On Dual Empirical Likelihood Inference under Semiparametric Density Ratio Models in the Presence of Multiple Samples

With Applications to Long Term Monitoring of Lumber Quality

by

Song Cai

B.Sc. in Atmospheric Science, Peking University, 1999
M.Sc. in Atmospheric Science, The University of British Columbia, 2008
M.Sc. in Statistics, The University of British Columbia, 2010

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

in

The Faculty of Graduate and Postdoctoral Studies

(Statistics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

May 2014

Abstract

Maintaining a high quality of lumber products is of great social and economic importance. This thesis develops theories as part of a research program aimed at developing a long term program for monitoring change in the strength of lumber. These theories are motivated by two important tasks of the monitoring program, testing for change in strength populations of lumber produced over the years and making statistical inference on strength populations based on Type I censored lumber samples. Statistical methods for these inference tasks should ideally be efficient and nonparametric. These desiderata lead us to adopt a semiparametric density ratio model to pool the information across multiple samples and use the nonparametric empirical likelihood as the tool for statistical inference.

We develop a dual empirical likelihood ratio test for composite hypotheses about the parameter of the density ratio model based on independent samples from different populations. This test encompasses testing differences in population distributions as a special case. We find the proposed test statistic to have a classical chi–square null limiting distribution. We also derive the power function of the test under a class of local alternatives. It reveals that the local power is often increased when strength is borrowed from additional samples even when their underlying distributions are unrelated to the hypothesis of interest. Simulation studies show that this test has better power properties than all potential competitors adopted to the multiple sample problem under the investigation, and is robust to model misspecification. The proposed test is then applied to assess strength properties of lumber with intuitively reasonable implications for the forest industry.

We also establish a powerful inference framework for performing empirical likelihood inference under the density ratio model when Type I censored samples are present. This inference framework centers on the maximization of a concave dual partial empirical likelihood function, and features an easy computation. We study the properties of this dual partial empirical likelihood, and find its corresponding likelihood ratio test to have a simple chi–square limiting distribution under the null model and a non–central chi–square limiting distribution under local alternatives.

Preface

This thesis is written up under the supervisions of Dr. Jiahua Chen and Dr. James V. Zidek. Chapter 3 and 4 are based on a submitted manuscript coauthored with them. Dr. Chen initiated the idea of using the density ratio model as the platform for inference and suggested the direction of studying the dual empirical likelihood ratio test. Following his advice, I worked out all the theoretical results, wrote the initial draft of the manuscript, and conducted most of the follow–up revisions. During this process, Dr. Chen had given me valuable constructive criticism, helped me to organize my ideas and improve my proofs, and taken up a few rounds of revisions. Dr. Zidek supported, encouraged and inspired me during the writing of the manuscript, gave me helpful suggestions, and revised and greatly improved the final draft.

Chapter 5 is based on a manuscript in preparation coauthored with Dr. Jiahua Chen. Dr. Chen suggested the topic of the manuscript. I independently developed all the theory and wrote the manuscript. Dr. Chen helped me to improve some details of the theory.

Chapter 6 is based on an R software package called *drmdel* that has been published on The Comprehensive R Archive Network. I independently designed and wrote the entire package.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AD Anderson–Darling test

ANOVA Analysis of variance

ASTM American Society for Testing and Materials

BFGS Broyden–Fletcher–Goldfarb–Shanno algorithm

CDF Cumulative distribution function

CRAN The Comprehensive R Archive Network

DEL Dual empirical likelihood

DELR Dual empirical likelihood ratio

DELRT Dual empirical likelihood ratio test

DPEL Dual partial empirical likelihood

DRM Density ratio model

EL Empirical likelihood

ELR Empirical likelihood ratio

IQR Interquartile range

KW Kruskal–Wallis rank–sum test

LHS Left hand side

MELE Maximum empirical likelihood estimator

MPELE Maximum partial empirical likelihood estimator

MWELE Maximum weighted empirical likelihood estimator

MOR Modulus of rupture

MOT Modulus of tension

PEL Partial empirical likelihood

RHS Right hand side

WEL Weighted empirical likelihood

Acknowledgements

I am deeply indebted to my principle supervisor, Dr. Jiahua Chen, who has inspired and guided me during the course of my Ph.D. study. Dr. Chen gave me lots of constructive criticism that cannot be heard from anyone else, and helped me to sharpen my view and form the shape of my research. From him, I not only have learned skills and knowledge but also a critical and serious attitude toward research. Thank you, Jiahua, for guiding me and pushing me toward a high research standard.

I also owe my deepest gratitude to my co–supervisor, Dr. James V. Zidek. Since the day I started my study in Statistics as a Master's student, he has been a great supervisor, a mentor, and a friend. He has been supporting me spiritually, financially, and in every other way. Thank you, Jim, for all you have done for me!

In addition, I would like to thank Dr. Yukun Liu, a friend and colleague, who had helped me with the theory of empirical likelihood inference and discussed with me frequently about my research when I started my work presented in this thesis.

Most importantly, I thank my dear wife and parents–in–law for their wholehearted love and support. They have dedicated all their time and energy to help me take care of my young children, one born when I started my PhD, and the other, when I was in the third year of my PhD study. They also have encouraged me and brought me much joy and happiness.

Without the help from my academic fathers, Dr. Zidek and Dr. Chen, or the help from my dear friend, Dr. Liu, or the support from my beloved family, the writing of this thesis would never have been possible.

To Zhao, Tiejun, Tieli, Brendan, and Heather

Chapter 1

Introduction

1.1 Application background

With nearly half of Canada's entire land surface covered by trees, lumber has been a vital natural resource and major construction material for this country. Maintaining a high quality of lumber products hence is of great social and economic importance: it is crucial to construction safety as well as to the Canadian lumber industry. This thesis develops statistical methods as part of a research project aimed at developing a program for monitoring change in the strength of lumber. Interest in such a program also has been sparked by climate change, which will affect the way trees grow, as well by the changing resource mix, for example due to increasing reliance on plantation lumber.

The monitoring program includes the following two important tasks. First, it is crucial to examine whether the overall strength of lumber changes over time, which translates into a hypothesis testing problem for detecting differences in the population distributions of the strength of lumber from different years. Second, it is desirable to make smart strength test plans to maximize the scientific value of each piece of lumber: collecting and testing lumber costs time and money; for example lumber must be conditioned in the lab over a period of months before being destructively tested. To achieve this goal, one way would be to collect Type I right–censored lumber strength samples: we stop a destructive strength test, say bending strength test, at a prespecified level such that not all the lumber is broken; the unbroken lumber then can be used afterwards for other strength tests, say tension strength

test, etc. The sample collect in the first test, although censored, is a representative sample of bending strength, and can be used to infer population characteristics of that strength. The sample collect in the second test is useful for studying the relationship between the two different strengths.

Echoing the above tasks, this thesis develops statistical methods aimed at the following inference goals: (i) testing hypotheses about parameters of a number of different population distributions, given a random sample from each; and (ii) establishing an inference framework for estimation and hypothesis testing problems concerning several different populations based on Type I censored multiple samples.

1.2 Density ratio models: concepts and examples

Desiderata for the statistical methods used in the long term monitoring program of lumber includes two key goals. First, the methods must be efficient to reduce the sizes of the required samples. Moreover, the reduced effective sample size of Type I censored samples demands highly efficient statistical methods to reach required estimation accuracy and hypothesis testing power. Toward the goal of efficiency, this thesis proposes methods that borrow strength among lumber samples by exploiting an obvious feature of the resource, that distinct populations of lumber over years, species, regions and so on will share some latent strength characteristics. Second, the methods should ideally be nonparametric in accordance with the well–ingrained practice in setting standards for forest products like those in American Society for Testing and Materials (ASTM) protocols (ASTM D1990 – 07).

These desiderata, lead to the semiparametric *density ratio model* (DRM) adopted in this thesis. In the targeted application, we have multiple lumber strength samples collected from different years. While the size of the sample from each population may be small because of the high data collection costs,

the total sample size could be large. If we can pool the information from different samples, efficient inference based on the pooled sample then can be expected. Since the lumber samples from different years share some common physical characteristics, it is reasonable to assume that the corresponding distribution functions have a certain relationship. In particular, we assume that the lumber quality populations connect with each other through their density functions. Suppose that we have $m+1$ independent random samples from populations with cumulative distribution functions (CDFs) $F_k(x)$, $k = 0, 1, \ldots, m$, with the same support. The DRM assumes that

$$dF_k(x) = \exp\left\{\alpha_k + \boldsymbol{\beta}_k^{\mathsf{T}}\boldsymbol{q}(x)\right\}dF_0(x), \text{ for } k = 1, 2, \ldots, m,$$

where $\boldsymbol{q}(x)$, which we call the *basis function* of the DRM, is a prespecified d–dimensional function, and $(\alpha_k, \boldsymbol{\beta}_k)$ are model parameters. But the baseline distribution $F_0(x)$ is completely unspecified.

The DRM assumption serves as a device for pooling information across samples, while the nonparametric baseline distribution keeps the model flexible. In fact, the DRM covers a large range of distribution families, and we demonstrate this with typical examples in the following subsections.

1.2.1 Exponential families of distributions

Every exponential family of distributions satisfies the assumption of the DRM. A family of distributions is called an exponential family if it has a density of the form

$$f(x;\, \boldsymbol{\vartheta}) = k(x)\exp\{\boldsymbol{\eta}^{\mathsf{T}}(\boldsymbol{\vartheta})\boldsymbol{t}(x) - A(\boldsymbol{\vartheta})\},\ x \in \mathcal{S},$$

where $\boldsymbol{\vartheta}$ is a parameter vector, $k(\cdot)$, $\boldsymbol{\eta}(\cdot)$, $\boldsymbol{t}(\cdot)$ and $A(\cdot)$ are given functions, and $\mathcal{S}$, the support of X, does not depend upon $\boldsymbol{\vartheta}$. Densities, $\{f_k(x)\}$, of the

same exponential family with parameter values $\{\boldsymbol{\vartheta}_k\}$ satisfy

$$f_k(x) = \exp\left\{\left(\boldsymbol{\eta}^\intercal(\boldsymbol{\vartheta}_k) - \boldsymbol{\eta}^\intercal(\boldsymbol{\vartheta}_0)\right)\boldsymbol{t}(x) + \left(A(\boldsymbol{\vartheta}_0) - A(\boldsymbol{\vartheta}_k)\right)\right\} f_0(x).$$

This relationship shows that distributions from a same exponential family fulfill the DRM assumption with basis function $\boldsymbol{q}(x) = \boldsymbol{t}(x)$ and parameters

$$\alpha_k = A(\boldsymbol{\vartheta}_0) - A(\boldsymbol{\vartheta}_k), \quad \boldsymbol{\beta}_k = \boldsymbol{\eta}^\intercal(\boldsymbol{\vartheta}_k) - \boldsymbol{\eta}^\intercal(\boldsymbol{\vartheta}_0).$$

In order to define an exponential family, the function $k(x)$ must be completely specified. In DRM, the baseline density function is the counterpart of $k(x)$, however, it is a non–parametric component of the model. This is the fundamental difference between the parametric exponential family and the semiparametric DRM, and also the reason that the DRM encompasses each exponential family as a special case.

Example 1.1. *The gamma distribution family with shape λ and rate κ, $\Gamma(\lambda,\, \kappa)$, is an exponential family with*

$$\boldsymbol{\eta}(\lambda,\, \kappa) = (-\kappa,\, \lambda - 1)^\intercal\,,\ \boldsymbol{t}(x) = \left(x,\, \ln x\right)^\intercal,\ A(\lambda,\, \kappa) = \ln\Gamma(\lambda) - \lambda \ln \kappa,$$

where $\Gamma(\cdot)$ is the gamma function. Therefore, gamma distributions with parameter values $\{(\lambda_k,\, \kappa_k)\}$ satisfy the DRM assumption with basis function $\boldsymbol{q}(x) = (x,\, \ln x)^\intercal$ and parameters

$$\alpha_k = \ln\frac{\Gamma(\lambda_0)}{\Gamma(\lambda_k)} + \lambda_k \ln \kappa_k - \lambda_0 \ln \kappa_0, \quad \boldsymbol{\beta}_k = (\kappa_0 - \kappa_k,\ \lambda_k - \lambda_0)^\intercal.$$

Example 1.2. *The normal distribution family with mean μ and standard deviation σ, $N(\mu,\, \sigma^2)$, is an exponential family with*

$$\boldsymbol{\eta}(\mu,\, \sigma) = \left(\frac{\mu}{\sigma^2},\, -\frac{1}{2\sigma^2}\right)^\intercal,\ \boldsymbol{t}(x) = \left(x,\, x^2\right)^\intercal,\ A(\mu,\, \sigma) = \frac{\mu^2}{2\sigma^2} + \ln \sigma.$$

Hence, normal distributions with parameter values $\{(\mu_k,\, \sigma_k)\}$ *satisfy the DRM assumption with basis function* $\boldsymbol{q}(x) = (x,\, x^2)^\intercal$ *and parameters*

$$\alpha_k = \ln\frac{\sigma_0}{\sigma_k} + \frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_k^2}{2\sigma_k^2}, \quad \boldsymbol{\beta}_k = \left(\frac{\mu_k}{\sigma_k^2} - \frac{\mu_0}{\sigma_0^2},\, \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_k^2}\right)^\intercal.$$

Example 1.3. *The log–normal distribution family with mean* μ *and standard deviation* σ *on the log–scale,* $LN(\mu,\, \sigma^2)$*, is an exponential family with*

$$\boldsymbol{\eta}(\mu,\, \sigma) = \left(\frac{\mu}{\sigma^2},\, -\frac{1}{2\sigma^2}\right)^\intercal,\ \boldsymbol{t}(x) = \big(\ln x,\, (\ln x)^2\big)^\intercal,\ A(\mu,\, \sigma) = \frac{\mu^2}{2\sigma^2} + \ln\sigma.$$

Hence, log–normal distributions with parameter values $\{(\mu_k,\, \sigma_k)\}$ *saftisfies the DRM assumption with basis function* $\boldsymbol{q}(x) = \big(\ln x,\, (\ln x)^2\big)^\intercal$ *and parameters*

$$\alpha_k = \ln\frac{\sigma_0}{\sigma_k} + \frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_k^2}{2\sigma_k^2}, \quad \boldsymbol{\beta}_k = \left(\frac{\mu_k}{\sigma_k^2} - \frac{\mu_0}{\sigma_0^2},\, \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_k^2}\right)^\intercal.$$

Example 1.4. *The Weibull distribution family with known shape* λ *and unknown scale parameter* κ*,* $W(\kappa)$*, is an exponential family with*

$$\boldsymbol{\eta}(\kappa) = -\frac{1}{\kappa^\lambda},\ \boldsymbol{t}(x) = x^\lambda,\ A(\kappa) = \lambda\ln\kappa - \ln\lambda.$$

Weibull distributions with known common shape λ *and different scales* $\{\kappa_k\}$ *then satisfy the DRM assumption with basis function* $\boldsymbol{q}(x) = x^\lambda$ *and parameters*

$$\alpha_k = \lambda\ln\frac{\kappa_0}{\kappa_k}, \quad \boldsymbol{\beta}_k = \frac{1}{\kappa_0^\lambda} - \frac{1}{\kappa_k^\lambda}.$$

Example 1.5. *The Pareto distribution family with location* x_{min} *and shape* λ*,* $P(x_{min},\, \lambda)$*, has the density of the form*

$$f(x) = \exp\big\{(\ln\lambda + \lambda\ln x_{min}) - (\lambda+1)\ln x\big\},\ \textit{for}\ \ x \geq x_{min} > 0\ \textit{and}\ \ \lambda > 0.$$

When x_{min} is fixed, the Pareto family is an exponential family. Pareto distributions with common location x_{min} and different shapes $\{\lambda_k\}$ satisfy the DRM assumption with basis function $\boldsymbol{q}(x) = \ln x$ and parameters

$$\alpha_k = \ln \frac{\lambda_k}{\lambda_0} + (\lambda_k - \lambda_0) \ln x_{min}, \quad \boldsymbol{\beta}_k = \lambda_0 - \lambda_k.$$

In addition to the exponential families, the DRM also naturally arises from many other statistical models, such as logistic regression models and biased sampling models, as described in the following subsections.

1.2.2 Relationship between logistic regression models and DRMs

The logistic regression model in case–control studies has a close relationship with the two–sample DRM (Qin and Zhang, 1997). A case–control study is to identify factors that may contribute to a medical condition by comparing two groups of individuals: those with the disease (case group) and those without the disease (control group). Let Y be the group indicator variable with 0 being control and 1 being case, and X be the random vector representing the exposures for an individual. A classical model for case–control data is the logistic regression model, e.g. used by Prentice and Pyke (1979), Farewell (1979) and Mantel (1973),

$$\Pr(Y = 1|X = x) = \frac{\exp(a + \boldsymbol{b}^\intercal x)}{1 + \exp(a + \boldsymbol{b}^\intercal x)},$$

where a and $\boldsymbol{b}$ are unknown parameters. Suppose the exposures X has unspecified marginal density $f(x)$. Denote the conditional density of exposures

X given group $Y = i$, $i = 0, 1$, as $f_i(x)$. Then by Bayes' rule, we have,

$$f_0(x) = \frac{\Pr(Y = 0|X = x)f(x)}{\Pr(Y = 0)} = \frac{f(x)}{\Pr(Y = 0)\{1 + \exp(a + \boldsymbol{b}^\intercal x)\}}$$
$$f_1(x) = \frac{\Pr(Y = 1|X = x)f(x)}{\Pr(Y = 1)} = \frac{\exp(a + \boldsymbol{b}^\intercal x)f(x)}{\Pr(Y = 1)\{1 + \exp(a + \boldsymbol{b}^\intercal x)\}}.$$

Hence the conditional densities of exposures X given the group $Y = 0, 1$ constitute a two–sample DRM with basis function $\boldsymbol{q}(x) = x$, $\alpha = a + \log\{\Pr(Y = 0)/\Pr(Y = 1)\}$, and $\boldsymbol{\beta} = \boldsymbol{b}$:

$$f_1(x) = \exp(\{a + \log\{\Pr(Y = 0)/\Pr(Y = 1)\}\} + \boldsymbol{b}^\intercal x)f_0(x).$$

As in the case of binary case–control data, if categorical data with categories $Y = 0, 1, \ldots, m$ and covariates X satisfy the multinomial logit model assumption,

$$\frac{\Pr(Y = k|X = x)}{\Pr(Y = 0|X = x)} = \exp(a_k + \boldsymbol{b}_k^\intercal x),\ k = 1, 2, \ldots, m,$$

where a_k and $\boldsymbol{b}_k$, $k = 1 \ldots, m$, are unknown parameters, then the conditional density $f_k(x)$ of covariates X given category $Y = k$, fulfills the DRM assumption with basis function $\boldsymbol{q}(x) = x$, $\alpha_k = a_k + \log\{\Pr(Y = 0)/\Pr(Y = k)\}$, $k = 1, 2, \ldots, m$, and $\boldsymbol{\beta}_k = \boldsymbol{b}_k$:

$$f_k(x) = \exp(\{a_k + \log\{\Pr(Y = 0)/\Pr(Y = k)\}\} + \boldsymbol{b}_k^\intercal x)f_0(x).$$

1.2.3 Biased sampling and DRM

Another example of the DRM is the biased sampling model. Selection bias happens in survey sampling if not every unit in population is given an equal chance to enter the sample. For example, in a survey of hospital patients, patients with longer visits may have a greater chance to be sampled than

those with shorter visits; in a survey of life times of light bulbs, the bulbs with longer lives may be more likely to be chosen than those with shorter lives. In such cases, the distribution of the units in a sample differs from the population distribution. Let $F(x)$ denote the population distribution function. Let $w(x;\, \boldsymbol{\vartheta})$ be a weighting function specifying the "chance" of a unit with value x being chosen, where $\boldsymbol{\vartheta}$ is an unknown parameter. The density, $dG(x)$, of the sampled units is

$$dG(x) = \frac{w(x;\, \boldsymbol{\vartheta})dF(x)}{\int w(x;\, \boldsymbol{\vartheta})dF(x)}.$$

This weighted density model can be extended to the multiple sample case. Suppose we have $m+1$ samples, labeled as $0,\, 1,\, \ldots,\, m$, from unknown distributions, and the population distribution function of the kth sample, F_k, is a weighted version of the population distribution F_0 of sample 0, i.e.

$$F_k(x) = \frac{\int_{-\infty}^{x} w_k(t;\, \boldsymbol{\vartheta})dF_0(t)}{\int_{-\infty}^{\infty} w_k(t;\, \boldsymbol{\vartheta})dF_0(t)},\ k = 1,\, \ldots,\, m, \tag{1.1}$$

where $w_k(x;\, \boldsymbol{\vartheta})$ is a given weighting function with parameter $\boldsymbol{\vartheta}$. Vardi (1982 and 1985), Gill et al. (1988) studied the estimation of the $\{F_k\}$ under this multi–sample biased sampling model with a given parameter in the weighting function, and Gilbert et al. (1999) and Gilbert (2000) studied the same estimation problem in a more general case where the parameter $\boldsymbol{\vartheta}$ in the weighting function is unknown. It is easily seen that the $\{F_k\}$ with weighting function $w_k(x;\, \boldsymbol{\beta}) = \exp\{\boldsymbol{\beta}_k^{\mathsf{T}}\boldsymbol{q}(x)\}$ satisfy the DRM assumption.

1.2.4 Other examples

Density ratio models are also used for life data modeling. For example (Marshall and Olkin, 2007), given a distribution function $F(x)$, the moment pa-

rameter family of distributions,

$$F_b(x) = \frac{1}{\mu_b} \int_0^x t^b dF(t),\ x \geq 0,$$

where $\mu_b = \int_0^\infty t^b dF(t) < \infty$, and the Laplace transform parameter family of distributions,

$$F_s(x) = \frac{1}{L(s)} \int_{-\infty}^x e^{-st} dF(t),$$

where $L(s) = \int_{-\infty}^\infty e^{-st} dF(t) < \infty$ is the Laplace transform of F, both satisfy the DRM assumption.

Moreover, specific forms of DRMs have been used by Anderson (1972) for logistic discrimination, Anderson (1979) for multivariate logistic compounds modeling, and Efron and Tibshirani (1996) for density estimation of exponential families.

1.3 Empirical likelihood inference under DRMs: historical and recent development

Although the use of the density ratio model can be traced back to Anderson (1972) for logistic discrimination, it is not until 1990's that the DRM started to gain popularity after Qin et al. had published a series of papers (Qin 1993, Qin and Zhang 1997, Qin 1998) on inference problems under two–sample DRMs using the *empirical likelihood* (EL). Indeed, on the one hand, the EL, given its nonparametric characteristic, is a natural inference framework for the semiparametric DRM; on the other, the theoretical foundation of EL established by Owen (2001) and Qin's (1994) extension of EL for estimating equations makes it a ready–to–use tool for inference under DRMs.

Qin (1998) formally introduced EL to inference problems under the DRM, and in a two–sample case, established the asymptotic normality of the max-

imum EL estimator of DRM parameters. After that, EL became a standard inference tool under the DRM, and numerous papers about various aspects of inference under the DRM using EL have been published. For estimation problems, Cheng and Chu (2004) and Fokianos (2004) studied density estimation under two–sample and multi–sample DRMs, respectively; Zhang (2000) and Chen and Liu (2013) studied quantile estimation under two–sample and multi–sample DRMs, respectively. For hypothesis testing problems, Qin and Zhang (1997) and Zhang (2002) studied goodness–of–fit tests for logistic models and generalized logit models based on case–control data using DRM formulation, respectively; Fokianos et al. (2001) proposed a simple Wald–type test for linear hypotheses about the parameters of multi–sample DRMs; Keziou and Leoni-Aubin (2008) studied the EL ratio test for testing the equality of two distributions that satisfy a two–sample DRM. The effect of misspecification of the basis function of the DRM was assessed by Fokianos and Kaimi (2006), and the basis function selection problem is studied by Fokianos (2007).

The DRM has been adopted for inference based on censored samples. Ren (2008) proposed a weighted EL approach for inference under a two–sample DRM based on randomly censored observations. Wang et al. (2011) studied EL inference under a two–sample DRM for randomly right–censored data. Shen et al. (2012) studied EL inference under the DRM with randomly censored biased–sampling data.

The DRM is also used in the context of finite mixture models for its flexibility and robustness against model misspecification. Zou et al. (2002) and Zou (2002) proposed a finite mixture model whose each mixing component is a further mixture of two distributions that satisfy a two–sample DRM, for the modeling of genetic loci influencing quantitative traits. They discovered that the DRM is not a regular model: when the slope DRM parameter $\boldsymbol{\beta} = \mathbf{0}$, the normalization constant $\boldsymbol{\alpha}$ must also be $\mathbf{0}$, which declares that when the true DRM parameter $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*) = \mathbf{0}$, the EL function is not well–defined in a

neighbourhood of the true parameter — the violation of an important regularity condition for likelihood type inference. Zou et al. hence proposed the so–called *partial empirical likelihood* (PEL) to get around the non–regularity of the DRM, and studied the asymptotic properties of the PEL. Zhang's (2006) score test for testing homogeneity of population distributions under the same mixture model is also based on the PEL. A main criticism of the PEL is that it does not use all the information contained in data, hence is less efficient than the full EL. For the same mixture model as used by Zou et al. (2002), Tan (2009) proposed to treat α_k as a function of $\boldsymbol{\beta}_k$ and baseline distribution F_0. The resulting EL function then does not contain the normalization constant $\boldsymbol{\alpha}$ and hence the non–regularity issue is avoided. However, that EL is a very complicated function of $\boldsymbol{\beta}$ and the computation of the maximum EL estimator is cumbersome. The work by Qin and Liang (2011) on hypothesis testing in a mixture case–control model is also along the lines of Tan (2009).

Luo and Tsai (2012) considered an extension of DRM where the slope parameter $\boldsymbol{\beta}$ is regressed on a set of covariates. On the other hand, the DRM is also used for generalizing regression models. For example, Rathouz and Gao (2009), Huang and Rathouz (2012) and Huang (2014) adopted the DRM for mean regression when data are generated from a generalized linear model.

There is a lot more work on the topic of DRM, including the applications to the estimation and comparison of the receiver operating characteristic (ROC) curve (Wan and Zhang 2008, Guan et al. 2012), to the inference about measurement of treatment effects (Fokianos and Troendle 2007, Jiang and Tu 2012), etc.

The above literature review is meant as a general overview of the work that has been done on EL inference under the DRM. Specific and more detailed reviews of literature that is closely related to the theory established in this thesis are given in each subsequent chapter.

1.4 Highlights of the contributions of this thesis

As noted in previous sections, this thesis develops a theory of EL inference for parameter estimation and hypotheses testing concerning a number of different population distributions under the DRM, with multiple complete or Type I censored random samples from each. In particular, it presents the following new results.

(i) A dual EL ratio test for a general composite hypothesis about the DRM parameter based on multiple samples is developed. It embraces testing for change in population distributions as a special case. The limiting distributions of the corresponding test statistic under the null model and also a local alternative model are derived. The null limiting distribution is useful for approximating the p–values of the proposed test; the limiting distribution under the local alternative model is useful for approximating the power of the proposed test, calculating the sample size required for achieving a given power, and comparing the local asymptotic powers of dual EL ratio tests formulated in different ways.

(ii) The effects of information pooling by the DRM on the estimation accuracy of the maximum EL estimator of the DRM parameter and on the local asymptotic power of the dual EL ratio test are assessed in theory. It is shown that when additional samples are incorporated by the DRM, the estimation accuracy of the maximum EL estimator of the DRM parameter is usually increased and the local asymptotic power of the dual EL ratio test is often improved, even if the underlying distributions of the additional samples are not related to the population distributions of direct interest.

(iii) A general EL inference framework under the DRM based on multiple Type I censored samples is established. This inference framework is

computationally efficient and equipped with rich asymptotic results. The theory of hypothesis testing for the DRM parameter under this framework is also developed. Moreover, this inference framework can potentially be used to extend any EL inference result that is available for complete samples under the DRM to the case of Type I censored samples.

In addition to the above contributions in theory, this thesis also presents an application of the proposed dual EL ratio test to assessing bending and tension strengths of lumber produced in year 2007, 2010 and 2011 with intuitively reasonable implications for the forest industry.

Moreover, the thesis introduces a user friendly R software package we developed, which is called "*drmdel*", for the dual EL inference under the DRM. This software package is fast because its core is written in C. It covers a broad range of methods including the ones developed in this thesis as well as those developed by Chen and Liu (2013) for quantile estimation and by Fokianos (2004) for density estimation.

1.5 Outline of the thesis

The rest of the thesis is organized as follows. Chapter 2 introduces the concept of dual empirical likelihood and presents some preliminary results that are essential for the development of the theories in the succeeding chapters. Chapter 3 proposes a dual EL ratio test for a general composite hypothesis about the DRM parameter, and presents the asymptotic properties of that test. Chapter 4 studies the effects of information pooling by DRM on the estimation accuracy of the maximum EL estimator and on the local asymptotic power of the dual EL ratio test. Chapter 5 establishes an EL inference framework under the DRM based on multiple Type I censored samples, and presents the theory of EL ratio test under that framework. Chapter 6 introduces our software package *drmdel* for dual EL inference under the DRM

and demonstrates its use. The last chapter summarizes the results of this thesis and discusses some future work.

Chapter 2

Fundamentals of Dual Empirical Likelihood Inference under the DRM

This chapter lays the foundation of the dual empirical likelihood (DEL), which is the framework this thesis adopts for inferences under the DRM. We first review some basics about the EL inference for a single sample and for multiple samples under the DRM, then define the DEL and summarize its important properties. All the theorems presented in this chapter are stemmed from the literature, although we present them under more general settings and conditions. We highlight a previously unnoticed result, Theorem 2.2, on the relationship between information matrix and asymptotic variance of the score function evaluated at the so–called true parameter value under the DRM, which is a key to proving some of our results in succeeding chapters. The proofs are given in the last section.

We use bold symbols for vectors, normal symbols for scalars or matrices, upper case letters for random variables and lower case letters for the corresponding realized values.

2.1 EL for a single sample

Let $x_1, x_2, \ldots, x_n$ be an independent sample of size n from a population with cumulative distribution function (CDF) $F(x)$. When F is a discrete distribution, we have $dF(x_i) = F(x_i) - F(x_i^-)$. The EL of the distribution

function F is defined as if F is discrete,

$$L_n(F) = \prod_{i=1}^{n} dF(x_i), \text{ subject to } \sum_{i=1}^{n} dF(x_i) = 1.$$

When there are no ties in the observations, $L_n(F)$ is maximized when $dF(x_i) = 1/n$, for $i = 1, 2, \ldots, n$. This maximum corresponds to the empirical distribution of $\{x_i\}_{i=1}^n$, $F_n = n^{-1}\sum_{i=1}^n \mathbb{1}(x_i \leq x)$, where $\mathbb{1}(\cdot)$ is the indicator function.

Many classical nonparametric inferences about F, e.g. quantile estimation, are based on this unconstrained maximum EL estimator of the distribution function, or equivalently the empirical distribution, F_n. A breakthrough comes from a result on the so–called *profile EL* of the population mean. The profile log EL of the population mean μ, $l_n(\mu)$, is defined as the supremum of $\log L_n(F)$ over the class of distribution functions with $\{x_i\}_{i=1}^n$ as support, i.e. the distribution functions of the form

$$F(x) = \sum_{i=1}^{n} p_i \mathbb{1}(x_i \leq x) \text{ with } \sum_{i=1}^{n} p_i = 1,$$

such that $\int x dF(x) = \mu$ for fixed μ. In other words,

$$l_n(\mu) = \sup\{\log L_n(F) : \; F(x) = \sum_{i=1}^{n} p_i \mathbb{1}(x_i \leq x), \; \sum_{i=1}^{n} p_i = 1, \; \sum_{i=1}^{n} x_i p_i = \mu\}.$$

The maximum of $l_n(\mu)$ as a function of μ is easily found to be obtained at $\mu = \bar{x} = n^{-1}\sum_{i=1}^n x_i$. Owen (1988) showed that the likelihood ratio statistic

$$2\big\{l_n(\bar{x}) - l_n(\mu^*)\big\}$$

where μ^* is the true population mean, converges in distribution to a chi–square random variable with one degree of freedom, under mild moment conditions. This elegant Theorem is the foundation of various EL based

inferences.

Another remarkable piece of work was done by Qin and Lawless (1994) which extends the empirical likelihood framework to incorporate a set of estimating functions. Suppose a parameter vector of interest, $\boldsymbol{\vartheta}$, can be defined as the solution to

$$\mathrm{E}\big\{\boldsymbol{g}(X;\, \boldsymbol{\vartheta})\big\} = \mathbf{0},$$

where $\boldsymbol{g}$ is a smooth function of $\boldsymbol{\vartheta}$. The profile log EL, $l_n(\boldsymbol{\vartheta})$, of $\boldsymbol{\vartheta}$ is defined to be the supremum of $\log L_n(F)$ over the class of distribution functions with $\{x_i\}_{i=1}^n$ as support, subject to $\int \boldsymbol{g}(x;\, \boldsymbol{\vartheta}) dF(x) = \mathbf{0}$ for fixed $\boldsymbol{\vartheta}$, i.e.

$$l_n(\boldsymbol{\vartheta}) = \sup\{\log L_n(F) : F(x) = \sum_{i=1}^n p_i \mathbb{1}(x_i \le x),\ \sum_{i=1}^n p_i = 1,$$
$$\sum_{i=1}^n p_i \boldsymbol{g}(x_i;\, \boldsymbol{\vartheta}) = \mathbf{0}\}.$$

This profile log EL $l_n(\boldsymbol{\vartheta})$ is again found to be useful for inference on $\boldsymbol{\vartheta}$. In particular, the maximum profile log EL estimator of $\boldsymbol{\vartheta}$ is asymptotically normal, and the EL ratio statistics still has a chi–square limiting distribution just as in parametric case.

2.2 EL for multiple samples under the DRM

Suppose we have $m+1$ independent random samples denoted as

$$\{x_{kj} : j = 1,\, 2, \ldots,\, n_k\}_{k=0}^m$$

with $n_k > 0$ being the size of the kth sample, which are collected from populations with distribution functions F_k, $k = 0,\, 1, \ldots,\, m$. Denote the total sample size as $n = \sum_k n_k$. Let $dF_k(x) = F_k(x) - F_k(x^-)$, and put $p_{kj} = dF_0(x_{kj})$.

Suppose the $\{F_k\}$ satisfy the DRM assumption postulated in Section 1.2:

$$dF_k(x) = \exp\left\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\right\} dF_0(x), \text{ for } k = 1, 2, \ldots, m, \tag{2.1}$$

where the basis function $\boldsymbol{q}(x)$ is a prespecified d–dimensional function, and $\boldsymbol{\theta}_k^\intercal = (\alpha_k, \boldsymbol{\beta}_k^\intercal)$ are model parameters. We denote $\boldsymbol{\theta}_0 = \mathbf{0}$ for ease of exposition. This assumption implies that the $\{F_k\}$ satisfy

$$\int dF_k(x) = \int \exp\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\} dF_0(x) = 1. \tag{2.2}$$

Under the DRM assumption, the EL of the $\{F_k\}$ is given by

$$\begin{aligned} L_n(F_0, F_1, \ldots, F_m) &= \prod_{k,j} dF_k(x_{kj}) = \prod_{k,j} \exp\left\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x_{kj})\right\} dF_0(x_{kj}) \\ &= \left\{\prod_{k,j} p_{kj}\right\} \cdot \exp\left\{\sum_{k,j} \left(\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x_{kj})\right)\right\}, \end{aligned} \tag{2.3}$$

where the sum and product are over all possible (k, j) combinations. Let $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)^\intercal$, $\boldsymbol{\beta}^\intercal = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_m^\intercal)$, and $\boldsymbol{\theta}^\intercal = (\boldsymbol{\alpha}^\intercal, \boldsymbol{\beta}^\intercal)$. We may also write the EL as $L_n(\boldsymbol{\theta}, F_0)$.

The maximum EL estimator (MELE) of $\boldsymbol{\theta}$ and F_0 is the maximum point of $L_n(\boldsymbol{\theta}, F_0)$ over the space of $\boldsymbol{\theta}$ and F_0 such that (2.2) is satisfied. As in the case of a single sample, for both theoretical discussion and numerical computation, the maximization is carried out in two steps. First, we define the profile log EL:

$$\tilde{l}_n(\boldsymbol{\theta}) = \sup_{F_0} \left\{ \log L_n(\boldsymbol{\theta}, F_0) : \sum_{k,j} \exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\} p_{kj} = 1, \ r = 0, \ldots, m. \right\}$$

where the supremum is taken over the space of F_0 with fixed $\boldsymbol{\theta}$. This supremum can be obtained by the method of Lagrange multipliers. For a fixed $\boldsymbol{\theta}$,

define the Lagrange function:

$$\Phi(\{p_{kj}\},\, \{\lambda_r\}_{r=0}^m) = \log L_n(\boldsymbol{\theta},\, F_0) + n\sum_{r=0}^m \lambda_r\big\{1 - \sum_{k,\,j} p_{kj}\exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\}\big\}.$$

The point $\{p_{kj}\}$ at which $L_n(\boldsymbol{\theta},\, F_0)$ is maximized must be on a stationary point of $\Phi(\{p_{kj}\},\, \{\lambda_r\})$ satisfying

$$\partial\Phi(\{p_{kj}\},\, \{\lambda_r\})/\partial p_{kj} = 0, \tag{2.4}$$

$$\partial\Phi(\{p_{kj}\},\, \{\lambda_r\})/\partial \lambda_r = 0. \tag{2.5}$$

Note that, at this stationary point,

$$\begin{aligned}
0 &= \sum_{k,\,j} p_{kj}\big\{\partial\Phi(\{p_{kj}\},\, \{\lambda_r\})/\partial p_{kj}\big\} \\
&= \sum_{k,\,j} p_{kj}\big\{1/p_{kj} - n\sum_{r=0}^m \lambda_r \exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\}\big\} \\
&= n - n\sum_{r=0}^m \lambda_r\big\{\sum_{k,\,j}\exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\}p_{kj}\big\} \\
&= n - n\sum_{r=0}^m \lambda_r,
\end{aligned}$$

where the last equality is obtained by the constraint $\sum_{k,\,j}\exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\}p_{kj} = 1$ for all $r = 0,\, 1,\, \ldots,\, m$. Solving equations (2.4) and using the relationship $0 = n - n\sum_{r=0}^m \lambda_r$, we find that the supremum of $L_n(\boldsymbol{\theta},\, F_0)$ is attained when $\lambda_0 = 1 - \sum_{r=1}^m \lambda_r$ and

$$p_{kj} = n^{-1}\Big\{1 + \sum_{r=1}^m \lambda_r\big[\exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\} - 1\big]\Big\}^{-1}, \tag{2.6}$$

where the Lagrange multipliers $\{\lambda_r\}_{r=1}^m$ solve, for $t = 0, 1, \ldots, m$,

$$\sum_{k,j} \exp\{\alpha_t + \boldsymbol{\beta}_t^\intercal \boldsymbol{q}(x_{kj})\} p_{kj} = 1. \tag{2.7}$$

The profile log EL can hence be written as

$$\begin{aligned}\tilde{l}_n(\boldsymbol{\theta}) = &-\sum_{k,j} \log\Big\{1 + \sum_{r=1}^m \lambda_r \big[\exp\{\alpha_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(x_{kj})\} - 1\big]\Big\} \\ &+ \sum_{k,j} \{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x_{kj})\}. \end{aligned} \tag{2.8}$$

The MELE $\hat{\boldsymbol{\theta}}$ is then the point at which $\tilde{l}_n(\boldsymbol{\theta})$ is maximized. Given $\hat{\boldsymbol{\theta}}$, we solve for the Lagrange multipliers $\hat{\lambda}_r$ through (2.7). The MELE must satisfy $\partial l_n(\hat{\boldsymbol{\theta}})/\partial \alpha_k = 0$ for $k = 1, \ldots, m$. We see that

$$\begin{aligned}\frac{\partial l_n(\hat{\boldsymbol{\theta}})}{\partial \alpha_k} &= n_k - \sum_{t,j} \frac{\hat{\lambda}_k \exp\{\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^\intercal \boldsymbol{q}(x_{tj})\}}{1 + \sum_{r=1}^m \hat{\lambda}_r \big[\exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^\intercal \boldsymbol{q}(x_{tj})\} - 1\big]} \\ &= n_k - \hat{\lambda}_k n \sum_{t,j} \exp\{\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^\intercal \boldsymbol{q}(x_{tj})\} \hat{p}_{kj} \\ &= n_k - \hat{\lambda}_k n, \end{aligned}$$

where the second equality is by (2.6) and the last equality is by (2.7). Therefore $\partial l_n(\hat{\boldsymbol{\theta}})/\partial \alpha_k = n_k - \hat{\lambda}_k n = 0$, and so, when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we have

$$\hat{\lambda}_k = n_k/n.$$

Subsequently, we obtain $\hat{p}_{kj}$ by plugging $\hat{\boldsymbol{\theta}}$ and $\hat{\lambda}_k$ into (2.6):

$$\hat{p}_{kj} = n^{-1}\Big\{1 + \sum_{r=1}^m \hat{\lambda}_r \big[\exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^\intercal \boldsymbol{q}(x_{kj})\} - 1\big]\Big\}^{-1}. \tag{2.9}$$

Finally, the MELE of F_k, $k = 0, 1, \ldots, m$ is given by

$$\hat{F}_k(x) = n^{-1} \sum_{r,j} \exp\{\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^{\intercal} \boldsymbol{q}(x_{rj})\} \hat{p}_{rj} \mathbb{1}(x_{rj} \leq x). \tag{2.10}$$

2.3 Non–regularity of the DRM and dual empirical likelihood

In applications such as that described in the Introduction to the forestry products industry, giving a point estimation is a minor part of the data analysis. Assessing the uncertainty in the point estimator and testing hypotheses would be judged of greater practical importance. Asymptotic properties of the point estimator and the likelihood function enable more such in–depth data analyses. However, classical asymptotic theories usually rely on differential properties of the likelihood function in the neighbourhood of the true parameter value. Consequently these results are applicable only if this neighbourhood lies in the parameter space.

According to (2.2), we have

$$\alpha_k = -\log \int \exp\{\boldsymbol{\beta}_k^{\intercal} \boldsymbol{q}(x)\} dF_0(x).$$

Thus, $\alpha_k = 0$ whenever $\boldsymbol{\beta}_k = \mathbf{0}$. When the true value $\boldsymbol{\theta}_1 = \mathbf{0}$, its neighborhood will not be contained in the parameter space. In statistical terminology, DRM is not regular at this $\boldsymbol{\theta}$, as noticed by Zou et al. (2002). Clearly, the regularity is also violated when $\boldsymbol{\beta}_k = \boldsymbol{\beta}_j$, $k \neq j$, which implies $\alpha_k = \alpha_j$. In our targeted applications, $\boldsymbol{\theta}_k$ would be the parameter of the lumber population, F_k, at year k and we are particularly concerned about the stability of lumber quality. Note that $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$ would signify the stability of the wood quality over these two years, because under the DRM, $F_k = F_0$ is equivalent to $\boldsymbol{\theta}_k = \mathbf{0}$ and $F_k = F_j$, $k \neq j$, is equivalent to $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$. In a statistical context, we test this stability by detecting the differences among lumber

distributions:

$$H_0: \ F_k = F_j \text{ for all } k, j \in \{0, \ldots, m\} \ \text{ against } \ H_1: \ F_k \neq F_j \text{ for some k, j,}$$

or equivalently,

$$H_0: \ \boldsymbol{\theta}_k = \boldsymbol{\theta}_j \text{ for all } k, j \in \{0, \ldots, m\} \ \text{ against } \ H_1: \ \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_j \text{ for some k, j.}$$

Non–regularity denies a simplistic application of the straightforward EL ratio test to this important hypothesis. This creates a need for other effective inferential methods.

To enable likelihood type inference in the presence of non–regularity of the DRM, Keziou and Leoni-Aubin (2008) proposed to use a "dual" form of the EL in the case of two samples. We extend their notion to the case of multiple samples and refer to it as the *dual empirical likelihood* function. Recall that when the profile log EL (2.8) is maximized, i.e. when $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we have $\hat{\lambda}_r = n_r/n$. We define DEL by replacing the $\{\lambda_r\}$ with $\{\hat{\lambda}_r\}$ in (2.8),

$$l_n(\boldsymbol{\theta}) = -\sum_{k,j} \log\Big\{\sum_{r=0}^{m} \hat{\lambda}_r \exp\big\{\alpha_r + \boldsymbol{\beta}_r^{\intercal}\boldsymbol{q}(x_{kj})\big\}\Big\} + \sum_{k,j}\big\{\alpha_k + \boldsymbol{\beta}_k^{\intercal}\boldsymbol{q}(x_{kj})\big\}. \tag{2.11}$$

Clearly, the MELE is also the point at which the DEL is maximized,

$$\hat{\boldsymbol{\theta}} = \operatorname*{argmax}_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta}),$$

and the profile log EL and the DEL have the same maximal values, $l_n(\hat{\boldsymbol{\theta}}) = \tilde{l}_n(\hat{\boldsymbol{\theta}})$.

The DEL has the following appealing features: (i) it is well–defined for all values of $\boldsymbol{\theta}$ in the corresponding Euclidean space, thus can be "safely" used for likelihood type inference even when $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$ for some $k \neq j$, k, $j = 0, \ldots, m$; (ii) it has a much simpler analytical form than the profile log EL since the

$\{\lambda_k\}$ in the profile log EL are now replaced by the data value independent $\{\hat{\lambda}_k\}$; and (iii) as we will prove in the next section, the DEL is a smooth concave function of the DRM parameter $\boldsymbol{\theta}$ (while the log–profile empirical likelihood is not), which leads to nice theoretical properties of the DEL and makes numerical computation of the MELE a pleasant task. The inference methods developed in this thesis are based on the DEL because of these attractive characteristics.

2.4 Properties of the DEL

This section presents some properties of the DEL that are useful for the development of the theory in the sequel. Most of these properties have been given in literature under a two–sample DRM or/and under slightly different conditions. We highlight an unnoticed result, the relationship between the information matrix and the asymptotic variance of the score function, which is a key to our study of the asymptotic properties of the DEL ratio statistic in subsequent chapters. The well–known asymptotic normality of the MELE is also included for the self–containedness of the thesis. The proofs of the results are given in the last section of this chapter.

For a matrix A, we will use $A > 0$ to denote that A is positive definite, and $A \geq 0$ to denote that A is positive semidefinite. For a differentiable function $g(x)$ and a particular value x_0, we use $\partial g(x_0)/\partial x$ to denote $(\partial g(x)/\partial x)|_{x=x_0}$.

Theorem 2.1 (Properties of the information matrix)**.** *Suppose we have $m+1$ random samples from populations with distributions of the DRM form given in (2.1) and a true parameter value $\boldsymbol{\theta}^*$ such that*

$$\int \exp\{\boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\} dF_0(x) < \infty$$

for $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}^*$, $\int \boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x) dF_0(x) > 0$ with $\boldsymbol{Q}^\intercal(x) = (1, \boldsymbol{q}^\intercal(x))$, and $\hat{\lambda}_k = n_k/n = \rho_k + o(1)$ for some constant $\rho_k \in (0, 1)$.

The empirical information matrix $U_n = -n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal$ *converges almost surely to a positive definite matrix* $U = \lim_{n\to\infty} U_n$.

Remark 2.1. The condition that $\int \boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x)dF_0(x) > 0$ is equivalent to say that $\boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x) > 0$ on a set (of x values) of positive probability with respect to F_0. It is a model identifiability condition. If any two components of the extended basis function $\boldsymbol{Q}(x)$ are linear dependent with probability one, say one component of $\boldsymbol{q}(x)$ is a constant, or x^2 and x^2+2 are two components of $\boldsymbol{q}(x)$, then the DRM is clearly not identifiable. This assumption ensures that we do not use an non–identifiable model.

The limiting matrix U may be regarded as an *information matrix*. We partition the entries of U in agreement with $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ and represent them as $U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}$, $U_{\boldsymbol{\alpha}\boldsymbol{\beta}}$, $U_{\boldsymbol{\beta}\boldsymbol{\alpha}}$ and $U_{\boldsymbol{\beta}\boldsymbol{\beta}}$. Let $\varphi_k(\boldsymbol{\theta},\, x) = \exp\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\}$, $k = 0, \ldots, m$, and

$$\begin{aligned} \boldsymbol{h}(\boldsymbol{\theta},\, x) &= (\rho_1\varphi_1(\boldsymbol{\theta},\, x),\, \ldots,\, \rho_m\varphi_m(\boldsymbol{\theta},\, x))^\intercal, \\ s(\boldsymbol{\theta},\, x) &= \rho_0 + \sum_{k=1}^{m} \rho_k\varphi_k(\boldsymbol{\theta},\, x), \\ H(\boldsymbol{\theta},\, x) &= \mathrm{diag}\{\boldsymbol{h}(\boldsymbol{\theta},\, x)\} - \boldsymbol{h}(\boldsymbol{\theta},\, x)\boldsymbol{h}^\intercal(\boldsymbol{\theta},\, x)/s(\boldsymbol{\theta},\, x). \end{aligned} \tag{2.12}$$

Let $\mathrm{E}_k(\cdot)$, $k = 0, 1, \ldots,\, m$, be the expectation operator with respect to F_k, i.e. $\mathrm{E}_k\{g(x)\} = \int g(x)dF_k(x)$ for a measurable function $g(x)$. Then, the block-wise algebraic expressions of the information matrix U in terms of $H(\boldsymbol{\theta}^*,\, x)$ and $\boldsymbol{q}(x)$ can be written as

$$\begin{aligned} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} &= -\lim_{n\to\infty} n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\intercal = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x)\big\}, \\ U_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\lim_{n\to\infty} n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal = \mathrm{E}_0\Big\{H(\boldsymbol{\theta}^*,\, x) \otimes \big(\boldsymbol{q}(x)\boldsymbol{q}^\intercal(x)\big)\Big\}, \\ U_{\boldsymbol{\alpha}\boldsymbol{\beta}} &= -\lim_{n\to\infty} n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^\intercal = U_{\boldsymbol{\beta}\boldsymbol{\alpha}}^\intercal = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x) \otimes \boldsymbol{q}^\intercal(x)\big\}, \end{aligned} \tag{2.13}$$

where $\otimes$ is the Kronecker product operator.

Put

$$\boldsymbol{v} = n^{-1/2}\{\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\}.$$

Let $\mathrm{E}(\cdot)$ be the usual expectation operator. Let

$$T = \rho_0{}^{-1}\mathbf{1}_m\mathbf{1}_m^{\intercal} + \mathrm{diag}\{\rho_1^{-1},\, \rho_2^{-1},\, \ldots,\, \rho_m^{-1}\} \text{ and } W = \begin{pmatrix} T & 0_{m\times md} \\ 0_{md\times m} & 0_{md\times md} \end{pmatrix},$$

where $\mathbf{1}_k$, in general, is a vector of 1s with length k.

Theorem 2.2 (Asymptotic properties of the score function)**.** *Under the conditions of Theorem 2.1, we have* $\mathrm{E}\boldsymbol{v} = \mathbf{0}$ *and* $\boldsymbol{v}$ *is asymptotically multivariate normal with mean* $\mathbf{0}$ *and covariance matrix*

$$V = U - UWU. \tag{2.14}$$

Remark 2.2. In the parametric likelihood setting, the information matrix equals the asymptotic variance of the $1/\sqrt{n}$–scaled score function evaluated at the true parameter. In the DEL framework under the DRM, the relationship between the information matrix and the asymptotic variance of the score function evaluated at the true value, as shown in (2.14), is different. This is a previously unnoticed result although it has been implicitly used by Chen and Liu (2013) and Zhang (2002). Due to the complicated algebraic expressions of U and V, this relationship is not obvious. However, once observed, as we will see later, we can intentionally "forget" the algebraic expression of U given by (2.13), and use this relationship solely for deriving the asymptotic distribution of the DEL ratio statistic.

The following two lemmas are useful for establishing the consistency of the MELE $\hat{\boldsymbol{\theta}}$.

Lemma 2.3. *The DEL function* $l_n(\boldsymbol{\theta})$ *defined in* (2.11) *is a concave function. Moreover, when* $\sum_{k,j}\boldsymbol{Q}(x_{kj})\boldsymbol{Q}^{\intercal}(x_{kj}) > 0$, $l_n(\boldsymbol{\theta})$ *is strictly concave.*

Remark 2.3. The condition that $\sum_{k,j} \boldsymbol{Q}(x_{kj})\boldsymbol{Q}^{\intercal}(x_{kj}) > 0$ is a sample version of the condition that $\int \boldsymbol{Q}(x)\boldsymbol{Q}^{\intercal}(x)dF_0(x) > 0$ in Theorem 2.1. This condition is usually fulfilled if number of distinct data points is larger than $d+1$ and the components of $\boldsymbol{Q}(x)$ are not functionally linearly dependent.

The concavity of the DEL in the case of two samples was pointed out by Keziou and Leoni-Aubin (2008). Since the DEL is concave, whenever it has a maximum, the maximum is a global one, and if the DEL is strictly concave, this maximum is unique. The next lemma states that when the total sample size n goes to infinity, the DEL $l_n(\boldsymbol{\theta})$ has a maximum with probability tending one, and this maximum is in a $n^{-1/3}$ neighbourhood of the true parameter $\boldsymbol{\theta}^*$. Let $\|\cdot\|$ denote the Euclidean norm of a vector.

Lemma 2.4. *Adopt the conditions postulated in Theorem 2.1. As $n \to \infty$, with probability tending one, the DEL $l_n(\boldsymbol{\theta})$ attains its maximum at some point in the interior of the closed ball, $B_{\boldsymbol{\theta}^*} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq n^{-1/3}\}$, which is centered on the true parameter value $\boldsymbol{\theta}^*$.*

A two–sample version of Lemma 2.4 was given by Keziou and Leoni-Aubin (2008).

Remark 2.4. Lemma 2.3 and 2.4 together confirm that when sample size is large, the MELE is well–defined and easy to compute: it is a global maximal point of a concave function whose maximum exists with probability tending one. Furthermore, they dictate that the MELE $\hat{\boldsymbol{\theta}}$ is $\sqrt[3]{n}$–consistent: by concavity, all the maximal points of the DEL must be interior points of the closed ball $B_{\boldsymbol{\theta}^*}$; hence the MELE $\hat{\boldsymbol{\theta}}$, as a maximal point of the DEL, must also be an interior point of $B_{\boldsymbol{\theta}^*}$.

With Theorem 2.1, 2.2, Lemma 2.3 and 2.4, the asymptotic normality of the MELE $\hat{\boldsymbol{\theta}}$ is an easy consequence.

Theorem 2.5 (Asymptotic normality of the MELE)**.** *Under the conditions of Theorem 2.1, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ has an asymptotic multivariate normal distribution*

with mean $\mathbf{0}$ *and covariance matrix* $U^{-1} - W$*, where* W *is given in Theorem 2.2.*

The asymptotic normality of the $\hat{\boldsymbol{\theta}}$ was also established by Chen and Liu (2013) and by Zhang (2002) under slightly different conditions. Theorem 2.5 reveals that the MELE is root–n consistent, an important fact that we will use in the subsequent chapters.

