PENALIZED MULTIVARIATE LOGISTIC REGRESSION WITH A LARGE DATA SET

By

Fangyu Gao

A dissertation submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN – MADISON

1999

Abstract

We combine a smoothing spline ANOVA model and a log-linear model to build a partly flexible model for multivariate Bernoulli data. The joint distribution conditioning on the predictor variables is estimated. The conditional log odds ratio is used to measure the association between outcome variables. A numerical scheme based on the block one-step SOR-Newton-Ralphson algorithm is proposed to obtain an approximate solution for the variational problem. It is proved for a special case that the approximate solution can achieve the same statistical convergence rate as the exact solution, but is much more computing efficient. We extend $GACV$ (Generalized Approximate Cross Validation) to the case of multivariate Bernoulli responses. Its randomized version is fast and stable to compute. Simulation studies show that it is an excellent computational proxy for the CKL (Comparative Kullback-Leibler) distance. It is used to adaptively select smoothing parameters in each block one-step SOR iteration. Approximate Bayesian confidence intervals are obtained for the flexible estimates of the conditional logit functions. Simulation studies are conducted to check the performance of the proposed method. Finally, the model is applied to two-eye observational data from the Beaver Dam Eye Study to examine the association of pigmentary abnormalities and various covariates.

Acknowledgements

I would like to express my deepest gratitude to my advisor, Professor Grace Wahba. She initiated the research described in this dissertation and her dedication to statistics has been a tremendous inspiration to me. During the course of this study we had many fruitful discussions and she provided me numerous insightful suggestions. I shall always appreciate her guidance which led me into the wonderful world of smoothing spline.

I would also like thank my final committee members, Professors Yi Lin, Barbara Klein, Wei-Yin Loh and Michael Kosorok, for their reading of this dissertation and valuable comments.

Fellow graduate students Xiwu Lin, Alan Chiang, Helen Zhang, Yoonkyung Lee, Xiaoyin Fan, Peter Hoff, Paul Bradley, Yong Zeng, Chunyang Gai, Hongyu Jiang, Kin-Yee Chan, Chen Wang, Yonghua Chen, Lei Shen have helped me in various ways in the course of this study and the actual writing of the thesis. These and other graduate students made my life in Madison an enjoyable one.

Finally, special thanks go to my parents and sister for their love and support.

This research is supported in part by NIH Grant EY09946 and NSF Grant DMS9704758.

Contents

Chapter 1

Introduction

1.1 Motivation

The original motivation of this study comes from many typical data from ophthalmological studies. One characteristic of such kind of data set is that we have outcomes from both eyes of the same person. Usually, they are correlated Bernoulli outcomes, $Y_{ij}, i = 1, 2, ..., n, j = 1, 2$. $Y_{ij} = 1$ indicates that the jth eye of the ith person has a certain disease. Both person-specific and eye-specific covariates may be available as predictor variables.

As in many medical data, it is not sufficient to directly predict the outcome based on the available covariates, since even people with the same covariate values do not necessarily have the same medical outcomes. Instead, we are interested in finding the relation between outcome variables and predictor variables, i.e. (1) what is the probability p of a certain outcome conditioning on some given predictor variable values; (2) how will the changes of predictor variables affect the conditional probability p; (3) how strong are the correlations between those multiple outcomes.

The first question is to build a predictive model for future observations.

Their covariate variable values may not appear in the training set. Consequently we need some smoothing technique which not only provides estimate of p on those data points available for model building, but also provides prediction between those data points.

The second question is related to interpretability of our model. Unlike a black box, it should have readily interpretable result for multivariate function estimate and reasonable assessment of accuracy after the model has been fitted. This property is especially important for medical researchers, since the investigators are usually interested in understanding the cause of certain outcomes. In computer sciences, neural networks have been one of the most popular techniques for predictive model building, but the result is difficult to interpret.

The third question is related to the special structure of typical ophthalmology data sets and many other data sets. When analyzing data from a typical ophthalmology study, we must take into account the fact that the measurements made on both eyes of the same person are highly correlated. Hence, we can not treat them as independent outcomes. Multiple outcomes for the same person (or group) may also arise from two-period cross-over designs (Jones & Kenward 1989), twin studies (Cessie & Houwelingen 1994) and typical longitudinal studies. It is also of interest to model several closely related endpoints simultaneously. For example, in Liang, Zeger & Qaqish (1992), two endpoints from the Indonesian Children's Study, respiratory and diarrheal infections were considered in the same model. To address the third question, it is not enough

to simply estimate the marginal distribution separately for individual outcome variables. Instead, we want to treat those outcome variables together and estimate their joint distribution. The dependence structure can be useful for the efficient estimation of the mean values, or it can be of direct scientific interest. Numerous schemes have been proposed to study it. For example, Cox (1972) expressed the likelihood function in terms of the multivariate exponential family distribution. Qu, Williams, Beck & Goormastic (1987) considered conditional logistic models. McCullagh & Nelder (1989) proposed multivariate marginal logistic regression model. Lipsitz, Laird & Harrington (1991) and Williamson, Kim & Lipsitz (1995) considered marginal models and used the (global) odds ratio as a measurement of association. Liang et al. (1992) had a discussion about the difference between log-linear and marginal models. Molenberghs & Ritter (1996) proposed a likelihood based marginal model and established the connection with the second order generalized estimating equations (GEE2).

Classical log-linear models have been widely used to estimate joint conditional probabilities. See Bishop, Fienberg & Holland (1975). People used to assume linear parametric forms for all the conditional logit functions to be estimated. However, it is not always adequate to make linear or even quadratic or cubic assumptions. When the linear assumption is far away from the truth, the result obtained under such an assumption may even be misleading.

On the other hand, the nonparametric approach can give us more flexibility

for model building. In the past time, one fact prevented nonparametric regression from wide application was the limited computing resource. However, the computing speeds of modern computers have been improved dramatically, and they are equipped with much larger high speed random access memory (RAM) nowadays. Various new algorithms have also been developed to speed up the computation. The nonparametric approach will be very useful when a parametric model is not sufficient. In the mean time, it can also serve as an automated diagnostic tool for parametric fitting. We will not review the general literature here, other than to note that the additive smoothing spline has been used by Heagerty & Zeger (1998) and Lin & Zhang (1999) for this purpose. Heagerty & Zeger (1998) used log odds ratio as a measurement of dependence and smoothing splines with fixed degrees of freedom. Their model was fitted by using Generalized Estimating Equation. Lin & Zhang (1999) proposed generalized additive mixed effect model and used smoothing splines to estimate the additive fixed effect terms.

Smoothing spline analysis of variance (SS-ANOVA) provides a general framework for multivariate nonparametric function estimation. It allows both main effects and interaction terms. These models have been studied extensively for Gaussian data. Recently, Lin (1998*b*) obtained some general convergence results for tensor product space ANOVA model and showed that smoothing spline ANOVA model achieves the optimal convergence rate. Wahba, Wang, Gu, Klein and Klein (1995, referred as WWGKK) gave a general setting for applying

smoothing spline ANOVA to data from exponential families. They successfully applied their method to analyze demographic medical data with Bernoulli outcomes. Lin (1998*a*) proposed to use SS-ANOVA to model data with polychotomous responses. Wang (1998*a*) developed mixed effect smoothing spline model for correlated Gaussian data. In this thesis, we will explore how to use smoothing spline ANOVA to model correlated multivariate Bernoulli data.

We will combine log-linear model and smoothing spline ANOVA model to obtain a partly flexible estimate of the joint distribution for multivariate Bernoulli data. It is of particular interest to us to explore the nonlinearity of the conditional logit functions. Conditional log odds ratio will be used to model the association among multivariate Bernoulli outcomes. We will still let log odds ratio take a simple parametric form and estimate it by using maximum likelihood estimation. An extension of $GACV$ proposed by Xiang & Wahba (1996) to multivariate responses will be used to choose smoothing parameters. We will iteratively estimate the conditional logit functions and log odds ratio until convergence.

1.2 Outline of the Thesis

In Chapter 2, we will review the log-linear model for multivariate Bernoulli observations and propose a smoothing spline ANOVA model to relax the parametric assumption. The existence and uniqueness of the nonlinear solution is investigated.

In Chapter 3, we discuss how to fit the penalized multivariate logistic regression model for a large data set. A numerical method combining the block one-step SOR-Newton-Ralphson algorithm and approximate smoothing spline is used to solve the variational problem for fixed smoothing parameters. We also proposed to use the iterated $ranGACV$ for multivariate Bernoulli data to select smoothing parameters adaptively. Simulation studies are conducted to illustrate the reasonable performance of the proposed algorithm.

In Chapter 4, we apply the proposed method to investigate the association between the pigmentary abnormalities and some risk factors for women in the Beaver Dam Eye Study. Finally, some discussions are given in Chapter 5.

Chapter 2

Penalized Multivariate Logistic Regression using Smoothing Spline ANOVA

2.1 Log-linear Model for Multivariate Bernoulli Data

Assuming there are J different endpoints, and K_j repeated measurements for the jth endpoint, let Y_{jk} denote the kth measurement of the jth endpoint. For example, in ophthalmological studies, we have two repeated measurement for each disease: left eye and right eye. In a typical longitudinal study, we have repeated measurements over the time. $Y = (Y_{jk}, j = 1, ..., J, k = 1, ..., K_j)$ is a multivariate Bernoulli outcome variable. Let $X_{jk} = (X_{jk1}, X_{jk2}, ..., X_{jkD})$ be a vector of predictor variables ranging over the subset $\mathcal{X}$ of $\mathcal{R}^D$, where X_{jkd} denotes the dth predictor variable for the kth measurement of the jth endpoint. Some predictor variables may take different values for different measurements

while others may be the same for all Y_{jk}'s. For example, in ophthalmology studies, there may be present both person-specific predictors and eye-specific predictors. The person-specific predictors are the same for each person while the eye-specific predictors may be different for the left and right eyes. Let $X = (X_{jk}, j = 1, ..., J, k = 1, ..., K_j)$. Then (X, Y) is a pair of random vectors. For a response vector $y = (y_{jk}, j = 1, ..., J, k = 1, ..., K_j)$, its joint probability distribution conditioning on the predictor variables X can be written as

$$\begin{aligned} P(Y = y|X) &= \exp\{\sum_{j=1}^{J}\sum_{k=1}^{K_j} f_{jk}y_{jk} + \sum_{j=1}^{J}\sum_{k_1<k_2} \alpha_{jk_1,jk_2}y_{jk_1}y_{jk_2} \\ &\quad + \sum_{j_1<j_2}\sum_{k_1,k_2} \alpha_{j_1k_1,j_2k_2}y_{j_1k_1}y_{j_2k_2} + ... \\ &\quad + \alpha_{11,12,...,JK_j}y_{11}y_{12}....y_{JK_j} - b(f,\alpha)\} \end{aligned} \tag{2.1.1}$$

where

$$\begin{aligned} b(f,\alpha) &= \log(1 + \sum_{j,k}\exp(f_{jk}) + \sum_{j_1,k_1}\sum_{j_2,k_2}\exp(f_{j_1k_1} + f_{j_2k_2} + \alpha_{j_1k_1,j_2k_2}) \\ &\quad + ... + \exp(\sum_{\text{all f}} f + \sum_{\text{all } \alpha} \alpha)) \end{aligned} \tag{2.1.2}$$

Let $M = \sum_{j=1}^{J} K_j$ be the length of the vector Y, there are in total $2^M - 1$ parameters: $(f, \alpha) = (f_{11}, f_{12}, ..., f_{JK_j}, \alpha_{11,12}, ..., \alpha_{11,12,...,JK_j})$, which may depend on X. The parameter space is unconstrained. They have straightforward interpretations in terms of conditional probabilities. For example,

$$f_{jk} = logit(P(Y_{jk} = 1|Y^{(-jk)} = 0, X)) \tag{2.1.3}$$

is the conditional logit function;

$$\alpha_{j_1k_1,j_2k_2} = \log OR(Y_{j_1k_1}, Y_{j_2k_2}|Y^{(-j_1k_1,-j_2k_2)} = 0, X) \tag{2.1.4}$$

is the conditional log odds ratio, which is a meaningful way to measure pairwise association;

$$\begin{aligned} & \alpha_{j_1k_1,j_2k_2,j_3k_3} \\ = \ & \log OR(Y_{j_1k_1}, Y_{j_2k_2}|Y_{j_3k_3}=1, Y^{(-j_1k_1,-j_2k_2,-j_3k_3)}=0, X) \\ - \ & \log OR(Y_{j_1k_1}, Y_{j_2k_2}|Y_{j_3k_3}=0, Y^{(-j_1k_1,-j_2k_2,-j_3k_3)}=0, X) \end{aligned} \tag{2.1.5}$$

is measuring three way association. Here $Y^{(-*)}$ denotes the subset of vector Y except Y_*, and

$$logit(p) = \log \frac{p}{1-p}, \tag{2.1.6}$$

$$OR(v,w) = \frac{P(v=1,w=1)P(v=0,w=0)}{P(v=1,w=0)P(v=0,w=1)}. \tag{2.1.7}$$

Now assume that we have n independent observations $(x_i, y_i), i = 1, ..., n$, where $y_i = (y_{i11}, y_{i12}, ..., y_{iJK_j})$ and $x_i = (x_{i11}, x_{i12}, ..., x_{iJK_j})$. Here y_{ijk} and $x_{ijk} = (x_{ijk1}, x_{ijk2}, ..., x_{ijkD})$ are the outcome variable and predictor vector for the kth measurement of the jth endpoint of the ith subject. From now on, we will use f_i and α_i to denote the parameters for the ith subject, while $y = (y_1, ..., y_n)$, $f = (f_1, ..., f_n)$ and $\alpha = (\alpha_1, ..., \alpha_n)$. We can write down the negative log likelihood function based on the observed data.

$$\begin{aligned} \mathcal{L}(y,f,\alpha) \ = \ & \sum_{i=1}^{n}\{\sum_{j=1}^{J}\sum_{k=1}^{K_j} f_{ijk}y_{ijk} + \sum_{j=1}^{J}\sum_{k_1<k_2} \alpha_{ijk_1,ijk_2}y_{ijk_1}y_{ijk_2} \\ & + \sum_{j_1<j_2}\sum_{k_1,k_2} \alpha_{ij_1k_1,ij_2k_2}y_{ij_1k_1}y_{ij_2k_2} + ... \\ & + \alpha_{i11,i12,...,iJK_j}y_{i11}y_{i12}....y_{iJK_j} - b(f_i,\alpha_i)\} \end{aligned} \tag{2.1.8}$$

We refer to equation (2.1.8) as the log-linear model for multivariate logistic regression. f_{ijk} is the conditional logit function for the kth measurement of the jth endpoint of the ith subject. Scientifically, except for that they may take different predictor values from measurement to measurement, there is little reason to believe they will take different functional form for the same endpoint. Hence we can assume $f_{ijk} = f_j(x_{ijk})$. Same reasoning applies to the association terms. For example, we can assume $\alpha_{ij_1k_1,ij_2k_2} = \alpha_{j_1j_2}(x_{ij_1k_1}, x_{ij_2k_2})$. The traditional parametric approach to fit the log-linear model is to assume linear relation between the parameters and predictors

$$f_{ijk} = f_j(x_{ijk}) = \beta_{j0} + \beta_{j1}x_{ijk1} + ... + \beta_{jD}x_{ijkD} \tag{2.1.9}$$

and so on. The model can be fitted efficiently by iterative proportional fitting (Bishop et al. 1975).

In practice, there are many ways to reduce the number of parameters to be estimated. For example, under many situations, scientific interest will be primarily focused on the conditional logit function f_{ijk} and log odds ratio $\alpha_{ij_1k_1,ij_2k_2}$, which measures pairwise association. The existence of three way association $\alpha_{ij_1k_1,ij_2k_2,ij_3k_3}$ and higher order associations are usually difficult to verify in practical situations, and may attract less scientific interest. Hence it is possible to set all higher order associations to be zero and only fit a parsimonious model instead of the saturated one described in (2.1.8). The reduced model is a member of the quadratic exponential model in Zhao & Prentice (1990).

2.2 The Variational Problem

In this thesis, we are interested in building flexible log-linear models. We are particularly interested in exploring the nonlinearity of the conditional logit functions f_j's. On the other hand, since it will take a very large number of observations to estimate many multivariate smooth functions simultaneously, this approach will still let the α's take a simple parametric form. In this section, to simplify the notation, we will consider a parsimonious model. Without loss of generality, except for the pairwise association, we will assume all higher order associations to be zero. Then the negative log likelihood function can be written as

$$\begin{aligned}
& \mathcal{L}(y, f, \alpha) \\
= \; & -\sum_{i=1}^{n} l_i(f(x_i), \alpha(x_i)) \\
= \; & -\sum_{i=1}\{\sum_{j=1}^{J}\sum_{k=1}^{K_j} f_j(x_{ijk})y_{ijk} + \sum_{j=1}^{J}\sum_{k_1<k_2} \alpha_{jj}(x_{ijk_1}, x_{ijk_2})y_{ijk_1}y_{ijk_2} \\
& + \sum_{j_1<j_2}\sum_{k_1,k_2} \alpha_{j_1j_2}(x_{ij_1k_1}, x_{ij_2k_2})y_{ij_1k_1}y_{ij_2k_2} - b(f_i, \alpha_i)\}
\end{aligned} \tag{2.2.1}$$

where

$$\begin{aligned}
& b(f_i, \alpha_i) \\
= \; & \log(1 + \sum_{j,k} \exp(f_j(x_{ijk})) \\
& + \sum_{j_1,k_1}\sum_{j_2,k_2} \exp(f_{j_1}(x_{ij_1k_1}) + f_{j_2}(x_{ij_2k_2}) + \alpha_{j_1j_2}(x_{ij_1k_1}, x_{ij_2k_2})) \\
& + \cdots + \exp(\sum_{j,k} f_j(x_{ijk}) + \sum_{j_1k_1}\sum_{j_2k_2} \alpha_{j_1j_2}(x_{ij_1k_1}, x_{ij_2k_2})))
\end{aligned} \tag{2.2.2}$$

We propose to use the penalized likelihood method to achieve greater flexibility in log-linear models. To relax the linear assumption, the penalized likelihood method (O'Sullivan 1983) only assumes the function to be estimated is smooth in some sense and imposes a certain roughness penalty on the function. Technically, a reproducing kernel Hilbert space (RKHS) is a Hilbert space of functions on $\mathcal{X}$ in which the evaluation functional is continuous (Aronszajn 1950). We will then assume $f_j \in \mathcal{H}^j$, where $\mathcal{H}^j$ is a reproducing kernel Hilbert space. The penalized multivariate logistic regression estimate of $f = (f_1, f_2, ..., f_J)$ and $\alpha = (\alpha_{11}, \alpha_{12}, ..., \alpha_{J,J})$ is the minimizer of the following variational problem

$$\mathcal{L}_\lambda(y, f, \alpha) = -\sum_{i=1}^{n} l_i(f(x_i), \alpha(x_i)) + \frac{n}{2}\mathbf{J}_\lambda(f_1, ..., f_J), \tag{2.2.3}$$

where the first part is the negative log likelihood and the second part is the roughness penalty. We will assume additive form of the penalty function for simplicity and easy interpretation:

$$\mathbf{J}_\lambda(f_1, ..., f_J) = \sum_{j=1}^{J} \lambda_j \mathbf{J}^j(f_j) \tag{2.2.4}$$

We consider the orthogonal decomposition $\mathcal{H}^j = \mathcal{H}_0^j \oplus \mathcal{H}_1^j$. Here $\mathcal{H}_0^j$ is finite dimensional (the "parametric" part, usually polynomials), and $\mathcal{H}_1^j$ (the "smooth" part) is the ortho-complement of $\mathcal{H}_0^j$ in $\mathcal{H}^j$. The penalty function will only be related to the smooth part of the function: $\mathbf{J}^j(f_j) = ||P_1^j f_j||^2$, where P_1^j is the orthogonal projection operator in $\mathcal{H}^j$ onto $\mathcal{H}_1^j$. The penalized likelihood

has the following expression:

$$\mathcal{L}_\lambda(y, f, \alpha) = -\sum_{i=1}^{n} l_i(f(x_i), \alpha(x_i)) + \sum_{j=1}^{J} \lambda_j ||P_1^j f_j||^2 \qquad (2.2.5)$$

The following theorem will show the existence and uniqueness of the solution to the variational problem (2.2.3). Denoting $\mathcal{H}_0 = \mathcal{H}_0^1 \times \cdots \times \mathcal{H}_0^J$ be the null space of $\mathcal{H}^1 \times \cdots \times \mathcal{H}^J$ with respect to the penalty function J_λ, the following theorem is true.

Theorem 2.1 *If the minimizer of (2.2.5) exists in $\mathcal{H}_0$, it uniquely exists in $\mathcal{H}^1 \times \cdots \times \mathcal{H}^J$*

Before we prove this theorem, we will first state two lemmas.

Lemma 2.1 *Let f_{ijk} denote $f_j(x_{ijk})$ and $\alpha_{ij_1k_1,ij_2k_2}$ denote $\alpha_{j_1j_2}(x_{ij_1k_1}, x_{ij_2k_2})$. $\mathcal{L}(y, f, \alpha)$ in (2.2.1) is a strictly convex function of f_{ijk}'s and $\alpha_{ij_1k_1,ij_2k_2}$'s.*

Proof We need to show the Hessian is positive definite. To simplify the notation, we will relabel $Y_i = (Y_{ijk})$ to be $(Y_{i1}, ..., Y_{iM})$, where $M = \sum_{j=1}^{J} K_j$. We simplify the notation for f's and α's similarly. From the property of exponential families, we know the Hessian with respect to f's and α's is $H = diag\{H_1, H_2, ..., H_n\}$, where H_i is the covariance matrix of $\tilde{Y}_i = (Y_{i1}, Y_{i2}, ..., Y_{iM}, Y_{i1}Y_{i2}, Y_{i1}Y_{i3}, ..., Y_{i,M-1}Y_{iM})^T$. Denoting $a_i = (a_{i1}, a_{i2}, ..., a_{iM}, a_{i12}, a_{i13}, ..., a_{i,M-1,M})^T$, if $a_i^T H_i a_i = var(a_i^T \tilde{Y}_i) = 0$, then we have $a_i^T \tilde{Y}_i = constant$. We will show a_i must be a zero vector. First, the constant here must be zero since we can let all Y_{im}'s be zero. To show $a_{im} = 0$, we will let $Y_{im} = 1$ and the rest

of vector $\tilde{Y}_i$ be zeroes. Afterwards, to derive $a_{im_1m_2} = 0$, we will let the only nonzero elements of the $\tilde{Y}_i$ vector be $Y_{im_1} = 1, Y_{im_2} = 1$ and $Y_{im_1}Y_{im_2} = 1$. This proof also extends to the saturated model. ∎

The following Lemma is Theorem 4.1 from Gu & Qiu (1993)

Lemma 2.2 *Suppose L(g) is a continuous and strictly convex functional in a Hilbert space* $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$, *where* $\mathcal{H}_1$ *has a square norm* $\mathbf{J}(g)$ *and* $\mathcal{H}_0$ *is the null space of* $\mathbf{J}(g)$ *of finite dimension. If* $L(g)$ *has a minimizer in* $\mathcal{H}_0$, *then* $L(g) + \mathbf{J}(g)$ *has a unique minimizer in* $\mathcal{H}$.

Proof of Theorem 2.1

Define

$$g(x_i, j_1, k_1, j_2, k_2) = \begin{cases} f_j(x_{ijk}), & 1 \le j = j_1 = j_2 \le J, \\ & 1 \le k = k_1 = k_2 \le K_j, f_j \in H^j \\ \alpha_{jj}(x_{ijk_1}, x_{ijk_2}) & 1 \le j = j_1 = j_2 \le J, \\ & 1 \le k_1 < k_2 \le K_j \\ \alpha_{j_1j_2}(x_{ij_1k_1}, x_{ij_2k_2}) & 1 \le j_1 < j_2 \le J, \\ & 1 \le k_1 \le K_{j_1}, 1 \le k_2 \le K_{j_2} \end{cases}.$$

Let $\mathcal{H} = \{g(x_i, j_1, k_1, j_2, k_2) : x_{ijk} \in \mathcal{X},\ 1 \le j_1 \le j_2 \le J,\ 1 \le k_1 \le K_{j_1},\ 1 \le k_2 \le K_{j_2}\}$. Then $\mathcal{H}$ is a Hilbert space with square semi-norm $\mathbf{J}_\lambda(g) = \mathbf{J}_\lambda(f_1, \ldots, f_J)$. Let $\mathcal{L}^*(g) = \mathcal{L}(y, f, \alpha)$. By Lemma 2.2, it suffices to show that $\mathcal{L}^*(g)$ is continuous and strictly convex in $\mathcal{H}$. Continuity is obvious. Strict convexity follows from Lemma 2.1. ∎

2.3 Smoothing Spline Analysis of Variance

Given a smooth multivariate function f on some domain $\mathcal{X}$, we are interested in decompose it into some component functions for the reason of easy interpretation and model building. A general ANOVA type decomposition is described in Chapter 10 of Wahba (1990) and Wahba, Wang, Gu, Klein & Klein (1995). To make any decomposition well defined, we assume that $\mathcal{M}$, a linear space of functions of x (the model space) which we assume contains f, can be decomposed as a direct sum of its subspaces.

$$\mathcal{M} = \mathcal{H}_0 \oplus \mathcal{H}_1 \oplus ... \oplus \mathcal{H}_q \tag{2.3.1}$$

Hence the decomposition of any $f \in \mathcal{M}$ into component functions in these subspaces is unique.

A unique ANOVA type decomposition can always be defined provided f satisfies some measurability conditions. Let $\mathcal{X}^{(\alpha)}$ be a measurable space. $d\mu_\alpha$ be a probability measure on $\mathcal{X}^{(\alpha)}$. Define the averaging operator $\mathcal{E}_\alpha$ on $\mathcal{X} = \mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(D)}$ as

$$(\mathcal{E}_\alpha)(x) = \int_{\mathcal{X}^{(\alpha)}} f(x_1, x_2, ..., x_D) d\mu_\alpha(x_\alpha) \tag{2.3.2}$$

Then the identity is decomposed as

$$\begin{aligned} I &= \prod_\alpha (\mathcal{E}_\alpha + (I - \mathcal{E}_\alpha)) \\ &= \prod_\alpha \mathcal{E}_\alpha + \sum_\alpha (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta + \sum_{\alpha < \beta} (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha, \beta} \mathcal{E}_\gamma \\ &\quad + ... + \prod_\alpha (I - \mathcal{E}_\alpha) \end{aligned} \tag{2.3.3}$$

The components of this decomposition generate the ANOVA decomposition of f in the following form

$$f(x_1, ..., x_d) = \mu + \sum_{\alpha=1}^{d} f_\alpha(x_\alpha) + \sum_{\alpha<\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + ... + f_{1,...,D}(x_1, ..., x_D), \tag{2.3.4}$$

where we have $\mu = (\prod_\alpha \mathcal{E}_\alpha)f$, $f_\alpha = ((I - \mathcal{E}_\alpha)\prod_{\beta\neq\alpha} \mathcal{E}_\beta)f$, $f_{\alpha\beta} = ((I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta} \mathcal{E}_\gamma)f$, and so forth.

The idea behind Smoothing Spline ANOVA is to construct a Reproducing Kernel Hilbert Space (RHKS) $\mathcal{H}$ of functions on $\mathcal{X}$ so that the components of the SS-ANOVA decomposition represent an orthogonal decomposition of f in $\mathcal{H}$. Let $\mathcal{H}^{(\alpha)}$ be an RKHS of functions on $\mathcal{X}^{(\alpha)}$ with $\int_{\mathcal{X}^{(\alpha)}} f_\alpha(x_\alpha)d\mu_\alpha = 0$ for $f_\alpha(x_\alpha) \in \mathcal{H}^{(\alpha)}$, and let $[1^{(\alpha)}]$ be the one dimensional space of constant functions on $\mathcal{X}^{(\alpha)}$. Construct $\mathcal{H}$ as the tensor product space

$$\mathcal{H} = \prod_{j=1}^{D}(\{[1^{(\alpha)}]\} \oplus \{\mathcal{H}^{(\alpha)}\}) = [1] \oplus \sum_\alpha \mathcal{H}^{(\alpha)} \oplus \sum_{\alpha<\beta}[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] \oplus \dots \tag{2.3.5}$$

where $[1]$ denotes the constant function on $\mathcal{X}$. With some abuse of notation, factors of the form $[1^\alpha]$ are omitted whenever they multiply a term of a different form. Thus $\mathcal{H}^{(\alpha)}$ is a shorthand for $[1^{(1)}]\otimes\cdots\otimes[1^{(\alpha-1)}]\otimes\mathcal{H}^{(\alpha)}[1^{(\alpha+1)}]\otimes\cdots\otimes[1^{(D)}]$ (which is a subspace of $\mathcal{H}$). The components of the ANOVA decomposition are now in mutually orthogonal subspaces of $\mathcal{H}$. Note that the components will depend on the measures $d\mu_\alpha$ and these should be chosen in specific application so that the fitted mean, main effects, two factor interactions, etc. have reasonable interpretations.

Next, $\mathcal{H}^{(\alpha)}$ is decomposed into a parametric part and a smooth part, by letting $\mathcal{H}^{(\alpha)} = \mathcal{H}_\pi^{(\alpha)} \oplus \mathcal{H}_S^{(\alpha)}$, where $\mathcal{H}_\pi^{(\alpha)}$ is finite dimensional (the "parametric" part) and $\mathcal{H}_S^{(\alpha)}$ (the "smooth" part) is the ortho-complement of $\mathcal{H}_\pi^{(\alpha)}$ in $\mathcal{H}^{(\alpha)}$. Elements of $\mathcal{H}_\pi^{(\alpha)}$ are not penalized through the device of letting $J_\alpha(f_\alpha) = \|P_S^{(\alpha)} f_\alpha\|^2$ where $P_S^{(\alpha)}$ is the orthogonal projector onto $\mathcal{H}_S^{(\alpha)}$. Now $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}]$ is a direct sum of four orthogonal subspaces: $[\mathcal{H}^{(\alpha)} \otimes \mathcal{H}^{(\beta)}] = [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_S^{(\beta)}] \oplus [\mathcal{H}_S^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}] \oplus [\mathcal{H}_S^{(\alpha)} \otimes \mathcal{H}_S^{(\beta)}]$. By convention the elements of the finite dimensional space $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}]$ will not be penalized. Continuing this way results in an orthogonal decomposition of $\mathcal{H}$ into sums of products of unpenalized finite dimensional subspaces, plus main effects "smooth" subspaces, plus two factor interaction spaces of the form parametric $\otimes$ smooth $[\mathcal{H}_\pi^{(\alpha)} \otimes \mathcal{H}_S^{(\beta)}]$, smooth $\otimes$ parametric $[\mathcal{H}_S^{(\alpha)} \otimes \mathcal{H}_\pi^{(\beta)}]$ and smooth $\otimes$ smooth $[\mathcal{H}_S^{(\alpha)} \otimes \mathcal{H}_S^{(\beta)}]$ and similarly for three and higher factor subspaces.

In practice, the series of ANOVA decomposition in (2.3.4) will be truncated at some point. Assuming that we have already decided which subspaces will be included in our model $\mathcal{M}(\subset \mathcal{H})$, we can regroup and write the model space as in (2.3.1). Usually we will let $\mathcal{H}_0$ be a finite dimensional space containing functions which are not going to be penalized. The norms on the composite $\mathcal{H}_l, 1 \leq l \leq q$ are the tensor product norms induced by the norms on the component subspaces, and $||f||^2 = ||P_0 f||^2 + \sum_{l=1}^q ||P_l f||^2$, where P_l is the orthogonal projector in $\mathcal{M}$ onto $\mathcal{H}_l$. Now we can use RKHS methods to explicitly impose roughness penalties. The smoothing spline ANOVA estimate of f in the Gaussian case is

the solution to the following variational problem

$$\min_{f\in\mathcal{M}}\{\sum_{i=1}^{n}(y_i - f(x_i))^2 + \frac{n}{2}\sum_{l=1}^{q}\lambda_l||P_l f||^2\}. \tag{2.3.6}$$

The first term in (2.3.6) is the sum of squared residuals which measures the goodness of fit while the second part is the penalty on roughness of the estimate. The λ_l's are smoothing parameters controlling the trade-off between goodness of fit and roughness. These smoothing parameters can be estimated from the data by the generalized cross validation (GCV) method or by the unbiased risk method (UBR) (see Wahba 1990).

2.4 Penalized Log-linear Model using Smoothing Spline ANOVA

We will use tensor product space and Smoothing Spline ANOVA to obtain a multivariate function estimate based on the variational problem (2.2.3). It is a direct generalization of (2.3.6) to multivariate Bernoulli observations.

