motivated by the need to analyze complex dependence among variables and objects. GEE does not deal with this question seriously. Furthermore, the correlation often is not the "best" notion of dependence for some multivariate non-normal variables.

iii. With the GEE approach, there is no clear way to assess the assumptions, such as common $\boldsymbol{\beta}$ for different univariate margins. The effective use of the GEE resides on the correct specification of marginal mean function. If the specifications are not correct, it would not be adequate to use the estimation results for the inference purposes. With GEE, when the inference is wrong, it is not easy to tell where is wrong, and to what extent the results are useful. The GEE has a direct representation within the exponential family, but may not be true for models not in the exponential family. Notice that many "interpretable" multivariate models are not in the exponential family.

iv. With the GEE approach, it may be difficult to have sound interpretations in some situations. A situation is the Example 6.3, where it is not possible to get an estimate of the parameter (or consistent estimate of the parameter) of interest through simple GEE, even under all favorable conditions for GEE, such as the correct specification of the mean functions, variance functions and the correlation function.

Then why GEE? GEE is a simple estimating approach in multivariate situations where only some knowledge of lower order moment characteristics are used. It may be considered as an appropriate approach when the relationships between covariates and marginal probabilities are of prime interest,

and when a proper multivariate model is not available or mathematically difficult to deal with. GEE may lead to some gain in estimation efficiency for the marginal regression parameters from a sound specification of the dependence among the response variables. In some practical situations, we may have some rough knowledge about the dependence structure among response variables; this knowledge can be appropriately incorporated into GEE. The GEE approach provides a way to avoid the difficulty of dealing with the complex relationship between some model parameters of interests and the joint probabilities that define the likelihood in multivariate (longitudinal) situation, while still estimating some parameters of interests.

However, in multivariate analysis, the marginal behaviour is only one of the possible features of interest. Others include the dependence structure among the response variables, the prediction of the probability for an outcome, the changes within subjects, etc. Within the general multivariate analysis framework, GEE should be considered only as a set of estimating equations for some parameters in a multivariate situation. Its usefulness is limited without incorporating it properly into the data analysis cycle, which mainly consist of initial data analysis, statistical modelling, disgnostics and inference.

Some technical problems related to the GEE approach are:

1. How efficient is the GEE approach? How it compares with the ML approach? (when a full model can be specified.)

2. How important is the correct specification of the response correlation matrix? If the correlations of the response variables do depend on the marginal regression parameters (see Example 6.2), how does GEE work?

3. What is the effect of the (correct) specification of variance function?

4. What is the practical meaning of "large sample size" with GEE to achieve the estimation consistency?

Item 1 is a natural question, since GEE is an approach which uses only the partial information of a likelihood model. Item 2 is related to the fact that the true correlation structure for the response variables is rarely known in practice. If different specifications of the correlation matrix make a difference on the marginal regression parameter estimate, what could we really say about the regression parameters? If different specifications of the correlation matrix do not make a difference, what else can we say (about the regression parameter estimates and correlations)? Item 4 is also a natural question for many statistical methodologies where their good properties are only established

in an asymptotic sense. For item 3, the point at issue is best introduced via a simple example. Suppose we have a Poisson-lognormal model. In Example 6.3, we have demonstrated that the GEE in (6.6) can not lead to an estimate of ν and η. We now simply limit our discussion to the univariate Poisson-lognormal model to illustrate our points. The GEE for the univariate case (with no covariates) is

$$\sum_{i=1}^{n} \frac{\partial a}{\partial \boldsymbol{\beta}} \frac{y_i - a}{b} = \mathbf{0}, \tag{6.7}$$

where $\boldsymbol{\beta} = (\nu, \eta)'$, $a = \mathrm{E}(Y_i) = \exp\{\nu + \eta^2/2\}$ and $b = \mathrm{Var}(Y_i) = a + a^2[\exp(\eta^2) - 1] = a + \tau a^2$, $\tau = \exp(\eta^2) - 1$. (6.7) is equivalent to

$$\sum_{i=1}^{n} (y_i - a) = 0. \tag{6.8}$$

To estimate ν as well as η, an additional equation to (6.8) is needed. One of a such equation (see McCullagh and Nelder 1989) is

$$\sum_{i=1}^{n} \frac{(y_i - a)^2}{b} - (n-1) = 0. \tag{6.9}$$

We see that in this simple univariate situation, (6.8) and (6.9) lead to the use of sample response mean and sample response variance to estimate the response mean and response variance. In the quasi-likelihood literature, it is usually assumed that the variance of Y_i has the form $\mathrm{Var}(Y_i) = \phi V(\mu_i)$, where ϕ is an unknown dispersion parameter and $V(\mu_i)$ is a function of $\mu_i = \mathrm{E}(Y_i)$. This is certainly not the case for the Poisson-lognormal model. In the Poisson-lognormal model, if we let $\mu_i = \mathrm{E}(Y_i)$, then $\mathrm{Var}(y_i) = \mu_i + \mu_i^2 \tau_i$, which cannot be identified with $\mathrm{Var}(y_i) = \phi V(\mu_i)$. If we always assume $\mathrm{Var}(y_i) = \phi V(\mu_i)$, in some situations, it is not possible to have a correct specification of the variance function. These arises a question of the effect of the variance function specification on the consistency of the marginal parameter estimates. It would be interesting to see how (6.8) and (6.9) estimate ν and η under different specifications of the variance functions.

In the next section, we will address these issues. For points 1, 2 and 4, we will use multivariate probit model for the investigation. For point 3, we will study the univariate Poisson-lognormal model.

6.3 GEE compared with the ML and IFM approaches

In this section, we will study some of the questions concerning GEE arisen at the end of section 6.2. We compare the GEE approach with the ML approach and the IFM approach by simulation with the knowledge of true models.

Comparison and simulation schemes

We study the cases where the regression parameters are common across all margins. We compare the GEE estimates with MLEs and IFMEs (from the pool-marginal-likelihood approach). Except for the Poisson-lognormal model where we investigate the effect of the specification of variance function, in the GEE estimation, we always assume we have the correct specification of the marginal variance functions. With the Poisson-lognormal model, we specify different variance functions to investigate the importance of correctly specifying the variance functions.

We use the mean-square error (MSE) of the estimate for a parameter from different approaches as the basis of the comparison. For an estimator $\tilde{\theta} = \tilde{\theta}(X_1, \ldots, X_n)$, where $X_1, \ldots, X_n$ is a random sample of size n from a distribution indexed by θ, the MSE of $\tilde{\theta}$ about the true value θ is

$$\text{MSE}(\tilde{\theta}) = E(\tilde{\theta} - \theta)^2 = Var(\tilde{\theta}) + [E(\tilde{\theta}) - \theta]^2.$$

Suppose that $\tilde{\theta}$ has a sampling distribution F, and suppose $\tilde{\theta}_1, \ldots, \tilde{\theta}_m$ are iid of F, then one obvious estimator of $\text{MSE}(\tilde{\theta})$ is

$$\widehat{\text{MSE}}(\tilde{\theta}) = \frac{\sum_{i=1}^{m}(\tilde{\theta}_i - \theta)^2}{m}. \tag{6.10}$$

The average of the parameter estimate is $mean(\tilde{\theta}) = \sum_{i=1}^{m} \tilde{\theta}_i/m$. Assume $\tilde{\theta}_{gee}$ is from the GEE approach, $\tilde{\theta}_{pmla}$ is from the IFM approach (with the pool-marginal-likelihood approach) and $\tilde{\theta}_{mle}$ is the MLE. We examine $r_1^2 = \widehat{\text{MSE}}(\tilde{\theta}_{mle})/\widehat{\text{MSE}}(\tilde{\theta}_{gee})$ and $r_2^2 = \widehat{\text{MSE}}(\hat{\theta}_{pmla})/\widehat{\text{MSE}}(\tilde{\theta}_{gee})$ (in all tables, r_1 and r_2 are reported). For a fixed sample size, $\hat{\theta}$ need not be the optimal estimate of θ in term of MSE, since now the bias of the estimate is also taken into consideration. The above two ratios may indicate how GEE performs in comparison with the other approaches, and particularly how it compares with the MLE. The approach is mainly computational, based on the computer implementation of specific models and then the subsequent intensive simulation and parameter estimation.

We will first use the multivariate probit model for investigating the relative efficiency of GEE estimates versus MLEs and IFMEs. We describe our simulation scheme and comparison scheme here in general terms. We simulate d-dimensional binary observations $\mathbf{y}_i$ $(i = 1, \ldots, n)$ from a multivariate probit model

$$Y_{ij} = I(Z_{ij} < z_{ij}), \quad j = 1, \ldots, d, \quad i = 1, \ldots, n,$$

where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{id})' \sim MVN_d(\mathbf{0}, \Theta_i)$ with $z_{ij} = \boldsymbol{\beta}_j' \mathbf{x}_{ij}$. The response correlation matrix for

the ith subject is $R_i = (r_{i,jk})$ where $r_{i,jj} = 1$ and

$$r_{i,jk} = \frac{P_{i,jk}(11) - P_{ij}(1)P_{ik}(1)}{\sqrt{P_{ij}(1)(1-P_{ij}(1))P_{ik}(1)(1-P_{ik}(1))}}, \quad \text{for } j \neq k. \tag{6.11}$$

where $P_{i,jk}(11) = \Phi_2(z, z; \theta_{jk})$, $P_{ij}(1) = P_{ik}(1) = \Phi(z)$ when there is no covariates, and $P_{i,jk}(11) = \Phi_2(\boldsymbol{\beta}_j' \mathbf{x}_{ij}, \boldsymbol{\beta}_k' \mathbf{x}_{ik}; \theta_{jk})$, $P_{ij}(1) = \Phi(\boldsymbol{\beta}_j' \mathbf{x}_{ij})$ and $P_{ik}(1) = \Phi(\boldsymbol{\beta}_k' \mathbf{x}_{ik})$ when there is a covariate vector. In the expression of $P_{i,jk}(11)$, we may have $\theta_{jk} = \rho$ or $\theta_{jk} = \rho^{|j-k|}$, depending on the dependence structure of the latent variables.

