Contributions to Functional Data Analysis and High-Throughput Screening Assay Analysis

by

Toshiya Hoshikawa

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
(Statistics)
in The University of Michigan
2012

Doctoral Committee:

Professor Tailen Hsing, Co-Chair
Professor Kerby A. Shedden, Co-Chair
Professor Bin Nan
Professor Naisyin Wang

To Koko

ACKNOWLEDGEMENTS

I am sincerely grateful to my Ph.D. advisors, Prof. Tailen Hsing, Prof. Kerby Shedden, and Prof. Naisyin Wang, for the support and guidance they showed me throughout my Ph.D. study. Without their helps, this dissertation would not have been possible.

I would also like to thank Prof. Shedden for his guidance beyond research. Without his direction, I would have never reached this goal.

My special thanks go to Prof. Bin Nan for participating in my dissertation committee in this busiest time.

I am truly indebted and thankful to my office friends, Kohinoor Dasgupta, Ming-Chi Hsu, and Joel Vaughan for their helps and encouragement. Without their friendship, I would not have stayed sane through these difficult years. I learned various things beyond statistics from them, which immeasurably enriched my Ph.D. life.

I also wish to thank my Japanese friend at Michigan, Yasuhiro Kyono, for all the time we spent together on entertainment. My life would have been dry and colorless without a friend like him who shares so many common interests with me, particularly, sports, running, and beer.

I owe earnest thankfulness to my parents. They bore me, raised me, and were always supporting and encouraging me with their best wishes.

Finally, and most importantly, I would like to show my best gratitude to my fiancée, Koko. She was always cheering me up from Japan without giving up on me for five years. Without her, it would not have had any meaning to finish the degree.

TABLE OF CONTENTS

CHAPTER

LIST OF FIGURES

Figure

LIST OF TABLES

Table

CHAPTER I

Introduction

Modern data analysis is characterized by the complex nature of available data sets; modern technology has brought us access to huge amount of data in a variety of fields, ranging from natural sciences to social sciences and from industry to public sectors, which had either not simply been available or been ignored before because of the lack of sufficient computational power. While these data sets are mines of information, traditional statistical methodologies have become insufficient to effectively extract information from them. Thus statisticians are now developing the new statistical techniques for such large and complex data sets. This dissertation contributes to the statistical analysis of several of such challenging types of data sets. In particular, I will present the results from three projects concerning functional data analysis and high-throughput screening assay data analysis.

In Chapter II, we consider a method that clusters a sample and estimates the regression models for the clusters simultaneously. In a regression analysis, suppose we suspect that there are several heterogeneous groups in the population that a sample represents. Mixture regression models have been applied to address such problems [DeSarbo and Cron (1988), McLachlan and Peel (2000), Naik et al. (2007), Yao et al. (2011)]. By modeling the conditional distribution of the response given the covariate as a mixture, the sample can be clustered into groups and the individual regression models for the groups can be estimated simultaneously. This approach treats the covariate as deterministic so that the covariate carries no information as to which group the subject is likely to belong to. Although this assumption may be reasonable in experiments where the covariate is completely determined by the experimenter, in observational data the covariate may behave differently across the groups. Thus the model should also incorporate the heterogeneity of the covariate, which allows us to estimate the membership of the subject from the covariate.

In this chapter, we consider a mixture regression model where the *joint* distribution of the response and the covariate is modeled as a mixture. Given a new observation of the covariate, this approach allows us to compute the posterior probabilities that the subject belongs to each group. Using these posterior probabilities, the prediction of the response can adaptively use the covariate. We introduce an inference procedure for this approach and show its properties concerning estimation and prediction. The model is explored for the functional covariate as well as the multivariate covariate. We present a real-data example where our approach outperforms the traditional approach, using the well-known Berkeley growth study data.

In Chapter III, we consider a regularization approach to functional linear models. Functional data analysis is a rapidly growing field as computer technology develops and data collection power dramatically improves. Within the field functional linear regression has drawn a particular attention in recent years, where the predictor is a random function and the response is a scalar. Generally speaking, the inference procedure for functional linear models can be divided into two ways. The first approach is to reduce the dimensionality of the functional predictor by mapping it onto a finite dimensional space. A space spanned by the leading eigenfunctions of the predictor covariance function is often used. There is, however, no compelling reason to believe that the eigenfunctions explain the regression structure well. The other approach is to regularize some aspect of a slope function, often its smoothness, by putting a penalty on it. This approach can avoid the choice of the finite dimensional space to map onto by assuming that the regression slope function lies in some reproducing kernel Hilbert space (RKHS). The property of the RKHS, then, makes the problem essentially a finite dimensional computation problem. The regularization approach has been explored by several authors, including Crambes et al. (2009) and Yuan and Cai (2010), but the derivation is often very technical and difficult to see what is behind the approach. Our goal in this project is two fold. The first goal is to provide much simpler derivation of the optimal prediction convergence rate in a general setting. It is revealed that the key idea behind the approach is to choose a penalty term so that the objective function has a unique solution. The second goal is to extend the practical aspects of the model by accommodating discrete observations and multiple predictors. The effects on the convergence rates are explored.

In Chapter IV, we consider a method to assess the overall toxicity of a chemical compound from high-throughput screening assays. Conventionally, toxicity has been measured by cell culture assays or animal testing to assess the direct effects on living organisms. However, these methods are costly in terms of time and money. Recently, advances in assay technology has made it possible to conduct millions of biochemical tests simultaneously under an automated system using machines. This experimental approach is called high-throughput screening (HTS) [Zhang et al. (1999), Zhang (2011)]. A promising way to enhance the exploration of the chemical space is to use HTS assays as less expensive alternatives to conventional approaches. The prediction results from the results of the HTS assays can be used to, for example, prioritize the compounds in terms of risk to save time and cost [Dix et al. (2007), Judson et al. (2010)]. The HTS assays are, however, limited to those that do not require careful human attention, such as those focusing on the molecular features of the compounds, thus it cannot directly evaluate the phenotypic effects. In this context, a prediction relationship between the HTS and conventional assays must be defined.

In some applications, the lowest value among the conventional assays is of primary interest, in which case it may be advantageous to predict this minimum value directly rather than in two stages following prediction of each assay separately. We introduce a method to directly specify and estimate a model for the parameter of interest through a profile likelihood function. Through a variety of simulation studies, the property of this profile likelihood approach is evaluated. In particular, we explore in what situations the use of the profile likelihood approach can be beneficial. The idea is also extend to the interval estimation. We apply this method to the ToxCast data of the EPA and to the 60 cell line screen of the NCI.

CHAPTER II

Mixture Regression for Observational Data, with Application to Functional Regression Models

2.1 Introduction

In a regression analysis, suppose we suspect that there are several heterogeneous groups in the population that a sample represents. Mixture regression models have been applied to address such problems [DeSarbo and Cron (1988), McLachlan and Peel (2000)]. It is assumed in mixture regression that, given a p-dimensional covariate X whose subject belongs to the kth group, the conditional mean of the response Y is related to the linear function of X through a link function h in the format of $h\{\mathbb{E}(Y|X, \delta_k = 1)\} = \alpha_k + \beta_k^T X$, where δ_k is the membership variable that returns one if the subject belongs to the kth group and zero otherwise. For simplicity, we focus on the normal, identity link model where the conditional density is given by

$$f_{Y|X,\delta_k=1}(y|x) = \varphi(y; \alpha_k + \beta_k^T x, \sigma_k^2),$$

where $\varphi(\cdot; \mu, \sigma^2)$ is the normal density with mean μ and variance σ^2. The EM algorithm can be used to compute the maximum likelihood estimator (MLE), as often done for finite mixture models [McLachlan and Peel (2000)]. Information criteria such as Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC) can be used to estimate the number of groups. Naik et al. (2007) introduced a modified

AIC that is tailored for mixture regression models.

Recently, Yao et al. (2011) introduced a mixture regression model where the covariate is given by functional data. They conducted a real-data analysis and claimed that the mixture regression approach works better than the (usual) linear regression approach in terms of prediction. We reconsider this analysis and show that the mixture regression approach works no better than the linear regression approach when the membership of a new observation is not available. For the overview on functional data analysis, readers may refer to excellent monographs written by Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006).

The mixture regression model introduced above treats the covariate as deterministic or its distribution as invariant across the groups. Thus the covariate carries no information as to which group the subject is likely to belong to. Consider the prediction of the response from a new observation of the covariate; the best we can do is to take the average of the linear predictors over the groups with certain fixed weights. Although this assumption may be reasonable in experiments where the covariate is determined in a completely deterministic way, in observational data the covariate may behave differently across the groups. Thus the model should incorporate the heterogeneity of the covariate as well so that we can estimate the membership of the subject from the covariate.

In this chapter, we introduce a mixture regression model, where the *joint* distribution of the response and the covariate is modeled as a mixture. In particular, we assume that the joint density of X and Y is given by

$$f_{Y,X|\delta_k=1}(y,x) = \varphi(y;\alpha_k + \beta_k^T x, \sigma_k^2)\varphi(x,\mu_k,\Sigma_k).$$

This is a generalization of the traditional mixture regression model; when the covariate distribution is identical across the groups, this model becomes equivalent to the traditional model. Our new approach allows the covariate to behave differently across the groups as its marginal distribution becomes a mixture. This covariate heterogeneity allows us to compute the posterior probabilities that the subject belongs to each group; using these posterior probabilities, the prediction of the response can adaptively use the covariate. This assumption is particularly reasonable in functional data analysis; in many practical situations, functional data appears in observational studies. We introduce one of such examples in Section 2.5.

The rest of the chapter proceeds as follows. In Section 2.2, we explore our new approach in more details and introduce an inference procedure. We first consider the multivariate covariate model, followed by the functional covariate model. Furthermore, we introduce a couple of simple but very effective ways to extend the model to improve the prediction performance; these tricks are used in Section 2.5. Section 2.3 discusses the properties of the estimator and the predictor in the joint mixture regression model. In Section 2.4, we explore the properties of our new approach by simulation studies. Section 2.5 presents a real-data analysis where we show how our approach can improve the prediction performance from the traditional approach, by using the well-analyzed Berkeley growth study data. Finally, we conclude the chapter with some remarks in Section 2.6.

2.2 Joint Mixture Regression

2.2.1 The Multivariate Covariate Model

Let us first consider the model for the multivariate covariate. Denote the response by Y and the p-dimensional covariate by X. We consider the model where the joint distribution of (Y, X) is a mixture whose density is given by

$$f(y, x) = \sum_{k=1}^{K} \pi_k \varphi(y; \alpha_k + \beta_k^T x, \sigma_k^2) \varphi(x; \mu_k, \Sigma_k), \tag{2.1}$$

where $\varphi(\cdot;\mu,\Sigma)$ is the (multivariate) normal density with mean μ and variance(-covariance matrix) Σ. α_k and β_k are respectively the regression intercept and slope for the kth mixture component. Within each mixture component, which represents a group, the marginal distribution of the covariate is given by a normal distribution whose parameters vary across the components. As noted in Introduction, this model differs from the traditional mixture regression models in that the traditional approach does not incorporate the covariate distribution into the model [McLachlan and Peel (2000), Naik et al. (2007), Yao et al. (2011)]. In particular, the traditional approach is based on the *conditional* distribution while in our approach the *joint* distribution is assumed to be a mixture. We call the former *ordinary mixture regression* (OMR) and the latter *joint mixture regression* (JMR). $(\pi_1,\ldots,\pi_K)$ are mixing proportions, i.e., $\pi_k > 0$ and $\sum_k \pi_k = 1$, and K is the number of the mixture components. To avoid identifiability issues, we assume that K is the smallest in the sense that there is no expression that has fewer components than K but still retains the identical distribution. We also treat the parameter space as the quotient space with respect to permutation in axes.

An alternative expression which is equivalent to (3.1) and will become useful when exploring the functional covariate model in the next section is given by

$$\begin{aligned} Y &= \sum_{k=1}^{K} \delta_k(\alpha_k + \beta_k^T X + \varepsilon_k) \\ X &= \sum_{k=1}^{K} \delta_k X_k, \end{aligned} \tag{2.2}$$

where $\Delta = (\delta_1,\ldots,\delta_K)$ follows a multinomial distribution with parameters $n = 1$ and $(\pi_1,\ldots,\pi_K)$, $X_k \sim N(\mu_k,\Sigma_k)$, $\varepsilon_k \sim N(0,\sigma_k^2)$, and X_k, δ_k and ε_k are jointly independent. Δ is often called the membership vector, which indicates the group to which the subject belongs. Note that we observe X but not Δ, i.e., the membership.

Let $(x_1,y_1),\ldots,(x_n,y_n)$ be n independent observations from (3.1), or equivalently

(4.7). For a fixed positive integer K, we estimate the parameters by maximizing the log-likelihood

$$\ell_n(\Psi; y, x) = \sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} \pi_k \varphi(y_i; \alpha_k + \beta_k^T x_i, \sigma_k^2) \varphi(x_i; \mu_k, \Sigma_k) \Big\},$$

where the parameters to be estimated are

$$\Psi = \{\pi_1, \ldots, \pi_{K-1}, \alpha_1, \beta_1, \sigma_1^2, \mu_1, \Sigma_1, \ldots, \alpha_K, \beta_K, \sigma_K^2, \mu_K, \Sigma_K\}.$$

As commonly used in finite mixture models, the EM algorithm can be used to compute the MLE, where Δ is treated as missing values. The explicit formula of the algorithm is given in Appendix 2.8.

The number of the mixture components K can be estimated through Bayesian Information Criterion (BIC), i.e.,

$$\widehat{K} = \underset{K}{\operatorname{argmax}} \Big\{ \max_{\Psi} \ell_n(\Psi; y, x) - \frac{|\Psi|}{2} \log n \Big\},$$

where $|\Psi|$ is the number of the parameters. Under some regularity conditions such as the compactness of the parameter space, BIC provides a consistent estimator for K. For the details, see Keribin (2000).

Given a new observation of the covariate X, the posterior probabilities of the membership are given by $\mathbb{E}(\delta_k|X)$. The best predictor of Y is then given by

$$\mathbb{E}(Y|X) = \sum_{k=1}^{K} p_k(X)(\alpha_k + \beta_k^T X),$$

where

$$p_k(X) := \mathbb{E}(\delta_k|X) = \frac{\pi_k \varphi(X; \mu_k, \Sigma_k)}{\sum_{k=1}^{K} \pi_k \varphi(X; \mu_k, \Sigma_k)}.$$

Note that the averaging weights for the conditional group means are given by the function of X, so that the prediction of the response can adaptively use the covariate information to adjust the weights. In OMR, in contrast, the best predictor is given by the weighted average with π_k used as the fixed weights, i.e., $\sum_{k=1}^{K} \pi_k(\alpha_k + \beta_k^T X)$. Intuitively speaking, the more separated the covariate distribution is across the groups, the better the prediction performance of the JMR approach will be by adaptively changing the weights, compared to the OMR approach.

The sample can be clustered by assigning a subject to the group whose empirical posterior is the largest. For instance, the ith subject is assigned to the group

$$\underset{k}{\operatorname{argmax}} \frac{\widehat{\pi}_k \varphi(y_i; \widehat{\alpha}_k + \widehat{\beta}_k^T x_i, \widehat{\sigma}_k^2) \varphi(x_i; \widehat{\mu}_k, \widehat{\Sigma}_k)}{\sum_{k=1}^{K} \widehat{\pi}_k \widehat{\varphi}(y_i; \widehat{\alpha}_k + \widehat{\beta}_k^T x_i, \widehat{\sigma}_k^2) \varphi(x_i; \widehat{\mu}_k, \widehat{\Sigma}_k)}.$$

In Section 2.4, JMR and OMR are numerically compared. We show that not only JMR performs better when the covariate distribution is heterogeneous across the groups, OMR possesses little advantage over JMR even when the covariate distribution is homogeneous and OMR is the correctly specified approach. Furthermore, it is shown that OMR works no better than fitting a linear regression model in terms of the prediction performance. We further confirm these properties with a real data in Section 2.5.

Remark: One may claim that the assumption of the normality for the covariate distribution is too restrictive. Another way to see JMR is to treat it as flexible approximation to the unknown population distribution by a mixture that can account for the heterogeneity of the covariate distribution as well. Under this interpretation, the number of the mixture components work as a tuning parameter [Genovese and Wasserman (2000), Ghosal and van der Vaart (2001)]. In this chapter, we leave this aspect of the problem aside and assume that the population model is (3.1) and the number of the mixture components have a physical meaning.

2.2.2 The Functional Covariate Model

Let us now extend the joint mixture regression (JMR) model to incorporate the functional covariate into the model. Replacing the multivariate covariate in (4.7) with a random function $X(t) \in L^2[0,1]$ (for simplicity, assume its domain is $[0,1]$), we have

$$\begin{aligned} Y &= \sum_{k=1}^{K} \delta_k (\alpha_k + \langle \beta_k, X \rangle + \varepsilon_k) \\ X &= \sum_{k=1}^{K} \delta_k X_k, \end{aligned} \tag{2.3}$$

where $\beta_k(t) \in L^2[0,1]$. The inner product is defined by the usual L^2 inner product, i.e., $\langle f, g \rangle := \int_0^1 f(t) g(t) dt$. Let $X_k(t)$ be a Gaussian process with mean function $\mu_k(t)$ and covariance function $\Gamma_k(s,t)$. Assume that the covariance function of X, say Γ, allows the eigen-decomposition

$$\Gamma(s,t) = \sum_{j=1}^{\infty} \lambda_j \psi_j(s) \psi_j(t),$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\{\psi_1, \psi_2, \dots\}$ forms a complete orthonormal basis in $L^2[0,1]$ [Mercer's theorem, see Ash and Gardner (1975)]. Then, X allows the Karhunen-Loève decomposition

$$X(t) = \mu(t) + \sum_{j=1}^{\infty} \xi_j \psi_j(t), \tag{2.4}$$

where $\mu(t) := \mathbb{E} X(t) = \sum_{k=1}^{K} \pi_k \mu_k(t)$ and $\xi_j = \langle X - \mu, \psi_j \rangle$ [see Ash and Gardner (1975)]. ξ_j has mean 0 and variance λ_j, and ξ_j and $\xi_{j'}$ are independent for $j \neq$

j'; without the Gaussianity assumption, they are uncorrelated but not necessarily independent. Plugging (2.4) into (2.3) yields

$$Y = \sum_{k=1}^{K} \delta_k (a_k + \sum_{j=1}^{\infty} b_{kj} \xi_j + \varepsilon_k), \tag{2.5}$$

where $a_k = \alpha_k + \sum_{j=1}^{\infty} b_{kj} \langle \mu, \psi_j \rangle$ and $b_{kj} = \langle \beta_k, \psi_j \rangle$. Note that $\xi_j = \sum_{k=1}^{K} \delta_k \langle X_k - \mu, \psi_j \rangle$ and $(\langle X_k - \mu, \psi_1 \rangle, \langle X_k - \mu, \psi_2 \rangle, \dots)$ is a discrete Gaussian process, so that $(\xi_1, \xi_2, \dots) = \sum_{k=1}^{K} \delta_k (\langle X_k - \mu, \psi_1 \rangle, \langle X_k - \mu, \psi_2 \rangle, \dots)$ is a finite mixture of discrete Gaussian processes. Thus the model (2.5) can be viewed as generalization of the multivariate model (4.7) to the infinite-dimensional covariate model.

Unlike the model (4.7), the problem is now infinite dimensional and we do not directly observe ξ_j. To reduce the dimensionality we follow the commonly used approach in the functional data analysis literature [Müller and Stadtmüller (2005), Cai and Hall (2006), Hall and Horowitz (2007), Yao et al. (2011)]. With sufficiently large positive integer M, assume that β_k can be spanned by M leading eigenfunctions, i.e., $\beta_k(t) = \sum_{j=1}^{M} b_{kj} \psi_k(t)$ for $k = 1, \dots, K$. This assumption turns (2.5) into

$$Y = \sum_{k=1}^{K} \delta_k (a_k + b_k^{*T} \xi^* + \varepsilon_k), \tag{2.6}$$

where $b_k^* = (b_{k1}, \dots, b_{kM})^T$ and $\xi^* = (\xi_1, \dots, \xi_M)^T$. This is essentially equivalent to the multivariate covariate model, except that ξ^* is not directly observable. We estimate ξ^* and use its estimate as a surrogate.

In practice, the functional covariate is not continuously observable; only a finite number of observations at discrete points per curve are available. Suppose there are n realizations $(Y_1, X_1), \dots, (Y_n, X_n)$. The form of the sample available to us is

$$\{Y_1, X_1(t_{1,1}), \dots, X_1(t_{1,m_1})\}, \dots, \{Y_n, X_n(t_{n,1}), \dots, X_n(t_{n,m_n})\},$$

where $m_1, \dots, m_n$ are the numbers of observation points per curve, and the sets of the observation points are not necessarily synchronized nor equally discretized. From these observations, we have to estimate $\xi_1^*, \dots, \xi_n^*$. The analysis of the components of the Karhunen-Loève decomposition, i.e., $\lambda_j, \psi_j, \xi_{ij}$ $(i = 1, \dots, n,\ j = 1, 2, \dots)$, is called functional principal component analysis (FPCA), and has been developed for the past two decades by many authors. To save space, we avoid going into the details on the FPCA techniques, but interested readers may refer to Yao et al. (2005), Hall et al. (2006), Benko et al. (2009), and the references therein. The point is that we can estimate $\xi_1^*, \dots, \xi_n^*$ by using a technique in FPCA. In Section 2.5 where we apply the functional covariate model to a real-data analysis, we follow Ramsay and Silverman's paradigm of mapping a curve onto the space spanned by a finite number of basis functions [Ramsay and Silverman (2002, 2005); an excellent package "fda" is available for R and MATLAB].

Given estimates $\widehat{\xi_i^*}$, we can estimate the parameters in the model (2.6) by maximizing the estimated log-likelihood, where ξ_i^* is replaced with $\widehat{\xi_i^*}$ in the true log-likelihood, i.e.,

$$\ell_n(\Psi; y, \widehat{\xi^*}) = \sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} \pi_k \varphi(y_i; a_k + b_k^{*T} \widehat{\xi_i^*}, \sigma_k^2) \varphi(\widehat{\xi_i^*}; \mu_k, \Sigma_k) \Big\}, \tag{2.7}$$

where the parameters to be estimated are

$$\Psi = \{\pi_1, \dots, \pi_{K-1}, a_1, b_1^*, \sigma_1^2, \mu_1, \Sigma_1, \dots, a_K, b_K^*, \sigma_K^2, \mu_K, \Sigma_K\}.$$

Finally, the regression slope functions β_k can be estimated by

$$\widehat{\beta}_k(t) = \sum_{j=1}^{M} \widehat{b}_{kj} \widehat{\psi}_k(t),$$

which is a consistent estimator (Section 2.3).

The procedures for prediction and clustering are similar to the multivariate covariate model. In Section 2.5, we apply the functional covariate model to the Berkley growth study data.

2.2.3 Tricks to Improve the Model

In this section, we introduce two ways to improve the model. First, recall that the heterogeneity of the covariate distribution plays a crucial role in prediction because it allows us to estimate the membership from the covariate. The more separated the covariance distribution is across the groups, the better the prediction performance will be, because it becomes easier to differentiate the membership. It is well known in the functional data analysis literature that sometimes higher order derivatives, $X', X'', \cdots$, shows a clearer difference by group than the original X. A famous example given in Ramsay and Silverman (2005) is a velocity or acceleration curve, which shows much clearer distinction between gender than a growth curve. One way to incorporate, say X', into the model is to apply integration by parts to (2.3), which yields

$$\begin{aligned} Y &= \sum_{k=1}^{K} \delta_k(\alpha_k + \langle \beta_k, X \rangle + \varepsilon_k) \\ &= \sum_{k=1}^{K} \delta_k(\alpha_k - \gamma_k(1)X(1) + \langle \gamma_k, X' \rangle + \varepsilon_k), \end{aligned}$$

where $\gamma_k(t) = -\int_0^t \beta_k$. These two expressions are identical in theory, but in practice the latter may perform better if X' is more distinguishable by group than X. As the constraint between the regression coefficient of $X(1)$ and the regression slope function of X' is inconvenient to estimate the model, we may avoid it by treating the regression coefficient as a free parameter, i.e.,

$$Y = \sum_{k=1}^{K} \delta_k(\alpha_k + \zeta_k X(1) + \langle \gamma_k, X' \rangle + \varepsilon_k),$$

where ζ_k is a free parameter. This model includes (2.3) as a submodel.

Another way to extend the model (2.3) is to allow two kinds of covariates: one that behaves similarly across the groups or is deterministic, and the other that behaves differently across the groups—sometimes we know beforehand that a certain covariate, say Z, has an invariant distribution or is deterministic such as covariates in experiments. Adding Z to the model (4.7), it becomes

$$Y = \sum_{k=1}^{K} \delta_k(\alpha_k + \zeta_k^T Z + \beta_k^T X + \varepsilon_k).$$

Under this model, the inference should be based on the conditional distribution given Z so that we can exclude unnecessary parameters from the model that does not contribute to the membership estimation. The EM algorithm can be straightforwardly modified to accommodate this model. These two extensions are simple, yet very effective to improve the prediction performance, as demonstrated in Section 2.5.

2.3 Theoretical Properties

2.3.1 Consistency of the MLE in the Functional Covariate Model

The maximizer of (2.7) does not coincide with the actual maximum likelihood estimator because ξ^* is replaced with $\widehat{\xi}^*$. Fortunately, the theory in Yao et al. (2011) straightforwardly applies to the current problem as well; under some regularity conditions the maximum likelihood estimator in (2.7) is still consistent. We assume that $\widehat{\xi}_k, \widehat{\psi}_k, k = 1, \ldots, M$, are obtained by using the technique in Yao et al. (2005).

Proposition II.1. *Assume that the population model is (2.6) and the assumptions*

A1 to A4 in Yao et al. (2011) hold. For any fixed compact set containing the true parameter Ψ *as an interior point, let* $\widehat{\Psi}$ *be the maximizer of (2.7) over the compact set. Then,* $\widehat{\Psi}$ *converges to* Ψ *in probability. Furthermore,* $\widehat{\beta}_k, k = 1, \ldots, K$ *is uniformly consistent, i.e.,* $\sup_{t \in [0,1]} |\widehat{\beta}_k(t) - \beta_k(t)|$ *converges to 0 in probability,*

There are two aspects of the model that involve the proof: the proximity of $\widehat{\xi}_k$ to ξ_k and the local behavior of the log-likelihood function

$$\ell(\Psi; y, \xi) = \log \Big\{ \sum_{k=1}^{K} \pi_k \varphi(y; a_k + b_k^T \xi, \sigma_k^2) \varphi(\xi_k; \mu_k, \Sigma_k) \Big\}.$$

Since the covariate distribution continues to satisfy the assumptions in Yao et al. (2011), the conditions concerning the first aspect are satisfied. On the other hand, since the likelihood function in JMR has a different form than the one in OMR (Yao et al. (2011) considered the functional covariate model for OMR), we need to check whether the current likelihood still retains appropriate local behavior. In Appendix 2.7.1, we verify that the log-likelihood function in JMR also satisfies the regularity conditions.

2.3.2 Asymptotic Mean Squared Prediction Error

As mentioned before, it is essential to estimate the membership from the covariate in order to predict the response well. In this section, we compare the asymptotic mean square prediction error (MSPE) between JMR and OMR. Recall that we predict the response by the empirical best predictor

$$\widehat{Y} = \sum_{k=1}^{K} \widehat{p}_k(X)(\widehat{\alpha}_k + \widehat{\beta}_k^T X), \tag{2.8}$$

where in JMR

$$\widehat{p}_k(X) = \frac{\widehat{\pi}_k \varphi(X; \widehat{\mu}_k, \widehat{\Sigma}_k)}{\sum_{k=1}^{K} \widehat{\pi}_k \varphi(X; \widehat{\mu}_k, \widehat{\Sigma}_k)}, \tag{2.9}$$

while in OMR $\widehat{p}_k(X) = \widehat{\pi}_k$. As seen in the last section, the MLE is consistent under the JMR model. Now, suppose that the population model is the JMR model, but the MLE is obtained by applying the OMR approach. It may be reasonable to suspect that the resulting MLE is no longer consistent. However, several numerical explorations that the author conducted including those given in Section 2.4 suggest that the MLE obtained by applying the OMR approach is also consistent. (We have not been successful in proving either this conjecture is true or false.) We will come back to this point again in the next section. In the following, we consider two cases concerning the OMR approach: one where the parameters are consistently estimated, and the other where the parameters are not consistently estimated.