Assume that we have already chosen a model space $\mathcal{M}_j = \mathcal{H}_0^j \oplus \sum_{l=1}^{q_j}\mathcal{H}_l^j$ for each conditional logit function f_j in (2.2.3), we can rewrite (2.2.5) as

$$\min_{f_j\in\mathcal{M}_j,\alpha_{j_1 j_2}}\{-\sum_{i=1}^{n} l_i(f(x_i),\alpha(x_i)) + \frac{n}{2}\sum_{j=1}^{J}\sum_{l=1}^{q_j}\lambda_{jl}||P_l^j f_j||^2\} \tag{2.4.1}$$

The first part in (2.4.1) measures the goodness of fit while the second part is roughness penalty in SS-ANOVA model. In the second part, P_l^j denotes the orthogonal projector in $\mathcal{M}_j$ onto the penalized subspace $\mathcal{H}_l^j$. The roughness

penalty for that subspace is the squared norm defined on that subspace $||P_l^j f||^2$. The λ_{jl}'s are the smoothing parameters which controls the bias-variance trade-off. Larger smoothing parameters will force the estimate into the parametric subspace while smaller ones will lead to more flexible estimate.

Let's define

$$||f||_{\Lambda_j}^2 = ||P_0^j f||^2 + \sum_{l=1}^{q_j} \lambda_{jl} ||P_l^j f_j||^2. \tag{2.4.2}$$

This is a modified but topologically equivalent norm on $\mathcal{M}_j$. indexed by Λ_j. Denoting the reproducing kernel for the subspace $\mathcal{H}_l^j$ associated with the original norm is R_l^j, we can show that $\lambda_{jl}^{-1} R_l^j$ is the RK for $\mathcal{H}_l^j$ associated with the modified norm.

The RK of the direct sum of orthogonal subspaces is the sum of the individual RK's. The RK of the tensor product space is the product of the RK's of the component spaces. Hence the computation for each R_l^j is straightforward. For example, if $R_{\mathcal{H}_\pi^{(d_1)}}(\cdot,\cdot)$ and $R_{\mathcal{H}_S^{(d_2)}}(\cdot,\cdot)$ are the RK corresponding to the Hilbert spaces $\mathcal{H}_\pi^{(d_1)}$ and $\mathcal{H}_S^{(d_2)}$ respectively, the RK corresponding to the tensor product space $\mathcal{H}_\pi^{(d_1)} \otimes \mathcal{H}_S^{(d_2)}$ is

$$R_{\mathcal{H}_\pi^{(d_1)}}(x_{d_1}(i_1), x_{d_1}(i_2)) R_{\mathcal{H}_S^{(d_2)}}(x_{d_2}(i_1), x_{d_2}(i_2)),$$

where $x_u(v)$ denotes the uth coordinate of the vth design point. Consequently, it can be shown that the RK for $\mathcal{M}_j$ under the modified norm is equal to

$$R_{j,\Lambda_j} = R_0^j(\cdot,\cdot) + \sum_{l=1}^{q_j} \lambda_{jl}^{-1} R_l^j(\cdot,\cdot). \tag{2.4.3}$$

In principle any positive-definite function may play the role of a reproducing kernel. Conditionally positive-definite functions as occur in thin plate spline can also be accommodated. One of the most commonly used penalty on $[0,1]$ is the square integral of the second derivative $\int_0^1 (f''(x))^2 dx$. Let $\mathcal{H}$ denote the Sobolev space $\{f | f, f' \text{absolutely continuous}, f'' \in \mathcal{L}_2\}$. We can decompose $\mathcal{H}$ into the direct sum of the unpenalized subspace $\mathcal{H}_0$ and the penalized subspace $\mathcal{H}_1$. A reproducing kernel of $\mathcal{H}_1$ with respect to the above penalty function can be written as

$$R(x, x') = k_2(x)k_2(x') - k_4([x - x']), \tag{2.4.4}$$

here $[\cdot]$ takes the fractional part of a number and

$$\begin{aligned} k_1(x) &= x - 1/2 \\ k_2(x) &= (k_1^2(x) - 1/12) \\ k_4(x) &= (k_1^4(x) - k_1^2(x)/2 + 7/240)/24. \end{aligned} \tag{2.4.5}$$

Furthermore, we have the relation

$$\int_0^1 (\frac{d^2}{dx^2}(\sum_{i=1}^n c_i R(x, x_i)))^2 dx = \sum_{i1=1}^n \sum_{i2=1}^n c_{i1} c_{i2} R(x_{i1}, x_{i2}). \tag{2.4.6}$$

Let $\phi_1(x) = 1$, $\phi_2(x) = k_1(x)$, then $\mathcal{H}_0 = span\{\phi_1(x), \phi_2(x)\}$. This penalty function and reproducing kernel is particularly useful in biostatistical applications. In practice, we can always rescale the original data points to the interval $[0,1]$.

Next, we will show that the minimizer of the variational problem (2.4.1) is actually within a finite dimensional linear space.

Theorem 2.2 *The solution to (2.4.1) has the form*

$$f_j(x) = \phi^j(x)^T d^j + \xi^j(x)^T c^j, \tag{2.4.7}$$

where c^j *and* d^j *are vectors of coefficients. Here* $\{\phi_v^j\}_{v=1}^{p_j}$ *is a set of basis functions spanning the null space* $\mathcal{H}_0^j$. $\phi^j(\cdot)^T = (\phi_1^j(\cdot), \cdots, \phi_{p_j}^j(\cdot))$. $\xi^j(\cdot)^T = (R_{j,\Lambda_j}(x_{1j1}, \cdot), \cdots, R_{j,\Lambda_j}(x_{1jK_j}, \cdot),\ R_{j,\Lambda_j}(x_{2j1}, \cdot), \cdots, R_{j,\Lambda_j}(x_{njK_j}, \cdot))$.

Proof See Wahba (1990). ∎

The above theorem states the fact that the minimizer in an infinite dimensional function space is actually a linear combination of a finite number of basis functions. Hence the computation of the minimizer is feasible. Substituting (2.4.7) into (2.4.1), we can estimate c^i and d^i by minimizing

$$\begin{aligned} & I_\lambda(c, d, \alpha) \\ = \ & -\sum_{i=1}^{n} l_i(\phi^1(x_i)^T d^1 + \xi^1(x_i)^T c^1, ..., \phi^J(x_i)^T d^J + \xi^J(x_i)^T c^J, \\ & \alpha_{11}(x_i), \alpha_{12}(x_i), ..., \alpha_{J,J}(x_i)) + \frac{n}{2} \sum_{j=1}^{J} c^{jT} Q_{j,\Lambda_j} c^j \end{aligned} \tag{2.4.8}$$

where Q_{j,Λ_j} is an $(nK_j \times nK_j)$ matrix

$$Q_{j,\Lambda_j} = \begin{pmatrix} Q_{j,11} & Q_{j,12} & \cdots & Q_{j,1n} \\ Q_{j,21} & Q_{j,22} & \cdots & Q_{j,2n} \\ \vdots & \vdots & \ddots & \vdots \\ Q_{j,n1} & Q_{j,n2} & \cdots & Q_{j,nn} \end{pmatrix}. \tag{2.4.9}$$

The definition of the $K_j \times K_j$ submatrix $Q_{j,i_1 i_2}$ is as following

$$Q_{j,i_1 i_2} = \begin{pmatrix} R_{j,\Lambda_j}(x_{i_1 j 1}, x_{i_2 j 1}) & R_{j,\Lambda_j}(x_{i_1 j 1}, x_{i_2 j 2}) & \dots & R_{j,\Lambda_j}(x_{i_1 j 1}, x_{i_2 j K_j}) \\ R_{j,\Lambda_j}(x_{i_1 j 2}, x_{i_2 j 1}) & R_{j,\Lambda_j}(x_{i_1 j 2}, x_{i_2 j 2}) & \dots & R_{j,\Lambda_j}(x_{i_1 j 2}, x_{i_2 j K_j}) \\ \vdots & \vdots & \ddots & \vdots \\ R_{j,\Lambda_j}(x_{i_1 j K_j}, x_{i_2 j 1}) & R_{j,\Lambda_j}(x_{i_1 j K_j}, x_{i_2 j 2}) & \dots & R_{j,\Lambda_j}(x_{i_1 j K_j}, x_{i_2 j K_j}) \end{pmatrix} \tag{2.4.10}$$

Since l_i's are not quadratic, solution of (2.4.8) does not have a closed form. In the next chapter, we will discuss how to obtain the estimate numerically. When the sample size is large, an approximate solution instead of the exact one will be obtained.

Chapter 3

Fitting the Penalized Multivariate Logistic Regression

3.1 Introduction

In this chapter, we will discuss how to numerically obtain the solution to the penalized multivariate logistic regression. Technically, Newton-Raphson algorithm can be used to obtain the solution because it is a quadratic convergent algorithm. However, the computational burden is extremely heavy. The parameters to be estimated according to Theorem 2.2 is about $\sum_{j=1}^{J}(p_j + nK_j)$. Consequently, the complexity for one step in Newton-Ralphson iteration is about $O((\sum_{j=1}^{J}(p_j + nK_j)^3)$, and the memory required to store the matrix is about $O((\sum_{j=1}^{J}(p_j + nK_j)^2)$. To reduce the computational burden, two methods are proposed here. The first one (Section 3.2) is an iterative method called block one-step SOR-Newton-Ralphson method. The convergence is super-linear. The complexity for one iteration is about $O(\sum_{j=1}^{J}(p_j + nK_j)^3)$. We sacrifice the convergent rate a little to reduce the computational complexity in each iteration. The second method (Section 3.4) is to obtain an approximate solution. Only a

small number of basis functions will be chosen for the final penalized regression step. It is shown in some special case that the approximate solution by using a subset of basis functions can achieve the same statistical convergence rate as the exact solution.

We will discuss a data-driven method to choose smoothing parameters in Section 3.5. For Gaussian data, two of the commonly recognized methods are the generalized cross validation (GCV) and the unbiased risk (UBR) methods (Wahba 1990). For general exponential family, Wahba et al. (1995) used iterated UBR method to choose smoothing parameters. Xiang & Wahba (1996) proposed generalized approximate cross validation ($GACV$). They reported that $GACV$ outperformed iterated UBR. This is further confirmed in Lin (1998a). In this thesis, we will extend $GACV$ to the case of multivariate Bernoulli responses. A randomized version for easy computation is also proposed. Combined with the block one-step SOR-Newton-Ralphson algorithm, $GACV$ will be used to choose smoothing parameters iteratively. Simulation studies show that it is an excellent computational proxy for the Comparative Kullback-Leibler (CKL) distance.

Bayesian "confidence intervals" were first proposed for smoothing spline with Gaussian data by Wahba (1983) and their properties were studied by Nychka (1988, 1990). Silverman (1985) provided another look at the Bayesian problem. Wahba et al. (1995) developed the componentwise approximate Bayesian "confidence intervals" for the non-Gaussian SS-ANOVA model. In Section 3.6, we

will identify the penalized likelihood estimation for multivariate logistic regression with a Bayesian problem. Based on this observation, approximate Bayesian "confidence intervals" were proposed for cross-validated smoothing spline estimates.

In the last section, to demonstrate the reasonable performance of smoothing spline estimates, we will show results from some simulation studies.

3.2 Block One-Step SOR Iteration

We will review how to use block one-step Successive Overrelaxation (SOR) method to solve a large nonlinear system in this section. Some convergence properties will also be discussed.

Assuming a large linear or nonlinear system we want to solve has m equations and m variables

$$\begin{cases} f_1(x_1, \cdots, x_m) &= 0 \\ \quad \vdots & \vdots \quad \vdots \\ f_m(x_1, \cdots, x_m) &= 0. \end{cases} \tag{3.2.1}$$

First let us assume this is a linear system. The Successive Overrelaxation Method, or SOR, is devised by applying extrapolation to the Gauss-Seidel method. This extrapolation takes the form of a weighted average between the previous iterate and the computed Gauss-Seidel iterate successively for each

component:

$$x_i^{(k+1)} = \omega \bar{x}_i^{(k+1)} + (1-\omega) x_i^{(k)} \tag{3.2.2}$$

where $\bar{x}_i^{(k+1)}$ is from a Gauss-Seidel iterate. This algorithm reduces to Gauss-Seidel algorithm when the relaxation (extrapolation) factor $\omega = 1$.

To derive the block SOR method, we need regroup the unknown $x = (x_1, x_2, \cdots, x_m)$ into p groups $(x^1, x^2, \cdots, x^p)$. Correspondingly, the m equations are also regrouped into p groups

$$\begin{cases} F_1(x^1, \cdots, x^p) &= 0 \\ \quad \vdots & \vdots \quad \vdots \\ F_p(x^1, \cdots, x^p) &= 0. \end{cases} \tag{3.2.3}$$

The updating formula for block SOR algorithm is

$$(x^i)^{(k+1)} = \omega (\bar{x}^i)^{(k+1)} + (1-\omega)(x^i)^{(k)} \tag{3.2.4}$$

where $(\bar{x}^i)^{(k+1)}$ is the successive Gauss-Seidel update for the ith linear system in (3.2.3)

$$F_i((x^1)^{(k+1)}, \cdots, (x^{i-1})^{(k+1)}, x^i, (x^{i+1})^{(k)}, \cdots, (x^p)^{(k)}) = 0. \tag{3.2.5}$$

In each iteration, we successively update the block component of x by the above method. This is repeated until some convergence criteria is met.

Now assuming that (3.2.1) is a nonlinear system. Hence in the updating formula (3.2.4), the successive Gauss-Seidel solution $(\bar{x}^i)^{(k+1)}$ of (3.2.5) can not be obtained explicitly. To solve the smaller nonlinear system (3.2.5), we need to

use some iterative method like Newton-Ralphson method. In this case, the process to solve a nonlinear system is called block nonlinear SOR-Newton-Ralphson method. See Ortega & Rheinboldt (1970) for details.

To simplify the nonlinear algorithm, we may only run the Newton-Ralphson iteration for one step to approximate the exact solution of (3.2.5), and use that as $(\bar{x}^i)^{(k+1)}$ in (3.2.4). This nonlinear SOR process is called block one-step SOR-Newton-Ralphson method. Specifically, the updating formula (3.2.4) now has the following expression

$$(x^i)^{(k+1)} = (x^i)^{(k)} - \omega \left[\frac{\partial F_i}{\partial x^i}(y^{(k,i)})\right]^{-1} F_i(y^{(k,i)}), \tag{3.2.6}$$

where

$$y^{(k,i)} = ((x^1)^{(k+1)}, \cdots, (x^{i-1})^{(k+1)}, (x^i)^{(k)}, \cdots, (x^l)^{(k)}).$$

In the statistics literature, the nonlinear system usually arises from a minimization or maximization problem in which we need to find a set of parameters to minimize (or maximize) a function. Specifically, suppose we are going to find $x \in R^m$ to minimize a twice differentiable multivariate function $g(x)$, then the updating formula for the block one-step SOR-Newton-Raphson method will become

$$(x^i)^{(k+1)} = (x^i)^{(k)} - \omega[\nabla^2_{ii} g(y^{(k,i)})]^{-1} \nabla_i g(y^{(k,i)}), \tag{3.2.7}$$

where $\nabla^2_{ii} g$ is the submatrix of the Hessian and $\nabla_i g$ is the sub-vector of the gradient.

In the next part, we will discuss some convergence properties for the general block nonlinear SOR and the block one-step SOR-Newton method. Define $F'(x) = D(x) - L(x) - U(x)$ to be the decomposition of $F'(x) = \partial F/\partial x$ into block diagonal, strictly block lower-triangular and strictly block upper-triangular parts, where

$$D(x) = \begin{pmatrix} \frac{\partial F_1}{\partial x^1} & 0 & \cdots & 0 \\ 0 & \frac{\partial F_2}{\partial x^2} & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \frac{\partial F_p}{\partial x^p} \end{pmatrix}. \tag{3.2.8}$$

For $\omega > 0$, let

$$H_\omega(x) = [D(x) - \omega L(x)]^{-1}[(1-\omega)D(x) + \omega U(x)]. \tag{3.2.9}$$

The local convergence of the block nonlinear SOR procedures is stated in the following lemma. The proof of this lemma can be found in Ortega & Rheinboldt (1970).

Lemma 3.3 (Local Convergence and Rate of Convergence) *Assume* $F : R^m \to R^m$ *be continuously differentiable over a compact set* S_0, *and* $x^* \in S_0$ *such that* $F(x^*) = 0$. *If* $D(x^*)$ *is nonsingular and* $\rho(H_\omega(x^*)) < 1$, *then there exists an open ball* $S = S(x^*, \delta)$ *in* S_0 *such that for any* $x^0 \in S$, *both the Block nonlinear SOR and the Block one-step SOR-Newton sequence converge to* x^*, *and they share the same convergent factor* $R_1(x^k, x^*) = \rho(H_\omega(x^*))$.

We will state the global convergence result in term of the minimization problem.

Lemma 3.4 (Global Convergence) *Assume* $g \in C^2(R^m)$, $\nabla^2 g(x) > 0$ *and* $S_0 = \{x | g(x) \leq g(x^0)\}$ *is bounded, then for suitable chosen relaxation parameter* ω*, the iterative sequence from the block one-step SOR-Newton method converges to the unique solution* x^**.*

The proof of the above lemma can be found in Schechter (1968). From the above lemma, we can see that in general the block one-step SOR-Newton-Raphson method with fixed ω is not guaranteed to converge globally. In practice, we can either change the initial value or tune the relaxation parameter to make the algorithm converge. The following lemma adapted from Varga (1984) can be used to check the conditions for the local convergence.

Lemma 3.5 *Let* $A = D - E - E^T$ *be a symmetric positive definite matrix, and* D *is also positive definite. Denote* $H_\omega = (D - \omega E)^{-1}((1-\omega)D + \omega E)$*. If* $D - \omega E$ *is nonsingular for* $0 \leq \omega \leq 2$*, then* $\rho(H_\omega) < 1$ *for* $0 < \omega < 2$*.*

The following Corollary is from Lin (1998*a*). It is obtained by directly applying the above lemma.

Corollary 3.1 *If* $A = D - E - E^T$ *is symmetric positive define and* D *is block diagonal matrix,* E *is strictly block lower triangular matrix. If* D *is nonsingular, then for* $0 < \omega < 2$*, we have* $\rho(H_\omega) < 1$*.*

According to Corollary 3.1, we note that if A is Hessian of a twice differentiable convex function, we will always have $\rho(H_\omega) < 1$ for $0 < \omega < 2$. Specifically, the local convergent property holds if we use block nonlinear SOR-Newton-Raphson

or block one-step SOR-Newton-Raphson method to find the minimizer of a twice differentiable convex function.

3.3 Implementation

In our implementation, we will keep $\alpha_{j_1 j_2}$'s as simple parametric forms. Consequently, we assume $\alpha_{j_1 j_2}$'s are depending on a set of parameters β's, which are to be estimated. Recall f_j depends on the coefficient vectors c^j and d^j. For simplicity reason, ω will be taken to be 1. The block one-step SOR-Newton-Ralphson algorithm for minimizing (2.4.8) is as following:

$\begin{pmatrix} c^j \\ d^j \end{pmatrix} \leftarrow$ initial values , $j = 1, ..., J$
$\beta \leftarrow$ initial values
do
    do j=1 to J
        $\begin{pmatrix} c^j \\ d^j \end{pmatrix} \leftarrow$ one-step Newton-Ralphson update for $\begin{pmatrix} c^j \\ d^j \end{pmatrix}$
    end
    $\beta \leftarrow$ Newton-Ralphson update for β
until (convergence)

Table 1: Block one-step SOR-Newton-Ralphson Algorithm

Notice that we only utilize one-step updating formula for f_j part in this implementation. Compared to the smoothing functions f_j's, the computational burden for the parametric part β's is relatively low. Therefore, we decide to run the Newton-Ralphson iteration until convergence in each step for β's.

Since the implementation of updating β's is straightforward, we will mainly describe how to update c^j and d^j in each step. To update c^j and d^j, the only relevant part of the likelihood in (2.2.1) is

$$l_j(f_j) = -\sum_{i=1}^{n}\{\sum_{k=1}^{K_j} f_j(x_{ijk})y_{ijk} - b(f_i, \alpha_i)\}. \tag{3.3.1}$$

The only relevant penalty term in (2.4.8) is

$$\mathbf{J}^j_{\Lambda_j}(f_j) = \frac{n}{2}{c^j}^T Q_{j,\Lambda_j} c^j. \tag{3.3.2}$$

With some abuse of notations, let $f_{ijk} = f_j(x_{ijk})$, $b_i = b(f_i, \alpha_i)$ and $Q_{j,\Lambda_j} = Q_j$. According to the property of exponential family, the following relations are true

$$\begin{aligned}
& \mu_{ijk} \\
= \;& \frac{\partial b_i}{\partial f_{ijk}} = EY_{ijk} \\
= \;& (e^{f_{ijk}} + \sum_{k_3 \neq k} e^{f_{ijk}+f_{ijk_3}+\alpha_{ijk,ijk_3}} + \sum_{j_3 \neq j}\sum_{k_3} e^{f_{ijk}+f_{ij_3k_3}+\alpha_{ijk,ij_3k_3}} + \cdots \\
& + e^{\sum_{j_3,k_3} f_{ij_3k_3} + \sum_{j_3,k_3}\sum_{j_4,k_4} \alpha_{ij_3k_3,ij_4k_4}})/ \\
& (1 + \sum_{j_3,k_3} e^{f_{ij_3k_3}} + \sum_{j_3,k_3}\sum_{j_4,k_4} e^{f_{ij_3k_3}+f_{ij_4k_4}+\alpha_{ij_3k_3,ij_4k_4}} + \cdots \\
& + e^{\sum_{j_3,k_3} f_{ij_3k_3} + \sum_{j_3,k_3}\sum_{j_4,k_4} \alpha_{ij_3k_3,ij_4k_4}})
\end{aligned} \tag{3.3.3}$$

$$\begin{aligned}
& w_{ijk,ijk} \\
= \;& \frac{\partial^2 b_i}{\partial f_{ijk}^2} = VarY_{ijk} \\
= \;& \mu_{ijk}(1-\mu_{ijk}),
\end{aligned} \tag{3.3.4}$$

$$
\begin{aligned}
& w_{ijk_1,ijk_2} \\
= \quad & \frac{\partial^2 b_i}{\partial f_{ijk_1} \partial f_{ijk_2}} = Cov(Y_{ijk_1}, Y_{ijk_2}) \\
= \quad & E(Y_{ijk_1} Y_{ijk_2}) - EY_{ijk_1} \cdot EY_{ijk_2} = \frac{\partial b_i}{\partial \alpha_{ijk_1,ijk_2}} - \mu_{ijk_1} \mu_{ijk_2} \\
= \quad & (e^{f_{ijk_1}+f_{ijk_2}+\alpha_{ijk_1,ijk_2}} + \cdots + e^{\sum_{j_3,k_3} f_{ij_3k_3} + \sum_{j_3,k_3}\sum_{j_4,k_4} \alpha_{ij_3k_3,ij_4k_4}})/ \\
& (1 + \sum_{j_3,k_3} e^{f_{ij_3k_3}} + \sum_{j_3,k_3}\sum_{j_4,k_4} e^{f_{ij_3k_3}+f_{ij_4k_4}+\alpha_{ij_3k_3,ij_4k_4}} + \cdots \\
& + e^{\sum_{j_3,k_3} f_{ij_3k_3} + \sum_{j_3,k_3}\sum_{j_4,k_4} \alpha_{ij_3k_3,ij_4k_4}}) - \mu_{ijk_1}\mu_{ijk_2}. \qquad (3.3.5)
\end{aligned}
$$

We introduce the following notations

$$
\begin{aligned}
u_{ijk} &= \frac{dl_j}{df_{ijk}} = -y_{ijk} + \mu_{ijk}, \\
u_j &= (u_{1j1}, u_{1j2}, ..., u_{1jK_j}, u_{2j1}, ..., u_{njK_j})^T \\
W_{ij} &= \begin{pmatrix} w_{ij1,ij1} & w_{ij1,ij2} & \cdots & w_{ij1,ijK_j} \\ w_{ij2,ij1} & w_{ij2,ij2} & \cdots & w_{ij2,ijK_j} \\ \vdots & \vdots & \ddots & \vdots \\ w_{ijK_j,ij1} & w_{ijK_j,ij2} & \cdots & w_{ijK_j,ijK_j} \end{pmatrix}, \\
W_j &= diag(W_{1j}, W_{2j}, ..., W_{nj}), \\
S_j &= \begin{pmatrix} \phi_1^j(x_{1j1}) & \phi_1^j(x_{1j1}) & \ldots & \phi_{p_j}^j(x_{1j1}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1^j(x_{1jK_j}) & \phi_1^j(x_{1jK_j}) & \ldots & \phi_{p_j}^j(x_{1jK_j}) \\ \phi_1^j(x_{2j1}) & \phi_1^j(x_{2j1}) & \ldots & \phi_{p_j}^j(x_{2j1}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1^j(x_{njK_j}) & \phi_1^j(x_{njK_j}) & \ldots & \phi_{p_j}^j(x_{njK_j}) \end{pmatrix}. \qquad (3.3.6)
\end{aligned}
$$

To update c^j and d^j, we only need to minimize part of the penalized likelihood

in (2.4.8), which is a summation of (3.3.1) and (3.3.2)

$$I_j = -\sum_{i=1}^{n}\{\sum_{k=1}^{K_j} f_{ijk}y_{ijk} - b_i\} + \frac{n}{2}{c^j}^T Q_j c^j. \tag{3.3.7}$$

Notice that in this expression, the smoothing parameters have already been absorbed into Q_j. This is a convex problem. According to Theorem (2.1), the minimizer of the above equation has the representation $f_j = S_j d^j + Q_j c^j$. Here $f_j = (f_{1j1},\ f_{1j2}, ...,\ f_{1jK_j}, f_{2j1}, ..., f_{njK_j})^T$ and S_j defined above is the collection of the parametric basis functions in (2.4.7). For one-step Newton-Ralphson updating formula, we need the following derivatives

$$\begin{aligned}
\frac{\partial I_j}{\partial c^j} &= Q_j u_j + nQ_j c^j, \\
\frac{\partial I_j}{\partial d^j} &= S_j^T u_j, \\
\frac{\partial^2 I_j}{\partial c^j \partial {c^j}^T} &= Q_j W_j Q_j + nQ_j, \\
\frac{\partial^2 I_j}{\partial d^j \partial {d^j}^T} &= S_j^T W_j S_j, \\
\frac{\partial^2 I_j}{\partial c^j \partial {d^j}^T} &= Q_j W_j S_j.
\end{aligned} \tag{3.3.8}$$

Hence the Block one-step SOR-Newton-Ralphson updating formula for coefficients (c^j, d^j) is

$$\begin{pmatrix} c^j \\ d^j \end{pmatrix} = \begin{pmatrix} c_-^j \\ d_-^j \end{pmatrix} - \begin{pmatrix} Q_j W_{j-} Q_j + nQ_j & Q_j W_{j-} S_j \\ S_j^T W_{j-} Q_j & S_j^T W_{j-} S_j \end{pmatrix}^{-1} \begin{pmatrix} Q_j u_{j-} + nQ_j c_-^j \\ S_j^T u_{j-} \end{pmatrix}, \tag{3.3.9}$$

where the subscript minus indicates the quantities evaluated at the latest update. By rearranging the above formula, c^j and d^j is the solution of the following

linear system

$$\begin{pmatrix} Q_j W_{j-} Q_j + nQ_j & Q_j W_{j-} S_j \\ S_j^T W_{j-} Q_j & S_j^T W_{j-} S_j \end{pmatrix} \begin{pmatrix} c^j - c_-^j \\ d^j - d_-^j \end{pmatrix} = \begin{pmatrix} -Q_j u_{j-} - nQ_j c_-^j \\ -S_j^T u_{j-} \end{pmatrix}. \tag{3.3.10}$$

Another equivalent representation is

$$\begin{pmatrix} Q_j W_{j-} Q_j + nQ_j & Q_j W_{j-} S_j \\ S_j^T W_{j-} Q_j & S_j^T W_{j-} S_j \end{pmatrix} \begin{pmatrix} c^j \\ d^j \end{pmatrix} = \begin{pmatrix} Q_j W_{j-} f_{j-} - Q_j u_{j-} \\ S_j^T W_{j-} f_{j-} - S_j^T u_{j-} \end{pmatrix}. \tag{3.3.11}$$

According to Theorem (2.1), $f_j = S_j d^j + Q_j c^j$ is always unique as long as S_j's are of full column rank. If Q_j is nonsingular, the above linear systems are equivalent to

$$\begin{pmatrix} W_{j-} Q_j + nI & W_{j-} S_j \\ S_j^T & 0 \end{pmatrix} \begin{pmatrix} c^j \\ d^j \end{pmatrix} = \begin{pmatrix} W_{j-} f_{j-} - u_{j-} \\ 0 \end{pmatrix}. \tag{3.3.12}$$

If Q_j is singular, any solution to (3.3.12) is also a solution to (3.3.10) and (3.3.11). Define $\tilde{Q}_j = W_{j-}^{1/2} Q_j W_{j-}^{1/2}$, $\tilde{S}_j = W_{j-}^{1/2} S_j$, $\tilde{c}^j = W_{j-}^{-1/2} c^j$, $\tilde{d}^j = d^j$ and $\tilde{\tilde{y}}_j = W_{j-}^{1/2}(f_{j-} - W_{j-}^{-1} u_{j-})$, (3.3.12) can be simplified as

$$\begin{cases} (\tilde{Q}_j + nI)\tilde{c}^j + \tilde{S}_j \tilde{d}^j &= \tilde{\tilde{y}}_j \\ \tilde{S}_j^T \tilde{c}^j &= 0 \end{cases} \tag{3.3.13}$$

It is easy to see that the solution of (3.3.12) gives the minimizer of

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} (\tilde{\tilde{y}}_{ij} - W_{ij-}^{1/2} f_{ij})^T (\tilde{\tilde{y}}_{ij} - W_{ij-}^{1/2} f_{ij}) + \tilde{c}^{jT} \tilde{Q}_j \tilde{c}^j \\ = \; & \frac{1}{n} \sum_{i=1}^{n} (\tilde{y}_{ij} - f_{ij})^T W_{ij-} (\tilde{y}_{ij} - f_{ij}) + c^{jT} Q_j c^j. \end{aligned} \tag{3.3.14}$$

With abuse of notations, we use u_{ij} to denote $(u_{ij1}, ..., u_{ijK_j})^T$ and f_{ij} to denote $(f_{ij1}, ..., f_{ijK_j})^T$, etc. $\tilde{y}_{ij} = f_{ij-} - W_{ij-}^{-1} u_{ij-}$ are called the **pseudo-data**. The block one-step SOR-Newton-Ralphson procedure iteratively reformulates the problem to estimate f_j from the pseudo-data by weighted penalized least squares.

The following theorem will show the pseudo-data approximately have the usual data structure if f_{j-} are not far away from f_j. This observation will later be used to construct the approximate Bayesian confidence intervals.

Theorem 3.3 *For fixed j, if $|f_{ijk-} - f_{ijk}| = o(1)$ uniformly in $i = 1, 2, ..., n$ and $k = 1, ..., K_j$, $|\alpha_- - \alpha| = o(\mathbf{1})$ uniformly, $\mu_j(x)$ is uniformly bounded away from 0 and 1, α's are uniformly bounded away from $-\infty$ and ∞, then*

$$\tilde{y}_{ij} = f_{ij} + \epsilon_{ij} + o_p(\mathbf{1})$$

where $\epsilon_{ij} = (\epsilon_{ij1}, ..., \epsilon_{ijK_j})^T$ has mean 0 and covariance matrix W_{ij}^{-1}, and $\epsilon_{1j}, \cdots, \epsilon_{nj}$ are independent.