The following simulation scheme is used:

1. The sample size is n, the number of simulations is N. Both are reported in the tables.

2. Situations with covariates for $d = 2$ and $d = 3$ and with no covariates for $d = 3$ and $d = 4$ are considered:

   (a) With no covariates: $z_{ij} = z$ for $i = 1, \ldots, n$, $j = 1, \ldots, d$. Two chosen values of z are: 0.5 and 1.5.

   (b) With covariates:

      i. There are two situations for $d = 2$: $z_{ij} = \beta_0 + \beta_1 x_{ij}$ with $\beta_0 = 0.5$, $\beta_1 = 1$ and $z_{ij} = \beta_0 + \beta_1 w_i + \beta_2 x_{ij}$ with $\beta_0 = -0.5$, $\beta_1 = 0.5$, $\beta_2 = 1$, where x_{ij} is margin-dependent and w_i is margin-independent covariate

      ii. For $d = 3$, only $z_{ij} = \beta_0 + \beta_1 x_{ij}$ with $\beta_0 = 0.5$, $\beta_1 = 1$ is considered.

   Situations with x_{ij} discrete and continuous and w_i discrete are studied. For w_i discrete, we choose $w_i = I(U \leq 0)$ where $U \sim$ uniform$(-1, 1)$. For x_{ij} discrete, we choose $x_{ij} = I(U \leq j/(2d))$ where $U \sim$ uniform$(-1, 1)$; for x_{ij} continuous, we choose $x_{ij} \sim N(j/(2d), 1/4)$. The continuous covariate case is only studied for $d = 2$.

3. We assume the latent correlation matrix Θ_i free of covariates. For $d = 2$, ρ is chosen to be 0.9 and 0.5. For $d \geq 3$, Θ is chosen to be exchangeable with all correlation equal to ρ, and an AR(1) with (j, k) component equal to $\rho^{|j-k|}$. In both exchangeable and AR(1) cases, ρ is chosen to be 0.9.

In GEE, the "working" correlation matrix is chosen by the investigator. There is arbitrariness in the choice of the "working" correlation matrix. We want to see how the choice of the "working" correlation matrix affects the estimation of the regression parameters in situations where the mean functions (also variance functions) are correctly specified. For GEE estimation, we study two type of "working" correlation matrix specification:

1. The correct specification of correlation matrix of the response variables, that is, $r_{i,jk}$ is calculated from (6.11) with the true parameter values. In the tables, we use η_g for GEE specification of $r_{i,jk}$, and $\eta_g = c$ for correct specification of correlation matrix.

2. The wrong specification of correlation matrix of the response variables:

   (a) For $d = 2$, let $r_{i,12} = \eta_g$. We select $\eta_g = 0.9,\ 0.5,\ 0,\ -0.5,\ -0.9$.

   (b) When the latent correlation matrix is exchangeable: (i) the "working" correlation matrix has exchangeable structure with $r_{i,jk} = \eta_g$, where when $d = 3$, $\eta_g = 0$ and $\eta_g = -0.4$, and when $d = 4$, $\eta_g = 0$ and $\eta_g = -0.3$; and (ii) the "working" correlation matrix has AR(1) structure when $d = 3$, with $r_{i,jk} = \eta_g^{|j-k|}$ where $\eta_g = 0.9$ and $\eta_g = -0.9$.

   (c) When the latent correlation matrix is AR(1): (i) the "working" correlation matrix has AR(1) structure with $r_{i,jk} = \eta_g^{|j-k|}$ where $\eta_g = 0$ and $\eta_g = -0.9$ for both $d = 3$ and $d = 4$; and (ii) the "working" correlation matrix has exchangeable structure when $d = 3$, with $r_{i,jk} = \eta_g$ where $\eta_g = 0.0$ and $\eta_g = -0.4$.

In the computer implementation, we first simulate d-dimensional binary data from a given d-dimensional probit model with or without covariates. We then use the GEE, ML and IFM approaches to estimate the parameters from each simulation, and then compute the MSE in (6.10) of the estimates from each parameter estimation approach. We also compute the mean of the parameter estimates.

Next we discuss the simulation and computation scheme with the univariate Poisson-lognormal model to investigate the effects of the specification of variance functions on the estimation consistency of the marginal regression parameters. The GEE that we use are (6.8) and (6.9). The true variance for Y_i is $\mathrm{Var}(Y_i) = a + a^2\tau$, where $a = \mathrm{E}(Y_i) = \exp(\nu + \eta^2/2)$ and $\tau = \exp(\eta^2) - 1$ for the situation with no covariate, and $\mathrm{Var}(Y_i) = a_i + a_i^2\tau$, where $a_i = \mathrm{E}(Y_i) = \exp(\nu_i + \eta_i^2/2)$ and $\tau_i = \exp(\eta_i^2) - 1$ for the situation with covariates. In the comparison study, we compare MLE to GEE with 1) correct specification of $\mathrm{Var}(Y_i)$, 2) $\mathrm{Var}(Y_i) = \tau a$, 3) $\mathrm{Var}(Y_i) = \tau a^2$, 4) $\mathrm{Var}(Y_i) = \tau a^3$. Let a be the mean and b be the variance from GEE specifications. The simulation scheme is as follows:

1. The sample size is n; the number of simulation is N. Both numbers are given within the tables.

2. We considered the situations of the parameter ν independent of covariates and depending on a covariate x. The parameters are:

i. With no covariates: $(\nu, \eta) = (0.99995, 0.01)$. In this case, $a = 2.718282$, and for the 4 different variance function specifications above, we have $b = 2.719$, 0.00027, 0.0007, 0.002 respectively, where $b = 2.719$ corresponds to the correct specification of the variance function.

ii. With no covariates: $(\nu, \eta) = (-0.1, 1.48324)$. In this case, $a = 2.718282$, and for the 4 different variance function specification above, we have $b = 62.02$, 21.81, 59.30, 161.19 respectively, $b = 62.02$ corresponds to the correct specification of the variance function.

iii. With covariate: $\nu = \alpha + \beta x$, where $\alpha = 0.5$, $\beta = 0.5$ and $x = I(U \leq 0)$ with $U \sim$ uniform$(-1, 1)$. The parameter $\eta = 0.01$.

Next we provide the numerical results for the situations outlined above.

Bivariate probit model

For the bivariate probit model with one covariate, the marginal linear regressions are $z_{ij} = \beta_0 + \beta_1 x_{ij}$, where β_0, β_1 are the common marginal regression parameters. In GEE, we have $\boldsymbol{\mu}_i = (P_{i1}(1), P_{i1}(1))' = (\Phi(\beta_0 + \beta_1 x_{i1}), \Phi(\beta_0 + \beta_1 x_{i2}))'$. The numerical results for the bivariate probit model with the covariate x_{ij} continuous and discrete are presented in Table 6.1 and Table 6.2. The numerical results for the bivariate probit model with the marginal linear regressions $z_{ij} = \beta_0 + \beta_1 w_i + \beta_2 x_{ij}$ for w_i and x_{ij} discrete are presented in Table 6.3. The results for w_i and x_{ij} being continuous are quite similar, so they are not presented. In Table 6.4, we also present a case with the marginal linear regressions are $z_{ij} = \beta_0 + \beta_1 x_{ij}$ for the situation where the true parameter $\rho = 0.5$. From these tables, two clear conclusions emerge: i) the specification of the "working" correlation has an effect on the estimation efficiency of GEE, with a major loss of efficiency when the specified "working" correlation is far from the true correlation. In fact, when the working correlation parameter is far away (particular with the wrong sign of the correlation) from the true correlation parameter, the GEE estimator performs poorly, and in some cases, the efficiency can be as low as 50%; ii) MLEs are always more efficient than GEE, but GEE is slightly more efficient than estimate from IFM when the the "working" correlation is correctly specified; iii) the observations in i) and ii) are consistent for the sample size from large to moderate.

Trivariate probit model

We first study the trivariate probit model with no covariate. We have $P(111) = \Phi_3(z, z, z; \rho_{12}, \rho_{13}, \rho_{23})$, where z is the common cut-off points for all three margins. In GEE, we have $\boldsymbol{\mu}_i = (P_1(1), P_2(1), P_3(1))' =$

Table 6.1: GEE assessment: $d = 2$, $\beta_0 = 0.5, \beta_1 = 1$, x_{ij} discrete, $\rho = 0.9$, $N = 1000$

		$n = 1000$			$n = 200$		
		mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	β_0	0.501 (0.0525)			0.504 (0.1191)		
	β_1	1.003 (0.0699)			1.011 (0.1553)		
IFME	β_0	0.502 (0.0570)			0.507 (0.1293)		
	β_1	1.001 (0.0788)			1.007 (0.1792)		
η_g =c	β_0	0.501 (0.0531)	0.989	1.074	0.505 (0.1198)	0.994	1.079
	β_1	1.003 (0.0707)	0.989	1.116	1.010 (0.1563)	0.993	1.146
$\eta_g = 0.9$	β_0	0.500 (0.0566)	0.929	1.009	0.501 (0.1280)	0.930	1.011
	β_1	1.005 (0.0776)	0.901	1.016	1.014 (0.1713)	0.907	1.046
$\eta_g = 0.5$	β_0	0.501 (0.0532)	0.987	1.072	0.504 (0.1199)	0.993	1.079
	β_1	1.003 (0.0707)	0.988	1.115	1.010 (0.1561)	0.995	1.148
$\eta_g = 0$	β_0	0.502 (0.0570)	0.921	1.0	0.507 (0.1293)	0.921	1.0
	β_1	1.001 (0.0788)	0.887	1.0	1.007 (0.1791)	0.867	1.0
$\eta_g = -0.5$	β_0	0.503 (0.0687)	0.765	0.831	0.508 (0.1575)	0.756	0.821
	β_1	1.001 (0.1017)	0.687	0.775	1.008 (0.2381)	0.652	0.752
$\eta_g = -0.9$	β_0	0.503 (0.0818)	0.642	0.698	0.506 (0.1906)	0.625	0.678
	β_1	1.000 (0.1262)	0.554	0.625	1.011 (0.3029)	0.513	0.591