Consider the multivariate covariate model (4.7). For simplicity, let $\alpha_1 = \cdots = \alpha_K = 0$ and use the inner-product notation, i.e., $\langle x, y \rangle = x^T y$. If the covariate distribution varies across the groups, (2.8) provides the smallest asymptotic MSPE among any possible predictors because it is the MSPE of the population conditional mean. The asymptotic MSPE is then given by the error variance, $\mathit{\Sigma} := \sum_{k=1}^{K} \pi_k \sigma_k^2$, plus

$$\mathbb{E}\Big[\sum_{k=1}^{K} \mathbb{E}(\delta_k|X)\Big(\sum_{\ell=1}^{K} \mathbb{E}(\delta_\ell|X)\langle \beta_k - \beta_\ell, X\rangle\Big)^2\Big]. \tag{2.10}$$

If we use $\widehat{p}_k(X) = \widehat{\pi}_k$, the asymptotic MSPE becomes $\mathit{\Sigma}$ plus

$$\mathbb{E}\Big[\sum_{k=1}^{K} \mathbb{E}(\delta_k|X)\Big(\sum_{\ell=1}^{K} \mathbb{E}(\delta_\ell)\langle \beta_k - \beta_\ell, X\rangle\Big)^2\Big], \tag{2.11}$$

which is strictly greater than (2.10) unless $\mathbb{E}(\delta_k|X) = \mathbb{E}(\delta_k)$ for all k, nor $\langle \beta_1, X\rangle =$

$\cdots = \langle \beta_K, X\rangle$ almost surely. The former case implies that the covariate distribution is invariant across the groups, which contradicts the assumption. In the latter case, (2.10) = (2.11) = 0; but in this case the ability to differentiate the group is not necessary because there is no harm by assuming a wrong group.

Now, suppose that the MLE is asymptotically biased and $\widehat{\beta}_k$, $\widehat{\pi}_k$ converges to some β_k^*, π_k^*, respectively. Then, the asymptotic MSPE becomes Σ plus

$$\sum_{k=1}^{K} \pi_k \mathbb{E}\Big[\Big\langle \sum_{\ell=1}^{K} \pi_\ell^*(\beta_k - \beta_\ell^*), X_k \Big\rangle^2\Big]. \tag{2.12}$$

This quantity is in fact greater than (2.11) at least when $\mathbb{E}(X_1X_1^T) = \cdots = \mathbb{E}(X_KX_K^T)$. Without this assumption, the effect of the bias is rather involved as it is easy to create an example where (2.12) is smaller than (2.11). The proofs are given in Appendix 2.7.2.

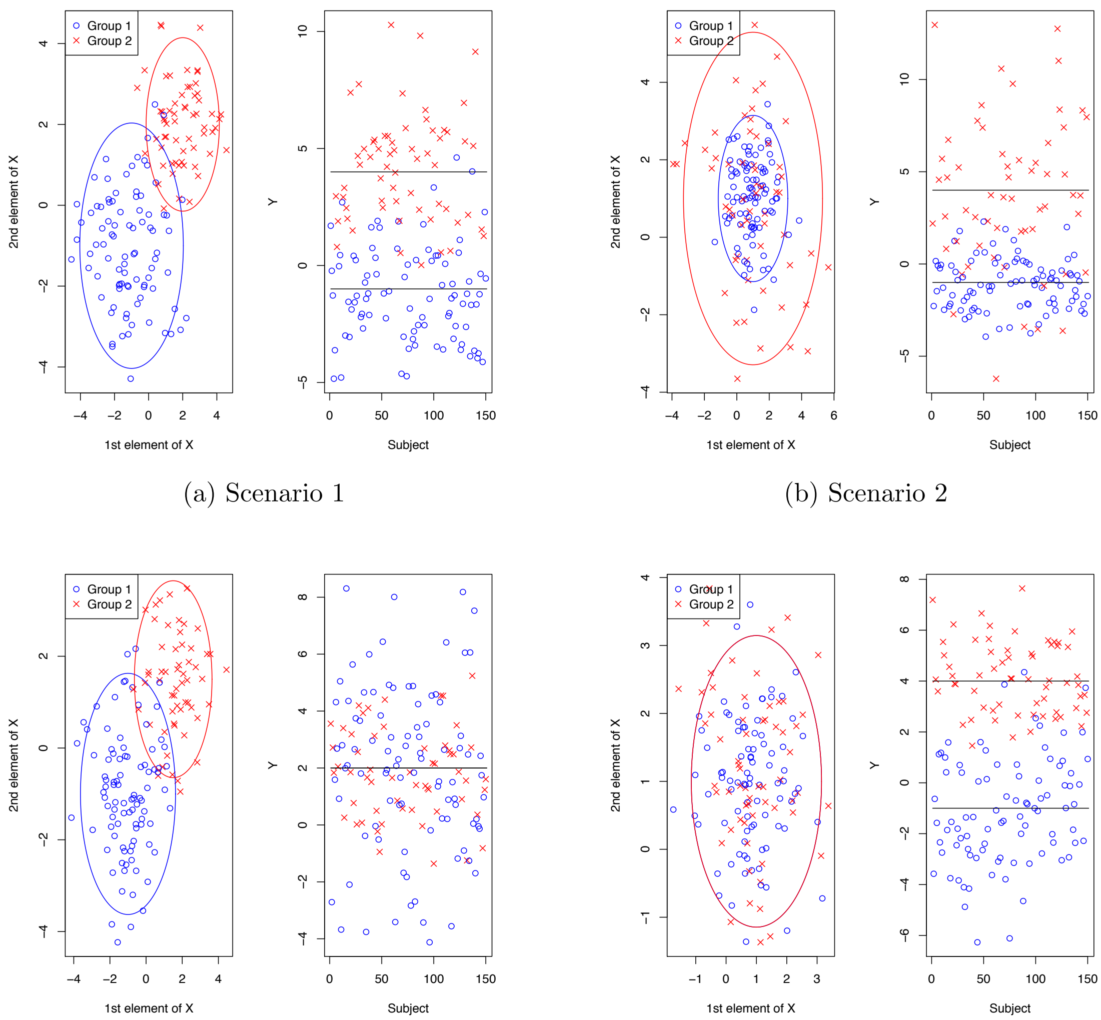
2.4 Simulation Study

This section illustrates how the JMR approach works in comparison to alternative methods. We generate a sample from the two-dimensional covariate, two-group model

$$\begin{aligned} Y &= \delta_1(\alpha_1 + \beta_1^T X + \varepsilon_1) + (1-\delta_1)(\alpha_2 + \beta_2^T X + \varepsilon_2) \\ X &= \delta_1 X_1 + (1-\delta_1) X_2, \end{aligned}$$

where the mixing proportion is $(\pi_1, \pi_2) = (0.6, 0.4)$ and the error variances are both 0.3^2. The training sample size is considered for 100 and 300, and the testing sample size is 500. The other parameters—regression coefficients, covariate means, and covariate variance-covariance matrices—are determined to construct the following four scenarios:

1. X and Y are both well separated by group.



(a) Scenario 1

(b) Scenario 2

(c) Scenario 3

(d) Scenario 4

Figure 2.1: A realization of the training sample of n=150 for each scenario. The covariate is plotted in the left figure where the two circles show .95th quantile contours. The response is plotted in the right figure where the two horizontal lines are the population means.

2. X has the common group means, and Y is well separated by group.

3. X is well separated by group, but Y is not.

4. X has the common cluster distributions, and Y is well separated by group.

Figure 2.1 shows a realization of the sample in each scenario. We calculate the mean squared prediction error (MSPE), the average misclassification rate (MCR), and the mean squared error (MSE) for part of the parameters over 500 iterations. We compare JMR to three alternative approaches: linear regression by ordinary least squares (OLS), ordinary mixture regression (OMR), and the two-step model-based clustering approach (MBC). MCR cannot be computed for OLS as it does not cluster a sample. MBC works as follows. First, it clusters a sample into two groups by fitting a mixture of normal to the covariate (model-based clustering); these two groups are used to compute the MCR. Second, in each cluster the linear regression model is estimated by fitting OLS. To predict the response, it uses the weighted average of the linear predictors from the two estimated linear models with the posterior probabilities calculated from a new observation of the covariate used as weights. We used Fraley and Raftery's R package "mclust" for this approach [Fraley and Raftery (2002)]. Note that JMR is the correctly specified approach in Scenarios 1–3 while OMR is the correctly specified approach in Scenario 4. It is well known that the clusters obtained in MBC are not identically-distributed samples of the component distributions, so that the estimates based on the resulting clusters are inevitably biased.

The results are given in Tables 2.1 and 2.2. We first look at the prediction performance (Table 2.1). When the covariate distribution is well separated across the groups (Scenarios 1 and 3), JMR and MBC outperform the other two methods. When it is difficult to differentiate the group by the covariate (Scenario 2) or the covariate distribution is homogeneous (Scenario 4), the overall performance deteriorates and the relative advantage of JMR reduces. Note that OMR is not even as good as OLS (Scenarios 1–3), and in Scenario 4 where OMR is the correctly specified approach, it is not noticeably better than the other approaches. The reason why OMR is no better than OLS is that computing the average over the linear predictors of the groups with fixed weights is essentially equivalent to fitting the linear model globally; then OLS tends to have a smaller variation because it needs to estimate much fewer parameters than OMR.

Method	OLS	OMR	JMR	MBC
MSPE	1.66	2.25	0.35	0.44
MCR		.074	.012	.057

(a) Scenario 1 (n=100)

Method	OLS	OMR	JMR	MBC
MSPE	1.62	2.23	0.31	0.33
MCR		.067	.009	.042

(b) Scenario 1 (n=300)

Method	OLS	OMR	JMR	MBC
MSPE	10.46	12.35	9.50	11.05
MCR		.022	.023	.331

(c) Scenario 2 (n=100)

Method	OLS	OMR	JMR	MBC
MSPE	10.11	12.21	9.22	10.43
MCR		.019	.022	.274

(d) Scenario 2 (n=300)

Method	OLS	OMR	JMR	MBC
MSPE	4.08	4.79	1.19	1.35
MCR		.048	.063	.080

(e) Scenario 3 (n=100)

Method	OLS	OMR	JMR	MBC
MSPE	3.97	4.73	0.88	0.98
MCR		.041	.013	.061

(f) Scenario 3 (n=300)

Method	OLS	OMR	JMR	MBC
MSPE	8.54	8.33	8.94	8.89
MCR		.015	.016	.441

(g) Scenario 4 (n=100)

Method	OLS	OMR	JMR	MBC
MSPE	8.37	8.31	8.52	8.57
MCR		.015	.016	.449

(h) Scenario 4 (n=300)

Table 2.1: The mean squared prediction error (MSPE) and the average misclassification rate (MCR) for four scenarios.

The results with respect to misclassfication seem a little different. The clustering performance by JMR is fairly well throughout the scenarios, including Scenario 4. OMR also works well when the covariate distribution is not much separated (Scenarios 2 and 4). It is even slightly better than JMR in Scenario 2 where OMR is a misspecified approach. For scenarios 1 and 3, in contrast, JMR works much better than OMR, though the overall performance of OMR is still comparable to MBC regardless of the fact that OMR does not take into account the heterogeneity of the covariate distribution. This implies that to cluster a sample whose clustering structure lies in the regression structure, clustering based on the regression is at least as

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.056	.048	.074
$\beta_{12}[1]$	.063	.052	.279
$\beta_{22}[2]$	.064	.049	.206

(a) Scenario 1 (n=100)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.037	.029	.039
$\beta_{12}[1]$	.033	.028	.166
$\beta_{22}[2]$	.039	.029	.150

(b) Scenario 1 (n=300)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.049	.051	.245
$\beta_{12}[-1]$	.026	.026	2.78
$\beta_{22}[2]$	.025	.026	4.54

(c) Scenario 2 (n=100)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.029	.031	.188
$\beta_{12}[-1]$	.015	.014	.362
$\beta_{22}[2]$	.015	.015	.992

(d) Scenario 2 (n=300)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.057	.089	.096
$\beta_{12}[-2]$	.164	.489	.618
$\beta_{22}[1]$	.183	.584	.368

(e) Scenario 3 (n=100)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.030	.032	.049
$\beta_{12}[-2]$	.032	.133	.429
$\beta_{22}[1]$	.030	.151	.204

(f) Scenario 3 (n=300)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.051	.051	.207
$\beta_{12}[-2]$	.053	.054	2.50
$\beta_{22}[1]$	.053	.053	3.82

(g) Scenario 4 (n=100)

Method	OMR	JMR	MBC
$\pi_1[0.6]$	.030	.031	.176
$\beta_{12}[-2]$	.028	.028	2.57
$\beta_{22}[1]$	.028	.028	2.48

(h) Scenario 4 (n=300)

Table 2.2: The square root of the mean squared error for some of the parameters in the four scenarios. The true parameter is given in the square brackets next to the symbol.

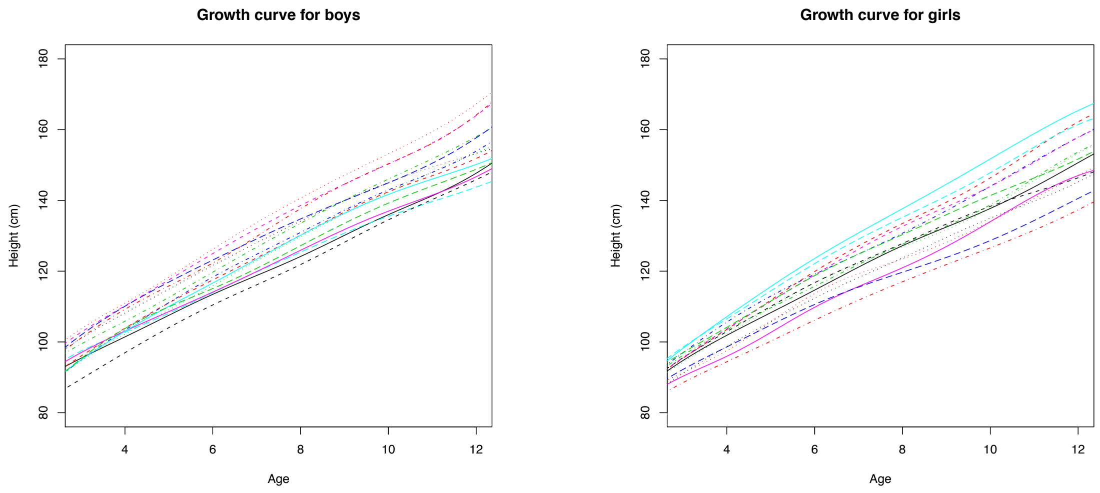
equally important as taking into account the covariate heterogeneity.

One may wonder whether the differences in the prediction performance in fact attribute to the estimability of the group. Table 2.2 shows the square root of the mean squared error for some of the parameters. Note that there is not much difference in the estimation performance between OMR and JMR; in some cases, OMR is even better than JMR. As the sample size increases, the MSE of OMR reduces at a similar rate to JMR. This raises the question as to the consistency of OMR; although OMR is a misspecified approach under the JMR model, the MLE by OMR may be still consistent for the parameters under the JMR model. We numerically investigated this conjecture, and the results seem to support it. (We have not been able to prove analytically whether this claim is true or not.) Because the parameter estimation by the two methods seems similar, we claim that the difference in the prediction performance mostly attributes to the estimability of the group. In other words, whether we can predict the response well largely depends on whether we can estimate the group that the subject of a new observation is likely to belong to from the covariate. Otherwise, we cannot expect much beyond simply fitting a linear regression model.

2.5 Berkeley Growth Study, Revisited

In this section, we present a real-data example where the joint mixture regression (JMR) approach improves the prediction performance of the traditional ordinary mixture regression (OMR) approach. We use the Berkeley growth study data [Tuddenham and Snyder (1954)], which contains the recorded height of boys and girls from age 1–18 years old; this is a well-analyzed data set and has been repeatedly used as an illustrating example in the functional data analysis literature. Recent examples using this data set include Chiou and Li (2007), Tang and Müller (2008), Hall et al. (2009), and Yao et al. (2011). The data set contains 39 boys and 54 girls whose height was measured quarterly from 1–2 years old, annually from 2–8 years old, and biannually from 8–18 years old. We reconsider the analysis given in Yao et al. (2011), where they considered the problem of predicting the height at the age of 18 from the height transition during the juvenile period.

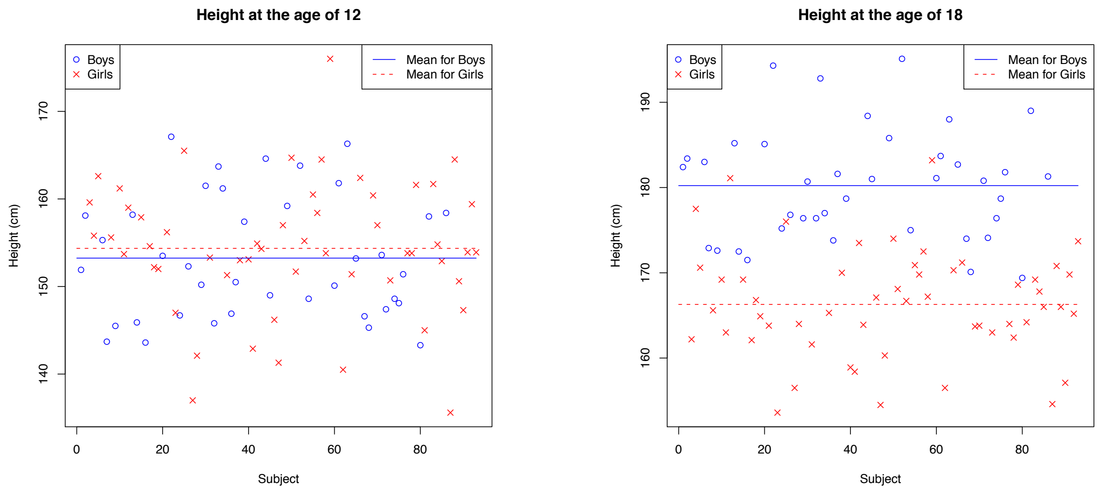
We first consider the model where the predictor is a growth curve from 3–12 years old (see Figure 2.2), which is the model that Yao et al. (2011) considered (referred



(a) Boys

(b) Girls

Figure 2.2: The growth curves for randomly selected 15 boys and 15 girls from age 3-12 years old. The curves are obtained by mapping observations onto B-splines of order 5.



(a) Age 12

(b) Age 18

Figure 2.3: Heights at the age of 12 (Left) and 18 (Right). The order is determined randomly. The vertical lines are the sample means for boys and girls.

as Model 1). This age period usually contains female pubertal growth peaks near the end of the range; male pubertal growth peaks usually come several years later. Given the juvenile growth curve of a new subject, we wish to predict the height at his or her age of 18. Figure 2.3 shows the height at the age of 12 and 18. It can be seen that predicting the height at the age of 18 from the height at the age of 12 is challenging as there is no significant difference in the height distribution at the age of 12 between boys and girls. Thus to predict the height well it is crucial to differentiate gender from the growth curve; recall that we do not assume that gender information is available (Yao et al. (2011) claims that JMR works better than simply fitting a linear regression model, but we suspect that they used gender information when predicting the response even though they did not use it when fitting the model). We predict the response by the empirical best predictor (2.8). In addition to OMR and JMR, we also consider functional principal component regression (PCR) as an alternative approach for comparison [Cai and Hall (2006), Hall and Horowitz (2007)]. PCR estimates the linear model so that it does not cluster a sample. For these three methods—PCR, OMR, and JMR—we calculated leave-one-curve-out cross validation (CV),

$$CV = \frac{1}{93}\sum_{i=1}^{93}(Y_i - \widehat{Y}_{(-i)})^2,$$

where $\widehat{Y}_{(-i)}$ is calculated by: first estimating the parameters from the entire sample except the ith subject, and then computing the predictor from X_i. The results are given in Table 2.3a; there are several points that are consistent with what we saw in the simulation study. Note that JMR displays its advantage over the other methods when using four or more eigenfunctions while using only two or three eigenfunctions it is not as good as PCR. Now, looking at Figure 2.4a where the scatterplots for the estimated standardized principal component (PC) scores labeled by gender are shown, it can be seen that the first three PC scores are not well separated by gender while the fourth PC score seems to show some heterogeneity between gender. Also, looking at Table 2.4a, which shows the number of the misclassification for gender, it can be seen that JMR clusters the sample by gender very well no matter how many eigenfunctions are used while OMR suddenly behaves poorly when using the fourth eigenfunction whose

# of eigenfunctions	2	3	4	5
PCR	48.785	40.460	27.695	26.465
OMR	53.783	47.521	27.940	28.421
JMR	50.412	42.369	26.618	22.901
CumVar	(0.9857)	(0.9932)	(0.9975)	(0.9993)

(a) Model 1

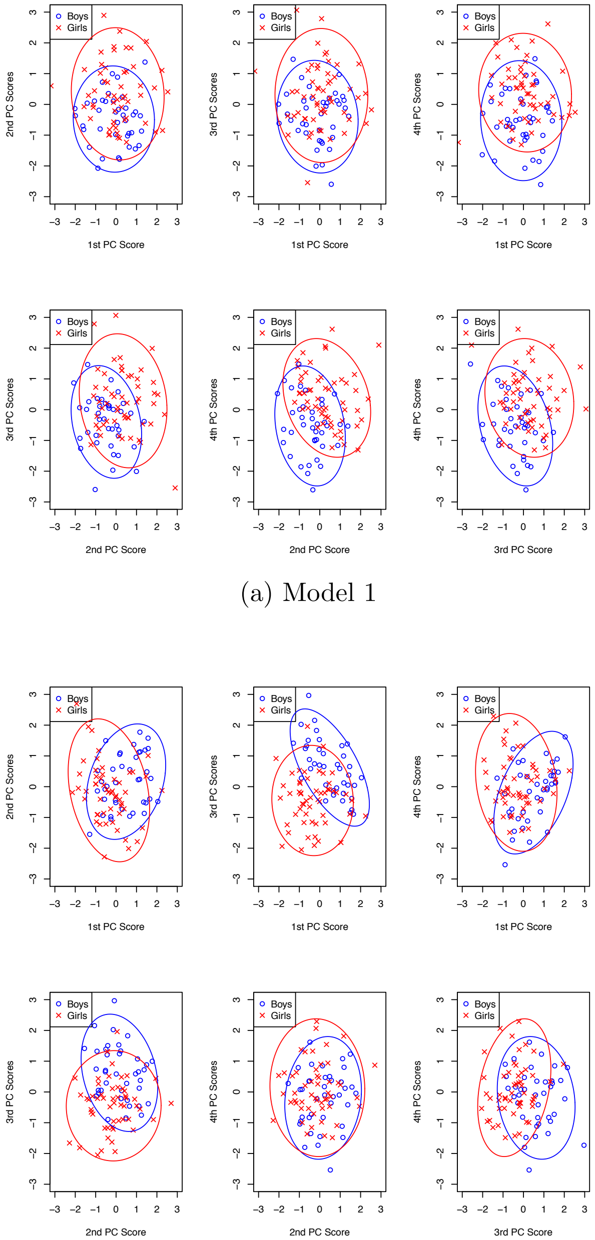
# of eigenfunctions	2	3	4	5
PCR	34.241	21.634	20.682	20.976
OMR	36.617	21.638	22.505	23.012
JMR	32.888	17.889	16.929	15.293
CumVar	(0.6584)	(0.7830)	(0.8908)	(0.9882)

(b) Model 2

Table 2.3: Cross Validation using the Berkeley Growth Data. The last row shows the proportion of the cumulative variance explained by the used eigenfunctions in the total variation (the sum of the eigenvalues).

PC score shows the differentiability between gender. These observations support the theory that the prediction performance of JMR depends on the heterogeneity of the covariance distribution. As we saw in the simulation study, OMR performs no better than PCR. We may wonder if there is a way to improve the model so that JMR performs the best regardless of the number of the eigenfunctions to be used. In fact, as seen in Figure 2.5a, which shows the cross-validated predictors from the leave-one-curve-out samples using three leading eigenfunctions, JMR suffers from a bias by gender (most of the male heights locate below the diagonal line while most of the female heights locate above it). We want JMR to perform in the way that it reduces this group bias by estimating the membership well. In the second part of this section, we explore an alternative model that uses the tricks we introduced in Section 2.2.3.

As mentioned in Ramsay and Silverman (2005), a velocity curve, or an acceleration curve, shows much clearer distinction by gender than the original growth curve does. We incorporate the velocity curve into the model by the way we introduced

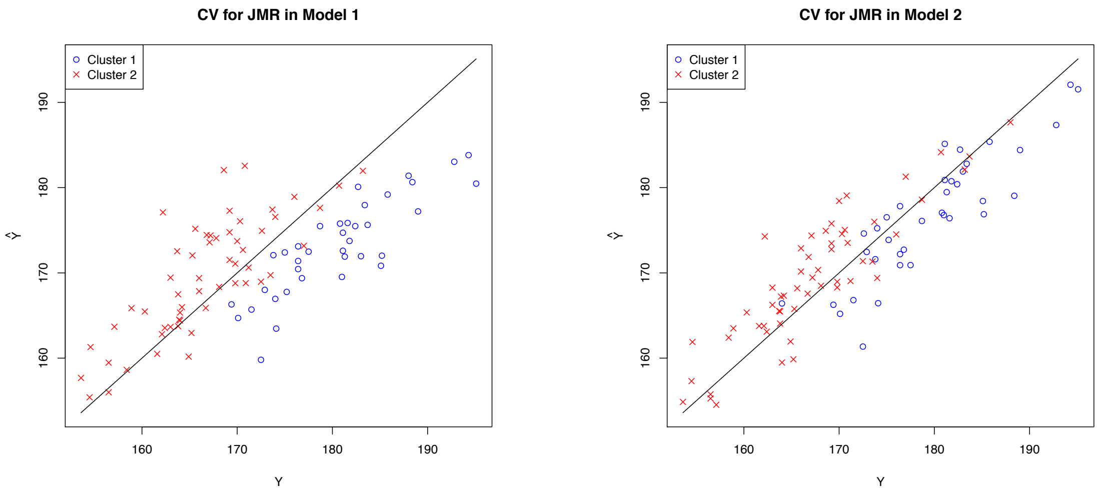


(a) Model 1

(b) Model 2

Figure 2.4: Scatterplots for the combinations of the standardized PC scores. The circles indicate .95th normal-quantile contours transformed by the sample mean and variance-covariance matrix.

in Section 2.2.3. In particular, we use the velocity curve from 3–12 years old as the functional covariate and the height at the age of 12 as the scalar covariate (referred as Model 2). We do, however, treat the latter covariate as an invariant covariate since



(a) Model 1

(b) Model 2

Figure 2.5: The cross-validated predictors from the one-curve-out samples using joint mixture regression with three leading eigenfunctions.

