ASPECTS OF PENALIZED SPLINES

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Yingxing Li

January 2011

ASPECTS OF PENALIZED SPLINES

Yingxing Li, Ph.D.

Cornell University 2011

Penalized splines approach has very important applications in statistics. The idea is to fit the unknown regression mean using a high-dimensional spline basis, subject to penalization on the roughness. Such an approach avoids the stringency of a parametric model and enables a considerable reduction in computational cost without a loss of statistical precision. Moreover, the idea can also be connected with ridge regression and mixed models, thus allowing more flexible handling of longitudinal and spatial correlation.

This thesis focuses on nonparametric and semiparametric estimation and inference using penalized splines. First, we consider the penalized splines approach proposed by Eilers and Marx (1996), which is also called P-splines approach. We derive its asymptotic property when the number of spline basis increases as the sample size does. For both the univariate model and the additive models, we establish the asymptotic distribution of the estimators and give simple expressions for the asymptotic mean and variance. Such an asymptotic theory allows P-splines estimators to be compared theoretically with other nonparametric estimators and offers guidance for practitioners when considering the choice of the penalty and basis functions.

Next, we turn to the global inferential problems for functional data. We model the population mean function using polynomial splines. By utilizing the mixed model based penalized splines approach, we treat some of the spline coefficients as random effects with a single variance component and relate hy-

potheses of interest to tests with this variance component being zero. To take into account the dependent structure or within subject correlation, we propose a pseudo likelihood test statistic and derive its null distribution. This work extends existing results on pseudo likelihood by allowing the use of nonparametric smoothing (usually with a slower convergence rate). Its effectiveness is demonstrated via simulations and the empirical application from the Sleep Health Heart Study.

BIOGRAPHICAL SKETCH

Yingxing Li was born in Guangzhou, China. In 2003, she received her B.Sc. degree in Mathematics from Nanjing University, China. After that, she entered the Statistics and Actuarial Science department at the University of Hong Kong, where she received her M.Phil. degree in 2005. She then continued her study at Cornell University and earned her Ph.D. degree in 2011.

This document is dedicated to my parents.

ACKNOWLEDGEMENTS

I wish to express my sincere gratitude to my advisor, Prof. David Ruppert, for his guidance, support and inspiration. I am very lucky to have learnt so much from him and I deeply appreciate his help. This thesis would not have been completed without him.

I thank Prof. Giles Hooker, for his discussions and encouragement as well as serving on my committee. I thank Prof. James Booth for his advices and serving on my committee. I thank Prof. Martin T. Wells for his suggestions and serving in my B-exam. I thank Prof. Ciprian M. Crainiceanu at Johns Hopkins University for providing me the Sleep Heart Health Study data and support. I thank Prof. Christine A. Shoemaker for her support. I thank Prof. Yongmiao Hong for his comments. I thank Prof. Bruce W. Turnbull for providing advices. I thank Prof. J. T. Gene Hwang for discussing researches with me. I thank Prof. Robert Strawderman for his help.

Especially, I would like to thank Prof. Tatiyana V. Apanasovich at Thomas Jefferson University, Prof. Ana-Maria Staicu at North Carolina State University, and Luo Xiao for sharing their ideas with me, which helps improve this thesis a lot.

Special thanks go to my friends at Cornell. I greatly appreciate Haiqiang Chen for his help and patience. I would also like to thank my parents for their unconditional love and always being there for me through ups and downs.

I acknowledge the financial support from NIH grant R01NS060910.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

Regression analysis is one of the most commonly used techniques in statistics, whose aim is to explore the relationship between dependent and independent variables. The most classical and simple form is linear regression, where one tries to fit a line through the given pairs (x_i, y_i), $i = 1, \ldots, n$. Although the linear regression technique is very useful, the assumption of linearity might not always be granted. Alternatively, one could consider nonlinear models. However, all these approaches are parametric, where the functional form is known. They are subjected to the risk of potential model mis-specification that could lead to inconsistent estimations of the regression coefficients.

To repair this drawback, nonparametric approaches are considered. In the univariate case, the model is written as $y_i = \mu(x_i) + \epsilon_i$, where conditionally on x_i, ϵ_i has mean zero and variance $\sigma^2(x_i)$, and the function form of $\mu(x)$ is unknown. Different methods are used to estimate the unknown function $\mu(x)$. One idea is to consider the data points that are near x. Intuitively, data points that are far from x should contain little information about the value of $\mu(x)$. Hence one can estimate the mean function by running local average. This idea can be further improved by incorporating kernel weights, which leads to the Nadaraya-Watson kernel estimator, see Nadaraja (1964) and Watson (1964),

$$\hat{\mu}(x) = \frac{\sum_{i=1}^{n} H_h(x_i - x) y_i}{\sum_{i=1}^{n} H_h(x_i - x)},$$

where $H_h(x) = H(x/h)$ is determined by some symmetric function $H(x)$ and some positive scalar h. The function $H(x)$ is called the kernel function and h is the bandwidth. By controlling the size of the local neighborhood, h plays the role of smoothing and balances between model flexibility and variability.

The asymptotic property of Nadaraya-Watson estimator is discussed in Härdle (1990) and its asymptotic bias depends on the design density, i.e. the density of x_i's. The choice of $H(x)$ does not affect the convergence rate of the kernel estimation, but the choice of h does.

Improvement can be made by approximating $\mu(x)$ by a locally weighted linear regression instead of a local constant. To obtain an estimate of $\mu(x)$ at a given point x, one estimate $\hat{\alpha}_x$ and $\hat{\beta}_x$ by minimizing

$$\sum_{i=1}^{n}\{y_i - \alpha_x + \beta_x(x_i - x)\}^2 H_h(x_i - x).$$

With the extra parameter β_x, the local linear fit is able to reduce the asymptotic bias without increasing the asymptotic variance. Moreover, it has the automatic boundary correction such that the estimation at the boundary has the same convergence rate as that in the interior.

Rather than fitting the mean function locally, one can consider a global approximation of the regression mean. For example, one can increase the number of parameters in polynomial regression. However, this approach is not very flexible and has two drawbacks. First, the observation that is far from x can still have a large influence on the estimated $\hat{\mu}(x)$. Second, the fitted curve is extremely smooth and possesses all orders of derivatives everywhere.

Alternatively, one might model $\mu(x)$ by the pth degree splines, whose pth derivatives have discontinuity at certain locations (knots). There are two popular choices of spline bases. One is B-splines family by de Boor (1978) and the other is the polynomial splines family. Take $p = 1$ as an example. The linear B-splines bases are formed by piecewise linear functions, usually defined as,

$$B_k^{[1]}(x) \;=\; K(x - \kappa_{k-1})I_{\kappa_{k-1}\le x\le\kappa_k} + K(\kappa_{k+1} - x)I_{\kappa_k\le x\le\kappa_{k+1}},\ k = 1, \ldots, K + p$$

where $\kappa_k = k/K$ for all k are equally spaced knots that record all jump locations of the first derivative. In contrast, the linear polynomial splines bases are defined as :

$$1, x, (x-\kappa_1)_+, (x-\kappa_2)_+, \ldots, (x-\kappa_K)_+,$$

where $(x)_+ = \max(x, 0)$ captures the departure from the linear function. We can simply transform these two families from one to the other. Generally speaking, B-splines are more numerical stable, while polynomial splines provide a more natural connection with standard polynomial regression. In both cases, the regression mean function $\mu(x)$ can be expressed as a linear combination of all spline bases once the knots are determined. The target is to estimate the spline coefficients.

There are three approaches to spline fitting. One is to estimate all spline coefficients via ordinary least squares criterion. This is the idea of regression splines. In order to maintain the modelling flexibility, the number of spline bases should increase as the sample size does. In practice, the choice of the number of knots as well as their locations are important. Sophisticated algorithms on the selection as well as placement of knots have been proposed; see Turbo algorithm by Friedman and Silverman (1989), and MARS algorithm by Friedman (1991). To avoid the difficulty of knot selection, one might consider adopting sufficiently many knots and introduce a penalty term to avoid overfitting. In the extreme case when each datum point is viewed as a knot, this is equivalent to another fitting approach, smoothing splines (Wahba, 1990). The penalty is imposed on the curvature and the approach minimizes

$$\sum_{i=1}^{n} \{y_i - \mu(x_i)\}^2 + \lambda \int \{\mu^{(2)}(t)\}^2 dt.$$

Without the penalty term, the naive least square criterion will interpolate the data, thus resulting in a wiggly curve with lots of variation. By incorporating the penalty on the roughness, we can trade off between the bias and variance. The parameter λ serves as the smoothing parameter and plays the most important role. It can be chosen objectively by data-driven approaches. For example, one can consider minimizing the cross validation criterion,

$$CV(\lambda) = n^{-1} \sum_{i=1}^{n} \{y_i - \hat{\mu}_\lambda, i(x_i)\},$$

where $\hat{\mu}_{\lambda,i}(x_i)$ is the estimator without using the ith observation. Silverman (1984) shows that smoothing splines approach is asymptotically equivalent to the kernel approach.

The third fitting approach is Penalized splines, which estimate the unknown regression function by least squares plus a penalty term imposed on the spline coefficients. This idea can be traced back at least to O'Sullivan (1986), but they became popular after the seminal paper of Eilers and Marx (1996) and, later, with the book by Ruppert et al. (2003). Penalized spline approach can be viewed as a compromise between regression splines and smoothing splines. Either the number of the knots or the penalty term can play the role of smoothing. In practice, it is often suggested to choose a sufficiently large number of knots and let the penalty term behaves as the key smoother to offset the risk of overfitting. Moreover, the number of knots can be chosen much less than the sample size without any sacrifice in the accuracy. The Penalized splines approach can hence gain more computation efficiency over smoothing splines.

However, unlike smoothing splines, little was known about the asymptotic property of penalized splines. This motivated our theoretical study on the P-spline estimation proposed by Eilers and Marx (1996). Such an asymptotic the-

ory allows P-splines estimators to be compared theoretically with other nonparametric estimators. In practice, one needs to specify the choice of the number of knots, the order of penalty and the spline basis. Our study will examine the impact of these subjective choices on the estimated regression function, and advocate further improvements over existing approach. All these issues will be address in the Chapter 2, where we focus on the univariate regression model using P-splines. Some of the results are also discussed in the papers by Li and Ruppert (2008), and Apanasovich et al. (2010).

The techniques in univariate nonparametric regression can be similarly extended for the case when multiple predictors are present. However, there is a high price to pay as the dimension of the covariates increases. This effect is known as the curse of dimensionality by Bellman (1961), which refers to the inherent sparsity of high-dimensional data. For instance, to maintain 10 data along each of the q axes, 10^q data are required. Such an exponential growth decreases the precision of nonparametric estimation and the lack of sufficient data drags down the application of multivariate nonparametric techniques.

Several dimension reduction techniques have hence been proposed to overcome this caveat. One popular way is to project all covariates onto a linear space and then fit a nonparametric curve to the linear combination. This leads to the single-index model by Ichimura (1993),

$$y_i = \mu(\alpha^T x_i) + \epsilon_i,$$

where $x_i = (x_{i1}, \ldots, x_{iq})$ and ϵ_i is zero mean. For identifiability, the norm of the projection vector satisfies $||\alpha|| = 1$. Härdle et al. (1993) suggest estimating the function $\mu(\cdot|\alpha)$ by a Nadaraya-Watson type estimator and they use cross-validation techniques to simultaneously select the bandwidth and α.

However, this approach might not be suitable for componentwise analysis or for direct interpretations on the marginal change, i.e. the effect of one variable when we keep the other fixed. To repair this drawback, separable models are considered,

$$y_i = \alpha + \sum_{d=1}^{q} \mu_d(x_{di}) + \epsilon_i,$$

where, conditionally given $x_i = (x_{d1}, \ldots x_{dq})$, ϵ_i has mean 0. For identifiability, it is often assumed that $E\{\mu_d(X_d)\} = 0$ for all d's. Stone (1985) shows that additive models avoid the curse of dimensionality and the optimal convergence rate for nonparametric additive model is the same as as that in the univariate model. However, if the additivity assumption is violated, for example when interaction exists, the fitted surface is just a additive approximation to the true surface.