Table 6.2: GEE assessment: $d = 2$, $\beta_0 = 0.5, \beta_1 = 1$, x_{ij} continuous, $\rho = 0.9$, $N = 1000$

		$n = 1000$		
		mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	β_0	0.501 (0.0416)		
	β_1	1.000 (0.0651)		
IFME	β_0	0.501 (0.0427)		
	β_1	1.001 (0.0730)		
η_g =c	β_0	0.501 (0.0416)	1.0	1.027
	β_1	1.000 (0.0653)	0.997	1.117
$\eta_g = 0.9$	β_0	0.502 (0.0496)	0.839	0.861
	β_1	0.998 (0.0766)	0.854	0.957
$\eta_g = 0.5$	β_0	0.501 (0.0421)	0.990	1.016
	β_1	0.999 (0.0675)	0.964	1.081
$\eta_g = 0$	β_0	0.501 (0.0427)	0.974	1.0
	β_1	1.001 (0.0730)	0.892	1.0
$\eta_g = -0.5$	β_0	0.500 (0.0463)	0.899	0.923
	β_1	1.004 (0.0952)	0.685	0.767
$\eta_g = -0.9$	β_0	0.499 (0.0512)	0.813	0.835
	β_1	1.007 (0.1209)	0.539	0.604

Table 6.3: GEE assessment: $d = 2$, $\beta_0 = -0.5, \beta_1 = 0.5, \beta_2 = 1$, w_i, x_{ij} discrete, $\rho = 0.9$, $N = 1000$

		$n = 1000$			$n = 200$		
		mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	β_0	-0.496 (0.0647)			-0.498 (0.1445)		
	β_1	0.497 (0.0782)			0.4986 (0.1793)		
	β_2	0.996 (0.0601)			1.005 (0.1289)		
IFME	β_0	-0.496 (0.0676)			-0.497 (0.1527)		
	β_1	0.497 (0.0788)			0.498 (0.1806)		
	β_2	0.997 (0.0664)			1.004 (0.1488)		
η_g =c	β_0	-0.495 (0.0652)	0.993	1.038	-0.495 (0.1447)	0.999	1.056
	β_1	0.498 (0.0783)	0.998	1.005	0.498 (0.1786)	1.003	1.012
	β_2	0.995 (0.0605)	0.994	1.098	1.001 (0.1294)	0.996	1.150
$\eta_g = 0.9$	β_0	-0.493 (0.0693)	0.934	0.977	-0.493 (0.1522)	0.950	1.003
	β_1	0.498 (0.0827)	0.946	0.953	0.503 (0.1905)	0.941	0.949
	β_2	0.992 (0.0676)	0.889	0.982	0.999 (0.1429)	0.903	1.042
$\eta_g = 0.5$	β_0	-0.495 (0.0655)	0.988	1.033	-0.495 (0.1451)	0.996	1.052
	β_1	0.497 (0.0786)	0.994	1.001	0.497 (0.1797)	0.998	1.005
	β_2	0.994 (0.0605)	0.994	1.097	1.001 (0.1294)	0.996	1.150
$\eta_g = 0$	β_0	-0.497 (0.0677)	0.957	1.0	-0.496 (0.1527)	0.946	1.0
	β_1	0.497 (0.0788)	0.993	1.0	0.498 (0.1806)	0.992	1.0
	β_2	0.997 (0.0664)	0.905	1.0	1.004 (0.1488)	0.866	1.0
$\eta_g = -0.5$	β_0	-0.499 (0.0768)	0.843	0.881	-0.499 (0.1762)	0.820	0.867
	β_1	0.498 (0.0789)	0.991	0.998	0.498 (0.1816)	0.987	0.995
	β_2	1.000 (0.0857)	0.701	0.774	1.008 (0.1977)	0.652	0.753
$\eta_g = -0.9$	β_0	-0.500 (0.0880)	0.736	0.769	-0.502 (0.2039)	0.709	0.749
	β_1	0.498 (0.0790)	0.989	0.997	0.498 (0.1823)	0.983	0.991
	β_1	1.003 (0.1067)	0.563	0.622	1.012 (0.2486)	0.519	0.599

Table 6.4: GEE assessment: $d = 2$, $\beta_0 = 0.5, \beta_1 = 1$, x_{ij} discrete, $\rho = 0.5$, $N = 1000$

		$n = 1000$			$n = 200$		
		mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	β_0	0.502 (0.053)			0.504 (0.119)		
	β_1	1.002 (0.075)			1.011 (0.155)		
IFME	β_0	0.502 (0.055)			0.507 (0.129)		
	β_1	1.001 (0.077)			1.007 (0.179)		
η_g =c	β_0	0.502 (0.053)	0.999	1.022	0.505 (0.120)	0.994	1.079
	β_1	1.002 (0.075)	0.994	1.028	1.010 (0.156)	0.993	1.146
$\eta_g = 0.9$	β_0	0.500 (0.060)	0.893	0.914	0.500 (0.135)	0.885	0.845
	β_1	1.005 (0.088)	0.849	0.878	1.017 (0.194)	0.898	0.875
$\eta_g = 0.5$	β_0	0.501 (0.054)	0.983	1.005	0.503 (0.122)	0.982	0.973
	β_1	1.003 (0.077)	0.967	1.001	1.011 (0.168)	0.997	1.008
$\eta_g = 0$	β_0	0.502 (0.055)	0.977	1.0	0.506 (0.121)	0.986	1.0
	β_1	1.001 (0.077)	0.966	1.0	1.007 (0.169)	0.965	1.0
$\eta_g = -0.5$	β_0	0.504 (0.063)	0.850	0.870	0.508 (0.139)	0.860	0.872
	β_1	0.999 (0.092)	0.807	0.835	1.005 (0.207)	0.788	0.817
$\eta_g = -0.9$	β_0	0.504 (0.074)	0.725	0.742	0.508 (0.164)	0.728	0.738
	β_1	0.998 (0.112)	0.664	0.688	1.006 (0.257)	0.635	0.658

Table 6.5: GEE assessment: $d = 3$, $z = 0.5$, latent exchangeable, $\rho = 0.9$, "working" exchangeable, $N = 1000$

	$n = 1000$			$n = 200$			$n = 100$		
	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	0.499 (0.0363)			0.503 (0.0839)			0.506 (0.1228)		
IFME	0.498 (0.0366)			0.503 (0.0841)			0.506 (0.1233)		
η_g =c	0.498 (0.0366)	0.992	1.0	0.503 (0.0841)	0.997	1.0	0.506 (0.1233)	0.996	1.0
$\eta_g = 0$	0.498 (0.0366)	0.992	1.0	0.503 (0.0841)	0.997	1.0	0.506 (0.1233)	0.996	1.0
$\eta_g = -0.4$	0.498 (0.0366)	0.992	1.0	0.503 (0.0841)	0.997	1.0	0.506 (0.1233)	0.996	1.0

Table 6.6: GEE assessment: $d = 3$, $z = 1.5$, latent exchangeable, $\rho = 0.9$, "working" exchangeable, $N = 1000$

	$n = 1000$			$n = 200$		
	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	1.500 (0.0525)			1.510 (0.1254)		
IFME	1.500 (0.0525)			1.510 (0.1256)		
η_g =c	1.500 (0.0525)	0.999	1.0	1.510 (0.1256)	0.998	1.0
$\eta_g = 0$	1.500 (0.0525)	0.999	1.0	1.510 (0.1256)	0.998	1.0
$\eta_g = -0.4$	1.500 (0.0525)	0.999	1.0	1.510 (0.1256)	0.998	1.0

$(\Phi(z), \Phi(z), \Phi(z))'$. The numerical results are presented in Table 6.5 to Table 6.8. The numerical results show that the specification of the correlation of the response variables in these simple situations have little effect on the parameter estimates from GEE. GEE is efficient in all cases.

For the trivariate probit model with one covariate, we have $P(111) = \Phi_3(\beta_0+\beta_1 x_1, \beta_0+\beta_1 x_2, \beta_0+\beta_1 x_3; \rho_{12}, \rho_{13}, \rho_{23})$, where β_0, β_1 are the common marginal regression parameters. In GEE, we have $\boldsymbol{\mu}_i = (P_{i1}(1), P_{i1}(1), P_{i3}(1))' = (\Phi(\beta_0+\beta_1 x_{i1}), \Phi(\beta_0+\beta_1 x_{i2}), \Phi(\beta_0+\beta_1 x_{i3}))'$. The numerical results for the trivariate probit model with covariate are presented in Table 6.9 and Table 6.10. We studied models with discrete covariate. Now the specification of the "working" correlation matrix has some effect on the estimation efficiency of GEE, with a major loss of efficiency when the specified "working" correlation matrix is far from the true correlation matrix. We also notice that GEE is slightly more

Table 6.7: GEE assessment: $d = 3$, $z = 0.5$, latent AR(1), $\rho = 0.9$, "working" AR(1), $N = 1000$

	$n = 1000$			$n = 200$			$n = 100$		
	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	0.499 (0.0355)			0.503 (0.0822)			0.506 (0.1192)		
IFME	0.498 (0.0357)			0.503 (0.0822)			0.505 (0.1198)		
η_g =c	0.498 (0.0357)	0.996	1.0	0.503 (0.0821)	1.0	1.0	0.506 (0.1192)	1.0	1.0
$\eta_g = 0$	0.498 (0.0357)	0.994	1.0	0.503 (0.0822)	1.0	1.0	0.505 (0.1198)	0.995	1.0
$\eta_g = -0.9$	0.498 (0.0362)	0.982	0.988	0.503 (0.0832)	0.989	0.988	0.505 (0.1219)	0.978	0.983

Table 6.8: GEE assessment: $d = 3$, $z = 1.5$, latent AR(1), $\rho = 0.9$, "working" AR(1), $N = 1000$

