DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1159
April 27, 2010

Penalized Likelihood Regression in Reproducing Kernel Hilbert Spaces with Randomized Covariate Data

Xiwen Ma[1]
Department of Statistics
University of Wisconsin, Madison

[1]Research supported in part by NIH Grant EY09946, NSF Grant DMS-0604572, NSF Grant DMS-0906818 and ONR Grant N0014-09-1-0655.

PENALIZED LIKELIHOOD REGRESSION IN REPRODUCING KERNEL HILBERT SPACES WITH RANDOMIZED COVARIATE DATA

by

Xiwen Ma

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

(Statistics)

at the

UNIVERSITY OF WISCONSIN–MADISON

2010

ACKNOWLEDGMENTS

First and foremost, I would like to express my deepest gratitude to my advisor Professor Grace Wahba, for her guidance and support through the course of this dissertation. Grace is a brilliant statistician and an outstanding, caring mentor. She always have her door open to listen to my latest ideas and results. Her dedication to statistics, her passion and her greatest personality have been inspiring me very much. It has been an honor to have her as my preceptor.

I want to thank Professor Zhengjun Zhang for his encouragement and guidance in my research. He helped me through difficult times during my PhD study. I am also grateful to Professor Kam-Wah Tsui, Professor Sündüz Keles and Professor Sijian Wang for their service in my thesis committee. Professor Kam-Wah Tsui has been a mentor to me in various ways. I am grateful to his influence on me about passions in science. Professor Sündüz Keles and Professor Sijian Wang have given me many valuable feedback and ideas in the Thursday group. They also set up for me the best examples of young researchers.

I want to thank Bin Dai for his effort within our collaborated projects. The great and past fellow graduate students of the Thursday group also helped me in various ways: Pei-fen Kuan, Shilin Ding, Kevin Eng, Zhigeng Geng, Dongjun Chung, Xin Li, Héctor Corrada Bravo, Hyonho Chun and Weiliang Shi. These and other graduate students made my life in Madison an enjoyable one.

My most heartfelt gratitude must go to my parents Li and Juan, and my wife Xiaodan for their love and support.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Figure Page

ABSTRACT

Penalized likelihood regression consists of a category of commonly used regularization methods, including regression splines with RKHS penalty and the LASSO. When the observed data comes from a non-Gaussian exponential family distribution, a penalized log-likelihood is commonly used to estimate of the regression function. This technique allows a flexible form of the estimator and aims at an appropriate balance between the goodness-of-fit and the flexibility of the estimator. This thesis is composed of two major parts, both of which are within the framework of penalized likelihood regression.

The first part of the thesis presents a direct extension of penalized likelihood regression with RKHS penalty to the situation when the observed covariates are probability spaces. In order to estimate the regression function, we use a penalized likelihood that incorporates the covariate distribution information. We prove that the penalized likelihood estimate exists under a mild condition. In the computation, we propose a dimension reduction technique to minimize the penalized likelihood and derive a GACV (Generalized Approximate Cross Validation) to choose the smoothing parameter. A direct implementation of our methods is to handle incomplete data problems such as covariate measurement error and partially missing covariates.

The second part of the thesis concerns estimating the degrees of freedom for penalized likelihood regression including regression splines and the LASSO. For non-Gaussian data the degrees of freedom can be defined in the framework of Efron's optimism theory. We show that the degrees of freedom for penalized likelihood regression can be estimated by: (1) the trace of the influence matrix of the mean; and (2) the GACV. With these results on hand, various model selection criteria–AIC, BIC, GACV and BGACV– are available to select the regularization parameter. We also generalize our methods to treat the variable selection problem with multivariate Bernoulli observations–a more complicated penalized likelihood regression problem.

Chapter 1

General overview

1.1 Overview

In this thesis we are concerned with two important issues of penalized likelihood regression with non-Gaussian data.

Classical penalized likelihood regression problems deal with the case that the independent variables data are known exactly. In practice, however, it is common to observe data with incomplete covariate information. In this case estimating a regression function nonparametrically is extremely difficult. We are concerned with a fundamentally important case where some of the observations do not represent the exact covariate information, but only a probability distribution. By working with randomized covariate data, we aim to provide a general framework to handle incomplete covariate data. Our methods have been extended to treat the situations of covariate measurement error and partially missing covariates.

Degrees of freedom is commonly used to quantify the model complexity of a statistical fitting procedure. Estimating the degrees of freedom is fundamentally important in penalized likelihood regression as it plays an important role in model assessment and selection. In the framework of Efron's optimism theory, we propose a convenient estimation of the degrees of freedom and discuss the relationship with the generalized approximate cross validation (GACV). Our methods can be extended to

treat the variable selection problem with multivariate Bernoulli observations–a more complicated penalized likelihood regression problem.

1.2 Outline of the Thesis

The rest of the thesis is organized as follows. Chapter 2 discusses the penalized likelihood regression with randomized covariate data. We first briefly reviewed the classical penalized likelihood regression and the related incomplete data problems. Then we derived the penalized likelihood for randomized covariate data. A series of lemmas under weak lower-semicontinuity are derived to show the existence of penalized likelihood estimate. Numerically based on quadrature integration formulas for probability measures, a novel computational scheme is developed to obtain an approximate estimator in a finite dimensional subspace. The corresponding GACV function is derived to select the smoothing parameters. After that we extend our methods to handle other incomplete data problems including covariate measurement error and partially missing covariates. Finally we illustrate our methods by several simulation studies and a real data analysis.

Chapter 3 concerns estimating the degrees of freedom for penalized likelihood regression. We first review the definition of the degrees of freedom and multivariate Bernoulli distribution. Then we derive a convenient estimation of the degrees of freedom for penalized likelihood regression. The relationship with the GACV is discussed. It can be shown that the GACV provides an alternative estimation of the degrees of freedom. After that we consider the variable selection problem with multivariate Bernoulli observations. We first propose a truncated log-linear model with fewer parameters to be estimated. Then we derive a GACV based on an augmented response technique. Afterwards we estimate the degrees of freedom. Finally we illustrate our methods via several simulated examples.

Chapter 4 provides some concluding remarks. Appendix A includes all the technical proofs. Appendix B derives the GACV for penalized likelihood regression with randomized covariate data. Appendix C extends the methods of Chapter 2 to smoothing spline analysis of variance (SS-ANOVA) models.

Chapter 2

Penalized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate data

2.1 Introduction

2.1.1 Penalized likelihood regression in reproducing kernel Hilbert spaces

We are concerned with non or semi parametric regression for data from a non-Gaussian exponential family. Suppose that we have n independent observations $(y_i, x_i), i = 1, ..., n$, where each y_i denotes the response and each x_i denotes the covariate information. The goal is to fit a probability mechanism, assuming that the conditional distribution of y_i given x_i has a density in the exponential family with the form

$$p(y_i|x_i, f) = \exp\{(y_i \cdot f(x_i) - b(f(x_i)))/a(\phi) + c(y_i, \phi)\}, \tag{2.1}$$

where $b(\cdot)$ and $c(\cdot)$ are given functions with $b(\cdot)$ strictly convex, ϕ is the scale parameter and f is the regression function to be estimated. We assume throughout this chapter that ϕ is known, as, for example, Binomial data and Poisson data. In this case, (2.1) can be simplified by

$$p(y_i|x_i, f) = \exp\{y_i \cdot f(x_i) - b(f(x_i)) + c(y_i)\}. \tag{2.2}$$

Note that the methods of this chapter can also be extended to the situation when ϕ is unknown, but may be more computationally complicated.

The regression function f will be estimated non or semi parametrically in some reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ by minimizing the penalized likelihood

$$I_\lambda(f) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, f) + \frac{\lambda}{2} J(f), \tag{2.3}$$

where the penalty $J(\cdot)$ is a norm or semi-norm in $\mathcal{H}$ with finite dimensional null space $\mathcal{H}_0 = \{f \in \mathcal{H} \mid J(f) = 0\}$ and λ is the smoothing parameter which balances the tradeoff between model fitting and smoothness. In this case if the null space $\mathcal{H}_0$ satisfies some condition, saying that $I_\lambda(f)$ has a unique minimizer in $\mathcal{H}_0$, then the minimizer of $I_\lambda(f)$ in $\mathcal{H}$ exists in a known n-dimensional subspace spanned by $\mathcal{H}_0$ and functions of the reproducing kernel. See, for example, Kimeldorf and Wahba (1971)[28], O'Sullivan, Yandell and Raynor (1983)[36], Wahba (1990)[39] and Xiang and Wahba (1996)[44]. This model building technique, known as penalized likelihood regression with RKHS penalty, allows for more flexibility than parametric regression models. We will not review the general literature, other than to note two books and references therein. Wahba (1990)[39] offers a general introduction of spline models. Gu (2002)[20] comprehensively reviews the smoothing spline analysis of variance (SS-ANOVA), an important implementation of penalized likelihood regression in multivariate function estimation.

2.1.2 Randomized covariate data and related problems

In this chapter the issue we are concerned about is the situation where components of x_i are not observable but only known to have come from a particular probability distribution. This concept of randomized covariate, without the requirement of any actual measure of x_i, is more flexible than the common sense of covariate measurement error. In this case a natural likelihood-based approach is to treat x_i's

as latent variables and minimize a randomized version of penalized likelihood that integrates x_i's out of the likelihood. This approach, however, typically leads to a non-convex and infinite dimensional optimization problem in RKHS. Therefore we shall first prove that the randomized penalized likelihood is minimizable. This is the subject of Section 2.2. Afterwards, two computational issues will be addressed in Section 2.3: (1) how to numerically compute an estimator; and (2) how to select the smoothing parameter.

Randomized covariate data can be treated as a basic version of incomplete data. Our methods can be extended to other incomplete data problems. For example, in the survey or medical research, it is common to obtain data where the covariates are measured with error. More specifically, x_i is not directly observed but instead $x_i^{err} = x_i + u_i$ is observed, where $u_i, i = 1, ..., n$ are iid random errors. Fan and Truong (1993)[15] regarded this measurement error problem in the context of nonparametric regression, using the methods based on kernel deconvolution. Their technique was later studied and extended by, for example, Ioannides and Alevizo (1997)[24], Schennach (2004)[34], Carroll, Ruppert, Stefanski and Crainiceanu (2006)[7] and Delaigle, Fan and Carroll (2009)[12]. Carroll, Maca and Ruppert (1999)[6] suggested to use the SIMEX method (Cook and Stefanski, 1994[10]) to build nonparametric regression models including both kernel regression and penalized likelihood regression. Berry, Carroll and Ruppert (2001)[3] described Bayesian approaches for smoothing splines and regression P-splines. More recently, Cardot, Crambes, Kneip and Sarda (2007)[5] used the total least square method (Van Huffel and Vandewalle, 1991[37]) to compute a smoothing spline estimator from noisy covariates. As a sequel to these works, in this chapter we treat measurement error as a special case of randomized covariates since each x_i can be viewed as a random

variable (vector) distributed as $x_i^{err} - u_i$. Therefore the methodology of randomized penalized likelihood estimate can be employed.

We will as well be able to make another modest extension to treat the important situation where some components of some x_i's are completely missing. In this case we may write $x_i = (x_i^{obs}, x_i^{mis})$, where x_i^{obs} and x_i^{mis} denote the observed and the missing components. It is well-known (Little and Rubin, 2002[32]) that a complete case analysis that deletes the cases with missing information often leads to bias or inefficient estimates. Various methods for missing covariate data have been developed in the context of parametric regression models, but to date few methods have been proposed for nonparametric penalized likelihood regression in RKHS. For parametric regression, one popular approach is the method of weights initially proposed by Ibrahim (1990)[21]. His suggestion is to assume the x_i's to be independent observations from a marginal distribution depending on some parameters and to maximize the joint distribution of (y_i, x_i) by the expectation-maximization (EM) algorithm. Discussions and extensions of this method appear in Ibrahim, Lipsitz and Chen (1999)[22], Horton and Laird (1999)[25], Huang, Chen and Ibrahim (2005)[27], Ibrahim, Chen, Lipsitz and Herring (2005)[23], Horton and Kleinman (2007)[26], Chen and Ibrahim (2006)[8], Chen, Zeng and Ibrahim (2007)[9] and elsewhere. Ibrahim's method can also be employed to build penalized likelihood regression models in RKHS. In this chapter we will treat the missing components x_i^{mis} as a random vector depending on both the observed components x_i^{obs} and the covariate marginal distribution. Then the methodology of randomized covariate data can be extended to handle missing covariate data.

2.1.3 Outline of the chapter

The rest of the chapter is organized as follows. In Section 2.2 we prove the existence of the randomized covariate penalized likelihood estimation in the general smoothing spline set-up. Computational techniques are presented in Section 2.3. Sections 2.4 and 2.5 extend our methods to the problem of covariate measurement error. Sections 2.6 and 2.7 describe penalized likelihood regression with missing covariate data. Section 2.8 provides some numerical results. We conclude the chapter in Section 2.9.

2.2 Randomized covariate penalized likelihood estimation (theory)

Consider the general smoothing spline set-up, where x is allowed to be from some arbitrary index set $\mathscr{T}$ on which an RKHS can be defined. Randomized covariate data is defined in the way that we "observe" for each subject i a *probability space* $(\mathcal{X}_i, \mathcal{F}_i, P_i)$, rather than a realization of x_i, where $\mathcal{X}_i \subseteq \mathscr{T}$ denotes the domain of x_i, $\mathcal{F}_i$ is a $\sigma-$algebra and P_i is a probability measure over $(\mathcal{X}_i, \mathcal{F}_i)$.

In this case each x_i can be treated as a latent random variable. Thus, given a regression function f, the distribution of $[y_i|f]$ has a density

$$p(y_i|f) = \int_{\mathcal{X}_i} p(y_i|x_i, f) dP_i. \tag{2.4}$$

Note that, throughout this chapter, we use the labels $[A|B]$ and $p(A|B)$ to denote the conditional distribution of A given B and the density function for this distribution. According to (2.4), the penalized likelihood estimate of f is the minimizer of

$$I_\lambda^R(f) = -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{X}_i} p(y_i|x_i, f) dP_i + \frac{\lambda}{2} J(f), \tag{2.5}$$

where R denotes the "randomness" of the covariates and f is restricted on the Borel measurable subset

$$\mathcal{H}_B = \{f \in \mathcal{H} \ : \ f \text{ is Borel measurable on } (\mathcal{X}_i, \mathcal{F}_i), i = 1, ..., n\} \tag{2.6}$$

in which the Lebesgue integrals in (2.5) can be defined. It can be shown that $\mathcal{H}_B$ is a subspace of $\mathcal{H}$.

PROPOSITION 2.1. *$\mathcal{H}_B$ is a subspace of $\mathcal{H}$.*

Proof See Appendix A. □

This methodology can be referred to as **randomized covariate penalized likelihood estimation** or **RC-PLE**. Note that RC-PLE includes the classical penalized likelihood regression where x_i's are observed exactly. Actually, $I_\lambda^R(f)$ equals $I_\lambda(f)$ if every $(\mathcal{X}_i, \mathcal{F}_i, P_i)$ stands for a single point probability.

However, computation of RC-PLE is extremely difficult. Firstly, since each $p(y_i|x_i, f)$ is log-concave as a function of f, $I_\lambda^R(f)$ is in general not convex due to the presence of the integrals. Secondly, if at least one $(\mathcal{X}_i, \mathcal{F}_i, P_i)$ has infinite support, then there is no finite dimensional subspace in which f_λ is known a *priori* to lie, as can be concluded from the arguments in Kimeldorf and Wahba (1971)[28]. Therefore, we shall first prove that $I_\lambda^R(f)$ is minimizable and hence the phrase "penalized likelihood estimate" is meaningful. Computational techniques will be described in Section 2.3.

Recall that for the classical penalized likelihood regression, the unique solution in the null space is sufficient to ensure the existence of the penalized likelihood estimate. In the case of randomized covariate data, we extend this condition as follows:

ASSUMPTION A.1 (Null space condition). There exist exactly observed subjects $(y_{k_1}, x_{k_1}), (y_{k_2}, x_{k_2}), ..., (y_{k_s}, x_{k_s})$ such that $\sum_{i=1}^{s} \log p(y_{k_i}|x_{k_i}, f)$ has a unique maximizer in $\mathcal{H}_0$.

Now we state our main theorem.

THEOREM 2.2. *Under A.1,* $\exists f_\lambda \in \mathcal{H}_B$ *such that* $I_\lambda^R(f_\lambda) = \inf_{f \in \mathcal{H}_B} I_\lambda^R(f)$.

Theorem 2.2 guarantees the existence of the RC-PLE estimate, which justifies the title of the chapter. In particular, if the null space of the penalty functional $J(\cdot)$ contains only constants, then A.1 can be ignored. In this case, the penalized likelihood estimate always exists. Our proof of the theorem is based on the lower-semicontinuity in the weak topology. We first recall some definitions.

DEFINITION 1. A sequence $\{f_k\}_{k \in \mathbb{N}}$ in a Hilbert space $\mathcal{H}$ is said to **converge weakly** to f if $\langle f_k, g \rangle \to \langle f, g \rangle$ for all $g \in \mathcal{H}$. Here $\langle \cdot, \cdot \rangle$ denotes the inner product of $\mathcal{H}$.

DEFINITION 2. Let $\mathcal{H}$ be a Hilbert space, a functional $\gamma : \mathcal{H} \to \mathbb{R}$ is **(weakly) sequentially lower semicontinuous** at $f \in \mathcal{H}$ if $\gamma(f) \leq \liminf \gamma(f_k)$ for any sequence $\{f_k\}_{k \in \mathbb{N}}$ that (weakly) converges to f.

DEFINITION 3. Let $\mathcal{H}$ be a Hilbert space, a functional $\gamma : \mathcal{H} \to \mathbb{R}$ is **positively coercive** if $||f||_{\mathcal{H}} \to +\infty$ implies $\gamma(f) = +\infty$. Here $|| \cdot ||_{\mathcal{H}}$ denotes the norm of $\mathcal{H}$.

Theorem 2.2 can be shown by combining Proposition 2.3 and Lemmas 2.4-2.6 below. Note that Proposition 2.3 is obtained from Theorem 7.3.7 in Kurdila and Zabarankin (2005)[30], Page 217. The proofs of lemmas are given in Appendix A.

PROPOSITION 2.3. *Let $\mathcal{H}$ be a Hilbert space. Suppose that $\gamma : \mathcal{M} \subseteq \mathcal{H} \to \mathbb{R}$ is positively coercive and weakly sequentially lower semicontinuous over the closed and convex set $\mathcal{M}$, then $\exists f_0 \in \mathcal{M}$ such that $\gamma(f_0) = \inf_{f\in\mathcal{M}} \gamma(f)$.*

LEMMA 2.4. *Under A.1, the penalized likelihood $I_\lambda^R(f)$ is positively coercive over $\mathcal{H}_B$.*

LEMMA 2.5. *The functional $\log \int_{\mathcal{X}_i} p(y_i|x_i, f)dP_i : \mathcal{H}_B \to \mathbb{R}$ is weakly sequentially continuous.*

LEMMA 2.6. *The penalty functional $J(\cdot)$ is weakly sequentially lower semicontinuous.*

Proof of Theorem 2.2. Consider the functional $I_\lambda^R : \mathcal{H}_B \subseteq \mathcal{H} \to \mathbb{R}$. Theorem 2.2 follows from Proposition 2.2, Lemma 2.4-2.6 and Proposition 2.3 $\square$.

2.3 Randomized covariate penalized likelihood estimation (computation)

In the preceding section, we theoretically extended penalized likelihood regression in RKHS to randomized covariate data, where f was restricted on the Borel measurable subspace $\mathcal{H}_B$. In practical applications, however, we often face the case that all functions in the RKHS are Borel measurable. In this case, we no longer need

the restriction mentioned in (2.6). Thus, we would like to proceed our discussion under the following condition:

ASSUMPTION A.2. Consider the Borel-σ field of $\mathcal{H}$ (generated by the open sets). Mapping:

$$\begin{aligned} \mathscr{T} &\rightarrow \mathcal{H} \\ x &\mapsto K_x(\cdot) = K(\cdot, x) \end{aligned}$$

is Borel measurable for all $(\mathcal{X}_i, \mathcal{F}_i)$, $i = 1, ...n$. Here $K(\cdot, \cdot)$ denotes the reproducing kernel of $\mathcal{H}$.

Under A.2, by Theorem 90 of Berlinet and Thomas-Agnan (2004)[2], Page 195, every function in $\mathcal{H}$ is Borel measurable. It can be verified that if the domain $\mathscr{T} \subseteq \mathbb{R}^d$ and every $\mathcal{F}_i$ is a Borel σ-field, then A.2 is satisfied with

- Every continuous kernel;
- Kernels built from tensor sums or products of continuous kernels;
- Any radial basis kernel $K(x, z) = r(||x - z||_d)$ such that $r(\cdot)$ is continuous at 0. Here $|| \cdot ||_d$ denotes the usual Euclidian norm.

2.3.1 Quadrature penalized likelihood estimates

As previously discussed, there is in general no finite dimensional subspace in which the RC-PLE estimate f_λ is known *a priori* to lie, so direct computation is not attractive. In this case we shall find a finite dimensional approximating subspace and compute an estimator in this space. We consider the following penalized likelihood:

$$I_\lambda^{Z,\Pi}(f) = -\frac{1}{n}\sum_{i=1}^{n} \log \sum_{j=1}^{m_i} \pi_{ij} p(y_i|z_{ij}, f) + \frac{\lambda}{2} J(f) \tag{2.7}$$

where $Z = \{z_{11}, ..., z_{1m_1}, z_{21}, ..., z_{nm_n}\}$ with $z_{ij} \in \mathscr{T}$ and $\Pi = \{\pi_{11}, ..., \pi_{1m_1}, \pi_{21}, ..., \pi_{nm_n}\}$ with $\pi_{ij} > 0$. In words, when we evaluate the integrals on the right hand side of (2.5), each $(\mathcal{X}_i, \mathcal{F}_i, P_i)$ is replaced by a discrete probability distribution defined over $\{z_{i1}, z_{i2}, ..., z_{im_i}\}$ with probability mass function $P(x_i = z_{ij}) = \pi_{ij},\ j = 1, ..., m_i$. Thus $z_{ij}, 1 \leq j \leq m_i$ and $\pi_{ij}, 1 \leq j \leq m_i$ are referred to as **nodes** and **weights** of a **quadrature rule** for probability measure P_i.

In (2.7), f is only evaluated on a finite number of quadrature nodes. Under A.1, it can be seen from Theorem 2.2 and the arguments in Kimeldorf and Wahba (1971)[28] that the minimizer of $I_\lambda^{Z,\Pi}(f)$ in $\mathcal{H}$ is in a finite dimensional subspace $\mathcal{H}_Z$ spanned by $\mathcal{H}_0$ and $\{K(\cdot, z_{ij}) : z_{ij} \in Z\}$. Thus, $I_\lambda^{Z,\Pi}(f)$ can be formulated as a parametric penalized likelihood. Green (1990)[19] gave a general discussion on the use of the EM algorithm for parametric penalized likelihood estimation with incomplete data. His method can be extended to minimize $I_\lambda^{Z,\Pi}(f)$. It can be shown that the E-step at iteration $t+1$ has the form of

$$Q(f|f^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^{(t)} \cdot \log p(y_i|z_{ij}, f) - \frac{\lambda}{2}J(f), \tag{2.8}$$

where $f^{(t)}$ is estimated at iteration t and the weight

$$w_{ij}^{(t)} = \frac{\pi_{ij}p(y_i|z_{ij}, f^{(t)})}{\sum_k \pi_{ik}p(y_i|z_{ik}, f^{(t)})} \tag{2.9}$$

indicates the conditional probability of $[z_{ij}|y_i, f^{(t)}]$. The M-step updates f by maximizing $Q(f|f^{(t)})$ in $\mathcal{H}$. This is straightforward because $-Q(f|f^{(t)})$ is seen to be a weighted complete data penalized likelihood.