2.5 Proofs

This section gives proofs for the theorems and lemmas presented in the last section. We first introduce more notations applicable to $k = 0, \ldots, m$. Recall that $\varphi_k(\boldsymbol{\theta}, x) = \exp\{\alpha_k + \boldsymbol{\beta}_k \boldsymbol{q}(x)\}$. We write

$$\mathcal{L}_{n,k}(\boldsymbol{\theta}, x) = -\log\Big\{\sum_{r=0}^{m} \hat{\lambda}_r \varphi_r(\boldsymbol{\theta}, x)\Big\} + \big\{\alpha_k + \boldsymbol{\beta}_k^{\intercal} \boldsymbol{q}(x)\big\}$$

with $\hat{\lambda}_r = n_r/n$ being the sample proportion. Hence, the DEL $l_n(\boldsymbol{\theta}) = \sum_{k,j} \mathcal{L}_{n,k}(\boldsymbol{\theta}, x_{kj})$ where the summation is over all possible (k, j). Let $\mathcal{L}_k(\boldsymbol{\theta}, x)$ be the "population" version of $\mathcal{L}_{n,k}(\boldsymbol{\theta}, x)$ by replacing $\hat{\lambda}_r$ with its limit ρ_r in the above definition. Let $\boldsymbol{e}_k$ be a vector of length m with the kth entry being 1 and the others being 0s, and let $\delta_{ij} = 1$ when $i = j$, and 0 otherwise. Recall the definitions (2.12) of $\boldsymbol{h}(\boldsymbol{\theta}, x)$, $s(\boldsymbol{\theta}, x)$ and $H(\boldsymbol{\theta}, x)$. The first order derivatives of $\mathcal{L}_k(\boldsymbol{\theta}, x)$ can be written as

$$\begin{aligned} \partial\mathcal{L}_k(\boldsymbol{\theta}, x)/\partial\boldsymbol{\alpha} &= (1 - \delta_{k0})\boldsymbol{e}_k - \boldsymbol{h}(\boldsymbol{\theta}, x)/s(\boldsymbol{\theta}, x), \\ \partial\mathcal{L}_k(\boldsymbol{\theta}, x)/\partial\boldsymbol{\beta} &= \{\partial\mathcal{L}_k(\boldsymbol{\theta}, x)/\partial\boldsymbol{\alpha}\} \otimes \boldsymbol{q}(x). \end{aligned} \tag{2.15}$$

Similarly, we have

$$\begin{aligned}
&\partial^2\mathcal{L}_k(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\intercal = -H(\boldsymbol{\theta},\, x)/s(\boldsymbol{\theta},\, x),\\
&\partial^2\mathcal{L}_k(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal = -\big\{H(\boldsymbol{\theta},\, x)/s(\boldsymbol{\theta},\, x)\big\}\otimes\big\{\boldsymbol{q}(x)\boldsymbol{q}^\intercal(x)\big\},\\
&\partial^2\mathcal{L}_k(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^\intercal = -\big\{H(\boldsymbol{\theta},\, x)/s(\boldsymbol{\theta},\, x)\big\}\otimes\boldsymbol{q}^\intercal(x).
\end{aligned} \tag{2.16}$$

The algebraic expressions of the derivatives of $\mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)$ are similar to those of $\mathcal{L}_k(\boldsymbol{\theta},\, x)$, with ρ_r replaced by the sample proportion $\hat{\lambda}_r$. Note that all entries of $\boldsymbol{h}(\boldsymbol{\theta},\, x)$ are non–negative, and $s(\boldsymbol{\theta},\, x)$ exceeds the sum of all entries of $\boldsymbol{h}(\boldsymbol{\theta},\, x)$. Thus, $\|\boldsymbol{h}(\boldsymbol{\theta},\, x)/s(\boldsymbol{\theta},\, x)\| \leq 1$, and the absolute value of each entry of $H(\boldsymbol{\theta},\, x)/s(\boldsymbol{\theta},\, x)$ is bounded by 1. By examining the algebraic expressions closely, this result implies

$$\begin{aligned}
&\big|\partial^2\mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\theta_i\partial\theta_j\big| \leq 1 + \boldsymbol{q}^\intercal(x)\boldsymbol{q}(x),\\
&\big|\partial^3\mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\theta_i\partial\theta_j\partial\theta_k\big| \leq \{1 + \boldsymbol{q}^\intercal(x)\boldsymbol{q}(x)\}^{3/2},
\end{aligned} \tag{2.17}$$

where θ_i in general denotes the ith entry of $\boldsymbol{\theta}$.

We also observed the following important relationships between the first and second order derivatives of $\mathcal{L}_k(\boldsymbol{\theta},\, x)$:

$$\mathrm{E}_0\left\{\frac{\partial\mathcal{L}_0(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\alpha}}\right\} = -\rho_0^{-1}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\mathbf{1}_m, \quad \mathrm{E}_0\left\{\frac{\partial\mathcal{L}_0(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\beta}}\right\} = -\rho_0^{-1}U_{\boldsymbol{\beta}\boldsymbol{\alpha}}\mathbf{1}_m, \tag{2.18}$$

and, for $k = 1,\, 2,\, \ldots,\, m$,

$$\begin{aligned}
&\mathrm{E}_k\left\{\frac{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\alpha}}\right\} = \rho_k^{-1}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\boldsymbol{e}_k, \quad E_k\left\{\frac{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\alpha}}\boldsymbol{q}^\intercal(x)\right\} = \rho_k^{-1}U_{\boldsymbol{\alpha}\boldsymbol{\beta}}(\boldsymbol{e}_k\otimes I_d),\\
&\mathrm{E}_k\left\{\frac{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\beta}}\right\} = \rho_k^{-1}U_{\boldsymbol{\beta}\boldsymbol{\alpha}}\boldsymbol{e}_k, \quad E_k\left\{\frac{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\beta}}\boldsymbol{q}^\intercal(x)\right\} = \rho_k^{-1}U_{\boldsymbol{\beta}\boldsymbol{\beta}}(\boldsymbol{e}_k\otimes I_d).
\end{aligned} \tag{2.19}$$

As a reminder, $\mathrm{E}_k(\cdot)$, $k = 0, 1, \ldots, m$, is the expectation operator with re-

spect to F_k.

The assumption that $\int \exp\{\boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\} dF_0(x) < \infty$ for $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}^*$ implies that the moment generating function of $\boldsymbol{q}(x)$ with respect to each F_k, exists in a neighbourhood of $\mathbf{0}$. Hence, all finite order moments of $\boldsymbol{q}(x)$ with respect to each F_k are finite. This fact and inequalities (2.17) reveal that the second and third order derivatives of $l_n(\boldsymbol{\theta})$ are bounded by an integrable function.

With the above preparation, we are ready to prove the theorems given in the chapter.

2.5.1 Theorem 2.1: Properties of the information matrix

We now show that the empirical information matrix $U_n = -n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal$ converges almost surely to a positive definite information matrix, and give its algebraic expression.

Recalling that $l_n(\boldsymbol{\theta}) = \sum_{k,\,j} \mathcal{L}_{n,k}(\boldsymbol{\theta}; x_{kj})$ and $\hat{\lambda} = n_k/n$, we have

$$U_n = -\frac{1}{n}\frac{\partial^2 l_n(\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} = -\sum_{k=0}^{m} \hat{\lambda}_k \left\{ \frac{1}{n_k} \sum_{j=1}^{n_k} \frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta}^*,\, x_{kj})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} \right\}.$$

Each term in the curly brackets is the average of the sum of independent and identically distributed (iid) random variables. And by bound (2.17) and the fact that $\boldsymbol{q}(x)$ has finite second moments, these random variables have finite covariance matrices. Hence, by the strong law of large numbers (Chow and Teicher, 1997, 5.4, Theorem 1), each term in the curly brackets has an almost sure limit. Along with $\lim_{n\to\infty} n_k/n = \rho_k$, we have that $\{U_n\}$ has an almost sure limit

$$U = \lim_{n\to\infty} U_n = -\sum_{k=0}^{m} \rho_k \mathrm{E}_k\{\partial^2 \mathcal{L}_k(\boldsymbol{\theta}^*,\, x)\ \partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal\}.$$

By expressions (2.16) of $\partial^2\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal$, we easily get the blockwise algebraic expressions of U as given in (2.13):

$$
\begin{aligned}
U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} &= -\lim_{n\to\infty} n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\intercal = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x)\big\},\\
U_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\lim_{n\to\infty} n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal = \mathrm{E}_0\Big\{H(\boldsymbol{\theta}^*,\, x)\otimes\big(\boldsymbol{q}(x)\boldsymbol{q}^\intercal(x)\big)\Big\},\\
U_{\boldsymbol{\alpha}\boldsymbol{\beta}} &= -\lim_{n\to\infty} n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\beta}^\intercal = U_{\boldsymbol{\beta}\boldsymbol{\alpha}}^\intercal = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x)\otimes\boldsymbol{q}^\intercal(x)\big\}.
\end{aligned}
$$

We now show that for any given $\boldsymbol{\theta}^*$, $U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x)\big\}$ is positive definite, which is implied if for any given value of $\boldsymbol{\theta}$ and x, $H(\boldsymbol{\theta},\, x)$ is positive definite. Let $\boldsymbol{a}$ be a nonzero vector of length m and a_i be its ith component. Recalling the definition (2.12) of $H(\boldsymbol{\theta},\, x)$, we have

$$
\begin{aligned}
\boldsymbol{a}^\intercal H(\boldsymbol{\theta},\, x)\boldsymbol{a} = s^{-1}(\boldsymbol{\theta},\, x)\Bigg\{&\sum_{i=1}^m a_i^2\rho_i\varphi_i(\boldsymbol{\theta},\, x)\left(s(\boldsymbol{\theta},\, x)-\rho_i\varphi_i(\boldsymbol{\theta},\, x)\right)\\
&-2\sum_{1\le i<j}^m a_i a_j\rho_i\varphi_i(\boldsymbol{\theta},\, x)\rho_j\varphi_j(\boldsymbol{\theta},\, x)\Bigg\}.
\end{aligned}
$$

Note that since $s(\boldsymbol{\theta},\, x)-\rho_i\varphi_i(\boldsymbol{\theta},\, x) = \rho_0+\sum_{j\neq i}^m\rho_j\varphi_j(\boldsymbol{\theta},\, x)$, the above equality can be further written as

$$
\begin{aligned}
&\boldsymbol{a}^\intercal H(\boldsymbol{\theta},\, x)\boldsymbol{a}\\
=&s^{-1}(\boldsymbol{\theta},\, x)\Bigg\{\sum_{i=1}^m a_i^2\rho_0\rho_i\varphi_i(\boldsymbol{\theta},\, x)+\sum_{i\neq j}^m a_i^2\rho_i\varphi_i(\boldsymbol{\theta},\, x)\rho_j\varphi_j(\boldsymbol{\theta},\, x)\\
&\qquad -2\sum_{1\le i<j}^m a_i a_j\rho_i\varphi_i(\boldsymbol{\theta},\, x)\rho_j\varphi_j(\boldsymbol{\theta},\, x)\Bigg\}\\
=&s^{-1}(\boldsymbol{\theta},\, x)\Bigg\{\sum_{i=1}^m a_i^2\rho_0\rho_i\varphi_i(\boldsymbol{\theta},\, x)+\sum_{1\le i<j}^m (a_i-a_j)^2\rho_i\varphi_i(\boldsymbol{\theta},\, x)\rho_j\varphi_j(\boldsymbol{\theta},\, x)\Bigg\}
\end{aligned}
$$

Since $s(\boldsymbol{\theta},\, x)$ is positive, the first term in the curly brackets on right hand side (RHS) of the above equality is positive and the second term is nonnegative,

we have $\boldsymbol{a}^\intercal H(\boldsymbol{\theta},\, x)\boldsymbol{a} > 0$ and so $H(\boldsymbol{\theta},\, x) > 0$ for any value of $\boldsymbol{\theta}$ and x. Therefore $U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x)\big\} > 0$.

Finally we show that $U > 0$. Recall that $\boldsymbol{Q}(x) = (1, \boldsymbol{q}(x)^\intercal)^\intercal$. By expressions (2.16), we see that U can be obtained from the matrix

$$\mathrm{E}_0\big\{H(\boldsymbol{\theta}^*,\, x) \otimes \big\{\boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x)\big\}\big\} \tag{2.20}$$

by simply permuting rows and columns respectively. Since $H(\boldsymbol{\theta}^*,\, x) > 0$ for any value of x and $\boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x) > 0$ on a set of positive probability with respect to F_0, by a property of Kronecker product, the matrix (2.20) is positive definite, and so is U. This completes the proof.

2.5.2 Theorem 2.2: Asymptotic properties of the score function

We now show the asymptotic normality of $\boldsymbol{v} = n^{-1/2}\{\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\}$, the scaled score function evaluated at the true parameter value $\boldsymbol{\theta}^*$. Recall that

$$\mathcal{L}_{n,k}(\boldsymbol{\theta},\, x) = -\log\Big\{\sum_{r=0}^{m} \hat{\lambda}_r \varphi_r(\boldsymbol{\theta},\, x)\Big\} + \big\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\big\}$$

and $l_n(\boldsymbol{\theta}) = \sum_{k,\,j} \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x_{kj})$. We have

$$\boldsymbol{v} = n^{-1/2}\{\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\} = n^{-1/2}\sum_{k,\,j}\{\partial\mathcal{L}_{n,k}(\boldsymbol{\theta},\, x_{kj})/\partial\boldsymbol{\theta}\}.$$

We first show that $\mathrm{E}\boldsymbol{v} = 0$. Denote $\boldsymbol{\mu}_{n,k} = \mathrm{E}_k\{\partial\mathcal{L}_{n,k}(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\theta}\}$. Partition $\boldsymbol{v}$ to subvectors $\boldsymbol{v}_{\boldsymbol{\alpha}}$ and $\boldsymbol{v}_{\boldsymbol{\beta}}$ in agreement with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. Let $\hat{\boldsymbol{\lambda}} = (\hat{\lambda}_1,\, \ldots,\, \hat{\lambda}_m)^\intercal$. Let $\boldsymbol{h}_n(\boldsymbol{\theta},\, x)$, $s_n(\boldsymbol{\theta},\, x)$ and $H_n(\boldsymbol{\theta},\, x)$ be the sample versions of $\boldsymbol{h}(\boldsymbol{\theta},\, x)$, $s(\boldsymbol{\theta},\, x)$ and $H(\boldsymbol{\theta},\, x)$ defined in (2.12) with ρ_k replaced by $\hat{\lambda}_k$. By expression (2.15) and noticing that $\mathrm{E}_k\{g(x)\} = \mathrm{E}_0\{g(x)\varphi_k(\boldsymbol{\theta}^*, x)\}$

for a measurable function $g(x)$, we have

$$\begin{aligned}
\mathrm{E}\boldsymbol{v}_{\boldsymbol{\alpha}} &= n^{1/2}\sum_{k=0}^{m}\hat{\lambda}_k \mathrm{E}_k\{\partial\mathcal{L}_{n,k}(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha}\} \\
&= n^{1/2}\Big\{\hat{\boldsymbol{\lambda}} - \mathrm{E}_0\big\{\boldsymbol{h}_n(\boldsymbol{\theta}^*,\, x)\big(\sum_{k=0}^{m}\hat{\lambda}_k\varphi_k(\boldsymbol{\theta}^*, x)\big)/s_n(\boldsymbol{\theta}^*,\, x)\big\}\Big\} \\
&= n^{1/2}\big\{\hat{\boldsymbol{\lambda}} - \mathrm{E}_0\{\boldsymbol{h}_n(\boldsymbol{\theta}^*,\, x)\}\big\} \\
&= \mathbf{0},
\end{aligned}$$

where the second last equality holds because $s_n(\boldsymbol{\theta}^*,\, x) = \sum_{k=0}^{m}\hat{\lambda}_k\varphi_m(\boldsymbol{\theta}^*, x)$ by definition, and the last equality holds because the kth entry of $\mathrm{E}_0\{\boldsymbol{h}_n(\boldsymbol{\theta}^*,\, x)\}$ is $\mathrm{E}_0\{\hat{\lambda}_k\varphi_k(\boldsymbol{\theta}^*, x)\} = \mathrm{E}_k\hat{\lambda}_k = \hat{\lambda}_k$. Similarly,

$$\begin{aligned}
\mathrm{E}\boldsymbol{v}_{\boldsymbol{\beta}} &= n^{1/2}\sum_{k=0}^{m}\hat{\lambda}_k \mathrm{E}_k\big\{\{\partial\mathcal{L}_{n,k}(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha}\}\otimes\boldsymbol{q}(x)\big\} \\
&= n^{1/2}\big\{\hat{\boldsymbol{\lambda}}\otimes\mathrm{E}_0\{\boldsymbol{q}(x)\} - \mathrm{E}_0\{\boldsymbol{h}_n(\boldsymbol{\theta}^*,\, x)\otimes\boldsymbol{q}(x)\}\big\} \\
&= \mathbf{0}.
\end{aligned}$$

Hence $\mathrm{E}\boldsymbol{v} = n^{1/2}\sum_{k=0}^{m}\hat{\lambda}_k\boldsymbol{\mu}_{n,k} = \mathbf{0}$.

Given the above result, we have

$$\boldsymbol{v} = \boldsymbol{v} - \mathrm{E}\boldsymbol{v} = \sum_{k=0}^{m}\hat{\lambda}_k^{1/2}\big\{n_k^{-1/2}\sum_{j=1}^{n_k}\big(\partial\mathcal{L}_{n,k}(\boldsymbol{\theta}^*,\, x_{kj})/\partial\boldsymbol{\theta} - \boldsymbol{\mu}_{n,k}\big)\big\}.$$

Clearly, each term in curly brackets is a centered sum of iid random variables with finite covariance matrices. Thus, by a triangular array version of central limit theorem (Chow and Teicher, 1997, 9.1, Corollary 1), they are all asymptotically normal with appropriate covariance matrices. In addition, these terms are independent of each other, $\hat{\lambda}_k = n_k/n$ are non–random with limits ρ_k. Therefore, the linear combination is also asymptotically normal.

What left is to verify the form of the asymptotic covariance matrix. The

asymptotic covariance matrix of each term in curly brackets is given by

$$V_k = \mathrm{E}_k\left\{(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\theta})(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\theta}^\intercal)\right\} - \boldsymbol{\mu}_k\boldsymbol{\mu}_k^\intercal, \tag{2.21}$$

where $\boldsymbol{\mu}_k = \lim_{n\to\infty}\boldsymbol{\mu}_{n,k} = \mathrm{E}_k\{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\theta}\}$. Hence the overall asymptotic variance matrix is $V = \sum_{k=0}^m \rho_k V_k$.

We now show that $V = U - UWU$. First we show

$$\sum_{k=0}^m \rho_k \mathrm{E}_k\left\{(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha})(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha}^\intercal)\right\} = U. \tag{2.22}$$

By (2.15), we have

$$\begin{aligned}
&\sum_{k=0}^m \rho_k \mathrm{E}_k\left\{(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha})(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha}^\intercal)\right\}\\
=&\sum_{k=1}^m \rho_k \boldsymbol{e}_k\boldsymbol{e}_k^\intercal + \sum_{k=0}^m \rho_k \mathrm{E}_k\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s^2(\boldsymbol{\theta}^*,\, x)\}\\
&\qquad - \sum_{k=1}^m \rho_k \mathrm{E}_k\{\boldsymbol{e}_k\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s(\boldsymbol{\theta}^*,\, x)\} - \sum_{k=1}^m \rho_k \mathrm{E}_k\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{e}_k^\intercal/s(\boldsymbol{\theta}^*,\, x)\}.
\end{aligned}$$

Note that $\sum_{k=1}^m \rho_k \boldsymbol{e}_k\boldsymbol{e}_k^\intercal = \mathrm{E}_0\{\mathrm{diag}\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\}\}$,

$$\begin{aligned}
&\sum_{k=0}^m \rho_k \mathrm{E}_k\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s^2(\boldsymbol{\theta}^*,\, x)\}\\
=&\mathrm{E}_0\big\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)\{\sum_{k=0}^m \rho_k\varphi(\boldsymbol{\theta}^*,\, x)\}/s^2(\boldsymbol{\theta}^*,\, x)\big\}\\
=&\mathrm{E}_0\big\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s(\boldsymbol{\theta}^*,\, x)\big\},
\end{aligned}$$

$$\begin{aligned}
\sum_{k=1}^m \rho_k \mathrm{E}_k\{\boldsymbol{e}_k\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s(\boldsymbol{\theta}^*,\, x)\} &= \mathrm{E}_0\big\{\{\sum_{k=1}^m \rho_k\varphi(\boldsymbol{\theta}^*,\, x)\boldsymbol{e}_k\}\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s(\boldsymbol{\theta}^*,\, x)\}\big\}\\
&= \mathrm{E}_0\big\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{h}^\intercal(\boldsymbol{\theta}^*,\, x)/s(\boldsymbol{\theta}^*,\, x)\big\},
\end{aligned}$$

and similarly

$$\sum_{k=1}^{m} \rho_k \mathrm{E}_k\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{e}_k^{\mathsf{T}}/s(\boldsymbol{\theta}^*,\, x)\} = \mathrm{E}_0\big\{\boldsymbol{h}(\boldsymbol{\theta}^*,\, x)\boldsymbol{h}^{\mathsf{T}}(\boldsymbol{\theta}^*,\, x)/s(\boldsymbol{\theta}^*,\, x)\big\}.$$

By the above expressions and the definition, (2.12), of $H(\boldsymbol{\theta}^*,\, x)$, we have

$$\sum_{k=0}^{m} \rho_k \mathrm{E}_k \left\{(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha})(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha}^{\mathsf{T}})\right\} = \mathrm{E}_0\{H(\boldsymbol{\theta}^*,\, x)\} = U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}.$$

Similarly, we get

$$\sum_{k=0}^{m} \rho_k \mathrm{E}_k \left\{(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha})(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\beta}^{\mathsf{T}})\right\} = U_{\boldsymbol{\alpha}\boldsymbol{\beta}}$$

and

$$\sum_{k=0}^{m} \rho_k \mathrm{E}_k \left\{(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\beta})(\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\beta}^{\mathsf{T}})\right\} = U_{\boldsymbol{\beta}\boldsymbol{\beta}}.$$

Therefore, identity (2.22) holds.

Lastly we show that $\sum_{k=0}^{m} \rho_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathsf{T}} = UWU$. By observation (2.18), we have

$$\boldsymbol{\mu}_0\boldsymbol{\mu}_0^{\mathsf{T}} = \frac{1}{\rho_0^2}\begin{pmatrix} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\left\{\mathbf{1}_m\mathbf{1}_m^{\mathsf{T}}\right\}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\left\{\mathbf{1}_m\mathbf{1}_m^{\mathsf{T}}\right\}U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \\ U_{\boldsymbol{\beta}\boldsymbol{\alpha}}\left\{\mathbf{1}_m\mathbf{1}_m^{\mathsf{T}}\right\}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\beta}\boldsymbol{\alpha}}\left\{\mathbf{1}_m\mathbf{1}_m^{\mathsf{T}}\right\}U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \end{pmatrix},$$

and by observation (2.19), we have, for any $k = 1,\, 2,\, \ldots,\, m$,

$$\boldsymbol{\mu}_k\boldsymbol{\mu}_k^{\mathsf{T}} = \frac{1}{\rho_k^2}\begin{pmatrix} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\left\{\mathrm{diag}(\boldsymbol{e}_k)\right\}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}\left\{\mathrm{diag}(\boldsymbol{e}_k)\right\}U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \\ U_{\boldsymbol{\beta}\boldsymbol{\alpha}}\left\{\mathrm{diag}(\boldsymbol{e}_k)\right\}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\beta}\boldsymbol{\alpha}}\left\{\mathrm{diag}(\boldsymbol{e}_k)\right\}U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \end{pmatrix}.$$

Hence,

$$\sum_{k=0}^{m} \rho_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathsf{T}} = \begin{pmatrix} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} T U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} T U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \\ U_{\boldsymbol{\beta}\boldsymbol{\alpha}} T U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\beta}\boldsymbol{\alpha}} T U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \end{pmatrix} = UWU. \tag{2.23}$$

By (2.21), (2.22) and (2.23), we have

$$V = \sum_{k=0}^{m} \rho_k \mathrm{E}_k \left\{ (\partial \mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial \boldsymbol{\theta})(\partial \mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial \boldsymbol{\theta}^{\mathsf{T}}) \right\} - \sum_{k=0}^{m} \rho_k \boldsymbol{\mu}_k \boldsymbol{\mu}_k^{\mathsf{T}} = U - UWU,$$

which completes the proof.

2.5.3 Lemma 2.3: Concavity of the DEL

In this subsection, we show that $l_n(\boldsymbol{\theta})$ is a concave function. To show the concavity of $l_n(\boldsymbol{\theta})$, it suffices to show that $\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}} \leq 0$ for all values of $\boldsymbol{\theta}$, and since $\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}} = \sum_{k,\,j} \left\{ \partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x_{kj})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}} \right\}$, it is enough to show that

$$\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}} \leq 0,$$

for any given $\boldsymbol{\theta}$, x and k.

Recall that $\boldsymbol{\theta}$ is composed of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$. We first show that $\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathsf{T}} < 0$. By (2.16),

$$\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathsf{T}} = -H_n(\boldsymbol{\theta},\, x)/s_n(\boldsymbol{\theta},\, x).$$

Noticing that $H_n(\boldsymbol{\theta},\, x)$ has a similar expression as $H(\boldsymbol{\theta},\, x)$, which is positive definite as we have shown in the proof of Theorem 2.1, only with ρ_k replace by $\hat{\lambda}_k$, we know that $H_n(\boldsymbol{\theta},\, x)$ is also positive definite. Along with the fact that $s_n(\boldsymbol{\theta},\, x) > 0$, we have $\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^{\mathsf{T}} = -H_n(\boldsymbol{\theta},\, x)/s_n(\boldsymbol{\theta},\, x) < 0$.

Secondly, by expression (2.16), we see that $\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\mathsf{T}}$ is just a

row and column permuted version of the matrix

$$\left\{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\intercal\right\} \otimes \left\{\boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x)\right\},$$

which is negative semidefinite for any value of $\boldsymbol{\theta}$ and x because $\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x)/\partial\boldsymbol{\alpha}\partial\boldsymbol{\alpha}^\intercal < 0$ and $\boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x) \geq 0$. Therefore

$$\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal = \sum_{k,\,j} \left\{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\theta},\, x_{kj})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal\right\} \leq 0$$

for any value of $\boldsymbol{\theta}$, and so $l_n(\boldsymbol{\theta})$ is concave.

Lastly, when $\boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x) > 0$ for a given x, $-\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal > 0$. Hence, when $\sum_{k,j} \boldsymbol{Q}(x_{kj})\boldsymbol{Q}^\intercal(x_{kj}) > 0$, $-\partial^2 l_n(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal > 0$ and the DEL is strictly concave. The proof is complete.

2.5.4 Lemma 2.4: $\sqrt[3]{n}$–consistency of the MELE

We show in this subsection that the MELE $\hat{\boldsymbol{\theta}}$ is attained in an interior point of a $\sqrt[3]{n}$–neighbourhood of the true parameter value $\boldsymbol{\theta}^*$ with probability tending one. Note that $\hat{\boldsymbol{\theta}}$ is a maximum point of the DEL $l_n(\boldsymbol{\theta})$. The idea is to show that for any $\boldsymbol{\theta}$ on the surface of the closed ball $B_{\boldsymbol{\theta}^*} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\| \leq n^{-1/3}\}$, $l_n(\boldsymbol{\theta}) < l_n(\boldsymbol{\theta}^*)$ with probability tending one. Then, by concavity of $l_n(\boldsymbol{\theta})$, all the maximum points, including $\hat{\boldsymbol{\theta}}$, must be interior points of $B_{\boldsymbol{\theta}^*}$ with probability tending to one.

We first expand $l_n(\boldsymbol{\theta})$ around $\boldsymbol{\theta}^*$. Recalling that $\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta} = \sqrt{n}\boldsymbol{v}$ and $\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal = -nU_n$, we get

$$l_n(\boldsymbol{\theta}) = l_n(\boldsymbol{\theta}^*) + \sqrt{n}\boldsymbol{v}^\intercal(\boldsymbol{\theta} - \boldsymbol{\theta}^*) - (1/2)n(\boldsymbol{\theta} - \boldsymbol{\theta}^*)^\intercal U_n(\boldsymbol{\theta} - \boldsymbol{\theta}^*) + \epsilon_n,$$

where

$$\epsilon_n = \frac{n}{6}\sum_{i,j,k}\frac{1}{n}\frac{\partial^3 l_n(\tilde{\boldsymbol{\theta}})}{\partial\theta_i\partial\theta_j\partial\theta_k}(\hat{\theta}_i-\theta_i^*)(\hat{\theta}_j-\theta_j^*)(\hat{\theta}_k-\theta_k^*),$$

with $\tilde{\boldsymbol{\theta}}$ being some parameter value. Notice that, for any value of $\boldsymbol{\theta}$,

$$\begin{aligned}\left|n^{-1}\partial^3 l_n(\boldsymbol{\theta})/\partial\theta_i\partial\theta_s\partial\theta_t\right| &= \left|n^{-1}\sum_{k=0}^{m}\sum_{j=1}^{n_k}\partial^3\mathcal{L}_{n,k}(\boldsymbol{\theta},x_{kj})/\partial\theta_i\partial\theta_s\partial\theta_t\right| \\ &\le \sum_{k=0}^{m}\left\{n_k^{-1}\sum_{j=1}^{n_k}|\partial^3\mathcal{L}_{n,k}(\boldsymbol{\theta},x_{kj})/\partial\theta_i\partial\theta_s\partial\theta_t|\right\} \\ &\le \sum_{k=0}^{m}\{n_k^{-1}\sum_{j=1}^{n_k}\|\boldsymbol{Q}(x_{kj})\|^3\},\end{aligned}$$

where the last inequality is by (2.17). Since $\|\boldsymbol{Q}(x)\|^3$ is integrable, and x_{kj} are iid across j for each k, by strong law of large numbers, the last term on the RHS of the above inequality is of $O(1)$, and so is $n^{-1}\partial^3 l_n(\boldsymbol{\theta})/\partial\theta_i\partial\theta_j\partial\theta_l$. This implies that for any $\boldsymbol{\theta}=\boldsymbol{\theta}^*+O_p(n^{-1/3})$, we have $\epsilon_n = O_p(1)$.

By the above result, for any $\boldsymbol{\theta}$ on the surface of the closed ball $B_{\boldsymbol{\theta}^*}$, i.e. for any $\boldsymbol{\theta}=\boldsymbol{\theta}^*+\boldsymbol{a}n^{-1/3}$ with $\|\boldsymbol{a}\|=1$, we have

$$\begin{aligned}l_n(\boldsymbol{\theta}^*+\boldsymbol{a}n^{-1/3})-l_n(\boldsymbol{\theta}^*) &= n^{1/6}\boldsymbol{v}^\intercal\boldsymbol{a}-(1/2)n^{1/3}\boldsymbol{a}^\intercal U_n\boldsymbol{a}+O(1) \\ &= n^{1/6}\boldsymbol{v}^\intercal\boldsymbol{a}-(1/2)n^{1/3}\boldsymbol{a}^\intercal U\boldsymbol{a}+o(n^{1/3}),\end{aligned}$$

where the last equality is by the fact that $U_n = U+o(1)$. Let c be the smallest eigenvalue of U. By Theorem 2.2, $\boldsymbol{v}=O_p(1)$, so we get

$$\begin{aligned}l_n(\boldsymbol{\theta}^*+\boldsymbol{a}n^{-1/3})-l_n(\boldsymbol{\theta}^*) &= n^{1/6}\boldsymbol{v}^\intercal\boldsymbol{a}-(1/2)n^{1/3}\boldsymbol{a}^\intercal U\boldsymbol{a}+o(n^{1/3}), \\ &\le O_p(n^{1/6})-(1/2)cn^{1/3}+o(n^{1/3}) \\ &= -(1/2)cn^{1/3}+o_p(n^{1/3}),\end{aligned}$$

uniformly in $\boldsymbol{a}$ that satisfies $\|\boldsymbol{a}\| = 1$. Since U is positive definite, $c > 0$. Clearly, with probability tending one, the last term on RHS is strictly smaller than 0 and hence $l_n(\boldsymbol{\theta}^* + \boldsymbol{a}n^{-1/3}) < l_n(\boldsymbol{\theta}^*)$, as $n \to \infty$. By the continuity of $l_n(\boldsymbol{\theta})$, $l_n(\boldsymbol{\theta})$ must have a maximum in the interior of the ball $B_{\boldsymbol{\theta}^*}$ with probability tending one.

2.5.5 Theorem 2.5: Asymptotic normality of the MELE

We now show the asymptotic normality of the MELE $\hat{\boldsymbol{\theta}}$. The idea is to show that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is well approximated by $U^{-1}\boldsymbol{v}$. As a reminder, U is the information matrix, and $\boldsymbol{v} = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}$.

Expanding $n^{-1/2}\partial l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}$ at $\boldsymbol{\theta}^*$, we get

$$n^{-1/2}\partial l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta} = \boldsymbol{v} - U_n\{\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} + \boldsymbol{\epsilon}_n,$$

where $U_n = n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal$ is the empirical information matrix and $\boldsymbol{\epsilon}_n$ is a vector of length $m(d+1)$ whose i^{th} entry is

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\intercal \left\{ \frac{1}{n}\frac{\partial^3 l_n(\tilde{\boldsymbol{\theta}})}{\partial\theta_i\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal} \right\} (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*),$$

with θ_i being the ith component of $\boldsymbol{\theta}$ and $\tilde{\boldsymbol{\theta}}$ being some parameter value. We have shown in the proof of Lemma 2.4 that the third order derivatives of $l_n(\boldsymbol{\theta})$ are uniformly bounded by an integrable function, so $n^{-1}\partial^3 l_n(\tilde{\boldsymbol{\theta}})/\partial\theta_i\partial\theta_j\partial\theta_l = O_p(1)$. This, along with the fact that $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = O_p(n^{-1/3})$, implies that $\boldsymbol{\epsilon}_n = o_p(1)$ and the expansion can be written as

$$n^{-1/2}\partial l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta} = \boldsymbol{v} - U_n\{\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)\} + o_p(1).$$

Note that $\partial l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta} = 0$ because the MELE $\hat{\boldsymbol{\theta}}$ is the point at which the smooth function $l_n(\boldsymbol{\theta})$ is maximized. By equating the left hand side (LHS)

of the above expansion to $\mathbf{0}$ and reorganizing terms, we get

$$U_n\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \boldsymbol{v} + o_p(1). \tag{2.24}$$

By Theorem 2.2, $\boldsymbol{v}$ is asymptotically normal, hence of $O_p(1)$, so the LHS of the above equality must be $O_p(1)$. Note that, on the LHS, the first factor U_n has a positive definite limit by Theorem 2.1. We then deduce that the second factor, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, must also be of $O_p(1)$. Furthermore, by Theorem 2.1, $U_n = U + o_p(1)$. Hence,

$$U_n\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \big(U + o_p(1)\big)\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = U\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + o_p(1).$$

Substituting the LHS of (2.24) by the RHS of the above equality and reorganizing terms, we get

$$U\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = \boldsymbol{v} + o_p(1).$$

Since U is positive definite, we can left multiply U^{-1} on both sides of the above equality to get

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = U^{-1}\boldsymbol{v} + o_p(1). \tag{2.25}$$

Combining the above equality and asymptotic normality of $\boldsymbol{v}$, we get the claimed result that $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) \to N(\mathbf{0}, U^{-1} - W)$ in distribution as $n \to \infty$.

Chapter 3

Dual Empirical Likelihood Ratio Test for Hypotheses about DRM Parameters

This chapter develops a DEL ratio test for composite hypotheses about the parameter of the DRM based on independent samples from different populations. The proposed test encompasses testing differences in population distributions as a special case. The DEL ratio test statistic is found to have a classical chi–square null limiting distribution and a non–central chi–square limiting distribution under a class of local alternatives. The null limiting distribution is useful for approximating the p–values of the proposed test; the limiting distribution under the local alternative model is useful for approximating the power of the proposed test, calculating the sample size required for achieving a given power, and comparing the local asymptotic powers of DEL ratio tests formulated in different ways. Simulation studies show that this test has better power properties than all potential competitors adopted to the multiple sample problem under the investigation, and is robust to model misspecification. The proposed test is then applied to assess strength properties of lumber with intuitively reasonable implications for the forest industry.

3.1 Introduction

An important task of the long term monitoring project is to monitor change in population distributions of the strength of lumber produced over the years. Recall Section 2.3 that, under the DRM (2.1), the equality of two distribution functions is equivalent to the equality of the corresponding DRM slope parameters: $F_k = F_0$ is equivalent to $\boldsymbol{\beta}_k = 0$ and $F_k = F_j$ is equivalent to $\boldsymbol{\beta}_k = \boldsymbol{\beta}_j$, $k, j = 1, \ldots, m$. Hence under the DRM, a hypothesis about the differences in distribution functions ultimately translates to a hypothesis about the DRM parameter $\boldsymbol{\beta}$.

In principle, for a linear hypothesis about the DRM parameter $\boldsymbol{\beta}$:

$$H_0: \ A\boldsymbol{\beta} = \boldsymbol{c} \quad \text{against} \quad H_1: \ A\boldsymbol{\beta} \neq \boldsymbol{c}$$

for some given matrix A and vector $\boldsymbol{c}$, a Wald type test (Fokianos et al., 2001, (17)) can be easily constructed based on the asymptotic normality of the MELE $\hat{\boldsymbol{\theta}}$. According to Theorem 2.5,

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*) \longrightarrow N(0, \Sigma_{\boldsymbol{\beta}})$$

in distribution for some positive definite covariance matrix $\Sigma_{\boldsymbol{\beta}}$. Under the null of the above linear hypothesis, $A\boldsymbol{\beta}^* = \boldsymbol{c}$, so

$$\sqrt{n}(A\hat{\boldsymbol{\beta}} - A\boldsymbol{\beta}^*) = \sqrt{n}(A\hat{\boldsymbol{\beta}} - \boldsymbol{c}) \longrightarrow N(0, A\Sigma_{\boldsymbol{\beta}}A^{\intercal})$$

in distribution. Let $\hat{\Sigma}_{\boldsymbol{\beta}}$ be a consistent estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. When A is a full rank $q \times md$ matrix with $q \leq md$, the dimension of $\boldsymbol{\beta}$, the test statistic

$$W_n = n(A\hat{\boldsymbol{\beta}} - \boldsymbol{c})^{\intercal}(A\hat{\Sigma}_{\boldsymbol{\beta}}A^{\intercal})^{-1}(A\hat{\boldsymbol{\beta}} - \boldsymbol{c})$$

has a chi–square limiting distribution, χ^2_q, of q degrees of freedom. Such a

test, however, suffers from a few drawbacks. First, it is usually not very powerful when the sample size is not very large because the estimation accuracy of the asymptotic covariance estimator $\hat{\Sigma}_{\boldsymbol{\beta}}$ in the denominator could be low in that case. Second, it is not invariant to transformations: if we transform the parameter and the hypothesis accordingly, the value of the test statistic and the corresponding p–value of the test may be different form those based on the original scale.

In contrast to Wald tests, likelihood ratio tests are usually more powerful because they do not need an estimation of the asymptotic covariance matrix, and are invariant to transformation. However, as described in Section 2.3, a simplistic application of the straightforward EL ratio test to hypotheses that compare the slope DRM parameter $\boldsymbol{\beta}_k$ of different population distributions is negated by the non–regularity of the DRM, under which the EL function is not well–defined in a neighbourhood of the true parameter value $\boldsymbol{\beta}^*$ if $\boldsymbol{\beta}_k^* = \boldsymbol{\beta}_j^*$ for some $k, j \in \{0, 1, \ldots, m\}$. Therefore we look for a likelihood ratio test based on the DEL because it is well–defined for all values of $\boldsymbol{\theta}$ in the corresponding Euclidean space, has a simple analytical form, and is concave.

The next section presents a *DEL ratio* (DELR) test for a general composite hypothesis about the DRM slope parameter $\boldsymbol{\beta}$, which encompasses testing differences in population distributions as a special case, and gives the limiting distributions of the proposed test statistic under both the null and a class of local alternatives of that hypothesis testing problem. The proofs of these properties are given in Section 3.6. Section 3.3 assesses, via simulation, the finite sample distributions of the DELR statistic under the null and local alternative models, as well as the power of the DELR test. The robustness of the proposed test against the misspecification of the DRM is studied via simulation in Section 3.4. In Section 3.5, we apply the DELR test to lumber bending strength data and find that the outcome leads to intuitively reasonable implications for the forest industry.

3.2 DELR statistic and its limiting distributions

Recall Section 2.3 that we defined the DEL as

$$l_n(\boldsymbol{\theta}) = -\sum_{k,j} \log\Big\{\sum_{r=0}^{m} \hat{\lambda}_r \exp\big\{\alpha_r + \boldsymbol{\beta}_r^{\intercal}\boldsymbol{q}(x_{kj})\big\}\Big\} + \sum_{k,j}\big\{\alpha_k + \boldsymbol{\beta}_k^{\intercal}\boldsymbol{q}(x_{kj})\big\},$$

where $\hat{\lambda}_r = n_r/n$, n_r is the size of the rth sample and n is the total sample size. The MELE $\boldsymbol{\theta}$ is the point at which the DEL is maximized, $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\, l_n(\boldsymbol{\theta})$. This DEL, unlike the EL, is well–defined for all values of $\boldsymbol{\theta}$, so we expect to derive the limiting distribution of the corresponding likelihood ratio statistic using classical techniques. Under a two–sample DRM ($m+1=2$), Keziou and Leoni-Aubin (2008) found that for simple hypothesis $H_0 : \boldsymbol{\beta}_1 = \mathbf{0}$, or equivalently $H_0 : F_1 = F_0$, the corresponding likelihood ratio test statistic, $2l_n(\hat{\boldsymbol{\theta}})$, has the usual chi–square limiting distribution, χ^2_d, with d degrees of freedom.

The success of Keziou and Leoni-Aubin leads us to wonder if the result is more generally applicable. In the long term monitoring program for lumber quality, we may encounter similar situations as follows: we have five ($m+1 = 5$) lumber samples, with the first two being spruce samples, the third and fourth being pine samples, and the fifth being a Douglas fir sample; we are interested in testing if the two spruce populations (F_0 & F_1) have the same overall quality, which amounts to the hypothesis testing problem of

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0}.$$

If we also concerned about the stability of the qualities of the two pine populations (F_2 & F_3) simultaneously, the hypothesis testing problem would

be

$$H_0 : \boldsymbol{\beta}_1 = \mathbf{0} \text{ and } \boldsymbol{\beta}_2 = \boldsymbol{\beta}_3 \qquad \text{against} \qquad H_1 : \boldsymbol{\beta}_1 \neq \mathbf{0} \text{ or } \boldsymbol{\beta}_2 \neq \boldsymbol{\beta}_3.$$

The first hypothesis testing problem above, despite of its simple appearance, is a composite hypothesis testing problem that is fundamentally different from the two–sample problem that Keziou and Leoni-Aubin has studied. It is clear that three other distributions F_2, F_3 and F_4 are also modeled by the DRM, but their corresponding slope parameters $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$ and $\boldsymbol{\beta}_4$ are nuisance parameters that are not specified in the hypothesis. The second hypothesis testing problem above is more complicated and also has a nuisance parameter $\boldsymbol{\beta}_4$ for the fifth population which is not specified in the hypothesis. Both testing problems are not covered by the results of Keziou and Leoni-Aubin. In addition, their proof of the result does not readily extend to more complicated hypotheses, because it is tailored for true parameter $\boldsymbol{\beta}^* = 0$, in which case the analytical expression of a key quadratic form that approximates the corresponding DELR statistic is much simpler than that for a composite hypothesis.

The above limitation of Keziou and Leoni-Aubin's result leads us to investigate the properties of the DEL ratio in a much more general setting.

3.2.1 DELR statistic and its null limiting distribution

All the above hypothesis testing problems can be abstractly stated as testing

$$H_0 : \ \boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0} \quad \text{against} \quad H_1 : \ \boldsymbol{g}(\boldsymbol{\beta}) \neq \mathbf{0} \tag{3.1}$$

for some smooth function $\boldsymbol{g} : \mathbb{R}^{md} \to \mathbb{R}^q$, with $q \leq md$, the length of $\boldsymbol{\beta}$. *We will always assume that $\boldsymbol{g}$, is thrice differentiable with a full rank Jacobian matrix $\partial \boldsymbol{g}/\partial \boldsymbol{\beta}$.* The parameters $\{\alpha_k\}$ are usually not a part of the hypothesis, because, by (2.2), their values are fully determined by the $\{\boldsymbol{\beta}_k\}$ and F_0 under

the DRM assumption:

$$\alpha_k = -\log \int \exp\{\boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\} dF_0,$$

although they are regarded as independent parameters in the DEL.

Let $\tilde{\boldsymbol{\theta}}$ be the point at which the maximum of the DEL $l_n(\boldsymbol{\theta})$ is attained under the null constraint $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$. Recall that the MELE $\hat{\boldsymbol{\theta}}$ is the point at which the $l_n(\boldsymbol{\theta})$ is maximized without the null constraint. The DELR test statistic is defined to be

$$R_n = 2\{l_n(\hat{\boldsymbol{\theta}}) - l_n(\tilde{\boldsymbol{\theta}})\}.$$

Does R_n have the properties of a regular likelihood ratio test statistic? The answer is positive and we state the result as follows, the proof of which is given in Section 3.6.

Recall the conditions of Theorem 2.1: we have $m+1$ random samples of sizes n_k, $k = 0, 1, \ldots, m$, from populations with distributions of the DRM form given in (2.1) and a true parameter value $\boldsymbol{\theta}^*$ such that $\int \exp\{\boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\} dF_0 < \infty$ for $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}^*$, $\int \boldsymbol{Q}(x)\boldsymbol{Q}^\intercal(x) dF_0$ is positive definite with $\boldsymbol{Q}^\intercal(x) = (1, \boldsymbol{q}^\intercal(x))$, and $\hat{\lambda}_k = n_k/n = \rho_k + o(1)$, where $n = \sum_{k=0}^m n_k$ is the total sample size, for some $\rho_k \in (0, 1)$.

Theorem 3.1 (Null limiting distribution of the DELR statistic)**.** *Adopt the conditions posited in Theorem 2.1. Under the null hypothesis $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$, $R_n \to \chi_q^2$ in distribution as $n \to \infty$, where q is the dimension of the null mapping $\boldsymbol{g}(\cdot)$ and χ_q^2 is a chi–squared random variable with q degrees of freedom.*

When $m = 1$, Theorem 3.1 reduces to the result of Keziou and Leoni-Aubin (2008) for $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta}_1$. This Theorem covers additional ground, for instance, the two composite hypothesis testing examples given at the beginning of this section and the case when we test the hypothesis $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \mathbf{0}$

based on all $m + 1 = 5$ samples.

The null limiting distribution of R_n is most useful for approximating the p–value, p, of a DELR test:

$$p \approx \Pr(\chi_q^2 \geq R_n).$$

At the significance level of α, we reject the null hypothesis of $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$ when $R_n \geq \chi^2_{q,\,1-\alpha}$, where $\chi^2_{q,\,p}$ in general is the pth quantile of χ_q^2 distribution.

3.2.2 Limiting distribution of the DELR statistic under local alternatives

Theorem 3.1 provides an approximation to the p–value of a test but it does not give the power of the test. As is well known, most sensible tests are consistent: the asymptotic power at any fixed alternative model goes to 1 as the sample size $n \to \infty$; this is true for DELR test. Hence, instead of looking at a fixed alternative, we here study the asymptotic power of the DELR test under a class of local alternatives, under which the limiting distribution of R_n usually is not a point mass. The finite–sample power properties of the test are studied by simulation in Section 3.3.2.

Let $\{\boldsymbol{\beta}_k^*\}$ be a set of parameter values which form a null model satisfying $H_0 : \boldsymbol{g}(\boldsymbol{\beta}) = 0$ under the DRM assumption. Let

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^* + n_k^{-1/2}\boldsymbol{c}_k, \tag{3.2}$$

for some constants $\{\boldsymbol{c}_k\}$, be a set of parameter values which form a local alternative. We denote the distribution functions corresponding to $\boldsymbol{\beta}_k^*$ and $\boldsymbol{\beta}_k$ as F_k and G_k with $G_0 = F_0$, respectively. Note that the $\{G_k\}$ are placed at $n^{-1/2}$ distance from the $\{F_k\}$. As $n \to \infty$, the limiting distribution of R_n under this local alternative is usually non–degenerate and provides useful information on the power of the test.

We now express the null hypothesis $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$ in an equivalent form. Recall that $\boldsymbol{g} : \mathbb{R}^{md} \to \mathbb{R}^q$ is thrice differentiable in a neighbourhood of $\boldsymbol{\beta}^*$ with a full rank Jacobian matrix evaluated at $\boldsymbol{\beta}^*$. Denote $\bigtriangledown = \partial \boldsymbol{g}(\boldsymbol{\beta}^*)/\partial \boldsymbol{\beta}$ and partition $\bigtriangledown$ into $(\bigtriangledown_1, \bigtriangledown_2)$, with q and $md-q$ columns respectively. When $q < md$, by the implicit function theorem (Zorich, 2004, 8.5.4, Theorem 1), there exists a unique function $\mathcal{G}$: $\mathbb{R}^{md-q} \to \mathbb{R}^{md}$, such that $\boldsymbol{g}(\boldsymbol{\beta}) = 0$ if and only if $\boldsymbol{\beta} = \mathcal{G}(\boldsymbol{\gamma})$ for some $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ in a corresponding neighbourhoods of $\boldsymbol{\beta}^*$ and $\boldsymbol{\gamma}^*$ respectively. In addition, $\mathcal{G}$ is also thrice differentiable in a neighbourhood of $\boldsymbol{\gamma}^*$, and its Jacobian matrix evaluated at $\boldsymbol{\gamma}^*$,

$$J = \partial \mathcal{G}(\boldsymbol{\gamma}^*)/\partial \boldsymbol{\gamma},$$

has a full rank. Furthermore, if $\bigtriangledown_1$ has a full rank, then

$$J = (-(\bigtriangledown_1^{-1} \bigtriangledown_2)^\intercal,\ I_{md-q})^\intercal, \tag{3.3}$$

where I_k is an identity matrix of size $k \times k$.

Let U be the information matrix (2.13) under the null model H_0 represented by the $\{\boldsymbol{\beta}_k^*\}$ and $\{F_k\}$.

Theorem 3.2 (Limiting distribution of the DELR under local alternatives)**.** *Under the conditions of Theorem 2.1 and local alternative defined by (3.2),*

$$R_n \to \chi_q^2(\delta^2)$$

in distribution as $n \to \infty$, where $\chi_q^2(\delta^2)$ is a non–central chi–square random variable with q degrees of freedom and a nonnegative non–central parameter

$$\delta^2 = \begin{cases} \boldsymbol{\eta}^\intercal \big\{\Lambda - \Lambda J \big(J^\intercal \Lambda J\big)^{-1} J^\intercal \Lambda\big\} \boldsymbol{\eta} & \text{if } q < md \\ \boldsymbol{\eta}^\intercal \Lambda \boldsymbol{\eta} & \text{if } q = md \end{cases}$$

where $\boldsymbol{\eta}^\intercal = (\rho_1^{-1/2} \boldsymbol{c}_1^\intercal,\ \rho_2^{-1/2} \boldsymbol{c}_2^\intercal,\ \ldots, \rho_m^{-1/2} \boldsymbol{c}_m^\intercal)$ and $\Lambda = U_{\boldsymbol{\beta}\boldsymbol{\beta}} - U_{\boldsymbol{\beta}\boldsymbol{\alpha}} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1} U_{\boldsymbol{\alpha}\boldsymbol{\beta}}$.

Moreover, $\delta^2 > 0$ unless $\boldsymbol{\eta}$ is in the column space of J.

The proof is given in Section 3.6. This result is useful for: (1) computing local power of the DELR test under specific distributional settings, (2) calculaing required sample size for achieving a certain power at a given alternative, and (3) comparing the powers of DELR tests formulated in different ways, which helps us to determine the most efficient use of information contained in multiple samples. We illustrate the first two points using the examples below, and discuss the last point in Chapter 4.

Example 3.1 (Computing the local asymptotic power of DELR test for a composite hypothesis)**.** *Consider the situation where* $m+1=3$*, samples are from a DRM with basis function* $\boldsymbol{q}(x)=(x,\,\log x)^{\intercal}$*, and the sample proportions are* $(0.4,\,0.3,\,0.3)$*. Let* F_k*,* $k=1,\,2$*, be the distributions with parameters* $\boldsymbol{\beta}_1^*=(-1,\,1)^{\intercal}$ *and* $\boldsymbol{\beta}_2^*=(-2,\,2)^{\intercal}$*. Let* H_0 *be* $\boldsymbol{g}(\boldsymbol{\beta})=2\boldsymbol{\beta}_1-\boldsymbol{\beta}_2=0$*. Consider the local alternative*

$$\boldsymbol{\beta}_k=\boldsymbol{\beta}_k^*+n_k^{-1/2}\boldsymbol{c}_k,\ \textit{for } k=1,\,2, \tag{3.4}$$

with $\boldsymbol{c}_1=(2,\,3)^{\intercal}$ *and* $\boldsymbol{c}_2=(-1,\,0)^{\intercal}$*.*

We find $\bigtriangledown=(2I_2,\,-I_2)$ *so* $J=((1/2)I_2,\,I_2)$*, and* $\boldsymbol{\eta}\approx(3.65,\,5.48,\,-1.83,\,0)^{\intercal}$*. The information matrix* U *is* F_0 *dependent. When* F_0 *is* $\Gamma(2,\,1)$*, where in general* $\Gamma(\lambda,\,\kappa)$ *denotes the gamma distribution with shape* λ *and rate* κ*, we obtain the information matrix* (2.13) *and hence* Λ*, based on numerical computation. We therefore get* $\delta^2\approx 10.29$ *based on formula given in the above theorem.*

The null limiting distribution of R_n *is* χ^2_2*. At the 5% level, the null is rejected when* $R_n\geq\chi^2_{2,\,0.95}\approx 5.99$*. Hence under the local alternative, the power of the DELR test is approximately* $Pr(\chi^2_2(10.29)\geq 5.99)\approx 0.83$*.*

Example 3.2 (Sample size calculation for Example 3.1)**.** *Adopt the settings of Example 3.1. Suppose we require the power of the DELR test to be at least* 0.8 *at the alternative of* $\boldsymbol{\beta}_1=\boldsymbol{\beta}_1^*+(0.5,\,1.5)^{\intercal}$ *and* $\boldsymbol{\beta}_2=\boldsymbol{\beta}_2^*+(0.5,\,0.5)^{\intercal}$ *at the* 5% *significance level. Recall that the sample proportions*

are (0.4, 0.3, 0.3). *This alternative corresponds to a local alternative of the form* (3.4) *with with* $\boldsymbol{c}_1 = (0.5\sqrt{n_1},\, 1.5\sqrt{n_1})^\intercal = 0.5(\sqrt{0.3n},\, 3\sqrt{0.3n})^\intercal$ *and* $\boldsymbol{c}_2 = (0.5\sqrt{n_2},\, 0.5\sqrt{n_2})^\intercal = 0.5(\sqrt{0.3n},\, \sqrt{0.3n})^\intercal$.

Using the above $\boldsymbol{c}_1$ *and* $\boldsymbol{c}_2$, *we obtain* $\boldsymbol{\eta} = (0.3^{-1/2}\boldsymbol{c}_1^\intercal,\, 0.3^{-1/2}\boldsymbol{c}_2^\intercal)^\intercal = 0.5\sqrt{n}(1,\, 3,\, 1,\, 1)^\intercal$ *as a function of the total sample size* n. *With the same* J, F_0 *and* U *as obtained in Example 3.1, and applying the formula given in Theorem 3.2, we obtain the non–central parameter* δ^2 *as a function of* n, *which we denote as* $\delta^2(n)$. *To attain a minimal power of* 0.8, *we solve*

$$Pr(\chi_2^2(\delta^2(n)) \geq \chi^2_{2,\,0.95}) \geq 0.8$$

for the total sample size n *and get* $n \geq 50$.