The most well-known estimating technique for additive models is backfitting by Hastie and Tibshirani (1990), which was first proposed by Buja et al. (1989). The idea is to iteratively update the existing estimators and the approach has been very successful in applications. The asymptotic property of backfitting estimation was studied in Opsomer and Ruppert (1997) and Mammen et al. (1999). However, the final estimates may depend on the initial values or the convergence criterion. If strong correlation exists among covariates, backfitting might break down due to slow convergence of the iterative algorithm Jiang et al. (2010).

Alternatively, non-iterative approach can be used to fit additive model. One example is the marginal integration (MI) approach discussed in Newey (1994), Tjøstheim and Auestad (1994), Linton and Nielsen (1995). Unlike backfitting, which looks for best projections of the data onto the additive space, MI estimates the marginal effect by averaging a preliminary multivariate fit. However, the

preliminary fit might suffer a lot from sparseness of observations. Moreover, MI is not fully efficient (Linton, 1997). As Nielsen and Linton (1998) pointed out, the choice between MI and backfitting is reminiscent of the choice between ordinary least squares and generalized least squares in regression. Generally speaking, MI are better in estimating the components as opposed to that backfitting are better in estimating the regression itself.

Since both backfitting and MI approaches have their limitation, we are going to discuss a third approach, i.e. the additive P-spline approach proposed by Marx and Eilers (1998). This approach is non-iterative and easy to be implemented, but its asymptotic property is less explored. Interestingly, we establish the connection between additive P-spline approach and local constant backfitting smoothing. Unfortunately, the estimator might not have oracle property, i.e. one obtains the same asymptotic distribution of the dth component as if the other were known. We then suggest a weighted approach to improve the additive P-splines fitting. All these are addressed in Chapter 3.

All discussions above are mainly about nonparametric estimations. Although nonparametric approach is model-free, it has several limitations. Compared with parametric approach, the convergence rate of the estimator is generally slower. Moreover, it does not allow extrapolation, thus creating potential difficulty in forecasting. Sometimes, one might suspect whether a family of parametric or nonparametric models fit adequately the given data. This motivates the idea of testing parametric versus nonparametric models. Much efforts have been devoted to this problem. In a stimulating paper, Härdle and Mammen (1993) propose to test the model adequacy by measuring the L_2 distance between the parametric fit and the nonparametric fit using kernel approach. An

alternative approach is to compare the sum of square residuals under the null parametric model, see Hong and White (1995), Fan and Li (1996). Recently, Fan et al. (2001) develop the generalized likelihood ratio test (GLRT). GLRT extends the idea of likelihood principals by replacing the nonparametric maximum likelihood estimator (MLE) with a reasonable smoothed estimator. By doing so, it avoids the difficulty in obtaining the nonparametric MLE and it also improves the test efficiency. GLRT is a general and unified approach, but it is designed for independent data.

An open question is how to test the adequacy of parametric models for correlated data. As technology advances, this question attracts more and more attention since it is very easy to collect repeated measurements for the same subject by high frequency sensing machine at very fine gradations in time or space. Ramsay and Silverman (2005) term these observations as functional data. Due to the intrinsic high dimensionality, functional data analysis relies a lot on dimension reduction and nonparametric smoothing to extract the inherent feature and account for within correlation. Because of technical difficulties, most studies are exploratory rather than confirmatory.

To fill in this gap, we consider global inference for functional data in Chapter 4. Our tool is a mixed model based penalized splines. In particular, we model the regression mean by the pth degree polynomial splines basis, i.e. $\{1, x, \ldots, x^p, (x-\kappa_1)_+^p, \ldots, (x-\kappa_K)_+^p\}$ for K given knots κ_i's, treat the spline coefficients as random departure from the polynomial regression and use their variance to control the smoothness. This approach allows one to consider estimation and inference via the Maximum Likelihood (ML) or Restricted Maximum Likelihood (REML) principles. In particular, global inference on a polyno-

mial fit versus a general alternative, can be related to Likelihood Ratio type Tests (LRT) on the variance components being 0. These tests are non-standard in the sense that the parameter under the null is on the boundary of its space. The null distribution of the LRT statistic is derived in Crainiceanu and Ruppert (2004) for the single variance component case. However, their results are not directly applicable for functional observations. A remedy is to estimate the within subject covariance and consider the pseudo LRT. In Chapter 4, a pseudo LRT statistic based on a nonparametric covariance estimation is proposed and its asymptotic null distribution is derived. Effectiveness of this approach is studied through simulations and an empirical study.

CHAPTER 2

ASYMPTOTICS OF THE P-SPLINE ESTIMATIONS

2.1 Univariate P-splines

Suppose we have a univariate regression model

$$y_i = \mu(x_i) + \epsilon_i, \qquad i = 1, \ldots, n, \tag{2.1}$$

where ϵ_i has mean zero and variance $\sigma^2(x_i)$. For simplicity, we assume that $x_i \in [0, 1]$. This section presents an asymptotic theory of penalized spline estimation.

Following Eilers and Marx (1996), we use the pth degree B-splines to model the regression mean $\mu(x)$ and estimate the spline coefficients using a penalized least square criterion. To be specific, first, we calculate the pth degree B-splines for $K+p$ given knots using the following recursive formula proposed by de Boor (1978):

$$\begin{aligned} B_k^{[0]}(x) &= I_{\kappa_k, \kappa_{k+1}}(x), \\ B_k^{[p]}(x) &= \frac{x - \kappa_k}{\kappa_{k+p} - \kappa_k} B_k^{[p-1]}(x) + \frac{\kappa_{k+1} - x}{\kappa_{k+p+1} - \kappa_{k+1}} B_{k+1}^{[p-1]}(x). \end{aligned}$$

It can easily be seen that the pth degree B-splines are piecewise polynomials and every $B_k^{[p]}$ is positive on a domain spanned by $(p + 2)$ knots. Second, we express the regression mean as $\mu(x) = \sum_{k=1}^{K+p} b_k B_k^{[p]}(x)$. The coefficients $\mathbf{b} = (b_1, \ldots, b_{K+p})^T$ are then estimated by minimizing

$$\sum_{i=1}^{N} \left\{ y_i - \sum_{k=1}^{K+p} b_k B_k^{[p]}(x_i) \right\}^2 + \lambda^* \sum_{k=m+1}^{K+p} \{\Delta^m(b_k)\}^2, \ \lambda^* \geq 0, \tag{2.2}$$

where Δ is the difference operator, that is, $\Delta(b_k) = b_k - b_{k-1}$ and $\Delta^m = \Delta(\Delta^{m-1})$ for any positive integer m. The P-spline estimator is defined as $\hat{\mu}(x) = \sum_{k=1}^{K+p} \widehat{b}_k B_k^{[p]}(x)$.

As mentioned in Chapter 1, Penalized splines use less knots than a smoothing splines and is more efficient in computation. Like other nonparametric approaches, the smoothing parameters, which are the penalty λ^* and the number of knots K of the spline, play an important role. It is an interesting but challenging problem to tune the smoothing parameters and to derive theoretical properties of the estimator. Yu and Ruppert (2002) and Wand (1999) discuss the case when the number of knots is held fixed as the sample size increases. But these approaches using a fixed K are indeed parametric. Hall and Opsomer (2005) was the first to consider the situation when K is infinite. By assuming a continuum of knots and give the expression for mean integrated square error of $\hat{\mu}(x)$, i.e. a knot at every x in some interval, they give a very complicated expression for the mean integrated square error of $\hat{\mu}(x)$. But their assumption on K is too restrictive because in practice, K can converge to ∞ at a slower rate without causing severe modelling bias.

Recently, Kauermann et al. (2009) considers when K can only increase at a moderate rate. Though they do not obtain an explicit expression for the asymptotic bias and variance, they generalize their results for non-normal responses. Claeskens et al. (2009) show that depending on whether $K \to \infty$ at a sufficiently fast or sufficiently slow rate, the asymptotic distribution of a P-spline is either close to that of a smoothing spline or an OLS regression spline. Correspondingly, they refer these as two scenarios. The former, i.e. the large K scenario, is closest to current practice where a relatively large number of knots is used and overfitting is controlled by a careful choice of λ^*. In contrast, the latter, i.e. the small K scenario would require a data-based choice of K, a relatively unexplored but potentially interesting and important problem. In the next section, we will focus on the large K scenario.

2.2 Equally Distributed Covariates

We now derive the asymptotic distribution for estimators using p-th degree splines and m-th order penalty for arbitrary integers p and m. For simplicity, we assume equally spaced design points and knots, i.e. $x_i = i/n$'s and $\kappa_k = k/K$'s. This assumption will be relaxed later and we will discuss unequally spaced design points and knots in the next section.

Let B_p be the $n \times (K+p)$ matrix whose (i,j)th element is $B_j^{[p]}(x_i)$. Let D_m be the $(K+p-m) \times (K+p)$ differencing matrix such that $D_m\mathbf{b}$ is the $K+p-m$ vector of mth differences of $\mathbf{b}$. Define $P_m = D_m^T D_m$. Let $M = n/K$. Without loss of generality, we assume that it is an integer. If n/K is not integer, the number of data points in the last bin would be less than in other bins, but such a boundary effect is asymptotically negligible boundary. Here the kth bin is defined to be $(\ (k-1)/K, k/K]$.

Note that the minimizer to (2.2) satisfies

$$\hat{\mathbf{b}} = (B_p^T B_p/M + \lambda P_m)^{-1} B_p^T \mathbf{Y}/M, \tag{2.3}$$

where $\lambda = \lambda^*/M$. Denote the (k,j) element and the kth row in $(B_p^T B_p/M + \lambda P_m)^{-1}$ as v_{kj} and $\mathbf{v}_k$ respectively. Let $\tilde{y}_j$ be the jth element of $B_p^T\mathbf{Y}/M$. Then $\hat{b}_k = \sum_{j=1}^{K+p} v_{kj}\tilde{y}_j$. Our goal is to approximate v_{kj} sufficiently accurately to derive the asymptotics of $\hat{b}_k$. To avoid the boundary effect, we consider estimation at interior points where k satisfies $k/K \to x \in (0,1)$.

First, the term $B_p^T\mathbf{Y}/M$ in (2.3) can be viewed as a binned version of $\mathbf{Y}$. If $p = 0$, then $B_p^T\mathbf{Y}/M$ is the vector of bin averages of the Y_i. If $p = 1$, each Y_i is split between two adjacent bins which is called linear binning by Hall and Wand (1996). In the general case, Y_i is split between $p+1$ adjacent bins. We

will show that $(B_p^T B_p/M + \lambda P_m)^{-1}$ is asymptotically equivalent to a Nadaraya-Watson smoother matrix with $2m$-th degree kernel. We see from (2.3) that the smoothing is applied to the binned Y. Because we will let the number of bins increase sufficiently rapidly that binning effects, e. g., binning bias, are asymptotically negligible, the main result will be that penalized spline estimators have the asymptotic distributions of Nadaraya-Watson kernel estimators with equivalent kernel depending only on m, not on p. As it will be shown later, the only effect of p is that it determines the minimum rate at which the number of bins should increase; higher values $p \geq 1$ allow the number of bins to increase more slowly than $p = 0$.

Next, consider the term $B_p^T B_p/M + \lambda P_m$. Notice that B_p depends on the distribution of the covariates. When x_i's are equally distributed, $B_p^T B_p/M$ only depends on p rather than n. Hence we can define

$$\Sigma_p = B_p^T B_p/M \text{ and } \Omega_{m,p} = \Sigma_p + \lambda P_m. \tag{2.4}$$

When $p = 0$, Σ_p becomes the identity matrix. Let $q = \max(m, p)$. Interestingly, $\Omega_{m,p}$ has a special pattern that its ith row, except the first and last q rows, has the common form

$$(0, \ldots, 0, \omega_q, \ldots, \omega_1, \omega_0, \omega_1, \ldots, \omega_q, 0, \ldots, 0),$$

where ω_0 is in the ith place. Be cautious that all ω_i's depend on both m and p, and $\omega_0, \ldots, \omega_m$ also depend on λ, though the notation does not reflect this.

