Average effects for regression models with misspecifications and diffuse interaction models

by

Juxin Liu

B.Sc., Anhui University, 2000
M.Sc., Beijing University, 2003
Ph.D., University of British Columbia, 2007

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

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Statistics)

The University Of British Columbia

June, 2007

Abstract

In epidemiological studies, how best to assess and interpret interaction of risk factors of interest has been the subject of a lively debate. In statistical regression models, the interaction between two putative risk factors is described by the regression coefficient of the product of the risk factors. What happens if a linear regression model without pairwise interaction terms is used to fit the data actually generated from a linear regression model with all possible pairwise interactions? We apply the idea of average effect to evaluate the consequence of misspecified models and find out that the average effect estimates are still consistent if the joint distribution of risk factors satisfy some certain conditions. It is known that pairwise interaction models encounter intractable problems especially when the number of risk factor under consideration is quite large. The number of pairwise interaction terms is $p(p-1)/2$, if the number of risk factors is p. As an alternative strategy, we introduce diffuse interaction model with only one parameter to reflect the interactions among all the risk factors, without specifying which of the risk factors do indeed interact. We compare the two kinds of interaction models in terms of ability to detect interactions. Another issue investigated in the thesis is to devise MCMC algorithms to estimate diffuse interaction models. This is done not only for the diffuse interaction model assuming all risk factors interact in the same direction, either synergistically or antagonistically, but also for extended diffuse interaction models which relaxing this strong assumption.

Table of Contents

Appendices

List of Tables

List of Figures

I would like to thank my supervisor, Professor Paul Gustafson, for his guidance, invaluable advice and great patience. Each time when I am doing "random walk", it is my supervisor who shows me an efficient "direction" to make meaningful movement.

Special thanks to Professors Harry Joe and Lang Wu for their helpful comments and suggestions on my proposal and table-tennis playing as well. And extra thanks to Harry for the tips on Latex, which makes the notations in my thesis well-exhibited.

I am very grateful to the helps from my fellow graduate students and the ladies in our main office. I feel lucky to have had the chance to study in our department, which is just like a big family.

Juxin Liu

The University of British Columbia

June 2007

To Ameng

Chapter 1

Introduction

1.1 Motivation

The term *interaction* is used in epidemiology to describe a situation in which two or more risk factors modify the effect of each other with regard to the occurrence or value of a given health outcome, denoted by Y.

For dichotomous variables, interaction means that the effect of one risk factor, say A, on the outcome differs depending on whether another variable B (*effect modifier*) is present. Moreover, if the presence of B, the effect modifier, potentiates/accentuates the effect of risk factor A, this variable and risk factor are said to be *synergistic* (positive interaction); if the presence of B diminishes or eliminates the effect of risk factor A, the two variables are *antagonistic* (negative interaction). For continuous variables, the phenomenon of interaction means that the effect of one risk factor on outcome differs depending on the value of another variable (effect modifier). A mathematical definition comes in Section 1.2. Later in Section 3.2.1, we will use a more general definition of synergism/antagonism when *diffuse interaction* models are introduced. In epidemiological studies, synergistic/antagonistic interaction among risk factors is common. For example, if people suffering from obesity have high blood cholesterol, then they have higher chances to get heart diseases. Another example (for antagonism) is the interaction between smoking and intake of Vitamin A for the risk of lung cancer. People who smoke a lot but take Vitamin A in daily dietary have lower risk of lung cancer than people who

seldom smoke but lack of Vitamin A.

Interaction can be described in two different but compatible ways. Each definition leads to a specific strategy for the assessment of interaction.

The first, definition is based on homogeneity/heterogeneity of effects. Interaction occurs when the effect of a risk factor A on outcome Y is not homogeneous across strata formed by a third variable B. When this definition is used, variable B is often referred to as an effect modifier.

The second, definition is based on the comparison between observed and expected joint effects of risk factor A and third variable B. Interaction occurs when the observed joint effect of A and B differs from that expected on the basis of the independent effects of A and B.

How does one assess interactions? In the thesis, we only focus on the situation where the relationship between outcome Y (continuous) and risk factors is of interest. Commonly the product of the two variables A and B is used to describe interaction effects. That is,

$$\mathrm{E}(Y|A,B) = \beta_0 + \beta_1 A + \beta_2 B + \beta_{12} AB,$$

where β_{12} reflects magnitude of interaction effect between A and B. If $\beta_{12} > 0$, the interaction between risk factors A and B is synergistic (positive) interaction; otherwise it is an antagonistic (negative) interaction. For simplicity, the pure quadratic terms A^2 and B^2 are omitted in the above model (These terms are also omitted in the linear regression models which are mentioned later and an analysis with quadratic terms is mentioned briefly in Section 2.5). Note that if the outcome of interest is discrete, we could replace $\mathrm{E}(Y|A,B)$ by $g(\mathrm{E}(Y|A,B))$ in the above model, i.e., a generalized linear model. For instance, if Y is binary, logistic regression can be used in terms of $g(\mathrm{E}(Y|A,B)) = \log(P(Y=1)/P(Y=0))$.

More and more attention has been focused on model misspecifications since statisticians realize that unfortunately misspecified models are not uncommon in practice. Box (1979) and also McCullagh and Nelder (1983) mentioned "all models are wrong", though some fit data better than the others. Since we never know what the true model is in reality, by "true" model we assuming that is the true structure or closer to the truth than the others. It is natural to ask whether the properties of the estimator derived from misspecified models are affected. Does the estimator still converge to some limit asymptotically, and does this limit have any meaning? If the estimator is approximately consistent, is it also asymptotically normal? White (1982) provides answers to these questions by using maximum likelihood techniques for estimation and inference of regression coefficients. Also in White (1981), the consequences and detection of misspecified nonlinear regression models are explored.

Under the general topic of interactions, the goal of our work is to explore the consequences of a particular scenario of misspecified models . To be specific, what happens if we apply an additive model ignoring pairwise interactions to data which are actually generated from a pairwise interaction model? In this context, to make clear how those ignored interaction effects affect the results, we apply the *average effect* idea (definition given later in Chapter 2), while not applying the results from White (1982) directly to the regression coefficients. As is known, the interpretability of regression coefficients of risk factors is rather limited when models include interaction terms.

The idea of average effect is proposed by Gelman and Pardoe (2007) and Gustafson et al. (2005) as well. Basically, it is the average of *predictive effect*, which is the expected change in outcome associated with a unit change in one of the risk factors. In a linear regression model without interactions, the average effect of any putative risk factor is simply the regression coefficient. However, in a model with interaction terms, the predictive effect in general depends on the value of risk factors. There are various definitions

based on different distributions to average over. Three versions are defined in Chapter 2. The main advantage of the average effect idea is to make comparisons possible between different parametric models with sets of parameters that have incomparable interpretations. The average effect idea could also be used in other contexts. For example, Xu and O'Quigley (2000), also Gustafson (2007), gives a definition of average effect in survival analysis. In the future, we could also apply the idea of average effect to explore the consequences of model misspecifications under the framework of survival analysis.

Nowadays another common issue arising in epidemiological studies is that a large (sometimes larger than sample size) number of potential risk factors should be considered in modelling. We could imagine the challenge to model the interaction effects when p, the number of risk factors, is relatively large. In particular, if the model under consideration is a pairwise interaction model, the number of all possible pairwise interaction terms is $p(p-1)/2$. For instance, if 12 risk factors are involved in the study, 66 pairs of possible interactions would be investigated besides the possibility of higher order interactions among three or more risk factors. *Stepwise* procedures are the most widely used approaches to select the important pairwise interaction terms in applied medical statistics. The basic idea of stepwise procedure is to find a "best" subset of potential risk factors by subsequently adding or dropping one risk factor at a time. Take forward stepwise regression procedure for instance, it starts off by choosing a model containing the single best risk factor variable and then attempts to build up with subsequent additions of other risk factors one at a time, as long as these additions are worthwhile.

There are, however, a number of limitations with stepwise procedure. In particular, if possible models under consideration are nested pairwise interaction models, the stepwise procedure does not scale up to the number of potential risk factors (suppose all the risk factors involved at this stage are all important). As the number p of risk factors increases, the number of submodels, $2^{(p(p-1)/2)}$, increases dramatically, making the computational

burden enormous. Also, the fitting of full model sometimes may not be suitable, because only a few of the p risk factors are typically included in the final model. And the fitting of the full model increases the numerical complexity of the methods unnecessarily.

Another problem is that the model selected by a stepwise procedures includes only those variables entered in that final model, and ignores the variables not selected and the uncertainty due to the model selection procedure. In the worst possible scenario, such procedures may underestimate uncertainty about the variables, overestimate confidence in a particular model being selected, and may lead to sub-optimal decisions and limited predictability (Raftery (1996); Draper (1995)). There are also other drawbacks of stepwise selection. For example, a small change in the data can result in very different models being selected and this can reduce prediction accuracy, as discussed in Breiman (1996).

To overcome the difficulties caused by large number of risk factors, we use *diffuse interaction* models as an alternative way to model interactions. These kind of models are proposed by Gustafson et al. (2005) in context of binary response. In this context, by *synergism/antagonism*, we mean that the effect of a putative risk factor increases/decreases in magnitude as all the other risk factors move from absent to present. (Or as all the other risk factors increases if they are continuous.) Under this particular probability model, only one parameter is used to describe interaction among all risk factors. That is, the parameter can tell the overall interaction direction but without indicating which of the risk factors actually interact in that direction and which of them not. Hence the model is a bit simplistic, but we postulate it has more power, compared to pairwise interaction model, to detect interactions.

1.2 Classes of regression models

When we are concerned with the dependence of a response variable Y on observed risk factors $X_1, \ldots, X_p$, an equation that relates Y to $X_1, \ldots, X_p$ is usually called as *regression equation*. Denote the regression equation by

$$E(Y|X_1 = x_1, X_2 = x_2, \ldots, X_p = x_p) = g(x_1, x_2, \ldots, x_p) = g(\mathbf{x}).$$

First of all, we give a more precise definition of synergism/antagonism in terms of mathematical languages. For any pair $j < k$ and all $\epsilon, \delta \geq 0$, if

$$g(\mathbf{x} + \epsilon \mathbf{1}_j + \delta \mathbf{1}_k) - g(\mathbf{x} + \delta \mathbf{1}_k) - \left[g(\mathbf{x} + \epsilon \mathbf{1}_j) - g(\mathbf{x})\right] \geq (\leq)\, 0,$$

the interaction among $\mathbf{X}$ is synergistic (antagonistic). Note $\mathbf{1}_j$ means a $p \times 1$ vector of zero except that the jth element is 1.

Equivalently, if g is twice differentiable, the above inequality can be rewritten as

$$\frac{\partial^2}{\partial x_j \partial x_k} g(x_1, \ldots, x_p) \geq (\leq)\, 0, \; \forall j \neq k.$$

Note that there are other names for the above definition, such as supermodular, directionally convex and lattice-superadditive for nonnegative derivatives and submodular, directionally concave, lattice-subadditive for nonpositive derivatives. They are used to compare the dependence structure of random vectors having the same marginal distributions. More details and discussions are in Müller and Stoyan (2002).

In the following we list all the regression equations considered in the thesis.

1. Linear regression models.

$$g(x_1, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j.$$

Due to the simplicity and interpretability, linear regression models are the most widely used for either experimental data or observational data. The regression coefficient β_j implies how much change in the response variable is associated with a unit change in X_j when keeping all the other risk factors unchanged.

It is acknowledged that this model is a linear approximation of the relationship between Y and $\mathbf{X}$. By first-order Taylor expansion, it is easy to derive the model expression. Therefore, it may work well only for a local region, where the surface does not have curvature. That's the reason why prediction of $\mathbf{X}$ values outside of the range where we fit the model is usually dangerous and not reliable.

By second-order Taylor expansion, we have

$$g(x_1, \ldots, x_p) = \beta_0 + \sum_{j=1}^{p} \beta_j x_j + \sum_{1 \le i \le j \le p} \beta_{ij} x_i x_j.$$

Also this is good for a local region, where the surface has some curvature. Note that for a binary variable x_j, the term $\beta_{jj} x_j^2$ is not necessary. One thing worth mentioning is that the explanation of coefficients β_j's are different from that under the additive model. The expected change in response variable Y after one unit change in X_j (while keeping all the others unchanged) now breaks down into several pieces, i.e., $\beta_j x_j$ and $\beta_{ij} x_i x_j (i = 1, \ldots, p)$. For this model, for any pair $i < j$, we have synergistic interaction if $\beta_{ij} \geq 0$ and antagonistic interaction if $\beta_{ij} \leq 0$.

2. Spline regression models.

$$g(x_1, \ldots, x_p) = \sum_{j=1}^{p} m_j(x_j), \tag{1.1}$$

where m_j is a smoothing function applied to x_j. Note this model is additive and does not include interaction terms.

Taylor expansion allows to write m_j as

$$\begin{aligned} m_j(x) &= \sum_{k=1}^{D+1} \alpha_{jk} x^{k-1} + Rem_j(x), \ a \le x \le b, \ j = 1, \ldots, p \\ Rem_j(x) &= (D!)^{-1} \int_a^b m_j^{(D+1)}(x)(x-\xi)_+^D d\xi, \end{aligned}$$

where $(x - t_k)_+ = x - t_k$, if $x \ge t_k$ and zero otherwise.

Note that for fixed j, if $Rem_j(x)$ are uniformly small for all observed x_{ji} in magnitude, polynomial regression of X_j may provide a reasonable analysis. Otherwise, we need some other methods to take account of the item Rem_j. One method of estimation which attempts to guard against departures from polynomial models is smoothing splines. The basic premise is the integral in Rem_j can be approximated using the quadrature formula

$$Rem_j(x) \approx \sum_{k=1}^{L_j} a_{jk}(x - t_{jk})_+^D$$

for coefficients $a_{j1}, \ldots, a_{jL_j}$ and $t_{j1} < t_{j2} < \cdots < t_{jL_j}$ (t_{ji}'s are *knots* in the definition of *spline*, which is given in Section 2.4.1). Combining this with the original polynomial approximation leads to an overall approximation of the regression function by

$$m_j(x) = \alpha_{j1} x + \cdots + \alpha_{jD} x^D + \sum_{k=1}^{L_j} a_{jk}(x - t_{jk})_+^D,$$

which is the spline function under consideration in Chapter 2.

To introduce the interactions between $x_1, \ldots, x_j$, we use the following regression function, similar to pairwise interaction models,

$$g(x_1, \ldots, x_p) = \sum_{j=1}^{p} m_j(x_j) + \sum_{1 \le i,j \le p} [x_i m_j(x_j) + m_i(x_i) x_j] \,. \tag{1.2}$$

Note that there are other possibilities to determine the interactions between any pair of risk factors. Here for simplicity we only use the summation of the products of one risk factor and smoothing function of the other risk factor, which is different from that discussed in Gu (2002). In Gu (2002), the multivariate function is given by an ANOVA decomposition, that is, it is expressed as a constant plus the sum of functions of one variable (main effects), plus the sum of functions of two variables (two-factor interactions)and so on. Note the interactions are assumed to be in tensor product spaces.

Note that if we have numerous number of risk factors, the additive model (1.1) may be too generous allowing a few degrees of freedom per X_j and it does not take into account the interactions between X_j's yet. Classes of function are not "dense" or a universal approximation to class of smooth function in a rectangular region of $\prod_{j=1}^{p}[x_{jL}, x_{jU}]$, where x_{jL}, x_{jU} are the lower and upper bounds of jth risk factor respectively. In the interaction model (1.2), the number of interaction terms increases dramatically when the number of risk factor increases. In such a case, we may use projection pursuit regression (Friedman and Stuezle, 1981) or multivariate adaptive regression splines (Friedman, 1991). These two methods consist of universal approximation to smooth functions by a sum of nonlinear functions of linear combinations of x_j's, i.e.,

$$g(x_1, \ldots, x_p) \approx \sum_{i=1}^{M} g_i(a_{i1}x_1 + \cdots + a_{ip}x_p).$$

More discussion is in Venables and Ripley (1999) and Diaconis and Shahshahani (1984). Note that projection pursuit regression models and multivariate adaptive regression splines are not considered in the thesis, but it is possible and worthwhile to be studied in the future.

3. Diffuse interaction models.

As discussed in the previous section, some problems arise when fitting a pairwise interaction model especially with a large number of risk factors. For instance, high blood pressure is a warning signal for health problems. There are many risk factors which may cause high blood pressure.

Age: The risk of high blood pressure increases as you get older.

Gender: Women are more likely to develop high blood pressure after menopause.

Family history: High blood pressure tends to run in families.

Body weight: The greater your body mass, more risk of high blood pressure.

Tobacco use: The chemicals in tobacco can increase the risk of high blood pressure.

Sodium intake: Too much sodium can lead to increased blood pressure.

Excessive alcohol: Heavy drinking can damage your heart.

Stress: High levels of stress can lead to a (temporary) increase in blood pressure.

Therefore, for all the risk factors listed except gender, larger value means more risk of high blood pressure. There might be synergistic interactions among those risk factors. How to depict the magnitude of the synergism? As proposed in Gustafson et al. (2005), the *diffuse* interaction model is defined as

$$g(x_1, \ldots, x_p) = \beta_0 + \left\{ \sum_{j=1}^{p} (\beta_j x_j)^{\lambda} \right\}^{1/\lambda}, \tag{1.3}$$

with $\beta_j \geq 0, x_j \geq 0, \lambda > 0$. It is easy to verify that $\lambda < (>)1$ leads to

$$\frac{\partial g(\mathbf{x})}{\partial x_i \partial x_j} > (<) 0.$$

That is, $\lambda < (>)1$ means synergistic(antagonistic) interaction. Clearly the magnitude of λ is a measurement of degree of synergism/antagonism and curvature of the surface as well.

Note that when $\lambda = 1$, the above regression equation just reduces to an additive linear regression model. Note that

$$g(x_j, \mathbf{x}_{(j)} = 0) = \beta_0 + \beta_j x_j,$$

where $\mathbf{x}_{(j)} = (x_1, \ldots, x_{j-1}, x_{j+1}, \ldots, x_p)$ and $x_j = 0$ means the absence of jth risk factor. That is, no matter what λ is, β_j is the increase in response variable associated with a unit change in X_j, in the absence of all the other risk factors. Now the interpretation of β_j is different from that under linear regression model. The models are increasing function hence in absence of all risk factors, that is, $x_j = 0, j = 1, \ldots, p$, β_0 stands for the smallest expected response (usually risk of some diseases). Therefore, we may rewrite the diffuse interaction models in a more general sense as

$$g(\mathbf{x}) = \beta_0 + \|\mathbf{x}\|, \tag{1.4}$$

where $\|\cdot\|$ is a norm in $\mathbb{R}^p$. Even more generally, to get a regression function that is not monotone increasing, we may consider

$$g(\mathbf{x}) = \beta_0 + \|\mathbf{x} - \mathbf{a}\|,$$

where **a** is the vector of location parameters, standing for the values of risk factors that leads to the smallest expected response.

The diffuse interaction models (3.5) are a special class of (1.4) with the choice of norm being

$$\|\mathbf{x}\|_\beta = \left\{ \sum_{j=1}^{p} (\beta_j x_j)^\lambda \right\}^{1/\lambda}, \ \beta_j \geq 0, \ x_j \geq 0, \ \lambda > 0. \tag{1.5}$$

Here β_j's are inverse scale parameters. Naturally, $\|x\|$ can be interpreted as the distance between $\mathbf{x}$ and $\mathbf{0}$.

More general classes of norms can be used.

(a) If the variables x_j's interact in different directions, one may want to partition those variables depending on the direction of interactions among them, that is

$$g(\mathbf{x}) = \beta_0 + \sum_{j \in \text{ADD}} (\beta_j x_j) + \left\{ \sum_{j \in \text{SYN}} (\beta_j x_j)^{\lambda_1} \right\}^{1/\lambda_1} + \left\{ \sum_{j \in \text{ANT}} (\beta_j x_j)^{\lambda_2} \right\}^{1/\lambda_2},$$

with $\beta_j \geq 0, x_j \geq 0$, for all j and $0 < \lambda_1 < 1 < \lambda_2$. It is easy to verify that for any pair $j < k \in$SYN(ANT),

$$\frac{\partial g(\mathbf{x})}{\partial x_j \partial x_k} > (<)\, 0.$$

(b) Assuming there is a nesting of groups of variables, one might use the 2-level nested model

$$g(\mathbf{x}) = \left\{ \left[\sum_{i \in S_1} (\beta_i x_i)^{\lambda_1} \right]^{\lambda/\lambda_1} + \left[\sum_{i \in S_2} (\beta_i x_i)^{\lambda_2} \right]^{\lambda/\lambda_2} \right\}^{1/\lambda}, \tag{1.6}$$

where $\lambda_1 < 1$ and $\lambda_2 > 1$ for a synergistic set S_1 and an antagonistic set S_2 with some interaction between the two sets. Also this function can be extended to multiple levels of nesting. Now the second derivative of $g(\mathbf{x})$ in (1.6) is more complicated. For $\lambda > 1$, any pair (j, k) with $j \in S_1, k \in S_2$, the second derivative is negative. For $\lambda < 1$, any pair (j, k) with $j \in S_1, k \in S_2$, the second derivative is positive. If $1 < \lambda < \lambda_2$, for any pair

$(j,k) \in S_2$, the second derivative is negative. If $\lambda_1 < \lambda < 1$, for any pair $(j,k) \in S_1$, the second derivative is positive.

The following example illustrate one scenario where we might consider to use (1.6). Suppose X_1, X_2, Z are the predictors relating to response variable Y. However, Z can not be measured and is replaced by two surrogate variables X_3 and X_4. In this case, one might use

$$g(x_1,x_2,x_3,x_4) = \left\{(\beta_1 x_1)^\lambda + (\beta_2 x_2)^\lambda + (\beta_3 x_3 + \beta_4 x_4)^\lambda\right\}^{1/\lambda}.$$

1.3 Outline of thesis

The following chapters are organized as follows.

In Chapter 2, we study the consequences of fitting an additive regression model to data generated from a pairwise interaction model. We find out that under some particular situations, the average effect estimates based on misspecified model can still be consistent with true values. Further, under the framework of spline regression, we investigate the consequences of model misspecifications by failing to include interaction terms into model, when such terms exist.

In Chapter 3, we introduce diffuse interaction models as an alternative to pairwise interaction models when the number of risk factors of interest is rather large. And we compare its ability to detect interactions with a pairwise interaction model.

In Chapter 4, we propose a MCMC algorithm to estimate the parameters in diffuse interaction model, introduced in chapter 3. Further, we propose other MCMC algorithms for more general versions of diffuse interaction model by relaxing the assumption that all the risk factors interact in the same way, either synergistically or antagonistically.

In chapter 5, we summarize the results of the above three chapters and also discuss some possible problems to consider in future.

Chapter 2

Average effects for regression models with misspecifications

How bad is the estimation when the real relationship between response variable and predictors (i.e.risk factors in epidemiological studies) does involve interactions, but a model without interaction is fitted? More specifically, if the actual data generating mechanism is

$$Y = \sum_{j=1}^{p} h_j(X_j) + \sum_{1 \le i < j \le p} h_{ij}(X_i, X_j) + \epsilon, \tag{2.1}$$

what will the result be if we fit an additive model

$$Y = \sum_{j=1}^{p} h_j(X_j) + \eta, \tag{2.2}$$

where both ϵ and η denote random errors, postulated to follow normal distributions.

2.1 Average effect

To evaluate the performance of the estimation under misspecified models, we apply the idea of *average effect*, which is proposed in Gelman and Pardoe (2007) and also in Gustafson et al. (2005). The reasons to introduce the average effect are listed as the following.

(a) For comparing models with different parametric forms, different sets of parameters

have different interpretations, so it is hard to explain the difference between the parameters from different models. For example, postulate the parametric forms of models (2.1) and (2.2) as below.

In model (2.1):

$$\begin{aligned} h_j(X_j) &= \beta_j X_j, \\ h_{ij}(X_i, X_j) &= \beta_{ij} X_i X_j. \end{aligned}$$

In model (2.2):

$$h_j(X_j) = \alpha_j X_j.$$

Then β_{ij}'s in model (2.1) are the parameters describing the pairwise interactions between the predictors, while no such parameters appear in model (2.2). It is hard to interpret the meaning of the difference between β_j and α_j. More precisely, the expected change in response variable caused by one unit change of X_j keeping other predictors unchanged in model (2.1) is $\beta_j + \sum_{i \neq j} \beta_{ij} X_i$, while in model (2.2) it is α_j.

(b) More generally, if the two models of interest are quite different, no common parameters could be compared. For example, say one model is a linear regression model while the other is a nonparametric regression model. Now it is impossible to compare the estimates of coefficients from the former model and the estimate of a function/curve from the latter model.

To handle the difficulty mentioned above, we need a quantity which is associated with something in common among different models. The average effect is such a quantity. The definition of average effect is as follows.

If the predictor X_j takes value in a continuous space, then the *predictive effect* of X_j

is defined as

$$\Delta_j(\mathbf{x};\boldsymbol{\theta}) = \frac{\partial \mathrm{E}\left(Y|X_j = x_j, \mathbf{X}_{(j)} = \mathbf{x}_{(j)};\boldsymbol{\theta}\right)}{\partial x_j},$$

where $\mathbf{X}_{(j)} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$, i.e., a random vector comprising all the explanatory variables except the j-th one, and $\boldsymbol{\theta}$ denotes the parameter vector in the model. If the predictor X_j can only take finitely many values, then the definition of its predictive effect is

$$\Delta_j(\mathbf{x}_{(j)}, x_j^{(1)}, x_j^{(2)};\boldsymbol{\theta}) = \frac{1}{x_j^{(1)} - x_j^{(2)}}\left\{\mathrm{E}\left(Y|X_j = x_j^{(1)}, \mathbf{X}_{(j)} = \mathbf{x}_{(j)};\boldsymbol{\theta}\right) - \mathrm{E}\left(Y|X_j = x_j^{(2)}, \mathbf{X}_{(j)} = \mathbf{x}_{(j)};\boldsymbol{\theta}\right)\right\},$$

where $x_j^{(1)}, x_j^{(2)}$ are a pair of different possible values of X_j. The quantity of predictive effect reflects the change in $\mathrm{E}(Y|\mathbf{X})$ associated with asmall change in the j-th predictor X_j conditioned on a specific value of $\mathbf{X}_{(j)} = \mathbf{x}_{(j)}$.

The reason to use the notation $\Delta_j(\mathbf{x};\boldsymbol{\theta})$ and not $\Delta_j(\mathbf{x}_{(j)};\boldsymbol{\theta})$ is that in general predictive effect is a function of x_j as well. For example, in the quadratic model with continuous predictors, that is

$$\mathrm{E}(Y|X) = \beta_0 + \sum_{j=1}^{p}\beta_j X_j + \sum_{j=1}^{p}\beta_{jj}X_j^2 + \sum_{1\le i<j\le p}\beta_{ij}X_iX_j,$$

the predictive effect of X_j is $2\beta_{jj}X_j + \sum_{i\neq j}\beta_{ij}X_i$. However, in some special situations such as model (2.10) and model (2.11), appearing in the next section, $\Delta_j(\mathbf{x};\boldsymbol{\theta})$ does not depend on x_j.

Based on the definition of predictive effect, we can define different versions of *average effect* as the expected value of $\Delta_j(\mathbf{x};\boldsymbol{\theta})$ with respect to different distributions.

Definition 1: If averaging the predictive effect over the joint distribution of $\mathbf{X}_{(j)}$, all the predictors except X_j, then the average effect of the j-th predictor is defined as

$$\delta_j(x_j;\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{X}_{(j)}}\{\Delta_j(X_j = x_j, \mathbf{X}_{(j)};\boldsymbol{\theta})\}. \tag{2.3}$$

Definition 2: If averaging the predictive effect over the joint distribution of all predictors $X_1, \ldots, X_p$, then the average effect of the j-th predictor is defined as

$$\delta_j(\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{X}}\{\Delta_j(\mathbf{X};\boldsymbol{\theta})\}. \tag{2.4}$$

Definition 3: If averaging the predictive effect over the conditional distribution of $\mathbf{X}_{(j)}|X_j$, then the average effect of the j-th predictor is defined as

$$\delta_j(x_j;\boldsymbol{\theta}) = \mathrm{E}_{\mathbf{X}_{(j)}|X_j}\{\Delta_j(X_j = x_j, \mathbf{X}_{(j)};\boldsymbol{\theta})\}. \tag{2.5}$$

In general the three definitions are *not* identical to each other. But in the special cases when the predictive effect of X_j does not depend on the value of X_j, as in model (2.10) and model (2.11), the first two definitions are the same because both are obtained by averaging over the joint distribution of $\mathbf{X}_{(j)}$. In the following sections, we stick to use Definition 2 unless specified particularly.

Note that the predictive effect of X_j is based on the conditional distribution of $Y|X$, while the average effect of X_j is defined with respect to the joint distribution of $(Y, \mathbf{X}_{(j)})$ or (Y, X). We should also keep in mind that the idea of average effect is *not* just confined within regression context. For example, Xu and O'Quigley (2000) and Gustafson (2007) apply this concept in survival analysis by averaging hazard function over the joint distribution of $(T, \mathbf{X})$, where T is failure time.

2.2 General results

Suppose we have p predictors, $X_1,\ldots,X_p$, and response variable Y, whose relationship with X_j's is of interest. For the general framework, let

$$\mathbf{T} = \big(T_1(X_1,\ldots,X_p),\ldots,T_t(X_1,\ldots,X_p)\big)',$$

whose components are the functions of predictors involved in the "true" relationship between response variable and predictors, and let

$$\mathbf{S} = \big(S_1(X_1,\ldots,X_p),\ldots,S_s(X_1,\ldots,X_p)\big)',$$

denote the functions of the predictors in the fitted model.

By allowing general forms of functions, even nonparametric forms, involved in the components of $\mathbf{S}$ and $\mathbf{T}$, we do have a rather broad possibilities of h_j and h_{ij} in (2.2) and (2.1). In the forthcoming sections, we will see how this general setting applies in different regression models.

Hence, models (2.2) and (2.1) can be rewritten as follows, respectively.

$$\begin{aligned} Y &= \mathbf{S}'\boldsymbol{\alpha} + \tilde{\epsilon}, \\ Y &= \mathbf{T}'\boldsymbol{\beta} + \epsilon, \end{aligned}$$

where ϵ follows $N(0,\sigma^2)$ and are independent of $\mathbf{X}$. Denote by $\delta_j(\boldsymbol{\beta})$, $\delta_j^*(\boldsymbol{\beta})$ the average effects of the j-th predictor under the "true" model and fitted one respectively.

Assuming X_j's are continuous and $\mathbf{T}, \mathbf{S}$ are differentiable, let

$$\begin{aligned}\widetilde{\mathbf{T}}_j &= \left(\frac{\partial}{\partial X_j}T_1(X_1,\ldots,X_p),\ldots,\frac{\partial}{\partial X_j}T_t(X_1,\ldots,X_p)\right)',\\ \widetilde{\mathbf{S}}_j &= \left(\frac{\partial}{\partial X_j}S_1(X_1,\ldots,X_p),\ldots,\frac{\partial}{\partial X_j}S_s(X_1,\ldots,X_p)\right)'.\end{aligned}$$

So that the true average effect of X_j, denoted by $\delta_j(\boldsymbol{\beta})$, is $[\mathrm{E}\,(\widetilde{\mathbf{T}}_j)]'\boldsymbol{\beta}$. Now let n be a sample size with $(x_{i1},\ldots,x_{ip},y_i)$ being the observations of predictors and response variable for the ith subject. $(x_{i1},\ldots,x_{ip}), i=1,\ldots,n$ are assumed to be iid replications of $(X_1,\ldots,X_p)$ and $\epsilon_1,\ldots,\epsilon_n$ are iid $N(0,\sigma^2)$. Let $\mathbb{T}$ be the $n\times t$ design matrix with (i,k) element equal to $T_k(X_{i1},\ldots,X_{ip})$. Let $\widetilde{\mathbb{T}}_j$ be $n\times t$ design matrix with (i,k) element equal to $\frac{\partial}{\partial X_j}T_k(X_{i1},\ldots,X_{ip})$. Assuming $\mathbb{T}$ is of full column rank, we have

$$\begin{aligned}\widehat{\boldsymbol{\beta}} &= (\mathbb{T}'\mathbb{T})^{-1}\mathbb{T}'\mathbb{Y},\\ \widehat{\mathrm{E}\,(\widetilde{\mathbf{T}}_j)} &= n^{-1}\widetilde{\mathbb{T}}_j'\mathbf{1}_{n\times 1}.\end{aligned}$$

where $\mathbb{Y}=(y_1,y_2,\ldots,y_n)'$.

Hence the estimate of the average effect of X_j is obtained by

$$\widehat{\delta_j(\boldsymbol{\beta})} = n^{-1}\mathbf{1}'\widetilde{\mathbb{T}}_j(\mathbb{T}'\mathbb{T})^{-1}\mathbb{T}'\mathbb{Y}.$$

With

$$\mathbb{Y}=\mathbb{T}\boldsymbol{\beta}+\boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon}=(\epsilon_1,\ldots,\epsilon_n)'$, we get

$$\widehat{\delta_j(\boldsymbol{\beta})} = n^{-1}\mathbf{1}'\widetilde{\mathbb{T}}_j\boldsymbol{\beta} + n^{-1}\mathbf{1}'\widetilde{\mathbb{T}}_j(\mathbb{T}'\mathbb{T})^{-1}\mathbb{T}'\boldsymbol{\epsilon}.$$

Therefore, we have

$$\sqrt{n}\left(\widehat{\delta_j(\boldsymbol{\beta})} - \delta_j(\boldsymbol{\beta})\right) = \sqrt{n}\left(n^{-1}\mathbf{1}'\widetilde{\mathbb{T}}_j - \mathrm{E}\left(\widetilde{\mathbf{T}}_j\right)'\right)\boldsymbol{\beta} + n^{-1}\mathbf{1}'\widetilde{\mathbb{T}}_j(\mathbb{T}'\mathbb{T}/n)^{-1}n^{-1/2}\mathbb{T}'\boldsymbol{\epsilon}.$$

The first term in the right side of the above equation can be rewritten as

$$n^{-1/2}\sum_{i=1}^{n}\left((\widetilde{\mathbb{T}}_j)_{i\cdot} - \mathrm{E}\left(\widetilde{\mathbf{T}}_j\right)'\right)\boldsymbol{\beta},$$

where $(\widetilde{\mathbb{T}}_j)_{i\cdot}$ denotes the ith row of $\widetilde{\mathbb{T}}_j$. Since $(\widetilde{\mathbb{T}}_j)_{i\cdot}$ are i.i.d., the above term converges in distribution to $N(0, \boldsymbol{\beta}'\mathrm{Var}\left(\widetilde{\mathbf{T}}_j\right)\boldsymbol{\beta})$ by the multivariate central limit theorem.

With the conditional variance identity and the assumption that $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$, we have

$$\begin{aligned}\mathrm{Var}\left(\mathbb{T}_{i\cdot}{}'\epsilon_i\right) &= \mathrm{E}\left(\mathrm{Var}\left(\mathbb{T}_{i\cdot}{}'\epsilon_i|X_i\right)\right) + \mathrm{Var}\left(\mathrm{E}\left(\mathbb{T}_{i\cdot}{}'\epsilon_i|X_i\right)\right)\\ &= \mathrm{E}\left(\mathbf{T}\mathbf{T}'\right)\sigma^2.\end{aligned}$$

Thus, by the multivariate central limit theorem,

$$n^{-1/2}\sum_{i=1}^{n}\mathbb{T}_{i\cdot}{}'\epsilon_i \xrightarrow{D} N\left(0, \mathrm{E}\left(\mathbf{T}'\mathbf{T}\right)\sigma^2\right).$$

Since the $(\widetilde{\mathbb{T}}_j)_{i\cdot}$ are i.i.d., by the strong law of large numbers, we have

$$n^{-1}\mathbf{1}'\widetilde{\mathbb{T}}_j \to \left\{\mathrm{E}\left(\widetilde{\mathbf{T}}_j\right)\right\}'.$$

If $E(\mathbf{T}\mathbf{T}')$ is invertible, then

$$\left(n^{-1}\mathbb{T}'\mathbb{T}\right)^{-1} \xrightarrow{a.s.} \left\{\mathrm{E}\left(\mathbf{T}\mathbf{T}'\right)\right\}^{-1}.$$

With the above results, a straightforward conclusion can be summarized as the following.