3.2.3 On the condition for the positiveness of the non–central parameter

A meaningful test should be unbiased: at the significance level of α, for any given alternative, the power of the test should be at least as large as α. Is the DELR test asymptotically unbiased under the local alternatives of the form (3.2)? The answer is positive. Recall that at the significance level of α, we reject the null hypothesis of $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{0}$ when $R_n \geq \chi^2_{q,\,1-\alpha}$. Hence, by Theorem 3.2, at any given local alternative of the form (3.2), the asymptotic power of the test is

$$\lim_{n\to\infty} \Pr(R_n \geq \chi^2_{q,1-\alpha}) = \Pr(\chi^2_q(\delta^2) \geq \chi^2_{q,1-\alpha}). \tag{3.5}$$

By a result about non–central chi–square distribution (Johnson et al., 1995, (29.25a)), if $0 \leq \delta_1^2 < \delta_2^2$, then for any $x > 0$,

$$\Pr(\chi^2_d(\delta_1^2) \geq x) < \Pr(\chi^2_d(\delta_2^2) \geq x). \tag{3.6}$$

Note that a non–central chi–square distribution with a 0 non–central parameter is just a usual chi–square distribution. Therefore, in view of (3.5), the local asymptotic power of the DELR test satisfy

$$\Pr(\chi_q^2(\delta^2) \geq \chi_{q,1-\alpha}^2) \geq \Pr(\chi_q^2 \geq \chi_{q,1-\alpha}^2) = \alpha$$

with equality if and only if $\delta^2 = 0$. Thus the DELR test is asymptotically unbiased under the local alternative model (3.2).

In practice, we always hope that the power of a test at an alternative is strictly larger than the significance level. We now take one step further to study in what situations the local asymptotic power of DELR test is strictly larger than the significance level α, or equivalently, $\delta^2 > 0$. Roughly, the answer lies in whether a $\boldsymbol{\beta}$ defined by the local alternative model (3.2) is truly a local alternative.

We first look at the case that q, the dimension of the $\boldsymbol{g}$ function in the hypothesis (3.1), equals md, the dimension of the DRM slope parameter $\boldsymbol{\beta}$. In this case, by the inverse function theorem (Zorich, 2004, 8.6.1, Theorem 1), $\boldsymbol{g}$ is invertible at $\boldsymbol{\beta}^*$, i.e. $\boldsymbol{\beta}^* = \boldsymbol{g}^{-1}(\mathbf{0})$. Hence $\boldsymbol{g}$ defines a simple hypothesis testing problem with $\boldsymbol{\beta}$ being fully specified to be $\boldsymbol{g}^{-1}(\mathbf{0})$ in the null. Then any $\boldsymbol{\beta}$ value defined by the local alternative model (3.2), as long as not all the $\{\boldsymbol{c}_k\}$ are 0, is a real alternative, i.e. does not satisfy the null constraint of $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$ for any given sample size n. In this case, by Theorem 3.2, the non–central parameter is $\delta^2 = \boldsymbol{\eta}^\intercal \Lambda \boldsymbol{\eta}$. Since Λ is positive definite and $\boldsymbol{\eta} \neq \mathbf{0}$, we have $\delta^2 > 0$.

When $q < md$, for some choices of $\boldsymbol{c}_k$, the $\boldsymbol{\beta}$ defined by the local alternative model may still satisfy the null model of $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$. For example, let $m+1 = 3$, the null hypothesis be $\boldsymbol{g}(\boldsymbol{\beta}) = 2\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 = \mathbf{0}$, and $\boldsymbol{\beta}^*$ be a parameter value satisfying this null model. Then, when $n_1 = n_2$, the parameter $\boldsymbol{\beta}$ satisfying (3.2) with $\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + n_1^{-1/2}\boldsymbol{v}$ and $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_1^* + 2n_2^{-1/2}\boldsymbol{v}$ for some vector $\boldsymbol{v}$ always satisfies the null model for any given n. In this case, $\delta^2 = 0$ as we will see soon.

The mathematical condition that $\delta^2 > 0$ is given in Theorem 3.2: $\boldsymbol{\eta}$ is not in the column space of J, the Jacobian matrix $\partial\mathcal{G}(\boldsymbol{\gamma}^*)/\partial\boldsymbol{\gamma}$. We now show that this condition is equivalent to: *for the $\boldsymbol{\beta}$ defined by the local alternative model* (3.2)*, the speed that $\boldsymbol{g}(\boldsymbol{\beta})$ converges to $\mathbf{0}$ is no faster than the speed that $\boldsymbol{\beta}$ converges to $\boldsymbol{\beta}^*$, which is on the order of $O(n^{-1/2})$.*

Let $\boldsymbol{\beta}^*$ be a parameter value satisfying the null model $\boldsymbol{g}(\boldsymbol{\beta}) = 0$, and $\boldsymbol{\beta}$ satisfy the local alternative model (3.2). Recall that we have defined $\bigtriangledown = \partial\boldsymbol{g}(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}$. Expanding $\boldsymbol{g}(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^*$ and using the fact that $\boldsymbol{g}(\boldsymbol{\beta}^*) = 0$, we get

$$\begin{aligned}\boldsymbol{g}(\boldsymbol{\beta}) &= \boldsymbol{g}(\boldsymbol{\beta}^*) + \{\partial\boldsymbol{g}(\boldsymbol{\beta}^*)/\partial\boldsymbol{\beta}\}(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2)\\ &= \bigtriangledown(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2).\end{aligned}$$

Notice that, when $\boldsymbol{\beta}$ satisfies the local alternative model (3.2), we have

$$\boldsymbol{\beta} - \boldsymbol{\beta}^* = n^{-1/2}(\hat{\lambda}_1^{-1/2}\boldsymbol{c}_1^\intercal,\, \hat{\lambda}_2^{-1/2}\boldsymbol{c}_2^\intercal,\, \ldots,\, \hat{\lambda}_m^{-1/2}\boldsymbol{c}_m^\intercal)^\intercal = n^{-1/2}\boldsymbol{\eta} + o(n^{-1/2}),$$

Consequently

$$\boldsymbol{g}(\boldsymbol{\beta}) = \bigtriangledown(\boldsymbol{\beta} - \boldsymbol{\beta}^*) + O(\|\boldsymbol{\beta} - \boldsymbol{\beta}^*\|^2) = n^{-1/2} \bigtriangledown \boldsymbol{\eta} + o(n^{-1/2}).$$

Without loss of generality, assume that the submatrix $\bigtriangledown_1$ of $\bigtriangledown = (\bigtriangledown_1,\, \bigtriangledown_2)$ is of full rank. Recall that, in this case, the Jacobian matrix, J, of $\mathcal{G}(\cdot)$ evaluated at $\boldsymbol{\gamma}^*$ is given by (3.3): $J = (-(\bigtriangledown_1^{-1}\bigtriangledown_2)^\intercal,\, I_{md-q})^\intercal$. We find that the column space of J is exactly the null space of $\bigtriangledown$. Hence, $\boldsymbol{\eta}$ is in the column space of J if and only if $\bigtriangledown\boldsymbol{\eta} = \mathbf{0}$, in which case, by the above expansion, we have $\boldsymbol{g}(\boldsymbol{\beta}) = o(n^{-1/2})$. Now, by Theorem 3.2, $\delta^2 = 0$ if and only if $\boldsymbol{\eta}$ is in the column space of J, so we conclude that $\delta^2 = 0$ if and only if $\boldsymbol{g}(\boldsymbol{\beta}) = o(n^{-1/2})$. On the other hand, if we know $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$ for all n like the case in the previous example, then we must have $\bigtriangledown\boldsymbol{\eta} = \mathbf{0}$ and so $\delta^2 = 0$.

The above analysis shows that $\delta^2 = 0$ if and only if $\boldsymbol{g}(\boldsymbol{\beta})$ converges to

$\mathbf{0}$ faster than the order of $O(n^{-1/2})$, which is the speed that $\boldsymbol{\beta}$ converges to $\boldsymbol{\beta}^*$. In this case, such a $\boldsymbol{\beta}$ defined by the local alternative model should be considered to be under the null model of $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$ in asymptotic sense.

3.3 Simulation studies

In this section, we conducted simulations to study: (1) the approximation accuracy of the limiting distributions to the finite–sample distributions of the DELR statistic under both the null and the alternative models, and (2) the power of the DELR test under correctly specified DRMs. The number of simulation runs in this and the next section (Section 3.4) is set to $10,000$. Our simulation is more extensive than what are presented in terms of hypothesis, population distribution, and sample sizes. We selected the most representative ones and included them here; but the other results are similar. All computations are carried out by our R package *drmdel* for EL inference under DRMs, which is introduced in Chapter 6 and available on The Comprehensive R Archive Network (CRAN).

3.3.1 Approximation to the distribution of the DELR

Null limiting distribution

We first study how well the chi–square distribution approximates the finite–sample distribution of the DELR statistic under the null hypothesis of (3.1). Set $m + 1 = 6$ and consider the hypothesis (3.1) with $\boldsymbol{g}(\boldsymbol{\beta}) = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_3^\intercal) - (\boldsymbol{\beta}_2^\intercal, \boldsymbol{\beta}_4^\intercal)$. The null hypothesis is equivalent to $F_1 = F_2$ and $F_3 = F_4$. We generated six samples of sizes (90, 60, 120, 80, 110, 30) from two distribution families. The first one is from normal distributions with means (0, 2, 2, 1, 1, 3.2) and standard deviations (1, 1.5, 1.5, 3, 3, 2). The second one is from gamma distributions with shapes (3, 4, 4, 5, 5, 3.2) and rates (0.5, 0.8, 0.8, 1.1, 1.1, 1.5).

When the basis function $\boldsymbol{q}(x)$ is correctly specified, i.e. $\boldsymbol{q}(x) = (x,\, x^2)^\intercal$ for the normal family and $\boldsymbol{q}(x) = (\log x,\, x)^\intercal$ for gamma family, the DELR statistic, R_n, has a χ^2_4 null limiting distribution in both cases. The quantile–quantile (Q–Q) plots of the distribution of R_n and χ^2_4 are shown in Figure 3.1. In both cases, the approximations are very accurate. The type I error rates of R_n at 5% level are 0.056 and 0.058 for normal and gamma data respectively.

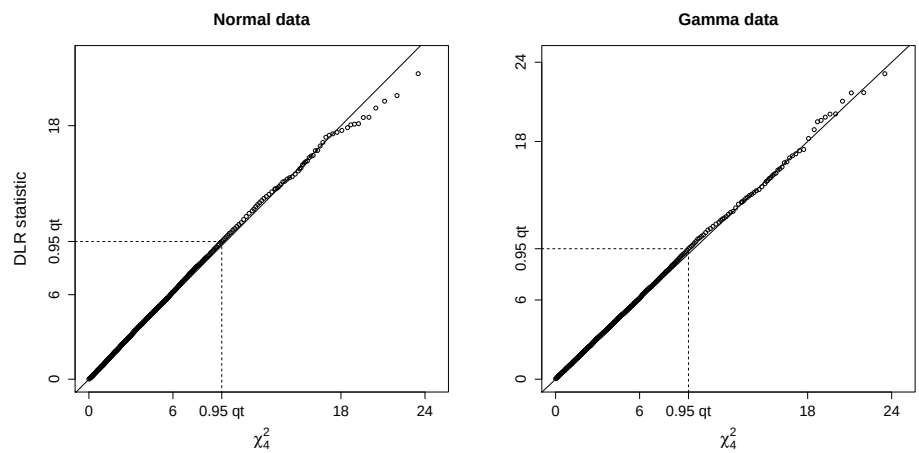

Figure 3.1: Q–Q plots of the simulated and the null limiting distributions of the DELR statistic.

In general, for data from distributions such as Weibull and log–normal, the chi–square approximation has satisfactory precision when $n_k \geq 70$.

Distribution under local alternatives

We next examine the precision of the non–central chi–square distribution under the local alternative model (3.2). We set $m + 1 = 4$ with sample sizes 120, 160, 80 and 60.

In the first scenario, we test the hypothesis (3.1) with $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta}_1^\intercal - \boldsymbol{\beta}_2^\intercal$. The perceived null model is specified by $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_2^* = (0.25,\, 1.875)^\intercal$, $\boldsymbol{\beta}_3^* =$

$(0.125,\ 1.97)^\intercal$ with basis function $\boldsymbol{q}(x) = (x, x^2)^\intercal$. The data were generated from $G_0 = N(0,\ 0.5^2)$, G_1 and G_3 with $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_3^*$ respectively, and G_2 with $\boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^* + n_2^{-1/2}(1,\ 0)^\intercal$. According to Theorem 3.2, the limiting distribution of R_n is $\chi_2^2(2.67)$.

In the second scenario, we test (3.1) with $\boldsymbol{g}(\boldsymbol{\beta}) = (\boldsymbol{\beta}_1^\intercal,\ \boldsymbol{\beta}_3^\intercal) - (\boldsymbol{\beta}_2^\intercal,\ (-6,\ 9)^\intercal)$. The perceived null model is specified by $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_2^* = (-4,\ 5)^\intercal$, $\boldsymbol{\beta}_3^* = (-6,\ 9)^\intercal$ with basis function $\boldsymbol{q}(x) = (\log x, x)^\intercal$. We generated data from $G_0 = \Gamma(3,\ 2)$ and G_k, $k = 1,\ 2,\ 3$, specified by (3.2) with $\boldsymbol{c}_1 = (0.5,\ 0.5)^\intercal$, $\boldsymbol{c}_2 = (1,\ 1)^\intercal$ and $\boldsymbol{c}_3 = (2,\ 2)^\intercal$. According to Theorem 3.2, the limiting distribution of R_n is $\chi_4^2(1.80)$.

The Q–Q plots under the two scenarios are shown in Figure 3.2. It is clear the non–central chi–square limiting distributions approximate these of of R_n very well. In unreported simulation studies under various settings, we find the approximate of the non–central chi–square is generally satisfactory when $n_k \approx 100$.

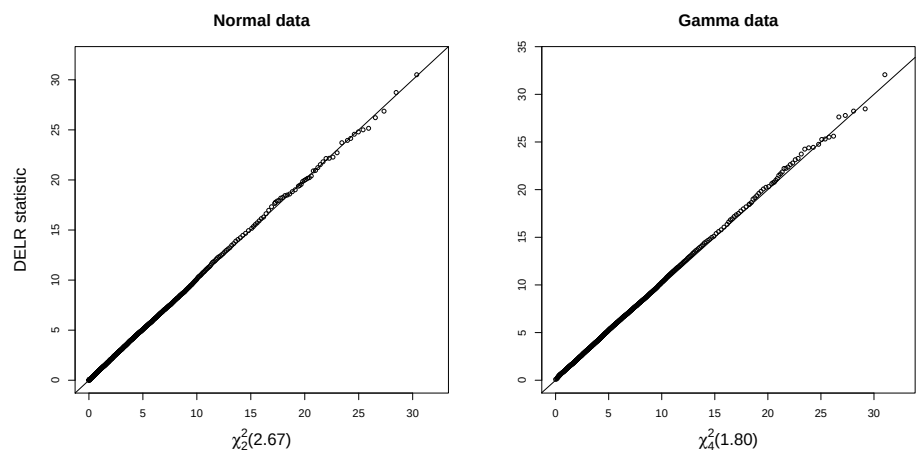

Figure 3.2: Q–Q plots of the distributions of the DELR statistics under the local alternative model against the corresponding asymptotic theoretical distributions.

3.3.2 Power comparison

We now compare the power of the DELR test (DELRT) with a number of popular methods for detecting differences between distribution functions, testing $H_0 : F_0 = F_1 = \ldots = F_m$. This is the same as (3.1) with $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta}$. We use the nominal level of 5%.

The competitors include the Wald test based on DRM (Wald) (Fokianos et al., 2001, (17)), analysis of variance (ANOVA), the Kruskal–Wallis rank–sum test (KW) (Wilcox, 1995), and the k–sample Anderson–Darling test (AD) (Scholz and Stephens, 1987).

The Wald test, as described in Section 3.1, is based on test statistic $n\hat{\boldsymbol{\beta}}^{\mathsf{T}}\hat{\Sigma}^{-1}\hat{\boldsymbol{\beta}}$ with $\hat{\Sigma}$ being a consistent estimator of the asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$. It uses a chi–square reference distribution. KW is a rank–based nonparametric test for equal population medians. AD is a nonparametric test based on the quadratic distances of empirical distribution functions for equal population distributions.

We first compare their powers based on normal data with $m + 1 = 2$ and sample sizes $n_0 = 30$ and $n_1 = 40$. In this case, the two–sample t–test is the most powerful unbiased test when the two populations have the same variance.

We consider two different scenarios for alternatives both having $F_0 = N(0, 2^2)$. In the first scenario, $F_1 = N(\mu, 2^2)$ with μ increasing in absolute value in a sequence of simulation experiments. In the second scenario, we consider seven parameter settings (settings 0–6) for $F_1 = N(\mu, \sigma^2)$ with μ and σ taking values in (0, 0.05, 0.1, 0.15, 0.25, 0.36, 0.55) and (2, 1.9, 1.8, 1.7, 1.62, 1.56, 1.50) respectively.

The power curves are shown in Figure 3.3. In the equal-variance scenario, DELR test is comparable to the optimal two–sample t–test. In the unequal variance scenario 2, the DELR test clearly has much higher power than its competitors, and its type I error is close to the nominal 0.05.

We next compare these tests on non–normal samples with $m + 1 = 5$ and

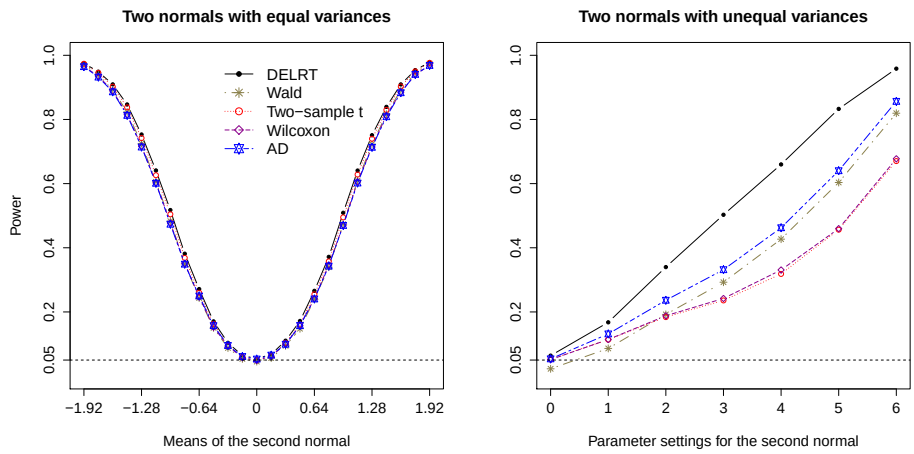

Figure 3.3: Power curves for normal data. The parameter setting 0 corresponds to the null model and the settings 1–6 correspond to alternative models.

sample sizes to be 30, 40, 25, 45 and 50. We generated data from four families of distributions: gamma, log–normal, Pareto with common support, and Weibull distributions with shape parameter equaling 0.8. The log–normal, Pareto and Weibull distributions satisfy DRMs with basis functions $\boldsymbol{q}(x) = (\log x, \log^2 x)^\mathsf{T}$, $\boldsymbol{q}(x) = \log x$, and $\boldsymbol{q}(x) = x^{0.8}$, respectively.

We used six DRM parameter settings (settings 0–5; shown in Table 3.10 in 3.7). Setting 0 satisfies the null hypothesis and settings 1–5 do not. The simulated rejection rates are shown in Figure 3.4. It is clear that the DELR test has the highest power while its type I error rates are close to the nominal.

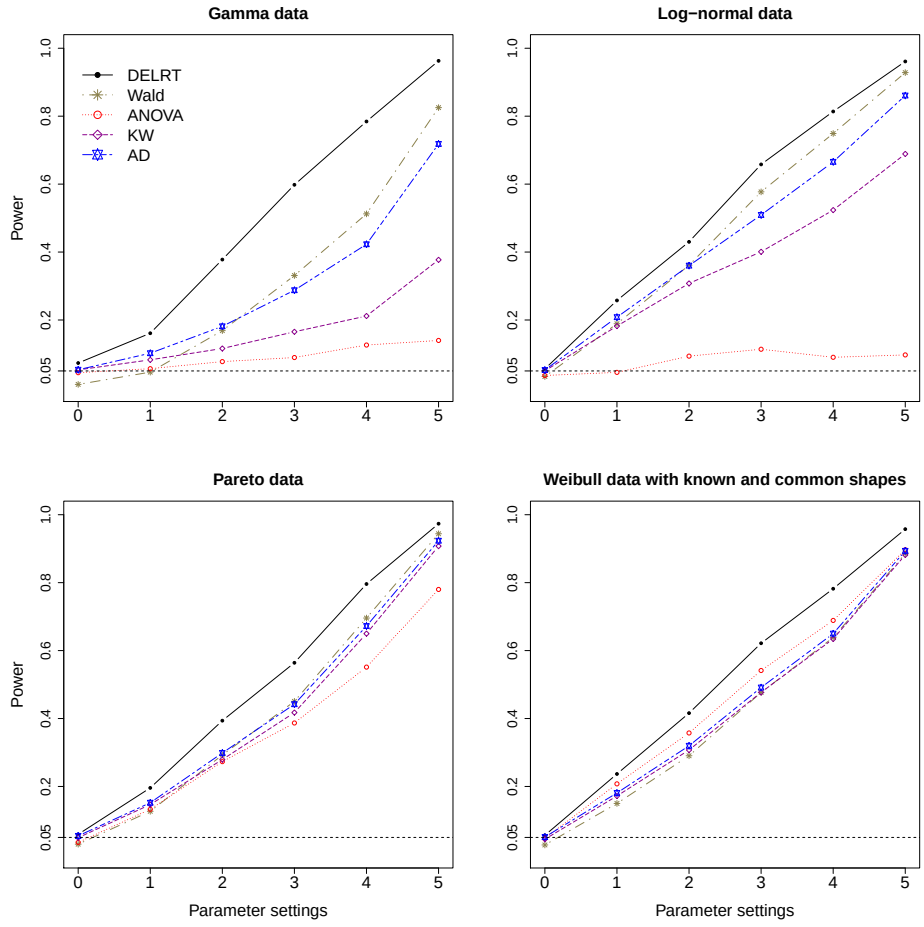

Figure 3.4: Power curves for non–normal data. The parameter setting 0 corresponds to the null model and the settings 1–5 correspond to alternative models.

3.4 Robustness of DELR test against model misspecification

The DRM is very flexible and includes a large number of distribution families as special cases. The risk of misspecification thus is considered low. Nevertheless, examining the effect of misspecification remains an important topic. Fokianos and Kaimi (2006) suggested that misspecifying the basis function $\boldsymbol{q}(x)$ has an adverse effect on estimating $\boldsymbol{\beta}$. Chen and Liu (2013) found that estimation of population quantiles is robust against misspecification. In this section, we use simulation studies to demonstrate that, even if the DRM is misspecified, the null distribution of the DELR statistic is still well approximated by the chi–square distribution for large sample sizes when a high dimensional basis function $\boldsymbol{q}(x)$ is utilized, and the DELR test remains to have a high power and reasonable type I error rate for testing the hypothesis of equal population distributions.

3.4.1 Null limiting distribution of the DELR statistic

We first study the chi–square approximation to the finite–sample distribution of the DELR statistic based on misspecified DRMs under the null hypothesis of (3.1). As for the simulation under the correctly specified DRM in Section 3.3.1, we set $m + 1 = 6$ and consider the hypothesis (3.1) with $\boldsymbol{g}(\boldsymbol{\beta}) = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_3^\intercal) - (\boldsymbol{\beta}_2^\intercal, \boldsymbol{\beta}_4^\intercal)$, which is equivalent to $F_1 = F_2$ and $F_3 = F_4$. We generated six samples of the same size under two different distributional settings respectively. The first setting consists of Weibull distributions with shapes (2.5, 1, 1, 2, 2, 1.8) and scales (1.2, 2.8, 2.8, 4, 4, .0.9). The second setting consists of distributions from different families:

$$X_0 \sim \text{Gamma}(3,\, 0.5), \quad X_1 \sim \text{log–normal}(0,\, 0.6), \quad X_2 \sim \text{log–normal}(0,\, 0.6),$$
$$X_3 \sim \text{Weibull}(2,\, 4), \quad X_4 \sim \text{Weibull}(2,\, 4), \quad X_5 \sim \text{Gamma}(2,\, 0.8).$$

The distributions under neither of the above settings satisfy a DRM, however we still fit a DRM to them under each setting. For the first setting, we fit a DRM with the basis function that is suitable for gamma family, $\boldsymbol{q}(x) = (x, \log x)^\intercal$, to the Weibull samples, because the shapes of Weibull densities are similar to those of gamma densities. For the second setting of mixed families of distributions, we fit DRMs with the following different basis functions:

DRM 1: $\boldsymbol{q}(x) = \log x$,

DRM 2: $\boldsymbol{q}(x) = (x, x^2)^\intercal$,

DRM 3: $\boldsymbol{q}(x) = (x, \log x)^\intercal$,

DRM 4: $\boldsymbol{q}(x) = (\log x, \sqrt{x}, x)^\intercal$,

DRM 5: $\boldsymbol{q}(x) = (\log x, x, x^2)^\intercal$,

DRM 6: $\boldsymbol{q}(x) = (\log x, \sqrt{x}, x, x^2)^\intercal$.

For each DRM, the theoretical limiting distribution of the DELR test for the null hypothesis of $\boldsymbol{g}(\boldsymbol{\beta}) = (\boldsymbol{\beta}_1^\intercal, \boldsymbol{\beta}_3^\intercal) - (\boldsymbol{\beta}_2^\intercal, \boldsymbol{\beta}_4^\intercal) = \mathbf{0}$ is χ^2_{2d}, where d is the dimension of the basis function.

Under the first setting, we calculate the DELR test statistic for different sample sizes: $n_k = 20, 40, 70, 100, 150, 300$ respectively for $k = 0, \ldots, 5$. The Q–Q plots of the DELR statistics against the quantiles of the theoretical limiting χ^2_4 distribution are shown in Figure 3.5 and the corresponding type–I error rates of the DELR tests at the nominal sizes of 0.10 and 0.05 are shown in Table 3.1. For small sample sizes, the Q–Q plots are always above the diagonal line, and the type–I errors are higher than the nominal sizes. As the sample size increases, the Q–Q plot slowly move towards the diagonal line. When the size of each sample reaches 150 and higher, the χ^2_4 distribution approximates the distribution of R_n well, and the type–I errors are close to

the nominal sizes. These tell us, although the DRM is misspecified, the chi–square approximation of the DERL statistic in this case is still useful for samples of large sizes.

Table 3.1: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for Weibull samples under a misspecified DRM.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$
0.10	0.1483	0.1322	0.128	0.123	0.1178	0.1102
0.05	0.0839	0.0728	0.0664	0.0664	0.0624	0.0543

For the second setting of mixed families of distributions, the Q–Q plots of the DELR statistics against the theoretical quantiles of the limiting distribution under DRM 1 – DRM 6 are shown in Figure 3.6 – 3.11. The corresponding type–I error rates at the nominal sizes of 0.10 and 0.05 are shown in Table 3.2 – 3.7.

For the DRM with the simplest basis function (DRM 1), the Q–Q line is always slightly under the diagonal line for all sample sizes, implying conservative DELR tests of all nominal sizes. For the two DRMs with two–dimensional basis functions (DRM 2 and 3), the Q–Q lines are always above the diagonal line for all sample sizes, which indicate anti–conservative DLR tests of all nominal sizes. For the two DRMs with three–dimensional basis functions (DRM 4 and 5), the Q–Q line is above the diagonal line when the sample size is small ($n_k = 20$, 40). It moves close to the diagonal line when the sample size becomes moderately large ($n_k = 70$, 100), and below the diagonal line when the sample size gets larger ($n_k = 150$, 300). If we increase the size of each sample to 500 or $1,000$, the Q–Q line stays slightly below and a little further away from the diagonal line. For all the above cases (DRM 1 – 5), the Q–Q plots show that the DLR test has some bias that does not diminish as the sample size becomes larger.

The DELR test under DRM 6, which has a four–dimensional basis function, has a different behaviour with respect to the sample size. The Q–Q line

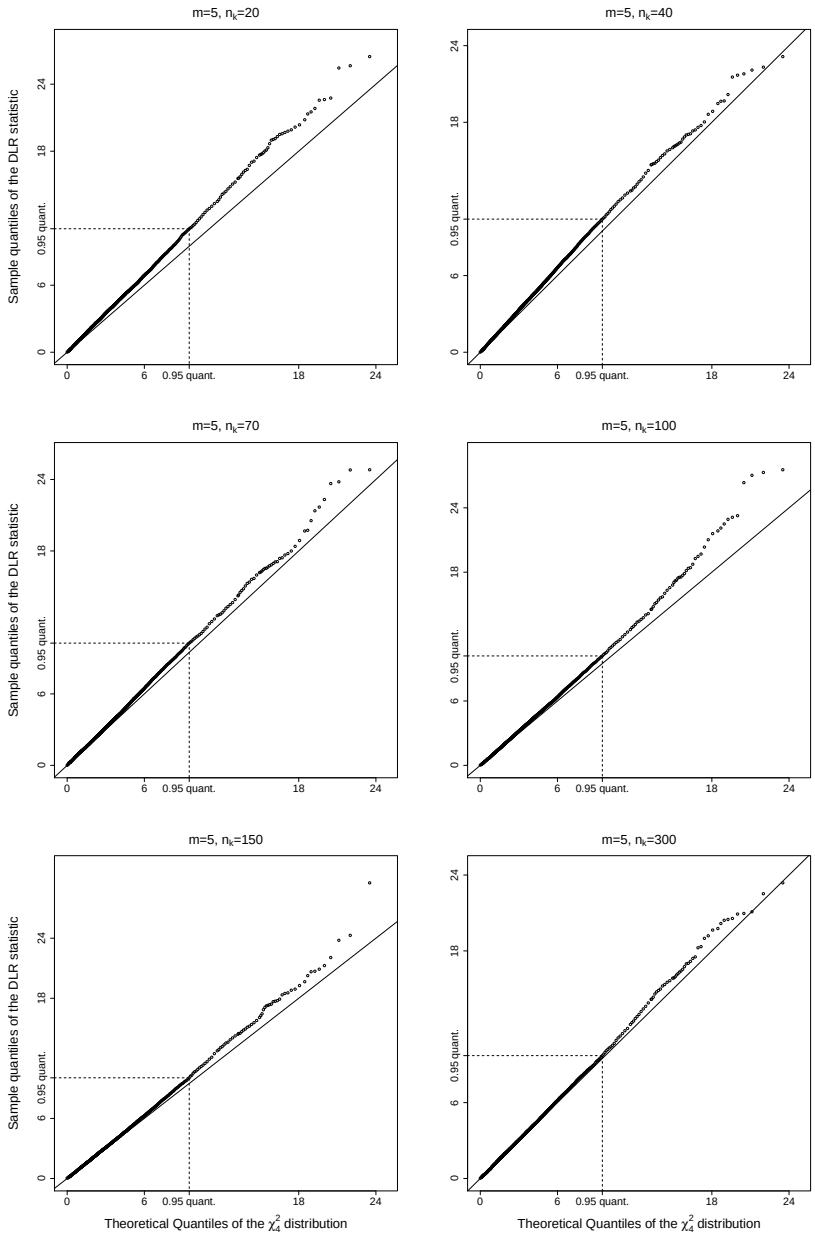

Figure 3.5: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for Weibull samples under a misspecified DRM.

is well above the diagonal line when the sample size is small. As the sample size increases, the Q–Q line moves gradually towards the diagonal line. When the size of each sample reaches 300, the Q–Q line is very close to the diagonal line. And if we further increase the size of each sample to 500 and $1,000$, the Q–Q line approaches even closer to the diagonal line from above, and it never goes below as in the case under DRMs with three–dimensional basis functions. This behaviour is similar to that of the DELR test under a correctly specified DRM, only the speed of convergence to the Chi–squared distribution is slower. It indicates that a DRM with a high dimensional basis function is more likely to fit the samples better.

Table 3.2: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for samples from different families of distributions under the misspecified DRM 1.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$
0.10	0.0867	0.0749	0.071	0.0739	0.07	0.0704
0.05	0.0419	0.0348	0.0311	0.0358	0.0321	0.0335

Table 3.3: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for samples from different families of distributions under the misspecified DRM 2.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$
0.10	0.143	0.1251	0.1254	0.1282	0.1201	0.1215
0.05	0.0759	0.0683	0.0644	0.0702	0.064	0.0649

Table 3.4: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for samples from different families of distributions under the misspecified DRM 3.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$
0.10	0.149	0.1419	0.1367	0.1408	0.1325	0.1364
0.05	0.0815	0.0794	0.0738	0.0778	0.0712	0.0742

Table 3.5: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for samples from different families of distributions under the misspecified DRM 4.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$
0.10	0.1554	0.1204	0.0987	0.0981	0.0877	0.0819
0.05	0.0848	0.0616	0.0472	0.0501	0.0391	0.0387

Table 3.6: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for samples from different families of distributions under the misspecified DRM 5.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$
0.10	0.1608	0.1198	0.1019	0.1006	0.0872	0.0817
0.05	0.0871	0.0642	0.0503	0.0517	0.0443	0.0416

Table 3.7: The type–I error rates of the DELR tests at nominal sizes of 0.10 and 0.05 for samples from different families of distributions under the misspecified DRM 6.

Nominal size	$n_k = 20$	$n_k = 40$	$n_k = 70$	$n_k = 100$	$n_k = 150$	$n_k = 300$	$n_k = 500$	$n_k = 1,000$
0.10	0.2228	0.1602	0.1336	0.1343	0.1201	0.1109	0.1132	0.1038
0.05	0.1343	0.0866	0.0708	0.0716	0.0628	0.0583	0.0578	0.0529

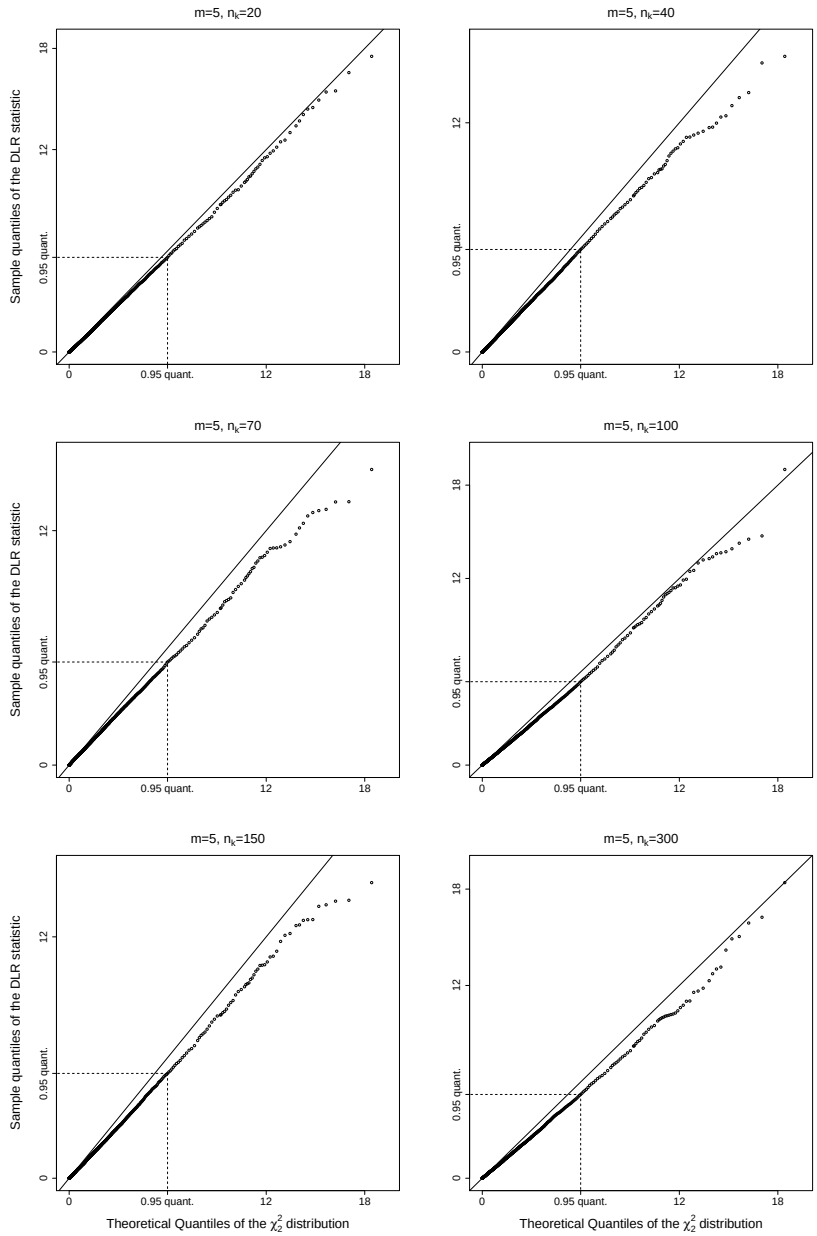

Figure 3.6: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for samples from different families of distributions under the misspecified DRM 1.

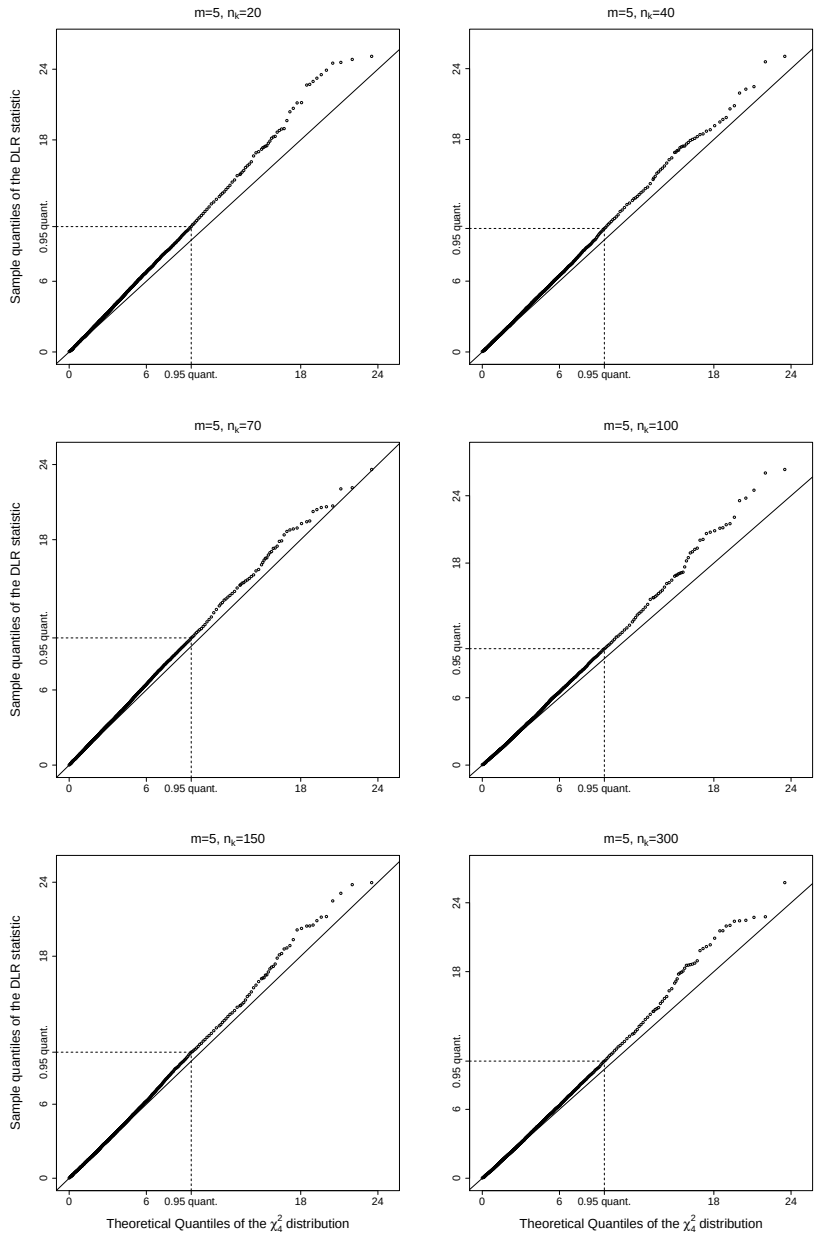

Figure 3.7: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for samples from different families of distributions under the misspecified DRM 2.

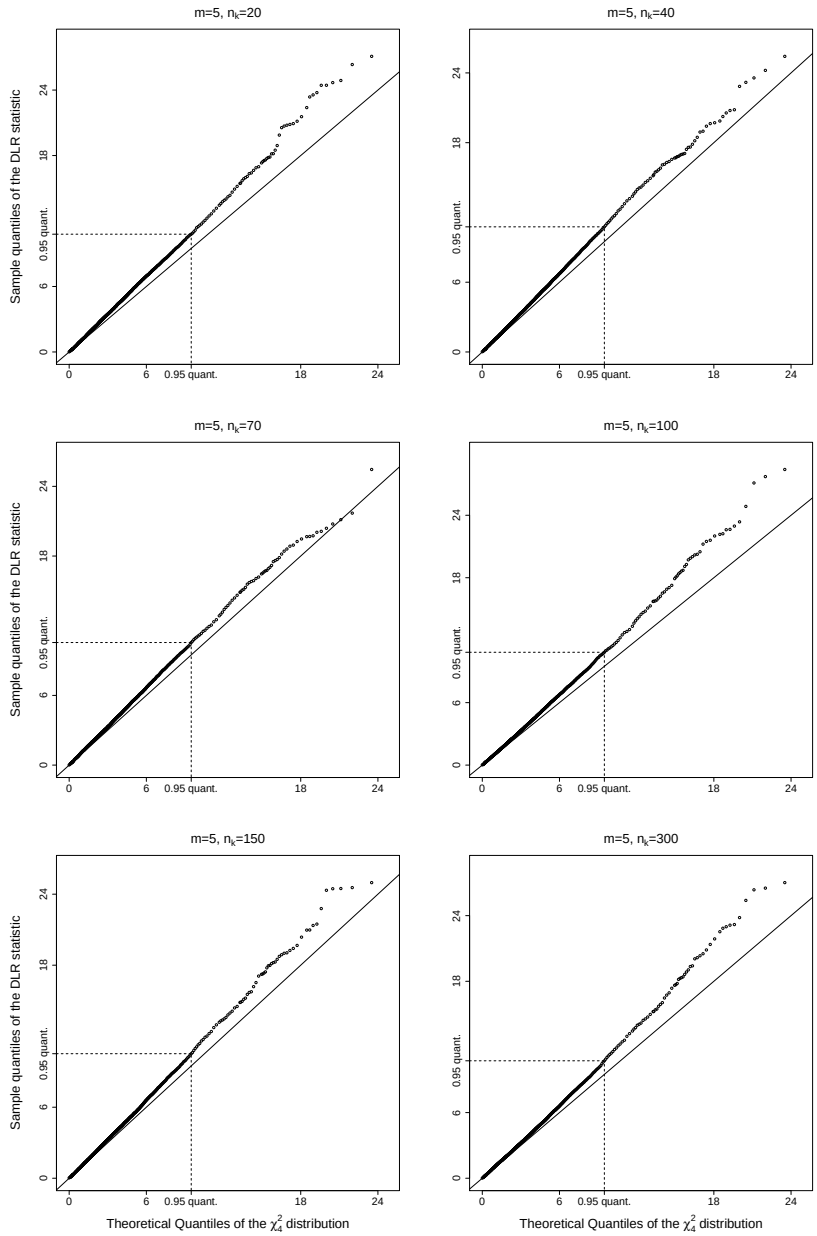

Figure 3.8: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for samples from different families of distributions under the misspecified DRM 3.

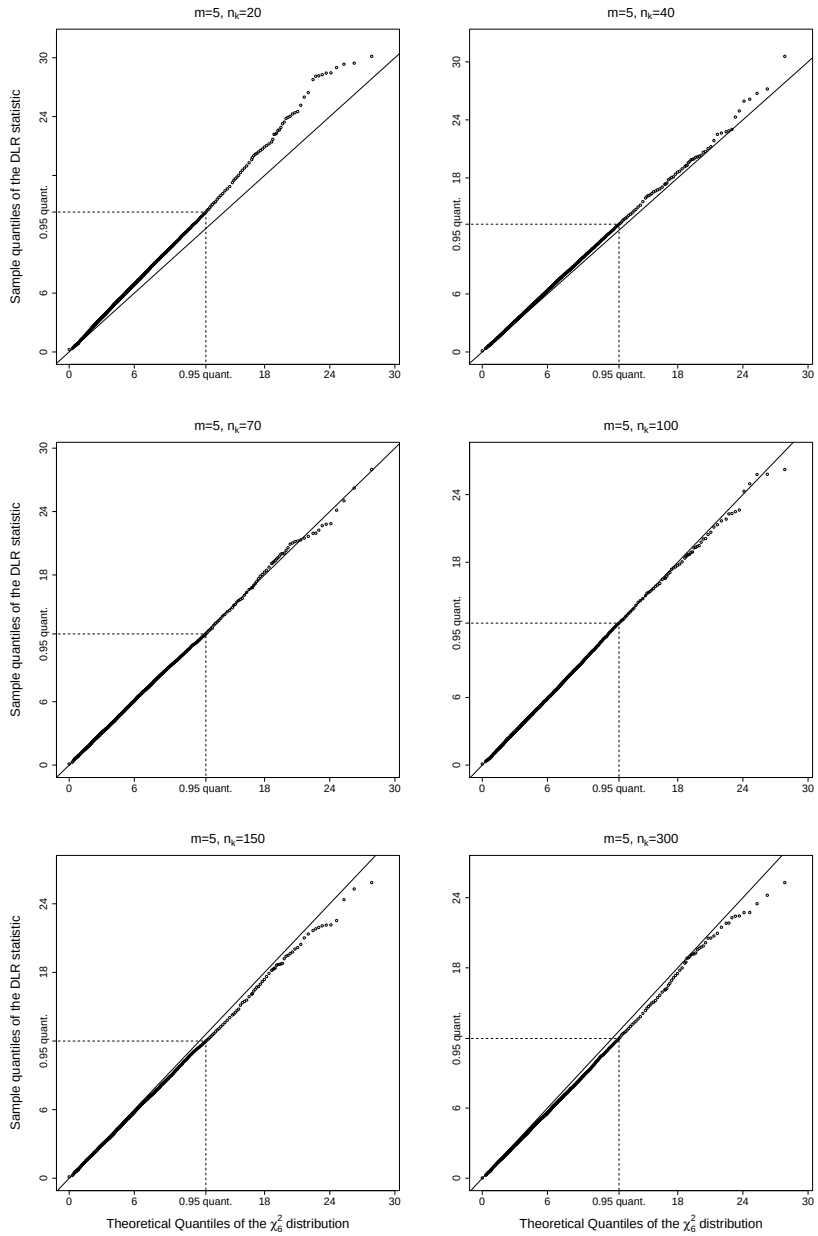

Figure 3.9: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for samples from different families of distributions under the misspecified DRM 4.

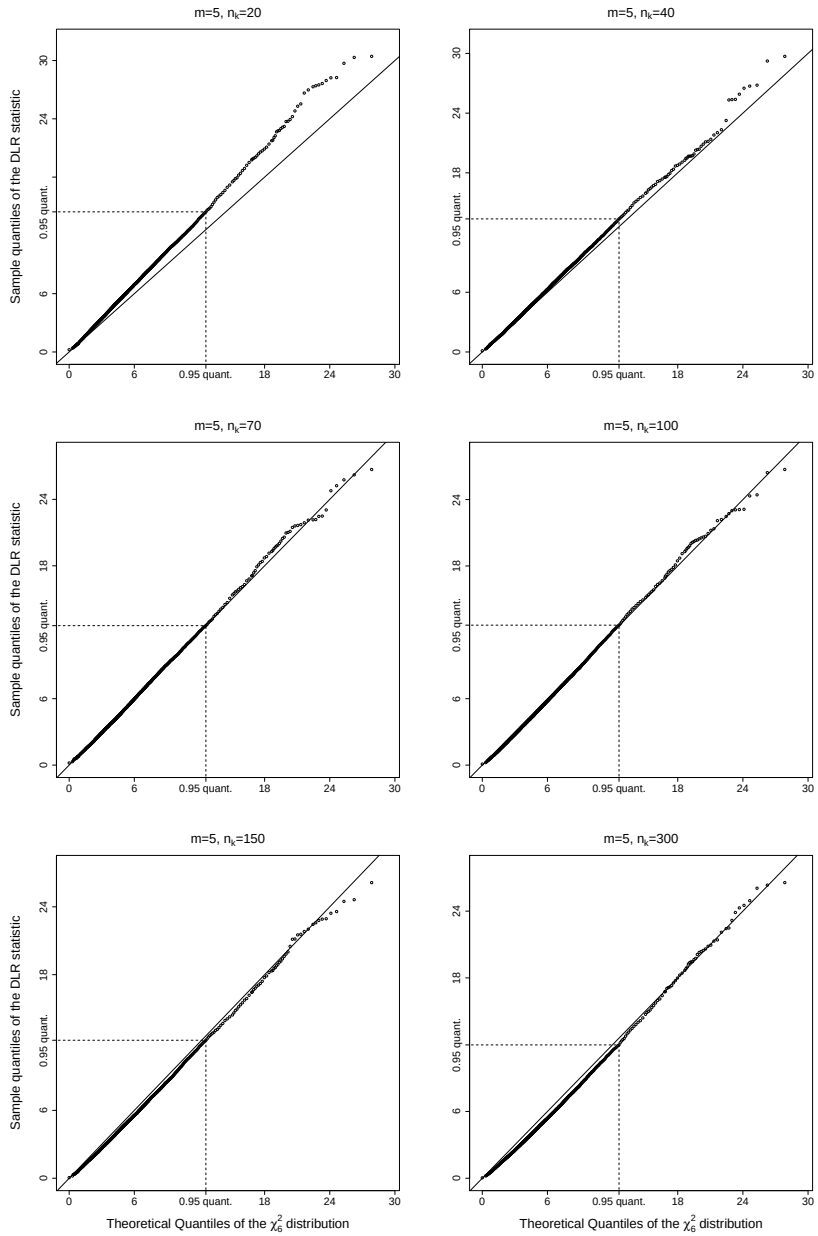

Figure 3.10: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for samples from different families of distributions under the misspecified DRM 5.

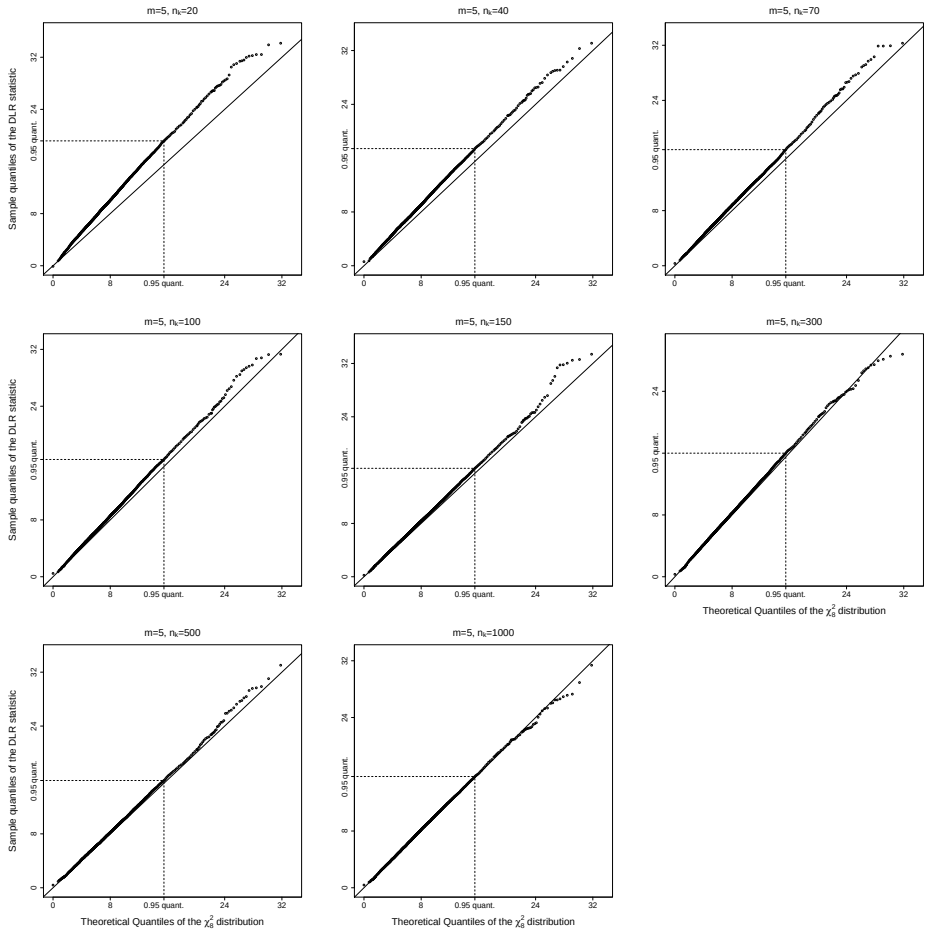

Figure 3.11: Q–Q plots of the simulated and the null limiting distribution of the DELR statistics for samples from different families of distributions under the misspecified DRM 6.

3.4.2 Power of the DELR test

We now study the power of the DELR test under misspecification of the DRM. As in Section 3.3.2, we consider the null hypothesis of $H_0 : F_0 = F_1 = \ldots = F_m$, which is the same as (3.1) with $\boldsymbol{g}(\boldsymbol{\beta}) = \boldsymbol{\beta}$. We use the nominal level of 5%.

We put $m + 1 = 5$ with sample sizes 90, 120, 75, 135 and 150. In the first simulation experiment, we generated data from Weibull distributions. We use DELR test and Wald test to test the same distribution hypothesis based on DRM assumption with $\boldsymbol{q}(x) = (x, \log x)^\intercal$. We used six parameter settings with the 0th set satisfying the null hypothesis. Note that the basis function is misspecified.

We repeated the experiment with data from mixtures of two normals, non–central t distributions, and and mixtures of a gamma and a Weibull. We conducted DELR test based on DRM assumption with $\boldsymbol{q}(x) = (x, x^2)^\intercal$ for non–central t data, for the normal mixture data, and $\boldsymbol{q}(x) = (x, \log x)^\intercal$ for data from gamma–Weibull mixture. All these DRMs are misspecified. The detailed parameter settings 0–5 are given in Table 3.11 in 3.7.

We also applied ANOVA, KW and AD tests. The results are summarized as power curves in Figure 3.12. Although all the DRMs are misspecified, we notice that, under all these data settings, the DELR test has close to nominal type I error rates, and superior power in detecting distributional differences.