When the EM algorithm converges, we will obtain an estimator $\hat{f}_\lambda$ which approximates the RC-PLE estimate f_λ. Note that $\hat{f}_\lambda$ can be interpreted as the minimizer of $I_\lambda^R(f)$ when the integrals are approximated by quadrature rules. Hence, this computational technique is referred to as **quadrature penalized likelihood estimation** or

QPLE. The motivation behind this approach is that an efficient quadrature rule often requires only a few nodes for a good approximation to the integral. This convenient property eases the computation burden at each M-step.

2.3.2 Construction of quadrature rules

Construction of quadrature rules is a practical issue. In order to derive more applicable results, we further assume that each $x_i = (x_{i1}, ..., x_{id})^T$ is a random vector, i.e., $\mathscr{T} \subset \mathbb{R}^d$.

2.3.2.1 Univariate quadrature rules

Suppose that x_i is univariate (i.e., $d = 1$). In this case if x_i is a categorical random variable or exactly observed, then $(\mathcal{X}_i, P_i)$ itself can be used as a quadrature rule. Otherwise, if x_i is a continuous random variable, we will construct a **Gaussian quadrature rule**. Development of computational methods and routines of Gaussian quadrature integration formulae for probability measures is a mathematical research topic. We will not survey the general literature here, other than to say that the methods considered in this chapter can be obtained from, Golub and Welsch (1969)[18], Fernandes and Atchley (2006)[16], Bosserhoff (2008)[4] and Rahman (2009)[33]. Though a k-node Gaussian quadrature rule typically requires the first $2k$ moments of the measure P_i to be finite, this convention can be satisfied by most popular probability distributions including normal, uniform, exponential, gamma, beta and others. Besides Gaussian quadrature rules, if x_i has a density with respect to the Lebesgue measure, we also consider a quadrature rule with equally-spaced points. More specifically, suppose that x_i ranges over $[a, b]$, then we take equally-spaced points in $[a, b]$ as quadrature nodes while the quadrature weights are proportional to the density evaluated at the chosen nodes. Note that if $a = -\infty$ (or $b = +\infty$), we set $a = \mu_i - 3\sigma_i$

(or $b = \mu_i + 3\sigma_i$) where μ_i and σ_i denote the first and second moments of P_i. We refer to this simple quadrature rule as the **grid quadrature rule**.

2.3.2.2 Multivariate quadrature rules

Suppose that $x_i = (x_{i1}, ..., x_{id})^T$ is a multivariate random vector (i.e., $d > 1$). In this case a quadrature rule can be generated recursively with one-dimensional conditional quadrature rules. The algorithm is summarized as follows:

1. Set $s = 1$. Compute the marginal distribution of x_{i1} and generate a quadrature rule for x_{i1} by using the method for univariate random variables.

2. Let $\{z_1^{(s)}, ..., z_{m_s}^{(s)}\}$ and $\{\pi_1^{(s)}, ..., \pi_{m_s}^{(s)}\}$ be the quadrature rule generated for the marginal distribution of $(x_{i1}, ..., x_{is})^T$. For each $z_j^{(s)}, 1 \leq j \leq m_s$, compute the one-dimensional conditional distribution of $[x_{i(s+1)}|(x_{i1}, ..., x_{is})^T = z_j^{(s)}]$. Then generate a quadrature rule for this distribution, denoted by $\{z_{j1}^*, ..., z_{jn_j}^*\}$ and $\{\pi_{j1}^*, ..., \pi_{jn_j}^*\}$. Then $\{((z_j^{(s)})^T, z_{jr}^*)^T, 1 \leq r \leq n_j, 1 \leq j \leq m_s\}$ and $\{\pi_j^{(s)} \cdot \pi_{jr}^*, 1 \leq r \leq n_j, 1 \leq j \leq m_s\}$ compose a quadrature rule for the marginal distribution of $(x_{i1}, ..., x_{is}, x_{i(s+1)})^T$.

3. Set $s = s + 1$. Repeat step 2 until $s = d$.

The order that x_{ij}'s jump into the algorithm is not important. One may rearrange the order to simplify the computation of the quadrature rules. From our experience, a quadrature rule with 7 to 12 nodes for each component of x_i usually yields a very good approximation. In this case, the above EM algorithm usually converges very rapidly.

2.3.3 Choice of the smoothing parameter

2.3.3.1 The comparative KL distance and leaving-out-one-subject CV

So far the smoothing parameter λ, is assumed to be fixed. Choice of λ is a key problem in the penalized likelihood regression. For non-Gaussian data, Kullback-Leibler (KL) distance is commonly used as the risk function for the estimator f_λ

$$\mathrm{KL}(f^*, f_\lambda) = \frac{1}{n}\sum_{i=1}^{n} E_{y_i^0|f^*}\left\{\log\frac{p(y_i^0|f^*)}{p(y_i^0|f_\lambda)}\right\}, \tag{2.10}$$

where f^* denotes the true regression function and the expectation is taken over $y_i^0 \sim p(y|f^*)$ independent of y_i. In order to estimate $\mathrm{KL}(f^*, f_\lambda)$, Xiang and Wahba (1996)[44] proposed **generalized approximate cross validation** (GACV) beginning with a leaving-out-one argument to choose the smoothing parameter, which works well for Bernoulli data. Lin, Wahba, Xiang, Gao, Klein and Klein (2000)[31] derived a randomized version of GACV (ranGACV) which is more computationally friendly for large data sets. In this section we obtain a convenient form of leaving-out-one-subject CV for randomized covariate data and extend GACV and randomized GACV to randomized covariate data in subsequent sections.

In the situation when each observed covariate is actually a probability space $(\mathcal{X}_i, \mathcal{F}_i, P_i)$, $[y_i^0|f]$ has a density of

$$p(y_i^0|f) = \int_{\mathcal{X}_i} p(y_i^0|x_i, f)dP_i. \tag{2.11}$$

Following (2.10) and leaving out the quantities which do not depend on λ, the comparative KL (CKL) distance can be written as

$$\mathrm{CKL}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n} E_{y_i^0|f^*}\left\{\log\int_{\mathcal{X}_i}\exp\left\{y_i^0 f_\lambda(x_i) - b(f_\lambda(x_i))\right\}dP_i\right\}. \tag{2.12}$$

To simplify the notation, let's denote

$$L(y, f, P_i) = \log\int_{\mathcal{X}_i}\exp\left\{yf(x_i) - b(f(x_i))\right\}dP_i \tag{2.13}$$

the log-likelihood function for randomized covariate data. Using first order Taylor expansion to expand L at the point y_i, we have that

$$L(y_i^0, f_\lambda, P_i) \approx L(y_i, f_\lambda, P_i) + (y_i^0 - y_i)\frac{\partial L}{\partial y}(y_i, f_\lambda, P_i). \tag{2.14}$$

Direct calculation yields

$$\begin{aligned}\frac{\partial L}{\partial y}(y_i, f_\lambda, P_i) &= \frac{\int_{\mathcal{X}_i} f_\lambda(x_i)\exp\{y_i f_\lambda(x_i) - b(f_\lambda(x_i))\}\, dP_i}{\int_{\mathcal{X}_i}\exp\{y_i f_\lambda(x_i) - b(f_\lambda(x_i))\}\, dP_i} \\ &= E_{x_i|y_i,f_\lambda} f_\lambda(x_i).\end{aligned} \tag{2.15}$$

Plugging (2.14) and (2.15) into (2.12), we have that

$$\begin{aligned}\text{CKL}(\lambda) &\approx \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^n E_{y_i^0|f^*}(y_i - y_i^0)E_{x_i|y_i,f_\lambda} f_\lambda(x_i) \\ &= \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^n (y_i - \mu_i^*)E_{x_i|y_i,f_\lambda} f_\lambda(x_i),\end{aligned} \tag{2.16}$$

where $\mu_i^* = E_{y_i^0|f^*} y_i^0$ is the true mean response and

$$\text{OBS}(\lambda) = -\frac{1}{n}\sum_{i=1}^n \log\int_{\mathcal{X}_i}\exp\{y_i f_\lambda(x_i) - b(f_\lambda(x_i))\}\, dP_i \tag{2.17}$$

is the observed log-likelihood. Denote $f_\lambda^{[-i]}$ the leaving-out-one estimator, i.e., the minimizer of $I_\lambda^R(f)$ with the ith subject omitted. Since $E_{x_i|y_i,f_\lambda} f_\lambda(x_i)$ is the posterior mean estimate of $f^*(x_i)$, following Xiang and Wahba (1996)[44], we may replace $\mu_i^* E_{x_i|y_i,f_\lambda} f_\lambda(x_i)$ by $y_i E_{x_i|y_i,f_\lambda^{[-i]}} f_\lambda^{[-i]}(x_i)$ and define the leaving-out-one-subject cross validation (CV) by

$$\text{CV}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^n y_i(E_{x_i|y_i,f_\lambda} f_\lambda(x_i) - E_{x_i|y_i,f_\lambda^{[-i]}} f_\lambda^{[-i]}(x_i)). \tag{2.18}$$

It can be seen that (2.16) and (2.18) generalize the complete data CKL and CV formulas proposed in Xiang and Wahba (1996)[44]. If a QPLE estimate $\hat{f}_\lambda$ is computed,

we may further approximate (2.18) by quadrature rules. More specifically, $\mathrm{OBS}(\lambda)$ can be evaluated by

$$\widehat{\mathrm{OBS}}(\lambda) = -\frac{1}{n}\sum_{i=1}^{n}\log\sum_{j=1}^{m_i}\pi_{ij}\exp\left\{y_i\hat{f}_\lambda(z_{ij}) - b(\hat{f}_\lambda(z_{ij}))\right\} \tag{2.19}$$

where z_{ij}'s and π_{ij}'s represent nodes and weights of the quadrature rules given in the preceding section. Define the weight functions

$$w_{ij}(\tau) = \frac{\pi_{ij}\exp\{y_i\tau_j - b(\tau_j)\}}{\sum_k \pi_{ik}\exp\{y_i\tau_k - b(\tau_k)\}},\ j = 1, ..., m_i \tag{2.20}$$

where $\tau = (\tau_1, ..., \tau_{m_i})^T$ is an arbitrary vector of length m_i. Let us use the notations

$$\vec{f}_{\lambda i} = (\hat{f}_\lambda(z_{i1}), ..., \hat{f}_\lambda(z_{im_i}))^T \tag{2.21}$$

$$\vec{f}_{\lambda i}^{[-i]} = (\hat{f}_\lambda^{[-i]}(z_{i1}), ..., \hat{f}_\lambda^{[-i]}(z_{im_i}))^T. \tag{2.22}$$

Then (2.15) yields

$$E_{x_i|y_i,\hat{f}_\lambda}\hat{f}_\lambda(x_i) \approx \sum_{j=1}^{m_i} w_{ij}(\vec{f}_{\lambda i})\hat{f}_\lambda(z_{ij}) = \sum_{j=1}^{m_i} w_{\lambda,ij}\hat{f}_\lambda(z_{ij}) \tag{2.23}$$

$$E_{x_i|y_i,\hat{f}_\lambda^{[-i]}}\hat{f}_\lambda^{[-i]}(x_i) \approx \sum_{j=1}^{m_i} w_{ij}(\vec{f}_{\lambda i}^{[-i]})\hat{f}_\lambda^{[-i]}(z_{ij}) = \sum_{j=1}^{m_i} w_{\lambda,ij}^{[-i]}\hat{f}_\lambda^{[-i]}(z_{ij}) \tag{2.24}$$

where $w_{\lambda,ij} = w_{ij}(\vec{f}_{\lambda i})$ and $w_{\lambda,ij}^{[-i]} = w_{ij}(\vec{f}_{\lambda i}^{[-i]})$ equal the weights at the final iteration of the EM algorithm, respectively, when $\hat{f}_\lambda$ and $\hat{f}_\lambda^{[-i]}$ were computed. Therefore a more convenient version of CV can be obtained as

$$\mathrm{CV}(\lambda) \approx \widehat{\mathrm{OBS}}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i \sum_{j=1}^{m_i}(w_{\lambda,ij}\hat{f}_\lambda(z_{ij}) - w_{\lambda,ij}^{[-i]}\hat{f}_\lambda^{[-i]}(z_{ij})). \tag{2.25}$$

Based on (2.25) and by using several first order Taylor expansions, a generalized approximate cross validation (GACV) can be derived for randomized covariate data. Before we proceed, we would like to establish some notations.

2.3.3.2 Parametric formulation of $I_\lambda^{Z,\Pi}$

As we previously discussed, $I_\lambda^{Z,\Pi}(f)$ can be formulate parametrically as

$$I_\lambda^{Z,\Pi}(\vec{y},\vec{f}\,) = -\frac{1}{n}\sum_{i=1}^{n}\log\sum_{j=1}^{m_i}\pi_{ij}p(y_i|f_{ij}) + \frac{\lambda}{2}\vec{f}^{\,T}\Sigma_\lambda\vec{f}, \tag{2.26}$$

where $\vec{f} = (f_{11},...,f_{1m_1},f_{21},...,f_{nm_n})^T$ denotes the vector of f evaluated at $\{z_{ij}, 1 \le i \le n, 1 \le j \le m_i\}$, $\vec{y} = (\vec{y}_1^{\,T}...,\vec{y}_n^{\,T})^T$ with $\vec{y}_i = (y_i,...,y_i)^T$ being m_i replicates of y_i and Σ_λ is the positive semi-definite matrix satisfying $\lambda J(f) = \vec{f}^{\,T}\Sigma_\lambda\vec{f}$. Note that minimizing $I_\lambda^{Z,\Pi}(f)$ in $\mathcal{H}$ is equivalent to minimizing $I_\lambda^{Z,\Pi}(\vec{y},\vec{f}\,)$ in $\mathbb{R}^{m_1+\cdots+m_n}$. Hence

$$\vec{f}_\lambda = (\hat{f}_\lambda(z_{11}),...,\hat{f}_\lambda(z_{1m1}),\hat{f}_\lambda(z_{21}),...,\hat{f}_\lambda(z_{nm_n}))^T \tag{2.27}$$

minimizes (2.26). Similarly, we can denote

$$\vec{f}_\lambda^{[-i]} = (\hat{f}_\lambda^{[-i]}(z_{11}),...,\hat{f}_\lambda^{[-i]}(z_{1m_1}),\hat{f}_\lambda^{[-i]}(z_{21}),...,\hat{f}_\lambda^{[-i]}(z_{nm_n}))^T \tag{2.28}$$

the minimizer of (2.26) with ith subject omitted.

2.3.3.3 Generalized average of submatrices, randomized estimator

To define the GACV and randomized GACV we use the concept of **generalized average** of submatrices and its **randomized estimator** introduced in Gao, Wahba, Klein and Klein (2001)[17] for the multivariate outcomes case. Let A be a square matrix with submatrices $A_{ii}, 1 \le i \le n$ on the diagonal. Denote $A_{ii} = (a_{st}^i)_{m_i\times m_i}, 1 \le s,t \le m_i$. Because A_{ii}'s may have different dimensions, we calculate for each A_{ii}

$$\delta_i = \frac{1}{nm_i}\sum_{k=1}^{n}\sum_{j=1}^{m_k}a_{jj}^k = \frac{1}{nm_i}tr(A) \tag{2.29}$$

and

$$\gamma_i = \begin{cases} 0, & \text{if } m_i = 1 \\ 1/(nm_i(m_i-1))\sum_{k=1}^{n}\sum_{s\neq t}a_{st}^k, & \text{if } m_i > 1. \end{cases} \tag{2.30}$$

Then the generalized average of A_{ii} is defined by

$$\bar{A}_{ii} = (\delta_i - \gamma_i) I_{m_i \times m_i} + \gamma_i \cdot e_i e_i^T = \begin{pmatrix} \delta_i & \gamma_i & \cdots & \gamma_i \\ \gamma_i & \delta_i & \cdots & \gamma_i \\ \vdots & \vdots & \ddots & \vdots \\ \gamma_i & \gamma_i & \cdots & \delta_i \end{pmatrix} \tag{2.31}$$

where $e_i = (1, 1..., 1)^T$ is the unit vector of length m_i. In this case, the inverse of $\bar{A}_{ii}$ can be easily obtained by

$$\bar{A}_{ii}^{-1} = \frac{1}{\delta_i - \gamma_i} I_{m_i \times m_i} - \frac{\gamma_i}{(\delta_i - \gamma_i)(\delta_i + (m_i - 1)\gamma_i)} \, e_i e_i^T. \tag{2.32}$$

Now we discuss how to obtain a randomized estimator of $\bar{A}_{ii}$. Let $\epsilon = (\epsilon_1^T, ..., \epsilon_n^T)^T$, where $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{im_i})^T$ with each ϵ_{ij} generated independently from $N(0, \sigma^2)$. Denote $\bar{\epsilon} = (\bar{\epsilon}_1, ..., \bar{\epsilon}_1, \bar{\epsilon}_2, ..., \bar{\epsilon}_n)^T$ the corresponding mean vector with m_i replicates of $\bar{\epsilon}_i$ for each $1 \leq i \leq n$, where $\bar{\epsilon}_i = 1/\sqrt{m_i} \sum_{j=1}^{m_i} \epsilon_{ij}$. Then we observe the following facts

$$E\epsilon^T A \epsilon = \sigma \cdot tr(A) \tag{2.33}$$

$$E\left\{\bar{\epsilon}^T A \bar{\epsilon} - \epsilon^T A \epsilon\right\} = \sigma \cdot \sum_{k=1}^{n} \sum_{s \neq t} a_{st}^k. \tag{2.34}$$

Thus, a randomized estimate of $\bar{A}_{ii}$ can be obtained by replacing δ_i and γ_i with their unbiased estimates $\frac{1}{nm_i\sigma} \epsilon^T A \epsilon$ and $\frac{1}{nm_i(m_i-1)\sigma} (\bar{\epsilon}^T A \bar{\epsilon} - \epsilon^T A \epsilon)$.

2.3.3.4 The GACV and randomized GACV

We now present the result of GACV as follow. Details of the derivation can be found in Appendix B. Denote H the influence matrix of (2.26) with respect to $\vec{f}$

evaluated at $\vec{f_\lambda}$. Write

$$H = \begin{pmatrix} H_{11} & * & * & * \\ * & H_{22} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & H_{nn} \end{pmatrix}_{\sum m_i \times \sum m_i} \tag{2.35}$$

where each H_{ii} is a $m_i \times m_i$ submatrix on the diagonal with respect to to $(f_{i1}, ..., f_{im_i})^T$. Define $W_i = \text{diag}(b''(\hat{f}_\lambda(z_{i1})), ..., b''(\hat{f}_\lambda(z_{im_i})))$ the diagonal matrix of estimated variances. Let $W = \text{diag}(W_1, ..., W_n)$ be the "big" variance matrix for all the observations. Denote $G = I - HW$ with submatrices $G_{ii} = I_{m_i \times m_i} - H_{ii}W_i, 1 \leq i \leq n$ on the diagonal. Now let $\bar{H}_{ii}$ and $\bar{G}_{ii}$ denote the generalized average of submatrices H_{ii} and G_{ii}. Then the generalized approximate cross validation (GACV) can be written as

$$\text{GACV}(\lambda) = \widehat{\text{OBS}}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i(d_{i1}, ..., d_{im_i})\bar{G}_{ii}^{-1}\bar{H}_{ii} \begin{pmatrix} y_i - \hat{\mu}_\lambda(z_{i1}) \\ \vdots \\ y_i - \hat{\mu}_\lambda(z_{im_i}) \end{pmatrix} \tag{2.36}$$

where $\hat{\mu}_\lambda(z_{ij}) = b'(\hat{f}_\lambda(z_{ij}))$ denote the estimated mean response and

$$d_{ij} = w_{\lambda,ij}\left[(y_i - \hat{\mu}_\lambda(z_{ij}))(\hat{f}_\lambda(z_{ij}) - \sum_{k=1}^{m_i} w_{\lambda,ik}\hat{f}_\lambda(z_{ik})) + 1\right]. \tag{2.37}$$

In practice, however, computation of the influence matrix H for large data sets is expensive and may be unstable. Note that, in order to compute $\bar{H}_{ii}$ and $\bar{G}_{ii}$, we only need the sum of traces and the sum of off-diagonal entries of H_{ii}'s and G_{ii}'s. Therefore, the exact computation of H and G can be avoided using randomized estimates of $\bar{H}_{ii}$ and $\bar{G}_{ii}$. To do this, we first generate a random perturbation vector $\epsilon = (\epsilon_1^T, ..., \epsilon_n^T)^T$, where $\epsilon_i = (\epsilon_{i1}, ..., \epsilon_{im_i})^T$ and ϵ_{ij}'s are iid from $N(0, \sigma^2)$. Then compute the mean vector $\bar{\epsilon} = (\bar{\epsilon}_1, ..., \bar{\epsilon}_1, \bar{\epsilon}_2, ..., \bar{\epsilon}_n)^T$ where $\bar{\epsilon}_i = 1/\sqrt{m_i}\sum_{j=1}^{m_i}\epsilon_{ij}$.

Denote $\vec{f}_\lambda^{\vec{y}+\epsilon}$ and $\vec{f}_\lambda^{\vec{y}+\bar{\epsilon}}$ the minimizers of (2.26) with the perturbed data $\vec{y}+\epsilon$ and $\vec{y}+\bar{\epsilon}$. Similarly, denote $\vec{f}_\lambda^{\vec{y}}(=\vec{f}_\lambda)$ the minimizer with the original data. To ease the computational burden, we can set $\vec{f}_\lambda^{\vec{y}}$ as the initial value for the EM algorithm of $\vec{f}_\lambda^{\vec{y}+\epsilon}$ and $\vec{f}_\lambda^{\vec{y}+\bar{\epsilon}}$. Because H is the influence matrix, we have that

$$\vec{f}_\lambda^{\vec{y}+\epsilon} \approx \vec{f}_\lambda^{\vec{y}} + H\epsilon, \;\; \vec{f}_\lambda^{\vec{y}+\bar{\epsilon}} \approx \vec{f}_\lambda^{\vec{y}} + H\bar{\epsilon}. \tag{2.38}$$

This yields

$$\epsilon^T H\epsilon \approx \epsilon^T(\vec{f}_\lambda^{\vec{y}+\epsilon} - \vec{f}_\lambda^{\vec{y}}), \;\; \bar{\epsilon}^T H\bar{\epsilon} \approx \bar{\epsilon}^T(\vec{f}_\lambda^{\vec{y}+\bar{\epsilon}} - \vec{f}_\lambda^{\vec{y}}). \tag{2.39}$$

Thus, a randomized estimate of $\bar{H}_{ii}$ can be obtained as we previously described. Also it is straightforward to show that

$$\epsilon^T G\epsilon \approx \epsilon^T\epsilon - \epsilon^T W(\vec{f}_\lambda^{\vec{y}+\epsilon} - \vec{f}_\lambda^{\vec{y}}), \;\; \bar{\epsilon}^T G\bar{\epsilon} \approx \bar{\epsilon}^T\bar{\epsilon} - \bar{\epsilon}^T W(\vec{f}_\lambda^{\vec{y}+\bar{\epsilon}} - \vec{f}_\lambda^{\vec{y}}) \tag{2.40}$$

which implies a randomized estimate of $\bar{G}_{ii}$. In order to reduce the variance of randomized trace estimates, one may draw R independent perturbation vectors $\epsilon^1, ..., \epsilon^R$ and compute for each $\epsilon^r, 1 \le r \le R$ the randomized estimates $\hat{\bar{H}}_{ii}^r, 1 \le i \le n$ and $\hat{\bar{G}}_{ii}^r, 1 \le i \le n$. Then the ($R$-replicated) ranGACV function is

$$\text{ranGACV}(\lambda) = \widehat{\text{OBS}}(\lambda) + \frac{1}{nR}\sum_{r=1}^{R}\sum_{i=1}^{n} y_i(d_{i1}, ..., d_{im_i})(\hat{\bar{G}}_{ii}^r)^{-1}\hat{\bar{H}}_{ii}^r \begin{pmatrix} y_i - \hat{\mu}_\lambda(z_{i1}) \\ \vdots \\ y_i - \hat{\mu}_\lambda(z_{im_i}) \end{pmatrix}. \tag{2.41}$$

2.4 Covariate measurement error (model)

Covariate measurement error is a common occurrence in many experimental settings including surveys, clinical trials and medical studies. Suppose that $x_i = (x_{i1}, ..., x_{id})^T$ takes values in the real space $\mathbb{R}^d$. In the presence of measurement error,

x_i is not directly observed but instead $x_i^{err} = x_i + u_i$ is observed, where $u_i, 1 \leq i \leq n$ are iid random errors, independent of (y_i, x_i). To estimate the regression function, our idea is to treat covariate measurement error as a special case of randomized covariate data. More specifically, each x_i is considered as a random vector distributed as $x_i^{err} - u_i$. When the error distribution is known, the distribution for x_i can be obtained immediately, and therefore RC-PLE can be directly employed without any extra effort.