We will make use of the special structure of $\Omega_{m,p}$. Define $P(x)$ as

$$P(x) = \omega_q + \omega_{q-1} x + \cdots + \omega_0 x^m + \cdots + \omega_q x^{2q}. \tag{2.5}$$

We can show that there are exactly q distinct roots, say $\rho_1, \ldots, \rho_q$, of $P(x)$ with

modulus less than 1 (see Proposition 2.2.1). Let

$$\mathbf{T}_i(\rho) = \left(\rho^{|1-i|}, \ldots, \rho, 1, \rho, \ldots, \rho^{|K-i|}\right),$$

where '1' is in the ith position. Since $P(\rho_k) = 0$, $\mathbf{T}_i(\rho_k)$ is orthogonal to the columns of $\Omega_{m,p}$ except the first q columns, last q columns, and the jth column for $|i-j| < q$. Because $\rho_1, \ldots, \rho_q$ are distinct, $\mathbf{T}_i(\rho_1), \ldots, \mathbf{T}_i(\rho_q)$ are linearly independent and we can obtain a unique linear combination, $\mathbf{S}_i = \sum_{k=1}^{q} a_k \mathbf{T}_i(\rho_k)$, such that

$$\mathbf{S}_i(\Omega_{m,p})_i^T = 1 \quad \text{and} \quad \mathbf{S}_i(\Omega_{m,p})_j^T = 0, \quad \text{for} \quad 0 < |i-j| \leq q-1. \tag{2.6}$$

Therefore, $\mathbf{S}_i$ is orthogonal to all columns of $\Omega_{m,p}$ except the ith column and the first and last q columns. Moreover, suppose $i/K \to x \in (0,1)$ to avoid boundary effects. Since the ρ_k's all have modulus less than 1, if λ is chosen properly as specified below, then

$$\mathbf{S}_i(\Omega_{m,p})_j^T \approx 0, \qquad \text{for} \quad 1 \leq j \leq q \quad \text{and} \quad K-q+p \leq j \leq K+p,$$

where $\approx$ means the convergence is exponentially fast as $n \to \infty$, or more precisely, $O(r^{n^\alpha})$ where $0 \leq r < 1$, $\alpha > 0$, r and α do not depend on n; see Remark 2.2.1. Therefore, the estimators satisfy $\hat{b}_i \approx \mathbf{S}_i B_p^T \mathbf{Y}/M$.

To derive the asymptotics of P-splines, we will relate the elements in $\mathbf{S}_i$ to weights provided by a specific kernel function. Instead of directly solving for the roots ρ_k's and the coefficients a_k's in $\mathbf{S}_i$, which is not possible if $q > 2$, we find asymptotic approximation to these quantities as $\lambda \to \infty$. The following propositions originally came from Dr. Tatiyana Apanasovich.

PROPOSITION 2.2.1. *We have the following conclusions.*

(I) There are $q = \max(m,p)$ roots of $P(x)$ in (2.5) with modulus less than 1.

(II) When $p \leq m$, there are m roots converging to 1. To be specific, for $k = 1, \ldots, m$,

$$\begin{aligned} \rho_k &= 1 - \left(\frac{1}{\lambda}\right)^{\frac{1}{2m}} (\alpha_k + \beta_k \imath) + O(\lambda^{-1/m}) \\ &= \exp\left\{-\frac{\alpha_k + \beta_k \imath}{\lambda^{1/(2m)}}\right\} + O(\lambda^{-1/m}), \qquad \text{as} \quad \lambda \to \infty, \end{aligned} \tag{2.7}$$

where $\imath = \sqrt{-1}$ and $\alpha_k + \beta_k \imath$'s are the roots of $x^{2m} + (-1)^m = 0$ that satisfy $\alpha_k > 0$.

(III) When $p > m$, there are still m roots satisfying equation (2.7). In addition, there are $p-m$ roots, denoted as $\rho_{m+1}, \ldots, \rho_p$ converging to 0, i.e., for $k = 1, \ldots, p-m$,

$$\rho_{m+k} = \left\{\frac{P(0)}{\lambda}\right\}^{\frac{1}{p-m}} \tilde{\psi}_k + O(\lambda^{-2/(p-m)}), \qquad \text{as} \quad \lambda \to \infty, \tag{2.8}$$

where $\tilde{\psi}_1, \ldots, \tilde{\psi}_{q-m}$ are the roots of $x^{p-m} + (-1)^m = 0$.

REMARK 2.2.1. *Suppose that $\lambda \sim (Kh)^{2m}$ where K is the number of knots and $a \sim b$ means that $a/b \to 1$ as $n \to \infty$. Then the dominant term in $\log(\rho_k^K)$ is $-\frac{\alpha_k + \beta_k \imath}{h}$ for $k = 1, \ldots, m$. Therefore, if $h = O(n^{-\nu})$ for some $\nu > 0$, then ρ_k^K converges to 0 exponentially fast.*

By approximating the coefficients a_k's in $\mathbf{S}_i$, we will soon show that only the roots in equation (2.7) play an important role in deriving the kernel function and the asymptotic property. The roots in (2.8) are asymptotically negligible. We have the following proposition.

PROPOSITION 2.2.2. *Let $a_1, \ldots, a_q$ be the coefficients in $\mathbf{S}_i$ such that equation (2.6) holds. Assume $\lambda \to \infty$. In either case, $p \leq m$ or $p > m$, we have*

$$a_k = \frac{\alpha_k + \beta_k \imath}{2m\lambda^{1/(2m)}} + O\{\lambda^{-1/m}\}, \qquad \text{for} \quad k = 1, \ldots, m. \tag{2.9}$$

In addition, if $p > m$,

$$a_k = O(\lambda^{-p/(p-m)}), \; m < k \leq p \tag{2.10}$$

Suppose $\lambda \sim (Kh)^{2m}$. Proposition 2.2.2 motivates us to relate $\mathbf{S}_i$ to

$$\sum_{k=1}^{m} \left(\frac{\alpha_k + \beta_k \imath}{2mKh} \right) \mathbf{T}_i \left[\exp\left\{ -\frac{\alpha_k + \beta_k \imath}{\lambda^{1/(2m)}} \right\} \right].$$

Since $(i-j)/K = \bar{x}_i - \bar{x}_j$, we will show below that the dominant term in the jth element of $\mathbf{S}_i$ is asymptotically equivalent to

$$\sum_{k=1}^{m} \left(\frac{\alpha_k + \beta_k \imath}{2mKh} \right) \exp\left\{ \frac{-(\alpha_k + \beta_k \imath)|\bar{x}_i - \bar{x}_j|}{h} \right\},$$

where $\bar{x}_i = (2i-1)/(2K)$ is the center of the ith "bin," which is the interval $\Big((i-1)/K, i/K \Big]$. This allows us to connect P-spline estimators with kernel estimators using a $2m$-th kernel function spanned by double exponential or double exponentially damped sin and cos functions. To be specific, consider the kernel function

$$H_m(x) = \sum_{i=1}^{m} \frac{\alpha_i + \beta_i \imath}{2m} \exp\left\{ -(\alpha_i + \beta_i \imath)|x| \right\}. \tag{2.11}$$

A direct calculate show that $H_m(x)$ satisfies that $\int x^{2k} H_m(x) dx = 0$ for $k = 1, \ldots, m-1$ and $\int H_m(x) dx = 1$. Furthermore we can arrange the $\alpha_i + \beta_i \imath$ such that $\alpha_{2i-1} + \beta_{2i-1}\imath$ and $\alpha_{2i} + \beta_{2i}\imath$ are conjugates, for $i \leq m/2$. When m is odd, there is an additional root of $x^{2m} + (-1)^m = 0$ equal to 1. Without loss of generality, assume $\beta_{2i} > 0$. Then $H_m(x)$ can be rewritten as follows: if m is even, then

$$H_m(x) = \sum_{i=1}^{m/2} \left\{ \frac{\alpha_{2i}}{m} \exp(-\alpha_{2i}|x|) \cos(\beta_{2i}|x|) + \frac{\beta_{2i}}{m} \exp(-\alpha_{2i}|x|) \sin(\beta_{2i}|x|) \right\};$$

if m is odd, then $H_m(x)$ equals

$$\frac{\exp(-|x|)}{2m} + \sum_{i=1}^{\frac{m-1}{2}} \left\{ \frac{\alpha_{2i}}{m} \exp(-\alpha_{2i}|x|) \cos(\beta_{2i}|x|) + \frac{\beta_{2i}}{m} \exp(-\alpha_{2i}|x|) \sin(\beta_{2i}|x|) \right\}.$$

We have the following propositions.

PROPOSITION 2.2.3. *Let $h = \lambda^{1/(2m)}/K$. Then for nonboundary x, we have*

$$\hat{\mu}(x) = n^{-1} \sum_{i=1}^{n} h^{-1} H_m\left(\frac{x - x_i}{h}\right) y_i + O_p(K^{-1}) I_{p=0} + O_p\{(Kh)^{-2}\} I_{p>0}. \tag{2.12}$$

Together with asymptotic results for kernel estimators such as those obtained by Fan (1992) or Simonoff (1998), we have the following theorem.

THEOREM 2.2.1. *Assume the followings are true.*

1) *There exists* $l > 0$ *such that* $\sup_n \sup_{i \le n} E(|Y_i|^{2+l}) < \infty$.

2) *The variance* $\sigma^2(x)$ *is continuous.*

3) *The regression function* $\mu(x)$ *in model (2.1) has a continuous* $2m$*-th derivative.*

4) $\epsilon_1, \ldots, \epsilon_n$ *are mutually independent.*

5) *The covariates satisfy* $x_i = i/n$.

Suppose that $\lambda \sim (Kh)^{2m}$ *where the equivalent bandwidth is* $h \sim h_0 n^{-\frac{1}{4m+1}}$ *for some positive constant* h_0. *Suppose that* $K \sim K_0 n^{\gamma}$, *where* $\gamma > 2m/(4m+1)$ *if* $p = 0$ *or* $\gamma > (m+1)/(4m+1)$ *if* $p \ge 1$, *and* $K_0 > 0$ *is a constant. Let* $\hat{\mu}(x) = \sum_{k=1}^{K+p} B_k^{[p]}(x)\hat{b}_k$ *denote the penalized estimator using a* m*-th order penalty and a* p*-th degree spline with equally spaced knots. Then for any* $x \in (0,1)$, *we have*

$$n^{\frac{2m}{4m+1}}\{\hat{\mu}(x) - \mu(x)\} \Rightarrow N\{\beta(x), \Psi(x)\}, \qquad as \qquad n \to \infty,$$

where $\beta(x) = \frac{1}{(2m)!}\mu^{(2m)}(x)h_0^{2m}\int t^{2m}H_m(t)dt$, $\Psi(x) = h_0^{-1}\sigma^2(x)\int H_m^2(t)dt$, *and* $\Rightarrow$ *means convergence in distribution.*

2.3 Unequally Spaced Covariates

Now we want to generalize model (2.1) with unequally distributed covariates x_i's. As mentioned in Li and Ruppert (2008), when x_i are not equally distributed

and the knots are at quantiles of $x_1, \ldots, x_n$, the P-spline estimator is not design-adaptive in the sense of Fan (1992), i.e., the asymptotic bias depends on the design density. To achieve design-adaptivity, we propose to use a weighted penalized least-square approach. Instead of using (2.2), we use $\hat{\mathbf{b}} = (\hat{b}_1, \ldots, \hat{b}_{K+p})^T$ that minimizes a weighted and penalized sum of squares

$$\sum_{t=1}^{n} \left\{ y_t - \sum_{j=1}^{K+p} b_j B_j^{[p]}(x_t) \right\}^2 w(x_t) + \lambda \sum_{k=m+1}^{K+p} \{\Delta^m(b_k)\}^2, \ \lambda^* \geq 0. \tag{2.13}$$

The choice of the weights $w(x_t)$ will be discussed later. If $w(x_t)$ is a constant, the minimizer yields an ordinary penalized spline estimator. Denote W as the diagonal matrix with elements $\{w(x_1), \ldots, w(x_n)\}$. Then $\hat{b}$ solves

$$\left\{ B_p^T W B_p + \lambda D_m^T D_m \right\} \hat{\mathbf{b}} = B_p^T W \mathbf{Y}.$$

Let $n_k = \sum_{t=1}^{n} I\{\kappa_{k-1} < x_t \leq \kappa_k\}$ be the number of the data in the k-th bin. Choose the weight function such that $w(x) = n_k^{-1}$ if x falls in the k-th bin, i.e. $(k-1)/K < x \leq k/K$. A further calculation shows that when $p = 0$, $B_p^T W B_p$ equals the identity matrix I_k, and $B_p^T W \mathbf{Y} = \bar{\mathbf{Y}} = (\bar{y}_1, \ldots, \bar{y}_K)'$, where $\bar{y}_k$ is the average of all y_t in k-th bin. For higher degree splines, we have

$$(B_p^T W B_p)_{i,j} - K \int B_i^{[p]}(t) B_j^{[p]}(t) dt = o_p(1).$$

Let Σ_p and $\Omega_{m,p}$ remain the same as in equation (2.4). Note that the term $K \int B_i^{[p]}(t) B_j^{[p]}(t) dt$ equals the (i,j)th element in Σ_p. We have

1) when p is odd,

   $(B_p^T W B_p + \lambda P_m - \Omega_{m,p})_{i,j} = \sum_{j=-(p+1)/2}^{(p-1)/2} o_p(n_{i-j}^{-1})$, for $|i-j| \leq p$;

   $(B_p^T W B_p + \lambda P_m - \Omega_{m,p})_{i,j} = 0$, for $|i-j| > p$.