	$n = 1000$			$n = 200$		
	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	1.499 (0.0515)			1.508 (0.1213)		
IFME	1.500 (0.0516)			1.509 (0.1220)		
η_g =c	1.500 (0.0515)	1.0	1.0	1.509 (0.1213)	1.0	1.0
$\eta_g = 0$	1.500 (0.0517)	0.997	1.0	1.509 (0.1220)	0.994	1.0
$\eta_g = -0.9$	1.500 (0.0527)	0.978	0.981	1.510 (0.1247)	0.973	0.979

Table 6.9: GEE assessment: $d = 3$, $\beta_0 = 0.5, \beta_1 = 1$, x_{ij} discrete, latent exchangeable, $\rho = 0.9$, "working" exchangeable, $N = 1000$

		$n = 1000$			$n = 200$			$n = 100$		
		mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	β_0	0.501 (0.0462)			0.499 (0.1057)			0.504 (0.1515)		
	β_1	0.998 (0.0582)			1.005 (0.1338)			1.015 (0.1915)		
IFME	β_0	0.502 (0.0502)			0.502 (0.1148)			0.509 (0.1680)		
	β_1	0.996 (0.0677)			1.001 (0.1588)			1.008 (0.2260)		
η_g =c	β_0	0.501 (0.0463)	0.997	1.084	0.500 (0.1066)	0.991	1.077	0.505 (0.1528)	0.992	1.100
	β_1	0.998 (0.0588)	0.991	1.153	1.004 (0.1369)	0.977	1.160	1.012 (0.1939)	0.987	1.165
$\eta_g = 0$	β_0	0.502 (0.0502)	0.920	1.0	0.502 (0.1148)	0.920	1.0	0.509 (0.1680)	0.902	1.0
	β_1	0.996 (0.0677)	0.860	1.0	1.001 (0.1588)	0.843	1.0	1.008 (0.2261)	0.847	1.0
$\eta_g = -0.4$	β_0	0.504 (0.0765)	0.604	0.657	0.503 (0.1764)	0.599	0.651	0.503 (0.2673)	0.567	0.628
	β_1	0.994 (0.1219)	0.478	0.556	1.001 (0.2886)	0.464	0.551	1.022 (0.4419)	0.433	0.512

efficient than the estimate from IFM when the "working" correlation matrix is correctly specified. From the tables, we also notice that GEE and IFME have the same parameter estimate when $\eta_g = 0$ and in exchangeable situations. In the following subsection, we will prove this is always true. We particularly notice that the GEE behaved very similarly to MLE and IFM, both in terms of marginal regression parameter estimates and their MSEs, when the working correlation matrix is chosen to be I.

4-variate probit model

The 4-variate probit model is considered for the situation with no covariate. We have $P(1111) = \Phi_4(z, z, z, z; \rho_{12}, \rho_{13}, \rho_{14}, \rho_{23}, \rho_{24}, \rho_{34})$, where z is the common cut-off points for all four margins. In GEE, we have $\boldsymbol{\mu}_i = (P_1(1), P_2(1), P_3(1), P_4(1))' = (\Phi(z), \Phi(z), \Phi(z), \Phi(z))'$. The numerical results are presented in Table 6.11 and Table 6.12. The numerical results show that the specification of the correlation of the response variables in these simple situations have little effect on the parameter estimates from GEE. The GEE approach is efficient in all cases.

Our simulation results also indicate that for estimation purposes, the estimating equations based

Table 6.10: GEE assessment: $d = 3$, $\beta_0 = 0.5, \beta_1 = 1$, x_{ij} discrete, latent AR(1), $\rho = 0.9$, "working" AR(1), $N = 1000$

		$n = 1000$			$n = 200$			$n = 100$		
		mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	β_0	0.501 (0.0455)			0.500 (0.1042)			0.505 (0.1517)		
	β_1	0.999 (0.0575)			1.004 (0.1330)			1.011 (0.1879)		
IFME	β_0	0.502 (0.0494)			0.502 (0.1123)			0.509 (0.1665)		
	β_1	0.997 (0.0666)			1.001 (0.1560)			1.006 (0.2195)		
η_g =c	β_0	0.501 (0.0455)	1.0	1.087	0.500 (0.1048)	0.994	1.071	0.507 (0.1530)	0.992	1.088
	β_1	0.999 (0.0580)	0.992	1.149	1.003 (0.1355)	0.981	1.151	1.007 (0.1899)	0.990	1.156
$\eta_g = 0$	β_0	0.502 (0.0494)	0.921	1.0	0.502 (0.1123)	0.928	1.0	0.509 (0.1665)	0.911	1.0
	β_1	0.997 (0.0666)	0.864	1.0	1.001 (0.1560)	0.852	1.0	1.006 (0.2195)	0.856	1.0
$\eta_g = -0.4$	β_0	0.504 (0.0688)	0.661	0.718	0.502 (0.1584)	0.658	0.709	0.505 (0.2367)	0.641	0.703
	β_1	0.995 (0.1076)	0.535	0.619	1.004 (0.2573)	0.517	0.606	1.020 (0.3789)	0.496	0.580

Table 6.11: GEE assessment: $d = 4$, $z = 0.5$, latent exchangeable, $\rho = 0.9$, "working" exchangeable, $N = 1000$

	$n = 1000$			$n = 200$		
	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	0.500 (0.0365)			0.503 (0.0820)		
IFME	0.500 (0.0366)			0.503 (0.0829)		
η_g =c	0.500 (0.0366)	0.997	1.0	0.503 (0.0829)	0.989	1.0
$\eta_g = 0$	0.500 (0.0366)	0.997	1.0	0.503 (0.0829)	0.989	1.0
$\eta_g = -0.3$	0.500 (0.0366)	0.997	1.0	0.503 (0.0829)	0.989	1.0

on an independence working correlation structure behave quite well.

IFME or GEE

We have seen in the preceding subsection that GEE has better performance than IFME when the response correlation matrix is correctly specified, but IFME has better performance than GEE in general. Now we will see some situations where GEE and IFME are equivalent.

Table 6.12: GEE assessment: $d = 4$, $z = 0.5$, latent AR(1), $\rho = 0.9$, "working" AR(1), $N = 1000$

	$n = 1000$			$n = 200$		
	mean ($\sqrt{\text{MSE}}$)	r_1	r_2	mean ($\sqrt{\text{MSE}}$)	r_1	r_2
MLE	0.500 (0.0349)			0.503 (0.0781)		
IFME	0.500 (0.0350)			0.503 (0.0791)		
η_g =c	0.500 (0.0350)	0.997	1.001	0.503 (0.0787)	0.992	1.005
$\eta_g = 0$	0.500 (0.0350)	0.997	1.0	0.502 (0.0791)	0.987	1.0
$\eta_g = -0.9$	0.500 (0.0354)	0.985	0.989	0.502 (0.0803)	0.972	0.985

Result 6.1 *For a multivariate probit model with common cut-off points across margins, GEE with $R_i(\boldsymbol{\alpha}) = R$, where R has exchangeable structure, is equivalent to IFM.*

Proof: For a multivariate probit model with common cut-off points across margins, the IFM leads to an estimating equation $\sum_{i=1}^{n}\sum_{j=1}^{d}(y_{ij}-\mu_j)=0$. This is equivalent to the GEE with $R_i(\boldsymbol{\alpha}) = R$, where R has an exchangeable structure. □

Result 6.2 *For a multivariate probit model with covariates, GEE with $R_i(\boldsymbol{\alpha}) = I$, where I is the identity matrix, is equivalent to IFM.*

Proof: Assume $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_d)'$ and $\mu_j = \Phi(\boldsymbol{\beta}x_j)$. For a multivariate probit model with common cut-off points across margins, IFM leads to the estimating equations

$$\sum_{i=1}^{n}\sum_{j=1}^{d}\frac{\partial\mu_j}{\partial\boldsymbol{\beta}}\frac{y_{ij}-\mu_j}{\mu_j(1-\mu_j)}=\mathbf{0}.$$

This is equivalent to the GEE with $R_i(\boldsymbol{\alpha}) = I$, I is the identity matrix. □

In this Chapter, we limit study to the GEE for common regression parameters across all margins. But GEE can also extended to the situations where parameters differ from margin to margin. We here introduce a result about the equivalency of GEE and IFM in some special situations with parameters differing from margin to margin.

Result 6.3 *For the multivariate probit model with parameters differing from margin to margin and with one margin-independent binary covariate, the GEE with $R_i(\boldsymbol{\alpha}) = R$ is equivalent to IFM for the marginal regression parameters.*

Proof: Assume x is the margin-independent binary covariate taking two values a and b. The marginal mean vector $\boldsymbol{\mu}_i = \{\mu_1(\alpha_1+\beta_1 x), \ldots, \mu_d(\alpha_d+\beta_d x)\}'$, $i = 1, \ldots, n$, takes two distinct vector values:

$$\begin{cases}\boldsymbol{\mu}_a = \{\mu_1(\alpha_1+\beta_1 a), \ldots, \mu_d(\alpha_d+\beta_d a)\}'\\ \boldsymbol{\mu}_b = \{\mu_1(\alpha_1+\beta_1 b), \ldots, \mu_d(\alpha_d+\beta_d b)\}'.\end{cases}$$

Assume there are n_a observations for $x = a$ and n_b for $x = b$, and let $\mathcal{I}_a = \{i \,|\, x_i = a\}$ and $\mathcal{I}_b = \{i \,|\, x_i = b\}$. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_d)'$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)'$. For $i \in \mathcal{I}_a$, $D_{i,\boldsymbol{\alpha}} = \partial\boldsymbol{\mu}_a/\partial\boldsymbol{\alpha}'$ does not depend on i, thus $D_{i,\boldsymbol{\alpha}}^T V_{i,\boldsymbol{\alpha}}^{-1}$ is a $d \times d$ invertible matrix which does not depend on i. Let us denote this matrix by A. For $i \in \mathcal{I}_a$, we also have $D_{i,\boldsymbol{\beta}}^T V_{i,\boldsymbol{\beta}}^{-1} = aD_{i,\boldsymbol{\alpha}}^T V_{i,\boldsymbol{\alpha}}^{-1} = aA$. Similarly, for $i \in \mathcal{I}_b$, $D_{i,\boldsymbol{\alpha}}^T V_{i,\boldsymbol{\alpha}}^{-1}$ does not depend on i. If we denote $D_{i,\boldsymbol{\alpha}}^T V_{i,\boldsymbol{\alpha}}^{-1}$ by B for $i \in \mathcal{I}_b$, we also have