Result 1: Assuming the existence of $\{\mathrm{E}(\mathbf{TT}')\}^{-1}$ and ϵ is independent of $\mathbf{X}$,

$$\sqrt{n}\left(\widehat{\delta_j(\boldsymbol{\beta})} - \delta_j(\boldsymbol{\beta})\right) \xrightarrow{D} N(0, v_j(\boldsymbol{\beta})),$$

where

$$v_j(\boldsymbol{\beta}) = \sigma^2[\mathrm{E}(\widetilde{\mathbf{T}}_j)]'[\mathrm{E}(\mathbf{TT}')]^{-1}\mathrm{E}(\widetilde{\mathbf{T}}_j) + \boldsymbol{\beta}'\mathrm{Var}(\widetilde{\mathbf{T}}_j)\boldsymbol{\beta}. \tag{2.6}$$

Similarly, in the fitted model, let $\mathbb{S}$ be the $n \times s$ design matrix with (i,k) element equal to $S_k(X_{i1}, \ldots, S_{ip})$. Let $\widetilde{\mathbb{S}}_j$ be $n \times s$ design matrix with (i,k) element equal to $\frac{\partial}{\partial X_j} S_k(X_{i1}, \ldots, X_{ip})$. Assuming that $\mathbb{S}$ is of full column rank, the estimate of the average effect of X_j with the "misspecified" model is

$$\widehat{\delta_j^*(\boldsymbol{\beta})} \;=\; n^{-1}\mathbf{1}'\widetilde{\mathbb{S}}_j(\mathbb{S}'\mathbb{S})^{-1}\mathbb{S}'\mathbb{Y}.$$

Assuming $E(\mathbf{SS}')$ is invertible,

$$\begin{aligned}
\widehat{\boldsymbol{\alpha}} &= (\mathbb{S}'\mathbb{S})^{-1}\mathbb{S}'\mathbb{Y} \\
&\xrightarrow{a.s.} \{\mathrm{E}(\mathbf{SS}')\}^{-1}\mathrm{E}(\mathbf{S}Y) \\
&= \{\mathrm{E}(\mathbf{SS}')\}^{-1}\{\mathrm{E}(\mathbf{ST}')\}\boldsymbol{\beta} \\
&\triangleq \boldsymbol{\alpha}_*.
\end{aligned}$$

Thus, in the limit as $n \to \infty$, $\delta_j^*(\boldsymbol{\beta}) = \lim \widehat{\delta_j^*(\boldsymbol{\beta})} = \mathrm{E}(\widetilde{\mathbf{S}}_j)'\boldsymbol{\alpha}_*$. Hence, with $\mathbb{Y} = \mathbb{T}\beta + \epsilon$, we have

$$\begin{aligned}
\sqrt{n}\big(\widehat{\delta_j^*(\boldsymbol{\beta})} - \delta_j^*(\boldsymbol{\beta})\big) \;=\; & n^{-1/2}\sum_{i=1}^{n}\left((\widetilde{\mathbb{S}}_j)_{i\cdot} - \mathrm{E}(\widetilde{\mathbf{S}}_j)'\right)(\mathbb{S}'\mathbb{S})^{-1}\mathbb{S}'\mathbb{Y} + \\
& \{\mathrm{E}(\widetilde{\mathbf{S}}_j)\}'(\mathbb{S}'\mathbb{S}/n)^{-1}n^{-1/2}\mathbb{S}'(\mathbb{T}\boldsymbol{\beta} - \mathbb{S}\boldsymbol{\alpha}_* + \boldsymbol{\epsilon}),
\end{aligned} \tag{2.7}$$

where $(\widetilde{\mathbb{S}}_j)_{i\cdot}$ is the ith row of $\widetilde{\mathbb{S}}_j$. By the strong law of large numbers,

$$n^{-1}\mathbb{S}'\mathbb{Y} \to \mathrm{E}(\mathbf{S}Y)(=\{\mathrm{E}(\mathbf{S}\mathbf{T}')\}\boldsymbol{\beta}),$$
$$(\mathbb{S}'\mathbb{S}/n)^{-1} \xrightarrow{a.s.} \{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}.$$

With the multivariate central limit theorem and the above facts, the first term in (2.7) asymptotically follows a multivariate normal distribution with mean vector of $\mathbf{0}$ and covariance matrix of $\boldsymbol{\alpha}_*'\mathrm{Var}(\widetilde{\mathbf{S}}_j)\boldsymbol{\alpha}_*$. Also by multivariate central limit theorem, we have

$$n^{-1/2}\sum_{i=1}^{n}\mathbb{S}_{i\cdot}'\left(\mathbb{T}_{i\cdot}\boldsymbol{\beta} - \mathbb{S}_{i\cdot}\boldsymbol{\alpha}_* + \epsilon_i\right) \xrightarrow{D} N(0, V^*),$$

where $\mathbb{S}_{i\cdot}$ is the ith row of $\mathbb{S}$, $\mathbb{T}_{i\cdot}$ is the ith row of $\mathbb{T}$ and

$$\begin{aligned} V^* &= \mathrm{E}\left\{\left(\mathbb{T}_{i\cdot}\boldsymbol{\beta} - \mathbb{S}_{i\cdot}\boldsymbol{\alpha}_* + \epsilon_i\right)^2\mathbb{S}_{i\cdot}'\mathbb{S}_{i\cdot}\right\} \\ &= \sigma^2\mathrm{E}(\mathbf{S}\mathbf{S}') + \mathrm{E}\left((\mathbf{T}'\boldsymbol{\beta} - \mathbf{S}'\boldsymbol{\alpha}_*)^2\mathbf{S}\mathbf{S}'\right). \end{aligned} \tag{2.8}$$

Note that the first part is due to random error and the second part is due to model misspecification. Therefore, the asymptotic distribution of the second term in (2.7) is a multivariate normal with mean vector of $\mathbf{0}$ and covariance matrix of

$$\{\mathrm{E}(\widetilde{\mathbf{S}}_j)\}'\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}V^*\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}\{\mathrm{E}(\widetilde{\mathbf{S}}_j)\}.$$

Immediately the combination of the above results leads to the following result.

Result 2: Assuming the existence of $\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}$ and $\boldsymbol{\epsilon}$ is independent of $\mathbf{X}$,

$$\sqrt{n}\left(\widehat{\delta_j^*(\boldsymbol{\beta})} - \delta_j^*(\boldsymbol{\beta})\right) \xrightarrow{D} N(0, v_j^*(\boldsymbol{\beta})),$$

where

$$v_j^*(\boldsymbol{\beta}) = \{\mathrm{E}(\widetilde{\mathbf{S}}_j)\}'\{\mathrm{E}(\mathbf{SS}')\}^{-1}V^*\{\mathrm{E}(\mathbf{SS}')\}^{-1}\{\mathrm{E}(\widetilde{\mathbf{S}}_j)\} + 2\boldsymbol{\alpha}_*'\mathrm{Cov}\left(\widetilde{\mathbf{S}}_j, \mathbf{S}(\mathbf{T}'\boldsymbol{\beta} - \mathbf{S}'\boldsymbol{\alpha}_*)\right)\{\mathrm{E}(\mathbf{SS}')\}^{-1}\mathrm{E}(\widetilde{\mathbf{S}}_j) + \boldsymbol{\alpha}_*'\mathrm{Var}(\widetilde{\mathbf{S}}_j)\boldsymbol{\alpha}_*. \quad (2.9)$$

Remark. In some instances $\widetilde{\mathbf{T}}_j$ and/or $\widetilde{\mathbf{S}}_j$ might be constant, in which case the second term in (2.6) and the last two terms in (2.9) vanishes. Particularly, if the "true" regressors include pairwise products from $\mathbf{X}$, but the fitted model includes only linear terms from $\mathbf{X}$, then the second term in (2.6) does not vanish, but the last two terms in (2.9) do vanish.

Result 1 and Result 2 give the asymptotic distributions of the average effect estimates based on "true" model and "misspecified" model, respectively. Combining the two results, we can get a consistent average effect estimator from the misspecified model as long as $\delta_j(\boldsymbol{\beta}) = \delta_j^*(\boldsymbol{\beta})$. Some easily-studied cases where the above equality establishes are shown in *Result 3* later. Also Result 1 and 2 make it possible to compare the efficiencies of the two estimators from "true" model and "misspecified" model. More discussion is given in Section 2.3.2.

One thing worth investigating here is to find the consistent estimator, under the "misspecified" model fitting, of the mean squared error $\sigma^2(\boldsymbol{\alpha}) = \mathrm{E}(Y - \mathbf{S}'\boldsymbol{\alpha})^2$.

A least squares estimator $\widehat{\boldsymbol{\alpha}}$ is a parameter vector that solves the problem

$$\min_{\boldsymbol{\alpha}} s_n^2(\boldsymbol{\alpha}) = (n-p-1)^{-1}\|\mathbb{Y} - \mathbb{S}\widehat{\boldsymbol{\alpha}}\|^2.$$

Now $s_n^2(\widehat{\boldsymbol{\alpha}})$ may be rewritten in terms of matrices,

$$s_n^2(\widehat{\boldsymbol{\alpha}}) = \frac{n}{n-p-1}\left\{n^{-1}\mathbb{Y}'\mathbb{Y} - (n^{-1}\mathbb{Y}'\mathbb{S})(n^{-1}\mathbb{S}\mathbb{S})^{-1}(n^{-1}\mathbb{S}'\mathbb{Y})\right\}.$$

By the strong law of large numbers,

$$(n^{-1}\mathbb{Y}'\mathbb{Y}) \xrightarrow{a.s.} \mathrm{E}(Y^2),$$
$$(n^{-1}\mathbb{S}'\mathbb{Y}) \xrightarrow{a.s.} \mathrm{E}(\mathbf{S}Y).$$

Therefore,

$$s_n^2(\widehat{\boldsymbol{\alpha}}) \xrightarrow{a.s.} \mathrm{E}(Y^2) - \mathrm{E}(\mathbf{S}'Y)\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}\mathrm{E}(\mathbf{S}Y).$$

Note that the $\boldsymbol{\alpha}_*$ is the unique minimizer of mean squared error $\sigma^2(\boldsymbol{\alpha})$. Now

$$\begin{aligned}
\sigma^2(\boldsymbol{\alpha}_*) &= \mathrm{E}(Y)^2 - 2\mathrm{E}(Y\mathbf{S}'\boldsymbol{\alpha}_*) + \mathrm{E}(\mathbf{S}'\boldsymbol{\alpha}_*)^2 \\
&= \mathrm{E}(Y^2) - \mathrm{E}(\mathbf{S}'\boldsymbol{\alpha}_*)^2 \ \left(\mathrm{E}(Y|\mathbf{S}) = \mathbf{S}'\boldsymbol{\alpha}_*\right) \\
&= \mathrm{E}(Y^2) - \boldsymbol{\alpha}_*'\mathrm{E}(\mathbf{S}\mathbf{S}')\boldsymbol{\alpha}_* \\
&= \mathrm{E}(Y^2) - \mathrm{E}(\mathbf{S}'Y)\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}\mathrm{E}(\mathbf{S}Y).
\end{aligned}$$

That is, $s_n^2(\widehat{\boldsymbol{\alpha}})$ is a consistent estimator of $\sigma^2(\boldsymbol{\alpha}_*)$. However,

$$(\mathbb{S}'\mathbb{S}/n)^{-1}s_n^2(\widehat{\boldsymbol{\alpha}}) \xrightarrow{a.s.} \{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}\sigma^2(\boldsymbol{\alpha}_*),$$

which is not equal to $\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}V^*\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}$ in general. Note the fact that, by the previous analyses and (2.8), we have

$$\begin{aligned}
\lim_{n\to\infty} \mathrm{Var}(\sqrt{n}\widehat{\boldsymbol{\alpha}}) &\triangleq v(\boldsymbol{\alpha}_*) \\
&= \{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}V^*\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1} \\
&= \{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}\sigma^2(\boldsymbol{\alpha}_*) + \{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}\mathrm{E}\left((\mathbf{T}'\boldsymbol{\beta} - \mathbf{S}'\boldsymbol{\alpha}_*)^2\mathbf{S}\mathbf{S}'\right)\{\mathrm{E}(\mathbf{S}\mathbf{S}')\}^{-1}.
\end{aligned}$$

Therefore, the true variances of the coefficients estimates from the fitted model are big-

ger than the reported standard errors of those estimates, which is caused by the model misspecification, the second piece in the last equality of the above equation. If no misspecification occurred, then $V^* = \mathrm{E}(\mathbf{SS}')$, then $(\mathbb{S}'\mathbb{S})s_n^2(\widehat{\boldsymbol{\alpha}})$ is the consistent estimator of $\mathrm{var}(\widehat{\boldsymbol{\alpha}})$. Let

$$\widehat{V}_{\boldsymbol{\alpha}} = n^{-1}\sum_{i=1}^{n}\left(Y_i - \mathbb{S}_{i\cdot}\widehat{\boldsymbol{\alpha}}\right)^2\mathbb{S}_{i\cdot}'\mathbb{S}_{i\cdot},$$

then it is not hard to derive that $\widehat{V}_{\boldsymbol{\alpha}}$ is a consistent estimator of V^* by a similar derivation as before. Therefore a large difference between $\widehat{V}_{\boldsymbol{\alpha}}$ and $\mathbb{S}'\mathbb{S}s_n^2(\widehat{\boldsymbol{\alpha}})$ can be an evidence for model misspecification (see White (1980) for a formal test for misspecification). Hence $v(\boldsymbol{\alpha}_*)$ can be consistently estimated by

$$n(\mathbb{S}'\mathbb{S})^{-1}\left\{\sum_{i=1}^{n}\left(Y_i - \mathbb{S}_{i\cdot}\widehat{\boldsymbol{\alpha}}\right)^2\mathbb{S}_{i\cdot}'\mathbb{S}_{i\cdot}\right\}(\mathbb{S}'\mathbb{S})^{-1}.$$

There are other ways to get consistent estimator of $v(\boldsymbol{\alpha}_*)$ as well. One way is so-called "sandwich" method, which is based on the derivatives of log-likelihood function. Actually it is exactly the same as the estimator used in White (1982), where White studied the asymptotic distribution of the *maximum likelihood estimate* in case of model misspecification. As known, the least squares estimate is the same as the maximum likelihood estimate in the linear regression. Under a simple setting with only two predictors involved, we can also apply the methodology in White (1982) and get the same result as Result 2. Details are shown in Appendix II.

2.3 Linear regression

In this section, we assume that the response variable Y depends linearly on the predictors X_i's and that Y given $X = (X_1, \ldots, X_p)$ is normally distributed with a constant variance.

Suppose the true relationship between Y and X_i's is

$$Y|\mathbf{X} \sim N\big(\beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \beta_{12} X_1 X_2 + \cdots + + \beta_{p-1,p} X_{p-1} X_p, \sigma^2\big), \qquad (2.10)$$

while the fitted model is

$$Y|\mathbf{X} \sim N\big(\alpha_0 + \alpha_1 X_1 + \cdots + \alpha_p X_p, \tau^2\big). \qquad (2.11)$$

Here $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{12}, \ldots, \beta_{p-1,p})$, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_p)$. By Definition 2 of average effect, it is direct to derive the following average effects based respectively on (2.10) and (2.11):

$$\delta_j(\boldsymbol{\beta}) = \beta_j + \sum_{i \neq j} \beta_{ij} \mathrm{E}\,(X_i), \qquad (2.12)$$

$$\delta_j^*(\boldsymbol{\beta}) = \alpha_j. \qquad (2.13)$$

Naturally, as an informative summary we study the estimates of the average effect defined as

$$\begin{aligned} \widehat{\delta_j(\boldsymbol{\beta})} &= n^{-1} \sum_{i=1}^{n} \Delta_j(\mathbf{x}_{i(j)}, \widehat{\boldsymbol{\beta}}) = \hat{\beta}_j + \sum_{i \neq j} \hat{\beta}_{ij} \bar{x}_i, \\ \widehat{\delta_j^*(\boldsymbol{\beta})} &= n^{-1} \sum_{i=1}^{n} \Delta_j(\mathbf{x}_{i(j)}, \widehat{\boldsymbol{\alpha}}) = \hat{\alpha}_j, \end{aligned}$$

where $\mathbf{x}_{i(j)}$ is the i-th observation of $\mathbf{X}_{(j)}$ and $\bar{x}_i$ is the sample mean of the observations of X_i. The estimates of parameters $\widehat{\boldsymbol{\beta}}$ and $\widehat{\boldsymbol{\alpha}}$ are all least squares estimates. Note that in the normal linear regression context, the maximum likelihood estimates of the parameters are the same as the least squares estimates.

In the following three sections, we will compare the estimators of two average effects.

2.3.1 Difference in large sample limits of the two average effect estimators

Without loss of generality, we can think about centering predictors in the fitted model, that is,

$$\mathrm{E}(Y|\mathbf{X}) = \alpha_0 + \alpha_1\widetilde{X}_1 + \cdots + \alpha_p\widetilde{X}_p,$$

where $\widetilde{X}_i = X_i - \mathrm{E}(X_i)$, i.e., $\mathrm{E}(\widetilde{\mathbf{X}}) = \mathbf{0}$. Note that the estimates of $\alpha_1, \ldots, \alpha_p$ are unchanged by centering.

The large-sample limit of $\widehat{\boldsymbol{\alpha}}$, denoted by $\boldsymbol{\alpha}_*$, satisfies

$$\mathrm{E}\left\{\begin{pmatrix}1\\ \widetilde{X}_1\\ \vdots\\ \widetilde{X}_p\end{pmatrix}(1, \widetilde{X}_1, \ldots, \widetilde{X}_p)\right\}\boldsymbol{\alpha}_* = \mathrm{E}\left\{Y\begin{pmatrix}1\\ \widetilde{X}_1\\ \vdots\\ \widetilde{X}_p\end{pmatrix}\right\}.$$

Since the equation is symmetric in the p predictors, it suffices to determine the relationship between α_{1*} and β_1.

By Cramer's Rule, we can solve the equation above and get

$$\boldsymbol{\alpha}_{*1} = \beta_1 + \sum_{j>1}\beta_{1j}\mathrm{E}(X_j) + \frac{\sum_{i<j}\beta_{ij}[\sum_{k=1}^{p}\Sigma^{k1}\mathrm{E}(\widetilde{X}_i\widetilde{X}_j\widetilde{X}_k)]}{|\Sigma|}, \tag{2.14}$$

where $\Sigma = (\sigma_{ij})_{p\times p}$ is the covariance matrix of $(X_1, \cdots, X_p)$ and Σ^{k1} is the cofactor of Σ. Details of the proof of (2.14) are in Appendix I. Moreover, based on (2.3), expressions of average effect, the above results can be summarized as below.

Result 3. Assume that the variables are centred, i.e., $\mathrm{E}(\mathbf{X}) = 0$. Let $\mathbf{T} = (1, \mathbf{X}', \mathbf{W}')'$ where $\mathbf{W} = (X_1X_2, X_1X_3, \ldots, X_{p-1}X_p)$, i.e., the true relationship involves pairwise interactions. Also let $\mathbf{S} = (1, \mathbf{X}')'$, i.e., the interactions are undetected or ignored in the

modelling process.

Then (i)

$$\delta_j^*(\boldsymbol{\beta}) = \delta_j(\boldsymbol{\beta}) + \frac{\sum_{i<l} \beta_{il}[\sum_{k=1}^{p} \Sigma^{kj} \mathrm{E}\,(X_i X_l X_k)]}{|\Sigma|}, \tag{2.15}$$

where Σ is the covariance matrix of $\mathbf{X}$, and Σ^{kj} is the cofactor of the determinant corresponding to element σ_{kj}.

Consequently

(ii) if $\mathbf{X}$ has a multivariate normal distribution, or if the components of $\mathbf{X}$ are independent, then

$$\delta_j^*(\boldsymbol{\beta}) - \delta_j(\boldsymbol{\beta}) \;=\; 0.$$

Particularly, when $p = 2$, the connection equation (2.15) can be simplified as below

$$\delta_1^*(\boldsymbol{\beta}) = \delta_1(\boldsymbol{\beta}) + \beta_{12}\left[\frac{\mathrm{Var}\,(X_2)\mathrm{E}\,(X_1^2 X_2) - \mathrm{Cov}(X_1, X_2)\mathrm{E}\,(X_1 X_2^2)}{\mathrm{Var}\,(X_1)\mathrm{Var}\,(X_2) - \mathrm{Cov}^2(X_1, X_2)}\right]. \tag{2.16}$$

Remarks: The two conditions, both independence and multivariate normality, are sufficient but not necessary to get the consistency of "wrong-model" estimate $\widehat{\delta_j^*(\boldsymbol{\beta})}$. The condition that $\mathbf{X}$ follows a multivariate normal distribution can be replaced by an elliptical distribution. The latter actually is an extension of the former. The equation (2.15) shows that the bias is controlled by $\mathrm{E}\,(X_i X_l X_k)$. Suppose the joint distribution of $\mathbf{X}$ is an elliptical distribution $EC(\mathbf{0}, \Sigma, \phi)$ and its characteristic function is $\exp(i\mathbf{t}'\boldsymbol{\mu})\phi(\mathbf{t}'\Sigma\mathbf{t})$, where ϕ is a scalar function and called characteristic generator.

If i, j, k are different to one another, the joint distribution of (X_i, X_l, X_k) is also an elliptical distribution $EC(\mathbf{0}, \Sigma_{ilk}, \phi)$ where Σ_{ilk} is the corresponding sub-matrix of Σ

associated with (X_i, X_l, X_k). By the definition we have

$$(X_i, X_l, X_k)' = A'(Z_i, Z_l, Z_k)',$$

where (Z_i, Z_l, Z_k) follows a spherical distribution (which is a special case of elliptical distribution by setting $\Sigma = \mathbf{I}$). A is the lower triangular matrix in the Cholesky decomposition of Σ_{ilk}. Hence we can rewrite the above as

$$\begin{aligned} X_i &= aZ_i, \\ X_l &= bZ_i + cZ_l, \\ X_k &= dZ_i + eZ_l + fZ_k. \end{aligned}$$

It is straightforward to get $\mathrm{E}(X_i X_l X_k) = 0$ because that $\mathrm{E}(Z_i^3) = 0, \mathrm{E}(X_i^2 Z_l) = 0$ and $\mathrm{E}(Z_i Z_l Z_k) = 0$. If $k = i$ or $k = l$, we can also show the expectation to be zero by replacing the third equation by either the first one $(k = i)$ or the second one $(k = l)$.

The second part of Result 3 says that for certain distributions on $\mathbf{X}$, a model ignoring interactions will yield consistent estimates of average effects, even though the true regression relationship involve interactions. In addition to being of conceptual interest, this suggests some practical modelling strategies. For instance, in applications where one wishes to avoid modelling interactions explicitly, one might attempt to pre-transform the predictors to approximate normality before fitting a linear model. If one is willing to think of average effects as targets of inference, then such transformations should reduce bias in estimating these effects via a model without interaction terms. Of course transformations applied to predictor and response variables are an important part of regression modelling in practice, and the desirability of transforming a response variable to approximate normality is clear. Transformations on predictors, however, are typically

argued for on the basis of compatibility with linearity and homoscedasticity assumptions, without regard for the resulting shape of the predictor distribution. Result 3 suggests that the shape of the distribution could also be a consideration in assessing possible transformations of the predictors.

However, we also find that the consistency does not hold generally, with the following two examples.

First, assume X_1 follows standard normal and $X_2|X_1$ follows $Poisson(c_1|X_1|)$. Since when $p = 2$, transformation of (2.16) gives the quantity $\beta_{12}^{-1}(\delta_1^*(\boldsymbol{\beta}) - \delta_1(\boldsymbol{\beta}))$, nothing to do with true values of $\boldsymbol{\beta}$ and only depending on the distribution of $\mathbf{X}$. Moreover, it can also somehow indicate the discrepancy of the two large sample limits. Note that if $\beta_{12} = \beta_1 = \beta_2$, under the setting of Result 3, this quantity is just the *relative bias*, i.e., $\left|\delta_1^{-1}(\boldsymbol{\beta})\left[\delta_1^*(\boldsymbol{\beta}) - \delta_1(\boldsymbol{\beta})\right]\right|$. Based on the property of the mean and variance of *Poisson* distribution, by some algebra we can derive that

$$\mathrm{Var}\,(X_2) = c_1^2\left(1 - \frac{2}{\pi}\right) + c_1\sqrt{\frac{2}{\pi}}.$$

Solve c_1 by setting the above to be 1. Note the fact that $\rho = \mathrm{E}\,(X_1X_2) - \mathrm{E}\,(X_1)\mathrm{E}\,(X_2) = c_1\mathrm{E}\,(X_1|X_1|) = 0$. This is caused by the fact that $X_1|X_1|$ is an odd function and the integral interval is symmetric about zero. Hence, based on (2.16), we have

$$\begin{aligned}
\frac{\delta_1^*(\boldsymbol{\beta}) - \delta_1(\boldsymbol{\beta})}{\beta_{12}} &= \mathrm{E}\,\{\widetilde{X}_1^2\widetilde{X}_2\} \\
&= c_1\mathrm{E}\,|X_1^3| - c_1\sqrt{\frac{2}{\pi}} \\
&= c_1\sqrt{\frac{2}{\pi}} \\
&\approx 0.718.
\end{aligned}$$

The last equality follows from

$$\begin{aligned} \mathrm{E}\left(|X_1^3|\right) &= 2\frac{1}{\sqrt{2\pi}}\int_0^\infty x^3 e^{-x^2/2}dx \\ &\overset{y=x^2}{=} \frac{1}{\sqrt{2\pi}}\int_0^\infty y e^{-y/2}dy = 2\sqrt{\frac{2}{\pi}}. \end{aligned}$$

Second, suppose equi-correlated predictors, each following log-normal distribution, defined as

$$X_j = \frac{e^{\tau Z_i} - \mu(\tau)}{V^{1/2}(\tau)},\ j = 1, \ldots, p,$$

where

$$\begin{aligned} Z &= (Z_1, \ldots, Z_p)' \sim N\left(\mathbf{0}, (1-\rho)\mathbf{I}_p + \rho\mathbf{J}_p\right), \\ \mu(\tau) &= \mathrm{E}\left(e^{\tau Z_j}\right) = e^{\tau^2/2}, \\ V(\tau) &= \mathrm{Var}\left(e^{\tau Z_j}\right) = e^{2\tau^2} - e^{\tau^2}. \end{aligned}$$

Here $\mathbf{I}_p$ is p-dimensional identity matrix and $\mathbf{J}_p$ is p-dimensional square matrix with all p^2 elements equal to 1. Note that as $\sigma \to 0$, X_j converges to a standard normal. So that larger τ corresponds to more non-normality and larger ρ corresponds to more dependence among predictors. Partition the (true) regression coefficients as $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_M', \boldsymbol{\beta}_I')'$ according to intercept, main, and interaction effects respectively. Using Result 3 we can compare the vector of true average effects $\delta(\boldsymbol{\beta}) = \boldsymbol{\beta}_M$ (since $\mathrm{E}(\mathbf{X}) = \mathbf{0}$) to the large-sample limit of estimated effects from a model without interactions, i.e., $\delta^*(\boldsymbol{\beta})$. From Result 3 we know the relative bias is zero if $\rho = 0$ or $\tau = 0$, so the question of interest is how fast the bias grows as the components of X becomes both correlated and skewed. For $p = 10$ and selected values of $(\boldsymbol{\beta}_M, \boldsymbol{\beta}_I)$, Figure 2.1 depicts the relative bias as ρ and

τ vary, where

$$RB_j = \left|\{\delta_j(\boldsymbol{\beta})\}^{-1}\{\delta_j^*(\boldsymbol{\beta}) - \delta_j(\boldsymbol{\beta})\}\right|.$$

This illustration is based on fixing the direction of $\boldsymbol{\beta}_M$ and the relative length of $\boldsymbol{\beta}_I$ compared to $\boldsymbol{\beta}_M$, i.e., $\boldsymbol{\beta}_M \propto \mathbf{1}_p$ and $\|\boldsymbol{\beta}_I\| = \|\boldsymbol{\beta}_M\|$. For convenience, we index the components of $\boldsymbol{\beta}_I$ by (u, v) pair, where $u < v$. In Figure 2.1, four choices for the direction of $\boldsymbol{\beta}_I$ are considered:

(a) $\beta_{I,uv} \propto 1$, which involves all predictor pairs interacting positively.

(b) $\beta_{I,uv} \propto (-1)^{v-u+1}$, which means about half the pairs interact positively and the other half negatively.

(c) $\beta_{I,uv} \propto I\{|v \pmod p - u| = 1\}$, which means only a few of positive interactions.

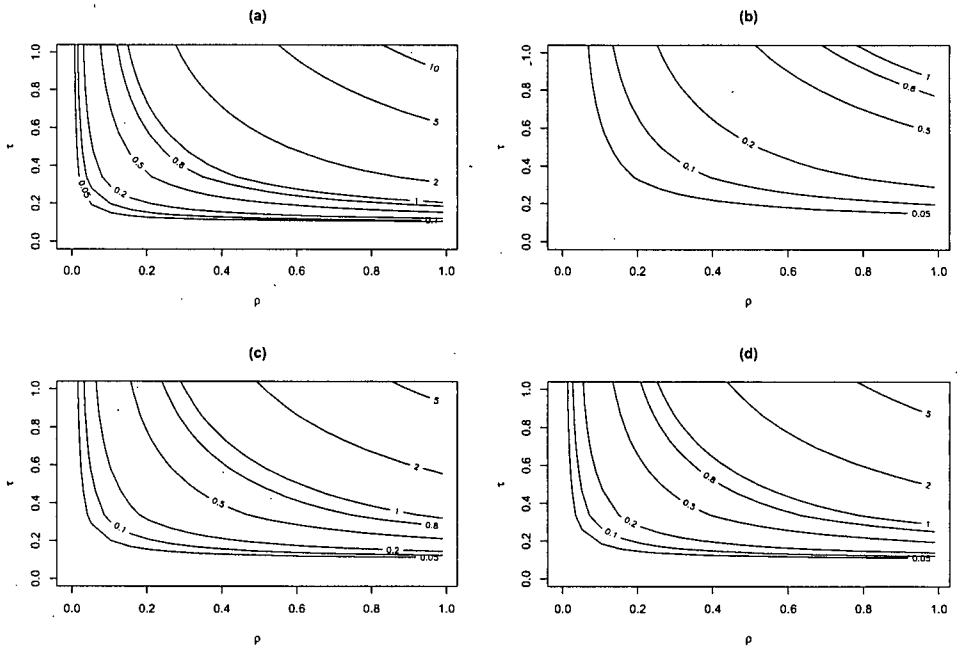
(d) $\beta_{I,uv} \propto (-1)^{I\{|v \pmod p - u| = 1\}}$, which means a majority (minority) of positive (negative) interactions.

In each case RB_1, the relative bias in estimating the average effect of X_1, is calculated. Note that each of these choices except (b) involves sufficient symmetry so that the relative bias is the same for estimating all p average effects, i.e., $RB_1 = \ldots = RB_p$. We can see that the bias increase when σ grows, which means that more non-normality would cause larger bias. Moreover, the size of bias also changes in an increasing trend as ρ increases. Therefore, more non-normality of and more dependence among $\mathbf{X}$ give larger bias. The general impression from Figure 1 is that very large biases are possible when estimating average effects, if predictors are substantially skewed and dependent. Note also that slight skewness and strong correlation tends to induce a bigger bias than strong skewness and slight correlation.

Another thing worth notice is that the relative biases are considerably smaller in panel (b) compared to the other three cases. It may represent a 'cancellation effect'

between positive and negative interaction terms. Moreover, we have similar magnitudes of relative biases under the other three cases. The common property of the three cases is the "overwhelming" strength of interaction in one direction over the other.

Figure 2.1: Magnitude of relative bias as a function of (ρ, τ) for multivariate log-normal distribution of $p = 10$ predictors. The cases (a) through (d) are as described in the text.



2.3.2 Relative efficiency of the two average effect estimators

A natural question of interest is which one is more efficient between the two estimators $\widehat{\delta_j(\boldsymbol{\beta})}$ and $\widehat{\delta_j^*(\boldsymbol{\beta})}$, which means to compare $v_j(\boldsymbol{\beta})$ in Result 1 and $v_j^*(\boldsymbol{\beta})$ in Result 2. We take the ratio of the latter to the former as the relative efficiency, with values larger than one representing the inefficiency occurring as a result of model misspecifications.

We apply Result 1 and 2 directly to linear regression context, where an additive model is fitted to the data generated by pairwise interaction model. Due to the symmetry in

the p predictors under both models, we can only take the average effect estimator of X_1 for example. Now we have

$$\begin{aligned}
\mathbf{T} &= (1, X_1, \ldots, X_p, X_1X_2, \ldots, X_{p-1}X_p)', \\
\widetilde{\mathbf{T}}_1 &= (0, 1, 0, \ldots, 0, X_2, \ldots, X_p, 0, \ldots, 0)' \\
\mathbf{S} &= (1, X_1, X_2, \ldots, X_p)' \\
\widetilde{\mathbf{S}}_1 &= (0, 1, 0, \ldots, 0)'.
\end{aligned}$$

With Result 1 and 2, we have

$$\begin{aligned}
n\mathrm{Var}\{\hat{\delta}_1^*\} \rightarrow\ & \sigma^2\{\mathrm{E}(\widetilde{\mathbf{S}}_1)\}'\Sigma_{\mathbf{S}}^{-1}\{\mathrm{E}(\widetilde{\mathbf{S}}_1)\} + \\
& \{\mathrm{E}(\widetilde{\mathbf{S}}_1)\}'\Sigma_{\mathbf{S}}^{-1}\mathrm{E}\left\{(\mathbf{T}'\boldsymbol{\beta} - \mathbf{S}'\boldsymbol{\alpha}_*)^2\mathbf{S}\mathbf{S}'\right\}\Sigma_{\mathbf{S}}^{-1}\{\mathrm{E}(\widetilde{\mathbf{S}}_1)\}, \qquad (2.17) \\
n\mathrm{Var}\{\hat{\delta}_1\} \rightarrow\ & \sigma^2\{\mathrm{E}(\widetilde{\mathbf{T}}_1)\}'\Sigma_{\mathbf{T}}^{-1}\{\mathrm{E}(\widetilde{\mathbf{T}}_1)\} + \\
& (\beta_{12}, \ldots, \beta_{1p})\mathrm{Var}(X_2, \ldots, X_p)(\beta_{12}, \ldots, \beta_{1p})', \qquad (2.18)
\end{aligned}$$

where

$$\begin{aligned}
\Sigma_{\mathbf{S}} &= \mathrm{E}\{(1, X_1, \ldots, X_p)'(1, X_1, \ldots, X_p)\}, \\
\Sigma_{\mathbf{T}} &= \mathrm{E}\{(1, X_1, \ldots, X_p, X_1X_2, \ldots, X_1X_p)' \\
&\qquad (1, X_1, \ldots, X_p, X_1X_2, \ldots, X_1X_p)\}.
\end{aligned}$$

Based on the expressions it is clear that the difference of the two unconditional variances depends on the true value of parameters $\boldsymbol{\beta}$. Therefore generally speaking, the comparison of the two variances could depend on the true values of coefficients of those interaction terms involving X_1.

In particular, we are interested in situations where the additive model can yield con-

sistent estimator under the true relationship involving all pairs of interactions. More precisely, if $\mathbf{X}$ has independence of components or multivariate normal distribution (discussed in Result 3), what is the relative efficiency?

We take the former case first. Assuming $\mathrm{E}(X_j) = 0, \mathrm{Var}(X_j) = 1, j = 1, \ldots, p$, and the independence of $X_1, \ldots, X_p$, therefore we have

$$\begin{aligned} \Sigma_{\mathbf{S}} &= \mathbf{I}_{p+1}, \\ \Sigma_{\mathbf{T}} &= \begin{pmatrix} \Sigma_{\mathbf{S}} & \mathbf{0} \\ \mathbf{0} & \Sigma_{22} \end{pmatrix}, \end{aligned}$$

where Σ_{22} is the covariance matrix of $(X_1X_2, \ldots, X_1X_p)'$.