However, in practical applications, we often face the case that the error distribution is unknown. One common approach in the measurement error literature is to assume a parametric model for the error density and to estimate the unknown parameters from the data. Let $p(u_i|\theta)$ denote the specified error density indexed by a real vector θ ranging over $\Theta \subseteq \mathbb{R}^q$ and let $F(u_i|\theta)$ denote the corresponding c.d.f. function. Since our goal is to estimate the regression function, θ is treated as a nuisance parameter. Given (f, θ), y_i has a marginal density of

$$p(y_i|f, \theta) = \int_{\mathbb{R}^d} p(y_i|x_i^{err} - u_i, f)p(u_i|\theta)du_i. \tag{2.42}$$

Thus RC-PLE can be extended by

$$I_\lambda^E(f, \theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathbb{R}^d} p(y_i|x_i^{err} - u_i, f)p(u_i|\theta)du_i + \frac{\lambda}{2}J(f). \tag{2.43}$$

In this case, we still need Assumption A.1 to obtain the existence of the penalized likelihood estimate. In addition we state the following extra assumption which can be satisfied with most parametric models for the error distribution.

ASSUMPTION B.1. The c.d.f. function $F(u|\theta)$ is continuous in θ for any $u \in \mathbb{R}^d$ and the parameter space Θ is compact.

Now we can show the existence of penalized likelihood estimate by the following Theorem which is actually a corollary to Theorem 2.2.

THEOREM 2.7. *Under A.1, A.2 and B.1, there exist* $f_\lambda \in \mathcal{H}$ *and* $\theta_\lambda \in \Theta$ *such that* $I^E_\lambda(f_\lambda, \theta_\lambda) = \inf_{f\in\mathcal{H},\theta\in\Theta} I^E_\lambda(f, \theta)$.

Proof See Appendix A. □

2.5 Covariate measurement error (computation)

In order to compute an estimator, we extend QPLE described in Section 2.3.1 as follows. Denote $(f^{(t)}, \theta^{(t)})$ the parameters estimated at iteration t. Let $z_j^{(t)}, 1 \le j \le m$ and $\pi_j^{(t)}, 1 \le j \le m$ denote the quadrature rule based on the density function $p(u|\theta^{(t)})$. Note that the quadrature rules can be generated using the method introduced in Section 2.3.2. It is not hard to see that the E-step at iteration $t+1$ is to compute the expectation of the penalized likelihood $-\frac{1}{n}\sum_{i=1}^n \log p(y_i|x_i^{err} - u_i, f)p(u_i|\theta) + \frac{\lambda}{2}J(f)$ with respect to the conditional distributions $[u_i|y_i, x_i^{err}, f^{(t)}, \theta^{(t)}]$, $1 \le i \le n$. Using the quadrature rule, each $[u_i|y_i, x_i^{err}, f^{(t)}, \theta^{(t)}]$ can be approximated by a discrete distribution with support $\{z_j^{(t)}, 1 \le j \le m\}$ and mass function $P(u_i = z_j^{(t)}) = w_{ij}^{(t)}$, where

$$w_{ij}^{(t)} = \frac{\pi_j^{(t)} p(y_i|x_i^{err} - z_j^{(t)}, f^{(t)})}{\sum_k \pi_k^{(t)} p(y_i|x_i^{err} - z_k^{(t)}, f^{(t)})}. \tag{2.44}$$

Thus the E-step can be written as

$$Q(f, \theta|f^{(t)}, \theta^{(t)}) = \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m w_{ij}^{(t)} \cdot \log p(y_i|x_i^{err} - z_j^{(t)}, f) - \frac{\lambda}{2}J(f) + \frac{1}{n}\sum_{i=1}^n\sum_{j=1}^m w_{ij}^{(t)} \cdot \log p(z_j^{(t)}|\theta). \tag{2.45}$$

Then the M-step maximizes $Q(f, \theta | f^{(t)}, \theta^{(t)})$, which can be done by separately maximizing a complete data penalized likelihood of f

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^{(t)} \cdot \log p(y_i | x_i^{err} - z_j^{(t)}, f) - \frac{\lambda}{2} J(f) \tag{2.46}$$

and a complete data log likelihood of θ

$$\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m} w_{ij}^{(t)} \cdot \log p(z_j^{(t)} | \theta). \tag{2.47}$$

Therefore the M-step becomes a standard problem which can be solved by much existing software. When the EM algorithm converges, we will obtain the QPLE estimate $(\hat{f}_\lambda, \hat{\theta}_\lambda)$.

Finally, we show how to select the smoothing parameter λ in the case of covariate measurement error. Note that our goal is to construct a good estimator of f, and θ is treated as a nuisance parameter. In other words, we only care about the goodness of fit of the $\hat{f}_\lambda$. Therefore λ can be selected in the same way as randomized covariate data. To do this, we first estimate the error distribution by $p(u_i | \hat{\theta}_\lambda)$ and then determine each covariate distribution P_i according to the relation $x_i = x_i^{err} - u_i$. After that, the method introduced in Section 2.3.3 can employed directly for the choice of λ.

Correcting for measurement error is a broad statistical research topic. In the interest of space, we only discuss the situation when we have a parametric model for the error distribution. It would be possible to extend our method to other situations of measurement error. For example, when additional data is available, such as a sample from the error distribution or repeated observations for some x_i, we may estimate the error distribution more accurately by using other approaches. Also, sometimes, the parametric model $p(u_i | \theta)$ may not be available and in this case, we may want to estimate the error distribution nonparametrically. These are interesting topics for future research.

2.6 Missing covariate data (model)

Now we describe penalized likelihood regression with missing covariate data. We assume the missing mechanism to be missing at random.

2.6.1 Notations and model

Let $x_i = (x_{i1}, ..., x_{id})$ denote the vector of covariates ranging over a subspace of $\mathbb{R}^d$. By the idea of Ibrahim's method of weights (Ibrahim, 1990[21] and Ibrahim, Lipsitz and Chen, 1999[22]), we first assume a parametric model for the marginal density of x_i, denoted as $p(x_i|\theta) > 0$, where $\theta \in \Theta \subseteq \mathbb{R}^q$ is a real vector of indexing parameters. Here θ is treated as a nuisance parameter.

Write $x_i = (x_i^{obs}, x_i^{mis})$ where x_i^{obs} is a vector of observed components and x_i^{mis} is a $d_i \times 1$ vector of missing components. Following Little and Rubin, (2002)[32], the likelihood of (f, θ) can be obtained by integrating or summing out the missing components in the joint density for (y_i, x_i)

$$L(f, \theta) = \sum_{i=1}^{n} \log \int_{\mathbb{R}^{d_i}} p(y_i|x_i, f)p(x_i|\theta)dx_i^{mis} \tag{2.48}$$

where $\int_{\mathbb{R}^{d_i}} p(y_i|x_i, f)p(x_i|\theta)dx_i^{mis} \equiv p(y_i|x_i, f)p(x_i|\theta)$ if x_i is completely observed. Then (f, θ) can be estimated by minimizing the following missing data penalized likelihood:

$$I_\lambda^M(f, \theta) = -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathbb{R}^{d_i}} p(y_i|x_i, f)p(x_i|\theta)dx_i^{mis} + \frac{\lambda}{2}J(f). \tag{2.49}$$

We note that this method can be viewed as an extension of RC-PLE. Define $P_{i,mis}^{\theta}$ the probability measure over $\mathbb{R}^{d_i}$, with respect to the conditional density of $[x_i^{mis}|x_i^{obs}, \theta]$

$$p(x_i^{mis}|x_i^{obs}, \theta) = \frac{p(x_i|\theta)}{\int_{\mathbb{R}^{d_i}} p(x_i|\theta)dx_i^{mis}}, \;\; x_i^{mis} \in \mathbb{R}^{d_i}. \tag{2.50}$$

Note that (2.50) is well-defined since $\int_{\mathbb{R}^{d_i}} p(x_i|\theta)dx_i^{mis} < \infty$ from the Fubini's Theorem. Let

$$\delta_{x_i^{obs}}(A) = \begin{cases} 1 & \text{if } x_i^{obs} \in A \\ 0 & \text{if } x_i^{obs} \notin A \end{cases} \tag{2.51}$$

denote the dirac measure defined for x_i^{obs}. Consider the product measure $P_i^\theta = \delta_{x_i^{obs}} \times P_{i,mis}^\theta$ which satisfies that for any Borel sets $A_1 \subset \mathbb{R}^{d-d_i}$, $A_2 \subset \mathbb{R}^{d_i}$ and their Cartesian product $A_1 \times A_2$, we have

$$P_i^\theta(A_1 \times A_2) = \delta_{x_i^{obs}}(A_1) \cdot P_{i,mis}^\theta(A_2). \tag{2.52}$$

Then it is not hard to see that

$$I_\lambda^M(f,\theta) = -\frac{1}{n}\sum_{i=1}^n \log \int_{\mathbb{R}^d} p(y_i|x_i,f)dP_i^\theta + \frac{\lambda}{2}J(f) - \frac{1}{n}\sum_{i=1}^n \log \int_{\mathbb{R}^{d_i}} p(x_i|\theta)dx_i^{mis} \tag{2.53}$$

is composed of a randomized covariate penalized likelihood of f and a log-likelihood of θ. Hence missing covariate data can be treated as a special case of randomized covariate data, allowing covariate distributions to be flexible.

2.6.2 Existence of the estimator

The following assumptions can be easily satisfied in the most experimental settings.

ASSUMPTION M.1. $\mathcal{D}_i^\theta = \{x_i^{mis} \in \mathbb{R}^{d_i} \ : \ p(x_i|\theta) > 0\}$ is compact for all $1 \le i \le n$ and $\theta \in \Theta$.

ASSUMPTION M.2. The density function $p(x|\theta)$ is continuous in θ for any $x \in \mathbb{R}^d$ and the parameter space Θ is compact.

The existence of the penalized likelihood estimate can be guaranteed by the following Theorem which is actually a corollary to Theorem 2.2.

THEOREM 2.8. *Under A.1, A.2, M.1 and M.2, there exist* $f_\lambda \in \mathcal{H}$ *and* $\theta_\lambda \in \Theta$ *such that* $I_\lambda^M(f_\lambda, \theta_\lambda) = \inf_{f\in\mathcal{H},\theta\in\Theta} I_\lambda^M(f,\theta)$.

Proof See Appendix A. □

2.7 Missing covariate data (computation)

In order to compute an estimator, we can extend QPLE in the same way as covariate measurement error. Denote $(f^{(t)}, \theta^{(t)})$ the parameters estimated at iteration t. Let $z_{ij}^{(t)}, 1 \le j \le m_i$ and $\pi_{ij}^{(t)}, 1 \le j \le m_i$ denote the quadrature rule based on the probability measure $P_i^{\theta^{(t)}}$ defined in (2.52). Then the E-step at iteration $t+1$ can be written as

$$Q(f,\theta|f^{(t)},\theta^{(t)}) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^{(t)} \cdot \log p(y_i|z_{ij}^{(t)}, f) - \frac{\lambda}{2}J(f) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^{(t)} \cdot \log p(z_{ij}^{(t)}|\theta) \tag{2.54}$$

where

$$w_{ij}^{(t)} = \frac{\pi_{ij}^{(t)} p(y_i|z_{ij}^{(t)}, f^{(t)})}{\sum_k \pi_{ik}^{(t)} p(y_i|z_{ik}^{(t)}, f^{(t)})}. \tag{2.55}$$

Then the M-step can be done by separately maximizing

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^{(t)} \cdot \log p(y_i|z_{ij}^{(t)}, f) - \frac{\lambda}{2}J(f) \tag{2.56}$$

and

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}^{(t)} \cdot \log p(z_{ij}^{(t)}|\theta) \tag{2.57}$$

which is computationally straightforward assuming the log-concavity of $p(x|\theta)$ as a function of θ. Again, when the EM algorithm converges, the QPLE estimate $(\hat{f}_\lambda, \hat{\theta}_\lambda)$ can be obtained.

In order to select the smoothing parameter, we note that θ is a nuisance parameter and the choice of λ only depends on the goodness of fit of $\hat{f}_\lambda$. Therefore, we may select λ in the same way as randomized covariate data. This is straightforward, since we can take $P_i^{\hat{\theta}_\lambda}$ defined in (2.52) as the covariate distribution. After that the method in Section 2.3.3 can employed directly.

Following Ibrahim, Lipsitz and Chen (1999)[22], our method can also be extended to the non-ignorable missing data mechanism. In this case, we may specify a parametric model for the missing data mechanism and incorporate it into the penalized likelihood. The extension is similar but more complicated. Thus this is another topic for future research.

2.8 Numerical Studies

In this section, we illustrate our method by several simulated examples with covariate measurement error and missing covariates. For each simulated data set, we will compare: (a) RC-PLE (QPLE); (b) full data analysis before measurement error or missing covariates; and (c) naive estimator that ignores measurement error or leaves out the observations with missing covariates. Note that the choice of the smoothing parameter has strong effect on the penalized likelihood estimator. Hence in order to show the potential gain of our method, for each data set, λ is selected by both ranGACV and the optimal value that minimizes the Theoretical Kullback-Leibler distance (TKL), which does not depend on the nuisance parameter θ.

$$\text{TKL} = \frac{1}{n}\sum_{i=1}^{n} E_{y_i^0|x_i,f^*}\left\{\log\frac{p(y_i^0|x_i,f^*)}{p(y_i^0|x_i,\hat{f})}\right\} \tag{2.58}$$

where f^* is the true regression function, $\hat{f}$ denotes its estimator and x_i denotes the true covariate vector before measurement error or 'missing'. Note that tuning by minimizing TKL is only available in a simulation study when the "truth" is known.

Our numerical studies focus on Poisson distribution and Bernoulli distribution which are also the cases in our real data set. The goal is to illustrate:

- the gain of RC-PLE (QPLE);
- the performance of ranGACV;
- the robustness of QPLE to the choice of quadrature rules.

All the simulations are conducted using R-2.9.1 installed in Red Hat Enterprise Linux 5.

2.8.1 Examples of measurement error

Cubic spline regression is perhaps the most popular case of penalized likelihood regression. We consider the following examples from Binomial and Poisson distributions:

(i) $p(y|x) = \begin{pmatrix} 2 \\ y \end{pmatrix} p(x)^y (1-p(x))^{2-y},\ y = 0, 1, 2$, where

$$p(x) = 0.63x\cos(2\pi x) + 0.36;$$

(ii) $p(y|x) = \Lambda(x)^y e^{-\Lambda(x)}/y!,\ y = 0, 1, 2...$, where

$$\Lambda(x) = 16e^{-18(x-0.4)^2} - 5e^{-7(x-0.5)^2} + 5;$$

(iii) Same distribution as (ii) except

$$\Lambda(x) = 10^6(x^{11}(1-x)^6) + 10^4(x^3(1-x)^{10}) + 2$$

which is a modification of Example 5.5 of Gu (2002)[20].

In each case, we take $X \sim U[0,1]$ and generate a sample of $n = 101$ (x, y) pairs. For each sample generated, measurement errors are created with the following scheme. We first randomly select five (x, y) pairs as complete observations and then in the rest of the 96 pairs, random errors are generated by $x_i + u_i$, where u_i's are iid either N$(0, \sigma^2)$ or U$[-\delta, \delta]$ for various values of the noise-to-signal ratio var(u)/var(X). For each generated data set, QPLE is conducted using either the Gaussian quadrature rule or the grid quadrature rule, where the Gaussian quadrature rule is computed by the *statmod* package in R-2.9.1. Note that we generate the same number of nodes for each noisy x_i. Simulation results are summarized by the following figures.

Figure 2.1 shows the estimated curves from one simulated data set of case (i) with normal error and var(u)/var$(X) = 0.25$. QPLE is computed via Gaussian quadrature where 11 nodes are created for each noisy x_i. Panel (c) plots for each regression method the box plot of TKL distances (2.58) calculated from 100 repeated simulations. We also report in (d) the TKL distances calculated in the same simulation setting except that u is uniform (with the same noise-to-signal ratio).

Remark 1 *Throughout Section 2.8, the choice of the curves to display from the various 100 simulations is primarily subjective but deemed to be typical of the bulk of the visual images of the comparisons between the estimates. An idea of the scatter in the TKL distances over the 100 simulations may be seen in the box plots.*

Figure 2.2 shows the estimated curves from one simulation for case (ii) with uniform error and var(u)/var$(X) = 0.3$. We assume that δ is unknown when QPLE is conducted. At each EM iteration, we use Gaussian quadrature and create 9 nodes for each noisy x_i. Panel (c) shows the TKL distances from 100 simulations. Panel

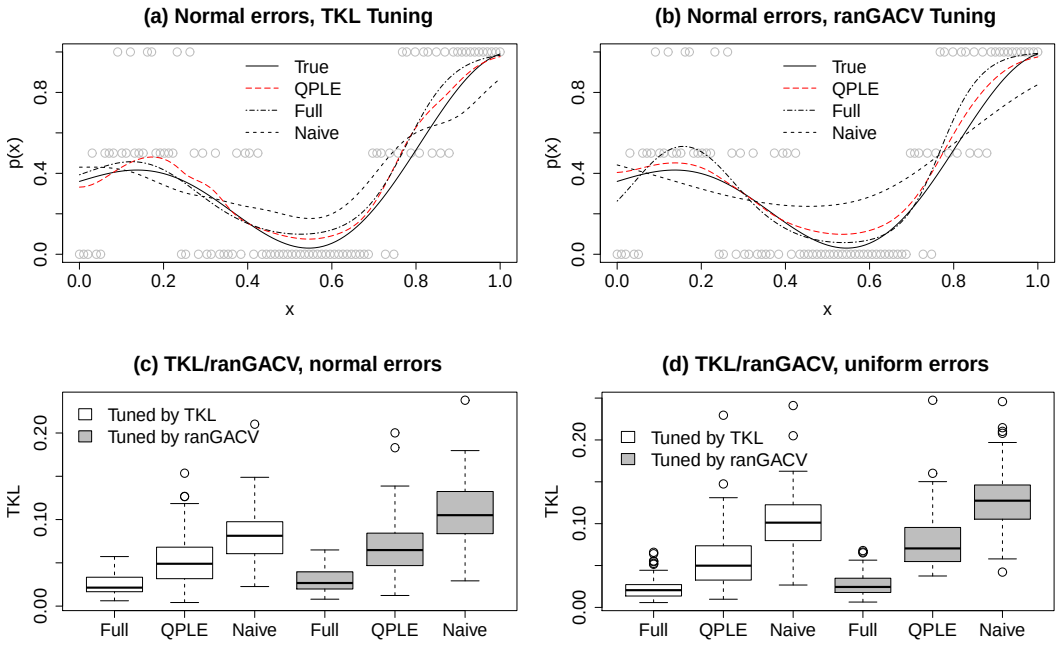

Figure 2.1: Estimated curves and TKL distances for case (i). Panels (a) and (b) compare the target (True) curve, and three estimated curves obtained from the full data analysis (Full), the QPLE estimate, and the Naive estimate. (a) Tuning: TKL, (b) Tuning: ranGACV. In (a) and (b) $u \sim N(0, 0.145^2)$, assumed known. Panels (c) and (d) provide plots of TKL distances. (c) $u \sim N(0, 0.145^2)$, assumed known. (d) $u \sim U[-0.25, 0.25]$, assumed known.

(d) is obtained in the same simulation setting except that u is normal (with the same noise-to-signal ratio), σ is unknown.

Our results indicate the significant gain of QPLE, when the smoothing parameter is selected by either TKL or ranGACV. As we previously discussed, QPLE incorporates the information about the error distribution and hence is more informative. Generally speaking, when measurement errors are ignored, the estimated curve of naive method tends to be oversmoothed and more biased near the modes and boundaries. Similar phenomenon has been noted for other nonparametric regression methods, for example, Local polynomial estimate, as in Delaigle, Fan and Carroll (2009)[12]. For the choice of smoothing parameter, the proposed ranGACV inherits the property of traditional ranGACV. As simulations suggest, it is capable of picking λ close to its optimal value even when θ is estimated.

We summarize the influence of quadrature rules on QPLE at Figure 2.3, using case (iii) with normal error and $\text{var}(u)/\text{var}(X) = 0.25$. In the computation, $\text{var}(u)$ is assumed to be unknown and λ is selected by TKL. We consider four QPLE estimators (QPLE1, QPLE2, QPLE3 and QPLE4) computed via, respectively, Gaussian quadrature, grid quadrature, Gaussian quadrature when u is wrongly assumed to be uniform and grid quadrature when u is wrongly assumed to be uniform. We first compare these quadrature rules by setting the number of nodes (for each noisy x_i) to be 11. The top two panels show the estimated curves from one simulation and panel (c) reports the TKL distances calculated from 100 simulations. Then we study the influence of the number of the nodes. On panel (d), we plot for each quadrature the mean TKL distance (based on 100 simulations) versus the number of nodes. From the simulation results, we observed no significant difference between Gaussian quadrature and grid quadrature, though, as we expected, Gaussian quadrature is more

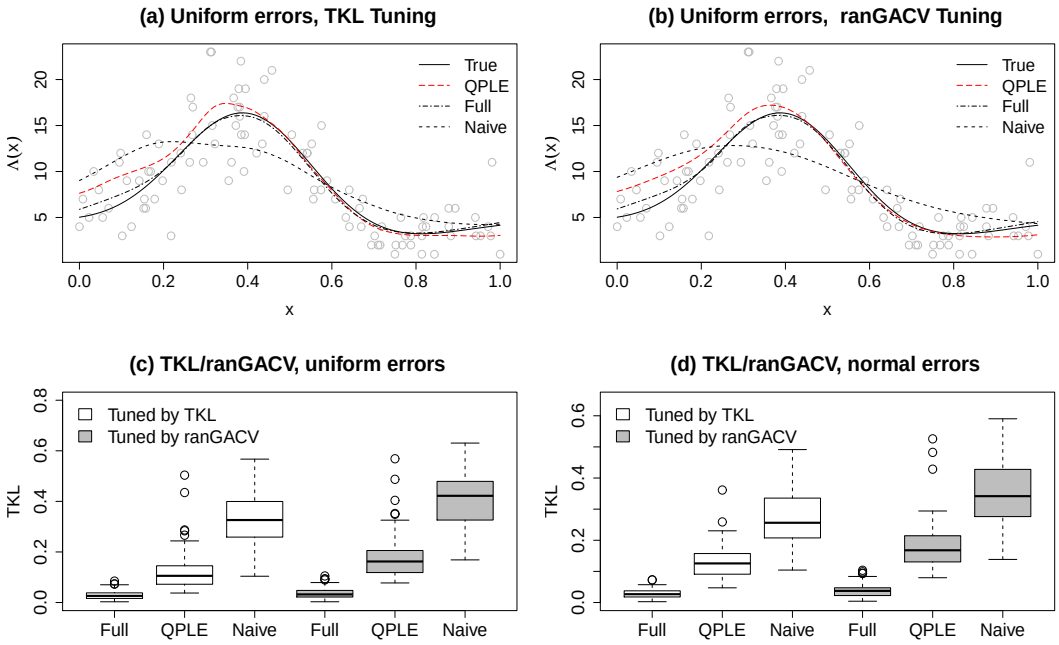

Figure 2.2: Estimated curves and TKL distances for case (ii). Panels (a) and (b) compare the target (True) curve, and three estimated curves obtained from the full data analysis (Full), the QPLE estimate, and the Naive estimate. (a) Tuning: TKL, (b) Tuning: ranGACV. In (a) and (b) $u \sim U[-0.273, 0.273]$, $\delta = 0.273$ assumed unknown. Panels (c) and (d) provide plots of TKL distances. (c) $u \sim U[-0.273, 0.273]$, $\delta = 0.273$ assumed unknown. (d) $u \sim N(0, 0.158^2)$, $\sigma = 0.158$ assumed unknown.

efficient. Surprisingly, even with a wrong error distribution prespecified, the potential gain of QPLE is still significant. Hence we may say that QPLE is robust to the choice of the quadrature. We also note that QPLE does not require a large number of quadrature nodes to compute a good estimator. There is not much gain to create more nodes if we already have enough. Hence, in our numerical experiments, we generally compute 7-12 nodes for each noisy or missing component in the covariates.