Table 6.13: Estimates of ν and η under different variance specification

Spec. of Var(Y_i)	$\tilde{\eta}$	$\tilde{\nu}$
$a + a^2\tau$	$\tilde{\eta}_1$=$\{\log[(s^2 - \bar{y})/\bar{y}^2 + 1]\}^{1/2}$	$\log \bar{y} - 0.5(\tilde{\eta}_1)^2$
$a\tau$	$\tilde{\eta}_2$=$\{\log[s^2/\bar{y} + 1]\}^{1/2}$	$\log \bar{y} - 0.5(\tilde{\eta}_2)^2$
$a^2\tau$	$\tilde{\eta}_3$=$\{\log[s^2/\bar{y}^2 + 1]\}^{1/2}$	$\log \bar{y} - 0.5(\tilde{\eta}_3)^2$
$a^3\tau$	$\tilde{\eta}_4$=$\{\log[s^2/\bar{y}^3 + 1]\}^{1/2}$	$\log \bar{y} - 0.5(\tilde{\eta}_4)^2$

$D^T_{i,\boldsymbol{\beta}} V^{-1}_{i,\boldsymbol{\beta}} = bD^T_{i,\boldsymbol{\alpha}} V^{-1}_{i,\boldsymbol{\alpha}} = bB$. The GEE for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are

$$\begin{cases} A \sum_{i_a=1}^{n_a} (\mathbf{y}_{i_a} - \boldsymbol{\mu}_a) + B \sum_{i_b=1}^{n_b} (\mathbf{y}_{i_b} - \boldsymbol{\mu}_b) = \mathbf{0} \\ aA \sum_{i_a=1}^{n_a} (\mathbf{y}_{i_a} - \boldsymbol{\mu}_a) + bB \sum_{i_b=1}^{n_b} (\mathbf{y}_{i_b} - \boldsymbol{\mu}_b) = \mathbf{0} \end{cases},$$

which simplify to

$$\begin{cases} \sum_{i_a=1}^{n_a} (\mathbf{y}_{i_a} - \boldsymbol{\mu}_a) = \mathbf{0} \\ \sum_{i_b=1}^{n_b} (\mathbf{y}_{i_b} - \boldsymbol{\mu}_b) = \mathbf{0}. \end{cases} \tag{6.12}$$

It is straightforward to see that (6.12) is also the estimating equations for $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ from IFM approach. □

Poisson-lognormal model

Let $\bar{y} = \sum y_i/n$ be the sample mean, and $s^2 = \sum(y_i - \bar{y})^2/(n-1)$ be the sample variance. Using the estimating equations (6.8) and (6.9), the estimates of ν and η based on different specification of the variance functions are listed in Table 6.13 ($a = \exp(\nu + \eta^2/2)$ and $\tau = \exp(\eta^2) - 1$). Tables 6.14 - 6.16 contain numerical results based on the simulation scheme outlined for the Poisson-lognormal model previously. From the results in Tables 6.14 - 6.16, we see that quasi-likelihood estimates may be fine when the variance function is correctly specified, but may be asymptotically inconsistent if the variance function specification is not correct. A similar problem occurs in the multivariate case. It is thus critical to assess the form of Var(Y) as a function of E(Y) before choosing GEE as the estimation method.

Table 6.14: GEE assessment: $(\nu, \eta) = (0.99995, 0.01)$, $\mathrm{E}(Y) = 2.718282$, $\mathrm{Var}(Y) = 2.719$, $n = 1000$, $N = 500$

	$\tilde{\bar{\eta}}$ ($\sqrt{\mathrm{MSE}}$)	r_1	$\tilde{\bar{\nu}}$ ($\sqrt{\mathrm{MSE}}$)	r_1
MLE	0.038 (0.0595)		0.997 (0.0190)	
$a + a^2\tau$	0.047 (0.0705)	0.844	0.996 (0.0190)	1.0
$a\tau$	0.831 (0.8214)	0.072	0.653 (0.3472)	0.055
$a^2\tau$	0.559 (0.5491)	0.108	0.843 (0.1587)	0.120
$a^3\tau$	0.356 (0.3461)	0.172	0.935 (0.0675)	0.282

Table 6.15: GEE assessment: $(\nu, \eta) = (-0.1, 1.48324)$, $\mathrm{E}(Y) = 2.718282$, $\mathrm{Var}(Y) = 62.02$, $n = 1000$, $N = 100$

	$\tilde{\bar{\eta}}$ ($\sqrt{\mathrm{MSE}}$)	r_1	$\tilde{\bar{\nu}}$ ($\sqrt{\mathrm{MSE}}$)	r_1
MLE	1.481 (0.0591)		-0.094 (0.0559)	
$a + a^2\tau$	1.418 (0.1488)	0.397	-0.016 (0.1790)	0.313
$a\tau$	1.724 (0.2743)	0.215	-0.497 (0.4361)	0.128
$a^2\tau$	1.436 (0.1347)	0.439	-0.042 (0.1621)	0.345
$a^3\tau$	1.126 (0.3767)	0.157	0.357 (0.4741)	0.118

Table 6.16: GEE assessment: $\alpha = 0.5$, $\beta = 0.5$, $\eta = 0.01$, $n = 1000$, $N = 500$

	$\tilde{\bar{\alpha}}$ ($\sqrt{\mathrm{MSE}}$)	r_1	$\tilde{\bar{\beta}}$ ($\sqrt{\mathrm{MSE}}$)	r_1	$\tilde{\bar{\eta}}$ ($\sqrt{\mathrm{MSE}}$)	r_1
MLE	0.492 (0.037)		0.502 (0.045)		0.063 (0.078)	
$a + a^2\tau$	0.492 (0.038)	0.985	0.502 (0.045)	0.994	0.071 (0.084)	0.921
$a\tau$	0.153 (0.349)	0.107	0.499 (0.049)	0.916	0.832 (0.822)	0.095
$a^2\tau$	0.262 (0.242)	0.154	0.580 (0.096)	0.469	0.624 (0.614)	0.127
$a^3\tau$	0.342 (0.164)	0.227	0.593 (0.107)	0.419	0.458 (0.448)	0.173

Appendix: Newton-Raphson method for GEE

We perform the model simulations and all MLE, IFM and GEE computations using programs in C written by the author. The code for the probit model incorporates the cases with covariates and with no covariates. For completeness, we provide here some mathematical details about the Newton-Raphson method that we used in the GEE estimation. To apply Newton-Raphson method, we need to evaluate both the estimating functions and the derivative of the estimating functions at arbitrary points of parameter vector.

When the same regression parameter vector is common to all margins, the marginal mean function vector is $\boldsymbol{\mu}_i = (\mu_{i1}(\boldsymbol{\beta}), \ldots, \mu_{id}(\boldsymbol{\beta}))'$ where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)'$. Assume the correlation matrix in GEE for the ith subject is

$$R_i = \begin{pmatrix} 1 & a_{i12} & \cdots & a_{i1d} \\ & 1 & \cdots & a_{i2d} \\ & & \ddots & \vdots \\ & & & 1 \end{pmatrix}.$$

Then the estimating functions in GEE are

$$\sum_{i=1}^{n} \left(\frac{\partial \boldsymbol{\mu}_i}{\partial \boldsymbol{\beta}'}\right)^T A_i^{-1/2} R_i^{-1} A_i^{-1/2} (\mathbf{y}_i - \boldsymbol{\mu}_i) = \begin{pmatrix} \sum_{i=1}^{n} \sum_{j=1}^{d} \left(\frac{1}{\sigma_{ij}} \frac{\partial \mu_{ij}}{\partial \beta_0} \sum_{k=1}^{d} \frac{y_{ik} - \mu_{ik}}{\sigma_{ik}} a_{i,jk} \right) \\ \vdots \\ \sum_{i=1}^{n} \sum_{j=1}^{d} \left(\frac{1}{\sigma_{ij}} \frac{\partial \mu_{ij}}{\partial \beta_p} \sum_{k=1}^{d} \frac{y_{ik} - \mu_{ik}}{\sigma_{ik}} a_{i,jk} \right) \end{pmatrix}.$$

The estimating function corresponding to the mth regression parameter for the ith subject ($i = 1, \ldots, n$, $m = 0, 1, \ldots, p$) is

$$\psi_{im} = \sum_{j=1}^{d} \left[\frac{1}{\sigma_{ij}} \frac{\partial \mu_{ij}}{\partial \beta_m} \sum_{k=1}^{d} \frac{y_{ik} - \mu_{ik}}{\sigma_{ik}} a_{i,jk} \right].$$

After a few lines of calculations, we then have

$$\begin{aligned} \frac{\partial \psi_{im}}{\partial \beta_q} = & \sum_{j=1}^{d} \left[\frac{2\mu_{ij} - 1}{2\sigma_{ij}^3} \frac{\partial \mu_{ij}}{\partial \beta_m} \frac{\partial \mu_{ij}}{\partial \beta_q} \sum_{k=1}^{d} \frac{y_{ik} - \mu_{ik}}{\sigma_{ik}} a_{i,jk} \right. \\ & \left. + \frac{1}{\sigma_{ij}} \frac{\partial^2 \mu_{ij}}{\partial \beta_m \partial \beta_q} \sum_{k=1}^{d} \frac{y_{ik} - \mu_{ik}}{\sigma_{ik}} a_{i,jk} + \frac{1}{\sigma_{ij}} \frac{\partial \mu_{ij}}{\partial \beta_m} \sum_{k=1}^{d} \left(\frac{(y_{ik} - \mu_{ik})(2\mu_{ik} - 1) - 2\sigma_{ik}^2}{2\sigma_{ik}^3} \frac{\partial \mu_{ik}}{\partial \beta_q} a_{i,jk} \right) \right]. \end{aligned}$$

6.4 A combination of GEE and IFM estimation approach

In section 6.3, we have observed that in some situations GEE provides a slightly more efficient marginal regression parameter estimation than IFM when the correlations of the responses are correctly specified. With the assumption of models, a natural specification of $R_i(\boldsymbol{\alpha})$ is possible. If $R_i(\boldsymbol{\alpha})$

can also be reasonably estimated, then GEE can be applied to obtain the marginal regression parameter estimates. This leads to the new approach for estimating the marginal regression parameters (for some models): i) use IFM approach to estimate model parameters, thus obtain $\tilde{R}_i(\boldsymbol{\alpha})$, ii) use GEE to re-estimate marginal regression parameters. In the following, we provide a few numerical results to illustrate this new approach.