# of eigenfunctions	2	3	4	5
OMR	5	6	27	25
JMR	5	6	6	5

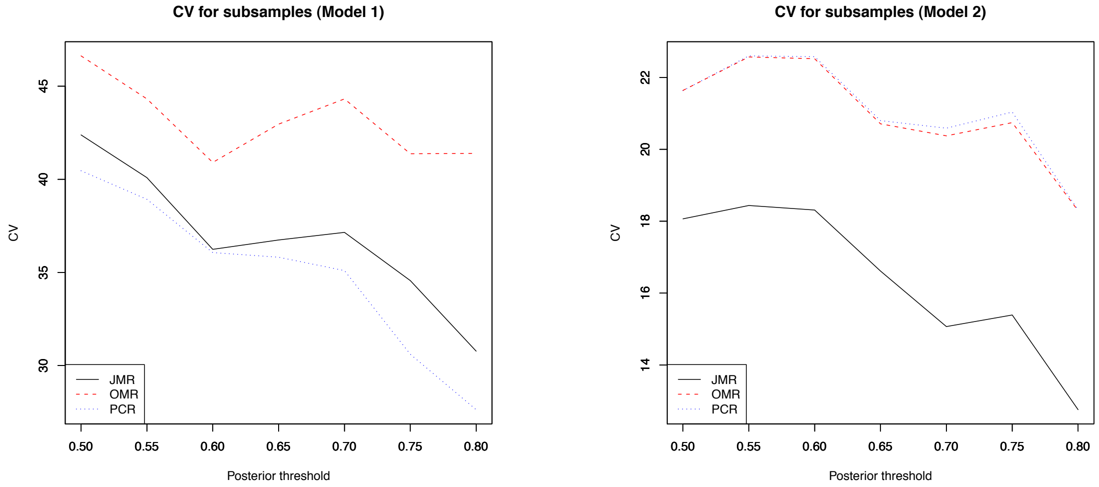
(a) Model 1

# of eigenfunctions	2	3	4	5
OMR	14	27	31	34
JMR	8	8	4	6

(b) Model 2

Table 2.4: The number of misclassifications based on the gender (n=93).

the heights at the age of 12 for boys and girls are very similar and almost impossible to differentiate (recall Figure 2.3a). Thus it is crucial to estimate gender from the velocity curve to improve the prediction performance. For PCR and OMR, we simply use these two variables as covariates. The difference between OMR and JMR under this model is whether we incorporate the distribution of the velocity curve into the model. The results are given in Table 2.3b. First, we notice that the overall prediction performance has dramatically improved from Model 1. In particular, JMR outperforms the other two approaches regardless of how many eigenfunctions are used. Looking



(a) Model 1

(b) Model 2

Figure 2.6: Cross validation comparison for subsamples possessing the estimated posteriors greater than or equal to threshold values.

at Figure 2.4b where the scatterplots for the estimated standardized PC scores of the velocity curve are shown, the leading PC scores are much more differentiable by gender than those of the growth curve (cf. Figure 2.4a). Also, Table 2.4b shows that JMR clusters the sample by gender fairly well while OMR no longer do so no matter how many eigenfunctions are uses. OMR again performs no better than PCR. Note that JMR keeps improving the prediction performance with more eigenfunctions used while PCR and JMR are stuck at the use of three of four eigenfunctions. Finally, Figure 2.5b shows that JMR under Model 2 considerably reduces the bias by gender.

Now, we may wonder how large posterior probabilities in the JMR approach actually contribute to improve the prediction. To see this, we calculate CV for the subsamples whose estimated posteriors are larger than a certain threshold. In this analysis, we first estimate the parameters from a leave-one-curve-out sample and compute the posteriors (2.9) for the subject that is left out. Then, we compute the mean squared prediction errors by collecting only those subjects whose greater posterior is larger than a predetermined threshold. Figure 2.6 shows the transition of the CV along different thresholds for the two models using three leading eigenfunctions. The thresholds used here are from 0.5 through 0.8 by 0.05 with which the resulting subsample sizes are respectively 93, 80, 69, 60, 50, 41, 33 for Model 1 and 93, 86, 84, 79, 73, 69, 59 for Model 2 (0.5 corresponds to the whole sample). Overall, Model 2 provides larger posteriors than Model 1 (at each threshold, the subsample size in Model 2 is larger than that in Model 2). This is consistent with the fact that the three PC scores in Model 2 behaves more differently by gender than those in Model 1 as seen in Figure 2.4. As seen in Figure 2.6b, in Model 2, JMR improves the prediction performance with a faster rate than the other two methods as the threshold increases. In contrast, Figure 2.6a does not display such behavior; in fact, PCR performs always better than the other two. This implies that under Model 2, JMR improves the prediction performance more than the other two by estimating the membership from the covariate that is heterogeneous by gender.

2.6 Discussion

In this chapter, we introduced a mixture regression model where the joint distribution of the response and the covariate is modeled as a mixture. We call it joint mixture regression in contrast to the traditional mixture regression, which we call ordinary mixture regression. By incorporating the covariate distribution into the model, the heterogeneity of the covariate distribution across the groups is also taken into account. From a new observation of the covariate, we can compute the posterior probabilities that the subject belongs to each group. Using these posterior probabilities, the prediction of the response can adaptively use the covariate. Through the simulation studies and the real-data analysis using the Berkeley growth study data, we showed that in order to predict the response well, it is crucial that the covariate behaves differently across the groups. If the covariate behaves similarly or is deterministic, the mixture regression approach performs no better than simply fitting a linear regression model. By including the covariate that behaves differently across the group, we showed that our approach can significantly improve the prediction performance from the traditional mixture regression approach.

We conclude this chapter with two question. First, as we saw in the simulation study the MLE obtained by fitting the ordinary mixture regression model may be consistent even under the joint mixture regression model. We conducted a large number of simulation studies, including the one given in this chapter, and they all seem to support this conjecture. Can we analytically examine the genuineness of this conjecture? Second, in the functional covariate model we used the eigenfunctions of the observed data as basis functions onto which the functional covariate is mapped. However, any basis functions can be used in this procedure. The best basis functions should be the ones where the projections have the distribution most separable across the groups so that it becomes easy to estimate the membership from them. Though using the eigenfunctions of the observed covariate makes an intuitive sense, analytical justification is lacking. What basis functions yield the best projection in the joint mixture regression model? We leave these two questions to be solved in the future.

2.7 Appendix: Proofs

2.7.1 Consistency of the MLE in the Functional Covariate Model

We need to verify if the likelihood function under the joint mixture regression model behaves appropriately. Recall that the log-likelihood function is given by

$$\ell(\Psi; y, \xi) = \log \Big\{ \sum_{k=1}^{K} \pi_k \varphi(y; a_k + b_k^T \xi, \sigma_k^2) \varphi(\xi; \mu_k, \Sigma_k) \Big\}.$$

The regularity conditions given in Yao et al. (2011) are as follows. For any Ψ_1 in a pre-fixed compact set defined in the proposition:

(B1) There exist some functions $g(y, \xi, \Psi)$ and $c(\Psi)$ such that, for all possible values of y, ξ', ξ'' and $\Psi \in N_{\Psi_1}$, where N_{Ψ_1} is some neighborhood of Ψ_1,

$$\|\ell(\Psi; y, \xi') - \ell(\Psi; y, \xi'')\| \leq g(y, \xi, \Psi)\|\xi' - \xi''\| + c(\Psi)\|\xi' - \xi''\|^2,$$

and $g(y, \xi, \Psi)$ and $c(\Psi)$ satisfy

$$\sup_{\Psi \in N_{\Psi_1}} \mathbb{E}[g^2(y, \xi, \Psi)] < \infty,$$
$$\sup_{\Psi \in N_{\Psi_1}} c(\Psi) < \infty,$$

where the integration is defined by the true parameters.

(B2.1) $\ell(\Psi; y, \xi)$ is upper semicontinuous in $\Psi \in N_{\Psi_1}$ for all (y, ξ).

(B2.2) There exists a function $D(y, \xi)$ such that $\mathbb{E}D(y, \xi) < \infty$ and $\ell(\Psi; y, \xi) \leq D(y, \xi)$ for all (y, ξ) and $\Psi \in N_{\Psi_1}$.

(B2.3) For $\Psi \in N_{\Psi_1}$ and sufficiently small $r > 0$, $\sup_{\Psi': \|\Psi' - \Psi\| < r} q(y, \xi, \Psi')$ is measurable in (y, ξ).

It is easy to see that (B2.1)–(B2.3) are satisfied. By setting

$$\begin{aligned} g(y, \xi, \Psi) &= \sum_{k=1}^{K} \Big[\frac{\|b_k\|}{\sigma_k^2}\{|y - a_k| + \|b_k\|\|\xi^{*\prime}\|\} + \lambda_{\max}(\Sigma_k)\{\|\xi^{*\prime}\| + \|\mu_k\|\}\Big] \\ c(\Psi) &= \sum_{k=1}^{K} \Big(\frac{\|b_k\|^2}{\sigma_k^2} + \lambda_{\max}(\Sigma_k)\Big), \end{aligned}$$

where $\lambda_{\max}(\Sigma_k)$ is the maximum eigenvalue of Σ_k, (B1) is also satisfied, and all the regularity conditions are satisfied by the likelihood in problem as well.

2.7.2 Asymptotic Mean Squared Prediction Error

We first confirm that (2.11) $\geq$ (2.10) where the equality holds only when $\mathbb{E}(\delta_k|X) = \mathbb{E}(\delta_k)$ for all k, or $\langle \beta_1, X\rangle = \cdots = \langle \beta_K, X\rangle$ almost surely. Denoting $\mathbb{E}(\delta_k|X)$ by p_k

and $\langle \beta_k, X\rangle$ by e_k, the inside of the expectation operator in (2.11) $-$ (2.10) is given by

$$L := \sum_{k=1}^{K} p_k\{\sum_{\ell=1}^{K} \pi_\ell(e_k - e_\ell)\}^2 - \sum_{k=1}^{K} p_k\{\sum_{\ell=1}^{K} p_\ell(e_k - e_\ell)\}^2.$$

Since $\pi_K = 1 - \pi_1 - \cdots - \pi_{K-1}$, for $j = 1, \ldots, K-1$,

$$\begin{aligned}\frac{\partial L}{\partial \pi_j} &= 2\sum_{k=1}^{K} p_k(e_k - e_j)\sum_{\ell=1}^{K} \pi_\ell(e_k - e_\ell) - 2\sum_{k=1}^{K} p_k(e_k - e_K)\sum_{\ell=1}^{K} \pi_\ell(e_k - e_\ell)\\ &= 2(e_K - e_j)(\sum_{k=1}^{K} p_k e_k - \sum_{\ell=1}^{K} \pi_\ell e_\ell),\end{aligned}$$

which is zero at $\pi_j = p_j$. Furthermore,

$$\frac{\partial^2 L}{\partial \pi_j \partial \pi_{j'}} = 2(e_K - e_j)(e_K - e_{j'}),$$

thus $[\frac{\partial^2 L}{\partial \pi_j \partial \pi_{j'}}]_{j,j'=1,\ldots,K-1}$ is strictly positive definite unless $e_1 = \cdots = e_K$, and the conclusion follow.

We now confirm the other claim. Note that (2.12) can be rewritten to

$$\mathrm{I} := \sum_{k=1}^{K} \pi_k \mathbb{E}[\langle \sum_{\ell=1}^{K} \pi_\ell^*(\beta_k - \beta_\ell) + \mathcal{B}, X_k\rangle^2],$$

where $\mathcal{B} := \sum_{\ell=1}^{K} \pi_\ell^*(\beta_\ell - \beta_\ell^*)$, while (2.11) can be rewritten to

$$\mathrm{II} := \sum_{k=1}^{K} \pi_k \mathbb{E}\Big[\Big\langle \sum_{\ell=1}^{K} \pi_\ell(\beta_k - \beta_\ell), X_k\Big\rangle^2\Big]$$

We will prove $\mathrm{I} - \mathrm{II} \geq 0$ under the assumption $\Gamma = \mathbb{E}(X_1 X_1^T) = \cdots = \mathbb{E}(X_K X_K^T)$.

Observe

$$\begin{aligned}
\mathrm{I} &= \sum_{k=1}^{K} \pi_k \mathbb{E}[\langle \sum_{\ell=1}^{K} \pi_\ell^*(\beta_k - \beta_\ell) + \mathcal{B}, X_k \rangle^2] \\
&= \sum_{k=1}^{K} \pi_k \mathbb{E}[\langle (\beta_k - \beta_K) + \sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell) + \mathcal{B}, X_k \rangle^2] \\
&= \sum_{k=1}^{K} \pi_k \{(\beta_k - \beta_K) + \sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell)\}^T \Gamma \{(\beta_k - \beta_K) + \sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell)\} \\
&\quad + 2\mathcal{B}^T \Gamma \sum_{k=1}^{K} \pi_k \{(\beta_k - \beta_K) + \sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell)\} + \mathcal{B}^T \Gamma \mathcal{B}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
\mathrm{II} &= \sum_{k=1}^{K} \pi_k \mathbb{E}[\langle \sum_{\ell=1}^{K} \pi_\ell(\beta_k - \beta_\ell), X_k \rangle^2] \\
&= \sum_{k=1}^{K} \pi_k \{(\beta_k - \beta_K) + \sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}^T \Gamma \{(\beta_k - \beta_K) + \sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}
\end{aligned}$$

Observe that the first term of I minus II is given by

$$\begin{aligned}
&- 2\{\sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell)\}^T \Gamma \{\sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}^T + \{\sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell)\}^T \Gamma \{\sum_{\ell=1}^{K-1} \pi_\ell^*(\beta_K - \beta_\ell)\}^T \\
&+ 2\{\sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}^T \Gamma \{\sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}^T - \{\sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}^T \Gamma \{\sum_{\ell=1}^{K-1} \pi_\ell(\beta_K - \beta_\ell)\}^T \\
&\quad = \{\sum_{\ell=1}^{K-1} (\pi_\ell^* - \pi_\ell)(\beta_K - \beta_\ell)\}^T \Gamma \{\sum_{\ell=1}^{K-1} (\pi_\ell^* - \pi_\ell)(\beta_K - \beta_\ell)\}.
\end{aligned}$$

The second term of I can be rewritten as

$$2\mathcal{B}^T \Gamma \sum_{\ell=1}^{K-1} (\pi_\ell^* - \pi_\ell)(\beta_K - \beta_\ell).$$

Thus, we have

$$\mathrm{I} - \mathrm{II} = \{\sum_{\ell=1}^{K-1}(\pi_\ell^* - \pi_\ell)(\beta_K - \beta_\ell) + \mathcal{B}\}^T\Gamma\{\sum_{\ell=1}^{K-1}(\pi_\ell^* - \pi_\ell)(\beta_K - \beta_\ell) + \mathcal{B}\} \geq 0.$$

2.8 Appendix: The EM Algorithm for Joint Mixture Regression

Denote the data matrix by $X = [x_1, \ldots, x_n]^T$, and let $\overline{X} = [\overline{x}_1, \ldots, \overline{x}_n]^T$ where $\overline{x}_i = [1, x_i^T]^T$, so that $\overline{X}$ is a $n \times (1+p)$ matrix. Also, let $\overline{\beta}_k = [\alpha_k, \beta_k^T]^T$. Once the M-step is done, the next E-step is given by

$$\tau_{ik}^{(+)} = \frac{\widehat{\pi}_k\varphi(y_i; \langle\widehat{\overline{\beta}}_k, \overline{x}_i\rangle, \widehat{\sigma}_k^2)\varphi(\overline{x}_i; \widehat{\mu}_k, \widehat{\Sigma}_k)}{\sum_{k=1}^K \widehat{\pi}_k\varphi(y_i; \langle\widehat{\overline{\beta}}_k, \overline{x}_i\rangle, \widehat{\sigma}_k^2)\varphi(\overline{x}_i; \widehat{\mu}_k, \widehat{\Sigma}_k)},$$

where the inner product is the usual inner product in $\mathbb{R}^{p+1}$ and the hat denotes the estimate obtained in the last M-step.

The M-step is obtained as follows. Define for $k = 1, \ldots, K$,

$$\begin{aligned}
\widehat{W}_k &= \mathrm{diag}\{\widehat{\tau}_{1k}, \ldots, \widehat{\tau}_{nk}\}, \quad \widetilde{X}_k = \widehat{W}_k^{1/2}X, \quad \widetilde{\overline{X}}_k = \widehat{W}_k^{1/2}\overline{X}, \\
\mathbb{1} &= [1, \ldots, 1]^T \in \mathbb{R}^n, \quad \widetilde{\mathbb{1}} = \widehat{W}_k^{1/2}\mathbb{1}, \\
\widetilde{y}_k &= \widehat{W}_k^{1/2}y, \quad y = [y_1, \ldots, y_n]^T \\
H(\widetilde{\mathbb{1}}) &= \widetilde{\mathbb{1}}(\widetilde{\mathbb{1}}^T\widetilde{\mathbb{1}})^{-1}\widetilde{\mathbb{1}}^T, \quad H(\widetilde{\overline{X}}_k) = \widetilde{\overline{X}}_k(\widetilde{\overline{X}_k}^T\widetilde{\overline{X}}_k)^{-1}\widetilde{\overline{X}_k}^T,
\end{aligned}$$

where $\widehat{\tau}_{ik}$ are the estimates obtained in the last E-step. For $k = 1, \ldots, K$, the new M-step is then given by

$$\begin{aligned}
&\mu_k^{(+)} = \{(\widetilde{\mathbb{1}}^T\widetilde{\mathbb{1}})^{-1}\widetilde{\mathbb{1}}^T\widetilde{X_k}\}^T, \quad \Sigma_k^{(+)} = (\widetilde{\mathbb{1}}^T\widetilde{\mathbb{1}})^{-1}\widetilde{X_k}^T\{I - H(\widetilde{\mathbb{1}})\}\widetilde{X_k}, \\
&\pi_k^{(+)} = \frac{1}{n}\sum_{i=1}^n \widehat{\tau}_{ik}, \quad \overline{\beta}_k^{(+)} = (\widetilde{\overline{X}_k}^T\widetilde{\overline{X}_k})^{-1}\widetilde{\overline{X}_k}^T\widetilde{y}_k, \quad \widehat{\sigma}_k^{2(+)} = (\widetilde{\mathbb{1}}^T\widetilde{\mathbb{1}})^{-1}\widetilde{y}_k^T\{I - H(\widetilde{\overline{X}_k})\}\widetilde{y}_k.
\end{aligned}$$

CHAPTER III

Regularized Smoothing in Functional Linear Models

3.1 Introduction

For the last two decades, interests have been rapidly growing in functional data analysis (FDA)—the analysis of data where the observations are modeled to be part of latent curves. Data of this sort are found in various fields, such as medical science, chemistry, linguistics, geosciences, finance, marketing, and others. For an excellent review, see Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006).

Since the name first appeared in Ramsay and Dalzell (1991), researchers have developed many FDA techniques based on traditional statistical procedures, such as principal component analysis (PCA) [Rice and Silverman (1991), Silverman (1996),James et al. (2000), Yao et al. (2005), Hall et al. (2006)], linear regression models [Cardot et al. (1999), Cardot et al. (2003), Cai and Hall (2006), Hall and Horowitz (2007), Crambes et al. (2009), Hall and Yang (2010), Yuan and Cai (2011)], the generalized linear models (GLM) [James (2002), James and Silverman (2005), Müller and Stadtmüller (2005), Dou et al. (2011)], nonparametric regression models [Ferraty and Vieu (2003, 2004), Ferraty et al. (2007), Hall et al. (2009)], and others.

Among these topics, functional linear regression has particularly drawn the attention of researchers during this decade, making significant progress. Generally, the approach to functional linear regression is divided into two ways. The first approach is based on functional PCA, where the regression is implemented on the basis of the principal component scores of the predictor covariance function [Cardot et al. (1999), Cai and Hall (2006), Hall and Horowitz (2007), Hall and Yang (2010)]. In this approach, predictors are projected onto the space spanned by a finite number of the leading eigenfunctions. Cai and Hall (2006) showed that this approach attains the optimal prediction convergence rate. Hall and Horowitz (2007) revealed that it also attains the optimal estimation convergence rate for the regression slope function in L^2-distance. To the best of our knowledge, this work has been so far the only prominent result focusing on the estimation error for the slope function. Despite these attractive, theoretical properties, however, there is no compelling reason why the leading eigenfunctions are the most relevant in terms of the regression structure; Hall and Yang (2010) provide the conditions under which the PCA approach in fact makes a perfect sense.

The second approach is based on the penalized least squares [Cardot et al. (2003), Crambes et al. (2009), Yuan and Cai (2010)]. This approach constrains the space where the regression slope function belongs by putting a certain penalty—often the smoothness penalty—resulting in a finite dimensional optimization problem. Crambes et al. (2009) used the smoothing-spline approach and demonstrated that the optimal prediction rate is achieved by this approach. We should mention that these two approaches—the FPCA approach and the regularization approach—require different kinds of assumptions and are not directly comparable. Yuan and Cai (2010) considered the problem in a more general framework using the reproducing kernel Hilbert space theory. They obtained the optimal convergence rate for the general norm that accounts for both estimation error and prediction error.

In this chapter, we also consider the regularization approach to the functional linear model. Our goal is two-fold. The first goal is to generalize their setups. Our setup yields much simpler derivation of the optimal convergence rate, so that one can see what is essentially behind the approach more clearly. The key idea behind it is to choose a penalty term so that an objective function has a unique solution. The second goal is to extend the practical aspects of the model by accommodating discrete observations and multiple predictors. The effects on the convergence rates are explored. We show that the optimal prediction convergence rates can be still attained but is dominated by the worst parameters characterizing the regression slope functions and the predictors.

The outline of this chapter is as follows. In Section 3.2, we set up the functional linear regression problem, introduce our approach to the problem, and present some properties regarding the estimator. Section 3.3 presents the asymptotic theory for the fully observed model. The optimal convergence prediction rate is given. Section 3.4 connects the fully observed model with the discretely observed model. Section 3.5 extends the single predictor model to the multiple predictor model. The chapter closes with some concluding remarks in Section 3.6.

3.2 Functional Linear Regression

3.2.1 Regularization Approach

Let $\{X(t), t \in [0,1]\}$ be a random function such that $\mathbb{E}X = 0$. Assume $\mathbb{E}\|X\|_{L^2}^2 < \infty$ where $\|f\|_{L^2}^2 := \int_0^1 f(t)^2 dt$, which implies that a sample path lies in $L^2[0,1]$ almost surely. Denote the covariance function of X by

$$R(s,t) = \mathbb{E}[X(s)X(t)], \qquad s,t \in [0,1].$$

Assume that $R(s,t)$ is continuous on $[0,1]^2$. Suppose that $(X_1, Y_1), \ldots, (X_n, Y_n)$ constitute a random sample of X and

$$Y_i = \langle f, X_i \rangle_{L^2} + \varepsilon_i, \qquad 1 \le i \le n, \tag{3.1}$$

where ε_i is random noise with mean 0 and variance σ^2, and f is an unknown regression slope function and the object of inference. For simplicity we remove the intercept, but the generalization is straightforward. We assume that f lies in the Sobolev-Hilbert space

$$W_m^2[0,1] := \{f : f, f', \ldots, f^{(m-1)} \text{ are absolutely continuous, } f^{(m)} \in L^2[0,1]\}.$$

There are several norms that can be assigned to $W_m^2[0,1]$. We assign the norm

$$\|f\|_{W_m^2}^2 = \|f\|_{L^2}^2 + \|f^{(m)}\|_{L^2}^2, \qquad f \in W_m^2[0,1]. \tag{3.2}$$

It is well known that $W_m^2[0,1]$ is a reproducing kernel Hilbert space [Wahba (1990)]. Denote the reproducing kernel associated with (3.2) by K.

Define

$$L_i f = \langle X_i, f\rangle_{L^2}, \qquad 1 \le i \le n,$$

and

$$Lf = (L_1 f, \ldots, L_n f)^T.$$

L_i is a random linear functional defined by X_i. It is bounded since, by using the Cauchy-Schwarz inequality,

$$|L_i f| \le \|X_i\|_{L^2} \|f\|_{L^2} \le \|X_i\|_{L^2} \|f\|_{W_m^2}.$$

Assign to $\mathbb{R}^n$ the norm

$$\|y\|_{\mathbb{R}^n}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2, \qquad y \in \mathbb{R}^n.$$

Then, L becomes a linear bounded operator from $W_m^2[0,1]$ to $\mathbb{R}^n$. From now on, $\mathbb{R}^n$ refers to the Hilbert space associated with the inner product induced by this norm. Denote the conjugate operator of L by L^*, i.e., L^* is a linear operator from $W_m^2[0,1]$ to $\mathbb{R}^n$ that satisfies the equality $\langle Lf, x\rangle_{\mathbb{R}^n} = \langle f, L^*x\rangle_{W_m^2}$ for all $f \in W_m^2[0,1]$ and $x \in \mathbb{R}^n$. Let

$$Y = (Y_1, \ldots, Y_n)^T \text{ and } \varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^T.$$

Then, (3.1) becomes

$$Y = Lf + \varepsilon.$$

Our goal is to estimate f.

For now, we assume that $X_1, \ldots, X_n$ are fully observed. In other words, L is fully observed. We estimate the regression slope function by the regularized least squares method. Define the objective function by

$$\mathcal{L}_\lambda(f) = \|Y - Lf\|_{\mathbb{R}^n}^2 + \lambda\|f\|_{W_m^2}^2, \qquad f \in W_m^2[0,1],$$

where $\lambda \in (0,\infty)$. We estimate f by minimizing $\mathcal{L}_\lambda(f)$.

Proposition III.1. *The unique minimizer of $\mathcal{L}_\lambda(f)$ is given by*

$$\widehat{f_\lambda} := (L^*L + \lambda I)^{-1}L^*Y, \tag{3.3}$$

where I is an identity operator.

Proof. For simplicity, let $G_\lambda = L^*L + \lambda I$. Express a candidate solution by $\widetilde{f} = \widehat{f}_\lambda + g$. Then,

$$\mathcal{L}_\lambda(\widetilde{f}) = \|Y\|^n_{\mathbb{R}^n} + \langle L\widetilde{f}, L\widetilde{f}\rangle_{\mathbb{R}^n} - 2\langle Y, L\widetilde{f}\rangle_{\mathbb{R}^n} + \lambda\|\widetilde{f}\|^2_{W^2_m}.$$

Since $L^*Y = G_\lambda \widehat{f}_\lambda$, we have

$$\langle Y, L\widetilde{f}\rangle_{\mathbb{R}^n} = \langle L^*Y, \widetilde{f}\rangle_{W^2_m} = \langle G_\lambda \widehat{f}_\lambda, \widetilde{f}\rangle_{W^2_m}.$$

Thus,

$$\begin{aligned}\mathcal{L}_\lambda(\widetilde{f}) &= \|Y\|^n_{\mathbb{R}^n} + \langle L^*L\widetilde{f}, \widetilde{f}\rangle_{\mathbb{R}^n} - 2\langle G_\lambda \widehat{f}_\lambda, \widetilde{f}\rangle_{W^2_m} + \lambda\|\widetilde{f}\|^2_{W^2_m} \\ &= \|Y\|^n_{\mathbb{R}^n} + \langle G_\lambda(\widetilde{f} - 2\widehat{f}_\lambda), \widetilde{f}\rangle_{W^2_m}.\end{aligned}$$

But,

$$\begin{aligned}\langle G_\lambda(\widetilde{f} - 2\widehat{f}_\lambda), \widetilde{f}\rangle_{W^2_m} &= \langle G_\lambda(\widetilde{f} - \widehat{f}_\lambda), \widetilde{f}\rangle_{W^2_m} - \langle G_\lambda \widehat{f}_\lambda, \widetilde{f}\rangle_{W^2_m} \\ &= \langle G_\lambda(\widetilde{f} - \widehat{f}_\lambda), \widetilde{f} - \widehat{f}_\lambda\rangle_{W^2_m} + \langle G_\lambda(\widetilde{f} - \widehat{f}_\lambda), \widehat{f}_\lambda\rangle_{W^2_m} - \langle G_\lambda \widehat{f}_\lambda, \widetilde{f}\rangle_{W^2_m} \\ &= \langle G_\lambda g, g\rangle_{W^2_m} - \langle G_\lambda \widehat{f}_\lambda, \widehat{f}_\lambda\rangle_{W^2_m},\end{aligned}$$

where the last equality follows because G_λ is self-adjoint. Since G_λ is positive definite, $\mathcal{L}_\lambda$ is minimized uniquely at $g = 0$. □

Throughout the chapter, we continue to use the notation $G_\lambda = L^*L + \lambda I$. Note that the existence of I, or $\|f\|^2_{L^2}$ in the penalty term, makes G_λ invertible. However, there are many other penalties that makes G_λ invertible. For example, Crambes et al. (2009) uses a penalty concerning the projection onto the pth order polynomial space.

By carefully following the derivation, however, it can be seen that the convergence rate does not depend much on a specific penalty.

We will derive the rate of convergence of $\widehat{f}_\lambda$ in terms of the prediction mean square error. Note that L in (3.3) is random. Also, $f \in W_m^2[0,1]$ while $X \in L^2[0,1]$. These facts make the derivation complicated. First, we compute the conditional squared bias and variance in estimation error.