Proof Denote $E(y_{ijk}) = \mu_{ijk}$, $Var(y_{ij}) = W_{ij}$ and $u_{ij} = -y_{ij} + \mu_{ij}$. Here μ_{ijk} is the shorthand for $\mu_j(x_{ijk})$. Then we have $E(W_{ij}^{-1} u_{ij}) = 0$ and $Var(W_{ij}^{-1} u_{ij}) = W_{ij}^{-1}$. Take the difference

$$\begin{aligned} \gamma &= (f_{ij-} - W_{ij-}^{-1} u_{ij-}) - (f_{ij} - W_{ij}^{-1} u_{ij}) \\ &= (f_{ij-} - f_{ij}) - (W_{ij-}^{-1}(-y_{ij-} + \mu_{ij-}) - W_{ij}^{-1}(-y_{ij} + \mu_{ij})) \end{aligned}$$

The expectation of γ is $(f_{ij-} - f_{ij}) - W_{ij-}^{-1}(\mu_{ij-} - \mu_{ij})$. Since $\mu_j(x)$ is uniformly bounded away from 0 and 1, α's are uniformly bounded away from $-\infty$ and

∞, it is easy to see that f_{ijk}'s are also uniformly bounded away from $-\infty$ and ∞. From $|f_{ijk-} - f_{ijk}| = o(1)$, $|\alpha_- - \alpha| = o(\mathbf{1})$ uniformly, we also have $|\mu_{ijk-} - \mu_{ijk}| = o(1)$ uniformly. The element of W_{ij-} also converges to the corresponding element of W_{ij} uniformly $|w_{ijk_1,ijk_2-} - w_{ijk_1,ijk_2}| = o(1)$.

Next, we will show there exist two constants $0 < c_1 < c_2 < \infty$ such that all eigenvalues of W_{ij} are in the interval (c_1, c_2). W_{ij} as a covariance matrix is positive definite. All of its eigenvalues are positive. Its trace is less than $K_j/4$. Note the trace of a matrix equals the summation of all of its eigenvalues. Hence its largest eigenvalue is also less than $K_j/4$. The smallest eigenvalue of W_{ij} as a function of f and α is continuous and always greater than zero. Its domain $(\mathcal{F}, \mathcal{A})$ is bounded and closed hence a compact set. There exists $c_1 > 0$ such that the smallest eigenvalue of W_{ij} is greater than c_1 for all $(f, \alpha) \in (\mathcal{F}, \mathcal{A})$.

Hence for $n > n_1$ (n_1 does not depend on f and α), the smallest eigenvalue of W_{ij-} is also greater than $c_1/2$. Consequently the largest eigenvalue of W_{ij-}^{-1} is bounded away from ∞. As a result, we have $E(\gamma) = o(\mathbf{1})$. Meanwhile, $Var(\gamma_k) < tr(Cov(\gamma)) < K_j||Cov(\gamma)||$. And for $n > n_1$,

$$\begin{aligned}
||Cov(\gamma)|| &= ||(W_{ij-}^{-1} - W_{ij}^{-1})W_{ij}(W_{ij-}^{-1} - W_{ij}^{-1})|| \\
&= ||W_{ij-}^{-1}(W_{ij} - W_{ij-})W_{ij}^{-1} \cdot W_{ij} \cdot W_{ij}^{-1}(W_{ij} - W_{ij-})W_{ij-}^{-1}|| \\
&\leq ||W_{ij-}^{-1}||^2 \cdot ||W_{ij}^{-1}|| \cdot ||W_{ij} - W_{ij-}||^2 \\
&\leq \frac{4}{c_1^3}||(W_{ij} - W_{ij-})^2|| \\
&\leq \frac{4}{c_1^3} tr(W_{ij} - W_{ij-})^2 = o(1). \qquad (3.3.15)
\end{aligned}$$

Hence the diagonal elements of $Cov(\gamma)$ go to zero uniformly. Consequently,

$$\tilde{y}_{ij} = f_{ij-} - W_{ij-}^{-1} u_{ij-} = f_{ij} - W_{ij}^{-1} u_{ij} + \gamma = f_{ij} + \epsilon_{ij} + o_p(\mathbf{1}),$$

where $\epsilon_{ij} = -W_{ij}^{-1} u_{ij}$ has mean 0 and covariance matrix W_{ij}^{-1}. The independence of ϵ_{ij}'s follows from the independence of y_{ij}'s. ∎

All of the previous discussions assume no special structure in the design points. The algorithm is specifically designed to handle the unstructured case. However, when special structure is available, the above algorithm can be simplified. One common case is the presence of person-specific covariates only. Hence $x_{ijk} = x_{ij}$ for all $k = 1, ..., K_j$. Similarly $f_{ijk} = f_j(x_{ijk}) = f_j(x_{ij}) = f_{ij}$. To update f_j, the part of the penalized likelihood needs to be minimized has the simplified form

$$I_j = -\sum_{i=1}^{n} \{(\sum_{k=1}^{K_j} y_{ijk}) f_{ij} - b_i\} + \frac{n}{2} c^{jT} Q_j c^j. \tag{3.3.16}$$

Now define

$$Q_j = \begin{pmatrix} R_{j,\Lambda_j}(x_{1j}, x_{1j}) & R_{j,\Lambda_j}(x_{1j}, x_{2j}) & \dots & R_{j,\Lambda_j}(x_{1j}, x_{nj}) \\ R_{j,\Lambda_j}(x_{2j}, x_{1j}) & R_{j,\Lambda_j}(x_{2j}, x_{2j}) & \dots & R_{j,\Lambda_j}(x_{2j}, x_{nj}) \\ \vdots & \vdots & \ddots & \vdots \\ R_{j,\Lambda_j}(x_{nj}, x_{1j}) & R_{j,\Lambda_j}(x_{nj}, x_{2j}) & \dots & R_{j,\Lambda_j}(x_{nj}, x_{nj}) \end{pmatrix},$$

$$S_j = \begin{pmatrix} \phi_1^j(x_{1j}) & \phi_2^j(x_{1j}) & \dots & \phi_{p_j}^j(x_{1j}) \\ \phi_1^j(x_{2j}) & \phi_2^j(x_{2j}) & \dots & \phi_{p_j}^j(x_{2j}) \\ \vdots & \vdots & \ddots & \vdots \\ \phi_1^j(x_{nj}) & \phi_2^j(x_{nj}) & \dots & \phi_{p_j}^j(x_{nj}) \end{pmatrix}. \tag{3.3.17}$$

The minimizer of (3.3.16) has the representation $f_j = S_j d^j + Q_j c^j$, too. Denote $y_{ij} = \sum_{k=1}^{K_j} y_{ijk}$, $\mu_{ij} = E(\sum_{k=1}^{K_j} Y_{ijk})$, $W_{ij} = Var(\sum_{k=1}^{K_j} Y_{ijk})$, $u_{ij} = -\sum_{k=1}^{K_j} y_{ijk} + \mu_{ij}$ and $u_j = (u_{1j}, u_{2j}, \cdots, u_{nj})^T$. Except for the above changes, all of the previous formulae and discussions remain true. But in the each iteration, we only need to solve an $n \times n$ system instead of an $(nK_j \times nK_j)$ one.

3.4 Approximate Smoothing Spline for Large Data Sets

As mentioned before, in each block-one-step SOR-Newton-Ralphson iteration, we need about $O(n^3)$ computing time and $O(n^2)$ memory space. However, the "true" function f_j to be estimated may not be very "complex". Hence it may be well approximated in the span of a much smaller subset of the basis functions. Therefore, this approach will take much smaller computer memory and shorter running time. This approach is particularly useful for analyzing medical data, where the underlying truth is believed to be quite "smooth".

3.4.1 An Approximate Solution

To obtain an approximate solution, a subset of basis functions needs to be chosen carefully. The variational problem is then solved in this lower dimensional subspace. This approach was proposed by Wahba (1980) for thin-plate splines. Luo & Wahba (1997) proposed hybrid adaptive spline. Xiang (1996) proposed

to use clustering method to choose the subset of basis functions. We will follow Xiang's approach here for selecting basis functions.

The basis function $\xi_{ijk}^{j}(\cdot) = R_{j,\Lambda_j}(x_{ijk}, \cdot)$ in (2.4.7) is the representer of design point x_{ijk} in the Reproducing Kernel Hilbert Space $\mathcal{M}_j^1$. Usually, when the design points are close, their representers are also very close. Hence, when the data set is large, it is very likely that lots of the basis functions will be nearly linearly dependent. On the other hand, if by some "prior" knowledge, it is believed that the structure of the true f_j is not very complicated, then it may be well approximated by a small number of basis functions. As a result, if we select the design points having maximum separation, their corresponding representers are expected to have less correlation.

Considering this problem from another point of view, the object is to group design points into several groups. Ideally, those groups should be spaced as far as possible from each other. Thus, we can borrow the classical cluster analysis technique to solve this problem. There are many algorithms for clustering the data. Even though there is no natural separation among design points in our case, we still can force the algorithm to run. SAS procedure FASTCLUS is designed for the disjoint clustering of very large data sets in minimum time. We will use it to separate the data sets into several clusters. Within each cluster, we randomly select the representer of one data point to form the approximating subspace.

Hence, as an iterative procedure, the algorithm for approximate spline is

as follows. When the number of basis functions V increases, the approximate solution converges to the exact solution.

$V \leftarrow$ initial value
do
    Cluster the data points into V groups
    Randomly select one data point from each group
    Generate the corresponding basis functions

    $f_j \leftarrow$ initial values, $j = 1, 2, ..., J$
    $\beta \leftarrow$ initial values
    do
        do $j = 1$ to J
            $f_j \leftarrow$ updated values in the approximating subspace
        end
        $\beta \leftarrow$ Newton-Raphson update for β
    until (convergence)

    $V \leftarrow 2 \times V$
until ($\frac{||f_{new} - f_{old}||}{||f_{old}||} < prec_1$ and $\frac{||\beta_{new} - \beta_{old}||}{||\beta_{old}||} < prec_2$)

Table 2: Iterative Algorithm for Approximate Spline

Here $prec_1$ and $prec_2$ are pre-specified thresholds. We suggest that the initial value for V be at least 25. The above algorithm usually converge very rapidly. From our experience, for medical data, 50 to 100 basis functions usually yield very good approximation.

Next, we will discuss the block one-step SOR updating formula for approximate spline. Assume for fixed V, we have selected V data points, which are indexed as $x_{j,v}$ for $v = 1, ..., V$. Their corresponding basis functions are $\xi_{j,v}(\cdot) = R_{j,\Lambda_j}(x_{j,v}, \cdot)$. We will still use S_j to denote the collection of basis

functions for parametric subspace. For approximating smooth subspace, denote

$$Q_{j,V} = \begin{pmatrix} \xi_{j,1}(x_{1j1}) & \xi_{j,2}(x_{1j1}) & \cdots & \xi_{j,V}(x_{1j1}) \\ \xi_{j,1}(x_{1j2}) & \xi_{j,2}(x_{1j2}) & \cdots & \xi_{j,V}(x_{1j2}) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{j,1}(x_{njK_j}) & \xi_{j,2}(x_{njK_j}) & \cdots & \xi_{j,V}(x_{njK_j}) \end{pmatrix},$$

$$Q^*_{j,V} = \begin{pmatrix} \xi_{j,1}(x_{j,1}) & \xi_{j,2}(x_{j,1}) & \cdots & \xi_{j,V}(x_{j,1}) \\ \xi_{j,1}(x_{j,2}) & \xi_{j,2}(x_{j,2}) & \cdots & \xi_{j,V}(x_{j,2}) \\ \vdots & \vdots & \ddots & \vdots \\ \xi_{j,1}(x_{j,V}) & \xi_{j,2}(x_{j,V}) & \cdots & \xi_{j,V}(x_{j,V}) \end{pmatrix}. \tag{3.4.1}$$

Let $c^j_V = (c_{j,1}, c_{j,2}, ..., c_{j,V})^T$. With abuse of notation, the approximate solution has the representation $f_j = S_j d^j + Q_{j,V} c^j_V$. It is easy to verify that the penalty for f_j has the quadratic form $||P_1 f_j||^2_{\Lambda_j} = {c^j_V}^T Q^*_{j,V} c^j_V$. Therefore, to update f_j, the variational problem is to minimize

$$I_{j,V} = -\sum_{i=1}^{n} \{\sum_{k=1}^{K_j} f_{ijk} y_{ijk} - b_i\} + \frac{n}{2} {c^j_V}^T Q^*_{j,V} c^j_V. \tag{3.4.2}$$

The one-step updating formula corresponding to (3.3.10) is to solve

$$\begin{pmatrix} Q^T_{j,V} W_{j-} Q_{j,V} + n Q^*_{j,V} & Q^T_{j,V} W_{j-} S_j \\ S^T_j W_{j-} Q_{j,V} & S^T_j W_{j-} S_j \end{pmatrix} \begin{pmatrix} c^j_V - c^j_{V-} \\ d^j - d^j_- \end{pmatrix} = \begin{pmatrix} -Q^T_{j,V} u_{j-} - n Q^*_{j,V} c^j_{V-} \\ -S^T_j u_{j-} \end{pmatrix}. \tag{3.4.3}$$

In practice, it is highly possible that the coefficient matrix of the linear system (3.4.3) would be computationally singular even if it is nonsingular in theory. In order to obtain a numerically stable solution, QR factorization with pivoting

is performed. In the meantime, a cutoff parameter τ (such as the machine precision times the largest absolute diagonal element of the R matrix) is specified. Let r_{ii} denote the diagonal element of the R matrix in the QR decomposition. Whenever $|r_{ii}| < \tau$, the corresponding solution in the coefficients vector c_V^j is set to be zero.

3.4.2 The Convergence Rate

In this section, we will prove in a special case, to achieve the same statistical convergence rate, the approximate spline only need a small portion of the basis functions compared to the exact solution. More general result is also believed to be true and it is one of my future research topic.

The special case treated here is the one dimensional smoothing spline estimate for Gaussian data. The classical variational problem to be solved is

$$\min_f \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||P_1 f||^2. \tag{3.4.4}$$

It is well known that for the penalty function $||P_1 f||^2 = \int_0^1 f^{(m)}(x)^2 dx$, for roughly equally spaced data on $(0,1)$, the statistical convergence rate for smoothing spline estimate is $O_p(n^{-\frac{2m}{2m+1}})$. We will demonstrate that in order to match the same convergence rate, V, the number of basis functions in the approximating space, only need to grow at a rate of $O(n^{\frac{2m}{(2m+1)(2m-1)}})$. This is a much smaller number compared to n when n is large. The proof is based on the following two lemmas. However, these lemmas are more general. They do not require the one dimensional assumption.

Assume the functional space can be decomposed into the direct sum of a parametric subspace and a smooth subspace. $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. We will use the following notations. Let the columns of S be the parametric basis functions in $\mathcal{H}_0$, the columns of Q be the smooth basis functions which are the representers of the evaluation functionals of all data points in $\mathcal{H}_1$. Hence the solution of (3.4.4) lies in the finite dimensional space $span\{S, Q\}$. Let Q_V denote the collection of a subset of all basis functions in Q. The approximating subspace is $span\{S, Q_V\}$. We will use P_1 to denote the projection into $\mathcal{H}_1$ under the original norm $||\cdot||$. P_{V*} is the projection into the approximating subspace $span\{S, Q_V\}$ under the modified norm $||\cdot||_*$. We will use $<\cdot,\cdot>$ to denote the inner product induced by the original norm while $<\cdot,\cdot>_*$ is used to denote the inner product induced by the modified norm.

The following lemma shows given the exact solution, how to calculate the approximate solution.

Lemma 3.6 *For fixed λ, denote $f = Sd+Qc$ to be the exact solution of the variational problem (3.4.4). Define a new norm $||f||_*^2 = \frac{1}{n}\sum_{i=1}^n f(x_i)^2 + \lambda||P_1 f||^2$. The approximate spline solution of (3.4.4) in the subspace $span\{S, Q_V\}$ is $f_* = P_{V*}(f)$, where P_{V*} denotes the projection into the subspace $span\{S, Q_V\}$ under the norm $||\cdot||_*$.*

Proof It is easy to check $||\cdot||_*$ is a valid norm in the space $span\{S, Q\}$. Under this norm, we have the following decomposition $f = f_* \oplus \rho_*$, where $< f_*, \rho_* >_* =$

0. Hence,

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 + \lambda||P_1 f||^2 \\
= \; & \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i f(x_i) + \frac{1}{n}\sum_{i=1}^{n} f(x_i)^2 + \lambda||P_1 f||^2 \\
= \; & \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i f(x_i) + ||f||_*^2 \\
= \; & \frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i (f_*(x_i) + \rho_*(x_i)) + ||f_*||_*^2 + ||\rho_*||_*^2 \\
= \; & (\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i f_*(x_i) + ||f_*||_*^2) \\
& +(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i \rho_*(x_i) + ||\rho_*||_*^2) \\
& -\frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
= \; & (\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i f_*(x_i) + \frac{1}{n}\sum_{i=1}^{n} f_*(x_i)^2 + \lambda||P_1 f_*||^2) \\
& +(\frac{1}{n}\sum_{i=1}^{n} y_i^2 - \frac{2}{n}\sum_{i=1}^{n} y_i \rho_*(x_i) + \frac{1}{n}\sum_{i=1}^{n} \rho_*(x_i)^2 + \lambda||P_1 \rho_*||^2) \\
& -\frac{1}{n}\sum_{i=1}^{n} y_i^2 \\
= \; & (\frac{1}{n}\sum_{i=1}^{n}(y_i - f_*(x_i))^2 + \lambda||P_1 f_*||^2) \\
& +(\frac{1}{n}\sum_{i=1}^{n}(y_i - \rho_*(x_i))^2 + \lambda||P_1 \rho_*||^2) - \frac{1}{n}\sum_{i=1}^{n} y_i^2 \qquad (3.4.5)
\end{aligned}
$$

Therefore if f is the minimizer of (3.4.4) in $span\{S, Q\}$, then f_* must be the minimizer of (3.4.4) in $span\{S, Q_V\}$, ρ_* must be the minimizer of (3.4.4) in in $\mathcal{H}_*$, where $span\{S, Q\} = span\{S, Q_V\} \oplus \mathcal{H}_*$ w.r.t. the norm $||\cdot||_*$. ∎

The next lemma gives an easy to handle upper bound for the difference ρ_* appeared in the above lemma.

Lemma 3.7 *For fixed λ, suppose f is the exact solution of (3.4.4), $f_* = P_{V*}(f)$ is the approximate solution in the subspace span$\{S, Q_V\}$, let $f^* = P_V(f)$ where P_V is the projection of f into the subspace span$\{S, Q_V\}$ under the original norm $||\cdot||$. Let $\rho^* = f - f^*$, $\rho_* = f - f_*$, we have the following relation:*

$$\frac{1}{n}\sum_{i=1}^{n} \rho_*(x_i)^2 \leq \frac{1}{n}\sum_{i=1}^{n} \rho^*(x_i)^2$$

Proof From Lemma 3.6, we know that $f_* = P_{V*}f$. Since $f = f_* + \rho_*$, let $\rho_{*0} = P_V(\rho_*)$ and $\rho_{*1} = \rho_* - \rho_{*0}$, then $f = (f_* + \rho_{*0}) + \rho_{*1}$, where $(f_* + \rho_{*0}) \in$ *span*$\{S, Q_V\}$, ρ_{*1} is orthogonal to *span*$\{S, Q_V\}$ under the original norm $||\cdot||$. Hence by a different way, we obtain the same decomposition as $f = f^* + \rho^*$. Therefore, actually $\rho_{*1} = \rho^*$ and $\rho_* = \rho_{*0} \oplus \rho^*$ under the original norm. Thus we conclude $||\rho_*||^2 \geq ||\rho^*||^2$. In fact, $||P_1(\rho_*)||^2 \geq ||P_1(\rho^*)||^2$ is also true since $P_1(\rho_*) = P_1(\rho_{*0}) \oplus P_1(\rho^*)$ under the original norm.

Similarly, we have $||\rho^*||_*^2 \geq ||\rho_*||_*^2$. By combining these two facts, we have

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} \rho_*(x_i)^2 + \lambda||P_1(\rho_*)||^2 &\leq \frac{1}{n}\sum_{i=1}^{n} \rho^*(x_i)^2 + \lambda||P_1(\rho^*)||^2 \\ ||P_1(\rho^*)||^2 &\leq ||P_1(\rho_*)||^2 \end{aligned} \tag{3.4.6}$$

Hence $(1/n)\sum_{i=1}^{n} \rho_*(x_i)^2 \leq (1/n)\sum_{i=1}^{n} \rho^*(x_i)^2$. ∎

Before we prove the next theorem, we will review some basic properties of the projection f^*.

Let $Q = (\xi_1, \xi_2, ..., \xi_n)$. Without loss of generality, let $Q_V = (\xi_1, \xi_2, ..., \xi_V)$ be the collection of first V columns in Q. ξ_i is the representer of the evaluation functional of the ith data point in the reproducing kernel Hilbert space $\mathcal{H}_1$. $\xi_i = P_1\eta_i$, where η_i is the representer of the evaluation functional of the ith data point in the reproducing kernel Hilbert space $\mathcal{H}$. Hence for any function $f \in \mathcal{H}$, $f(x_i) =< f, \eta_i >$ and $P_1 f(x_i) =< f, \xi_i >$.

In the above lemma, f is decomposed into the direct sum of f^* and ρ^*, where ρ^* is orthogonal to the approximate subspace $span\{S, Q_V\}$. Hence, we have $< \xi_i, \rho^* >= 0$ for $i = 1, 2, ..., V$. Meanwhile, since $\rho^* \in \mathcal{H}_1$,

$$\rho^*(x_i) =< \eta_i, \rho^* >=< \eta_i, P_1\rho^* >=< P_1\eta_i, \rho^* >=< \xi_i, \rho^* >= 0.$$

Hence the values of f at the data point x_i $(1 \leq i \leq V)$ remain unchanged after the projection. However, $||P_1 f^*|| \leq ||P_1 f||$. Intuitively, f^* is smoother than f. Some detail is lost during the projection, while the values of f on certain chosen design points are preserved. So it raises an interesting question as how to select a good subset of representers.

In the following proof, without any knowledge of the underlying true function, we will select V roughly equally spaced design points in $[0, 1]$.

Theorem 3.4 *Assume $f \in W_2[0, 1]$ and $||P_1 f||^2 = \int_0^1 [f''(x)]^2 dx$. For n roughly equally spaced design points, by selecting V basis functions corresponding to V roughly equally spaced design points, we only need $V = O(n^{\frac{4}{15}})$ to achieve the same convergence rate as the exact cubic spline estimate.*

Proof Let f be the exact solution. $f = f^* \oplus \rho^*$, where ρ^* is orthogonal to $span\{S, Q_V\}$. Hence, we have $\rho^*(x_i) = 0$ for $i = 1, 2, ..., V$. Without loss of generality, we assume $x_i \approx i/V$ for $1 \leq i \leq V$.

The following relation is true for any $x_i \leq a \leq b \leq x_{i+1}$, $i = 0, 2, ..., V-1$,

$$\int_{x_i}^{x_{i+1}} \rho^{*\prime\prime}(x)^2 dx \geq \int_a^b \rho^{*\prime\prime}(x)^2 dx \geq \left(\int_a^b \rho^{*\prime\prime}(x) dx \right)^2 = (\rho^{*\prime}(a) - \rho^{*\prime}(b))^2. \tag{3.4.7}$$

Since $\rho^*(x_i) = 0$ for $i = 1, 2, ..., V$ and ρ is smooth, there must be some point $b \in (x_i, x_{i+1})$ such that $\rho^{*\prime}(b) = 0$. Therefore, for any point $a \in (x_i, x_{i+1})$, we have $\rho^{*\prime}(a)^2 \leq \int_{x_i}^{x_{i+1}} \rho^{*\prime\prime}(x)^2 dx$. Combining with the fact $\rho^*(x_i) = 0$, we have

$$\rho^*(a)^2 \leq \left(\frac{1}{V}\right)^2 \left(\max_{x \in (x_i, x_{i+1})} \rho^{*\prime}(x) \right)^2 \leq \left(\frac{1}{V}\right)^2 \int_{x_i}^{x_{i+1}} \rho^{*\prime\prime}(x)^2 dx \tag{3.4.8}$$

for all $a \in (x_i, x_{i+1})$. Consequently,

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \rho^*(x_i)^2 dx &= \frac{1}{n} \sum_{i=1}^{V} \left(\sum_{x_i < x_j < x_{i+1}} \rho^*(x_j)^2 \right) \\ &\leq \frac{1}{n} \sum_{i=1}^{V} \frac{n}{V} \left(\frac{1}{V}\right)^2 \int_{x_i}^{x_{i+1}} \rho^{*\prime\prime}(x)^2 dx \\ &= \frac{1}{V^3} \int_0^1 \rho^{*\prime\prime}(x)^2 dx \\ &\leq \frac{1}{V^3} \int_0^1 f''(x)^2 dx \end{aligned} \tag{3.4.9}$$

Meanwhile, we know when n is large, $\int_0^1 f''(x)^2 dx$ is bounded in probability by some constant. Hence, to match the same converge rate of cubic spline $O(n^{-4/5})$, we only need $\frac{1}{V^3} = O(n^{-4/5})$. Hence, it is sufficient for $V = O(n^{4/15})$.

From Lemma 3.7, we know that ρ^* is an upper bound for ρ_*, which is the difference between the exact solution and the approximate solution. Hence the result is proved. ∎

The following corollary extends the above result to $f \in W_m$, $m \geq 2$ case.

Corollary 3.2 *For $f \in W_m[0,1]$, let the penalty $\mathbf{J}(f) = \int_0^1 f^{(m)}(x)^2 dx$. For roughly equally spaced design points on $[0,1]$, to match the same convergence rate rate as the exact solution, V only needs to grow at a rate of $O(n^{\frac{2m}{(2m+1)(2m-1)}})$.*

Proof Notice that there will be a point such that ρ^* has zero ith derivative $(1 \leq i \leq m)$ with i adjacent intervals separated by the grid points $x_1, x_2, ..., x_V$. Hence the maximum absolute value of ρ^* within the interval (x_i, x_{i+m-1}) is bounded.

$$\begin{aligned}
\max_{x_i<x<x_{i+m-1}} \rho^*(x)^2 &\leq (\frac{m-1}{V})^2 \max_{x_i<x<x_{i+m-1}} \rho^{*\prime}(x)^2 \leq \cdots \\
&\leq (\frac{m-1}{V})^{2(m-1)} \max_{x_i<x<x_{i+m-1}} \rho^{*(m-1)}(x)^2 \\
&\leq (\frac{m-1}{V})^{2(m-1)} \int_{x_i}^{x_{i+m-1}} f^{(m)}(x)^2 dx \qquad (3.4.10)
\end{aligned}$$

The proof of the above theorem extends immediately here. To achieve the same convergence rate, we must have

$$\frac{1}{V}\left(\frac{m-1}{V}\right)^{2(m-1)} = O(n^{\frac{2m}{2m+1}}).$$

Therefore, $V = O(n^{\frac{2m}{(2m+1)(2m-1)}})$. ∎

3.5 Adaptive Choice of the Smoothing Parameters

So far, all smoothing parameters are considered as fixed. When λ is small, the estimate tends to follow the data and hence appears to be wiggly. The estimated value has small bias but large variance. As $\lambda \to \infty$, f_j is forced into the null space $\mathcal{H}_0^j$ of the penalty function, which is usually a parametric space. Hence it has small variance but large bias. When λ varies, we have a family of flexible models. Tuning the smoothing parameters manually in low dimensional situations may be possible. Alternatively, pre-specified generalized degrees of freedom may be useful. However, to make this method more practical, an automated data-driven method to choose smoothing parameters is highly desirable.

3.5.1 Comparative Kullback-Leibler Distance

Certain risk function has to be chosen to measure the average closeness of an estimator to the truth. In Gaussian case, a popular choice is the expected squared loss function.

$$L(\mu, \hat{\mu}) = E_\mu(\hat{\mu} - \mu)^2. \tag{3.5.1}$$

Here, the observed data are distributed as $N(\mu, \sigma^2)$ with σ^2 known. It can be shown the above loss function is in fact a special case of the more general so called Kullback-Leibler distance.

Let $p(y)$ denote the true density function to be estimated. $\hat{p}(y)$ is our estimated density function. The Kullback-Leibler distance is defined by

$$KL(p, \hat{p}) = E_p \log \left(\frac{p(y)}{\hat{p}(y)} \right). \tag{3.5.2}$$

where E_p denotes the expectation under the truth p. Note the Kullback-Leibler distance is not a distance in fact since it is not symmetric. The comparative Kullback-Leibler distance CKL is defined by

$$\begin{aligned} CKL(p, \hat{p}) &= KL(p, \hat{p}) - E_p \log p(y) \\ &= -E_p \log \hat{p}(y), \end{aligned} \tag{3.5.3}$$

which differs from the Kullback-Leibler distance by a quantity which does not depend on the estimator. One way to look at the comparative Kullback-Leibler distance is to view it as the expected negative log-likelihood based on the estimated density function. To minimize the CKL distance is equivalent to maximize the expected log-likelihood for the future observations.

In many practical problems, except for the observed outcome variable y_i, we also observe a set of covariates $x_i \in \mathcal{X} \subseteq \mathcal{R}^D$, which can be used as predictors. Considering the random pair (Y_i, X_i), we are interested in estimating the conditional probability $p(y|x)$. Hence, conditioning on the value of X, the CKL distance of $p(y|X)$ and $\hat{p}(y|X)$ is

$$\begin{aligned} CKL(p, \hat{p}|X) &= -E_p \left(\log \hat{p}(y|X) | X \right) \\ &= -\int \log \hat{p}(y|X) p(y|X) dy \end{aligned} \tag{3.5.4}$$

Hence, the object function desired to be minimized should be the expectation of $CKL(p, \hat{p}|X)$ with respect to X

$$\begin{aligned} E(CKL(p,\hat{p}|X)) &= -E(E_p(\log \hat{p}(y|X)|X)) \\ &= -\int_x \left(\int_y \log \hat{p}(y|x) p(y|x) dy \right) p(x) dx \\ &= -\int_x \int_y \log \hat{p}(y|x) p(y,x) dy dx. \end{aligned} \tag{3.5.5}$$

Unfortunately, this quantity is unknown if we do not know the true $p(y, x)$. If we have n pairs of observed data (y_i, x_i), a consistent estimate of the above quantity is

$$CKL = \frac{1}{n} \sum_{i=1}^{n} E_p(\log \hat{p}(y|x_i)|x_i) = \frac{1}{n} \sum_{i=1}^{n} \int_y \log \hat{p}(y|x_i) p(y|x_i) dy. \tag{3.5.6}$$

This expression is useful when we are not interested in the distribution of X. However, it still depends on the unknown quantity $p(y|x_i)$. Therefore, it is desired to have a good estimate or proxy for it. In the Gaussian case, we can show that the UBR and AIC criterias are equivalent to the unbiased risk estimates for the above quantities. For complex modeling procedures, Ye (1998) defines the generalized degrees of freedom (GDF), and by an interesting theorem shows that it is the key to model fitting and selection when the goal is to minimize the CKL. The GDF generalizes the degrees of freedom for signal for the Gaussian penalized likelihood estimates, given in Wahba (1983). Interesting examples of Gaussian Case are given in Ye (1998), where randomization techniques are used in the estimation process. However, for Bernoulli data, it is known that no exact unbiased risk estimate exists (Wong 1992). Thus we can only have

approximately unbiased estimates. This, no doubt, explains why smoothing parameter selection with Bernoulli data has resisted a final, definitive answer so far.

Xiang & Wahba (1996) proposed the generalized approximate cross validation ($GACV$). Simulation studies show that it is an excellent computational proxy for CKL distance. We will give a heuristic argument here to support this observation. For Bernoulli outcomes, the CKL distance has the form $(1/n)\sum_{i=1}^{n}(-\mu_i\hat{f}_i + b(\hat{f}_i))$, where $\hat{f}_i$ is the estimated logit function for the ith observation. However, the true mean μ_i is unknown. One approach is to substitute it with the observed y_i and calculate $OBS = (1/n)\sum_{i=1}^{n}(-y_i\hat{f}_i + b(\hat{f}_i))$, which is the observed negative log-likelihood function for $\hat{f}$. But it is well known that OBS tends to underestimate CKL because that y_i and $\hat{f}_i$ are usually positively correlated for any meaningful modeling procedure. Hence $E(CKL - OBS) = (1/n)\sum E(y_i - \mu_i)\hat{f}_i = (1/n)\sum Cov(y_i, \hat{f}_i)$, which tends to be a positive number. See Efron (1986) for reference. To correct this bias, leave-out-one cross validation will also substitute $\hat{f}_i$ by $\hat{f}_i^{(-i)}$ in CKL, which only depends on the observations other than y_i. Thus $\hat{f}_i^{(-i)}$ is independent of y_i, and for large n, is expected to be close to $\hat{f}_i$. $Ey_i\hat{f}_i^{(-i)} = Ey_iE\hat{f}_i^{(-i)} \approx \mu_iE\hat{f}_i$. Therefore we expect CV to be a computable proxy for CKL distance.

3.5.2 *GACV* for Multivariate Bernoulli Responses

We will extend *GACV* to multivariate Bernoulli distribution to choose smoothing parameters adaptively. Before we proceed, we need to generalize the leave-out-one lemma in Craven & Wahba (1979) first. This time, we need to leave out one independent unit at a time.