To be more general, we study the situation where regression parameters differ from margin to margin. GEE is extended to this situation. We basically compare GEE marginal estimates (when $\tilde{R}_i(\boldsymbol{\alpha})$ from IFM estimation is used) to IFM estimates. The comparison is carried out by simulation. We assume a multivariate probit model, $Y_{ij} = I(Z_{ij} < \beta_{j0} + \beta_{j1}x_{ij})$, as in section 6.3. The simulation parameters are $d = 3, 4, 5$, $\boldsymbol{\beta}_0 = (0.7, 0, -0.7, 0, 0.5)'$, $\boldsymbol{\beta}_1 = (1, 1.5, 2, 0.5, -0.5)$ with the first 3 components of $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ for $d = 3$, and the first 4 components of $\boldsymbol{\beta}_0, \boldsymbol{\beta}_1$ for $d = 4$, so on. The two situations of covariate x_{ij} are i) discrete where $x_{ij} = I(U \leq 0)$, $U \sim$ uniform$(-1, 1)$, ii) continuous where $x_{ij} \sim N(0, 1/4)$. The latent correlation matrix is an exchangeable correlation matrix with all correlations equal to $\rho = 0.5$. The number of observations is 1000 and the number of simulations for each scenario is 1000. Table 6.17 contains the ratio r ($r^2 = \widehat{\text{MSE}}(\tilde{\theta}_{ifm})/\widehat{\text{MSE}}(\tilde{\theta}_{gee-ifm})$) for a parameter θ, where $\tilde{\theta}_{gee-ifm}$ means the estimate of θ from combined GEE and IFM approaches. The calculation of MSE is defined in section 6.2. The Table 6.17 shows that there is some gain of efficiency with the new approach, since all $r \geq 1$.

Table 6.17: A comparison of IFM to GEE with $\tilde{R}_i(\boldsymbol{\alpha})$ given

margin	1		2		3		4		5	
	β_{10}	β_{11}	β_{20}	β_{21}	β_{30}	β_{31}	β_{40}	β_{41}	β_{50}	β_{51}
	x_{ij} discrete									
$d = 3$	1.036	1.044	1.018	1.028	1.019	1.038				
$d = 4$	1.046	1.047	1.057	1.060	1.042	1.064	1.040	1.082		
$d = 5$	1.063	1.061	1.045	1.046	1.032	1.063	1.058	1.094	1.062	1.111
	x_{ij} continuous									
$d = 3$	1.000	1.055	1.006	1.053	1.003	1.027				
$d = 4$	0.999	1.087	1.005	1.078	1.011	1.064	0.999	1.104		
$d = 5$	1.004	1.101	1.009	1.063	1.011	1.082	1.002	1.101	1.002	1.110

6.5 Summary

In this chapter, we discussed the drawbacks of the GEE in a multivariate analysis framework and examined the efficiency of GEE approach relative to a model based likelihood approach. The purpose of such a study is to partially fill in what is lacking in the statistical literature. Our conclusion is that

GEE is sensitive to the specification of dependence (or correlation) structure; when the specification of dependence is far from the correct one, there is a substantial loss of efficiency with GEE parameter estimation.

The application of GEE to multivariate analysis (longitudinal studies and repeated measures) seems to have grown in relative importance in recent years, but the GEE method does have drawbacks, possible inefficiency, and some assumptions that may be too strong. One should be cautious in the use of GEE, particularly for count data, unless one has a way to assess the assumptions.

Chapter 7

Some further research topics

Many new ideas associated with the construction of multivariate non-normal models, and for estimation and data analysis in multivariate models are advanced in this thesis. The IFM theory for dealing with multivariate models makes the parameters estimation and inference in multivariate non-normal models possible in many situations. More importantly, the research in this thesis may lead to some more potentially fruitful avenues of research. There is much room for further extensions of the ideas in this thesis to general multivariate analysis.

In this final chapter, we mention a variety of research topics of special interest directly related to this thesis work. These topics may include:

1. Comparison of different models and inferences for short and long discrete time series. Long discrete time series situations may include (a) n independent long time series $\mathbf{Y}_i$ $(i = 1, 2, \ldots, n)$, where $\mathbf{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{it_i})$ has length t_i; (b) m correlated times series from a single subject; (c) n independent subjects $(i = 1, \ldots, n)$, and m_i repeated measures are observed on each subject i over a long time period. General MCD and MMD models with IFM inference approach may not be efficient in investigating either the marginal behaviour or dependence structure with these long time series. Adaptation of MCD and MMD to general random effects models together with a relative of the IFM approach can be used in these cases of long time series for each subject. Some applications would be the modelling in environmental studies and health studies of longitudinal or time series nature.

2. Models and inference for mixed multivariate responses (some continuous and some discrete variables). To analyze jointly multivariate discrete and continuous response data, appropriate multivariate models with desirable properties (see Chapter 1) are required as the foundation for inferences.

The analysis of dependence (or associations) between the discrete and continuous response variables would be interesting and important part of the modelling and inference process. There are some obvious extensions of MCD and MMD models for mixed multivariate response variables. Other classes of models based on specified conditional distributions for mixed multivariate response variables may also be promising. The extension of the inference procedures based on IFM for mixed multivariate response variables is also possible. There is interesting potential to develop applications for real life situation. Some recent references on this topic are Catalano and Ryan (1992) and Fitzmaurice and Laird (1995).

3. Models for multinominal categorical responses with covariates. When the polytomous response variables do not have an ordered marginal structure, the existence of a MCD model becomes hard to justify since we are not able to justify the existence of latent continuous variables associated to the response variables. In the univariate situation, Cox (1970) proposed a model for unordered polytomous response variable. When the response variable Y takes m distinct values $y_1, y_2, \ldots, y_m$ and p regressor variables $\mathbf{x} = (x_1, \ldots, x_p)'$, then a model for Y is

$$Pr(Y = y_i|\mathbf{x}) = \frac{\exp(\alpha_i + \boldsymbol{\beta}_i'\mathbf{x})}{\sum_{i=1}^{m} \exp(\alpha_i + \boldsymbol{\beta}_i'\mathbf{x})}, \quad i = 1, \ldots, m,$$

where $\alpha_1 + \boldsymbol{\beta}_1'\mathbf{x}$ is assigned the value 0 for all $\mathbf{x}$ to make the parameters identifiable. Now suppose we have d correlated polytomous response variables (assume the dependence is well defined). Is there any suitable multivariate model for appropriately modelling the marginal behaviour as well as the multivariate dependence structure? What about the extension of MCD models?

4. Extension to multivariate compositional data. Sometimes, the analytical problems of interest to scientists produce data sets that consist essentially of relative proportions and thus are subject to nonnegativity and constant sum constraints. These situations lead to the compositional data. The Dirichlet distribution provides the parametric model of choice when analyzing such data. But the covariance structure associated with Dirichlet random vectors is well-known to be limited to nonpositive. Hence compositional data that exhibit positive correlations cannot be modeled with the Dirichlet. Aitchison (1986) developed classes of logistic normal models partly in response to this shortcoming. Unfortunately, Aitchison's logistic normal classes do not contain the Dirichlet distribution as a special case. As a result, they exhibit interesting dependence structures but are unable to model extreme independence. It is possible to relate the compositional data modelling to the big family of multivariate copula model. The questions are: Can we have models which can model the complicated dependence or complicated independence structure (see Aitchison, 1986)? And what about the appropriate estimation and inference procedures?

Other research topics include: (i) modelling of unequally spaced longitudinal data; (ii) modelling of multivariate data with spatial patterns; (iii) modelling of multivariate directional data; (iv) adaptation of MCD and MMD models and the IFM approach to missing data; and (v) further studies of families of copulas with given bivariate margins (such as Molenberghs-Lesaffre construction).

The above topics may accept some obvious extensions of this thesis work. The research approaches for the above topics to be taken may make use of copula models, latent variables, mixtures, stochastic processes, and point process modelling. Inference can be based on the expansion of IFM approach.

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data.* Chapman and Hall, New York.

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-log normal distribution. *Biometrika*, **76**, 643–653.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *2nd Inter. Symp. on Information Theory*, Petrov, B. N. and Csaki, F. (eds.), Akademiai Kiado, Budapest, 267–281.

Akaike, H. (1977). On entropy maximization principle. In *Application of Statistics*, Krishnaiah (ed.), North-Holland, 27–41.

Al-Osh, M. A. and Aly, E. A. A. (1992). First order autoregressive time series with negative binomial and geometric marginals. *Commun. Statist. A*, **21**, 2483–2492.

Al-Osh, M. A. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR(1)) process. *J. Time Series Anal.*, **8**, 261–275.

Anderson, J. A. and Pemberton, J. D. (1985). The grouped continuous model for multivariate ordered categorical variables and covariate adjustment. *Biometrics*, **41**, 875–885.

Ashford, J. R. and Sowden, R. R. (1970). Multivariate probit analysis. *Biometrics*, **26**, 535–546.

Bahadur, R. R. (1961). A representation of the joint distribution of responses to n dichotomous items. In *Studies in Item Analysis and Prediction*, H. Solomon (ed.). Stanford Mathematical Studies in the Social Sciences VI. Stanford, California: Stanford University Press.

Bonney, G. E. (1987). Logistic regression for dependent binary observations. *Biometrics*, **43**, 951–973.

Bradley, R. A. and Gart, J. J. (1962). The asymptotic properties of ML estimators when sampling from associated populations. *Biometrika*, **49**, 205–213.