3.5 Analysis of lumber quality data

We now turn to the primary application of this thesis: analysis of the lumber strength. As members of the Forest Products Stochastic Modeling Group centered at the University of British Columbia, we are helping develop methods for assessing the engineering strength properties of lumber. A primary goal, one noted in Chapter 1, is an effective but relatively inexpensive long term monitoring program for the strength of lumber. Two strengths of

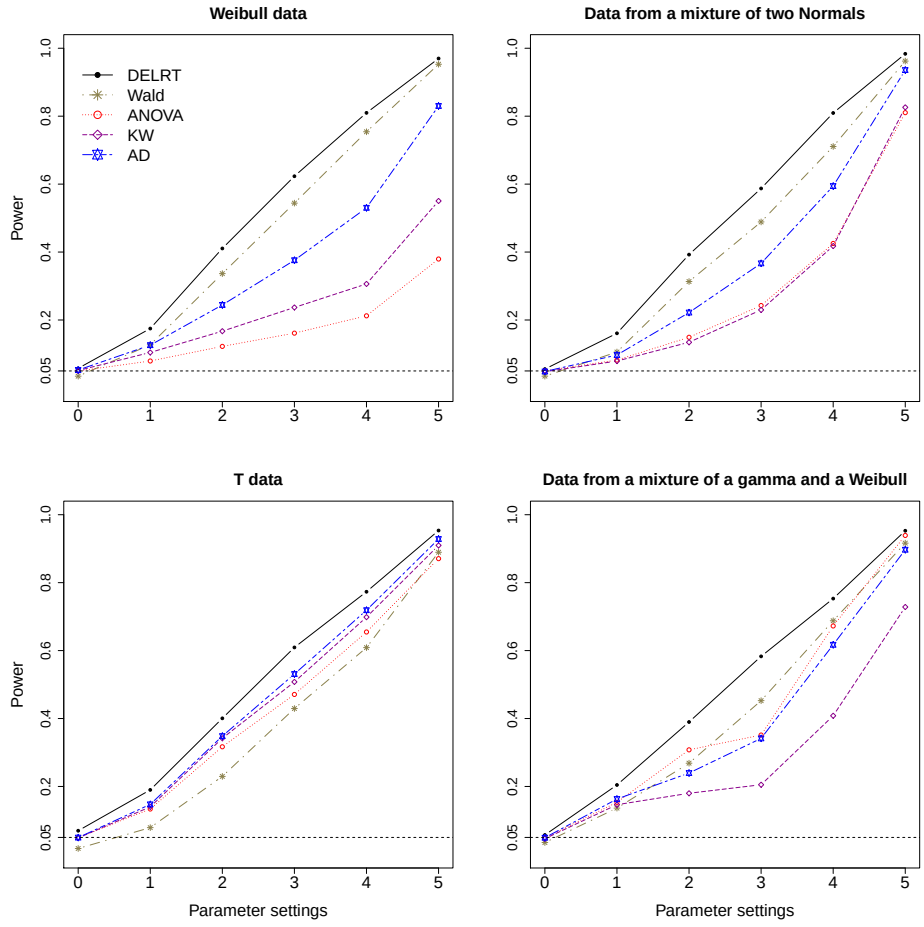

Figure 3.12: Power curves of the five tests with DELR and Wald tests based on misspecified DRMs. The parameter setting 0 corresponds to the null model and the settings 1–5 correspond to alternative models.

great importance are the so–called modulus of rupture (MOR) or "bending strength", and modulus of tension (MOT) or "tension strength", both of which are measured in units of 10^3 pound–force per square inch (psi). The Forest Products Stochastic Modeling Group collected three MOR samples in year 2007, 2010 and 2011 with sample sizes 98, 282 and 445 respectively, and two MOT samples in year 2007 and 2011 with sample sizes 98 and 425 respectively. Our interest in change of lumber quality over time, leads us to test the hypothesis that the MOR and MOT samples respectively come from the populations of the same distributions. We do so using the DELR test, Wald test, ANOVA and Kruskal–Wallis rank–sum test.

3.5.1 Assessing the DRM fit: an exploratory approach

A good DRM fit to the data is crucial to the effectiveness of the DELR test. We now assess this by comparing the EL density estimates obtained using the fitted DRM to the histograms of the observed samples.

The kernel density plots of the MOR and MOT samples are shown in Figure 3.13. The three densities with modes around $6,100$ psi correspond to the MOR samples, while the two with modes around $3,600$ psi correspond to the MOT samples. The MOR samples of the year 2007 (MOR07) and the year 2010 (MOR10) seem to have similar density plots: both are slightly right–skewed, although MOR10 has a small lump to the right of its mode. Both seem well approximated by gamma distributions. The MOR sample of the year 2011 (MOR11) seems to have a quite different characteristic from MOR07 and MOR10: it has a clearly larger spread and looks more symmetric, which mimics the shape of the density of a normal distribution. The density plots of the MOT samples of the year 2007 (MOT07) and the year 2011 (MOT11) look similar: both are roughly bell shape with a little right–skewed, looking like something between a normal and a gamma distribution.

Since the above density plots look like either gamma or normal densi-

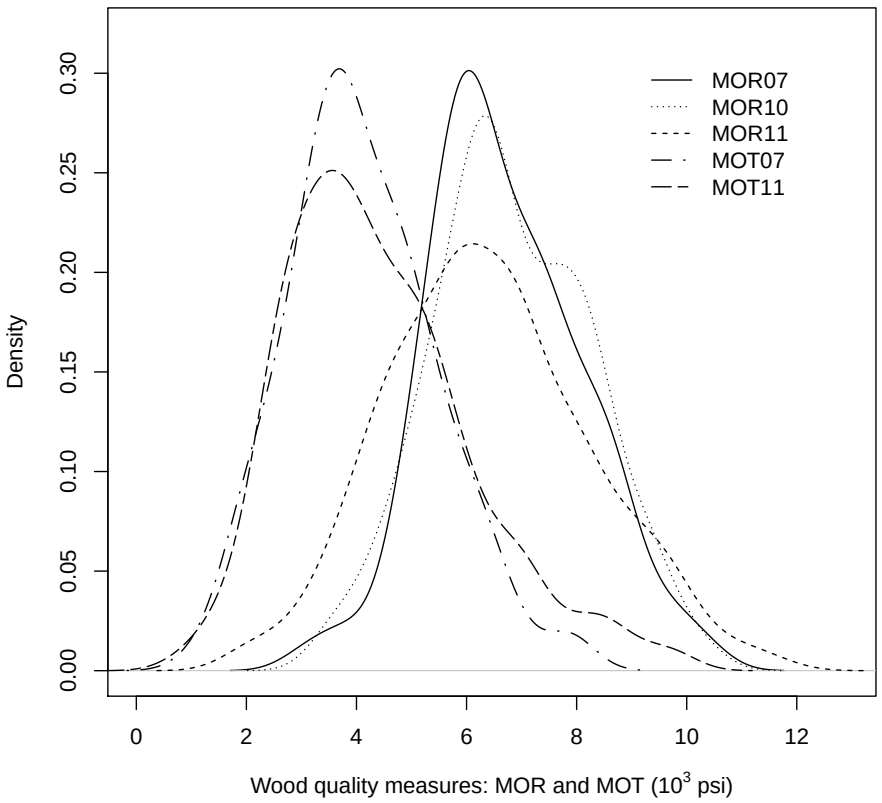

Figure 3.13: Kernel density plots of the MOR and MOT samples.

ties, when fitting a DRM to MOR or MOT samples, we use a basis function $\boldsymbol{q}(x) = (\log x,\, x,\, x^2)^\intercal$ which combines the basis function for the gamma distributions, $\boldsymbol{q}(x) = (\log x,\, x)^\intercal$, and that for the normal distributions, $\boldsymbol{q}(x) = (x,\, x^2)^\intercal$. The corresponding DRM is understood as a generalization of both gamma and normal distributions. Below we assess how this DRM fits to the MOR and MOT samples respectively. The DRM fit can be assessed using goodness–of–fit tests like those developed by Qin and Zhang (1997) and Zhang (2002). Here we use a less formal but visually straightforward

exploratory approach: checking whether the EL kernel density estimators of the different populations based on DRM agree with the histograms of the observed samples.

Recall that under the DRM the EL estimator $\hat{p}_{kj}$ for $dF_0(x_{kj})$ of the baseline distribution is given in (2.9). Applying the DRM assumption (2.1), we obtain the EL estimators for $dF_l(x_{kj})$, $l = 1, \ldots, m$, as

$$\hat{p}_{kj}^{(l)} = \exp\left\{\hat{\alpha}_l + \hat{\boldsymbol{\beta}}_l^{\mathsf{T}} \boldsymbol{q}(x_{kj})\right\} \hat{p}_{kj}.$$

An *EL kernel density estimator* of the population l is defined to be the kernel density estimator with the $\{\hat{p}_{kj}^{(l)}\}$ as weights and all the $\{x_{kj}\}$ as observations. Fokianos (2004) showed that under mild conditions this EL kernel density estimator is consistent and more efficient than the classical kernel density estimator with empirical weight $1/n_k$ for every data point within sample k.

The DRM fit to the MOR samples

We fit the DRM with basis function $\boldsymbol{q}(x) = (\log x,\, x,\, x^2)^{\mathsf{T}}$ to the three MOR samples, and compute the EL kernel density estimates of the MOR populations. We compare the EL kernel density estimates based on the DRM to the histograms, classical kernel density estimates with empirical weights, and parametric density estimates based on a three parameter Weibull distribution with density function

$$f(x;\, \lambda,\, \kappa,\, c) = \left\{\kappa(x-c)^{\kappa-1}/\lambda^{\kappa}\right\} \cdot \exp\left\{-\{(x-c)/\lambda\}^{\kappa-1}\right\},$$

for $x > c$ and $\lambda,\, \kappa > 0$. Such a Weibull distribution seems to be a flexible model for distributions with slightly heavy tails on the right. We calculate the maximum likelihood estimates (MLE) of the Weibull parameters based on each MOR sample, then estimate the density of the corresponding population by plugging in the MLEs to the above density function. Note that when the shape parameter $\kappa < 1$ and the location parameter c tends to the smallest

observation from below, the likelihood tends to infinity. Thus the MLEs of the parameters for this Weibull distribution are not well–defined. In practice, when calculating the MLE, we constrained the range of location parameter c to be smaller than the smallest observation minus a small positive constant, 10^{-6}.

The histograms and density estimates of the MOR samples are shown in Figure 3.14. For all samples, the EL kernel density estimates agree with the histograms well. Compared to the classical kernel density estimates with empirical weights, the EL kernel density estimates based on the DRM are more smooth, especially for MOR10. The mode of the EL kernel density estimate for MOR07 is slightly to the right of the mode of the corresponding classical kernel density estimate. Such differences exist because the EL kernel density estimators are obtained using combined data from all samples while the classical kernel density estimators are obtained using single samples. Compared to the Weibull density estimator, the EL kernel density estimator looks more flexible: it captures the small lumps on the right of the modes of MOR10 and MOR11, while the Weibull estimates do not. Overall, the DRM seems to fit the three MOR samples well.

The EL kernel density estimates is also found to be quite robust to the choice of the DRM basis function. We fit DRMs with other basis functions, including that for the normal family, that for the gamma family, $\boldsymbol{q}(x) = (\log x,\ \sqrt{x},\ x)^\intercal$, $\boldsymbol{q}(x) = (\log x,\ \sqrt{x},\ x,\ x^2)^\intercal$, etc. The corresponding EL kernel density estimates are very similar to the one we get using the current basis function $\boldsymbol{q}(x) = (\log x,\ x,\ x^2)^\intercal$.

The DRM fit to the MOT samples

As for the MOR samples, we fit the DRM with basis function $\boldsymbol{q}(x) = (\log x,\ x,\ x^2)^\intercal$ to the two samples of MOT for the year 2007 and 2010, and calculate the EL kernel density estimates based on the DRM for the corresponding populations. Again, we compare these density estimates to the

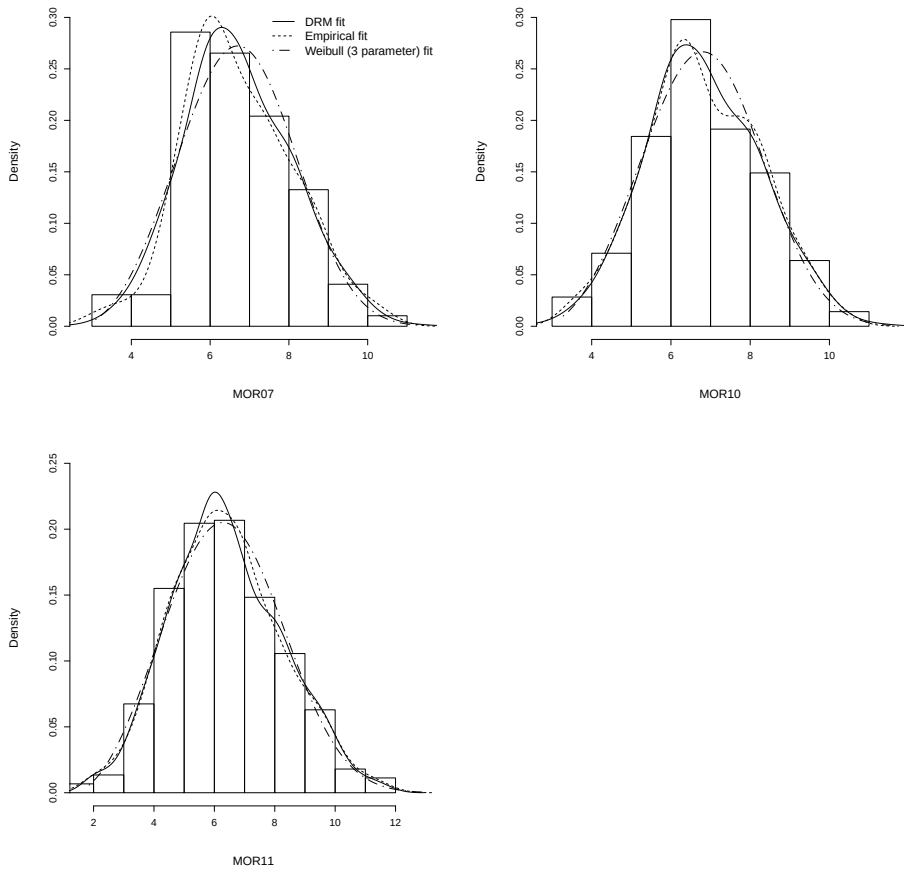

Figure 3.14: The histograms, EL kernel density estimates (solid curves), classical kernel density estimates (dashed curves) and three parameter Weibull density estimates (dot–dashed curves) for MOR samples.

histograms, the classical kernel density estimates with empirical weights and the three parameter Weibull density estimates (Figure 3.15). The EL kernel density estimates agree with the histograms well and are quite close to the classical kernel density estimates. Both the EL and classical kernel density estimators are more flexible and show more curvature than the Weibull density estimator.

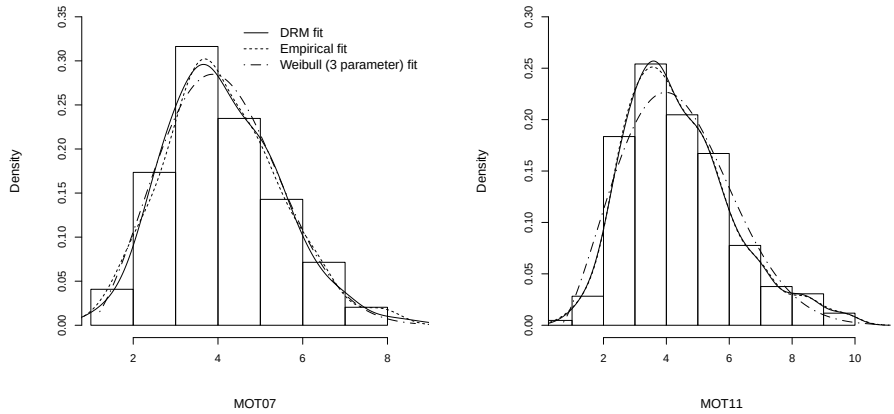

Figure 3.15: The histograms, EL kernel density estimates (solid curves), classical kernel density estimates (dashed curves) and three parameter Weibull density estimates (dot–dashed curves) for MOT samples.

3.5.2 Testing for equality of strength populations

Comparing the MOR populations

We now test the hypothesis that all MOR samples are from populations of the same distributions. As mentioned, we use the DELR test, Wald test, ANOVA and KW test for this hypothesis. The first two tests were carried under the DRM assumption with basis function $\boldsymbol{q}(x) = (\log x,\, x,\, x^2)^\intercal$. The DELR test and Wald test based on other basis functions lead to the same conclusion below, although the p–values are not identical.

The p–values obtained using the DELR test, Wald test, ANOVA and KW test are respectively 3.05e-8, 2.04e-6, 0.0029 and 0.00108. All tests strongly reject the null hypothesis of equal MOR distributions. The DRM–based tests, especially the DELR test, have much smaller p–values.

Following the rejection of that hypothesis it is natural to look for its cause through pairwise comparisons. For comparing pairs, we use the two sample t–test adjusted for unequal variances (also known as Welch's t–test)

and the Wilcoxon rank–sum test with continuity correction, instead of the plain ANOVA and KW test. We note that the plain ANOVA and KW test give slightly different p–values but same the conclusions. Let F_{MOR07}, F_{MOR10} and F_{MOR11} denote the MOR population distributions for years 2007, 2010 and 2011, respectively. The p–values for pairwise comparisons are given in Table 3.8. Note that the two DRM–based tests strongly suggest F_{MOR11} is markedly different from F_{MOR07} and F_{MOR10}, while F_{MOR07} and F_{MOR10} are not significantly different. The other two tests arrive at the same conclusion, but without statistical significance at 5% level. We also remark that the conclusion does not change at the 5% level when a Bonferroni correction is applied to account for the multiple comparison.

In addition, if the 5% size is strictly observed, t–test and KW test would imply $F_{\text{MOR07}} = F_{\text{MOR10}}$ and $F_{\text{MOR07}} = F_{\text{MOR11}}$, but $F_{\text{MOR10}} \neq F_{\text{MOR11}}$. This is much harder to interpret in applications.

Table 3.8: The p–values of pairwise comparisons among three MOR populations.

	DELRT	Wald	t–test	Wilcoxon
H_0: $F_{\text{MOR07}} = F_{\text{MOR10}}$	0.871	0.875	0.516	0.431
H_0: $F_{\text{MOR07}} = F_{\text{MOR11}}$	5.40e-4	7.01e-3	0.0579	0.0604
H_0: $F_{\text{MOR10}} = F_{\text{MOR11}}$	4.54e-8	1.82e-6	6.09e-4	3.95e-4

The three–sample DRM fit versus two–sample DRM fit. In the above pairwise comparisons, the DELR and Wald tests are based on a DRM fitted to all three MOR samples. We can also fit the DRM to the two samples in comparison only. How do the tests based on the two–sample DRMs compared to those based on the three–sample DRM. The p–values of such two–sample tests for pairwise comparisons are shown in Table 3.9. The p–values are very close to those based on the DRM for all the three MOR

populations, and the corresponding conclusions are the same. This observation agrees with the conclusion of Theorem 4.3 that we will present in the next chapter: when comparing the equality of a set of distribution functions, incorporating more samples, whose population distributions are unrelated to the hypothesis, into the DRM does not change the local asymptotic power of the DELR test.

Table 3.9: The p–values of the DELR and Wald tests based on two–sample DRMs for pairwise comparisons among the three MOR populations.

	DELRT	Wald
H_0: $F_{\mathrm{MOR07}} = F_{\mathrm{MOR10}}$	0.856	0.859
H_0: $F_{\mathrm{MOR07}} = F_{\mathrm{MOR11}}$	6.51e-04	7.8e-03
H_0: $F_{\mathrm{MOR10}} = F_{\mathrm{MOR11}}$	4.75e-08	2.04e-06

Comparing the MOT populations

We now test the hypothesis that the distribution functions of the two MOT populations are equal. The p–values of the DELR test, Wald test, Welch's t–test and Wilcoxon rank–sum test for this hypothesis are 0.0877, 0.228, 0.0749 and 0.295, respectively. No test rejects the null hypothesis at the 5% significance level. The DELR test and the t–test provide marginal evidence against the null hypothesis, showing that they are picking up the vague difference between the two samples. These results agree with our observation from the kernel density plots of the MOT samples (Figure 3.13).

3.6 Proofs

3.6.1 Theorem 3.1: Null limiting distribution of the DELR statistic

In this subsection, we show that the DELR test statistic, R_n, has a simple chi–square limiting distribution under the null hypothesis of (3.1). The idea is to show that the R_n is well approximated by a quadratic form that is asymptotically chi–square. We first give two important lemmas.

Lemma 3.3 (Block matrix inversion formula)**.** *Let M be a $(s+t) \times (s+t)$ nonsingular matrix with partition*

$$M = \begin{pmatrix} \underset{s\times s}{A} & \underset{s\times t}{B} \\ \underset{t\times s}{C} & \underset{s\times t}{D} \end{pmatrix}.$$

If A is nonsingular, then so is $S_A = D - CA^{-1}B$ and

$$M^{-1} = \begin{pmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{pmatrix}.$$

This is the conclusion of Theorem 8.5.11 of Harville (2008). We sketch the proof here.

Proof of Lemma 3.3. We first show that if A is nonsingular, then $S_A = D - CA^{-1}B$ is also nonsingular. Observe the following equality,

$$\begin{pmatrix} I_s & 0 \\ -CA^{-1} & I_t \end{pmatrix} M = \begin{pmatrix} I_s & 0 \\ -CA^{-1} & I_t \end{pmatrix} \begin{pmatrix} A & B \\ C & D \end{pmatrix} = \begin{pmatrix} A & B \\ 0 & S_A \end{pmatrix}.$$

The first matrix on the LHS is lower block–triangular, square, and has a full rank, so the rank of the second matrix on the LHS, M, is the same as the rank of the matrix on the RHS. Note that the matrix on the RHS is an

upper block–triangular matrix, so its rank equals the sum of the ranks of its diagonal blocks A and S_A. Therefore,

$$\operatorname{rank}(M) = \operatorname{rank}\begin{pmatrix} A & B \\ 0 & S_A \end{pmatrix} = \operatorname{rank}(A) + \operatorname{rank}(S_A).$$

Since M and A are both of full rank, S_A also has a full rank by the above equality. Hence S_A is nonsingular.

Since the above RHS matrix is nonsingular, we can solve for M^{-1} as

$$\begin{aligned} M^{-1} &= \begin{pmatrix} A & B \\ 0 & S_A \end{pmatrix}^{-1} \begin{pmatrix} I_m & 0 \\ -CA^{-1} & I_n \end{pmatrix} \\ &= \begin{pmatrix} A^{-1} & -A^{-1}CS_A^{-1} \\ 0 & S_A^{-1} \end{pmatrix} \begin{pmatrix} I_m & 0 \\ -CS_A^{-1} & I_n \end{pmatrix} \\ &= \begin{pmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{pmatrix}. \end{aligned}$$

□

Lemma 3.4 (Quadratic form decomposition formula)**.** *Adopt the settings of Lemma 3.3. Let* $\boldsymbol{z}^\intercal = (\boldsymbol{z}_1^\intercal, \boldsymbol{z}_2^\intercal)$ *be a vector of length* $s + t$*, partitioned in agreement with* s *and* t*. If* A *is nonsingular, then*

$$\boldsymbol{z}^\intercal M^{-1}\boldsymbol{z} = \big(\boldsymbol{z}_2 - B^\intercal (A^{-1})^\intercal \boldsymbol{z}_1\big)^\intercal \big(D - CA^{-1}B\big)^{-1}\big(\boldsymbol{z}_2 - CA^{-1}\boldsymbol{z}_1\big) + \boldsymbol{z}_1^\intercal A^{-1}\boldsymbol{z}_1.$$

Lemma 3.4 is an easy consequence of Lemma 3.3. Its proof thus is omitted.

Proof of Theorem 3.1. We now prove the asymptotic chi–squareness of the DELR statistic $R_n = 2\{l_n(\hat{\boldsymbol{\theta}}) - l_n(\tilde{\boldsymbol{\theta}})\}$ under the null model of (3.1). We first work on quadratic expansions of $l_n(\hat{\boldsymbol{\theta}})$ and $l_n(\tilde{\boldsymbol{\theta}})$ under the null model. The difference of the two quadratic forms is then shown to have a chi–square limiting distribution.

Recall $\boldsymbol{v} = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$ is the true DRM parameter. By expanding $l_n(\hat{\boldsymbol{\theta}})$ at $\boldsymbol{\theta}^*$, we get

$$l_n(\hat{\boldsymbol{\theta}}) = l_n(\boldsymbol{\theta}^*) + \sqrt{n}\boldsymbol{v}^\intercal(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) - (1/2)n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)^\intercal U_n(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) + \epsilon_n$$

where $\epsilon_n = O_p(n^{-1/2})$ since $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* = O_p(n^{-1/2})$ and the third derivative is bounded by an integrable function shown in (2.17). Also, in the proof of Theorem 2.5, we have shown in (2.25) that

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = U^{-1}\boldsymbol{v} + o_p(1).$$

Plugging the above expression of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ into the expansion of $l_n(\hat{\boldsymbol{\theta}})$, we get

$$l_n(\hat{\boldsymbol{\theta}}) = l_n(\boldsymbol{\theta}^*) + (1/2)\boldsymbol{v}^\intercal U^{-1}\boldsymbol{v} + o_p(1). \tag{3.7}$$

Next, we work on an expansion for $l_n(\tilde{\boldsymbol{\theta}})$ under the null model $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$. Recall Section 3.2.2 that, when $q < md$, the null model can be equivalently represented as $\boldsymbol{\beta} = \mathcal{G}(\boldsymbol{\gamma})$ for some function $\mathcal{G}$: $\mathbb{R}^{md-q} \to \mathbb{R}^{md}$ and parameter $\boldsymbol{\gamma}$ of dimension $md - q$. In addition, $\mathcal{G}$ is thrice differentiable in a neighbourhood of $\boldsymbol{\gamma}^*$, and its Jacobian matrix $J = \partial\mathcal{G}(\boldsymbol{\gamma}^*)/\partial\boldsymbol{\gamma}$ is of full rank. With this representation, the DRM parameter under the null hypothesis is $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \mathcal{G}(\boldsymbol{\gamma}))$. Hence, we may write the likelihood function under null model as

$$\ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma}) = l_n(\boldsymbol{\alpha}, \mathcal{G}(\boldsymbol{\gamma})).$$

Let $(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}})$ be the maximal point of $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma})$. Note that $\tilde{\boldsymbol{\theta}} = (\tilde{\boldsymbol{\alpha}}, \mathcal{G}(\tilde{\boldsymbol{\gamma}}))$ and so $l_n(\tilde{\boldsymbol{\theta}}) = \ell_n(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}})$. Clearly, $\ell_n(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ has the same properties as $l_n(\boldsymbol{\theta})$ and $\ell_n(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}})$ has a similar expansion as (3.7):

$$l_n(\tilde{\boldsymbol{\theta}}) = \ell_n(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}}) = \ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*) + (1/2)\tilde{\boldsymbol{v}}^\intercal \tilde{U}^{-1}\tilde{\boldsymbol{v}} + o_p(1),$$

where $\tilde{\boldsymbol{v}} = n^{-1/2}\partial l_n(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*)/\partial(\boldsymbol{\alpha}, \boldsymbol{\gamma})$ and $\tilde{U}$ is the corresponding information matrix. Partition $\boldsymbol{v}$ into $\boldsymbol{v}_1 = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\alpha}$ and $\boldsymbol{v}_2 = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\beta}$. Note that

$$n^{-1/2}\partial\ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*)/\partial\boldsymbol{\alpha} = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\alpha} = \boldsymbol{v}_1.$$

By the chain rule,

$$n^{-1/2}\partial\ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*)/\partial\boldsymbol{\gamma} = n^{-1/2}J^\intercal\{\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\beta})\} = J^\intercal\boldsymbol{v}_2. \tag{3.8}$$

Similarly, the new information matrix is found to be

$$\tilde{U} = \begin{pmatrix} I_m & 0 \\ 0 & J^\intercal \end{pmatrix} \begin{pmatrix} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\alpha}\boldsymbol{\beta}} \\ U_{\boldsymbol{\beta}\boldsymbol{\alpha}} & U_{\boldsymbol{\beta}\boldsymbol{\beta}} \end{pmatrix} \begin{pmatrix} I_m & 0 \\ 0 & J \end{pmatrix} = \begin{pmatrix} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}} & U_{\boldsymbol{\alpha}\boldsymbol{\beta}}J \\ J^\intercal U_{\boldsymbol{\beta}\boldsymbol{\alpha}} & J^\intercal U_{\boldsymbol{\beta}\boldsymbol{\beta}}J \end{pmatrix}.$$

Consequently, we have

$$l_n(\tilde{\boldsymbol{\theta}}) = \ell_n(\tilde{\boldsymbol{\alpha}}, \tilde{\boldsymbol{\gamma}}) = \ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*) + (1/2)(\boldsymbol{v}_1^\intercal, \boldsymbol{v}_2^\intercal J)\tilde{U}^{-1}(\boldsymbol{v}_1^\intercal, \boldsymbol{v}_2^\intercal J)^\intercal + o_p(1).$$

Combining (3.7) and the above expansion, and noticing that $\ell_n(\boldsymbol{\alpha}^*, \boldsymbol{\gamma}^*) = l_n(\boldsymbol{\theta}^*)$, we have

$$R_n = 2\{l_n(\hat{\boldsymbol{\theta}}) - l_n(\tilde{\boldsymbol{\theta}})\} = \boldsymbol{v}^\intercal U^{-1}\boldsymbol{v} - (\boldsymbol{v}_1^\intercal, \boldsymbol{v}_2^\intercal J)\tilde{U}^{-1}(\boldsymbol{v}_1^\intercal, \boldsymbol{v}_2^\intercal J)^\intercal + o_p(1).$$

Applying the quadratic form decomposition formula given in Lemma 3.4 to the two quadratic forms on the RHS of the above expansion, we get

$$\begin{aligned} &\boldsymbol{v}^\intercal U^{-1}\boldsymbol{v} = \boldsymbol{\xi}^\intercal\Lambda^{-1}\boldsymbol{\xi} + \boldsymbol{v}_1^\intercal U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}\boldsymbol{v}_1, \\ &(\boldsymbol{v}_1^\intercal, \boldsymbol{v}_2^\intercal J)\tilde{U}^{-1}(\boldsymbol{v}_1^\intercal, \boldsymbol{v}_2^\intercal J)^\intercal = \boldsymbol{\xi}^\intercal J(J^\intercal\Lambda J)^{-1}J^\intercal\boldsymbol{\xi} + \boldsymbol{v}_1^\intercal U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}\boldsymbol{v}_1, \end{aligned} \tag{3.9}$$

where $\boldsymbol{\xi} = (-U_{\boldsymbol{\beta}\boldsymbol{\alpha}}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1},\ I_{md})\boldsymbol{v}$ and $\Lambda = U_{\boldsymbol{\beta}\boldsymbol{\beta}} - U_{\boldsymbol{\beta}\boldsymbol{\alpha}}U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1}U_{\boldsymbol{\alpha}\boldsymbol{\beta}}$ is defined in

Theorem 3.2. We then obtain the following expansion

$$R_n = 2\{l_n(\hat{\boldsymbol{\theta}}) - l_n(\tilde{\boldsymbol{\theta}})\} = \boldsymbol{\xi}^\intercal\{\Lambda^{-1} - J(J^\intercal\Lambda J)^{-1}J^\intercal\}\boldsymbol{\xi} + o_p(1). \tag{3.10}$$

Recall that, by Theorem 2.2, $\boldsymbol{v}$ is asymptotically $N(\mathbf{0},\, U - UWU)$, where $W = \mathrm{diag}\{T,\, 0_{md\times md}\}$ and $T = {\rho_0}^{-1}\mathbf{1}_m\mathbf{1}_m^\intercal + \mathrm{diag}\{\rho_1^{-1}, \ldots, \rho_m^{-1}\}$ as given in Theorem 2.2. Thus, $\boldsymbol{\xi}$ is asymptotic normal with mean $\mathbf{0}$ and covariance matrix

$$(-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\, I_{md})(U - UWU)(-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\, I_{md})^\intercal.$$

Noting that

$$(-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\, I_{md})U(-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\, I_{md})^\intercal = U_{\boldsymbol{\beta\beta}} - U_{\boldsymbol{\alpha\beta}}^\intercal U_{\boldsymbol{\alpha\alpha}}U_{\boldsymbol{\alpha\beta}} = \Lambda$$

and

$$(-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\ I_{md})UW = \Big(-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1}U_{\boldsymbol{\alpha\alpha}}T + U_{\boldsymbol{\beta\alpha}}T,\quad 0\Big) = 0,$$

we get that the asymptotic covariance matrix of $\boldsymbol{\xi}$ is Λ.

The last step is to verify the quadratic form in the above expansion of R_n has the claimed limiting distribution. We can easily check that

$$\Lambda^{1/2}\{\Lambda^{-1} - J(J^\intercal\Lambda J)^{-1}J^\intercal\}\Lambda^{1/2} = I_{md} - \Lambda^{1/2}J(J^\intercal\Lambda J)^{-1}J^\intercal\Lambda^{1/2}$$

is idempotent. Moreover, by the additivity and commutativity of the trace

operation, we find the trace of the above idempotent matrix to be

$$\begin{aligned}
&\operatorname{tr}\left\{I_{md} - \Lambda^{1/2} J (J^\intercal \Lambda J)^{-1} J^\intercal \Lambda^{1/2}\right\} \\
=&\operatorname{tr}(I_{md}) - \operatorname{tr}\left\{\Lambda^{1/2} J (J^\intercal \Lambda J)^{-1} J^\intercal \Lambda^{1/2}\right\} \\
=&md - \operatorname{tr}\left\{(J^\intercal \Lambda J)^{-1} (J^\intercal \Lambda J)\right\} \\
=&md - \operatorname{tr}(I_{md-q}) \\
=&q.
\end{aligned}$$

Therefore, by Theorem 5.1.1 of Mathai (1992), the quadratic form in expansion (3.10), and hence also R_n, has a χ^2_q limiting distribution.

The above proof is applicable to $q < md$. When $q = md$, the value of $\boldsymbol{\beta}$ is fully specified. Hence, the maximization under null is solely with respect to $\boldsymbol{\alpha}$ and we easily find

$$l_n(\tilde{\boldsymbol{\theta}}) = l_n(\boldsymbol{\theta}^*) + (1/2)\boldsymbol{v}_1^\intercal U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1} \boldsymbol{v}_1 + o_p(1).$$

This, along with the expansion (3.7) of $l_n(\hat{\boldsymbol{\theta}})$ and expression of $\boldsymbol{v}^\intercal U^{-1} \boldsymbol{v}$ given in (3.9), implies that $R_n = \boldsymbol{\xi}^\intercal \Lambda^{-1} \boldsymbol{\xi} + o_p(1)$. Just as the proof for the case of $q < md$, the limiting distribution of the above R_n is seen to be χ^2_{md}. □

3.6.2 Theorem 3.2: Limiting distribution of the DELR under local alternatives

In this subsection, we prove that the DELR test statistic, R_n, has a non–central chi–square limiting distribution under the local alternative model (3.2). We first sketch out the proof. Let $\boldsymbol{\beta}^*$ be a specific parameter value under the null hypothesis and $\{F_k\}$ be the corresponding distribution functions. Let $\{G_k\}$ be the set of distribution functions satisfying the DRM with parameter given by $\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^* + n_k^{-1/2} \boldsymbol{c}_k$, $k = 1, \ldots, m$, and $G_0 = F_0$. When the samples are generated from the $\{G_k\}$, we still have that the DELR statistic is approximated by the quadratic form on the RHS of (3.10). The limiting

distribution of R_n is therefore determined by that of $\boldsymbol{v} = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}$. According to Le Cam's third lemma (van der Vaart 2000, 6.7), $\boldsymbol{v}$ has a specific limiting distribution under the $\{G_k\}$ if $\boldsymbol{v}$ and $\sum_{k,j}\log\{dG_k(x_{kj})/dF_k(x_{kj})\}$, under the $\{F_k\}$, are jointly normal with a particular mean and variance structure. The core of the proof then is to establish that structure.

For each $k = 0, 1, \ldots, m$, let $\textsc{Var}_k(\cdot)$ and $\textsc{Cov}_k(\cdot, \cdot)$ be the variance and covariance operators with respect to F_k, respectively.

Lemma 3.5. *Under the conditions of Theorem 3.1 and the distribution functions $\{G_k\}$, $\boldsymbol{v}$ is asymptotically normal with mean*

$$\boldsymbol{\tau} = \sum_{k=1}^{m}\sqrt{\rho_k}\textsc{Cov}_k\{\partial\mathcal{L}_k(\boldsymbol{\theta}^*, x)/\partial\boldsymbol{\theta}, \boldsymbol{q}^\intercal(x)\}\boldsymbol{c}_k$$

and covariance matrix $V = U - UWU$ as given in Theorem 2.2.

Proof of Lemma 3.5. We first expand $\boldsymbol{w}_k = \sum_{j=1}^{n_k}\log\{dG_k(x_{kj})/dF_k(x_{kj})\}$. Notice that

$$\begin{aligned} dG_k(x)/dF_k(x) &= \exp\{\alpha_k + \boldsymbol{\beta}_k\boldsymbol{q}(x)\}/\exp\{\alpha_k^* + \boldsymbol{\beta}_k^*\boldsymbol{q}(x)\} \\ &= \exp\{\alpha_k - \alpha_k^* + n_k^{-1/2}\boldsymbol{c}_k\boldsymbol{q}(x)\}. \end{aligned}$$

Because α_k is normalization constants that can be expressed as

$$\alpha_k = -\log\int\exp\{\boldsymbol{\beta}_k^\intercal\boldsymbol{q}(x)\}dF_0(x) = -\log\int\exp\{(\boldsymbol{\beta}_k^{*\intercal} + n_k^{-1/2}\boldsymbol{c}_k^\intercal)\boldsymbol{q}(x)\}dF_0(x),$$

we have

$$\begin{aligned} \exp\{\alpha_k^* - \alpha_k\} &= \int\exp\{\alpha_k^* + (\boldsymbol{\beta}_k^{*\intercal} + n_k^{-1/2}\boldsymbol{c}_k^\intercal)\boldsymbol{q}(x)\}dF_0(x). \\ &= \int\exp\{n_k^{-1/2}\boldsymbol{c}_k^\intercal\boldsymbol{q}(x)\}\exp\left\{\boldsymbol{\alpha}_k^* + \boldsymbol{\beta}_k^{*\intercal}\boldsymbol{q}(x)\right\}dF_0(x) \\ &= \int\exp\{n_k^{-1/2}\boldsymbol{c}_k^\intercal\boldsymbol{q}(x)\}dF_k(x) \end{aligned}$$

Expanding the exponential term on the RHS, we get

$$\exp\{\alpha_k^* - \alpha_k\} = \int \left\{1 + n_k^{-1/2}\boldsymbol{c}_k^\intercal \boldsymbol{q}(x) + (2n_k)^{-1}(\boldsymbol{c}_k^\intercal \boldsymbol{q}(x))^2\right\} dF_k(x) + \epsilon_n,$$

where $\epsilon_n \propto n_k^{-3/2} \int \|\boldsymbol{q}(x)\|^3 dF_k(x) = O(n^{-3/2})$ uniformly in x, because the third order moment of $\boldsymbol{q}(x)$ is finite. Denote $\boldsymbol{\nu}_k = \mathrm{E}_k \boldsymbol{q}(x)$. Then, the above equality is further simplified to

$$\exp\{\alpha_k^* - \alpha_k\} = 1 + n_k^{-1/2}\boldsymbol{c}_k^\intercal \boldsymbol{\nu}_k + (2n_k)^{-1}\boldsymbol{c}_k^\intercal \mathrm{E}_k(\boldsymbol{q}^2(x))\boldsymbol{c}_k + O(n^{-3/2}).$$

Hence, ignoring the $O(n^{-3/2})$ term, which is uniform in x, we have

$$\log\{dG_k(x)/dF_k(x)\} \approx n_k^{-1/2}\boldsymbol{c}_k \boldsymbol{q}(x) - \log\{1 + n_k^{-1/2}\boldsymbol{c}_k^\intercal \boldsymbol{\nu}_k + (2n_k)^{-1}\boldsymbol{c}_k^\intercal \mathrm{E}_k(\boldsymbol{q}^2(x))\boldsymbol{c}_k\}.$$

Write $\boldsymbol{\sigma}_k = \textsc{Var}_k(\boldsymbol{q}(x))$. Expanding the logarithmic term on the RHS, we get

$$\begin{aligned}
&\log\{1 + n_k^{-1/2}\boldsymbol{c}_k^\intercal \boldsymbol{\nu}_k + (2n_k)^{-1}\boldsymbol{c}_k^\intercal \mathrm{E}_k(\boldsymbol{q}^2(x))\boldsymbol{c}_k\} \\
=&n_k^{-1/2}\boldsymbol{c}_k^\intercal \boldsymbol{\nu}_k + (2n_k)^{-1}\boldsymbol{c}_k^\intercal \mathrm{E}_k(\boldsymbol{q}^2(x))\boldsymbol{c}_k - n_k\boldsymbol{c}_k^\intercal\{\boldsymbol{\nu}_k\boldsymbol{\nu}_k^\intercal\}\boldsymbol{c}_k + O(n^{-3/2}) \\
=&n_k^{-1/2}\boldsymbol{c}_k^\intercal \boldsymbol{\nu}_k + (2n_k)^{-1}\boldsymbol{c}_k^\intercal \boldsymbol{\sigma}_k\boldsymbol{c}_k + O(n^{-3/2}),
\end{aligned}$$

where the remainder $O(n^{-3/2})$ is again uniform in x. Therefore

$$\log\{dG_k(x)/dF_k(x)\} = n_k^{-1/2}\boldsymbol{c}_k^\intercal\{\boldsymbol{q}(x) - \boldsymbol{\nu}_k\} - (2n_k)^{-1}\boldsymbol{c}_k^\intercal \boldsymbol{\sigma}_k\boldsymbol{c}_k + O(n^{-3/2}),$$

uniformly in x. Summing over j, we get, for each k,

$$\begin{aligned}
\boldsymbol{w}_k &= \sum_{j=1}^{n_k} \log\{dG_k(x_{kj})/dF_k(x_{kj})\} \\
&= n_k^{-1/2}\boldsymbol{c}_k^\intercal \sum_{j=1}^{n_k}\{\boldsymbol{q}(x_{kj}) - \boldsymbol{\nu}_k\} - (1/2)\boldsymbol{c}_k^\intercal \boldsymbol{\sigma}_k\boldsymbol{c}_k + O(n^{-1/2}).
\end{aligned}$$

When $k = 0$, we have $\boldsymbol{c}_0 = 0$.

Recall that $\boldsymbol{v} = n^{-1/2}\partial l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}$, $l_n(\boldsymbol{\theta}^*) = \sum_{k,j}\mathcal{L}_{n,k}(\boldsymbol{\theta}^*, x_{kj})$ and $\hat{\lambda}_k = n_k/n$ whose limit is ρ_k, we have

$$\begin{pmatrix} \boldsymbol{v} \\ \sum_k \boldsymbol{w}_k \end{pmatrix} \approx \sum_{k=0}^{m} \frac{1}{\sqrt{n_k}} \sum_{j=1}^{n_k} \begin{pmatrix} \sqrt{\rho_k}\,\{\partial\mathcal{L}_{n,k}(\boldsymbol{\theta}^*, x_{kj})/\partial\boldsymbol{\theta} - \boldsymbol{\mu}_k\} \\ \boldsymbol{c}_k^\intercal\{\boldsymbol{q}(x_{kj}) - \boldsymbol{\nu}_k\} \end{pmatrix} - \sum_{k=0}^{m} \begin{pmatrix} \boldsymbol{0} \\ \frac{1}{2}\boldsymbol{c}_k^\intercal\boldsymbol{\sigma}_k\boldsymbol{c}_k \end{pmatrix},$$

which is seen to be jointly asymptotically normal under the null distributions $\{F_k\}$. The corresponding mean vector and variance matrix are given by

$$\left(\boldsymbol{0}^\intercal,\ -\frac{1}{2}\sum_k \boldsymbol{c}_k^\intercal\boldsymbol{\sigma}_k\boldsymbol{c}_k\right)^\intercal \text{ and } \begin{pmatrix} V & \boldsymbol{\tau} \\ \boldsymbol{\tau}^\intercal & \sum_k \boldsymbol{c}_k^\intercal\boldsymbol{\sigma}_k\boldsymbol{c}_k \end{pmatrix},$$

where $\boldsymbol{\tau}$ is the one given in the Lemma, and $V = U - UWU$ is the asymptotic covariance matrix of $\boldsymbol{v}$ as given in Theorem 2.2. Because the second entry of the mean vector equals negative half of the lower–right entry of the covariance matrix, the condition of Le Cam's third lemma is satisfied. By that lemma, we conclude that $\boldsymbol{v}$ has a normal limiting distribution with mean $\boldsymbol{\tau}$ and covariance matrix V under the local alternative distributions $\{G_k\}$. □

Proof of Theorem 3.2. We now show that the DELR statistic R_n has a non–central chi–square limiting distribution under the distributions $\{G_k\}$ that satisfy the local alternative model (3.2). We first show that, under the $\{G_k\}$, R_n is still approximated by the quadratic form on the RHS of (3.10).

Under the $\{G_k\}$, we still have $-n^{-1}\partial^2 l_n(\boldsymbol{\theta}^*)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal \to U$ and, by Lemma 3.5, $\boldsymbol{v} = O_p(1)$. In addition, $\hat{\boldsymbol{\theta}}$ still admits the expansion

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*) = U^{-1}\boldsymbol{v} + o_p(1) = O_p(1),$$

and hence it is root–n consistent for $\boldsymbol{\theta}^*$. Similarly, the constrained MELE $\tilde{\boldsymbol{\theta}}$ is also root–n consistent for $\boldsymbol{\theta}^*$ under the $\{G_k\}$. The root–n consistency of

$\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$ imply

$$R_n = \boldsymbol{\xi}^\intercal\{\Lambda^{-1} - J(J^\intercal\Lambda J)^{-1}J^\intercal\}\boldsymbol{\xi} + o_p(1)$$

when $q < md$, and $R_n = \boldsymbol{\xi}^\intercal\Lambda^{-1}\boldsymbol{\xi} + o_p(1)$ when $q = md$. The matrix in the quadratic form of the expansion of R_n is the same as that in (3.10). What has changed is the distribution of $\boldsymbol{\xi} = (-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\ I_{md})\boldsymbol{v}$.

By Lemma 3.5, under the local alternative $\{G_k\}$, $\boldsymbol{v}$ is asymptotically $N(\boldsymbol{\tau},\, V)$. Hence $\boldsymbol{\xi}$ also has a normal limiting distribution. Since the asymptotic covariance matrix of $\boldsymbol{v}$ is the same as that under the $\{F_k\}$, the asymptotic covariance matrix of $\boldsymbol{\xi}$ is still Λ as we have shown in the proof of Theorem 3.1. The mean of the limiting distribution of $\boldsymbol{\xi}$ now is $\boldsymbol{\mu} = (-U_{\boldsymbol{\beta\alpha}}U_{\boldsymbol{\alpha\alpha}}^{-1},\ I_{md})\boldsymbol{\tau}$. Partition $\boldsymbol{\tau}$ into $(\boldsymbol{\tau}_{\boldsymbol{\alpha}}^\intercal,\, \boldsymbol{\tau}_{\boldsymbol{\beta}}^\intercal)^\intercal$ as

$$\begin{aligned}
\boldsymbol{\tau}_{\boldsymbol{\alpha}} &= \sum_{k=1}^{m}\sqrt{\rho_k}\mathrm{Cov}_k\{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\alpha}, \boldsymbol{q}^\intercal(x)\}\boldsymbol{c}_k,\\
\boldsymbol{\tau}_{\boldsymbol{\beta}} &= \sum_{k=1}^{m}\sqrt{\rho_k}\mathrm{Cov}_k\{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)/\partial\boldsymbol{\beta}, \boldsymbol{q}^\intercal(x)\}\boldsymbol{c}_k.
\end{aligned}$$

By relationship (2.18) and (2.19), we have

$$\begin{aligned}
\mathrm{Cov}_k\left(\frac{\partial\mathcal{L}_k(\boldsymbol{\theta}^*,\, x)}{\partial\boldsymbol{\alpha}},\ \boldsymbol{q}^\intercal(x)\right) &= \frac{1}{\rho_k}\left\{U_{\boldsymbol{\alpha\beta}}(\boldsymbol{e}_k\otimes I_d) - U_{\boldsymbol{\alpha\alpha}}\boldsymbol{e}_k\mathrm{E}_k\big(\boldsymbol{q}^\intercal(x)\big)\right\},\\
\mathrm{Cov}_k\left(\left\{\boldsymbol{e}_k - \frac{\boldsymbol{h}(\boldsymbol{\theta}^*, x)}{s(\boldsymbol{\theta}^*, x)}\right\}\otimes\boldsymbol{q}(x),\ \boldsymbol{q}^\intercal(x)\right) &= \frac{1}{\rho_k}\left\{U_{\boldsymbol{\beta\beta}}(\boldsymbol{e}_k\otimes I_d) - U_{\boldsymbol{\beta\alpha}}\boldsymbol{e}_k\mathrm{E}_k\big(\boldsymbol{q}^\intercal(x)\big)\right\}.
\end{aligned}$$

Plugging the above expressions into the expressions of $\boldsymbol{\tau}_{\boldsymbol{\alpha}}$ and $\boldsymbol{\tau}_{\boldsymbol{\beta}}$, we get

$$\begin{aligned}
\boldsymbol{\tau}_{\boldsymbol{\alpha}} &= \sum_{k=1}^{m}{\rho_k}^{-1/2}\left\{U_{\boldsymbol{\alpha\beta}}(\boldsymbol{e}_k\otimes I_d) - U_{\boldsymbol{\alpha\alpha}}\boldsymbol{e}_k\mathrm{E}_k\big(\boldsymbol{q}^\intercal(x)\big)\right\}\boldsymbol{c}_k,\\
\boldsymbol{\tau}_{\boldsymbol{\beta}} &= \sum_{k=1}^{m}{\rho_k}^{-1/2}\left\{U_{\boldsymbol{\beta\beta}}(\boldsymbol{e}_k\otimes I_d) - U_{\boldsymbol{\beta\alpha}}\boldsymbol{e}_k\mathrm{E}_k\big(\boldsymbol{q}^\intercal(x)\big)\right\}\boldsymbol{c}_k.
\end{aligned} \tag{3.11}$$

Consequently,

$$\boldsymbol{\mu} = -U_{\boldsymbol{\beta\alpha}} U_{\boldsymbol{\alpha\alpha}}^{-1} \boldsymbol{\tau}_{\boldsymbol{\alpha}} + \boldsymbol{\tau}_{\boldsymbol{\beta}} = \Lambda \sum_{k=1}^{m} {\rho_k}^{-1/2} (\boldsymbol{e}_k \otimes I_d) \boldsymbol{c}_k = \Lambda \boldsymbol{\eta},$$

where the second last equality is by (3.11) and $\boldsymbol{\eta}$ is defined in Theorem 3.2.

In the proof of Theorem 3.1, we have verified that the matrix

$$A = \Lambda^{1/2} \{ \Lambda^{-1} - J(J^{\intercal} \Lambda J)^{-1} J^{\intercal} \} \Lambda^{1/2}$$

is idempotent with rank q. Hence, by Corollary 5.1.3a of Mathai (1992), the quadratic form in the above expansion of R_n, and hence R_n, has a non–central chi–square limiting distribution with q degrees of freedom and non–central parameter

$$\delta^2 = \boldsymbol{\mu}^{\intercal} \{ \Lambda^{-1} - J(J^{\intercal} \Lambda J)^{-1} J^{\intercal} \} \boldsymbol{\mu} = \boldsymbol{\eta}^{\intercal} \{ \Lambda - \Lambda J (J^{\intercal} \Lambda J)^{-1} J^{\intercal} \Lambda \} \boldsymbol{\eta}$$

in the case of $q < md$, and

$$\delta^2 = \boldsymbol{\mu}^{\intercal} \Lambda^{-1} \boldsymbol{\mu} = \boldsymbol{\eta}^{\intercal} \Lambda \boldsymbol{\mu}$$

in the case of $q = md$.

In the last step we verify the condition for positiveness of the non–central parameter δ^2. When $q = md$, $\delta^2 = \boldsymbol{\eta}^{\intercal} \Lambda \boldsymbol{\eta} > 0$ because Λ is positive definite. When $q < md$, $\delta^2 = (\boldsymbol{\eta}^{\intercal} \Lambda^{1/2}) A (\Lambda^{1/2} \boldsymbol{\eta})$. We verified that A is an idempotent matrix. Hence, A is positive semidefinite and $\delta^2 \geq 0$. Moreover, $\delta^2 = 0$ if and only if $\Lambda^{1/2} \boldsymbol{\eta}$ is in the null space of A. The null space of A is the column space of $I - A = \Lambda^{1/2} J \big(J^{\intercal} \Lambda J \big)^{-1} J^{\intercal} \Lambda^{1/2}$, which is just the column space of $\Lambda^{1/2} J$. It is easily verified that $\Lambda^{1/2} \boldsymbol{\eta}$ is in the column space of $\Lambda^{1/2} J$ if and only if $\boldsymbol{\eta}$ is in the column space of J. Hence $\Lambda^{1/2} \boldsymbol{\eta}$ is in the null space of A and $\delta^2 = 0$ if and only if $\boldsymbol{\eta}$ is in the column space of J. $\square$

3.7 Appendix: Parameter values in simulation studies

Table 3.10: Parameter values for power comparison under non–normal distributions (Section 3.3.2).

$\Gamma(\lambda, \kappa)$: gamma distribution with shape λ and rate κ;
$LN(\mu, \sigma)$: log–normal distribution with mean μ and standard deviation σ on log scale;
$Pa(\gamma)$: Pareto distribution with shape γ and common support of $x > 1$;
$W(b)$: Weibull distribution with scale b and common shape of 0.8.
F_0 remains unchanged across parameter settings 0–5.

		Parameter settings									
F_0		1		2		3		4		5	
		λ	κ	λ	κ	λ	κ	λ	κ	λ	κ
$\Gamma(0.2, 0.8)$	F_1:	0.18	0.7	0.17	0.6	0.16	0.5	0.155	0.45	0.14	0.4
	F_2:	0.22	0.85	0.24	0.95	0.255	1.05	0.18	0.7	0.17	0.6
	F_3:	0.23	0.95	0.255	1.2	0.275	1.25	0.29	1.4	0.33	1.6
	F_4:	0.24	1.05	0.27	1.3	0.29	1.4	0.31	1.55	0.35	1.85
$LN(0, 1.5)$		μ	σ	μ	σ	μ	σ	μ	σ	μ	σ
	F_1:	0.44	1.3	0.7	1.2	0.9	1.15	1	1	1.2	0.85
	F_2:	0.22	1.32	0.57	1.30	0.62	1.25	0.67	1.20	0.87	1
	F_3:	0.18	1.35	0.63	1.33	0.73	1.30	0.83	1.28	0.85	1.28
	F_4:	0.37	1.38	0.60	1.35	0.70	1.33	0.75	1.32	0.95	1.30
$Pa(2)$		γ		γ		γ		γ		α	
	F_1:	1.9		1.85		1.8		1.75		1.7	
	F_2:	2.1		2.2		2.3		1.85		1.75	
	F_3:	2.35		2.55		2.70		2.85		3.25	
	F_4:	2.5		2.78		2.98		3.2		3.75	
$W(1)$		b		b		b		b		b	
	F_1:	0.76		0.65		0.59		0.53		0.42	
	F_2:	1.2		1.26		1.31		1.35		1.42	
	F_3:	1.08		1.05		1.10		1.12		1.14	
	F_3:	0.90		0.89		0.85		0.82		0.78	

Table 3.11: Parameter values for power comparison under misspecified DRMs (Section 3.4.2).

$W(a,\, b)$: Weibull distribution with shape a and scale b;
$t(\nu,\, c)$: non–central t distribution with ν degrees of freedom and non–central parameter c;
F_0 remains unchanged across parameter settings 0–5.