Thus the asymptotic variance of the estimates from the right model is

$$\sigma^2(0, 1, 0, \ldots, 0)\Sigma_{\mathbf{S}}^{-1}(0, 1, 0, \ldots, 0)' + \sum_{j=2}^{p} \beta_{1j}^2.$$

The first term is right the first term in (2.17). Note the fact that

$$\mathbf{T}'\boldsymbol{\beta} - \mathbf{S}'\boldsymbol{\alpha}_* = \sum_{1 \le i < j \le p} \beta_{ij} X_i X_j,$$

where

$$\begin{aligned} \boldsymbol{\alpha}_* &= \Sigma_{\mathbf{S}}^{-1}\mathrm{E}(\mathbf{S}\mathbf{T}') \\ &= (\mathbf{I}_{p+1}, \mathbf{0})\boldsymbol{\beta}. \end{aligned}$$

Therefore, the second term in (2.17) is

$$\begin{aligned} & \{\mathrm{E}(\widetilde{\mathbf{S}}_1)\}'\Sigma_{\mathbf{S}}^{-1}\mathrm{E}\left\{\left(\sum_{i<j}\beta_{ij}X_iX_j\right)^2\mathbf{S}\mathbf{S}'\right\}\Sigma_{\mathbf{S}}^{-1}\{\mathrm{E}(\widetilde{\mathbf{S}}_1)\} \\ = \; & \mathrm{E}\left\{(\sum_{i<j}\beta_{ij}X_iX_j)^2X_1^2\right\} \end{aligned}$$

Hence, the ratio of the two asymptotic variances is

$$\begin{aligned} \frac{v_1^*}{v_1} &= \frac{\sigma^2+\mathrm{E}\left\{(\sum_{i<j}\beta_{ij}X_iX_j)^2X_1^2\right\}}{\sigma^2+\sum_{j=2}^p\beta_{1j}^2} \\ &= \frac{1+3\|\gamma_{1\cdot}\|^2+\|\gamma_{-1\cdot}\|^2}{1+\|\gamma_{1\cdot}\|^2}, \end{aligned}$$

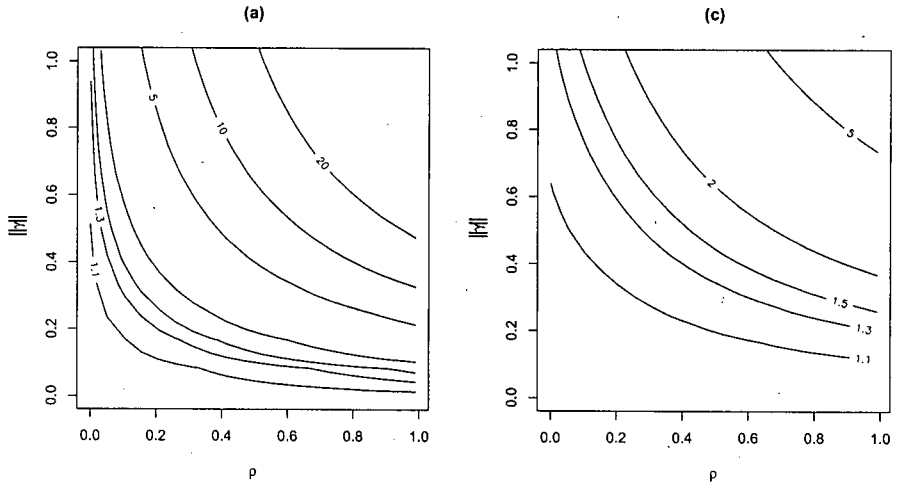
where $\gamma=\sigma^{-1}\boldsymbol{\beta}_I$ and $(\gamma_{1\cdot},\gamma_{-1\cdot})$ is a partition of γ into those interaction terms which do and don't involve X_1 respectively. Therefore, the asymptotic variance of the estimator based on "misspecified" model is larger. This is consistent with our intuition that more uncertainty in $\widehat{\delta_j^*(\boldsymbol{\beta})}$ is caused by model misspecification.

Now the question of interest is how large/small the relative efficiency would be if there is some dependence among the components of $\mathbf{X}$? Assuming that $\mathbf{X}$ has a multivariate normal distribution with mean $\mathbf{0}$ and equi-correlated covariance matrix, that is, $\mathbf{X}\sim N_p(\mathbf{0},(1-\rho)\mathbf{I}_p+\rho\mathbf{J}_p)$. From (2.18) and (2.17) for given p, it is easy to justify that the relative efficiency depends only on ρ and $\gamma=\sigma^{-1}\boldsymbol{\beta}_I$, with the latter one describing the interaction 'signal' relative to noise.

For p =10, Figure 2.2 shows the relative efficiency as a function of ρ and $\|\gamma\|$, where two certain directions for $\boldsymbol{\beta}_I$, i.e., case (a) and (c) defined in Section 2.3.1, are considered. We can see that if the true interactions among $\mathbf{X}$ are rather "sparse", like the choice of $\boldsymbol{\beta}_I$ in case (c), the efficiency loss is much smaller compared to that with a "dense" interaction structure in panel (a). Also note that the misspecified-model estimator can

be very inefficient with strongly correlated predictors, but the efficiency loss tends to be slight with independent predictors. As a related point, the rate at which the efficiency loss grows with the strength of the underlying interaction signal is governed by the strength of correlation.

Figure 2.2: Relative efficiency of the "misspecified"-model average-effect estimator as a function of ρ and $\|\gamma\|$, for an equi-correlated multivariate normal distribution of $p = 10$ predictors. The cases (a) and (c) are as described in the text.



2.4 Nonparametric regression and smoothing

The above analyses in Section 2.3 are based on a simple scenario, where a parametric model in terms of a linearity of response variable Y in continuous predictors X_i's is postulated. What if the linear regression model is not appropriate? Fitting a linear model to the data actually containing a nonlinear structure can given very misleading results, even worse than useless. A more general alternative to linear regression is nonparametric regression model. The distinguishing property of nonparametric regression is that there is no (or very little) a priori knowledge about the form of the true structure of the regression

function. It allows the class of functions which the model can represent to be very broad.

When shall we use nonparametric regression model? In many real problems, there is no information from the data nor scientific knowledge to suggest a parametric form, so that a parametric model is often specified from a casual graphical summary of the data or chosen for convenience (for example linear regression models). A predetermined parametric model might be too restricted or too low-dimensional to fully model a "rich" data set containing many unexpected features. In such a case, we would like the nonparametric (smoothing) approach, which offers a flexible tool in analyzing unknown regression relationships between response variable and predictors. Also parametric vs non-parametric depends a lot on sample size, especially when there are many predictors.

Smoothing methods, widely used in nonparametric modelling, deserve a respectable place in statistics. There are many papers and a number of books study on this topic (Silverman 1986; Eubank, 1988; Hastie and Tibshirani, 1990; Wahba, 1990; Green and Silverman, 1994; Gu, 2002). As a matter of fact, smoothing methods provide a bridge/compromise between making no assumptions on the underlying process that generated the data (a purely nonparametric approach) and making very strong assumptions (a parametric approach).

In the following subsections, we mainly focus on consequences of model misspecification omitting interactions under the context of least squares regression in smoothing.

2.4.1 Spline regression models

Spline functions are very flexible and thus are often used in smoothing regression. A spline function is a piecewise or segmented polynomial. More precisely, it is defined as below.

Definition: The function ϕ is a *spline* on $[a, b]$ of degree D with *knots* $t_1, \ldots, t_L$ ($a < t_1 < \cdots < t_L < b$) if ϕ is a polynomial of degree D on the subintervals $[a, t_1], [t_1, t_2], \ldots, [t_L, b]$ and ϕ has $D - 1$ continuous derivatives on $[a, b]$. Denote the collection of these splines by $S_p(t_1, \ldots, t_L; D)$. Take $D = 1$ for example. A spline of degree 1 is continuous and piecewise linear, with breaks in slope occurring at $t_1, \cdots, t_L$.

Based on the definition of spline, it is not hard to show that the collection of the splines of degree D is a linear space of functions. Then we can talk about its dimension and construct bases for the space. The dimension is the number of parameters needed to describe a member of the space, which is $D + L + 1$(or $D + L$ if forcing the spline space to not include a constant term). The number of functions in a basis will simply be equal to the dimension.

For simplicity and ease-of-understanding, we introduce the *power basis*. Note that for other forms of bases, the analysis discussed later is also applicable but with more difficult computational problems.

Power Basis: Denote the power basis by $\phi_1, \ldots, \phi_{L+D+1}$:

$$\begin{aligned}
\phi_1(x) &= 1, \\
\phi_2(x) &= x, \\
&\cdots \\
\phi_{D+1}(x) &= x^D, \\
\phi_{D+2}(x) &= (x - t_1)_+^D, \\
&\cdots \\
\phi_{D+L+1}(x) &= (x - t_L)_+^D,
\end{aligned}$$

where

$$(x-t)_+^D = \begin{cases} (x-t)^D & \text{for x} \geq t, \\ 0 & \text{for x} < t. \end{cases}$$

It is easy to show that each ϕ_j above is a spline and all the ϕ_j's are linearly independent. Therefore for any function h in the spline space with degree D, it can be written as $h(x) = \sum_{j=1}^{L+D+1} \beta_j \phi_j(x)$.

Thus what happens if the fitted model is misspecified? Note that the main concern now is about the impact of model misspecifications. Therefore, how to choose the degree of spline (D), number of knots (L) and locations of these knots $(t_1, \ldots, t_L)$ is left aside for the time being, although this is always concerned in smoothing. That is, we assume they are already appropriately chosen in the following studies.

We start with a regression model having only two predictors. Say the fitted model is

$$(Y|X_1 = x_1, X_2 = x_2) \sim N\left(m_1(x_1) + m_2(x_2), \sigma^2\right),$$

while the "true" model is

$$(Y|X_1 = x_1, X_2 = x_2) \sim N\left(g_1(x_1) + g_2(x_2) + g_{12}(x_1, x_2), \tau^2\right),$$

where m_1, m_2, g_1, g_2 are splines. Here g_{12} accounts for the interactions between the two predictors. In general, there are many plausible possibilities for the form of g_{12}. For simplicity and interpretability, we use the form of $g_{12}(X_1, X_2) = X_2 t_1(X_1) + X_1 t_2(X_2)$ in the following, where t_1, t_2 are splines.

Note that generally the basis functions for different predictors could be different. For concise notations without loss of clarity, we suppress the subscript of Φ since each of its

components is expressed as as a function of X_i. Suppose

$$\begin{aligned} m_i &= \mathbf{\Phi}'(X_i)\boldsymbol{\alpha}_i, \\ g_i &= \mathbf{\Phi}'(X_i)\boldsymbol{\beta}_i, \\ t_i &= \mathbf{\Phi}'(X_i)\boldsymbol{\beta}_i^{int}, i = 1, 2. \end{aligned}$$

Similar to the linear regression case, without loss of generality we assume that for $i = 1, \ldots, p$, $\mathrm{E}(X_i) = 0, \mathrm{E}[\mathbf{\Phi}(X_i)] = \mathbf{0}$, where $\mathbf{\Phi}(X_i) = (\phi_1(X_1), \ldots, \phi_{K_i}(X_i))'$ (K_i denotes the number of basis functions for X_i). Otherwise, we can replace X_i (or $\mathbf{\Phi}(X_i)$) by $X_i - \mathrm{E}(X_i)$ (or $\mathbf{\Phi}(X_i) - \mathrm{E}[\mathbf{\Phi}(X_i)]$), and those centering constants would be included into the intercept, keeping the coefficients of basis functions unchanged. Thus the mean function of the fitted model can be rewritten as

$$\mathrm{E}(Y|X_1, X_2) = \alpha_0 + \left(\mathbf{\Phi}'(X_1), \mathbf{\Phi}'(X_2)\right) \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \end{pmatrix},$$

where α_0 denotes the intercept.

Let $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2)$. Based on White (1982), we have $\boldsymbol{\alpha}_*$, the large-sample limit of the $\widehat{\boldsymbol{\alpha}}$, as the solution to

$$\mathrm{E}\left(\frac{\partial \log f(Y|X_1, X_2)}{\partial \boldsymbol{\alpha}}\right) = 0,$$

where $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1', \boldsymbol{\alpha}_2')'$.

Note that the f function in above equality denotes the density function of the fitted model and the expectation operation is with respect to the true joint distribution of Y and $\mathbf{X}$.

Therefore, we derive that

$$\mathrm{E}\left\{\begin{pmatrix}1\\ \boldsymbol{\Phi}(X_1)\\ \boldsymbol{\Phi}(X_2)\end{pmatrix}(1,\boldsymbol{\Phi}'(X_1),\boldsymbol{\Phi}'(X_2))\right\}\begin{pmatrix}\alpha_0\\ \boldsymbol{\alpha}_1\\ \boldsymbol{\alpha}_2\end{pmatrix}=\mathrm{E}\left\{Y\begin{pmatrix}1\\ \boldsymbol{\Phi}(X_1)\\ \boldsymbol{\Phi}(X_2)\end{pmatrix}\right\}.$$

Note the fact that

$$\mathrm{E}(Y|\mathbf{X})=\beta_0+\left(\boldsymbol{\Phi}'(X_1),\boldsymbol{\Phi}'(X_2),X_2\boldsymbol{\Phi}'(X_1),X_1\boldsymbol{\Phi}'(X_2)\right)\begin{pmatrix}\boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\\ \boldsymbol{\beta}_1^{int}\\ \boldsymbol{\beta}_2^{int}\end{pmatrix}. \tag{2.19}$$

Therefore we can link the above two equalities to derive that

$$\begin{pmatrix}\alpha_{0*}\\ \boldsymbol{\alpha}_{1*}\\ \boldsymbol{\alpha}_{2*}\end{pmatrix}=\begin{pmatrix}\beta_0\\ \boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}+C_{11}^{-1}C_{12}\begin{pmatrix}\boldsymbol{\beta}_1^{int}\\ \boldsymbol{\beta}_2^{int}\end{pmatrix}, \tag{2.20}$$

where

$$\begin{aligned}
C_{11} &= \mathrm{E}\left\{\left(1,\boldsymbol{\Phi}'(X_1),\boldsymbol{\Phi}'(X_2)\right)'\left(1,\boldsymbol{\Phi}'(X_1),\boldsymbol{\Phi}'(X_2)\right)\right\},\\
C_{12} &= \mathrm{E}\left\{\left(1,\boldsymbol{\Phi}'(X_1),\boldsymbol{\Phi}'(X_2)\right)'\left(X_2\boldsymbol{\Phi}'(X_1),X_1\boldsymbol{\Phi}'(X_2)\right)\right\},\\
C_{22} &= \mathrm{E}\left\{\left(X_2\boldsymbol{\Phi}'(X_1),X_1\boldsymbol{\Phi}'(X_2)\right)'\left(X_2\boldsymbol{\Phi}'(X_1),X_1\boldsymbol{\Phi}'(X_2)\right)\right\}.
\end{aligned}$$

We also use the idea of average effect as that in the linear scenario to evaluate the impact of model misspecifications. That is, we are interested in the average effect by averaging over the joint distribution of all predictors $\mathbf{X}$. Note that the predictive effect

of X_j here is function of all predictors while just a function of $\mathbf{X}_{(j)}$ in linear regression settings.

Without loss of generality, we only take the average effect of X_1 for example. Here the large sample limit of average effect of X_1 estimated from the fitted model is

$$\mathrm{E}\left(\boldsymbol{\alpha}_{1*}^{'}\frac{d\boldsymbol{\Phi}(X_1)}{dX_1}\right), \tag{2.21}$$

while that in the "true" model is

$$\mathrm{E}\left(\boldsymbol{\beta}_1^{'}\frac{d\boldsymbol{\Phi}(X_1)}{dX_1}+\boldsymbol{\beta}_1^{int'}X_2\frac{d\boldsymbol{\Phi}(X_1)}{dX_1}+\boldsymbol{\beta}_2^{int'}\boldsymbol{\Phi}(X_2)\right). \tag{2.22}$$

Similar to the situations considered in linear regression, we also focus on some special cases here. Say, if X_1 and X_2 are independent of each other, then we get $\mathrm{E}(C_{12})=0$. Hence, the equality (2.20) now becomes

$$\begin{pmatrix}\alpha_{0*}\\ \boldsymbol{\alpha}_{1*}\\ \boldsymbol{\alpha}_{2*}\end{pmatrix}=\begin{pmatrix}\beta_0\\ \boldsymbol{\beta}_1\\ \boldsymbol{\beta}_2\end{pmatrix}.$$

That is, if X_1 and X_2 are independent, we get the consistency of the coefficient estimates $\widehat{\boldsymbol{\alpha}}$. Subsequently we also get the agreement between the two average effects defined by (2.21) and (2.22).

In the following we study the uncertainties of the two average effect estimates. Let $\boldsymbol{\Phi}(\mathbb{X}_l)$ be an n by K_l matrix with $\Phi_{ij}=\phi_j(x_{li})$ and $\mathbf{x}_l$ be an n by 1 vector with components x_{li}, $l=1,2, i=1,\ldots,n, j=1,\ldots,K_l$. Define the sample version of $\Sigma_{\mathbf{T}}(=\mathrm{E}(\mathbf{T}\mathbf{T}'))$ as

follows.

$$
\begin{aligned}
D_{\mathbf{T}} &= \begin{pmatrix} D_{11} & D_{12} \\ D_{12}' & D_{22} \end{pmatrix}, \\
D_{11} &= \begin{pmatrix} 1 \\ \mathbf{\Phi}'(\mathbb{X}_1) \\ \mathbf{\Phi}'(\mathbb{X}_2) \end{pmatrix} \big(1, \mathbf{\Phi}(\mathbb{X}_1), \mathbf{\Phi}(\mathbb{X}_2)\big), \\
D_{12} &= \begin{pmatrix} 1 \\ \mathbf{\Phi}'(\mathbb{X}_1) \\ \mathbf{\Phi}'(\mathbb{X}_2) \end{pmatrix} \big(\mathbb{X}_2\mathbf{\Phi}(\mathbb{X}_1), \mathbb{X}_1\mathbf{\Phi}(\mathbb{X}_2)\big), \\
D_{22} &= \begin{pmatrix} \mathbb{X}_2\mathbf{\Phi}'(\mathbb{X}_1) \\ \mathbb{X}_1\mathbf{\Phi}'(\mathbb{X}_2) \end{pmatrix} \big(\mathbb{X}_2\mathbf{\Phi}(\mathbb{X}_1), \mathbb{X}_1\mathbf{\Phi}(\mathbb{X}_2)\big).
\end{aligned}
$$

Again similar to the linear regression case, we have

$$
\begin{pmatrix} \hat{a}_0 \\ \hat{\alpha}_1 \\ \hat{\alpha}_2 \end{pmatrix} = \begin{pmatrix} \hat{b}_0 \\ \widehat{\boldsymbol{\beta}}_1 \\ \widehat{\boldsymbol{\beta}}_2 \end{pmatrix} + D_{11}^{-1} D_{12} \begin{pmatrix} \widehat{\boldsymbol{\beta}}_1^{int} \\ \widehat{\boldsymbol{\beta}}_2^{int} \end{pmatrix}.
$$

Hence the estimate of the fitted average effect can be rewritten as

$$
\widehat{\delta_1^*(\boldsymbol{\beta})} = \widehat{[\mathrm{d}\,\mathbf{\Phi}]}'\widehat{\boldsymbol{\beta}}_1 + \left(\widehat{[\mathrm{d}\,\mathbf{\Phi}]}' D_{11}^*\right) \widehat{\boldsymbol{\beta}}_1^{int} + \left(\widehat{[\mathrm{d}\,\mathbf{\Phi}]}' D_{12}^*\right) \widehat{\boldsymbol{\beta}}_2^{int},
$$

where

$$\begin{aligned}\widehat{\mathrm{d}\,\Phi} &= n^{-1}\sum_i \frac{\mathrm{d}\,\Phi(x)}{\mathrm{d}\,x}\Big|_{x=x_{1i}},\\ D^* &= \left(D_{11}^{-1}D_{12}\right)[-1,]\\ &= \begin{pmatrix} D_{11}^* & D_{12}^* \\ D_{21}^* & D_{22}^* \end{pmatrix}.\end{aligned}$$

Notationally, $\#[-1,]$ denotes a sub-matrix of $\#$ after deleting the first row. The estimate of the average effect of X_1 based on the "true" model is

$$\widehat{\delta_1(\boldsymbol{\beta})} = [\widehat{\mathrm{d}\,\Phi}]'\widehat{\boldsymbol{\beta}}_1 + \left[n^{-1}\sum_i x_{2i}\frac{\mathrm{d}\,\Phi(x)}{\mathrm{d}\,x}\Big|_{x=x_{1i}}\right]'\widehat{\boldsymbol{\beta}}_1 + \left[n^{-1}\sum_i \Phi(x_{2i})\right]'\widehat{\boldsymbol{\beta}}_2.$$

With Result 1 and 2, it is possible to compare the asymptotic variances of the two estimates. However, the calculation of (2.9) now is rather complicated due to $\mathrm{E}\,((\mathbf{T}'\boldsymbol{\beta} - \mathbf{S}'\boldsymbol{\alpha}_*)^2\mathbf{S}\mathbf{S}')$ and the last two non-vanishing terms in (2.9).

Thus we compare the asymptotic *conditional* variances in the following. Recall that $\mathbf{T}$ is the vector of the predictor functions in the true relationship and $\mathbf{S}$ is the vector the predictor functions in the fitted model. Recall that $\widetilde{\mathbf{T}}_1, \widetilde{\mathbf{S}}_1$ denote the derivative with respect to X_1 of $\mathbf{T}$ and $\mathbf{S}$, respectively. Using essentially the same derivation as that in the previous subsection of linear regression scenario, we get that as $n \to \infty$,

$$\begin{aligned} n\,\mathrm{Var}\left(\hat{\delta}_1(\boldsymbol{\alpha})|\mathbf{X}\right) &\to \sigma^2[\mathrm{E}\,(\widetilde{\mathbf{S}}_1)]'\left\{\mathrm{E}\,(\mathbf{S}\mathbf{S}')\right\}^{-1}\mathrm{E}\,(\widetilde{\mathbf{S}}_1)\\ n\,\mathrm{Var}\left(\hat{\delta}_1(\boldsymbol{\beta})|\mathbf{X}\right) &\to \sigma^2[\mathrm{E}\,(\widetilde{\mathbf{T}}_1)]'\left\{\mathrm{E}\,(\mathbf{T}\mathbf{T}')\right\}^{-1}\mathrm{E}\,(\widetilde{\mathbf{T}}_1),\end{aligned}$$

where

$$\begin{aligned}
\mathbf{T} &= (1, \mathbf{\Phi}'(X_1), \mathbf{\Phi}'(X_2), X_1\mathbf{\Phi}'(X_2), X_2\mathbf{\Phi}'(X_1))', \\
\widetilde{\mathbf{T}}_1 &= \left\{0, \left[(\frac{\mathrm{d}\,\mathbf{\Phi}(X_1)}{\mathrm{d}\,X_1})\right]', \mathbf{0}_{1\times K_2}, [(\mathbf{\Phi}(X_2))]', \left[\mathrm{E}\,(X_2\frac{\mathrm{d}\,\mathbf{\Phi}}{\mathrm{d}\,X_1})\right]'\right\}', \\
\mathbf{S} &= (1, \mathbf{\Phi}'(X_1), \mathbf{\Phi}'(X_2))', \\
\widetilde{\mathbf{S}}_1 &= (0, \frac{\partial}{\partial X_1}\mathbf{\Phi}'(X_1), \mathbf{0}_{1\times K_2}).
\end{aligned}$$

As we discussed before, in the linear regression scenario, the joint normality of the predictors can also yield the consistent estimates of average effects. Will this good property still hold now? To explore the effect of the correlation between predictors on the difference of the two average effects, we set up predictors and basis functions as follows.

A1. Suppose X_1 and X_2 are bivariate normal with mean vector 0 and covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

A2. For simplicity, a common set of quadratic power basis functions for two predictors is used:

$$\begin{aligned}
\phi_1(x) &= x - k_1, \\
\phi_2(x) &= x^2 - k_2, \\
\phi_3(x) &= (x - t_1)^2 I\{x > t_1\} - k_3, \\
\phi_4(x) &= (x - t_2)^2 I\{x > t_2\} - k_4,
\end{aligned}$$

where t_1, t_2 are the knots, which are set to be the 25% and 75% percentile of standard normal, respectively. The k_i's are the centering constants to make $\mathrm{E}\,(\phi_i(X_1)) = 0, i = 1, \ldots, 4$. Although the choice of basis functions is simple, it mimics the generality of

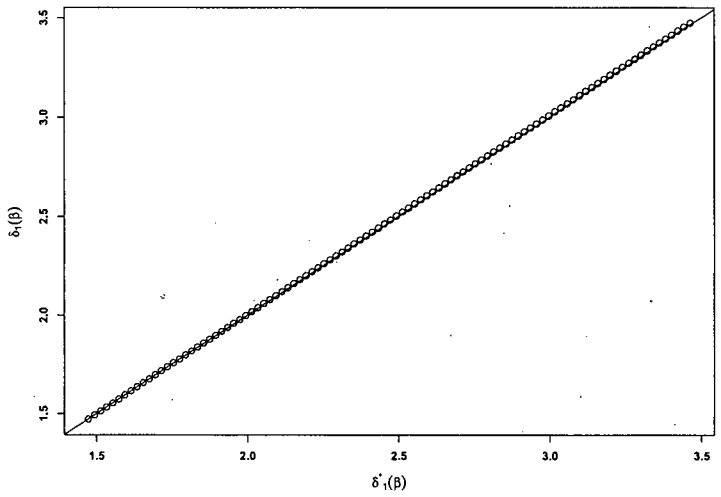
splines expressed as linear combinations of the power basis functions. As for the number of knots, we can also easily incorporate more knots into the basis functions. Here for computational convenience, we just use two knots for demonstration.

A3. Set the true values of parameters as $\beta_0 = 0, \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 = \boldsymbol{\beta}_1^{int} = \boldsymbol{\beta}_2^{int} = (0.5, 0.5, 0.5, 0.5)'$.

Under the above settings, we can get the values of the elements of matrices C_{11}, C_{12} by a numerical approach. The details of the numerical approach for the required integrals are in Appendix III.

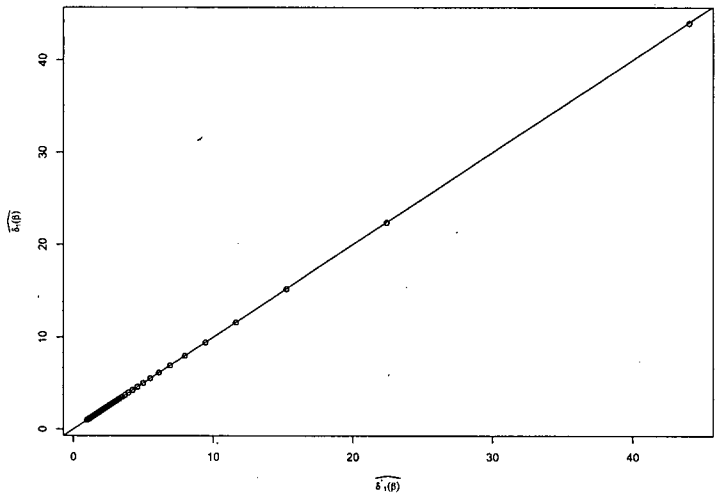
One fact worth noticing is that when $\rho = 1$, that is $X_1 = X_2$, C_{11} is not invertible. Therefore we can think of the difference between the two average effects as a function of ρ, the correlation coefficient, confined within $[0, 1)$. As the correlation increases, the discrepancy between the two average effects is quite small compared to the range of average effects. The graphical summarization is shown in Figure 2.3, which implies that the bias of the average effect is close to zero even if the correlation of the two predictors is high.

Figure 2.3: Comparison of two average effect estimators $\delta_1(\boldsymbol{\beta})$ and $\delta_1^*(\boldsymbol{\beta})$ for spline regression models with two predictors involved. The x-coordinate of each circle is $\delta_1^*(\boldsymbol{\beta})$ and the y-coordinate is $\delta_1(\boldsymbol{\beta})$. Different circles are produced by different values of ρ, the correlation coefficient between X_1 and X_2.



As shown in Figure 2.4, we can see that the difference of the two conditional variances is rather small compared to magnitude of the conditional variances. This means that even though the model is misspecified, the precision of the estimates for given values of predictors does not seem to be affected that much by the misspecification.

Figure 2.4: Comparison of asymptotic conditional variances of two average effect estimators for spline regression models with two predictors involved. The x-coordinate of each circle is the conditional variance of misspecified-model estimator and the y-coordinate is that of right-model estimator. Different circles are produced by different values of ρ, the correlation coefficient between X_1 and X_2.



2.4.2 Penalized regression models

Penalized spline regression (often referred to as *P-splines*) has received attention as a powerful smoothing method. Originally suggested by O'Sullivan (1986), the method provides a range of practical modelling tools in applied statistics, with the books by Green and Silverman (1994) and more recently by Ruppert, Wand, and Carroll (2003).

The main principle of penalized spline regression is to estimate the unknown regression function by a compromise between sum of squares of residuals (represent the fidelity to the data) and smoothness of the estimate.

We start with a univariate case. Suppose that we have data (x_i, y_i) (for now x_i is a

scalar not a vector),

$$Y_i = m(x_i; \boldsymbol{\alpha}) + \epsilon_i,$$

where m is a smooth function denoting the conditional mean of Y_i given x_i, and $\{\epsilon_i\}_{i=1}^n$ are independent, mean zero random errors with a constant variance. To estimate m we use a spline regression model

$$m(x; \boldsymbol{\alpha}) = \alpha_0 + \alpha_1 x + \cdots + \alpha_D x^D + \sum_{k=1}^{L} a_k (x - t_k)_+^D,$$

where $d \geq 1$ is an integer, $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \ldots, \alpha_D, a_1, \ldots, a_L)'$ is a vector of regression coefficients, $t_1 < \cdots < t_L$ are fixed knots and $(x - t_k)_+^D = (x - t_k)^D I\{x \geq t_k\}$. Actually, m is expressed by a set of power spline basis functions. Recall that $\mathbb{S}$ denotes the "design matrix" for the model so that the i-th row of $\mathbb{S}$ is

$$\begin{aligned} \mathbb{S}_{i\cdot} &= (1, x_i, \ldots, x_i^D, (x_i - t_1)_+^D, \ldots, (x_i - t_L)_+^D) \\ &= \left(1, \boldsymbol{\Phi}'(x_i)\right). \end{aligned}$$

However, simple parametric fitting of $\boldsymbol{\alpha}$ would lead to unsatisfactory results due to the high dimensionality of basis functions. Instead, $\boldsymbol{\alpha}$ is estimated in a penalized manner by imposing a penalty on the coefficients in m. A roughness penalty is placed on $\{a_k\}_{k=1}^L$ which is the set of jumps at the knots in the p-th derivative of $m(x;, \boldsymbol{\alpha})$. This leads to the penalized least-squares estimator

$$\begin{aligned} \widehat{\boldsymbol{\alpha}(\lambda)} &= \min_{\boldsymbol{\alpha}} \left\{ \sum_{i=1}^{n} \{y_i - m(x; \boldsymbol{\alpha})\}^2 + \lambda \sum_{k=1}^{L} a_k^2 \right\} \\ &= \min_{\boldsymbol{\alpha}} \left\{ (\mathbb{Y} - \mathbb{S}\boldsymbol{\alpha})'(\mathbb{Y} - \mathbb{S}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}' P \boldsymbol{\alpha} \right\}, \end{aligned}$$

with λ as penalty parameter controlling the trade-off between fidelity to the data and

smoothness of the fitted spline and P as the corresponding penalty matrix. To put the roughness penalty mentioned above, we choose P to be a diagonal matrix whose first $(1+D)$ diagonal elements are 0 and whose remaining diagonal elements are 1. By simple algebra, the penalized least squares estimate of $\boldsymbol{\alpha}$ is given by

$$\begin{aligned}\widehat{\boldsymbol{\alpha}(\lambda)} &= (\mathbb{S}'\mathbb{S}+\lambda P)^{-1}\mathbb{S}'\mathbb{Y}, \\ &= (n^{-1}\mathbb{S}'\mathbb{S}+n^{-1}\lambda P)^{-1}\{n^{-1}\mathbb{S}'\mathbb{Y}\}.\end{aligned}$$

Since $n^{-1}\lambda P$ goes to a zero matrix, the asymptotic behavior of $\widehat{\boldsymbol{\alpha}(\lambda)}$ is the same as that in the standard spline regression without penalty.

To work out the least squares estimates $\widehat{\boldsymbol{\alpha}(\lambda)}$, we need to determine an appropriate value of λ first. Generalized cross-validation (GCV) (Craven and Wahba, 1979) is one method of smoothing parameter selection that has proven effective and has good theoretical properties. Here we follow Ruppert (2002) closely. Let

$$ASR(\lambda) = n^{-1}\sum_{i=1}^{n}\left\{y_i - m(x_i;\widehat{\boldsymbol{\alpha}(\lambda)})\right\}^2$$

be the average squared residuals using λ. Let

$$\mathbf{S}(\lambda) = \mathbb{S}(\mathbb{S}'\mathbb{S}+\lambda P)^{-1}\mathbb{S}'$$

be the "smoother" or "hat" matrix. Then

$$GCV(\lambda) = \frac{ASR(\lambda)}{(1-n^{-1}\mathrm{tr}\{\mathbf{S}(\lambda)\})^2} \tag{2.23}$$

is the generalized cross validation statistic. Here $\mathrm{tr}\{\mathbf{S}(\lambda)\}$ is the "effective degrees of freedom" of the fit. One chooses λ minimizing GCV statistic over a grid of values of

λ. Computation can be sped up and stabilized numerically with the following diagonalization method that is variation on the Demmler-Reinsch algorithm used to compute smoothing splines.

Let B be a square matrix satisfying $B^{-1}(B^{-1})' = \mathbb{S}'\mathbb{S}$, for example, B^{-1} is a Cholesky factor of $\mathbb{S}'\mathbb{S}$. Let U be orthogonal and let C be diagonal such that $UCU' = BPB'$. For example, we can use the eigen-decomposition of BPB' to find U and C. Then by some algebra, we get

$$\mathrm{tr}\{\mathbf{S}(\lambda)\} = \sum_i (1 + \lambda C_i)^{-1}, \tag{2.24}$$

where C_i is the ith diagonal element of C. Details of proof is given in Appendix IV.

The elegance of this method is that the work of calculating B and (U, C) needs to be done only once and then these quantities can be used for all values of λ. Therefore, we have an efficient way to evaluate GCV defined by (2.23).