Catalano, P. J. and Ryan, L. M. (1992). Bivariate latent variable models for clustered discrete and continuous outcomes. *J. Amer. Statist. Assoc.*, **87**, 651–658.

Chandrasekar, B. (1988). An optimality criterion for vector unbiased statistical estimation functions. *J. Statist. Plann. Inference*, **18**, 115–117.

Chandrasekar, B. and Kale, B. K. (1984). Unbiased statistical estimation functions for the parameters in the presence of nuisance parameters. *J. Statist. Plann. Inference*, **9**, 45–54.

Char, B. W., Geddes, K. O., Gonnet, G. H., Monagan, M. B. and Watt. S. M. (1992). *Maple Reference Manual.* Watcom, Waterloo, Canada.

Conaway, M. R. (1989). Analysis of repeated categorical measurements with conditional likelihood methods. *J. Amer. Statist. Assoc.*, **84**, 53–62.

Connolly, M.A. and Liang, K.-Y. (1988). Conditional logistic regression models for correlated binary data. *Biometrika* **75** 501-506.

Consul, P. C. (1989). *Generalized Poisson Distributions.* Marcel Dekker, New York.

Cox, D. R. (1970). *The Analysis of Binary Data.* Methuen, London.

Cox, D. R. (1972). The analysis of multivariate binary data. *Appl. Statist.*, **21**, 113–120.

Cramèr, H. (1946). *Mathematical Methods of Statistics.* Princeton University Press, Princeton, NJ.

Darlington, G. A. and Farewell, V. T. (1992). Binary longitudinal data analysis with correlation a function of explanatory variables. *Biometrical J.*, **34**, 899–910.

Davis, P. J. and Rabinowitz, P. (1984). *Methods of Numerical Integration*, second edition. Academic Press, Orlando.

Efron, B. and Stein, C. (1981). The jackknife estimate of variance. *Ann. Statist.*, 9, 586–596.

Efron, B. (1982). *The Jackknife, the Bootstrap and Other Resampling Plans.* Society for Industrial and Applied Mathematics, Philadelphia.

Fahrmeir, L. and Kaufmann, H. (1987). Regression models for non-stationary categorical time

series, *J. Time Series Anal.*, **8**, 147–160.

Ferreira, P. E. (1982). Sequential estimation through estimating equations in the nuisance parameter case. *Ann. Statist.*, **10**, 167–173.

Fienberg, S. E., Bromet, E. J., Follman, D., Lambert, D. and May, S. M. (1985). Longitudinal analysis of categorical epidemiological data: a study of Three Mile Island. *Environ. Health Perspectives*, **63**, 241–248.

Fisher, R. A. (1924). The conditions under which χ^2 measures the discrepancy between observation and hypothesis. *J. Roy. Statist. Soc.*, **87**, 442–450.

Fitzmaurice, G. M. and Laird, N. M. (1993). A likelihood-based method for analyzing longitudinal binary responses. *Biometrika*, **80**, 141–151.

Fitzmaurice, G. M. and Laird, N. M. (1995). Regression models for a bivariate discrete and continuous outcome with clustering. *J. Amer. Statist. Assoc.*, **90**, 845–852.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, **13**, 317–322.

Gardner, W. (1990). Analyzing sequential categorical data: individual variation in Markov chains. *Psychometrika*, **55**, 263–275.

Genest, C. and MacKay, R. J. (1986). Copules archimédiennes et familles de lois bidimensionelles dont les marges sont données. *Canad. J. Statist.*, **14**, 145–159.

Glonek, G. F. V. and McCullagh, P. (1995). Multivariate logistic models. *J. R. Statist. Soc. B*, **57**, 533–546.

Godambe, V. P. (1960). An optimal property of regular maximum likelihood estimation. *Ann. Math. Statist.*, **31**, 1208–1211.

Godambe, V. P. (1976). Conditional likelihood and unconditional optimum estimating equations. *Biometrika*, **63**, 277–284.

Godambe, V. P. (1991). *Estimating Functions*. Oxford University Press, New York.

Goodman, L. A. and Kruskal, W. H. (1954). Measures of association for cross classifications. *J. Amer. Statist. Assoc.*, **49**, 732–764.

Hoadley, B. (1971). Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case. *Ann. Math. Statist.*, **42**, 1977–1991.

Joe, H. (1993). Parametric family of multivariate distributions with given margins. *J. Multivariate Anal.*, **46**, 262–282.

Joe, H. (1994). *Lecture notes*, Course given at Department of Mathematics and Statistics, University of Pittsburgh, Pittsburgh, USA.

Joe, H. (1994). Multivariate extreme-value distributions with applications to environmental data. *Canad. J. Statist.*, **22**, 47–64.

Joe, H. (1995). Approximations to multivariate normal rectangle probabilities based on conditional expectations. *J. Amer. Statist. Assoc.*, **90**, 957–964.

Joe, H. (1996). *Multivariate Models and Dependence Concepts*, Draft book and Stat 521 course notes. Department of Statistics, University of British Columbia, Vancouver, Canada.

Joe, H. (1996a). Families of m-variate distributions with given margins and $m(m-1)/2$ bivariate dependence parameters. In *Distributions with fixed marginals, doubly stochastic measures and Markov operators*, Sherwood, H. and Taylor, M. (eds.), IMS Lecture Notes – Monograph Series, Hayward, CA.

Joe, H. (1996b). Time series models with univariate margins in the convolution–closed infinitely divisible class. *J. Appl. Probab.*, to appear.

Joe, H. and Hu, T. (1996). Multivariate distributions from mixtures of max-infinitely divisible distributions. *J. Multivariate Anal.*, **57**, 240–265.

Jϕrgensen B. and Labouriau, R. S. (1995). *Exponential Families and Theoretical Inference*, Lecture notes. Department of Statistics, University of British Columbia, Vancouver, Canada.

Johnson, N. L. and S. Kotz (1975). On some generalized Farlie-Gumbel-Morgenstern distributions. *Communication in Statistics*, **4**, 415–427.

Johnson, N. L. and S. Kotz (1977). On some generalized Farlie-Gumbel-Morgenstern distributions – II Regression, correlation and further generalizations. *Communication in Statistics*, A, **6**, 485–496.

Joseph, B. and Durairajan, T. M. (1991). Equivalence of various optimality criteria for estimating functions. *J. Statist. Plann. Inference*, **27**, 355–360.

Kimeldorf, G. and Sampson, A. R. (1975). Uniform representations of bivariate distributions. *Comm. Stat.-Theor. Meth.*, **4**, 617–628.

Lesaffre, E. and Molenberghs, G. M. (1991). Multivariate probit analysis: a neglected procedure in medical statistics. *Stat. in Medicine*, **10**, 1391–1403.

Kocherlakota, S. and Kocherlakota, K. (1992). *Bivariate Discrete Distributions*. Dekker, New York.

Lawless, J. F. (1987). Negative binomial and mixed Poisson regression. *Canad. J. Statist.*, **15**, 209–225.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Liang, K. Y., Zeger, S. L. and Qaqish, B. (1992). Multivariate regression analysis for categorical data. *J. Roy. Statist. Soc.*, **B, 54**, 3–40.

Lipsitz, S. R., Dear, K. B. G. and Zhao, L. (1994). Jackknife estimators of variance for parameter estimates from estimating equations with applications to clustered survival data. *Biometrics*, **50**, 842–846.

Mahamunulu, D. M. (1967). A note on regression in the multivariate Poisson distribution. *J. Amer. Statist. Assoc.*, **62**, 251–258.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, second edition. Chapman and Hall, London.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions. *Adv. Appl. Probab.*, **18**, 679–705.

McKenzie, E. (1988). Some ARMA models for dependent sequences of Poisson counts. *Adv. Appl.*

Probab., **20**, 822–835.

McLeish, D. L. and Small, C. G. (1988). *The Theory and Applications of Statistical Inference Functions.* Lecture Notes in Statistics 44, Springer-Verlag, New York.

Meester, S. G. and MacKay, J. (1994). A parametric model for cluster correlated categorical data. *Biometrics*, **50**, 954–963.

Miller, R. G. (1974). The jackknife – a review. *Biometrika*, **61**, 1–15.

Molenberghs, G. M. and Lesaffre, E. (1994). Marginal modeling of correlated ordinal data using a multivariate Plackett distribution. *J. Amer. Statist. Assoc.*, **89**, 633–644.

Morgenstern, D. (1956). Einfache Beispeile zweidimensionaler Verteilungen. *Mitteilungsblatt für Mathematische Statistik*, **8**, 234–235.

Muenz, L.R .and Rubinstein, L.V. (1985). Markov models for covariate dependence of binary sequences. *Biometrics* **41** 91-101.

Nash, J. C. (1990). *Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation*, second edition. Hilger, New York.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized linear models. *J. Roy. Statist. Soc. A*, **135**, 370–384.

Prentice, R. L. (1986). Binary regression using an extended beta-binomial distribution, with discussion of correlation induced by covariate measurement errors. *J. Amer. Statist. Assoc.*, **81**, 321–327.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, **44**, 1033–1048.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory.* Clarendon Press, Oxford.

Quenouille, M. H.(1956). Notes on bias in estimation. *Biometrika*, **43**, 353–360.

Rao, C. R. (1973). *Linear statistical inference and its applications.* 2nd ed. Wiley, New York.

Read, T. R. C. and Cressie, N. A. C. (1988). *Goodness-of-fit Statistics for Discrete Multivariate*

Data. Springer-Verlag, New York.

Rousseeuw, P. J. and Molenberghs, G. (1994). The shape of correlation matrices. *The American Statistician*, **48**, 276–279.

Sakamoto, Y., Ishiguro, M. and Kitagawa, G. (1986). *Akaike Information Criterion Statistics*. KTK Scientific Publishers, Tokyo.

Schervish, M. J. (1984). Multivariate normal probabilities with error bound. *Appl. Statist.*, **33**, 81–87.

Schweizer, B. and Wolff, E. F. (1981). On nonparametric measures of dependence for random variables. *Ann. Statist.*, **9**, 879–885.