		Parameter settings									
F_0		1		2		3		4		5	
		a	b	a	b	a	b	a	b	a	b
$W(1,\, 1)$	F_1:	0.9	0.95	0.85	0.94	0.82	0.92	0.79	0.91	0.75	0.88
	F_2:	0.98	0.98	0.96	0.96	0.95	0.95	0.94	0.94	0.91	0.92
	F_3:	1.03	1.04	1.05	1.06	1.07	1.07	1.09	1.08	1.12	1.12
	F_4:	1.01	0.95	1.02	0.92	1.03	0.90	1.05	0.89	1.07	0.85
		ν	c	ν	c	ν	c	ν	c	ν	c
$t(4,\, 0)$	F_1:	3.85	-0.05	3.75	-0.1	3.65	-0.1	3.45	-0.15	3.4	-0.2
	F_2:	5	0.1	6	0.05	6.5	0.1	3.8	0.1	3.7	0.15
	F_3:	4.5	0.05	4.8	0.1	5.5	0.15	6	0.15	8	0.2
	F_4:	6	0.1	6.5	0.15	8.5	0.2	8	0.23	10	0.28
$0.6N(1,1.3)+$ $0.4N(0.8,1)$	F_1:	$0.8N(1,1.3)+$ $0.2N(0.8,1)$		$0.8N(1.05,1.3)+$ $0.2N(0.8,1)$		$0.8N(1.1,1.2)+$ $0.2N(0.8,0.95)$		$0.8N(1.15,1.18)+$ $0.2N(0.85,0.9)$		$0.8N(1.2,1.15)+$ $0.2N(0.9,0.9)$	
	F_2:	$0.6N(1,1.3)+$ $0.4N(0.8,1)$		$0.5N(1,1.3)+$ $0.5N(0.75,0.95)$		$0.5N(1,1.25)+$ $0.5N(0.7,1)$		$0.5N(1.1,1.25)+$ $0.5N(0.7,0.95)$		$0.5N(1.1,1.3)+$ $0.5N(0.65,0.92)$	
	F_3:	$0.3N(1,1.3)+$ $0.7N(0.8,1)$		$0.3N(1,1.25)+$ $0.7N(0.8,0.95)$		$0.3N(0.9,1.15)+$ $0.7N(0.8,87)$		$0.3N(0,8,1.07)+$ $0.7N(0.8,0.83)$		$0.3N(0.75,1.02)+$ $0.7N(0.75,0.78)$	
	F_4:	$0.6N(1,1.3)+$ $0.4N(0.8,1)$		$0.2N(1,1.3)+$ $0.8N(0.8,1)$		$0.1N(1,1.3)+$ $0.9N(0.95,1)$		$0.1N(0.95,1.3)+$ $0.9N(1,1)$		$0.1N(0.9,1.25)+$ $0.9N(1.1,0.95)$	
$0.3\Gamma(0.8,1)+$ $0.7W(1,1.3)$	F_1:	$0.45\Gamma(0.85,1)+$ $0.55W(1.15,1.22)$		$0.45\Gamma(0.85,1)+$ $0.55W(1.3,1.15)$		$0.4\Gamma(0.9,1)+$ $0.6W(1.4,1.15)$		$0.4\Gamma(1,0.9)+$ $0.6W(1.35,0.9)$		$0.4\Gamma(1.4,0.8)+$ $0.6W(1.5,0.8)$	
	F_2:	$0.3\Gamma(1.05,0.9)+$ $0.7W(1,1.25)$		$0.3\Gamma(1.15,0.9)+$ $0.7W(1,1.23)$		$0.3\Gamma(1.25,0.9)+$ $0.7W(1,1.2)$		$0.3\Gamma(1.4,0.8)+$ $0.7W(1,1.2)$		$0.3\Gamma(2,0.7)+$ $0.7W(1,1)$	
	F_3:	$0.25\Gamma(0.8,1.4)+$ $0.75W(1,1.3)$		$0.25\Gamma(0.85,1.65)+$ $0.75W(1.05,1.28)$		$0.25\Gamma(0.85,1.85)+$ $0.75W(1.1,1.25)$		$0.25\Gamma(0.95,2.05)+$ $0.75W(1.1,1.25)$		$0.25\Gamma(1.2,2.5)+$ $0.75W(1.2,1.2)$	
	F_4:	$0.3\Gamma(0.8,0.93)+$ $0.7W(1.05,1.3)$		$0.3\Gamma(0.8,0.88)+$ $0.7W(1,1.3)$		$0.3\Gamma(0.8,0.85)+$ $0.7W(1,1.3)$		$0.3\Gamma(0.8,0.7)+$ $0.7W(1.3,1)$		$0.3\Gamma(1.6,0.5)+$ $0.7W(1.4,0.9)$	

Chapter 4

Effects of Information Pooling by DRM

Our use of DRM is motivated by its ability to pool information across a number of samples. We believe the resulting inferences are more efficient than inferences based on individual samples. Moreover, strong evidence about this improved efficiency already exists: Fokianos (2004) obtained more efficient density estimators under DRM than the classical kernel density estimators based on individual samples; Chen and Liu (2013) found DRM–based quantile estimators to be more efficient than the empirical quantile estimators. We also anticipate that there will be a gain on the estimation accuracy of the MELE of the DRM parameter $\boldsymbol{\theta}$ and on the power the DELR test if we combine information from additional samples using the DRM — a topic that is not studied in literature. This chapter provides both theoretical and simulation supports for this conjecture.

4.1 Effects on the estimation accuracy of the MELE

Suppose we have $m+1$ independent samples whose distributions satisfy the DRM (2.1). Yet our interest may well focus on the inference about a subset of these distributions, without loss of generality, the first $r+1$ distributions F_0, F_1, $\ldots$, F_r with $1 \leq r < m$. Shall we base the inference on a DRM fitted to the first $r+1$ samples or on one fitted to all the $m+1$ samples? We here

prove that the latter DRM yields a higher estimation accuracy for the MELE of the DRM parameter that corresponds to the distributions of interest.

Let $\boldsymbol{\nu} = (\alpha_1, \ldots, \alpha_r, \boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_r^\intercal)^\intercal$ denote the DRM parameter for the first r non–baseline distributions. Denote the MELEs of $\boldsymbol{\nu}$ based on the DRM for the first $r+1$ distributions and that for all distributions as $\hat{\boldsymbol{\nu}}^{(1)}$ and $\hat{\boldsymbol{\nu}}^{(2)}$ respectively. Let $\Sigma^{(1)}$ and $\Sigma^{(2)}$ be the corresponding asymptotic covariance matrices. Recall that the size of the kth sample is n_k, $k = 0, 1, \ldots, m$. Put $n^{(1)} = \sum_{k=0}^{r} n_k$ and $\rho = \lim_{n\to\infty} n^{(1)}/n$, where n is the total size of all the $m+1$ samples. By the asymptotic normality of the MELE (Theorem 2.5), we have

$$\sqrt{n^{(1)}}(\hat{\boldsymbol{\nu}}^{(1)} - \boldsymbol{\nu}^*) \xrightarrow{(d)} N(0, \Sigma^{(1)}),$$
$$\sqrt{n}(\hat{\boldsymbol{\nu}}^{(2)} - \boldsymbol{\nu}^*) \xrightarrow{(d)} N(0, \Sigma^{(2)}),$$

where $\boldsymbol{\nu}^*$ is the true parameter value. Hence the asymptotic covariance matrix of $\hat{\boldsymbol{\nu}}^{(1)}$ is approximated by $\Sigma^{(1)}/n^{(1)}$ and that of $\hat{\boldsymbol{\nu}}^{(2)}$ is approximated by $\Sigma^{(2)}/n$. The estimation accuracy of an estimator is measured by the inverse of its covariance matrix. The scaling factors in the above results are different, so, to ensure fairness, we should compare $\Sigma^{(1)}$ to $\lim_{n\to\infty}(n^{(1)}/n)\Sigma^{(2)} = \rho\Sigma^{(2)}$. The following theorem tells us that the estimation accuracy of $\hat{\boldsymbol{\nu}}^{(2)}$ is never lower than that of $\hat{\boldsymbol{\nu}}^{(1)}$.

Recall that, for a matrix A, we use $A > 0$ ($A \geq 0$) to denote that A is positive definite (positive semidefinite). For matrices A and B, we will use $A > B$ ($A \geq B$) to represent $A - B > 0$ ($A - B \geq 0$).

Theorem 4.1. *Under the conditions of Theorem 2.1,* $\Sigma^{(1)} \geq \rho\Sigma^{(2)}$.

Example 4.1. *Consider the situation where* $m+1 = 4$*, samples are from a DRM with basis function* $\boldsymbol{q}(x) = (x, \log x)^\intercal$*, and the sample proportions are* $(3/11, 3/22, 2/11, 9/22)$*. Let* F_k*,* $k = 1, 2, 3$*, be the distributions with parameters* $\boldsymbol{\beta}_1^* = (-2, 2)^\intercal$*,* $\boldsymbol{\beta}_2^* = (-.25, 0.2)^\intercal$ *and* $\boldsymbol{\beta}_2^* = (-0.5, 0.1)^\intercal$*, respectively.*

Let $r+1=2$ *and* $\boldsymbol{\nu}=(\alpha_1,\,\boldsymbol{\beta}_1^\intercal)^\intercal$*, then* $\rho=9/22$*. When* F_0 *is* $\Gamma(2,\,1)$*, we numerically compute the information matrices* (2.13) *for the parameters of the DRM based on the first* $r+1=2$ *samples and that based on all the* $m+1=4$ *samples. Based on the information matrices, we obtain the asymptotic covariance matrices,* $\Sigma^{(1)}$ *and* $\Sigma^{(2)}$*, of the MELEs for* $\boldsymbol{\nu}$ *under these two DRMs by* (2.14)*. The smallest eigenvalue of* $\Sigma_1-r\Sigma_2$ *is then found to be approximately* 0.237*, so* $\Sigma_1>r\Sigma_2$*.*

Example 4.2. *Consider the situation where* $m+1=5$*, samples are from a DRM with basis function* $\boldsymbol{q}(x)=x^{1.5}$*, and the sample proportions are* (0.2, 0.1, 0.1, 0.2, 0.2)*. Let* F_k*,* $k=1,\,\ldots,\,4$*, be the distributions with parameters* $\boldsymbol{\beta}_1^*=-9.78$*,* $\boldsymbol{\beta}_2^*=0.72$*,* $\boldsymbol{\beta}_3^*=1.11$*, and* $\boldsymbol{\beta}_4^*=-1.43$*, respectively.*

Let $r+1=3$ *and* $\boldsymbol{\nu}=(\alpha_1,\,\alpha_2,\,\boldsymbol{\beta}_1^\intercal,\,\boldsymbol{\beta}_2^\intercal)^\intercal$*, then* $\rho=0.4$*. When* F_0 *is the Weibull distribution with shape of* 1.5 *and scale of* 0.8*, we numerically compute the information matrices* (2.13) *for the parameters of the DRM based on the first* $r+1=3$ *samples and that based on all the* $m+1=5$ *samples. Based on the information matrices, we obtain the asymptotic covariance matrices,* $\Sigma^{(1)}$ *and* $\Sigma^{(2)}$*, of the MELEs for* $\boldsymbol{\nu}$ *under these two DRMs by* (2.14)*. The smallest eigenvalue of* $\Sigma_1-r\Sigma_2$ *is then found to be approximately* 0.013*, so* $\Sigma_1>r\Sigma_2$*.*

4.2 Effects on the power of the DELR test

In this section, we show that the local asymptotic power of the DELR test is often increased when strength is borrowed from additional samples even when their underlying distributions are unrelated to the hypothesis of interest.

We adopt the setting posited in the last section for multiple samples from distributions satisfying the DRM assumption. A hypothesis of interest may also focus on a characteristic of just a subset of these populations. If so, why should our tests be based on all the samples? One answer is found in their improved local power as we now demonstrate.

Without loss of generality, consider a null hypothesis regarding subpopulations $F_0, F_1, \ldots, F_r$ with $r < m$ and let $\boldsymbol{\zeta}^\intercal = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_r^\intercal)$. The composite hypotheses are specified as

$$H_0 : \ \boldsymbol{g}(\boldsymbol{\zeta}) = \mathbf{0} \quad \text{against} \quad H_1 : \ \boldsymbol{g}(\boldsymbol{\zeta}) \neq \mathbf{0} \tag{4.1}$$

for some smooth function $\boldsymbol{g} : \mathbb{R}^{rd} \to \mathbb{R}^q$ with $q \leq rd$. A DELR test can be based either on samples from just $F_0, F_1, \ldots, F_r$, or on the samples from all the populations $F_0, F_1, \ldots, F_m$. We denote the corresponding test statistics as $R_n^{(1)}$ and $R_n^{(2)}$, respectively.

Theorem 3.1 implies that $R_n^{(1)}$ and $R_n^{(2)}$ have the same χ_q^2 distribution in the limit under the null model of (4.1). Theorem 3.2, on the other hand, provides a useful tool for comparing their local asymptotic powers. It implies that $R_n^{(1)}$ and $R_n^{(2)}$ have non–central chi–square limiting distributions of the same q degrees of freedom, however with possibly different non–central parameter values at a local alternative. By the result (3.6) in Section 3.2.3, a power comparison can therefore be made using these non–central parameter values. The following two theorems, whose proofs are given in Section 4.4, implement this idea and provide that power comparison both in a general and a special situation.

Theorem 4.2. *Adopt the conditions of Theorem 2.1, the hypotheses (4.1) and the local alternatives*

$$\boldsymbol{\beta}_k = \begin{cases} \boldsymbol{\beta}_k^* + n_k^{-1/2}\boldsymbol{c}_k, & \text{if } k = 1, \ldots, r, \\ \boldsymbol{\beta}_k^*, & \text{otherwise,} \end{cases} \tag{4.2}$$

for some given set of constants $\{\boldsymbol{c}_k\}$. Let δ_1^2 and δ_2^2 be non–central parameter values of the limiting distribution of $R_n^{(1)}$ and $R_n^{(2)}$ under the local alternative model. Then $\delta_2^2 \geq \delta_1^2$.

Example 4.3. *Consider the situation where $m + 1 = 3$, samples are from a DRM with basis function $\boldsymbol{q}(x) = (x, x^2)^\intercal$, and the sample proportions*

are $(0.5,\, 0.25,\, 0.25)$. *Let* F_k, $k = 1,\, 2$, *be the distributions with parameters* $\boldsymbol{\beta}_1^* = (6,\, -1.5)^\intercal$ *and* $\boldsymbol{\beta}_2^* = (-0.25,\, 0.375)^\intercal$. *Suppose* H_0 *is given as* $\boldsymbol{g}(\boldsymbol{\zeta}) = \boldsymbol{\beta}_1 - (6,\, -1.5)^\intercal = \mathbf{0}$, *and the local alternative is*

$$\boldsymbol{\beta}_1 = \boldsymbol{\beta}_1^* + n_1^{-1/2}\boldsymbol{c}_1;\ \boldsymbol{\beta}_2 = \boldsymbol{\beta}_2^*$$

with $\boldsymbol{c}_1 = (2,\, 2)^\intercal$.

Let $R_n^{(1)}$ *and* $R_n^{(2)}$ *be the DELR test statistics based on* F_0, F_1, *and on* F_0, F_1, F_2, *respectively. When* F_0 *is,* $N(0,\, 1)$, *the standard normal distribution, we obtain information matrices* (2.13), *and hence* $\Lambda = U_{\boldsymbol{\beta}\boldsymbol{\beta}} - U_{\boldsymbol{\beta}\boldsymbol{\alpha}} U_{\boldsymbol{\alpha}\boldsymbol{\alpha}}^{-1} U_{\boldsymbol{\alpha}\boldsymbol{\beta}}$, *for* $R_n^{(1)}$ *and* $R_n^{(2)}$ *based on numerical computation. For* $R_n^{(1)}$, *we have* $\boldsymbol{\eta} = (4,\, 4)^\intercal$ *and* $q = d = 2$. *Then by Theorem 3.2, we find* $\delta_1^2 = \boldsymbol{\eta}^\intercal \Lambda \boldsymbol{\eta} \approx 5.90$. *For* $R_n^{(2)}$, *we find* $\boldsymbol{\eta} = (4,\, 4,\, 0, 0)^\intercal$, $\bigtriangledown = (I_2,\, 0_{2\times 2})$, *and* $J = (0_{2\times 2},\, I_2)^\intercal$. *By Theorem 3.2, we get* $\delta_2^2 \approx 6.67$. *Now, since* $\delta_1^2 \approx 5.90 < \delta_2^2 \approx 6.67$ *and the degrees of freedom of the limiting distributions of both* $R_n^{(1)}$ *and* $R_n^{(2)}$ *are* 2, $R_n^{(2)}$ *is more powerful than* $R_n^{(1)}$ *even though the null hypothesis concerns the parameter of just population 1.*

Note that at the 5% *level, the powers of* $R_n^{(1)}$ *and* $R_n^{(2)}$ *are approximately* 0.577 *and* 0.633, *respectively.*

The asymptotic power of $R_n^{(2)}$ (based on all samples) is not always higher than that of $R_n^{(1)}$. We demonstrate this fact in the following special case. Partition $\{1,\, \ldots,\, r\}$ into K parts denoted by $\mathbb{S}_k$, $k = 1,\, \ldots,\, K$, such that the size, s_k, of $\mathbb{S}_k$ satisfies $s_1 \geq 0$ and $s_k \geq 2$ for $k = 2,\, \ldots,\, K$. Let $\boldsymbol{\zeta}_k$ be the vector consisting of all the $\boldsymbol{\beta}_j$ with $j \in \mathbb{S}_k$. Consider the null hypothesis H_0 composed of

$$\boldsymbol{g}_k(\boldsymbol{\zeta}) = A_k \boldsymbol{\zeta}_k = 0 \text{ for } k = 1,\, \ldots,\, K, \tag{4.3}$$

with $A_1 = I_{s_1 d}$ and $A_k = (\mathbf{1}_{(s_k-1)} \otimes I_d,\, -I_{(s_k-1)d})$ for $k = 2, \ldots, K$. In other words, H_0 posits that the distributions within the first group are all identical to F_0 and those within any other given group are identical to each other.

When s_1, the size of $\mathbb{S}_1$, is 0, no non–baseline distribution is compared to the baseline F_0.

Theorem 4.3. *Adopt the conditions postulated in Theorem 4.2. For testing the null hypothesis (4.3), the limiting distributions of $R_n^{(1)}$ and $R_n^{(2)}$ under local alternative (4.2) have equal non–central parameters: $\delta_1^2 = \delta_2^2$.*

Example 4.4. *Consider the situation where $m + 1 = 6$, samples are from a DRM with basis function $\boldsymbol{q}(x) = (\log x,\, x)^\intercal$, and the sample proportions are $(0.25,\, 0.22,\, 0.16,\, 0.10,\, 0.17,\, 0.10)$. Let F_k, $k = 1, \ldots, 5$, to be the distributions with parameters $\boldsymbol{\beta}_1^* = (0,\, 0)^\intercal$, $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}_3^* = (-1,\, -0.5)^\intercal$, $\boldsymbol{\beta}_4^* = (0.3,\, 1.2)^\intercal$, and $\boldsymbol{\beta}_5^* = (-1.2,\, -0.4)^\intercal$. We partition $\{1,\, 2,\, 3\}$ into $\mathbb{S}_1 = \{1\}$ and $\mathbb{S}_2 = \{2,\, 3\}$. The null hypothesis (4.3) postulates $F_0 = F_1$ and $F_2 = F_3$. Let the local alternative be*

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^* + n_k^{-1/2}\boldsymbol{c}_k, \text{ for } k = 1, 2, 3,$$

with $\boldsymbol{c}_1 = \boldsymbol{c}_2 = (1,\, 1)^\intercal$ and $\boldsymbol{c}_3 = (-1,\, 2)^\intercal$.

Let $R_n^{(1)}$ and $R_n^{(2)}$ be the DELR test statistics based on $F_0, \ldots, F_3$, and on $F_0, \ldots, F_5$, respectively. When F_0 is $\Gamma(2,\, 1)$, we obtained information matrices (2.13), and hence Λ, for $R_n^{(1)}$ and $R_n^{(2)}$ based on numerical computation. For $R_n^{(1)}$, we find $\boldsymbol{\eta} \approx (2.13,\, 2.13,\, 2.5,\, 2.5,\, -3.16,\, 6.32)^\intercal$, $\bigtriangledown = diag\{I_2,\, (I_2, -I_2)\}$, and $J = (0_{2\times 2},\, I_2,\, I_2)^\intercal$. For $R_n^{(2)}$, we get $\boldsymbol{\eta} \approx (2.13,\, 2.13,\, 2.5,\, 2.5,\, -3.16,\, 6.32,\, 0,\, 0,\, 0,\, 0)^\intercal$, $\bigtriangledown = (diag\{I_2,\, (I_2, -I_2)\},\, 0_{4\times 4})$, and $J = (0_{6\times 2},\, (I_2,\, 0_{2\times 4})^\intercal,\, I_6)^\intercal$. By Theorem 3.2, we confirm that $\delta_1^2 = \delta_2^2 \approx 2.72$. Hence $R_n^{(1)}$ and $R_n^{(2)}$ are asymptotically equally powerful.

The scenario presented in Theorem 4.3 is similar to the one–way ANOVA used in experimental design (Wu and Hamada, 2009). Suppose there are five treatments under investigation and we want to test the equal mean hypothesis of the first two treatments. One may use pooled variance estimator from all samples to construct the two–sample t–test. This test gains in the degrees

of freedom comparing to the t–test based on the first two samples alone, but not in the first order asymptotics.

4.3 Simulation studies

4.3.1 Comparison of estimation accuracy: $\hat{\boldsymbol{\nu}}^{(1)}$ versus $\hat{\boldsymbol{\nu}}^{(2)}$

We now conduct simulation studies to compare the estimation accuracy of the MELE based on a subset of the samples to that based on all the samples. The number of repetitions for simulation is set to $100,000$ for a high simulation accuracy.

Recall that we have used $\boldsymbol{\nu}$ to denote the DRM parameter of interest, and $\hat{\boldsymbol{\nu}}^{(1)}$ and $\hat{\boldsymbol{\nu}}^{(2)}$ to denote the MELEs of $\boldsymbol{\nu}$ based on the samples from the populations of interest and the samples from all the populations, respectively. Consider the data settings of Example 4.1 and 4.2. Recall that, for Example 4.1, $m+1=4$ and $\boldsymbol{\nu}=(\alpha_1,\, \boldsymbol{\beta}_1)^\intercal$, and for Example 4.2, $m+1=5$ and $\boldsymbol{\nu}=(\alpha_1,\, \alpha_2,\, \boldsymbol{\beta}_1,\, \boldsymbol{\beta}_2)^\intercal$. The bias, standard deviation (sd), and root mean squared error (rmse) of $\hat{\boldsymbol{\nu}}^{(1)}$ and $\hat{\boldsymbol{\nu}}^{(2)}$ under the two settings are shown in Tables 4.1 and 4.2, respectively. We see that under both settings, $\hat{\boldsymbol{\nu}}^{(2)}$ always has smaller absolute bias, sd and rmse, thereby a higher estimation accuracy. This observation agrees with our theoretical conclusion given in Theorem 4.1.

4.3.2 Comparison of testing power: $R_n^{(1)}$ versus $R_n^{(2)}$

Recall that $R_n^{(1)}$ and $R_n^{(2)}$ are DELR statistics based on partial data sets and full data sets, respectively. We now conduct simulations to compare the power of $R_n^{(1)}$ and $R_n^{(2)}$. The number of simulation repetitions is set to $10,000$.

We first let $m+1=4$ and consider a hypothesis test for $\boldsymbol{\beta}_1$. The DELR test can be conducted based on the first two samples ($R_n^{(1)}$) or based on all

Table 4.1: Comparison of the estimation accuracies of $\hat{\boldsymbol{\nu}}^{(1)}$ and $\hat{\boldsymbol{\nu}}^{(2)}$ under the setting of Example 4.1. $\boldsymbol{\beta}_1[1]$, $\boldsymbol{\beta}_1[2]$: the two components of $\boldsymbol{\beta}_1$.

	$\hat{\alpha}_1^{(1)}$	$\hat{\alpha}_1^{(2)}$	$\hat{\boldsymbol{\beta}}_1^{(1)}[1]$	$\hat{\boldsymbol{\beta}}_1^{(2)}[1]$	$\hat{\boldsymbol{\beta}}_1^{(1)}[2]$	$\hat{\boldsymbol{\beta}}_1^{(2)}[2]$
bias	0.44	0.392	-0.403	-0.362	0.477	0.422
sd	1.31	1.2	1.15	1.05	1.46	1.34
rmse	1.39	1.26	1.22	1.11	1.54	1.41

Table 4.2: Comparison of the estimation accuracies of $\hat{\boldsymbol{\nu}}^{(1)}$ and $\hat{\boldsymbol{\nu}}^{(2)}$ under the setting of Example 4.2.

	$\hat{\alpha}_1^{(1)}$	$\hat{\alpha}_1^{(2)}$	$\hat{\boldsymbol{\beta}}_1^{(1)}$	$\hat{\boldsymbol{\beta}}_1^{(2)}$	$\hat{\alpha}_2^{(1)}$	$\hat{\alpha}_2^{(2)}$	$\hat{\boldsymbol{\beta}}_2^{(1)}$	$\hat{\boldsymbol{\beta}}_2^{(2)}$
bias	0.071	0.0376	-0.627	-0.424	-0.0339	-0.0121	0.0364	0.0184
sd	0.346	0.254	2.51	2.09	0.246	0.218	0.242	0.218
rmse	0.354	0.257	2.59	2.13	0.248	0.218	0.244	0.219

four samples ($R_n^{(2)}$). The Wald test can also be conducted in two different ways. We denote them as Wald$^{(1)}$ and Wald$^{(2)}$ respectively.

We generated samples with sizes (60, 30, 40, 90), from gamma distributions under six parameter settings (Table 4.3).

Table 4.3: Gamma parameter values for power comparison of $R_n^{(1)}$ and $R_n^{(1)}$.

Common parameter settings						
$F_0 : \Gamma(2, 1)$, $F_2 : \Gamma(2.2, 1.25)$, $F_3 : \Gamma(2.1, 1.5)$						
Parameter settings for F_1						
	0	1	2	3	4	5
Case 1:	$\Gamma(4, 3)$	$\Gamma(5.3, 4.3)$	$\Gamma(6.3, 5.3)$	$\Gamma(7.1, 6.1)$	$\Gamma(8.3, 7.3)$	$\Gamma(10, 9)$
Case 2:	$\Gamma(2, 1)$	$\Gamma(2.45, 1.45)$	$\Gamma(2.9, 1.9)$	$\Gamma(3.3, 2.3)$	$\Gamma(3.8, 2.8)$	$\Gamma(5, 4)$

We consider two null hypotheses: one is $\boldsymbol{\beta}_1 = (-2, 2)^\intercal$ and the other is $\boldsymbol{\beta}_1 = \mathbf{0}$. The first hypothesis asks whether F_1 differs from F_0 in a specific

way; while the second one asks whether $F_0 = F_1$. By Theorems 4.2 and 4.3, $R_n^{(2)}$ is more powerful than $R_n^{(1)}$ for testing the first hypothesis, and two tests are asymptotically equally powerful for the second.

The simulated power curves are shown in Figure 4.1. It is seen that Wald tests are generally not as powerful. The results on $R_n^{(1)}$ and $R_n^{(2)}$ closely match the predictions of Theorems 4.2 and 4.3.

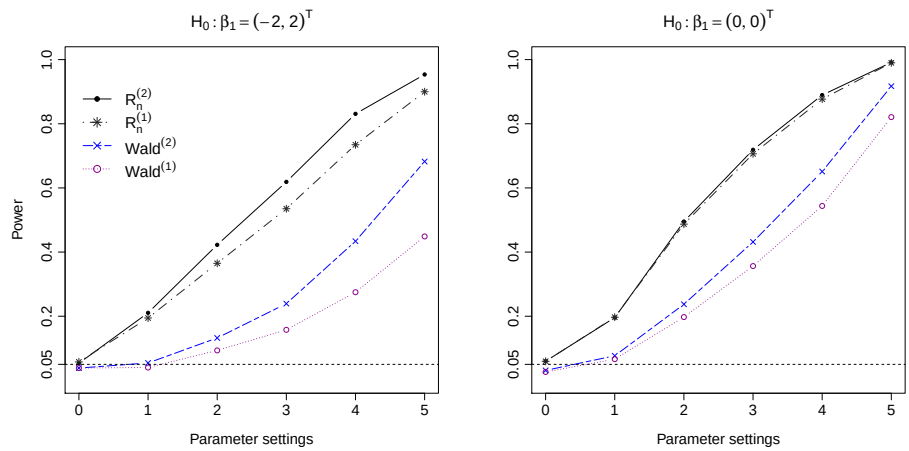

Figure 4.1: Power curves of $R_n^{(1)}$, $R_n^{(2)}$, Wald$^{(1)}$ and Wald$^{(2)}$. The parameter setting 0 corresponds to the null model and the settings 1–5 correspond to alternative models.

We conduct a second set of simulations where the data settings are taken from Example 4.3 and 4.4. For the setting of Example 4.3, we set the total sample size n to be 240, and for that of Example 4.4, we set $n = 600$. Under each data setting, we again calculate the powers of $R_n^{(1)}$, $R_n^{(2)}$, Wald$^{(1)}$ and Wald$^{(2)}$ under six different DRM parameters as shown in Table 4.4 and 4.5. The corresponding power curves of are shown in Figure 4.2. We see that our expectation again meets the observation from simulation.

Table 4.4: Parameter values for power comparison of $R_n^{(1)}$ and $R_n^{(1)}$ under the setting of Example 4.3.

Common parameter settings $F_0 : N(0, 1)$, $F_2 : N(-1, 2)$					
Parameter settings for F_1					
0	1	2	3	4	5
$N(1.5, 0.5)$	$N(1.57, 0.45)$	$N(1.58, 0.41)$	$N(1.6, 0.39)$	$N(1.62, 0.36)$	$N(1.64, 0.31)$

Table 4.5: Parameter values for power comparison of $R_n^{(1)}$ and $R_n^{(1)}$ under the setting of Example 4.4.

Common parameter settings $F_0 : \Gamma(2, 1)$, $F_4 : \Gamma(3.2, 0.7)$, $F_5 : \Gamma(1.6, 2.2)$						
Parameter settings for F_1, F_2 and F_3						
	0	1	2	3	4	5
F_1:	$\Gamma(2, 1)$	$\Gamma(2.3, 1.1)$	$\Gamma(2.3, 1)$	$\Gamma(2.4, 1)$	$\Gamma(2.4, 1)$	$\Gamma(2.4, 0.9)$
F_2:	$\Gamma(1.5, 2)$	$\Gamma(1.6, 2.1)$	$\Gamma(1.5, 2)$	$\Gamma(1.5, 2)$	$\Gamma(1.4, 1.9)$	$\Gamma(1.4, 1.9)$
F_3:	$\Gamma(1.5, 2)$	$\Gamma(1.9, 2.2)$	$\Gamma(1.9, 2.4)$	$\Gamma(1.9, 2.3)$	$\Gamma(2, 2.2)$	$\Gamma(2.1, 2.2)$

4.4 proofs

4.4.1 Theorem 4.1: Estimation accuracy comparison

We first introduce a useful notation for Schur complements that will be frequently encountered in the subsequent proofs. Let square matrix

$$M = \begin{pmatrix} A & B \\ C & D \end{pmatrix}$$

be nonsingular. We write $M/A = D - CA^{-1}B$ and call it the *Schur complement of M with respect to its upper–left block A*. Also, we write $M/D = A - BD^{-1}C$ and call it the *Schur complement of M with respect to its lower–right block D*.

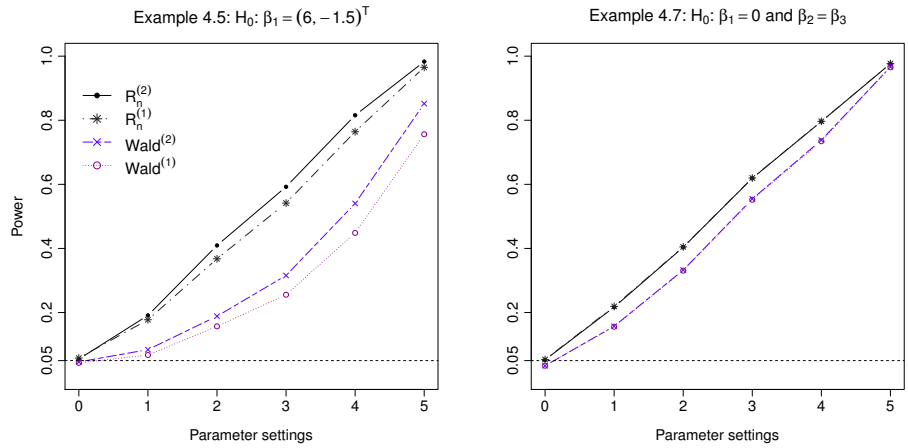

Figure 4.2: Power curves of $R_n^{(1)}$, $R_n^{(2)}$, Wald$^{(1)}$ and Wald$^{(2)}$ under the data settings of Example 4.3 and 4.4. The parameter setting 0 corresponds to the null model and the settings 1–5 correspond to alternative models.

Lemma 4.4. *Adopt the above partition for a symmetric matrix M of size $(s+t)\times(s+t)$. When $A>0$, $M\geq 0$ if and only if $M/A\geq 0$. When $D>0$, $M\geq 0$ if and only if $M/D\geq 0$.*

This is Theorem 1.4 of Zhan (2002). The outline of the proof is given below.

Proof. We prove the claimed result for the case that $A>0$. The proof for the case that $D>0$ is similar and so omitted.

Note that since matrix

$$N=\begin{pmatrix} I_m & -A^{-1}B \\ 0 & I_n \end{pmatrix}$$

is of full rank, any non–zero vector $\boldsymbol{u}$ of length $s+t$ can be written as $\boldsymbol{u}=N\boldsymbol{w}$

for some non–zero vector $\boldsymbol{w}$. Hence, we have

$$\begin{aligned}\boldsymbol{u}^\intercal M\boldsymbol{u} &= \boldsymbol{w}^\intercal N^T M N\boldsymbol{w}\\ &= \boldsymbol{w}^\intercal \begin{pmatrix} I_m & 0\\ -B^\intercal A^{-1} & I_n\end{pmatrix}\begin{pmatrix} A & B\\ B^\intercal & D\end{pmatrix}\begin{pmatrix} I_m & -A^{-1}B\\ 0 & I_n\end{pmatrix}\boldsymbol{w}\\ &= \boldsymbol{w}^\intercal \begin{pmatrix} A & 0\\ 0 & M/A\end{pmatrix}\boldsymbol{w}.\end{aligned}$$

Therefore, $M \geq 0$ if and only

$$\begin{pmatrix} A & 0\\ 0 & M/A\end{pmatrix} \geq 0.$$

Since $A > 0$, the latter condition is equivalent to $M/A \geq 0$. □

Recall that we defined $\boldsymbol{\theta}_k^\intercal = (\alpha_k,\, \boldsymbol{\beta}_k^\intercal)$. Denote the information matrix with respect to $(\boldsymbol{\theta}_1^\intercal,\, \ldots,\, \boldsymbol{\theta}_r^\intercal)^\intercal$ under the DRM based on the first $r+1$ samples as U_1, and that with respect to $(\boldsymbol{\theta}_1^\intercal,\, \ldots,\, \boldsymbol{\theta}_m^\intercal)^\intercal$ under the DRM based on all $m+1$ samples as U_2. Let $U_{2,c}$ be the lower–right $(m-r)(d+1) \times (m-r)(d+1)$ block of U_2. Recall that $\rho = \lim_{n\to\infty}(\sum_{k=0}^{r} n_k)/n$.

Lemma 4.5. *Under the conditions of Theorem 2.1, $U_2/U_{2,c} \geq \rho U_1$.*

Proof of Lemma 4.5. We prove the result for $m = r+1$, namely we compare the DRM based on the first $r+1 = m$ samples and that based on all the $m+1$ samples. The general result is true by mathematical induction.

Let $U_{2,a}$ be the upper–left $r(d+1) \times r(d+1)$ block, and $U_{2,b}$ be the upper–right $r(d+1) \times (m-r)(d+1)$ block, of U_2. Note that $U_2/U_{2,c} = U_{2,a} - U_{2,b}U_{2,c}^{-1}U_{2,b}^\intercal$, so to show the claimed result of $U_2/U_{2,c} \geq \rho\tilde{U}_1$, it suffices to show that

$$(U_{2,a} - \rho U_1) - U_{2,b}U_{2,c}^{-1}U_{2,b}^\intercal$$

is positive semidefinite. Notice that the above matrix is the Shur complement of

$$D = \begin{pmatrix} U_{2,a} - \rho U_1 & U_{2,b} \\ U_{2,b}^{\mathsf{T}} & U_{2,c} \end{pmatrix} = U_2 - \mathrm{diag}(\rho U_1,\, 0). \tag{4.4}$$

By Lemma 4.4, the positive semidefiniteness is implied by that of D.

We now show D is positive semidefinite. We first give useful algebraic expressions for U_2 and ρU_1. Notice that $(\boldsymbol{\theta}_1^{\mathsf{T}},\, \ldots,\, \boldsymbol{\theta}_m^{\mathsf{T}})$ is just permuted $\boldsymbol{\theta}^{\mathsf{T}} = (\boldsymbol{\alpha}^{\mathsf{T}}, \boldsymbol{\beta}^{\mathsf{T}})$, the information matrix (2.13) of which helps us to obtain algebraic expressions for U_1 and U_2. Recall $\boldsymbol{Q}^{\mathsf{T}}(x) = (1,\, \boldsymbol{q}^{\mathsf{T}}(x))$. For the DRM based on all the $m+1$ samples, we get

$$U_2 = \mathrm{E}_0\big\{H(\boldsymbol{\theta}^*, x) \otimes \{\boldsymbol{Q}(x)\boldsymbol{Q}^{\mathsf{T}}(x)\}\big\}.$$

For the DRM based on the first $r+1 = m$ samples, we find

$$\rho U_1 = \mathrm{E}_0\big\{H_r(\boldsymbol{\theta}^*, x) \otimes \{\boldsymbol{Q}(x)\boldsymbol{Q}^{\mathsf{T}}(x)\}\big\},$$

where $H_r(\boldsymbol{\theta}, x)$ is the H matrix defined in (2.12) based on the first $r+1$ samples:

$$H_r(\boldsymbol{\theta},\, x) = \mathrm{diag}\{\boldsymbol{h}_r(\boldsymbol{\theta},\, x)\} - \boldsymbol{h}_r(\boldsymbol{\theta},\, x)\boldsymbol{h}_r^{\mathsf{T}}(\boldsymbol{\theta},\, x)/s_r(\boldsymbol{\theta},\, x).$$

with $\boldsymbol{h}_r(\boldsymbol{\theta},\, x) = \big(\rho_1\varphi_1(\boldsymbol{\theta},\, x), \ldots, \rho_r\varphi_r(\boldsymbol{\theta},\, x)\big)^{\mathsf{T}}$ and $s(\boldsymbol{\theta},\, x) = \rho_0 + \sum_{k=1}^{r} \rho_k\varphi_k(\boldsymbol{\theta},\, x)$. As a reminder $\varphi_k(\boldsymbol{\theta},\, x) = \exp\{\alpha_k + \boldsymbol{\beta}_k^{\mathsf{T}}\boldsymbol{q}(x)\}$, $k = 0,\, 1,\, \ldots,\, m$. Put $s^*(x) = s(\boldsymbol{\theta}^*,\, x)$ for simplicity and similarly define $s_r^*(x)$, $\boldsymbol{h}_r^*(x)$ and $\varphi_k^*(x)$ for $k = 0,\, 1,\, \ldots,\, m$. Substituting the above expressions of U_2 and ρU_1 into the ex-

pression (4.4) of D and noticing $s^* - s_r^* = \rho_m \varphi_m^*$ by $m = r+1$, we get

$$
\begin{aligned}
D &= \mathrm{E}_0 \left\{ \begin{pmatrix} \rho_m \varphi_m^*(x) \boldsymbol{h}_r^*(x) \boldsymbol{h}_r^{*\mathsf{T}}(x) / \{s^*(x) s_r^*(x)\} & \rho_m \varphi_m^*(x) \boldsymbol{h}_r^*(x) / s^*(x) \\ \rho_m \varphi_m^*(x) \boldsymbol{h}_r^{*\mathsf{T}}(x) / s^*(x) & \rho_m \varphi_m^*(x) - \{\rho_m \varphi_m^*(x)\}^2 / s^*(x) \end{pmatrix} \otimes \right. \\
&\qquad \left. \{\boldsymbol{Q}(x) \boldsymbol{Q}^{\mathsf{T}}(x)\} \right\} \\
&= \rho_m \mathrm{E}_0 \left\{ \varphi_m^*(x) \begin{pmatrix} \boldsymbol{h}_r^*(x) \boldsymbol{h}_r^{*\mathsf{T}}(x) / \{s^*(x) s_r^*(x)\} & \boldsymbol{h}_r^*(x) / s^*(x) \\ \boldsymbol{h}_r^{*\mathsf{T}}(x) / s^*(x) & s_r^*(x) / s^*(x) \end{pmatrix} \otimes \{\boldsymbol{Q}(x) \boldsymbol{Q}^{\mathsf{T}}(x)\} \right\} \\
&= \rho_m \mathrm{E}_0 \big\{ \{\boldsymbol{w}(x) \boldsymbol{w}^{\mathsf{T}}(x)\} \otimes \{\boldsymbol{Q}(x) \boldsymbol{Q}^{\mathsf{T}}(x)\} \big\},
\end{aligned}
$$

with

$$
\boldsymbol{w}(x) = \sqrt{\varphi_m^*(x)} \left(\boldsymbol{h}_r^{*\mathsf{T}}(x),\, s_r^*(x) \right)^{\mathsf{T}} / \sqrt{s^*(x) s_r^*(x)}.
$$

Since D is the expectation of the Kronecker product of two squares of vectors, it is positive semidefinite. This completes the proof. □

Proof of Theorem 4.1. We now show that the asymptotic covariance matrix of the MELE based on the samples from the first $r+1$ populations is no smaller than that based on the sample from all the populations. We prove the result for the estimation of $(\boldsymbol{\theta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\theta}_r^{\mathsf{T}})^{\mathsf{T}}$. The claimed result for $\boldsymbol{\nu} = (\alpha_1, \ldots, \alpha_r, \boldsymbol{\beta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\beta}_r^{\mathsf{T}})^{\mathsf{T}}$ is then true because $\boldsymbol{\nu}$ is just a permutation of $(\boldsymbol{\theta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\theta}_r^{\mathsf{T}})^{\mathsf{T}}$.

The asymptotic covariance matrix, $\Sigma^{(1)}$, of the MELE for $(\boldsymbol{\theta}_1^{\mathsf{T}}, \ldots, \boldsymbol{\theta}_r^{\mathsf{T}})^{\mathsf{T}}$ under the DRM based on the first $r+1$ samples, by Theorem 2.5, is found to be

$$
\Sigma^{(1)} = U_1^{-1} - \rho W_1
$$

where U_1 is the information matrix under this DRM and

$$W_1 = \left\{ \left({\rho_0}^{-1} \mathbf{1}_r \mathbf{1}_r^\intercal + \operatorname{diag}\left({\rho_1}^{-1}, \ldots, {\rho_r}^{-1} \right) \right) \right\} \otimes \operatorname{diag}\{\tilde{\boldsymbol{e}}_1\},$$

with $\tilde{\boldsymbol{e}}_k$, in general, being the vector of length d with the kth entry being 1 and the others being 0s.

Similarly, the asymptotic covariance matrix of the MELE for $(\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_m^\intercal)^\intercal$ under the DRM based on all the $m+1$ samples is found to be

$$U_2^{-1} - W_2$$

where U_2 is the information matrix with respect to $(\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_m^\intercal)^\intercal$ under the current DRM and

$$W_2 = \left\{ \left({\rho_0}^{-1} \mathbf{1}_m \mathbf{1}_m^\intercal + \operatorname{diag}\left({\rho_1}^{-1}, \ldots, {\rho_m}^{-1} \right) \right) \right\} \otimes \operatorname{diag}\{\tilde{\boldsymbol{e}}_1\}.$$

Therefore, the asymptotic covaraince matrix, $\Sigma^{(2)}$ of the MELE for the sub-vector $(\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_r^\intercal)^\intercal$ under this DRM based on all the $m+1$ sampels, is just the upper–left $rd \times rd$ submatrix of $U_2^{-1} - W_2$. By the block matrix inversion formula (Lemma 3.3), the upper–left $rd \times rd$ submatrix of U_2^{-1} is $(U_2/U_{2,c})^{-1}$. Also notice that the correspond upper–left submatrix of W_2 is just W_1. Hence,

$$\Sigma^{(2)} = (U_2/U_{2,c})^{-1} - W_1.$$

By the above expressions of $\Sigma^{(1)}$ and $\Sigma^{(2)}$, to show the claimed result of $\Sigma^{(1)} \geq \rho\Sigma^{(2)}$, it suffices to show that

$$U_1^{-1} \geq \rho(U_2/U_{2,c})^{-1},$$

which is equivalent to

$$\rho U_1 \leq (U_2/U_{2,c})$$

because the matrices on both sides of the inequality are positive definite. The latter inequality is true by Lemma 4.5, so the claimed result of the theorem is true.

□

4.4.2 Theorem 4.2: Local power comparison in general

Lemma 4.6. *Let M and N be $(s+t)\times(s+t)$ positive definite matrices with partition*

$$M = \begin{pmatrix} \underset{s\times s}{M_a} & \underset{s\times t}{M_b} \\ \underset{t\times s}{M_c} & \underset{t\times t}{M_d} \end{pmatrix} \quad \textit{and} \quad N = \begin{pmatrix} \underset{s\times s}{N_a} & \underset{s\times t}{N_b} \\ \underset{t\times s}{N_c} & \underset{t\times t}{N_d} \end{pmatrix}.$$

If $M \geq N$, then $M/M_a \geq N/N_a$ and $M/M_d \geq N/N_d$.

Let U be the information matrix with respect to $\boldsymbol{\theta}$ based on all $m+1$ samples (corresponding to $R_n^{(2)}$), and $\tilde{U}$ be that with respect to $\boldsymbol{\nu} = (\alpha_1, \ldots, \alpha_r, \boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_r^\intercal)^\intercal$ based on the first $r+1$ samples (corresponding to $R_n^{(1)}$). Similar to the partition of U, we partition $\tilde{U}$ to $\tilde{U}_{\boldsymbol{\alpha\alpha}}$, $\tilde{U}_{\boldsymbol{\alpha\beta}}$, $\tilde{U}_{\boldsymbol{\beta\alpha}}$ and $\tilde{U}_{\boldsymbol{\beta\beta}}$, and similar to the definition $\Lambda = U/U_{\boldsymbol{\alpha\alpha}}$ given in Theorem 3.2, we define $\tilde{\Lambda} = \tilde{U}/\tilde{U}_{\boldsymbol{\alpha\alpha}}$. Moreover, we partition Λ into

$$\Lambda = \begin{pmatrix} \Lambda_a & \Lambda_b \\ \Lambda_b^\intercal & \Lambda_c \end{pmatrix}$$

with Λ_a being the upper–left $rd \times rd$ block of Λ.

Lemma 4.7. *Under the conditions of Theorem 2.1, $\Lambda/\Lambda_c \geq \rho\tilde{\Lambda}$.*

Lemma 4.8. *Let M be a $s \times s$ positive definite matrix and N be a $s \times s$ positive semidefinite matrix. Also let X and Y be $s \times t$ matrices, and suppose the column space of Y is contained in that of B. Then*

$$(X+Y)^\intercal (M+N)^{-1}(X+Y) \leq X^\intercal M^{-1} X + Y^\intercal N^\dagger Y$$

where $N^\dagger$ is the Moore–Penrose pseudoinverse of N.

The proofs of the above lemmas, being lengthy, are given after the proof of Theorem 4.2.

Proof of Theorem 4.2. We now show that under the local alternative model (4.2), the non–central parameter of the limiting distribution of the DELR statistic based on the samples from the first $r+1$ populations in general is not greater than that based on the samples from all the populations.

We have defined two DELR statistics $R_n^{(1)}$ and $R_n^{(2)}$ which are constructed using the samples from only the first $r+1$ populations $F_0, \cdots, F_r$, and the samples from all the populations, respectively. Recall that the null hypothesis of (4.1) contains a constraint $\boldsymbol{g}(\boldsymbol{\zeta}) = \mathbf{0}$ with $\boldsymbol{\zeta}^\intercal = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_r^\intercal)$ related only to populations $F_0, \cdots, F_r$. By Theorem 3.2, under the $\{G_k\}$ defined by the local alternative model (4.2), $R_n^{(1)}$ and $R_n^{(2)}$ both have non–central chi–square limiting distributions of q degrees of freedom, but with different non–central parameters δ_1^2 and δ_2^2. We also know that

$$\delta_1^2 = \rho \tilde{\boldsymbol{\eta}}^\intercal \left\{ \tilde{\Lambda} - \tilde{\Lambda} J (J^\intercal \tilde{\Lambda} J)^{-1} J^\intercal \tilde{\Lambda} \right\} \tilde{\boldsymbol{\eta}},$$

where $\tilde{\boldsymbol{\eta}} = (\rho_1^{-1/2} \boldsymbol{c}_1^\intercal, \ldots, \rho_r^{-1/2} \boldsymbol{c}_r^\intercal)$ is a subvector of $\boldsymbol{\eta}$ defined in Theorem 3.2. Moreover, noticing that for the local alternative (4.2) under investigation, $\boldsymbol{\eta}^\intercal = (\tilde{\boldsymbol{\eta}}^\intercal, 0_{m-r}^\intercal)$, we get

$$\delta_2^2 = \tilde{\boldsymbol{\eta}}^\intercal \left\{ (\Lambda/\Lambda_c) - (\Lambda/\Lambda_c) J (J^\intercal (\Lambda/\Lambda_c) J)^{-1} J^\intercal (\Lambda/\Lambda_c) \right\} \tilde{\boldsymbol{\eta}}.$$

by applying Theorem 3.2 and the quadratic form decomposition formula

given in Lemma 3.4. Hence, to show the claimed result $\delta_2^2 \geq \delta_1^2$, it suffices to show that

$$(\Lambda/\Lambda_c) - (\Lambda/\Lambda_c)J(J^\intercal(\Lambda/\Lambda_c)J)^{-1}J^\intercal(\Lambda/\Lambda_c) \geq \rho\left\{\tilde{\Lambda} - \tilde{\Lambda}J(J^\intercal\tilde{\Lambda}J)^{-1}J^\intercal\tilde{\Lambda}\right\}. \tag{4.5}$$

Define $M = \Lambda/\Lambda_c - \rho\tilde{\Lambda}$. In Lemma 4.8, let $A = \rho J^\intercal\tilde{\Lambda}J$, $B = J^\intercal MJ$, $X = \rho J^\intercal\tilde{\Lambda}$, $Y = J^\intercal M$. Then $A + B = J^\intercal(\Lambda/\Lambda_c)J$ and $X + Y = J^\intercal(\Lambda/\Lambda_c)$. Matrix A is positive definite because $\tilde{\Lambda}$ is positive definite and J is of full rank. B is positive semidefinite because M is positive semidefinite by Lemma 4.7. Moreover, it is easily seen that the column space of Y is the same as that of B. Hence the conditions of Lemma 4.8 are satisfied, and we have

$$(\Lambda/\Lambda_c)J(J^\intercal(\Lambda/\Lambda_c)J)^{-1}J^\intercal(\Lambda/\Lambda_c) \leq \rho\tilde{\Lambda}J(J^\intercal\tilde{\Lambda}J)^{-1}J^\intercal\tilde{\Lambda} + MJ(J^\intercal MJ)^\dagger J^\intercal M.$$

The above inequality and $\Lambda/\Lambda_c = \rho\tilde{\Lambda} + M$ imply that

$$\begin{aligned}&(\Lambda/\Lambda_c) - (\Lambda/\Lambda_c)J(J^\intercal(\Lambda/\Lambda_c)J)^{-1}J^\intercal(\Lambda/\Lambda_c)\\ \geq&\rho\{\tilde{\Lambda} - \tilde{\Lambda}J(J^\intercal\tilde{\Lambda}J)^{-1}J^\intercal\tilde{\Lambda}\} + \{M - MJ(J^\intercal MJ)^\dagger J^\intercal M\}.\end{aligned}$$

The term $M - MJ(J^\intercal MJ)^\dagger J^\intercal M$ is positive semidefinite because

$$M - MJ(J^\intercal MJ)^\dagger J^\intercal M = M^{1/2}\{I - M^{1/2}J(J^\intercal MJ)^\dagger J^\intercal M^{1/2}\}M^{1/2},$$

and $I - M^{1/2}J(J^\intercal MJ)^\dagger J^\intercal M^{1/2}$ is easily verified to be idempotent, hence positive semidefinite. Therefore inequality (4.5) holds and the claimed result is true. □

Proof of Lemma 4.6. By Lemma 3.3, the upper–left $s\times s$ submatrices of M^{-1} and N^{-1} are $(M/M_d)^{-1}$ and $(N/N_d)^{-1}$ respectively, and the lower–right $t\times t$ submatrices of M^{-1} and N^{-1} are $(M/M_a)^{-1}$ and $(N/N_a)^{-1}$ respectively.

Since both M and N are positive definite and $M \geq N$, we have $M^{-1} \leq$

N^{-1}. The corrsponding submatrices of M^{-1} and N^{-1} satisfy the same inequality, so we have $(M/M_d)^{-1} \leq (N/N_d)^{-1}$ and $(M/M_a)^{-1} \leq (N/N_a)^{-1}$. By the positive definiteness of these Schur complements, we can take inverses on both sides and reverse the directions of the above two inequalities, which gives us the claimed result.

□

To prove Lemma 4.7, partition $U_{\boldsymbol{\alpha\alpha}}$, $U_{\boldsymbol{\alpha\beta}}$ and $U_{\boldsymbol{\beta\beta}}$ as follows:

$$U_{\boldsymbol{\alpha\alpha}} = \begin{pmatrix} U_{\boldsymbol{\alpha\alpha},a} & U_{\boldsymbol{\alpha\alpha},b} \\ U_{\boldsymbol{\alpha\alpha},b}^{\intercal} & U_{\boldsymbol{\alpha\alpha},c} \end{pmatrix}, \; U_{\boldsymbol{\alpha\beta}} = \begin{pmatrix} U_{\boldsymbol{\alpha\beta},a} & U_{\boldsymbol{\alpha\beta},b} \\ U_{\boldsymbol{\alpha\beta},c} & U_{\boldsymbol{\alpha\beta},d} \end{pmatrix}, \; U_{\boldsymbol{\beta\beta}} = \begin{pmatrix} U_{\boldsymbol{\beta\beta},a} & U_{\boldsymbol{\beta\beta},b} \\ U_{\boldsymbol{\beta\beta},b}^{\intercal} & U_{\boldsymbol{\beta\beta},c} \end{pmatrix},$$

where $U_{\boldsymbol{\alpha\alpha},a}$, $U_{\boldsymbol{\alpha\beta},a}$ and $U_{\boldsymbol{\beta\beta},a}$ are the corrsponding upper–left $r \times r$, $r \times rd$ and $rd \times rd$ blocks.

We also introduce an important property of the Schur complement. Let

$$M = \begin{pmatrix} \underset{s\times s}{A} & \underset{s\times t}{B} \\ \underset{t\times s}{C} & \underset{t\times t}{D} \end{pmatrix} \quad \text{and} \quad D = \begin{pmatrix} \underset{u\times u}{E} & \underset{u\times v}{F} \\ \underset{v\times u}{G} & \underset{v\times v}{H} \end{pmatrix},$$

where $u + v = t$. Suppose M, A and D are nonsingular. By Theorem 1.4 of Zhang (2005), the lower–right $u \times u$ block of M/H is just D/H, and

$$M/D = (M/H)/(D/H). \tag{4.6}$$

The above equality is known as the *quotient formula*. Similar quotient formula holds for M/A.

Proof of Lemma 4.7. We first give an algebraic expression for Λ/Λ_c. Recall the definition $\Lambda = U_{\boldsymbol{\beta\beta}} - U_{\boldsymbol{\beta\alpha}} U_{\boldsymbol{\alpha\alpha}}^{-1} U_{\boldsymbol{\alpha\beta}}$, so

$$\Lambda = \Psi / U_{\boldsymbol{\alpha\alpha}},$$

where

$$\Psi = \begin{pmatrix} U_{\boldsymbol{\beta\beta}} & U_{\boldsymbol{\beta\alpha}} \\ U_{\boldsymbol{\alpha\beta}} & U_{\boldsymbol{\alpha\alpha}} \end{pmatrix}.$$

Let Ψ_1 be the lower–right $\{(m-r)d+m\} \times \{(m-r)d+m\}$ block of Ψ. Then Λ_c, the lower–right $(m-r)d \times (m-r)d$ block of $\Lambda = \Psi/U_{\boldsymbol{\alpha\alpha}}$, satisfies

$$\Lambda_c = \Psi_1/U_{\boldsymbol{\alpha\alpha}}.$$

Therefore

$$\Lambda/\Lambda_c = (\Psi/U_{\boldsymbol{\alpha\alpha}})/(\Psi_1/U_{\boldsymbol{\alpha\alpha}}) = \Psi/\Psi_1,$$

where the second equality above is by quotient formula (4.6).