Lemma 3.8 *(Leave-out-one-subject lemma) Let* $-l_j(y_{ij}, f_{ij}) = -\sum_k y_{ijk} f_{ijk} + b(f_{ij})$ *be the part of likelihood function related to the* j*th endpoint. All other parts of the likelihood function are considered as fixed.* $I_{\Lambda_j}(f_j, Y_j) = -\sum_i l_j(y_{ij}, f_{ij}) + \frac{n}{2}\mathbf{J}_{\Lambda_j}(f_j)$. *Suppose* $h(i, z, \cdot)$ *is the minimizer of* $I_{\Lambda_j}(f_j, Z)$, *where* $Z = (y_{1j}^T, ..., y_{i-1,j}^T, z^T, y_{i+1,j}^T, ..., y_{nj}^T)^T$, *then*

$$h(i, \mu^{(-i)}(x_{ij}), \cdot) = f_{\Lambda_j}^{(-i)}(\cdot),$$

where $f_{\Lambda_j}^{(-i)}$ *is the minimizer of* $\sum_{i_1 \neq i} l(y_{i_1 j}, f_{i_1,j}) + \frac{n}{2}\mathbf{J}_{\Lambda_j}(f_j)$, *and* $\mu^{(-i)}(x_{ij}) = (\mu^{(-i)}(x_{ij1}), ..., \mu^{(-i)}(x_{ijK_j}))^T$ *is the vector of means corresponding to* $f_{\Lambda_j}^{(-i)}(\cdot)$.

Proof We have

$$-l_j(\mu^{(-i)}(x_{ij}), f_{\Lambda_j}^{(-i)}(x_{ij})) \leq -l_j(\mu^{(-i)}(x_{ij}), f_j(x_{ij})). \tag{3.5.7}$$

This follows since setting

$$-\frac{\partial l_j(\mu^{(-i)}(x_{ij}), \tau)}{\partial \tau_k} = -\mu^{(-i)}(x_{ijk}) + \frac{\partial b(\tau)}{\partial \tau_k} = 0$$

and using the fact $\frac{\partial^2 b(\tau)}{\partial \tau^T \partial \tau} > 0$, implies that $-l_j(\mu^{(-i)}(x_{ij}), \tau)$ achieves its unique

minimum for $\frac{\partial b(\tau)}{\partial \tau_k} = \mu^{(-i)}(x_{ijk})$, hence $\tau_k = f_{\Lambda_j}^{(-i)}(x_{ijk})$. Therefore, for any f_j,

$$\begin{aligned} I_{\Lambda_j}(f_j, Z) &= -l_j(\mu^{(-i)}(x_{ij}), f_{ij}) - \sum_{i_1 \neq i} l_j(y_{i_1 j}, f_{i_1 j}) + \frac{n}{2}\mathbf{J}_{\Lambda_j}(f_j) \\ &\geq -l_j(\mu^{(-i)}(x_{ij}), f_{\Lambda_j}^{(-i)}(x_{ij})) - \sum_{i_1 \neq i} l_j(y_{i_1 j}, f_{i_1 j}) + \frac{n}{2}\mathbf{J}_{\Lambda_j}(f_j) \\ &\geq -l_j(\mu^{(-i)}(x_{ij}), f_{\Lambda_j}^{(-i)}(x_{ij})) - \sum_{i_1 \neq i} l_j(y_{i_1 j}, f_{\Lambda_j}^{(-i)}(x_{i_1 j})) + \frac{n}{2}\mathbf{J}_{\Lambda_j}(f_{\Lambda_j}^{(-i)}) \end{aligned}$$

The first inequality is due to (3.5.7), the second one is due to the fact that $f_{\Lambda_j}^{(-i)}(\cdot)$ is the minimizer of $-\sum_{i_1 \neq i} l(y_{i_1 j}, f_{i_1,j}) + \frac{n}{2}\mathbf{J}_{\Lambda_j}(f_j)$. Therefore we have $h(i, \mu^{(-i)}(x_{ij}), \cdot) = f_{\Lambda_j}^{(-i)}(\cdot)$. ∎

Let $Y_j^{(-i)} = (y_{1j}^T, ..., y_{i-1,j}^T, \mu^{(-i)}(x_{ij})^T, y_{i+1,j}^T, ..., y_{nj}^T)^T$. Because that (f_{Λ_j}, Y_j) and $(f_{\Lambda_j}^{(-i)}, Y_j^{(-i)})$ are two local minimizers of $I_{\Lambda_j}(f, Z)$, $\partial I_{\Lambda_j}/\partial f_j$ is equal to zero on those two points. Thus,

$$\frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}, Y_j) = 0, \frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}^{(-i)}, Y_j^{(-i)}) = 0. \tag{3.5.8}$$

It is also easy to verify that

$$\frac{\partial^2 I_{\Lambda_j}}{\partial f_j \partial f_j^T} = W_j(f) + n\Sigma_{\Lambda_j}, \frac{\partial^2 I_{\Lambda_j}}{\partial Y_j \partial f_j^T} = -I, \tag{3.5.9}$$

where $W_j(f) = diag(W_{1j}, W_{2j}, ..., W_{nj})$ is defined in (3.3.6). Σ_{Λ_j} is the semi-positive definite matrix satisfying $\mathbf{J}_{\Lambda_j}(f_j) = f_j^T \Sigma_{\Lambda_j} f_j$.

Using a first order Taylor expansion, we have the following equation

$$\begin{aligned} 0 &= \frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}^{(-i)}, Y_j^{(-i)}) \\ &= \frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}, Y_j) + \frac{\partial^2 I_{\Lambda_j}}{\partial f_j \partial f_j^T}(f^*, Y^*)(f_{\Lambda_j}^{(-i)} - f_{\Lambda_j}) \\ &\quad + \frac{\partial^2 I_{\Lambda_j}}{\partial Y_j \partial f_j^T}(f^*, Y^*)(Y_j^{(-i)} - Y_j) \\ &= \frac{\partial^2 I_{\Lambda_j}}{\partial f_j \partial f_j^T}(f^*, Y^*)(f_{\Lambda_j}^{(-i)} - f_{\Lambda_j}) + \frac{\partial^2 I_{\Lambda_j}}{\partial Y_j \partial f_j^T}(f^*, Y^*)(Y_j^{(-i)} - Y_j) \end{aligned} \quad (3.5.10)$$

Equivalently, this is

$$(f_{\Lambda_j} - f_{\Lambda_j}^{(-i)}) = (W_j(f^*) + n\Sigma_{\Lambda_j})^{-1}(Y_j - Y_j^{(-i)}), \quad (3.5.11)$$

where (f^*, Y^*) is a point somewhere between (f_{Λ_j}, Y_j) and $(f_{\Lambda_j}^{(-i)}, Y_j^{(-i)})$. Approximate $W(f^*)$ by $W(f_{\Lambda_j})$ and note that $Y - Y^{(-i)} = (0, ..., 0, (y_{ij} - \mu^{(-i)}(x_{ij}))^T, 0, ..., 0)^T$. We have

$$\begin{pmatrix} f_{\Lambda_j}(x_{1j1}) - f_{\Lambda_j}^{(-i)}(x_{1j1}) \\ \vdots \\ f_{\Lambda_j}(x_{ij1}) - f_{\Lambda_j}^{(-i)}(x_{ij1}) \\ \vdots \\ f_{\Lambda_j}(x_{ijK_j}) - f_{\Lambda_j}^{(-i)}(x_{ijK_j}) \\ \vdots \\ f_{\Lambda_j}(x_{njK_j}) - f_{\Lambda_j}^{(-i)}(x_{njK_j}) \end{pmatrix} \approx (W_j(f_{\Lambda_j}) + n\Sigma_{\Lambda_j})^{-1} \begin{pmatrix} 0 \\ \vdots \\ y_{ij1} - \mu^{(-i)}(x_{ij1}) \\ \vdots \\ y_{ijK_j} - \mu^{(-i)}(x_{ijK_j}) \\ \vdots \\ 0 \end{pmatrix} \quad (3.5.12)$$

Denote $H^j = [W_j(f_{\Lambda_j}) + n\Sigma_{\Lambda_j}]^{-1}$, which is the inverse Hessian of $I_{\Lambda_j}(f_j, Y_j)$ with respect to f_j evaluated at f_{Λ_j}. H^j has the following structure

$$H^j = \begin{pmatrix} H^j_{11} & & * \\ & H^j_{22} & \\ * & & \ddots \\ & & & H^j_{nn} \end{pmatrix}, \tag{3.5.13}$$

where H^j_{ii} is the $K_j \times K_j$ submatrix on the diagonal. Hence, we have

$$\begin{pmatrix} f_{\Lambda_j}(x_{ij1}) - f^{(-i)}_{\Lambda_j}(x_{ij1}) \\ \vdots \\ f_{\Lambda_j}(x_{ijK_j}) - f^{(-i)}_{\Lambda_j}(x_{ijK_j}) \end{pmatrix} \approx H^j_{ii} \begin{pmatrix} y_{ij1} - \mu^{(-i)}(x_{ij1}) \\ \vdots \\ y_{ijK_j} - \mu^{(-i)}(x_{ijK_j}) \end{pmatrix}. \tag{3.5.14}$$

Starting with the ordinary leave-out-one cross validation function $CV(\Lambda_j)$, we will use the above relation and several first order Taylor expansions in our derivation.

$$\begin{aligned} CV(\Lambda_j) &= \frac{1}{n}\sum_{i=1}^{n}[-\sum_{k=1}^{K_j} y_{ijk} f^{(-i)}_{ijk} + b(f_{ij})] \\ &= \frac{1}{n}\sum_{i=1}^{n}[-\sum_{k=1}^{K_j} y_{ijk} f_{ijk} + b(f_{ij}) + \sum_{k=1}^{K_j} y_{ijk}(f_{ijk} - f^{(-i)}_{ijk})] \\ &= OBS(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{K_j} y_{ijk}(f_{ijk} - f^{(-i)}_{ijk}) \\ &= OBS(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix} \begin{pmatrix} f_{ij1} - f^{(-i)}_{ij1} \\ \vdots \\ f_{ijK_j} - f^{(-i)}_{ijK_j} \end{pmatrix} \end{aligned} \tag{3.5.15}$$

Next, we need to show the following relation is true. The first approximation is due to Taylor expansion for a function with vector responses.

$$
\begin{aligned}
& \begin{pmatrix} y_{ij1} - \mu_{ij1} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j} \end{pmatrix} \\
= & \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix} + \begin{pmatrix} \mu_{ij1}^{(-i)} - \mu_{ij1} \\ \vdots \\ \mu_{ijK_j}^{(-i)} - \mu_{ijK_j} \end{pmatrix} \\
= & \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix} + \begin{pmatrix} \frac{\partial b}{\partial f_{ij1}}(f_{\Lambda_j}^{(-i)}(x_{ij})) - \frac{\partial b}{\partial f_{ij1}}(f_{\Lambda_j}(x_{ij})) \\ \vdots \\ \frac{\partial b}{\partial f_{ijK_j}}(f_{\Lambda_j}^{(-i)}(x_{ij})) - \frac{\partial b}{\partial f_{ijK_j}}(f_{\Lambda_j}(x_{ij})) \end{pmatrix} \\
\approx & \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix} + W_{ij} \begin{pmatrix} f_{\Lambda_j}^{(-i)}(x_{ij}) - f_{\Lambda_j}(x_{ij}) \\ \vdots \\ f_{\Lambda_j}^{(-i)}(x_{ij}) - f_{\Lambda_j}(x_{ij}) \end{pmatrix} \\
\approx & \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix} - W_{ij} H_{ii}^{j} \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix} \\
= & (I - W_{ij} H_{ii}^{j}) \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix}
\end{aligned}
\tag{3.5.16}
$$

Hence, we have the following approximate relation. We will use it to define the approximate cross validation (ACV) function.

$$
\begin{aligned}
& CV(\Lambda_j) \\
= \;& OBS(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix} \begin{pmatrix} f_{ij1} - f_{ij1}^{(-i)} \\ \vdots \\ f_{ijK_j} - f_{ijK_j}^{(-i)} \end{pmatrix} \\
\approx \;& OBS(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix} H_{ii}^{j} \begin{pmatrix} y_{ij1} - \mu_{ij1}^{(-i)} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}^{(-i)} \end{pmatrix} \\
\approx \;& OBS(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix} H_{ii}^{j}(I - W_{ij}H_{ii}^{j})^{-1} \begin{pmatrix} y_{ij1} - \mu_{ij1} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j} \end{pmatrix} \\
\equiv \;& ACV(\Lambda_j). \qquad (3.5.17)
\end{aligned}
$$

Now define $G_{ii}^{j} = (I - W_{ij}H_i)$. In a step reminiscent of that used to get from leave-out-one cross validation to GCV in the Gaussian case, we will obtain a generalized form of the approximate cross validation. There, the diagonal elements of certain matrix was replaced by $1/n$ times its trace. Here, for any matrices $A_{ii}, 1 \le i \le n$,

$$A_i = \left(a_{i,k_1k_2}\right)_{K\times K}, 1 \le k_1, k_2 \le K,$$

we define

$$\bar{A} = (\delta - \gamma) I_{K\times K} + \gamma \cdot e e^T = \begin{pmatrix} \delta & \gamma & \cdots & \gamma \\ \gamma & \delta & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \delta \end{pmatrix}, \tag{3.5.18}$$

where $e = (11\cdots 1)^T$ is the unit vector, and δ and γ are the average values of corresponding elements in the matrices A_{ii}'s.

$$\begin{aligned} \delta &= \frac{1}{nK} \sum_{i=1}^{n} \sum_{k=1}^{K} a_{i,kk}, \\ \gamma &= \frac{1}{n \cdot K(K-1)} \sum_{i=1}^{n} \sum_{k_1 \neq k_2} a_{i,k_1 k_2}. \end{aligned} \tag{3.5.19}$$

Since $\bar{A}$ has a very special structure, it is very easy to obtain the closed form of its inverse

$$\begin{aligned} \bar{A}^{-1} &= \frac{1}{\delta - \gamma} I_{K\times K} - \frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} e e^T \\ & \begin{pmatrix} \frac{\delta+(K-2)\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \cdots & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} \\ -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \frac{\delta+(K-2)\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \cdots & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & -\frac{\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} & \cdots & \frac{\delta+(K-2)\gamma}{(\delta-\gamma)(\delta+(K-1)\gamma)} \end{pmatrix} \end{aligned} \tag{3.5.20}$$

Hence, we define the generalized form of approximate cross validation ($GACV$) for multivariate Bernoulli distribution as following

$$\begin{aligned}
& GACV(\Lambda_j) \\
= \; & OBS(\Lambda_j) + \frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix}\bar{H}^j(\bar{G}^j)^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j} \end{pmatrix} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}[-\sum_{k=1}^{K_j} y_{ijk}f_{ijk} + b(f_{ij})] \\
& +\frac{1}{n}\sum_{i=1}^{n}\begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix}\bar{H}^j(\bar{G}^j)^{-1}\begin{pmatrix} y_{ij1}-\mu_{ij1} \\ \vdots \\ y_{ijK_j}-\mu_{ijK_j}. \end{pmatrix} \qquad (3.5.21)
\end{aligned}$$

We remark that the above formula is reduced to (2.9) in Xiang & Wahba (1996) when $j = 1$ and $K_1 = 1$. In practice, we will iteratively choose smoothing parameters in each Block nonlinear SOR iteration in order to minimize $GACV$.

When only person-specific covariates exist, following the notation defined at the end of section (3.3), we can rewrite the above formula to a simpler form

$$\begin{aligned}
& GACV(\Lambda_j) \\
= \; & OBS(\Lambda_j) + \frac{tr(H^j)/n\cdot\sum_{i=1}^{n} y_{ij}(y_{ij}-\mu_{ij})}{n - tr(W_j^{1/2}H^jW_j^{1/2})} \\
= \; & \frac{1}{n}\sum_{i=1}^{n}[-y_{ij}f_{ij} + b(f_{ij})] + \frac{tr(H^j)/n\cdot\sum_{i=1}^{n} y_{ij}(y_{ij}-\mu_{ij})}{n - tr(W_j^{1/2}H^jW_j^{1/2})} \qquad (3.5.22)
\end{aligned}$$

3.5.3 The One-Step Randomized Estimate

The $GACV$ defined in the last section is very computing intensive. It involves the computation of the inverse Hessian, which is a large matrix in our case. However, this explicit calculation can be avoided by using a technique in the spirit of the randomized trace method, provided a solution, either exact or approximate, of the variational problem can be obtained at a lower cost. In this section, we will propose a one-step randomized estimate of $GACV$, which is fast and cheap to calculate.

The randomized trace technique was proposed in Girard (1987), Girard (1991), Girard (1998). Given any square matrix A, and ϵ is a zero mean random vector with independent components with variance σ^2, then $tr(A) = \frac{1}{\sigma^2} E\epsilon^T A\epsilon$. Hence we can estimate the trace of A by $\frac{1}{\sigma^2} \cdot \epsilon^T A\epsilon$. In practice, σ^2 is replaced by $\frac{1}{n}\sum_{i=1}^{n} \epsilon_i^2$.

Given a square matrix A with $A_{ii}(1 \leq i \leq n)$ being the $K \times K$ submatrices on the diagonal, we discuss how to obtain a randomized estimate of $\bar{A}$. First, a vector of $i.i.d.$ random variables distributed as $N(0,1)$ is generated. $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{iK_j})^T$ and $\epsilon = (\epsilon_1^T, ..., \epsilon_n^T)^T$. Hence, $\delta = tr(A)/(nK)$ can be estimated by $(\epsilon^T A\epsilon)/(nK)$. On the other hand, $\gamma = (\sum_i \sum_{k_1,k_2} a_{i,k_1k_2} - tr(A))/(nK(K-1))$. To estimate $\sum_i \sum_{k_1,k_2} a_{i,k_1k_2}$, let $\bar{\epsilon}_i = (1/\sqrt{K}) \sum_{k=1}^{K} \epsilon_{ik}$, $\bar{\epsilon} = (\bar{\epsilon}_1, ..., \bar{\epsilon}_1, \bar{\epsilon}_2, ..., \bar{\epsilon}_n)^T$. $\bar{\epsilon}$ is a column vector with K replicates of $\bar{\epsilon}_i$ for each $1 \leq i \leq n$. We notice that $E\bar{\epsilon}^T A\bar{\epsilon} = \sum_i \sum_{k_1,k_2} a_{i,k_1k_2}$. Hence, we can estimate γ by $(\bar{\epsilon}^T A\bar{\epsilon} - \epsilon^T A\epsilon)/(nK(K-1))$. Therefore, a randomized estimate of $\bar{A}$ can

be obtained.

In practice, the randomized estimate of $GACV$ is calculated by solving the nonlinear system on the perturbed data $Y_j + \epsilon$ and $Y_j + \bar{\epsilon}$. Denote $f_{\Lambda_j}^{Y_j}$ as the solution of (3.3.7) by using the original data and $f_{\Lambda_j}^{Y_j+\epsilon}$ as the solution by using the perturbed data. If we take $f_{\Lambda_j}^{Y_j}$ as the initial value to a Newton-Ralphson calculation of $f_{\Lambda_j}^{Y_j+\epsilon}$, and we run the iteration only once by using all matrix decompositions which have already been performed for calculating $f_{\Lambda_j}^{Y_j}$ in the last step, we obtain the one step solution $f_{\Lambda_j}^{Y_j+\epsilon,1}$. Since $\frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}^{Y_j}, Y_j) = 0$ and $\frac{\partial^2 I_{\Lambda_j}}{\partial f_j^T \partial f_j}(f_{\Lambda_j}^{Y_j}, Y_j) = \frac{\partial^2 I_{\Lambda_j}}{\partial f_j^T \partial f_j}(f_{\Lambda_j}^{Y_j}, Y_j + \epsilon)$, we observe the simple relation

$$\begin{aligned} f_{\Lambda_j}^{Y_j+\epsilon,1} &= f_{\Lambda_j}^{Y_j} - [\frac{\partial^2 I_{\Lambda_j}}{\partial f_j^T \partial f_j}(f_{\Lambda_j}^{Y_j}, Y_j + \epsilon)]^{-1} \frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}^{Y}, Y_j + \epsilon) \\ &= f_{\Lambda_j}^{Y_j} - [\frac{\partial^2 I_{\Lambda_j}}{\partial f_j^T \partial f_j}(f_{\Lambda_j}^{Y_j}, Y)]^{-1}(-\epsilon + \frac{\partial I_{\Lambda_j}}{\partial f_j}(f_{\Lambda_j}^{Y_j}, Y_j)) \\ &= f_{\Lambda_j}^{Y_j} + (W_j + n\Sigma_{\Lambda_j})^{-1}\epsilon. \end{aligned} \tag{3.5.23}$$

Hence, we have

$$f_{\Lambda_j}^{Y_j+\epsilon,1} - f_{\Lambda_j}^{Y_j} = H^j \epsilon. \tag{3.5.24}$$

Thus, $\epsilon^T(f_{\Lambda_j}^{Y_j+\epsilon,1} - f_{\Lambda_j}^{Y_j}) = \epsilon^T H^j \epsilon$ and $\bar{\epsilon}^T(f_{\Lambda_j}^{Y_j+\bar{\epsilon},1} - f_{\Lambda_j}^{Y_j}) = \bar{\epsilon}^T H^j \bar{\epsilon}$, we can obtain a randomized estimate of $\bar{H}^j$. Similarly $\epsilon^T G^j \epsilon = \epsilon^T \epsilon + \epsilon^T W_j(f_{\Lambda_j}^{Y_j+\epsilon,1} - f_{\Lambda_j}^{Y_j})$, and $\bar{\epsilon}^T G^j \bar{\epsilon} = \bar{\epsilon}^T \bar{\epsilon} + \bar{\epsilon}^T W_j(f_{\Lambda_j}^{Y_j+\bar{\epsilon},1} - f_{\Lambda_j}^{Y_j})$. We can calculate the randomized estimate of $\bar{G}^j$. This approach avoids the explicit calculation of inverse Hessian H^j, which is computational expensive and tends to be unstable for ill conditioned matrix. A randomized estimate can always be obtained provided a cheap and stable "black

box" exists to calculate the (approximate) one-step solution for perturbed data. The resulting $ranGACV$ function is

$$\begin{aligned} & ranGACV(\Lambda_j) \\ = \; & \frac{1}{n}\sum_{i=1}^{n}[-\sum_{j=1}^{K_j} y_{ijk} f_{ijk} + b(f_{ij})] \\ & + \frac{1}{n}\sum_{i=1}^{n} \begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix} \hat{\bar{H}}^j (\hat{\bar{G}}^j)^{-1} \begin{pmatrix} y_{ij1} - \mu_{ij1} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}. \end{pmatrix}, \quad (3.5.25) \end{aligned}$$

where $\hat{\bar{H}}^j$ and $\hat{\bar{G}}^j$ denote the randomized estimates. To reduce the variance in the term after "+" in (3.5.25), we may draw R independent random vectors $\epsilon^{(1)},...,\epsilon^{(R)}$, replace the term after "+" in (3.5.25) by

$$\frac{1}{nR}\sum_{r=1}^{R}\sum_{i=1}^{n} \begin{pmatrix} y_{ij1} & \cdots & y_{ijK_j} \end{pmatrix} \hat{\bar{H}}^{j^{(r)}} (\hat{\bar{G}}^{j^{(r)}})^{-1} \begin{pmatrix} y_{ij1} - \mu_{ij1} \\ \vdots \\ y_{ijK_j} - \mu_{ijK_j}. \end{pmatrix} \quad (3.5.26)$$

to obtain an R-replicated $ranGACV$ function. Combined with the approximate spline described in the last section, the computation of $ranGACV$ is fast and stable. We will iteratively minimize $ranGACV$ in each step of block one-step SOR iteration. This will be done repeatedly until some pre-specified convergence criteria is met, or the number of iterations exceeds the pre-specified limit.

The $GACV$ and $ranGACV$ function is derived by assuming that the minimizer of (3.3.7) is calculated at each block nonlinear SOR iteration. To speed up the algorithm, however, only one-step update will be calculated. We remark

that all favorable properties of $GACV$ and $ranGACV$ are preserved for Block one-step SOR algorithm and approximate spline estimate. It is very easy to carry out the computation as no additional matrix decomposition is required. By evaluating H^j and W_j at the latest updated value f_{j-}, most of the approximations in the derivation of $GACV$ becomes exact. If we take f_{j-} as the initial value, all matrix decompositions which have been done for calculating $f_{\Lambda_j}^{Y_j,1}$ is readily available for computing the one-step estimate $f_{\Lambda_j}^{Y_j+\epsilon,1}$ for the perturbed data. Moreover, the relation in (3.5.24) remains to be true for the block one-step SOR algorithm, which sets $f_{\Lambda_j}^{Y_j} = f_{\Lambda_j}^{Y_j,1}$ in every iteration.

Since it is difficult to write down the derivatives of $ranGACV$ with respect to the smoothing parameter(s) Λ, to search for the minimizer of $ranGACV$ function, optimization methods which do not require the explicit calculation of the derivatives are highly desired. For single smoothing parameters, we will use Golden section method. For multiple smoothing parameters, we will use downhill simplex method. See Press, Flannery, Teukolsky & Vetterling (1996) for reference.

3.5.4 Numerical Examples

(i) $ranGACV$ **vs. iterated** $ranGACV$

The first experiment is to compare the performances of $ranGACV$ and iterated $ranGACV$. For fixed smoothing parameters, Xiang & Wahba (1996) and Lin, Wahba, Xiang, Gao, Klein & Klein (1998) proposed to find the solution of the

variational problem, then evaluate the $GACV$ function. However, for multivariate Bernoulli data, when there present more than one logit functions to be estimated, or we assume the parametric form for the association terms, evaluating and minimizing $ranGACV$ for each logit function at the corresponding step of the block one-step SOR-Newton-Ralphson algorithm seems to be more convenient and natural. In this experiment, we will assume $j = 1$ and $K_j = 1$, the situation is reduced to the univariate Bernoulli distribution.

The first three univariate functions are taken from Xiang & Wahba (1996). We define the true logit functions to be estimated as

$$\begin{aligned} f_1(x) &= 3 - (5x - 2.5)^2 \\ f_2(x) &= 2\sin(10x) \\ f_3(x) &= 0.218 - 4.312x. \end{aligned} \tag{3.5.27}$$

Figure 1 shows the true probability functions determined by $p(x) = e^{f(x)}/(1 + e^{f(x)})$. The predictor variable x was taken to be uniformly distributed in $(0, 1)$. Two sample sizes $n = 100$ and $n = 400$ were used for this simulation. To compare the effectiveness of these two methods, 100 independent sets of data for each combination of logit function and sample size were generated. We used the same random perturbations and set $R = 5$ and computed the 5-replicated $ranGACV$ for both methods. Only 50 basis functions chosen by clustering method were used for approximate spline for all cases. The pairwise comparison of CKL distance is plotted in Figure 2. From this experiment, the performances of $ranGACV$ and iterated $ranGACV$ are almost the same. $ranGACV$ seems

to be slightly better than its iterated version for small sample sizes. However, this difference becomes negligible very quickly when the sample size increases. The iterated $ranGACV$ method is not guaranteed to converge, although this happens very rare. From extensive simulation studies, when the algorithm does not converge, very often, the value at the last step of the iteration is still an acceptable estimation.

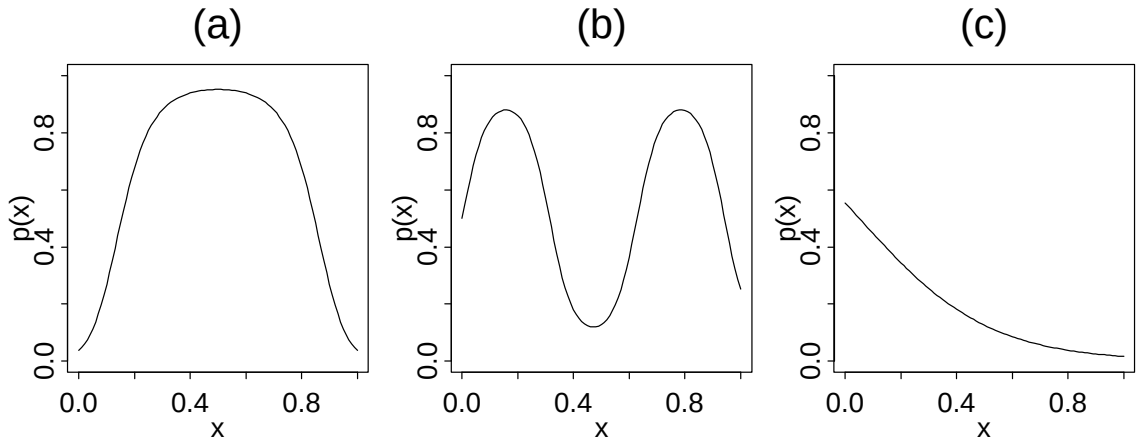

Figure 1: True probability function $p(x)$ determined by the logit functions in (3.5.27): (a) f_1 (b) f_2 (c) f_3

The next Monte Carlo simulation uses the WESDR (Wisconsin Epidemiology Study of Diabetes Retinopathy) data. See Wahba et al. (1995) and references cited there. Three covariates dur, gly and bmi are used as predictor variables. The outcome variable is the progression of retinopathy. The following ANOVA model is fitted by iterated UBR method by GRKPACK (Wang 1997),

$$logit(p(dur, gly, bmi) = c + f_1(dur) + f_2(gly) + f_3(bmi) + f_{12}(dur, bmi).$$

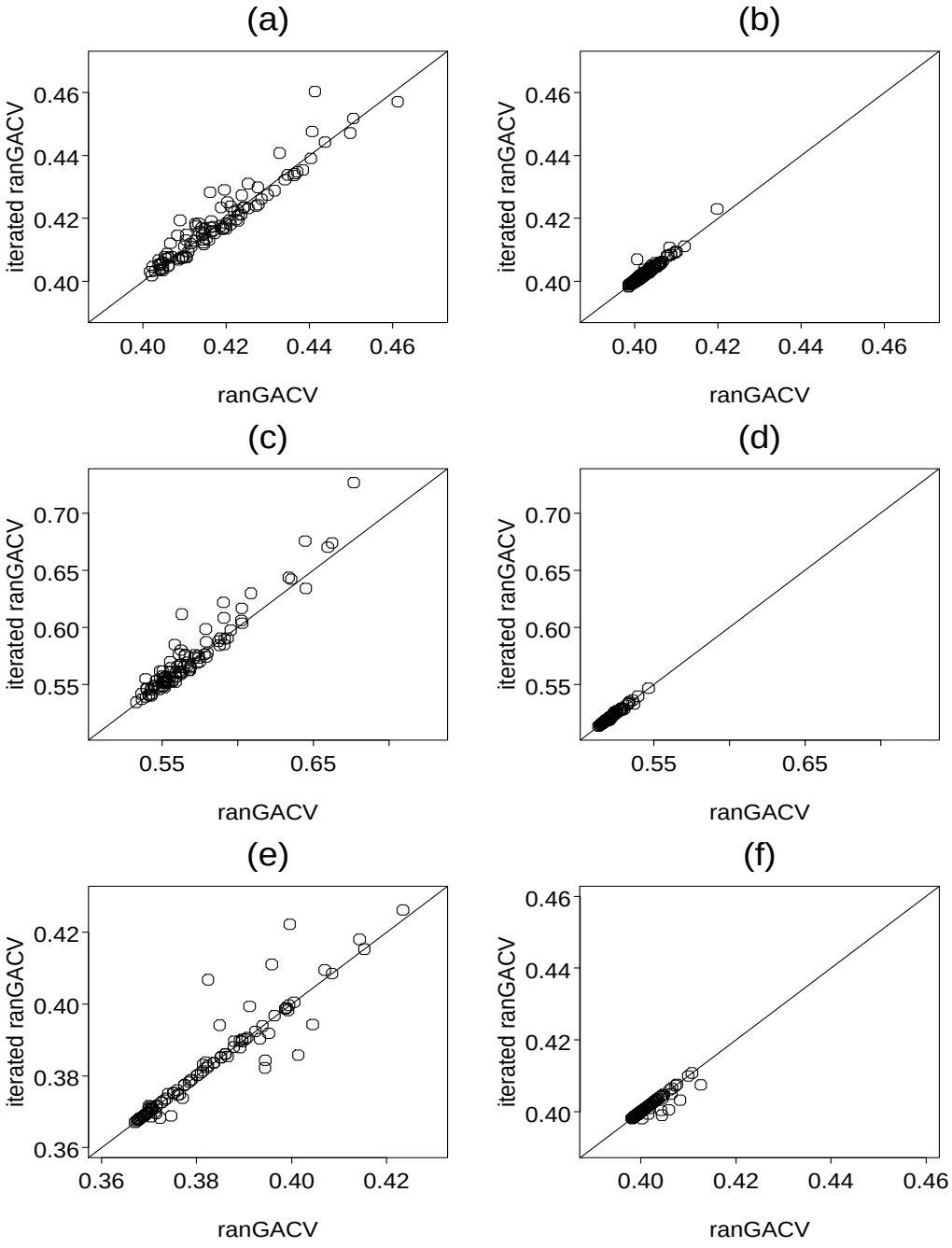

Figure 2: Pairwise comparison of CKL for $ranGACV$ and iterated $ranGACV$ for the cases in (3.5.27). (a) $f_1, n = 100$ (b) $f_1, n = 400$ (c) $f_2, n = 100$ (d) $f_2, n = 400$ (e) $f_3, n = 100$ (f) $f_3, n = 400$

The fitted logit function is then treated as the true test function in our simulation. 100 replicates of data are generated and fitted for the above ANOVA model by both $ranGACV$ and iterated $ranGACV$ methods. The number of replicates R for randomized estimate of $GACV$ is taken to be 5 for both methods. In the mean time, we used clustering method to obtain 50 basis functions for the approximate spline. For each run, the CKL distance between the true probability function used to generate the data and the estimated probability is computed. The pairwise comparison of the CKL distance is plotted in Figure 3. The $ranGACV$ method seems to be slightly better than the iterated $ranGACV$ algorithm. However, the difference is very small.