This method is easy to be extended to the additive model with more than one predictor, i.e., $p > 1$. Suppose we have data $(y_i, \mathbf{x}_i)$ $(\mathbf{x}_i = (x_{1i}, \ldots, x_{pi})')$,

$$Y_i = \alpha_0 + m_1(x_{1i}) + \cdots + m_p(x_{pi}) + \epsilon_i.$$

We will use a spline model for each m_k:

$$m_l(x; \boldsymbol{\alpha}_l) = \alpha_{l1}x + \cdots + \alpha_{lD}x^D + \sum_{k=1}^{L_l} a_{lk}(x - t_l)_+^D, \quad l = 1, \ldots, p,$$

where $\boldsymbol{\alpha}_l = (\alpha_{l1}, \ldots, \alpha_{lD}, a_{l1}, \ldots, a_{lL_l})'$. Hence the vector of whole parameters is $\boldsymbol{\alpha} = (\alpha_0, \boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_p')'$, and the penalty matrix P is set to be the sum of p diagonal matrices P_i $(i = 1, \ldots, p)$. Each P_i is a $(pK + 1)$ (K is the dimension of basis functions) by $pK + 1$ square matrix whose $(p \times i + 2)$th to $(p \times i + L + 1)$th diagonal elements are 1 and whose remaining diagonal elements are 0. Therefore the least squares estimate of $\boldsymbol{\alpha}$

is the minimizer of

$$\sum_{i=1}^{n}\{y_i - m(\mathbf{x}_i;\boldsymbol{\alpha})\}^2 + \sum_{l=1}^{p}\lambda_l\sum_{k=1}^{L_l}a_{lk}^2$$

$$\text{i.e.} \qquad (\mathbb{Y}-\mathbb{S}\boldsymbol{\alpha})'(\mathbb{Y}-\mathbb{S}\boldsymbol{\alpha}) + \boldsymbol{\alpha}'\left(\sum_{i=1}^{p}\lambda_i P_i\right)\boldsymbol{\alpha}.$$

Note $\mathbb{S}$ is the corresponding "design matrix" with i-th row as

$$\mathbb{S}_{i\cdot} = (1, \boldsymbol{\Phi}'(x_{1i}), \ldots, \boldsymbol{\Phi}'(x_{pi}))\,.$$

Consequently the least squares estimate of $\boldsymbol{\alpha}$ can be solved by

$$\widehat{\boldsymbol{\alpha}(\lambda)} = \left(\mathbb{S}'\mathbb{S} + \sum_{i=1}^{p}\lambda_i P_i\right)^{-1}\mathbb{S}'\mathbb{Y}.$$

If the components function $\{m_l\}_{l=1}^{p}$ require roughly the same amounts of smoothing, we may assume a common value of $\lambda_l, l = 1, \ldots, p$, which makes a quick access of the estimation of smoothing parameters by using the methodology introduced in the univariate P-spline case. More realistically, different component function m_l's require different amounts of smoothing and this can be accomplished by allowing λ_l values to vary rather than a common value. In the algorithm of Ruppert and Carroll (2000), $\lambda_1, \ldots, \lambda_p$ are chosen by GCV in two steps. Note that GCV is a function of $\lambda_1, \ldots, \lambda_p$. In the first step, GCV is minimized by assuming that $\lambda_1 = \cdots = \lambda_p = \lambda$. In the second step, set the common smoothing parameter as the starting value of each λ_i, $\lambda_1, \ldots, \lambda_p$ are selected one-at-a-time by minimizing the GCV criterion.

In the following simulation study, we see how the two average effect estimates based on penalized least squares would be affected by the smoothing parameter. Suppose we have two predictors involved in the model and both of them follows a uniform distribution

on [0,1] independently. Here a common set of centered cubic basis functions are used and hence a common smoothing parameter λ is assumed.

In the basis function, five knots are used, which are equi-spaced within [0,1]. Generate a data set of $n = 200$ realizations of (Y, X_1, X_2), where Y is generated from normal distribution with variance of 1 and mean is based on the model with interactions (2.19). The true values of all the coefficients are set to be 0.5. Repeat generating 200 data sets, and calculating the average effect estimate for each data set. We can vary the sample size n to see what happens as n gets larger. The top panel in Figure 2.5 shows the average effect estimates from the data sets of size $n = 200$ and the bottom panel shows that of size $n = 1000$. The solid horizontal line in both panel is the true value of the average effect. Therefore we can see that the estimates lie around the true value with a larger variability, which is caused by the larger variability of the smoothing parameter estimates. As sample size increases, the variability goes down, i.e., the estimates in the bottom panel are more concentrated on the true value than the top one. One problem from Figure 2.5 is that the range of the estimates is quite broad compared to the scale of the true value 2.0208, which could visually blur the concentration around the true value. To make the concentration around the true value more visual, we use only a subset of those estimates, those within -10 and 10, to plot the histograms, illustrated by Figure 2.6. Note that the number of estimates outside of this range is 32 for $n = 200$ and 8 for $n = 1000$, which are both small compared to 200, the number of estimates in all.

Figure 2.5: Scatter plots of average effect estimates in penalized spline regression with only two predictors: the top panel with sample size $n = 200$ and bottom with $n = 1000$. The solid horizontal line in each panel identifies the true value of average effect, and each circle represents a different simulated data set.

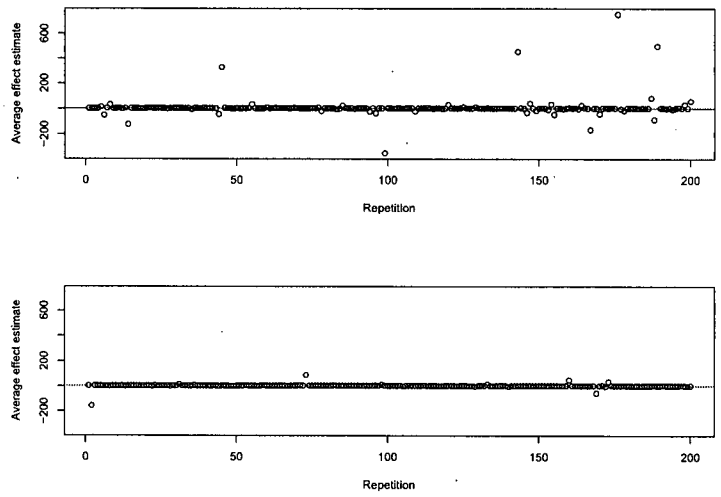
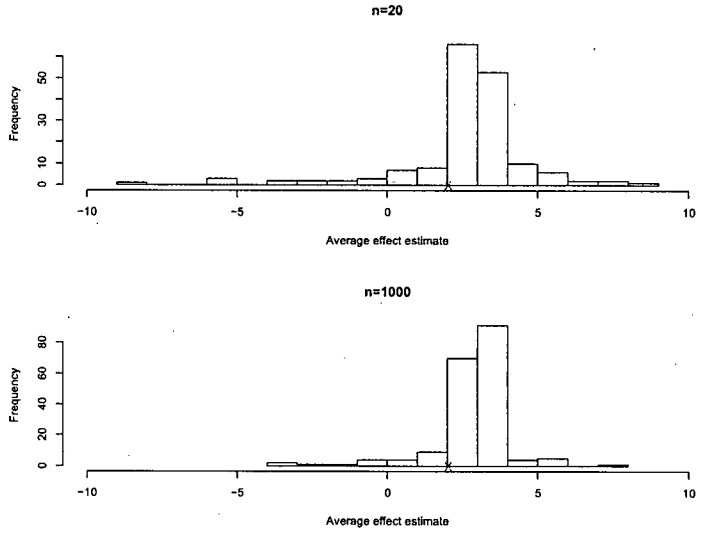


Figure 2.6: Histograms of the average effect estimates in penalized spline regression with only two predictors: the top panel with sample size $n = 200$ and bottom with $n = 1000$, and the symbol '$\times$' in each panel marks the true value of average effect.



2.5 A middle scenario

Although we do not get exactly consistent estimates in the spline regression scenario as discussed above, we still wonder whether there is an intermediate situation between straight-line fitting and curve fitting. In this section, we consider models involving the quadratic terms of the predictors. To explain, we assume the true model to be

$$\mathrm{E}(Y|X_1, X_2) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1^2 + \beta_4 X_2^2 + \beta_{12} X_1 X_2,$$

while the fitted model is

$$\mathrm{E}(Y|X_1, X_2) = \alpha_0 + \alpha_1 X_1 + \alpha_2 X_2 + \alpha_3 X_1^2 + \alpha_4 X_2^2.$$

Assume that the joint distribution of X_1 and X_2 is a bivariate normal distribution with mean vector $\mathbf{0}$, $\mathrm{Cov}(X_1, X_2) = \rho$ and $\mathrm{Var}(X_i) = 1, i = 1, 2$.

Based on the previous result (2.14), we have

$$\begin{aligned}
\alpha_{1*} &= \beta_1 + \beta_{12}\mathrm{E}(X_2) + \beta_{12}\frac{\sum_{k=1}^{4}\Sigma^{k1}\mathrm{E}(X_1X_2)\widetilde{X}_k}{|\Sigma|}, \\
\alpha_{3*} &= \beta_3 + \beta_{12}\frac{\sum_{k=1}^{4}\Sigma^{k3}\mathrm{E}(X_1X_2)\widetilde{X}_k}{|\Sigma|},
\end{aligned}$$

where Σ is the covariance matrix of (X_1, X_2, X_1^2, X_2^2).

Using moments of the bivariate normal distribution, it is easy to derive that

$$\begin{aligned}
\alpha_{1*} &= \beta_1, \\
\alpha_{3*} &= \beta_3 + \frac{\rho}{1+\rho^2}\beta_{12}.
\end{aligned}$$

With Definition 2 of average effect, that is, averaging over the joint distribution of

(X_1, X_2), we have $\delta_1(\boldsymbol{\beta}) = \delta_1^*(\boldsymbol{\beta})$. That is, we still have (exactly) consistent estimator under this setting.

However, interestingly, based on Definition 1, i.e., averaging over the marginal distribution of X_2, we get

$$\begin{aligned} \delta_1^*(x_1; \boldsymbol{\beta}) &= \alpha_1 + 2\alpha_{3*}x_1 \\ &= \beta_1 + 2\left(\beta_3 + \frac{\rho}{1+\rho^2}\beta_{12}\right)x_1, \end{aligned}$$

and

$$\delta_1(x_1; \boldsymbol{\beta}) = \beta_1 + 2\beta_3 x_1.$$

Hence, we get the difference between the fitted and true average effects

$$\delta_1^*(x_1; \boldsymbol{\beta}) - \delta_1(x_1; \boldsymbol{\beta}) = \frac{\rho}{1+\rho^2}\beta_{12}x_1.$$

It is easy to verify that the bias increases when ρ increases for given $X_1 = x_1$. One special case is that when ρ equals 0, the difference disappears. It is also clear that the independence of X_i's make consistent average effect estimates no matter whether the joint distribution of X_i's is normal or not.

Furthermore, if Definition 3 is used, averaging over the conditional distribution $X_2|X_1 = x_1$, we have

$$\delta_1(x_1; \boldsymbol{\beta}) = \beta_1 + 2\beta_3 x_1 + \beta_{12}\rho x_1.$$

Note that $\delta_1^*(x_1; \boldsymbol{\beta})$ does not subject to the change of definition because it does not

depend on X_2. Thus, the bias becomes correspondingly

$$\begin{aligned}\delta_1^*(x_1;\boldsymbol{\beta}) - \delta_1(x_1;\boldsymbol{\beta}) &= \left(\frac{\rho}{1+\rho^2} - \rho\right)\beta_{12}x_1 \\ &= \frac{\rho - \rho^3}{1+\rho^2}\beta_{12}x_1.\end{aligned}$$

It is easy to verify that for any given x_1, the bias increases if $|\rho| < \sqrt{\sqrt{5}-2}$ and decreases otherwise.

Therefore, conditional on different definitions of average effect, we may have different conclusions about the consistency of estimators based on a simpler model.

2.6 Summary

In section 2.2, Result 1 & 2 show the asymptotic distribution of average effect estimators under "true" model and "misspecified" model, respectively. Note we don't have to assume that the two models are nested, that is, our results can be applied to more general model misspecification situations. In the linear regression context, discussed in section 2.5, Result 3 gives the conditions to yield consistent estimator when fitting an additive model without interactions to the data generated from a pairwise interaction model without pure quadratic terms. Although the conditions, independence or joint normality (elliptical-contoured) of the components of $\mathbf{X}$ (centered), may not be satisfied in practice, we could try appropriate transformations to make the distributions of $\mathbf{X}$ close to either of the conditions. In section 2.4, spline regression context, we can still have consistent estimator when the predictors are independent and spline basis are centered. We also explore the bias under the condition of joint normality, and find out the magnitude of bias is quite small even when there is strong dependence between the predictors. That is, the estimator based on additive model without interaction is approximately consistent.

In section 2.5, we consider models with quadratic terms of predictors, where consistent estimator comes into being under the conditions in Result 3. We should be aware of the fact that the consistency of average effect estimator depends on the definition of average effect, as implicated by the example in section 2.5. All the results/conclusions we have in Chapter 2 is based on Definition 2, that is, averaging over the joint distribution of all components of $\mathbf{X}$. However, this definition may not always be appropriate to use. For example, if one is interested in comparison the risk of lung cancer between two groups of people having different smoking habit. Say one group never smoke and the other smoke everyday. There are also bunch of other risk factors, such as gender, age, resident and so on. Since we want to know how smoking makes difference in the risk of lung cancer, we should consider Definition 3, averaging over the distribution of all the other risk factors conditional on smoking factor. Thus, for different scenarios, we need to investigate the consistency of estimator equipped with different definitions. As for which definitions should be applied, it depends on the goal of study.

Chapter 3

Comparison of interaction detectability under different interaction models

This chapter will focus on comparison of power to detect interactions under different regression models, in particular, a pairwise interaction model and a diffuse interaction model. Section 3.1 has the background on asymptotic power under local alternatives for Wald (or quadratic form score) statistics. Section 3.2 gives a concrete example to show how powerful the diffuse interaction model is to detect interactions no matter what the true structure of interaction is diffuse or not.

3.1 General framework

In this section, we give a general result about the asymptotic power function of score (quadratic form) test for presence of interactions are derived based on two models.

Let $\mathcal{F} = \{f(y|\ \mathbf{x}, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \Theta\}$ and $\mathcal{G} = \{g(y|\ \mathbf{x}, \boldsymbol{\omega}), \boldsymbol{\omega} \in \Omega\}$ denote two different parametric families of densities under consideration when modelling the relationship between response variable Y and predictor variables $X_1, \ldots, X_p$. We assume an agreement between the two families at some certain values of their parameters. That is, one member

from G satisfies that

$$g(y|\,\mathbf{x},\boldsymbol{\omega}_0) = f(y|\,\mathbf{x},\boldsymbol{\theta}_0),\ \text{for all } y \text{ and } \mathbf{x},$$

so as $\boldsymbol{\omega}$ moves away from $\boldsymbol{\omega}_0$, $g(y|\,\mathbf{x},\boldsymbol{\omega})$ moves away from $\mathcal{F}$ in some specified way. Under model $\mathcal{F}$ ($\mathcal{G}$), $\boldsymbol{\theta} = \boldsymbol{\theta}_0(\boldsymbol{\omega} = \boldsymbol{\omega}_0)$ stands for the null model of no interactions.

Let f,g denote densities and F, G denote the corresponding distributions. Let $p_1 = \dim(\boldsymbol{\theta}), p_2 = \dim(\boldsymbol{\omega})$ be the dimensions of $\mathcal{F}$ and $\mathcal{G}$, with

$$s_F(\boldsymbol{\theta}, Y, \mathbf{X}) = \partial[\log\{f(Y|\,\mathbf{X},\boldsymbol{\theta})\}]/\partial\boldsymbol{\theta}$$

and

$$s_G(\boldsymbol{\omega}, Y, \mathbf{X}) = \partial[\log\{g(Y|\,\mathbf{X},\boldsymbol{\omega})\}]/\partial\boldsymbol{\omega}$$

being the respective score vectors, and

$$I_F(\boldsymbol{\theta}) = \mathrm{E}_{\boldsymbol{\theta}}\{s_F(\boldsymbol{\theta}, Y, \mathbf{X})s_F'(\boldsymbol{\theta}, Y, \mathbf{X})\}$$

and

$$I_G(\boldsymbol{\omega}) = \mathrm{E}_{\boldsymbol{\omega}}\{s_G(\boldsymbol{\omega}, Y, \mathbf{X})s_G'(\boldsymbol{\omega}, Y, \mathbf{X})\}$$

being the corresponding Fisher information matrices. Note the two above expectation operations are with respect to joint distribution of Y and $\mathbf{X}$. Let $f_{\mathbf{X}}(\mathbf{x})$ be density of $\mathbf{X}$ with respect to measure $\nu(\mathbf{x})$ (either Lebesgue or counting measure).

To compare the capability to detect interactions under the two families $\mathcal{F}$ and $\mathcal{G}$, we set up two sets of hypothesis tests with $\mathcal{F}$ as the fitted model and $\mathcal{G}$ as the true model. The reverse case with $\mathcal{F}$ being the true distribution while $\mathcal{G}$ being fitted model, is just an analog. Hence in the following, we just take the former case for demonstration.

To make the comparison tractable and also to produce some nontrivial asymptotic powers that are not all equal to 1, we use a Pitman-type local analysis (developed by Le Cam (1960)), focusing on $n^{-1/2}$ – neighborhoods of the true parameter values.

Say $\mathcal{F}$ is the fitted model while $\mathcal{G}$ is the true model with $\boldsymbol{\omega}_n = \boldsymbol{\omega}_0 + n^{-1/2}\Delta\boldsymbol{\eta}$, where Δ is a scalar and $\boldsymbol{\eta}$ is a vector of length p_2. The hypotheses we used here are $H_0 : C\boldsymbol{\theta} = \boldsymbol{\zeta}_0$ versus $H_a : C\boldsymbol{\theta} \neq \boldsymbol{\zeta}_0$, where C is a matrix $r \times \dim(\boldsymbol{\theta})$ of full row rank and $\boldsymbol{\zeta}_0$ is a certain vector (for example, zero vectors denotes an additive model if $C\boldsymbol{\theta}$ is the vector of interaction parameters). Note that the above hypotheses are more general than testing $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$, where C just reduces to an identity matrix as a special case. The above hypotheses consider any linear combination of the whole parameter $\boldsymbol{\theta}$ under the fitted model.

To compare the capability to detect interactions, the parameters of interest are just those related to interactions.

In particular, if the pairwise interaction model (defined later by (3.6)) is fitted, the whole parameter vector

$$\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \beta_{12}, \ldots, \beta_{(p-1)p})'.$$

Only the last $q = p(p-1)/2$ components relate to interactions. Hence let $C_1 = (\mathbf{0}_{q\times(p+1)}, I_{q\times q})$ so that $C_1\boldsymbol{\theta} = (\beta_{12}, \ldots, \beta_{(p-1)p})'$, which is what the parameter vector of interest for testing. If the fitted model is the diffuse interaction model (3.5), the whole parameter vector is $\boldsymbol{\omega} = (\beta_0, \beta_1, \ldots, \beta_p, \lambda)'$ and only λ relates to interaction. Thus let $C_2 = (\mathbf{0}_{1\times(p+1)}, 1)$ so that $C_2\boldsymbol{\omega} = \lambda$.

Let $\widehat{\boldsymbol{\theta}}_n$ be the maximum likelihood estimator based on $f(y|\,\mathbf{x}, \boldsymbol{\theta})$ from n observations from $g(y|\,\mathbf{x}, \boldsymbol{\omega}_n)$.

Then $n\big(C(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)' \left\{CI_F^{-1}(\boldsymbol{\theta}_0)C'\right\}^{-1} \big(C(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)$ is the Wald test statistic used

here, whose asymptotic distribution is χ^2 with degrees of freedom of rank(C) if $f(\cdot|\,\boldsymbol{\theta}_0)$ is the true model. Because $\mathcal{F}$ is not the true model,

$$n^{1/2}C(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = n^{1/2}C\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*(\boldsymbol{\omega}_n)\right) + n^{1/2}C\left(\boldsymbol{\theta}_*(\boldsymbol{\omega}_n) - \boldsymbol{\theta}_0\right), \tag{3.1}$$

where $\boldsymbol{\theta}_*(\boldsymbol{\omega})$ is the parameter vector which minimizes the Kullback-Leibler information criterion, that is

$$\boldsymbol{\theta}_*(\boldsymbol{\omega}) = \text{argmin}_{\boldsymbol{\theta}} \int \left\{\log \frac{g(y|\,\mathbf{x},\boldsymbol{\omega})}{f(y|\,\mathbf{x},\boldsymbol{\theta})}\right\} g(y|\,\mathbf{x},\boldsymbol{\omega}) f_{\mathbf{X}}(\mathbf{x}) dy d\nu(\mathbf{x}).$$

Note that the fact $g(\cdot|\,\boldsymbol{\omega}_0) = f(\cdot|\,\boldsymbol{\theta}_0)$ yields that $\boldsymbol{\theta}_*(\boldsymbol{\omega}_0) = \boldsymbol{\theta}_0$.

By White (1982) we know that the first item on the right side of (3.1) is asymptotically normal with mean $\mathbf{0}$ and covariance matrix $CI_P^{-1}(\boldsymbol{\theta}_0)C'$. So we only need to work on the second item on the right side in (3.1).

By the definition of $\boldsymbol{\theta}_*(\boldsymbol{\omega})$, we know that $\boldsymbol{\theta}_*(\boldsymbol{\omega})$ is the value of $\boldsymbol{\theta}$ solving

$$\int s_F(\boldsymbol{\theta}(\boldsymbol{\omega}), y, \mathbf{x}) g(y|\,\mathbf{x},\boldsymbol{\omega}) f_{\mathbf{X}}(\mathbf{x}) dy d\nu(\mathbf{x}) = 0. \tag{3.2}$$

Based on Gustafson (2001), implicit differentiation of (3.2) gives

$$\mathrm{E}_{\boldsymbol{\omega}}[s'_F(\boldsymbol{\theta}_*(\boldsymbol{\omega}), Y, \mathbf{X})]\boldsymbol{\theta}'_*(\boldsymbol{\omega}) + \mathrm{E}_{\boldsymbol{\omega}}[s'_F(\boldsymbol{\theta}_*(\boldsymbol{\omega}), Y, \mathbf{X}) s_G(\boldsymbol{\omega}; Y, \mathbf{X})] = 0.$$

Evaluated at $\boldsymbol{\omega} = \boldsymbol{\omega}_0$, the above equality yields

$$\frac{\partial \boldsymbol{\theta}_*}{\partial \boldsymbol{\omega}}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0} = I_F^{-1}(\boldsymbol{\theta}_0)\mathrm{E}_{\boldsymbol{\theta}_0}\{s'_F(\boldsymbol{\theta}_0; Y, \mathbf{X}) s_G(\boldsymbol{\omega}_0; Y, \mathbf{X})\}, \tag{3.3}$$

which is derived by the fact that $\boldsymbol{\theta}_*(\boldsymbol{\omega}_0) = \boldsymbol{\theta}_0$. Therefore, we have

$$\begin{aligned} n^{1/2}C\left\{\boldsymbol{\theta}_*(\boldsymbol{\omega}_n) - \boldsymbol{\theta}_0\right\} &= n^{1/2}C\left\{\frac{\partial \boldsymbol{\theta}_*}{\partial \boldsymbol{\omega}}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\Delta\boldsymbol{\eta}n^{-1/2} + O(n^{-1})\right\} \\ &\to \Delta C\frac{\partial \boldsymbol{\theta}_*}{\partial \boldsymbol{\omega}}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\boldsymbol{\eta}, \end{aligned}$$

where

$$\frac{\partial \mathbf{u}}{\partial \mathbf{v}}[i,j] = \frac{\partial u_i}{\partial v_j}.$$

Recall the equality (3.1). Based on O'Brien et al. (2006), assuming $\mathcal{G}$ being the true model, the asymptotic distribution of

$$n\big(C(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)'\left\{CI_F^{-1}(\boldsymbol{\theta}_0)C'\right\}^{-1}\big(C(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\big)$$

is a noncentral χ^2 with degrees of freedom of rank(C) and noncentrality parameter δ is calculated by

$$\delta = \left\{\Delta C\frac{\partial \boldsymbol{\theta}_*}{\partial \boldsymbol{\omega}}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\boldsymbol{\eta}\right\}'\left\{CI_F^{-1}(\boldsymbol{\theta}_0)C'\right\}^{-1}\left\{\Delta C\frac{\partial \boldsymbol{\theta}_*}{\partial \boldsymbol{\omega}}\Big|_{\boldsymbol{\omega}=\boldsymbol{\omega}_0}\boldsymbol{\eta}\right\}. \tag{3.4}$$

Suppose the equivalent null hypothesis for $\mathcal{G}$ is $C_G\boldsymbol{\omega} = \boldsymbol{\zeta}_{0G}$ where C_G is a $r_G \times \dim(\boldsymbol{\omega})$ matrix. The Wald statistic based on the fitted model $\mathcal{G}$ is

$$n\big(C_G(\widehat{\boldsymbol{\omega}}_n - \boldsymbol{\omega}_0)\big)'\left\{C_GI_G^{-1}(\boldsymbol{\omega}_0)C_G'\right\}^{-1}\big(C(\widehat{\boldsymbol{\omega}}_n - \boldsymbol{\omega}_0)\big).$$

Its asymptotic distribution is noncentral $\chi^2_{r_G}(\delta_G)$, where

$$\delta_G = \{\Delta C_G\boldsymbol{\eta}\}'\left\{C_GI_G^{-1}(\boldsymbol{\omega}_0)C_G'\right\}^{-1}\{\Delta C_G\boldsymbol{\eta}\}.$$

3.2 Comparison between pairwise interaction models and diffuse interaction models

3.2.1 Introduction of diffuse interaction model

Greenland (1983) pointed out that the powers of statistical tests to detect interactions are very low in some commonly encountered epidemiological studies. We could imagine even lower power in the situations where the number of risk factors is rather large and only a very small fraction of all possible (pairwise) interaction terms really play a role. Gustafson et al. (2005) proposed another kind of interaction model, the *diffuse interaction* model, to deal with difficulties caused by a large number of risk factor under pairwise interaction models. By diffuse interaction, we mean that the effect of a particular risk factor either increases (*synergism*) or decreases (*antagonism*) as all the other risk factors increase, without regard to which of the other risk factors get involved with the effect modification. The diffuse interaction model introduced in Gustafson et al. (2005) is defined as

$$\begin{aligned} \mathrm{E}\left(Y|X_1,\ldots,X_p\right) &= \mu_D \\ &= \beta_0 + \left\{\sum_{i=1}^{p}(\beta_i X_i)^{\lambda}\right\}^{1/\lambda}, \quad X_i \geq 0,\ \beta_i > 0. \end{aligned} \tag{3.5}$$

The parameter λ reflects the magnitude of the synergism/antagonism. Take a binary X_j for example, if $\lambda > 1$ then it is easily verified that the interaction is antagonistic, in the sense that the value of $\mathrm{E}\left(Y|X_j = 1, \mathbf{X}_{(j)} = \mathbf{x}_{(j)}\right) - \mathrm{E}\left(Y|X_j = 0, \mathbf{X}_{(j)} = \mathbf{x}_{(j)}\right)$ decreases in each component of $\mathbf{x}_{(j)}$. If X_j is continuous, it is also easy to show that $\lambda > 1$ gives

$$\frac{\partial \mathrm{E}\left(Y|\mathbf{X} = \mathbf{x}\right)}{\partial x_j \partial x_k} > 0, \forall k \neq j,$$

which means synergism based on the definition in Section 1.2. Conversely, $\lambda < 1$ corresponds to synergism. That is, the effect modification caused by any putative risk factor increases as other risk factors increase. The magnitude of the difference between λ and 1 implies how much synergism/antagonism is present. However, we should be aware that λ does not provide the information about which of those risk factors contribute in the effect modification for any putative risk factor.

3.2.2 Power comparison

Say the response variable Y is normally distributed and $X_1, \ldots, X_p$ are the corresponding explanatory variables. We study an example under the following two interaction models. Pairwise interaction model:

$$\begin{aligned} \mathrm{E}(Y|X_1, \ldots, X_p) &= \mu_P \\ &= \beta_0 + \sum_{i=1}^{p} \beta_i X_i + \sum_{1 \le i < j \le p} \beta_{ij} X_i X_j, \end{aligned} \tag{3.6}$$

Recall diffuse interaction model:

$$\begin{aligned} \mathrm{E}(Y|X_1, \ldots, X_p) &= \mu_D \\ &= \beta_0 + \left\{ \sum_{i=1}^{p} (\beta_i X_i)^{\lambda} \right\}^{1/\lambda}, \; X_i \ge 0, \; \beta_i \ge 0. \end{aligned}$$

Let $\boldsymbol{\beta}_M = (\beta_1, \ldots, \beta_p)'$ and $\boldsymbol{\beta}_I = (\beta_{ij})_{q \times 1}$ $(q = p(p-1)/2)$, the coefficients of pairwise interaction terms. Let $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \boldsymbol{\beta}_I', \sigma^2)'$, which is the whole parameter vector in the pairwise interaction model and let $\Omega = (\beta_0, \beta_1, \ldots, \beta_p, \lambda, \sigma^2)'$, which is the whole parameter vector in the diffuse interaction model.

If $\boldsymbol{\beta}_I = \boldsymbol{\beta}_{I0} = \mathbf{0}$ in the pairwise interaction model, the model reduces to be an additive model without interaction terms. Correspondingly, if $\lambda = \lambda_0 = 1$ in the diffuse

interaction model, the model reduces to an additive model as well. Therefore, the two interaction models are the same if evaluating at $\boldsymbol{\beta}_{I0}$ and λ_0 respectively. Denote by f_P the density function of $Y|X_1,\ldots,X_p$ under the pairwise interaction model, and by f_D the density function under the diffuse interaction model. Denote by I_P the information matrix under the pairwise interaction model and I_D under the diffuse interaction model. Denote by $s_P(\cdot, Y, \mathbf{X})$ the score function of f_P under pairwise interaction model and so $s_D(\cdot, Y, \mathbf{X})$ the score function of f_D under diffuse interaction model. That is,

$$s_P(\cdot, Y, \mathbf{X}) = \sigma^{-2}(Y - \mu_P)\begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \\ X_1X_2 \\ \vdots \\ X_{p-1}X_p \end{pmatrix},$$

$$s_D(\cdot, Y, \mathbf{X})|_{\lambda=1} = \sigma^{-2}(Y - \mu_D)\begin{pmatrix} 1 \\ X_1 \\ \vdots \\ X_p \\ \frac{\partial \mu_D}{\partial \lambda}|_{\lambda=1} \end{pmatrix},$$

where $$\frac{\partial \mu_D}{\partial \lambda}\Big|_{\lambda=1} = -\left(\sum_{i=1}^{p}\beta_i X_i\right)\log\left(\sum_{i=1}^{p}\beta_i X_i\right) + \sum_{i=1}^{p}\beta_i X_i \log(\beta_i X_i).$$

It is obvious that interaction can be measured by only one parameter λ in (3.5) while $p(p-1)/2$ parameters are used in (3.6). Hence in the diffuse interaction model (3.5), we assume that all the predictors interact in the same direction, either synergistically or antagonistically, which corresponds to $\lambda < 1$ or $\lambda < 1$ respectively. It is reasonable

to imagine that model (3.5) is more powerful to detect interaction since detecting the interaction effect in one direction could be easier than that in many possible directions. The comparison between two models will be explored by the following cases.

Case I: Assume that the diffuse interaction (3.5) is the true model with $\lambda_n = 1 + \Delta n^{-1/2}$, where Δ is a scalar.

(i) The fitted model is the pairwise interaction model shown as (3.6). To test whether there are interaction effects, set up the following hypotheses.

$H_0 : C_1\boldsymbol{\theta} = \mathbf{0}$ versus $H_a : C_1\boldsymbol{\theta} \neq \mathbf{0}$, i.e.,

$H_0 : \boldsymbol{\beta}_I = \mathbf{0}$ versus $H_a : \boldsymbol{\beta}_I \neq \mathbf{0}$.

Recall that $C_1 = (\mathbf{0}_{q\times p}, I_{q\times q})$. Let $\widehat{\boldsymbol{\theta}}_n$ denote the MLE of $\boldsymbol{\theta}$. Therefore, we have

$$\text{Power} = P\left(n\{C_1(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\}' \{C_1 I_P^{-1}(\boldsymbol{\theta}_0) C_1'\}^{-1} \{C_1(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\} \geq \chi^2_{q,\,\alpha}\right),$$

where α is the test level and $\chi^2_{q,\,\alpha}$ is the upper α quantile of χ^2_q. With the partition of $\boldsymbol{\theta} = ((\beta_0, \boldsymbol{\beta}_M')', \boldsymbol{\beta}_I')'$, we get the corresponding partition of $I_P^{-1}(\boldsymbol{\theta}_0)$.

$$I_P^{-1}(\boldsymbol{\theta}_0) = \begin{pmatrix} I_P^{11}(\boldsymbol{\theta}_0) & I_P^{12}(\boldsymbol{\theta}_0) \\ I_P^{21}(\boldsymbol{\theta}_0) & I_P^{22}(\boldsymbol{\theta}_0) \end{pmatrix}.$$

Hence,

$$C_1 I_P^{-1}(\boldsymbol{\theta}_0) C_1' = I_P^{22}(\boldsymbol{\theta}_0).$$

According to the results we discussed in the previous section, we know that

$$n(C_1(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0))'\{I_P^{22}(\boldsymbol{\theta}_0)\}^{-1}(C_1(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)) \xrightarrow{D} \chi^2_q(\delta),$$

where the noncentrality parameter is

$$\delta = \left\{ C_1 \frac{\partial \boldsymbol{\theta}_*}{\partial \lambda} \Big|_{\lambda=1} \Delta \right\}' \{ I_P^{22}(\boldsymbol{\theta}_0) \}^{-1} \left\{ C_1 \frac{\partial \boldsymbol{\theta}_*}{\partial \lambda} \Big|_{\lambda=1} \Delta \right\},$$

where

$$\frac{\partial \boldsymbol{\theta}_*}{\partial \lambda} \Big|_{\lambda=1} = I_P^{-1}(\boldsymbol{\theta}_0) \left\{ \mathrm{E}_{\boldsymbol{\theta}_0} \{ s_P(\boldsymbol{\theta}_0, Y, X) s_D(\lambda_0, Y, X) \} \right\}.$$

Note that $s_D(\lambda_0, Y, X)$ is the derivative of $\log f_D(\boldsymbol{\omega}, Y, \mathbf{X})$ with respect to λ evaluated at λ_0, i.e., the last component of the score vector $s_D(\boldsymbol{\omega}, Y, \mathbf{X})$.

Therefore the asymptotic power is $P(\chi_q^2(\delta) > \chi_{q,\,\alpha}^2)$, where $\chi_{q,\,\alpha}^2$ is the upper α quantile of χ^2- distribution with degrees of freedom q.

(ii) Fitted model is diffuse interaction model. Now the hypotheses to be tested are $H_0 : C_2\boldsymbol{\omega} = 1$ versus $H_a : C_2\boldsymbol{\omega} \neq 1$, i.e.,

$H_0 : \lambda = 1$ vs. $H_1 : \lambda \neq 1$.

Recall that $C_2 = (\mathbf{0}_{1\times p},\, 1)$. Therefore the asymptotic power is

$$P\left(\chi_1^2(\delta) > \chi_{1,\,\alpha}^2\right),$$

where

$$\begin{aligned}
\delta &= \Delta^2 V_D^{-1}(\lambda_0) \\
V_D(\lambda_0) &= C_2 I_D^{-1}(\boldsymbol{\omega}_0) C_2', \\
I_D(\boldsymbol{\theta}_0) &= \mathrm{E}_{\boldsymbol{\omega}_0} \left\{ \left(\frac{\partial \log f_D}{\partial \boldsymbol{\omega}} \right) \left(\frac{\partial \log f_D}{\partial \boldsymbol{\omega}} \right)' \right\}, \\
\boldsymbol{\omega}_0 &= (\beta_0, \beta_1, \ldots, \beta_p, \lambda_0)'.
\end{aligned}$$

Remark: As a matter of fact, the $\hat{\lambda}_n$ should be positive to make the operation $(\cdot)^{\hat{\lambda}_n}$ meaningful. That is, the above analyses work well when sample size n big enough since

$\hat{\lambda}_n$ converges to 1, which is away from the boundary value 0.

Case II: Assume the true model is a pairwise interaction model with

$$\begin{aligned}\boldsymbol{\beta}_{In} &= \begin{pmatrix} \beta_{12,\,n} \\ \vdots \\ \beta_{(p-1)p,\,n} \end{pmatrix}_{q\times 1} \\ &= n^{-1/2}\Delta\boldsymbol{\eta},\end{aligned}$$

where $\boldsymbol{\eta}$ is a $q \times 1$ vector. For this time being, we set each element of $\boldsymbol{\eta}$ to be 1, which means every pair of interactions is positive. Later we discuss the consequences of different choices of $\boldsymbol{\eta}$.

(i) Fitted model is diffuse interaction model (3.5). The hypotheses to be tested here are $H_0 : C_2\boldsymbol{\omega} = 1$ vs. $H_a : C_2\boldsymbol{\omega} \neq 1$. All the following is an analog of i) of Case I. Now $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \lambda)'$.