It is easily seen that $\Psi/\Psi_1 = \Omega/\Omega_1$, where

$$\Omega = \left(\begin{array}{cc:cc} U_{\boldsymbol{\beta\beta},a} & U_{\boldsymbol{\beta\alpha},a} & U_{\boldsymbol{\beta\beta},b} & U_{\boldsymbol{\beta\alpha},b} \\ U_{\boldsymbol{\alpha\beta},a} & U_{\boldsymbol{\alpha\alpha},a} & U_{\boldsymbol{\alpha\beta},b} & U_{\boldsymbol{\alpha\alpha},b} \\ \hdashline U^{\intercal}_{\boldsymbol{\beta\beta},b} & U_{\boldsymbol{\beta\alpha},c} & U_{\boldsymbol{\beta\beta},c} & U_{\boldsymbol{\beta\alpha},d} \\ U_{\boldsymbol{\alpha\beta},c} & U^{\intercal}_{\boldsymbol{\alpha\alpha},b} & U_{\boldsymbol{\alpha\beta},d} & U_{\boldsymbol{\alpha\alpha},c} \end{array}\right)$$

and Ω_1 is the lower–right block of Ω with the same size as that of Ψ_1. Thus we get

$$\Lambda/\Lambda_c = \Psi/\Psi_1 = \Omega/\Omega_1.$$

Let Ω_2 be the lower–right $(m-r)(d+1) \times (m-r)(d+1)$ block of Ω_1. Matrix Ω_1/Ω_2 is just the lower–right $r \times r$ block of Ω/Ω_2, and $\Omega/\Omega_1 = (\Omega/\Omega_2)/(\Omega_1/\Omega_2)$ by quotient formula (4.6). Hence, we finally get

$$\Lambda/\Lambda_c = \Omega/\Omega_1 = (\Omega/\Omega_2)/(\Omega_1/\Omega_2).$$

The above identity implies that our cliam of $\Lambda/\Lambda_c \geq \rho\tilde{\Lambda}$ is equivalent to

$$(\Omega/\Omega_2)/(\Omega_1/\Omega_2) \geq \rho\tilde{\Lambda}.$$

Further notice that $\tilde{\Lambda} = \check{U}/\tilde{U}_{\boldsymbol{\alpha\alpha}}$, where

$$\check{U} = \begin{pmatrix} \tilde{U}_{\boldsymbol{\beta\beta}} & \tilde{U}_{\boldsymbol{\beta\alpha}} \\ \tilde{U}_{\boldsymbol{\alpha\beta}} & \tilde{U}_{\boldsymbol{\alpha\alpha}} \end{pmatrix},$$

so, the above inequality is equivalent to

$$(\Omega/\Omega_2)/(\Omega_1/\Omega_2) \geq \rho(\check{U}/\tilde{U}_{\boldsymbol{\alpha\alpha}}). \tag{4.7}$$

In the last step, we prove the above inequality (4.7). Recall Lemma 4.6 that if matrices M and N are both positive definite and $M \geq N$, then the corresponding Schur complements satisfy the same inequality. Note that both Ω/Ω_2 and $\check{U}$ are positive definite and the terms $(\Omega/\Omega_2)/(\Omega_1/\Omega_2)$ and $\check{U}/\tilde{U}_{\boldsymbol{\alpha\alpha}}$ in (4.7) are corresponding Schur complements, so, by Lemma 4.6, to show (4.7), it is enough to show that

$$\Omega/\Omega_2 \geq \rho\check{U}.$$

Note that parameter $\boldsymbol{\phi}^\intercal = (\boldsymbol{\beta}_1^\intercal, \ldots, \boldsymbol{\beta}_r^\intercal, \alpha_1, \ldots, \alpha_r, \boldsymbol{\beta}_{r+1}^\intercal, \ldots, \boldsymbol{\beta}_m^\intercal, \alpha_{r+1}, \ldots, \alpha_m)$ is just permuted $(\boldsymbol{\theta}_1^\intercal, \ldots, \boldsymbol{\theta}_m^\intercal)$, so the conculsion of Lemma 4.5 also applies to the information matrix with respect to $\boldsymbol{\phi}$. The information matrix with respect to $\boldsymbol{\phi}$ for $R_n^{(2)}$ is just Ω, and its lower–right $(m-r)(d+1)\times(m-r)(d+1)$ block is Ω_2. For $R_n^{(1)}$, the infromation matrix is just $\check{U}$. Thus by Lemma 4.5, we have $\Omega/\Omega_2 \geq \rho\check{U}$. The proof is complete. □

Proof of Lemma 4.8. Notice that

$$\begin{pmatrix} M+N & X+Y \\ (X+Y)^\intercal & X^\intercal M^{-1}X + Y^\intercal N^\dagger Y \end{pmatrix} = \begin{pmatrix} M & X \\ X^\intercal & X^\intercal M^{-1}X \end{pmatrix} + \begin{pmatrix} N & Y \\ Y^\intercal & Y^\intercal N^\dagger Y \end{pmatrix}.$$

The first matrix on the RHS is positive semidefinite by Lemma 4.4 since $M > 0$ and the Schur complement of this matrix with respect to M is 0. By a generalized version of Lemma 4.4 (Zhang (2005)), observing that $N \geq 0$, Y is in the column space of N and the Schur complement of the second matrix on the RHS with respect to N is 0, we have that the second matrix on the RHS is also positive semidefinite. Therefore the matrix on the LHS is positive semidefinite. Also note that $M + N > 0$. Hence, by Lemma 4.4, the Schur complement of the LHS with respect to its upper–left block $M + N$,

$$X^\intercal M^{-1} X + Y^\intercal N^\dagger Y - (X + Y)^\intercal (M + N)^{-1} (X + Y),$$

must also be positive semidefinite. The claimed result then follows. □

4.4.3 Theorem 4.3: Local power comparison in a special case

In this section, we show that for the special hypothesis (4.3), under the local alternative model (4.2), the non–central parameters of the limiting distributions of $R_n^{(1)}$ and $R_n^{(2)}$ are equal.

Recall that δ_1^2 and δ_2^2 are the non–central parameters of the limiting distributions of $R_n^{(1)}$ and $R_n^{(2)}$ under the local alternative model (4.2). Under the special hypothesis (4.3) in this theorem, Jacobian J has a special structure. We use this structure to show the equality of δ_1^2 and δ_2^2. Without loss of generality, we assume that the indices in $\mathbb{S}_k$, $k = 1, \ldots, K$, are in natural order.

By Theorem 3.2, we have

$$\delta_1^2 = \rho \tilde{\boldsymbol{\eta}}^\intercal \big\{ \tilde{\Lambda} - \tilde{\Lambda} J_1 (J_1^\intercal \tilde{\Lambda} J_1)^{-1} J_1^\intercal \tilde{\Lambda} \big\} \tilde{\boldsymbol{\eta}}$$

where $\tilde{\boldsymbol{\eta}}$ is defined in the proof of Theorem 4.2 and

$$J_1 = \big(0_{(K-1)\times s_1},\, \mathrm{diag}(\mathbf{1}_{s_2}^{\intercal},\, \ldots,\, \mathbf{1}_{s_K}^{\intercal})\big)^{\intercal} \otimes I_d.$$

As the proof of Theorem 4.2, we find

$$\delta_2^2 = \tilde{\boldsymbol{\eta}}^{\intercal} A \tilde{\boldsymbol{\eta}},$$

where A is the upper–left $rd \times rd$ block of $\Lambda - \Lambda J_2 (J_2^{\intercal} \Lambda J_2)^{-1} J_2^{\intercal} \Lambda$ with $J_2 = \mathrm{diag}(J_1,\, I_{(m-r)d})$. Therefore, to show the claimed $\delta_1^2 = \delta_2^2$, it suffices to show that

$$\rho\left\{\tilde{\Lambda} - \tilde{\Lambda} J_1 (J_1^{\intercal} \tilde{\Lambda} J_1)^{-1} J_1^{\intercal} \tilde{\Lambda}\right\} = A \tag{4.8}$$

We first simplify the LHS of (4.8). Let $\boldsymbol{\varrho}_k$, $k = 1, \ldots, K$, be the vector consisting of ρ_i, $i \in \mathbb{S}_k$, and $\varsigma_k = \sum_{i \in \mathbb{S}_k} \rho_i$. We observe that

$$J_1^{\intercal} \tilde{\Lambda} = (J_1^{\intercal} \tilde{\Lambda} J_1) B$$

where

$$B = \big(-(\varsigma_1 + \rho_0)^{-1} \mathbf{1}_{K-1} \otimes \boldsymbol{\varrho}_1^{\intercal},\ \mathrm{diag}(\varsigma_2^{-1} \boldsymbol{\varrho}_2^{\intercal},\, \ldots,\, \varsigma_K^{-1} \boldsymbol{\varrho}_K^{\intercal})\big) \otimes I_d.$$

Thus, we have $(J_1^{\intercal} \tilde{\Lambda} J_1)^{-1} J_1^{\intercal} \tilde{\Lambda} = B$ and

$$\rho\left\{\tilde{\Lambda} - \tilde{\Lambda} J_1 (J_1^{\intercal} \tilde{\Lambda} J_1)^{-1} J_1^{\intercal} \tilde{\Lambda}\right\} = \rho\left\{\tilde{\Lambda} - \tilde{\Lambda} J_1 B\right\}. \tag{4.9}$$

We then simplify $\Lambda - \Lambda J_2 (J_2^{\intercal} \Lambda J_2)^{-1} J_2^{\intercal} \Lambda$ to get another expression of A. We find that

$$J_2^{\intercal} \Lambda = J_2^{\intercal} \Lambda J_2 \begin{pmatrix} B & 0 \\ C & I_{(m-r)d} \end{pmatrix},$$

where $C = \big(- \{\varsigma_1 + \rho_0\}^{-1}(\mathbf{1}_{m-r}\boldsymbol{\varrho}_1^\intercal \otimes I_d),\ 0\big)$. Thus,

$$\Lambda - \Lambda J_2(J_2^\intercal \Lambda J_2)^{-1} J_2^\intercal \Lambda = \Lambda - \Lambda J_2 \begin{pmatrix} B & 0 \\ C & I_{(m-r)d} \end{pmatrix}.$$

Recall that A is the upper–left $rd \times rd$ block of $\Lambda - \Lambda J_2(J_2^\intercal \Lambda J_2)^{-1} J_2^\intercal \Lambda$. Hence, the above identity gives another expression of A as

$$A = \Lambda_a - \Lambda_a J_1 B - \Lambda_b C, \tag{4.10}$$

where Λ_a and Λ_b are respectively the $rd \times rd$ and $rd \times (m-r)d$ blocks of Λ.

Finally, the proof is completed by showing the RHS expressions of (4.9) and (4.10) are equal. This is done by linking Λ–matrices to the information matrix (2.13), and applying the block matrix inversion formaula given in Lemma 3.3 and the quadratic form deceomposition formula given in Lemma 3.4.

Chapter 5

Empirical Likelihood Inference under the DRM Based on Multiple Type I Censored Samples

As noted in Chapter 1, it is desirable to make smart strength test plans to maximize the scientific value of each piece of lumber, and one such plan is to collect Type I right–censored strength samples such that some lumber can be used for multiple strength tests. However, the theory of EL inference under the DRM for complete samples does not carry automatically to the case of Type I censored samples due to the substantially more complicated analytical form of the EL in the latter case. This chapter creates a powerful EL inference framework for the latter case. In particular, we (1) show that the maximization of the EL can be reduced to the maximization of a concave function, which we call the *dual partial EL*, (2) give the asymptotic properties of the dual partial EL, and (3) study the properties of the EL ratio test for hypothesis about the DRM parameter $\boldsymbol{\beta}$. We argue that the inference framework established in the chapter can potentially be used to extend any EL inference result that is available for multiple complete samples under the DRM to the case of multiple Type I censored samples.

The chapter is organized as follows. Section 5.1 introduces the concept of Type I censoring and gives the definition of the corresponding EL for a single sample. Section 5.2 defines the EL function under the DRM based

on multiple Type I censored samples, and the the associated maximum EL estimator. Section 5.3 explores the relationship between the EL and the partial EL and reduces the constrained maximization problem for EL to a convex maximization problem. Section 5.4 gives an interpretation for the PEL and Section 5.5 study the asymptotic properties of the MELE for the DRM parameters. The theory of EL ratio test based on Type I censored samples is established in Section 5.6, followed by a short discussion in Section 5.7 about other inference tasks under this framework. The proofs are given in Section 5.8.

5.1 Type I censored single random samples and the corresponding EL

Let $\{x_i\}_{i=1}^n$ be an independent sample from a population with CDF $F(x)$. Let $\{C_i\}_{i=1}^n$ be a set of iid random variables from another population, and $\{c_i\}_{i=1}^n$ be a set of corresponding realizations. The sample $\{x_i\}_{i=1}^n$ is said to be *right–censored* if we only observe value $z_i = \min(x_i, c_i)$ and the indicator $\mathbb{1}(x_i \leq c_i)$, which shows whether an observation is censored or not. When $\mathbb{1}(x_i \leq c_i) = 1$, the ith observation is said to be *uncensored*, otherwise, *censored*. If the underlying distribution of C_i is a point mass at a given constant c, that is when $z_i = \min(x_i, c)$, then the sample is said to be *Type I right–censored*. Type I right censoring usually arises in reliability engineering and medical studies when a experiment stops at a prespecified value c and the values of the sample points that are larger than c are unknown to observers.

Similarly we can define other kinds of Type I censoring, for example,

Type I left–censored sample: we observe $z_i = \max(x_i, c)$ and the censoring indicator $\mathbb{1}(x_i \geq c)$ for some given constant c;

Type I left and right–censored sample: we observe the censoring indicators $\mathbb{1}(x_i \in [c_1, c_2])$ and $\mathbb{1}(x_i < c_1)$ for given constants $c_1 < c_2$, and $z_i =$

$$x_i \mathbb{1}(x_i \in [c_1,\, c_2]) + c_1 \mathbb{1}(x_i < c_1) + c_2 \mathbb{1}(x_i > c_2).$$

This chapter focuses on the above three kinds of Type I censored samples, and in the sequel, we simply refer to them as "Type I censored samples". Clearly, the Type I right–censored sample and Type I left–censored sample are both special cases of Type I left and right–censored sample with $c_1 = -\infty$ and $c_2 = \infty$ respectively. We focus on studying the most general one of the three: the Type I left and right–censored sample.

Let $\tilde{n}$ be the number of uncensored observations, i.e. $\tilde{n} = \sum_{j=1}^{n} \mathbb{1}(x_i \in [c_1,\, c_2])$, and denote the uncensored observations as $\tilde{x}_i$, $i = 1, \ldots, \tilde{n}$. Let $\check{n}$ be the number of left–censored observations, i.e. $\check{n} = \sum_{j=1}^{n} \mathbb{1}(x_i < c_1)$. Put

$$\begin{aligned} \varsigma &= \Pr(X \in [c_1,\, c_2]) = F(c_2) - F(c_1^-), \\ \iota &= \Pr(X < c_1) = F(c_1^-). \end{aligned}$$

We have $\Pr(X > c_2) = 1 - \iota - \varsigma$. Recall that we defined $dF(x) = F(x) - F(x^-)$ in Section 2.1. The EL based on a Type I censored sample is given to be

$$\begin{aligned} L_n(F) &= \prod_{i=1}^{n} \iota^{\mathbb{1}(x_i < c_1)} \{dF(x_i)\}^{\mathbb{1}(x_i \in [c_1,\, c_2])} (1 - \iota - \varsigma)^{1 - \mathbb{1}(x_i < c_1) - \mathbb{1}(x_i \in [c_1,\, c_2])} \\ &= \iota^{\check{n}} \Big\{ \prod_{i=1}^{\tilde{n}} dF(\tilde{x}_i) \Big\} (1 - \iota - \varsigma)^{n - \check{n} - \tilde{n}}. \end{aligned}$$

5.2 EL for multiple Type I censored samples under the DRM

Suppose we have $m + 1$ independent Type I censored samples

$$\{x_{kj} : j = 1,\, 2, \ldots,\, n_k\}_{k=0}^{m}$$

from populations $\{F_k(x)\}$ of the same support $\mathcal{S}$, which satisfy the DRM assumption (2.1). Our question is, based on such censored samples, how

should we estimate the DRM parameter and test hypotheses about it, an important inference task in our long term monitoring program for lumber strength as noted in Chapter 1.

As for complete data, we consider EL inference as it seems to be an effective and most natural inference tool under the semiparametric DRM. Let $c_{k,1}$ and $c_{k,2}$, $k = 0, 1, \ldots, m$, be the left and right censoring cutting points for the kth Type I censored sample, respectively. Let $\mathcal{S}_k = [c_{k,1}, c_{k,2}] \cap \mathcal{S}$. The set $\mathcal{S}_k$ is the support of the distribution of the uncensored observations in the kth sample. Let $\tilde{n}_k$ be the number of the uncensored observations in the kth sample and denote the uncensored observations as $\{\tilde{x}_{kj} : j = 1, \ldots, \tilde{n}_k\}$. Let $\check{n}_k$ be the number of left–censored observations in the kth sample. Define

$$\varsigma_k = \Pr(X_{k1} \in \mathcal{S}_k) = F_k(c_{k,2}) - F_k(c_{k,1}^-) \quad \text{and} \quad \iota_k = F_k(c_{k,1}^-).$$

We will always assume that, for all $k = 0, 1, \ldots, m$, $\varsigma_k > 0$ *and* $\mathcal{S}_k \subseteq \mathcal{S}_0$. When the samples are not both left and right–censored, we can always choose a baseline such that this assumption is satisfied. As in the case of a single sample, the EL of the $\{F_k\}$ based on these Type I censored samples is defined to be

$$L_n(F_0, \ldots, F_m) = \Big\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} dF_k(\tilde{x}_{kj}) \Big\} \Big\{ \prod_{k=0}^{m} \iota_k^{\check{n}_k} \Big\} \Big\{ \prod_{k=0}^{m} \{1 - \iota_k - \varsigma_k\}^{n_k - \check{n}_k - \tilde{n}_k} \Big\}.$$

The first factor on RHS in the above definition is the contribution of the uncensored observations to the likelihood; the second factor, the contribution of the left–censored observations; and the third, the contribution of the right–censored observations. When the $\{F_k\}$ satisfy the DRM assumption, the

above likelihood function can be further written as

$$L_n(F_0, \ldots, F_m) = \Big\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} dF_0(\tilde{x}_{kj}) \Big\} \Big\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} \exp\{\alpha_k + \boldsymbol{\beta}_k^{\intercal} \boldsymbol{q}(\tilde{x}_{kj})\} \Big\}$$
$$\Big\{ \prod_{k=0}^{m} \iota_k^{\check{n}_k} \{1 - \iota_k - \varsigma_k\}^{n_k - \check{n}_k - \tilde{n}_k} \Big\}, \tag{5.1}$$

where α_k and $\boldsymbol{\beta}_k$ satisfy

$$\int_{x \in \mathcal{S}_k} \exp\{\alpha_k + \boldsymbol{\beta}_k^{\intercal} \boldsymbol{q}(x)\} dF_0(x) = \varsigma_k.$$

Put $p_{kj} = dF_0(\tilde{x}_{kj})$. Define

$$\boldsymbol{p} = \{p_{kj} : j = 1, \ldots, \tilde{n}_k\}_{k=0}^{m}, \quad \boldsymbol{\iota} = \{\iota_k\}_{k=0}^{m}, \quad \text{and} \quad \boldsymbol{\varsigma} = \{\varsigma_k\}_{k=0}^{m}.$$

We see from (5.1) that, under the DRM, the EL is a function of $\boldsymbol{\theta}$, $\boldsymbol{p}$, $\boldsymbol{\iota}$ and $\boldsymbol{\varsigma}$. We therefore write the EL as $L_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\iota}, \boldsymbol{\varsigma})$. Recall that we have defined $\alpha_0 = 0$ and $\boldsymbol{\beta}_0 = \mathbf{0}$. Let

$$(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}})$$
$$= \underset{\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\iota}, \boldsymbol{\varsigma}}{\operatorname{argmax}} \Big\{ L_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\iota}, \boldsymbol{\varsigma}) : \sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \exp\{\alpha_r + \boldsymbol{\beta}_r^{\intercal} \boldsymbol{q}(\tilde{x}_{kj})\} p_{kj} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_r) = \varsigma_r,$$
$$p_{kj} \geq 0, \ 0 < \varsigma_r \leq 1, \ 0 \leq \iota_r \leq 1 - \varsigma_r, \ r = 0, \ldots, m \Big\}. \tag{5.2}$$

We call $\hat{\boldsymbol{\theta}}$ the MELE of the DRM parameter $\boldsymbol{\theta}$, and $(\hat{\boldsymbol{p}}, \hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}})$ the MELE of the baseline distribution F_0. The next section addresses how to calculate these MELEs.

5.3 Calculating the MELE

The constrained maximization (5.2) of the EL appears to be a complicated problem. We now show it can however be reduced to a simple concave maximization problem.

5.3.1 Partial EL and its relation to EL

The EL (5.1) of the multiple Type I censored samples can be factorized as

$$L_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma}) = PL_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma}) \cdot \mathbb{L}_n(\boldsymbol{\iota},\, \boldsymbol{\varsigma}), \tag{5.3}$$

where

$$PL_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma}) = \left\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} \frac{p_{kj}}{\varsigma_0} \right\} \left\{ \prod_{k=1}^{m} \prod_{j=1}^{\tilde{n}_k} \frac{\varsigma_0}{\varsigma_k} \exp\left\{ \alpha_k + \boldsymbol{\beta}_k^{\intercal} \boldsymbol{q}(\tilde{x}_{kj}) \right\} \right\},$$

$$\mathbb{L}_n(\boldsymbol{\iota},\, \boldsymbol{\varsigma}) = \Big\{ \prod_{k=0}^{m} \iota_k^{\check{n}_k} \varsigma_k^{\tilde{n}_k} \{1 - \iota_k - \varsigma_k\}^{n_k - \check{n}_k - \tilde{n}_k} \Big\}.$$

We call $PL_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma})$ the *partial empirical likelihood* (PEL) function. Our next proposition indicates that the EL attains its maximum under the corresponding constraints given in (5.2) when both the PEL under the same constraints and the $\mathbb{L}_n(\boldsymbol{\iota},\, \boldsymbol{\varsigma})$ attain their maxima independently.

Proposition 5.1. *Under the constraint*

$$\sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \exp\{\alpha_r + \boldsymbol{\beta}_r^{\intercal} \boldsymbol{q}(\tilde{x}_{kj})\} p_{kj} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_r) = \varsigma_r, \tag{5.4}$$

for $r = 0,\, 1,\, \ldots,\, m$, $\sup_{\boldsymbol{\theta},\boldsymbol{p}} PL_n(\boldsymbol{\theta},\, \boldsymbol{p}, \boldsymbol{\varsigma})$ *does not depend on* $\boldsymbol{\varsigma}$.

Proof of Proposition 5.1. With reorganizations of terms, the PEL can be

written as

$$
\begin{aligned}
&PL_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma}) \\
&= \left\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} \frac{p_{kj}}{\varsigma_0} \right\} \left\{ \prod_{k=1}^{m} \prod_{j=1}^{\tilde{n}_k} \exp\left\{ (\alpha_k + \log \varsigma_0 - \log \varsigma_k) + \boldsymbol{\beta}_k^{\intercal} \boldsymbol{q}(\tilde{x}_{kj}) \right\} \right\},
\end{aligned}
$$

and the corresponding constraint (5.4) can be written as

$$
\sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \exp\left\{ (\alpha_r + \log \varsigma_0 - \log \varsigma_r) + \boldsymbol{\beta}_r^{\intercal} \boldsymbol{q}(\tilde{x}_{kj}) \right\} \frac{p_{kj}}{\varsigma_0} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_r) = 1, \qquad (5.5)
$$

for $r = 0,\, 1,\, \ldots,\, m$. Now suppose $(\hat{\boldsymbol{p}},\, \hat{\boldsymbol{\varsigma}},\, \hat{\boldsymbol{\alpha}},\, \hat{\boldsymbol{\beta}})$ is a point at which the PEL is maximized under the constraint (5.5). Let $\check{\boldsymbol{\varsigma}}$ be an arbitrary vector each component, $\check{\varsigma}_k$, of which is in the interval $(0, 1]$. Put, for $k = 0,\, 1,\, \ldots,\, m$ and $j = 1,\, \ldots, n_k$,

$$
\check{p}_{kj} = \hat{p}_{kj} \frac{\check{\varsigma}_0}{\hat{\varsigma}_0} \quad \text{and} \quad \check{\alpha}_k = \hat{\alpha}_k + \log \frac{\hat{\varsigma}_0}{\check{\varsigma}_0} - \log \frac{\hat{\varsigma}_k}{\check{\varsigma}_k}.
$$

It is easily seen that $(\check{\boldsymbol{p}},\, \check{\boldsymbol{\varsigma}},\, \check{\boldsymbol{\alpha}},\, \hat{\boldsymbol{\beta}})$ also satisfies the constraint (5.5) and $PL_n(\check{\boldsymbol{\theta}},\, \check{\boldsymbol{p}},\, \check{\boldsymbol{\varsigma}})$ has the same value as $PL_n(\hat{\boldsymbol{\theta}},\, \hat{\boldsymbol{p}},\, \hat{\boldsymbol{\varsigma}})$. Hence the claimed result holds. □

Proposition 5.1 has a few important implications that help us calculating as well as studying the properties of the MELEs.

(i) The EL $L_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\iota},\, \boldsymbol{\varsigma})$ attains its maximum when both the PEL $PL_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma})$ and $\mathbb{L}_n(\boldsymbol{\iota},\, \boldsymbol{\varsigma})$ attain their maxima, respectively. Hence, the MELE $(\hat{\boldsymbol{\iota}},\, \hat{\boldsymbol{\varsigma}})$ is exactly the point at which the $\mathbb{L}_n(\boldsymbol{\iota},\, \boldsymbol{\varsigma})$ is maximized given that $0 < \varsigma_k \leq 1$ and $0 \leq \iota_k \leq 1 - \varsigma_k$. This point is easily seen to be

$$
\hat{\iota}_k = \check{n}_k / n_k \quad \text{and} \quad \hat{\varsigma}_k = \tilde{n}_k / n_k.
$$

(ii) The constrained maximization of PEL $PL_n(\boldsymbol{\theta},\, \boldsymbol{p},\, \boldsymbol{\varsigma})$ is overparameter-

ized: the maximum of PEL is independent of the value of $\boldsymbol{\varsigma}$. Let $\wp_{kj} = \varsigma_0{}^{-1} p_{kj}$, $\boldsymbol{\wp} = \{\wp_{kj} : j = 1, \ldots, \tilde{n}_k\}_{k=0}^m$, $\kappa_k = \alpha_k + \log \varsigma_0 - \log \varsigma_k$, and $\boldsymbol{\kappa} = (\kappa_1, \ldots, \kappa_m)^\intercal$. From the proof of Proposition 5.1, we see that the over–parameterization issue can be removed by re–parameterizing the PEL to

$$PL_n(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\wp}) = \left\{ \prod_{k=0}^{m} \prod_{j=1}^{\tilde{n}_k} \wp_{kj} \right\} \left\{ \prod_{k=1}^{m} \prod_{j=1}^{\tilde{n}_k} \exp\left\{ \kappa_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(\tilde{x}_{kj}) \right\} \right\}, \quad (5.6)$$

and the corresponding constraint (5.4) to

$$\sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \exp\left\{ \kappa_r + \boldsymbol{\beta}_r^\intercal \boldsymbol{q}(\tilde{x}_{kj}) \right\} \wp_{kj} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_r) = 1, \quad (5.7)$$

for $r = 0, 1, \ldots, m$. Define the maximum PEL estimator (MPELE) of $(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\wp})$ as

$$(\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\wp}}) = \underset{\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\wp}}{\operatorname{argmax}} \left\{ PL_n(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\wp}) : \text{ constraint (5.7)} \right\}.$$

Then the MELEs of $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ and $\boldsymbol{p}$ are just $\hat{\alpha}_k = \hat{\kappa}_k - \log \hat{\varsigma}_0 + \log \hat{\varsigma}_k$ for $k = 1, 2, \ldots, m$, $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{p}} = \hat{\varsigma}_0 \hat{\boldsymbol{\wp}}$, respectively.

(iii) All the information about the slope DRM parameter $\boldsymbol{\beta}$ is contained in the PEL, $PL_n(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\wp})$, so basing the inference about $\boldsymbol{\beta}$ on the PEL will cause no loss of efficiency for estimation or loss of statistical power for hypothesis testing.

5.3.2 Maximization of the PEL

We have seen that, to calculate the MELEs, the key is to maximize the PEL (5.6) under the constraint (5.7). Note that the PEL $PL_n(\boldsymbol{\kappa}, \boldsymbol{\beta}, \boldsymbol{\wp})$ and the corresponding constraint have similar mathematical expressions as

the EL (2.3) for uncensored data under the DRM and the corresponding constraint (2.2), except that in the current problem the $\{n_k\}$ are replaced by the $\{\tilde{n}_k\}$ and indicator terms are added to represent the supports of the uncensored observations. The constrained maximization of PEL can then be solved using a similar approach to that for the constrained maximization of EL in uncensored case. We first obtain the profile log–PEL

$$\tilde{\ell}_n(\boldsymbol{\kappa},\, \boldsymbol{\beta}) = \sup_{\boldsymbol{\wp}} \left\{ \log PL_n(\boldsymbol{\kappa},\, \boldsymbol{\beta},\, \boldsymbol{\wp}) :\ \text{constraint (5.7)} \right\}. \tag{5.8}$$

Let $\tilde{n} = \sum_{k=0}^{m} \tilde{n}_k$ be the total number of uncensored observations. This constrained maximization problem again can be solved by the method of Lagrange multipliers, and the supremum is found to be attained at

$$\wp_{kj} = \tilde{n}^{-1} \Big\{ 1 + \sum_{r=1}^{m} \lambda_r \big[\exp\big\{ \kappa_r + \boldsymbol{\beta}_r^{\intercal} \boldsymbol{q}(x_{kj}) \big\} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_r) - 1 \big] \Big\}^{-1}, \tag{5.9}$$

where the $\{\lambda_r\}_{r=1}^{m}$ are the solution to

$$\sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \wp_{kj} \exp\big\{ \kappa_t + \boldsymbol{\beta}_t^{\intercal} \boldsymbol{q}(\tilde{x}_{kj}) \big\} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_t) = 1,$$

for $t = 0,\, 1,\, \ldots,\, m$. The resulting profile log–PEL is

$$\begin{aligned} \tilde{\ell}_n(\boldsymbol{\kappa},\, \boldsymbol{\beta}) = &- \sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \log \Big\{ 1 + \sum_{r=1}^{m} \lambda_r \big[\exp\big\{ \kappa_r + \boldsymbol{\beta}_r^{\intercal} \boldsymbol{q}(\tilde{x}_{kj}) \big\} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_r) - 1 \big] \Big\} \\ &+ \sum_{k=1}^{m} \sum_{j=1}^{\tilde{n}_k} \big\{ \kappa_k + \boldsymbol{\beta}_k \boldsymbol{q}(\tilde{x}_{kj}) \big\}. \end{aligned}$$

As in the case of the profile log–EL for uncensored data, we found that the maximum of $\tilde{\ell}_n(\boldsymbol{\kappa},\, \boldsymbol{\beta})$ is attained when $\lambda_k = \tilde{n}_k/\tilde{n}$, for $k = 1,\, \ldots,\, m$. We define the *dual PEL* (DPEL) of $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$ by replacing the $\{\lambda_k\}$ in $\tilde{\ell}_n(\boldsymbol{\kappa},\, \boldsymbol{\beta})$

with $\{\tilde{n}_k/\tilde{n}\}$ as

$$\ell_n(\boldsymbol{\kappa},\, \boldsymbol{\beta}) = -\sum_{k=0}^{m}\sum_{j=1}^{\tilde{n}_k} \log\left\{\sum_{r=0}^{m}(\tilde{n}_r/\tilde{n})\exp\left\{\kappa_r + \boldsymbol{\beta}_r^{\intercal}\boldsymbol{q}(\tilde{x}_{kj})\right\}\mathbb{1}(\tilde{x}_{kj}\in\mathcal{S}_r)\right\} + \sum_{k=1}^{m}\sum_{j=1}^{\tilde{n}_k}\left\{\kappa_k + \boldsymbol{\beta}_k\boldsymbol{q}(\tilde{x}_{kj})\right\}. \tag{5.10}$$

Clealy, the DPEL $\ell_n(\boldsymbol{\kappa},\, \boldsymbol{\beta})$ shares the same maximal point and value as the profile log–PEL $\tilde{\ell}_n(\boldsymbol{\kappa},\, \boldsymbol{\beta})$. Therefore the MPELE $(\hat{\boldsymbol{\kappa}},\, \hat{\boldsymbol{\beta}})$ can be calculated as

$$(\hat{\boldsymbol{\kappa}},\, \hat{\boldsymbol{\beta}}) = \operatorname*{argmax}_{\boldsymbol{\kappa},\boldsymbol{\beta}} \ell_n(\boldsymbol{\kappa},\, \boldsymbol{\beta}).$$

The DPEL, just like the DEL for complete data, *has a simple analytical form and is concave*, so the above maximum can be easily computed.

Plugging $(\hat{\boldsymbol{\kappa}},\, \hat{\boldsymbol{\beta}})$ into (5.9) and replacing λ_r with $\tilde{n}_r/\tilde{n}$, we get

$$\hat{\wp}_{kj} = \left\{\sum_{r=0}^{m}\tilde{n}_r\exp\left\{\hat{\kappa}_r + \hat{\boldsymbol{\beta}}_r^{\intercal}\boldsymbol{q}(x_{kj})\right\}\mathbb{1}(\tilde{x}_{kj}\in\mathcal{S}_r)\right\}^{-1}. \tag{5.11}$$

5.4 Interpretation of the PEL

We now give an interpretation for the PEL (5.6): it is exactly the EL of the underlying distributions of the uncensored observations $\{\tilde{x}_{kj} : j = 1, \ldots, \tilde{n}_k\}$.

For a given $k \in \{0, 1, \ldots, m\}$, denote the CDF of the uncensored observations as $\tilde{F}_k$. Then $\tilde{F}_k$ is a truncated version of F_k:

$$d\tilde{F}_k(x) = {\varsigma_k}^{-1}\mathbb{1}(x\in\mathcal{S}_k)dF_k(x). \tag{5.12}$$

Recall that $\kappa_k = \alpha_k + \log\varsigma_0 - \log\varsigma_k$. Since the $\{F_k\}$ satisfy the DRM (2.1)

and $\mathcal{S}_k \subseteq \mathcal{S}_0$, we have

$$\begin{aligned} d\tilde{F}_k(x) &= \varsigma_k{}^{-1}\mathbb{1}(x \in \mathcal{S}_k)\exp\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\}dF_0(x). \\ &= \varsigma_0\varsigma_k{}^{-1}\exp\{\alpha_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\}\mathbb{1}(x \in \mathcal{S}_k)\{\varsigma_0{}^{-1}\mathbb{1}(x \in \mathcal{S}_0)dF_0(x)\} \\ &= \exp\{\kappa_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\}\mathbb{1}(x \in \mathcal{S}_k)d\tilde{F}_0(x). \end{aligned}$$

Note that we can add the factor $\mathbb{1}(x \in \mathcal{S}_0)$ in the second equality because $\mathcal{S}_k \subseteq \mathcal{S}_0$ implies $\mathbb{1}(x \in \mathcal{S}_k) = \mathbb{1}(x \in \mathcal{S}_k)\mathbb{1}(x \in \mathcal{S}_0)$.

We now see that the $\{\tilde{F}_k\}$ satisfy a DRM of the form

$$d\tilde{F}_k(x) = \exp\{\kappa_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\}\mathbb{1}(x \in \mathcal{S}_k)d\tilde{F}_0(x), \quad \text{for } k = 1, \ldots, m. \tag{5.13}$$

This model differs from the DRM (2.1) in that the supports of the non–baseline distributions are allowed to be different from each other, although all have to be contained in the support of the baseline distribution, while (2.1) requires all the distributions to have the same support. We therefore call model (5.13) the *varying support density ratio model* (VSDRM). Since $\tilde{F}_k$'s are distribution functions, $(\kappa_k, \boldsymbol{\beta}_k^\intercal)^\intercal$ must satisfy

$$\int \exp\{\kappa_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(x)\}\mathbb{1}(x \in \mathcal{S}_k)d\tilde{F}_0(x) = 1, \quad \text{for } k = 1, \ldots, m. \tag{5.14}$$

Just like the EL (2.3) for uncensored data, the EL of the $\{\tilde{F}_k\}$, is defined to be

$$\begin{aligned} L_n(\tilde{F}_0, \ldots, \tilde{F}_m) &= \prod_{k=0}^{m}\prod_{j=1}^{\tilde{n}_k} d\tilde{F}_k(\tilde{x}_{kj}) \\ &= \left\{\prod_{k=0}^{m}\prod_{j=1}^{\tilde{n}_k} \wp_{kj}\right\}\left\{\prod_{k=1}^{m}\prod_{j=1}^{\tilde{n}_k} \exp\left\{\kappa_k + \boldsymbol{\beta}_k^\intercal \boldsymbol{q}(\tilde{x}_{kj})\right\}\right\}, \end{aligned}$$

with $\wp_{kj} = d\tilde{F}_0(\tilde{x}_{kj}) = \varsigma_0^{-1}F_0(\tilde{x}_{kj}) = \varsigma_0^{-1}p_{kj}$. We see that this is exactly the PEL (5.6). The constraint (5.14) is exactly the constraint (5.7) corresponding

to the PEL, when we confine the support of $\tilde{F}_0$ on the $\{\tilde{x}_{kj}\}$. Therefore, maximizing the PEL (5.6) given constraint (5.7) is equivalent to maximizing the EL of the distribution functions for uncensored observations given the corresponding constraint (5.14).

5.5 Properties of the MPELE

Put $\boldsymbol{\vartheta} = (\boldsymbol{\kappa}, \boldsymbol{\beta})$. Given that MPELE $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\beta}})$ is the point at which the concave DPEL $\ell_n(\boldsymbol{\kappa}, \boldsymbol{\beta})$ — a function much like the DEL for complete data — is maximized, we wonder whether $\hat{\boldsymbol{\vartheta}}$ is asymptotically normal just as the MELE $\hat{\boldsymbol{\theta}}$ in the case of complete data. Recall that the asymptotic normality of $\hat{\boldsymbol{\theta}}$ in the case of complete data is determined by two facts: (1) the negative second–order derivative of DEL, which we call the empirical information matrix, has a limit; and (2) the score function evaluated at the true parameter value $\boldsymbol{\theta}^*$ has a normal limiting distribution. If these two properties also hold by the DPEL in the case of Type I censored data, the asymptotic normality of $\hat{\boldsymbol{\vartheta}}$ will follow.

The differences between the algebraic expressions of the DPEL (5.10) and the DEL (2.11) post a challenge for showing these properties when data are Type I censored. In the case of complete data, both limits are derived based on the fact that the DEL is a sum of iid random variables. However, the DPEL is no longer a sum of iid random variables, but rather a sum of dependent random variables, because of the repeated appearance of the same random number $\tilde{n}_k$ in each summand. Does the DPEL have the similar asymptotic properties as the DEL? The answer is positive as we summarize in the following lemmas. The proofs are given in Section 5.8

Let $\boldsymbol{\vartheta}^*$ denote the true parameter value $(\boldsymbol{\kappa}^*, \boldsymbol{\beta}^*)$. Call

$$\mathcal{U}_n = -n^{-1}\partial^2\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal$$

the *partial empirical information matrix*. Recall that $\boldsymbol{Q}^\intercal(x) = (1, \boldsymbol{q}^\intercal(x))$, n_k

is the size of the kth sample, and $n = \sum_{k=0}^{m} n_k$ is the total sample size.

Lemma 5.2 (Properties of the partial information matrix)**.** *Suppose we have $m+1$ random samples from populations with distributions of the DRM form given in (2.1) and a true parameter value $\boldsymbol{\theta}^*$ such that*

$$\int \exp\{\boldsymbol{\beta}_k^{\mathsf{T}} \boldsymbol{q}(x)\} dF_0(x) < \infty$$

for $\boldsymbol{\theta}$ in a neighbourhood of $\boldsymbol{\theta}^*$. Each sample is Type I censored with uncensored observations fall in the range of $\mathcal{S}_k = [c_{k,1},\, c_{k,2}]$ for given $c_{k,1} < c_{k,2}$, and $\mathcal{S}_k \subseteq \mathcal{S}_0$ for all k. Also, $\int \boldsymbol{Q}(x)\boldsymbol{Q}^{\mathsf{T}}(x) dF_0(x) > 0$, and $n_k/n = \rho_k + O(n^{-\delta})$ for some constants $\rho_k \in (0,1)$ and $\delta > 0$.

The partial empirical information matrix $\mathcal{U}_n$ converges almost surely to a positive definite matrix $\mathcal{U}$.

We call the limiting matrix $\mathcal{U}$ the *partial information matrix.* We partition its entries in agreement with $\boldsymbol{\kappa}$ and $\boldsymbol{\beta}$ and represent them as $\mathcal{U}_{\boldsymbol{\kappa}\boldsymbol{\kappa}}$, $\mathcal{U}_{\boldsymbol{\kappa}\boldsymbol{\beta}}$, $\mathcal{U}_{\boldsymbol{\beta}\boldsymbol{\kappa}}$ and $\mathcal{U}_{\boldsymbol{\beta}\boldsymbol{\beta}}$. Recall that $\varsigma_k = F_k(c_{k,2}) - F_k(c_{k,1}^-)$. Define $\varphi_k(\boldsymbol{\vartheta},\, x) = \exp\{\kappa_k + \boldsymbol{\beta}_k^{\mathsf{T}} \boldsymbol{q}(x)\}$ for $k = 0, \ldots, m$. Let

$$\begin{aligned}
\hbar(\boldsymbol{\vartheta},\, x) &= (\rho_1 \varsigma_1 \varphi_1(\boldsymbol{\vartheta},\, x) \mathbb{1}(x \in \mathcal{S}_1),\, \ldots,\, \rho_m \varsigma_m \varphi_m(\boldsymbol{\vartheta},\, x) \mathbb{1}(x \in \mathcal{S}_m))^{\mathsf{T}}, \\
\mathfrak{s}(\boldsymbol{\vartheta},\, x) &= \sum_{k=0}^{m} \rho_k \varsigma_k \varphi_k(\boldsymbol{\vartheta},\, x) \mathbb{1}(x \in \mathcal{S}_k), \qquad (5.15) \\
\mathcal{H}(\boldsymbol{\vartheta},\, x) &= \operatorname{diag}\{\hbar(\boldsymbol{\vartheta},\, x)\} - \hbar(\boldsymbol{\vartheta},\, x) \hbar^{\mathsf{T}}(\boldsymbol{\vartheta},\, x)/\mathfrak{s}(\boldsymbol{\vartheta},\, x).
\end{aligned}$$

Recall that $\mathrm{E}_k(\cdot)$, $k = 0, 1, \ldots,\, m$, be the expectation operator with respect to F_k. The blockwise algebraic expressions of the partial information matrix

$\mathcal{U}$ in terms of $\mathcal{H}(\boldsymbol{\vartheta}^*, x)$ and $\boldsymbol{q}(x)$ can be written as

$$\begin{aligned}
\mathcal{U}_{\boldsymbol{\kappa}\boldsymbol{\kappa}} &= -\lim_{n\to\infty} n^{-1}\partial^2\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}^\intercal = \varsigma_0^{-1}\mathrm{E}_0\big\{\mathcal{H}(\boldsymbol{\vartheta}^*, x)\big\},\\
\mathcal{U}_{\boldsymbol{\beta}\boldsymbol{\beta}} &= -\lim_{n\to\infty} n^{-1}\partial^2\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\intercal = \varsigma_0^{-1}\mathrm{E}_0\Big\{\mathcal{H}(\boldsymbol{\vartheta}^*, x)\otimes\big(\boldsymbol{q}(x)\boldsymbol{q}^\intercal(x)\big)\Big\},\\
\mathcal{U}_{\boldsymbol{\kappa}\boldsymbol{\beta}} &= -\lim_{n\to\infty} n^{-1}\partial^2\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\kappa}\partial\boldsymbol{\beta}^\intercal = \mathcal{U}_{\boldsymbol{\beta}\boldsymbol{\kappa}}^\intercal = \varsigma_0^{-1}\mathrm{E}_0\big\{\mathcal{H}(\boldsymbol{\vartheta}^*, x)\otimes\boldsymbol{q}^\intercal(x)\big\}.
\end{aligned} \tag{5.16}$$

Let $\boldsymbol{v} = n^{-1/2}\partial\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\vartheta}$. Define

$$\begin{aligned}
\mathcal{T} &= (\rho_0\varsigma_0)^{-1}\mathbf{1}_m\mathbf{1}_m^\intercal + \mathrm{diag}\{(\rho_1\varsigma_1)^{-1}, (\rho_2\varsigma_2)^{-1}, \ldots, (\rho_m\varsigma_m)^{-1}\},\\
\mathcal{W} &= \begin{pmatrix} \mathcal{T} & 0_{m\times md}\\ 0_{md\times m} & 0_{md\times md}\end{pmatrix}.
\end{aligned}$$

Lemma 5.3 (Asymptotic properties of the score function)**.** *Under the conditions of Theorem 5.2,* $\mathrm{E}\boldsymbol{v} = \mathbf{0}$ *and* $\boldsymbol{v}$ *is asymptotically multivariate normal with mean* $\mathbf{0}$ *and covariance matrix* $\mathcal{V} = \mathcal{U} - \mathcal{U}\mathcal{W}\mathcal{U}$.

The asymptotic normality of the MPELE, $\hat{\boldsymbol{\vartheta}} = (\hat{\boldsymbol{\kappa}}, \hat{\boldsymbol{\beta}})$, follows from Lemma 5.2 and 5.3.

Theorem 5.4 (Asymptotic normality of the MPELE)**.** *Under the conditions of Theorem 5.2,* $\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*)$ *has an asymptotic multivariate normal distribution with mean* $\mathbf{0}$ *and covariance matrix* $\mathcal{U}^{-1} - \mathcal{W}$.

5.6 EL ratio test for the DRM parameter

A primary inference problem of our interest, as noted in Chapter 1, is to test whether the underlying distributions of a few Type I censored lumber samples are equal. As commented in Chapter 3, such hypotheses can be translated into testing equalities among the $\{\boldsymbol{\beta}_k\}$ parameters under the DRM setting, which are special cases embraced by the general composite hypothesis testing

problem (3.1),

$$H_0: \ \boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0} \quad \text{against} \quad H_1: \ \boldsymbol{g}(\boldsymbol{\beta}) \neq \mathbf{0}$$

for some smooth function $\boldsymbol{g} : \mathbb{R}^{md} \to \mathbb{R}^q$, with $q \leq md$, the length of $\boldsymbol{\beta}$. The function $\boldsymbol{g}$ is assumed to be thrice differentiable with a full rank Jacobian matrix. Does the corresponding EL ratio statistic based on Type I censored samples have a chi–square limiting distribution just like the case for complete data? The answer is affirmative as we will seen in this section.

Let $(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\varsigma}})$ denote the point at which the maximum of the EL (5.1) is attained under the null constraint $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$. Recall that the MELE $(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}})$ defined in (5.2) is the point at which the EL is maximized without the null constraint. The EL ratio (ELR) test statistic is defined to be

$$R_n = 2\{\log L_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}}) - \log L_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\varsigma}})\}.$$

Factorizing the $L_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\iota}, \boldsymbol{\varsigma})$ in the above definition into the product of PEL and $\mathbb{L}_n(\boldsymbol{\iota}, \boldsymbol{\vartheta})$ as in (5.3), we get

$$R_n = 2\big(\{\log PL_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\varsigma}}) + \log \mathbb{L}_n(\hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}})\} - \{\log PL_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\varsigma}}) + \log \mathbb{L}_n(\tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\varsigma}})\}\big).$$

By Proposition 5.1, the maximum value of $PL_n(\boldsymbol{\theta}, \boldsymbol{p}, \boldsymbol{\varsigma})$ is independent of the value of $\boldsymbol{\varsigma}$, so the MELE $(\hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\vartheta}})$ is just the point at which $\mathbb{L}_n(\boldsymbol{\iota}, \boldsymbol{\vartheta})$ is maximized. Now, under the null constraint on $\boldsymbol{\beta}$, the conclusion of Proposition 5.1 still applies, because the proof of that proposition does not involve any algebraic operation on $\boldsymbol{\beta}$. Hence, $(\tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\varsigma}})$ is also the point at which $\mathbb{L}_n(\boldsymbol{\iota}, \boldsymbol{\vartheta})$ is maximized, thereby not influenced by the null constraint. Consequently, $(\tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\varsigma}}) = (\hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}})$, $\mathbb{L}_n(\hat{\boldsymbol{\iota}}, \hat{\boldsymbol{\varsigma}}) = \mathbb{L}_n(\tilde{\boldsymbol{\iota}}, \tilde{\boldsymbol{\varsigma}})$, and R_n becomes

$$R_n = 2\{\log PL_n(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{p}}, \hat{\boldsymbol{\varsigma}}) - \log PL_n(\tilde{\boldsymbol{\theta}}, \tilde{\boldsymbol{p}}, \tilde{\boldsymbol{\varsigma}})\}.$$

Recalling that the DPEL $\ell(\boldsymbol{\vartheta})$ shares the same maximum value with the

PEL, R_n hence further simplifies to

$$R_n = 2\{\ell_n(\hat{\boldsymbol{\vartheta}}) - \ell_n(\tilde{\boldsymbol{\vartheta}})\}.$$

In other words, the EL ratio statistic equals the DPEL ratio statistic, which agrees with the argument (iii) in Section 5.3.1, that basing the inference about $\boldsymbol{\beta}$ on the PEL will cause no loss of statistical power for hypothesis testing. Based on this fact, we find the asymptotic properties the ELR test as summarized in the theorem below.

Recall that, when $q < md$, the null hypothesis $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$ can be equivalently expressed as $\boldsymbol{\beta} = \mathcal{G}(\boldsymbol{\gamma})$ for some lower dimensional parameter $\boldsymbol{\gamma}$ and a unique function $\mathcal{G}$: $\mathbb{R}^{md-q} \to \mathbb{R}^{md}$, which is thrice differentiable with full rank Jacobian matrix $J = \partial\mathcal{G}(\boldsymbol{\gamma}^*)/\partial\boldsymbol{\gamma}$. When $q = md$, $\boldsymbol{g}$ is invertible and the null hypothesis is fully specified as $\boldsymbol{\beta} = \boldsymbol{g}^{-1}(\mathbf{0})$.

Theorem 5.5 (Asymptotic properties of the ELR test)**.** *Adopt the conditions postulated in Lemma 5.2.*

(i) Under the null hypothesis, $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$, of (3.1)*, $R_n \to \chi^2_q$ in distribution as $n \to \infty$.*

(ii) Under local alternative (3.2)*:*

$$\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^* + n_k^{-1/2}\boldsymbol{c}_k,$$

$R_n \to \chi^2_q(\delta^2)$ in distribution as $n \to \infty$, where δ^2 is a nonnegative non–central parameter with expression

$$\delta^2 = \begin{cases} \boldsymbol{\eta}^\intercal\{\tilde{\Lambda} - \tilde{\Lambda}J\big(J^\intercal\tilde{\Lambda}J\big)^{-1}J^\intercal\tilde{\Lambda}\}\boldsymbol{\eta} & \text{if } q < md \\ \boldsymbol{\eta}^\intercal\tilde{\Lambda}\boldsymbol{\eta} & \text{if } q = md \end{cases}$$

where $\tilde{\Lambda} = \mathcal{U}_{\boldsymbol{\beta\beta}} - \mathcal{U}_{\boldsymbol{\beta\kappa}}\mathcal{U}_{\boldsymbol{\kappa\kappa}}^{-1}\mathcal{U}_{\boldsymbol{\kappa\beta}}$. and $\boldsymbol{\eta} = \big(\rho_1^{-1/2}\boldsymbol{c}_1^\intercal, \rho_2^{-1/2}\boldsymbol{c}_2^\intercal, \ldots, \varsigma_m^{-1/2}\boldsymbol{c}_m^\intercal\big)^\intercal$ is the one given in Theorem 3.2.

Also, $\delta^2 > 0$ *unless* $\boldsymbol{\eta}$ *is in the column space of* J*.*

There are similar results to Theorem 4.2 and 4.3 that concerns local power comparison of ELR tests based on Type I censored data, but omitted since they are just straightforward extensions.

5.7 Other inference tasks

As we have seen, the above inference framework for multiple Type I censored samples under the DRM centers on the DPEL (5.10), a function looks remarkably similar to the DEL (2.11) for complete data under the DRM. In Section 5.5, we showed that the DPEL actually has similar asymptotic properties to the DEL. Based on these properties, the DEL ratio test that is available for complete data can be adapted to the case of Type I censored data. In fact, with the DPEL, we can show that the results on EL quantile estimation given by Chen and Liu (2013) and on density estimation given by Fokianos (2004), all of which are derived under the DRM for complete samples, also extend to the Type I censored case. Similarly, based on the EL inference framework given in this chapter, any EL inference result that is in effect for complete samples under the DRM may be extended to the case of Type I censored samples.