2.8.2 Examples of missing covariate data

In this section, we consider Franke's "principal test function"

$$T(x) = \frac{3}{4}e^{-((9x_1-2)^2+(9x_2-2)^2)/4} + \frac{3}{4}e^{-((9x_1+1)^2/49+(9x_2+1)^2/10)} + \frac{1}{2}e^{-((9x_1-7)^2+(9x_2-3)^2)/4} - \frac{1}{5}e^{-((9x_1-4)^2+(9x_2-7)^2)} \tag{2.59}$$

which was used as a test function of smoothing splines in Wahba (1983)[38]. $T(x)$ is shown in Figure 2.4. Consider the following examples

(i) Binomial distribution: $p(y|x) = \binom{5}{y} p(x)^y (1-p(x))^{5-y}$, where

$$p(x) = \frac{1}{1.24}(T(x) + 0.198); \tag{2.60}$$

(ii) Poisson distribution: $p(y|x) = \Lambda(x)^y e^{-\Lambda(x)}/y!$, where

$$\Lambda(x) = 15T(x) + 3. \tag{2.61}$$

In each case, we take $X = (X_1, X_2) \sim U[0,1] \times [0,1]$ and generate a sample of $n = 300$ observations from the distribution of (Y, X). Afterwards, a missing data is created in a way that if $y > 3$ in case (i) or $y > 10$ in case (ii), we randomly take one of the following actions with equal probability: (1) delete x_1 only; (2) delete x_2 only

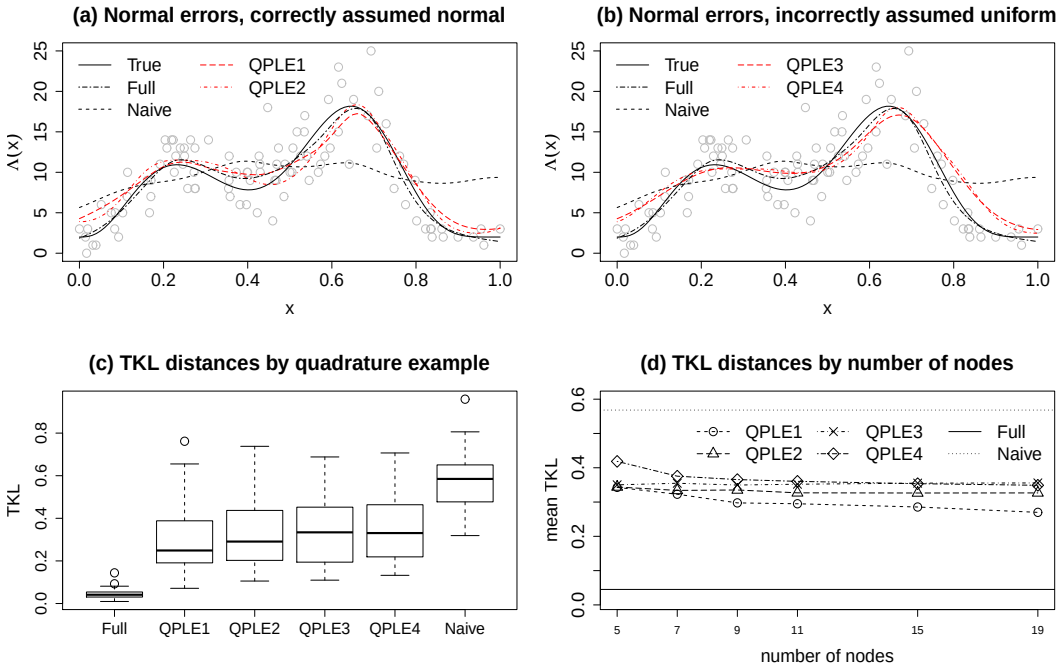

Figure 2.3: Estimated curves and TKL distances for case (iii). $u \sim N(0, 0145^2)$, assumed unknown. Tuning: TKL. Panels (a) and (b) give the target curve, and estimated curves from Full and Naive estimate. Panel (a) compares the Gaussian quadrature (QPLE1) and the grid quadrature (QPLE2) when the errors are correctly assumed to be zero-mean normal (with unknown variance), and panel (b) compares the Gaussian quadrature (QPLE3) and the grid quadrature (QPLE4) when the errors are incorrectly assumed to be uniform (with unknown range); (a) and (b) use 11 nodes. Panel (c) plots TKL distances, using 11 nodes. Panel (d) plots mean TKL versus number of nodes. The dotted upper and solid lower lines represent the mean TKL for the naive method and the full data analysis.

.

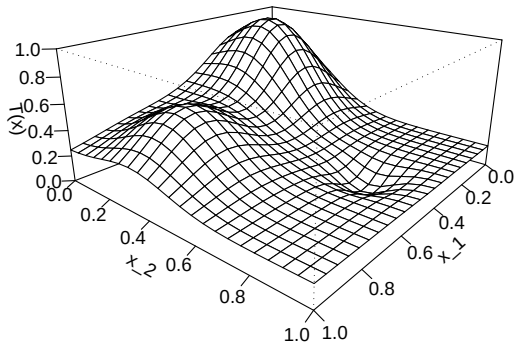

Figure 2.4: Franke’s principal test function

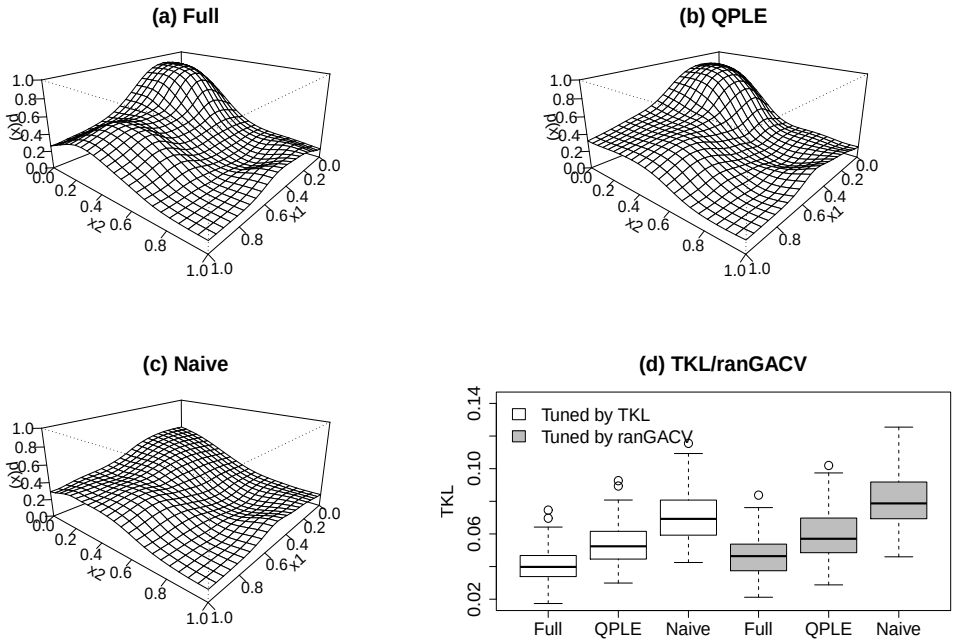

Figure 2.5: Estimated functions of $p(x_1, x_2)$ and TKL distances for case (i). (a) Full data estimate. (b) QPLE estimate. (c) Naive estimate. The λ's in (a), (b) and (c) are tuned by ranGACV. (d) Box plots of TKL distances when tuned by TKL and by ranGACV.

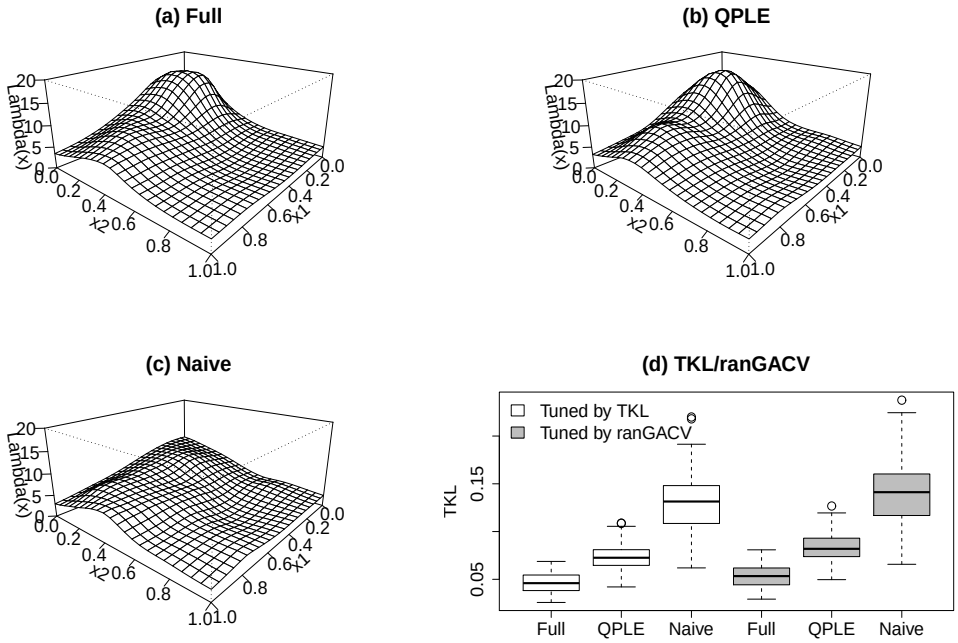

Figure 2.6: Estimated functions of $\Lambda(x_1, x_2)$ and TKL distances for case (ii). (a) Full data estimate. (b) QPLE estimate. (c) Naive estimate. The λ's in (a), (b) and (c) are tuned by ranGACV. (d) Box plots of TKL distances when tuned by TKL and by ranGACV.

and (3) delete both x_1 and x_2. On average, we create 47 incomplete observations (out of 300) in case (i) and 61 incomplete observations in case (ii).

We will test our method by thin plate spline regression. In order to implement QPLE, we specify for x a bivariate normal distribution $N(\mu, \Sigma)$, where $\mu = (\mu_1, \mu_2)^T$ and $\Sigma = \{\sigma_{ij}\}_{2\times 2}$ (an arbitrary covariance matrix) are to be estimated. At each EM iteration, we construct for each incomplete x_i a Gaussian quadrature rule, where 11 nodes are created for each missing component. Simulation results are summarized at Figure 2.5 and 2.6.

Figure 2.5 and 2.6 show the estimated functions where the smoothing parameter is tuned by ranGACV. The bottom right panel reports the TKL distances based on 100 simulations, when λ is selected by TKL and ranGACV. The performance of QPLE is also impressive in the case of missing covariate data. Note that most incomplete observations appeared near the 'peak' of the test function. In this case if these incomplete observations are left out, we will miss the information about the peak, as indicated by the naive estimator. On the other hand, by incorporating most information in the data including the observations with partially missing covariates, QPLE provides encouraging results, even though we actually specified a wrong covariate distribution.

2.8.3 Case study

In this section, we illustrate our method on an observational data set that has been previously analyzed, by deleting some covariates, and then comparing our method with the original analysis and the naive method of dropping files with missing covariates.

The Beaver Dam Eye Study is an ongoing population-based study of age-related ocular disorders. Subjects were a group of 4926 people aged 43-86 years at the

Attributes	unit	range	code
systolic blood pressure	$mmHg$	71-221	*sys*
serum total cholesterol	mg/dL	102-503	*chol*
age at baseline	$years$	43-86	*age*
body mass index	kg/m^2	15-64.8	*bmi*
taking hormone replacement therapy	yes/no	yes,no	*horm*
history of heavy drinking	yes/no	yes,no	*drin*

Table 2.1: Covariates for Pigmentary Abnormalities

start of the study who lived in Beaver Dam, WI and were examined at baseline, between 1988 and 1990. A description of the population and details of the study at baseline may be found in Klein, Klein, Linton and Demets (1991)[29]. Pigmentary abnormalities are one of the ocular disorders of interest in that study. Pigmentary abnormalities are an early sign of age-related macular degeneration and are defined by the presence of retinal depigmentation and/or increased retinal pigment.

Lin, Wahba, Xiang, Gao, Klein and Klein (2000)[31] and Gao, Wahba, Klein and Klein(2001)[17] considered only the $n = 2545$ women members of this cohort. 11.88% of them showed evidence of pigmentary abnormalities. They examined the association of pigmentary abnormalities with six other attributes at baseline, by fitting a Smoothing Spline ANOVA (SS-ANOVA) model. The six attributes are are listed in Table 2.1.

Let $p(x)$ be the probability that a subject with attribute vector x at baseline will be found to have a pigmentary abnormality in at least one eye, at baseline.

The model fitted was of the form

$$f(x) = \text{constant} + f_1(sys) + f_2(chol) + f_{12}(sys, chol) \qquad (2.62)$$
$$+ d_{age} \cdot age + d_{bmi} \cdot bmi + d_{horm} \cdot I_2(horm) + d_{drin} \cdot I_2(drin).$$

Here x denotes the vector of covariates listed in Table 2.1 and $f(x)$ is the logit form of the probability: $f(x) = \log \frac{p(x)}{1-p(x)}$.

The data analysis is summarized in Figure 2.7, which is adapted from Lin, Wahba, Xiang, Gao, Klein and Klein (2000)[31]. On each panel, we plot the estimated probability of pigmentary abnormalities as a function of *chol*, for various values of sys, age and $horm$. Note that we only plot for *bmi* = 27.5 and *drin* = no, because *bmi* has relatively small effect in the fitted model while only 152 out of 2585 subjects have *drin* = 1. Hence Figure 2.7 is adequate to demonstrate the estimated association patterns.

Generally speaking, higher *chol* was associated with a protective effect. However, when *chol* goes from 250 to 350 mg/dL, a "bump" appears on the estimated curves. This phenomenon provides us a good opportunity to test our method. In order to 'hide' the bump, we create a data set with missing covariates by deleting some attribute values for those subjects whose cholesterol is between 250 and 350. Consequently, 517 subjects with incomplete data are created with values of *sys*, *bmi* and *horm* randomly removed. More exactly, 30 subjects missed *sys*, *bmi* and *horm*, 109 subjects missed both *sys* and *bmi*, 118 subjects missed both *sys* and *horm* and 260 subjects missed only one attribute value.

We shall first claim that the methodology in this chapter can be extended to SS-ANOVA models without any extra effort, as illustrated in Appendix C. In this case, QPLE can be conducted following Ibrahim, Lipsitz and Chen (1999)[22]. We first model the joint covariate distribution via a sequence of one-dimensional conditional distributions. Note that (*age*, *chol*, *drin*) are always observed and hence we do not need to model them. Also, very few subjects have $drin = 1$, hence $drin$ will be

ignored in the modeling. Given (*age*, *chol*), we adopt a bivariate normal distribution (*sys*, *bmi*) $\sim N(\mu, \Sigma)$, where $\mu = (\mu_1,\ \mu_2)$ with $\mu_k = a_{k0} + a_{k1}age + a_{k2}chol,\ k = 1, 2$ and $\Sigma = \{\sigma_{ij}\}_{2\times 2}$ is an arbitrary covariance matrix, and the a's and Σ are to be estimated. Now conditionally on other attributes, $horm$ is modeled via a logistic regression model

$$p(horm = 1) = \frac{\exp\{a_{30} + a_{31}age + a_{32}chol + a_{33}sys + a_{34}bmi\}}{1 + \exp\{a_{30} + a_{31}age + a_{32}chol + a_{33}sys + a_{34}bmi\}}.$$

Following this construction of covariate distributions and using the method described in Section 2.3.2, a quadrature rule can be obtained recursively at each EM iteration. In the computation, the numbers of nodes generated for *sys*, *bmi* and *horm* are 10, 10 and 2 respectively. Results of QPLE are given at Figure 2.8. Figure 2.9 shows the naive estimator computed over the 2068 subjects without missing covariates.

Note that only the subjects with incomplete data contain information about the bumps. Consequently, the naive estimator omitted these bumps, leading to monotone decreasing probability curves. In words, high cholesterol appears to generally lower the risk of pigmentary abnormalities especially in the older, $horm = no$ group, aside from the "bump", from the full data analysis shown at Figure 2.7. However the naive estimator appears to make this risk decrease substantially more rapidly due to missing the "bump" completely, while the QPLE did an excellent job of recovering the original analysis–the QPLE estimated curves are very close to those of the full data analysis. This can be understood from the fact that most of the subjects with incomplete data missed only one or two (out of six) covariates. Hence most information is still retained in the missing data.

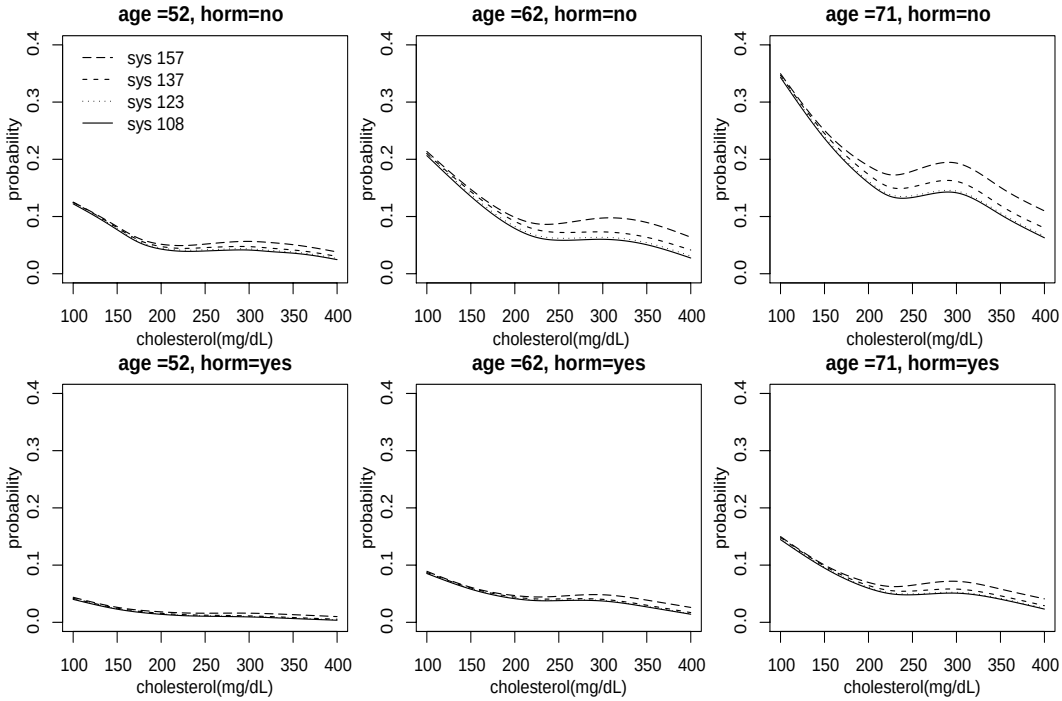

Figure 2.7: Probability curves estimated from the full data analysis. This figure is adapted from Figures 9 and 10 from Lin, Wahba, Xiang, Gao, Klein and Klein (2000)[31]. Each panel plots the estimated probability of pigmentary abnormalities as a function of cholesterol, for four different values of sys. The six panels correspond to different values of age and $horm$, when *drin*=no and *bmi*=27.5 are fixed.

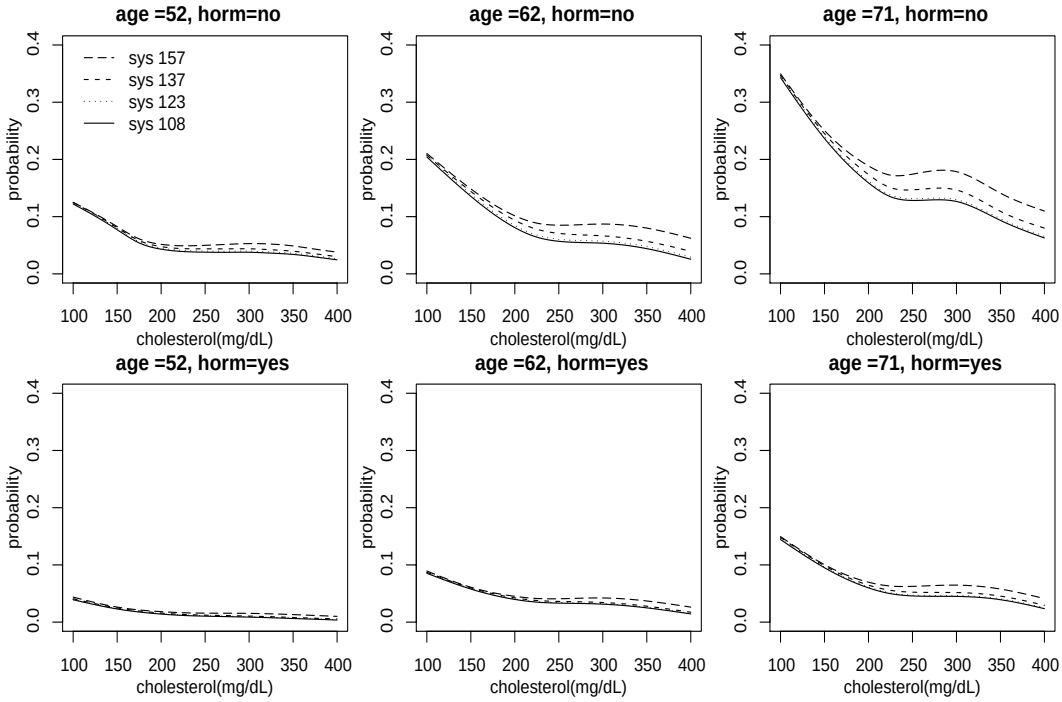

Figure 2.8: Probability curves obtained from QPLE. Each panel plots the estimated probability of pigmentary abnormalities as a function of cholesterol, for four different values of sys. The six panels correspond to different values of age and $horm$, when *drin*=no and *bmi*=27.5 are fixed.