Theorem III.2. *Let $\mathbb{E}_\varepsilon$ refer to the expectation with respect to ε. Then,*

$$\begin{aligned}\|L\mathbb{E}_\varepsilon(\widehat{f}_\lambda - f)\|^2_{\mathbb{R}^n} &\leq \lambda\|f\|^2_{W_m^2}\\ \mathbb{E}_\varepsilon\|L(\widehat{f}_\lambda - \mathbb{E}_\varepsilon\widehat{f}_\lambda)\|^2_{\mathbb{R}^n} &= \frac{\sigma^2}{n}\mathrm{tr}\{(G_\lambda^{-1}L^*L)^2\}.\end{aligned}$$

Proof. Let us first consider the bias. Note that

$$\mathbb{E}_\varepsilon\widehat{f}_\lambda = (L^*L + \lambda I)^{-1}L^*Lf$$

is the minimizer of

$$\mathcal{L}^*_\lambda(g) = \|Lf - Lg\|^2_{\mathbb{R}^n} + \lambda\|g\|^2_{W_m^2}.$$

Thus,

$$\|L\mathbb{E}_\varepsilon(\widehat{f}_\lambda - f)\|^2_{\mathbb{R}^n} + \lambda\|\mathbb{E}_\varepsilon\widehat{f}_\lambda\|^2_{W_m^2} = \mathcal{L}^*_\lambda(\mathbb{E}\widehat{f}_\lambda) \leq \mathcal{L}^*_\lambda(g) = \lambda\|f\|^2_{W_m^2}.$$

The non-negativity of $\|\mathbb{E}_\varepsilon\widehat{f}_\lambda\|^2_{W_m^2}$ leads to the conclusion.

For the variance, since $\widehat{f_\lambda} - \mathbb{E}_\varepsilon \widehat{f_\lambda} = G_\lambda^{-1} L^* \varepsilon$, we have

$$
\begin{aligned}
\mathbb{E}_\varepsilon \| L(\widehat{f_\lambda} - \mathbb{E}_\varepsilon \widehat{f_\lambda}) \|^2_{\mathbb{R}^n} &= \mathbb{E}_\varepsilon \langle L G_\lambda^{-1} L^* \varepsilon, L G_\lambda^{-1} L^* \varepsilon \rangle_{\mathbb{R}^n} \\
&= \frac{1}{n} \mathbb{E}_\varepsilon \mathrm{tr}\{\varepsilon^T L G_\lambda^{-1} L^* L G_\lambda^{-1} L^* \varepsilon\} \\
&= \frac{\sigma^2}{n} \mathrm{tr}\{L G_\lambda^{-1} L^* L G_\lambda^{-1} L^*\} \\
&= \frac{\sigma^2}{n} \mathrm{tr}\{(G_\lambda^{-1} L^* L)^2\}.
\end{aligned}
$$

□

We also have the following bound of $\widehat{f_\lambda}$ in $W_m^2[0,1]$.

Theorem III.3.

$$
\mathbb{E}_\varepsilon \| \widehat{f_\lambda} \|^2_{W_m^2} \leq 2 \| f \|^2_{W_m^2} + \frac{2\sigma^2}{n} \mathrm{tr}(G_\lambda^{-2} L^* L).
$$

Proof. Note

$$
\widehat{f_\lambda} = G_\lambda^{-1} L^* L f + G_\lambda^{-1} L^* \varepsilon,
$$

and so

$$
\| \widehat{f_\lambda} \|^2_{W_m^2} \leq 2 \| G_\lambda^{-1} L^* L f \|^2_{W_m^2} + 2 \| G_\lambda^{-1} L^* \varepsilon \|^2_{W_m^2}.
$$

For the first term,

$$
\| G_\lambda^{-1} L^* L f \|^2_{W_m^2} = \langle f, L^* L G_\lambda^{-2} L^* L f \rangle^2_{W_m^2} \leq \| f \|^2_{W_m^2},
$$

since $L^*LG_\lambda^{-2}L^*L \leq I$. For the second term,

$$\mathbb{E}_\varepsilon \|G_\lambda^{-1}L^*\varepsilon\|^2_{W^2_m} = \mathbb{E}_\varepsilon\langle \varepsilon, LG_\lambda^{-2}L^*\varepsilon\rangle_{\mathbb{R}^n} = \frac{\sigma^2}{n}\mathrm{tr}(G_\lambda^{-2}L^*L).$$

□

To evaluate the rate of the variance in Theorem III.2, it is crucial to explore the behavior of the eigenvalues of L^*L; if the eigenvalues of L^*L are $\rho_j, j \geq 1$, then

$$\mathrm{tr}\{(G_\lambda^{-1}L^*L)^2\} = \sum_{j=1}^{\infty}\left(\frac{\rho_j}{\rho_j+\lambda}\right)^2.$$

Recall that K is the reproducing kernel of $W^2_m[0,1]$. Define

$$R_n(s,t) = \frac{1}{n}\sum_{i=1}^n X_i(s)X_i(t),$$
$$Q_n(s,t) = \int_0^1 R_n(s,u)K(u,t)du.$$

Then, L^* and L^*L can be evaluated as follows.

Theorem III.4.

(a) Denote $K_t(\cdot) = K(t,\cdot)$. Then, for $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$,

$$L^*y = \frac{1}{n}\sum_{i=1}^n y_i \int_0^1 X_i(t)K_t(\cdot)dt.$$

(b) For $f \in W^2_m[0,1]$,

$$L^*Lf = \int_0^1 Q_n(s,\cdot)f(s)ds,$$

i.e., $L^*L = \mathfrak{K}\mathfrak{R}_n$ *where* $\mathfrak{K}$ *and* $\mathfrak{R}_n$ *are the integral operators with kernels* K *and* R_n*, respectively.*

Proof. For each $f \in W_m^2[0,1]$,

$$\begin{aligned}\langle L^*y, f\rangle_{W_m^2} = \langle y, Lf\rangle_{\mathbb{R}^n} &= \frac{1}{n}\sum_{i=1}^{n} y_i \int_0^1 X_i(t)f(t)dt \\ &= \frac{1}{n}\sum_{i=1}^{n} y_i \int_0^1 X_i(t)\langle K_t, f\rangle_{W_m^2} dt,\end{aligned}$$

where the last equality follows by the reproducing property of K. By interchanging the order of integration and inner product, we obtain

$$\langle L^*y, f\rangle_{W_m^2} = \left\langle \frac{1}{n}\sum_{i=1}^{n} y_i \int_0^1 X_i(t)K_t dt, f \right\rangle_{W_m^2},$$

from which (a) follows.

To prove (b), plugging $L_i f = \int_0^1 X_i(s)f(s)ds$ into y_i in (a), we obtain

$$\begin{aligned}L^*Lf &= \frac{1}{n}\sum_{i=1}^{n} \int_0^1 X_i(s)f(s)ds \int_0^1 X_i(t)K_t(\cdot)dt \\ &= \int_0^1 Q_n(s,\cdot)f(s)ds.\end{aligned}$$

□

3.2.2 The Eigen-System of $W_m^2[0,1]$

The following result is well known in the spline smoothing literature.

Proposition III.5. *There exists a complete orthonormal basis,* $\{\phi_j, j \geq 1\}$*, of*

$L^2[0,1]$ *such that*

$$\langle \phi_i^{(m)}, \phi_j^{(m)} \rangle_{L^2} = \rho_j \delta_{ij}, \quad i, j \geq 1,$$

where δ_{ij} *is a Dirac delta, i.e.,* $\delta_{ij} = 1$ *when* $i = j$ *and* $\delta_{ij} = 0$ *otherwise, and*

$$0 = \rho_1 = \cdots = \rho_m < \rho_{m+1} < \rho_{m+2} < \ldots,$$

and for some constants $C_1, C_2 \in (0, \infty)$,

$$C_1 j^{2m} \leq \rho_{j+m} \leq C_1 j^{2m}, \quad j \geq i.$$

For the proof, see Speckman (1985) and the references therein.

Define $\nu_j = 1 + \rho_j$. Then,

$$\langle \phi_i, \phi_j \rangle_{W_m^2} = \langle \phi_i, \phi_j \rangle_{L^2} + \langle \phi_i^{(m)}, \phi_j^{(m)} \rangle_{L^2} = \nu_j \delta_{ij}. \tag{3.4}$$

The following result concerning the eigen-system of $W_m^2[0,1]$ is straightforward.

Theorem III.6.

(a) The collection of $\{\phi_j / \sqrt{\nu_j}, j \geq 1\}$ *is a complete orthonormal basis of* $W_m^2[0,1]$. *Thus any* $f \in W_m^2[0,1]$ *can be written as*

$$f = \sum_{j=1}^{\infty} \nu_j^{-1/2} f_j \phi_j, \tag{3.5}$$

where $\sum_{j=1}^{\infty} f_j^2 < \infty$.

(b) For any $i, j \geq 1$, $\langle \mathfrak{K} \phi_i, \phi_j \rangle_{W_m^2} = \delta_{ij}$.

Proof. Since each function in $W_m^2[0,1]$ is a function in $L^2[0,1]$, we conclude that $\{\phi_j, j \geq 1\}$ is also a basis of $W_m^2[0,1]$. Then the conclusions of (a) follow easily from (3.4). To prove (b), by interchanging the order of integration and inner product,

$$\begin{aligned}
\langle \mathfrak{K}\phi_i, \phi_j \rangle_{W_m^2} &= \left\langle \int\limits_0^1 K(s,t)\phi_i(s)ds, \phi_j(t) \right\rangle_{W_m^2} \\
&= \int\limits_0^1 \langle K(s,t), \phi_j(t) \rangle_{W_m^2} \phi_i(s)ds \\
&= \int\limits_0^1 \phi_i(s)\phi_j(s)ds,
\end{aligned}$$

where the last equality follows from the reproducing property of K. □

3.3 Asymptotic Theory for Fully Observed Functional Data

In this section we assume that the functional data $X_i, 1 \leq i \leq n$, are fully observed, and use the results in Theorem III.2 to derive the prediction convergence rate procedure based on $\widehat{f}_\lambda$.

It is essential to understand the behavior of the eigenvalues of L^*L, or in other words the eigenvalues of $\mathfrak{K}\mathfrak{R}_n$ (Theorem III.4 (b)). It will be convenient to use expansions based on $\{\phi_j\}$. We can write

$$\mathfrak{R}_n = \sum_{j=1}^{\infty} \sum_{k=1}^{\infty} r_{jk} \phi_j \otimes_{L^2} \phi_k, \tag{3.6}$$

where

$$r_{jk} = \frac{1}{n} \sum_{i=1}^{n} \langle X_i, \phi_j \rangle_{L^2} \langle X_i, \phi_k \rangle_{L^2}.$$

Since $L^*L = \mathfrak{K}\mathfrak{R}_n$, representing $f \in W_m^2[0,1]$ as in (3.5),

$$L^*Lf = \sum_j \sum_\ell r_{j\ell} \nu_j^{-1/2} f_j \mathfrak{K}\phi_k. \tag{3.7}$$

The following result follows easily from (3.7) and Theorem III.6 (b).

Proposition III.7. *Let $f, g \in W_m^2[0,1]$ have the basis expansions*

$$f = \sum_{j=1}^\infty \nu_j^{-1/2} f_j \phi_j \text{ and } g = \sum_{k=1}^\infty \nu_j^{-1/2} g_k \phi_k.$$

Then

$$\langle L^*Lf, g\rangle_{W_m^2} = \sum_{j=1}^\infty \sum_{k=1}^\infty r_{jk} \nu_j^{-1/2} \nu_k^{-1/2} f_j g_k.$$

By Theorem III.6 (a), the mapping from $W_m^2[0,1]$ to ℓ^2 given by

$$f = \sum_{j=1}^\infty \nu_j^{-1/2} f_j \phi_j \mapsto (f_1, f_2, \dots) \tag{3.8}$$

is an isometric isomorphism. By the above proposition, the mapping in ℓ^2 that corresponds to L^*L is given by

$$\mathfrak{H} = \sum_{j=1}^\infty \sum_{k=1}^\infty r_{jk} \nu_j^{-1/2} \nu_k^{-1/2} e_j \otimes_{\ell^2} e_k,$$

where e_j is the element in ℓ^2 whose jth element is 1 and all the other elements are 0. Observe that we can write

$$\mathfrak{H} = \mathfrak{D}\mathfrak{S}\mathfrak{D} \tag{3.9}$$

where

$$\mathfrak{D} = \sum_j \nu_j^{-1/2} e_j \otimes_{\ell^2} e_j \quad \text{and} \quad \mathfrak{S} = \sum_j \sum_k r_{jk} e_j \otimes_{\ell^2} e_k.$$

Let $\rho_{n,1} \geq \rho_{n,2} \geq \dots$ be the eigenvalues of $\mathfrak{R}_n$ in $L^2[0,1]$ and so they are also the eigenvalues of $\mathfrak{S}$. Similarly, let $\rho_1 \geq \rho_2 \geq \dots$ be the eigenvalues of $\mathfrak{R}$ in $L^2[0,1]$, where $\mathfrak{R}$ is the theoretical covariance operator of X_i. For each k, let

$$t_{n,k} = \sum_{j>k} \rho_{n,j} \quad \text{and} \quad t_k = \sum_{j>k} \rho_j.$$

Lemma III.8. *For any sequence k_n,*

$$t_{n,k_n} = O_p(t_{k_n}).$$

Proof. $P_n^{(k)}$ be the projection on the span of the first k sample principal components and let $P^{(k)}$ be the projection on the span of the first k theoretical principal components. Thus,

$$t_{n,k} = \mathrm{tr}\{\mathfrak{R}_n(I - P_n^{(k)})\} \quad \text{and} \quad t_k = \mathrm{tr}\{\mathfrak{R}(I - P^{(k)})\}.$$

It follows that

$$\mathrm{tr}(\mathfrak{R}_n P_n^{(k)}) \geq \mathrm{tr}(\mathfrak{R}_n P^{(k)})$$

by the definition of $P_n^{(k)}$, and so

$$\mathrm{tr}\{\mathfrak{R}_n(I - P_n^{(k)})\} \leq \mathrm{tr}\{\mathfrak{R}_n(I - P^{(k)})\}.$$

By Chebyshev's inequality, using the fact $\mathbb{E}\mathrm{tr}\{\mathfrak{R}_n(I-P^{(k)})\} = \mathrm{tr}\{\mathfrak{R}(I-P^{(k)})\}$, we obtain for $K > 0$

$$\mathbb{P}(\mathrm{tr}\{\mathfrak{R}_n(I-P_n^{(k)})\} > K) \le \mathbb{P}(\mathrm{tr}\{\mathfrak{R}_n(I-P^{(k)})\} > K) \le \frac{\mathrm{tr}\{\mathfrak{R}(I-P^{(k)})\}}{K}.$$

□

Now we evaluate the trace of the variance in Theorem III.2.

Theorem III.9. *For any $j, k \ge 1$,*

$$\mathrm{tr}(G_\lambda^{-1}L^*L) \le k + j + \frac{t_{n,k}}{\lambda \nu_{j+1}}.$$

Proof. By the isometric isomorphism in (3.8),

$$\mathrm{tr}(G_\lambda^{-1}L^*L) = \mathrm{tr}((\mathfrak{H}+\lambda I)^{-1}\mathfrak{H}).$$

Let P be the projection of the span of the first k eigenvectors of $\mathfrak{S}$ and $P' = I - P$. Then

$$\mathrm{tr}\{(\mathfrak{H}+\lambda I)^{-1}\mathfrak{H}\} = \mathrm{tr}\{(\mathfrak{H}+\lambda I)^{-1}\mathfrak{D}\mathfrak{S}P\mathfrak{D}\} + \mathrm{tr}\{(\mathfrak{H}+\lambda I)^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D}\}.$$

Note that $\mathfrak{S}$ and P commute and so do $\mathfrak{S}$ and P'. Since $\mathfrak{H} \ge \mathfrak{D}\mathfrak{S}P\mathfrak{D}$ and $\mathfrak{H} \ge \mathfrak{D}\mathfrak{S}P'\mathfrak{D}$,

$$\begin{aligned}&\mathrm{tr}\{(\mathfrak{H}+\lambda I)^{-1}\mathfrak{H}\} \\ &\quad \le \mathrm{tr}\{(\mathfrak{D}\mathfrak{S}P\mathfrak{D}+\lambda I)^{-1}\mathfrak{D}\mathfrak{S}P\mathfrak{D}\} + \mathrm{tr}\{(\mathfrak{D}\mathfrak{S}P'\mathfrak{D}+\lambda I)^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D}\}.\end{aligned} \tag{3.10}$$

Observe that $(\mathfrak{D}\mathfrak{S}P\mathfrak{D} + \lambda I)^{-1}\mathfrak{D}\mathfrak{S}P\mathfrak{D}$ is self-adjoint and has norm bounded by 1. Thus,

$$\mathrm{tr}\{(\mathfrak{D}\mathfrak{S}P\mathfrak{D} + \lambda I)^{-1}\mathfrak{D}\mathfrak{S}P\mathfrak{D}\} \leq \dim(\mathrm{Im}(P)) = k. \tag{3.11}$$

Similarly, $(\mathfrak{D}\mathfrak{S}P'\mathfrak{D} + \lambda I)^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D}$ is self-adjoint, with norms bounded by 1, and

$$(\mathfrak{D}\mathfrak{S}P'\mathfrak{D} + \lambda I)^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D} \leq \lambda^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D}.$$

Thus,

$$\begin{aligned}\mathrm{tr}((\mathfrak{D}\mathfrak{S}P'\mathfrak{D} + \lambda I)^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D}) &= \sum_i \langle e_i, (\mathfrak{D}\mathfrak{S}P'\mathfrak{D} + \lambda I)^{-1}\mathfrak{D}\mathfrak{S}P'\mathfrak{D}e_i\rangle_{\ell^2} \\ &\leq j + \frac{1}{\lambda}\sum_{i>j}\langle e_i, \mathfrak{D}\mathfrak{S}P'\mathfrak{D}e_i\rangle_{\ell^2} \\ &\leq j + \frac{1}{\lambda\nu_{j+1}}\sum_{i>j}\langle e_i, \mathfrak{S}P'e_i\rangle_{\ell^2} \\ &\leq j + \frac{1}{\lambda\nu_{j+1}}\mathrm{tr}(\mathfrak{S}P'). \qquad (3.12)\end{aligned}$$

The result follows from (3.10)–(3.12). □

We now evaluate the estimation error.

Theorem III.10. *Assume that $t_k = O(k^{-\beta})$ for some $\beta > 0$. Then, setting $\lambda = n^{-\frac{2m+\beta+1}{2m+\beta+2}}$, we have*

$$\mathbb{E}_\varepsilon \|L(\widehat{f}_\lambda - f)\|^2_{\mathbb{R}^n} = O_p(n^{-\frac{2m+\beta+1}{2m+\beta+2}}).$$

Proof. By Lemma III.8 we conclude that $t_{n,k} = O_p(k^{-\beta})$. It follows from Theorem

III.9, with $k = j = [n^{\frac{1}{2m+\beta+2}}]$, that

$$\mathrm{tr}(G_\lambda^{-1}L^*L) = O_p(n^{\frac{1}{2m+\beta+2}}). \tag{3.13}$$

Note that $G_\lambda^{-1}L^*L$ is self-adjoint and bounded by I. Thus,

$$\mathrm{tr}((G_\lambda^{-1}L^*L)^2) \leq \mathrm{tr}(G_\lambda^{-1}L^*L),$$

and the result follows from Theorem III.2. □

Next, we evaluate the prediction error. Let X be a new independent observation. We consider

$$\mathbb{E}_{\varepsilon,X}\{L_X(\widehat{f}_\lambda - f)\}^2,$$

where $L_X f = \langle X, f\rangle_{L^2}$. Note that

$$\|Lf\|_{\mathbb{R}^n}^2 = \langle f, \mathfrak{R}_n f\rangle_{L^2} \quad \text{and} \quad \mathbb{E}_X(L_X f)^2 = \langle f, \mathfrak{R} f\rangle_{L^2}.$$

Thus,

$$\mathbb{E}_\varepsilon\|L(\widehat{f}_\lambda - f)\|_{\mathbb{R}^n}^2 = \mathbb{E}_\varepsilon\langle \widehat{f}_\lambda - f, \mathfrak{R}_n(\widehat{f}_\lambda - f)\rangle_{L^2},$$
$$\mathbb{E}_{\varepsilon,X}\{L_X(\widehat{f}_\lambda - f)\}^2 = \mathbb{E}_\varepsilon\langle \widehat{f}_\lambda - f, \mathfrak{R}(\widehat{f}_\lambda - f)\rangle_{L^2}.$$

The first expression was already considered by Theorem III.10. It therefore suffices to consider the difference of the two terms.

For any $f \in L^2[0,1]$,

$$\langle f, \mathfrak{R}_n f\rangle_{L^2} = \langle f, \mathfrak{R} f\rangle_{L^2} + \langle f, (\mathfrak{R}_n - \mathfrak{R}) f\rangle_{L^2}.$$

Below, we follow the argument in Crambes et al. (2009) to deal with the second term. Let

$$\mathfrak{R} = \sum_j \rho_j e_j \otimes e_j,$$

where the e_j are the principal components of $\mathfrak{R}$ (in $L^2[0,1]$). Define the scores by

$$\xi_{i,r} = \langle X_i, e_r\rangle,\ 1 \le i \le n, r \ge 1,$$

and let

$$\tau_{rs} = \frac{1}{\sqrt{n\rho_r\rho_s}} \sum_{i=1}^{n} (\xi_{i,r}\xi_{i,s} - \rho_r\delta_{rs}).$$

Lemma III.11. *Suppose that there exists a fixed $C < \infty$ such that*

$$\mathbb{V}\mathrm{ar}(\xi_{i,r}\xi_{i,s}) \le C\rho_r\rho_s \quad \textit{for all } r, s. \tag{3.14}$$

Then, for any $f \in L^2[0,1]$ and any $k \ge 1$,

$$\begin{aligned}&\langle f, \mathfrak{R} f\rangle_{L^2[0,1]} \\ &\qquad \le \langle f, \mathfrak{R}_n f\rangle_{L^2} + O_p(n^{-1})\|f\|_{L^2}^2 \sum_{r=1}^{k} t_{r-1} + O_p(n^{-1/2})\|f\|_{L^2}^2 t_k.\end{aligned}$$

Proof. Write

$$\langle f, (\mathfrak{R}_n - \mathfrak{R})f\rangle_{L^2} = \frac{1}{\sqrt{n}} \sum_{r=1}^{\infty} \sum_{s=1}^{\infty} f_r f_s \sqrt{\rho_r \rho_s} \tau_{rs},$$

where $f_r = \langle f, e_r\rangle$. Then

$$\begin{aligned}
&|\langle f, (\mathfrak{R}_n - \mathfrak{R})f\rangle_{L^2}| \\
&\leq \frac{2}{\sqrt{n}} \sum_{r=1}^{k} \sum_{s=r}^{\infty} |f_r f_s \sqrt{\rho_r \rho_s} \tau_{r,s}| + \frac{2}{\sqrt{n}} \sum_{r=k+1}^{\infty} \sum_{s=r}^{\infty} |f_r f_s \sqrt{\rho_r \rho_s} \tau_{r,s}| \\
&\leq \frac{2}{\sqrt{n}} \left(\sum_{r=1}^{k} \sum_{s=r}^{\infty} \rho_r f_r^2 f_s^2 \right)^{1/2} \left(\sum_{r=1}^{k} \sum_{s=r}^{\infty} \rho_s \tau_{r,s}^2 \right)^{1/2} \\
&\quad + \frac{2}{\sqrt{n}} \left(\sum_{r=k+1}^{\infty} \sum_{s=r}^{\infty} f_r^2 f_s^2 \right)^{1/2} \left(\sum_{r=k+1}^{\infty} \sum_{s=r}^{\infty} \rho_r \rho_s \tau_{r,s}^2 \right)^{1/2}.
\end{aligned}$$

Note that

$$\sum_{r=1}^{\infty} f_r^2 = \|f\|_{L^2}^2 \quad \text{and} \quad \sum_{r=1}^{\infty} \rho_r f_r^2 = \langle f, \mathfrak{R} f\rangle_{L^2[0,1]}.$$

Thus,

$$\begin{aligned}
\sum_{r=1}^{k} \sum_{s=r}^{\infty} \rho_r f_r^2 f_s^2 &\leq \langle f, \mathfrak{R} f\rangle_{L^2} \|f\|_{L^2}^2, \\
\sum_{r=k+1}^{\infty} \sum_{s=r}^{\infty} f_r^2 f_s^2 &\leq \|f\|_{L^2}^4.
\end{aligned}$$

Using the assumption that $\mathbb{E}\tau_{r,s}^2 \leq C$ for all r, s,

$$\begin{aligned}
\sum_{r=1}^{k} \sum_{s=r}^{\infty} \rho_s \tau_{r,s}^2 &= O_p(1) \sum_{r=1}^{k} t_{r-1} \\
\sum_{r=k+1}^{\infty} \sum_{s=r}^{\infty} \rho_r \rho_s \tau_{r,s}^2 &= O_p(1) \cdot t_k^2.
\end{aligned}$$

Combining these derivations,

$$\begin{aligned}&|\langle f, (\mathfrak{R}_n - \mathfrak{R})f\rangle_{L^2}| \\ &\qquad = O_p(n^{-1/2})\|f\|_{L^2} \left\{ \langle f, \mathfrak{R}f\rangle_{L^2}^{1/2} \left(\sum_{r=1}^{k} t_{r-1} \right)^{1/2} + t_k \|f\|_{L^2} \right\}.\end{aligned}$$

Now,

$$\begin{aligned}&\langle f, \mathfrak{R}f\rangle_{L^2} \\ &\qquad = \langle f, \mathfrak{R}_n f\rangle_{L^2} + O_p(n^{-1/2})\|f\|_{L^2} \left\{ \langle f, \mathfrak{R}f\rangle_{L^2}^{1/2} \left(\sum_{r=1}^{k} t_{r-1} \right)^{1/2} + t_k \|f\|_{L^2} \right\}.\end{aligned}$$

Note that if

$$\langle f, \mathfrak{R}f\rangle_{L^2} = O_p(n^{-1/2})\|f\|_{L^2} \langle f, \mathfrak{R}f\rangle_{L^2}^{1/2} \left(\sum_{r=1}^{k} t_{r-1} \right)^{1/2},$$

then

$$\langle f, \mathfrak{R}f\rangle_{L^2} = O_p(n^{-1})\|f\|_{L^2}^2 \sum_{r=1}^{k} t_{r-1}.$$

Thus,

$$\begin{aligned}&\langle f, \mathfrak{R}f\rangle_{L^2} \\ &\qquad \leq \langle f, \mathfrak{R}_n f\rangle_{L^2[0,1]} + O_p(n^{-1})\|f\|_{L^2}^2 \sum_{r=1}^{k} t_{r-1} + O_p(n^{-1/2})\|f\|_{L^2}^2 t_k.\end{aligned}$$

□

We can now compute the prediction convergence rate.

Theorem III.12. *Assume that (3.14) holds and $t_k = O(k^{-\beta})$ for some $\beta > 0$. Then, taking $\lambda = n^{-\frac{2m+\beta+1}{2m+\beta+2}}$ yields*

$$\mathbb{E}_{\varepsilon,X}\{L_X(\widehat{f}_\lambda - f)\}^2 = O_p(n^{-\frac{2m+\beta+1}{2m+\beta+2}} + n^{-\frac{\beta+1}{2}}).$$

Proof. It follows from the last lemma that

$$\begin{aligned}&\langle \widehat{f}_\lambda - f, \mathfrak{R}(\widehat{f}_\lambda - f)\rangle_{L^2} \\ &\quad \leq \langle \widehat{f}_\lambda - f, \mathfrak{R}_n(\widehat{f}_\lambda - f)\rangle_{L^2} + O_p(n^{-1})\|\widehat{f}_\lambda - f\|_{L^2}^2 \sum_{r=1}^{k} t_{r-1} + O_p(n^{-1/2})\|\widehat{f}_\lambda - f\|_{L^2}^2 t_k.\end{aligned} \tag{3.15}$$

By Theorem III.3,

$$\mathbb{E}_\varepsilon \|\hat{f}_\lambda\|_{W_m^2}^2 \leq 2\|f\|_{W_m^2}^2 + \frac{2\sigma^2}{n}\mathrm{tr}(G_\lambda^{-2}L^*L).$$

In view of the fact that $\mathrm{tr}(G_\lambda^{-2}L^*L) \leq \lambda^{-1}\mathrm{tr}(G_\lambda^{-1}L^*L)$ and (3.13),

$$\frac{1}{n}\mathrm{tr}(G_\lambda^{-2}L^*L) = \frac{1}{n\lambda}O_p(n^{\frac{1}{2m+\beta+2}}) = O_p(1).$$

Thus, $\|\widehat{f}_\lambda\|_{L^2} \leq \|\widehat{f}_\lambda\|_{W_m^2} = O_p(1)$. If we let $k = [n^{1/2}]$, then the second and third terms of the right-hand side of (3.15) are $O_p(n^{-(\beta+1)/2})$. The result follows from inserting the rate found in Theorem III.10 in (3.15). □

Finally, we finish this section by proving that $n^{-\frac{2m+\beta+1}{2m+\beta+2}}$ is the minimax rate.