The asymptotic power is

$$P\left(\chi_1^2(\delta) > \chi_{1,\,\alpha}^2\right),$$

where

$$\begin{aligned}\delta &= \left\{\left(C_2\frac{\partial\lambda_*}{\partial\boldsymbol{\beta}_I}\Big|_{\boldsymbol{\beta}_I=\mathbf{0}}\right)\Delta\boldsymbol{\eta}\right\}'\left\{C_2 I_D^{-1}(\boldsymbol{\theta}_0)C_2'\right\}^{-1}\left\{\left(C_2\frac{\partial\lambda_*}{\partial\boldsymbol{\beta}_I}\Big|_{\boldsymbol{\beta}_I=\mathbf{0}}\right)\Delta\boldsymbol{\eta}\right\}, \\ \frac{\partial\lambda_*}{\partial\boldsymbol{\beta}_I}\Big|_{\boldsymbol{\beta}_I=\mathbf{0}} &= \left\{I_D^{-1}(\boldsymbol{\omega}_0)\mathrm{E}_{\boldsymbol{\omega}_0}\left(s_D(\omega_0, Y, X)s_P'(\boldsymbol{\theta}_0, Y, X)\right)\right\}_{(p+2)\cdot},\end{aligned}$$

where $M_{r\cdot}$ denotes for the rth row vector of matrix M.

(ii) Fitted model is pairwise interaction model (3.6). The hypotheses to be tested here are $H_0 : C_1\boldsymbol{\theta} = \mathbf{0}$ versus $H_1 : C_1\boldsymbol{\theta} \neq \mathbf{0}$.

We have

$$n\{C_1(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\}'\{I_P^{22}(\boldsymbol{\theta}_0)\}^{-1}C_1(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{L} \chi_q^2(\delta),$$

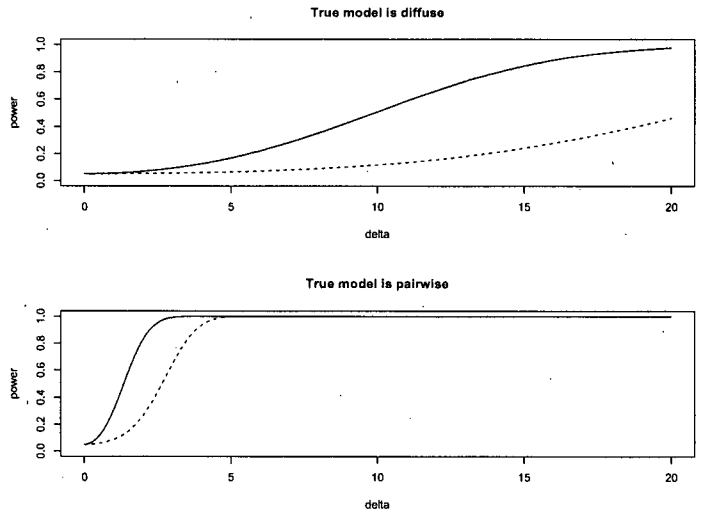
with $\delta = \Delta^2 \boldsymbol{\eta}' \{I_P^{22}(\boldsymbol{\theta}_0)\}^{-1} \boldsymbol{\eta}$. Hence the asymptotic power is $P\left(\chi_q^2(\delta) \geq \chi_{q,\,\alpha}^2\right)$.

As discussed above, we have derived four asymptotic power functions for the four subcases, corresponding to all possible patterns of either pairwise interaction or diffuse interaction model being fitted while the underlying data generation process is the other one. However, it is not easy to tell which power is bigger based on the expressions of asymptotic powers. To be more specific, we take an example as below. Suppose we have $p = 9$ predictors at hand, the true values of all components of $\boldsymbol{\beta}_M$ are all equal to 0.5 and $\beta_0 = 0$. The variance of random error is set to be $\sigma^2 = 1$. Also the predictors are identically independently distributed as Bernoulli with parameter ξ, the probability of success. Then we plot the power functions against the value of Δ, with $\xi = 0.5$.

As shown before, the key things to compute the power are $I_P = \mathrm{E}\left\{\mathrm{s_P}(\cdot, \mathrm{Y}, \mathbf{X})\mathrm{s_P'}(\cdot, \mathrm{Y}, \mathbf{X})\right\}$, $I_D = \mathrm{E}\left\{\mathrm{s_D}(\cdot, \mathrm{Y}, \mathbf{X})\mathrm{s_D'}(\cdot, \mathrm{Y}, \mathbf{X})\right\}$, $\mathrm{E}\left\{\mathrm{s_P}(\cdot, \mathrm{Y}, \mathbf{X})\mathrm{s_D'}(\cdot, \mathrm{Y}, \mathbf{X})\right\}$ and $\mathrm{E}\left\{\mathrm{s_D}(\cdot, \mathrm{Y}, \mathbf{X})\mathrm{s_P'}(\cdot, \mathrm{Y}, \mathbf{X})\right\}$.

Under the above settings, Figure 3.1 shows the result of the powers in the four subcases discussed above. The solid lines denote the power curves for diffuse interaction model fitting and the dashed lines denote those of pairwise interaction model fitting. We find that the diffuse interaction model is more powerful to detect interactions than pairwise interaction model, regardless of whether the true model we postulate is the diffuse interaction model or not.

Figure 3.1: Power Curves with X_i's $\overset{i.i.d.}{\sim}$ Bernoulli(0.5): the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting. In the bottom panel $\boldsymbol{\eta} = \mathbf{1}_{q\times 1}$.



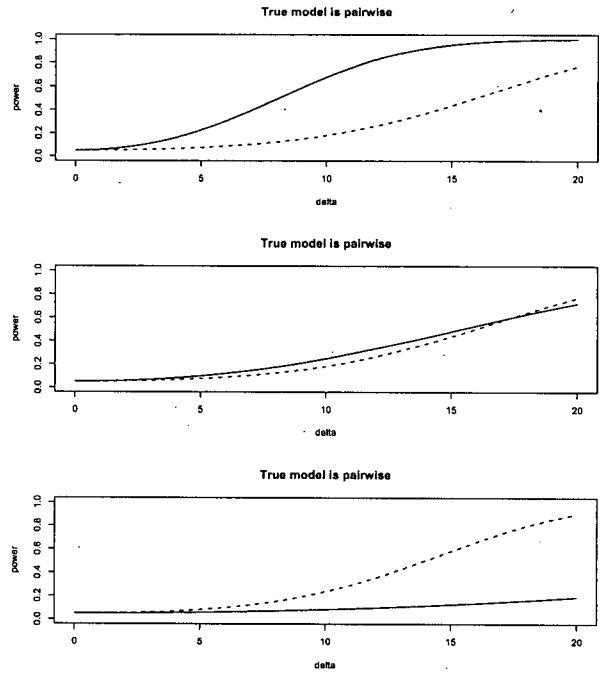
In case II, consider replacing $\boldsymbol{\eta} = (1, \ldots, 1)'$ with a vector of $\boldsymbol{\eta}$ having entries ± 1, that is, the pairwise interaction terms have different "directions". For those terms with positive coefficient, it implies that the impact of the interaction is to increase the value of response variable when the corresponding predictors increase while other predictors keeping unchanged. While for the terms with negative coefficient, the interaction effect causes decreased response if the involved predictors increase with other predictors unchanged.

As opposed to a pairwise interaction model, the role of λ in the diffuse interaction model is to measure the magnitude of the overall interaction among the predictors. Note that diffuse interaction models do not specify which of the predictors do or do not contribute to the interaction. Hence, we could expect the power of diffuse interaction model

would get worse when the true model under consideration involves more mixed directions of interaction effect.

Since the "overall" interaction is getting weaker and weaker as more other direction of interaction terms appear in $\boldsymbol{\eta}$, we would expect the performance of the diffuse interaction models to get worse and worse. That is what Figure 3.2 implies. In the first panel, with the choice of $\boldsymbol{\eta}$ stating that all interaction terms have the same direction, the diffuse interaction model performs better than the pairwise interaction model in terms of power. However, when $\boldsymbol{\eta}$ is changed to have only part of the interaction terms playing the role in the same direction, there is crossing of the two power curves as shown in the middle panel. That is, for some values of Δ, the diffuse interaction model works better while for some other values of Δ, the pairwise interaction model does better. Moreover, when $\boldsymbol{\eta}$ is set to have more mixed directions of interaction effect, the performance of diffuse interaction model is worse as implied in the last panel. In Figure 3.2, the top panel is the power curves obtained by setting $\boldsymbol{\eta}$ of a vector of 1's, the middle panel with $\boldsymbol{\eta}$ of a vector made up of 10 1's and 26 0's, and the bottom panel with $\boldsymbol{\eta}$ of a vector made up of 10 1's, 10 0's and 16 -1's. Note here to make the plots across different $\boldsymbol{\eta}$'s comparable, we normalize each $\boldsymbol{\eta}$ to have length equal to 1.

Figure 3.2: Different choices of $\boldsymbol{\eta}$ in Case II with binary predictors: the top panel involves all predictor pairs interacting positively, the middle panel has only a few of predictor pairs interacting positively, and the bottom panel has more mixed directions of interactions. The lengths of $\boldsymbol{\eta}$'s in different panels are normalized to be 1. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting.



Subsequently what happens if other settings, like ξ, $\boldsymbol{\beta}_M$, $\sigma^2(=\text{Var}(\epsilon))$ vary? Figure 3.3 shows the outcome of different ξ's with $\boldsymbol{\beta}_M = 0.5\mathbf{1}_p$ and $\sigma^2 = 1$. In the leftmost panel, ξ =0.2, while ξ =0.5 in the middle panel and ξ =0.8 in the rightmost one. The three plots in the top panel are power curves in case of the true model being diffuse interaction, and the other three in the bottom under the true model being pairwise interaction. As ξ varies, the performance of both models get worse. However, the difference among the other three plots in the bottom panel where the true model is pairwise interaction model is not great. In other words, the distribution of predictors has greater effect in the case

of diffuse interaction model being true model than it does in case of pairwise interaction model being true model.

Figure 3.3: Different choices of ξ in X_i's distribution: the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting. From left to right across the three columns, ξ=0.2,0.5 and 0.8 respectively.

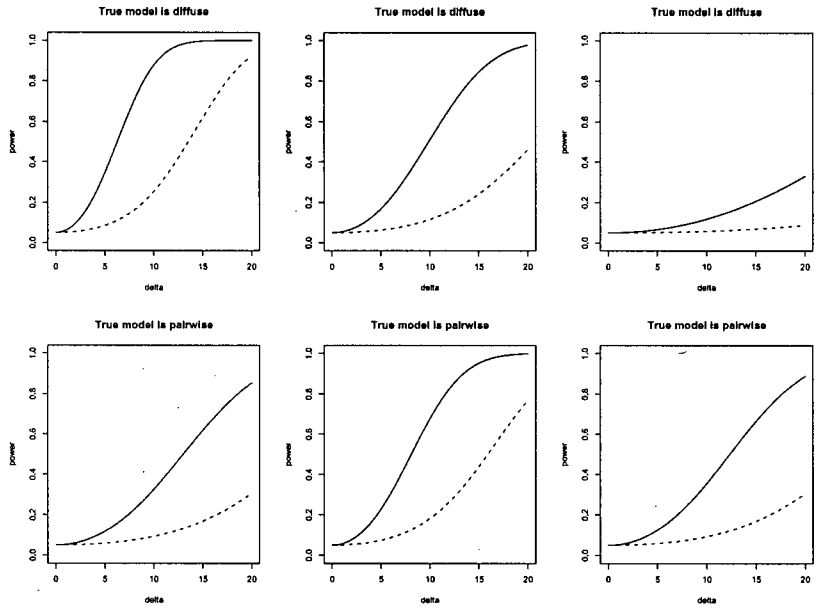


Figure 3.4 shows the result of different choices of $\boldsymbol{\beta}_M$. For simplicity, we set all the components of $\boldsymbol{\beta}_M$ equal to each other, that is $\boldsymbol{\beta}_M \propto b\mathbf{1}_p$. In the leftmost panel, b is set to be 0.1, while 0.5 in the middle panel and 1 in the rightmost one. From the three plots in the top panel, we can see that the performance of both models get better as the magnitude of b increases. However, the three plots in the bottom panel are almost the same. Therefore, we get the similar conclusion as above, where ξ varies. Change of b has more effect in the case that the diffuse interaction model is true.

Figure 3.4: Different choices of b in $\boldsymbol{\beta}_M = b\mathbf{1}_p$: the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting. From left to right across the three columns, b=0.1,0.5 and 1 respectively.

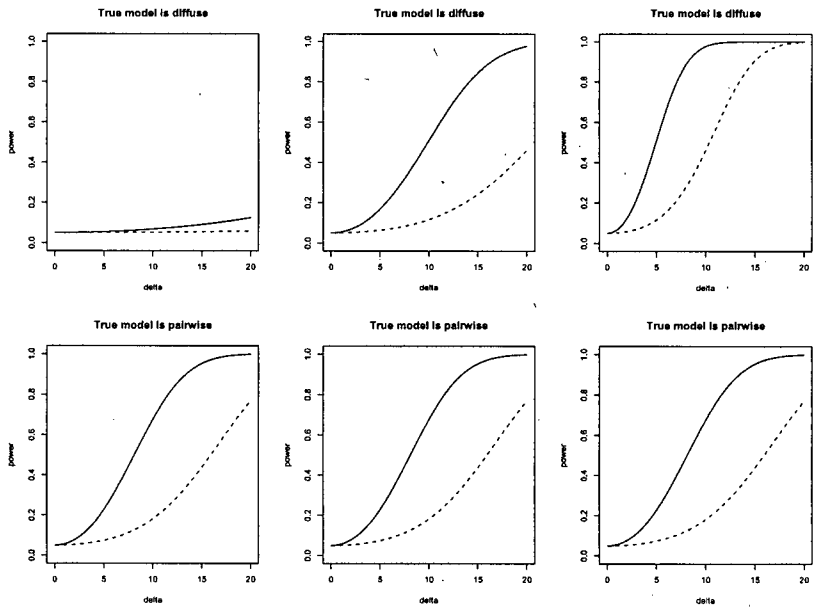


Figure 3.5 shows what happens if σ^2 varies. In the leftmost panel, σ^2 set to be 0.5, while 1 in the middle panel and 5 in the rightmost one. Now the three plots in the top panel are quite similar to those in the bottom respectively. In fact, based on the formula (3.4), we know that the noncentrality parameter δ is just proportional to σ^{-2}. Therefore if we change the scale of Δ according to the value of σ^2, the shapes of the power curves are the same across different σ^2. That is what Figure 3.6 implies.

Figure 3.5: Different choices of Var(ϵ):the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting. From left to right across the three columns, Var(ϵ)=0.5,1 and 5 respectively.

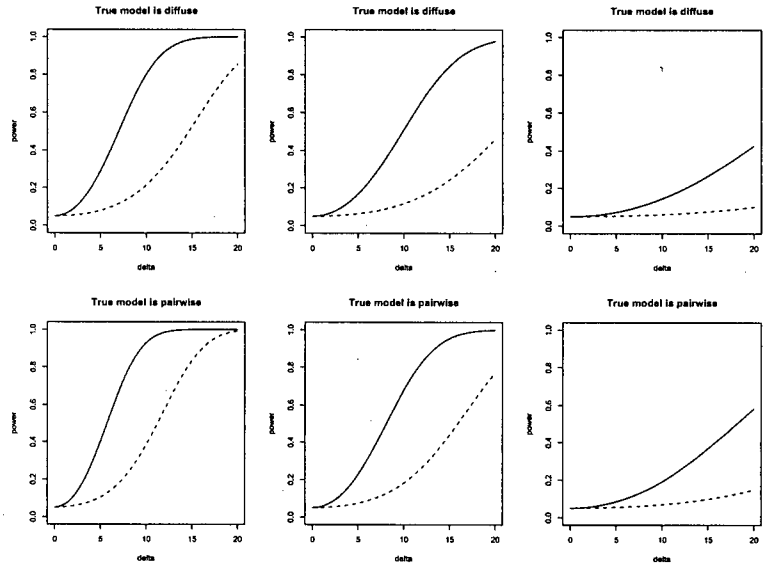
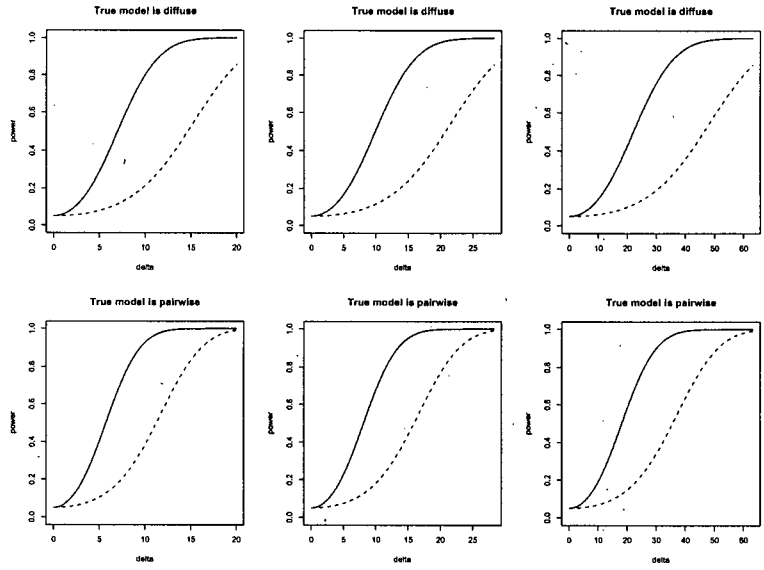


Figure 3.6: Different choices of $\mathrm{Var}(\epsilon)$ with scaled Δ: the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting.



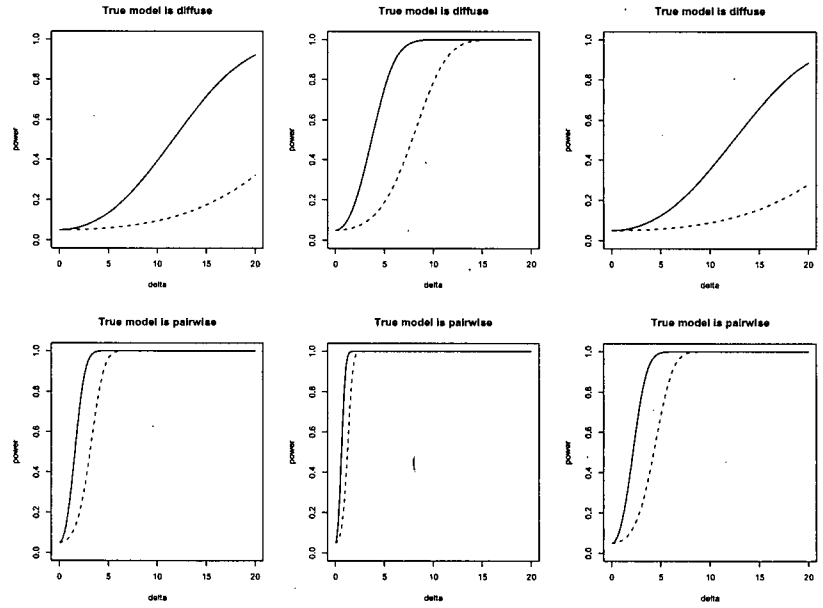
All the above plots are based on the distribution of $\mathbf{X}$ having independent components. In the situation that those predictors are dependent, what happens if the dependency changes? To construct a dependent structure between those predictors, we let

$$X_1, \ldots, X_p | Z \overset{i.i.d.}{\sim} Bernoulli(Z),$$
$$Z \sim Beta(a, b).$$

By a little algebra, we have $\rho = (1 + a + b)^{-1}$, as the correlation between any pair of $X_1, \ldots, X_p$. Hence by changing the values of a and b, we may get different correlations. In Figure 3.7, the first column is power curves obtained from $a = b = 100$, so that $\rho = 0.005$; for the second column, $a = b = 0.5$, hence $\rho = 0.5$ and for the third column $a = b = 0.01$, which leads to $\rho = 0.98$. The change of correlation does affect the power performance

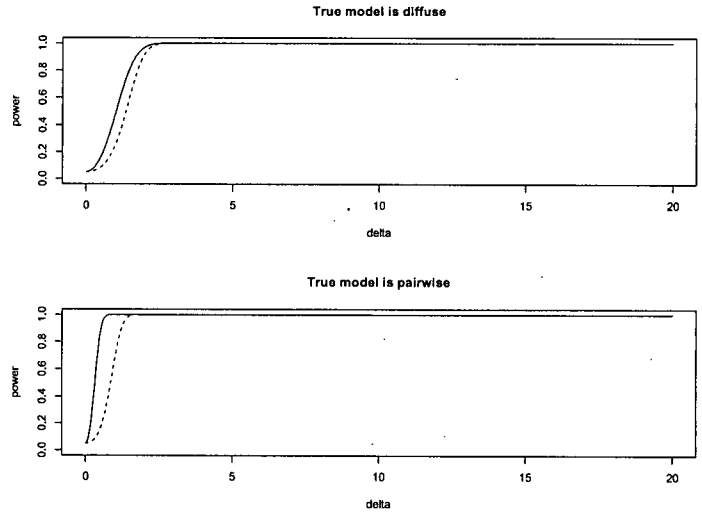
more in the case that diffuse model is true than that in the case that pairwise interaction model is true.

Figure 3.7: Different choices of ρ among X_i's: the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting. From left to right across the three columns, ρ=0.005,0.5 and 0.98 respectively.



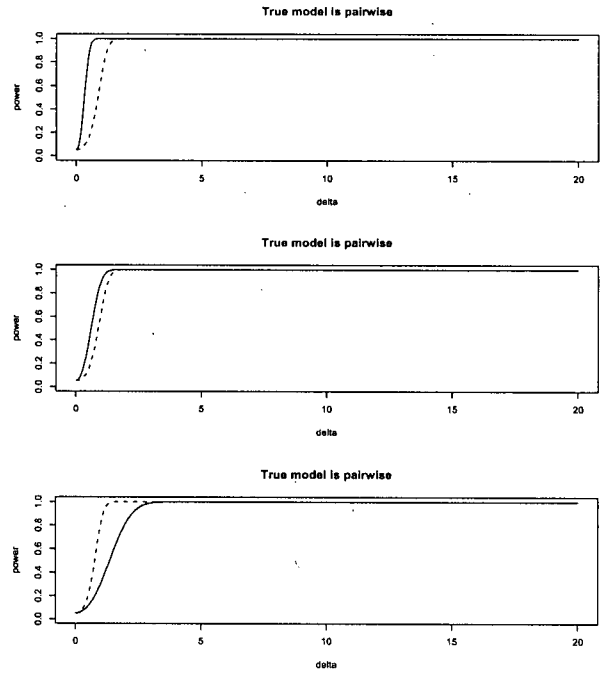
We also investigate what happens if the predictors are continuous. Suppose X_i's ($i = 1, \ldots, 9$) are i.i.d. following a log-normal distribution. That is for each i, $X_i = \exp(Z_i)$, where Z_i's are i.i.d. standard normal variables. Set $\beta_0 = 0, \boldsymbol{\beta}_M = 0.5\mathbf{1}_p$ and the variance of random error $\sigma^2 = 1$. Then we plot the asymptotic power against Δ (defined in the local alternatives) for different four subcases discussed before. Figure 3.8 shows a similar outcome to Figure 3.1, which is for binary predictors. That is, the diffuse interaction model is more powerful to detect interactions than pairwise interaction model, regardless of what the true structure of interaction is diffuse or not.

Figure 3.8: Power Curves with X_i's $\overset{i.i.d.}{\sim}$ Log-normal(0,1): the top panel with the true structure to be diffuse interaction model and bottom panel with the true structure to be pairwise interaction model. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting. In the bottom panel $\boldsymbol{\eta} = \mathbf{1}_{q\times 1}/\sqrt{q}$.



Similar to the previous example with binary predictors, we also change the direction of interactions when the pairwise interaction model is the true model. As shown in Figure 3.9, the diffuse interaction model lose the power to detect interactions as the "overall" strength of interaction gets weaker.

Figure 3.9: Different choices of η in Case II with continuous predictors: the top panel involves all predictor pairs interacting positively, the middle panel has only a few of predictor pairs interacting positively, and the bottom panel has more mixed directions of interactions. The lengths of η's in different panels are normalized to be 1. Solid lines denote power curves based on diffuse interaction model fitting and dashed lines denote power curves based on pairwise interaction model fitting.



3.3 Summary

By the two examples studied in section 3.2.2, we can see that diffuse interaction model is more powerful to detect the interactions no matter what the true interaction structure (which is postulated) is diffuse interaction or pairwise interaction. However, as the direction of true interactions among predictors is more mixed, the diffuse interaction model gets less powerful to detect interaction.

Chapter 4

MCMC algorithms for diffuse interaction models

4.1 Why MCMC?

We introduced a diffuse interaction model in Chapter 3, where it is proposed to be more powerful to detect interactions than pairwise interaction model by using only a single parameter to describe the interactions among numerous predictors. As a related point, the diffuse interaction model would be better for inferences on interactions. In this chapter, we are going to apply MCMC algorithms to implement the model fitting.

In this section, we enumerate reasons for using MCMC algorithms but not maximum likelihood estimation, which also may be feasible to be implemented.

First, in the diffuse interaction models, all the parameters β_j's (except the intercept β_0) and λ are designated to be positive. The statistical inference of constrained maximum likelihood estimates usually is more complicated. Standard asymptotic theory asserts that statistical inference regarding inequality constrained parameters does not require special techniques, because for a large enough sample there will always be a confidence interval at the selected level of confidence that avoids the constraint boundaries. Sufficiently large, however, can be quite large, in the cases when the true parameter values are very close to these boundaries. In practice, our finite samples may not be large enough for confidence intervals to avoid constraint boundaries, and this has implications

for all parameters in models with inequality constraints, even those that are not themselves constrained. Comparatively, MCMC sampler can automatically accommodate the constraints and yield appropriate interval estimate without extra efforts.

Second, sometimes the interested quantity may not be the parameters in the model directly and would be some complicated function of them. In particular, under diffuse interaction models, we apply the average effect idea to make inferences on interactions. Now the form of the first derivative of the regression function with respect to $x_j, j = 1, \ldots, p$ is somewhat intricate so that it takes efforts to approach the interval estimates for average effects based on standard MLE. However, the interval estimates can be effectively achieved based on the posterior samples of the parameters of diffuse interaction models.

Third, considering some extensions to the diffuse interaction model, the maximum likelihood estimates could be heavy in hand to calculate. For example, to relax the assumption that all the predictors interact in the same way, we could allow the predictors to be categorized into groups: within each group, the predictors interact in the same way. Here we assume that there is no overlapping among the predictors within different subgroups. That is, each predictor has an indicator variable denoting the group to which it belongs. The specific parametric form of this extended model will be shown in the coming section 4.2. Therefore, the parameter set is a mixture of continuous parameters (β_j's, the coefficients of predictors, and λ, the diffuse interaction parameter), and discrete parameters (the indicator variables associated with predictors). We have 2^p (or 3^p if three groups) different patterns of group allocation, hence the optimization procedure would be more complicated especially when p is large.

4.2 Details of MCMC algorithms

In Gustafson et al. (2005), an efficient MCMC sampler is developed for binary responses. Now, we apply a similar MCMC algorithm for continuous (normal) responses. We start with a simple case, that is, all the predictors interact in the same way.

4.2.1 One-group diffuse interaction models

In this subsection, we consider a one-group diffuse interaction model

$$Y|X = x \sim N\left(\beta_0 + \left\{\sum_{j=1}^{p}(\beta_j x_j)^{\lambda}\right\}^{1/\lambda}, \sigma^2\right), \tag{4.1}$$

where $\lambda > 0$ is the parameter accounting for interactions and $\beta_j \geq 0, x_j \geq 0$ for $j = 1, \ldots, p$.

We apply a *hybrid* MCMC (HY) approach, similar as that in Gustafson et al. (2005) to make inference. The reason for using this algorithm is to avoid the waste caused by the randomness introduced by the proposal of candidate value. Basically, the idea of HY is to incorporate derivative information of the target density and to suppress random walk behavior in the sampling simultaneously. Both of the strategies attempt to eliminate the inefficiency of *random walk*, which is commonly used in Metropolis-Hastings (MH) algorithms to generate candidate states. To explain, by using random walk, the direction of each movement about the target distribution is randomized. This can greatly increase the number of iterations required before achieving the equilibrium. The situation is getting even worse especially when the parameters involved in the target distribution are highly dependent to each other. More discussions in Gustafson et al. (2004) and Neal (1998). The pseudocode of the hybrid MCMC algorithm is provided in Appendix V.

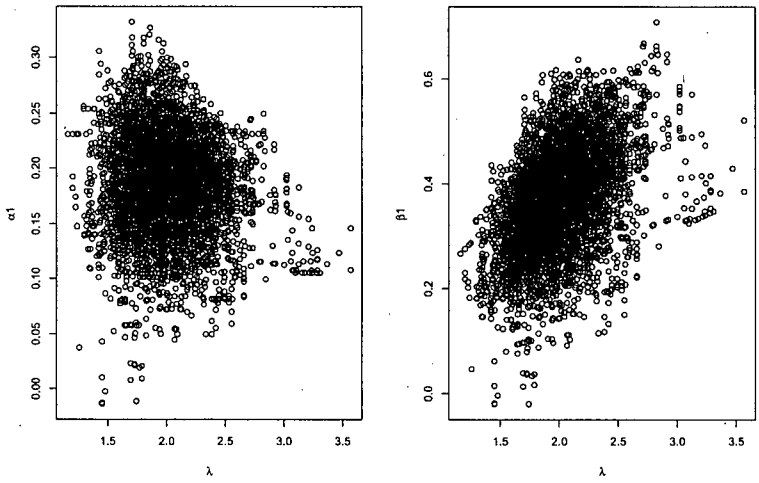
In our simulation study, the priors of the parameters are $\log \lambda \sim \mathrm{N}(0,\sigma_\lambda^2)$ with σ_λ^2=100;

$\beta_0 \sim \mathrm{N}(0,\sigma^2_{\beta_0})$ with $\sigma^2_{\beta_0}$=100; and $\beta_j \sim \mathrm{N}^+(0,\sigma^2_{\beta_j})$ with $\sigma^2_{\beta_j} = 100$ for $j = 1,\ldots,p$, where $\mathrm{N}^+(\mu,\sigma^2)$ denotes the $\mathrm{N}(\mu,\sigma^2)$ distribution truncated to non-negative values. Note that we also check the prior sensitivity by using informative priors with smaller hyper-parameters $\sigma^2_{\beta_j}, \sigma^2_\lambda$. The prior for σ^2 is inverse-gamma (a,b) with the shape parameter $a = 0.0001$ and the scale parameter $b = 0.0001$.

Set the number of predictors to be $p = 10$, the sample size to be n =2000. Generate the p binary predictor variables from Bernoulli(0.2) distribution independently. For simplicity, we set the true values of β_j's $(j = 0,\ldots,p)$ all equal to 0.5. The true value of λ is set to be 2, that is, all the predictors interact in antagonistic direction. Then for $i = 1,\ldots,n$,we generate the response variable Y_i from normal distribution with mean of $\beta_0 + \left\{\sum_{j=1}^p (\beta_j x_{ij})^\lambda\right\}^{1/\lambda}$ and variance $\sigma^2 = 1$. Note that in the model for binary responses, no variance component is involved. As a consequence, our algorithm here has one more step to update σ^2, compared to the MCMC algorithm used in Gustafson et al. (2005).

Following Hills and Smith (1992), the MCMC literature has considered changes in the parameterization of a model as a way to speed up convergence in a Gibbs Sampler. The general suggestion is to try to make the components as "independent" as possible. Thus we implement MCMC using the new parameters $(\boldsymbol{\alpha},\lambda)$, where $\alpha_0 = \beta_0, \alpha_j = \beta_j/\lambda, j = 1,\ldots,p$. From the scatter plot of posterior samples for λ and α_1 (left panel of Figure 4.1), a smaller correlation between reparametrized components can be clearly seen. That is, the reparametrization works well. Note that all the samples used in Figure 4.1 are based on the 30,000 iterations after 20,000 burn-in iterations.

Figure 4.1: Algorithm I: Posterior correlations for (λ, β_1) and (λ, α_1) respectively.



MCMC Algorithm I:

At the $t-$th iteration,

Step 1. For given $\lambda^{(t-1)}$, update from $(\beta_0^{(t-1)}, \alpha_1^{(t-1)}, \ldots, \alpha_p^{(t-1)})$ to $(\beta_0^{(t)}, \alpha_1^{(t)}, \ldots, \alpha_p^{(t)})$ as a block by using a hybrid MCMC.

Step 2. Update $\lambda^{(t-1)}$ by using a random walk Metropolis-Hastings to $\log(\lambda)$.

Step 3. Update $\sigma^{2(t-1)}$ via Gibbs sampler. Given the prior of σ^2 to be inverse-gamma (a, b), the posterior conditional distribution given all the other parameters is inverse gamma with shape parameter $a + n/2$ and scale parameter of $b + \text{RSS}/2$, where

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \beta_0^{(t-1)} - \left\{ \sum_{j=1}^{p} \{\beta_j^{(t-1)} x_{ij}\}^{\lambda^{(t-1)}} \right\}^{\frac{1}{\lambda^{(t-1)}}} \right)^2 .$$

Note that Step 2 can also be implemented by any other choices of prior for λ due to the generality of MH. However Step 3 needs conjugate prior setting for σ^2 because Gibbs sampler get samples from the corresponding full conditional distribution. It is actually

a particular choice of the proposal in MH leading to the acceptance ratio of 1.

The trace plots of MCMC outputs in Figure 4.2 shows that the above MCMC approach works well for the simulated data set under the specific setting of priors for parameters. In each panel of Figure 4.3, the true value, marked by the solid vertical line, is covered within the corresponding 95% equal-tailed credible interval. For some parameters though, the true value is somehow close to the lower/upper end of the credible interval.

Figure 4.2: Algorithm I: MCMC traceplots for $\beta_j, j = 0, 1, \ldots, p$ and λ with diffuse priors $\sigma^2_{\beta_0} = \sigma^2_{\beta_j} = \sigma^2_{\lambda} = 100$.