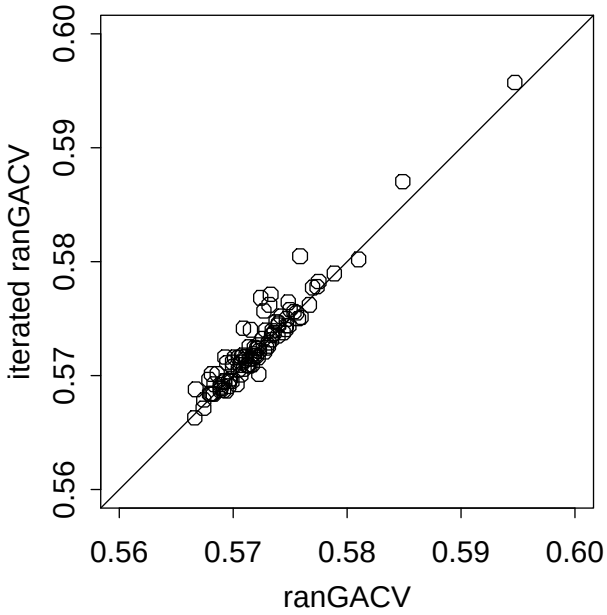

Figure 3: Pairwise comparison of CKL for $ranGACV$ and iterated $ranGACV$ for 100 runs of the simulated WESDR data.

In light of the results of the above simulation studies, we prefer to solve the variational problem for fixed smoothing parameters, then evaluate the $ranGACV$ function at the solution whenever possible. However, for more complicated situations, there may exist more than one logit functions to be estimated, or some functions to be estimated may take simple unpenalized parametric form. It is very difficult to write down the closed form of $ranGACV$ and to compute it directly. On the other hand, combined with some iterative algorithm to solve the variational problem, iterated $ranGACV$ is the natural alternative which is expected to be nearly as efficient as $ranGACV$ itself.

(ii) Iterated $ranGACV$ as a proxy for CKL distance

In this experiment, we will show that the iterated $ranGACV$ is an excellent computational proxy for CKL distance for multivariate Bernoulli data. Iterated $ranGACV$ is an estimator of CKL distance at every updating step of the Block one-step SOR-Newton-Ralphson algorithm.

We assume that $j = 1$ and $K_j = 2$. There are one endpoint of interest and two repeated measurements for it. The first example is for the single smoothing parameter case. The predictor variable x is assumed to be uniformly distributed on $(0, 1)$. For each subject, x is assumed to be the same for both measurements. The true conditional logit function to be estimated is

$$f(x) = logit(P(Y_k = 1|Y^{(-k)} = 0, x)) = 3\sin(2.7x^2) - 2. \qquad (3.5.28)$$

Odds ratio is used to measure the association between correlated observations.

We will let the conditional log odds ratio be a constant

$$\alpha = \log OR(Y_1, Y_2|x) = 1. \tag{3.5.29}$$

The sample size n is taken to be 500. The predictor variable x is assumed to be uniformly distributed on $(0, 1)$. The true marginal probability $p(x) = P(Y_k = 1|x) = (e^{f(x)} + e^{2f(x)+\alpha})/(1 + 2e^{f(x)} + e^{2f(x)+\alpha})$ and one set of randomly generated data according to the true joint distribution are plotted in Figure 4. This set of data is used in our simulation study. To compute the approximate spline estimate, only 50 basis functions are selected.

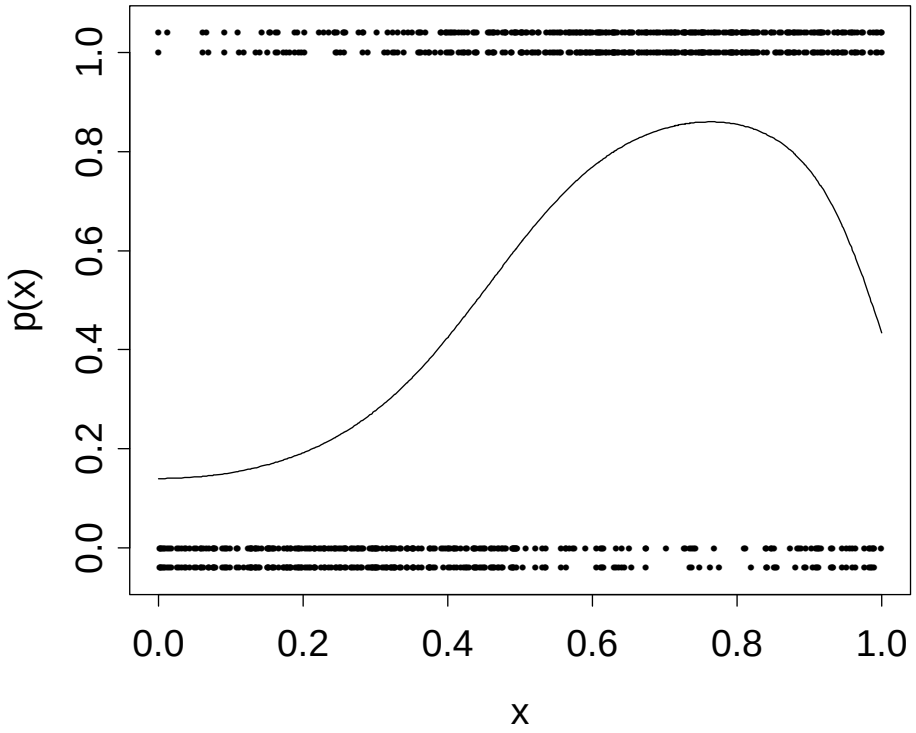

Figure 4: True marginal probability $P(Y_k = 1|x)$ and one set of generated data.

As proposed early, the algorithm we used to estimate the joint distribution

will iterative update f and α. We proposed to iteratively minimize $ranGACV$ whenever updating f by the one-step updating formula. The initial values for both f and α are taken to be 0. At different stage of this process, the true CKL distance and the $ranGACV$ function are computed and plotted in Figure 5-7. Figure 5 shows the comparison made at the first iteration step while $\hat{\alpha} = 0$. Figure 7 shows the comparison made at the converged value while $\hat{\alpha} = 1.53$. Figure 6 shows the comparison made in the middle of this iterative algorithm, while $\hat{\alpha} = 0.91$. Three different values are taken for R, the number of replicates used to evaluate the randomized estimate of $GACV$ in order to reduce variance. And for each value of R, 10 independent realizations of $ranGACV$ function are computed and plotted. The closed circle is the minimizer of the CKL distance while the open circles indicate the minimizers for each $ranGACV$ curve.

In terms of locating the best λ which yields the smallest CKL distance, $ranGACV$ is an excellent proxy to be minimized. When R increases, $ranGACV$ seems to have smaller variance and better performance. Since the iterated algorithm minimizes $ranGACV$ at every step, we really prefer it to have smaller variance. In the meanwhile, The computation of $ranGACV$ is in fact very fast since no additional matrix decomposition is necessary. Hence we suggest to let R be large enough, for example, $R = 20$.

The next example is for multiple smoothing parameters. Still, there is one endpoint of interest and paired observations for each subject. The predictor variables (x_1, x_2) are assumed to be uniformly distributed on the unit square

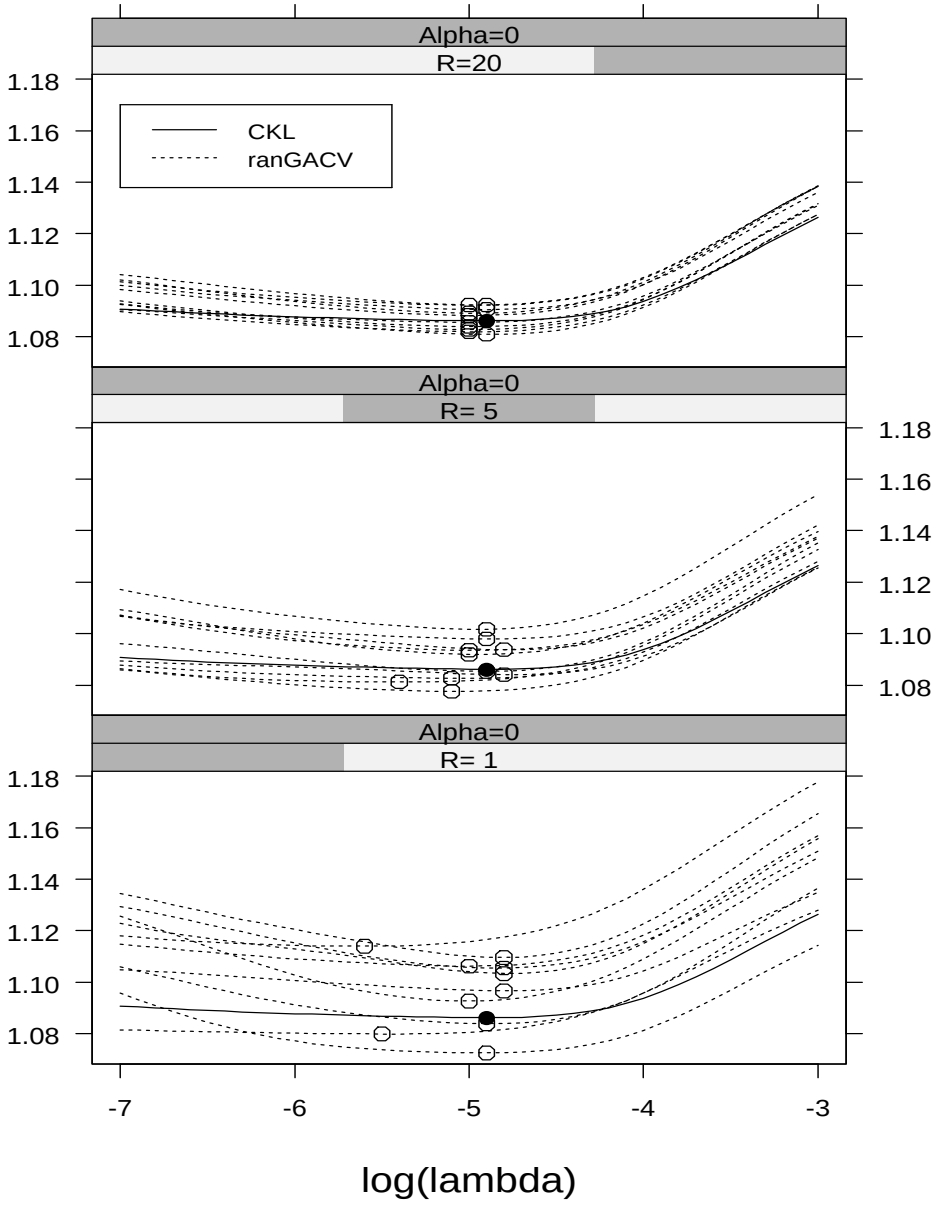

Figure 5: 10 replicates of $ranGACV$ curves compared to CKL when $\hat{\alpha} = 0$. R is the number of replicates used to evaluate the randomized estimate of $GACV$. Circles indicate the minimizers for each curve.

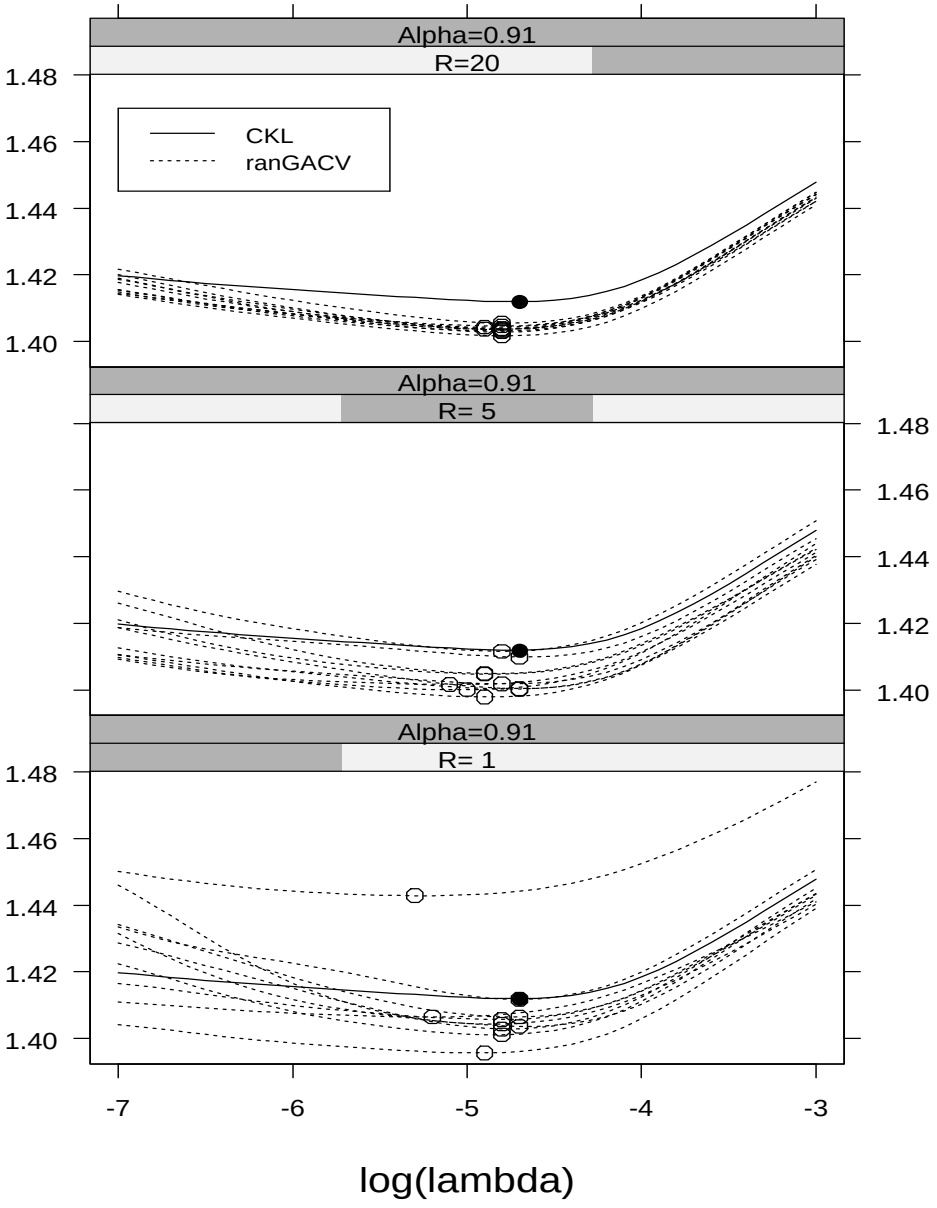

Figure 6: 10 replicates of $ranGACV$ curves compared to CKL when $\hat{\alpha} = 0.91$. R is the number of replicates used to evaluate the randomized estimate of $GACV$. Circles indicate the minimizers for each curve.

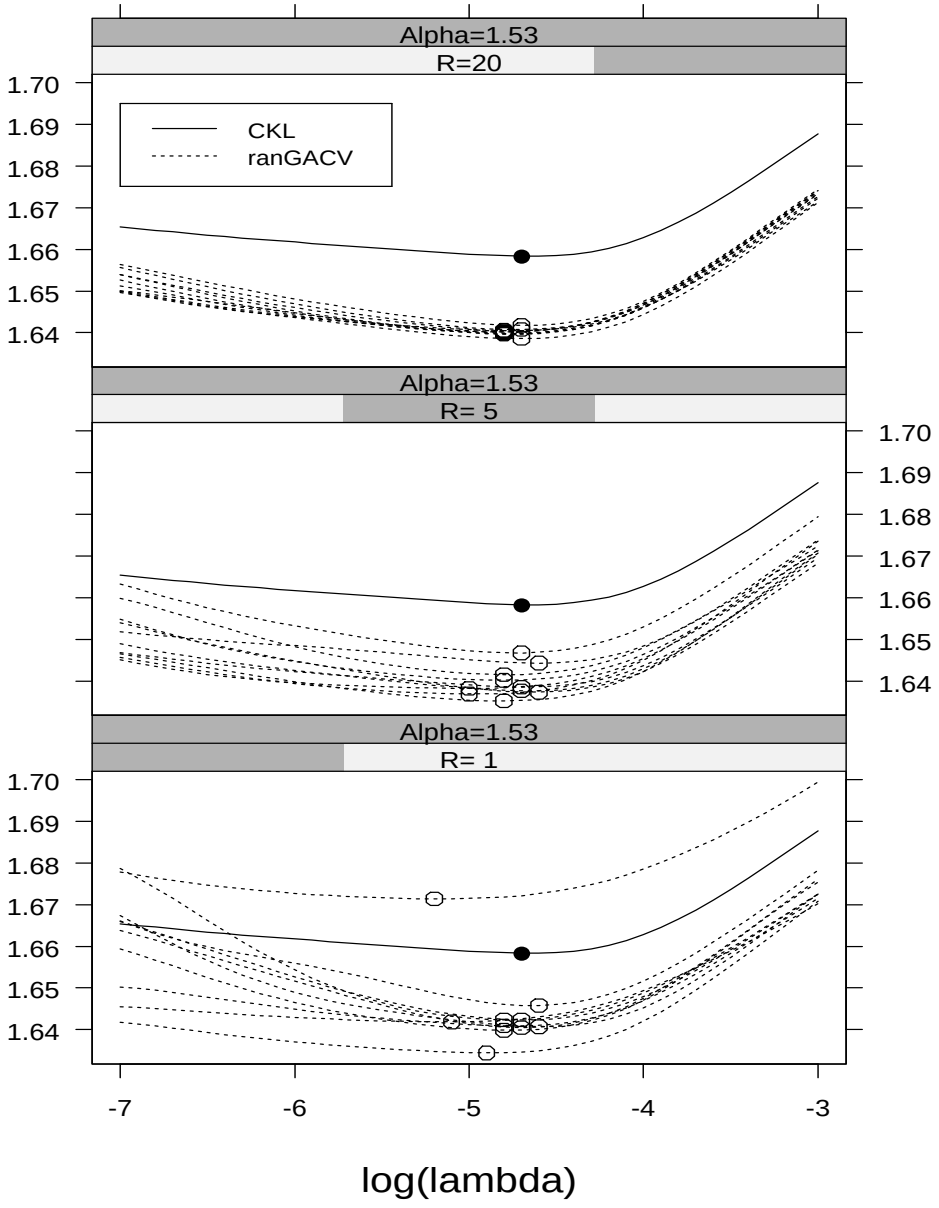

Figure 7: 10 replicates of $ranGACV$ curves compared to CKL when $\hat{\alpha} = 1.53$. R is the number of replicates used to evaluate the randomized estimate of $GACV$. Circles indicate the minimizers for each curve.

$(0,1) \times (0,1)$. We assume the true conditional logit function has an additive form

$$f(x_1, x_2) = logit(P(Y_k = 1|Y^{(-k)} = 0, x_1, x_2)) = 2\sin(2\pi x_1) - \sin(2\pi x_2). \tag{3.5.30}$$

As in the previous example, we let the conditional log odds ratio be a constant

$$\alpha = \log OR(Y_1, Y_2|x_1, x_2) = 1.5. \tag{3.5.31}$$

The true marginal probability is plotted in Figure 8(a).

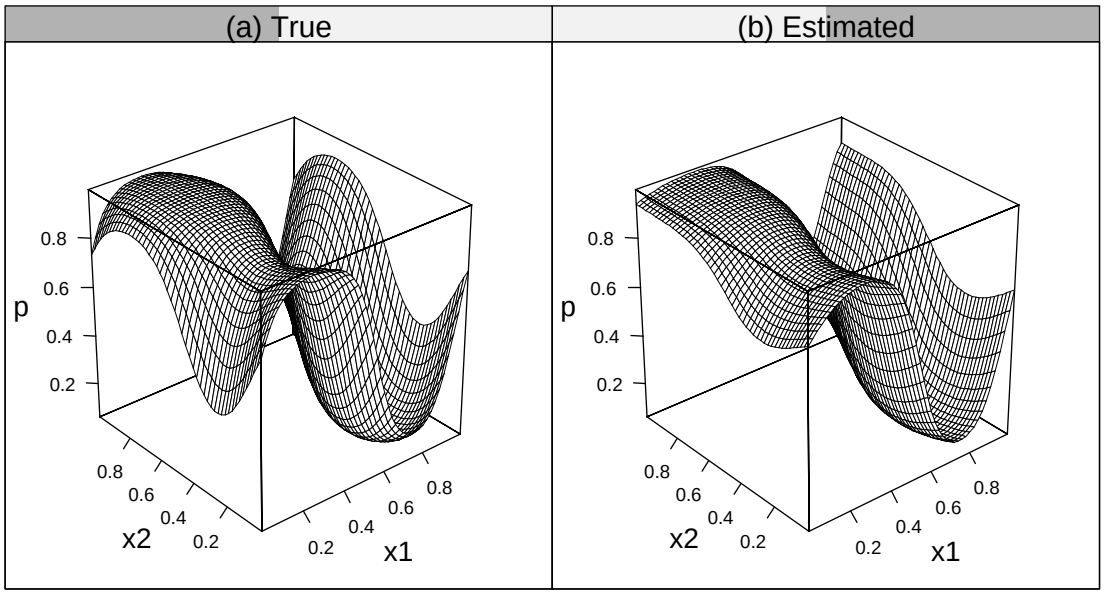

Figure 8: The true and estimated marginal probability function $p(x_1, x_2) = P(Y_k = 1|x_1, x_2)$.

For this simulation study, 500 pairs of observations are generated according to the joint distribution. 50 basis functions are selected by clustering method. We apply the Block one-step SOR algorithm combined with iterated $ranGACV$ to estimate the joint distribution $P(Y_1, Y_2|x_1, x_2)$. $R = 20$ replicates are used for estimating $ranGACV$. The estimated marginal probability is plotted in Figure 8(b). Figure 8 and Figure 9 show the perspective plots and contour plots for both $ranGACV$ and CKL surfaces. Three comparisons are made during the iteration process: at the first step (when $\hat{\alpha} = 0$), in the middle of the iterations (when $\hat{\alpha} = 0.77$) and at the converged value (when $\hat{\alpha} = 1.29$).

From the plots, iterated $ranGACV$ does an excellent job in terms of searching for the minimum value of CKL distance. Although the minimizers of $ranGACV$ are not the minimizers of CKL distance, considering the flat nature of CKL surface near its minima in this case, we notice that the CKL distances achieved by the minimizers of $ranGACV$ are very close to the minimum values of CKL distance. The comparison of the minimum CKL values and the one achieved by the minimizers of $ranGACV$ is listed in Table 3.

	$\min_{\lambda_1,\lambda_2} CKL(\lambda_1, \Lambda_2)$	$CKL(\hat{\lambda}_1, \hat{\lambda}_2)$
$\hat{\alpha} = 0$	0.88501	0.88912
$\hat{\alpha} = 0.77$	1.22738	1.23200
$\hat{\alpha} = 1.29$	1.50359	1.50903

Table 3: Comparison of the minimum CKL distances and CKL achieved by $(\hat{\lambda}_1, \hat{\lambda}_2)$, the minimizers of $ranGACV$ function.

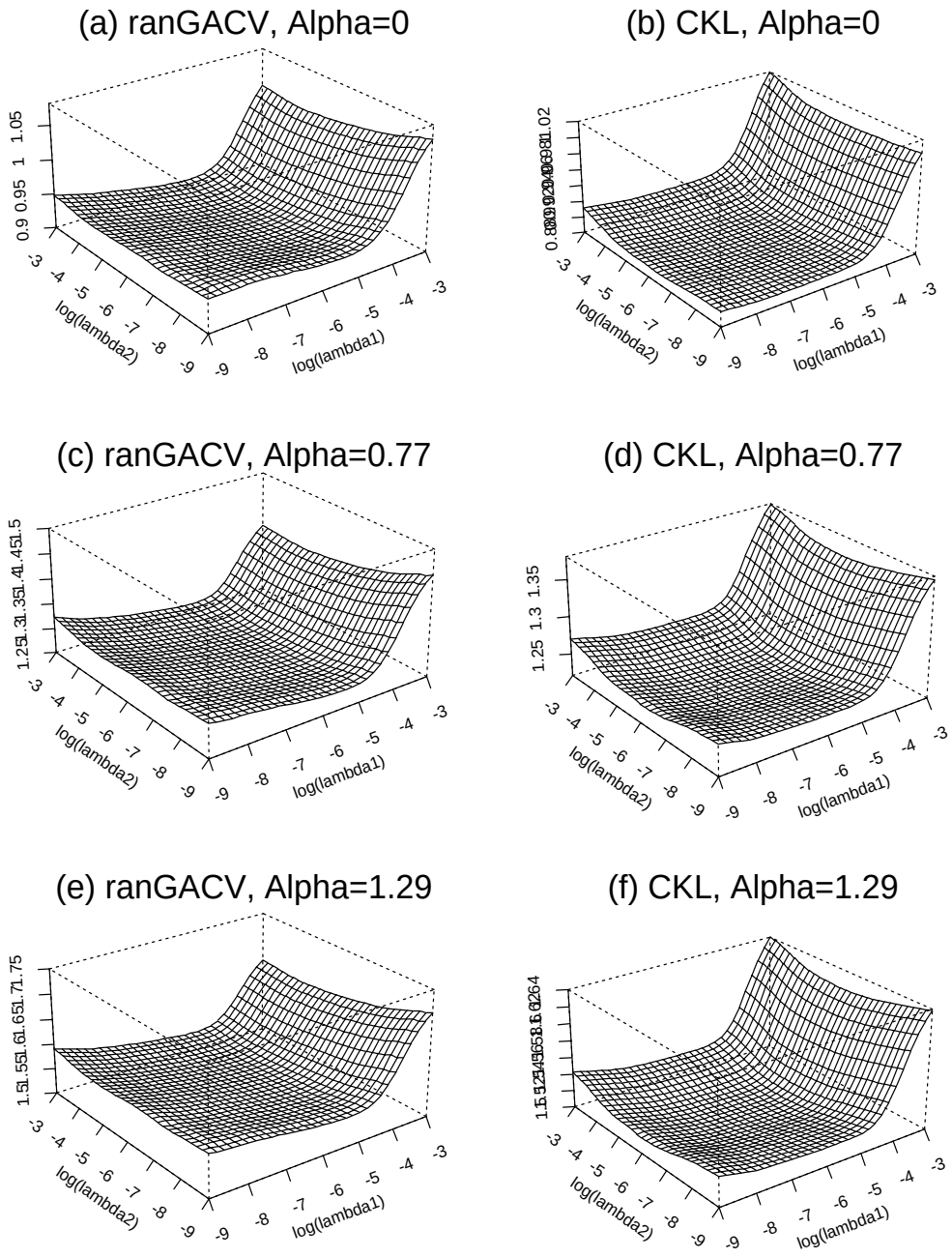

Figure 9: Comparison of iterated $ranGACV$ and CKL surfaces.

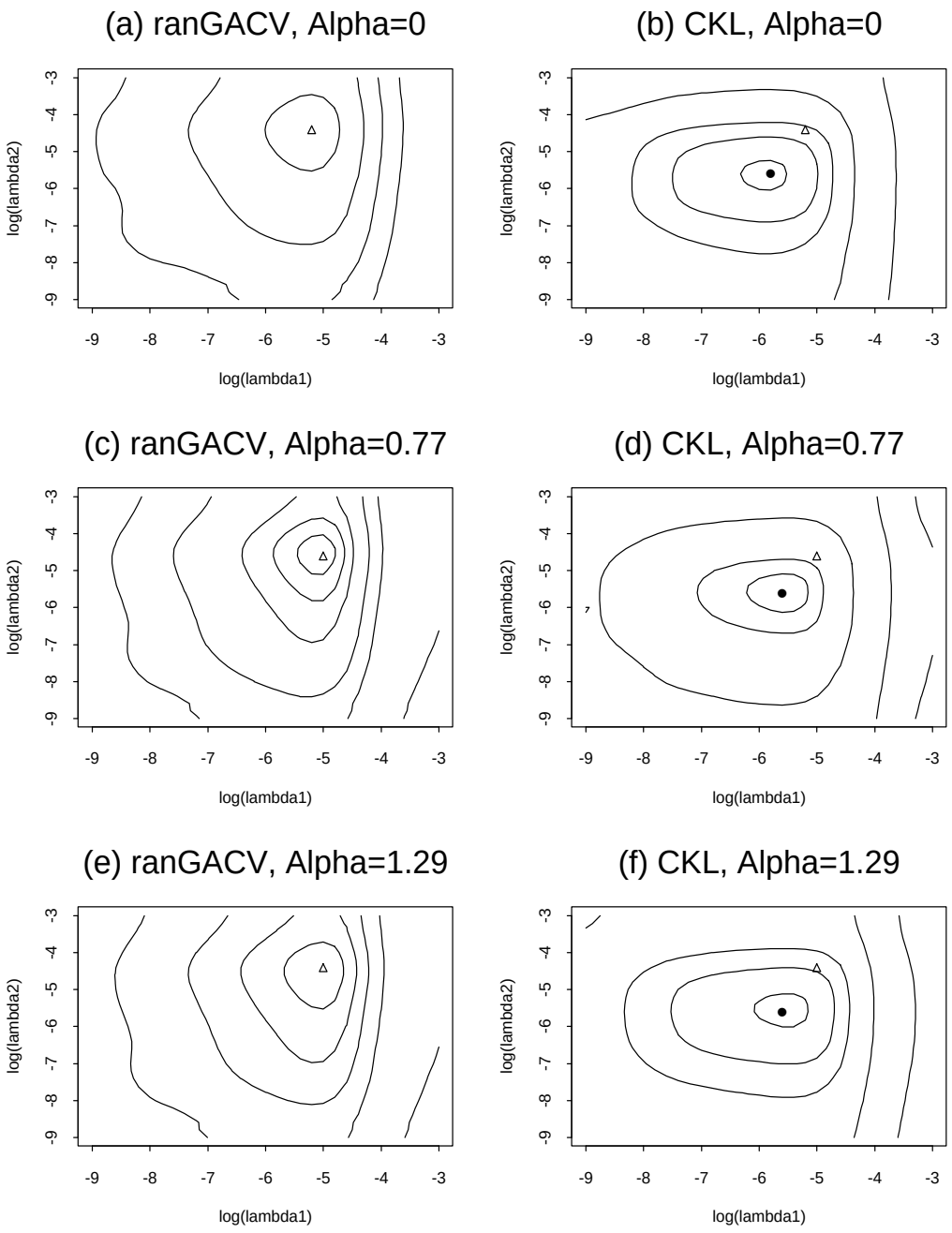

Figure 10: Contour plots of iterated $ranGACV$ and CKL. Solid dots denote the minimizer of CKL distance while the triangles denote the minimizer of $ranGACV$ functions.

3.6 Bayesian Inference and Approximate Confidence Intervals

Theorem 3.3 shows that the pseudo-data defined in section 3.3 have approximately the usual data structure. We will make use of such an observation in this section to construct the approximate Bayesian confidence interval. An approach similar to that used by Silverman (1985) is adapted for the approximate spline solution to the variational problem.