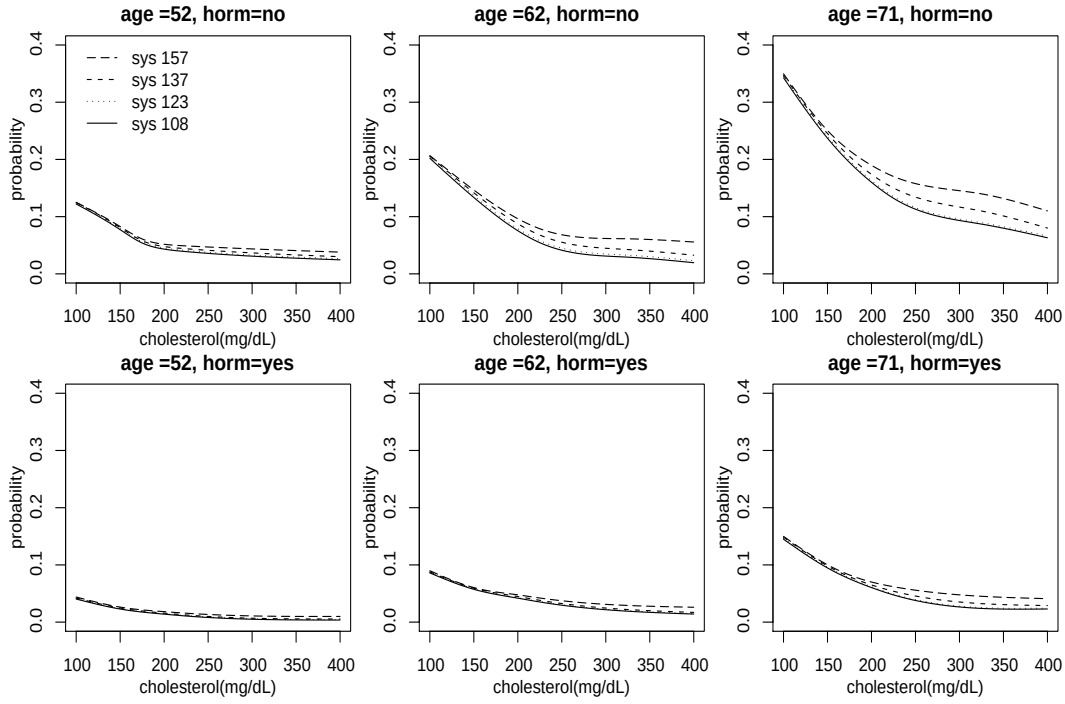

Figure 2.9: Probability curves obtained from the naive method. Each panel plots the estimated probability of pigmentary abnormalities as a function of cholesterol, for four different values of sys. The six panels correspond to different values of age and $horm$, when *drin*=no and *bmi*=27.5 are fixed.

2.9 Concluding remarks

We have presented a direct extension of penalized likelihood regression to the situation when the observed covariates are probability spaces. The regression function is estimated by minimizing a penalized likelihood that incorporates distributional information of the covariates. Numerically, we compute a finite dimensional estimator after approximating the integrals in the likelihood function by quadrature rules. Using the same approximation, GACV and its randomized version have been derived to select the smoothing parameters. Our method is computationally efficient, as it only requires a small number of quadrature nodes to obtain a good estimate. A direct implementation of our method is to handle incomplete covariate data such as covariate measurement error and partially missing covariates. In the examples we have investigated, the resulting estimator substantially outperformed the naive estimator and appeared to be close to the full data analysis.

Chapter 3

Estimating the degrees of freedom for penalized likelihood regression

3.1 Introduction

We are concerned with penalized likelihood regression for data from a non-Gaussian exponential family. Let $(y_i, x_i), i = 1, ..., n$ be n independent observations, where each y_i denotes the response and each x_i denotes the covariate information. In this chapter we assume that $(y_i, x_i), i = 1, ..., n$ are completely observed. The goal is to fit a probability mechanism, assuming that the conditional distribution of y_i given x_i has a density in the exponential family without the nuisance parameters

$$p(y_i|x_i, f) = \exp\{y_i \cdot f(x_i) - b(f(x_i)) + c(y_i)\}, \tag{3.1}$$

where $b(\cdot)$ and $c(\cdot)$ are given functions with $b(\cdot)$ strictly convex. Penalized likelihood regression estimates f in some function space by minimizing a penalized likelihood

$$I_\lambda(f) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, f) + \frac{\lambda}{2} J(f), \tag{3.2}$$

where $J(\cdot)$ is the penalty functional which controls the flexibility of f and λ is the regularization parameter which balances the tradeoff between model fitting and the penalty. In Chapter 2 we have reviewed the important situation when f is estimated in some reproducing kernel Hilbert space (RKHS) and $J(\cdot)$ is a quadratic norm or semi norm penalty. In this case (3.2) is commonly known as penalized likelihood

regression or regression splines with RKHS penalty. Besides regression splines, in this chapter, we also consider a LASSO model, assuming a parametric form of f

$$f(x) = \mu + \sum_{j=1}^{N_B} c_j \beta_j(x), \tag{3.3}$$

where $\beta_j(\cdot), 1 \leq j \leq N_B$ are known basis functions, N_B is the number of the basis functions and $c_j, 1 \leq j \leq N_B$ are the coefficients to be estimated. In this case $J(f)$ stands for the LASSO penalty

$$J(f) = \sum_{j=1}^{N_B} |c_j|. \tag{3.4}$$

Both regression splines and the LASSO require a good choice of the regularization parameter λ. For regression splines with quadratic (semi) norm penalty, Xiang and Wahba (1996)[44] proposed **generalized approximate cross validation** (GACV) which works well for Bernoulli data. Zhang, Wahba, Lin, Voelker, Ferris, Klein and Klein (2004)[47] extended the GACV to the case of Bernoulli data with continuous covariates and the LASSO penalty. The GACV begins with the Kullback-Leibler (KL) distance between the true regression function f^* and the penalized likelihood estimator f_λ

$$\mathrm{KL}(f^*, f_\lambda) = \frac{1}{n} \sum_{i=1}^{n} E_{y_i^0 | x_i, f^*} \left\{ \log \frac{p(y_i^0 | x_i, f^*)}{p(y_i^0 | x_i, f_\lambda)} \right\}, \tag{3.5}$$

where the expectation is taken over $y_i^0 \sim p(y|x_i, f^*)$ independent of y_i. In practice it suffices to work with the comparative KL (CKL) distance

$$\mathrm{CKL}(\lambda) = \frac{1}{n} \sum_{i=1}^{n} \left[-\mu_i^* f_{\lambda i} + b(f_{\lambda i}) \right], \tag{3.6}$$

where $\mu_i^* = E\{y_i^0 | x_i, f^*\}$ denotes the true mean response and $f_{\lambda i} = f_\lambda(x_i)$. Based on a leaving-out-one argument and several first order Taylor expansions, a GACV(λ)

can be derived as a proxy of CKL(λ)

$$\text{GACV}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\frac{\text{tr}(H)\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})}{n - \text{tr}(W^{1/2}HW^{1/2})}, \tag{3.7}$$

where

$$\text{OBS}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}(-y_i f_{\lambda i} + b(f_{\lambda i})) \tag{3.8}$$

denotes the observed log-likelihood, $\mu_{\lambda i} = b'(f_{\lambda i})$ denotes the expected mean response,

$$W = \text{diag}(b''(f_{\lambda 1}), ..., b''(f_{\lambda n})) \tag{3.9}$$

is the diagonal matrix of estimated variances and H is so-called the influence matrix of the penalized likelihood (3.2) (more details can be found in Xiang and Wahba (1996)[44] and Shi, Wahba, Wright, Lee, Klein and Klein (2008)[42]).

Alternatively, one may want to use some model selection criteria to choose λ, such as Akaike information criterion (AIC)(Akaike, 1973[1]) and Bayesian information criterion (BIC) (Schwarz, 1978[40]). In this case the key issue is to estimate the degrees of freedom as a model complexity measure. Stein's unbiased risk estimation (SURE) theory (Stein, 1981[35]) provided a rigorous definition of the degrees of freedom for Gaussian data. Given an arbitrary fitting procedure, let $\hat{\mu}_i^y, 1 \leq i \leq n$ represent the estimated mean responses, where $y = (y_1, ..., y_n)^T$ denotes the vector of responses. It is shown (Efron, 2004[14]) that the Stein's definition of the degrees of freedom is equivalent to

$$\text{df} = \sum_{i=1}^{n} cov(\hat{\mu}_i^y, y_i)/\sigma^2, \tag{3.10}$$

where σ^2 is the common variance. Efron's optimism theory (Efron, 2004[14]) further generalized this result to a $q\ class$ of error measures (Efron, 1986[13]) including the exponential family distributions. Following Efron's optimism theory, the degrees of

freedom can be defined by

$$\text{df} = \sum_{i=1}^{n} cov(\hat{f}_i^y, y_i), \tag{3.11}$$

where $\hat{f}_i^y$ is the estimated natural parameter. Note that

$$\sum_{i=1}^{n} cov(\hat{f}_i^y, y_i) = \sum_{i=1}^{n} E\Big\{ \hat{f}_i^y (y_i - \mu_i) \Big\}, \tag{3.12}$$

where $\mu_i = Ey_i$. Thus (3.11) equals the generalized degrees of freedom (Ye and Wong, 1997[46]), as noted in Lin, Wahba, Xiang, Gao, Klein and Klein (2000)[31]. In this chapter we derive an estimation of the degrees of freedom for penalized likelihood regression, using a first order Taylor expansion. In the LASSO model the degrees of freedom can be estimated by the number of nonzero coefficients. After that we discuss the relationship between the GACV and the degrees of freedom. We show that the GACV provides an alternative estimation of the degrees of freedom.

We will as well be able to generalize our methods to the variable selection problem with multivariate Bernoulli observations. Multiple binary outcomes arises frequently in the many experimental settings, such as surveys, environmental studies and medical researches. Commonly there is a suspicion that these outcomes are correlated. The most general way to account for all possible correlations is assuming a multivariate Bernoulli distribution. Let $Y = (Y_1, ..., Y_K)^T$ be a $K-$dimensional random vector. From Whittaker (1990)[43], the joint density of multivariate Bernoulli distribution can be written as

$$P(Y_1 = y_1, ..., Y_K = y_K) = p(0, 0, ..., 0)^{[\Pi_{j=1}^{K}(1-y_j)]} p(1, 0, ..., 0)^{[y_1 \Pi_{j=2}^{K}(1-y_j)]} \cdots p(1, 1, ..., 1)^{[\Pi_{j=1}^{K} y_j]}, \tag{3.13}$$

where $p(0, ..., 0), ..., p(1, ..., 1)$ are the probabilities for each possible observation of Y. In practice it is more convenient to use the logistic form of (3.13) in statistical model fitting. In this case the model parameters are the log odds ratios which

measure the associations between outcome variables. Gao, Wahba, Klein and Klein (2001)[17] combined a smoothing spline analysis of variance (SS-ANOVA) model and a log-linear model to build a partly flexible model for multivariate Bernoulli data. In this chapter we first derive a truncated log-linear model which ignores the higher order associations. The model parameters can be estimated via the LASSO algorithm. Then we propose an augmented response technique to obtain the GACV. We show that the degrees of freedom of the LASSO can be estimated either by the number of nonzero coefficients or the GACV.

The rest of the chapter is organized as follows. In Section 3.2 we describe how to estimate the degrees of freedom for penalized likelihood regression. In Section 3.3 we study the variable selection problem with multivariate Bernoulli observations. Section 3.4 estimates the degrees of freedom of the LASSO with multivariate Bernoulli observations. Simulation examples can be found in Section 3.5. We present a discussion in Section 3.6.

3.2 Estimating the degrees of freedom for penalized likelihood regression

3.2.1 Estimating the degrees of freedom for penalized likelihood regression

The degrees of freedom defined in (3.11) can be estimated as follows.

$$\begin{aligned}
\text{df} &= \sum_{i=1}^{n} cov(\hat{f}_i^y, y_i) \\
&= \sum_{i=1}^{n} E_{y_{-i}}\Big\{E_{y_i}\big\{\hat{f}_i^y(y_i-\mu_i) \mid y_{-i}\big\}\Big\} \\
&\approx \sum_{i=1}^{n} E_{y_{-i}}\Big\{E_{y_i}\big\{(\frac{\partial \hat{f}_i^y}{\partial y_i}\Big|_{\mu_i}(y_i-\mu_i) + \hat{f}_i^{(y_i=\mu_i, y_{-i})})(y_i-\mu_i) \mid y_{-i}\big\}\Big\} \\
&= \sum_{i=1}^{n} var(y_i) E_{y_{-i}}\Big\{\frac{\partial \hat{f}_i^y}{\partial y_i}\Big|_{\mu_i}\Big\} \\
&\approx \sum_{i=1}^{n} var(y_i) E_y\Big\{\frac{\partial \hat{f}_i^y}{\partial y_i}\Big\}, \qquad (3.14)
\end{aligned}$$

where the first approximation is given by first order Taylor expansion. Given a data set, (3.14) suggests to estimate df by

$$\widehat{\text{df}} = \sum_{i=1}^{n} \widehat{var}(y_i)\frac{\partial \hat{f}_i^y}{\partial y_i}. \qquad (3.15)$$

Now we describe our computational techniques. For penalized likelihood regression (3.2), it is straightforward to show that

$$\begin{aligned}
\widehat{\text{df}}(\lambda) &= \sum_{i=1}^{n} \widehat{var}(y_i)\frac{\partial f_{\lambda i}}{\partial y_i} \\
&= \sum_{i=1}^{n} b''(f_{\lambda i}) h_{ii} \\
&= \text{tr}(WH) \qquad (3.16) \\
&= \text{tr}(W^{1/2} H W^{1/2}), \qquad (3.17)
\end{aligned}$$

where h_{ii} is the ith diagonal element of the influence matrix H. Let $\vec{f_\lambda} = (f_{\lambda 1}, ..., f_{\lambda n})^T$ and $\vec{\mu}_\lambda = (\mu_{\lambda 1}, ..., \mu_{\lambda n})^T$. Then

$$\frac{\partial \vec{\mu}_\lambda}{\partial y} = \frac{\partial \vec{\mu}_\lambda}{\partial \vec{f_\lambda}} \frac{\partial \vec{f_\lambda}}{\partial y} = WH. \tag{3.18}$$

Thus we may conclude that the degrees of freedom can be estimated by the trace of the influence matrix of the mean response.

For penalized likelihood regression with quadratic norm penalty, computation of the influence matrix for large data sets is expensive and may be unstable. Following Lin, Wahba, Xiang, Gao, Klein and Klein (2000)[31], we may produce a randomized estimate of tr$(W^{1/2}HW^{1/2})$ without doing any explicit calculation. More specifically, we first put a small perturbation $\epsilon = (\epsilon_1, ..., \epsilon_n)^T$ on y. Let f_λ^y and $f_\lambda^{y+\epsilon}$ denote the penalized likelihood estimates with respect to y and $y + \epsilon$. And denote $\vec{f}_\lambda^{\,y} = (f_\lambda^y(x_1), ..., f_\lambda^y(x_n))^T$ and $\vec{f}_\lambda^{\,y+\epsilon} = (f_\lambda^{y+\epsilon}(x_1), ..., f_\lambda^{y+\epsilon}(x_n))^T$. It can be shown that

$$\epsilon^T W^{1/2} H W^{1/2} \epsilon \approx \epsilon^T W (\vec{f}_\lambda^{\,y+\epsilon} - \vec{f}_\lambda^{\,y}). \tag{3.19}$$

Therefore a randomized trace estimate of tr$(W^{1/2}HW^{1/2})$ is given by

$$\text{tr}(W^{1/2}HW^{1/2}) \approx \frac{n\epsilon^T W(\vec{f}_\lambda^{\,y+\epsilon} - \vec{f}_\lambda^{\,y})}{\epsilon^T \epsilon}. \tag{3.20}$$

In order to reduce the variance of randomized trace estimate, one may draw R independent perturbations $\epsilon^1, ..., \epsilon^R$ and obtain a $R-$replicated randomized trace estimate

$$\widehat{\text{df}}(\lambda) = \text{tr}(W^{1/2}HW^{1/2}) \approx \frac{1}{R} \sum_{j=1}^{R} \frac{n(\epsilon^j)^T W(\vec{f}_\lambda^{\,y+\epsilon^j} - \vec{f}_\lambda^{\,y})}{(\epsilon^j)^T \epsilon^j}. \tag{3.21}$$

For the penalized likelihood regression with the LASSO penalty, the degrees of freedom can be estimated by the following result:

PROPOSITION 3.1. *Let N denote the number of nonzero coefficients in the LASSO estimate, then*

$$\widehat{df}(\lambda) = tr(W^{1/2}HW^{1/2}) = N. \tag{3.22}$$

Proof of Proposition 3.1. Shi, Wahba, Wright, Lee, Klein and Klein (2008)[42] showed that for penalized likelihood regression with the LASSO penalty,

$$\text{tr}(W^{1/2}HW^{1/2}) = N.$$

The proposition now follows by (3.17).

Proposition 3.1 shows that the degrees of freedom of the LASSO can be estimated by the number of nonzero coefficients. This generalizes the result of Zou, Hastie and Tibshirani (2007)[48] for Gaussian data.

3.2.2 The relationship with the GACV

The GACV is a proxy to the CKL distance. Write

$$\text{CKL}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} f_{\lambda i}(y_i - \mu_i). \tag{3.23}$$

We note the fact

$$E\left\{\sum_{i=1}^{n} f_{\lambda i}(y_i - \mu_i)\right\} = \sum_{i=1}^{n} cov(f_{\lambda i}, y_i) \tag{3.24}$$

which is exactly the degrees of freedom. Therefore analogous to the AIC, the GACV estimates the degrees of freedom by

$$\widehat{\text{df}}(\lambda) = \frac{\text{tr}(H)\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})}{n - \text{tr}(W^{1/2}HW^{1/2})}. \tag{3.25}$$

As in Shi, Wahba, Wright, Lee, Klein and Klein (2008)[42]), we may also derive a BIC-type generalized approximate cross validation (BGACV)

$$\text{BGACV}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\frac{\log n}{2}\frac{\text{tr}(H)\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})}{n - \text{tr}(W^{1/2}HW^{1/2})}. \tag{3.26}$$

Most applications of penalized likelihood regression try to keep

$$\text{tr}(W^{1/2}HW^{1/2}) \ll n. \tag{3.27}$$

For example, in the LASSO estimation, the number of nonzero coefficients is usually expected to be much smaller than the sample size. In this case we observe from (3.25) that

$$\widehat{\text{df}}(\lambda) \approx \text{tr}(H)\frac{\sum_{i=1}^{n} y_i(y_i - \mu_{\lambda i})}{n} \tag{3.28}$$

which is close to $\text{tr}(W^{1/2}HW^{1/2}) = \text{tr}(HW)$.

3.3 Variable selection with multivariate Bernoulli observations

3.3.1 Log-linear models

In this section we are concerned with the variable selection problem with multivariate Bernoulli observations. It can be verified that the logistic model for the joint distribution (3.13) can be written as

$$l(y, \mathbb{f}) = \sum_{j=1}^{K} f^j B_j(y) + \sum_{j<k} f^{jk} B_{jk}(y) + \cdots + f^{12\ldots K} B_{12\ldots K}(y) - b(\mathbb{f}), \tag{3.29}$$

where $\mathbb{f} = (f^1, f^2, ..., f^{12\ldots K})^T$ denotes the vector of natural parameters, $B_{j_1 j_2 \ldots j_r}(y) = y_{j_1} \times y_{j_2} \cdots \times y_{j_r}$ denotes the observed main effects and interactions and

$$b(\mathbb{f}) = \log(1 + \sum_{j} e^{S^j} + \sum_{j<k} e^{S^{jk}} + \cdots + e^{S^{12\ldots K}}) \tag{3.30}$$

with

$$S^{j_1 j_2 \cdots j_r} = \sum_{1 \le s \le r} f^{j_s} + \sum_{1 \le s < t \le r} f^{j_s j_t} + \cdots + f^{j_1 j_2 \cdots j_r}. \tag{3.31}$$

Each parameter $f^{j_1 j_2 \cdots j_r}$ represents a log odds ratio

$$f^{j_1 j_2 \cdots j_r} = \log OR(Y_{j_1}, Y_{j_2}, ..., Y_{j_r} \mid Y_{-(j_1, j_2, ..., j_r)} = 0), \tag{3.32}$$

where Y_{-*} denotes the subset of vector Y except Y_* and (3.32) can be computed recursively by

$$\log OR(Y_1) = \log\{P(Y_1 = 1)\} - \log\{1 - P(Y_1 = 1)\} \tag{3.33}$$

and

$$\begin{aligned} \log OR(Y_1, ..., Y_{(k-1)}, Y_k) = &\log OR(Y_1, ..., Y_{(k-1)} \mid Y_k = 1) \\ &- \log OR(Y_1, ..., Y_{(k-1)} \mid Y_k = 0). \end{aligned} \tag{3.34}$$

Therefore $f^{12}, f^{13}, ..., f^{12...K}$ can be used to measure the association between outcome variables.

Note that it is not practical to estimate all the $2^K - 1$ model parameters when K is large. Since higher order associations are usually of less scientific interest, we may want to "truncate" the full model (3.29) by ignoring the higher order associations. Suppose that we are interested in the associations up to order m, a truncated log-linear model can be obtained by setting the higher order log odds ratios to be zero

$$l_m(y, \mathbb{f}) = \sum_{j=1}^{K} f^j B_j(y) + \sum_{j<k} f^{jk} B_{jk}(y) + \cdots + \sum_{j_1 < j_2 ... < j_m} f^{j_1 j_2 \cdots j_m} B_{j_1 j_2 \cdots j_m}(y) - b(\mathbb{f}), \tag{3.35}$$

where $\mathbb{f} = (f^1, ..., f^{K-m+1...K})^T$ and

$$b(\mathbb{f}) = \log(1 + \sum_j e^{S_m^j} + \sum_{j<k} e^{S_m^{jk}} + \cdots + e^{S_m^{12...K}}) \tag{3.36}$$

with

$$S_m^{j_1 j_2 \cdots j_r} = \begin{cases} S^{j_1 j_2 \cdots j_r}, & \text{if } r \le m \\ \sum_{1 \le s \le r} f^{j_s} + \cdots + \sum_{1 \le t_1 < ... < t_m \le r} f^{j_{t_1} j_{t_2} \cdots j_{t_m}}, & \text{if } r > m. \end{cases} \tag{3.37}$$

When $m = 1$, (3.35) reduces to the main effect model. In this case Y_i's are independent. When $m = K$, (3.35) is the same as the full model (3.29).

Clearly (3.35) has fewer parameters to be estimated. Considering the joint distribution (3.13), we will show that the truncated log-linear model only estimates the lower order probabilities (i.e., $p(\cdots)$ with no more than m 1's in the outcomes). To simplify the notation, let us denote

$$p_0 = P(Y = (0, 0, ..., 0)^T) \tag{3.38}$$

and

$$p_{j_1 j_2 ... j_r} = P(Y_{j_1} = Y_{j_2} = \cdots = Y_{j_r} = 1, Y_{-(j_1, j_2, ..., j_r)} = 0). \tag{3.39}$$

Now we state the following result which describes the joint distribution with respect to the truncated log-linear model.

PROPOSITION 3.2. *Suppose that $Y = (Y_1, ..., Y_K)^T$ is a multivariate Bernoulli random vector having a log density (3.35), then for any $r > m$ we have that*

$$\log(\frac{p_{12...r}}{p_0}) = \sum_{t=0}^{m-1} \left[(-1)^t \binom{r-m+t-1}{t} \times \sum_{1 \le j_1 < ... j_{m-t} \le r} \log(\frac{p_{j_1 ... j_{m-t}}}{p_0}) \right] \tag{3.40}$$

.

Proposition 3.2 can be shown by combining Lemmas 3.3-3.4 below. The proofs of lemmas are given in Appendix A.

LEMMA 3.3. *Under the condition of Proposition 3.2, for any $r \le K$, we have that*

$$\exp\{S_m^{12...r}\} = \frac{p_{12...r}}{p_0}. \tag{3.41}$$

LEMMA 3.4. *Under the condition of Proposition 3.2, if* $m < r \leq K$*, then*

$$S_m^{12\ldots r} = \sum_{t=0}^{m-1} \left[(-1)^t \binom{r-m+t-1}{t} \sum_{1 \leq j_1 < \ldots j_{m-t} \leq r} S^{j_1 \ldots j_{m-t}} \right], \quad (3.42)$$

where $S^{j_1 j_2 \cdots j_q}$ *is defined in (*3.*31).*

The proposition is now proved by combining Lemma 3.3 and Lemma 3.4. Note that Proposition 3.2 is true for any superscript $j_1 j_2 ... j_r$, since we can always permute $Y_{j_1}, Y_{j_2}, ..., Y_{j_r}$ to the first r positions. Proposition 3.2 show that in the joint distribution with respect to a truncated log-linear model, we may treat the lower order probability as free parameters. In this case the higher order probabilities can be determined by (3.40).