2) when p is even,

$(B_p^T W B_p + \lambda P_m - \Omega_{m,p})_{i,j} = \sum_{j=-p/2}^{p/2} o_p(n_{i-j}^{-1})$, for $|i-j| \le p$;

$(B_p^T W B_p + \lambda P_m - \Omega_{m,p})_{i,j} = 0$, for $|i-j| > p$.

where $(A)_{i,j}$ denotes the (i,j)th element in the matrix A. We will use these results to modify the proof of Theorem 2.2.1 and establish the following theorem.

THEOREM 2.3.1. *Assume the conditions 1)–4) in Theorem 2.2.1 are true. Assume that the empirical distribution of x_i converges weakly to a distribution with density $f(x)$ that is continuous and positive for $x \in [0,1]$. Let $h \sim h_0 n^{-\frac{1}{4m+1}}$ for some positive constant h_0. Suppose that $K \sim K_0 n^{\gamma}$, where $\gamma > 2m/(4m+1)$ if $p = 0$ or $\gamma > (m+1)/(4m+1)$ if $p \ge 1$, and $K_0 > 0$ is a constant. Let $\hat{\mu}(x)$ be the weighted penalized estimator using a m-th order penalty and a pth degree spline with equally spaced knots. Then for any $x \in (0,1)$, we have*

$$n^{\frac{2m}{4m+1}} \{\hat{\mu}(x) - \mu(x)\} \Rightarrow N\{\beta(x), \Psi(x)\}, \qquad \text{as} \qquad n \to \infty,$$

where $\beta(x) = \frac{1}{(2m)!}\mu^{(2m)}(x) h_0^{2m} \int t^{2m} H_m(t) dt$ and $\Psi(x) = \frac{\sigma^2(x)}{h_0 f(x)} \int H_m^2(t) dt$.

REMARK 2.3.1. *The asymptotic bias does not depend on the empirical distribution of x. Therefore, the weighted P-spline estimation is design adaptive at any x in $(0,1)$.*

2.4 Comparisons with O-splines

O'Sullivan (1986) introduced another class of penalized splines that are called O'Sullivan splines or O-splines by Wand and Ormerod (2008). O-splines use the same penalty as smoothing splines but, like P-splines, O-splines have a reduced number of knots. For simplicity, we will focus on cubic O-splines, denoted as $B_{3,j}(x)$ for $j = 1, \ldots, K+3$, as an example. Let B_3 be the design matrix with

(i,j)th entry $B_{3,j}(x_i)$. Define the (k,k')th element of the penalty matrix Γ as $\int B^{(2)}_{3,k}(x)B^{(2)}_{3,k'}(x)dx$. Then $\hat{\mu}_o(x) = \mathbf{B}_3(x)\hat{\mathbf{b}}$ is the penalized estimator using cubic O-splines, where $\mathbf{B}_3(x) = \{B_{3,1}(x), \ldots, B_{3,K+3}(x)\}$ and $\hat{\mathbf{b}} = (B_3^T B_3 + \lambda\Gamma)^{-1} B_3^T \mathbf{Y}$.

When there is an interior knot at each unique x_i, $\hat{\mu}_o(x)$ is a smoothing spline. A detailed study on the asymptotic distribution of smoothing splines is given by Silverman (1984). As Wand and Ormerod (2008) point out, P-splines and O-splines are similar when the knots are equally-spaced because the differencing penalty is equivalent to a discrete approximation to the integrated square of the mth derivatives of the B-spline smoother. The methodology in the previous sections can be used to derive the asymptotic distribution of O-splines.

THEOREM 2.4.1. *Assume the conditions 1)-5) in Theorem 2.2.1 hold. Suppose that* $\lambda \sim (Kh)^4$ *where the equivalent bandwidth is* $h \sim h_0 n^{-\frac{1}{9}}$ *for some positive constant* h_0*. Suppose that* $K \sim K_0 n^{\gamma}$*, where* $K_0 > 0$ *is a constant and* $\gamma > 1/3$*. Let* $\hat{\mu}(x)$ *be the penalized estimator using the cubic O-splines with equally spaced knots. Then for any* $x \in (0,1)$*, we have the same conclusion for* $\hat{\mu}(x)$ *as Theorem 2.2.1 with* $m = 2$*.*

Theorem 2.4.1 shows that for interior x, P-splines and O-splines have the same asymptotic distributions. Wand and Ormerod (2008) show, however, that O-splines and P-splines have different finite-sample behavior at the boundaries. The asymptotics of P-splines at the boundaries is studied by Wang et al. (2010). For generalization and applications of O-splines, one can refer to Wahba (1990).

CHAPTER 3

THE ADDITIVE MODELS

3.1 Additive P-splines

We now extend our study on P-splines from the univariate setting to the multivariate setting. We first assume that the marginal distributions of the covariates are uniform. This assumption will be relaxed in the next section.

Consider the bivariate additive model

$$y_i = \alpha + \sum_{d=1}^{2} \mu_d(x_{di}) + \epsilon_i, \qquad 1 \le i \le n, \tag{3.1}$$

where, conditionally given $x_i = (x_{1i}, x_{2i})$, ϵ_i has mean 0. For simplicity, the domain of the covariates is set to be $[0,1]^2$. We first consider the additive P-splines approach proposed by Marx and Eilers (1998). Let $0 = \kappa_0 < \kappa_1 < \cdots < \kappa_{K+p_d}$ be equally-spaced knots and $B_k^{[p_d]}(x)$ be the p_dth degree B-spline defined on $(\kappa_{k-1}, \kappa_{k+p_d-1}]$. Then we can model the dth component as,

$$\mu_d(x) = \sum_{k=1}^{K+p_d} B_k^{[p_d]}(x) b_{dk}.$$

For identifiability, it is often assumed that $E\{\mu_d(X_d)\} = 0$ for all d's. Correspondingly, we impose the following constraints on the P-spline coefficients:

$$n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{K+p_d} B_k^{[p_d]}(x_{di}) b_{dk} = 0, \qquad \text{for } d = 1, 2. \tag{3.2}$$

Under this constraint, we choose $\hat{\alpha}$, $\hat{b}_{dk}$ that minimize

$$\sum_{i=1}^{n} \left\{ y_i - \alpha - \sum_{d=1}^{2} \sum_{k=1}^{K+p_d} b_{dk} B_k^{[p_d]}(x_{di}) \right\}^2 + \sum_{d=1}^{2} \lambda_d^* \sum_{k=m+1}^{K+p_d} (\Delta^m b_{dk})^2, \tag{3.3}$$

where λ_d^* is the penalty parameter, Δ is the difference operator so that $\Delta b_{dk} = b_{dk} - b_{d,k-1}$, m is a positive integer, and $\Delta^m = \Delta(\Delta^{m-1})$. Then $\hat{\mu}_d(x_d) = \sum_{k=1}^{K+p_d} \hat{b}_{dk} B_k^{[p_d]}(x_d)$ is called the P-spline estimator for the dth component.

By taking the derivative of (3.3) with respect to α and setting it to be 0, we have

$$\hat{\alpha} = n^{-1} \sum_{i=1}^{n} \left\{ y_i - \sum_{d=1}^{2} \sum_{k=1}^{K+p_d} b_{dk} B_k^{[p_d]}(x_{di}) \right\}. \tag{3.4}$$

Because of the constraints (3.2), we have $\hat{\alpha} = \bar{y}$. The randomness in $\hat{\alpha}$ is of smaller order than any nonparametric estimators and can be effectively ignored. The objective function (3.3) now becomes

$$\sum_{i=1}^{n} \left\{ y_i - \bar{y} - \sum_{d=1}^{2} \sum_{k=1}^{K+p_d} b_{dk} B_k^{[p_d]}(x_{di}) \right\}^2 + \sum_{d=1}^{2} \lambda_d^* \sum_{k=m+1}^{K+p_d} (\Delta^m b_{dk})^2. \tag{3.5}$$

We now derive the minimizer of (3.5) under the constraints (3.2). Let B_d be an $n \times (K+p)$ matrix with (t,j) th entry equal to $B_j^{[p_d]}(x_{dt})$. The design matrix becomes $(B_1, \ldots, B_d)$. Let $\tilde{Y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$. Define D_m as the $(K - m + p_d) \times (K + p)$ differencing matrix satisfying

$$D^m b = \begin{pmatrix} \Delta^m(b_{m+1}) \\ \vdots \\ \Delta^m(b_{K+p}) \end{pmatrix}.$$

Let $P_m = D_m^T D_m$. Denote $M = n/K$, $\lambda_d = \lambda_d^*/M$, $C_{ij} = B_i^T B_j / M$ and $V_i = B_i^T \tilde{Y}/M$. Consider the following linear equations.

$$\begin{pmatrix} C_{11} + \lambda_1 P_m & C_{12} \\ C_{21} & C_{22} + \lambda_2 P_m \end{pmatrix} \begin{pmatrix} \tilde{b}_1 \\ \tilde{b}_2 \end{pmatrix} = \begin{pmatrix} V_1 \\ V_2 \end{pmatrix}. \tag{3.6}$$

Note that the matrix on the left hand side of equation (3.6) is singular. For any of its solution $\tilde{b}_1^T$ and $\tilde{b}_2^T$, $\tilde{b}_1^T + c1_{K+p_1}$ and $\tilde{b}_2^T - c1_{K+p_2}$ are also its solution for any constant c. However, there exists a constant c such that $\tilde{b}_1^T + c1_{K+p_1}$ and

$\tilde{b}_2^T - c1_{K+p_2}$ satisfy constraints (3.2). Note that this solution is unique for any $\tilde{b}_d$'s. They are the estimated P-splines coefficients.

The following proposition shows that $\hat{b}_d$'s can also be obtained by the following iterative approach A.1. For notation simplification, we use $\hat{b}_d^{(j)}$ to indicate the estimators in the jth step.

PROPOSITION 3.1.1. *Set $\hat{b}_d^{(0)}$ as the zero vector. For step $j+1$, update the estimates by*

$$\hat{b}_d^{(j+1)} = S_d V_d - \sum_{d' \neq d} \{S_d C_{dd'} \hat{b}_{d'}^{(j*)}\},\ d = 1, 2,$$

where $S_d = (C_{dd} + \lambda_d P_m)^{-1}$ and $\hat{b}_d^{(j*)} = S_d V_d - \sum_{d' \neq d} \{S_d C_{dd'} \hat{b}_{d'}^{(j)}\},\ d = 1, 2.$

Then $\lim_{j \to \infty} \hat{b}_d^{(j)}$ exist for all d and they satisfy (3.6) and the constraints (3.2).

Proposition 3.1.1 indicates that the P-spline estimators can be obtained by an iterative algorithm. Motivated by the connection between P-splines and kernel approach established in Chapter 1, we now relate Approach A.1 to the iterative approach A.2.