Theorem III.13. *There exists a constant d such that*

$$\liminf_{n\to\infty} \inf_{\widetilde{f}\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}(\beta), f\in W_m^2} \mathbb{P}(\mathbb{E}_X\{L_X(\widetilde{f}-f)\}^2 > dn^{-\frac{2m+\beta+1}{2m+\beta+2}}) > 0,$$

where $\mathcal{F}$ is a collection of measurable functions of $(X_1, Y_1), \dots, (X_n, Y_n)$, and $\mathcal{P}$ is a collection of probability measures such that X is a centered L^2-random function satisfying $t_k = O(k^{-\beta})$ and ε is a random noise with a mean 0 and a variance σ^2.

Proof. The proof uses the argument based on the Neyman-Pearson lemma, as seen in the papers such as Hall and Horowitz (2007) and Yuan and Cai (2010). Define

$$\mathcal{W}_n := \Big\{ f : f(t) = \sum_{k=L_n+1}^{2L_n} L_n^{-1/2} k^{-m} \theta_k \phi_k(t), \theta_k \in \{0,1\}, L_n = \lfloor n^{\frac{1}{2m+\beta+2}} \rfloor \Big\}.$$

For a sufficiently large $C > 0$, $\|f\|_{\mathcal{W}_n}^2 < C$ for all $f \in \mathcal{W}_n$ and all n. Thus, $\mathcal{W}_n \subset W_m^2$. Note that there are 2^{s_n} elements in $\mathcal{W}_n$. Next, define a collection of probability measures $\mathcal{P}^*$ such that ϵ follows $N(0, \sigma^2)$ and X is defined by the expansion $X = \sum_{k=1}^{\infty} \xi_k k^{-\frac{\beta+1}{2}} \phi_k$ where ξ_k's are independent uniform random variables on $[-\sqrt{3}, \sqrt{3}]$. $\mathcal{P}^*$ actually consists of only one element and obviously $\mathcal{P}^* \subset \mathcal{P}$. For notational simplicity, denote $\mathcal{G} := (\mathcal{P}^*, \mathcal{W}_n)$. For a given $G \in \mathcal{G}$, define G_{k0} such that it is exactly same as G except that θ_k in G is 0, and similarly define G_{k1}.

Define an arbitrary estimator by $\widetilde{f} = \sum_{k=s_n+1}^{2s_n} s_n^{-1/2} k^{-m} \widetilde{\theta}_k \phi_k$. Then,

$$
\begin{aligned}
\sup_{G\in\mathcal{G}} \mathbb{E}_G\{L_X(\widetilde{f}-f)\}^2 &= \sup_{G\in\mathcal{G}} \sum_{k=s_n+1}^{2s_n} s_n^{-1} k^{-(2m+\beta+1)} \mathbb{E}_G(\widetilde{\theta}_k - \theta_k)^2 \\
&\geq \frac{1}{2^{s_n}} \sum_{G\in\mathcal{G}} \sum_{k=s_n+1}^{2s_n} s_n^{-1} k^{-(2m+\beta+1)} \mathbb{E}_G(\widetilde{\theta}_k - \theta_k)^2 \\
&= s_n^{-1} \sum_{k=s_n+1}^{2s_n} k^{-(2m+\beta+1)} \frac{1}{2^{s_n}} \sum_{G\in\mathcal{G}} \mathbb{E}_G(\widetilde{\theta}_k - \theta_k)^2 \\
&= s_n^{-1} \sum_{k=s_n+1}^{2s_n} k^{-(2m+\beta+1)} \frac{1}{2^{s_n}} \sum_{G\in\mathcal{G}} \frac{1}{2}\Big\{\mathbb{E}_{G_{k0}}(\widetilde{\theta}_k - \theta_k)^2 + \mathbb{E}_{G_{k1}}(\widetilde{\theta}_k - \theta_k)^2\Big\} \\
&= s_n^{-1} \sum_{k=s_n+1}^{2s_n} k^{-(2m+\beta+1)} \frac{1}{2^{s_n}} \sum_{G\in\mathcal{G}} \mathbb{E}_{G_k}(\widetilde{\theta}_k - \theta_k)^2,
\end{aligned}
$$

where $P_{G_k} := \frac{1}{2}\{P_{G_{k0}} + P_{G_{k1}}\}$. Now, define k-th loglikelihood by

$$
R_{G_k} = \frac{L_{G_{k1}}}{L_{G_{k0}}}.
$$

Then,

$$
\begin{aligned}
\mathbb{E}_{G_k}(\widetilde{\theta}_k - \theta_k)^2 &= \mathbb{E}_{G_k}[P_{G_k}(\theta_k = 0|Y,X)1\{\widetilde{\theta}_k = 1\} + P_{G_k}(\theta_k = 1|Y,X)1\{\widetilde{\theta}_k = 0\}] \\
&= \mathbb{E}_{G_k}[(R_{G_k}+1)^{-1}1\{\widetilde{\theta}_k = 1\} + R_{G_k}(R_{G_k}+1)^{-1}1\{\widetilde{\theta}_k = 0\}] \\
&\geq \mathbb{E}_{G_k}[(R_{G_k}+1)^{-1}1\{R_{G_k} \geq 1\} + R_{G_k}(R_{G_k}+1)^{-1}1\{R_{G_k} < 1\}] \\
&= \mathbb{E}_{G_k}\Big[\min\Big\{\frac{R_{G_k}}{1+R_{G_k}}, \frac{1}{1+R_{G_k}}\Big\}\Big] \\
&\geq \frac{1}{2}\mathbb{E}_{G_{k0}}\Big[\min\Big\{\frac{R_{G_k}}{1+R_{G_k}}, \frac{1}{1+R_{G_k}}\Big\}\Big] \\
&\geq \frac{1}{64}\Big(\mathbb{E}_{G_{k0}} R_{G_k}^2\Big)^{-2},
\end{aligned}
$$

where the last inequality comes from Hall (1989). Now, again following Hall's argument on the evaluation of a normal likelihood ratio, we conclude

$$\mathbb{E}_{G_{k0}} R_{G_k}^2 = \{1 + O(s_n^{-1} k^{-(2m+\beta+1)})\}^n = \{1 + O(s_n^{-1} s_n^{-(2m+\beta+1)})\}^n = \{1 + O(n^{-1})\}^n,$$

because $s_n + 1 \le k \le 2s_n$. Thus, there exists a constant $C_1 > 0$, which is independent of G and k, such that for all sufficiently large n

$$\mathbb{E}_{G_k}(\widetilde{\theta}_k - \theta_k)^2 > C_1.$$

Now, we conclude that there exist another constant $C_2 > 0$ such that

$$\sup_{G \in \mathcal{G}} \mathbb{E}_G \{L_X(\widetilde{f} - f)\}^2 \ge C_2 n^{-\frac{2m+\beta+1}{2m+\beta+2}}.$$

To consider the lower bound, we only have to consider estimators such that $0 \le \widetilde{\theta}_k \le 1$. Thus,

$$\begin{aligned}\sum_{k=s_n+1}^{2s_n} s_n^{-1} k^{-(2m+\beta+1)} (\widetilde{\theta}_k - \theta_k)^2 &\le \sum_{k=s_n+1}^{2s_n} s_n^{-1} k^{-(2m+\beta+1)} \\ &\le n^{-\frac{2m+\beta+1}{2m+\beta+2}}.\end{aligned}$$

Now, we conclude that

$$\begin{aligned}&\liminf_{n\to\infty} \inf_{\widetilde{f}\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}(\beta), f_o\in W_m^2} \mathbb{P}(\mathbb{E}_X\{L_X(\widetilde{f} - f_o)\}^2 > dn^{-\frac{2m+\beta+1}{2m+\beta+2}}) \\ &\ge \liminf_{n\to\infty} \inf_{\widetilde{f}\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}^*, f_o\in\mathcal{W}_n} \mathbb{P}(\mathbb{E}_X\{L_X(\widetilde{f} - f_o)\}^2 > dn^{-\frac{2m+\beta+1}{2m+\beta+2}}) \\ &> 0,\end{aligned}$$

where $d > 0$ is some small constant.

□

3.4 Asymptotic Theory for Discretely Observed Functional Data

In this section, we consider the case where each X_i is observed at $t_{i,1}, \dots, t_{i,p_i}$. Let $I_{i,j}$ be an interval containing $t_{i,j}$ and $I_{i,1}, \dots, I_{i,p_i}$ form a partition of $[0,1]$. Define

$$\widetilde{X}_i(t) = \sum_{j=1}^{p_i} X_i(t_{i,j}) I(t \in I_{i,j}).$$

Accordingly, let

$$\widetilde{L}_i f = \langle \widetilde{X}_i, f \rangle_{L^2} \quad \text{and} \quad \widetilde{L} f = (\widetilde{L}_1 f, \dots, \widetilde{L}_n f)^T.$$

In the asymptotic theory below, assume that $p_i = p_{n,i}$. Denote by $\widetilde{f}_\lambda$ the penalized least squares solution:

$$\widetilde{f}_\lambda = \underset{f \in W_m^2}{\operatorname{argmin}} \left\{ \|Y - \widetilde{L} f\|_{\mathbb{R}^n}^2 + \lambda \|f\|_{W_m^2}^2 \right\}, \tag{3.16}$$

Thus

$$\widetilde{f}_\lambda := \widetilde{G}_\lambda^{-1} \widetilde{L}^* Y,$$

where $\widetilde{G}_\lambda = \widetilde{L}^* \widetilde{L} + \lambda I$.

Let $\widetilde{\mathfrak{R}}_n$ be the sample covariance operator of $\widetilde{X}_1, \dots, \widetilde{X}_n$ in $L^2[0,1]$, and $\widetilde{\mathfrak{R}} = \mathbb{E}(\widetilde{\mathfrak{R}}_n)$. Note that $\widetilde{\mathfrak{R}}$ also depends on n. Also, let $\widetilde{\rho}_1 \geq \widetilde{\rho}_2 \geq \dots$ be the eigenvalues of $\widetilde{\mathfrak{R}}$ and $\widetilde{t}_k = \sum_{j>k} \widetilde{\rho}_j$. Similarly, define

$$\widetilde{\xi}_{i,s} = \langle \widetilde{X}_i, \widetilde{e}_s \rangle_{L^2},$$

where $\widetilde{e}_s$ is the eigenfunction of $\widetilde{\mathfrak{R}}$ that corresponds to the eigenvalue $\widetilde{\rho}_s$. Let

$$L_X f = \langle X, f\rangle_{L^2} \quad \text{and} \quad L_{\widetilde{X}} f = \langle \widetilde{X}, f\rangle_{L^2}.$$

Theorem III.14. *Assume that $\widetilde{t}_k = O(k^{-\beta})$ for some $\beta > 0$ and there exists a fixed $C < \infty$ such that*

$$\mathbb{V}\mathrm{ar}(\widetilde{\xi}_{i,r}\widetilde{\xi}_{i,s}) \leq C\widetilde{\rho}_r\widetilde{\rho}_s \quad \textit{for all } r, s. \tag{3.17}$$

Then, taking $\lambda = n^{-\frac{2m+\beta+1}{2m+\beta+2}}$ yields

$$\mathbb{E}_{\varepsilon,X}\{L_X(\widetilde{f}_\lambda - f)\}^2 = O_p\left(n^{-\frac{2m+\beta+1}{2m+\beta+2}} + n^{-\frac{\beta+1}{2}} + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\|X_i - \widetilde{X}_i\|_{L^2}^2\right).$$

Proof. First, we show that

$$\mathbb{E}_\varepsilon\|\widetilde{L}(\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2 = O_p\left(n^{-\frac{2m+\beta+1}{2m+\beta+2}} + \frac{1}{n}\sum_{i=1}^n \mathbb{E}\|X_i - \widetilde{X}_i\|_{L^2}^2\right). \tag{3.18}$$

Write

$$\begin{aligned}\mathbb{E}_\varepsilon\|\widetilde{L}(\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2 &= \mathbb{E}_\varepsilon\|\widetilde{L}(\widetilde{f}_\lambda - \mathbb{E}_\varepsilon\widetilde{f}_\lambda + \mathbb{E}_\varepsilon\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2 \\ &\leq 2\mathbb{E}_\varepsilon\|\widetilde{L}(\widetilde{f}_\lambda - \mathbb{E}_\varepsilon\widetilde{f}_\lambda)\|^2 + 2\|\widetilde{L}\mathbb{E}_\varepsilon(\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2. \end{aligned}\tag{3.19}$$

The first term on the right of (3.19) is a variance computation, and derivations similar to the continuous case show that

$$\mathbb{E}_\varepsilon\|\widetilde{L}(\widetilde{f}_\lambda - \mathbb{E}_\varepsilon\widetilde{f}_\lambda)\|^2 = \frac{\sigma^2}{n}\mathrm{tr}\{(\widetilde{G}_\lambda\widetilde{L}^*\widetilde{L})^2\} = O_p\left(n^{-\frac{2m+\beta+1}{2m+\beta+2}}\right). \tag{3.20}$$

The second term on the right of (3.19) is a square bias computation and can be handled as follows. First observe that $\mathbb{E}_\varepsilon \widetilde{f}_\lambda = \widetilde{G}_\lambda^{-1} \widetilde{L}^* L f$ is the minimizer of

$$\|Lf - \widetilde{L}f\|_{\mathbb{R}^n}^2 + \lambda \|f\|_{W_m^2}^2.$$

Thus,

$$\|Lf - \widetilde{L}\mathbb{E}_\varepsilon \widetilde{f}_\lambda\|_{\mathbb{R}^n}^2 + \lambda \|\mathbb{E}_\varepsilon \widetilde{f}_\lambda\|_{W_m^2}^2 \leq \|Lf - \widetilde{L}f\|_{\mathbb{R}^n}^2 + \lambda \|f\|_{W_m^2}^2. \tag{3.21}$$

Since

$$\|\widetilde{L}\mathbb{E}_\varepsilon(\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2 \leq 2\|\widetilde{L}f - Lf\|_{\mathbb{R}^n}^2 + 2\|Lf - \widetilde{L}\mathbb{E}_\varepsilon \widetilde{f}_\lambda\|_{\mathbb{R}^n}^2,$$

using (3.21) we arrive at the inequality

$$\|\widetilde{L}\mathbb{E}_\varepsilon(\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2 \leq 3\|Lf - \widetilde{L}f\|_{\mathbb{R}^n}^2 + 2\lambda \|f\|_{W_m^2}^2. \tag{3.22}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} \|Lf - \widetilde{L}f\|_{\mathbb{R}^n}^2 &\leq \|f\|_{L^2}^2 \frac{1}{n} \sum_{i=1}^n \|X_i - \widetilde{X}_i\|_{L^2}^2 \\ &= O_p\left(\frac{1}{n} \sum_{i=1}^n \mathbb{E}\|X_i - \widetilde{X}_i\|_{L^2}^2 \right), \end{aligned}$$

and, by (3.22), we obtain

$$\|\widetilde{L}\mathbb{E}_\varepsilon(\widetilde{f}_\lambda - f)\|_{\mathbb{R}^n}^2 = O_p\left(\lambda + \frac{1}{n} \sum_{i=1}^n \mathbb{E}\|X_i - \widetilde{X}_i\|_{L^2}^2 \right). \tag{3.23}$$

Then, (3.18) follows from (3.19), (3.20) and (3.23).

Next let $X_1', \ldots, X_n'$ be a random sample from the same distribution as X_1, and let Z be a discrete uniform random variable with possible values $1, \ldots, n$. Assume that $X_1, \ldots, X_n$, $X_1', \ldots, X_n', Z$ are all independent. Define

$$\widetilde{X}_i'(t) = \sum_{j=1}^{p_i} X_i'(t_{i,j}) I(t \in I_{i,j}),$$

and

$$\widetilde{X} = \sum_{i=1}^{n} I(Z = i) \widetilde{X}_i'.$$

By (3.18) and following the lines of proof in Lemma III.11 and Theorem III.12,

$$\begin{aligned}
&\mathbb{E}_{\varepsilon,\widetilde{X}}\{L_{\widetilde{X}}(\widetilde{f}_\lambda - f)\}^2 \\
&\quad = O_p\left(n^{-\frac{2m+\beta+1}{2m+\beta+2}} + n^{-\frac{\beta+1}{2}} + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\|X_i - \widetilde{X}_i\|_{L^2}^2\right). \qquad (3.24)
\end{aligned}$$

To relate it to $\mathbb{E}_{\varepsilon,X}\{L_X(\widetilde{f}_\lambda - f)\}^2$, consider the X defined by

$$X = \sum_{i=1}^{n} I(Z = i) X_i',$$

which clearly has the same distribution as X_1. Since

$$L_X(\widetilde{f}_\lambda - f) = L_{\widetilde{X}}(\widetilde{f}_\lambda - f) + \langle \widetilde{f}_\lambda - f, X - \widetilde{X} \rangle_{L^2},$$

we have

$$\{L_X(\widetilde{f}_\lambda - f)\}^2 \leq 2\{L_{\widetilde{X}}(\widetilde{f}_\lambda - f)\}^2 + 2\langle \widetilde{f}_\lambda - f, X - \widetilde{X} \rangle_{L^2}^2.$$

Thus,

$$
\begin{aligned}
&\mathbb{E}_{\varepsilon,X}\{L_X(\widetilde{f}_\lambda - f)\}^2 \\
&\quad = \mathbb{E}_{\varepsilon,X'_1,\ldots,X'_n,Z}\{L_X(\widetilde{f}_\lambda - f)\}^2 \\
&\quad \le 2\mathbb{E}_{\varepsilon,X'_1,\ldots,X'_n,Z}\{L_{\widetilde{X}}(\widetilde{f}_\lambda - f)\}^2 + 2\mathbb{E}_{\varepsilon,X'_1,\ldots,X'_n,Z}\langle \widetilde{f}_\lambda - f, X - \widetilde{X}\rangle_{L^2}^2 \\
&\quad = 2\mathbb{E}_{\varepsilon,\widetilde{X}}\{L_{\widetilde{X}}(\widetilde{f}_\lambda - f)\}^2 + 2\mathbb{E}_{\varepsilon,X,\widetilde{X}}\langle \widetilde{f}_\lambda - f, X - \widetilde{X}\rangle_{L^2}^2.
\end{aligned}
$$

The first term on the right has been handled by (3.24). By the Cauchy-Schwarz inequality,

$$\mathbb{E}_{\varepsilon,X,\widetilde{X}}\langle \widetilde{f}_\lambda - f, X - \widetilde{X}\rangle_{L^2}^2 \le \mathbb{E}_\varepsilon\|\widetilde{f}_\lambda - f\|_{L^2}^2 \mathbb{E}\|X - \widetilde{X}\|_{L^2}^2.$$

As in the proof of Theorem III.12 the first term on the right-hand side is $O_p(1)$. The second term is equal to $\frac{1}{n}\sum_{i=1}^n \mathbb{E}\|X_i - \widetilde{X}_i\|_{L^2}^2$, and this completes the proof. □

Note that f_λ and $\widetilde{f}_\lambda$ are not natural splines. To compute these estimators one can use the representer theorem for the reproducing kernel Hilbert space $W_m^2[0,1]$. We focus on $\widetilde{f}_\lambda$. Let ξ_i be the representer of the functional $f \mapsto \langle \widetilde{X}_i, f\rangle_{L^2}, f \in W_m^2[0,1]$, namely, $\langle \xi_i, f\rangle_{W_m^2} = \langle \widetilde{X}_i, f\rangle_{L^2}$. Let K be a reproducing kernel of W_m^2. By the reproducing property,

$$\xi_i(t) = \langle \xi_i, K_t\rangle_{W_m^2} = \langle \widetilde{X}_i, K_t\rangle_{L^2}.$$

Let U be the matrix contains the elements

$$\widetilde{L}_i\xi_j = \langle \widetilde{X}_i, \xi_j\rangle_{L^2} = \int_0^1\int_0^1 K(s,t)\widetilde{X}_i(s)\widetilde{X}_j(t)dsdt.$$

By the representer theorem, the solution can be written as

$$\widetilde{f}_\lambda = \sum_{i=1}^{n} c_i \xi_i$$

where

$$(c_1, \ldots, c_n) = (U^T U + \lambda U)^{-1} U^T y = (U + \lambda I)^{-1} y.$$

3.5 Functional Linear Regression with Multiple Predictors

In this section, we consider multiple functional predictors. For simplicity, we consider the two-predictor case. Extension to more than two predictors is straightforward. Define a product space $W[0,1]^2 = W^2_{m_1}[0,1] \times W^2_{m_2}[0,1]$ and the linear operator $L : W[0,1]^2 \to \mathbb{R}^n$ by $Lf = L_1 f_1 + L_2 f_2$ where $f = (f_1, f_2) \in W[0,1]^2$ and $L_j f_j = (\langle X_{j,1}, f_j \rangle, \ldots, \langle X_{j,n}, f_j \rangle)$ for $j = 1, 2$. Suppose Y is generated by

$$Y = Lf + \varepsilon = L_1 f_1 + L_2 f_2 + \varepsilon.$$

We use a natural inner product in $W[0,1]^2$ such that for $f_1, g_1 \in W^2_{m_1}[0,1]$ and $f_2, g_2 \in W^2_{m_2}[0,1]$,

$$\langle (f_1, f_2), (g_1, g_2) \rangle = \langle f_1, g_1 \rangle + \langle f_2, g_2 \rangle.$$

It is not difficult to see that $L^* x = (L_1^* x, L_2^* x)$. Define a linear operator $\Lambda : W[0,1]^2 \to W[0,1]^2$ by $\Lambda f := (\lambda_1 f_1, \lambda_2 f_2)$. We use a loss function

$$\begin{aligned} \mathcal{L}_\lambda(f) &:= \|Y - Lf\|^2 + \langle f, \Lambda f \rangle \\ &= \|Y - L_1 f_1 - L_2 f_2\|^2 + \lambda_1 \|f_1\|_{W^2_{m_1}} + \lambda_2 \|f_2\|_{W^2_{m_2}}, \end{aligned}$$

whose minimizer is, according to Proposition III.1, given by

$$\widehat{f_\lambda} = (\widehat{f_1}, \widehat{f_2}) = (L^*L + \Lambda)^{-1}L^*y.$$

As before, define $G_\lambda = L^*L + \Lambda$. Theorems III.2 and III.3 hold without any change, i.e.,

$$\begin{aligned}
\mathbb{E}_\varepsilon \|L(\widehat{f_\lambda} - f)\|^2 &\le \langle f, \Lambda f\rangle, \\
\mathbb{E}_\varepsilon \|L(\widehat{f_\lambda} - \mathbb{E}_\varepsilon \widehat{f_\lambda})\|^2 &= \frac{\sigma^2}{n}\mathrm{tr}((G_\lambda^{-1}L^*L)^2),
\end{aligned}$$

and

$$\mathbb{E}_\varepsilon \|\widehat{f_\lambda}\|^2 \le 2\|f\| + \frac{2\sigma^2}{n}\mathrm{tr}(G_\lambda^{-2}L^*L).$$

Let K_1 and K_2 be reproducing kernels in $W^2_{m_1}[0,1]$ and $W^2_{m_2}[0,1]$, respectively. The result corresponding to Theorem III.4 becomes

$$\begin{aligned}
L^*y &= (L_1^*y, L_2^*y) \\
&= \Big(\frac{1}{n}\sum_{i=1}^n y_i\langle X_{1i}(t), K_1(t,\cdot)\rangle, \frac{1}{n}\sum_{i=1}^n y_i\langle X_{2i}(t), K_2(t,\cdot)\rangle\Big), \\
L^*Lf &= \Big(L_1^*(L_1f_1 + L_2f_2), L_2^*(L_1f_1 + L_2f_2)\Big) \\
&= \Big(\frac{1}{n}\sum_{i=1}^n \{\langle X_{1i}, f_1\rangle + \langle X_{2i}, f_2\rangle\}\langle X_{1i}(t), K_1(t,\cdot)\rangle, \\
&\qquad \frac{1}{n}\sum_{i=1}^n \{\langle X_{1i}, f_1\rangle + \langle X_{2i}, f_2\rangle\}\langle X_{2i}(t), K_2(t,\cdot)\rangle\Big) \\
&= \Big(\mathcal{K}_1\mathcal{R}_{n,(1,1)}f_1 + \mathcal{K}_1\mathcal{R}_{n,(1,2)}f_2, \mathcal{K}_2\mathcal{R}_{n,(2,1)}f_1 + \mathcal{K}_2\mathcal{R}_{n,(2,2)}f_2\Big),
\end{aligned}$$

where

$$\mathcal{R}_{n,(a,b)}f = \frac{1}{n}\sum_{i=1}^{n} X_{a,i}\langle X_{b,i}, f\rangle.$$

Define $\{\phi_{1,j}, \nu_{1,j}\}$ and $\{\phi_{2,j}, \nu_{2,j}\}$ as before. Then,

$$\mathcal{R}_{n,(a,b)} = \sum_j \sum_k r_{jk}^{(a,b)} \phi_{a,j} \otimes \phi_{b,k},$$

where

$$r_{jk}^{(a,b)} = \frac{1}{n}\sum_{i=1}^{n} \langle X_{a,i}, \phi_{a,j}\rangle \langle X_{b,i}, \phi_{b,k}\rangle.$$

Then,

$$\begin{aligned} L^*Lf = \Big(&\sum_j \sum_k (r_{jk}^{(1,1)} \nu_{1k}^{-1/2} f_{1k} + r_{jk}^{(1,2)} \nu_{2k}^{-1/2} f_{2k}) \mathcal{K}_1 \phi_{1j} \\ &\sum_j \sum_k (r_{jk}^{(2,1)} \nu_{1k}^{-1/2} f_{1k} + r_{jk}^{(2,2)} \nu_{2k}^{-1/2} f_{2k}) \mathcal{K}_2 \phi_{2j} \Big). \end{aligned}$$

Furthermore,

$$\begin{aligned} \langle g, L^*Lf\rangle &= \sum_j \sum_k \big(r_{jk}^{(1,1)} \nu_{1j}^{-1/2} \nu_{1k}^{-1/2} g_{1j} f_{1k} + r_{jk}^{(1,2)} \nu_{1j}^{-1/2} \nu_{2k}^{-1/2} g_{1j} f_{2k} \\ &\qquad + r_{jk}^{(2,1)} \nu_{2j}^{-1/2} \nu_{1k}^{-1/2} g_{2j} f_{1k} + r_{jk}^{(2,2)} \nu_{2j}^{-1/2} \nu_{2k}^{-1/2} g_{2j} f_{2k} \big) \\ &= \sum_j \sum_k \sum_{a=1}^{2} \sum_{b=1}^{2} r_{jk}^{(a,b)} \nu_{aj}^{-1/2} \nu_{bk}^{-1/2} g_{aj} f_{bk}. \end{aligned}$$

Therefore, as in the single predictor case, the mapping corresponding to L^*L in $\ell^2 \times \ell^2$ is

$$\mathcal{Z} = \mathcal{D}\mathcal{G}\mathcal{D},$$

where

$$\mathcal{D} = \begin{pmatrix} \mathcal{D}_1 & \\ & \mathcal{D}_2 \end{pmatrix}, \quad \mathcal{G} = \begin{pmatrix} \mathcal{G}_{11} & \mathcal{G}_{12} \\ \mathcal{G}_{21} & \mathcal{G}_{22} \end{pmatrix}$$

and

$$\mathcal{D}_a = \sum_j \nu_{aj}^{-1/2} e_j \otimes e_j, \quad \mathcal{G}_{a,b} = \sum_j \sum_k r_{jk}^{(a,b)} e_j \otimes e_j.$$

The mapping corresponding to Λ is given by

$$\mathcal{W}_\lambda = \begin{pmatrix} \lambda_1 I & \\ & \lambda_2 I \end{pmatrix}.$$

Define $t_{n,k}^{(a)} = \sum_{j>k} \rho_{n,j}^{(a)}$ and $t_k^{(a)} = \sum_{j>k} \rho_j^{(a)}$ for $a = 1, 2$ similarly to the single predictor case. The result corresponding to Theorem III.9 is given as follows.