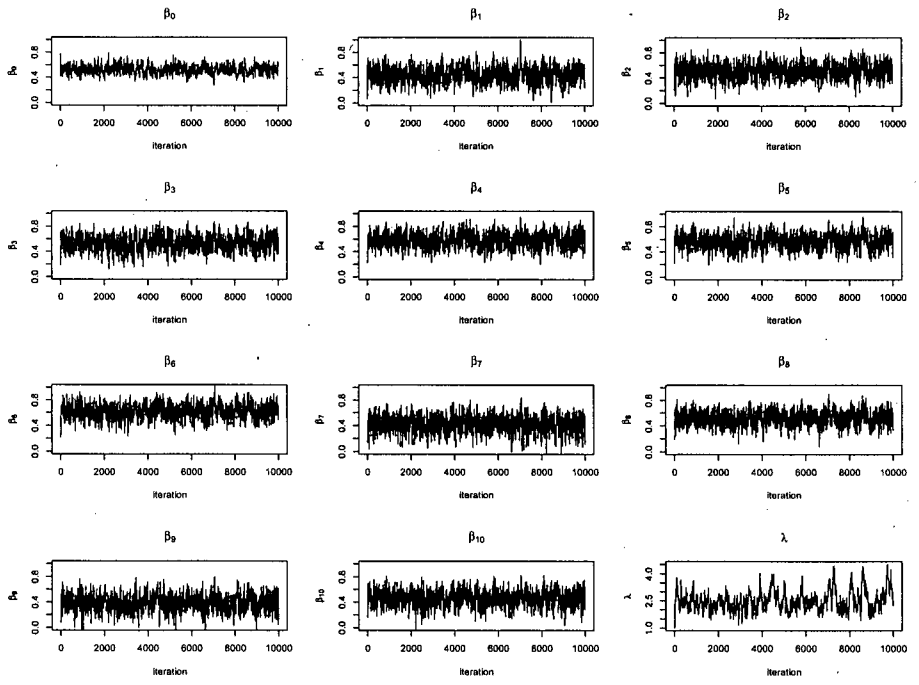
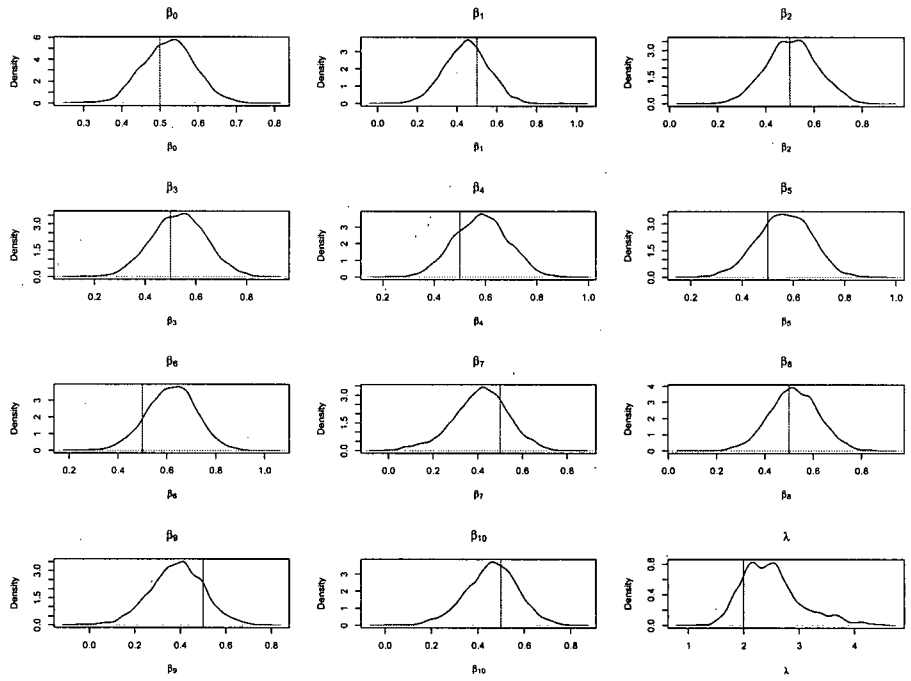


Figure 4.3: Algorithm I: Marginal posterior densities for $\beta_j, j = 0, 1, \ldots, p$, and λ with diffuse priors $\sigma^2_{\beta_0} = \sigma^2_{\beta_j} = \sigma^2_{\lambda} = 100$.

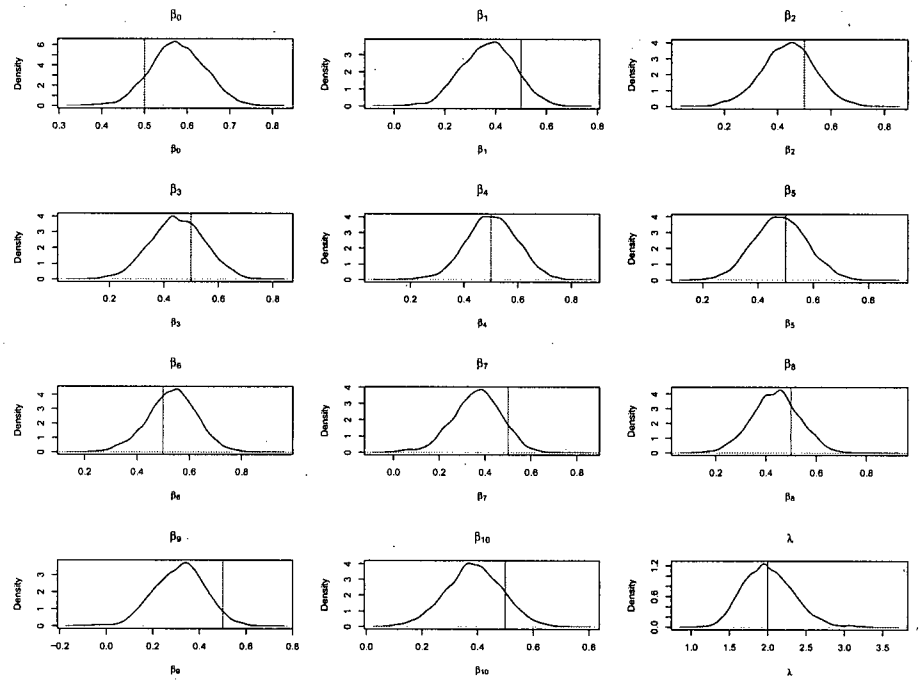


Remarks for the algorithm:

First, the step sizes used to produce candidate values in step 2 and 3 should be adjusted as sample size changes. Basically, we tune the step sizes to get relatively high acceptance rates, about 70% for step 2 and about 50% for step 3.

Second, when the priors of the parameters $\beta_j, j = 0, \ldots, p$ and λ are changed to be more informative (smaller variance), the outcomes do not change much. Comparing the density plots in Figure 4.3 and Figure 4.4, there is no serious difference between the distributions of posterior samples from different priors. This point is different from Gustafson et al. (2005), where the prior of β_0 does have influence on the Bayesian inference.

Figure 4.4: Algorithm I: Posterior densities for $\beta_j, j = 0, 1, \ldots, p$ and λ with informative priors $\sigma^2_{\beta_0} = \sigma^2_{\beta_j} = \sigma^2_{\lambda} = 0.4$.



Third, as suggested in Gustafson et al. (2005), the algorithm above can be easily extended to a more general case of no positive constraint of the sign of β_j, replacing $x_j\beta_j$, in model (4.1), by $g(x_j, \beta_j)$, where

$$g(x, \beta) = \begin{cases} |\beta|x; & \beta > 0, \\ |\beta|(1-x); & \beta < 0. \end{cases}$$

To remove the positive constraint on the β_j's, we need the assumption that X_j's are bounded within [0,1]. In practice, transformation of predictors may be necessary. In the following, for simplicity, we always focus on β_j's with positive constraints and leave the extension without the constraints to the future work.

4.2.2 Two-group diffuse interaction models

As one possible extension mentioned but not pursued in Gustafson et al. (2005), we study a more complicated model in this subsection. Say

$$Y|X = x \sim N\left(\beta_0 + \sum_{j\in \text{ADD}} (\beta_j x_j) + \left\{\sum_{j\in \text{ANT}} (\beta_j x_j)^{\lambda}\right\}^{1/\lambda}, \sigma^2\right), \qquad (4.2)$$

where ADD consists of indices of the predictors belonging to the additive group and ANT is the set of indices of the predictors interacting in antagonistic direction. (For sure, if we suspect that some of the explanatory variables interact in synergistic way, we would replace the antagonistic group by synergistic group by setting $\lambda < 1$.) In this subsection, we take λ to be fixed. Let $\mathbf{I} = \{I_k\}_{k=1,\ldots,p}$, where

$$I_k = \begin{cases} 1; & k \in \text{ADD}, \\ 2; & k \in \text{ANT}. \end{cases}$$

Therefore, the parameters in this model now are $(\{\beta_j\}_{j=0}^{p}, \{I_j\}_{j=1}^{p}, \sigma^2)$. So that we have $p-1$ more parameters,with λ fixed, to update than that in the simple case (4.1).

To explore the MCMC approach to model (4.2), we do a simulation study as the following. As before, set the number of predictors p =12, the sample size n =2000. Generate each predictor variable from a Bernoulli(0.2) distribution. For simplicity, we set the true values of β_j's (j=1,...,p) all equal to 0.7 and $\beta_0 = 0$. The true value of λ is set to be 2, that is, all the predictors involved in interaction group interact in antagonistic direction. The first six predictors $X_1 \ldots X_6$ are set to belong to ADD group while the others belong to ANT group. The starting values of the indicator variables are set to be the worst possible case, that is, each I_k is set to be the opposite value of the corresponding true value. Then we generate the response variable $Y_i, i = 1, \ldots, n$ from a

normal distribution with mean of

$$\beta_0 + \sum_{j\in \mathrm{ADD}} (\beta_j x_{ij}) + \left\{ \sum_{j\in \mathrm{ANT}} (\beta_j x_{ij})^{\lambda} \right\}^{1/\lambda},$$

and variance σ^2 =1. For each $j \in \{1,\ldots,p\}$, the prior of I_j is uniform distribution over $\{1,2\}$. The prior of β_0 is $N(0,\sigma^2_{\beta_0})$ with $\sigma^2_{\beta_0}$=100. For $j \in 1,\ldots,p$, the prior of β_j is $\beta_j \sim \mathrm{N}^+(0,\sigma^2_{\beta_j})$ with $\sigma^2_{\beta_j} = 100$ for $j = 1,\ldots,p$, where $\mathrm{N}^+(\mu,\sigma^2)$ denotes the $\mathrm{N}(\mu,\sigma^2)$ distribution truncated to non-negative values.

MCMC Algorithm II:

Step 1. Update the coefficients in the additive group, denoted by $\boldsymbol{\beta}_{\mathrm{ADD}}$, together with β_0.

Given $\mathbf{I}$ and coefficients of predictors in ANT group, denoted by $\boldsymbol{\beta}_{\mathrm{ANT}}$, it is easy to verify that the posterior of $(\beta_0, \boldsymbol{\beta}'_{\mathrm{ADD}})'$ is proportional to

$$\exp\Big\{-\sigma^{-2}\left((\beta_0,\boldsymbol{\beta}'_{\mathrm{ADD}})' - (\mathbb{X}_1'\mathbb{X}_1 + D)^{-1}\mathbb{X}_1'\mathbb{Y}_1\right)' (\mathbb{X}_1'\mathbb{X}_1 + D)\left((\beta_0,\boldsymbol{\beta}'_{\mathrm{ADD}})' - (\mathbb{X}_1'\mathbb{X}_1 + D)^{-1}\mathbb{X}_1'\mathbb{Y}_1\right)\Big\}\, \mathrm{I}\{\boldsymbol{\beta} > \mathbf{0}\}.$$

It implies that $(\beta_0, \boldsymbol{\beta}'_{\mathrm{ADD}})'$ follows a multivariate normal $N\left(\{\mathbb{X}_1'\mathbb{X}_1 + D\}^{-1}\mathbb{X}_1'\mathbb{Y}_1, \{\mathbb{X}_1'\mathbb{X}_1 + D\}^{-1}\sigma^2\right)$ truncated by $\boldsymbol{\beta}_{\mathrm{ADD}} > \mathbf{0}$, where

$$\begin{aligned}
\mathbb{X}_1 &= (\mathbf{1}, \mathbb{X}_{\mathrm{ADD}}),\\
\mathbb{Y}_1 &= \mathbb{Y} - \mathbb{Y}_2,\\
\mathbb{Y}_2 &= \left\{\mathbb{X}_{\mathrm{ANT}}{}^{\lambda}\boldsymbol{\beta}^{\lambda}_{\mathrm{ANT}}\right\}^{1/\lambda},\\
D &= \sigma^2 \mathrm{diag}\{\sigma^{-2}_{\beta_0}, \sigma^{-2}_{\beta_j}, \ldots, \sigma^{-2}_{\beta_j}\}.
\end{aligned}$$

Note that $\mathbb{X}_{\mathrm{ADD}}$ denotes the design matrix with column vectors of the values of the

explanatory variables in the additive group. Analogously, $\mathbb{X}_{\mathrm{ANT}}$ is the design matrix with column vectors of the observations of predictors in the antagonistic group.

Step 2. Update the intercept together with the coefficients in the antagonistic group.

Subtracting the contribution of predictors in the additive group, we can now pretend there is only one group and all predictors interact antagonistically. To be specific, we apply Algorithm I, devised for the one-group diffuse interaction model, to $\left(\mathbb{Y}', \mathbb{X}_{\mathrm{ANT}}, \boldsymbol{\beta}_{\mathrm{ANT}}, \beta_0\right)$ where

$$\mathbb{Y}' = \mathbb{Y} - \mathbb{X}_{\mathrm{ADD}}\boldsymbol{\beta}_{\mathrm{ADD}}.$$

Step 3. Update (β_k, I_k) together. We generate the candidate values for (β_k, I_k), denoted by (β_k^*, I_k^*), as the following:

$$\begin{aligned} I_k^* &= 3 - I_k, \\ \log(\beta_k^*) &\sim N\left(\log(\beta_k^0), \tau^2\right). \end{aligned}$$

That is, the candidate value of I_k is just opposite to the current value. Say $I_k = 1$, then $I_k^* = 2$. Using random walk on the log scale of β_k to fulfil the positive constraints in model (4.2).

There could be more than one criteria to determine β_k^0 and two of them are discussed after remarks on Algorithm II. At this stage, let's write β_k^0 in a general way as

$$\beta_k^0 = h_{I_k, I_k^*}(\beta_k),$$

where h is a deterministic function with reversibility property, that is, h is 1-1 mapping

on $[0, \infty)$ and $h^{-1} = h$, where h^{-1} is the inverse function of h. Thus, we have

$$h_{I_k^*, I_k}\left(h_{I_k, I_k^*}(\beta_k)\right) = \beta_k.$$

Then, the acceptance ratio is

$$\frac{\pi(I^*, \beta^*)}{\pi(I, \beta)} \frac{P(I_k|I_k^*)\phi\left(\frac{\log(\beta_k) - \log(h_{I_k^*, I_k}(\beta_k^*))}{\tau}\right)}{P(I_k^*|I_k)\phi\left(\frac{\log(\beta_k^*) - \log(h_{I_k, I_k^*}(\beta_k))}{\tau}\right)} \frac{\beta_k^*}{\beta_k}, \tag{4.3}$$

where $\pi(\cdot)$ is the joint posterior density of I and β for given σ^2 and ϕ is the density function of standard normal distribution. According to Step 3 in Algorithm II $P(I_k|I_k^*) = P(I_k^*|I_k) = 1$. Remarks:

First, the fraction of β_k^*/β_k in (4.3) comes out from the Jacobian of transformation because the proposed value is generated on the log scale of β_k.

Second, if $\tau^2 = 0$, then the proposed value of β_k is exactly β_k^0. However, the acceptance ratio now becomes somewhat intractable because it takes effort to figure out

$$\lim_{\tau \to 0} \frac{\phi\left(\frac{\log(\beta_k) - \log\left(h_{I_k^*, I_k}(\beta_k^*)\right)}{\tau}\right)}{\phi\left(\frac{\log(\beta_k^*) - \log\left(h_{I_k, I_k^*}(\beta_k)\right)}{\tau}\right)} = \frac{\phi\left(-\epsilon\beta_k^0 h'_{I_k^*, I_k}(\beta_k^0)/h_{I_k^*, I_k}(\beta_k^0)\right)}{\phi(\epsilon)}, \tag{4.4}$$

where ϵ follows standard normal distribution. Moreover, tuning the size of τ to be larger helps to speed up the convergence of samples for β_j's.

In the following, we demonstrate two ways to determine β_k^0, i.e., $h_{I_k, I_k^*}(\beta)$.

Proposal 1 : Choose β_k^0 to keep the average effect of X_k unchanged. Since all X_j's are binary in the simulation study, based on Definition 1, think of the average effect for

X_k as

$$\mathrm{E}\,(Y|X_k = 1, \mathbf{X}_{(k)}) - \mathrm{E}\,(Y|X_k = 0, \mathbf{X}_{(k)}).$$

Therefore, we have

$$\delta_k = \begin{cases} \beta_k, & \text{if } k \in \mathrm{ADD}; \\ \mathrm{E}\,\{\beta_k^\lambda + Z\}^{1/\lambda} - \mathrm{E}\,\{Z^{1/\lambda}\}, & \text{if } k \in \mathrm{ANT}, \end{cases}$$

where

$$Z = \sum_{j \in \mathrm{ANT}_{-\{k\}}} (\beta_j X_j)^\lambda.$$

Let $z_i = \sum_{j \in \mathrm{ANT}_{-\{k\}}} (\beta_j x_{ij})^\lambda$. If $I_k = 1$, then $I_k^* = 2$, and β_k^0 is the solution to

$$n^{-1} \sum_{i=1}^{n} \left\{ {\beta_k}^\lambda + z_i \right\}^{1/\lambda} - n^{-1} \sum_{i=1}^{n} z_i^{1/\lambda} = \beta_k. \tag{4.5}$$

If $I_k = 2$, then $I_k^* = 1$, and β_k^0 is the solution to

$$\beta_k = n^{-1} \sum_{i=1}^{n} \{\beta_k^\lambda + z_i\}^{1/\lambda} - n^{-1} \sum_{i=1}^{n} z_i^{1/\lambda}. \tag{4.6}$$

In above two equations, the left side is the estimated average effect of X_k after change of I_k to I_k^*, while the right side is the estimated average effect before the change. Here estimated average effect is the sample mean of average effect evaluated at each realization of $\mathbf{X}_{(k)}$. The reason for this proposal is that average effect estimate tends to be more robust, as discussed in Gustafson et al. (2005).

Proposal 2 : Choose β_k^0 to make $\alpha_k (= \beta_k / \lambda_{I_k})$ unchanged. (Note that $\lambda_1 = 1, \lambda_2 = \lambda$.) That is,

$$\beta_k^0 = \begin{cases} \lambda \beta_k, & \text{if } I_k = 1, \\ \lambda^{-1} \beta_k, & \text{if } I_k = 2. \end{cases}$$

Based on the following plots of the MCMC output, Figure 4.5 and Figure 4.6, using Proposal 1, the algorithm works well. For a couple of the β_j's, though the true value falls close to tail of the density plot.

Figure 4.5: Algorithm II: Using Proposal 1, MCMC trace plots for $\beta_j, j = 1, \ldots, p$ with diffuse priors $\sigma^2_{\beta_0} = \sigma^2_{\beta_j} = 100$.

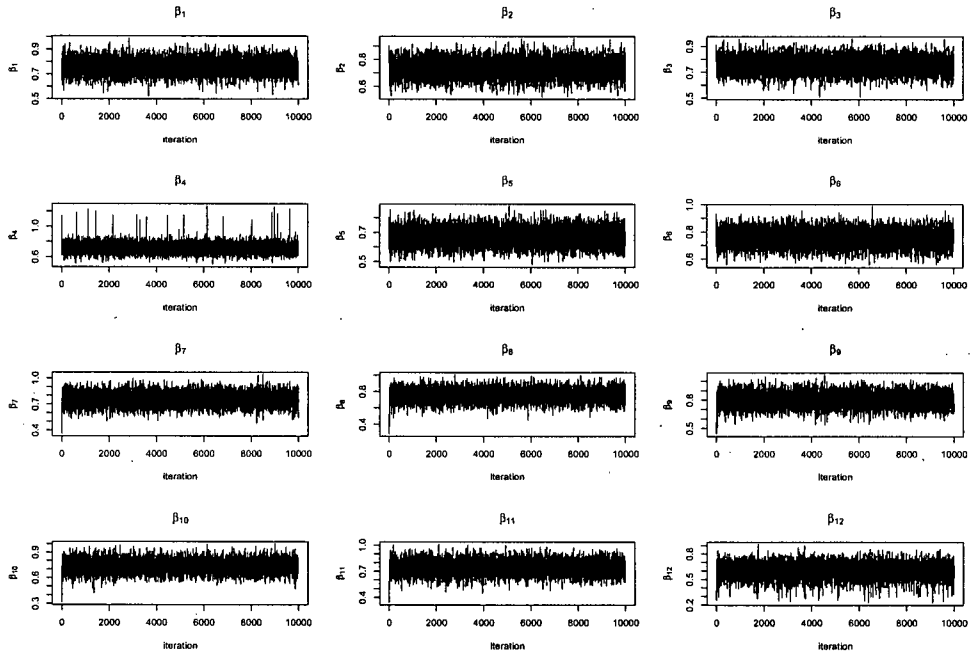
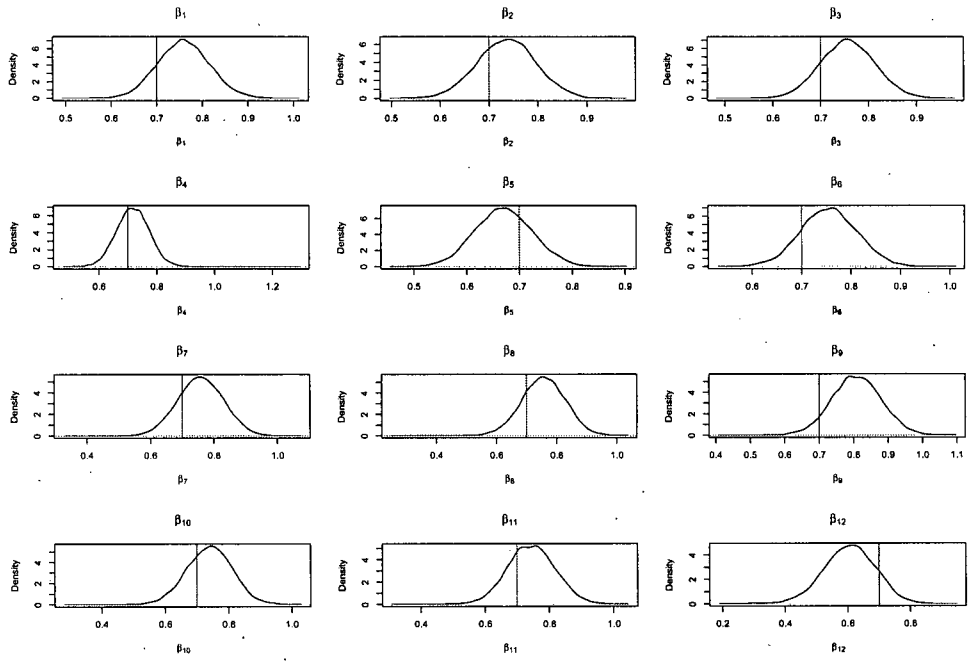


Figure 4.6: Algorithm II: Using Proposal 1, posterior densities for $\beta_j, j = 1, \ldots, p$ with diffuse priors $\sigma^2_{\beta_0} = \sigma^2_{\beta_j} = 100$.



To compare the performances of samples under different proposals in Step 3, we plot the number of correct group allocations, since the true group allocation is known in the simulation study. Figure 4.7 illustrates that Proposal 1, that is, solving for β_k^0 by keeping the average effect unchanged, gives the right group allocations only after a couple of iterations. However it takes many more iterations for the samples under Proposal 2, where β_k^0 is solved by keeping α_k unchanged. To make this point more clear, we also plot the autocorrelation coefficients up to lag 40 for MCMC samples under different proposals, as shown in Figures 4.8 and 4.9 respectively. In Figure 4.9, for some β_j's $(\beta_6 - \beta_{12})$, there is still some dependence even for lag 40. While in Figure 4.8, for all β_j's, the serial dependence drops close to zero after a small number of lags. It implies that the convergence rate under Proposal 1 is faster than that under Proposal 2.

Figure 4.7: Algorithm II: Comparison of two proposals: Number of correct group allocations based on posterior samples for I.

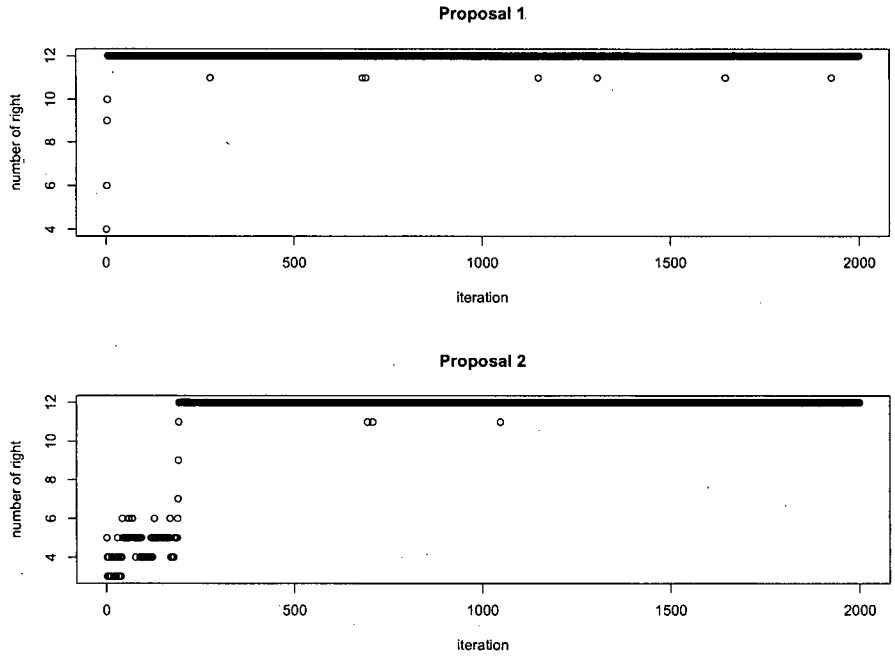


Figure 4.8: Algorithm II: With Proposal 1, the autocorrelation curves for posterior samples of $\beta_j, j = 1, \ldots, p$ respectively.

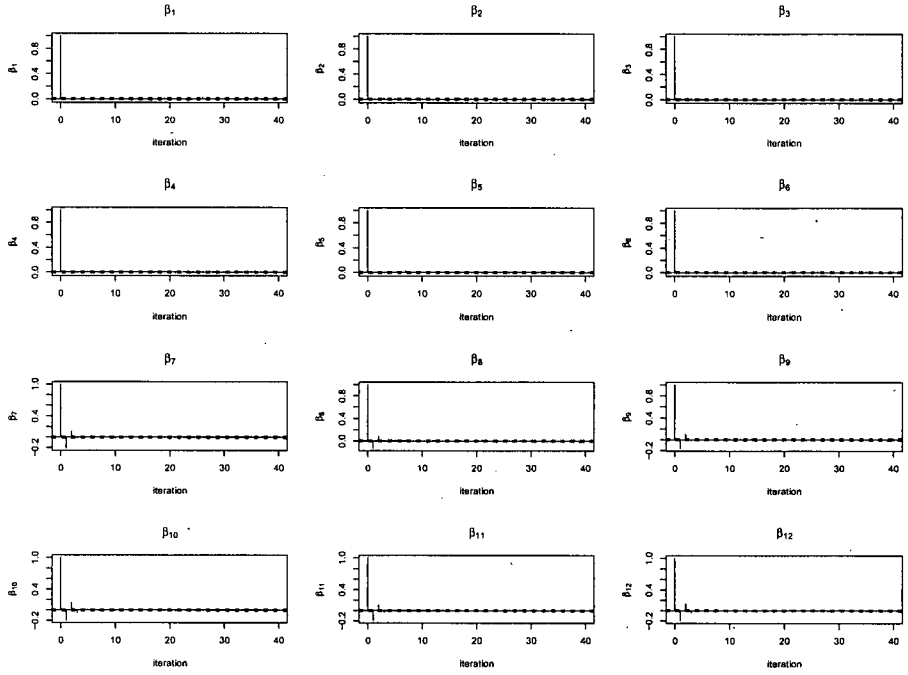
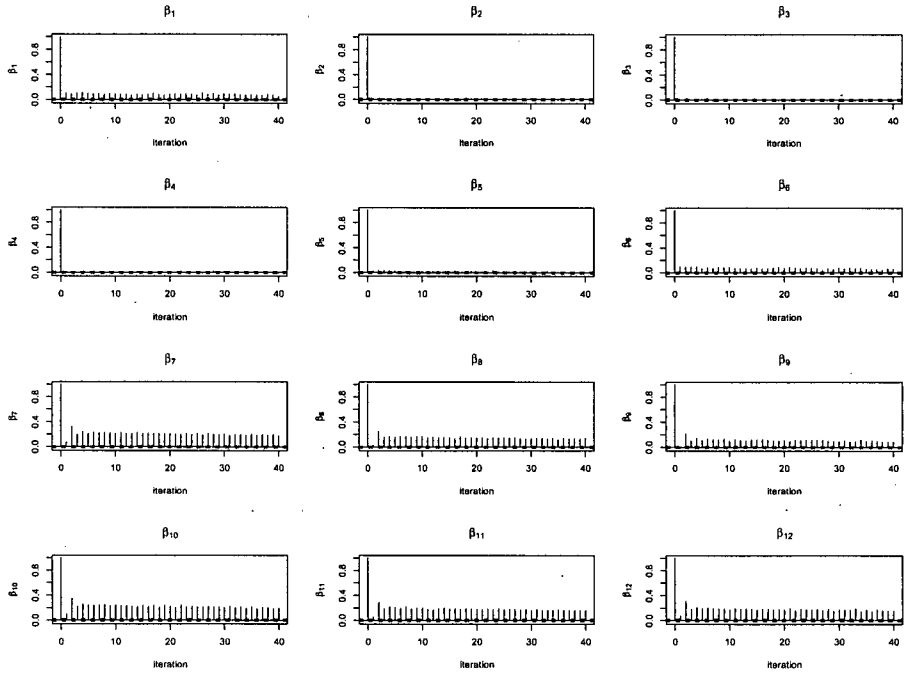


Figure 4.9: Algorithm II: With Proposal 2, the autocorrelation curves for posterior samples of $\beta_j, j = 1, \ldots, p$ respectively.



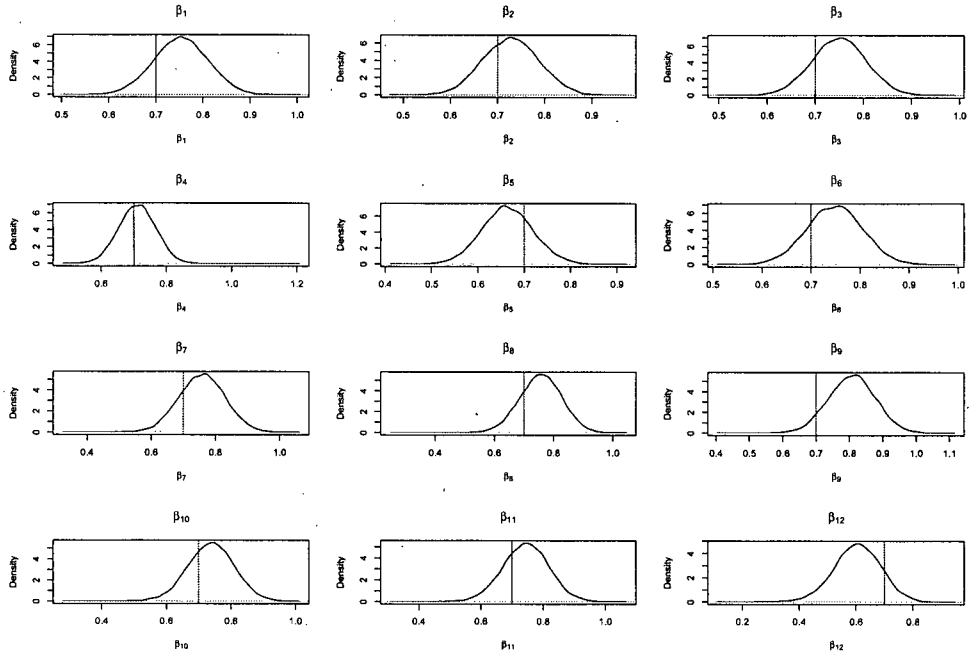
Remarks on MCMC algorithm II:

First, in step 3, we only allow the movement within the parameter space such that at least two predictors fall within the antagonistic/synectic interaction group. The reason of the constraint for now is that the terminology interaction has real meaning only when at least two predictors are involved. We could remove this constraint later and more would be discussed in the future work.

Second, the priors of β_j's do not strongly affect the posterior distribution. Change the value of hyper-parameter $\sigma^2_{\beta_j}$ to be 0.4, which is rather informative/narrow compared with the previous value 100. There is not much difference between the density plots in Figure 4.6, based on diffuse priors and those from the informative prior shown in Figure 4.10. In addition, for both kinds of priors, the posterior samples for $I_j, j = 1, \ldots, p$ have

high frequency of agreement with the true values.

Figure 4.10: Algorithm II: With Proposal 1, the posterior densities of samples from with informative priors :$\sigma^2_{\beta_0} = \sigma^2_{\beta_j} = 0.4$.



Third, the true value of $\beta_j (j \in \text{ANT})$ does have influence on the posterior information about group structure. Bigger β_j's ($j \in$ ANT) lead to more posterior concentration around the true value of I, which governing group allocation. While smaller β_j's seems to lead to less posterior concentration around the true group structure. Consider an extreme case that all the $\beta_j, j \in$ ANT are very small, say pretty close to 0, hence the conditional expectation of response variable given all explanatory variables are approximately equal to $\beta_0 + \sum_{j \in \text{ADD}} X_j \beta_j$. In other words, the diffuse interaction model is reduced to an additive model with explanatory variables of $X_j, j \in$ ADD. [1]

Let's make this point clearer by looking at a simulation study. Set the true value of

[1] In fact, if β_j ($j \in$ANT) are too small, the numerical solution of (4.6) would be zero, which cause difficulty in random walk on log scale of β_j.

each $\beta_j (i \in \text{ANT})$ to be 0.4, while the true value of coefficients in additive group is 0.7, same as the previous simulation study. Checking the posterior sample of each indicator variable for the first 10000 iterations, we get the following table. Clearly, as shown in Figure 4.11 and Table 4.1 as well, we get more posterior mass on wrong group allocations with smaller value of β_j's, but less with larger β_j's.

Figure 4.11: Algorithm II: Comparison of number of correct group allocations based on posterior samples for I with different values of $\beta_j (j \in \text{ANT})$: top panel with smaller value (=0.4), and bottom panel with bigger value (=0.7).

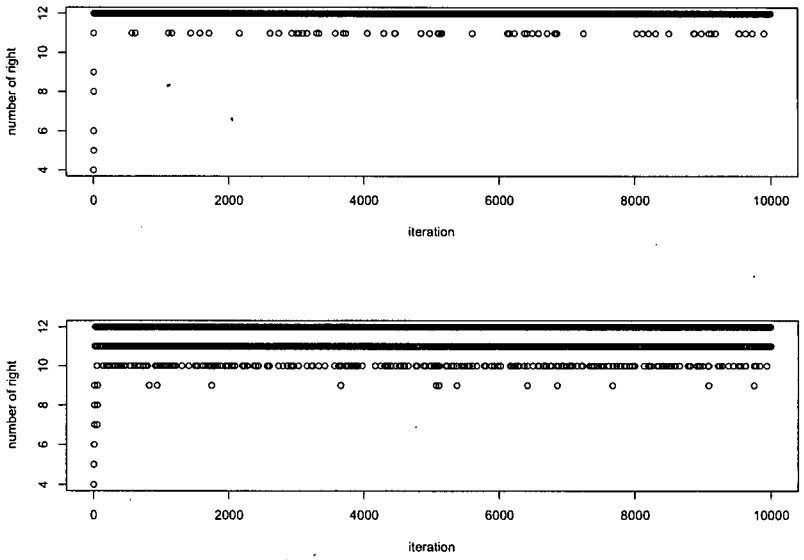


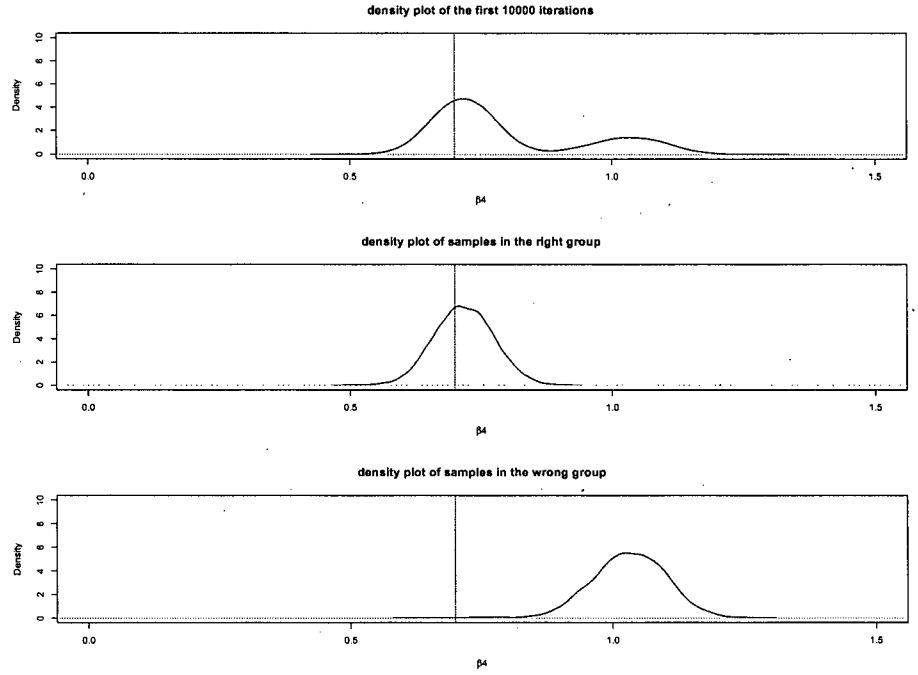
Table 4.1: Algorithm II: Frequency table based on posterior samples for each component of I with different values of $\beta_j, j \in \text{ANT}$.