First let us consider the Bayesian formulation of the variational problem associated with correlated Gaussian observations. For fixed smoothing parameter(s), we will identify the variational problem with a Bayesian problem. Assume there is only one endpoint, $J = 1$. On domain $\mathcal{X}$, $y_{ik} = f(x_{ik}) + \epsilon_{ik}$, $i = 1, ..., n$, $k = 1, ..., K$, where $(\epsilon_{i1}, ..., \epsilon_{iK})$, $i = 1, ..., n$ are $i.i.d.$ distributed as $N(0, \sigma^2 W^{-1})$, with W a known positive definite matrix. With abuse of notation, the approximate spline solution of $f(x)$ is a combination of the selected basis functions

$$f = Sd + Q_V c, \tag{3.6.1}$$

where $Q_V = (\phi_1, ..., \phi_V)$. Let Q_V^* denote the matrix with $(Q_V^*)_{ij} =< \phi_i, \phi_j >$. By assuming an improper prior distribution on the coefficients (c, d), we let their log-density function take the form

$$l_{prior}(c, d) \stackrel{c}{=} -\frac{1}{2} b c^T Q_V^* c, \tag{3.6.2}$$

where $b = (n\lambda)/\sigma^2$ and the notation "$\stackrel{c}{=}$" means "equals up to a constant". Following some standard Bayesian manipulation, the posterior log-likelihood has the following form

$$l_{post}(c,d) \stackrel{c}{=} -\frac{1}{2} b c^T Q_V^* c - \frac{1}{2\sigma^2}(y - Q_V c - Sd)^T W (y - Q_V c - Sd). \qquad (3.6.3)$$

Hence by minimizing the posterior negative log-likelihood of (c,d), we obtain exactly the same solution as solving the variational problem in the approximating subspace $span(S, Q_V)$.

From (3.6.3), (c,d) in fact has a proper posterior distribution as a multivariate normal with mean $(\hat{c},\hat{d})$ and covariance matrix $\sigma^2 M^{-1}$, where

$$M = \begin{pmatrix} Q_V^T W Q_V + n\lambda Q_V^* & Q_V^T W S \\ S^T W Q_V & S^T W S \end{pmatrix} \qquad (3.6.4)$$

and

$$\begin{pmatrix} \hat{c} \\ \hat{d} \end{pmatrix} = M^{-1} \begin{pmatrix} Q_V^T \\ S^T \end{pmatrix} WY. \qquad (3.6.5)$$

Hence, for $f = Sd + Q_V c$, the following is true

$$Var(f) = \sigma^2 \begin{pmatrix} Q_V^T \\ S^T \end{pmatrix} M^{-1} \begin{pmatrix} Q_V & S \end{pmatrix}. \qquad (3.6.6)$$

Define the influence matrix $A(\lambda)$ satisfying $f = A(\lambda)y$ to be

$$A(\lambda) = \begin{pmatrix} Q_V^T \\ S^T \end{pmatrix} M^{-1} \begin{pmatrix} Q_V & S \end{pmatrix} W. \qquad (3.6.7)$$

(3.6.6) can be re-written as

$$Var(f) = \sigma^2 A(\lambda) W^{-1}. \tag{3.6.8}$$

Therefore, Bayesian confidence intervals can be constructed once the posterior mean and covariance matrix are computed for (c, d).

The construction of Bayesian confidence intervals for multivariate Bernoulli data utilizes the fact that the pseudo-data have approximately multivariate normal distribution, which is based on the Taylor expansion of the penalized log-likelihood function centered at the mode (c, d). Denote the negative log-density function of y conditioning on f and α as $l(y|f, \alpha)$. To estimate the conditional logit function for the jth endpoint f_j, we will condition on the other estimated values for $f^{(-j)}$ and α. f_j is the minimizer of

$$l_j(f_j) + \frac{n}{2}\lambda \mathbf{J}^j(f_j) = l_j(f_j) + \frac{n}{2}\lambda c_j^T Q_V^* c_j. \tag{3.6.9}$$

At the converged step of the block one-step SOR iteration, we are actually solving a penalized weighted least square problem based on the pseudo-data

$$\frac{1}{n}\sum_{i=1}^{n}(\tilde{y}_{ij} - f_{ij})^T W_{ij-}(\tilde{y}_{ij} - f_{ij}) + \lambda {c^j}^T Q_V^* c^j. \tag{3.6.10}$$

Here W_{ij-}^{-1} is an estimated value of $Var(Y_j) = W_j^{-1}$. From Theorem 3.3, we know that $\tilde{y}_j$ is approximately distributed as $N(f_j, W_j^{-1})$. Hence by dealing with the pseudo-data $\tilde{y}_j$, similar to (3.6.5), we have

$$\begin{pmatrix} \hat{c} \\ \hat{d} \end{pmatrix} = M^{-1} \begin{pmatrix} Q_V^T \\ S^T \end{pmatrix} W\tilde{y}_j \tag{3.6.11}$$

where M is evaluated at the converged step of the iterations as in (3.6.4). To calculated the posterior variance of f, (3.6.6) remains to be true. Therefore, the pseudo-data can be used to construct the approximate Bayesian confidence interval for the multivariate Bernoulli data.

3.7 Monte Carlo Simulations

In this section, we will demonstrate results from some Monte Carlo simulations to evaluate the performance of the proposed method. The comparative Kullback-Leibler distance (CKL) is used to measure the performance of the estimated values.

3.7.1 Repeated Measurements for the Same Endpoint

The first example is about the single smoothing parameter situation. We will try to mimic the characteristic of possible ophthalmology data. There is one endpoint of interest and paired observations for each subject. There presents one observation-specific covariate $X_{ik}, (k = 1, 2)$. X_{i1}'s are assumed to be uniformly distributed on the interval $(0.05, 0.95)$. $X_{i2} = X_{i1} + \epsilon_i$, while ϵ_i's are uniformly distributed on $(-0.05, 0.05)$.

The true conditional logit function is assumed to be

$$\begin{aligned} f(x_{ik}) &= logit(P(Y_{ik} = 1 | Y_i^{(-k)} = 0, x_{ik})) \\ &= 2[\exp(-30(x_{ik} - 0.25)^2) + \sin(\pi x_{ik}^2)] - 2. \end{aligned} \tag{3.7.1}$$

And the conditional log odds ratio $\alpha = \log OR(Y_{i1}, Y_{i2}|x_i) = 0.8$. Three different sample sizes are used in this simulation: $n = 125$, $n = 250$, $n = 500$. For each sample size, 100 independent sets of data are randomly generated according to the true joint distribution.

Figure 11 shows the histogram plots of the estimated $\hat{\alpha}$ for the three different sample sizes. The dotted lines represent the true value of 0.8. The fitted values appear to converge to the truth while the sample size increases. The estimator of α appears to be approximately unbiased and normally distributed from the histogram.

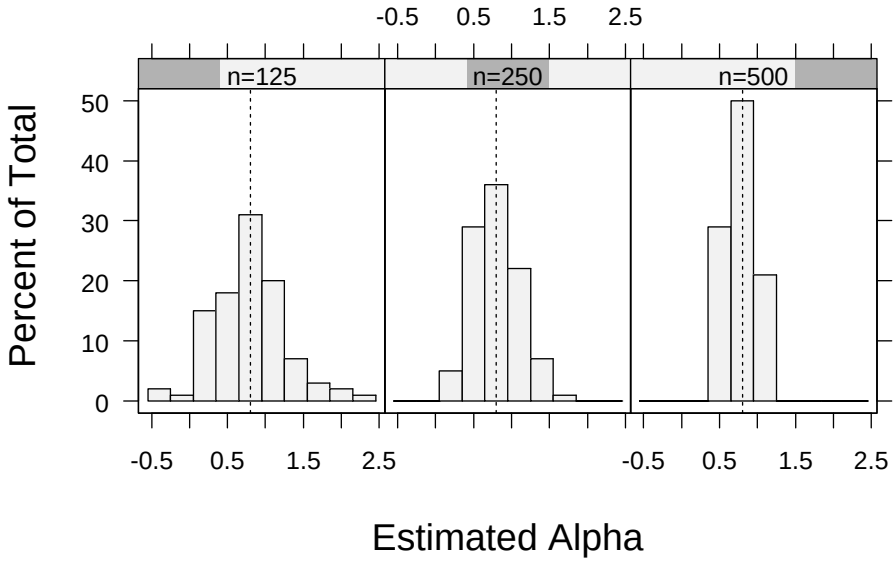

Figure 11: Histogram of $\hat{\alpha}$ for three different sample sizes. The dotted lines represent the true value of $\alpha = 0.8$.

In Figure 12, 13 and 14, we plot the true conditional probability function

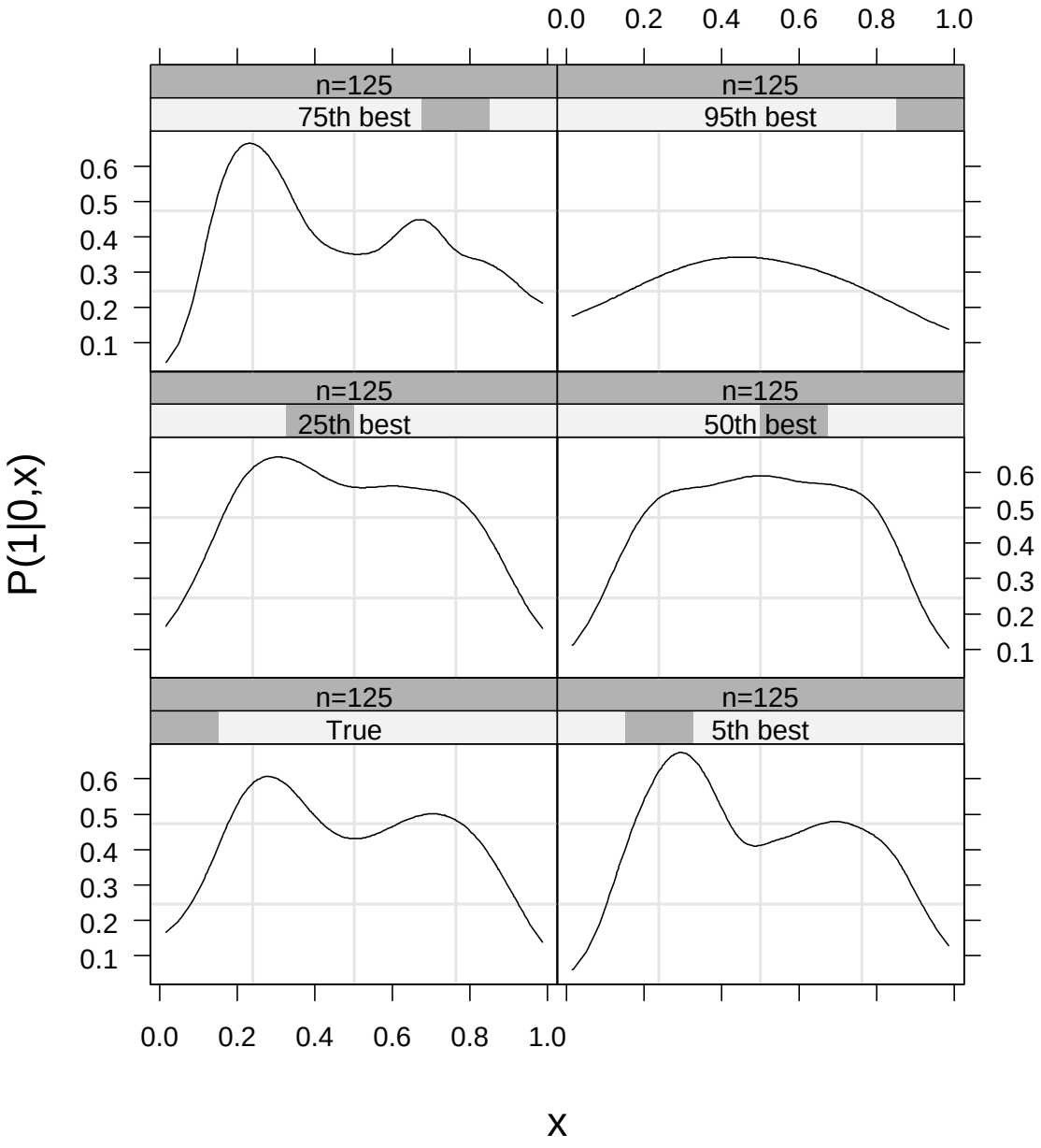

Figure 12: True and estimated conditional probability functions when $n = 125$.

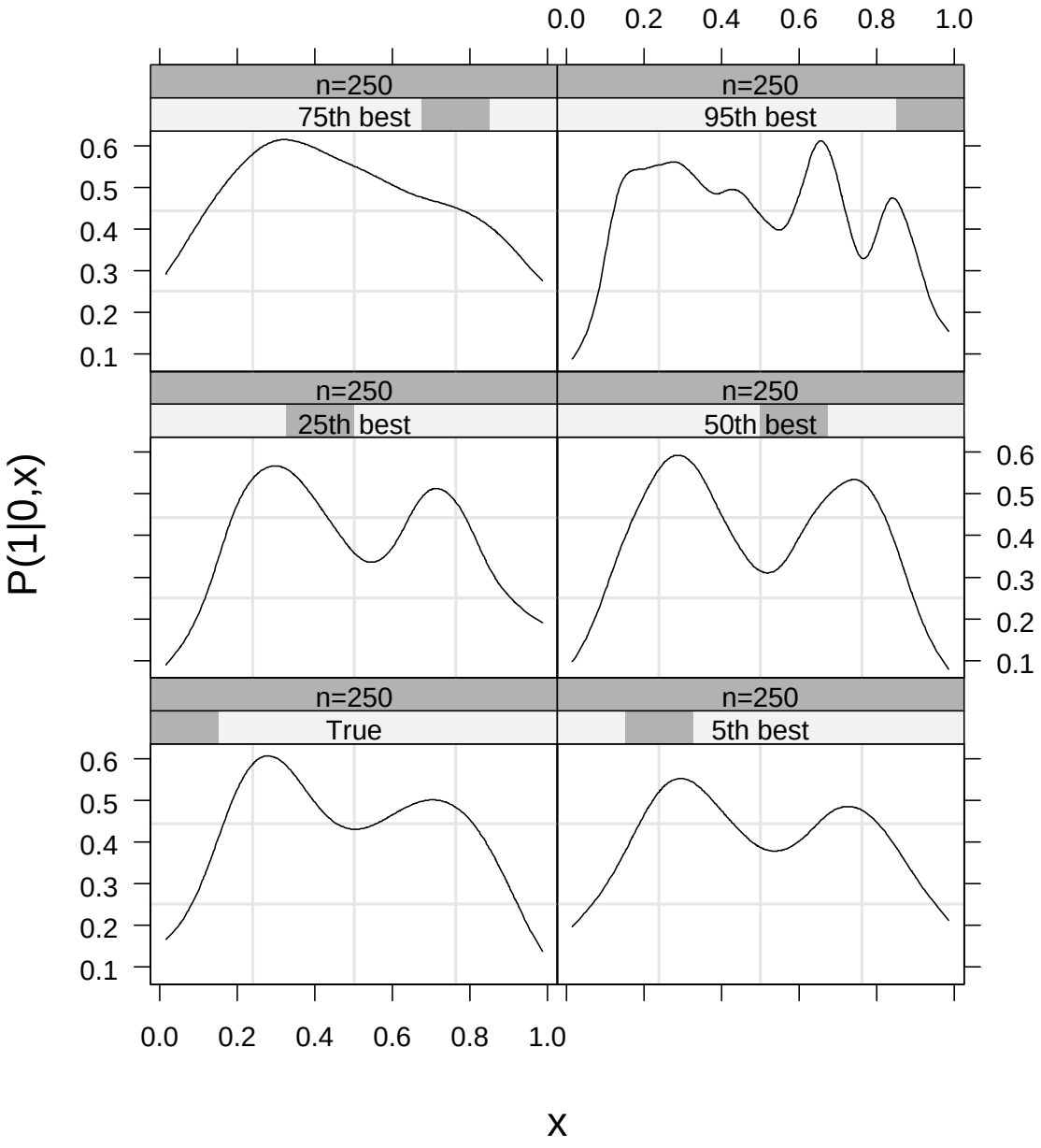

Figure 13: True and estimated conditional probability functions when $n = 250$.

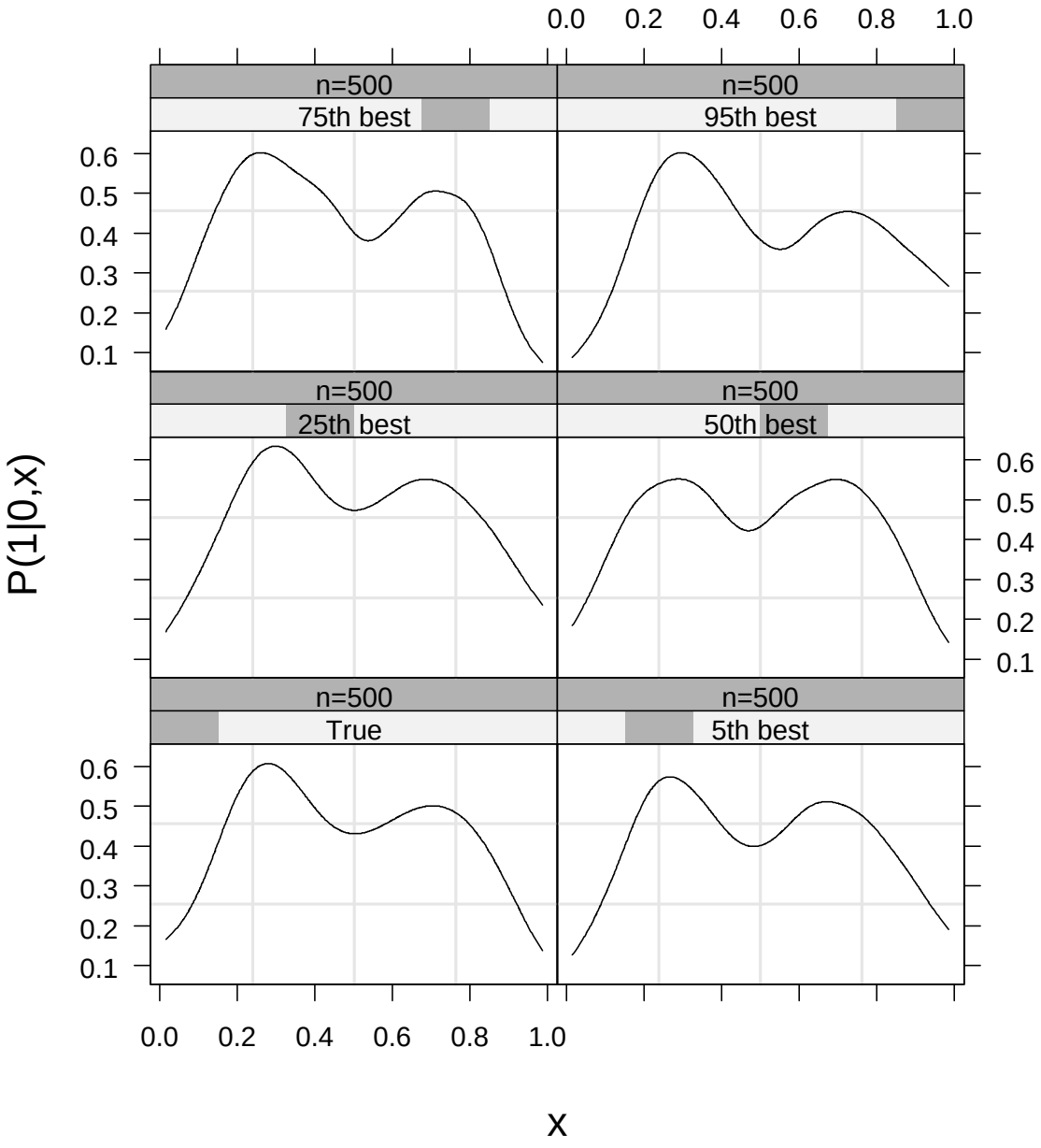

Figure 14: True and estimated conditional probability functions when $n = 500$.

and the estimated curves for each sample size, $P(Y_{ik} = 1|Y_i^{(-k)} = 0, x_{ik}) = e^{f(x_{ik})}/(1 + e^{f(x_{ik})})$. For each sample size, the 100 fitted values are ranked according to the CKL distances between the estimated joint distributions and the truth. The 5th, 25th, 50th, 75th and 95th best fits are plotted for each sample size. The true conditional logit function is a bi-modal function. The trend is clear that when the sample size increases, the estimated curves become more and more accurate. However, for parametric model, there might be no prior knowledge about the bi-modal nature of the truth. Hence a linear or even quadratic form will miss the true curve no matter how large the sample size is.

In the next experiment, we will compare the proposed new multivariate method to the univariate fit. In the ophthalmology studies, one question of interest is to estimate the probability of at least one eye developing a certain disease given the values of the predictor variables for a person. Assuming there is no eye-specific covariate. X_i's are uniformly distributed on $(0, 1)$. For each subject, there are paired observations (Y_{i1}, Y_{i2}). We want to estimate the probability $P(Y_{i1} = 1 \vee Y_{i2} = 1|x_i) = (2e^{f_i} + e^{2f_i+\alpha_i})/(1 + 2e^{f_i} + e^{2f_i+\alpha})$ from the observed data.

For this experiment, we assume

$$p(x_i) = P(Y_{i1} = 1 \vee Y_{i2} = 1|x_i) = 0.8 \sin(2.7x_i^2) + 0.1 \qquad (3.7.2)$$

The true $p(x)$ is plotted in Figure 15. Four different values are used for α: 0, 0.4, 0.8, 1.2. $\alpha = 0$ is corresponding to the case that Y_{i1} and Y_{i2} are independent. However we pretend that this fact is unknown, α is still estimated by the

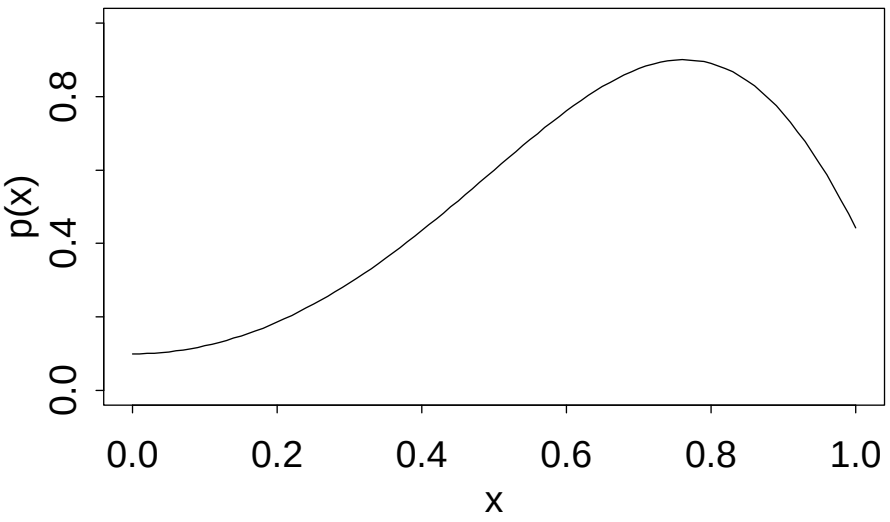

Figure 15: True $p(x_i) = P(Y_{i1} = 1 \vee Y_{i2} = 1|x_i)$ used for the simulation study.

proposed algorithm. Straightforward calculation yields the following formula to compute f_i for given α and $P(Y_{i1} = 1 \vee Y_{i2} = 1|x_i)$

$$f_i = \log \frac{(p(x_i) - 1) + \sqrt{(1 - p(x_i))^2 + e^{\alpha} p(x_i)(1 - p(x_i))}}{e^{\alpha}(1 - p(x_i))}. \tag{3.7.3}$$

The experiment is conducted as follows. First, for the univariate fit, the only information needed is $\check{Y}_i$ which is defined to be 0 when both $Y_{i1} = Y_{i2} = 0$ and 1 otherwise. $P(\check{Y}_i = 1|x_i) = p(x_i)$. We generate 100 sets of data according to the true distribution and fit the data by using univariate penalized logistic regression. For the bivariate fit, we first calculate the true joint distribution of (Y_{i1}, Y_{i2}) according to the previous formula. For each value of α, 100 sets of data are randomly generated and the joint distribution is estimated by the proposed multivariate method. Afterwards, the probability of $P(Y_{i1} = 1 \vee Y_{i2} = 1|x_i)$ can

be derived from the estimated joint distribution. For every run, CKL distance between the estimated $\hat{p}(x_i)$ and $p(x_i)$ is calculated.

The above procedure is performed for three different sample sizes: $n = 100$, $n = 200$ and $n = 400$. In Figure 16, we show the histograms of the estimated $\hat{\alpha}$'s for different sample sizes and true values of α. Dotted lines represent the true values of α. From the plot, the estimated values have an approximate bell-shaped distribution and are approximately unbiased. When sample size increases, the estimated values become closer to the true value.

In Figure 17, we compare the CKL distances between the fitted probability and the true probability $p(x_i) = P(Y_{i1} = 1 \vee Y_{i2} = 1|x_i)$ for different method. Obviously, for all true values of α, the bivariate fit, which estimates the joint distribution of (Y_{i1}, Y_{i2}), has a better efficiency than the univariate fit, which estimates $P(\check{Y}_i = 1)$ directly. This is not surprising since the univariate fit only needs to know $\check{Y}_i$, hence some information in (Y_{i1}, Y_{i2}) is not used in the estimation procedure.

The next experiment is similar to the previous one but for multiple smoothing parameters. Assume (X_{i1}, X_{i2})'s are uniformly distributed on the unit square $(0, 1) \times (0, 1)$. The true conditional logit function is taken to be

$$f(x_{i1}, x_{i2}) = 2\sin(3x_{i1} - 3x_{i1}x_{i2}) + \cos(2 - 2x_{i2}) - 3(x_{i1} - 0.35)^2 - 1.5 \tag{3.7.4}$$

and the conditional log odds ratio α is taken to be a constant 1. Each time, 500

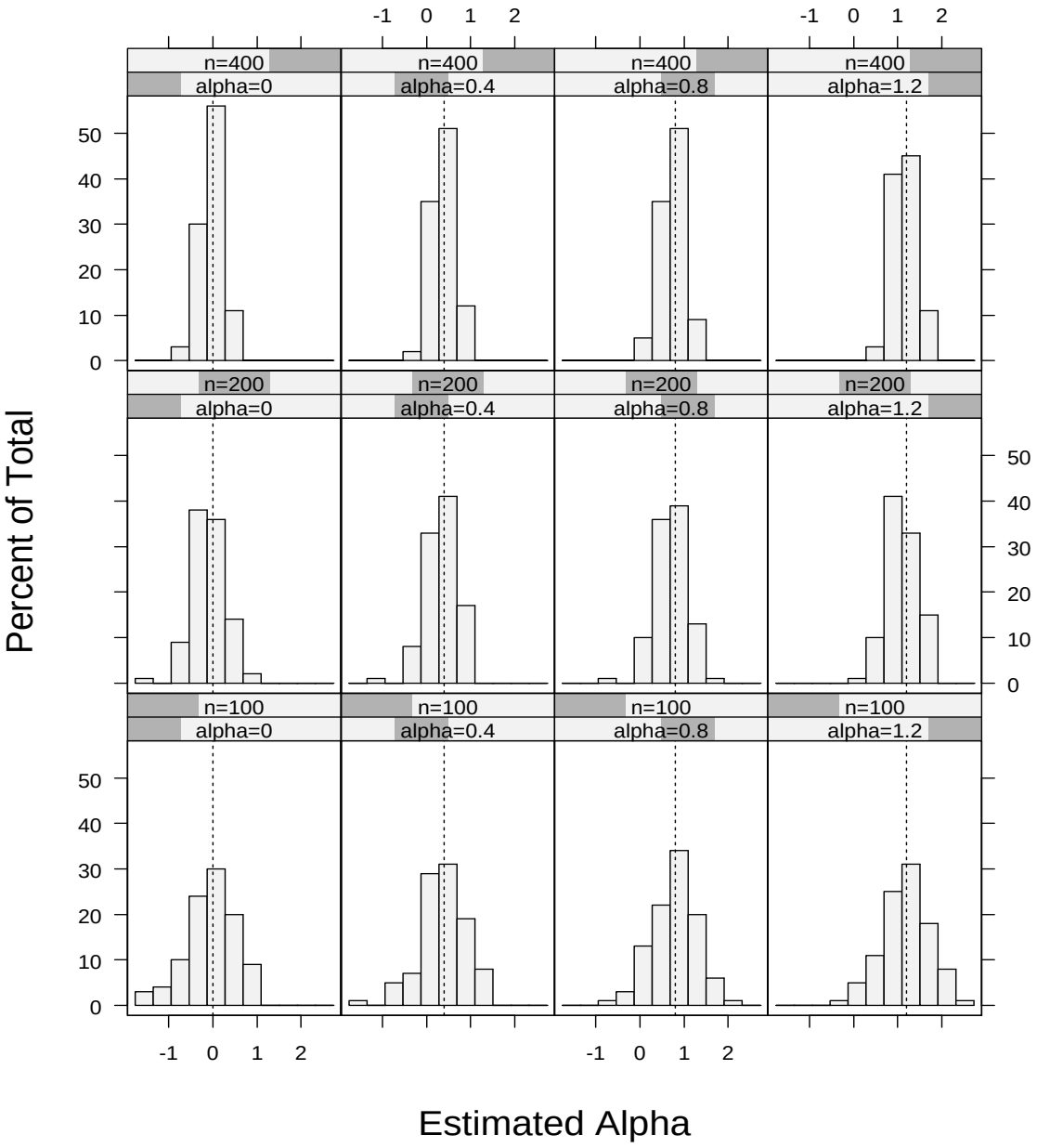

Figure 16: Histograms of estimated $\hat{\alpha}$'s for $n = 100$, $n = 200$ and $n = 400$. Dotted lines represent the true values of α.

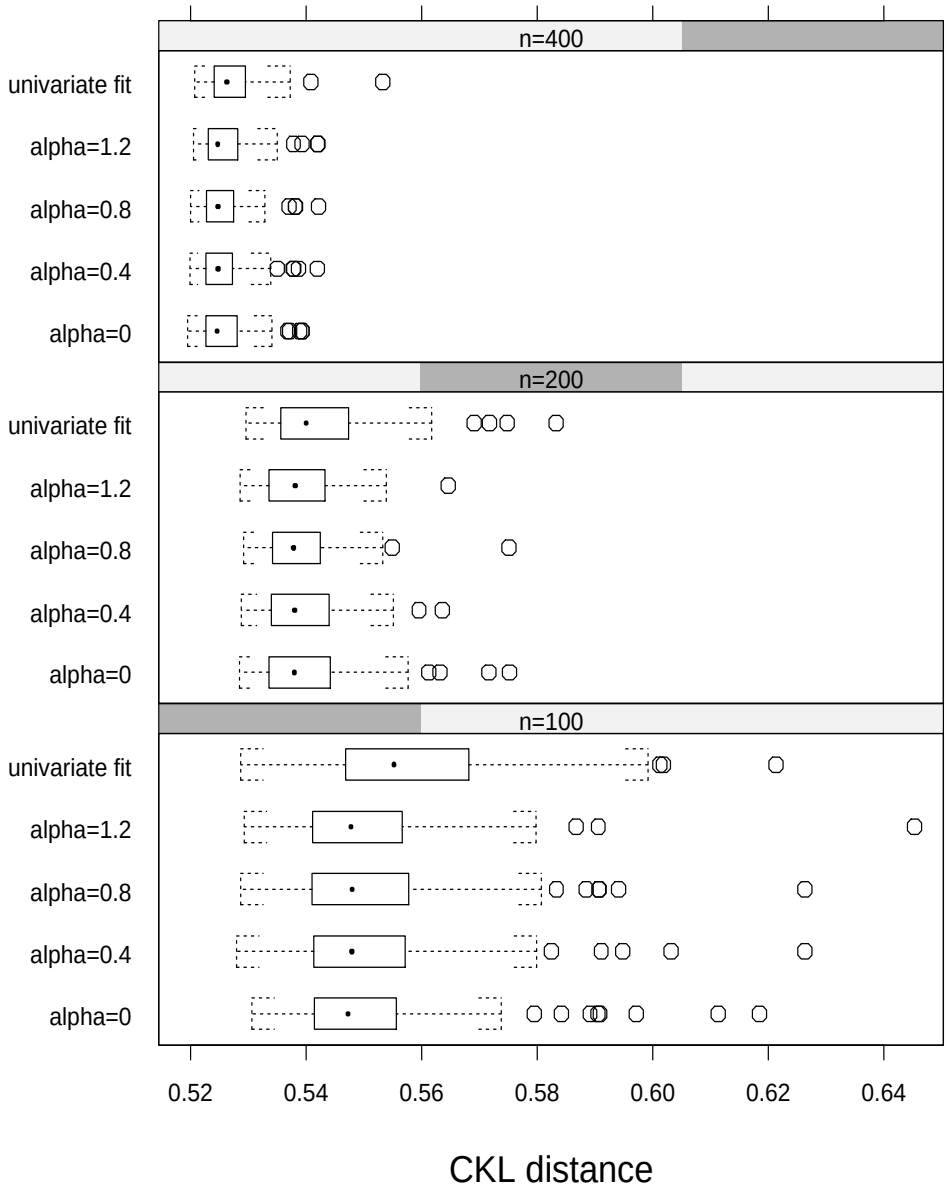

Figure 17: Boxplots of CKL for the univariate fit and the multivariate fits.

independent pairs of observations (Y_{i1}, Y_{i2})'s are simulated. The proposed penalized multivariate logistic regression is used to estimate the joint distribution. This is repeated for 100 times.