Theorem III.15. *For all* $j_1, j_2, k_1, k_2 \geq 1$,

$$\operatorname{tr}(G_\lambda^{-1} L^* L) \leq \sum_{a=1}^{2} \Big(k_a + j_a + \frac{t_{n,k_a}^{(a)}}{\lambda_a \nu_{a,j_a+1}} \Big).$$

Proof. Let P be the projection of the span of the first $k_1 + k_2$ eigenvectors of $\mathcal{G}$ and $P' = 1 - P$. Then,

$$\begin{aligned}
\operatorname{tr}(G_\lambda^{-1} L^* L) &= \operatorname{tr}((\mathcal{Z} + \mathcal{W}_\lambda)^{-1} \mathcal{Z}) \\
&\leq \operatorname{tr}((\mathcal{D}\mathcal{G}P\mathcal{D} + \mathcal{W}_\lambda)^{-1} \mathcal{D}\mathcal{G}P\mathcal{D}) + \operatorname{tr}((\mathcal{D}\mathcal{G}P'\mathcal{D} + \mathcal{W}_\lambda)^{-1} \mathcal{D}\mathcal{G}P'\mathcal{D}) \\
&\leq k_1 + k_2 + \operatorname{tr}((\mathcal{D}\mathcal{G}P'\mathcal{D} + \mathcal{W}_\lambda)^{-1} \mathcal{D}\mathcal{G}P'\mathcal{D})
\end{aligned}$$

Define $e_i^{(1)} = (e_i, 0)^T$ and $e_i^{(2)} = (0, e_i)^T$. Then,

$$\begin{aligned}
\operatorname{tr}((\mathcal{DG}P'\mathcal{D} + \mathcal{W}_\lambda)^{-1}\mathcal{DG}P'\mathcal{D}) &= \sum_a \sum_i \langle e_i^{(a)}, (\mathcal{DG}P'\mathcal{D} + \mathcal{W}_\lambda)^{-1}\mathcal{DG}P'\mathcal{D}e_i^{(a)}\rangle \\
&\le j_1 + j_2 + \sum_a \frac{1}{\lambda_a} \sum_{i>j_a} \langle e_i^{(a)}, \mathcal{DG}P'\mathcal{D}e_i^{(a)}\rangle \\
&\le j_1 + j_2 + \sum_a \frac{t_{n,k_a}^{(a)}}{\lambda_a \nu_{a,j_a+1}},
\end{aligned}$$

where the last step follows from the single predictor case. □

Assuming $t_k^{(a)} = O(k^{-\beta_a})$ and setting $\lambda_a = n^{-\frac{2m+\beta+1}{2m+\beta+2}}$, we can obtain

$$\mathbb{E}_\varepsilon \|L(\widehat{f}_\lambda - f)\|^2 = O_p\Big(\sum_{a=1}^{2} n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}}\Big).$$

Next, we consider the prediction convergence rate

$$\mathbb{E}_{\varepsilon,X}\{L_X(\widehat{f}_\lambda - f)\}^2 = \mathbb{E}_{\varepsilon,X}\{L_{X_1}(\widehat{f}_1 - f_1) + L_{X_2}(\widehat{f}_2 - f_2)\}^2$$

Define $\mathcal{R}_n = L^*L$ and $\mathcal{R} = \mathbb{E}_X L_X^* L_X$. Note

$$\begin{aligned}
\langle f, \mathcal{R}f\rangle &= \mathbb{E}_X \|L_1 f_1 + L_2 f_2\|^2 \\
&\le 2(\mathbb{E}_{X_1}\|L_{X_1} f_1\|^2 + \mathbb{E}_{X_2}\|L_{X_2} f_2\|^2).
\end{aligned}$$

Thus, using the result from the single predictor case, for any $k_1, k_2 \ge 1$,

$$\langle f, \mathcal{R}f\rangle \le \Big\{2\sum_{a=1}^{2} \langle f_a, \mathcal{R}_{n,(a,a)} f_a\rangle + O_p(n^{-1})\|f_a\|^2 \sum_{r=1}^{k_a} t_r^{(a)} + O_p(n^{-1/2})\|f_a\|^2 t_{k_a}^{(a)}\Big\}.$$

Thus again using the result from the single predictor case, we obtain the prediction convergence rate.

Theorem III.16. *Assume that (3.14) holds for both X_1 and X_2, and $t_k^{(a)} = O(k^{-\beta_a})$ for some $\beta_a > 0$ and $a = 1, 2$. Then, taking $\lambda_a = n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}}$ yields*

$$\mathbb{E}_{\varepsilon,X}\{L_X(\widehat{f}_\lambda - f)\}^2 = O_p\Big(\sum_{a=1}^{2}(n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}} + n^{-\frac{\beta_a+1}{2}})\Big).$$

Finally, we prove that $\sum_{a=1}^2 n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}}$ is the optimal rate. The proof is a straightforward extension of the one predictor case.

Corollary III.17. *There exists a constant $d > 0$ such that*

$$\liminf_{n\to\infty} \inf_{\widetilde{f}\in\mathcal{F}} \sup_{\mathbb{P}\in\mathcal{P}(\beta), f\in W_m^2} \mathbb{P}(\mathbb{E}_X\{L_X(\widetilde{f} - f)\}^2 > d\sum_{a=1}^{2} n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}}) > 0.$$

Proof. Using linearity, the proof can be derived straightforwardly from the one-predictor case. Create a set of probability measures $\mathcal{P}^*$ so that $X_{1,i}$ and $X_{2,i}$ are independent. □

The theory for discretely observed data also straightforwardly applies to the multiple predictor case.

Corollary III.18. *For $a = 1, 2$, assume that $\widetilde{t}_k^{(a)} = O(k^{-\beta_a})$ for some $\beta_a > 0$ and there exists a fixed $C < \infty$ such that*

$$\mathbb{V}\mathrm{ar}(\widetilde{\xi}_{i,r}^{(a)}\widetilde{\xi}_{i,s}^{(a)}) \le C\widetilde{\rho}_r^{(a)}\widetilde{\rho}_s^{(a)} \quad \textit{for all } r, s.$$

Then, setting $\lambda_a = n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}}$ yields

$$\mathbb{E}_{\varepsilon,X}\{L_X(\widetilde{f}_\lambda - f)\}^2 = O_p\Big(\sum_{a=1}^{2}(n^{-\frac{2m_a+\beta_a+1}{2m_a+\beta_a+2}} + n^{-\frac{\beta_a+1}{2}} + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\|X_{a,i} - \widetilde{X}_{a,i}\|_{L^2}^2)\Big).$$

3.6 Discussion

In this chapter, we explored a regularized approach to the functional linear regression under the framework of the reproducing Hilbert kernel space. Our derivation of the optimal convergence rate is much simpler than the existing literature, such as Crambes et al. (2009) or Yuan and Cai (2010), showing clearly what is behind the approach. It reveals that the choice of the penalty is not so important in terms of the convergence rate as long as the penalty term in the objective function makes it invertible to find the unique solution. We bypassed the theoretical setting where the functional predictors are fully observed to the practical setting where the functional predictors are only discretely observed by a very simple way. We also explored the model that accommodates the multiple predictors and its effect on the convergence rate.

This chapter focused primarily on the theoretical aspect of the problem, devoting all our efforts to the analytical derivation. For a practical purpose, more thorough numerical explorations are desired since, for example, the choice of the penalty term or the order of the functional space have a significant impact on the performance in reality. We leave these computational aspects of the problem for the future study.

CHAPTER IV

Estimation of the Composite Risk in High-Throughput Screening Assay Analysis

4.1 Introduction

Advances in assay technology have made it possible to conduct millions of biochemical tests using machines; an automated system saves a vast amount of time and cost as it no longer necessitates experimenters to manually conduct the tests by hand. This experimental approach is called high-throughput screening (HTS) [Zhang et al. (1999), Zhang (2011)]. The limitation of such screening method is that the type of tests are limited to only those which does not require careful human attention, such as those focusing on the molecular features of the chemical compounds, rather than more complicated tests that can directly assess the effects on living organisms or humans; the HTS assays require only a low effort to obtain but do not directly reveal the phenotypic features that we are most interested. Animal testing is conducted to assess the direct effects on living organisms, but it is costly in terms of both time and money. Furthermore, it also involves ethics issues. One promising way to overcome such issues is to utilize information from the HTS assay results to predict phenotypic effects, instead of conducting animal testing one by one. In this respect, it is essential to explore the prediction relationship between the HTS and conventional assays. What is challenging in the HTS assay analysis is that it has to deal with the large dimensionality or the weak signals.

The specific object of interest in this paper is the toxicity of the chemical compounds. A conventional approach to chemical toxicity testing uses phenotypic live-cells or animal testing of each chemical compound of interest. Multiple such assays are performed to characterize the effects of a compound on all organ systems. The overall risk of a compound can be then determined by the lowest concentration at which it has an adverse effect of any type. This approach is, however, expensive and time-consuming. Thus we want to use the results of the HTS assays to predict the overall toxicity. The prediction results can be used to prioritize the compounds in terms of risk to save time and cost [Dix et al. (2007), Judson et al. (2010)]. In the following, we formulate the problem in the statistical model.

Denote a p-dimensional explanatory variable by $X \in \mathbb{R}^p$ and a q multiple dependent variables by $Y = (Y_1, \ldots, Y_q) \in \mathbb{R}^q$. In our problem, X corresponds to the HTS assay results, which require a low effort to obtain, and Y corresponds to the conventional assay results such as cell-culture assays or animal testing. The unit of the analysis is a chemical compound. Let $\mu_j(X) = \mathbb{E}(Y_j|X)$, which corresponds to the assay-specific risk predictable by the HTS assays. We define the overall risk by $\nu(X) = \min_j \mu_j(X)$, which we call the composite risk. Our goal is to estimate the composite risk function ν.

There are a couple of naive approaches to estimate ν. The first is a two-step approach where we first regress Y_j on X and then compute the minimum of the estimated assay-specific risk functions. This approach however requires to estimate a large number of parameters when q is large. Also, the operation of taking the minimum may cause a serious bias. The second is to regress the observed minimum, i.e., $\min_j Y_j$, on X. The observed minimum, however, is biased for ν, and the bias can be a serious issue when q is large. To avoid these issues, we focus the modeling efforts directly on ν. In this paper, we introduce an approach that specifies an explicit model for ν through the profile likelihood function.

In Section 4.2, we introduce a method to directly specify and estimate a model for the composite risk function through the profile likelihood function. In Section 4.3, several simulation studies are conducted to explore the properties of the profile likelihood approach and naive alternative approaches. We investigate under what situations the profile likelihood approach can be advantageous over the other approaches. In Section 4.4, we extend the idea of directly specifying a model for the composite risk through the profile likelihood function to the interval estimation for it. An alternative approach—namely the Wald-type approach—is also introduced and compared to the profile likelihood approach. In Section 4.5, we illustrate the estimation of the composite risk function using the ToxCast data of the EPA and the 60 cell line screen of the NCI. Finally, we conclude this paper with some remarks in Section 4.6.

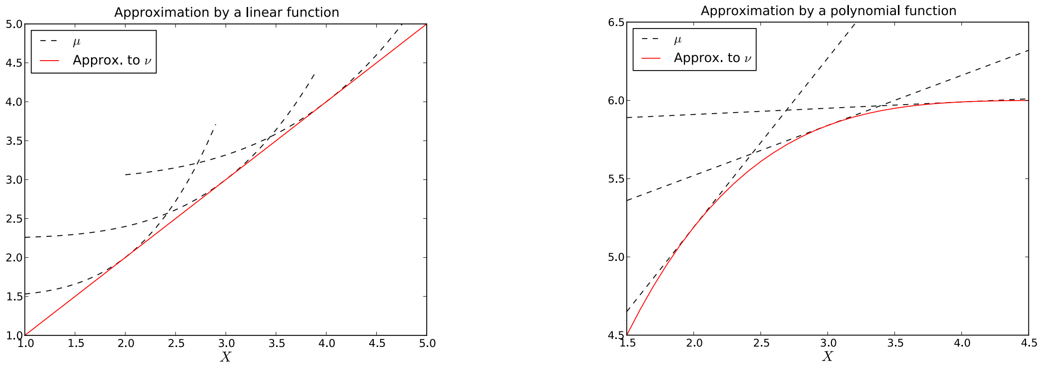
4.2 Estimation of the Composite Risk Function

4.2.1 Problem Set-Up

Let us recall the notation first. Denote the p-dimensional explanatory variable $X \in \mathbb{R}^p$ and the q multiple dependent variables $Y = (Y_1, \ldots, Y_q) \in \mathbb{R}^q$. Let $\mu_j(X) = \mathbb{E}(Y_j|X)$ and $\nu(X) = \min_j \mu_j(X)$. Given a random sample $((X_1, Y_1), \ldots, (X_n, Y_n))$ of (X, Y), our goal is to estimate the composite risk function ν. (By slightly abusing the notation, we also use the notation (X, Y_j) to refer to the data matrix, i.e., $X \in \mathbb{R}^{n \times p}, Y_j \in \mathbb{R}^{n \times 1}$.) This function can be used for two purposes. If we wish to predict the composite risk for a compound that has not been assessed for overall risk, but for which the HTS data X are available, $\nu(X)$ can be used as a prediction. Also, if we want to improve the experimental overall risk $\min_j Y_{ij}$, we can consider estimates of the form $w \min_j Y_{ij} + (1 - w)\nu(X_j)$ where $0 \leq w \leq 1$. The motivation for doing this is that ν borrows information from other assays and from other compounds that can improve the precision of the estimated composite risk. However, relying solely on $\nu(X_j)$, i.e. by setting $w = 0$, may lead to poor performance if the model is not sufficiently predictive. We call the first goal "out of sample prediction," and the second goal "in sample estimation."

The features of this problem are as follows. First, the object of interest involves an extreme, i.e., the minimum in the toxicity analysis. The involvement of the extreme inevitably causes a bias issue if not explicitly accommodated. Suppose that you have a good prospect about the model for each (X, Y_j) and can construct an unbiased estimator for $\mu_j(X)$, say $\widehat{\mu}_j(X)$. Then, the natural estimator for $\nu(X)$ is given by $\min_j \widehat{\mu}_j(X)$. However, this estimator incurs a downward bias due to the minimum. This naive two-step approach deteriorates even more when $\mu_1(X), \ldots, \mu_q(X)$ are similar to each other and/or when q is large. Second, there is a large number of nuisance (function) parameters. Note that although $\nu(X)$ can be computed from $\mu_1(X), \ldots, \mu_q(X)$, the entire shapes of $\{\mu_j\}$ do not appear in ν. In this sense, the effort to understand all of $\{\mu_j\}$ is redundant for the estimation of ν and may cause an unnecessarily large efficiency loss, in particular when q is large. One may consider regressing the observed minimum $\min_j Y_j$ on X, but $\min_j Y_j$ is unbiased for $\nu(X)$; hence this approach still suffers from a bias, again which can be a serious issue when q is large. In order to avoid these issues, we consider an approach that specifies an explicit model for ν, rather than $\mu_1, \ldots, \mu_q$.

In the following section, we introduce an approach that directly models and estimates ν through the profile likelihood function. By profiling the likelihood function with respect to ν, an explicit model for ν can be specified. Figure 4.1 provides two schematics to describe what we want to attain by this approach. The left panel of the figure represents the case where μ_j are polynomial functions but ν can be well approximated by a linear function; the right panel of the figure represents the case where μ_j are linear functions but ν can be well approximated by a polynomial func-



(a) Approximation by a linear function

(b) Approximation by a polynomial function

Figure 4.1: Schematics representing the idea of direct modeling.

tion. The key idea is that by directly specifying the model for ν, we can avoid a large number of nuisance parameters to estimate and the necessity of computing the minimum.

For simplicity, we focus on the normal additive error model throughout the paper:

$$Y_{ij} = \mu_j(X_i) + \varepsilon_{ij}, \quad \varepsilon_{ij} \sim N(0, \sigma_j^2),$$

where $i = 1, \dots, n$ and $j = 1, \dots, q$.

4.2.2 Estimation Through the Profile Likelihood Function

Denote the conditional log-density of Y_j given X by $\ell_j(y_j|\mu_j(x))$. Let $\mu(x) = (\mu_1(x), \dots, \mu_q(x))$. Then, we define the profile likelihood function of ν given X and Y by

$$g(\nu) = \max_{\mu(x) \in \mathcal{C}(\nu(x))} \sum_{j=1}^{q} \ell_j(y_j|\mu_j(x)), \tag{4.1}$$

where

$$\mathcal{C}(u) := \{(\mu_1, \dots, \mu_q) \in \mathbb{R}^q : \mu_j \geq u, \ \forall j \quad \wedge \quad \mu_j = u, \ \exists j\}.$$

The form of (4.1) allows us to specify the model for the target parameter ν directly; we can leave the form of μ_j completely arbitrary, so that we only need to specify a model for ν regardless of the size q and avoid the necessity of modeling and estimating all μ_j. We can also put a constraint on μ_j such as a smoothness condition if we wish to interpret each μ_j. In the following, we leave μ_j to be completely arbitrary functions.

For simplicity, we assume that ν can be approximated by a linear function. More complex functions such as spline functions or shape restricted functions can also be used, for example when there is prior information on the shape of ν. Let $\nu(X) = \beta^T X$. Given observations $((x_1, y_1), \ldots, (x_n, y_n))$, we estimate β by maximizing the profile likelihood function

$$\mathcal{G}(\beta) = \max_{\wedge_i \mu(x_i) \in \mathcal{C}(\beta^T x_i)} \sum_{i=1}^{n} \sum_{j=1}^{q} \ell_j(y_{ij} | \mu_j(x_i)), \tag{4.2}$$

i.e., $\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}\, \mathcal{G}(\beta)$. To compute this, we can rewrite (4.2) to

$$\mathcal{G}(\beta) = \sum_{i=1}^{n} \max_{1 \le j \le q} \{\ell_j(y_{ij} | \beta^T x_i) + \sum_{j' \ne j} \ell_j(y_{ij'} | y_{ij'} \vee \beta^T x_i)\},$$

because $\ell_j(y_{ij}|u)$ is maximized at $u = y_{ij}$. For each i, we have to calculate the maximum only when $y_{ij} > \beta^T x_i$ for all j, because otherwise the maximum is given by $\sum_{1 \le j \le q} \ell_j(y_{ij} | y_{ij} \vee \beta^T x_i)$. When $y_{ij} > \beta^T x_i$ for all j, the maximum is given by $\max_{1 \le j \le q}\{\ell_j(y_{ij}|\beta^T x_i) + \sum_{j' \ne j} \ell_j(y_{ij'}|y_{ij'})\}$. Summarizing these, we have

$$\begin{aligned}\mathcal{G}(\beta) = \sum_{i=1}^{n} \Big[&(1 - 1\{y_{ij} > \beta^T x_i \text{ for all } j\}) \sum_{1 \le j \le q} \ell_j(y_{ij} | y_{ij} \vee \beta^T x_i) \\ &+ 1\{y_{ij} > \beta^T x_i \text{ for all } j\} \max_{1 \le j \le q} \{\ell_j(y_{ij}|\beta^T x_i) + \sum_{j' \ne j} \ell_{j'}(y_{ij'}|y_{ij'})\}\Big].\end{aligned} \tag{4.3}$$

This can be maximized by using a general optimization package.

4.2.3 Four Methods to Compare

In Section 4.3, we numerically explore the properties of the profile likelihood approach in comparison to three alternative approaches. The first approach (referred as OLS(min)) is the ordinary least squares estimator for the regression of the observed minimum of Y on X,

$$\widehat{\nu}(x) = x^T(X^TX)^{-1}X^TY_{min},$$

where $x \in \mathbb{R}^p, X \in \mathbb{R}^{n\times p}$ and $Y_{min} \in \mathbb{R}^n$, and Y_{min} is the vector whose ith element is $\min_{1\leq j\leq q} Y_{ij}$. The second approach (referred as min(OLS)) is the minimum of the ordinary least squares estimators for the regression of Y_j on X for $j = 1, \ldots, q$. This is a two-step approach, and the estimator is given by

$$\widehat{\nu}(x) = \min_j\{x^T(X^TX)^{-1}X^TY_j\},$$

where $Y \in \mathbb{R}^{n\times q}$. The third approach (referred as MARS) is similar to the second approach, min(OLS), where instead of the OLS estimator the Multivariate Adaptive Regression Splines (MARS) is used to regress Y_j on X, i.e.,

$$\widehat{\nu}(x) = \min_j\{\widehat{\mu}_j(x)\},$$

where $\widehat{\mu}_j$ is obtained by applying MARS to (X, Y_j). For the details about MARS, see Friedman (1991).

4.3 Simulation Study for the Estimation of the Composite Risk Function

In this section, we numerically evaluate the properties of the profile likelihood approach and the other alternative approaches for the estimation of the composite risk function. There are several aspects that will affect the performance of the estimation methods. First, for the profile likelihood approach to work well it is essential to specify a good model for ν, but it may not be easy to specify a perfect model. We explore how the specification of ν influences the performance of the profile likelihood approach, compared with the other naive approaches. It will be shown that even if ν is not perfectly specified, the profile likelihood approach still begins to show an advantage over the correctly-specified two-step approach as the specification of ν becomes closer to the true ν. Second, the varying dimensionality of the dependent variable, q, may have different effects between the profile likelihood approach and the other approaches. As mentioned before, larger q produces more nuisance parameters, and the number of the elements among which the minimum is computed grows. Thus the naive approaches incur a large bias and variance. In contrast, the profile likelihood approach estimates ν directly, hence it can avoid such issues. Third, the estimability of μ_j may deeply affect the naive approaches that need to estimate all the μ_j. Because it need not estimate each μ_j, the profile likelihood approach is robust against this issue. Finally, the error variances are treated differently by the procedures. The profile likelihood approach incorporates the variance information in the procedure as part of the likelihood function while the naive approaches do not utilize the error variance information in the procedures. It will be shown that the heteroscedasticity of the error variances has a negative effect on the methods except the profile likelihood approach, which is robust against this issue. In the following, we evaluate these aspects of the problem. In particular, we focus on four aspects:

the dimensionality of the dependent variable, the specification of ν, the smoothness of μ_j, and the heteroscedasticity of the error variances.

4.3.1 The Specification of the Composite Risk

First, we explore how the profile likelihood approach behaves when the specification of ν gradually becomes closer to the true one, along with the effect of the dimensionality of the dependent variable. We generate a sample of size 300 from a linear model $Y_j = \beta_j^T X + \varepsilon_j$ where $p = 10$, $X_i \sim N(0, I_p)$, $\varepsilon_j \sim N(0, 2^2)$, and

- $\beta_1 = (1, 0, \ldots, 0)^T$,
- $\beta_j[1] = \sqrt{1-\alpha^2}$ and $\beta_j[2:p] = \alpha \frac{z_j}{\|z_j\|}$ for $j = 2, \ldots, q$,

where $z_j = (z_{2j}, \ldots, z_{pj})^T$ is a standard normal random vector, $\|z_j\|^2 = \sum_{i=2}^p z_{ij}^2$, and $\beta_j[k]$ means the kth element of β_j and I_p is the p-dimensional identity matrix. We consider three cases: $q = 5, 15, 45$. α controls the proximity of the assumed ν to the true ν; as α moves from 1 to 0, ν becomes closer to a linear function. When $\alpha = 1$, β_j are on average orthogonal, and ν is the farthest from the linear form; when $\alpha = 0$, ν becomes exactly linear. In fact, $\nu(X) = \beta_1^T X = \cdots = \beta_q^T X$ when $\alpha = 0$ (see a schematic in Figure 4.2). β_j is standardized to have a unit norm, i.e., $\|\beta_j\| = 1$ for all j. The covariate X_i is generated by the standard normal distribution independently from β_j. Thus, $\beta_j^T X_i$ is also standardized in the sense that $\mathbb{E}(\beta_j^T X_i)^2 = 1$ for all i and j. We assume a relatively low signal to noise ratio as that is usually the case in the HTS assay analysis (the individual HTS assays have weak signals in terms of predicting the outcomes of the traditional assays). The MARS approach is excluded from this analysis as min(OLS) is the correctly specified approach under this setting.

For each sample, we compute an estimation bias $\frac{1}{n}\sum_{i=1}^n \{\nu(X_i) - \widehat{\nu}(X_i)\}$ and a mean squared error (MSE) $\frac{1}{n}\sum_{i=1}^n \{\nu(X_i) - \widehat{\nu}(X_i)\}^2$, where $\widehat{\nu}$ is the estimate obtained by each approach. Repeating this procedure 300 times, we compute the average bias and the average MSE for $\alpha = (1, 0.7, 0.5, 0.3, 0.1, 0)$. The results are shown in Figure 4.3. At $\alpha = 1$, the profile likelihood approach has much worse MSE than min(OLS). As α approaches to 0, the profile likelihood approach becomes better in terms of both bias and MSE while min(OLS) becomes worse, and the superiority reverses from a certain point on. Why min(OLS) becomes worse is that as α moves toward 0, μ_j become similar each other, which makes the bias issue more serious; in fact, min(OLS) has a downward bias due to the minimum and the bias deteriorates as α approaches 0. The second approach (OLS(min)) is always worse than the other two approaches in terms of MSE.

Figure 4.4 shows the ratio of the MSE of min(OLS) to that of the profile likelihood approach. Note that as the specification of ν becomes closer to the true ν, the ratio increases and becomes higher than one from a certain point on, where the profile likelihood approach begins to have a smaller MSE than the two-step approach. As q increases, the ratio becomes larger and the point where the ratio exceeds one shifts toward earlier points, indicating that the profile likelihood approach becomes more advantageous for the larger dimensionality over min(OLS). Note that except $\alpha = 0$ the profile likelihood approach is always a misspecified approach while the third approach is always the correctly specified approach, including $\alpha = 0$, as μ_j are always linear functions regardless of the value of α. Thus, when a good approximation to ν is available, the profile likelihood approach may behave better than the correctly specified two-step approach.