1 Set $\bar{\mu}_d^{[0]}(x_d) = 0$.

2 For step $j+1$, update the estimates by

$$\bar{\mu}_d^{(j+1)}(x_d) = \tilde{\mu}_d(x_d) - \sum_{d' \neq d} \int \bar{\mu}_{d'}^{(j*)}(x_{d'}) \bar{f}(x_d, x_{d'}) / \bar{f}_d(x_d) dx_{d'} - \bar{y}, \qquad (3.7)$$

where

$$\tilde{\mu}_d(x_d) = \frac{n^{-1} \sum_{i=1}^n H_m\{(x_d - x_{di})/h_d\} y_i}{n^{-1} \sum_{i=1}^n H_m\{(x_d - xdi)/h_d\}}, \qquad (3.8)$$

$$\bar{\mu}_d^{(j*)}(x_d) = \tilde{\mu}_d(x_d) - \sum_{d' \neq d} \int \bar{\mu}_{d'}^{(j)}(x_{d'}) \bar{f}(x_d, x_{d'}) / \bar{f}_d(x_d) dx_{d'} - \bar{y},$$

$\bar{f}(x_d, x_{d'})$ is the bivariate kernel density estimator, i.e.,

$$\bar{f}(x_d, x_{d'}) = (nh_d h_{d'})^{-1} \sum_{i=1}^{n} H_m\{(x_d - x_{di})/h_d\} H_m\{(x_{d'} - x_{d'i})/h_{d'}\},$$

$H_m(\cdot)$ is the equivalent kernel defined in equation (2.11), and $\bar{f}(x_d)$ is the univariate marginal kernel density estimator of $f(x_d)$.

We can further conclude the following.

PROPOSITION 3.1.2. *Suppose that $K \sim \tau n^{\gamma}$, where $\gamma > 2m/(4m+1)$ if $\min_d(p_d) = 0$ and $\gamma > (m+1)/(4m+1)$ if $\min_d(p_d) \geq 1$. Suppose that λ_d's are chosen so that $\lambda_d \sim (Kh_d)^{2m}$, where $h_d = h'_d n^{-1/(4m+1)}$ for some constant h'_d. Let $\bar{\mu}_d^{(j)}(x_d)$ be defined as in Approach A.2. Let $\hat{b}_d^{(j)}$ be defined as in Approach A.1. Then for interior x_d,*

$$n^{2m/(4m+1)} \Big(\sum_{k=1}^{K+p_d} \hat{b}_{dk}^{(j)} B_k^{[p_d]}(x_d) - \bar{\mu}_d^{(j)}(x_d) \Big) \xrightarrow{p} 0, \ j = 0, 1, 2, \ldots.$$

Mammen et al. (1999) have studied Approach A.2 and concluded that the limit of $\bar{\mu}_d^{(j)}(x_d)$, denoted as $\bar{\mu}_d(x_d)$, is the solution of the

$$\begin{aligned} \bar{\mu}_1(x_1) &= \tilde{\mu}_1(x_1) - \int \bar{\mu}_2(x_2) \bar{f}(x_1, x_2)/\bar{f}_1(x_1) dx_2 - \bar{y} \\ \bar{\mu}_2(x_2) &= \tilde{\mu}_2(x_2) - \int \bar{\mu}_1(x_1) \bar{f}(x_1, x_2)/\bar{f}_2(x_2) dx_1 - \bar{y}, \end{aligned}$$

where $\tilde{\mu}_d(x_d)$ is defined as in equation (3.8). They further derive the asymptotic distribution of $\bar{\mu}_d(x_d)$. Applying their results, we can conclude the same for the P-spline estimators.

THEOREM 3.1.1. *Assume the followings are true:*

1) *The additive model (3.1) holds.*

2) *For some $\theta > 2 + \ell$ where ℓ is positive, it holds that $E(|y|^{\theta}) < \infty$.*

3) *The empirical distribution of* $\{(x_{1i}, x_{2i}\}_{i=1}^n$ *converges weakly to a distribution with density* $f(x_1, x_2)$*, and that* $f(x_1, x_2)$ *is continuous and positive on* $[0,1]^2$*. The* $(2m-1)$*th partial derivatives of* $f(x_1, x_2)$ *also exist and are continuous.*

4) *The regression function* $\mu_d(x_d)$*,* $d = 1, 2$ *has continuous* $2m$*th partial derivatives.*

5) *The conditional variances* $\sigma_d^2(x_d) = \mathrm{var}(y|x_d)$ *are continuous.*

6) *The marginal distribution of* x_d *is uniform.*

Let $h_d = h_d' n^{-1/(4m+1)}$ *for some positive constant* h_d'*. Suppose that the number of knots satisfies that* $K \sim \tau n^\gamma$*, where* $\gamma > 2m/(4m+1)$ *if* $\min_d(p_d) = 0$ *and* $\gamma > (m+1)/(4m+1)$ *if* $\min_d(p_d) \geq 1$*. Suppose that* λ_d *are chosen so that* $\lambda_d \sim (Kh_d)^{2m}$*. Use the* p_d*th degree B-splines with equally spaced knots and* m*th order penalty to fit the* d*th component and obtain the estimates* $\hat{\mu}_d(x_d)$*.*

Then for $x \in (0,1)^2$*,*

$$n^{\frac{2m}{4m+1}}\{\hat{\mu}_d(x_d) - \mu_d(x_d)\} \Rightarrow N\{(h_d')^{2m}\beta_d(x_d)\int t^{2m}H_m(t)dt, \Psi_d(x_d)\}, \quad (3.9)$$

where $\beta_d(x_d)$*'s are the* $L^2(p)$ *projections of* $\beta(x_1, x_2)$*, which equals*

$$\left[\frac{1}{(2m)!}\sum_{d=1}^2 \mu_d^{(2m)}(x_d) + \sum_{d=1}^2\sum_{i=1}^{2m-1}\left\{\frac{\mu_d^{(i)}(x_d)}{i!(2m-i)!}\frac{\partial^{2m-i}}{\partial^{2m-i}x_d}f(x_1,x_2)\right\}\right]/f(x_1,x_2),$$

onto the space of additive functions, i.e., they minimize $\int\{\beta(x_1,x_2) - \beta_0 - \sum_{d=1}^2\beta_d(x_d)\}^2 f(x_1,x_2)dx_1dx_2$ *and they satisfy* $\int \beta_d(x_d)dx_d = 0$*,*

$$\Psi_d(x) = (h_d')^{-1}\sigma_d^2(x_d)\int H_m^2(t)dt,$$

and $H_m(t)$ *is given by (2.11),*

REMARK 3.1.1. *When* $m = 1$*, the bias is more than* $\sum_{d=1}^2 \mu_d^{(2m)}(x_d)$ *unless the joint density* $f(x_1, x_2)$ *can be expressed as* $\prod_{d=1}^2 f_d(x_d)$*. So the penalized spline estimator is not design-adaptive in the sense of Fan (1992), i.e. the asymptotic bias depends on*

the design density. However, $\Psi_d(x_d)$'s are the asymptotic variance of the oracle estimators, i.e. the same as if we estimate $\mu_d(x_d)$ with prior knowledge of all other additive components.

3.2 Weighted P-splines

In this section, we are still considering the bivariate model (3.1), i.e. $y_i = \alpha + \mu_1(x_{1i}) + \mu_2(x_{2i}) + \epsilon_i$. Different from the above section, we do not require that the data are marginally uniformly distributed.

We first consider using the zero degree splines to model each of these two components, i.e., $\mu_d(x) = \sum_{k=1}^{K} b_{dk} B_k(x)$. For identifiability, we assume that $\sum_{k=1}^{K} b_{dk} = 0$. Under this restriction, the weighted estimator $\hat{\alpha}$, $\hat{b}_1 = (\hat{b}_{11}, \ldots, \hat{b}_{1K})$ and $\hat{b}_2 = (\hat{b}_{21}, \ldots, \hat{b}_{2K})$ are chosen to minimize

$$\sum_{i=1}^{n} \left\{ y_i - \alpha - \sum_{k=1}^{K} b_{1k} B_k(x_{1i}) - \sum_{k=1}^{K} b_{2k} B_k(x_{2i}) \right\}^2 w(x_{1i}, x_{2i}) + \lambda_1 \sum_{k=m+1}^{K} (\Delta^m b_{1k})^2 + \lambda_2 \sum_{k=m+1}^{K} (\Delta^m b_{2k})^2, \tag{3.10}$$

where $\lambda_1 \geq 0$, $\lambda_2 \geq 0$, Δ is the difference operator and m is the order of penalty.

The weighting matrix is chosen as follows: if (x_{1i}, x_{2i}) falls in the $(k, k')$the cell, then $w(x_{1i}, x_{2i}) = 1/n_{kk'}$, where $n_{kk'}$ is the count of data in the (k, k')th cell. Denote $\bar{y}_{kk'}$ be the average of the (k, k')th cell. By taking the derivative of (3.10) with respect to α and setting it to be 0, we have

$$\hat{\alpha} = \bar{y} =: K^{-1} K^{-1} \sum_{k=1}^{K} \sum_{k'=1}^{K} \bar{y}_{kk'}, \tag{3.11}$$

which is a weighted average of all responses. The randomness in $\hat{\alpha}$ is of smaller order than any nonparametric estimators and can be effectively ignored.

Let B_1 be an $n \times K$ matrix with (i,k) th entry equal to $B_k(x_{1i})$. Let B_2 be another $n \times K$ matrix with (i,k') entry $B_{k'}(x_{2i})$ and $\mathcal{B} = [B_1, B_2]$. Let $\tilde{Y} = (y_1 - \bar{y}, \ldots, y_n - \bar{y})^T$. Similar as the previous section, the P-splines coefficients $\hat{b}$ simultaneously satisfy $\sum_{k=1}^{K} \hat{b}_{dk} = 0$ and the following equation:

$$\left[\mathcal{B}^T W \mathcal{B} + \begin{pmatrix} \lambda_1 P_m & 0_{K\times K} \\ 0_{K\times K} & \lambda_2 P_m \end{pmatrix} \right] \hat{b} = \mathcal{B}^T W \tilde{Y},$$

where $0_{K\times K}$ is the $K \times K$ dimensional zero matrix. We first consider the case when $n_{kk'} > 0$ for $k = 1, \ldots, K$ and $k' = 1, \ldots, K$. Relaxation of this assumption will be given later. A further calculation shows that $\mathcal{B}^T W \mathcal{B} = \begin{pmatrix} KI_K & J_{K\times K} \\ J_{K\times K} & KI_K \end{pmatrix}$, where $J_{K\times K}$ is the matrix whose elements are all 1. Denote $\mathcal{B}^T W \tilde{Y}$ as $\tilde{V}$. Partition $\tilde{V}$ into two $K \times 1$ subvectors $\tilde{V}_1$ and $\tilde{V}_2$. Let $S_1 = (I_K + \lambda_1 P_m)^{-1}$ and $S_2 = (I_K + \lambda_2 P_m)^{-1}$. Hence the minimizer of (3.10) simultaneously solve

$$\begin{aligned} \hat{b}_1 &= S_1 \tilde{V}_1 - S_1 J_{K\times K}/K \hat{b}_2 \\ \hat{b}_2 &= S_2 \tilde{V}_2 - S_2 J_{K\times K}/K \hat{b}_1. \end{aligned} \tag{3.12}$$

Notice that we require $\sum_{k=1}^{K} \hat{b}_{1k} = 0$ and $\sum_{k'=1}^{K} \hat{b}_{2k'} = 0$. Therefore, the above equations can be further simplified as

$$\begin{aligned} \hat{b}_1 &= S_1 \tilde{V}_1 \\ \hat{b}_2 &= S_2 \tilde{V}_2. \end{aligned} \tag{3.13}$$

Let 1_a be the $a \times 1$ column vector whose components are all 1. Note that

$$1_K^T \tilde{V}_d =: 1_K^T B_d^T W \tilde{Y} = 1_n^T W \tilde{Y} = \sum_{k=1}^{K} \sum_{k'=1}^{K} (\bar{y}_{k,k'} - \hat{\alpha}) = 0,$$

where the last equality holds because of equation (3.11). Hence both $\tilde{V}_1$ and $\tilde{V}_2$ are centered. Therefore, $\hat{b}_1$ and $\hat{b}_2$ defined by (3.13) satisfies the constraints (3.2). Since they also minimize (3.10), they are the weighted penalized spline estimator. In contrary to last section, they can directly be obtained without using an

iterative approach. Using the same techniques in deriving the asymptotics of univariate P-splines, we can conclude that the asymptotic bias for the dth component only depends on the $2m$th derivatives of $\mu_d(x_d)$. Since the use of higher degree splines will not affect the asymptotic distribution of the estimators, we have the following conclusions in general.