We can derive $p(x_{i1}, x_{i2}) = P(\check{Y}_i = 1|x_{i1}, x_{i2})$ from the estimated joint distribution. Figure 18 shows the true $p(x_{i1}, x_{i2})$ and the 5th, 25th, 50th, 75th and 95th best estimated values ranked by the CKL distance. The proposed method gives very good estimations most of the times.

To make the comparison, we also use the univariate method to estimate $p(x_{i1}, x_{i2})$ directly for the same 100 sets of data. Only the derived outcome variable $\check{Y}_i$ is used in the estimation procedure. Assuming we are only interested in estimating $P(\check{Y}_i = 1|x_{i1}, x_{i2})$, the pairwise comparison of CKL distance is shown in Figure 19. About 2/3 of the times, the bivariate fit yields better estimation.

3.7.2 Different Endpoints

In this example, we assume that there are two correlated endpoints of interest. For each subject, there are two binary outcome variables: Y_{i1} for the first endpoint and Y_{i2} for the second endpoint. The proposed method will estimate the conditional joint distribution of $P(Y_{i1}, Y_{i2}|X_i)$. This model is also useful to predict the outcome of one endpoint, given the outcome of another endpoint is known. For example, if a person already has one disease, what is the probability of getting another disease?

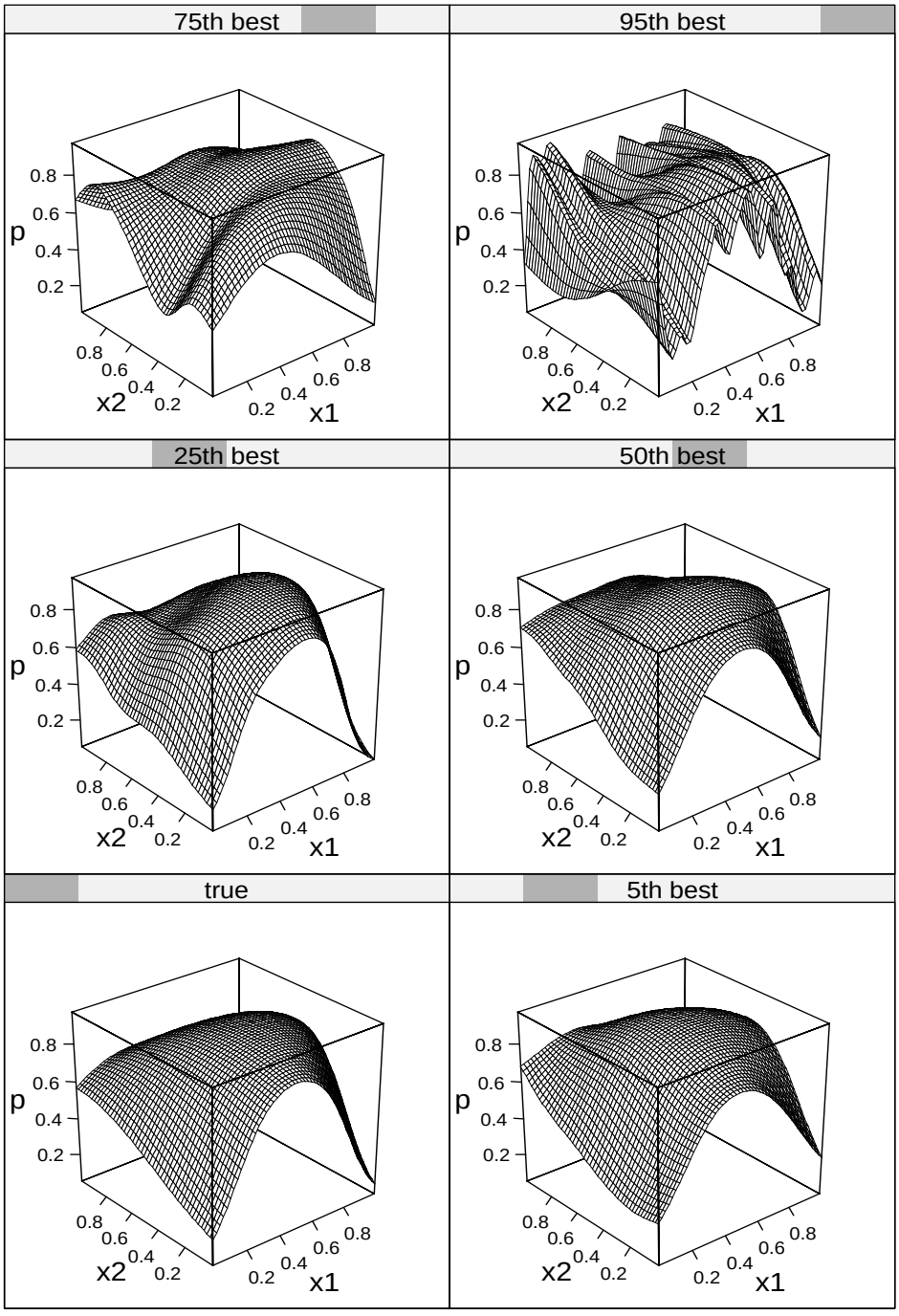

Figure 18: True $p(x_{i1}, x_{i2}) = P(\check{Y}_i = 1 | x_{i1}, x_{i2})$ and estimated surfaces.

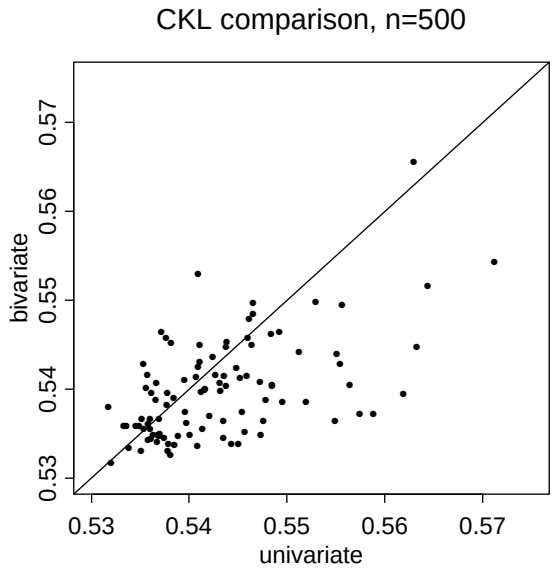

Figure 19: Pairwise comparison of CKL distance for the bivariate fit and the univariate fit.

The true association factor $\alpha = \log OR(Y_{i1}, Y_{i2})$ is taken to be 1.5 in this simulation. The true conditional logit functions for the two different endpoints are

$$f_1(x_i) = logit(P(Y_{i1} = 1|Y_{i2} = 0, x_i)) = 10\cos(2x_i) + 7e^{x_i^2} - 16 \qquad (3.7.5)$$

and

$$f_2(x_i) = logit(P(Y_{i2} = 1|Y_{i1} = 0, x_i)) = 2\cos(5x_i + 1.4) + x_i^2. \qquad (3.7.6)$$

Two sample sizes ($n = 200$ and $n = 500$) are used in this simulation. For each sample size, 100 sets of independent data are generated according to the true joint distribution. The predictor variables X_i are assumed to have uniform distribution over $(0, 1)$. Only 50 basis functions are selected to generate the

approximating subspace for the approximate spline solutions. To compute the randomized version of $GACV$, we use $R = 20$ replicates to reduce the variance of the estimated values.

In Figure 20, we present the histogram plots of the estimated $\hat{\alpha}$ for two different sample sizes. The dotted lines are the true value of $\alpha = 1.5$. The estimated values converge to the truth while sample size increases.

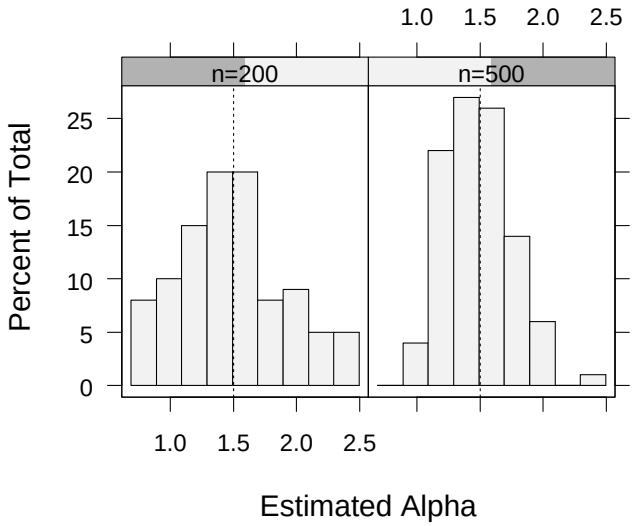

Figure 20: Histograms of estimated $\hat{\alpha}$ for two different sample sizes. The dotted lines represent the true values of $\alpha = 1.5$.

In Figure 21 and 22, we plot the true and estimated conditional probability functions for both endpoints. For each sample size, the 100 fitted values are ranked according to the CKL distance between the estimated joint distribution and the truth. The 5th, 25th, 50th, 75th, 95th best fits are plotted for both sample sizes. Figure 21 shows the conditional probability for the first endpoint

$P(Y_{i1} = 1|Y_{i2} = 0, x_i) = e^{f_1(x_i)}/(1 + e^{f_1(x_i)})$. Figure 22 shows the conditional probability for the second endpoint $P(Y_{i2} = 1|Y_{i1} = 0, x_i) = e^{f_2(x_i)}/(1 + e^{f_2(x_i)})$.

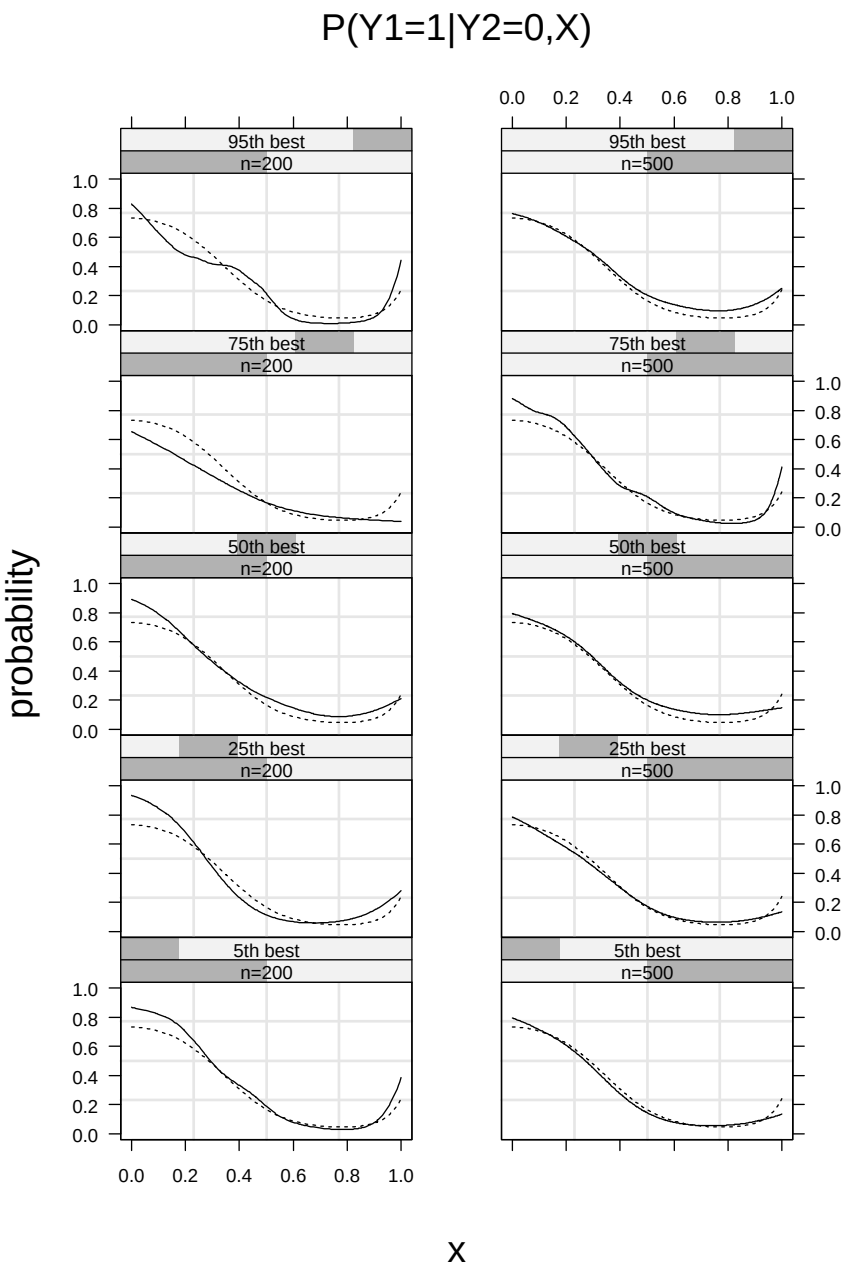

Figure 21: True and estimated conditional probability $P(Y_{i1} = 1|Y_{i2} = 0, X_i)$. Solid lines are the estimated functions while dotted lines represent the true function.

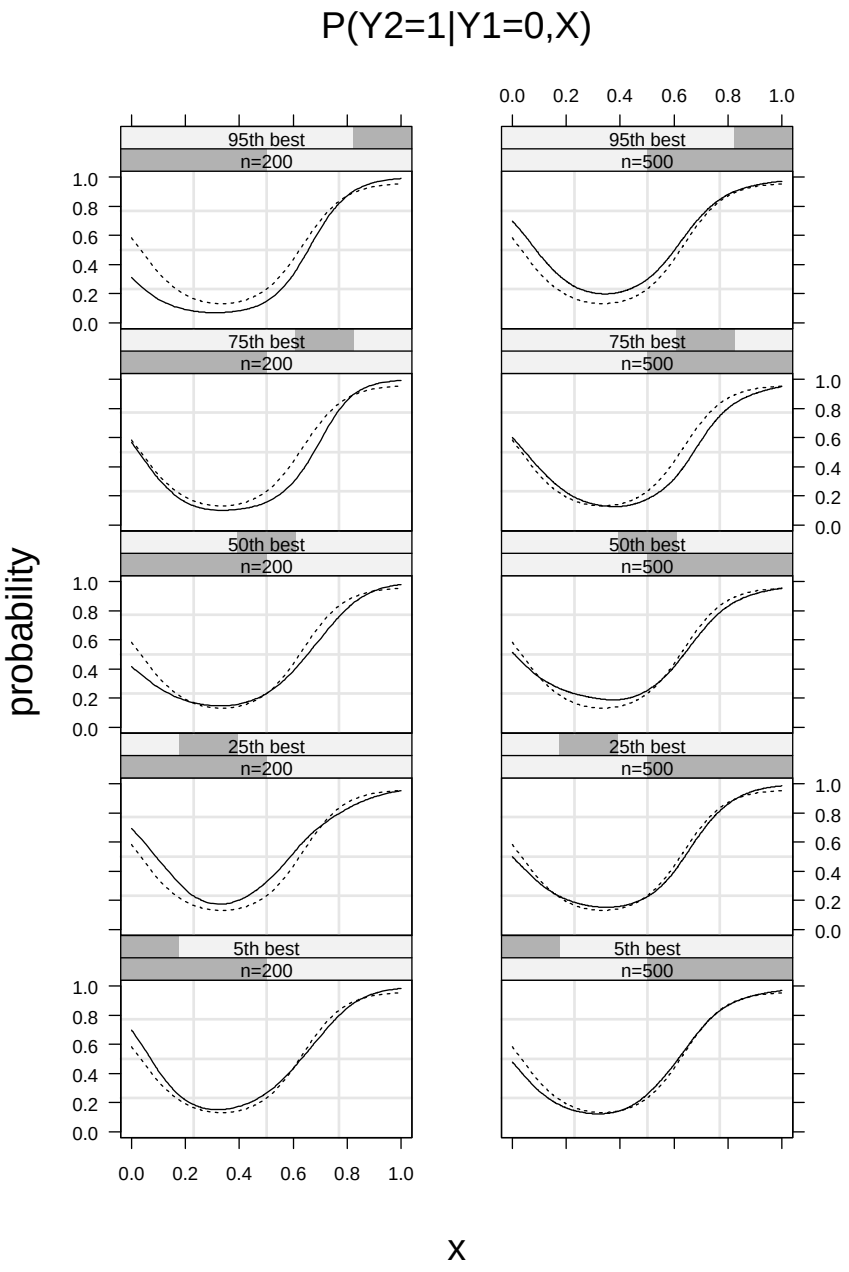

Figure 22: True and estimated conditional probability $P(Y_{i2} = 1|Y_{i1} = 0, X_i)$. Solid lines are the estimated functions while dotted lines represent the true function.

Chapter 4

Application to the Beaver Dam Eye Study

4.1 Introduction

The Beaver Dam Eye Study (BDES) is an ongoing population-based cohort study of age-related eye diseases, cataract and maculopathy. A description of the population and details of the study at the baseline may be found in Klein, Klein & Linton (1992). Five-year followup data has now been collected and analyzed, see, for example, Klein, Klein, Jensen & Meuer (1997*b*), and the ten-year followup of the cohort is in progress.

A private census of the population of Beaver Dam, Wisconsin was performed from September 15, 1987 to May 4, 1988 to identify the eligible population, which is defined as being 43 to 84 years of age at the time of census. Afterwards, the population was examined over a 30-month period. Of the 5925 eligible people, 4926 (83.1%) participated in the study. Photographs of each eye were taken and graded. An examination and a standardized questionnaire were also administrated.

4.2 The Pigmentary Abnormalities for Women

The association of pigmentary abnormalities with six other attributes at the baseline was studied by the "univariate" penalized logistic regression in Lin et al. (1998). Only the $n = 2585$ women members of the cohort in the baseline with no missing values were considered. Pigmentary abnormalities are an early sign of age-related macular degeneration and are defined by the presence of retinal depigmentation or increased retinal pigmentation in association with retinal drusen. Pigmentary abnormalities were found in 11.88% of the $n = 2585$ cohort studied. Here, the question of interest is to estimate the probability of at least one eye developing pigmentary abnormalities given the values of the predictor variables.

Based on the previous work, age is known to be a very strong risk factor for the presence of pigmentary abnormalities and other age-related maculopathy in the Beaver Dam Eye Study. The association between cardiovascular disease and its risk factors and the incidence of age-related maculopathy was examined in Klein, Klein & Jensen (1997*a*). Hormone replacement therapy was associated with a weak protective effect while a history of heavy alcohol consumption and beer drinking was associated with a deleterious effect for some endpoints. See Klein, Klein & Ritter (1994), Ritter, Klein, Klein, Mares-Perlman & Jensen (1995) and Moss, Klein, Klein, Jensen & Meuer (1998) for references. We used multiple linear logistic regression and contingency tables for the preliminary analysis. First, one predictor variable was examined at a time. Only those

variables whose p-values are below some threshold (0.1) were kept for further analysis. A forward selection procedure was then carried out for the linear logistic regression. Afterwards, several possible forms of the model were closely examined by the nonparametric method. If the fitted value of any term had no significant visual effect to the overall fit, that term was considered to have no practical importance. The six "predictor" variables selected for the final nonparametric model are listed in Table 4.

Variable	units	code
current usage of hormone replacement therapy	yes/no	`horm`
history of heavy drinking	yes/no	`drin`
body mass index	kg/m^2	`bmi`
age	$years$	`age`
systolic blood pressure	$mmHg$	`sys`
serum cholesterol	mg/dL	`chol`

Table 4: Predictor variables for the Beaver Dam Pigmentary abnormalities model.

The model fitted there is

$$\begin{aligned} f(x) &= C + f_1(\texttt{sys}) + f_2(\texttt{chol}) + f_{12}(\texttt{sys},\texttt{chol}) \\ &\quad + d_{\texttt{age}}\texttt{age} + d_{\texttt{bmi}}\texttt{bmi} + d_{\texttt{horm}}I_1(\texttt{horm}) + d_{\texttt{drin}}I_1(\texttt{drin}). \quad (4.2.1) \end{aligned}$$

I_1 and I_2 are indicator variables. Originally, `age` and `bmi` were fitted as smooth main effects, however visual inspection indicated that they are indistinguishable from linear terms, so that they were set to be linear in the final model. Thus, there are 5 smoothing parameters in the model, one for each of the main effects

of `sys` and `chol`, another 3 for the interaction term ($linear_{\texttt{sys}} \otimes smooth_{\texttt{chol}}$, $smooth_{\texttt{sys}} \otimes linear_{\texttt{chol}}$, $smooth_{\texttt{sys}} \otimes smooth_{\texttt{chol}}$). The results were reported in Lin et al. (1998).

In this section, we will re-examine the association by using the proposed penalized multivariate logistic regression. $n = 2495$ women with outcomes available for both eyes are included in the analysis. For reference, the percentiles of the continuous predictor variables are given in Table 5.

Percentile	Min	12.5	25	37.5	50	62.5	75	87.5	Max
`sys`$(mmHg)$	71	108	116	122	129	136	145	157	221
`chol`(mg/dL)	102	191	210	225	237	252	266.5	290	503
`bmi`(kg/m^2)	15	22.5	24.25	25.9	27.4	29.5	31.55	35.2	68.4
`age`$(years)$	43	48	52	58	62	66	71	76	86

Table 5: Percentiles of the predictor variables.

In Table 6, we summarize the relation between the outcome variable and the categorical predictor variables.

We apply the penalized multivariate logistic regression to analyze these data. Here $J = 1$ and $K_1 = 2$. All predictor variables took the same values for both eyes of the same person. The association between fellow eyes is assumed to be a constant $\alpha = \log \dfrac{P(1,1|x_i)P(1,1|x_i)}{P(1,0|x_i)P(0,1|x_i)}$. The final model takes the same functional form as in (4.2.1), although this time on the conditional logit scale. Only 50 basis functions selected by the clustering method is used to fit the final model. To estimated the $ranGACV$, the number of replicates R is taken to be 20. Upon convergence, the estimated $\hat{\alpha} = 2.8269$. The naive estimate

	pigmentary abnormalities		
horm	no	one eye	both eyes
no	1953	184	104
yes	245	6	3

	pigmentary abnormalities		
drin	no	one eye	both eyes
no	2073	174	100
yes	125	16	7

Table 6: Summaries of the relation between the pigmentary abnormalities and the current usage of hormone replacement therapy and the heavy drinking history

of odds ratio without adjustment for any covariates is 26.06. The estimated odds ratio from the multivariate model goes down to $OR = e^{2.8269} = 16.89$. Obviously, the common predictor values for the same person explain partly the strong association between fellow eyes. We plot the estimated main effects of all predictor variables in conditional logit scale in Figure 23. Not surprisingly, `age` turns out to be the most influential predictor.

From the estimated joint probability, we can calculate the probability of at least one eye developing the pigmentary abnormalities. Figures 24 and 25 give the estimated probability of finding pigmentary abnormalities in at least one eye as a function of `chol`, for various values of `sys`, `age` and `bmi`. In Figure 24, (`horm, drin`)=(no, no) and in Figure 25, (`horm, drin`)=(yes, no). A suggestion of a nonlinear protective effect of cholesterol, particularly for those who were older in the `horm`=no group, may be seen as a result of fitting this

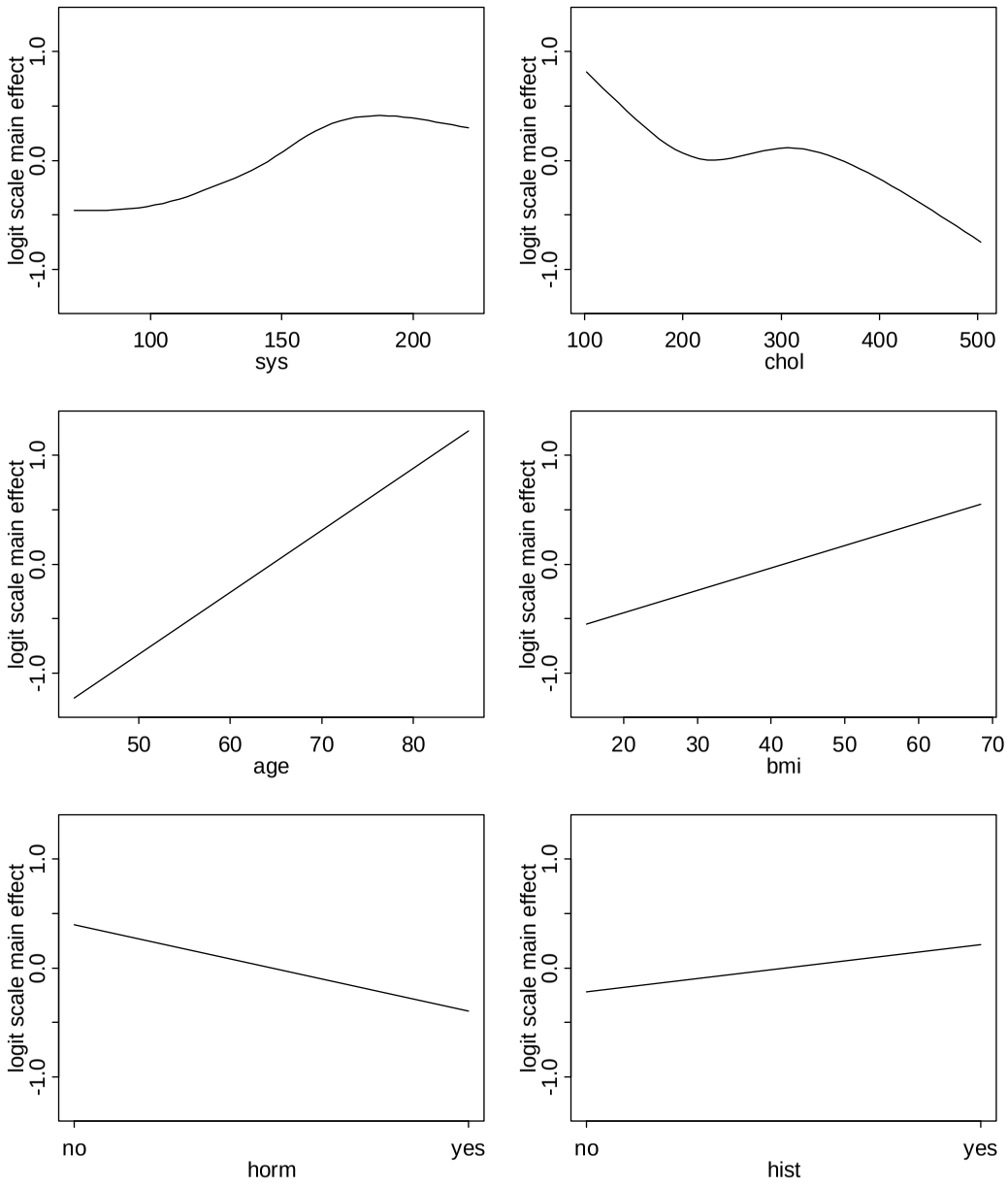

Figure 23: Estimated main effects in conditional logit scale for all predictor variables.

model. Figures 26 and 27 give the estimated probability of finding pigmentary abnormalities in at least one eye as a function of `sys`, for various values of `chol`, `age` and `bmi`. In Figure 26, (`horm, drin`)=(no, no) and in Figure 27, (`horm, drin`)=(yes, no). A protective effect of hormone replacement therapy is still evident from this bivariate model. Figure 28 gives cross sectional plots of the estimated probabilities along with the 90% Bayesian confidence intervals as a function of `chol` for both values of `horm` and four values of `age`, which are taken to be the middle of the four age groups defined in the Beaver Dam Eye Study.

The new analysis basically confirms the result obtained in Lin et al. (1998). The trend of the effect for each predictor variable remains the same. Compared to Figures 9-11 in Lin et al. (1998), we do notice some small difference between these two models. From the simulation studies, we expect that the new model is closer to the underlying truth. Besides, we notice that the outcomes for both eyes of the same person are highly correlated ($OR = e^{2.8269} = 16.89$), even after adjusted for all the predictor variables in this model. This partly explains why the results from the two models look very similar. When the outcomes are less correlated, or there are more repeated measurements for the same person, the "multivariate" method estimating the joint distribution is expected to extract more information from the data.

Another merit of this new approach is to estimate the probability $P(Y_k = 1|Y^{(-k)} = 1, X)$. Figure 29 shows this conditional probability as a function of `chol`. This conditional probability is medically meaningful to a patient who has

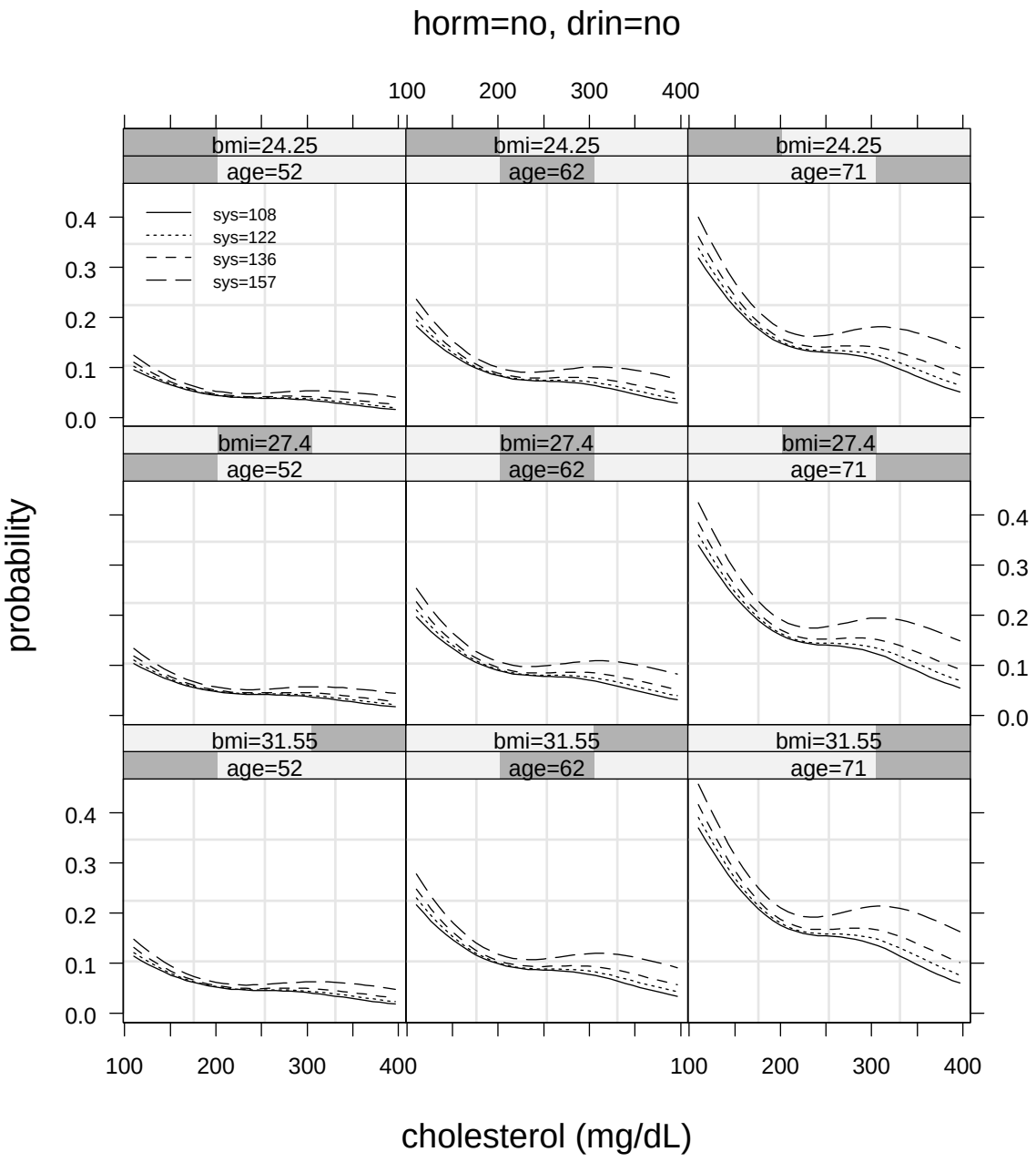

Figure 24: Estimated probability of at least one eye having the pigmentary abnormalities as a function of cholesterol by three levels of `age` and `bmi`. `horm`=no, `drin`=no.