3.3.2 The LASSO estimation and the GACV

Suppose that we have n independents observations $(y(i), x(i)), i = 1, ..., n$, where each $y(i) = (y_1(i), ..., y_K(i))^T$ stands for the response vector with K outcomes. For the truncated log-linear model (3.35), a LASSO estimator can be obtained by minimizing a penalized likelihood

$$-\frac{1}{n} \sum_{i=1}^{n} l_m(y(i), \mathbb{f}(x(i))) + \sum_{r=1}^{m} \left[\sum_{j_1 < j_2 \cdots < j_r} \frac{\lambda^{j_1 j_2 \ldots j_r}}{2} \sum_{t=1}^{N_B} |c_t^{j_1 j_2 \ldots j_r}| \right], \quad (3.43)$$

where $\mathbb{f} = (f^1, ..., f^{K-m+1 \ldots K})^T$ and each $f^{j_1 j_2 \ldots j_r}$ has a specific linear model

$$f^{j_1 j_2 \ldots j_r}(x(i)) = \mu^{j_1 j_2 \ldots j_r} + \sum_{t=1}^{N_B} c_t^{j_1 j_2 \ldots j_r} \beta_t(x(i)). \quad (3.44)$$

Here $\beta_t(\cdot), 1 \leq t \leq N_B$ are known basis functions, N_B is the number of the basis functions and $c_t^{j_1 j_2 \ldots j_r}, 1 \leq t \leq N_B$ are the coefficients to be estimated.

Now we derive the GACV for the LASSO model. Let $\mathbb{f}^*$ denote the true model and $\mathbb{f}_\lambda$ denote the LASSO estimator. It is straightforward to show that the CKL distance with respect to $l_m(y, \mathbb{f})$ can be written as

$$\begin{aligned}\text{CKL}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\Big[-\sum_{j=1}^{K} f_\lambda^j(x(i))\mu_j(i) - \sum_{j<k} f_\lambda^{jk}(x(i))\mu_{jk}(i) - \\ \cdots - \sum_{j_1<\cdots<j_m} f_\lambda^{j_1\ldots j_m}(x(i))\mu_{j_1\ldots j_m}(i) + b(\mathbb{f}_\lambda(x(i)))\Big],\end{aligned} \tag{3.45}$$

where $\mu_{j_1\ldots j_r}(i) = E\{B_{j_1\ldots j_r}(Y)|x(i), \mathbb{f}^*\}$ denotes the true mean response. Let us define

$$B(y(i)) = (B_1(y(i)), B_2(y(i)), ..., B_{12}(y(i)), ..., B_{K-m+1\ldots K}(y(i)))^T \tag{3.46}$$

the **augmented response** of $y(i)$ and denote

$$\mu(i) = (\mu_1(i), \mu_2(i), ..., \mu_{12}(i), ..., \mu_{K-m+1\ldots K}(i))^T \tag{3.47}$$

the mean of the augmented response. Then

$$\text{CKL}(\lambda) = \frac{1}{n}\sum_{i=1}^{n}\Big[-\mu(i)^T\mathbb{f}_\lambda(x(i)) + b(\mathbb{f}_\lambda(x(i)))\Big]. \tag{3.48}$$

Hence the leaving-out-one-subject cross validation can be obtained by

$$\begin{aligned}\text{CV}(\lambda) &= \frac{1}{n}\sum_{i=1}^{n}\Big[-B(y(i))^T\mathbb{f}_\lambda^{[-i]}(x(i)) + b(\mathbb{f}_\lambda(x(i)))\Big] \\ &= \text{OBS}(\lambda) + \frac{1}{n}B(y(i))^T(\mathbb{f}_\lambda(x(i)) - \mathbb{f}_\lambda^{[-i]}(x(i))),\end{aligned} \tag{3.49}$$

where $\mathbb{f}_\lambda^{[-i]}$ denotes the estimation with the ith subject left out and

$$\text{OBS}(\lambda) = \frac{1}{n}\Big[-B(y(i))^T\mathbb{f}_\lambda(x(i)) + b(\mathbb{f}_\lambda(x(i)))\Big] \tag{3.50}$$

is the observed log-likelihood. To evaluate $\text{CV}(\lambda)$, We firstly introduce a new version of leaving-out-one-subject lemma with respect to the augmented response (3.46).

LEMMA 3.5. *(Leaving-out-one-subject lemma) For fixed i and a new augmented response $\mathcal{Y}$, let $h_\lambda[i,\mathcal{Y}]$ be the minimizer of*

$$-\sum_{k\neq i} l_m(y(k),\mathbb{f}(x(k))) - \mathcal{Y}^T\mathbb{f}(x(i)) + b(\mathbb{f}(x(i))) + nJ_\lambda(\mathbb{f}). \tag{3.51}$$

Then $h_\lambda\left[i,\mu_\lambda^{[-i]}(i)\right] = \mathbb{f}_\lambda^{[-i]}$. Here $J_\lambda(\mathbb{f})$ is the second term in (3.43) and $\mu_\lambda^{[-i]}(i) = E\{B(Y)|x(i),\mathbb{f}_\lambda^{[-i]}\}$.

Proof See Appendix A. □

Let

$$\vec{\mathbb{f}} = (\mathbb{f}(x(1))^T,\mathbb{f}(x(2))^T,...,\mathbb{f}(x(n))^T)^T \tag{3.52}$$

be the evaluation of $\mathbb{f}$ and

$$\mathbb{c} = (c_1^1,...,c_{N_B}^1,c_1^2,...,c_{N_B}^2,...,c_1^{K-m+1...K},...,c_{N_B}^{K-m+1...K})^T \tag{3.53}$$

be the vector of coefficients. Denote

$$\beta(x(i)) = (\beta_1(x(i)),\beta_2(x(i)),...,\beta_{N_B}(x(i)))^T \tag{3.54}$$

the evaluation of basis functions over $x(i)$. Then

$$D = \begin{pmatrix} \beta(x(1))^T & 0 & \cdots & 0 \\ 0 & \beta(x(1))^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta(x(1))^T \\ \vdots & \vdots & \ddots & \vdots \\ \beta(x(n))^T & 0 & \cdots & 0 \\ 0 & \beta(x(n))^T & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \beta(x(n))^T \end{pmatrix}_{(qn)\times(qN_B)} \tag{3.55}$$

is the grand design matrix, where $q = \begin{pmatrix} K \\ 1 \end{pmatrix} + \begin{pmatrix} K \\ 2 \end{pmatrix} + \cdots + \begin{pmatrix} K \\ m \end{pmatrix}$ is the length of $\mathbb{c}$. Thus we have that $\vec{\mathbb{f}}(x) = D\mathbb{c}$. Denote

$$W_i = var\left(B(y(i)) \,\middle|\, x(i), \mathbb{f}_\lambda\right) \tag{3.56}$$

the estimated covariance matrix for the augmented response. Write

$$W_{(qn)\times(qn)} = \text{diag}(W_1, ..., W_n). \tag{3.57}$$

Define

$$B(y) = (B(y(1))^T, ..., B(y(n))^T)^T \tag{3.58}$$

the $qn \times 1$ vector including all the augmented responses. Then penalized likelihood (3.43) can be written as

$$I_\lambda(B(y), \mathbb{c}) = \frac{1}{n}\Big[-B(y)^T D\mathbb{c} + \sum_{i=1}^{n} b(\mathbb{f}(x(i)))\Big] + \sum_{r=1}^{m}\left[\sum_{1\le j_1<...<j_r\le K} \lambda^{j_1...j_r} \sum_{t=1}^{N_B} |c_t^{j_1...j_r}|\right]. \tag{3.59}$$

Suppose that $\mathbb{c}_\lambda$ is the minimizer of (3.59). Let us put a small perturbation on the augmented response $B(y)+\epsilon$, and let $\mathbb{c}_{\lambda,\epsilon}$ denotes the new minimizer of (3.59). From the KKT condition of (3.59), if ϵ is small enough, zero elements of $\mathbb{c}_\lambda$ will remain to be zero in $\mathbb{c}_{\lambda,\epsilon}$. Suppose there are N nonzero elements in $\mathbb{c}_\lambda$ at the locations of $\{a_1, ..., a_N\}$. Let $\tilde{\mathbb{c}}_\lambda$ and $\tilde{\mathbb{c}}_{\lambda,\epsilon}$ be the sub-vector of $\mathbb{c}_\lambda$ and $\mathbb{c}_{\lambda,\epsilon}$ at these locations. And let $\widetilde{D}_{(qn)\times N}$ be the matrix composed by the $a_1, ..., a_N$th columns of D. Now write $\vec{\mathbb{f}}_\lambda = D\mathbb{c}_\lambda = \widetilde{D}\tilde{\mathbb{c}}_\lambda$ and $\vec{\mathbb{f}}_{\lambda,\epsilon} = D\mathbb{c}_{\lambda,\epsilon} = \widetilde{D}\tilde{\mathbb{c}}_{\lambda,\epsilon}$. We have that

$$\vec{\mathbb{f}}_{\lambda,\epsilon} - \vec{\mathbb{f}}_\lambda = \widetilde{D}(\tilde{\mathbb{c}}_{\lambda,\epsilon} - \tilde{\mathbb{c}}_\lambda). \tag{3.60}$$

From a first order Taylor expansion, we have that

$$\begin{aligned} 0 &= \frac{\partial I_\lambda}{\partial \tilde{\mathbb{c}}}(B(y)+\epsilon, \tilde{\mathbb{c}}_{\lambda,\epsilon}) \\ &\approx \frac{\partial I_\lambda}{\partial \tilde{\mathbb{c}}}(B(y), \tilde{\mathbb{c}}_\lambda) + \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbb{c}} \partial \tilde{\mathbb{c}}^T}(B(y), \tilde{\mathbb{c}}_\lambda)(\tilde{\mathbb{c}}_{\lambda,\epsilon} - \tilde{\mathbb{c}}_\lambda) \\ &\qquad + \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbb{c}} \partial B(y)^T}(B(y), \tilde{\mathbb{c}}_\lambda)(B(y)+\epsilon - B(y)) \\ &= \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbb{c}} \partial \tilde{\mathbb{c}}^T}(B(y), \tilde{\mathbb{c}}_\lambda)(\tilde{\mathbb{c}}_{\lambda,\epsilon} - \tilde{\mathbb{c}}_\lambda) + \frac{\partial^2 I_\lambda}{\partial \tilde{\mathbb{c}} \partial B(y)^T}(B(y), \tilde{\mathbb{c}}_\lambda)\epsilon. \end{aligned} \tag{3.61}$$

The second equality is due to the fact that $\mathbb{c}_\lambda$ is the solution for $B(y)$. Direct calculation yields

$$\frac{\partial^2 I_\lambda}{\partial \tilde{\mathbb{c}}\ \partial \tilde{\mathbb{c}}^T}(B(y), \tilde{\mathbb{c}}_\lambda) = \widetilde{D}^T W \widetilde{D}. \tag{3.62}$$

and

$$\frac{\partial^2 I_\lambda}{\partial \tilde{\mathbb{c}}\ \partial B(y)^T}(B(y), \tilde{\mathbb{c}}_\lambda) = -\widetilde{D}^T. \tag{3.63}$$

Plug (3.62) and (3.63) into (3.61) and note that $\widetilde{D}^T W \widetilde{D}$ is always invertible, we have that

$$\tilde{\mathbb{c}}_{\lambda,\epsilon} - \tilde{\mathbb{c}}_\lambda \approx (\widetilde{D}^T W \widetilde{D})^{-1} \widetilde{D}^T \epsilon. \tag{3.64}$$

Therefore (3.60) implies that

$$\vec{\mathbb{f}}_{\lambda,\epsilon} - \vec{\mathbb{f}}_\lambda \approx H\epsilon, \tag{3.65}$$

where

$$H = \widetilde{D}(\widetilde{D}^T W \widetilde{D})^{-1} \widetilde{D}^T \tag{3.66}$$

is the influence matrix for the penalized likelihood (3.59). Let $\epsilon \to 0$, we have that

$$\frac{\partial \vec{\mathbb{f}}_\lambda}{\partial B(y)} = H. \tag{3.67}$$

Write

$$H = \begin{pmatrix} H_{11} & * & * & * \\ * & H_{22} & * & * \\ * & * & \ddots & * \\ * & * & * & H_{nn} \end{pmatrix}_{qn \times qn} \tag{3.68}$$

where H_{ii} is the $q \times q$ submatrix on the diagonal with respect to $B(y(i))$.

It can be seen from Lemma 3.5 that

$$\mathbb{f}_\lambda^{[-i]}(x(i)) - \mathbb{f}_\lambda(x(i)) \approx H_{ii}(\mu_\lambda^{[-i]}(i) - B(y(i))). \tag{3.69}$$

In addition it is easy to verified that

$$\mu_\lambda^{[-i]}(i) - \mu_\lambda(i) \approx W_i(\mathbb{f}_\lambda^{[-i]}(x(i)) - \mathbb{f}_\lambda(x(i))). \tag{3.70}$$

Combining (3.69) and (3.70) we have that

$$\mathbb{f}_\lambda(x(i)) - \mathbb{f}_\lambda^{[-i]}(x(i)) \approx (I - H_{ii}W_i)^{-1}H_{ii}(B(y(i)) - \mu_\lambda(i)). \tag{3.71}$$

Plug (3.71) into (3.49), we obtain the approximate cross validation (ACV)

$$\text{ACV}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} B(y(i))^T(I - H_{ii}W_i)^{-1}H_{ii}(B(y(i)) - \mu_\lambda(i)). \tag{3.72}$$

A generalized version of (3.72) can be obtained if we replace H_{ii} and $(I - H_{ii}W_i)$ by their average matrices. In this chapter we use the generalized average of submatrices introduced in Gao, Wahba, Klein and Klein (2001)[17]. Suppose we have n matrices $A_i = \{a_{i,st}\}_{q\times q}$, then the average of diagonal and off diagonal elements can be computed by

$$\delta = \frac{1}{nq}\sum_{i=1}^{n} tr(A_i), \quad \gamma = \frac{1}{nq(q-1)}\sum_{i=1}^{n}\sum_{k_1 \neq k_2} a_{i,k_1k_2}. \tag{3.73}$$

Thus the average matrix is obtained by

$$\bar{A} = (\delta - \gamma) I_{q\times q} + \gamma \cdot ee^T = \begin{pmatrix} \delta & \gamma & \cdots & \gamma \\ \gamma & \delta & \cdots & \gamma \\ \vdots & \vdots & \ddots & \vdots \\ \gamma & \gamma & \cdots & \delta \end{pmatrix}. \tag{3.74}$$

And its inverse can be computed by

$$\bar{A}^{-1} = \frac{1}{\delta - \gamma} I_{q\times q} - \frac{\gamma}{(\delta - \gamma)(\delta + (q-1)\gamma)} ee^T, \tag{3.75}$$

where $e = (1, 1..., 1)^T$ is the unit vector of length q.

Let $Q_i = I - H_{ii}W_i$, then the GACV can be written as

$$\text{GACV}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} B(y(i))^T \bar{Q}^{-1}\bar{H}(B(y(i)) - \mu_\lambda(i)), \tag{3.76}$$

where $\bar{Q}$ and $\bar{H}$ are the generalized averages of Q_i and H_{ii}.

3.4 Estimating the degrees of freedom of the LASSO with multivariate Bernoulli observations

In the framework of Efron's optimism theory, the degrees of freedom can be defined by

$$\text{df}(\lambda) = \sum_{i=1}^{n}\sum_{r=1}^{m}\sum_{j_1<\cdots<j_r} cov(f_\lambda^{j_1\ldots j_r}(x(i)), B_{j_1\ldots j_r}(y(i))), \tag{3.77}$$

Using a similar approximation in Section 3.2.1, (3.77) can be estimated by

$$\widehat{\text{df}}(\lambda) = \sum_{i=1}^{n}\sum_{r=1}^{m}\sum_{j_1<\cdots<j_r} \widehat{var}(B_{j_1\ldots j_r}(y(i)))\frac{\partial f_\lambda^{j_1\ldots j_r}(x(i))}{\partial B_{j_1\ldots j_r}(y(i))}. \tag{3.78}$$

From (3.56), (3.57) and (3.67), we can show that

$$\begin{aligned}\widehat{\text{df}}(\lambda) &= \text{tr}(WH) \\ &= \text{tr}(W\widetilde{D}(\widetilde{D}^T W\widetilde{D})^{-1}\widetilde{D}^T) \\ &= \text{tr}(\widetilde{D}^T W\widetilde{D}(\widetilde{D}^T W\widetilde{D})^{-1}) \\ &= N. \end{aligned} \tag{3.79}$$

where N is the number of nonzero coefficients in the LASSO estimates. Therefore we have the following result:

PROPOSITION 3.6. *Let N denote the number of nonzero coefficients in the LASSO estimates, then*

$$\widehat{df}(\lambda) = tr(WH) = N. \tag{3.80}$$

Now we look at the GACV estimation of $\text{df}(\lambda)$. Firstly the CKL distance can be written as

$$\text{CKL}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{f}_\lambda(x(i))^T(B(y(i)) - \mu(i)). \tag{3.81}$$

Note that

$$\begin{aligned} E\Big\{\sum_{i=1}^{n} \mathbb{f}_\lambda(x(i))^T(B(y(i)) - \mu(i))\Big\} &= \sum_{i=1}^{n}\sum_{r=1}^{m}\sum_{j_1<\cdots<j_r} cov(f_\lambda^{j_1\ldots j_r}(x(i)), B_{j_1\ldots j_r}(y(i))) \\ &= \text{df}(\lambda). \end{aligned} \tag{3.82}$$

Thus another estimation of $\text{df}(\lambda)$ can be obtained from the GACV

$$\widehat{\text{df}}(\lambda) = \sum_{i=1}^{n} B(y(i))^T\bar{Q}^{-1}\bar{H}(B(y(i)) - \mu_\lambda(i)). \tag{3.83}$$

Therefore the BGACV can be defined by

$$\text{BGACV}(\lambda) = \text{OBS}(\lambda) + \frac{1}{n}\frac{\log n}{2}\sum_{i=1}^{n} B(y(i))^T \bar{Q}^{-1}\bar{H}(B(y(i)) - \mu_\lambda(i)). \quad (3.84)$$

3.5 Numerical studies

We have presented two estimations of the degrees of freedom, which result in four tuning methods to select the regularization parameters:

(a) GACV;

(b) BGACV;

(c) AIC $= -\frac{1}{n}\sum_{i=1}^{n} \log p(y(i)|x(i), \mathbb{f}_\lambda) + \frac{1}{n}\text{tr}(WH)$;

(d) BIC $= -\frac{1}{n}\sum_{i=1}^{n} \log p(y(i)|x(i), \mathbb{f}_\lambda) + \frac{1}{n}\frac{\log n}{2}\text{tr}(WH)$.

In this section we illustrate these methods by two simulated examples of bivariate Bernoulli observations. All the simulations are conducted using Matlab 7.9.0529 installed in Red Hat Enterprise Linux 5.

3.5.1 Simulation settings

Consider the bivariate Bernoulli distribution:

$$p(y|x, \mathbb{f}) = \exp\{f^1(x)y_1 + f^2(x)y_2 + f^{12}(x)y_1y_2 - b(\mathbb{f}(x))\},$$

where

$$b(\mathbb{f}(x)) = \log(1 + \exp\{f^1(x)\} + \exp\{f^2(x)\} + \exp\{f^1(x) + f^2(x) + f^{12}(x)\})$$

and $x = (x_1, x_2, ..., x_{25})^T$ is composed of 25 covariates. We consider the following two examples

$$\text{(i)} \begin{cases} f^1(x) = -3 + 2x_1 + 2x_2 \\ f^2(x) = -3 + 2x_3 + 2x_4 \\ f^{12}(x) = -3 + 2x_5 + 2x_6 \end{cases}$$

$$\text{(ii)} \begin{cases} f^1(x) = -3 + 2x_1 + 2x_2 \\ f^2(x) = -4 + 1.5x_3 + 1.5x_4 + 1.5x_7 + 1.5x_8 \\ f^{12}(x) = -3 + 2x_2 + 2x_3 \end{cases}$$

In each case we assume $X_j, j = 1, ..., 25$ to be iid random variables distributed as Bernoulli(0.5) and generated a sample of $n = 500$ observations from the distribution of (X, Y). For each sample generated, the LASSO estimation is conducted as follows. We first assume that

$$f^r(x) = \mu^r + \sum_{t=1}^{25} c_t^r x_t, \quad r = 1, 2, 12. \tag{3.85}$$

Then estimate c_t^r's by minimizing the following penalized likelihood

$$-\frac{1}{n}\sum_{i=1}^{n} \log p(y(i)|x(i), \mathbb{f}) + \frac{\lambda^1}{2}\sum_{r=1}^{2}\sum_{t=1}^{25} |c_t^r| + \frac{\lambda^{12}}{2}\sum_{t=1}^{25} |c_t^{12}| \tag{3.86}$$

Note that in (3.86) we used the same λ^1 to penalize the main effects f^1 and f^2.

3.5.2 Results

Table 3.1 and 3.2 summarize our simulation results. The true regression functions (with constant omitted) are shown at the top of each table. The numbers show the relative frequencies of each covariate being captured by, respectively, GACV, BGACV, AIC and BIC, out of 100 repeated simulations. The last column presents the average number of false covariates captured in the 100 simulations.

f^1	$2x_1$	$2x_2$					$x_k, k \neq 1, 2$
f^2			$2x_3$	$2x_4$			$x_k, k \neq 3, 4$
f^{12}					$2x_5$	$2x_6$	$x_k, k \neq 5, 6$
GACV	0.98	0.99	0.97	0.99	1.00	1.00	8.07
BGACV	0.86	0.88	0.87	0.87	0.63	0.63	3.91
AIC	0.99	0.99	0.99	0.99	1.00	1.00	7.31
BIC	0.98	0.98	0.97	0.96	1.00	1.00	3.20

Table 3.1: The relative frequency of each covariate being captured by, respectively, GACV, BGACV, AIC and BIC, out of the 100 simulations. The true regression functions (with constant omitted) are shown at the top of the table. The last column presents the average number of false covariates captured in the 100 simulations.

f^1	$2x_1$	$2x_2$							$x_k, k \neq 1, 2$
f^2			$1.5x_3$	$1.5x_4$	$1.5x_7$	$1.5x_8$			$x_k, k \neq 3, 4, 7, 8$
f^{12}							$2x_2$	$2x_3$	$x_k, k \neq 2, 3$
GACV	0.94	0.98	0.48	0.43	0.50	0.46	1.00	1.00	3.55
BGACV	0.45	0.58	0.24	0.10	0.14	0.11	0.96	0.93	1.43
AIC	0.93	0.98	0.46	0.44	0.45	0.49	1.00	1.00	3.88
BIC	0.83	0.93	0.41	0.32	0.32	0.36	1.00	1.00	2.45

Table 3.2: The relative frequency of each covariate being captured by, respectively, GACV, BGACV, AIC and BIC, out of the 100 simulations. The true regression functions (with constant omitted) are shown at the top of the table. The last column presents the average number of false covariates captured in the 100 simulations.

Generally speaking, BGACV and BIC assign a large penalty on the degrees of freedom, leading to a sparse estimation. This property is usually desirable when the true model is of low dimension while the sample size is large. As a result, in both examples, BGACV and BIC outperformed GACV and AIC with much fewer false covariates captured. On the other hand, there is no significant difference between two estimations of the degrees of freedom. We can see that BIC slightly outperformed BGACV on Table 1 but GACV slightly outperformed AIC on Table 2.

3.6 Discussion

In this chapter we have shown that the degrees of freedom for penalized likelihood regression can be estimated by: (1) the trace of the influence matrix of the mean; and (2) the GACV. The two estimations can be extended to the variable selection problem with multivariate Bernoulli observations. In this case the trace of the influence matrix of the mean is equal to the number of nonzero coefficients. In our numerical studies the two estimations of the degrees of freedom are comparable. In some simulations AIC (or BIC) outperformed GACV (or BGACV) while in other simulations we observed the opposite.

The techniques developed in this chapter can be extended to other situations. For example, we may want to estimate the degrees of freedom for data with missing covariates or covariate measurement error. Though the GACV estimation is immediately available from Chapter 2, we still need to do a lot of work to fully understand the degrees of freedom in the presence of incomplete data. Motivated from the real data sets, we also want to extend our methods to deal with the variable selection problem with multiple continuous outcomes. These could be interesting topics for future research.