4.3.2 Varying Smoothness of the Assay-Specific Risk Functions

Next, we consider the case where ν is linear while μ_j are nonlinear. We explore how varying smoothness of μ_j affects the procedures. The MARS approach is now included in the analysis in addition to the three approaches. Our aim is to investigate the effect of the smoothness of μ_j, in particular, on the profile likelihood approach and

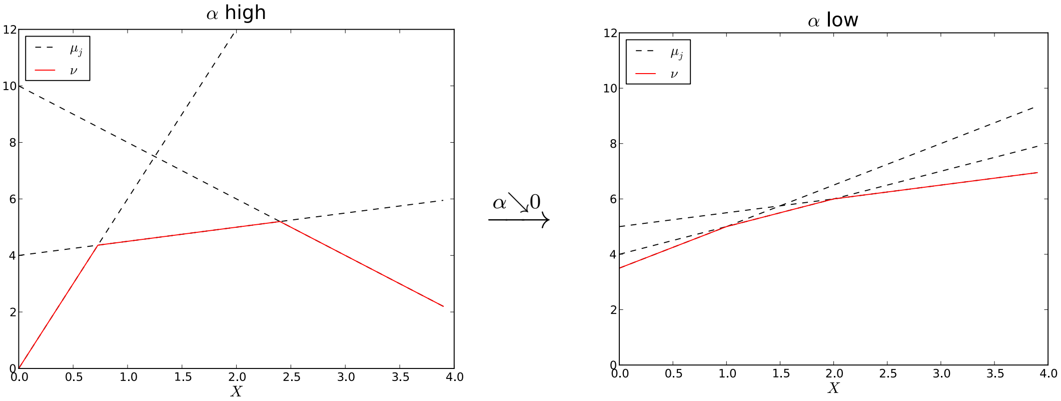
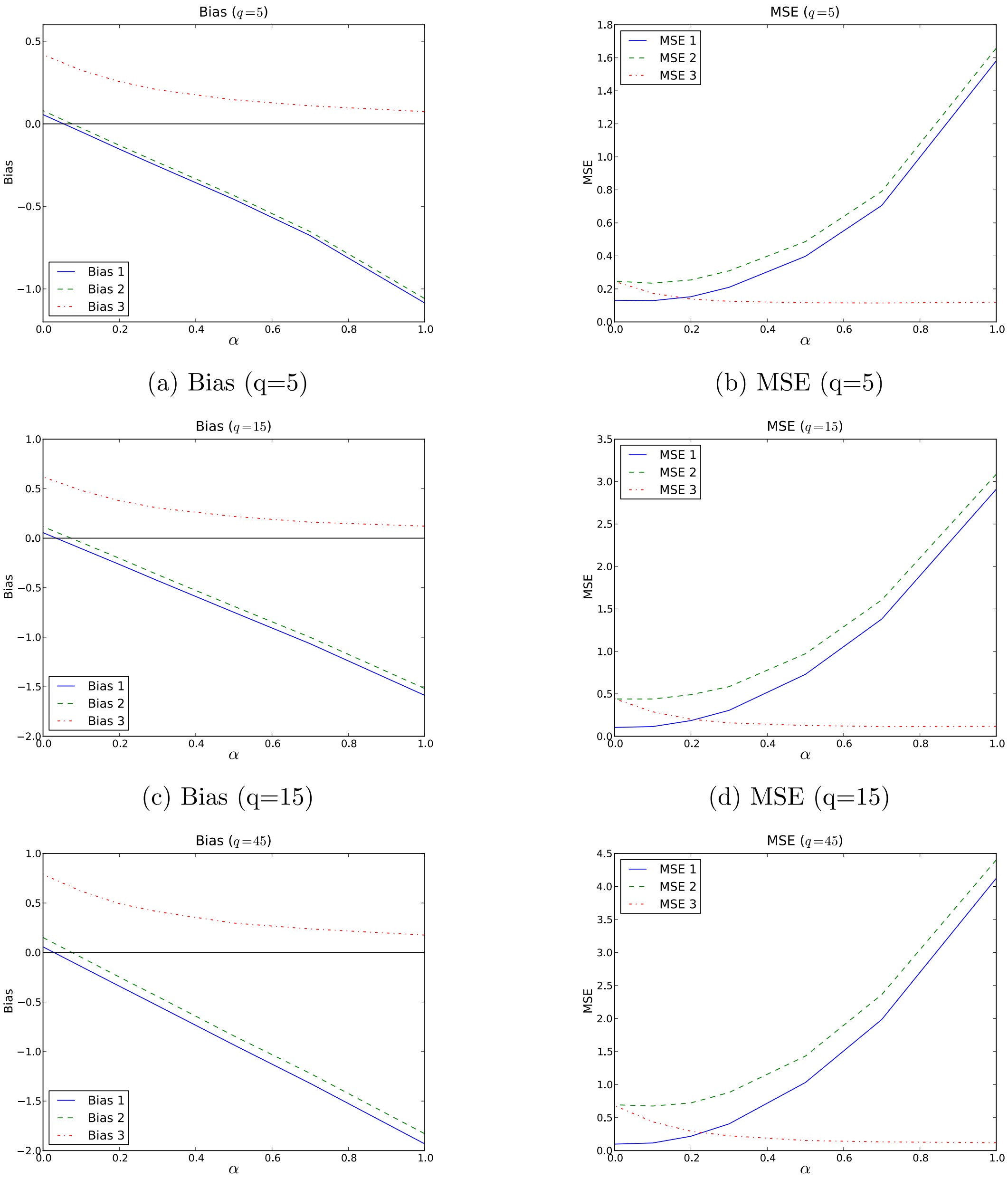


Figure 4.2: Schematic for the simulation study concerning the specification of the composite risk

the MARS approach. First, we consider the case where μ_j are wiggly and difficult to estimate. We generate a sample of size 300 from $Y_j = \mu_j(X) + \varepsilon_j$ where $X_i \sim N(0, I_p)$, $p = 10$, $\varepsilon_j \sim N(0, 2^2)$, and setting $\nu(X_i) = \beta^T X_i$ where $\beta = (1, 0, \ldots, 0)$:

- Generate $n = 300$ random integers from 1 through q, and denote them by $(k_1, \ldots, k_n)$.
- Set $\mu_{k_i}(X_i) = \beta^T X_i$.
- Set $\mu_j(X_i)$ to be $\beta^T X_i$ plus a χ^2_1-random variable independently for $j \neq k_i$.

Under this model, $\nu(X_i) = \min_{1 \leq j \leq q} \mu_j(X_i) = \beta^T X_i$ while μ_j becomes extremely wiggly (see the schematic in Figure 4.5a). We again consider the three cases $q = 5, 15, 45$ to see the effect of the dimensionality of the dependent variable. With 300 repetitions, we compute the average MSE and the average bias. The results are given in Table 4.1. As can be seen, the profile likelihood approach has much better MSE than all the other methods. It may even be surprising that the performance improves as the dimension of the dependent variable grows. Since the bias remains the same, this improvement is solely due to the reduction in variance. That implies that for the profile likelihood approach a larger sample due to a larger q means simply that more information is available rather than that it is the cause of the bias. The two



(a) Bias (q=5)

(b) MSE (q=5)

(c) Bias (q=15)

(d) MSE (q=15)

(e) Bias (q=45)

(f) MSE (q=45)

Figure 4.3: The transition of a bias (Left) and MSE (Right) with varying α. 1: PL, 2: OLS(min), 3: min(OLS)

approaches using OLS become worse as q increases. Most of the MSE of MARS is attributable to bias, which implies that MARS basically fails to estimate such wiggly functions.

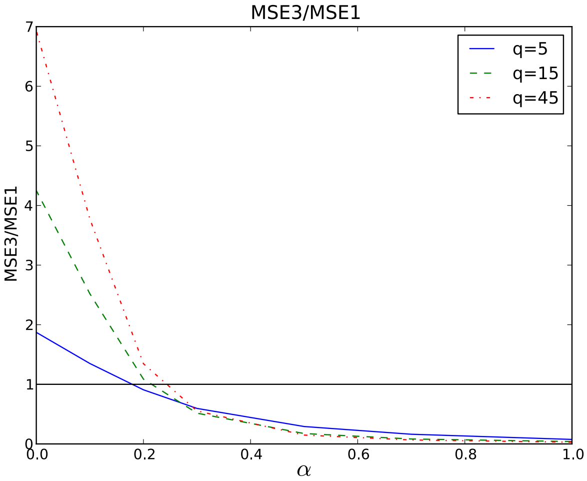
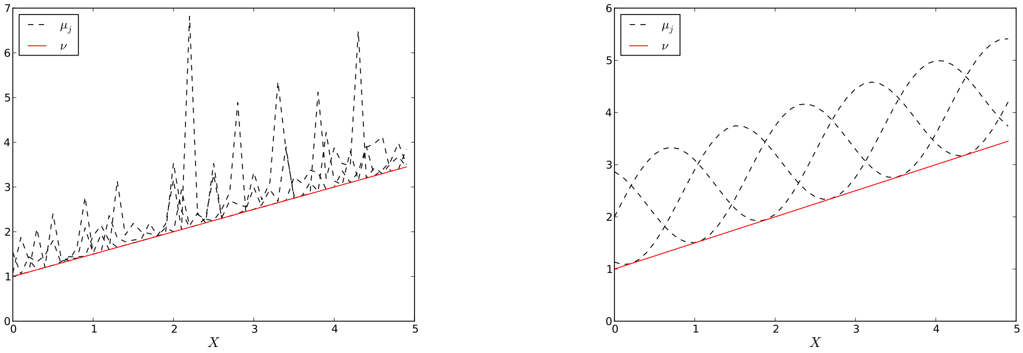


Figure 4.4: The ratio of the MSE of min(OLS) to that of PL with varying α.



(a) Wiggly μ_j

(b) Smooth μ_j

Figure 4.5: Schematic for the simulation study concerning the smoothness of the assay-specific risks

Second, we consider the case where μ_j is smooth. The above simulation setting continues to be used except that μ_j are now given as follows:

- $\mu_1(X_i) = \beta^T X_i$ where $\beta = (1, 0, \ldots, 0)$
- $\mu_j(X_i) = \beta^T X_i + \sin\{(\|X_i\|/p + (j-1)/q)\pi\} + 1$

Under this model, $\nu(X_i) = \beta^T X_i$ and all μ_j become smooth functions (see the schematic in Figure 4.5b). The results are given in Table 4.2. Now, the MARS approach performs best. Unlike the last setting where μ_j were wiggly, the MARS

Approach	PL	OLS(min)	min(OLS)	MARS
$q = 5$	0.116	0.179	0.338	0.344
	(0.049)	(0.062)	(0.472)	(-0.534)
$q = 15$	0.095	0.331	0.632	0.374
	(0.050)	(0.097)	(0.721)	(-0.569)
$q = 45$	0.081	0.528	0.985	0.325
	(0.049)	(0.127)	(0.933)	(-0.521)

Table 4.1: MSE and bias for the simulation study where the assay-specific risks are wiggly. The bias is given in the parentheses.

Approach	PL	OLS(min)	min(OLS)	MARS
$q = 5$	0.117	0.169	0.259	0.054
	(0.047)	(0.057)	(0.405)	(0.010)
$q = 15$	0.084	0.289	0.477	0.045
	(0.047)	(0.090)	(0.623)	(0.010)
$q = 45$	0.076	0.495	0.733	0.045
	(0.049)	(0.125)	(0.801)	(0.015)

Table 4.2: MSE and bias for the simulation study where the assay-specific risks are smooth. The bias is given in the parentheses.

approach has fairly small bias and seems robust against the dimensionality of the dependent variable. The profile likelihood approach again improves as the dimensionality grows. The two approaches using OLS become worse as the dimensionality grows.

4.3.3 Heteroscedasticity of the Error Variances

In the previous sections, the variance errors were assumed to be homoscedastic. Now, we show that the heteroscedasticity of the variance errors negatively affect the MARS approach while it has a positive effect on the profile likelihood approach. We continue to use the last simulation setting except that:

- $\sigma_j = |2 + s \cdot z_j|$ where $z_j \sim N(0, 1)$
- $s = 0, 0.5, 1$

Approach	PL	OLS(min)	min(OLS)	MARS
$s = 0$	0.084	0.289	0.477	0.045
	(-0.047)	(-0.057)	(-0.405)	(-0.010)
$s = 0.5$	0.074	0.302	0.541	0.047
	(-0.043)	(-0.09)	(-0.644)	(-0.029)
$s = 1$	0.049	0.403	0.684	0.063
	(-0.032)	(-0.099)	(-0.711)	(-0.037)

Table 4.3: MSE and bias for the case where ν is linear while μ_j are smooth, but the error variances are heterogeneous ($q = 15$). The bias is given in the parentheses.

s controls the degree of the heteroscedasticity; as s increases, the error variances become more heteroscedastic. The results for $q = 15$ are given in Table 4.3. As the error variances become more heteroscedastic, all the approaches become worse except the profile likelihood approach, which actually becomes better in terms of both MSE and bias. This may be because only the profile likelihood approach automatically accounts for the heteroscedasticity by including the variances in the profile likelihood function while the other approaches do not take into account the heteroscedasticity in their procedures.

4.3.4 Summary

In this section, we numerically evaluated the properties of the profile likelihood approach and the other naive approaches to estimate the composite risk function. Because the profile likelihood function need not estimate each assay-specific risk by specifying an explicit model for the composite risk, it can reduce the dimensionality of the parameters and avoid the computation of the minimum. Through a variety of simulation studies, we confirmed what situations the use of the profile likelihood approach can be beneficial. First, even if the composite risk function is not perfectly specified, the profile likelihood approach can still display better performance than the correctly-specified two-step approaches when the specification is relatively well. Second, the smoothness of the assay-specific risk functions affects the naive two-step approach that relies on the estimability of these intermediate quantities; the profile likelihood approach is not influenced by this issue because it directly estimates the composite risk. When the assay-specific risk functions are wiggly and difficult to estimate, the profile likelihood approach performs much better than the two-step approach. When the assay-specific risks are smooth, the MARS approach works fairly well though it does not display the performance improvement when q grows unlike the profile likelihood approach. Third, the heteroscedasticity of the error variances has a positive effect on the profile likelihood approach, while it has a negative effect on the naive approaches. These distinct effects seem to be due to whether a procedure accounts for the heteroscedasticity in its estimation procedure; the profile likelihood approach incorporates it into the likelihood function while the other approaches do not explicitly use it in their procedures. Finally, a large dimensionality due to large q reflects the availability of more information for the profile likelihood approach, which improves the performance with increasing q, while it can be the cause of a serious bias for the other approaches.

4.4 Interval Estimation for the Composite Risk

In this section, we extend the idea of directly specifying a model for $\nu(x)$ through the profile likelihood function to construct a confidence interval (CI) for the composite risk. Recall that our motivating problem is the toxicity analysis. In terms of health protective perspective, interval estimation may be more appropriate than point estimation as we have control over how conservative we want to be. We introduce two approaches: the profile likelihood CI and the Wald-type CI. We show their similarities and dissimilarities both numerically and analytically.

4.4.1 Profile Likelihood Interval Estimation

We introduce an interval estimation method for $\nu(x)$ through the profile likelihood function. Denote the log-likelihood function of a sample of size n by $\ell_n(\beta)$ where $\beta = (\beta_1, \dots, \beta_q) \in \mathbb{R}^{p \times q}$. Then, define the profile log-likelihood function for $\nu := \nu(x) = \min_j\{\beta_j^T x\}$ by

$$\mathcal{G}(u) = \max_{\beta \in \mathcal{C}(u)} \ell_n(\beta),$$

where

$$\mathcal{C}(u) = \{\beta \in \mathbb{R}^p : \beta_j^T x \geq u \text{ for all } j \;\wedge\; \beta_j^T x = u \text{ for at least one } j\}. \tag{4.4}$$

Then, the $(1-\alpha)$-level confidence interval for ν is given by

$$\mathcal{I} = \{u \in \mathbb{R} : \mathcal{G}(u) \geq \max_v \mathcal{G}(v) - \chi^2_{1,\alpha}\},$$

where $\chi^2_{1,\alpha}$ is the upper αth quantile for the χ^2-distribution with one degree of freedom. In particular, the confidence lower bound is given by

$$\widehat{\nu}_{lb} = \min\{\mathcal{I}\},$$

and similarly the confidence upper bound $\widehat{\theta}_{ub}$ can be obtained. Note that $\mathcal{C}(u)$ is locally of dimension $pq-1$, except where more than one of $(\beta_1^T x, \dots, \beta_q^T x)$ simultaneously become the minimum. Therefore, assuming that there exists a unique minimum in the true parameters $(\beta_1^T x, \dots, \beta_q^T x)$, the likelihood ratio test (LRT) statistic

$$\mathcal{R}_n = 2\{\max_u \mathcal{G}(u) - \mathcal{G}(\nu)\}, \tag{4.5}$$

follows asymptotically a χ^2-distribution with one degree of freedom.

As before, we assume the normality of the errors. Thus, the underlying model becomes

$$Y_{ij} = \beta_j^T X_i + \varepsilon_{ij}, \tag{4.6}$$

where $\varepsilon_{ij} \sim N(0, \sigma_j^2)$ for $1 \leq i \leq n$ and $1 \leq j \leq q$. To use the profile likelihood approach for this model, we need to know the variances. We treat the maximum likelihood estimator of σ_j^2 as the proxy for the true parameters, and implement the profile likelihood approach as described above. In the next section, we present the algorithm to compute the confidence interval.

4.4.2 Algorithm for Profile Likelihood Interval Estimation

Assume that variances are known; in practice, we can substitute the MLE of the variances. We describe the algorithm with respect to the lower bound; the algorithm for the upper bound can be constructed similarly. The key observation is that $\mathcal{C}(u)$ in (4.4) can be decomposed into $\cup_{k=1}^q \mathcal{C}_k(u)$ where

$$\mathcal{C}_k(u) = \{\beta \in \mathbb{R}^{p \times q} : \beta_j^T x \geq u \text{ for all } j \neq k \ \wedge \ \beta_k^T x = u\}.$$

The algorithm proceeds as follows:

1. Fix $u < \min_j \widehat{\beta}_j^T x$ where $\widehat{\beta}$ is the unconstrained MLE.

2. Fix $k \in [1, 2, \ldots, q]$.

3. Calculate $M_k(u) = \max_{\beta \in \mathcal{C}_k(u)} \ell_n(\beta)$.

4. Repeat 2–3 to calculate $M(u) = \max_k M_k(u)$.

5. Repeating 1–4, find $\widehat{\nu}_{LB}$ that satisfies $2\{\ell_n(\widehat{\beta}) - M(\widehat{\nu}_{LB})\} = (1.96)^2$.

The step 3 can be achieved by constrained quadratic programming under the normality assumption [Nocedal and Wright (2006)]. Note that we have to calculate the constrained maximum likelihood for k and only $j \neq k$ such that $\widehat{\beta}_j^T x < u$ because the remaining parts are canceled out in the step 5. In particular, when calculating the lower bound, we only need to calculate the constrained maximum likelihood for the kth variable. For the step 5, we use simple bisection algorithm, which turns out to be fairly fast in our analysis.

4.4.3 Censored Dependent Variables

In this section, we consider the case where dependent variables are censored. In the toxicity analysis, the response may be upper-censored because a compound may cause no significant change within the range of doses. The model for (Y_{ij}, X_i) is then given by

$$\begin{aligned} Y_{ij}^* &= \beta_j^T X_i + \varepsilon_{ij}, \\ Y_{ij} &= \min\{Y_{ij}^*, T_j\}. \end{aligned} \tag{4.7}$$

The censor thresholds T_j are assumed known. Note that this is the well-known Tobit model, a classic model in econometrics [Tobin (1958)], hence a variety of computational packages to compute the maximum likelihood estimator is available. The difference between this model and the non-censored model occurs in the form of the likelihood function. In the censored model, the likelihood is no longer quadratic, hence we cannot use quadratic programming in the step 3 of the algorithm. We use a nonlinear conjugate gradient algorithm for unconstrained optimization, and the constrained optimization by linear approximation for constrained optimization [Nocedal and Wright (2006), Powell (1994)]. In the following, we numerically confirm that the profile likelihood CI has the correct coverage probability and the null likelihood ratio statistic does approximately follow a chi-squared distribution with one degree of freedom for both non-censored and censored models.

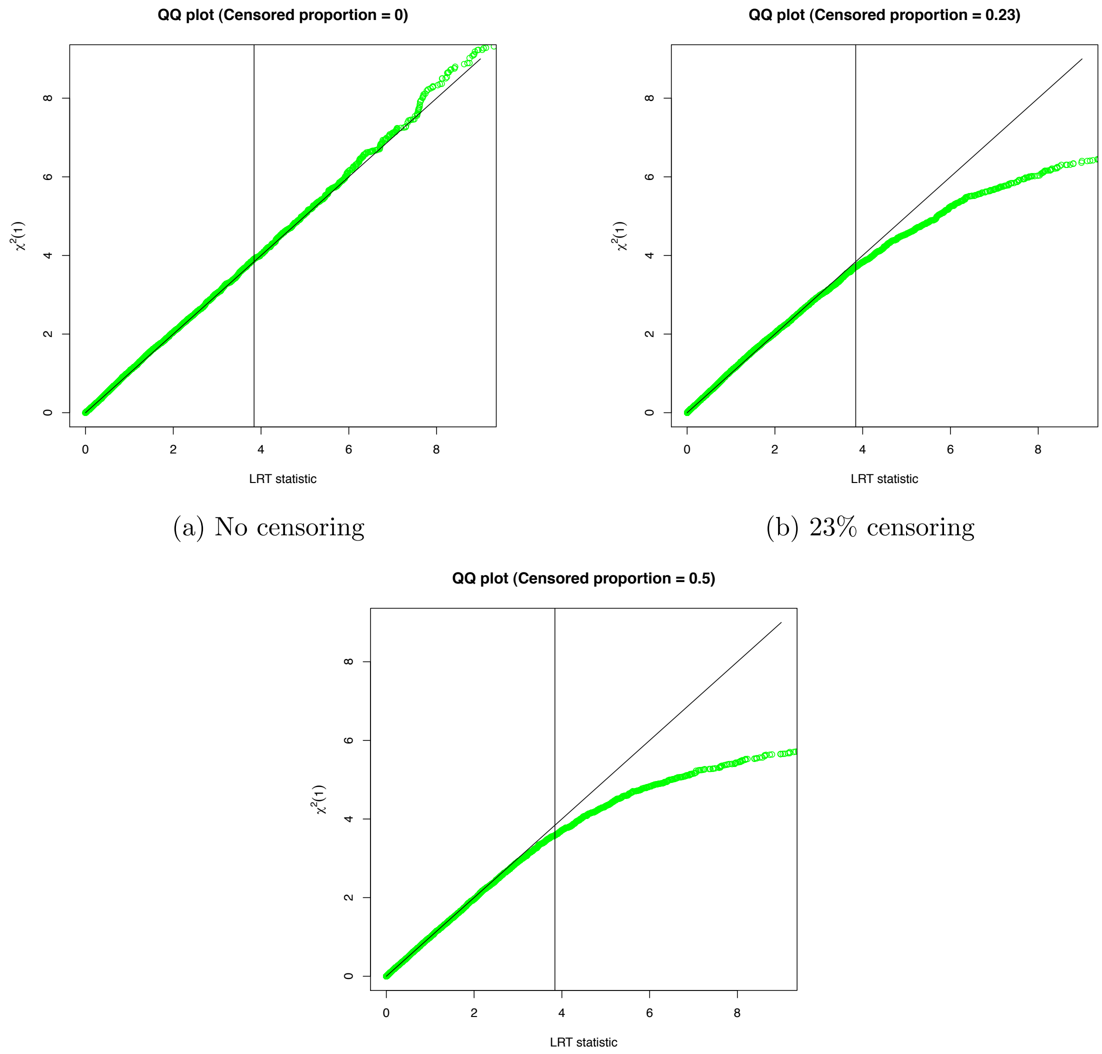
Let us check numerically whether the likelihood ratio test statistic (4.5) under the model (4.7) has a χ^2-distribution with one degree of freedom under the hypothesis that the conditional means $\beta_1^T X_i, \ldots, \beta_q^T X_i$ have no ties. We generate 5000 samples from the model (4.7) where we set $n = 300, p = 10, q = 3, \sigma = 1.5$. β and X are generated from the standard normal distribution, and we computed 5000 LRT statistics. We considered three cases: non-censoring case, $T_j \sim U(2.5, 3.5)$, and $T_j \sim U(0.5, 1.5)$. The latter two cases cause censoring, respectively, 23% and 50% of the times. Figure 4.6 shows the Q-Q plots between the LRT statistic and the χ^2-distribution with degrees of freedom 1. Note that the more censoring occurs the more the tail of the LRT statistic deviates from $\chi^2(1)$. However, the two distributions are quite similar within 95% region, hence the approximate coverage of the profile likelihood approach remains 95%.

4.4.4 Alternative Approach: the Wald-Type Confidence Interval

As can be seen, the profile likelihood approach can construct the confidence interval for any point, whether a point is within the sample or outside of the sample. When the goal is to create an estimation interval for a point within the sample, another approach is possible. Note that the minimum function $f(y_1, \ldots, y_q) = \min\{y_1, \ldots, y_q\}$ is differentiable everywhere except there are ties, and its derivative is one. Thus, if there is no tie among $\mu_1(X_i), \ldots, \mu_q(X_i)$, the usual Delta-method approach can be applied, and it can be easily seen that conditional on X_i,

$$\min_j(Y_{ij}) - \min_j(\mu_j(X_i)) = Y_{j'} - \mu_{j'}(X_i) \sim N(0, \sigma_j^2),$$

where $j' = \operatorname*{argmin}_j \mu_j(X_i)$. Unfortunately, j' is not observable so we use $j^* = \operatorname*{argmin}_j(Y_{ij})$ as a surrogate. Then, the Wald-type 95% confidence interval for $\mu_{j'}(X_i)$



(a) No censoring

(b) 23% censoring

(c) 50% censoring

Figure 4.6: QQ plots between the likelihood ratio test statistic and χ^2-distribution with degrees of freedom one. The horizontal line is 0.95th quantile of $\chi^2(1)$.

is given by

$$\min_j(Y_{ij}) \pm 1.96\sigma_j.$$

Since $\min_j(Y_{ij}) \le Y_{ij^*}$, the coverage probability of this CI will be smaller than 95%. In particular, as $\mu_{j'}$ becomes closer to the others or the error variances become larger, the coverage probability will become worse because they cause $j' \neq j^*$ more often.

4.4.5 The Profile Likelihood CI vs. the Wald-Type CI

In this section, we compare the profile likelihood confidence interval with the Wald-type confidence interval both numerically and analytically. It turns out that the Wald-type is much more robust against the tie issue than the profile likelihood approach while the both performs similarly when we focus on only the lower bound. The advantage of the profile likelihood approach is that it can construct a confidence interval for any point while using the Wald-type approach limits to a point within the sample.

To see the structure more clearly, we simplify the problem. We generate a sample from

$$Y_j = \mu_j + \varepsilon_j,$$

where $q = 15$, $\mu_j \in \mathbb{R}^q$ and $\varepsilon_j \sim N(0, 1)$. In particular, μ_j is determined as follows:

- $\ell = 3, 6, 12$
- For $j = 1, \ldots, q - \ell + 1$, generate μ_j independently from a standard normal distribution.
- For $j = q - \ell + 2, \ldots, q$, set $\mu_j = \min\{\mu_1, \ldots, \mu_{q-\ell+1}\} + \alpha Z_j$ where Z_j is generated independently from standard normal.

α controls the proximity of the ties around the minimum. ℓ is the number of the means which coincide with the minimum or are near it, thus it is roughly the number of ties. As α moves from 2 to 0, $\ell - 1$ means become closer to the minimum, i.e., there are ℓ means including the minimum that are close to each other. This setting simplifies the original model (4.7) as if we know the true β_j. Repeating 10000 times, we plot the coverage probability and the coverage width for the observed points. The results are given in Figure 4.7. As can be seen in the left figures, the closer the means around the minimum become to one another (α approaches to 0), the coverage probabilities of both the profile likelihood CI and the Wald-type CI deteriorate. This is even exacerbated as the number of ties, ℓ, increases. Note that the profile likelihood CI deteriorates much worse than the Wald-type in any case. This can be explained by the coverage width, which is given in the figures to the right; the width of the Wald-type CI remains the same while that of the profile likelihood CI becomes narrower as α becomes closer to 0. This shrinking width makes the deteriorating coverage probability even worse.