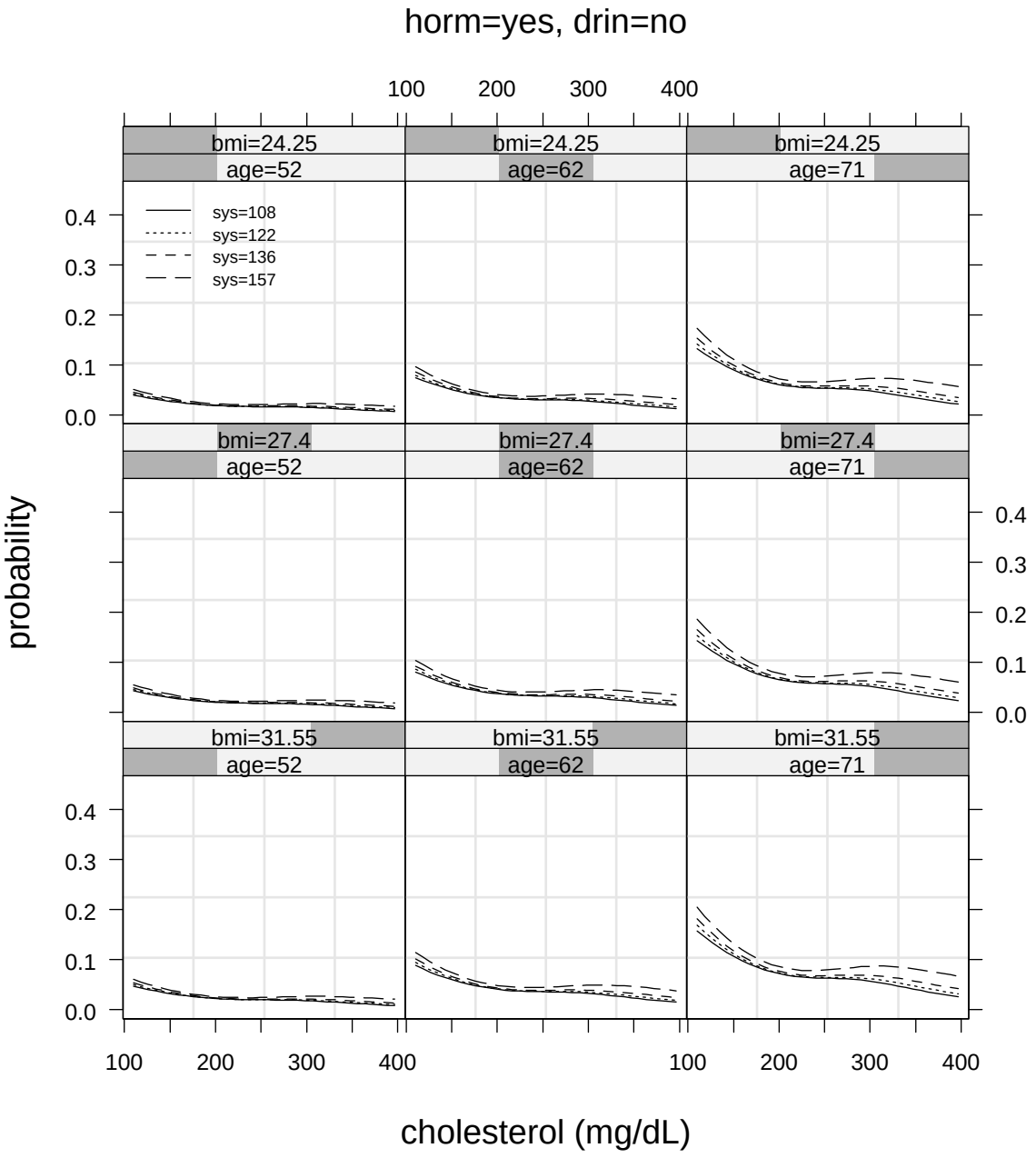

Figure 25: Estimated probability of at least one eye having the pigmentary abnormalities as a function of cholesterol by three levels of `age` and `bmi`. `horm`=yes, `drin`=no.

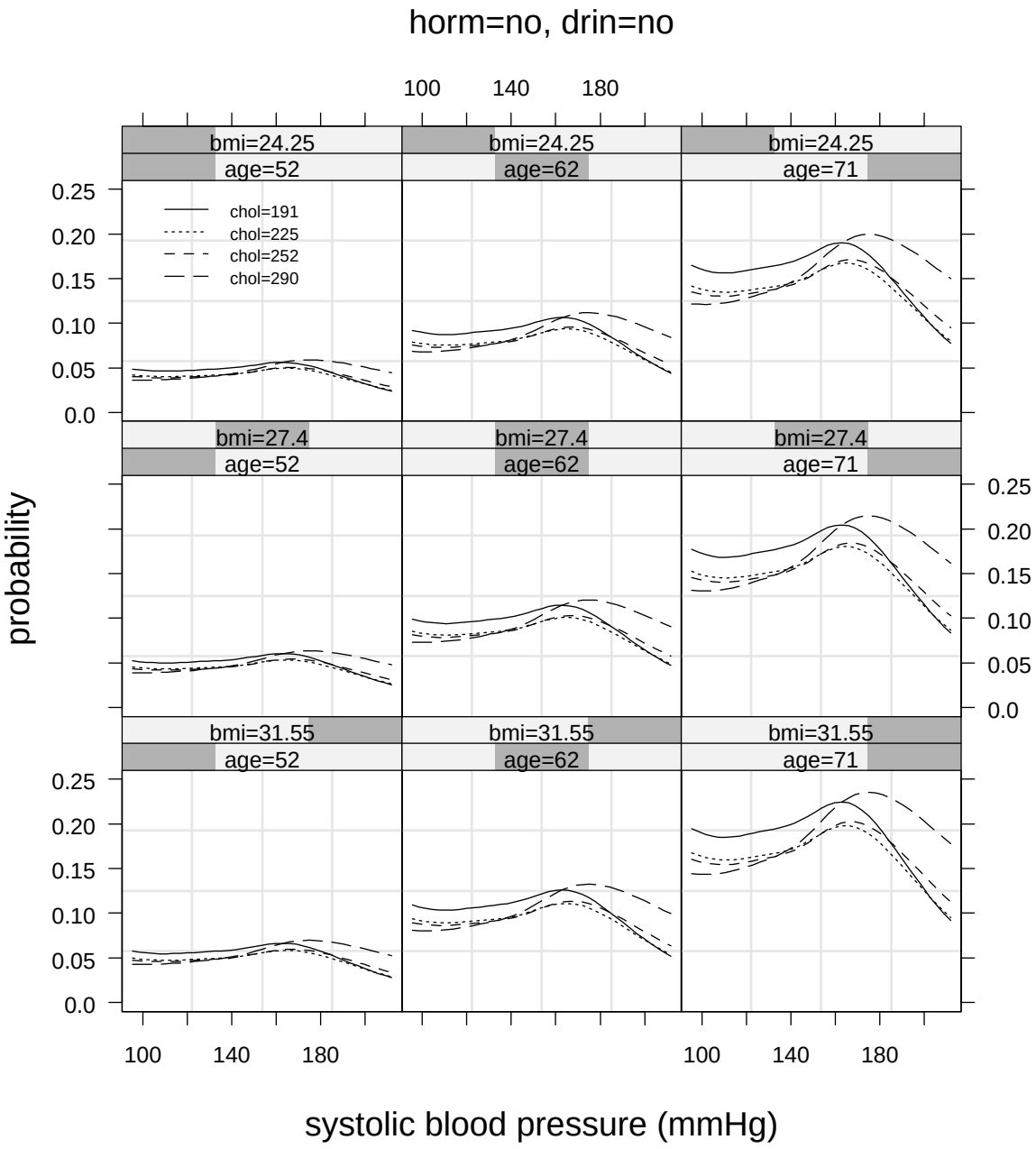

Figure 26: Estimated probability of at least one eye having the pigmentary abnormalities as a function of systolic blood pressure by three levels of `age` and `bmi`. `horm`=no, `drin`=no.

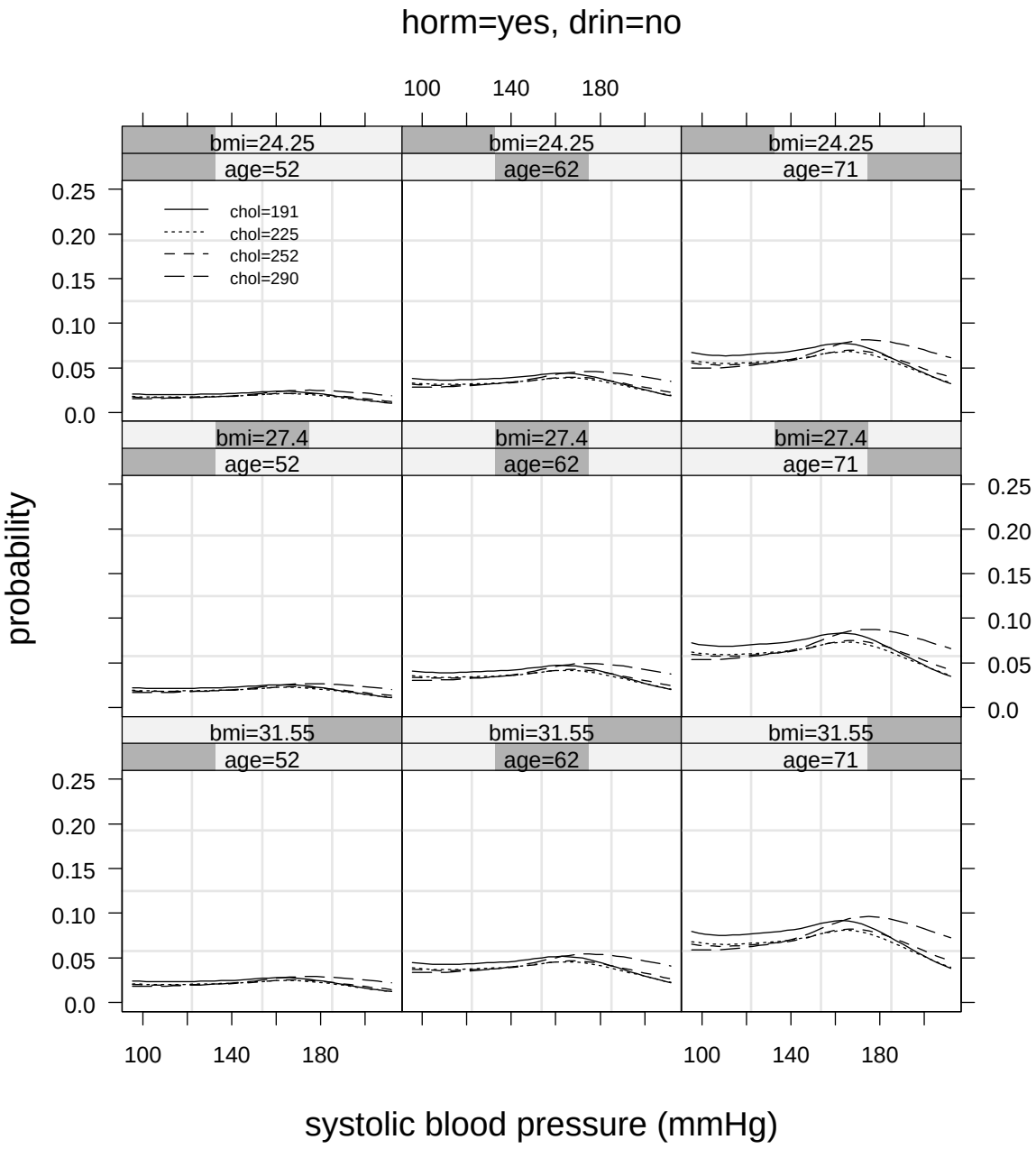

Figure 27: Estimated probability of at least one eye having the pigmentary abnormalities as a function of systolic blood pressure by three levels of `age` and `bmi`. `horm`=yes, `drin`=no.

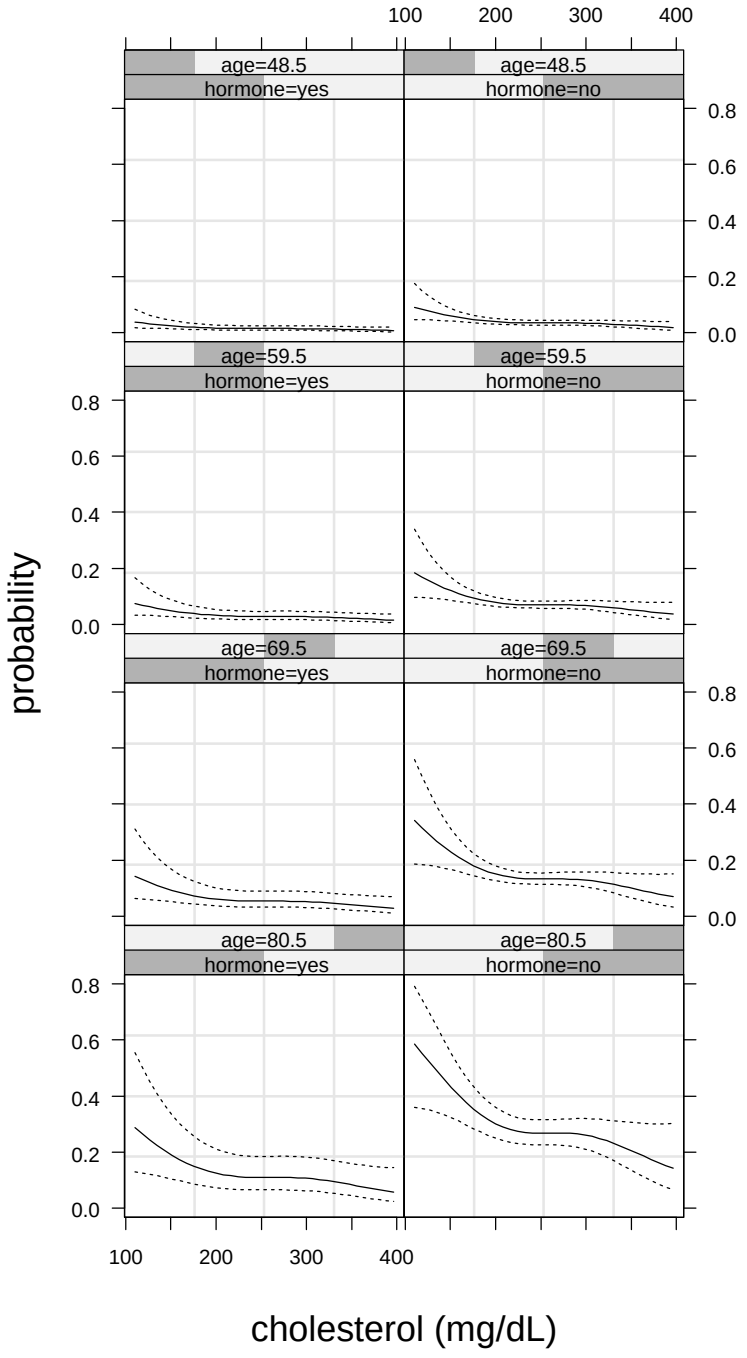

Figure 28: Bayesian confidence intervals for the probability of at least one eye having the pigmentary abnormalities. `bmi` and `sys` are fixed at their median. `drin`=no.

been diagnosed to have a certain disease for one eye. It provides a guideline as how to reduce the risk of the same disease for the other healthy eye.

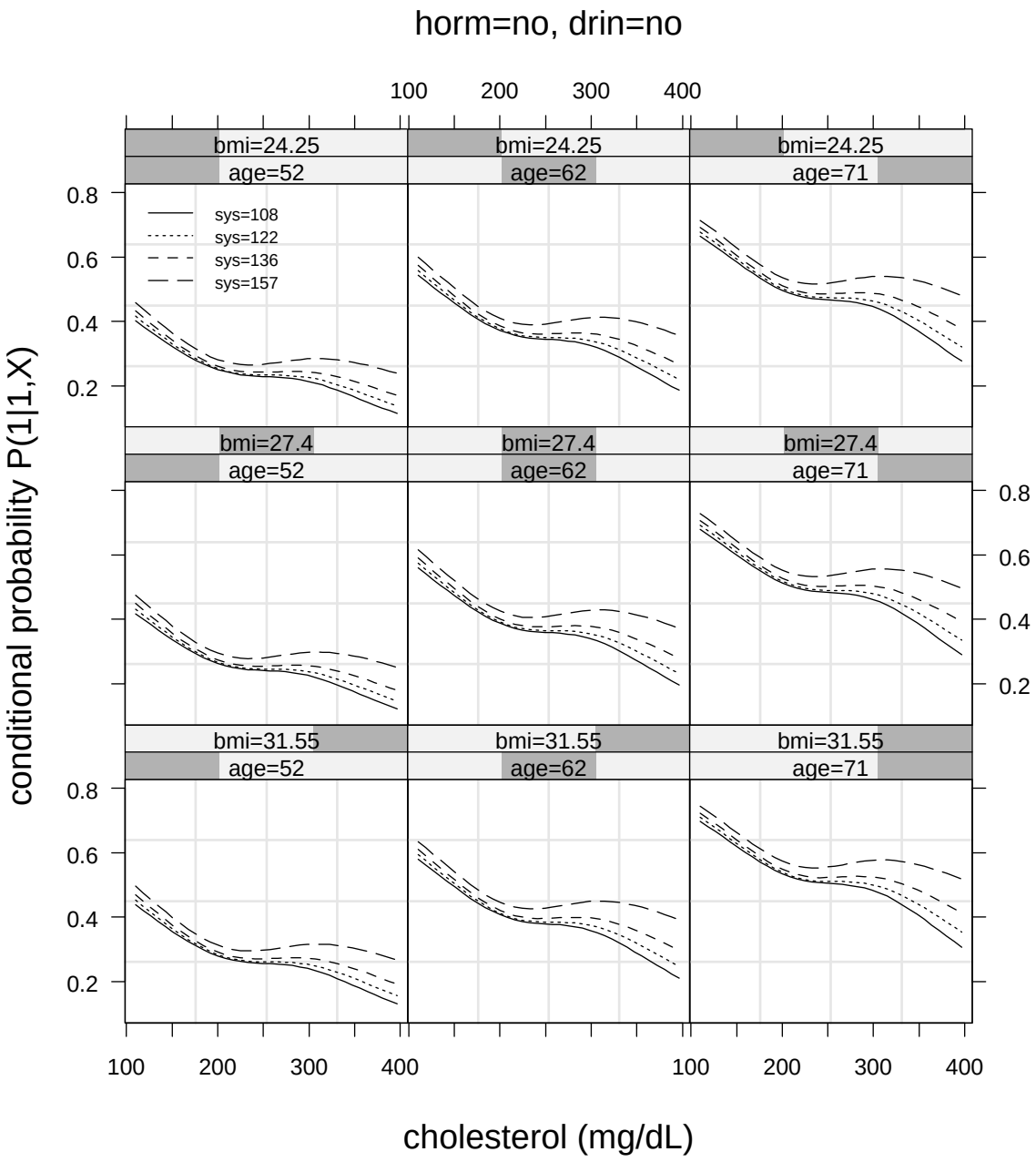

Figure 29: Estimated probability of one eye developing pigmentary abnormalities conditioning on the other eye already having this disease as a function of cholesterol by three levels of `age` and `bmi`. `horm`=no, `drin`=no.

Chapter 5

Summarizing Remarks

5.1 Conclusion

Penalized multivariate logistic regression using smoothing spline ANOVA model has been proposed to estimate the joint distribution for multivariate Bernoulli data, given the values of the predictor variables. The estimate is obtained by solving a variational problem involving the penalized likelihood.

Numerically, an approximate solution of the minimization problem is obtained by using the block one-step SOR-Newton-Ralphson algorithm. It has been proved in some special case, the approximate solution requires much less computing resources to achieve the same statistical convergence rate as the exact solution. Extensive Monte-Carlo experiments demonstrate that the performance of the approximate solution is very close to the exact one. Hence, we can deal with much larger data set by using the approximate solution instead of the exact one. $GACV$ for multivariate Bernoulli data has been derived. Its randomized version has been used to adaptively select smoothing parameters in every step of the block one-step SOR iteration. From the simulation studies, the iterated $ranGACV$ is an excellent computational proxy for the CKL distance.

The association terms are still kept as simple parametric forms in this model. They are estimated iteratively by maximum likelihood estimation in each block one-step SOR updating step.

By taking the dependence structure into consideration, we can obtain a partly flexible estimate of the joint probability, conditioning on the predictor variables. This approach is particular useful when the correct form of the function to be estimated is unknown. We successfully applied this method to analyze a medical data set. Some interesting features of this data set are brought to our attention by the nonparametric model, while more conventional parametric approach is unlikely to reveal such a relationship without more prior knowledge of the data set.

5.2 Log-linear vs. Marginal Model, and Future Research

The model we considered in this thesis is a conditional logistic regression model. The parameters f's and α's in our model have straightforward interpretations in terms of conditional probabilities. They are the canonical parameters in the log-linear model. Another class of model is the marginal model. The joint distribution is parameterized in terms of marginal means and odds ratio rather than conditional means and odds ratio.

The conditional model is very useful for prediction. In practice, for a vector of correlated outcomes, we may not observe all of them at the same time. However, we want to predict the outcomes of the unobserved variables conditioning on the predictor variables and observed outcomes. The conditional model addresses this problem more directly than the marginal model.

The computation of the marginal model is more difficult than the conditional model, since it involves re-parameterization of the canonical parameters. However, it also enjoys the reproducibility property, especially when the numbers of repeated measurements for each subject vary. Although it is argued that when the association factor is of interest, this will be most likely genuine multivariate data of equal cluster size, it will be interesting to build a marginal model by using a SS-ANOVA model. When the cluster sizes are unequal, like some longitudinal studies, the association factor can be viewed as a nuisance parameter. Data-driven method to select the smoothing parameters need to be developed.

Another interesting problem is to develop a semi-parametric model for time-to-event data using a smoothing spline model. We will assume a nonparametric form for baseline hazard function. The outcome variable could be correlated multivariate responses, for example, the time to developing a certain eye disease for each eye of the same person. Alternatively, there may exist correlated competing or semi-competing risks or informative censoring. Full penalized likelihood may be useful for model building. As always, a central question is how

to adaptively choose the amount of smoothing.

Bibliography

Aronszajn, N. (1950), 'Theory of reproducing kernels', *Trans. Am. Math. Soc.* **68**, 337–404.

Bishop, Y. M. M., Fienberg, S. E. & Holland, P. W. (1975), *Discrete Multivariate Analyses: Theory and Practice*, MITPress:MA.

Brumback, B. A. & Rice, J. (1998), 'Smoothing spline models for the analysis of nested and crossed samples of curves', *JASA* **93**, 961–975.

Carey, V., Zeger, S. L. & Diggle, P. (1993), 'Modelling multivariate binary data with alternating logistic regressions', *Biomtrka* **80**, 517–526.

Cessie, S. L. & Houwelingen, J. C. V. (1994), 'Logistic regression for correlated binary data', *Applied Statistics* **43**, 95–108.

Cox, D. D. & O'Sullivan, F. (1990), 'Asymptotic analysis of penalized likelihood and related estimators', *The Annals of Statistics* **18**, 1676–1695.

Cox, D. R. (1972), 'The analysis of multivariate binary data', *Applied Statistics* **21**, 113–120.

Craven, P. & Wahba, G. (1979), 'Smoothing noisy data with spline functions', *Numer. Math.* **31**, 377–403.

Efron, B. (1986), 'How biased is the apparent error rate of a prediction rule?', *JASA* **81**, 461–470.

Fitzmaurice, G. M. & Laird, N. M. (1993), 'A likelihood-based method for analysing longitudinal binary responses', *Biometrika* **80**, 141–151.

Friedman, J. H. (1991), 'Multivariate Adaptive Regression Splines (disc: P67-141)', *The Annals of Statistics* **19**, 1–67.

Girard, D. (1987), A fast 'Monte Carlo cross validation' procedure for large least squares problems with noisy data, Technical Report RR 687-M, IMAG, Grenoble, France.

Girard, D. (1991), 'Asymptotic optimality of the fast randomized versions of Gcv and C_L in ridge regression and regularization', *The Annals of Statistics* **19**, 1950–1963.

Girard, D. (1998), 'Asymptotic comparison of (partial) cross-validation, GCV and randomized GCV in nonparametric regression', *The Annals of Statistics* **26**, 315–334.

Glonek, G. F. V. & McCullagh, P. (1995), 'Multivariate logistic models', *JRSS-B* **57**, 533–546.

Gu, C. (1990), 'Adaptive spline smoothing in non-Gaussian regression models', *JASA* **85**, 801–807.

Gu, C. & Qiu, C. (1993), 'Smoothing spline density estimation: Theory', *The Annals of Statistics* **21**, 217–234.

Gu, C. & Wahba, G. (1993), 'Smoothing spline ANOVA with component-wise Bayesian "confidence intervals"', *Journal of Computational and Graphical Statistics* **2**, 97–117.

Gu, C. & Xiang, D. (1999), Cross validating non Gaussian data: GACV revisited, manuscript.

Hastie, T. J. & Tibshirani, R. J. (1990), *Generalized Additive Models*, Chapman and Hall: London.

Heagerty, P. J. & Zeger, S. L. (1996), 'Marginal regression models for clustered ordinal measurements', *JASA* **91**, 1024–1036.

Heagerty, P. J. & Zeger, S. L. (1998), 'Lorelogram: A regression approach to exploring dependence in longitudinal categorical responses', *JASA* **93**, 150–162.

Jones, B. & Kenward, M. G. (1989), *Design and Analysis of Cross-over Trials*, Chapman Hall:London.

Katz, J., Zeger, S. & Liang, K.-Y. (1994), 'Appropriate statistical methods to account for similarities in binary outcomes between fellow eyes', *Investigative Ophthalmology and Visual Science* **35**, 2461–2465.

Klein, R., Klein, B. E. & Jensen, S. (1997*a*), 'The relation of cardiovascular disease and its risk factors to the 5-year incidence of age-related maculopathy: The Beaver Dam Eye Study.', *Ophthalmology* **104**, 1804–1812.

Klein, R., Klein, B. E. & Linton, K. (1992), 'Prevalence of age-related maculopathy: The Beaver Dam Eye Study', *Ophthalmology* **99**, 933–943.

Klein, R., Klein, B. E. & Ritter, L. (1994), 'Is there evidence of an estrogen effect on age-related lens opacities? the Beaver Dam Eye Study.', *Arch. Ophthalmology* **112**, 85–91.

Klein, R., Klein, B. E., Jensen, S. & Meuer, S. (1997*b*), 'The five-year incidence progression of age-related maculopathy: The Beaver Dam Eye Study', *Ophthalmology* **104**, 7–21.

Li, K.-C. (1986), 'Asymptotic optimality of C_L and generalized cross-validation in ridge regression with application to spline smoothing', *The Annals of Statistics* **14**, 1101–1112.

Liang, K.-Y., Zeger, S. L. & Qaqish, B. (1992), 'Multivariate regression analyses for categorical data (disc: P24-40)', *JRSS-B* **54**, 3–24.

Lin, X. (1998*a*), Smoothing Spline ANOVA For Polychotomous Response Data, PhD thesis, Dept. of Statistics, University of Wisconsin-Madison, Madison, WI. Technical Report 1003.

Lin, X. & Zhang, D. (1999), 'Inference in generalized additive mixed model using smoothing splines', *Journal of the Royal Statistical Society, Serie B* **61**, 381–400.

Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R. & Klein, B. (1998), Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV, Technical Report 998, University of Wisconsin, Madison, WI.

Lin, Y. (1998*b*), Tensor product space ANOVA models, Technical Report 996, University of Wisconsin, Madison, WI.

Lipsitz, S. R., Laird, N. M. & Harrington, D. P. (1991), 'Generalized estimating equations for correlated binary data: Using the odds ratio as a measure of association', *Biomtrika* **78**, 153–160.

Luo, Z. & Wahba, G. (1997), 'Hybrid adaptive splines', *JASA* **92**, 107–116.

McCullagh, P. & Nelder, J. A. (1989), *Generalized Linear Models (Second Edition)*, Chapman Hall:London.

Molenberghs, G. & Ritter, L. L. (1996), 'Methods for analyzing multivariate binary data, with association between outcomes of interest', *Biometrics* **52**, 1121–1133.

Moss, S., Klein, R., Klein, B. E., Jensen, S. & Meuer, S. (1998), 'Alcohol

consumption and the 5-year incidence of age-related maculopathy: the Beaver Dam Eye Study.', *Ophthalmology* **105**, 789–794.

Nychka, D. (1988), 'Bayesian confidence intervals for smoothing splines', *JASA* **83**, 1134–1143.

Nychka, D. (1990), 'The average posterior variance of a smoothing spline and a consistent estimate of the average squared error', *The Annals of Statistics* **18**, 415–428.

Ortega, J. & Rheinboldt, W. (1970), *Iteration Solution of Nonlinear Equations in Several Variables*, Academic Press.

O'Sullivan, F. (1983), The analysis of some penalized likelihood estimation schemes, PhD thesis, Dept. of Statistics, University of Wisconsin, Madison, WI. Technical Report 726.

Press, W., Flannery, B., Teukolsky, S. & Vetterling, W. (1996), *Numerical Recipes in Fortran 90*, Cambridge Univ:UK.

Qu, Y., Williams, G. W., Beck, G. J. & Goormastic, M. (1987), 'A generalized model of logistic regression for clustered data', *CommStA* **16**, 3447–3476.

Ritter, L., Klein, R., Klein, B. E., Mares-Perlman, J. & Jensen, S. (1995), 'Alcohol use and age-related maculopathy in the Beaver Dam Eye Study.', *Am. J. Ophthalmology* **120**, 190–196.

Schechter, S. (1968), 'Relaxation methods for convex problems', *SIAM J. Numer. Anal.* **5**, 601–612.

Silverman, B. W. (1985), 'Some aspects of the spline smoothing approach to non-parametric regression curve fitting', *JRSS-B* **47**, 1–21.

Varga, R. S. (1984), *Matrix Iterative Analysis*, Prentice-Hall.

Wahba, G. (1980), Spline bases, regularization, and generalized cross validation for solving approximation problems with large quantities of noisy data, *in* W. Cheney, ed., 'Approximation Theory III', Academic Press, pp. 905–912.

Wahba, G. (1983), 'Bayesian "confidence intervals" for the cross-validated smoothing spline', *JRSS-B* **45**, 133–150.

Wahba, G. (1990), *Spline Models for Observational Data*, SIAM:PA.

Wahba, G., Wang, Y., Gu, C., Klein, R. & Klein, B. (1995), 'Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy', *The Annals of Statistics* **23**, 1865–1895.

Wang, Y. (1997), 'GRKPACK: Fitting smoothing spline analysis of variance models to data from exponential families', *Commun. Statist. Simulation and Computation* **26**, 765–782.

Wang, Y. (1998*a*), 'Mixed-effects smoothing spline ANOVA', *Journal of the Royal Statistical Society B* **60**, 159–174.

Wang, Y. (1998*b*), 'Smoothing spline models with correlated random errors', *JASA* **93**, 341–348.

Wang, Y. & Brown, M. B. (1996), 'A flexible model for human circadian rhythms', *Biometrics* **52**, 588–596.

Williamson, J. M., Kim, K. & Lipsitz, S. R. (1995), 'Analyzing bivariate ordinal data using a global odds ratio', *JASA* **90**, 1432–1437.

Wong, W. (1992), Estimation of the loss of an estimate, Technical Report 356, Dept. of Statistics, University of Chicago, Chicago, Il.

Xiang, D. (1996), Model fitting and testing for non-Gaussian data with a large data set, Technical Report 957, Ph.D. thesis, Department of Statistics, University of Wisconsin-Madison.

Xiang, D. & Wahba, G. (1996), 'A generalized approximate cross validation for smoothing splines with non-Gaussian data', *Statistica Sinica* **6**, 675–692.

Ye, J. (1998), 'On measuring and correcting the effects of data mining and model selection', *JASA* **93**, 120–131.

Zhao, L. & Prentice, R. L. (1990), 'Correlated binary regression using a quadratic exponential model', *Biometrika* **77**, 642–648.

Zhao, L. P., Prentice, R. L. & Self, S. G. (1992), 'Multivariate mean parameter estimation by using a partly exponential model', *JRSS-B* **54**, 805–811.