THEOREM 3.2.1. *Assume conditions 1)-5) in Theorem 3.1.1.*

Let $h_d = h'_d n^{-1/(4m+1)}$ *for some positive constant* h'_d*. Suppose that the number of knots satisfies that* $K \sim \tau n^{\gamma}$*, where* $\gamma > 2m/(4m+1)$ *if* $\min_d(p_d) = 0$ *and* $\gamma > (m+1)/(4m+1)$ *if* $\min_d(p_d) \geq 1$*. Suppose that* λ_d *are chosen so that* $\lambda_d \sim (Kh_d)^{2m}$*. Use the* p_d*th degree B-splines with equally spaced knots and* m*th order penalty to fit the* d*th component and obtain the estimates* $\hat{\mu}_d(x_d)$*. Then for* $(x_1, x_2) \in (0,1) \times (0,1)$*,*

$$n^{2m/(4m+1)}(\hat{\mu}_d(x_d) - \mu_d(x_d)) \Rightarrow N\{(h'_d)^{2m}\mu_d^{(2m)}(x_d) \int t^{2m} H_m(t)dt, \Psi_d(x_d)\}$$

where $\Psi_d(x_d) = \sigma_d^2(x_d)\{h'_d f(x_d)\}^{-1} \int H_m^2(t)dt$ *and* $f(x_d)$ *is the marginal density for* x_d*,* $d = 1, 2$*.*

Theorem 3.2.1 implies that the weighted P-spline estimators preserve the requires oracle property, i.e. the estimator has the same asymptotics as if other components are known, if the joint density $f(x_1, x_2)$ is positive over the domain. However, if the joint density does not satisfy this assumption, the weighted approach might not be oracle. Alternatively, we suggest a two-stage approach A.3, which uses the weighted P-spline approach as its first step.

1 Obtain the weighted P-spline coefficients estimators $\hat{b}_j^s$, by using a suboptimal penalty choice where $\lambda_s = o\{(Kh_s)^{2m}\}$ and $h_s = o(n^{-1/(4m+1)})$.

2 Conduct a one step updated procedure by

$$\hat{b}_d \;=\; S_d V_d - \sum_{j \neq d} (S_d C_{dj} \hat{b}_j^s), \tag{3.14}$$

where $S_d = (C_{dd} + \lambda_d P_m)^{-1}$, $\lambda_d = (K h_d)^{2m}$ and $h_d \sim h'_d n^{-2m/(4m+1)}$.

Let $B_d(x_d) = \{B_{d1}(x_d), \ldots, B_{dK}(x_d)\}$. Denote $\hat{\mu}_d^s(x_d) = B_d(x_d)\hat{b}_d^s$ be the P-spline estimator using the suboptimal penalty λ_s. Note that equation (3.14) can be written as

$$\hat{b}_d = S_d B_d^T W\{\tilde{Y} - \sum_{d' \neq d} \mu_{d'}(X_{d'})\} + S_d B_d^T W \sum_{d' \neq d} \{\mu_{d'}(X_{d'}) - \hat{\mu}_{d'}^s(X_{d'})\}, \tag{3.15}$$

where the first term is the estimator as if all the other components are known. Since we use an undersmooth estimator, the second term is asymptotically negligible and the two-stage estimator can achieve full efficiency.

THEOREM 3.2.2. *Assume conditions 1)-5) in Theorem 3.1.1.*

Suppose that the number of knots satisfies that $K \sim \tau n^{\gamma}$, *where* $\gamma > 2m/(4m+1)$ *if* $\min_d(p_d) = 0$ *and* $\gamma > (m+1)/(4m+1)$ *if* $\min_d(p_d) \geq 1$. *Consider the* p_d*th degree B-splines with equally spaced knots and* m*th order penalty. Obtain the estimator* $\hat{\mu}_d(x_d)$ *by Approach A.3. Then for* $(x_1, x_2) \in (0,1) \times (0,1)$,

$$n^{2m/(4m+1)}\{\hat{\mu}_d(x_d) - \mu_d(x_d)\} \Rightarrow N\{(h'_d)^{2m}\mu_d^{(2m)}(x_d) \int t^{2m} H_m(t)dt, \Psi_d(x_d)\}$$

where $\Psi_d(x_d) = \sigma_d^2(x_d)\{h'_d f(x_d)\}^{-1} \int H_m^2(t)dt$ *and* $f(x_d)$ *is the marginal density for* x_d, $d = 1, 2$.

The idea of Approach A.3 is similar to the two step approach proposed by Linton (1997). The former uses weighted P-spline estimators while the latter uses marginal integration in its first step. In terms of asymptotic distribution,

these two approaches are equivalent. In practice, we propose to use weighted P-splines in the first step because marginal integration might not be very stable and efficient since it seeks projection with respect to a product measure rather than the joint (Nielsen and Linton, 1998).

CHAPTER 4

INFERENCES UNDER LINEAR MIXED MODEL FRAMEWORK

4.1 General Methodology

We now focus on inferential problems. Consider the univariate model with repeated measurements, i.e.

$$y_{ij} = \mu(t_{ij}) + e_{ij},\ i = 1, \dots n, j = 1, \dots, m_i \tag{4.1}$$

where $\mu(t)$ is a smooth function, i is the subject index and m_i is the number of observations associated with subject i. We are interested in global inference on $\mu(t)$. Examples include testing (1) whether $\mu(t) \equiv 0$; (2) whether $\mu(t)$ is a qth degree polynomial. To define an alternative that is flexible enough and suitable for testing, we express $\mu(t)$ using the pth degree polynomial splines where $p \geq q$, i.e.

$$\mu(t) = x_t\beta + z_t u, \tag{4.2}$$

where $x_t = (1, t, \dots, t^p)$, $z_t = \{(t-\kappa_1)_+^p, \dots, (t-\kappa_K)_+^p\}$, $\beta = (\beta_0, \dots, \beta_p)$, $u = (u_1, \dots, u_K)$, $\kappa_1, \dots, \kappa_K$ are K given knots and K does not depend on n. The coefficients u measures the departure of $\mu(t)$ from the pth polynomial.

We consider the mixed-model based penalized splines approach. Under this framework, we treat β and u as fixed effect and random effect parameters respectively. Let $Y_i = (y_{i1}, \dots, y_{im_i})^T$ and $Y = (Y_1^T, \dots, Y_n^T)^T$. Similarly define e based on all e_{ij}'s. Define X_i such that its jth row is $x_{t_{ij}}$ and let $X = (X_1^T, \dots, X_n^T)^T$. Similarly define Z based on all $z_{t_{ij}}$'s. Then we can rewrite model 4.1 as $Y = X\beta + Zu + e$. Further assume that $\mathrm{cov}(e) = \Sigma$ for some positive

definite matrix Σ, and

$$\begin{pmatrix} u \\ e \end{pmatrix} \sim N\left\{ \mathbf{0}_{(K+n)\times 1}, \begin{pmatrix} \sigma_u^2 I_K & \mathbf{0} \\ \mathbf{0} & \Sigma \end{pmatrix} \right\}, \tag{4.3}$$

where $\mathbf{0}_{s\times t}$ is the $s \times t$ dimensional zero matrix. Global inferences on the mean function $\mu(t)$ can be considered via the hypothesis test on β and σ_u^2. In general, define Q be a subset of $\{0, 1, \ldots, p\}$. We are interested in testing the null hypothesis as

$$H_0 : \beta_q = 0 \quad \text{for all} \quad q \in Q \qquad \text{and} \quad \sigma_u^2 = 0 \tag{4.4}$$

versus the composite alternative

$$H_A : \exists\, q_0 \in Q \quad \text{such that} \quad \beta_{q_0} \neq \beta_{q_0}^0 \quad \text{or} \quad \sigma_u^2 > 0. \tag{4.5}$$

Such a hypothesis can be formulated to different tests. When $Q = \{0, 1, \ldots, p\}$, we are testing whether $\mu(t) \equiv 0$. When $Q = \{q+1, \ldots, p\}$ for some given q, we are testing whether $\mu(t)$ is a qth degree polynomial. Note that if $q = p$, Q is the null set and we are only testing the null of zero variance of random effects, i.e. $\sigma_u^2 = 0$.

The test of (4.4) vs (4.5) is not standard because, under the null, the parameter σ_u^2 is on the boundary of the parameter space. Large sample properties of LRT statistic when the true parameter value may be on the boundary of the parameter space were first discussed in Self and Liang (1987). However, their results of having an asymptotic $0.5\chi_\tau^2 : 0.5\chi_{\tau+1}^2$ null distribution might lead to a very conservative test because one of their assumptions is violated due to the fact that the response variable vector Y can not be partitioned into i.i.d. subvectors under the alternative. Crainiceanu and Ruppert (2004) consider this problem and relax Self and Liang's assumption. They derive the finite sample and

asymptotic null distribution of the LRT when assuming the covariance Σ is proportional to the identity matrix. We now consider further extend their approach for more complex covariance structure.

Provided that Σ in equation (4.3) is known, we can define twice of the log-likelihood of Y as

$$2\log L_Y(\beta,\lambda) = -\log(|\Sigma+\lambda ZZ^T|) - (Y-X\beta)^T(\Sigma+\lambda ZZ^T)^{-1}(Y-X\beta),$$

and the LRT statistic as

$$LRT_N = \sup_{H_0\cup H_A} 2\log L_Y(\beta,\lambda) - \sup_{H_0} 2\log L_Y(\beta,\lambda).$$

In practice, Σ is unknown. Hence we will consider the pseudo LRT, which replaces Σ with an estimate $\hat{\Sigma}$. Let $A^{-1/2}$ be the matrix square root of A^{-1} for any positive definite matrix A. By replacing Σ with $\hat{\Sigma}$, we define the pseudo loglikelihood as

$$2\log \hat{L}_{\hat{Y}}(\beta,\lambda) = -\log|\hat{H}_\lambda| - (\hat{Y}-\hat{X}\beta)^T\hat{H}_\lambda^{-1}(\hat{Y}-\hat{X}\beta), \tag{4.6}$$

where

$$\hat{Y} = \hat{\Sigma}^{-1/2}Y, \qquad \hat{X} = \hat{\Sigma}^{-1/2}X, \qquad \hat{Z} = \hat{\Sigma}^{-1/2}Z, \tag{4.7}$$

and $\hat{H}_\lambda = I_N + \lambda\hat{Z}\hat{Z}^T$. The pseudo LRT statistic to test (4.4) versus (4.5) is

$$pLRT_N = \sup_{H_0\cup H_A} 2\log \hat{L}_{\hat{Y}}(\beta,\lambda) - \sup_{H_0} 2\log \hat{L}_{\hat{Y}}(\beta,\lambda).$$

Proposition 4.1.1 below derives the asymptotic distribution of the pseudo LRT under the null hypothesis.

PROPOSITION 4.1.1. *Assume the following conditions:*

(C1) The null hypothesis H_0 defined in (4.4) is true.

(C2) The covariance matrix $\Sigma = \mathrm{cov}(e)$ in equation (4.3) is positive definite. Let $\hat{\Sigma}$ be its consistent estimate satisfying that

$$a^T\hat{\Sigma}^{-1}a - a^T\Sigma^{-1}a = o_p(1) \qquad a^T\hat{\Sigma}^{-1}e - a^T\Sigma^{-1}e = o_p(1), \tag{4.8}$$

where a is any $N \times 1$ non random normalized vector, and N is the sample size, i.e. the length of $Y = (Y_1^T, \ldots, Y_n^T)^T$.

(C3) There exists positive constants ϱ' and ϱ such that $N^{-\varrho'}X^TX$ and $N^{-\varrho}Z^TZ$ converge to nonzero matrices respectively. For every eigenvalue $\tilde{\xi}_{k,N}$ and $\tilde{\zeta}_{k,N}$ of the matrices $N^{-\varrho}Z^T\Sigma^{-1}Z$ and $N^{-\varrho}\{Z^T\Sigma^{-1}Z - Z^T\Sigma^{-1}X(X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}Z\}$, we have

$$\tilde{\xi}_{k,N} \to \xi_k \quad \text{and} \quad \tilde{\zeta}_{k,N} \to \zeta_k,$$

for some $\xi_1, \ldots, \xi_K$, $\zeta_1, \ldots, \zeta_K$, not all of which are 0.