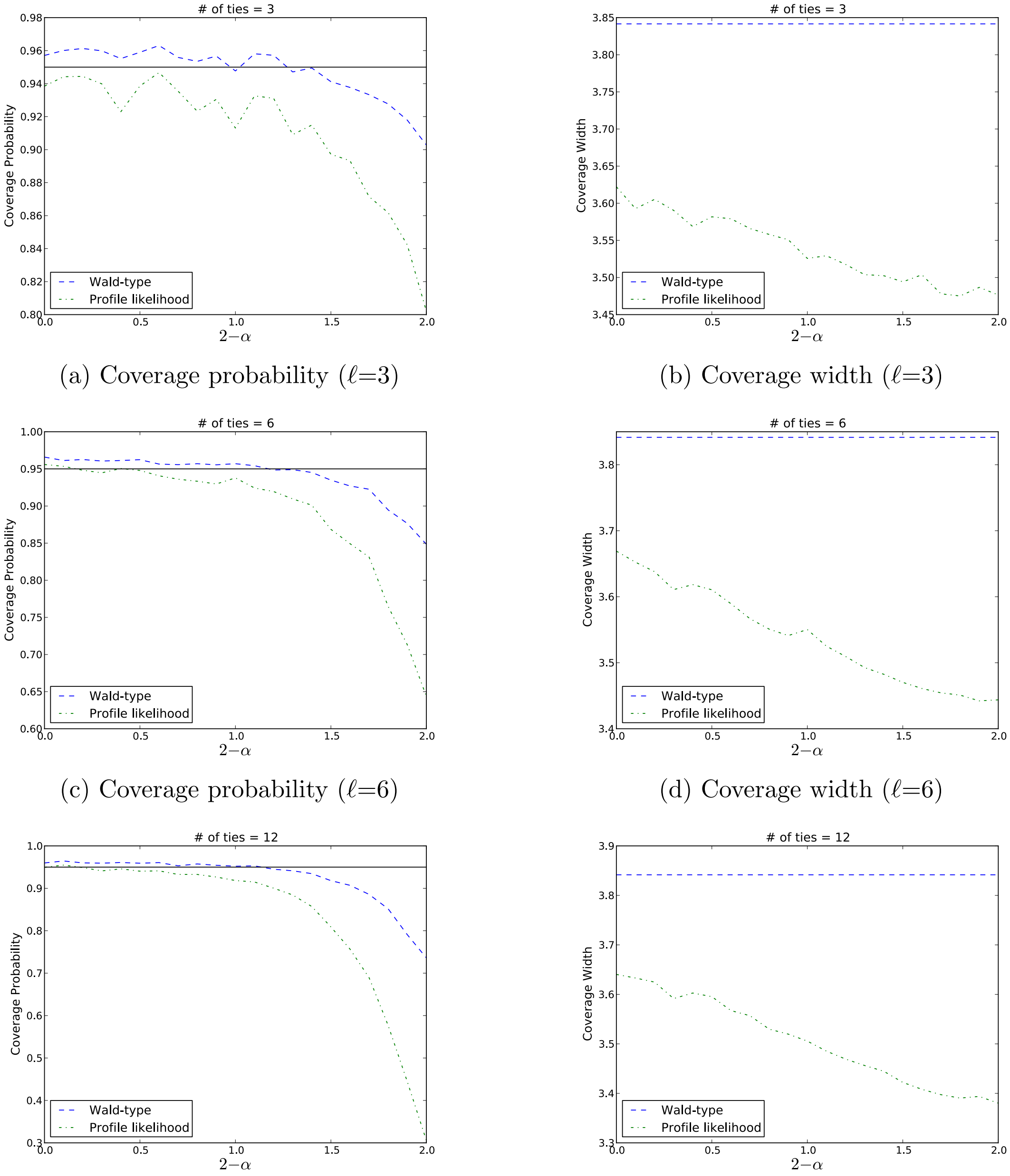
We explain analytically why these happen. Recall that the Wald-type CI for μ_j is given by $\min_j Y_j \pm 1.96$, which has the constant width regardless of $j^* = \underset{j}{\operatorname{argmax}} Y_j$. The reduction of the coverage probability is due to the bias caused by the events $j^* \neq j' := \underset{j}{\operatorname{argmin}} \mu_j$, which occurs more frequently as α approaches to 0 and/or ℓ increases. Now, the profile likelihood function is given by

$$g(u) = \max_{\mu \in \mathcal{C}(u)} \Big\{ -\frac{1}{2} \sum_j (Y_j - \mu_j)^2 \Big\},$$

where $\mathcal{C}(u) = \{\mu : \mu_j \geq u \text{ for all } j \wedge \mu_j = u \text{ for at least one } j\}$. The upper and lower bounds (μ_*, μ^*) are given by those that satisfy $g(\mu_*) = g(\mu^*) = -\frac{(1.96)^2}{2}$ and $\mu_* < \mu^*$. It is not difficult to see that

$$\mu_* = \min_j Y_j - 1.96$$
$$\mu^* \leq \min_j Y_j + 1.96,$$

where the equality happens only when $\min_j Y_j + 1.96 \leq Y_{[2]}$, which is the second smallest number. The discrepancy becomes larger as more μ_j get closer to the minimum. Therefore, while the lower bounds of the profile likelihood CI and the Wald-type CI are similar, the upper bound of the profile likelihood CI tends to be smaller than that



(a) Coverage probability (ℓ=3)

(b) Coverage width (ℓ=3)

(c) Coverage probability (ℓ=6)

(d) Coverage width (ℓ=6)

(e) Coverage probability (ℓ=12)

(f) Coverage width (ℓ=12)

Figure 4.7: Coverage probability (Left) and coverage width (Right) by the profile likelihood CI and the Wald-type CI. The horizontal lines in the left figures indicate the target 0.95 line. α controls the proximity of the means around the minimum. ($\alpha = 0$ means complete tie.)

of the Wald-type CI, which makes the overall coverage probability by the profile likelihood CI worse than that of the Wald-type CI. Fortunately, when the lower bound is only of interest (which is the case in the toxicity analysis), the both works similarly.

4.4.6 Summary

In this section, we considered the interval estimation for the composite risk. The idea of directly specifying a model for the composite risk through the profile likelihood function can be naturally extended to construct a confidence interval. We introduced the algorithm to compute the interval by using constrained quadratic programming. As is often the case in the toxicity analysis, we also considered the censored-response model. Computationally, the censoring issue can be handled by assuming a censored likelihood and using more general constrained optimization; constrained optimization by linear approximation can be a good candidate to use. We also introduced an alternative approach, namely the Wald-type approach. The Wald-type CI is limited to a point within the sample while the profile likelihood approach can be used for any point. We showed that the Wald-type CI is more robust against ties at the minimum in terms of coverage probability both numerically and analytically . However, when interest lies solely on the lower bound, they both work similarly.

4.5 Data Analysis

In this section, we apply the procedures to estimate the composite risk functions to the ToxCast data of the EPA and to the 60 cell line screen of the NCI.

4.5.1 Application to the ToxCast Data

The ToxCast data is the publicly available HTS assay data from the EPA (http://www.epa.gov/ncct/toxcast/). The chemical library of the ToxCast contains 320 chemical compounds, which are explored in 25 cell-based HTS assays and 484 cell-free HTS assays. The outcome of the assays is measured by the lowest effective level, i.e., the lowest level of the dose where there is a statistically significant change from the control. The assays are measured in log10 micromolar (mM) unit. Since many compounds cause no significant change in many assays within the range of doses, we removed the assays whose outcomes are missing for at least 10% of the compounds. This reduces the number of the cell-based assays to 15 and the number of the cell-free assays to 78. Also, because of the discreteness of the dose points many compounds show similar outcomes in the cell-based assays. Thus we removed the cell-based assays whose number of the unique outcomes is less than 100. This further reduces the number of the cell-based assays to 3, which are `CLM_CellLoss_72hr`, `CLM_MitoticArrest_72hr`, and `ACEA_IC50`. Following Judson et al. (2010), the outcomes of the cell-free assays are converted to binary response according to whether it displays a significant change below a certain amount of dose (30mM). We reduce the dimensionality of the cell-free assays through principal component (PC) decomposition and use 9 leading PC scores as explanatory variables, which is the smallest number that explains at least 50% of the total variation in the cell-free assays. The PC scores are standardized to have unit variances. Overall, the data used in this analysis consists of the 9 leading PC scores of the cell-free assays, which are used as explanatory variable, and 3 cell-based assays, which are used as dependent variables. The proportion of the compounds that display the minimum per each cell-based assay is respectively .1625, .13125, and .70625.

The lowest effective levels may be upper-censored as a cell line may not display significant change within the range of dose concentrations. Let $t_j, j = 1, \ldots, q$ be the thresholds and F_j be the cumulative distribution function of ε_j. The censored log-likelihood is given by

$$\ell_j^*(y_{ij}|\mu_j(x_i)) = 1\{y_{ij} > t_j\} \log F_j(\mu_j(x_i) - t_j) + 1\{t_j \leq y_{ij}\}\ell_j(y_{ij}|\mu_j(x_i))$$

Then, the profile likelihood function (4.3) becomes

$$\mathcal{G}(\beta) = \sum_{i=1}^{n} \Big[1\{y_{ij} \leq \beta^T x_i, y_{ij} < t_j \text{ for some } j\} \sum_{j=1}^{q} 1\{y_{ij} < t_j\} \ell_j(y_{ij}|y_{ij} \vee \beta^T x_i)$$

$$+ 1\{\text{otherwise}\} \max_{1 \leq j \leq q} \{ 1\{y_{ij} < t_j\} \ell_j(y_{ij}|\beta^T x_i) + 1\{y_{ij} \geq t_j\} \log F_j(\beta^T x_i - t_j)$$

$$+ \sum_{j' \neq j} 1\{y_{ij'} \geq t_{j'}\} \ell_{j'}(y_{ij'}|y_{ij'}) \} \Big].$$

The maximum likelihood estimator of β is given by $\widehat{\beta} = \underset{\beta}{\operatorname{argmax}}\, \mathcal{G}(\beta)$. As the variances are unknown, we substitute the estimates obtained by fitting the Tobit model to each cell-based assay; the cross-validated prediction r^2 is given by $0.60, 0.29, 0.34$ respectively. We compare this approach to the other two linear-regression based approaches. Figure 4.8 shows the regression coefficients of the Tobit model for each cell-based assay and those of the profile likelihood approach. Only the intercept (not shown) and the coefficient of the first PC score are significant at 5% significance level. The Wald statistic testing the zero coefficients show significance at 5% level for all the three assays. Note that all the three coefficients for the cell-based assays are similar; this is a favorable condition for the profile likelihood approach as we saw in section 4.3.1. Figure 4.9 shows the norm-standardized regression coefficients, where the weighted average of the three regression coefficients are shown for the min(OLS) approach with the proportions of the compounds whose minimum occurs at each assay used as weights. This figure also shows the similarity of all the coefficients. On the other hand, the number of the dependent variables is just three, which may not encourage using the profile likelihood approach. We computed the cross validation (CV) score using the observed minimum, i.e.,

$$CV = \frac{1}{n} \sum_{i=1}^{n} \{ \min_j Y_{ij} - \widehat{Y}_j^{(-i)} \},$$

where $\widehat{Y}_j^{(-i)}$ is the estimate given by each of the three methods. The three CV scores

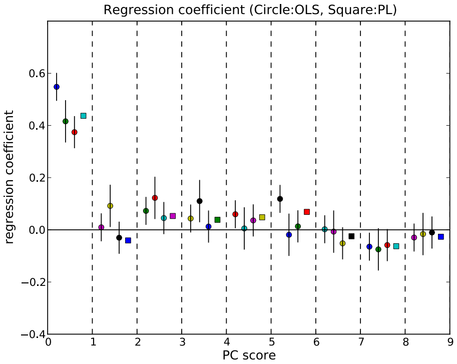


Figure 4.8: Regression coefficients of the Tobit model for each cell-based assay and those of the profile likelihood approach. The estimated intercepts are respectively 2.19, 2.53, 2.27, and 1.99 (not shown). The vertical lines show the 95% confidence intervals.

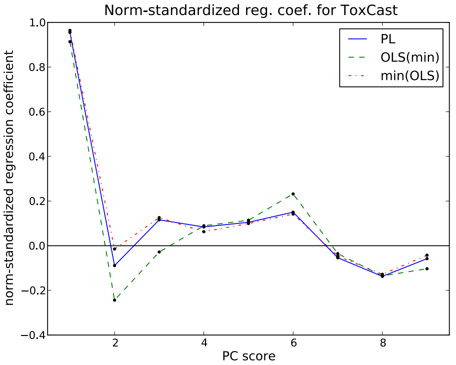


Figure 4.9: The norm-standardized regression coefficients for the three approaches. For the min(OLS) approach, the weighted average of the three regression coefficients are shown, where the proportions of the compounds whose minimum occur per assay are used as weights.

are respectively 0.65, 0.42, and 0.33. Note that $\min_j Y_{ij}$ is not an unbiased proxy for the object of interest that we aim to estimate—$\min_j \mathbb{E}(Y_{ij}|X_i)$. Still, we presume that the CV provides some insights as to how profile likelihood approach works.

4.5.2 Application to the NCI60 Cell Line Screen Data

In this section, we apply the methods to the 60 cell line screen of the NCI. The data consists of 60 different human cell lines, each representing one of nine disease/cancer types: leukemia, melanoma and cancers of the lung, colon, brain, ovary, breast,

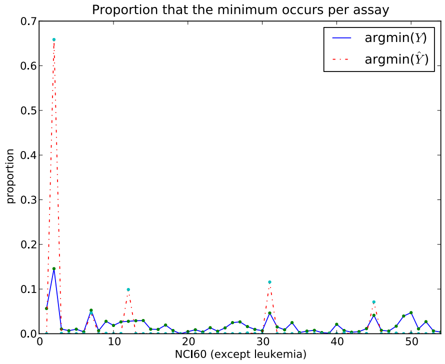


Figure 4.10: The proportion of the compounds whose minimum occurs per cell line in terms of the raw observations and the OLS estimates.

prostate, and kidney. The assay outcome is given by GI50, the drug concentration resulting in 50% inhibition of the growth relative to control cells. The measurement unit is log10 micro Molar. See Shoemaker (2006) for the details.

We first explore the estimation of the composite risk represented by the cell line assays from the molecular features of the compounds. We use 10 molecular features such as diameter, mass, or charge as explanatory variables. The original NCI60 data we obtained contains 37327 compounds screened for 60 cell lines (the data is publicly available at the NCI website). We removed six assays that are related to leukemia as the average level of the leukemia assays are lower than the others so that most of the times the minimum occurs at one of the leukemia assays (the other assays contribute little to the evaluation of the overall risk). Thus, dependent variables comprise 54 cell line assays. We further removed the compounds that contain missing values and the censored values. This reduces the data to 2376 compounds. The mean of cross-validated prediction r^2 of the OLS regression for the 54 cell line assays is 0.15 and the standard deviation is 0.028. Figure 4.10 shows the proportions of the compounds whose minimum occur at each cell line assay in terms of the raw observations and the OLS estimates. There are four assays where the estimated minimum often occurs, which are related to: non-small cell of the lung cancer, the colon cancer, melanoma, and the breast cancer. Among the assays that represent each type of disease, there

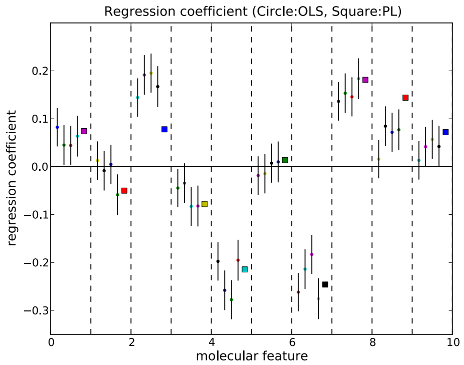


Figure 4.11: The OLS regression coefficients for the four assays regarding non-small cell of a lung, colon, melanoma, and breast, as well as the profile likelihood estimates. The vertical lines show the 95% confidence intervals.

seem to be a typical assay for each disease where the minimum occurs for the OLS estimates. Figure 4.11 shows the OLS regression coefficients for these four assays and the profile likelihood estimates. The four OLS regression coefficients are fairly similar, which is a favorable condition for the profile likelihood approach to use. Fairly high dimensionality of the dependent variables is also favorable to the profile likelihood approach. Despite these conditions, though, very low predictive r^2 may cast some doubt on the legitimacy in extracting a firm conclusion from the analysis. Figure 4.12 shows the norm-standardized regression coefficients for the three approaches. The three coefficients are, again, fairly similar though there are a couple of variables where the coefficients are different. For example, the third variable (b_rotN) is considered important by the two OLS-based methods while the ninth variable (density) is considered relatively important by the profile likelihood approach. If there is prior information that suggests that either of them is likely to be the case, we may be able to diagnose which method is working better. The cross-validation scores using the observed minimum as explored in the ToxCast data analysis for the three approaches are respectively given by 1.53, 1.27, and 1.71. In contrast to the ToxCast data analysis, the third approach now becomes worst.

Next, we explore the use of leukemia assays to estimate the composite risk rep-

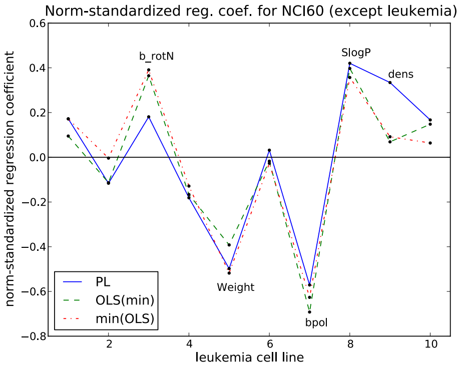


Figure 4.12: The norm-standardized regression coefficients for the three approaches. For the min(OLS) approach, the weighted average of the three regression coefficients are shown, where the proportions of the compounds whose minimum occur per assay are used as weights.

resented by the other cell line assays—the same dependent variables as before, but different explanatory variables. This setting aims to investigate whether the use of only a fraction of cell line assays possess a predictive power for the overall toxicity level. Similarly to the previous setting, we removed all the compounds that contain missing values. This reduces the number of compounds to 2066. The mean of the cross-validated prediction r^2 by OLS regression is 0.76 and the standard deviation is 0.065. This high r^2 makes sense as the cell line assays should roughly behave similarly against toxic compounds, though there are some selective compounds that are toxic to some types of cells while innocuous to others. Figure 4.13 shows the proportions of the compounds whose minimum occur at each cell line assay in terms of the raw observations and the OLS estimates. The proportions for the raw observations are exactly same as the previous setting. The proportions for the estimates are also quite similar to the previous analysis, but there is one difference; the minimum occurs more often in the two of the assays for the breast cancer (the two peaks near 50 on the horizontal axis). The four disease assays where the minimum occurs most often are related to: non-small cell of the lung cancer, the colon cancer, melanoma, and the breast cancer. Figure 4.14 shows the OLS regression coefficients for these four as-

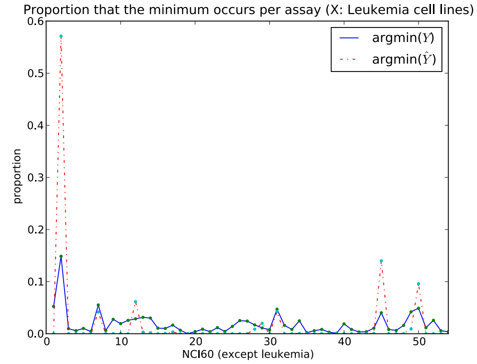


Figure 4.13: The proportion of the compounds whose minimum occurs per cell line in terms of the raw observations and the OLS estimates.

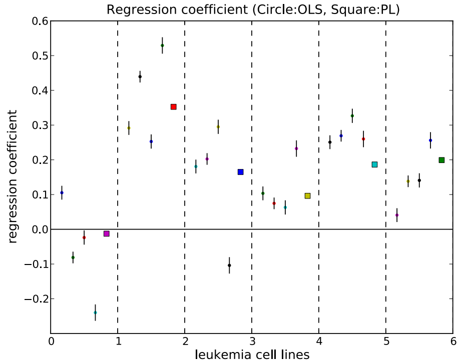


Figure 4.14: The OLS regression coefficients for the four assays regarding non-small cell of a lung, colon, melanoma, and breast, as well as the profile likelihood estimates. The vertical lines show the 95% confidence intervals.

says and the profile likelihood estimates. Now, the four OLS regression coefficients are not as similar as the previous two analyses. However, Figure 4.15, the norm-standardized regression coefficients for the three approaches, shows similarity among the three approaches except the last two assays. The cross-validation scores are 0.27, 0.26, and 0.46. The profile likelihood approach looks relatively good compared to the previous two analyses, though we cannot make a firm conclusion as the score is computed for the observed minimum, which is biased for the target—the minimum of the conditional regression means.

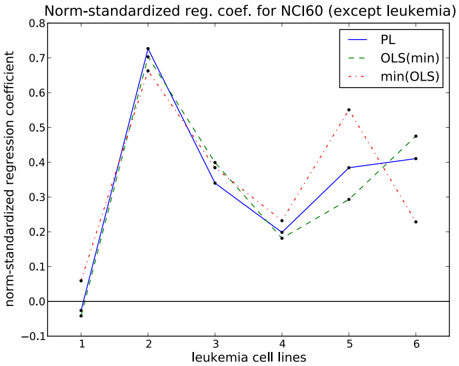


Figure 4.15: The norm-standardized regression coefficients for the three approaches. For the min(OLS) approach, the weighted average of the three regression coefficients are used, where the proportions of the compounds whose minimum occur per assay are used as weights.

4.6 Discussion

In this paper, we considered the estimation of the composite risk function, the minimum of the assay-specific risk functions. To avoid the necessity of estimating all the assay-specific functions, we introduced a method of directly specifying a model for the composite risk function through the profile likelihood function. Under this approach, we only need to specify one target parameter function regardless of the dimensionality of the explanatory variables, hence it also avoids the operation of computing the minimum.

Through several simulation studies, we explored under what situations it is beneficial to use the profile likelihood approach rather than naive approaches. First, the performance of the profile likelihood approach greatly depends on whether we can find a good model for the composite risk; as long as the specified model approximates the true model relatively well, the profile likelihood approach can perform better than the correctly specified two-step approach. Second, the high dimensionality of the dependent variables works simply as a large sample size for the profile likelihood approach. On the other hand, for the other naive approaches the high dimensionality requires a large number of parameters to estimate and the operation of computing the minimum among a large number of elements, hence the performance deteriorates. Third, the smoothness of the assay-specific risks deeply influences the performance of the two-step approaches while the profile likelihood approach is robust against this issue. For the two-step approach to perform well, all the assay-specific risks need to be easily estimated. Finally, the heteroscedasticity of the error variances has a negative effect on the procedures except the profile likelihood approach, on which the heteroscedasticity shows a positive effect. This may be because the profile likelihood approach incorporates the variance information in its procedure while the other approaches do not take it into account in their procedures.

We extended the idea of directly specifying a model for the composite risk through the profile likelihood function to the interval estimation. An alternative approach, the Wald-type approach, is also introduced. It was shown that the Wald-type CI is more robust against the ties in the assay-specific risks, but when the goal is to estimate the minimum, which is actually the case in the toxicity analysis, the both approaches work similarly. The advantage of the profile likelihood approach is that it can construct an interval for any point while the Wald-type approach limits to a point within the sample, hence it cannot be used for prediction.

We demonstrated the estimation of the composite risk function using the ToxCast data and the NCI60 cell line screen data. Although the inclusion of the profile likelihood approach provides another option for the analysis, there is no suitable way to diagnose the model fitting by the profile likelihood approach. This is mainly because there is no unbiased proxy for the composite risk unlike usual regression problems where the response is unbiased for the regression function, hence it can be used for the model diagnosis, for example, by checking the cross-validation score. We leave the model diagnosis issue for the profile likelihood approach as a future research problem.

BIBLIOGRAPHY

BIBLIOGRAPHY

R. B. Ash and M. F. Gardner. *Topics in stochastic processes.* Academic Press, New York, 1975.

M. Benko, W. Härdle, and A. Kneip. Common functional principal components. *Ann. Statist.*, 37(1):1–34, 2009.

T. T. Cai and P. Hall. Prediction in functional linear regression. *Ann. Statist.*, 34(5): 2159–2179, 2006.

H. Cardot, F. Ferraty, and P. Sarda. Functional linear model. *Statist. Probab. Lett.*, 45(1):11–22, 1999.

H. Cardot, F. Ferraty, and P. Sarda. Spline estimators for the functional linear model. *Statist. Sinica*, 13(3):571–591, 2003.

J.-M. Chiou and P.-L. Li. Functional clustering and identifying substructures of longitudinal data. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 69(4):679–699, 2007.

C. Crambes, A. Kneip, and P. Sarda. Smoothing splines estimators for functional linear regression. *Ann. Statist.*, 37(1):35–72, 2009.

W. S. DeSarbo and W. L. Cron. A maximum likelihood methodology for clusterwise linear regression. *J. Classification*, 5(2):249–282, 1988.

D. J. Dix, K. A. Houck, M. T. Martin, A. M. Richard, R. W. Setzer, and R. J. Kavlock. The toxcast program for prioritizing toxicity testing of environmental chemicals. *Toxicological Sciences*, 95(1):5–12, 2007.

F. Ferraty and P. Vieu. Curves discrimination: a nonparametric functional approach. *Comput. Statist. Data Anal.*, 44(1-2):161–173, 2003.

F. Ferraty and P. Vieu. *Nonparametric functional data analysis.* Springer Series in Statistics. Springer, New York, 2006.

F. Ferraty, A. Mas, and P. Vieu. Nonparametric regression on functional data: inference and practical aspects. *Aust. N. Z. J. Stat.*, 49(3):267–286, 2007.

C. Fraley and A. E. Raftery. Model-based clustering, discriminant analysis, and density estimation. *J. Amer. Statist. Assoc.*, 97(458):611–631, 2002.

J. H. Friedman. Multivariate adaptive regression splines (with discussion). *Ann. Statist.*, 19(1):1–141, 1991.

C. R. Genovese and L. Wasserman. Rates of convergence for the Gaussian mixture sieve. *Ann. Statist.*, 28(4):1105–1127, 2000.

S. Ghosal and A. W. van der Vaart. Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. *Ann. Statist.*, 29 (5):1233–1263, 2001.

P. Hall. On convergence rates in nonparametric problems. *Internat. Statist. Rev.*, 57 (1):45–58, 1989.

P. Hall and J. L. Horowitz. Methodology and convergence rates for functional linear regression. *Ann. Statist.*, 35(1):70–91, 2007.

P. Hall, H.-G. Müller, and J.-L. Wang. Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.*, 34(3):1493–1517, 2006.

P. Hall, H.-G. Müller, and F. Yao. Estimation of functional derivatives. *Ann. Statist.*, 37(6A):3307–3329, 2009.

G. M. James. Generalized linear models with functional predictors. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):411–432, 2002.

G. M. James and B. W. Silverman. Functional adaptive model estimation. *J. Amer. Statist. Assoc.*, 100(470):565–576, 2005.

G. M. James, T. J. Hastie, and C. A. Sugar. Principal component models for sparse functional data. *Biometrika*, 87(3):587–602, 2000.

R. S. Judson, K. A. Houck, R. J. Kavlock, T. B. Knudsen, M. T. Martin, H. M. Mortensen, D. M. Reif, D. M. Rotroff, I. Shah, A. M. Richard, and D. J. Dix. *In vitro* screening of environmental chemicals for targeted testing prioritization: The toxcast project. *Environ. Health Perspect.*, 118(4):485–492, 4 2010.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhyā Ser. A*, 62(1):49–66, 2000.

G. McLachlan and D. Peel. *Finite mixture models.* Wiley-Interscience, New York, 2000.

H.-G. Müller and U. Stadtmüller. Generalized functional linear models. *Ann. Statist.*, 33(2):774–805, 2005.

P. A. Naik, P. Shi, and C.-L. Tsai. Extending the Akaike information criterion to mixture regression models. *J. Amer. Statist. Assoc.*, 102(477):244–254, 2007.

J. Nocedal and S. J. Wright. *Numerical optimization.* Springer Series in Operations Research and Financial Engineering. Springer, New York, second edition, 2006.

M. J. D. Powell. A direct search optimization method that models the objective and constraint functions by linear interpolation. In *Advances in optimization and numerical analysis (Oaxaca, 1992)*, volume 275 of *Math. Appl.*, pages 51–67. Kluwer Acad. Publ., Dordrecht, 1994.

J. O. Ramsay and C. J. Dalzell. Some tools for functional data analysis. *J. Roy. Statist. Soc. Ser. B*, 53(3):539–572, 1991. With discussion and a reply by the authors.

J. O. Ramsay and B. W. Silverman. *Applied functional data analysis.* Springer Series in Statistics. Springer-Verlag, New York, 2002.

J. O. Ramsay and B. W. Silverman. *Functional data analysis.* Springer Series in Statistics. Springer, New York, second edition, 2005.

J. A. Rice and B. W. Silverman. Estimating the mean and covariance structure nonparametrically when the data are curves. *J. Roy. Statist. Soc. Ser. B*, 53(1): 233–243, 1991.

R. Shoemaker. The NCI60 human tumour cell line anticancer drug screen. *Nature Rev.*, 6:813–823, October 2006.

B. W. Silverman. Smoothed functional principal components analysis by choice of norm. *Ann. Statist.*, 24(1):1–24, 1996.

P. Speckman. Spline smoothing and optimal rates of convergence in nonparametric regression models. *Ann. Statist.*, 13(3):970–983, 1985.

R. Tang and H.-G. Müller. Pairwise curve synchronization for functional data. *Biometrika*, 95(4):875–889, 2008.

J. Tobin. Estimation of relationships for limited dependent variables. *Econometrica*, 26:24–36, 1958.

R. D. Tuddenham and M. M. Snyder. Physical growth of california boys and girls from birth to age 18. *Calif. Publ. Child Develop.*, 1:183–364, 1954.

G. Wahba. *Spline models for observational data.* Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 1990.

F. Yao, H.-G. Müller, and J.-L. Wang. Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.*, 100(470):577–590, 2005.

F. Yao, Y. Fu, and T. C. M. Lee. Functional mixture regression. *Biostat.*, 12(2): 341–353, 2011.

M. Yuan and T. T. Cai. A reproducing kernel hilbert space approach to functional linear regression. *Ann. Statist.*, 38(6):3412–3444, 2010.

J.-H. Zhang, T. D. Y. Chung, and K. R. Oldenburg. A simple statistical parameter for use in evaluation and validation of high throughput screening assays. *J Biomol Screen*, 4(2):67–73, April 1999.

X. D. Zhang. *Optimal High-Throughput Screening.* Cambridge University Press, Pennsylvania, 2011.