β_j=0.4	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	1.00	1.00	1.00	.73	1.00	1.00	0.01	0.01	0.00	0.00	0.00	.11
2	0.00	0.00	0.00	.27	0.00	0.00	0.99	0.99	1.00	1.00	1.00	.89
β_j=0.7	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	1.00	1.00	1.00	1.00	1.00	1.00	0.00	0.00	0.00	0.00	0.00	0.00
2	0.00	0.00	0.00	0.00	0.00	0.00	1.00	1.00	1.00	1.00	1.00	1.00

Fourth, there is high posterior dependence between I_k and β_k for each k. The blocking of (I_k, β_k) in Step 3 is motivated by this fact. We tried to update I_k and β_k separately (for each k), and found the sampler got stuck so that we could not move efficiently around the whole parameter space. The reason is that in the separate update case, posterior distribution of I_k for given β_k highly concentrates on one certain value and vice versa. In other word, the update of β_k and the update of I_k do affect each other a lot. Therefore, we should block the two parameters.

In the simulation study for MCMC Algorithm II, changing the true values of $\beta_j, j \in$ ANT to be 0.4 while keeping $\beta_j, j \in$ ADD unchanged, Figure 4.12 depicts the dependence between I_k and β_k more clearly. The top panel is the unconditional posterior density plot of posterior samples for β_4. The middle panel is the posterior density plot of $\beta_4|I_4 = 1$, that is the posterior sample of β_4 given posterior samples for I_4 equal to 1. The bottom panel is the posterior density plot of samples for $\beta_4|I_4 = 2$. Clearly the modes of the samples within different groups are different by the latter two plots. Due to the considerable distance between the two modes, we get a bi-mode posterior density curve estimated by all the samples of β_4, as indicated by the top panel.

Figure 4.12: Algorithm II: MCMC output of β_4 with true value set to be 0.4: the top panel is the posterior density for entire sequence of 10000 iterations, the middle is the posterior density for subsequence $\beta_4|I_4 = 1$, and the bottom is the posterior density for subsequence $\beta_4|I_4 = 2$.



Fifth, $\hat{I}_k$, the estimate of indicator variable I_k may be determined by the one having higher frequency to be sampled. Here we should be careful about the Bayesian inference to β_k's. For each β_k, the point estimate and 95% equal-tailed credible interval could be obtained based on the samples belonging to the $\hat{I}_k$ group only, while not taking into account the whole sample sequence of β_k. To be more explicit,

$$\begin{aligned}
\hat{\beta}_k &= \frac{\sum_{t:I_k^{(t)}=\hat{I}_k} \beta_k^{(t)}}{\left|\left\{t : I_k^{(t)} = \hat{I}_k\right\}\right|}, \\
L_{0.025} &= 2.5\% \text{ quantile of} \left\{\beta_k^{(t)} : I_k^{(t)} = \hat{I}_k\right\} \\
U_{0.025} &= 97.5\% \text{ quantile of} \left\{\beta_k^{(t)} : I_k^{(t)} = \hat{I}_k\right\}
\end{aligned}$$

where $\beta_k^{(t)}, I_k^{(t)}$ are the t-th samples for β_k, I_k respectively, and $|\{A\}|$ means the number of element in the set A, and $L_{0.025}$ /$U_{0.025}$ is the lower/upper bound of a 95% credible interval.

Or there is another possibility to make the inference based all the samples since the true I is unknown in practice. In such a case, we could imagine that if $\hat{I}_k$ with much higher frequency than other sampled values of I_k, the estimate of β_k based on samples from $\hat{I}_k$ group only should be close to that based on samples from all possible groups. The credible intervals may be close to each other as well. On the other hand, if $I_k = \hat{I}_k$ does not have a superior frequency, then the estimate based on all samples for β_k would be somewhat different that from the $\hat{I}_k$ group only. The credible interval in the former case would be wider.

4.3 Discussion

Other than the two diffuse interaction models discussed above, we could consider more complicated models.

1. As a direct extension to model (4.2), we could allow λ to vary as well. However, a corresponding direct extension to MCMC algorithm II is not trivial to achieve. We are aware of the fact that $(I_k, \beta_k, \lambda_k)$ are closely associated together for each $k \in \{1, \ldots, p\}$. Therefore we need to devise a good joint proposal for the triple set. To make use of the previous algorithm for two groups with λ known, we propose the following algorithm.

MCMC Algorithm III:

Step 1. Update $\boldsymbol{\beta}_{\mathrm{ADD}}$, coefficients in the additive group, together with intercept β_0.

Step 2. Update $\boldsymbol{\beta}_{\mathrm{ANT}}$, coefficients in the antagonistic group, together with intercept β_0.

(Note that the two steps have exactly the same structure as in that Algorithm II.)

Step 3. For $k = 1, \ldots, p$, update (β_k, I_k) as a block in the same way used in Algorithm II, that is, I_k^* is the opposite to the current value of I_k and β_k^* is proposed by adding noise to β_k^0, which leaves the average effect of X_k unchanged for given value λ.

Step 4. Update λ, for given other parameters, by using MH update of $\log(\lambda)$.

Step 5. Update σ^2 via Gibbs sampler.

From the following trace plot of λ samples, Figure 4.13, we can see that the MCMC samples for λ have not achieved the stationary distribution within the first 10000 iterations. We increased the number of iterations and the MCMC still do not mix very well for λ. We may adjust the stepsize for the update of λ or better figure out other ways to propose the candidates of λ more efficiently. Moreover, according to the autocorrelation plot of samples for λ, the convergence speed is really slow since the autocorrelation of lag 40 is still rather large. It is similar for the posterior samples for some β_j's, as displayed in Figure 4.16. This implies that Algorithm III still need to be improved in terms of speeding up the convergence. But still the algorithm seems promising since we get eight out of twelve correct estimates of indicator variables in the simulated example, as shown in Table 4.2.

Figure 4.13: Algorithm III: Trace plots for $\beta_j, j = 1, \ldots, p$ based on two-group diffuse interaction model with λ unknown, respectively.

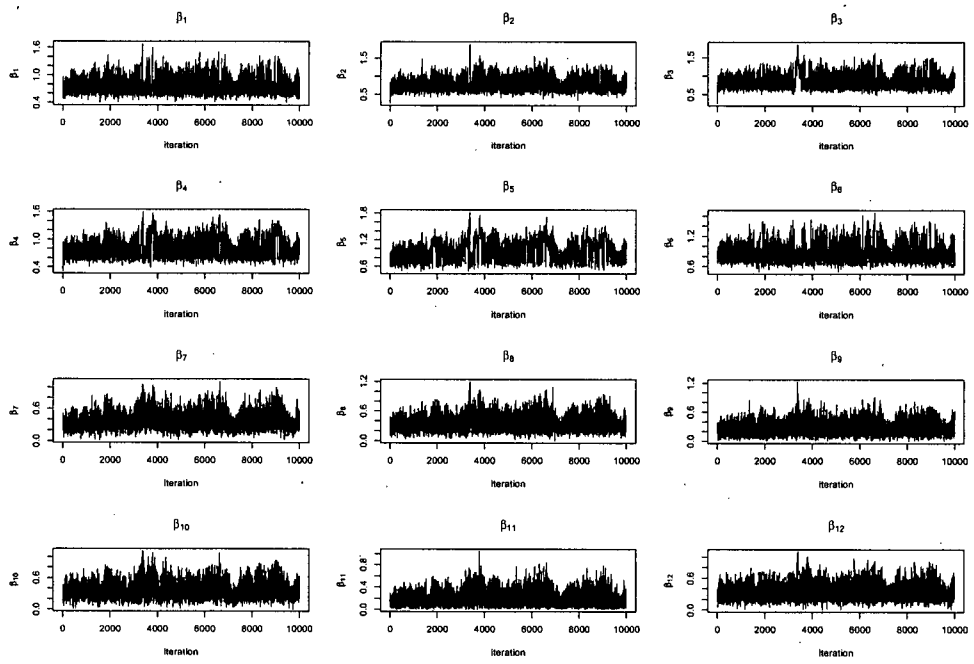


Figure 4.14: Algorithm III: Marginal posterior densities for $\beta_j, j = 1, \ldots, p$ based on two-group diffuse interaction model with λ unknown, respectively.

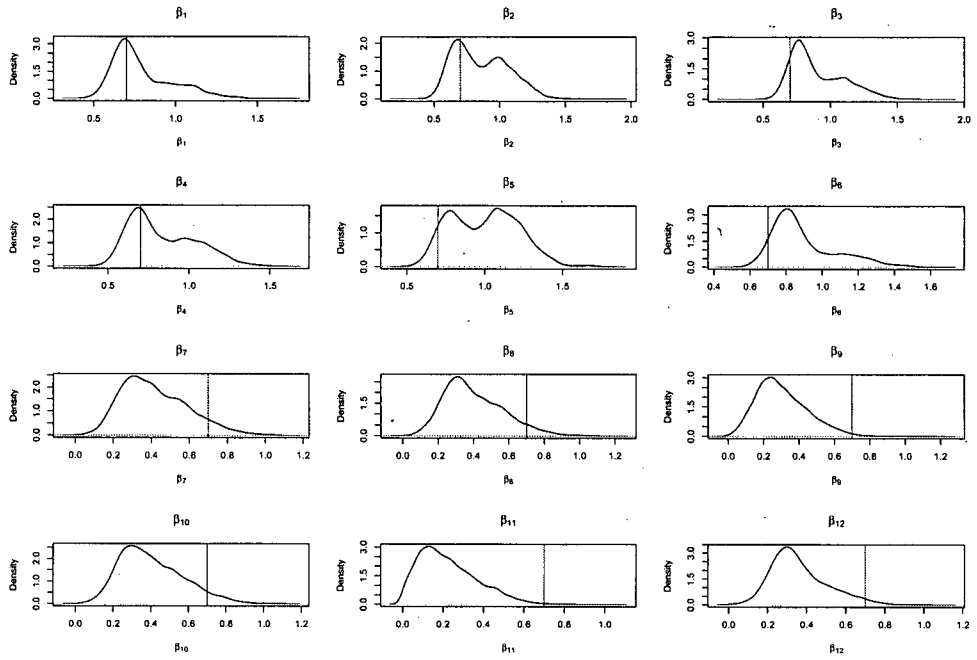


Figure 4.15: Algorithm III: Plots of posterior samples for λ, the top panel is the trace plot, the middle is the posterior density plot, and the bottom is the autocorrelation curve.

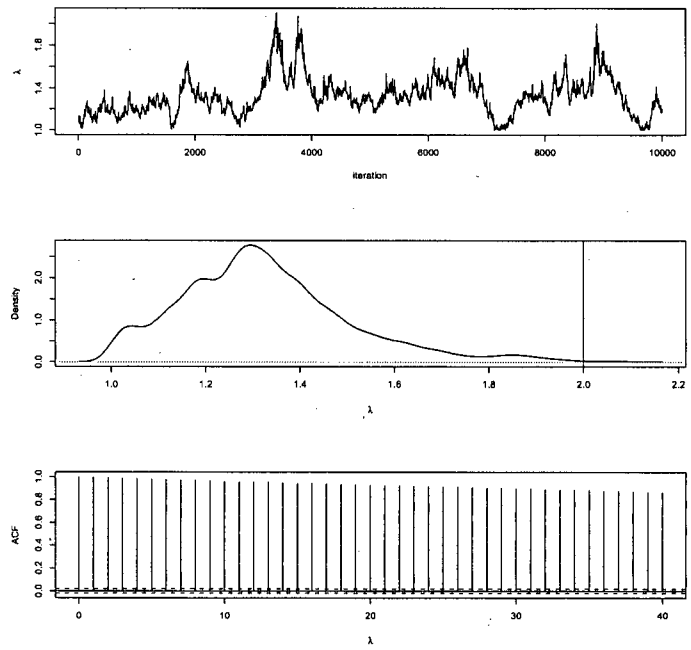
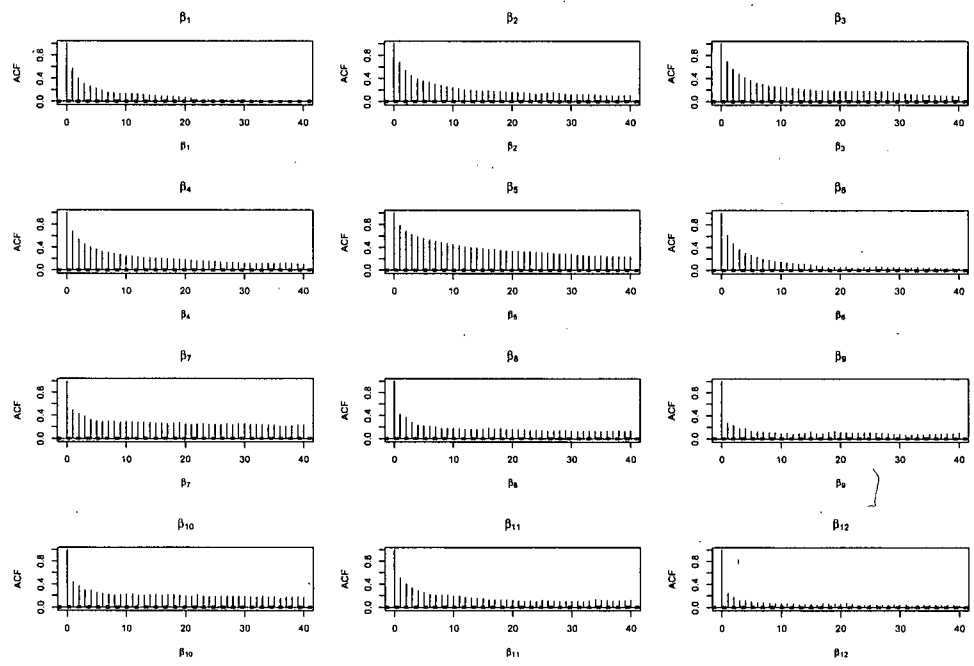


Table 4.2: Algorithm III: Frequency table based on posterior samples for each component of I under two-group diffuse interaction model with λ unknown.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	0.63	0.40	0.55	0.48	0.28	0.66	0.34	0.40	0.46	0.39	0.34	0.54
2	0.37	0.60	0.45	0.52	0.72	0.34	0.66	0.60	0.54	0.61	0.66	0.46
true value	1	1	1	1	1	1	2	2	2	2	2	2

Figure 4.16: Algorithm III: Autocorrelation curves for $\beta_j, j = 1, \ldots, p$ based on two-group diffuse interaction model with λ unknown, respectively.



2. Including more groups into the model, say additive, synergistic and antagonistic group, with the corresponding λ's known, for instance,$\lambda_1 = 1, \lambda_2 = 4/5, \lambda_3 = 5/4$ respectively. (The reason to choose λ_2/λ_3 close to 1 is stated later.) In this scenario, there are three possible values of each I_k. Hence we cannot just flip the current value of I_k when in the joint proposal for update of (I_k, β_k). We need more complicated jumping rule for the update of I_k. To apply the previous algorithm for two groups to the current situation with three groups, an easy-to-extend algorithm is as below.

MCMC Algorithm IV:

Step 1. Update the coefficients in the additive group together with intercept.

Step 2. Update the coefficients in the synergistic group together with intercept.

Step 3. Update the coefficients in the antagonistic group together with intercept.

Step 4. For $k \in \text{SYN} \bigcup \text{ADD}$, update (β_k, I_k) by applying step 3 in Algorithm II to

$$\mathbb{Y} - \left\{ \mathbb{X}_{\text{ANT}}{}^{\lambda_3} \beta_{\text{ANT}}^{\lambda_3} \right\}^{1/\lambda_3},$$

where $\mathbb{X}_{\text{ANT}}$ is the design matrix with column vectors of the observations of predictors in the antagonistic group. That is, by subtracting the contribution of the antagonistic group from the response variable, we could pretend that we have only two groups, additive and synergistic group.

Step 5. For $k \in \text{ANT} \bigcup \text{ADD}$, update (β_k, I_k) by applying step 3 in Algorithm II to

$$\mathbb{Y} - \left\{ \mathbb{X}_{\text{SYN}}{}^{\lambda_2} \beta_{\text{SYN}}^{\lambda_2} \right\}^{1/\lambda_2}.$$

where $\mathbb{X}_{\text{SYN}}$ is the design matrix with column vectors of the observations of predictors in the synergistic group. Again, by taking away the given contribution of the synergistic group, we may imagine having only two groups, additive and antagonistic group.

Step 6. Update σ^2 via Gibbs sampler.

The step 4 and 5 in the above algorithm allows jumps between additive and synergistic groups, and jumps between additive and antagonistic groups. In other words, the big jumps between synergistic and antagonistic groups are broken down into two small jumps, which would be easier to achieve. Unfortunately, one obvious drawback is that at each iteration we need to update $|\text{ADD}| + p$ ($|\text{ADD}|$ denotes for the number of elements in ADD group) of (I_k, β_k) pairs, since the index k runs over the indices in ADD twice as stated in step 4 and 5. Thus we propose the following algorithm which need to propose just p of (β_k, I_k) pairs at each iteration.

MCMC Algorithm V:

Step 1. Update the coefficients in the additive group together with intercept.

Step 2. Update the coefficients in the synergistic group together with intercept.

Step 3. Update the coefficients in the antagonistic group together with intercept.

Step 4. For $k = 1, \ldots, p$, update (β_k, I_k) as a block. Now the proposed value of I_k, different from I_k, is drawn with probability

$$\frac{\pi(I_k^*|I_{[-k]}, \boldsymbol{\beta}, \sigma^2)}{1 - \pi(I_k|I_{[-k]}, \boldsymbol{\beta}, \sigma^2)},$$

where π is the joint posterior density function of $I, \boldsymbol{\beta}, \sigma^2$. This is referring to Metropolized Gibbs sampler in Liu (1996), which proves that the way used in Step 4 above is more efficient than the Gibbs sampling method, that is, the proposed value of I_k, possibly the same as the current value of I_k, is sampled from $\pi(\cdot|I_{[-k]}, \boldsymbol{\beta}, \sigma^2)$. Then β_k^* is proposed by leaving the average effect of X_k unchanged, which is the same idea as before. Hence, the acceptance probability is the same as (4.3), with

$$\begin{aligned} P(I_k|I_k^*) &= \frac{\pi(I_k|\boldsymbol{\beta}^*)}{\sum_{j \neq I_k^*} \pi(I_k = j|\boldsymbol{\beta}^*)}, \\ P(I_k^*|I_k) &= \frac{\pi(I_k^*|\boldsymbol{\beta})}{\sum_{j \neq I_k} \pi(I_k = j|\boldsymbol{\beta})}. \end{aligned}$$

Figure 4.17: Algorithm V: Trace plots of MCMC samples for $\beta_j, j = 1, \ldots, p$ based on three-group diffuse interaction models with $\lambda_2 = 4/5, \lambda_3 = 5/4$.

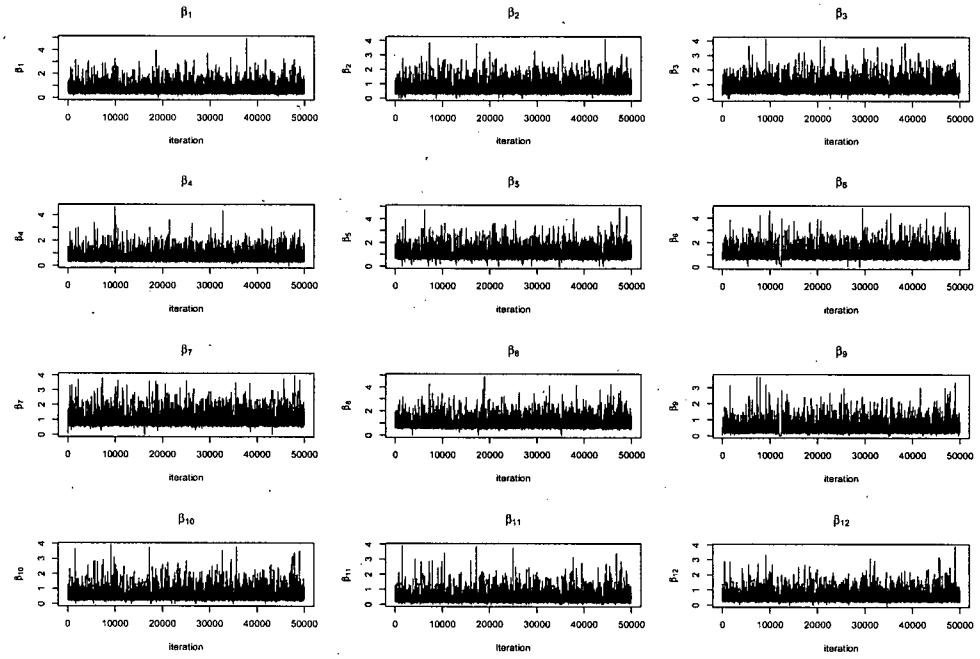


Figure 4.18: Algorithm V: Posterior densities of MCMC samples for $\beta_j, j = 1, \ldots, p$ based on three-group diffuse interaction models with $\lambda_2 = 4/5, \lambda_3 = 5/4$.

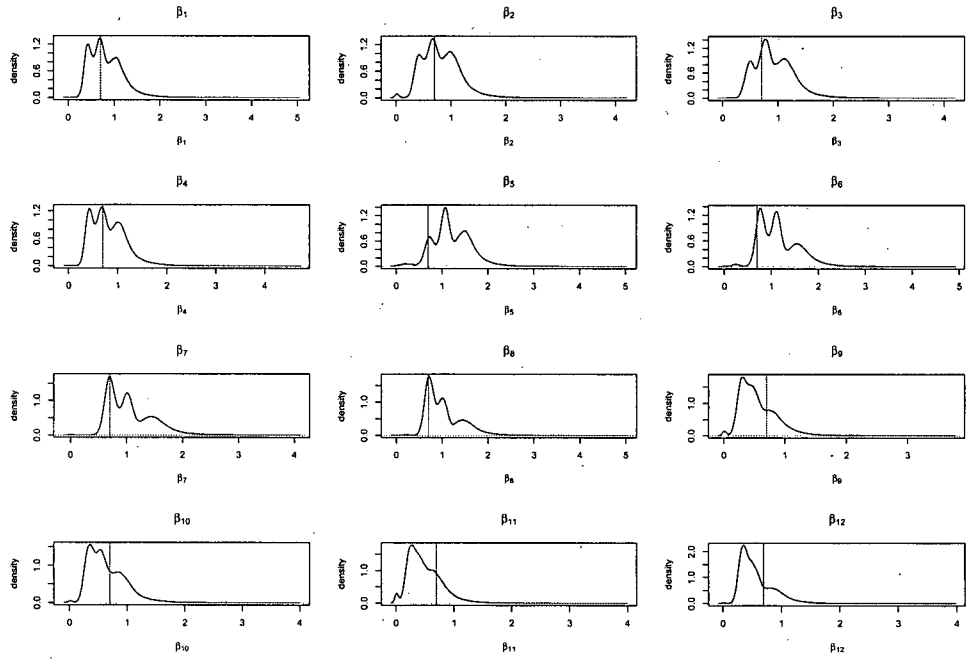
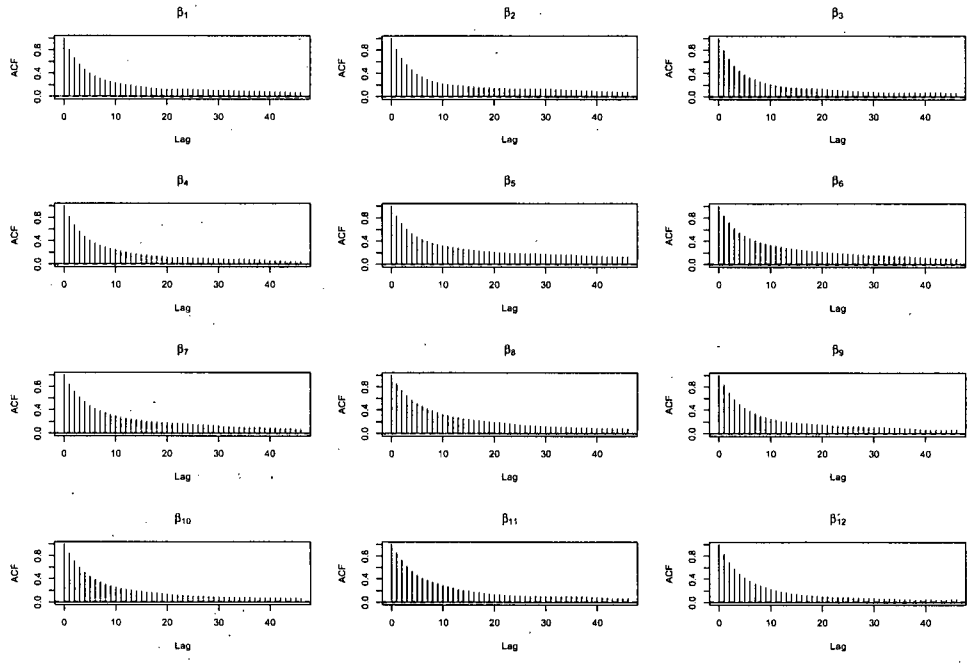


Figure 4.19: Algorithm V: Autocorrelation curves of MCMC samples for $\beta_j, j = 1, \ldots, p$ based on three-group diffuse interaction models with $\lambda_2 = 4/5, \lambda_3 = 5/4$.



Remarks on MCMC Algorithm V:

First, note that the magnitudes of λ_2 and λ_3 have an effect. If λ_2 and λ_3 getting closer to 1, the jump between different groups, i.e., update in Step 4, is easier to make. Otherwise, the change of group may cause big change in the posterior density, which leads to a small acceptance ratio. So the proposed values are more likely to be rejected, which makes the algorithm inefficient. We also check the performance of Algorithm V by setting all λ's equal to 1. Under this setting, the posterior distribution of I_k,for given $k \in \{1, \ldots, p\}$, is almost uniform distribution over {1,2,3}. This is consistent with our intuition, since when $\lambda_1 = \lambda_2 = \lambda_3 = 1$, the three group are actually identical to each other.

Second, according to the trace plots, as shown in Figure 4.17, the sampler mix well.

However, Figure 4.19 tells us that there is some scope to make improvement in convergence rate. The sample dependence does not drop close to zero after lag 40 for some β_k's, like β_5 and β_{11}. One thing worth noticing is the triple-mode in almost all density plots of β_k's. The reason is that for each k, the posterior distributions of β_k conditional on different values of I_k may have different (up to three) modes. Actually this is another evidence to show the high dependence between β_k and I_k.

Table 4.3: Algorithm III: Frequency table based on posterior samples for each component of I under three-group diffuse interaction group.

	I_1	I_2	I_3	I_4	I_5	I_6	I_7	I_8	I_9	I_{10}	I_{11}	I_{12}
1	0.33	0.31	0.34	0.31	0.33	0.33	0.29	0.28	0.31	0.29	0.26	0.28
2	0.23	0.19	0.18	0.24	0.17	0.34	0.38	0.43	0.31	0.28	0.26	0.43
3	0.44	0.51	0.49	0.46	0.50	0.33	0.32	0.29	0.38	0.43	0.48	0.29
true value	1	1	1	1	2	2	2	2	3	3	3	3

Third, we also plot the number of correct group allocations. It seems like that most of the samples for I have 4-6 components correctly valued. It looks promising as implied by Figure 4.20, some of the samples for I do achieve 10 or 11 correct allocations. Moreover, Figure 4.21 shows the posterior density plots of samples for β_k conditional on the true group allocation of k-th predictor, $k = 1, \ldots, p$. We find that for each β_k, the samples highly concentrates around its true value. It is a good sign to see that Algorithm V somehow works.

Figure 4.20: Algorithm V: Number of correct group allocations based on posterior samples of I under three-group diffuse interaction models.

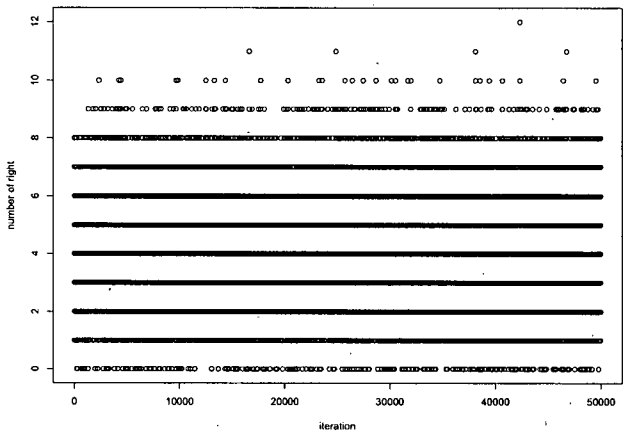
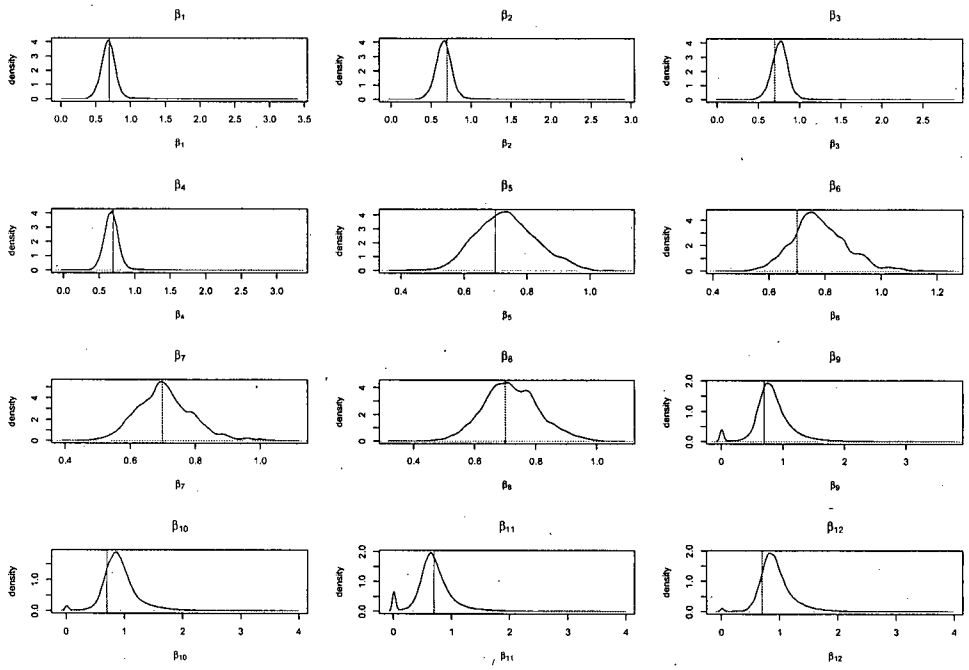


Figure 4.21: Algorithm V: Posterior densities of samples for β_j conditional on I_j correctly allocated ($j = 1, \ldots, p$) under three-group diffuse interaction models.



4.4 Example

To provide an illustrative example of fitting the diffuse interaction model to a real data set, we consider the abalone growth dataset available from the UCI Machine Learning Repository (Newman *et. al.* 1998). The response variable Y is the age of abalone. Ignoring the single categorical explanatory variable (sex), we take the dependent variables $\mathbf{X}$ to be the seven continuous explanatory variables (length, diameter, height, whole weight, shucked weight, viscera weight, shell weight). The data are observed on $n = 4177$ specimens. To find out the overall direction of interaction among those dependent variables, we apply Algorithm I to make inference. One thing worth noting is the release of positive constraints on β_j's by using the following model mentioned in Section 4.2.

$$\mathrm{E}\,(Y|\mathbf{X}) = \beta_0 + \left\{ \sum_{j=1}^{p} \left[g(x_j, \beta_j) \right]^{\lambda} \right\}^{1/\lambda},$$

where

$$g(x, \beta) = \begin{cases} |\beta| x; & \beta > 0, \\ |\beta|(1-x); & \beta < 0. \end{cases}$$

As a consequence, all the observations of X_j's are scaled to [0,1].

The chosen priors are $\beta_j \sim N(0, 0.5)(j = 0, 1, \ldots, p)$, $\log \lambda \sim N(0, 0.5)$, and $\sigma^2 \sim$ inverse-gamma $(0.0001, 0.0001)$. We also tried rather diffuse priors by replacing the small variance 0.5 with a larger value 50 and we did not see any serious difference in the output. Figure 4.22 gives trace plots for posterior samples of $\beta_j, j = 0, 1, \ldots, p$, which shows the MCMC algorithm worked well. The solid curves in Figure 4.23 are posterior densities for average effects $\delta_1, \ldots, \delta_p$ and the dashed lines are posterior densities for $\beta_1, \ldots, \beta_p$. We can see that most of δ_j's are slightly smaller than the corresponding β_j's, although the case for X_5 is the other way around. This means the overall interaction among $\mathbf{X}$ is

antagonistic, which is consistent with the inference we can make based on the posterior sample for λ. The posterior mean of λ is 1.19, and the 95% equal-tailed credible interval is (0.92,1.46). The posterior probability that $\lambda > 1$ is 0.93. Hence, we have evidence of the presence of antagonism among $\mathbf{X}$. The posterior density plot for λ is given in Figure 4.24.

For a better understanding of the content of antagonism, we refer to *relative antagonistic effect* as

$$\frac{g(\mathbf{x}) - g(\mathbf{0}) - \sum_{j=1}^{p} \left(g(x_j \mathbf{1}_j) - g(\mathbf{0})\right)}{\sum_{j=1}^{p} \left(g(x_j \mathbf{1}_j) - g(\mathbf{0})\right)}, \tag{4.7}$$

where $\mathbf{1}_j$ means a $p \times 1$ vector of zero except that the jth element is 1. Note that the numerator is the difference of the joint effect of $\mathbf{X}$ and the sum of independent effect of each $X_j, j = 1, \ldots, p$, and the denominator is the sum of independent effect. Averaging the ratio over the joint distribution of $\mathbf{X}$, we get *average* relative antagonistic effect (ARAE). By using the empirical distribution of $\mathbf{X}$ (the true distribution of X is unknown), we get ARAE based on the posterior samples of $\beta_j, j = 0, \ldots, p$ and λ and the graphical summary is given by the bottom panel in Figure 4.24. We can also see the evidence in favor of antagonism since all samples of ARAE are negative. Moreover, most of samples of ARAE are valued in (-0.6,-1), which indicates a considerable size of antagonism. The bi-mode of ARAE in the plot might be caused by the complexity of (4.7), because the denominator now is also a function of λ other than the simple form $\sum_j \beta_j x_j$ as before when β_j's are confined to be positive. One thing worth mention is that MCMC can provide an effective way to make inference on this complicated function ARAE, based on the posterior samples. That's one of the main reasons why we pursue MCMC approaches for model fitting, as mentioned in the beginning of this chapter.

Note that both of the definitions above can be easily applied to synergistic effect case.

Figure 4.22: Abalone data: trace plots for $\beta_j, j = 0, 1, \ldots, p$, for the whole sequence of 100,000 iterations including the burn-in period.