Denote τ the cardinality of the set Q in the null hypothesis (4.4). Then

$$pLRT_N \Rightarrow \sup_{\lambda \geq 0} LRT_\infty(\lambda) + \sum_{j=1}^{\tau} \nu_j^2, \tag{4.9}$$

$$LRT_\infty(\lambda) = \sum_{k=1}^{K} \frac{\lambda}{1+\lambda\zeta_k} w_k^2 - \sum_{k=1}^{K} \log(1+\lambda\xi_k),$$

$w_k \sim N(0, \zeta_k)$ for $k = 1, \ldots, K$, $\nu_j \sim N(0,1)$ for $j = 1, \ldots, \tau$, and all w_k's ν_j's are mutually independent.

REMARK 4.1.1. *Condition (C2) discusses when the estimator $\hat{\Sigma}$ can work as well as the true Σ such that quantities defined using Σ^{-1} can be well approximated by those defined similarly using $\hat{\Sigma}^{-1}$. In the special case when X_i and Z_i remain the same for all i, and $\Sigma = I_n \otimes \Sigma_0$ where $\otimes$ is the kronecker product, we can replace condition (C2) by the following:*

(C2′) That the minimum eigenvalue of Σ_0 is bounded away from 0. Let $\hat{\Sigma}_0$ be its consistent estimate satisfying that

$$a^T\hat{\Sigma}_0^{-1}a - a^T\Sigma_0^{-1}a = o_p(1) \qquad a^T\hat{\Sigma}_0^{-1}e_0 - a^T\Sigma_0^{-1}e_0 = o_p(1), \tag{4.10}$$

where a is any $m \times 1$ non random normalized vector and $e_0 = n^{-1/2}\sum_{i=1}^n e_i$.

REMARK 4.1.2. *The assumption in condition (C3) is very mild. In the special case when the observation time points are equally-spaced and Z is the associated design matrix using polynomial splines with equally spaced knots (see section 4.2), we can choose $\varrho = 1$, see Crainiceanu (2003), page 70. The asymptotic null distribution of $pLRT$ is not standard, but can be simulated effectively by precalculating the eigenvalues ξ_k's and ζ_k's (see Algorithm A.4 below). In practice, Σ is unknown. Lemma C.1.1 in the Appendix indicates that we can replace ξ_k by $\hat{\xi}_k$ and ζ_k by $\hat{\zeta}_k$, where $\hat{\xi}_k$ and $\hat{\zeta}_k$ are the kth eigenvalues of $N^{-\varrho}Z^T\hat{\Sigma}^{-1}Z$ and $N^{-\varrho}\{Z^T\hat{\Sigma}^{-1}Z - Z^T\hat{\Sigma}^{-1}X(X^T\hat{\Sigma}^{-1}X)^{-1}X^T\hat{\Sigma}^{-1}Z\}$ respectively.*

We leave more discussions on the key condition (C2) or (C2′) later and end this section by describing how to simulate the null distribution of pseudo LRT using the following Algorithm A.4:

Step 1 Define a grid $0 = \lambda_1 < \lambda_2 < \cdots < \lambda_L$ of possible values for λ.

Step 2 Simulate independent $N(0, \zeta_k)$ random variables w_k, for $k = 1, \ldots, K$.

Step 3 Compute $LRT_\infty(\lambda)$ in (4.9) and determine its maximizer $\lambda_{\max}$ on the grid.

Step 4 Compute $pLRT = LRT_\infty(\lambda_{\max}) + \sum_{j=1}^{\tau}\nu_j^2$, where ν_j's are i.i.d. $N(0,1)$.

Step 5 Repeat Steps 2–4.

4.2 Pseudo LRT for Functional Data

In this section, we will apply the pseudo LRT approach described in Section 4.1 for global inference for functional data. As we mentioned above, the pseudo LRT approach requires that the working covariance satisfying condition (C2) or (C2′). First, we will start with a brief overview on functional data and discuss its covariance structure.

Functional data can be modelled as noisy repeated measurements from a collection of smooth curves. Without loss of generality, we assume that the domain is a closed time interval $\mathcal{T} = [0, 1]$. The model is written as

$$y_i(t_{ij}) = \mu(t_{ij}) + \eta_i(t_{ij}) + \varepsilon_i(t_{ij}), \tag{4.11}$$

where the noises $\varepsilon_i(t_{ij})$ are assumed to be i.i.d. $N(0, \sigma_\varepsilon^2)$ distributed and they are independent of the random subject-specific deviation $\eta_i(t)$. More discussions on how to model $\eta_i(t)$'s will be provided later.

There are several approaches that investigate whether the subject-specific deviation $\eta_i(t)$ exists such as Guo (2002) and Antoniadis and Sapatinas (2007). Though these approaches differ in their choices of the basis functions to model $\eta_i(t)$, they all use the LMM framework and relate the problem to LRT for one variance component. Provided that the subject-specific deviation exists, there are few discussions on global inference of the overall mean function $\mu(t)$. Greven et al. (2008) consider this problem. They suggest subtracting $y_i(t)$ from an estimated $\hat{\eta}_i(t)$, treat the rest as if generated from $\mu(t) + \varepsilon_i(t_{ij})$ and perform LRT. Their approach requires that $\hat{\eta}_i(t)$ approximates $\eta_i(t)$ well enough for all i. But they did not justify what estimators of $\eta_i(t)$'s are good enough.

In contrast to Greven et al. (2008) approach, we consider the inferential

problem of $\mu(t)$ by applying pseudo LRT described in section 4.1. Let $e_i(t) = \eta_i(t) + \varepsilon_i(t)$ and $e_{ij} = e_i(t_{ij})$. We can rewrite the functional data model (4.11) as $Y = X\beta + Zu + e$. Now we will take a closer look at the term e and discuss its covariance Σ. Follow the idea of functional principal component analysis (PCA), we express $\eta_i(t)$ using the Karhunen-Loève representation, i.e. $\eta_i(t) = \sum_k \gamma_{ik}\theta_k(t)$, where $\theta_k(t)$'s are the orthogonal eigenfunctions with unit L^2 norms, and γ_{ik}'s are independent normally distributed with variance σ_k^2. These random variables γ_{ik}'s are also called principal component scores and the associated variance σ_k^2's are the eigenvalues of the covariance function $\Gamma(t,t') = \sum_k \sigma_k^2\theta(t)\theta(t')$. Note that the jth element in e_i is

$$e_i(t_{ij}) = \eta_i(t_{ij}) + \varepsilon_i(t_{ij}) = \sum_k \gamma_{ik}\theta_k(t_{ij}) + \varepsilon_i(t_{ij}). \tag{4.12}$$

Hence Σ, the variance of $(e_1, \ldots, e_n)$, is a block diagonal matrix and the (j,j')th element of its ith block Σ_i is

$$\mathrm{cov}\{e_i(t_{ij}), e_i(t_{ij'})\} = \Gamma(t_{ij}, t_{ij'}) + \sigma_\varepsilon^2 I(t_{ij} = t_{ij'}). \tag{4.13}$$

In the FDA literature, there are already extensive discussions on how to estimate σ_ε^2 and the covariance function $\Gamma(t,t')$, or equivalently, all of its eigenvalues σ_k^2's and eigenfunctions $\theta_k(t)$, for example, see Rice and Silverman (1991), Yao et al. (2005). In the coming subsections, we will further justify under what conditions, the working covariance based on these estimates can satisfy condition (C2) or (C2′). Since the estimating procedures differ from each other when the data are densely observed or sparsely observed, we will separately discuss these two situations.

4.2.1 Dense setting

Dense design is the classical setting for functional data, where a sufficiently large number of regularly spaced observations over individuals are available or can be obtained after pre-processing or alignment. These observations are often recorded by high frequency machine and can be modelled by equation (4.11). Specifically, we assume that each subject is observed at common time points, i.e. $t_{ij} = (j - 1/2)/m$ for $j = 1, \ldots, m$ and $m \to \infty$.

As explained in subsection 4.2, we can conduct pseudo LRT for global inference on $\mu(t)$ under the framework of LMM. This requires to construct a working covariance based on estimators of the variance σ_ε^2, all eigenvalues σ_k^2's and eigenfunctions $\theta_k(t)$'s. Since the observation time points are the same for all subjects, we can consider condition (C2') instead of (C2).

PROPOSITION 4.2.1. *Assume the following conditions for model (4.11):*

(C4) There are m common design points $t_1, \ldots, t_m$ and $m \sim C_1 n^\delta$ for some positive constants C_1 and δ.

(C5) The variance $\sigma_\varepsilon^2 > 0$. The mean function $\mu(t)$ and the covariance function $\Gamma(t, t')$ have continuous second derivatives. Moreover, $\Gamma(t, t')$ has $M \geq 1$ distinct eigenvalues that are positive.

(C6) Assume that the estimated covariance converges at the rate of n^α,

$$\hat{\sigma}_k^2 - \sigma_k^2 = O_p(n^{-\alpha}), \quad \text{and} \quad \sup_{t \in \mathcal{T}} |\hat{\theta}_k(t) - \theta_k(t)| = O_p(n^{-\alpha}),$$

for all k and $\hat{\sigma}_\varepsilon^2 - \sigma_\varepsilon^2 = O_p(n^{-\alpha})$. Moreover, α satisfies

$$\alpha > \delta/2. \tag{4.14}$$

Then condition (C2') holds. Further assume conditions (C1) and (C3) with $\varrho = 1$. Then the pseudo likelihood ratio test statistic $pLRT_N$ has the same asymptotic null distribution as in Proposition 4.1.1.

REMARK 4.2.1. *Condition (C5) guarantees that the covariance function $\Gamma(t, t')$ is smooth, its eigenfunctions are identified, and Σ is invertible. In practice, M is unknown and can be chosen using AIC or BIC. As shown in Yao et al. (2005), in many simulations, the number of eigenvalues M can be correctly chosen. The key assumption in Proposition 4.2.1 is condition (C6). Equation (4.14) indicates that α must exceed some lower bound so that the working covariance can approximate the true Σ well enough. Equation (4.14) can also be written as $\delta < 2\alpha$. According to Hall et al. (2006), $\alpha \leq 1/2$. Hence m has an upper bound such that $m = o(n)$.*

REMARK 4.2.2. *In practice, one may have missing observations per subject. Assume that the missingness is at random. Define $\bar{Y}$ by letting its jth component be the average of all observed y_{ij} and use the model $\bar{Y} = X\beta + Zu + \bar{e}$, where X and Z are the design matrices based on $t_1, \ldots, t_m$. Let n_j be the number of non missing y_{ij}. Our conclusion still holds if $n_j/n \to 1$ for all j.*

Condition (C5) imposes sparseness on the eigenvalues of the covariance. Under condition (C5), we can construct $\hat{\Sigma}$ using the estimated eigenvalues and eigenfunctions obtained from functional PCA, and doing this helps us control $||\hat{\Sigma}^{-1} - \Sigma^{-1}||$ when $m \to \infty$ at a fast rate. It is more important to bound $||\hat{\Sigma}^{-1} - \Sigma^{-1}||$ than $||\hat{\Sigma} - \Sigma||$ itself, because quantities in the pseudo likelihood function have the form $a^T\hat{\Sigma}^{-1}b$ and we need them to converge to $a^T\Sigma^{-1}b$. If we do not assume (C5) and choose the sample covariance $\hat{\Sigma}_{Sam}$ as the working covariance, then standard perturbation theory [see Fan *et al.* (2007)] implies that

$$a^T(\hat{\Sigma}_{Sam}^{-1} - \Sigma^{-1})a = o_p\left(m^2\sqrt{\log(n)/n}\right)$$

for any $||a|| = O(1)$, so we require $m = o(n^{1/4})$. In contrast, assuming (C5) and using PCA, $a^T(\hat{\Sigma}^{-1} - \Sigma^{-1})a = o_p(1)$ even when m converges to infinity faster than $n^{1/4}$. In the simulation results, $\hat{\Sigma}$ based on PCA performs better than $\hat{\Sigma}_{Sam}$. It is possible that (C5) could be weakened by applying PCA with the number of eigenvalues increasing with n, but this is beyond the scope of this thesis and might not be too relevant to practice where the number of eigenvalues is usually kept rather small.