Chapter 4

Concluding remarks

Penalized likelihood regression with RKHS penalty is widely used as a powerful non or semi parametric regression tool in data analysis. In this thesis we present an important extension to randomized covariate data. Though the term "randomized covariate" has little practical meaning, it provides us a theoretical foundation in the incomplete data analysis. In Chapter 2 we have shown the existence of the penalized likelihood estimate for randomized covariate data in the general smoothing spline setup. This result can be treated as a necessary condition for the penalized likelihood regression with incomplete covariate data since any other type of incomplete covariate can be treated as a special case of randomized covariate. In order to minimize the penalized likelihood, we suggest to use an EM algorithm based on quadrature rules. This implementation of the EM algorithm is computationally friendly as it does not require a large number of quadrature nodes to get a good estimate. From our experience a quadrature rule with 7 to 12 nodes for each incomplete covariate usually yields a very good approximation. In this case, the EM algorithm usually converges very rapidly.

In this thesis our discussion of missing covariate data is limited to the missing mechanism of missing at random. In practical applications, however, it is more likely to face a non-ignorable missing data mechanism. In this case serious biases in the estimation may result if we do not model the missing data mechanism. Thus we need

to extend our methods to the non-ignorable missing data mechanism. A straightforward solution is to specify a parametric model for the missing data mechanism and incorporate it into the missing data penalized likelihood. This is an interesting topic for future research.

Our work on the degrees of freedom of penalized likelihood regression can be extended in several directions. One interesting direction is to estimate the degrees of freedom for data with missing covariates or covariate measurement error. Though the GACV estimation is immediately available from Chapter 2, we still need to do a lot of work to fully understand the degrees of freedom in the presence of incomplete data. Another interesting future direction is to deal with the variable selection problem with multiple continuous outcomes, which is motivated from the real data sets.

Appendix A: Technical proofs

Proof of Proposition 2.1. Any linear combination of measurable functions is still measurable. Therefore it suffices to prove that $\mathcal{H}_B$ is complete. Let $f_1, f_2, ...$ be a Cauchy sequence in $\mathcal{H}_B$ and f^* be its limit in $\mathcal{H}$. Then $f_1, f_2, ...$ converge pointwise to f^*. Note that the pointwise limit of measurable functions is still a measurable function. Therefore $f^* \in \mathcal{H}_B$. □

To simply the notation in the proofs of Lemma 2.4-2.6, let's define

$$l_i(t) = y_i \cdot t - b(t) + c(y_i) \tag{A.1}$$

the log-density as a function of the natural parameter. Then $l_i(t)$ is strictly concave and bounded from above. Therefore there are three possible cases of the limit of $l_i(t)$:

$$\text{(1)} \quad \lim_{t\to-\infty} l_i(t) = \bar{l}_i \text{ and } \lim_{t\to+\infty} l_i(t) = -\infty; \tag{A.2}$$

$$\text{(2)} \quad \lim_{t\to-\infty} l_i(t) = -\infty \text{ and } \lim_{t\to+\infty} l_i(t) = \bar{l}_i; \tag{A.3}$$

$$\text{(3)} \quad \lim_{t\to-\infty} l_i(t) = -\infty \text{ and } \lim_{t\to+\infty} l_i(t) = -\infty \tag{A.4}$$

where $\bar{l}_i = \sup_t l_i(t) < \infty$.

Proof of Lemma 2.4. Without loss of generality, we suppose that A.1 is satisfied with the first m cases (hence they are completely observed). In order to show Lemma 2.4, we first prove that under A.1, $-\sum_{i=1}^{m} \log p(y_i|x_i, f)$ is positively coercive over $\mathcal{H}_0$. Suppose to the contrary that this is not true. Then there exists a constant $U > 0$ and a sequence $\{g_k\}_{k\in\mathbb{N}} \subseteq \mathcal{H}_0$ with $||g_k||_{\mathcal{H}} = 1$ such that

$$-\sum_{i=1}^{m} l_i(k \cdot g_k(x_i)) \leq U, \quad k \in \mathbb{N}. \tag{A.5}$$

Since the unit sphere $\{g \in \mathcal{H}_0 \ : \ ||g||_{\mathcal{H}} = 1\}$ is sequence compact, there exists a subsequence $\{g_{k_j}\}_{j\in\mathbb{N}}$ converging to some g^* with $||g^*||_{\mathcal{H}} = 1$. We claim that

$$g^*(x_i) \begin{cases} \leq 0, & \text{if } i \text{ belongs to Case 1 as (A.2)} \\ \geq 0, & \text{if } i \text{ belongs to Case 2 as (A.3)} \\ = 0, & \text{if } i \text{ belongs to Case 3 as (A.4).} \end{cases} \tag{A.6}$$

Suppose to the contrary that (A.6) is not true. If i belong to case (1), then $g^*(x_i) = a > 0$. Since $\{g_{k_j}\}_{j\in\mathbb{N}}$ converges to g^*, there exists $N > 0$ such that

$$g_{k_j}(x_i) \geq a/2, \quad \text{for all } j > N. \tag{A.7}$$

From (A.5), we have

$$l_i(k_j \cdot g_{k_j}(x_i)) \geq -U - \sum_{s\neq i} \overline{l}_s > -\infty, \quad j \in \mathbb{N}. \tag{A.8}$$

This is a contradiction of (A.2) since when $j > N$

$$k_j \cdot g_{k_j}(x_i) \geq k_j \cdot a/2 \to +\infty. \tag{A.9}$$

Similar contradiction can be observed when i belongs to case (2) or case (3). Therefore the claim in Equation (A.6) follows.

Now let g_0 be the unique maximizer of $\sum_{i=1}^m l_i(g(x_i))$ in $\mathcal{H}_0$. Consider $g_0 + rg^*$ with $r > 0$. Combining (A.2)–(A.4) and (A.6), we can see that

$$\sum_{i=1}^m l_i(g_0(x_i) + rg^*(x_i)) \geq \sum_{i=1}^m l_i(g_0(x_i)), \ \ \forall r > 0. \tag{A.10}$$

But this is a contradiction. Hence $-\sum_{i=1}^m \log p(y_i|x_i, f)$ is positively coercive over $\mathcal{H}_0$, which means that

$$||g||_{\mathcal{H}} \to \infty \Rightarrow -\sum_{i=1}^m l_i(g(x_i)) \to +\infty, \quad g \in \mathcal{H}_0. \tag{A.11}$$

Since $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$ where $\mathcal{H}_1$ denotes the subspace of smooth functions, we have the orthogonal decomposition $f = g + h$ where $g \in \mathcal{H}_0 \bigcap \mathcal{H}_B$ and $h \in \mathcal{H}_1 \bigcap \mathcal{H}_B$. The Lemma can be proved in steps.

(i) $||h||_{\mathcal{H}} \to +\infty$. In this case

$$I_\lambda^R(f) \geq -\frac{1}{n}\sum_{i=1}^{n} \bar{l}_i + \frac{1}{2}\lambda||h||_{\mathcal{H}} \to +\infty. \tag{A.12}$$

(ii) $||h||_{\mathcal{H}} \leq U$ for some $U > 0$ but $||g||_{\mathcal{H}} \to +\infty$. In this case

$$|h(x_i)| = |\langle h, K(\cdot, x_i)\rangle| \leq ||h||_{\mathcal{H}} K^{1/2}(x_i, x_i) \leq U \cdot K^{1/2}(x_i, x_i), \;\; i = 1, 2, ...m$$

which implies that

$$f(x_i) = g(x_i) + h(x_i) = g(x_i) + O(1), \;\; i = 1, ..., m, \; ||h||_{\mathcal{H}} \leq U.$$

Let $||g||_{\mathcal{H}} \to \infty$, we have

$$\begin{aligned} I_\lambda^R(f) &\geq -\frac{1}{n}\sum_{i=1}^{n} \log \int_{\mathcal{X}_i} p(y_i|x_i, f) dP_i \\ &\geq -\frac{1}{n}\sum_{i=1}^{m} l_i(g(x_i) + h(x_i)) - \frac{1}{n}\sum_{j=m+1}^{n} \bar{l}_j \\ &= -\frac{1}{n}\sum_{i=1}^{m} l_i(g(x_i) + O(1)) - \frac{1}{n}\sum_{j=m+1}^{n} \bar{l}_j \\ &\to +\infty \end{aligned} \tag{A.13}$$

where (A.13) follows from the claim in Equation (A.11).

The Lemma is now proved by combining (i) and (ii). □

Proof of Lemma 2.5. Let $\{f_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}_B$ which converges weakly to f^*. It is easy to see that $\{f_k\}_{k\in\mathbb{N}}$ also pointwise converges to f^*. Since pointwise limit of measurable functions is still a measurable function, $f^* \in \mathcal{H}_B$. From the

continuity of $l_i(t)$, $\{e^{l_i(f_k(x_i))}\}_{k\in\mathbb{N}}$ pointwise converges to $e^{l_i(f^*(x_i))}$ over $\mathcal{X}_i$. Note that $e^{l_i(f_k(x_i))} \leq e^{\bar{l}_i}$ and every constant is integrable with respect to $(\mathcal{X}_i, \mathcal{F}_i, P_i)$. By the Dominated Convergence Theorem, we have that

$$\lim_{k\to\infty} \int_{\mathcal{X}_i} e^{l_i(f_k(x_i))} dP_i = \int_{\mathcal{X}_i} e^{l_i(f^*(x_i))} dP_i. \tag{A.14}$$

The Lemma now follows since $\log(\cdot)$ is continuous. □

Proof of Lemma 2.6. Let $\{f_k\}_{k\in\mathbb{N}}$ be a sequence in $\mathcal{H}_B$ which weakly converges to f^*. Consider the orthogonal decomposition of each f_k by $f_k = g_k + h_k$ with $g_k \in \mathcal{H}_0 \bigcap \mathcal{H}_B$ and $h_k \in \mathcal{H}_1 \bigcap \mathcal{H}_B$. It is straightforward to see that $\{h_k\}_{k\in\mathbb{N}}$ weakly converges to h^*, the smooth part of f^*. Therefore we can write

$$0 \leq ||h_k - h^*||^2_{\mathcal{H}} = ||h_k||^2_{\mathcal{H}} + ||h^*||^2_{\mathcal{H}} - 2\langle h_k, h^*\rangle. \tag{A.15}$$

Let $k \to \infty$, we observe that

$$0 \leq \lim_k \inf ||h_k||^2_{\mathcal{H}} - ||h^*||^2_{\mathcal{H}} \tag{A.16}$$

and the Lemma is proved by definition. □

Proof of Theorem 2.7. For any fixed $\theta \in \Theta$, by Theorem 2.2, $I^E_\lambda(f, \theta)$ is minimizable in $\mathcal{H}$. Let

$$T(\theta) \triangleq \min_{f\in\mathcal{H}} I^E_\lambda(f, \theta) \tag{A.17}$$

denote the minimum penalized likelihood given θ. We claim that $T(\theta)$ is continuous.

For any sequence $\{\theta_k\}_{k\in\mathbb{N}} \in \Theta$ that converges to θ^*, let P_{θ_k} and P_{θ^*} denote the probability measures on $\mathbb{R}^d$ with density functions $p(u|\theta_k)$ and $p(u|\theta^*)$. Since $F(u|\theta_k) \to F(u|\theta^*)$ for any $u \in \mathbb{R}^d$, P_{θ_k} weakly converges to P_{θ^*}. Note that, for

any fixed $f \in \mathcal{H}$, $G(u) \triangleq p(y_i|x_i^{err} - u, f)$ is a real-valued, continuous and bounded function on $\mathbb{R}^d$. Thus $\int G(u)dP_{\theta_k} \to \int G(u)dP_{\theta^*}$. Equivalently, that is

$$\int_{\mathbb{R}^d} p(y_i|x_i^{err} - u_i, f)p(u_i|\theta_k)du_i \to \int_{\mathbb{R}^d} p(y_i|x_i^{err} - u_i, f)p(u_i|\theta^*)du_i \qquad \text{(A.18)}$$

which implies that $I_\lambda^E(f, \theta)$ is continuous in θ for any fixed f. This is sufficient to prove the continuity of $T(\theta)$. The theorem now follows from the compactness of Θ. $\square$.

Proof of Theorem 2.8. For any fixed $\theta \in \Theta$, by (2.53) and Theorem 2.2, $I_\lambda^M(f, \theta)$ is minimizable in $\mathcal{H}$. Thus, we can define

$$T(\theta) \triangleq \min_{f \in \mathcal{H}} I_\lambda^M(f, \theta). \qquad \text{(A.19)}$$

We claim that $T(\theta)$ is continuous.

By Assumption M.1 and M.2, there exists $U > 0$ such that $p(x_i|\theta) < U$ for all $x_i^{mis} \in \mathcal{D}_i^\theta$, $\theta \in \Theta$ and $1 \leq i \leq n$. Now for any sequence $\{\theta_k\}_{k\in\mathbb{N}} \in \Theta$ that converges to θ^*, $p(y_i|x_i, f)p(x_i|\theta_k)$ pointwise converges to $p(y_i|x_i, f)p(x_i|\theta^*)$. Note that $p(y_i|x_i, f)p(x_i|\theta_k) \leq e^{\bar{l}_i} \cdot U$ and any constant is integrable on the compact domain $\mathcal{D}_i^\theta$. By Dominated Convergence Theorem, we conclude that

$$\lim_{k\to\infty} \int_{\mathcal{D}_i^\theta} p(y_i|x_i, f)p(x_i|\theta_k)dx_i^{mis} = \int_{\mathcal{D}_i^\theta} p(y_i|x_i, f)p(x_i|\theta^*)dx_i^{mis} \qquad \text{(A.20)}$$

which implies that $I_\lambda^M(f, \theta)$ is continuous in θ for any fixed f. This is sufficient to prove the continuity of $T(\theta)$. The theorem now follows from the compactness of Θ. $\square$

Proof of Lemma 3.3. It is straightforward to show that

$$P(Y = (0, 0, ..., 0)^T) = \exp\{-b(\mathbb{f}_m)\}. \qquad \text{(A.21)}$$

By (3.35), the right hand side of (3.41) is

$$\begin{aligned} & \exp\{\sum_{j=1}^{r} f^{j} + \sum_{1\le j<k\le r} f^{jk} + \cdots \sum_{1\le j_1<\cdots<j_{\min(m,r)}\le r} f^{j_1\cdots j_{\min(m,r)}} - b(\mathbb{f}_m)\} \\ & \qquad\qquad \times \exp\{b(\mathbb{f}_m)\} \\ = \; & \exp\{\sum_{j=1}^{r} f^{j} + \sum_{1\le j<k\le r} f^{jk} + \cdots \sum_{1\le j_1<\cdots<j_{\min(m,r)}\le r} f^{j_1\cdots j_{\min(m,r)}}\} \\ = \; & \exp\{S_m^{12\cdots r}\} \;\; \square \end{aligned}$$

Proof of Lemma 3.4. The lemma can be proved by induction. Obviously, (3.42) holds for $m = 1$. Suppose that (3.42) holds for $m = k < r - 1$. For $m = k + 1$, we have that

$$\begin{aligned} S_{k+1}^{12\cdots r} &= \sum_{1\le j_1<\cdots<j_{k+1}\le r} f^{j_1\cdots j_{k+1}} + S_k^{12\cdots r} \\ &= \sum_{1\le j_1<\cdots<j_{k+1}\le r} (S^{j_1\cdots j_{k+1}} - S_k^{j_1\cdots j_{k+1}}) + S_k^{12\cdots r} \\ &= \sum_{1\le j_1<\cdots<j_{k+1}\le r} S^{j_1\cdots j_{k+1}} - \left[\sum_{1\le j_1<\cdots<j_{k+1}\le r} S_k^{j_1\cdots j_{k+1}} - S_k^{12\cdots r}\right] \end{aligned} \quad \text{(A.22)}$$

Note that (3.42) is true for $m = k$, so we observe that

$$\begin{aligned}
&\sum_{1\le j_1<\cdots<j_{k+1}\le r} S_k^{j_1\cdots j_{k+1}} - S_k^{12\cdots r} \\
= &\sum_{1\le j_1<\cdots<j_{k+1}\le r} \sum_{t=0}^{k-1}(-1)^t \binom{t}{t} \sum_{1\le h_1<\cdots<h_{k-t}\le k+1} S^{j_{h_1}\cdots j_{h_{k-t}}} \\
&\qquad - \sum_{t=0}^{k-1}(-1)^t \binom{r-k-1+t}{t} \sum_{1\le j_1<\cdots<j_{k-t}\le r} S^{j_1\cdots j_{k-t}} \\
= &\sum_{t=0}^{k-1}(-1)^t \Big[\sum_{1\le j_1<\cdots<j_{k+1}\le r} \sum_{1\le h_1<\cdots<h_{k-t}\le k+1} S^{j_{h_1}\cdots j_{h_{k-t}}} \\
&\qquad - \binom{r-k-1+t}{t} \sum_{1\le j_1<\cdots<j_{k-t}\le r} S^{j_1\cdots j_{k-t}} \Big] \\
= &\sum_{t=0}^{k-1}(-1)^t \Big[\binom{r-(k-t)}{k+1-(k-t)} \sum_{1\le j_1<\cdots<j_{k-t}\le r} S^{j_1\cdots j_{k-t}} \\
&\qquad - \binom{r-k-1+t}{t} \sum_{1\le j_1<\cdots<j_{k-t}\le r} S^{j_1\cdots j_{k-t}} \Big] \\
= &\sum_{t=0}^{k-1} \Big[(-1)^t \binom{r-k+t-1}{t+1} \sum_{1\le j_1<\cdots<j_{k-t}\le r} S^{j_1\cdots j_{k-t}} \Big] \qquad \text{(A.23)}
\end{aligned}$$

The last equality is due to the fact that

$$\binom{n}{m} = \binom{n}{m-1} + \binom{n-1}{m-1} \quad \forall\, n > m \qquad \text{(A.24)}$$

Now plug (A.23) into (A.22), we observe

$$S_{k+1}^{12\ldots r} = \sum_{s=0}^{k} \left[(-1)^s \binom{r-k+s-2}{s} \sum_{1\le j_1<\ldots<j_{k-s}\le r} S^{j_1\ldots j_{k-s}} \right] \qquad \text{(A.25)}$$

Hence (3.42) holds for $m = k+1$ and the lemma is proved by induction. $\square$

Proof of Lemma 3.5. For any $\mathbb{f}$ other than $\mathbb{f}_\lambda^{[-i]}$, it is straightforward to verify that

$$-\mu_\lambda^{[-i]}(i)^T \mathbb{f}_\lambda^{[-i]}(x(i)) + b(\mathbb{f}_\lambda^{[-i]}(x(i))) \leq -\mu_\lambda^{[-i]}(i)^T \mathbb{f}(x(i)) + b(\mathbb{f}(x(i))) \quad \text{(A.26)}$$

Now we have that

$$\begin{aligned} & -\sum_{k\neq i} l_m(y(k), \mathbb{f}_\lambda^{[-i]}(x(k))) - \mu_\lambda^{[-i]}(i)^T \mathbb{f}_\lambda^{[-i]}(x(i)) + b(\mathbb{f}_\lambda^{[-i]}(x(i))) + nJ_\lambda(\mathbb{f}_\lambda^{[-i]}) \\ \leq \quad & -\sum_{k\neq i} l_m(y(k), \mathbb{f}(x(k))) - \mu_\lambda^{[-i]}(i)^T \mathbb{f}_\lambda^{[-i]}(x(i)) + b(\mathbb{f}_\lambda^{[-i]}(x(i))) + nJ_\lambda(\mathbb{f}) \\ \leq \quad & -\sum_{k\neq i} l_m(y(k), \mathbb{f}(x(k))) - \mu_\lambda^{[-i]}(i)^T \mathbb{f}(x(i)) + b(\mathbb{f}(x(i))) + nJ_\lambda(\mathbb{f}) \end{aligned}$$

The first inequality is due to the fact that $\mathbb{f}_\lambda^{[-i]}$ minimizes

$$-\sum_{k\neq i} l_m(y(k), \mathbb{f}(x(k))) + nJ_\lambda(\mathbb{f}) \quad \text{(A.27)}$$

and the second one is due to (A.26). □

Appendix B: Derivation of GACV

Our GACV is derived based on the cross validation function (2.25). Let us use the notations (2.21) and (2.22). It can be seen from (2.24) that $\sum_{j=1}^{m_i} w_{\lambda,ij}^{[-i]} \hat{f}_\lambda^{[-i]}(z_{ij})$ can be treated as a function of $\vec{f}_{\lambda i}^{[-i]}$. Note that $\vec{f}_{\lambda i}^{[-i]}$ is expected to be close to $\vec{f}_{\lambda i}$, Thus using the first order Taylor expansion to expand $\sum_{j=1}^{m_i} w_{\lambda,ij}^{[-i]} \hat{f}_\lambda^{[-i]}(z_{ij})$ at $\vec{f}_{\lambda i}$, we have that

$$\begin{aligned} \mathrm{CV}(\lambda) &\approx \widehat{\mathrm{OBS}}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i (\vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]})^T \frac{\partial \sum_{j=1}^{m_i} w_{ij}(\tau)\tau_j}{\partial \tau}\Big|_{\vec{f}_{\lambda i}} \\ &= \widehat{\mathrm{OBS}}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i (\vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]})^T \begin{pmatrix} d_{i1} \\ \vdots \\ d_{im_i} \end{pmatrix} \end{aligned} \tag{B.1}$$

where $w_{ij}(\tau)$ and d_{ij} are defined by (2.20) and (2.37), respectively. Thus, it remains to estimate $\vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]}$. To do this, we first extend the leave-out-one lemma (Craven and Wahba,1979[11]) to randomized covariate data.

LEMMA B.1 *(leave-out-one-subject lemma) Let $l(y_i, t) = y_i \cdot t - b(t) + c(y)$ be the log-likelihood function and*

$$I_\lambda^{Z,\Pi}(\vec{y}, f) = -\sum_{i=1}^{n} \log \sum_{j=1}^{m_i} \pi_{ij} \exp\{l(y_i, f(z_{ij}))\} + \frac{n\lambda}{2} J(f) \tag{B.2}$$

, where $\vec{y} = (\vec{y}_1^{\,T} ..., \vec{y}_n^{\,T})^T$ with $\vec{y}_i^{\,T} = (y_i, ..., y_i)^T$ being m_i replicates of y_i. Suppose that $\tau = (\tau_1, ..., \tau_{m_i})^T$ is a $m_i \times 1$ vector and $h_\lambda(i, \tau, \cdot)$ is the minimizer in $\mathcal{H}$ of $I_\lambda^{Z,\Pi}(\vec{Y}, f)$, where $\vec{Y} = (\vec{y}_1^{\,T}, ..., \vec{y}_{i-1}^{\,T}, \tau^T, \vec{y}_{i+1}^{\,T}, ..., \vec{y}_n^{\,T})^T$. Then

$$h_\lambda(i, \vec{\mu}_{\lambda i}^{[-i]}, \cdot) = \hat{f}_\lambda^{[-i]} \tag{B.3}$$

where $\hat{f}_\lambda^{[-i]}$ minimizes $-\sum_{k \neq i} \log \sum_{j=1}^{m_k} \pi_{kj} \exp\{l(y_i, f(z_{kj}))\} + \frac{n\lambda}{2} J(f)$, and $\vec{\mu}_{\lambda i}^{[-i]} = (b'(\hat{f}_\lambda^{[-i]}(z_{i1})), ..., b'(\hat{f}_\lambda^{[-i]}(z_{im_i})))^T$ is the vector of means corresponding to

$\hat{f}_\lambda^{[-i]}$.

Proof of Lemma B.1. Firstly, we claim that

$$l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), \hat{f}_\lambda^{[-i]}(z_{ij})) \geq l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), f(z_{ij})),\ 1 \leq j \leq m_i,\ \forall f \in \mathcal{H}. \quad \text{(B.4)}$$

This follows since

$$\frac{\partial\, l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), t)}{\partial\, t} = b'(\hat{f}_\lambda^{[-i]}(z_{ij})) - b'(t)$$

and using the fact that $\frac{\partial^2 l(y,t)}{\partial t^2} = -b''(t) < 0$. Therefore $l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), t)$ achieves its unique maximum for $t = \hat{f}_\lambda^{[-i]}(z_{ij})$.