5.8 Proofs

We first introduce a few results and more notations applicable to $k = 0, \ldots, m$. A common condition for the theorems of this Chapter is $n_k/n = \rho_k + O(n^{-\delta})$ for some positive constant δ. Without loss of generality, we assume $\delta \leq 1/3$: if a result holds for a larger positive constant δ, it also holds for smaller δ. Recall that $\varsigma_k = \Pr(X_{k1} \in \mathcal{S}_k)$, $\tilde{n}_k$ is the number of uncensored observations

in the kth sample, and $\tilde{n} = \sum_{k=0}^{m} \tilde{n}_k$ By the law of the iterated logarithm,

$$\tilde{n}_k/n_k = n_k^{-1} \sum_{i=1}^{n_k} \mathbb{1}(x_{kj} \in \mathcal{S}_k) = \varsigma_k + O(n^{-1/2} \log\log n).$$

Therefore,

$$\frac{\tilde{n}_k}{n} = \frac{\tilde{n}_k}{n_k} \cdot \frac{n_k}{n} = \big(\varsigma_k + O(n^{-1/2} \log\log n)\big)\big(\rho_k + O(n^{-\delta})\big) = \rho_k \varsigma_k + O(n^{-\delta}). \tag{5.17}$$

We use symbol $\sum_{k,j}$ to denote $\sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k}$, the sum over all k and $j = 1, \ldots, \tilde{n}_k$ for each given k. Recall that $\varphi_k(\boldsymbol{\vartheta}, x) = \exp\{\kappa_k + \boldsymbol{\beta}_k \boldsymbol{q}(x)\}$ and $\tilde{\lambda}_k = \tilde{n}_k/\tilde{n}$. Write

$$\mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x) = -\log\Big\{\sum_{r=0}^{m} (\tilde{n}_r/\tilde{n}) \varphi_r(\boldsymbol{\vartheta}, x) \mathbb{1}(x \in \mathcal{S}_r)\Big\} + \{\kappa_k + \boldsymbol{\beta}_k^{\mathsf{T}} \boldsymbol{q}(x)\}.$$

The DPEL (5.10) can be written as $\ell_n(\boldsymbol{\vartheta}) = \sum_{k,\,j} \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, \tilde{x}_{kj})$. Let $\mathscr{h}_n(\boldsymbol{\vartheta}, x)$, $\mathscr{s}_n(\boldsymbol{\vartheta}, x)$ and $\mathscr{H}_n(\boldsymbol{\vartheta}, x)$ be defined as the $\mathscr{h}(\boldsymbol{\vartheta}, x)$, $\mathscr{s}(\boldsymbol{\vartheta}, x)$ and $\mathscr{H}(\boldsymbol{\vartheta}, x)$ in (5.15) with $\rho_k \varsigma_k$ replaced by $\tilde{n}_k/n$. Since $\lim_{n\to\infty}(\tilde{n}_k/n) \to \rho_k \varsigma_k$, $\mathscr{h}(\boldsymbol{\vartheta}, x)$, $\mathscr{s}(\boldsymbol{\vartheta}, x)$ and $\mathscr{H}(\boldsymbol{\vartheta}, x)$ are the limits of $\mathscr{h}_n(\boldsymbol{\vartheta}, x)$, $\mathscr{s}_n(\boldsymbol{\vartheta}, x)$ and $\mathscr{H}_n(\boldsymbol{\vartheta}, x)$, respectively, as n tends to infinity. The first order derivatives of $\mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)$ can be written as

$$\begin{aligned} \partial \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial \boldsymbol{\kappa} &= (1 - \delta_{k0}) \boldsymbol{e}_k - \mathscr{h}_n(\boldsymbol{\vartheta}, x)/\mathscr{s}_n(\boldsymbol{\vartheta}, x), \\ \partial \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial \boldsymbol{\beta} &= \{\partial \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial \boldsymbol{\kappa}\} \otimes \boldsymbol{q}(x). \end{aligned} \tag{5.18}$$

Similarly, we have

$$\begin{aligned} \partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial \boldsymbol{\kappa} \partial \boldsymbol{\kappa}^{\mathsf{T}} &= -\mathscr{H}_n(\boldsymbol{\vartheta}, x)/\mathscr{s}_n(\boldsymbol{\vartheta}, x), \\ \partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^{\mathsf{T}} &= -\big\{\mathscr{H}_n(\boldsymbol{\vartheta}, x)/\mathscr{s}_n(\boldsymbol{\vartheta}, x)\big\} \otimes \big\{\boldsymbol{q}(x) \boldsymbol{q}^{\mathsf{T}}(x)\big\}, \\ \partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial \boldsymbol{\kappa} \partial \boldsymbol{\beta}^{\mathsf{T}} &= -\big\{\mathscr{H}_n(\boldsymbol{\vartheta}, x)/\mathscr{s}_n(\boldsymbol{\vartheta}, x)\big\} \otimes \boldsymbol{q}^{\mathsf{T}}(x). \end{aligned} \tag{5.19}$$

Note that all entries of $\mathcal{H}_n(\boldsymbol{\vartheta}, x)$ are non–negative, and $\mathcal{s}_n(\boldsymbol{\vartheta}, x)$ exceeds the sum of all entries of $\mathcal{h}_n(\boldsymbol{\vartheta}, x)$. Thus, $\|\mathcal{h}_n(\boldsymbol{\vartheta}, x)/\mathcal{s}_n(\boldsymbol{\vartheta}, x)\| \leq 1$, and the absolute value of each entry of $\mathcal{H}_n(\boldsymbol{\vartheta}, x)/\mathcal{s}_n(\boldsymbol{\vartheta}, x)$ is bounded by 1. By examining the algebraic expressions closely, this result implies

$$\begin{aligned} &\left|\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial\vartheta_i\partial\vartheta_j\right| \leq 1 + \boldsymbol{q}^{\mathsf{T}}(x)\boldsymbol{q}(x), \\ &\left|\partial^3 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)/\partial\vartheta_i\partial\vartheta_j\partial\vartheta_k\right| \leq \{1 + \boldsymbol{q}^{\mathsf{T}}(x)\boldsymbol{q}(x)\}^{3/2}, \end{aligned} \tag{5.20}$$

where ϑ_i denotes the ith entry of $\boldsymbol{\vartheta}$. These inequalities are just a "censored" version of inequalities (2.17) for complete data.

Let $\mathcal{L}_k(\boldsymbol{\vartheta}, x)$ be the "population" version of $\mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)$ by replacing $\tilde{n}_r/n$ with its limit $\rho_r\varsigma_r$ in the above definition. The first and second order derivatives of $\mathcal{L}_k(\boldsymbol{\vartheta}, x)$ are the same as those of $\mathcal{L}_{n,k}(\boldsymbol{\vartheta}, x)$ with $\mathcal{h}_n(\boldsymbol{\vartheta}, x)$, $\mathcal{s}_n(\boldsymbol{\vartheta}, x)$ and $\mathcal{H}_n(\boldsymbol{\vartheta}, x)$ replaced by $\mathcal{h}(\boldsymbol{\vartheta}, x)$, $\mathcal{s}(\boldsymbol{\vartheta}, x)$ and $\mathcal{H}(\boldsymbol{\vartheta}, x)$. Also, $\mathcal{L}_k(\boldsymbol{\vartheta}, x)$ satisfy inequalities (5.20).

5.8.1 Lemma 5.2: Properties of the partial information matrix

By the fact that $\ell_n(\boldsymbol{\vartheta}) = \sum_{k,j} \mathcal{L}_{n,k}(\boldsymbol{\vartheta}, \tilde{x}_{kj})$ and $\tilde{n}_k/n = \rho_k\varsigma_k + o(1)$, we have

$$\mathcal{U}_n = -\frac{1}{n}\frac{\partial^2 \ell_n(\boldsymbol{\vartheta}^*)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\mathsf{T}}} = -\sum_{k=0}^{m}(\rho_k\varsigma_k + o(1))\left\{\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\mathsf{T}}}\right\}. \tag{5.21}$$

Note that, for any given k, the sum in the curly brackets is not a sum of iid random variables because of the presence of the random variable $\tilde{n}_k$ in the expressions of $\mathcal{h}_n(\boldsymbol{\vartheta}, x)$, $\mathcal{s}_n(\boldsymbol{\vartheta}, x)$ and $\mathcal{H}_n(\boldsymbol{\vartheta}, x)$, which appear in the expression (5.19) of $\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})/\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^{\mathsf{T}}$. This negates a simple application

of the law of large numbers. However, as we will show later,

$$\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\partial^2\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\,\tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}=\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,\tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}+o(1). \tag{5.22}$$

Now with $\{\tilde{n}_k/n\}$ replaced by constants $\{\rho_k\varsigma_k\}$ in the expression of the second order derivatives of $\mathcal{L}_k(\boldsymbol{\vartheta},\,x)$, the sum on the RHS of the above equality is a sum of an iid random variables. By inequalities (5.20), each summand on the RHS is dominated by an integrable function. Hence by the strong law of large numbers, the first term on the RHS satisfies

$$\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,\tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}=\tilde{\mathrm{E}}_k\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,x)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}+o(1),$$

where $\tilde{\mathrm{E}}_k(\cdot)$ is the expectation operator with respect to $\tilde{F}_k$. Consequently, (5.21) simplifies to

$$\begin{aligned}\mathcal{U}_n&=-\sum_{k=0}^{m}(\rho_k\varsigma_k+o(1))\Big\{\tilde{\mathrm{E}}_k\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,x)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}+o(1)\Big\}\\&=-\sum_{k=0}^{m}\rho_k\varsigma_k\tilde{\mathrm{E}}_k\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,x)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}+o(1).\end{aligned}$$

Therefore $\mathcal{U}_n$ converges almost surely to

$$\mathcal{U}=-\sum_{k=0}^{m}\rho_k\varsigma_k\tilde{\mathrm{E}}_k\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,x)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}.$$

Recall that when $\mathcal{S}_k\subseteq\mathcal{S}_0$, the $\{\tilde{F}_k\}$ satisfy, (5.13), the VSDRM, so the above expression of $\mathcal{U}$ can be further written as

$$\mathcal{U}=-\sum_{k=0}^{m}\rho_k\varsigma_k\tilde{\mathrm{E}}_0\left\{\frac{\partial^2\mathcal{L}_{k}(\boldsymbol{\vartheta}^*,\,x)}{\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal}\varphi_k(\boldsymbol{\vartheta}^*,\,x)\mathbb{1}(x\in\mathcal{S}_k)\right\}.$$

By expressions (5.19) of $\partial^2 \mathcal{L}_k(\boldsymbol{\vartheta}^*, x)/\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\intercal$, we have

$$\begin{aligned}
\mathcal{U}_{\boldsymbol{\kappa\kappa}} &= \sum_{k=0}^{m} \rho_k \varsigma_k \tilde{\mathrm{E}}_0 \left\{ \{\mathcal{H}(\boldsymbol{\vartheta}^*, x)/\boldsymbol{s}(\boldsymbol{\vartheta}^*, x)\} \varphi_k(\boldsymbol{\vartheta}^*, x) \mathbb{1}(x \in \mathcal{S}_k) \right\} \\
&= \tilde{\mathrm{E}}_0 \Big\{ \{\mathcal{H}(\boldsymbol{\vartheta}^*, x)/\boldsymbol{s}(\boldsymbol{\vartheta}^*, x)\} \Big\{ \sum_{k=0}^{m} \rho_k \varsigma_k \varphi_k(\boldsymbol{\vartheta}^*, x) \mathbb{1}(x \in \mathcal{S}_k) \Big\} \Big\} \\
&= \tilde{\mathrm{E}}_0 \{\mathcal{H}(\boldsymbol{\vartheta}^*, x)\},
\end{aligned}$$

where the last equality is by $\boldsymbol{s}(\boldsymbol{\vartheta}^*, x) = \sum_{k=0}^{m} \rho_k \varsigma_k \varphi_k(\boldsymbol{\vartheta}^*, x) \mathbb{1}(x \in \mathcal{S}_k)$. Recall that $d\tilde{F}_0(x) = \varsigma_0^{-1} \mathbb{1}(x \in \mathcal{S}_0) dF_0(x)$ by (5.12), and $\mathcal{S}_k \in \mathcal{S}_0$ for all k. The above expression of $\mathcal{U}_{\boldsymbol{\kappa\kappa}}$ therefore simplies to

$$\mathcal{U}_{\boldsymbol{\kappa\kappa}} = \tilde{\mathrm{E}}_0 \{\mathcal{H}(\boldsymbol{\vartheta}^*, x)\} = \varsigma_0^{-1} \mathrm{E}_0 \{\mathcal{H}(\boldsymbol{\vartheta}^*, x)\}.$$

As a reminder, $\mathrm{E}_k(\cdot)$ in general is the expectation operator with respect to F_k. Similarly, we found the expressions for $\mathcal{U}_{\boldsymbol{\beta\beta}}$ and $\mathcal{U}_{\boldsymbol{\kappa\beta}}$ as given in (5.16).

That $\mathcal{U}$ and $\mathcal{U}_{\boldsymbol{\kappa\kappa}}$ are positive definite, can be shown using a similar argument to that for showing the positive definiteness of U and $U_{\boldsymbol{\alpha\alpha}}$ in the proof of Lemma 2.1 (Section 2.5.1).

To complete the proof, we show the matrix equality (5.22) block by block. First we show that

$$\frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}^\intercal} = \frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\partial^2 \mathcal{L}_k(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial\boldsymbol{\kappa}\partial\boldsymbol{\kappa}^\intercal} + o_p(1). \tag{5.23}$$

Define, for $k = 0, 1, \ldots, m$,

$$\begin{aligned}
\phi_k(x) &= \varphi_k(\boldsymbol{\vartheta}^*, x) \mathbb{1}(x \in \mathcal{S}_k) \\
\Delta_k &= \tilde{n}_k/n - \rho_k \varsigma_k, \\
\Delta &= \max_{0 \le k \le m} |\Delta_k|.
\end{aligned}$$

By (5.17), $\Delta_k = O(n^{-\delta})$ for each k and $\Delta = O(n^{-\delta})$. By expressions (5.19), the element on the ith row and t^{th} column, $i,\, t \in \{1,\, \ldots,\, m\}$ and $i \neq t$, of the LHS matrix of the equality (5.23) can be written as

$$\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\partial\kappa_i\partial\kappa_t} = \frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{(\rho_i\varsigma_i + \Delta_i)\phi_i(\tilde{x}_{kj})(\rho_t\varsigma_t + \Delta_t)\phi_t(\tilde{x}_{kj})}{\boldsymbol{s}_n^2(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}. \tag{5.24}$$

Note that

$$\frac{1}{\boldsymbol{s}_n^2(\boldsymbol{\vartheta}^*,\, x)} = \frac{1}{\boldsymbol{s}^2(\boldsymbol{\vartheta}^*,\, x)}\frac{1}{\left\{1 + (\sum_{r=0}^m \Delta_r\phi_r(x)/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)\right\}^2}.$$

Now the key step is to perform a Taylor expansion for the second factor on the RHS of the above equality,

$$\frac{1}{\left\{1 + (\sum_{r=0}^m \Delta_r\phi_r(x))/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)\right\}^2} = 1 - \frac{2\Delta\left\{\sum_{r=0}^m(\Delta_r/\Delta)\phi_r(x)\right\}/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)}{\{1 + a_n(\sum_{r=0}^m \Delta_r\phi_r(x))/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)\}^3},$$

where a_n is a non–random number in the interval $[0,\, 1]$ that may change with the total sample size n. With this expansion and (5.24), we get

$$\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\partial\kappa_i\partial\kappa_t} = (\rho_i\varsigma_i + \Delta_i)(\rho_t\varsigma_t + \Delta_t)\left\{\frac{1}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\frac{\phi_i(\tilde{x}_{kj})\phi_t(\tilde{x}_{kj})}{\boldsymbol{s}^2(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})} - R_n\right\}, \tag{5.25}$$

where

$$R_n = \frac{2\Delta}{\tilde{n}_k}\sum_{j=1}^{\tilde{n}_k}\left\{\frac{\phi_i(\tilde{x}_{kj})\phi_t(\tilde{x}_{kj})}{\boldsymbol{s}^2(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}\frac{(\sum_{r=0}^m(\Delta_r/\Delta)\phi_r(\tilde{x}_{kj}))/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\{1 + a_n(\sum_{r=0}^m \Delta_r\phi_r(\tilde{x}_{kj}))/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})\}^3}\right\}. \tag{5.26}$$

The first term in the curly brackets on the RHS of (5.25) is the average of a

sum of iid random variables. Moreover, by the definition of $\boldsymbol{s}(\boldsymbol{\vartheta}, x)$, (5.15), and recalling $\phi_k(x) = \varphi_k(\boldsymbol{\vartheta}^*, x)\mathbb{1}(x \in \mathcal{S}_k)$, we have that, for all x and every $k \in \{0, 1, \ldots, m\}$,

$$0 < \phi_k(x)/\boldsymbol{s}(\boldsymbol{\vartheta}^*, x) < 1/(\rho_k \varsigma_k) \leq 1/(\min_{0 \leq i \leq m} \rho_i \varsigma_i). \tag{5.27}$$

Thus for all x and any $i, t \in \{0, 1, \ldots, m\}$,

$$0 < \phi_i(x)\phi_t(x)/\boldsymbol{s}^2(\boldsymbol{\vartheta}^*, x) < 1/(\min_{0 \leq i \leq m} \rho_i \varsigma_i)^2.$$

We can therefore invoke the strong law of large numbers to conclude that the first term in the curly brackets on the RHS of (5.25) is of $O(1)$. The second term R_n, as we will show soon, is of $o(1)$. Hence, with the fact that $\Delta_k = O(n^{-\delta}) = o(1)$, we have

$$\begin{aligned}
\frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial \kappa_i \partial \kappa_t} &= \frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\rho_i \varsigma_i \phi_i(\tilde{x}_{kj}) \rho_t \varsigma_t \phi_t(\tilde{x}_{kj})}{\boldsymbol{s}^2(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})} + o(1) \\
&= \frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\partial^2 \mathcal{L}_k(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial \kappa_i \partial \kappa_t} + o(1).
\end{aligned}$$

Similarly, for $i = 1, \ldots, m$, we can show

$$\frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial \kappa_i^2} = \frac{1}{\tilde{n}_k} \sum_{j=1}^{\tilde{n}_k} \frac{\partial^2 \mathcal{L}_k(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})}{\partial \kappa_i^2} + o(1).$$

Therefore (5.23) holds.

By expressions (5.19) and the fact that $\boldsymbol{q}(x)\boldsymbol{q}^\intercal(x)$ is an integrable function, we can show that similar equalities to (5.23) hold for $\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})/\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\intercal$ and $\partial^2 \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})/\partial \boldsymbol{\kappa} \partial \boldsymbol{\beta}^\intercal$. Hence (5.22) is true and the lemma is proved.

To finish up, we show R_n is of $o(1)$. Recalling (5.15) that $\boldsymbol{s}(\boldsymbol{\vartheta}^*, x)$ is the sum of the positive terms $\rho_k \varsigma_k \phi_r(x)$ over $k = 0, 1, \ldots, m$, we thus have, for

all x,

$$\left|\frac{\sum_{r=0}^{m}(\Delta_r/\Delta)\phi_r(x)}{\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)}\right| \leq \frac{\sum_{r=0}^{m}|\Delta_r/\Delta|\phi_r(x)}{\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)} \leq \frac{\sum_{r=0}^{m}\phi_r(x)}{\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)} \leq \frac{1}{\min_{0\leq i\leq m}\rho_i\varsigma_i}. \tag{5.28}$$

Consequently, for all x,

$$\left|\frac{\sum_{r=0}^{m}\Delta_r\phi_r(x)}{\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)}\right| = |\Delta|\left|\frac{\sum_{r=0}^{m}\Delta_r/\Delta_r\phi_r(x)}{\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)}\right| \leq \frac{\Delta}{\min_{0\leq i\leq m}\rho_i\varsigma_i}.$$

Since $\Delta = o(1)$ and $0 \leq a_n \leq 1$ for all n, we can find a N, such that whenever $n > N$ and uniformly in x,

$$|a_n(\sum_{r=0}^{m}\Delta_r\phi_r(x))/\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)| < 1/2,$$

and so

$$\frac{2}{3} < \frac{1}{1 + a_n(\sum_{r=0}^{m}\Delta_r\phi_r(x))/\mathcal{s}(\boldsymbol{\vartheta}^*,\, x)} < 2. \tag{5.29}$$

Therefore, by bounds (5.27), (5.28), (5.29) and the expression (5.26) of R_n, we have, for all n large enough,

$$|R_n| < \frac{16\Delta}{(\min_{0\leq i\leq m}\rho_i\varsigma_i)^3}.$$

Since $\Delta = o(1)$, we have $R_n = o(1)$. The proof now is complete.

5.8.2 Lemma 5.3: Asymptotic properties of the score function

For ease of presentation, we first proof the lemma for $\delta = 1/3$, i.e. when $n_k/n = O(n^{-1/3})$. Then we show that the lemma also holds for arbitrary

$0 < \delta < 1/3$.

Recall that $\boldsymbol{v} = n^{-1/2}\partial\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\vartheta} = n^{-1/2}\sum_{k,j}\{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})/\partial\boldsymbol{\vartheta}\}$. We first show $\boldsymbol{v}$ can be centered in a particular sense. For a function $g(x)$ that functionally involves the random variables $\{\tilde{n}_k\}$, we use $\tilde{\int} g(x)d\tilde{F}_r(x)$, $r = 0, \ldots, m$, to denote an integral that pretends $\tilde{n}_r$ to be a non–random constant. For example, for the previously defined $\boldsymbol{s}_n(\boldsymbol{\vartheta}^*,\, x) = \sum_{k=0}^{m}(\tilde{n}_k/n)\phi_k(x)$,

$$\tilde{\int} \boldsymbol{s}_n(x)d\tilde{F}_r(x) = \sum_{k=0}^{m}(\tilde{n}_k/n)\int \phi_k(x)d\tilde{F}_r(x).$$

By the VSDRM assumption (5.13) and expression (5.18), we find

$$\sum_{k=0}^{m}(\tilde{n}_k/n)\tilde{\int}\{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, x)/\partial\boldsymbol{\vartheta}\}d\tilde{F}_k = \mathbf{0}.$$

Hence, $\boldsymbol{v}$ can be "centered" as

$$\begin{aligned}\boldsymbol{v} &= n^{-1/2}\sum_{k,j}\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})/\partial\boldsymbol{\vartheta}\\ &= \sum_{k=0}^{m}\frac{\sqrt{\tilde{n}_k}}{\sqrt{n}}\left\{\frac{1}{\sqrt{\tilde{n}_k}}\sum_{j=1}^{\tilde{n}_k}\Big(\frac{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}} - \tilde{\int}\frac{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, x)}{\partial\boldsymbol{\vartheta}}d\tilde{F}_k\Big)\right\}. \quad (5.30)\end{aligned}$$

In the second step, we show that for each given k, $k = 0, \ldots, m$, the term in the curly brackets of (5.30) satisfies

$$\begin{aligned}&\frac{1}{\sqrt{\tilde{n}_k}}\sum_{j=1}^{\tilde{n}_k}\Big(\frac{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}} - \tilde{\int}\frac{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, x)}{\partial\boldsymbol{\vartheta}}d\tilde{F}_k\Big)\\ =&\frac{1}{\sqrt{\tilde{n}_k}}\sum_{j=1}^{\tilde{n}_k}\Big(\frac{\partial\mathcal{L}_k(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\partial\boldsymbol{\vartheta}} - \tilde{\mathrm{E}}_k\frac{\partial\mathcal{L}_k(\boldsymbol{\vartheta}^*,\, x)}{\partial\boldsymbol{\vartheta}}\Big) + o_p(1). \quad (5.31)\end{aligned}$$

By (5.18) and $\Delta_k = \tilde{n}_k/n - \rho_k\varsigma_k$, to show the above equality, it is enough to

show that for any given $i \in \{1, 2, \ldots, m\}$,

$$\begin{aligned}
&\frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{(\rho_i \varsigma_i + \Delta_i)\phi_i(\tilde{x}_{kj})}{s_n(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})} - \tilde{\int} \frac{(\rho_i \varsigma_i + \Delta_i)\phi_i(x)}{s_n(\boldsymbol{\vartheta}^*, x)} d\tilde{F}_k \right\} \\
=&\frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{\rho_i \varsigma_i \phi_i(\tilde{x}_{kj})}{s(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})} - \tilde{\mathrm{E}}_k \frac{\rho_i \varsigma_i \phi_i(x)}{s(\boldsymbol{\vartheta}^*, x)} \right\} + o_p(1). \qquad (5.32)
\end{aligned}$$

Note that

$$\frac{1}{s_n(\boldsymbol{\vartheta}^*, x)} = \frac{1}{s(\boldsymbol{\vartheta}^*, x)} \frac{1}{\left\{1 + (\sum_{r=0}^m \Delta_r \phi_r(x))/s(\boldsymbol{\vartheta}^*, x)\right\}}.$$

Recall that $\Delta = \max_{0 \le k \le m} |\Delta_k|$. The second factor on the RHS of the above equality admits the following expansion

$$\frac{1}{1 + (\sum_{r=0}^m \Delta_r \phi_r(x))/s(\boldsymbol{\vartheta}^*, x)} = 1 - \frac{\sum_{r=0}^m \Delta_r \phi_r(x)}{s(\boldsymbol{\vartheta}^*, x)} + \Delta^2 \mathbb{Q}_n(x), \qquad (5.33)$$

where

$$\mathbb{Q}_n(x) = \frac{\left(\sum_{r=0}^m (\Delta_r/\Delta)\phi_r(x)\right)^2}{s^2(\boldsymbol{\vartheta}^*, x)} \cdot \frac{1}{\left\{1 + a_n(\sum_{r=0}^m \Delta_r \phi_r(x))/s(\boldsymbol{\vartheta}^*, x)\right\}^3},$$

with a_n being a non–random number in the interval [0, 1]. With the above expansion, we then have

$$\begin{aligned}
&\frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{(\rho_i \varsigma_i + \Delta_i)\phi_i(\tilde{x}_{kj})}{s_n(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})} - \tilde{\int} \frac{(\rho_i \varsigma_i + \Delta_i)\phi_i(x)}{s_n(\boldsymbol{\vartheta}^*, x)} d\tilde{F}_k \right\} \\
=&(\rho_i \varsigma_i + \Delta_i)(a + \sum_{r=0}^m \Delta_r b_r + \Delta^2 c), \qquad (5.34)
\end{aligned}$$

where

$$a = \frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{\phi_i(\tilde{x}_{kj})}{\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})} - \tilde{\mathrm{E}}_k \frac{\phi_i(x)}{\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)} \right\},$$

$$b_r = \frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{\phi_i(\tilde{x}_{kj})\phi_r(\tilde{x}_{kj})}{\boldsymbol{s}^2(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})} - \tilde{\mathrm{E}}_k \frac{\phi_i(x)\phi_r(\boldsymbol{\vartheta}^*,\, x)}{\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)} \right\},$$

$$c = \frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{\phi_i(\tilde{x}_{kj})}{\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})} \mathbb{Q}_{\mathbb{n}}(\tilde{x}_{kj}) - \tilde{\int} \left(\frac{\phi_i(x)}{s(\boldsymbol{\vartheta}^*,\, x)} \mathbb{Q}_{\mathbb{n}}(x) \right) d\tilde{F}_k \right\}.$$

Now, for any given i and k, $\{\phi_i(x_{kj})\}_{j=1}^{\infty}$ is an iid sequence and, by (5.27),

$$\phi_i^2(x)/\boldsymbol{s}^2(\boldsymbol{\vartheta}^*,\, x) < 1/(\min_{0 \le r \le m} \rho_r \varsigma_r)^2,$$

so by the central limit theorem, term a is of $O_p(1)$. Similarly, b_r is of $O_p(1)$ for each r. Recalling that $\Delta_r = o(1)$, we have

$$\sum_{r=0}^{m} \Delta_r b_r = o_p(1).$$

We then look at term the c. By bound (5.27), (5.28) and (5.29), we have, for all large enough n and all x,

$$|\{\phi_i(x)/\boldsymbol{s}(\boldsymbol{\vartheta}^*,\, x)\}\mathbb{Q}_n(x)| < 8/(\min_{0 \le r \le m} \rho_r \varsigma_r)^3.$$

Consequently,

$$
\begin{aligned}
|c| &= \frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left| \frac{\phi_i(\tilde{x}_{kj})}{s(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})} \mathbb{Q}_{\mathbb{n}}(\tilde{x}_{kj}) - \tilde{\int} \left(\frac{\phi_i(x)}{s(\boldsymbol{\vartheta}^*,\, x)} \mathbb{Q}_{\mathbb{n}}(x) \right) d\tilde{F}_k \right| \\
&\leq \frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \left| \frac{\phi_i(\tilde{x}_{kj})}{s(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})} \mathbb{Q}_{\mathbb{n}}(\tilde{x}_{kj}) \right| + \tilde{\int} \left| \frac{\phi_i(x)}{s(\boldsymbol{\vartheta}^*,\, x)} \mathbb{Q}_{\mathbb{n}}(x) \right| d\tilde{F}_k \right\} \\
&< \frac{16\sqrt{n_k}}{(min_{0 \leq l \leq m} \rho_l)^3}.
\end{aligned}
$$

Recall that $\Delta = O(n^{-\delta})$. When $\delta = 1/3$, we have

$$
\Delta^2 c = O(n^{-2/3}) \cdot O(n_k^{1/2}) = o(1).
$$

With the above orders of a, b_r and c, and the expression (5.34), we know that equality (5.32) holds. The terms in the curly brackets on the RHS of (5.32) are iid across j, so the LHS of that equality has a normal limiting distribution. It follows that the item in the curly brackets of (5.30) has a normal limiting distribution:

$$
\frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{\partial \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, \tilde{x}_{kj})}{\partial \boldsymbol{\vartheta}} - \tilde{\int} \frac{\partial \mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*,\, x)}{\partial \boldsymbol{\vartheta}} d\tilde{F}_k \right\} \longrightarrow N(0,\, \mathcal{V}_k)
$$

in distribution, where

$$
\mathcal{V}_k = \tilde{\mathrm{E}}_k \left\{ \frac{\partial \mathcal{L}_k(\boldsymbol{\vartheta}^*,\, x)}{\partial \boldsymbol{\vartheta}} \frac{\partial \mathcal{L}_k(\boldsymbol{\vartheta}^*,\, x)}{\partial \boldsymbol{\vartheta}^\intercal} \right\} + \tilde{\mathrm{E}}_k \left\{ \frac{\partial \mathcal{L}_k(\boldsymbol{\vartheta}^*,\, x)}{\partial \boldsymbol{\vartheta}} \right\} \tilde{\mathrm{E}}_k \left\{ \frac{\partial \mathcal{L}_k(\boldsymbol{\vartheta}^*,\, x)}{\partial \boldsymbol{\vartheta}^\intercal} \right\}.
$$

By the above asymptotic nomality, (5.30) and $\sqrt{\tilde{n}_k/n} = \sqrt{\rho_k \varsigma_k} + o(1)$, we have

$$
\boldsymbol{v} \longrightarrow N(0,\, \mathcal{V})
$$

with $\mathcal{V} = \sum_{k=0}^{m} \rho_k \varsigma_k \mathcal{V}_k$. As the proof of Theorem 2.2 for complete data, we

find $\mathcal{V} = \mathcal{U} - \mathcal{U}\mathcal{W}\mathcal{U}$. The proof is complete.

Remark 5.1. In the above proof, we assumed $n_k/n = \rho_k + O(n^{-\delta})$ with $\delta = 1/3$. We now show the general case of $0 < \delta < 1/3$.

Recall that we have shown $\tilde{n}_k/n = \rho_k + O(n^{-\delta})$. Note that the only place we used the order of $O(n^{-1/3})$ for $\Delta_k = \tilde{n}_k/n - \rho_k$ is in the proof of equality (5.32). The key to showing (5.32) lies in the expansion (5.33). Now, suppose $\Delta_k = O(n^{-\delta})$ for some $1/4 > \delta > 0$. Let $\mathbb{T} = \lceil 1/(2\delta) \rceil + 1$, where $\lceil \cdot \rceil$ is the ceiling function. We expand $1/\{1 + (\sum_{r=0}^{m} \Delta_r \phi_r(x))/\boldsymbol{s}(\boldsymbol{\vartheta}^*, x)\}$ to the $\mathbb{T}^{th}$ order instead of the 2nd order as in (5.33). Then similar to expansion (5.34),

$$\frac{1}{\sqrt{\tilde{n}_k}} \sum_{j=1}^{\tilde{n}_k} \left\{ \frac{(\rho_i \varsigma_i + \Delta_i)\phi_i(\tilde{x}_{kj})}{\boldsymbol{s}_n(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})} - \tilde{\int} \frac{(\rho_i \varsigma_i + \Delta_i)\phi_i(x)}{\boldsymbol{s}_n(\boldsymbol{\vartheta}^*, x)} d\tilde{F}_k \right\}$$

correspondingly has a $\mathbb{T}^{th}$ order expansion of the form

$$(\rho_i \varsigma_i + \Delta_i)(a_1 + a_2 + \ldots + a_{\mathbb{T}} + r^{\mathbb{T}} a_{\mathbb{T}+1}). \tag{5.35}$$

The leading term a_1 is exactly the same as the leading term a in (5.34). Each a_t, $t = 2, 3, \ldots, \mathbb{T}$, just like term $\sum_{r=0}^{m} \Delta_r b_r$ in (5.34), by multinomial theorem, is a finite sum of $o_p(1)$ terms, so is also $o_p(1)$. Lastly, the residual term $a_{\mathbb{T}+1}$, similar to the c in (5.34), can be shown to be bounded by

$$\frac{2^{\mathbb{T}+2}\sqrt{n_k}}{(min_{0 \le r \le m} \rho_r \varsigma_r)^{\mathbb{T}+1}}.$$

This bound, along with $\Delta = \max_{0 \le r \le m} |\Delta_i| = O(n^{-\delta})$ and $\mathbb{T} = \lceil 1/(2\delta) \rceil + 1$, gives us

$$\Delta^{\mathbb{T}} a_{\mathbb{T}+1} = o(1).$$

Then, in view of (5.35), we conclude that, even when $n_k/n = \rho_k + O(n^{-\delta})$ for an arbitary $\delta > 0$, equality (5.32) is still true and so Theorem 2.2 still holds.

5.8.3 Theorem 5.4: Asymptotic normality of the MPELE

Based on properties of the partial information matrix (Lemma 5.2) and the score function (Lemma 5.3), as the proof of Lemma 2.4 for complete data (Section 2.5.2), we can show that the DPEL $\ell_n(\boldsymbol{\vartheta})$ attains a maximum in the interior of a $\sqrt[3]{n}$–neighbourhood of the true parameter value $\boldsymbol{\vartheta}^*$. Along with the fact that the DPEL is concave, we conclude that all the maxima of DPEL must be in the interior of that neighbourhood. Consequently, the MPELE, which is a maximum of the DPEL, must be $\sqrt[3]{n}$–consistent.

Recall that the partial empirical information matrix is defined as

$$\mathcal{U}_n = -n^{-1}\partial^2\ell_n(\boldsymbol{\vartheta}^*)/\partial\boldsymbol{\vartheta}\partial\boldsymbol{\vartheta}^\mathsf{T}.$$

Expanding $n^{-1/2}\partial\ell_n(\hat{\boldsymbol{\vartheta}})/\partial\boldsymbol{\vartheta}$ around $\boldsymbol{\vartheta}^*$, we get

$$n^{-1/2}\partial\ell_n(\hat{\boldsymbol{\vartheta}})/\partial\boldsymbol{\vartheta} = \boldsymbol{v} - \mathcal{U}_n\{\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*)\} + o_p(1),$$

where the third–order residue term is of $o_p(1)$ because $\hat{\boldsymbol{\vartheta}}$ is $\sqrt[3]{n}$–consistent and the third–order derivatives of the DPEL $\ell_n(\boldsymbol{\vartheta})$ are bounded by an integrable function as implied by (5.20). Since $\hat{\boldsymbol{\vartheta}}$ is the point at which $\ell_n(\boldsymbol{\vartheta})$ is maximized, the LHS of the above expansion is $\mathbf{0}$. Reorganizing terms and recalling that $\mathcal{U}_n$ converges to $\mathcal{U}$ almost surely by Lemma 5.2, we easily get

$$\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*) = \mathcal{U}^{-1}\boldsymbol{v} + o_p(1). \tag{5.36}$$

The claimed asymptotic normality of $\sqrt{n}(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*)$ then follows from the asymptotic normality of $\boldsymbol{v}$, which is given by Lemma 5.3.

5.8.4 Theorem 5.5: Asymptotic properties of the ELR test

Proof of Theorem 5.5 (i). Recall that the ELR statistic R_n equals the DPEL ratio statistic given by $R_n = 2\{\ell_n(\hat{\boldsymbol{\vartheta}}) - \ell_n(\tilde{\boldsymbol{\vartheta}})\}$. As the proof of Theorem 3.1 for complete data, the idea is to find suitable quadratic expansions for $\ell_n(\hat{\boldsymbol{\vartheta}})$ and $\ell_n(\tilde{\boldsymbol{\vartheta}})$ under the null model, and show that the difference of the two, which equals the ELR statistic R_n, has a chi–square limiting distribution.

Expanding $\ell_n(\hat{\boldsymbol{\vartheta}})$ around $\boldsymbol{\vartheta}^*$, we get

$$\ell_n(\hat{\boldsymbol{\vartheta}}) = \ell_n(\boldsymbol{\vartheta}^*) + \sqrt{n}\boldsymbol{v}^\intercal(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*) - (1/2)n(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*)^\intercal U_n(\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^*) + o_p(1),$$

where the last term is of $o_p(1)$ since $\hat{\boldsymbol{\vartheta}} - \boldsymbol{\vartheta}^* = O_p(n^{-1/2})$ and the third derivatives of $\ell_n(\hat{\boldsymbol{\vartheta}})$ are bounded by an integrable function. Combining the above expansion with (5.36) and using the fact that $\mathcal{U}_n = \mathcal{U} + o_p(1)$, we obtain

$$\ell_n(\hat{\boldsymbol{\vartheta}}) = \ell_n(\boldsymbol{\vartheta}^*) + (1/2)\boldsymbol{v}^\intercal \mathcal{U}^{-1}\boldsymbol{v} + o_p(1).$$

We then give an expansion of $\ell_n(\tilde{\boldsymbol{\vartheta}})$ under the null model $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$. As noted in Section 3.2.2, when $q < md$, the null model is equivalently to $\boldsymbol{\beta} = \mathcal{G}(\boldsymbol{\gamma})$ for some function $\mathcal{G}$: $\mathbb{R}^{md-q} \to \mathbb{R}^{md}$ and parameter $\boldsymbol{\gamma}$ of dimension $md - q$. In addition, $\mathcal{G}$ is thrice differentiable, and its Jacobian matrix $J = \partial\mathcal{G}(\boldsymbol{\gamma}^*)/\partial\boldsymbol{\gamma}$ is of full rank. Using exactly the same technique for the proof of Theorem 3.1, we find the follow expansion for the DPEL under the null model

$$\ell_n(\tilde{\boldsymbol{\vartheta}}) = \ell_n(\boldsymbol{\vartheta}^*) + (1/2)\tilde{\boldsymbol{v}}^\intercal \tilde{\mathcal{U}}^{-1}\tilde{\boldsymbol{v}} + o_p(1),$$

where $\tilde{\boldsymbol{v}} = \{\mathrm{diag}(I_m,\, J)\}^\intercal \boldsymbol{v}$ and $\tilde{\mathcal{U}} = \{\mathrm{diag}(I_m,\, J)\}^\intercal \mathcal{U}\{\mathrm{diag}(I_m,\, J)\}$.

With the above expansions of $\ell_n(\hat{\boldsymbol{\vartheta}})$ and $\ell_n(\tilde{\boldsymbol{\vartheta}})$, we then get

$$R_n = 2\{\ell_n(\hat{\boldsymbol{\vartheta}}) - \ell_n(\tilde{\boldsymbol{\vartheta}})\} = \boldsymbol{v}^\intercal \mathcal{U}^{-1} \boldsymbol{v} - \tilde{\boldsymbol{v}}^\intercal \tilde{\mathcal{U}}^{-1} \tilde{\boldsymbol{v}} + o_p(1).$$

Applying the quadratic form decomposition formula given in Lemma 3.4 to the above two quadratic forms and after cancelling terms, R_n finally simplifies to

$$R_n = \tilde{\boldsymbol{\xi}}^\intercal \{\tilde{\Lambda}^{-1} - J(J^\intercal \tilde{\Lambda} J)^{-1} J^\intercal\} \tilde{\boldsymbol{\xi}} + o_p(1), \tag{5.37}$$

where $\tilde{\boldsymbol{\xi}} = (-\mathcal{U}_{\boldsymbol{\beta\kappa}} \mathcal{U}_{\boldsymbol{\kappa\kappa}}^{-1},\ I_{md}) \boldsymbol{v}$ and $\tilde{\Lambda} = \mathcal{U}_{\boldsymbol{\beta\beta}} - \mathcal{U}_{\boldsymbol{\beta\kappa}} \mathcal{U}_{\boldsymbol{\kappa\kappa}}^{-1} \mathcal{U}_{\boldsymbol{\kappa\beta}}$ is defined in Theorem 5.5 (ii). The above expression of R_n is the same as that of DEL ratio statistic based on complete samples given in (3.10), so using the same proving technique, we find that R_n has the claimed chi–square limiting distribution. □

Proof of Theorem 5.5 (ii). Let $\boldsymbol{\beta}^*$ be a specific parameter value under the null hypothesis and $\{F_k\}$ be the corresponding distribution functions. Let $\{G_k\}$ be the set of distribution functions satisfying the DRM with parameter given by the alternative model $\boldsymbol{\beta}_k = \boldsymbol{\beta}_k^* + n_k^{-1/2} \boldsymbol{c}_k$, $k = 1, \ldots, m$, and $G_0 = F_0$. Denote the distributions of the uncensored observations under the null model and the local alternative model as $\{\tilde{F}_k\}$ and $\{\tilde{G}_k\}$, respectively. When the samples are generated from the $\{G_k\}$, ELR statistic still follows the expansion (5.37), just like what we have shown in the proof of Theorem 3.2 for complete samples. The limiting distribution of R_n is therefore again determined by that of $\boldsymbol{v} = n^{-1/2} \partial \ell_n(\boldsymbol{\vartheta}^*)/\partial \boldsymbol{\vartheta}$, which can be found by using Le Cam's third lemma.

We now derive the limiting distribution of $\boldsymbol{v}$ under the local alternative distributions $\{G_k\}$. Let $\tilde{\boldsymbol{w}}_k = \sum_{j=1}^{\tilde{n}_k} \log\{d\tilde{G}_k(\tilde{x}_{kj})/d\tilde{F}_k(\tilde{x}_{kj})\}$. Note that $\boldsymbol{v}$ involves only uncensored observations $\{\tilde{x}_{kj}\}$, so by Le Cam's third lemma, the key to find this limiting distribution lies in finding the joint limiting distribution of $\boldsymbol{v}$ and $\sum_{k=0}^{m} \tilde{\boldsymbol{w}}_k$ under the null limiting distributions for uncensored

observations $\{\tilde{F}_k\}$.

We first work on an expansion for $\sum_{k=0}^{m} \tilde{\boldsymbol{w}}_k$. For each $k = 0, 1, \ldots, m$, let $\tilde{\mathrm{VAR}}_k(\cdot)$ and $\tilde{\mathrm{COV}}_k(\cdot)$ be the variance and covariance operators with respect to $\tilde{F}_k$, respectively. Just as the proof of Lemma 3.5 for complete data, we find that

$$\log\{d\tilde{G}_k(x)/d\tilde{F}_k(x)\} = n_k^{-1/2}\boldsymbol{c}_k^{\intercal}\{\boldsymbol{q}(x) - \tilde{\boldsymbol{\nu}}_k\} - (2n_k)^{-1}\boldsymbol{c}_k^{\intercal}\tilde{\boldsymbol{\sigma}}_k\boldsymbol{c}_k + O(n^{-3/2})$$

uniformly in x, where $\tilde{\boldsymbol{\nu}}_k = \tilde{\mathrm{E}}_k\boldsymbol{q}(x)$ and $\tilde{\boldsymbol{\sigma}}_k = \tilde{\mathrm{VAR}}_k(\boldsymbol{q}(x))$. Therefore we have

$$\begin{aligned}
\tilde{\boldsymbol{w}}_k &= \sum_{j=1}^{\tilde{n}_k} \log\{d\tilde{G}_k(\tilde{x}_{kj})/d\tilde{F}_k(\tilde{x}_{kj})\} \\
&= (\tilde{n}_k/n_k)^{1/2}\boldsymbol{c}_k^{\intercal}\big\{\tilde{n}_k^{-1/2}\sum_{j=1}^{\tilde{n}_k}\{\boldsymbol{q}(\tilde{x}_{kj}) - \tilde{\boldsymbol{\nu}}_k\}\big\} - (1/2)(\tilde{n}_k/n_k)\boldsymbol{c}_k^{\intercal}\tilde{\boldsymbol{\sigma}}_k\boldsymbol{c}_k + O(n^{-1/2}) \\
&= \varsigma_k^{1/2}\boldsymbol{c}_k^{\intercal}\big\{\tilde{n}_k^{-1/2}\sum_{j=1}^{\tilde{n}_k}\{\boldsymbol{q}(\tilde{x}_{kj}) - \tilde{\boldsymbol{\nu}}_k\}\big\} - (1/2)\varsigma_k\boldsymbol{c}_k^{\intercal}\tilde{\boldsymbol{\sigma}}_k\boldsymbol{c}_k + o_p(1).
\end{aligned}$$

where the last equality is by $\tilde{n}_k/n_k = \varsigma_k + o(1)$ and $\tilde{n}_k^{-1/2}\sum_{j=1}^{\tilde{n}_k}\{\boldsymbol{q}(\tilde{x}_{kj}) - \tilde{\boldsymbol{\nu}}_k\} = O_p(1)$.

With the above expansion of $\tilde{\boldsymbol{w}}_k$, the expression (5.30) of $\boldsymbol{v}$, the expansion (5.31) and the fact that $\tilde{n}_k/n = \rho_k\varsigma_k + o(1)$, we get the following joint expansion for $\boldsymbol{v}$ and $\sum_{k=0}^{m}\tilde{\boldsymbol{w}}_k$,

$$\begin{aligned}
\begin{pmatrix} \boldsymbol{v} \\ \sum_k \tilde{\boldsymbol{w}}_k \end{pmatrix} &= \sum_{k=0}^{m}\frac{1}{\sqrt{\tilde{n}_k}}\sum_{j=1}^{\tilde{n}_k}\begin{pmatrix} (\rho_k\varsigma_k)^{1/2}\big\{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, \tilde{x}_{kj})/\partial\boldsymbol{\vartheta} - \tilde{\mathrm{E}}_k\{\partial\mathcal{L}_{n,k}(\boldsymbol{\vartheta}^*, x)/\partial\boldsymbol{\vartheta}\}\big\} \\ \varsigma_k^{1/2}\boldsymbol{c}_k^{\intercal}\{\boldsymbol{q}(x_{kj}) - \boldsymbol{\nu}_k\} \end{pmatrix} \\
&\quad - \sum_{k=0}^{m}\begin{pmatrix} \boldsymbol{0} \\ (1/2)\varsigma_k\boldsymbol{c}_k^{\intercal}\tilde{\boldsymbol{\sigma}}_k\boldsymbol{c}_k \end{pmatrix} + o_p(1).
\end{aligned}$$

Hence $\boldsymbol{v}$ and $\sum_{k=0}^{m}\tilde{\boldsymbol{w}}_k$ have a joint normal limiting distribution with mean

vector and covariance matrix given by

$$\left(\mathbf{0}^{\intercal},\ -\frac{1}{2}\sum_k \varsigma_k \boldsymbol{c}_k^{\intercal}\tilde{\boldsymbol{\sigma}}_k\boldsymbol{c}_k\right)^{\intercal} \text{ and } \begin{pmatrix} \mathcal{V} & \tilde{\boldsymbol{\tau}} \\ \tilde{\boldsymbol{\tau}}^{\intercal} & \sum_k \varsigma_k \boldsymbol{c}_k^{\intercal}\tilde{\boldsymbol{\sigma}}_k\boldsymbol{c}_k \end{pmatrix},$$

with

$$\tilde{\boldsymbol{\tau}} = \sum_{k=1}^{m} {\rho_k}^{1/2}\varsigma_k \mathrm{Cov}_k\{\partial\mathcal{L}_k(\boldsymbol{\vartheta}^*,\, x)/\partial\boldsymbol{\vartheta}, \boldsymbol{q}^{\intercal}(x)\}\boldsymbol{c}_k$$

Because the second entry of the mean vector equals negative half of the lower–right entry of the covariance matrix, the condition of Le Cam's third lemma is satisfied. By that lemma, we conclude that

$$\boldsymbol{v} \longrightarrow N(\tilde{\boldsymbol{\tau}},\, \mathcal{V})$$

in distribution, under the local alternative distributions $\{G_k\}$.

We have argued at the beginning of the proof that, under the $\{G_k\}$, the ELR statistics R_n is still approximated by $\tilde{\boldsymbol{\xi}}^{\intercal}\{\tilde{\Lambda}^{-1} - J(J^{\intercal}\tilde{\Lambda}J)^{-1}J^{\intercal}\}\tilde{\boldsymbol{\xi}}$ with $\tilde{\boldsymbol{\xi}} = (-\mathcal{U}_{\boldsymbol{\beta\kappa}}\mathcal{U}_{\boldsymbol{\kappa\kappa}}^{-1},\ I_{md})\boldsymbol{v}$. The vector $\tilde{\boldsymbol{\xi}}$ has a normal limiting distribution because $\boldsymbol{v}$ has one as we have just shown. Based on this result, just like the proof of Theorem 3.2 for complete samples, the above quadratic form is found to have the claimed non–central chis–square limiting distribution. This completes the proof. □

5.9 Appendix: Some thoughts on the weighted EL inference for Type I censored samples

For inference under the DRM for two samples with randomly censored observations, Ren (2008) proposed to use the so–called *weighted empirical likelihood* (WEL) function. This leads us to wonder whether this WEL is also

a useful tool for inference based on Type I censored samples. We find that in fact it results in an inconsistent estimator for the DRM scaling parameter $\boldsymbol{\alpha}$, as demonstrated in this section.

Suppose the observations are randomly right–censored. The idea of the WEL is to construct a likelihood type function based on the Kaplan–Meier estimator (Kaplan and Meier, 1958) of the $\{F_k\}$, which is defined as

$$\check{F}_k(t) = \sum_{j=1}^{s_k} w_{kj} \mathbb{1}\{y_{kj} \leq t\},$$

where s_k is the total number of distinct uncensored observations in the kth sample and $y_{k1} < y_{k2} < \cdots < y_{ks_k}$ are the ordered values of those distinct uncensored observations. The $\{w_{kj}\}_{j=1}^{s_k}$ are a set of positive weights. Let d_{kj}, $j = 1, \cdots, s_k$, be the number of failures at y_{kj} in the kth sample, and r_{kj} be the number at risk just prior to y_{kj} in that sample. The weights $\{w_{kj}\}$ are given by

$$w_{k1} = \frac{d_{k1}}{r_{k1}} \quad \text{and} \quad w_{kj} = \frac{d_{kj}}{r_{kj}} \prod_{l=1}^{j-1} \frac{r_{kl} - d_{kl}}{r_{kl}}, \; j = 2, \ldots, s_k.$$

and for a positive integer $t \leq s_k$, we have

$$\sum_{j=1}^{t} w_{kj} = 1 - \prod_{j=1}^{t} \frac{r_{kj} - d_{kj}}{r_{kj}}. \tag{5.38}$$

When the largest censored observation is larger than the largest uncensored observation, $\sum_{j=1}^{s_k} w_{kj} < 1$ and $\check{F}_k$ is a not proper CDF.

For the $\{F_k\}$ that satisfy the DRM assumption, Ren (2008) defined the

WEL to be

$$
\begin{aligned}
&L_n^{(w)}(F_0,\, \boldsymbol{\alpha},\, \boldsymbol{\beta}) \\
=&\prod_{k=0}^{m}\prod_{j=1}^{s_k}\{dF_k(y_{kj})\}^{n_k w_{kj}} \\
=&\left\{\prod_{k=0}^{m}\prod_{j=1}^{s_k}\left\{dF_0(y_{kj})\right\}^{n_k w_{kj}}\right\}\cdot\exp\left\{\sum_{k=1}^{m}\sum_{j=1}^{s_k} n_k w_{kj}\left\{\alpha_k+\boldsymbol{\beta}_k\boldsymbol{q}(y_{kj})\right\}\right\}.
\end{aligned}
\tag{5.39}
$$

The corresponding profile log–WEL is defined as

$$
\begin{aligned}
l_n^{(w)}(\boldsymbol{\alpha},\, \boldsymbol{\beta}) = \sup_{F_0}\{\log L_n^{(w)}(F_0,\, \boldsymbol{\alpha},\, \boldsymbol{\beta}) :\ & \sum_{k=1}^{m}\sum_{j=1}^{s_k}\exp\left\{\alpha_l+\boldsymbol{\beta}_l\boldsymbol{q}(y_{kj})\right\}dF_0(y_{kj}) = 1, \\
& l = 0,\, 1,\, \cdots,\, m\}.
\end{aligned}
$$

In the case of two samples with randomly right–censored observations, Ren showed that, under suitable conditions, the maximum WEL estimator (MWELE) of $(\boldsymbol{\alpha},\, \boldsymbol{\beta})$ is consistent and asymptotically normal, and for a special form of the density ratio, the corresponding likelihood ratio statistic has a scaled chi–square limiting distribution.

We now look at a direct adaptation of the WEL to the case of Type I right–censored samples. For Type I right–censored samples, $r_{kj}-d_{kj} = r_{k(j+1)}$ for $j = 1,\, \ldots,\, s_k - 1$, and $r_{ks_k} - d_{ks_k} = n_k - \tilde{n}_k$. Hence, by (5.38), we get

$$
\sum_{j=1}^{s_k} w_{kj} = \frac{\tilde{n}_k}{n_k}.
$$

Note that $\tilde{n}_k/n_k \to \varsigma_k$ almost surely as $n \to \infty$, so when $\varsigma_k < 1$, $\sum_{j=1}^{s_k} w_{kj} < 1$ almost surely for all large n, and consequently $\check{F}_k$ is not a proper CDF. In

this case, the constraint in the profile log–WEL should be changed to

$$\sum_{k=1}^{m}\sum_{j=1}^{s_k}\exp\left\{\alpha_l+\boldsymbol{\beta}_l\boldsymbol{q}(y_{kj})\right\}\mathbb{1}(y_{kj}\in\mathcal{S}_l)dF_0(y_{kj})=\varsigma_l, \tag{5.40}$$

because the WEL is essentially constructed using the uncensored observations only. Assuming no ties in the uncensored observations, then $s_k=\tilde{n}_k$ and $w_{kj}=1/n_k$ for all $j=1,\ldots,\tilde{n}_k$. Now, the WEL (5.39) can be written as

$$\begin{aligned}
&L_n^{(w)}(F_0,\,\boldsymbol{\alpha},\,\boldsymbol{\beta})\\
&=\left\{\prod_{k=0}^{m}\prod_{j=1}^{s_k}\left\{dF_0(y_{kj})\right\}^{n_kw_{kj}}\right\}\cdot\exp\left\{\sum_{k=1}^{m}\sum_{j=1}^{s_k}n_kw_{kj}\left\{\alpha_k+\boldsymbol{\beta}_k\boldsymbol{q}(y_{kj})\right\}\right\}\\
&=\left\{\prod_{k=0}^{m}\prod_{j=1}^{\tilde{n}_k}dF_0(\tilde{x}_{kj})\right\}\exp\left\{\sum_{k=1}^{m}\sum_{j=1}^{\tilde{n}_k}\left\{\alpha_k+\boldsymbol{\beta}_k\boldsymbol{q}(\tilde{x}_{kj})\right\}\mathbb{1}(\tilde{x}_{kj}\in\mathcal{S}_k)\right\},
\end{aligned}$$

and the corresponding constraint (5.40) can alternatively be written as

$$\sum_{k=0}^{m}\sum_{j=1}^{\tilde{n}_k}\exp\{\alpha_l+\boldsymbol{\beta}_l^{\mathsf{T}}\boldsymbol{q}(\tilde{x}_{kj})\}\mathbb{1}(\tilde{x}_{kj}\in\mathcal{S}_l)dF_0(\tilde{x}_{kj})=\varsigma_l.$$

Strikingly, this WEL has the same expression as the PEL (5.6) except that the $\wp_{kj}$, which is $d\tilde{F}_0(\tilde{x}_{kj})$, and κ_k in (5.6) are replaced by $dF_0(\tilde{x}_{kj})$ and α_k here. The profile log–WEL of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is then defined to be

$$\begin{aligned}
l_n^{(w)}(\boldsymbol{\alpha},\,\boldsymbol{\beta})=\sup_{F_0,\,\{\varsigma_k\}}\Big\{&\log L_n^{(w)}(F_0,\,\boldsymbol{\alpha},\,\boldsymbol{\beta}):\\
&\sum_{k=0}^{m}\sum_{j=1}^{\tilde{n}_k}\exp\{\alpha_l+\boldsymbol{\beta}_l^{\mathsf{T}}\boldsymbol{q}(\tilde{x}_{kj})\}\mathbb{1}(\tilde{x}_{kj}\in\mathcal{S}_l)dF_0(\tilde{x}_{kj})=\varsigma_l,\\
&dF_0(\tilde{x}_{kj})\geq 0,\ 0<\varsigma_l\leq 1,\quad l=0,\ldots,m\Big\}.
\end{aligned}$$

Clearly, the $L_n^{(w)}(F_0,\,\boldsymbol{\alpha},\,\boldsymbol{\beta})$ is functionally independent of the $\{\varsigma_l\}$, so the

above supremum is attained on the boundary of space of ς_l's when $\varsigma_l = 1$ for all $l = 0, 1, \ldots, m$. Hence,

$$
\begin{aligned}
l_n^{(w)}(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sup_{F_0} \Big\{ l_n^{(w)}(F_0, \boldsymbol{\alpha}, \boldsymbol{\beta}) : & \\
& \sum_{k=0}^{m} \sum_{j=1}^{\tilde{n}_k} \exp\{\alpha_l + \boldsymbol{\beta}_l^{\intercal} \boldsymbol{q}(\tilde{x}_{kj})\} \mathbb{1}(\tilde{x}_{kj} \in \mathcal{S}_l) dF_0(\tilde{x}_{kj}) = 1, \\
& dF_0(\tilde{x}_{kj}) \geq 0, \ l = 0, \ldots, m \Big\}.
\end{aligned}
$$

Again, $l_n^{(w)}(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has the same mathematical expression as that for the profile log–PEL (5.8), except that the $\wp_{kj}$ and κ_k in (5.8) are now replaced by $dF_0(\tilde{x}_{kj})$ and α_k. Hence the MWELE of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ has exactly the same value as the MPELE of $(\boldsymbol{\kappa}, \boldsymbol{\beta})$. Based on this result and the fact that the MPELE of $(\boldsymbol{\kappa}, \boldsymbol{\beta})$ is a consistent estimator, we conclude that the MWELE of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ converges to $(\boldsymbol{\kappa}^*, \boldsymbol{\beta}^*)$ as the total sample size n goes to infinity. Since in general $\kappa_k^* = \alpha_k^* + \log \varsigma_0^* - \log \varsigma_k^* \neq \alpha_k^*$, where ς_k^* is the true value of ς_k, the MWELE of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ is not a consistent estimator.