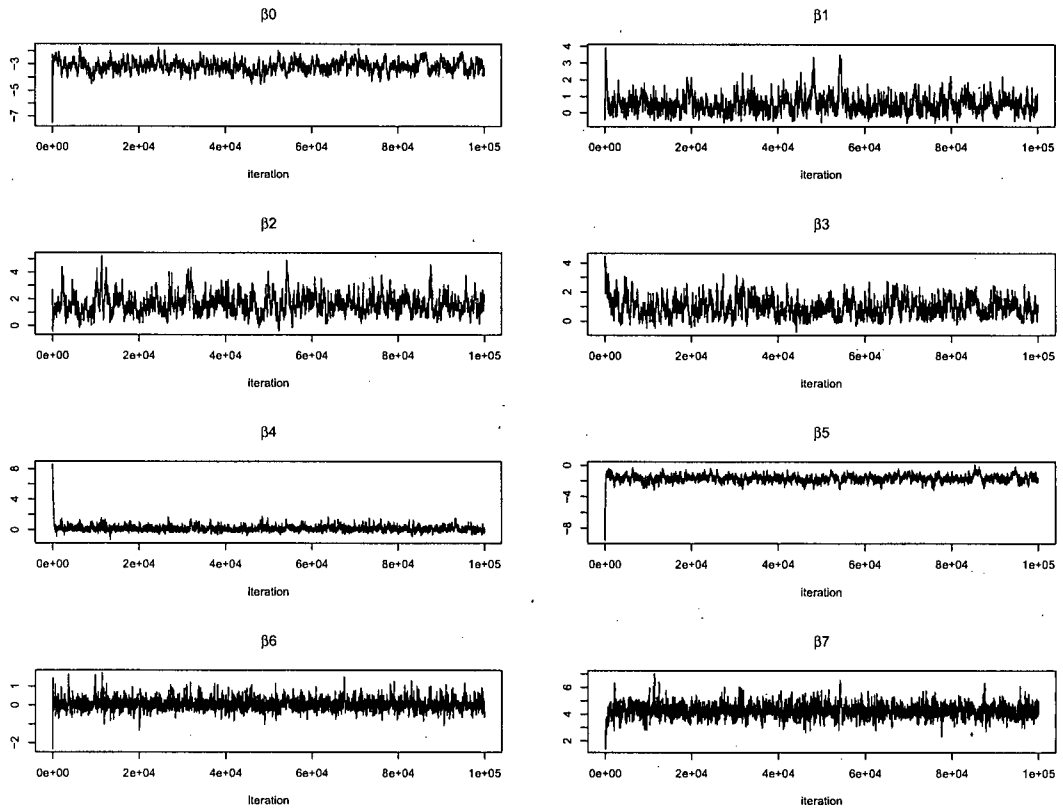


Figure 4.23: Abalone data: density plots for β_j and δ_j, $j = 1, \ldots, p$, for the sequence of 40,000 post burn-in iterations. The solid lines stand for average effect δ_j's and dashed lines stand for β_j's.

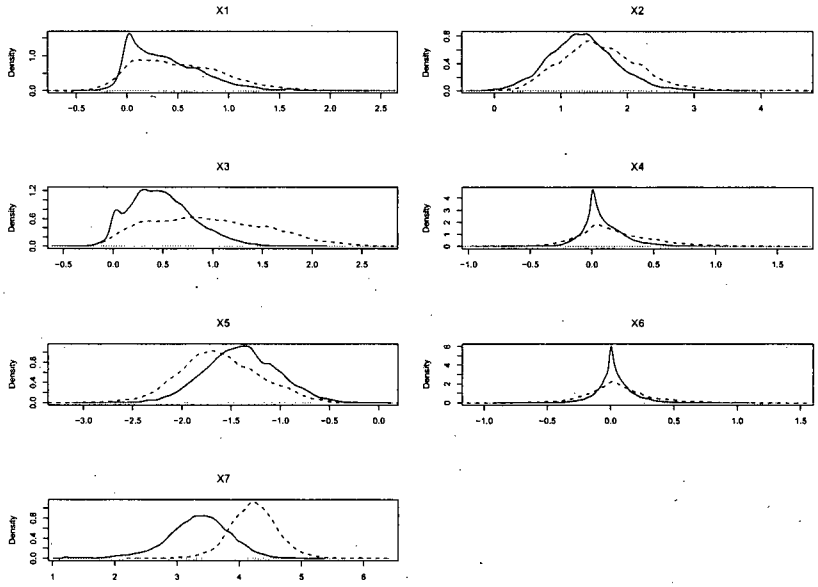
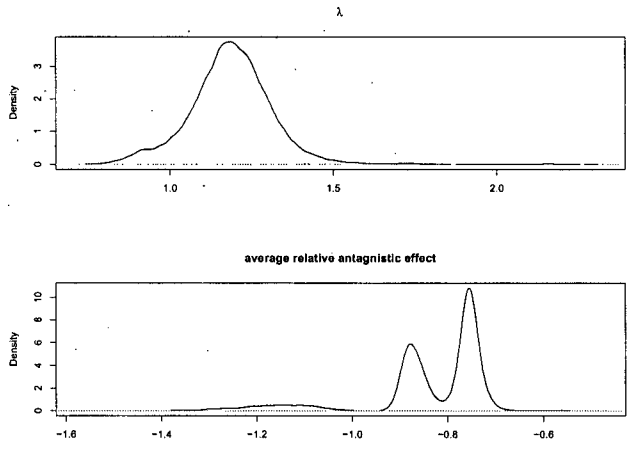


Figure 4.24: Abalone data: the top panel is the density plot of λ and the bottom panel is the density plot of relative antagonistic effect, for the sequence of 40,000 post burn-in iterations.



Chapter 5

Summarization and future work

5.1 Conclusions

Overall, the findings are as follows.

In chapter 2, we study the consequences of fitting an additive regression model a pairwise interaction model is assumed to be the true model. We obtained the asymptotic distribution of average effect estimates based on the "misspecified" model and "true" model as well, as shown in Result 1 and Result 2, respectively. With the two large sample limits achieved in the two results, we work out the consistency of average effect estimates from "misspecified" model under some easily-studied situations, which are given by Result 3. This result implies that the distribution of risk factors does influence the size of bias of estimates in cases of model misspecifications. Result 3 suggests that transformations of risk factors, aiming toward normality, may help to reduce bias of average effect estimates. More generally, under the framework of spline regression models, we investigate the consequences of model misspecifications by failing to incorporate the interaction terms, which are assumed to be included into the true model.

In chapter 3, we introduce the diffuse interaction model, which is more powerful to detect interaction than the pairwise interaction model especially when the number of risk factors of interest is rather large. We compare the power to detect interactions under the two models in situations where we assume the true model is either diffuse or pairwise interaction model. To make the comparison tractable, large sample and small

misspecification approximations are employed. To be specific, for either interaction model that is assumed to be true, the true values of the parameters standing for the magnitude of interactions (λ in the diffuse interaction model, β_{ij}'s in the pairwise interaction model) are just within a (local) $n^{-1/2}$ neighborhood of those values which imply no interactions at all, i.e., additive models. Therefore, as sample size n goes up to infinity, the model misspecification vanishes to zero. Under a set of specific settings (shown in Section 3.3), we find out that the power of diffuse interaction model is superior to pairwise interaction model no matter whether the true model is itself or not. However, if the true model is pairwise interaction, the detectability via diffuse interaction model decreases when the "overall" strength of interaction among the risk factors gets weaker.

In chapter 4, we develop an efficient MCMC algorithm for one-group diffuse interaction model. Also we investigate the possibility of generalizing the model away from the strong assumption that all risk factors interact in the same direction, i.e., synergistic or antagonistic. In the more generalized but complex model, we have more parameters since each risk factor has an corresponding indicator variable denoting to which interaction group it belongs. With λ fixed, we have an efficient algorithm for model with two groups of risk factors, one group for risk factors having no interactions and the other groups for risk factors interacting synergistically/antagonistically. And we also see some hope to develop a good MCMC sampler for a more general model with three groups of risk factors, i.e., a no-interaction/additive group, a synergistic group and an antagonistic group.

5.2 Future work

In this section we discuss some interesting problems that could be studied in the future.

1. In terms of model misspecification ignoring interactions:

(a). In Section 2.3, the linear regression context, we find that under some certain condition such as independence or joint normality of risk factors, the average effect estimators based on the misspecified model are still consistent with the true values. But we need to know whether there are more general (possibly weaker) conditions which can produce the consistency in the face of model misspecifications.

(b). We explored two examples in section 2.3 to see how far the bias can be away from zero. However, we need more general investigation of the magnitude of bias (or relative bias) as the joint distribution of $(X_1, \cdots, X_p)$ moves away from multivariate normally or independence.

(c). We could also study the consequences of omitting the interaction terms in the context of generalized linear models. Actually, in many epidemiological studies, the health outcome Y is often binary/categorical (for example, diseased or not).

(d) The main results in terms of average effects we have derived so far is referring to Definition 2, that is, averaging predictive effects over the joint distribution of all predictors. What could the result be if other versions of average effect are applied? Moreover, we are aware of the fact that idea of average effect could have more general applications, not just confined within the context of regression or generalized regression. For example, we can also apply the idea of average effect in survival analysis by averaging over the change in hazard function, instead of outcome/response variable, associated with a unit change in the putative risk factor.

2. In terms of MCMC algorithm development for more complex diffuse interaction models:

As suggested in Gustafson et al. (2005), a more general model can be built by partitioning the risk factors into three sets: additive, synergistic, and antagonistic, that

is

$$\begin{aligned} \mathrm{E}\,(Y|X_1,\cdots,X_p) &= \beta_0 + \sum_{j\in \mathrm{ADD}} \beta_j X_j \\ &\quad + \left\{ \sum_{j\in \mathrm{SYN}} (\beta_j X_j)^{\lambda_s} \right\}^{1/\lambda_s} + \left\{ \sum_{j\in \mathrm{ANT}} (\beta_j X_j)^{\lambda_a} \right\}^{1/\lambda_a}, \end{aligned} \tag{5.1}$$

where $0 < \lambda_s < 1 < \lambda_a$.

Starting with the simplest diffuse interaction model with all the risk factors in single interaction group, we have an efficient MCMC sampler which does a good job. And also MCMC performs well for a diffuse model with two groups, additive and synergistic or antagonistic with a fixed value of λ.

However, for a more complicated situation, model (5.1), how to implement an efficient MCMC approach is still in process. The challenge here is how to propose the indicator I_k and corresponding β_k together in an efficient way. In our algorithm, we only allow the transition between additive group and synergistic/antagonistic group, while the transition between two interaction groups are not allowed directly within one single movement. The reason is that the direct change between two interaction groups would cause rather big change in posterior density function, which leads to low acceptance rate of proposed values of (β_k, I_k) pair. Although we see some hope in the simulation studies that the sampler mix well, we still need to find better one which mixes faster and leads to better estimation.

To be more realistic, we would also add one more group, that is, no-effect group into the current three-group interaction model. This is more challenging because it is even harder to make efficient jumping schemes between the no-effect group and other three groups.

3. In terms of model selections :

In the analyses of linear regression with interactions, a stepwise strategy is often applied to choose a subset of interaction terms. The problem with stepwise procedure is that it is a computationally intractable technique when the number of risk factors is quite large, as discussed in Chapter 1. In the worst situation, $\frac{1}{2}\binom{p}{2}\left(\binom{p}{2}+1\right)$ possible cases must be evaluated before obtaining the final model. Therefore, stepwise regression is not always a practical technique and other selection techniques may have to be considered.

Suppose we have a linear model with pairwise and maybe even triple-wise interaction terms. We could take this model and reparametrize from the original parameters $\boldsymbol{\theta}$ to $(\boldsymbol{\phi}, \boldsymbol{\lambda})$, where $\boldsymbol{\phi}$ are the average effects and $\boldsymbol{\lambda}$ are nuisance parameters. Note that when $\boldsymbol{\lambda} = \mathbf{0}$, the interaction model is an additive model. Take a pairwise interaction model (3.6) for instance, reparametrize the parameters $\boldsymbol{\beta}$ as follows:

$$\begin{aligned} \phi_0 &= \beta_0, \\ \phi_j &= \beta_j + \sum_{i \neq j} \beta_{ij} \mathrm{E}(X_i), \\ \boldsymbol{\lambda} &= \{\beta_{ij}\}_{1 \leq i < j \leq p}. \end{aligned}$$

By setting up a prior for $\boldsymbol{\lambda}$ which is quite concentrated around zero, we can do the posterior inference to $\boldsymbol{\lambda}$ and then pick out those important interaction terms instead of doing stepwise procedures.

Appendix I

Proof of (2.14)in Section 2.3.1

By Cramer's rule to solve the $(p+1) \times (p+1)$ system of equations, we get

$$\alpha_1 |D| = |D_1|, \tag{.1}$$

where

$$D = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \mathrm{Var}\,(X_1) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \mathrm{Cov}(X_1, X_p) & \cdots & \mathrm{Var}\,(X_p) \end{pmatrix},$$

$$D_1 = \begin{pmatrix} 1 & EY & 0 & \cdots & 0 \\ 0 & \mathrm{E}\,(Y\widetilde{X}_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \mathrm{E}\,(Y\widetilde{X}_p) & \mathrm{Cov}(X_2, X_p) & \cdots & \mathrm{Var}\,(X_p) \end{pmatrix}.$$

Let

$$\Sigma = (\sigma_{ij})_{p\times p} = \begin{pmatrix} \mathrm{Var}\,(X_1) & \mathrm{Cov}(X_1, X_2) & \cdots & \mathrm{Cov}(X_1, X_p) \\ \mathrm{Cov}(X_1, X_2) & \mathrm{Var}\,(X_2) & \cdots & \mathrm{Cov}(X_2, X_p) \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}(X_1, X_p) & \mathrm{Cov}(X_2, X_p) & \cdots & \mathrm{Var}\,(X_p) \end{pmatrix}.$$

Since

$$\mathrm{E}\left\{Y\left(\begin{array}{cccc} 1, & \widetilde{X}_1, & \ldots, & \widetilde{X}_p \end{array}\right)'\right\} = \left(\begin{array}{c} \mathrm{E}(Y) \\ \sum_i \beta_i \mathrm{E}(X_i \widetilde{X}_1) + \sum_{i<j} \beta_{ij} \mathrm{E}(X_i X_j \widetilde{X}_1) \\ \vdots \\ \sum_i \beta_i \mathrm{E}(X_i \widetilde{X}_p) + \sum_{i<j} \beta_{ij} \mathrm{E}(X_i X_j \widetilde{X}_p) \end{array}\right),$$

with determinant expansion by the cofactors (first column), we may write $|D_1|$ as

$$\left|\begin{array}{c} \sum_i \beta_i \mathrm{E}(X_i \widetilde{X}_1) + \sum_{i<j} \beta_{ij} \mathrm{E}(X_i X_j \widetilde{X}_1) \\ \vdots \qquad\qquad , \boldsymbol{\sigma}_2, \cdots, \boldsymbol{\sigma}_p \\ \sum_i \beta_i \mathrm{E}(X_i \widetilde{X}_p) + \sum_{i<j} \beta_{ij} \mathrm{E}(X_i X_j \widetilde{X}_p) \end{array}\right|,$$

where $\boldsymbol{\sigma}_i$ denotes i-th column vector of Σ.

By the property of determinant, we can rewrite the expressions above as summation of two determinants $D_1^{(1)}$ and $D_1^{(2)}$, where

$$|D_1^{(1)}| = \left|\begin{array}{c} \sum_i \beta_i \mathrm{E}(X_i \widetilde{X}_1) \\ \vdots \qquad , \boldsymbol{\sigma}_2, \cdots, \boldsymbol{\sigma}_p \\ \sum_i \beta_i \mathrm{E}(X_i \widetilde{X}_p) \end{array}\right|,$$

$$|D_1^{(2)}| = \left|\begin{array}{c} \sum_{i<j} \beta_{ij} \mathrm{E}(X_i X_j \widetilde{X}_1) \\ \vdots \qquad , \boldsymbol{\sigma}_2, \cdots, \boldsymbol{\sigma}_p \\ \sum_{i<j} \beta_{ij} \mathrm{E}(X_i X_j \widetilde{X}_p) \end{array}\right|.$$

Let's consider $D_1^{(1)}$ at first. Note the fact that $\mathrm{E}(X_i\widetilde{X}_j) = \mathrm{Cov}(X_i, X_j)$.

$$\begin{aligned} |D_1^{(1)}| &= \sum_i \beta_i \mathrm{Cov}(X_i, X_1)\Sigma^{11} + \cdots + \sum_i \beta_i \mathrm{Cov}(X_i, X_p)\Sigma^{p1} \\ &= \sum_i \beta_i [\mathrm{Cov}(X_i, X_1)\Sigma^{11} + \cdots + \mathrm{Cov}(X_i, X_p)\Sigma^{p1}] \\ &= \beta_1 |\Sigma|. \end{aligned} \tag{.2}$$

The last equality is derived by the fact that when $i \neq 1$

$$\begin{aligned} &\mathrm{Cov}(X_i, X_1)\Sigma^{11} + \cdots + \mathrm{Cov}(X_i, X_p)\Sigma^{p1} \\ = &\begin{vmatrix} \mathrm{Cov}(X_i, X_1) & & \\ \vdots & , \boldsymbol{\sigma}_2, \cdots, \boldsymbol{\sigma}_p & \\ \mathrm{Cov}(X_i, X_p) & & \end{vmatrix} = 0 \end{aligned}$$

Similarly, we get

$$|D_1^{(2)}| = \sum_{i<j} \beta_{ij} [\Sigma^{11}\mathrm{E}(X_i X_j \widetilde{X}_1) + \cdots + \Sigma^{p1}\mathrm{E}(X_i X_j \widetilde{X}_p)].$$

Since

$$\mathrm{E}(X_i X_j \widetilde{X}_1) = \mathrm{E}(\widetilde{X}_i \widetilde{X}_j \widetilde{X}_1) + \mathrm{E}(X_j)\mathrm{Cov}(X_i, X_1) + \mathrm{E}(X_i)\mathrm{Cov}(X_j, X_1),$$

we can derive that

$$\begin{aligned}
|D_1^{(2)}| &= \sum_{i<j} \beta_{ij}(\Sigma^{11}\mathrm{E}\,(\widetilde{X}_i\widetilde{X}_j\widetilde{X}_1) + \cdots + \Sigma^{p1}\mathrm{E}\,(\widetilde{X}_i\widetilde{X}_j\widetilde{X}_p)) \\
&+ \sum_{i<j} \beta_{ij}\mathrm{E}\,(X_j)[\mathrm{Cov}(X_i, X_1)\Sigma^{11} + \cdots + \mathrm{Cov}(X_i, X_p)\Sigma^{p1}] \\
&+ \sum_{i<j} \beta_{ij}\mathrm{E}\,(X_i)[\mathrm{Cov}(X_j, X_1)\Sigma^{11} + \cdots + \mathrm{Cov}(X_j, X_p)\Sigma^{p1}] \\
&= \sum_{i<j} \beta_{ij}(\Sigma^{11}\mathrm{E}\,(\widetilde{X}_i\widetilde{X}_j\widetilde{X}_1) + \cdots + \Sigma^{p1}\mathrm{E}\,(\widetilde{X}_i\widetilde{X}_j\widetilde{X}_p)) + \sum_{j>1} \beta_{1j}\mathrm{E}\,(X_j)|\Sigma|. \quad (.3)
\end{aligned}$$

Combining equations (.1), (.2) and (.3), we get the equation (2.15) and so (2.16) when $p = 2$.

Appendix II

Another consistent estimator of V^* mentioned in Section 2.2

The following is to show that sandwich estimator is a consistent estimator of V^* as well. We only work on the simple case with two predictors and it is easy to expand the results of $p = 2$ to general p.

In the fitted model, the log-likelihood function of an observation is

$$\log f_\theta(Y|X_1, X_2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \tau^2 - \frac{1}{2\tau^2}(Y - \alpha_0 - \alpha_1 X_1 - \alpha_2 X_2)^2,$$

while the true log-likelihood function should be

$$\log g(Y|X_1, X_2) = -\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma^2 - \frac{1}{2\sigma^2}(Y - \beta_0 - \beta_1 X_1 - \beta_2 X_2 - \beta_{12} X_1 X_2)^2.$$

Define the matrices as below.

$$\begin{aligned} A_n(\boldsymbol{\theta}) &= \left\{ n^{-1} \sum_{i=1}^{n} \partial^2 \log f_\theta(Y_i|X_1 = x_{1i}, X_2 = x_{2i})/\partial\theta_i \partial\theta_j \right\}, \\ B_n(\boldsymbol{\theta}) &= \left\{ n^{-1} \sum_{i=1}^{n} \partial \log f_\theta(Y_i|X_1 = x_{1i}, X_2 = x_{2i})/\partial\theta_i \cdot \partial \log f_\theta(Y_i|X_1 = x_{1i}, X_2 = x_{2i})/\partial\theta_j \right\}. \end{aligned}$$

The expecations of them are defined as

$$\begin{aligned} A(\boldsymbol{\theta}) &= \mathrm{E}\left(\partial^2 \log f_{\boldsymbol{\theta}}(Y|X_1, X_2)/\partial\theta_i \partial\theta_j\right), \\ B(\boldsymbol{\theta}) &= \mathrm{E}\left(\partial \log f_{\boldsymbol{\theta}}(Y|X_1, X_2)/\partial\theta_i \cdot \partial \log f_{\boldsymbol{\theta}}(Y|X_1, X_2)/\partial\theta_j\right). \end{aligned}$$

By White (1982) we know that the MLE of parameter vector $\widehat{\boldsymbol{\theta}}_n = (\hat{\alpha}_0, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\tau}^2)$, which is a consistent estimator for $\boldsymbol{\theta}_*$, minimizing the Kullback-Leibler Information Criterion (KL). That is

$$\boldsymbol{\theta}_* = \arg\min_{\boldsymbol{\theta}} \mathrm{E}\left(\log \frac{g(Y|X_1, X_2)}{f_{\boldsymbol{\theta}}(Y|X_1, X_2)}\right), \tag{.1}$$

and

$$\sqrt{n}\,(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_*) \to N(0, C(\boldsymbol{\theta}_*)). \tag{.2}$$

Intuitively, KL measures our ignorance about the true model.

Remarks:

1) All the expectations here are calculated under the true density function unless specified.

2) $C(\boldsymbol{\theta})$ is defined as follows.

$$C(\boldsymbol{\theta}) = A(\boldsymbol{\theta})^{-1} B(\boldsymbol{\theta}) A(\boldsymbol{\theta})^{-1}.$$

And $A_n(\boldsymbol{\theta})^{-1} B_n(\boldsymbol{\theta}) A_n(\boldsymbol{\theta})^{-1}$ is a consistent estimator of $C(\boldsymbol{\theta})$. According to its sandwich-like shape, this estimator is also called "sandwich" estimator.

Note that when the fitted model is correct, $A(\boldsymbol{\theta}_*) + B(\boldsymbol{\theta}_*)$ equals zero, otherwise it may not. So $A(\boldsymbol{\theta}_*) + B(\boldsymbol{\theta}_*)$ is a useful indicator of misspecifications.

By some algebra, we have

$$A(\boldsymbol{\theta}_*) = \frac{-1}{\tau_*^2}\begin{pmatrix} \Sigma_S & \mathbf{0}_{3\times 1} \\ \mathbf{0}_{1\times 3} & \frac{1}{2\tau_*^2} \end{pmatrix}, \tag{.3}$$

where

$$\Sigma_{\mathbf{S}} = \mathrm{E}\,(\mathbf{S}\mathbf{S}') = \mathrm{E}\left\{(1, X_1, X_2)'(1, X_1, X_2)\right\}.$$

$$B(\boldsymbol{\theta}_*) = \frac{1}{\tau_*^4}\begin{pmatrix} V^* & \gamma \\ \gamma^T & \triangle \end{pmatrix},$$

where

$$\begin{aligned} \gamma &= \frac{1}{4\tau_*^2}\begin{pmatrix} \mathrm{E}\,(Y-\alpha_0-\alpha_1X_1-\alpha_2X_2)^3 \\ \mathrm{E}\left((Y-\alpha_0-\alpha_1X_1-\alpha_2X_2)^3X_1\right) \\ \mathrm{E}\left((Y-\alpha_0-\alpha_1X_1-\alpha_2X_2)^3X_2\right) \end{pmatrix}, \\ \triangle &= \mathrm{E}\left[\frac{1}{2\tau_*^2}(Y-\alpha_0-\alpha_1X_1-\alpha_2X_2)^2-\frac{1}{2}\right]^2. \end{aligned}$$

Note V^* is defined in (2.8).

Therefore by the definition of matrix C, we get

$$C(\boldsymbol{\theta}_*) = \begin{pmatrix} \Sigma_S^{-1}V^*\Sigma_S^{-1} & 2\tau_*^2\Sigma_S^{-1}\gamma \\ 2\tau_*^2\gamma'\Lambda^{-1} & 4\tau_*^4\triangle \end{pmatrix}.$$

The left upper sub-matrix is just $v(\boldsymbol{\alpha}_*)$.

Appendix III

Numerical approach to C_{11} and C_{12} in Section 2.4.1

We give details of the numerical approach how to get C_{11} and C_{12} in (2.20). The joint distribution of (X_1, X_2) is bivariate normal.

The elements of matrix C_{11} are easily obtained by one-dimensional integration. As for the elements of matrix C_{12}, the typical term is

$$\mathrm{E}\left\{\left((X_1 - c_1)^a (I\{X_1 > c_1\} - k_1\right)\left((X_2 - c_2)^b I\{X_2 > c_2\} - k_2\right)\right\}, \qquad (.1)$$

where

$$\begin{aligned} c_i &= \begin{cases} 0, & \text{if } c_i = -\infty, \\ t_i, & \text{if } c_i = t_i \neq -\infty. \end{cases} \\ k_i &= \mathrm{E}\left((X_i - c_i)^a (I\{X_i > c_i\}\right). \end{aligned}$$

To reduce the two-dimensional integration into one dimension, we just need to figure out

$$\mathrm{E}\left((X_2 - c_2)^b I\{X_2 > c_2\} - k_2 | X_1\right).$$

Note the fact conditional on $X_1 = x_1$, X_2 can be rewritten as

$$X_2 = \rho x_1 + \sqrt{1-\rho^2}\, Z,$$

where ρ is the correlation coefficient of X_1 and X_2 and Z stands for a standard normal variable. Therefore,

$$\begin{aligned} f(x_1) &= \mathrm{E}\left((X_2 - c_2)^b I\{X_2 > c_2\} | X_1 = x_1\right) \\ &= \left(1-\rho^2\right)^{b/2} \mathrm{E}\left\{\sum_{i=0}^{b} C_i^b Z^i \left(\frac{\rho x_1 - c_2}{\sqrt{1-\rho^2}}\right)^{b-i} I\left\{Z > \frac{c_2 - \rho x_1}{\sqrt{1-\rho^2}}\right\}\right\}. \end{aligned}$$

This integral can be easily evaluated since Z is standard normal and then we can regard the integrand in (.1), typical term of C_{12}, as a function of one variable:

$$g(X_1) = \left((X_1 - c_1)^a (I\{X_1 > c_1\} - k_1\right) \times \left(f(X_1) - k_2\right).$$

Thus, we can approximate the integral of the function by the following

$$\mathrm{E}\left\{g(X_1)\right\} \approx K^{-1} \sum_{j=1}^{K} g(b_j),$$

where b_i is $i/(K+1)$ quantile of the marginal distribution of X_1. In our numerical approach to the elements of C_{11} and C_{12}, K is set to be 2000.

Appendix IV

Proof of (2.24) in Section 2.4.2

With the facts that $B^{-1}(B^{-1})' = \mathbb{S}'\mathbb{S}$ and $UCU' = BPB'$, we have

$$\begin{aligned}\mathbb{S}'\mathbb{S} + \lambda P &= B^{-1}(B^{-1})' + \lambda P \\ &= B^{-1}(\mathbf{I} + \lambda BPB')(B^{-1})' \\ &= B^{-1}U(\mathbf{I} + \lambda C)U'(B^{-1})'.\end{aligned}$$

Thus

$$(\mathbb{S}'\mathbb{S} + \lambda P)^{-1} = B'U(\mathbf{I} + \lambda C)^{-1}U'B.$$

Let $V = \mathbb{S}B'U$, then $V'V = \mathbf{I}$, which gives

$$\begin{aligned}\mathrm{tr}(\mathbf{S}(\boldsymbol{\alpha})) &= \mathrm{tr}\{V(\mathbf{I} + \lambda C)^{-1}V'\} \\ &= \mathrm{tr}(\mathbf{I} + \lambda C)^{-1}.\end{aligned}$$

Appendix V

Pseudocode for the hybrid MCMC algorithm in Section 4.2.1

Let $\boldsymbol{\theta} \sim \Pi$ be the target distribution, having an unnormalized density function $\pi(\boldsymbol{\theta})$ on a subset of $\mathbb{R}^k$. In the following, we will abuse notation with the same symbol used to denote different functions if the meaning is clear from the context. The algorithm works by extending the state from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta}, z)$, and the unnormalized target density from $\pi(\boldsymbol{\theta})$ to

$$\begin{aligned} \pi(\boldsymbol{\theta}, \mathbf{z}) &= \pi(\boldsymbol{\theta})\pi(\mathbf{z}) \\ &= \pi(\boldsymbol{\theta}) \exp\left(-\frac{1}{2}\sum_{i=1}^{k} z_i^2\right). \end{aligned}$$

0. Set values for function g, constant ϵ, constant α, number of iterations I.

1. Initialize the value of $\boldsymbol{\theta}$. Could be generated from the prior of each component of $\boldsymbol{\theta}$ or by fitting a simpler model (for example an additive model in regression model context). Also initialize $\mathbf{z}$ by sampling from standard normal.

2. For each iteration, $i = 1, \ldots, I$:

a) Generate a candidate state $(\boldsymbol{\theta}^*, \mathbf{z}^*)$ as

$$\begin{aligned} \boldsymbol{\theta}^* &\leftarrow \boldsymbol{\theta} + \epsilon\{\mathbf{z} + (\epsilon/2)g(\boldsymbol{\theta})\}, \\ \mathbf{z}^* &\leftarrow -\mathbf{z} - (\epsilon/2)\{g(\boldsymbol{\theta}) + g(\boldsymbol{\theta}^*)\}, \end{aligned}$$

and draw u from unif$(0,1)$. Set $(\boldsymbol{\theta}, \mathbf{z}) \leftarrow (\boldsymbol{\theta}^*, \mathbf{z}^*)$ if

$$u < \frac{\pi(\boldsymbol{\theta}^*, \mathbf{z}^*)}{\pi(\boldsymbol{\theta}, \mathbf{z})},$$

otherwise, keep the original $(\boldsymbol{\theta}, \mathbf{z})$.

b) Unconditionally negate $\mathbf{z}$; that is,

$$\mathbf{z} \leftarrow -\mathbf{z}.$$

c) Perform an autoregressive update to $\mathbf{z}$; that is,

$$\mathbf{z} \leftarrow N(\alpha\mathbf{z}, (1-\alpha^2)\mathbf{I}_k).$$

d) Set $\boldsymbol{\theta}_i \leftarrow \boldsymbol{\theta}$.

3. Output $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_I$.

Remarks: Setting $g(\boldsymbol{\theta}) = \nabla \log \pi(\boldsymbol{\theta})$ and α close to 1 we obtain hybrid algorithm. Other choices of g and α gives different algorithms.

Bibliography

Box, G. (1979). Robustness in the strategy of scientific model building, *Robustness in Statistics* pp. 201–236.

Breiman, L. (1996). Heuristics of instability and stabilization in model selection, *Annal of Statistics* **24**(6): 2350–2383.

Diaconis, P. and Shahshahani, M. (1984). Nonlinear functions of linear combinations, *SIAM Journal of Scientific Statistical Computation* **5**(1): 175–191.

Draper, D. (1995). Assessment and propagation of model uncertainty, *Journal of the Royal Statistical Society. Series B (Methodological)* **57**(1): 45–97.

Eubank, R. (1999). *Nonparametric regression and spline smoothing*, Marcel Dekker.

Friedman, J. (1991). Multivariate adaptive regression splines, *The Annals of Statistics* **19**(1): 1–67.

Gelman, A. and Pardoe, I. (2007). Average predictive effects for models with nonlinearity, interactions, and variance components, *Sociological Methodology*.

Greenland, S. (1979). Limitations of the logistics analysis of epidemilogic data, *American Journal of Epidemiology* **110**(6): 693.

Greenland, S. (1983). Tests for interaction in epidemiologic studies: a review and a study of power, *Statistics in Medicine* **2**(2): 243–51.

Greenland, S. (1993). Basic problems in interaction assessment, *Environmental Health Perspectives* **101**: 59–66.

Gu, C. (2002). *Smoothing spline ANOVA models*, Springer.

Gustafson, P. (2000). Bayesian regression modeling with interactions and smooth effects, *Journal of the American Statistical Association* **95**(451): 795–806.

Gustafson, P. (2001). On measuring sensitivity to parametric model misspecification, *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **63**(1): 81–94.

Gustafson, P. (2007). On robustness and model flexibility in survival analysis: transformed hazard models and average effects, *Biometrics* **63**: 69–77.

Gustafson, P., Kazi, A. and Levy, A. (2005). Extending logistic regression to model diffuse interactions, *Statistics in medicine* **24**(13): 2089–2104.

Gustafson, P., MacNab, Y. and Wen, S. (2004). On the Value of derivative evaluations and random walk suppression in Markov Chain Monte Carlo algorithms, *Statistics and Computing* **14**(1): 23–38.

Hills, S. and Smith, A. (1992). Parameterization issues in Bayesian inference, *Bayesian Statistics* **4**: 227–246.

Hogan, M., Kupper, L., Most, B. and Haseman, J. (1978). Alternatives to Rothman's approach for assessing synergism (or antagonism) in cohort studies, *American Journal of Epidemiology* **108**(1): 60–67.

Karlin, S. and Studden, W. (1966). *Tchebycheff Systems: With Applications in Analysis and Statistics*, Interscience Publishers.

Kleinbaum, D., Kupper, L. and Morgenstern, H. (1982). Interaction, effect modification, and synergism, *Epidemiologic Research. New York: Van Nostrand Reinhold* pp. 407–417.

Koopman, J. (1981). Interaction between discrete causes, *American Journal of Epidemiology* **113**(6): 716–724.

Le Cam, L. (1960). Locally asymptotically normal families of distributions, *University of California Publications in Statistics* **3**(2): 37–98.

Liu, J. (1996). Metropolized Gibbs sampler: an improvement.

Madigan, D. and Raftery, A. (1994). Model selection and accounting for model uncertainly in graphical models using Occam's window, *Journal of the American Statistical Association*.

McCullagh, P. and Nelder, J. (1983). *Generalized linear models*, Chapman & Hall.

Müller, A. and Stoyan, D. (2002). *Comparison methods for stochastic models and risks*, John Wiley and Sons.

Neal, R. (1996). *Bayesian learning for neural networks*, Springer.

Neal, R. (1998). Suppressing random walks in Markov chain Monte Carlo using ordered overrelaxation, *Learning in Graphical Models* pp. 205–225.

O'Brien, S., Kupper, L. and Dunson, D. (2006). Performance of tests of association in misspecified generalized linear models, *Journal of statistical planning and inference* **136**(9): 3090–3100.

Raftery, A. (1996). Approximate Bayes factors and accounting for model uncertainty in generalised linear models, *Biometrika* **83**(2): 251–266.

Rothman, K. (1974). Synergy and antagonism in cause-effect relationships, *American Journal of Epidemiology* **99**(6): 385.

Rothman, K., Greenland, S. and Walker, A. (1980). Concepts of interaction, *American Journal of Epidemiology* **112**(4): 467.

Saracci, R. (1980). Interaction and synergism, *American Journal of Epidemiology* **112**(4): 465.

Venables, W. and Ripley, B. (1999). *Modern applied statistics with S-Plus*, Springer New York.

Wang, D., Zhang, W. and Bakhai, A. (2004). Comparison of Bayesian model averaging and stepwise methods for model selection in logistic regression, *Statistics in Medicine* **23**(22): 3451–3467.

White, H. (1980). Using least squares to approximate unknown regression functions, *International Economic Review* **21**(1): 149–170.

White, H. (1981). Consequences and detection of misspecified nonlinear regression models, *Journal of the American Statistical Association* **76**(374): 419–433.

White, H. (1982). Maximum likelihood estimation of misspecified models, *Econometrica* **50**(1): 1–26.

Xu, R. and O'Quigley, J. (2000). Estimating average regression effect under non-proportional hazards, *Biostatistics* **1**(4): 423–439.