Define $\vec{y}^{\,[-i]} = (\vec{y}_1^{\,T}, ..., \vec{y}_{i-1}^{\,T}, (\vec{\mu}_{\lambda i}^{[-i]})^T, \vec{y}_{i+1}^{\,T}, ..., \vec{y}_n^{\,T})^T$. Then for any $f \in \mathcal{H}$,

$$\begin{aligned}
I_\lambda^{Z,\Pi}(\vec{y}^{\,[-i]}, f) &= -\log \sum_{j=1}^{m_i} \pi_{ij} \exp\{l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), f(z_{ij}))\} \\
&\quad - \sum_{k \neq i} \log \sum_{j=1}^{m_k} \pi_{kj} \exp\{l(y_k, f(z_{kj}))\} + \frac{n\lambda}{2} J(f) \\
&\geq -\log \sum_{j=1}^{m_i} \pi_{ij} \exp\{l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), \hat{f}_\lambda^{[-i]}(z_{ij}))\} \\
&\quad - \sum_{k \neq i} \log \sum_{j=1}^{m_k} \pi_{kj} \exp\{l(y_k, f(z_{kj}))\} + \frac{n\lambda}{2} J(f) \\
&\geq -\log \sum_{j=1}^{m_i} \pi_{ij} \exp\{l(b'(\hat{f}_\lambda^{[-i]}(z_{ij})), \hat{f}_\lambda^{[-i]}(z_{ij}))\} \\
&\quad - \sum_{k \neq i} \log \sum_{j=1}^{m_k} \pi_{kj} \exp\{l(y_k, \hat{f}_\lambda^{[-i]}(z_{kj}))\} + \frac{n\lambda}{2} J(\hat{f}_\lambda^{[-i]}).
\end{aligned}$$

The first inequality is due to (B.4) and the second one is due to the fact that $\hat{f}_\lambda^{[-i]}$ minimizes $-\sum_{k \neq i} \log \sum_{j=1}^{m_k} \pi_{kj} \exp\{l(y_i, f(z_{kj}))\} + \frac{n\lambda}{2} J(f)$. Thus we have $h_\lambda(i, \vec{\mu}_{\lambda i}^{[-i]}, \cdot)$ $= \hat{f}_\lambda^{[-i]}$. $\square$

Consider the parametric form of the penalized likelihood in (2.26) and denote

$$\vec{y}^{\,[-i]} = (\vec{y}_1^{\,T}, ..., \vec{y}_{i-1}^{\,T}, (\vec{\mu}_{\lambda i}^{[-i]})^T, \vec{y}_{i+1}^{\,T}, ..., \vec{y}_n^{\,T})^T. \tag{B.5}$$

Then Lemma B.1 says that

$$\vec{f}_\lambda^{\,[-i]} = (\hat{f}_\lambda^{[-i]}(z_{11}), ..., \hat{f}_\lambda^{[-i]}(z_{1m_1}), \hat{f}_\lambda^{[-i]}(z_{21}), ..., \hat{f}_\lambda^{[-i]}(z_{nm_n}))^T \tag{B.6}$$

minimizes $I_\lambda^{Z,\Pi}(\vec{y}^{\,[-i]}, \vec{f}\,)$. Note that

$$\vec{f}_\lambda = (\hat{f}_\lambda(z_{11}), ..., \hat{f}_\lambda(z_{1m1}), \hat{f}_\lambda(z_{21}), ..., \hat{f}_\lambda(z_{nm_n}))^T \tag{B.7}$$

minimizes $I_\lambda^{Z,\Pi}(\vec{y}, \vec{f}\,)$. Thus

$$\frac{\partial\, I_\lambda^{Z,\Pi}}{\partial\, \vec{f}}(\vec{y}, \vec{f}_\lambda) = 0, \quad \frac{\partial\, I_\lambda^{Z,\Pi}}{\partial\, \vec{f}}(\vec{y}^{\,[-i]}, \vec{f}_\lambda^{\,[-i]}) = 0. \tag{B.8}$$

Using first order Taylor expansion, we have that

$$\begin{aligned}
0 &= \frac{\partial I_\lambda^{Z,\Pi}}{\partial \vec{f}}(\vec{y}^{\,[-i]}, \vec{f}_\lambda^{\,[-i]}) \\
&= \frac{\partial I_\lambda^{Z,\Pi}}{\partial \vec{f}}(\vec{y}, \vec{f}_\lambda) + \frac{\partial^2 I_\lambda^{Z,\Pi}}{\partial \vec{f}\,\partial \vec{f}^T}(\vec{y}^*, \vec{f}_\lambda^{\,*})(\vec{f}_\lambda^{\,[-i]} - \vec{f}_\lambda) + \frac{\partial^2 I_\lambda^{Z,\Pi}}{\partial \vec{y}\,\partial \vec{f}^T}(\vec{y}^*, \vec{f}_\lambda^*)(\vec{y}^{\,[-i]} - \vec{y}) \\
&= \frac{\partial^2 I_\lambda^{Z,\Pi}}{\partial \vec{f}\,\partial \vec{f}^T}(\vec{y}^*, \vec{f}_\lambda^*)(\vec{f}_\lambda^{\,[-i]} - \vec{f}_\lambda) + \frac{\partial^2 I_\lambda^{Z,\Pi}}{\partial \vec{y}\,\partial \vec{f}^T}(\vec{y}^*, \vec{f}_\lambda^*)(\vec{y}^{\,[-i]} - \vec{y})
\end{aligned} \tag{B.9}$$

where $(\vec{y}^*, \vec{f}_\lambda^*)$ is a point between $(\vec{y}, \vec{f}_\lambda)$ and $(\vec{y}^{\,[-i]}, \vec{f}_\lambda^{\,[-i]})$.

Consider any arbitrary vector $\vec{f} = (\vec{f}_1^{\,T}, ..., \vec{f}_n^{\,T})$ with $\vec{f}_i = (f_{i1}, ..., f_{im_i})^T$ being an $m_i \times 1$ vector. For $1 \le i \le n$ and $1 \le s, t \le m_i$, let's denote

$$b_{st}^i(\vec{f}\,) = \begin{cases} -w_{is}(\vec{f}\,)\left[1 + (1 - w_{is}(\vec{f}\,))f_{is}(y_i - b'(f_{is}))\right], & \text{if } s = t \\ w_{is}(\vec{f}\,)w_{it}(\vec{f}\,)f_{is}(y_i - b'(f_{it})), & \text{if } s \neq t \end{cases}$$

$$d_{st}^i(\vec{f}\,) = \begin{cases} w_{is}(\vec{f}\,)\left[b''(f_{is}) - (1 - w_{is}(\vec{f}\,))(y_i - b'(f_{is}))^2\right], & \text{if } s = t \\ w_{is}(\vec{f}\,)w_{it}(\vec{f}\,)(y_i - b'(f_{is}))(y_i - b'(f_{it})), & \text{if } s \neq t. \end{cases}$$

Define submatrices $B_i(\vec{f}\,) = \left(b^i_{st}(\vec{f}\,)\right)_{m_i\times m_i}$ and $D_i(\vec{f}\,) = \left(d^i_{st}(\vec{f}\,)\right)_{m_i\times m_i}$ and let $B(\vec{f}\,) = \text{diag}(B_1(\vec{f}\,), ..., B_n(\vec{f}\,))$ and $D(\vec{f}\,) = \text{diag}(D_1(\vec{f}\,), ..., D_n(\vec{f}\,))$ be block diagonal matrices. Then direct calculation yields

$$\frac{\partial^2 I^{Z,\Pi}_\lambda}{\partial\vec{f}\,\partial\vec{f}^T}(\vec{y}^*, \vec{f}^*_\lambda) = \frac{1}{n}D(\vec{f}^*_\lambda) + \Sigma_\lambda, \quad \frac{\partial^2 I^{Z,\Pi}_\lambda}{\partial\vec{y}\,\partial\vec{f}^T}(\vec{y}^*, \vec{f}^*_\lambda) = \frac{1}{n}B(\vec{f}^*_\lambda). \tag{B.10}$$

Therefore, from (B.9), we have

$$\vec{f}_\lambda - \vec{f}^{\,[-i]}_\lambda = -(D(\vec{f}^*_\lambda) + n\Sigma_\lambda)^{-1}B(\vec{f}^*_\lambda)(\vec{y} - \vec{y}^{\,[-i]}). \tag{B.11}$$

Approximate $B(\vec{f}^*_\lambda)$ and $D(\vec{f}^*_\lambda)$ by $B(\vec{f}_\lambda)$ and $D(\vec{f}_\lambda)$. Then denote

$$H = -(D(\vec{f}_\lambda) + n\Sigma_\lambda)^{-1}B(\vec{f}_\lambda) \tag{B.12}$$

the influence matrix of $I^{Z,\Pi}_\lambda(\vec{y}, \vec{f}\,)$ with respect to $\vec{f}$ evaluated at $\vec{f}_\lambda$. From (B.11), we have

$$\begin{pmatrix} \vec{f}_{\lambda 1} - \vec{f}^{\,[-i]}_{\lambda 1} \\ \vdots \\ \vec{f}_{\lambda i} - \vec{f}^{\,[-i]}_{\lambda i} \\ \vdots \\ \vec{f}_{\lambda n} - \vec{f}^{\,[-i]}_{\lambda n} \end{pmatrix} \approx H \begin{pmatrix} 0 \\ \vdots \\ \vec{y}_i - \vec{\mu}^{\,[-i]}_{\lambda i} \\ \vdots \\ 0 \end{pmatrix}_{\sum m_i\times 1}. \tag{B.13}$$

Write

$$H = \begin{pmatrix} H_{11} & * & * & * \\ * & H_{22} & \cdots & * \\ \vdots & \vdots & \ddots & \vdots \\ * & * & \cdots & H_{nn} \end{pmatrix}_{\sum m_i\times\sum m_i} \tag{B.14}$$

where each H_{ii} is a $m_i \times m_i$ submatrix matrix on the diagonal with respect to $(f_{i1}, ..., f_{im_i})^T$. We observe from (B.13) that

$$\vec{f}_{\lambda i} - \vec{f}^{\,[-i]}_{\lambda i} \approx H_{ii}(\vec{y}_i - \vec{\mu}^{\,[-i]}_{\lambda i}). \tag{B.15}$$

Recall that $\vec{\mu}_{\lambda i}^{[-i]} = (b'(\hat{f}_{\lambda}^{[-i]}(z_{i1})), ..., b'(\hat{f}_{\lambda}^{[-i]}(z_{im_i})))^T$ is a vector of $b'(\cdot)$ evaluated at $\vec{f}_{\lambda i}^{[-i]}$. Hence, using a first order Taylor expansion to expand $b'(\cdot)$ at $\vec{f}_{\lambda i}$, we have

$$\vec{\mu}_{\lambda i}^{[-i]} - \vec{\mu}_{\lambda i} \approx W_i(\vec{f}_{\lambda i}^{[-i]} - \vec{f}_{\lambda i}) \tag{B.16}$$

where $W_i = \text{diag}(b''(\hat{f}_{\lambda}(z_{i1})), ..., b''(\hat{f}_{\lambda}(z_{im_i})))$ is a diagonal matrix of variances.

Combining (B.15) and (B.16), we can show that

$$\begin{aligned} \vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]} &\approx H_{ii}(\vec{y}_i - \vec{\mu}_{\lambda i}^{[-i]}) \\ &= H_{ii}(\vec{y}_i - \vec{\mu}_{\lambda i} + \vec{\mu}_{\lambda i} - \vec{\mu}_{\lambda i}^{[-i]}) \\ &\approx H_{ii}(\vec{y}_i - \vec{\mu}_{\lambda i} + W_i(\vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]})). \end{aligned} \tag{B.17}$$

Now, an approximation of $\vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]}$ can be obtained by solving (B.17)

$$\vec{f}_{\lambda i} - \vec{f}_{\lambda i}^{[-i]} \approx (I_{m_i \times m_i} - H_{ii}W_i)^{-1}H_{ii}(\vec{y}_i - \vec{\mu}_{\lambda i}). \tag{B.18}$$

Plug (B.18) into the CV function (B.1), we obtain the approximate cross validation (ACV) function

$$\text{ACV}(\lambda) = \widehat{\text{OBS}}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i(d_{i1}, ..., d_{im_i})(I_{m_i \times m_i} - H_{ii}W_i)^{-1}H_{ii}(\vec{y}_i - \vec{\mu}_{\lambda i}) \tag{B.19}$$

where $\widehat{\text{OBS}}(\lambda)$ is given in (2.19). Define $G_{ii} = I_{m_i \times m_i} - H_{ii}W_i$. Then a generalized form of approximate cross validation (GACV) can be obtained by replacing each H_{ii} and G_{ii} with the generalized average of submatrices defined in (2.31). Let $\bar{H}_{ii}$ and $\bar{G}_{ii}$ denote the generalized average of H_{ii} and G_{ii}. Then the generalized approximate

cross validation (GACV) can be defined

$$\begin{aligned} \text{GACV}(\lambda) &= \widehat{\text{OBS}}(\lambda) + \frac{1}{n}\sum_{i=1}^{n} y_i(d_{i1},...,d_{im_i})\bar{G}_{ii}^{-1}\bar{H}_{ii}(\vec{y}_i - \vec{\mu}_{\lambda i}) \\ &= -\frac{1}{n}\sum_{i=1}^{n}\log\sum_{j=1}^{m_i}\pi_{ij}\exp\left\{y_i\hat{f}_\lambda(z_{ij}) - b(\hat{f}_\lambda(z_{ij}))\right\} \qquad \text{(B.20)} \\ &\quad + \frac{1}{n}\sum_{i=1}^{n} y_i(d_{i1},...,d_{im_i})\bar{G}_{ii}^{-1}\bar{H}_{ii}\begin{pmatrix} y_i - \hat{\mu}_\lambda(z_{i1}) \\ \vdots \\ y_i - \hat{\mu}_\lambda(z_{im_i}) \end{pmatrix}. \end{aligned}$$

We remark that if all the x_i's are exactly observed, then the above GACV function will reduce to the original GACV formula in Xiang and Wahba (1996)[44].

Appendix C: Extension to SS-ANOVA model

Smoothing spline analysis of variance (SS-ANOVA) provides a general framework for multivariate nonparametric function estimation. The application is very broad. To extend the methodologies of Chapter 2, it suffices to show that the penalized likelihood for SS-ANOVA model can be formulated in the form of (2.3). The following arguments are derived from Wahba (1990)[39].

The penalized likelihood of smoothing Spline ANOVA model takes the form of

$$I_\lambda(f) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, f) + \sum_{\beta=1}^{b} \lambda_\beta ||\mathbb{P}_1^\beta f||^2_{\mathcal{H}_1^\beta} \tag{C.1}$$

where $\mathcal{H}_1^\beta$ are nonparametric subspaces (smooth spaces) which are assumed to be RKHS with reproducing kernel $K_1^\beta(\cdot,\cdot)$ and $\mathbb{P}_1^\beta$ projects f onto $\mathcal{H}_1^\beta$. Now For $\lambda_\beta > 0$, define $\mathcal{H}_1 = \sum_{\beta=1}^{b} \oplus \mathcal{H}_1^\beta$ with norm

$$||\eta||^2_{\mathcal{H}_1} = \sum_{\beta=1}^{b} \lambda_\beta ||\mathbb{P}_1^\beta \eta||^2_{\mathcal{H}_1^\beta}, \eta \in \mathcal{H}_1. \tag{C.2}$$

It can be shown that $\mathcal{H}_1$ is a RKHS with a reproducing kernel $\sum_{\beta=1}^{b} \frac{1}{\lambda_\beta} K_1^\beta(s,t)$. Then we can write that

$$I_\lambda(f) = -\frac{1}{n}\sum_{i=1}^{n} \log p(y_i|x_i, f) + ||\mathbb{P}_1 f||^2_{\mathcal{H}_1} \tag{C.3}$$

where $\mathbb{P}_1$ projects $f \in \mathcal{H}$ onto $\mathcal{H}_1$. Set $J(f) = ||\mathbb{P}_1 f||^2_{\mathcal{H}_1}$. Then the above expression takes the form of (2.3). Therefore our discussion in Chapter 2 can be extended to SS-ANOVA model.

LIST OF REFERENCES

[1] AKAIKE, H. (1973). Information theory and an extension of the maximum likelihood principle. *In Second International Symposium on Information Theory*. 267–281.

[2] BERLINET, A. and THOMAS–AGNAN, C. (2004). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Kluwer Academic Publishers, Norwell, Massachusetts.

[3] BERRY, S. M., CARROLL, R. J. and RUPPERT, D. (2001). Bayesian smoothing and regression splines for measurement error problems. *J. Amer. Statist. Assoc.* **97** 160–169.

[4] BOSSERHOFF, V. (2008). The bit-complexity of finding nearly optimal quadrature rules for weighted integration. *Journal of Universal Computer Science* **14** 938–955.

[5] CARDOT, H., CRAMBES, C., KNEIP, A. and SARDA, P. (2007). Smoothing splines estimators in functional linear regression with errors-in-variables. *Computational Statistics and Data Analysis* **51** 4832–4848.

[6] CARROLL, R. J., MACA, J. D. and RUPPERT, D. (1999). Nonparametric regression in the presence of measurement error. *Biometrika* **86** 541–554.

[7] CARROLL, R. J., RUPPERT, D., STEFANSKI, L. A. and CRAINICEANU, C. M. (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall CRC Press, Boca Raton.

[8] CHEN, Q. and IBRAHIM, J. G. (2006). Semiparametric models for missing covariate and response data in regression models. *Biometrics* **62** 177–184.

[9] CHEN, Q., ZENG, D. and IBRAHIM, J. G. (2007). Sieve maximum likelihood estimation for regression models with covariates missing at random. *J. Amer. Statist. Assoc.* **102** 1309–1317.

[10] COOK, J. R. and STEFANSKI, L. A. (1994). Simulation-extrapolation estimation in parametric measurement error models. *J. Amer. Statist. Assoc.* **89** 1314–1328.

[11] CRAVEN, P. and WAHBA, G. (1979). Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation *Numer. Math.* **31** 377–403.

[12] DELAIGLE, A., FAN, J. and CARROLL, R. J. (2009). A design-adaptive local polynomial estimator for the errors-in-variables problem. *J. Amer. Statist. Assoc.* **104** 348–359.

[13] EFRON, B. (1986). How biased is the apparent error rate of a prediction rule? *J. Amer. Statist. Assoc.* **81** 461–470.

[14] EFRON, B. (2004). The estimation of prediction error: Covariance penalties and cross-validation (with discussion). *J. Amer. Statist. Assoc.* **99** 619–642.

[15] FAN, J. and TRUONG, Y. K. (1993). Nonparametric regression with errors in variables. *Ann. Statist.* **21** 1900–1925.

[16] FERNANDES, A. D. and ATCHLEY, W. R. (2006). Gaussian quadrature formulae for arbitrary positive measures. *Evolutionary Bioinformatics Online* **2** 251–259.

[17] GAO, F., WAHBA, G., KLEIN, R. and KLEIN, B. E. K. (2001). Smoothing spline ANOVA for multivariate Bernoulli observation, with application to ophthalmology data. *J. Amer. Statist. Assoc.* **96** 127–160.

[18] GOLUB, G. H. and WELSCH, J. H. (1969). Calculation of Gauss quadrature rules. *Mathematics of Computation* **23** 221–230.

[19] GREEN, P. J. (1990). On use of the EM for penalized likelihood estimation. *J. Roy. Statist. Soc. Ser. B* **52** 443–452.

[20] GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York.

[21] IBRAHIM, J. G. (1990). Incomplete data in generalized linear models. *J. Amer. Statist. Assoc.* **85** 765–769.

[22] IBRAHIM, J. G., LIPSITZ, S. R. and CHEN, M. (1999). Missing covariates in generalized linear models when the missing data mechanism is nonignorable. *J. Roy. Statist. Soc. Ser. B* **61** 173–190.

[23] IBRAHIM, J. G., CHEN, M., LIPSITZ, S. R. and HERRING, A. H. (2005). Missing data methods for generalized linear models: a comparative review. *J. Amer. Statist. Assoc.* **100** 332–346.

[24] IOANNIDES, D. A. and ALEVIZO, P. D. (1997). Nonparametric regression with errors in variables and applications. *Statist. Probab. Lett.* **32** 35–43.

[25] HORTON, N. J. and LAIRD, N. M. (1999). Maximum likelihood analysis of generalized linear models with missing covariates. *Statistical Methods in Medical Research* **8** 37–50.

[26] HORTON, N. J. and KEN, P. K. (2007). Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. *J. Amer. Statist. Assoc.* **61** 79–90.

[27] HUANG, L., CHEN, M. and IBRAHIM, J. (2005). Bayesian analysis for generalized linear models with nonignorably missing covariates. *Biometrics* **61** 767–780.

[28] KIMELDORF, G. and WAHBA, G. (1971). Some results on tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95.

[29] KLEIN, R., KLEIN, B. E. K., LINTON, K. L. and DEMETS, D. L. (1991). The Beaver Dam eye study: Visual acuity. *Ophthalmology* **98** 1310–1315.

[30] KURDILA, A. and ZABARANKIN, M. (2005). *Convex Functional Analysis (Systems and Control: Foundations and Applications)*. Birkhauser Basel, Switzerland.

[31] LIN, X., WAHBA, G., XIANG, D., GAO, F., KLEIN, R. and KLEIN, B. E. K. (2000). Smoothing Spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. *Ann. Statist.* **28** 1570–1600.

[32] LITTLE, R. J. A. and RUBIN, D. B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. Wiley, New York.

[33] RAHMAN, S. (2009). Extended polynomial dimensional decomposition for arbitrary probability distributions. *Journal of Engineering Mechanics* **135** 1439–1451.

[34] SCHENNACH, S. M. (2004). Nonparametric regression in the presence of measurement error. *Econometric Theory* **20** 1046–1093.

[35] STEIN, C. (1981). Estimation of the mean of a multivariate normal distribution. *Ann. Statist.* **9** 1135–1151.

[36] O'SULLIVAN, F. (1983). The analysis of some penalized likelihood estimation schemes. Technical Report 726, Dept. Statistics, Univ. Wisconsin-Madison.

[37] VAN HUFFEL, S. and VANDEWALLE, J. (1991). *The Total Least Squares Problem: Computational Aspects and Analysis*. SIAM, Philadelphia.

[38] WAHBA, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *J. Roy. Statist. Soc. Ser. B* **45** 133–150.

[39] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

[40] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.

[41] SHI, W. (2008). LASSO-Patternsearch Algorithm. Technical Report 1147, Dept. Statistics, Univ. Wisconsin-Madison.

[42] SHI, W., WAHBA, G., WRIGHT, S., LEE, K., KLEIN, R. and KLEIN, B. E. K. (2008). LASSO-Patternsearch algorithm with application to ophthalmology and genomic data. *Statistics and Its Interface* **1** 137–153.

[43] WHITTAKER, J. (1990). *Graphical Models in Applied Mathematical Multivariate Statistics*. Wiley.

[44] XIANG, D. and WAHBA, G. (1996). A generalized approximate cross validation for smoothing splines with non-Gaussian data. *Statist. Sinica* **6** 675–692.

[45] YE, J. (1998). On measuring and correcting the effects of data mining and model selection. *J. Amer. Statist. Assoc.* **93** 120–131.

[46] YE, J. and WONG, W. (1997). Model uncertainty and correcting for selection bias. Unpublished manuscript.

[47] ZHANG, H., WAHBA, G., LIN, Y., VOELKER, M., FERRIS, M., KLEIN, R. and KLEIN, B. (2004). Variable selection and model building via likelihood basis pursuit. *J. Amer. Statist. Assoc.* **99** 659–672.

[48] ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2007). On the "degrees of freedom" of the LASSO. *Ann. Statist.* **35** 217–2192.