Chapter 6

R software package "drmdel" for DEL inference under the DRM

This chapter introduces and illustrates the use of an R software package, *drmdel*, that we wrote for the DEL inference under the DRM based on multiple complete samples. This package can be used to calculate the MELE of the DRM parameter, perform the DELR test, estimate population distribution functions, estimate quantiles of the population distributions as found in Chen and Liu (2013), compare quantiles from different distributions using a Wald test, and estimate densities of different populations as found in Fokianos (2004). The package and its manual can be download from The Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/drmdel/index.html; alternatively, it can be installed within R using command `install.packages("drmdel")`.

An extension of this package to the case of Type I censored samples is under development by the time this thesis is written and will soon be available.

6.1 Under the hood: consideration and implementation

As we have seen in earlier chapters, the first and a key step of EL inference under the DRM is to compute the MELEs of the DRM parameter $\boldsymbol{\theta}$ and the baseline distribution F_0. For multiple complete samples, the MELE $\hat{\boldsymbol{\theta}}$

can be computed as the point at which the concave function DEL (2.11) is maximized. After obtaining $\hat{\boldsymbol{\theta}}$, the MELE $\{\hat{p}_{kj}\}$ of the baseline distribution can then be computed through (2.9).

The core of the above computational procedure is a smooth concave maximization problem, and for such a problem, the quasi–Newton methods (Nocedal and Wright, 2006) are probably the most popular methods because they are fast, reliable, and implemented in most computational software packages. A quasi–Newton method has a super–linear convergence rate, i.e. faster than linear rate but slower than quadratic rate, and for each iteration, its computational complexity is $O(p^2+p\mathcal{C}_f)$, where p is the dimension of the parameter of the function to be optimized and $\mathcal{C}_f$ is the number of operations needed for one function evaluation. Our particular choice is the famous Broyden–Fletcher–Goldfarb–Shanno (BFGS) algorithm, a quasi–Newton method simultaneously discovered by Broyden (1970), Fletcher (1970), Goldfarb (1970) and Shanno (1970), because it is well ingrained in the R `optim` function.

As noted above, the speed of the maximization procedure is influenced by the speed of function evaluation. And function evaluation becomes the main factor when the dimension p of the parameter is not very high. The evaluation of the object function DEL (2.11) is straightforward but involves a multiple summation. With such a structure, an R implementation of function evaluation will be either cumbersome without loops or very slow with loops. We hence implement the function and its Jacobian evaluation in C for both speed and convenience. The evaluated function and its Jacobian then are passed to the R `optim` function for maximization. In principle, the speed of the optimization can be increased even more if we use a "BFGS" procedure implemented in C to save the communication time between R and C. In practice, we found such an approach does not provides a noticeable speed gain. Moreover, since our intention is to write an R package, an R implementation of the algorithm is supposed to provide extra reliability, and R users who are familiar with `optim` function will find it more convenient.

These considerations lead to our final implementation of function and its Jacobian evaluation in C, and optimization in R.

The computation of $\hat{p}_{kj}$ (2.9) is also implemented in C for maximal computational efficiency. With this estimator of $dF_0(x_{kj})$, by applying the DRM assumption (2.1), we estimate $dF_r(x_{kj})$, $r = 1, \ldots, , m$, by

$$\hat{p}_{kj}^{(r)} = d\hat{F}_r(x_{kj}) = \exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^{\mathsf{T}} \boldsymbol{q}(x_{kj})\}\hat{p}_{kj}.$$

The other inferences tasks described in the sequel are all based on the values $\hat{\boldsymbol{\theta}}$ and $\{\hat{p}_{kj}^{(r)}\}$, so mostly implemented in R unless otherwise noted.

6.2 DRM fitting

The primary function of the package is `drmdel`, which fits a DRM to data, calculates the MELEs $\hat{\boldsymbol{\theta}}$ and $\{\hat{p}_{kj}\}$, and performs the DELR test of Chapter 3 for hypotheses about the DRM parameter $\boldsymbol{\beta}$. This function should be used prior to any other function in the package for performing EL inference, since all others depend on the output of this one.

The function has the following generic form:

```
drmdel(x, n_samples, basis_func, g_null=NULL,
       g_null_jac=NULL, par_dim_null=NULL, ...).
```

The arguments are:

`x:` a long vector containing all $m+1$ multiple samples in the order of x_0, $x_1, \ldots, x_m$, where x_0 is the sample from the baseline distribution and x_k, $k = 1, \ldots, m$, is the kth non–baseline sample.

`n_samples:` a vector indicating the size of each sample, i.e. $(n_0, n_1, \ldots, n_m)$.

`basis_func:` the basis function $\boldsymbol{q}(x)$ of the DRM to be used. It could be either an integer between 1 and 11 or an R function. The integers represent 11 different built–in basis functions as follows:

`1:` $\boldsymbol{q}(x) = x$.

`2:` $\boldsymbol{q}(x) = \log|x|$.

`3:` $\boldsymbol{q}(x) = \sqrt{|x|}$.

`4:` $\boldsymbol{q}(x) = x^2$.

`5:` $\boldsymbol{q}(x) = (x,\, x^2)^\intercal$.

`6:` $\boldsymbol{q}(x) = (x,\, \log|x|)^\intercal$.

`7:` $\boldsymbol{q}(x) = (\log|x|,\, \sqrt{|x|},\, x)^\intercal$.

`8:` $\boldsymbol{q}(x) = (\log|x|,\, \sqrt{|x|},\, x^2)^\intercal$.

`9:` $\boldsymbol{q}(x) = (\log|x|,\, x,\, x^2)^\intercal$.

`10:` $\boldsymbol{q}(x) = (\sqrt{|x|},\, x,\, x^2)^\intercal$.

`11:` $\boldsymbol{q}(x) = (\log|x|,\, \sqrt{|x|},\, x,\, x^2)^\intercal$.

`g_null:` the function $\mathcal{G}$ specifying the null hypothesis of (6.1), if there is one. The default value of NULL represents that there is no hypothesis specified.

`g_null_jac:` the Jacobian matrix (first–order derivatives) of $\mathcal{G}$, if available.

`par_dim_null:` dimension of the null parameter $\boldsymbol{\gamma}$ in the hypothesis testing problem (6.1), if there is one.

`...:` further arguments to be passed to the R function `optim` for maximizing the DEL. See `help(optim)` for details. The current default values for “control$method” and “control$maxit” are set to "BFGS" and 10000, respectively.

The output of the function is an R list object containing lots of elements. A complete list of these output elements can be found by using R command `help(drmdel)`. We here only describe the most important ones as follows:

`mele:` the MELE of the DRM parameter in the form of $(\hat{\boldsymbol{\theta}}_1^T, \ldots, \hat{\boldsymbol{\theta}}_m^T)$, where $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^\intercal)^\intercal$.

`p_est:` the MELE $\{\hat{p}_{kj}^{(r)}\}$ of $\{dF_r(x_{kj})\}$, $r = 0, 1, \ldots, m$. This is a data frame with the following three columns:

- `k:` label of the populations, $k = 0, 1, \ldots, m$.
- `x:` data points. It specifies the value of x_{kj} at which $dF_k(x)$ is estimated.
- `p_est:` values of $\{\hat{p}_{kj}^{(r)}\}$ for $r = 0, 1, \ldots, m$.

`info_mat:` the estimated information matrix, $\hat{U}$.

`mele_null:` the MELE of the DRM parameters $\boldsymbol{\gamma}$ under the null model of (6.1), if available. This is a list object with two elements:

- `alpha:` the MELE of the DRM parameter $\boldsymbol{\alpha}$ under the null model.
- `gamma:` the MELE of the null parameter $\boldsymbol{\gamma}$.

`delr:` the value of the DELR statistic for the hypothesis testing problem (6.1). When no hypothesis (`g_null`) is given, this is the value of the DELR statistic for testing the hypothesis that all the distribution functions are equal.

`df:` degrees of freedom of the chi–square limiting distribution of the DELR statistic under the null model.

`p_val:` p–value of the DELR test.

We now give an example to illustrate the use of the `drmdel` function for DRM fitting.

Example 6.1 (Fitting a DRM and calculating the MELE)**.** *Suppose we have* $m + 1 = 5$ *samples generated from gamma distributions as follow.*

```
set.seed(25)

# sample sizes
n_samples <- c(100, 200, 180, 150, 175)  # sample sizes

# data generation
x0 <- rgamma(n_samples[1], shape=5, rate=1.8)
x1 <- rgamma(n_samples[2], shape=12, rate=1.2)
x2 <- rgamma(n_samples[3], shape=12, rate=1.2)
x3 <- rgamma(n_samples[4], shape=18, rate=5)
x4 <- rgamma(n_samples[5], shape=25, rate=2.6)

# concatenating the samples to a long data vector
x <- c(x0, x1, x2, x3, x4)
```

We now fit a DRM to these samples. Recall that the appropriate DRM basis function for gamma distributions is $\boldsymbol{q}(x) = (x, \log x)^\intercal$*, which is the built–in basis function* `6` *of the* `drmdel` *function. We hence fit a DRM with such a basis function to the data as follows.*

```
# load the drmdel package into R
library(drmdel)

# fit the DRM to data
drmfit_ex1 <- drmdel(x=x, n_samples=n_samples, basis_func=6)

# checking the names of the outputs
names(drmfit_ex1)
```

As we have noted, the output of a fitted DRM object has lots of components, therefore hard to read. We hence provide a function, `summaryDRM`*, to accompany* `drmdel`*, which reads a fitted DRM object and gives a nicely formatted summary of its output.*

```
# checking the summary of the fitted DRM object
summaryDRM(drmfit_ex1)
```

The returned summary is:

```
Basic information about the DRM:
    Number of samples (m+1): 5
    Basis function: 6
    Dimension of the basis function (d): 2
    Sample sizes: 100 200 180 150 175
    Sample proportions (rho): 0.124 0.248 0.224 0.186 0.217

Maximum empirical likelihood estimator (MELE) of the DRM
parameter:
   alpha[] beta[,1] beta[,2]
F1  -22.14  -0.1935     13.5
F2  -19.49   0.0243     11.4
F3   -4.95  -4.6389     17.8
F4  -32.88  -1.2457     22.8

Default hypothesis testing problem:
    H_0: All distribution functions, F_k's, are equal.
    H_1: One of the distribution functions is different from
         the others.
Dual empirical likelihood ratio (DELR) statistc: 1035.261
Degree of freedom: 8
p-value: 0
```

```
Summary statistics of the estimated F_k's (mean, var --
variance, sd -- standard deviation, Q1 -- first quartile, Q3
-- third quartile, IQR -- inter-quartile range):
    mean  var    sd   Q1    Q3  IQR
F0  2.64 1.59 1.262 1.75  3.25 1.51
F1 10.20 7.70 2.776 8.11 11.82 3.71
F2 10.29 9.31 3.051 7.99 11.96 3.98
F3  3.61 0.71 0.843 3.06  4.20 1.14
F4  9.61 4.10 2.024 8.07 10.96 2.89
```

In the above DRM fitting, we have used the built–in basis function `6`. An alternative way to specify the basis function of the DRM is to directly pass an R function to the `drmdel` function. This way, we can use any basis function we want, but not limited to the built–in basis functions. However, if the basis function of our choice is one of the build–ins, we should always use the built–in for a higher computational efficiency. The following code illustrates the use of an outer basis function.

```
# specify a basis function
basis_gamma <- function(x) return(c(x, log(abs(x))))

# fit the DRM with this specified basis function
drmfit_ex1a <- drmdel(x=x, n_samples=n_samples,
                  basis_func=basis_gamma)

# One can see the summary of this DRM fit is exactly the
# same as that of the previous fit with basis_func=6
summaryDRM(drmfit_ex1a)
```

In addition to `summaryDRM`, we give another function, `meleCov`, to complement the `drmdel` function. The `meleCov` function estimates the asymptotic

covariance matrix of the centered and scaled MELE $\hat{\boldsymbol{\theta}}$, $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$. Recall that the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ is given by Theorem 2.5 as $U^{-1} - W$, where U is the information matrix and W, which is defined in Theorem 2.2, is a matrix determined by sample proportions. Thus, to estimate the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$, the key is to estimate U^{-1}. A consistent estimator of U is given by $\hat{U} = -n^{-1}\partial l_n(\hat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\intercal$. Although U is invertible, $\hat{U}$ is not garanteed to be invertible; and in such a case, R will return an error and stop the program. Hence we implement this in a separate function `meleCov` to increase the stability of the `drmdel` function. This function should be used in the form of `meleCov(drmfit)`, where `drmfit` is a fitted DRM object, i.e. an output from the `drmdel` function. For example, the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*)$ in Example 6.1 can be estimated with the command `meleCov(drmfit_ex1)`.

6.3 The DELR test

From the output of the DRM fit of Example 6.1, we see that the `drmdel` function by default tests the hypothesis

$$H_0: \ F_0 = F_1 = \ldots = F_m \quad \text{against} \quad H_1: \ F_i \neq F_j, \text{ for some } i \text{ and } j,$$

which is equivalent to

$$H_0: \ \boldsymbol{\beta} = \mathbf{0} \quad \text{against} \quad H_1: \ \boldsymbol{\beta} \neq \mathbf{0},$$

under the DRM. We now illustrate how to perform DELR tests for more complicated hypotheses.

In our package, a DELR test for a general composite hypothesis of the form (3.1) is also carried out by the `drmdel` function. Recall that an equiv-

alent form of that hypothesis is given by

$$H_0: \ \boldsymbol{\beta} = \mathcal{G}(\boldsymbol{\gamma}) \quad \text{against} \quad H_1: \ \boldsymbol{\beta} \neq \mathcal{G}(\boldsymbol{\gamma}), \tag{6.1}$$

where $\mathcal{G}$: $\mathbb{R}^{md-q} \to \mathbb{R}^{md}$ is a smooth function and $\boldsymbol{\gamma}$ is a lower dimensional parameter. With this equivalent representation of the hypothesis (3.1), the maximization of the DEL under the new null model is with respect to $\boldsymbol{\gamma}$ and therefore does not involve any constraint. This allows a more efficient implementation of the DELR test and hence we adopt this representation in our software.

Example 6.2 (The DELR test for composite hypothesis)**.** *Adopt the data setting of Example 6.1. Consider the hypothesis testing problem*

$$\begin{aligned} H_0: \ & \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \ \ \textit{and} \ \ \boldsymbol{\beta}_3 = (-3.2,\, 13)^\intercal \\ & \textit{against} \\ H_1: \ & \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \ \ \textit{or} \ \ \boldsymbol{\beta}_3 \neq (-3.2,\, 13)^\intercal. \end{aligned} \tag{6.2}$$

Under the null model, the free DRM slope parameter becomes $\boldsymbol{\gamma} = (\boldsymbol{\beta}_1^T,\, \boldsymbol{\beta}_4^\intercal)$*. And the hypothesis testing problem can be equivalently represented using the form* (6.1) *with*

$$\mathcal{G} = A\boldsymbol{\gamma} + b,$$

where

$$A = \begin{pmatrix} I_2 & I_2 & 0_{2\times 2} & 0_{2\times 2} \\ 0_{2\times 2} & 0_{2\times 2} & 0_{2\times 2} & I_2 \end{pmatrix}^\intercal \quad \textit{and} \quad b = (\mathbf{0}_4^\intercal,\, -3.2,\, 13,\, \mathbf{0}_2^\intercal)^\intercal$$

To test such a hypothesis, we need to pass the function $\mathcal{G}$ *to the argument* `g_null` *of the* `drmdel` *function.*

```
m <- 4  # number of non--baseline distributions
```

```
d <- 2  # dimension of the basis function
# dimension of the DRM parameter beta
dim_beta <- m*d
# dimension of the null parameter gamma
dim_gamma <- dim_beta - 2*d

# A matrix
A <- matrix(rep(0, dim_beta*dim_gamma), dim_beta, dim_gamma)
A[1:d, 1:d] <- diag(2)
A[(d+1):(2*d), 1:d] <- diag(2)
A[(3*d+1):(4*d), (d+1):(2*d)] <- diag(2)

# b vector
b <- numeric(dim_beta)
b[(2*d+1):(3*d)] <- c(-2.3, 13)

# null mapping
g_null <- function(par_gamma) {
  par_beta <- as.vector(A %*% par_gamma) + b
  return(par_beta)
}
```

Optionally, we can also pass the Jacobian matrix of $\mathcal{G}$ *to the* `drmdel` *function through* `g_null_jac` *argument to accelerate the maximization of the DEL under the null model. In general, this Jacobian matrix is a function of* $\boldsymbol{\gamma}$*, so has to be passed to* `drmdel` *as a function. In this particular example, the Jacobian matrirx of* $\mathcal{G}$ *is just a constant matrix,* A*.*

```
# Jacobian matrix of the null mapping
g_null_jac <- function(par_gamma) return(A)
```

With the above preparation, we are ready to perform a DELR test for the hypothesis testing problem (6.2)*.*

```
drmfit_ex2 <- drmdel(x=x, n_samples=n_samples, basis_func=6,
    g_null=g_null, g_null_jac=g_null_jac,
    par_dim_null=dim_gamma)

summaryDRM(drmfit_ex2)
```

The part of the output from `summaryDRM(drmfit_ex2)` *for the DELR test is*

```
Dual empirical likelihood ratio (DELR) statistc: 4.067291
Degree of freedom: 4
p-value: 0.397
```

showing that we find no strong evidence to reject the null hypothesis. The other part of the output is the same as the one given in Example 6.1 because we used the same data and basis function to fit the DRM.

6.4 EL population CDF estimation

As noted in Section 6.1, $dF_r(x_{kj})$, $r = 1, \ldots, , m$, is estimated by $\hat{p}_{kj}^{(r)}$. The corresponding estimator of $F_r(x)$ is then given by

$$\hat{F}_r(x) = n^{-1} \sum_{k,j} \hat{p}_{kj}^{(r)} \mathbb{1}(x_{kj} \leq x).$$

This is exactly the MELE of $F_r(x)$ defined in (2.10). This CDF estimator is a step function which jumps at each distinct point of the data values. One may linearly interpolate between every two adjacent distinct data values to get a continuous version of this CDF estimator.

The *drmdel* package implements the above population CDF estimator through the `cdfDRM` function:

```
cdfDRM(k, x=NULL, drmfit, interpolation=TRUE).
```

The arguments are:

`k`: a vector of labels of populations whose CDFs are to be estimated, with $k[i] = 0, 1, \ldots, , m$.

`x`: can be:

(1) a list whose length is the same as the argument "k". The ith component of this list must be a vector of values at which the CDF of population $k[i]$ is estimated.

(2) a single vector of values, in which case, each CDF is estimated at the same values given by this vector.

(3) NULL (default), in which case, each CDF is estimated at the values of all the observed data points.

`drmfit`: a fitted DRM object, i.e. an output from the `drmdel` function.

`interpolation`: a logical variable specifying whether to linearly interpolate the EL CDF estimator to make it continuous. The default value is `TRUE`.

The output of the function is an R list object whose length is the same as its argument "`k`". The ith component of this list is a data frame with the following two columns:

`x`: values at which the CDF of population $k[i]$ is estimated.

`cdf_est`: the corresponding estimated CDF values of population $k[i]$.

Example 6.3 (EL CDF estimation under the DRM)**.** *Adopt the data and DRM setting of Example 6.1.*

To estimate the $F_1(x)$ *evaluated at* (3, 7.5, 11) *and the* $F_3(x)$ *evaluated at* (2, 6)*, we use:*

```
cdf_est <- cdfDRM(k=c(1, 3), x=list(c(3, 7.5, 11), c(2, 6)),
                  drmfit=drmfit_ex1)
```

```
# show the output
names(cdf_est)
cdf_est$F1
cdf_est$F3
```

To estimate the $F_2(x)$ and $F_4(x)$ at the values of all the observed data points, we use:

```
cdf_est1 <- cdfDRM(k=c(2, 4), drmfit=drmfit_exp1)

# show the output
names(cdf_est1)
cdf_est1$F2
cdf_est1$F4
```

6.5 EL quantile estimation

Based on the EL CDF estimator $\hat{F}_r(x)$, Chen and Liu (2013) proposed to estimate the αth, $\alpha \in (0, 1)$, quantile ξ_r of $F_r(x)$ as

$$\hat{\xi}_r = \inf\{x : \hat{F}_r(x) \geq \alpha\}.$$

They showed that the EL quantile estimator of a vector of quantiles from possibly different distributions is jointly asymptotically normal and is more efficient than the sample quantile based on the rth sample alone.

As noted by Chen and Liu (2013), a quantile estimator based on a discrete distribution function has some disadvantage on estimating lower (or higher) quantiles of a continuous distribution: it tends to over (under) estimate a lower (higher) quantile. To adjust for this effect, they modified the $\hat{\xi}_r$ to $\hat{\xi}_r - (2n_r)^{-1}$ for lower quantile estimation, and commented that such a

modification does not change the first–order asymptotics of the EL quantile estimator. One can similarly modify $\hat{\xi}_r$ for higher quantile estimation.

Function `quantileDRM` calculates the above quantile estimator $\hat{\boldsymbol{\xi}}$ for a vector of quantiles of possibly different populations, and provides an estimate of the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)$, where $\boldsymbol{\xi}^*$ is the true value of $\boldsymbol{\xi}$. The form of `quantileDRM` is:

```
quantileDRM(k, p, drmfit, cov=TRUE, interpolation=TRUE,
            adj=FALSE, adj_val=NULL, bw=NULL,
            show_bw=FALSE).
```

Its arguments are:

`k`: a vector of labels of populations whose quantiles are to be estimated, with $k[i] = 0, 1, \ldots, , m$.

`p`: a vector of probabilities at which the quantiles are to be estimated. Three combinations of `k` and `p` are allowed:

(1) `k` and `p` have the same length: the $p[i]^{th}$ quantile of population $k[i]$ will be estimated for each i.

(2) `k` is a single integer but `p` is a vector: the $p[i]^{th}$ quantile of the same population k will be estimated for each i.

(3) `k` is a vector but `p` is single integer: the pth quantile of population $k[i]$ will be estimated for each i.

`drmfit`: a fitted DRM object.

`cov`: a logical variable specifying whether to estimate the asymptotic covariance matrix of the quantile estimator. With a `TURE` value, the speed of the quantile estimation will be slower. The default is `TRUE`.

`interpolation`: The EL quantile estimator is based on the EL CDF estimator. Hence the way the EL CDF estimate is calculated affects

the result of the quantile estimation. This argument is to be passed to the `cdfDRM` function for tweaking the EL CDF estimator. Its meaning and usage are explained when we introduce the `cdfDRM` function in Section 6.4.

`adj:` a logical variable specifying whether to adjust the EL quantile estimator by adding a term for lower or higher quantile estimation. The default value is `FALSE`.

`adj_val:` a vector of the same length as `k` (or as `p` if the length of `k` is 1) containing the values of adjustment terms for lower or higher quantile estimation, if `adj=TRUE`. The default value, NULL, uses $-(2n_{k[i]})^{-1}$, where $n_{k[i]}$ is the size of the $k[i]^{th}$ sample, for each i, to adjust the EL quantile estimator for lower quantile estimation.

`bw:` to estimate the asymptotic covariance matrix of the EL quantile estimator, the densities of the population distributions have to be estimated. An EL kernel density estimator (described in Section 6.7) is used for this purpose. The argument `bw` is a vector of the same length as `k` containing the bandwidths needed by that kernel density estimator. If `bw` is a single value, the same bandwidth will be used for each population $k[i]$. The default `bw` value, NULL, uses that given by (6.4) for each different $k[i]$. Note that `bw` is only needed for estimating the asymptotic covariance matrix of the EL quantile estimator, but not the population quantiles themselves.

`show_bw:` a logical variable specifying whether to output bandwidths when `cov=TRUE`. The default value is `FALSE`.

The output of the function is a list object containing the following components:

`est:` estimated quantiles.

`cov`: estimated asymptotic covariance matrix of the quantile estimator, available only if `cov=TRUE` at input.

`bw`: bandwidths used for EL kernel density estimation required for estimating the asymptotic covariance matrix of the EL quantile estimator, available only if `cov=TRUE` and `show_bw=TRUE` at input.

Example 6.4 (EL quantile estimation under the DRM)**.** *Adopt the data and DRM setting of Example 6.1. Denote the* α^{th}, $\alpha \in (0, 1)$, *quantile of the* k^{th}, $k = 0, 1, \ldots, m$, *population as* $\xi_{k,\alpha}$.

To estimate $\xi_{0,0.25}$, $\xi_{0,0.6}$, $\xi_{1,0.1}$ *and* $\xi_{2,0.1}$, *we do:*

```
# estimate quantiles and show the output
(qe <- quantileDRM(k=c(0, 0, 1, 2), p=c(0.25, 0.6, 0.1, 0.1),
                   drmfit=drmfit_ex1))
```

To estimate the 0.05^{th}, 0.2^{th} *and* 0.8^{th} *quantiles of* $F_3(x)$, *we do:*

```
(qe1 <- quantileDRM(k=3, p=c(0.05, 0.2, 0.8),
                    drmfit=drmfit_ex1))
```

To estimate the 0.05^{th} *quantiles of* $F_1(x)$, $F_3(x)$ *and* $F_4(x)$, *we do:*

```
(qe2 <- quantileDRM(k=c(1 , 3, 4), p=0.05,
                    drmfit=drmfit_ex1))
```

6.6 Quantile comparison

A bonus provided by the asymptotic theory of the EL quantile estimator is that it enables a Wald test for comparing the quantiles of different populations, an important task of our long term monitoring program for lumber quality as noted in Chapter 1.

Let $\boldsymbol{\xi}$ be a K–vector of quantiles of possibly different populations. Let A be a given $M \times K$ matrix with $M \leq K$, and b be a given vector of length K. For the following linear hypothesis about $\boldsymbol{\xi}$,

$$H_0: \ A\boldsymbol{\xi} = \boldsymbol{b} \quad \text{against} \quad H_1: \ A\boldsymbol{\xi} \neq \boldsymbol{b},$$

a Wald test statistic is defined to be

$$W_n = n(A\hat{\boldsymbol{\xi}} - \boldsymbol{c})^{\intercal}(A\hat{\Sigma}_{\boldsymbol{\xi}}A^{\intercal})^{-1}(A\hat{\boldsymbol{\xi}} - \boldsymbol{c}),$$

where $\hat{\Sigma}_{\boldsymbol{\xi}}$ is a consistent estimator of the asymptotic covariance matrix of $\sqrt{n}(\hat{\boldsymbol{\xi}} - \boldsymbol{\xi}^*)$. By the asymptotic normality of $\hat{\boldsymbol{\xi}}$, W_n has a χ^2_M limiting distribution.

The above Wald test for a linear hypothesis about population quantiles can be carried out by the `quantileCompWald` function:

```
quantileCompWald(quantileDRMObject, n_total, pairwise=TRUE,
                 p_adj_method="none", A=NULL, b=NULL).
```

The arguments are:

`quantileDRMObject:` an output from the `quantileDRM` function. It must contain an estimate of the asymptotic covariance matrix of the EL quantile estimator, `quantileDRMobject$cov`; that is, the argument `cov` at the input of the `quantileDRM` function must be set to `TRUE` when running `quantileDRM`.

`n_total:` the total sample size.

`pairwise:` a logical variable specifying whether to perform pairwise comparisons of the quantiles. The default is `TRUE`.

`p_adj_method:` the method for adjusting p–values for multiple comparisons, provided `pairwise=TRUE`. This is implemented through the R

function `p.adjust`, and the available methods are: `holm`, `hochberg`, `hommel`, `bonferroni`, `BH`, `BY`, `fdr` and `none`. See `help(p.adjust)` for details. The default value is `none`, i.e. no adjustment.

`A:` matrix A in the linear hypothesis. The default is `NULL`, i.e. no linear hypothesis to test.

`b:` vector b in the linear hypothesis. This must be given if the argument `A` is not `NULL`.

The output is a list object containing the following components:

`p_val_pair:` p–values of pairwise comparisons, in the format of a lower triangular matrix, available only if `pairwise=TRUE` at input.

`p_val:` p–value of the linear hypothesis test, available only if the arguments `A` and `b` are not NULL at input.

Example 6.5 (Comparison of popuation quantiles under the DRM)**.** *Adopt the data and DRM setting of Example 6.1 and the notation of Example 6.4.*

We compare the 5th *percentiles of population* 0, 1, 2 *and* 3, *i.e.*

$$\begin{aligned} H_0 :\ & \xi_{0,\,0.05} = \xi_{1,\,0.05} = \xi_{2,\,0.05} = \xi_{3,\,0.05} \\ & \textit{against} \\ H_1 :\ & \xi_{i,\,0.05} \neq \xi_{j,\,0.05} \ \ \textit{for some } i \textit{ and } j. \end{aligned} \tag{6.3}$$

We first estimate these quantiles and the asymptotic covariance matrix of the EL quantile estimator using

```
qe <- quantileDRM(k=c(0, 1, 2, 3), p=0.05,
                  drmfit=drmfit_ex1)
```

To compare the quantiles using a Wald test, we need to specify the contrast matrix A *and the vector* b*; and for the hypothesis testing problem* (6.3)*, one*

way of doing this would be to set

$$A = \begin{pmatrix} 1 & -1 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ 1 & 0 & 0 & -1 \end{pmatrix}$$

and correspondingly $b = \mathbf{0}$.

```
# specify the matrix A
A <- matrix(rep(0, 12), 3, 4)
A[1,] <- c(1, -1, 0, 0)
A[2,] <- c(1, 0, -1, 0)
A[3,] <- c(1, 0, 0, -1)

# specify the vector b
b <- rep(0, 3)
```

With the above preparation, we now can compare the quantiles as follows:

```
# Adjust the p-values for pairwise comparisons using the
# "holm" method.
(quantComp <- quantileCompWald(qe, n_total=sum(n_samples),
                               p_adj_method="holm", A=A,
                               b=b))
```

The output is

```
$p_val_pair
             q1       q2 q3 q4
q1           NA       NA NA NA
q2 0.000000e+00       NA NA NA
q3 0.000000e+00 0.537371 NA NA
q4 1.159073e-13 0.000000  0 NA
```

```
$p_val
[1] 0
```

indicating that overall the quantiles are different, but not enough evidence (p–value $= 0.537371$*) is found to reject the hypothesis that* $\xi_{1,0.05} = \xi_{2,0.05}$*, which, in truth, are equal.*

6.7 EL kernel density estimation

With the EL CDF estimator $\hat{F}_r(x)$, it is easy to construct an EL kernel density estimator for $F_r(x)$. Let $K(\cdot) \geq 0$ be a commonly used kernel function such that $\int K(x)dx = 1$, $\int xK(x)dx = 0$ and $\int x^2K(x)dx < \infty$. For a given bandwidth $b > 0$, put $K_b(x) = (1/b)K(x/b)$. The EL Kernel density estimator of $F_r(x)$ is defined as

$$\hat{f}_r(x) = \int K_b(x-y)d\hat{F}_r(x) = \sum_{k,j} K_b(x - x_{kj})\hat{p}_{kj}^{(r)}.$$

This estimator is originally proposed by Fokianos (2004). He showed that the asymptotic mean integrated square error of this estimator is smaller than that of the classical kernel density estimator with empirical weight of $1/n_r$ based on the rth sample alone.

Kernel density estimation requires a bandwidth b to be specified in advance. For the classical density estimation based on n iid observations, Deheuvels (1977) and Silverman (1986) suggested to use $b = 1.06n - 1/5 \min\{\hat{\sigma}, \hat{R}/1.34\}$, where $\hat{\sigma}$ and $\hat{R}$ are the estimated standard deviation and interquartile range (IQR) of the population, respectively. We adopt this formula in EL kernel density estimation and in our software uses the default bandwidth of

$$b_r = 1.06n^{-1/5} \min\{\hat{\sigma}_r,\ \hat{R}_r/1.34\} \tag{6.4}$$

for the density estimation of the rth population, where n is the total sample size, and $\hat{\sigma}_r$ and $\hat{R}_r$ are the standard deviation and IQR of the estimated CDF $\hat{F}_r(x)$ (2.10), respectively.

The above $\hat{f}_r(x)$ conforms with the general definition of a kernel density estimator with weights given by $\{\hat{p}_{kj}^{(r)}\}$, so can be easily implemented using the R function `density`. The `densityDRM` function of our package provides an easy interface for automatically passing the data, $\{x_{kj}\}$, and the weights, $\{\hat{p}_{kj}^{(r)}\}$, to the R `density` function. The `densityDRM` function has a generic form of

```
densityDRM(k, drmfit, interpolation=TRUE, ...).
```

The arguments are:

`k`: a single label of the population whose density is to be estimated, `k` $\in \{0, 1, \ldots, m\}$.

`drmfit`: a fitted DRM object.

`interpolation`: a logical variable to be passed to `quantileDRM` and then ultimately to `cdfDRM`, for estimating the population standard deviations and IQRs required for calculating the default bandwidth (6.4). The default value is `TRUE`.

`...`: further arguments to be passed to the R `density` function for kernel density estimation. See `help(density)` for details. One can customize bandwidth using the `bw` argument of the `density` function. The arguments `x` and `weights` should not be specified because they are supposed to be extracted automatically from the fitted DRM object `drmfit`. If specified, they will be automatically replaced by those extracted from `drmfit` and a warning message is returned.

The output is an object of class "density", which comes from the R `density` function.

Example 6.6 (EL kernel density estimation)**.** *Adopt the data and DRM setting of Example 6.1.*

We can estimate the density of F_3 using command

```
dens_pop3 <- densityDRM(k=3, drmfit=drmfit).
```

We compare the EL kernel density estimator, the classical kenel density estimator based on the third sample alone, and the true density curve by plotting them.

```
# Plotting the EL kernel density estimates
plot(dens_pop3, xlim=range(c(0, 10)), ylim=range(c(0, 0.5)),
     main=expression(
          paste("Kernel density estimators (KDE) of ", F[3],
                sep="")),
     xlab="x")

# Adding the classical kernel density curve of F_3 based
# on the third sample alone
lines(density(x3), col="blue", lty="28F8")

# Add the true density curve of F_3
lines(seq(0, 10, 0.01), dgamma(seq(0, 10, 0.01), 18, 5),
      type="l", col="red", lty="dotted")

legend(6.5, 0.48,
       legend=c("EL KDE", "Classical KDE", "True density"),
       col=c("black", "blue", "red"),
       lty=c("solid", "28F8", "dotted"))
```

The comparative plot is shown in Figure 6.1.

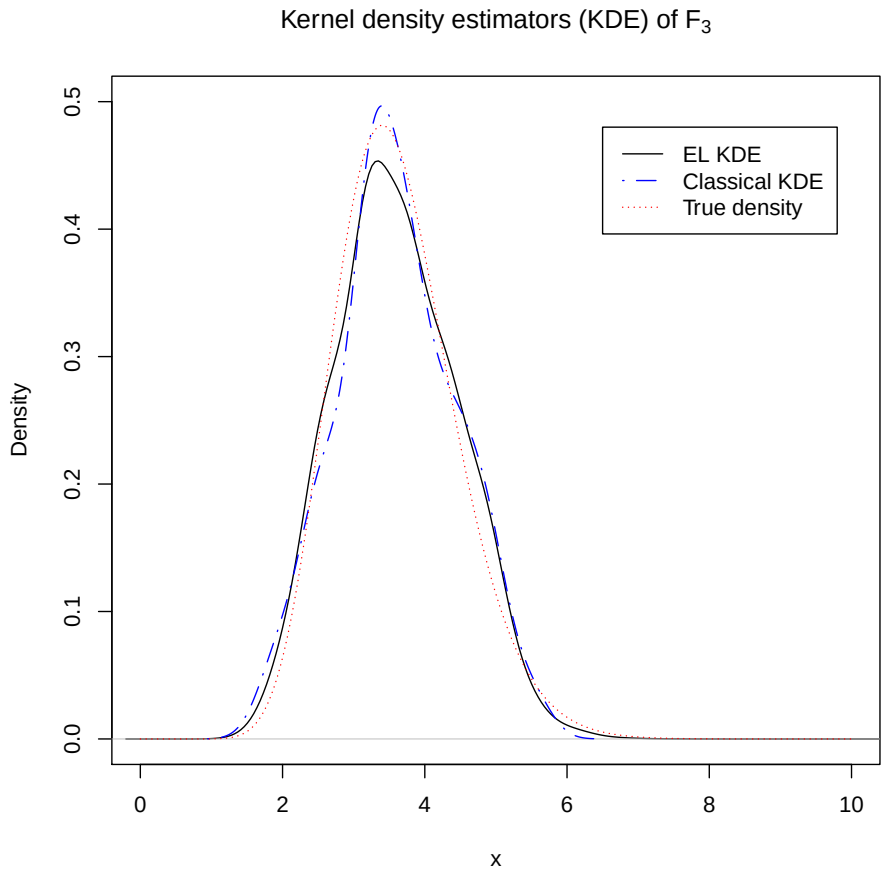

Figure 6.1: Comparative plot of the EL kernel density estimator, classical kernel density estimator and true density of F_3 in Example 6.6.

Chapter 7

Summary and Future Work

7.1 Summary of the present work

This thesis has presented new theories for inference, especially for hypothesis testing, concerning a number of populations from which multiple complete or Type I censored independent samples are observed. The work was motivated by an important application, the development of a new long term monitoring program for the North American lumber industry. Traditional reliance in that industry on standards based on nonparametric statistical procedures led to our adoption of a semiparametric approach, with a large nonparametric component. The need for efficiency and hence small sample sizes led to our density ratio model approach where common information across samples could be borrowed to gain strength.

7.1.1 Contribution I: DELR test for hypothesis about the DRM parameter

The first contribution of the thesis is a theory of dual EL ratio test for a general class of composite hypotheses about the DRM slope parameter $\boldsymbol{\beta}$, which encompasses testing differences among population distributions as a special case. The new theory is assessed, both through theoretical analysis for large samples and by simulation studies for small ones. The theoretical results are illustrated by examples and the use of the proposed test is demonstrated on a dataset collected by our group over five years.

Our overall conclusion is that the new theory works well and achieves its

intended objectives. We recommend its use in applications for comparing population parameters when independent samples from each are available.

More specifically the new theory is very general and flexible in that it embraces in the DRM a large family of familiar distributions such as the normal and gamma, making it quite robust against misspecification of population distributions. It comes equipped with an asymptotic theory that enables its properties to be assessed. In particular, easy to apply asymptotic approximations are available for the distributions of the test statistics involved. The asymptotic distribution of the proposed DELR test statistic is derived under the null model and also under a local alternative model. The null limiting distribution allows us to approximate the p–values of the DELR tests; the limiting distribution under the local alternative model enables us to approximate the power of a DELR test, to calculate the sample size required for attaining a given power, and to compare the local asymptotic powers of DELR tests constructed in different ways.

Simulation studies show that when the basis function $\boldsymbol{q}(x)$ is correctly specified, the distribution of the DELR test statistic, R_n, is well–approximated by chi–square distribution under the null model and by non–central chi–square distribution under the local alternative model; that for normal data with equal variances, the power of the DELR test is comparable to that of the optimal two–sample t–test; that for normal data with unequal variance and non–normal data, the DELR test has a much higher power than all its competitors and its type I error rate is close to the nominal size of 0.05. When the DRM is misspecified, we observe the similar results on power comparison and type I error rate. Also, Wald tests are generally not as powerful as the DELR tests.

The demonstration of the use of the method on three lumber samples, shows our method to give a more incisive assessment than competitors through paired comparisons of the populations.

7.1.2 Contribution II: Effects of information pooling by the DRM

The second contribution of the thesis is a theoretical assessment of the effects of information pooling by the DRM on the estimation accuracy of the MELE of the DRM parameter and on the local asymptotic power of the DELR test. We show that when additional samples are incorporated by the DRM, the estimation accuracy of the MELE $\hat{\boldsymbol{\theta}}$ is usually increased and the local asymptotic power of the DELR test is often improved, even if the underlying distributions of the additional samples are not related to the population distributions of direct interest. In the special case of testing equality among distribution functions within subgroups of populations, including extra samples does not change the local asymptotic power of the DELR test. This is similar to the classical t–test: if we construct a t–test using a pooled sample variance by including additional samples, the gain is only on the degree of freedom, but not in asymptotic sense.

Our simulations support these theoretical results, that the DRM does borrow strength as intended, to reduce the sample size needed to achieve required power even against local alternatives.

7.1.3 Contribution III: EL inference under the DRM based on Type I censored samples

The third contribution is an EL inference framework for population distributions under the DRM from which multiple Type I censored samples are observed. In particular, we solve the maximization problem of the EL based on Type I censored samples by reducing it to the maximization of a concave DPEL, study the asymptotic properties of the DPEL, and establish the theory of EL ratio test for hypotheses about the DRM parameter under this DPEL inference framework.

The proposed EL inference framework features a fast computation be-

cause of the concavity of the DPEL, and is very general in the sense that (1) it may be used to extend any EL inference result that is derived for complete samples under the DRM to the case of Type I censored samples, (2) it applies to the general Type I left and right–censored samples, and (3) it does not require the censoring cutting points for each sample to be the same.

7.1.4 Contribution IV: Software package "drmdel" for DEL inference under the DRM

The last but not least contribution of this thesis is a user friendly R software package *drmdel* for DEL inference under the DRM for multiple complete samples, which is available on CRAN at http://cran.r-project.org/web/packages/drmdel/index.html. The package is written in C from the core, so is fast. It can calculate the MELE of the DRM parameter, perform the DELR test as found in Chapter 3, estimate population distribution functions, estimate quantiles of the population distributions as found in Chen and Liu (2013), compare quantiles from different distributions using a Wald test, and estimate densities of different populations as found in Fokianos (2004). The use of the package is fully illustrated by examples in Chapter 6.

An implementation of this package for EL inference under the DRM based on Type I censored samples is currently under way and will soon be available.

7.2 Outlook on future work

7.2.1 EL ratio test for comparing quantiles under the DRM

Quantile estimation under the DRM based on complete samples has been studied by Chen and Liu (2013). A Wald test for comparing quantiles of different populations can be constructed based on the asymptotic normality

of that quantile estimator. However, there is no existing result on EL ratio test for comparing quantiles. A population quantile can be defined as the solution of a non–smooth estimation equation. My proposal is to construct an EL ratio test by incorporating this estimation equation as a constraint when profiling the EL function. The resulting profile EL then is a function of the quantiles to be compared, and so is the corresponding EL ratio statistic. The limiting distribution of this EL ratio statistic is hard to derive because the estimating equation defining a quantile is not smooth. However, when the population distribution functions are smooth, this estimating equation is asymptotically smooth. I have proved that, for parameters defined by smooth estimation equations, the EL ratio statistic under the DRM has a chi–square limiting distribution. Therefore, based on the "asymptotic smoothness" argument, I conjecture that, for quantile comparisons, the corresponding EL ratio statistic also has a chi–square limiting distribution. I aim to prove this conjecture.

Another way to construct an EL ratio test for quantile comparisons is to incorporate quantiles in EL using a kernel–smoothed estimating equation. In the one–sample case (not under the DRM), this approach is proposed by Chen and Hall (1993). One issue with this approach is that one has to choose a tuning constant "bandwidth" that affects the asymptotic properties of the EL ratio statistic. I plan to study this approach under the DRM for multiple samples, and study the optimal choice of the associated bandwidth.

7.2.2 Effects of information pooling on quantile estimation under the DRM

As we have shown in Chapter 4, incorporating extra samples using DRM in general increases the estimation accuracy of the MELE and improves the local asymptotic power of the DELR test. But dose it also improve the estimation accuracy of the Chen–Liu (2013) EL quantile estimator? This is a hard question due to the complicated algebraic expression of the asymptotic

covariance matrix of that estimator. I seek for an answer.

7.2.3 Inference under the DRM based on randomly censored samples

Chapter 5 establishes the theory of EL inference under the DRM based on type–I censored samples. What if the samples are randomly censored? In that case, the EL function under the DRM is very complicated, and the study of the properties of the EL is found hard. As outlined in the Appendix 5.9 of Chapter 5, Ren (2008) proposed to base the inference about randomly right–censored samples under the DRM on the WEL, an EL like function constructed upon the Kaplan–Meier estimators of the population distribution functions instead of the population distributions themselves. An advantage of the WEL over the EL is that it has a simple analytic form. Ren showed that under a two–sample DRM, the WEL ratio statistic has a scaled chi–square limiting distribution. The scaling factor depends on the distributions of the populations as well as those of the censoring variables, which are generally unknown. I found that, under a DRM of more than two samples, the limiting distribution of the WEL ratio statistic is a mixture of chi–squares. The mixing coefficients, which also depend on the distributions of the populations and those of the censoring variables, are unknown. Hence the WEL ratio is not useful for constructing confidence intervals or for testing hypotheses about the DRM parameters.

I aim to construct a usable WEL ratio statistic for randomly right–censored samples under the DRM. The reason the limiting distribution of the WEL is a mixture of chi–squares instead of a simple chi–square is that, under the WEL, the information matrix and the asymptotic covariance matrix of the score function do not have a neat relationship like that found for the DEL based on complete samples. My idea is to add weights to the Kaplan–Meier estimators of the distribution functions and to construct a WEL based on the tweaked Kaplan–Meier estimators such that, under the

new WEL, the information matrix and the asymptotic covariance matrix of the score function have a "nice" relationship. The corresponding WEL ratio may then have a simple chi–square limiting distribution.

7.2.4 Basis function selection in the DRM

Whether an EL inference under the DRM is effective relies heavily on whether the model fits data well, which in turn depends on the selection of the basis function $\boldsymbol{q}(x)$. Simulations show that when there is vague information about the functional form of $\boldsymbol{q}(x)$, using a long vector containing many linearly independent components as the basis function usually gives a good model fit. However, some components might be superfluous. Ideally, we want to choose a subset of this long vector as the final basis function such that it yields both a good model fit to the data and a high estimation accuracy. I plan to study the following two approaches for the selection of the basis function: (1) derive an information criterion similar to the Akaike's information criterion that estimates the Kullback–Leibler divergence between the assumed DRM and the true distributions; and (2) study the maximum penalized EL approach with an L_1 or nonconcave penalty function.

7.2.5 Random–effect DRMs

The lumber quality samples are collect from different mills across Canada each year. Although our purpose is to assess change in lumber strength over time, which does not involve the difference on the mill level, the mill–to–mill variability should be considered for a more realistic model. My rough idea is to add random effects to the DRM slope parameters, $\{\boldsymbol{\beta}_k\}$, to represent this mill–level variability. The model constructed this way is flexible, however my initial investigation showed that the resulting EL function is very complicated and the study of the properties and computation of the corresponding maximum EL estimator are hard. Given the above difficulties, I plan to start

from a simpler model, where the mill–level variability is assumed to be fully represented by a random effect on the population means only. Hopefully, this investigation can help me to ultimate overcome the difficulties in the former more sophisticated random–effect DRM.

7.2.6 Other projects: high dimensional DRM and finite sample corrections for the DELR test

The current theory of EL inference under the DRM assumes that the number of samples is a constant that does not change with the total sample size n. However, in our targeted applications, new lumber samples are added year by year. Hence, in reality, the number of DRM parameters increases as the sample size n increases, thereby nullifying the existing asymptotic results on the DELR test and the EL quantile estimation under the DRM. I plan to study the conditions under which these asymptotic results are still valid when the number of samples increases with n.

Another practical issue is that the DELR test for the DRM parameter is found to be anti–conservative when the total sample size is small. A finite sample correction may be useful for adjusting for this effect. I intend to study whether the DELR test is Bartlett correctable; if it is, I hope the corresponding Bartlett correction will have a second–order accuracy and can make the DELR test less anti–conservative.

References

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59(1):19–35.

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika*, 66(1):17–26.

ASTM (2007). *ASTM D1990 – 07: Standard Practice for Establishing Allowable Properties for Visually-Graded Dimension Lumber from In-Grade Tests of Full-Size Specimens.* ASTM Intenational, West Conshohocken, USA.

Broyden, C. G. (1970). The convergence of a class of double–rank minimization algorithms. *Journal of the Institute of Mathematics and Its Applications*, 6(1):76–90.

Chen, J. and Liu, Y. (2013). Quantile and quantile–function estimations under density ratio model. *The Annals of Statistics*, 41(3):1669–1692.

Chen, S. X. and Hall, P. (1993). Smoothed empirical likelihood confidence intervals for quantiles. *The Annuals of Statistics*, 21(3):1166–1181.

Cheng, K. F. and Chu, C. K. (2004). Semiparametric density estimation under a two–sample density ratio model. *Bernoulli*, 10(4):583–604.

Chow, Y. S. and Teicher, H. (1997). *Probability Theory – Independence, Interchangeability, Martingales.* Springer, New York, USA, 3rd edition.

Deheuvels, P. (1977). Estimation non paramétrique de la densité par histogrammes généralisés. *Revue de Statistique Appliquée*, 25(3):5–42.

Efron, B. and Tibshirani, R. (1996). Using specially designed exponential families for density estimation. *The Annals of Statstics*, 24(6):2431–2461.

Farewell, V. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, 66:27–32.

Fletcher, R. (1970). A new approach to variable metric algorithms. *Computer Journal*, 13(3):317–322.

Fokianos, K. (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 66(4):941–958.

Fokianos, K. (2007). Density ratio model selection. *Journal of Statistical Computation and Simulation*, 77(9):805–819.

Fokianos, K. and Kaimi, I. (2006). On the effect of misspecifying the density raio model. *Annals of the Institute of Statistical Mathematics*, 58:475–497.

Fokianos, K., Kedem, B., Qin, J., and Short, D. A. (2001). A semiparametric approach to the one–way layout. *Technometrics*, 43(1):56–65.

Fokianos, K. and Troendle, J. F. (2007). Inference for the relative treatment effect with the density ratio model. *Statistical Modelling*, 7(2):155–173.

Gilbert, P. B. (2000). Large sample theory of maximum likelihood estimates in semiparametric biased sampling models. *The Annals of Statistics*, 28(1):151–194.

Gilbert, P. B., Lele, S. R., and Vardi, Y. (1999). Maximum likelihood estimation in semiparametric selection bias models with application to aids vaccine trials. *Biometrika*, 86(1):27–43.

Gill, R. D., Varidi, Y., and Wellner, J. A. (1988). Large sample theory of empirical distributions in biased sampling models. *The Annals of Statistics*, 16(3):1069–112.

Goldfarb, D. (1970). A family of variable metric updates derived by variational means. *Mathematics of Computation*, 24(109):23–26.

Guan, Z., Qin, J., and Zhang, B. (2012). Information borrowing methods for covariate-adjusted roc curve. *The Canadian Journal of Statistics*, 40(2):1–19.

Harville, D. A. (2008). *Matrix Algebra From a Statistician's Perspective*. Springer, New York, USA, 1st edition.

Huang, A. (2014). Joint estimation of the mean and error distribution in generalized linear models. *Journal of the American Statistical Association*, 109(505):186–196.

Huang, A. and Rathouz, P. J. (2012). Proportional likelihood ratio models for mean regression. *Biometrika*, 99(1):223–229.

Jiang, S. and Tu, D. (2012). Inference on the probability $p(t_1 < t_2)$ as a measurement of treatment effect under a density ratio model and random censoring. *Computational Statistics and Data Analysis*, 56:1069–1078.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions, Volume 2*. Wiley–Interscience, Hoboken, USA, 2nd edition.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, 53(282):457–481.

Keziou, A. and Leoni-Aubin, S. (2008). On empirical likelihood for semiparametric two–sample density ratio models. *Journal of Statistical Planning and Inference*, 138:915–928.

Luo, X. and Tsai, W. Y. (2012). A proportional likelihood ratio model. *Biometrika*, 99(1):211–222.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics*, 29:479–486.

Marshall, A. W. and Olkin, I. (2007). *Life Distributions – Structure of Nonparametric, Semiparametric and Parametric Families*. Springer, New York, USA.

Mathai, A. M. (1992). *Quadratic Forms in Random Variables: Theory and Applications*. Marcel Dekker, Inc., New York, USA.

Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization*. Springer, New York, USA, 2nd edition.

Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75:237–249.

Owen, A. B. (2001). *Empirical Likelihood*. Chapman & Hall, New York, USA.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case–control studies. *Biometrika*, 66:403–422.

Qin, J. (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics*, 21(3):1182–1196.

Qin, J. (1998). Inferences for case–control and semiparametric two–sample density ratio models. *Biometrika*, 85(3):619–630.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics*, 22(1):300–325.

Qin, J. and Liang, K. Y. (2011). Hypothesis testing in a mixture case–control model. *Biometrics*, 67:182–193.

Qin, J. and Zhang, B. (1997). A goodness of fit test for the logistic regression model based on case–control data. *Biometrika*, 84:609–618.

Rathouz, P. J. and Gao, L. (2009). Generalized linear models with unspecified reference distribution. *Biostatistics*, 10(2):205–218.

Ren, J. (2008). Weighted empirical likelihood in some two–sample semi-parametric models with various types of censored data. *The Annals of Statistics*, 36(1):147–166.

Scholz, F. W. and Stephens, M. A. (1987). K–sample Anderson–Darling tests. *Journal of the American Statistical Association*, 82(339):918–924.

Shanno, D. F. (1970). Conditioning of quasi–Newton methods for function minimization. *Mathematics of Computation*, 24(111):647–656.

Shen, Y., Ning, J., and Qin, J. (2012). Likelihood approaches for the invariant density ratio model with biased-sampling data. *Biometrika*, 99(2):363–378.

Silverman, B. W. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall, Boca Raton, USA, 1st edition.

Tan, Z. (2009). A note on profile likelihood for exponential tilt mixture models. *Biometrika*, 96(1):229–236.

van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge University Press, Cambridge, UK.

Vardi, Y. (1982). Nonparametric estimation in the presence of length bias. *The Annals of Statistics*, 10(2):616–620.

Vardi, Y. (1985). Empirical distributions in selection bias models. *The Annals of Statistics*, 13(1):178–203.

Wan, S. and Zhang, B. (2008). Comparing correlated roc curves for continuous diagnostic tests under density ratio models. *Computational Statistics and Data Analysis*, 53:233–245.

Wang, C., Tan, Z., and Louis, T. A. (2011). Exponential tilt models for two-group comparison with censored data. *Journal of Statistical Planning and Inference*, 141:1102–1117.

Wilcox, R. R. (1995). Anova: A paradigm for low power and misleading measures of effect size. *Review of Educational Research*, 65(1):51–77.

Wu, C. F. J. and Hamada, M. S. (2009). *Experiments: Planning, Analysis, and Optimization*. Wiley, Hoboken, USA, 2nd edition.

Zhan, X. (2002). *Matrix Inequalities*. Springer–Verlag, Berlin, Germany.

Zhang, B. (2000). Quantile estimation under a two–sample semi–parametric model. *Bernoulli*, 6(3):491–511.

Zhang, B. (2002). Assessing goodness-of-fit of generalized logit models based on case–control data. *Journal of Multivariate Anlysis*, 82:17–38.

Zhang, B. (2006). A partial empirical likelihood based score test under a semiparametric finite mixture model. *Annals of the Institute of Statistical Mathematics*, 58:707–719.

Zhang, F., editor (2005). *The Schur complement and its applications*. Springer, New York, USA.

Zorich, V. A. (2004). *Mathematical Analysis II*. Springer–Verlag, Berlin, Germany.

Zou, F. (2002). A note on a partial empirical likelihood. *Biometrika*, 89(4):958–961.

Zou, F., Fine, J. P., and Yandell, B. S. (2002). On empirical likelihood for a semiparametric mixture model. *Biometrika*, 89(1):61–75.