Seber, G. A. F. (1984). *Multivariate Observation*. Wiley, New York.

Sen, P. K. and Singer, J. M. (1993). *Large Sample Methods in Statistics*. Chapman & Hall, New York.

Serfling, R. J. (1980). *Approximation Theorems of Mathematical Statistics*. Wiley, New York.

Sklar, A. (1959). Fonction de répartition à n dimensions et leurs marges. *Publ. Inst. Statist. Univ. Paris*, **8**, 229–231.

Stein, G. Z., Zucchini, W. and Juritz, J. M. (1987). Parameter estimation for the Sichel distribution and its multivariate extension. *J. Amer. Statist. Assoc.*, **82**, 938–944.

Stram, D. O., Wei, L. J. and Ware, J. H. (1988). Analysis of repeated ordered categorical outcomes with possibly missing observations and time-dependent covariates. *J. Amer. Statist. Assoc.*, **83**, 631–637.

Teicher, H. (1954). On the multivariate Poisson distribution. *Skandinavisk Aktuarietidskrift*, **37**, 1–9.

Thorburn, D. (1976). Some asymptotic properties of jackknife statistics. *Biometrika*, **63**, 305–313.

Tong, Y. L. (1990). *The Multivariate Normal Distribution*. Springer-Verlag, New York.

Tukey, J. W. (1958). Bias and confidence in not quite large samples. Abstract in *Ann. Math.*

Statist., **29**, 614.

Ware, J. H., Dockery, D. W., Spiro, A. Speizer, F. E. and Ferris, B. G., Jr. (1984). Passive smoking, gas Cooking, and respiratory health of children living in six cities. *American Review of Respiratory Disease*, **129**, 366–374.

Zeger, S. L. and Liang, K. Y. (1986). Longitudinal data analysis for discrete and continuous outcomes. *Biometrics*, **42**, 121–130.

Zeger, S. L., Liang, K. Y. and Albert, P. S. (1988). Models for longitudinal data: A generalized estimating equation approach. *Biometrics*, **44**, 1049–1060.

Zeger, S.L., Liang, K.-Y. and Self, S.G. (1985). The analysis of binary longitudinal data with time-independent covariates. *Biometrika*, **72**, 31–38.

Zhao, L. P. and Prentice, R. L. (1990). Correlated binary regression using a quadratic exponential model. *Biometrika*, **77**, 642–648.

Appendix A

Maple programs

This appendix contains a program written in Maple for Example 4.3 in Chapter 4.

```
g12p11 := 1/4+1/(2*pi)*arcsin(r12);
g12p00 := g12p11; g12p10 := 1/2 - g12p11; g12p01 := g12p10;
dg12p11 := diff(g12p11, r12); dg12p00 := dg12p11;
dg12p10 := diff(g12p10, r12); dg12p01 := dg12p10;

g13p11 := 1/4+1/(2*pi)*arcsin(r13);
g13p00 := g13p11; g13p10 := 1/2 - g13p11; g13p01 := g13p10;
dg13p11 := diff(g13p11, r13); dg13p00 := dg13p11;
dg13p10 := diff(g13p10, r13); dg13p01 := dg13p10;

g23p11 := 1/4+1/(2*pi)*arcsin(r23);
g23p00 := g23p11; g23p10 :=1/2 - g23p11;  g23p01 := g23p10;
dg23p11 := diff(g23p11, r23); dg23p00 := dg23p11;
dg23p10 := diff(g23p10, r23); dg23p01 := dg23p10;

gp111 := 1/8+1/(4*pi)*(arcsin(r12)+arcsin(r13)+arcsin(r23));
gp110 := g12p11-gp111; gp011 := g23p11-gp111; gp101 := g13p11-gp111;
gp001 := g23p01-gp101; gp100 := g12p10-gp101; gp010 := g12p01-gp011;
```

```
gp000 := 1-gp111-gp110-gp011-gp101-gp001-gp100-gp010;

I11 := 1/gp111*diff(gp111,r12)^2+1/gp110*diff(gp110,r12)^2
       +1/gp101*diff(gp101,r12)^2+1/gp011*diff(gp011,r12)^2
       +1/gp100*diff(gp100,r12)^2+1/gp001*diff(gp001,r12)^2
       +1/gp010*diff(gp010,r12)^2+1/gp000*diff(gp000,r12)^2;
I22 := 1/gp111*diff(gp111,r13)^2+1/gp110*diff(gp110,r13)^2
       +1/gp101*diff(gp101,r13)^2+1/gp011*diff(gp011,r13)^2
       +1/gp100*diff(gp100,r13)^2+1/gp001*diff(gp001,r13)^2
       +1/gp010*diff(gp010,r13)^2+1/gp000*diff(gp000,r13)^2;
I33 := 1/gp111*diff(gp111,r23)^2+1/gp110*diff(gp110,r23)^2
       +1/gp101*diff(gp101,r23)^2+1/gp011*diff(gp011,r23)^2
       +1/gp100*diff(gp100,r23)^2+1/gp001*diff(gp001,r23)^2
       +1/gp010*diff(gp010,r23)^2+1/gp000*diff(gp000,r23)^2;
I12 := 1/gp111*diff(gp111,r12)*diff(gp111,r13)
       +1/gp110*diff(gp110,r12)*diff(gp110,r13)
       +1/gp101*diff(gp101,r12)*diff(gp101,r13)
       +1/gp011*diff(gp011,r12)*diff(gp011,r13)
       +1/gp100*diff(gp100,r12)*diff(gp100,r13)
       +1/gp001*diff(gp001,r12)*diff(gp001,r13)
       +1/gp010*diff(gp010,r12)*diff(gp010,r13)
       +1/gp000*diff(gp000,r12)*diff(gp000,r13);
I13 := 1/gp111*diff(gp111,r12)*diff(gp111,r23)
       +1/gp110*diff(gp110,r12)*diff(gp110,r23)
       +1/gp101*diff(gp101,r12)*diff(gp101,r23)
       +1/gp011*diff(gp011,r12)*diff(gp011,r23)
       +1/gp100*diff(gp100,r12)*diff(gp100,r23)
       +1/gp001*diff(gp001,r12)*diff(gp001,r23)
       +1/gp010*diff(gp010,r12)*diff(gp010,r23)
       +1/gp000*diff(gp000,r12)*diff(gp000,r23);
I23 := 1/gp111*diff(gp111,r13)*diff(gp111,r23)
       +1/gp110*diff(gp110,r13)*diff(gp110,r23)
```

```
       +1/gp101*diff(gp101,r13)*diff(gp101,r23)
       +1/gp011*diff(gp011,r13)*diff(gp011,r23)
       +1/gp100*diff(gp100,r13)*diff(gp100,r23)
       +1/gp001*diff(gp001,r13)*diff(gp001,r23)
       +1/gp010*diff(gp010,r13)*diff(gp010,r23)
       +1/gp000*diff(gp000,r13)*diff(gp000,r23);
I11 := simplify(I11); I22 := simplify(I22); I33 := simplify(I33);
I12 := simplify(I12); I13 := simplify(I13); I23 := simplify(I23);

E11 := 1/g12p11*dg12p11^2+1/g12p10*dg12p10^2
       +1/g12p01*dg12p01^2+1/g12p00*dg12p00^2;
E22 := 1/g13p11*dg13p11^2+1/g13p10*dg13p10^2
       +1/g13p01*dg13p01^2+1/g13p00*dg13p00^2;
E33 := 1/g23p11*dg23p11^2+1/g23p10*dg23p10^2
       +1/g23p01*dg23p01^2+1/g23p00*dg23p00^2;
E12 := gp111/(g12p11*g13p11)*dg12p11*dg13p11
       +gp110/(g12p11*g13p10)*dg12p11*dg13p10+gp101/(g12p10*g13p11)*dg12p10*dg13p11
       +gp100/(g12p10*g13p10)*dg12p10*dg13p10+gp011/(g12p01*g13p01)*dg12p01*dg13p01
       +gp010/(g12p01*g13p00)*dg12p01*dg13p00+gp001/(g12p00*g13p01)*dg12p00*dg13p01
       +gp000/(g12p00*g13p00)*dg12p00*dg13p00;
E13 := gp111/(g12p11*g23p11)*dg12p11*dg23p11
       +gp110/(g12p11*g23p10)*dg12p11*dg23p10+gp101/(g12p10*g23p01)*dg12p10*dg23p01
       +gp100/(g12p10*g23p00)*dg12p10*dg23p00+gp011/(g12p01*g23p11)*dg12p01*dg23p11
       +gp010/(g12p01*g23p10)*dg12p01*dg23p10+gp001/(g12p00*g23p01)*dg12p00*dg23p01
       +gp000/(g12p00*g23p00)*dg12p00*dg23p00;
E23 := gp111/(g13p11*g23p11)*dg13p11*dg23p11
       +gp110/(g13p10*g23p10)*dg13p10*dg23p10+gp101/(g13p11*g23p01)*dg13p11*dg23p01
       +gp100/(g13p10*g23p00)*dg13p10*dg23p00+gp011/(g13p01*g23p11)*dg13p01*dg23p11
       +gp010/(g13p00*g23p10)*dg13p00*dg23p10+gp001/(g13p01*g23p01)*dg13p01*dg23p01
       +gp000/(g13p00*g23p00)*dg13p00*dg23p00;
E11 := simplify(E11); E22 := simplify(E22); E33 := simplify(E33);
E12 := simplify(E12); E13 := simplify(E13); E23 := simplify(E23);
```

```
with(linalg);
I := matrix(3,3, [I11,I12,I13,I12,I22,I23,I13,I23,I33]);
Iinv := evalm(I^(-1));
M:= matrix(3,3, [E11,E12,E13,E12,E22,E23,E13,E23,E33]);
Minv := evalm(M^(-1));
D := matrix(3,3,[E11,0,0,0,E22,0,0,0,E33]);
Dinv := evalm(D^(-1));
Jinv := evalm(Dinv &* M &* Dinv);
map(simplify,evalm(Jinv-Iinv));
det(evalm(Jinv-Iinv));
```