4.2.2 Sparse setting

Another important branch in FDA is to consider observations that are sparsely obtained. Such data can be found in e-commerce and auction bid prices (Jank and Shmueli, 2006), growth data (James et al., 2000) and some other social and life science studies. We still model these data using model (4.11). However, each subject is assumed to have a finite number of observations, i.e. m_i is bounded, and the observation time points vary from subject to subject. We can view the sparsely observed functional data as incomplete observations (with lots of missing values) from the complete data. For example, when the observation points are uniformly distributed, we can round them to the nearest $t_k = (k - 1/2)/m$, and viewed them as if they are sampled uniformly without replacement from $(t_1, \ldots, t_m)$ for some m that converges to infinity. Provided that m converges to infinity fast enough, these two sampling methods are asymptotically equivalent.

Similar as subsection 4.2, we can conduct pseudo LRT for global inference on $\mu(t)$ under the framework of LMM. Since the covariance function for the

sparse functional data still satisfies equation (4.13), the working covariance in the pseudo LRT approach can be constructed in the same way once the covariance estimation $\hat{\Gamma}(t, t')$ is obtained. We have the following conclusions.

PROPOSITION 4.2.2. *Assume the following conditions for model (4.11):*

(C4′) *There are m_i observations for subject i and $\max_{1\le i\le n} m_i$ is bounded as $n \to \infty$. The observation points are generated uniformly without replacement from $(t_1, \ldots, t_m)$, where $t_k = (k - 1/2)/m$ and $m \sim n^\delta$ for some positive δ.*

(C5′) *The variance $\sigma_\varepsilon^2 > 0$. The mean function $\mu(t)$ and the covariance function $\Gamma(t, t')$ have continuous second derivatives. Moreover, $\Gamma(t, t')$ has $M \ge 1$ distinct eigenvalues that are positive.*

(C6′) *Assume that the estimated covariance converges at the rate of n^α, i.e.,*

$$\hat{\sigma}_k^2 - \sigma_k^2 = O_p(n^{-\alpha}), \quad \text{and} \quad \sup_{t\in\mathcal{T}} |\hat{\theta}_k(t) - \theta_k(t)| = O_p(n^{-\alpha}),$$

for all k and $\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = O_p(n^{-\alpha})$. Moreover, α satisfies $\alpha > \delta/2$.

Then condition (C2) holds. Further assume condition (C1) and (C3) with $\varrho = 1$. Then the pseudo likelihood ratio test statistic $pLRT_N$ has the same asymptotic null distribution as in Proposition 4.1.1.

REMARK 4.2.3. *Similar as the dense setting, we require that the estimators converge fast enough, i.e. $\alpha > \delta/2$. We can use Yao et al. (2005)'s approach to construct a plug-in covariance estimator that satisfy condition (C6′).*

REMARK 4.2.4. *As mentioned, if the observation points are uniformly distributed, we can round them into the nearest t_k. Because of the smoothness intrinsic to functional data (ignoring the noise), the effect of this rounding is asymptotically negligible when $m \to \infty$ at a rate faster than n^α. If we relax condition (C4′) and assume that the t_{ij}'s are*

uniformly distributed between 0 and 1, Proposition 4.2.2 still holds if the last statement in condition (C6') is strengthened to $\delta/2 < \alpha < \delta$. *Since* $0 < \delta < 2\alpha$, *we cannot allow* m *to grow too fast, e.g.,* $m = o(n^{4/5})$. *Note that this differs from the dense case where the upper bound is* $m = o(n)$. *This is because the fastest convergence rates for estimators of the eigenfunctions in the sparse case and the dense case are* $2/5$ *and* $1/2$ *respectively.*

In summary, for both densely and sparsely observed functional data and under mild conditions, using the estimated covariance achieves the same asymptotic null distribution of the test statistic as using the true covariance. However, in order that the estimation errors in Σ have an asymptotically negligible effect, it is required that m not grow too fast.

4.3 Extensions: Multilevel FDA

A further extension in FDA is to consider longitudinally collected functional data or so-called multilevel FDA. For example, in follow-up or panel studies, we have multiple functional data for the same subject because he/she has follow-up exams/visits. This motivates the idea of multilevel FDA. Recent work on this area includes multilevel functional PCA by Di et al. (2009) and Di and Crainiceanu (2010), multilevel functional regression by Morris et al. (2008), Staicu et al. (2010), and Greven et al. (2010). But none of them consider testing whether developmental trends across visits exist. In this thesis, we address this issue by applying the pseudo LRT approach. Before we discuss the testing procedure, we provide a brief overview on how to model multilevel functional data.

4.3.1 A review on the multilevel FDA model

Let $y_{ijs} = y_{is}(t_{ijs})$ denote the jth observation for subject i at visit s and time t_{ijs}, where $i = 1, \ldots, n$, $j = 1, \ldots, m_{is}$, and $s = 1, \ldots, s_0$. Following Di et al. (2009) and Di and Crainiceanu (2010), we can model them as

$$y_{is}(t_{ijs}) = \mu_s(t_{ijs}) + \eta_i(t_{ijs}) + \Omega_{is}(t_{ijs}) + \epsilon_{is}(t_{ijs}), \tag{4.15}$$

where $\mu_s(t)$ is the visit-specific mean function, $\eta_i(t)$ and $\Omega_{is}(t)$ are the subject-specific and subject/visit-specific deviations respectively, and $\epsilon_{is}(t)$ is the measurement error. We assume that $\epsilon_{is}(t_{ijs})$ are i.i.d. $N(0, \sigma_\epsilon^2)$ and independent of $\eta_i(t)$ and $\Omega_{is}(t)$ that are modelled as below.

By Karhunen-Loève expansion, we have

$$\eta_i(t) = \sum_k \gamma_{ik}^{(1)} \theta_k^{(1)}(t), \quad \Omega_{is}(t) = \sum_l \gamma_{ils}^{(2)} \theta_l^{(2)}(t),$$

where $\theta_k^{(1)}$'s and $\theta_l^{(2)}$'s are level 1 (subject) and 2 (subject/visit) eigenfunctions, $\gamma_{ik}^{(1)}$'s and $\gamma_{ils}^{(2)}$'s are the principal component scores. Assume that $\gamma_{ik}^{(1)} \sim N(0, \sigma_{k1}^2)$ and $\gamma_{ils}^{(2)} \sim N(0, \sigma_{l2}^2)$ and all of them are independent. Denote the level 1 and 2 covariance function as

$$\Gamma_1(t, t') = \sum_k \sigma_{k1}^2 \theta_k^{(1)}(t) \theta_k^{(1)}(t'), \quad \Gamma_2(t, t') = \sum_l \sigma_{l2}^2 \theta_l^{(2)}(t) \theta_l^{(2)}(t').$$

Di et al. (2009) and Di and Crainiceanu (2010) discuss how to estimate σ_ϵ^2, $\Gamma_1(t, t')$, $\Gamma_2(t, t')$ as well as all σ_{k1}^2's, $\theta_k^{(1)}(t)$'s, σ_{l2}^2's, $\theta_l^{(2)}(t)$'s using functional PCA. In this thesis, we focus on testing whether developmental trends exist among different visits, i.e. $\mu_1(t) = \mu_2(t) = \cdots = \mu_{s_0}(t)$. Our tool is to formulate the multilevel functional data under the LMM framework and then apply pseudo LRT.

4.3.2 Test for equality of two curves

For simplicity, we first consider the case when each subject only has two visits and leave the discussions for more than 2 visits in subsection 4.3.3. We are interested in the significance test whether $\mu_{21}(t) =: \mu_2(t) - \mu_1(t) \equiv 0$. We will separately discuss the dense design and sparse design.

First consider the dense design. When the functional data are densely observed, we can subtract the function of visit 2 from that of visit 1 because they are both observed at time $(t_1, \ldots, t_m)$. Define the pairwise difference as $y_{i21}(t_j) = y_{i2}(t_j) - y_{i1}(t_j)$ for $i = 1, \ldots, n$. By model (4.15), we have

$$y_{i21}(t_j) = \mu_{21}(t_j) + \sum_l \gamma^{(2)}_{il21}\theta^{(2)}_l(t_j) + \epsilon_{i21}(t_j) = \mu_{21}(t_j) + e_{i21}(t_j), \tag{4.16}$$

where $\mu_{21}(t) = \mu_2(t) - \mu_1(t)$, $\gamma^{(2)}_{il21} = \gamma^{(2)}_{il2} - \gamma^{(2)}_{il1}$, $\epsilon_{i21}(t) = \epsilon_{i2}(t) - \epsilon_{i1}(t)$, and

$$e_{i21}(t) =: \sum_l \gamma^{(2)}_{il21}\theta^{(2)}_l(t) + \epsilon_{i21}(t). \tag{4.17}$$

After differencing, the variation in the direction of level 1 eigenfunctions is eliminated and the pairwise difference data can be treated as single level functional data.

Therefore, we can adopt the procedures described in subsection 4.2.1. We express the visit difference $\mu_{21}(t)$ using the pth spline polynomials as (4.2), i.e., $\mu_{21}(t) = x_t\beta + z_t u$. Define

$$Y_i = \{y_{i21}(t_1), \ldots, y_{i21}(t_m)\}$$

and similarly define e_i based on $e_{i21}(t_j)$'s. Then we can formulate model (4.16) as $Y = X\beta + Zu + e$. In particular, the variance Σ is a block diagonal matrix and the (j, j')th element of Σ_i equals $2\Gamma_2(t_j, t_{j'}) + 2\sigma^2_\epsilon I_{t_j = t_{j'}}$. Proposition 4.2.1

holds if we replace σ_k^2 by σ_{k2}^2, $\hat{\sigma}_k^2$ by $\hat{\sigma}_{k2}^2$, θ_k by $\theta_k^{(2)}$ and $\hat{\theta}_k$ by $\hat{\theta}_k^{(2)}$ in condition (C6). Therefore, we can apply the pseudo LRT approach to test whether the difference $\mu_{21}(t) \equiv 0$.

Next consider the sparse setting, where the sample curves are only intermittently or sparsely observed and the observation time points for each subject vary from visit to visit. Correspondingly, it is meaningless to consider pairwise differencing. Our idea is to consider the quasi residuals $y_{is}(t_{is}) - \bar{\mu}(t_{is})$, where $\bar{\mu}(t)$ is the average of the estimated mean in both visits, i.e. $\bar{\mu}(t) = \{\hat{\mu}_1(t) + \hat{\mu}_2(t)\}/2$, where $\hat{\mu}_s(t)$ is the mean function obtained by simply smoothing all observations obtained at the sth visit. Yao et al. (2005) derive the convergence rate of $\hat{\mu}_s(t)$. According to Kulasekera (1995), we can rewrite the model (4.15) as

$$\begin{aligned} y_{i1}(t_{ij1}) - \bar{\mu}(t_{ij1}) &= -2^{-1}\mu_{21}(t_{ij1}) + e_{i1}(t_{ij1}), \\ y_{i2}(t_{ij2}) - \bar{\mu}(t_{ij2}) &= 2^{-1}\mu_{21}(t_{ij2}) + e_{i2}(t_{ij2}), \end{aligned} \tag{4.18}$$

where $\mu_{21}(t) = \mu_2(t) - \mu_1(t)$ and $e_{is}(t) = \eta_i(t) + \Omega_{is}(t) + \epsilon_{is}(t)$.

Then we can rewrite model (4.18) under the LMM framework. We still use the pth spline polynomials to express the visit difference $\mu_{21}(t)$. Let T_{is} be the vector formed by all time points t_{ijs}'s for $j = 1, \ldots, m_{is}$. Let $x_{t_{ijs}} = (1, t_{ijs}, \ldots, t_{ijs}^p)$ and $z_{t_{ijs}} = \{(t_{ijs} - \kappa_1)_+, \ldots, (t_{ijs} - \kappa_K)_+\}$. Define X_{is} and Z_{is} based on $x_{t_{ijs}}$ and $z_{t_{ijs}}$. Let

$$Y_i = \{y_{i1}(t_{i11}), \ldots, y_{i1}(t_{im_{i1}1}), y_{i2}(t_{i12}), \ldots, y_{i2}(t_{im_{i2}2})\}^T,$$

and similarly define e_i based on all $e_{is}(t_{ijs})$'s. Let $X_i = (-X_{i1}^T, X_{i2}^T)^T$ and $Z_i = (-Z_{i1}^T, Z_{i2}^T)^T$. Then we can formulate model (4.18) as $Y = X\beta + Zu + e$. In particular, Σ is a block diagonal matrix. Its ith block is of dimension $(m_{i1} + m_{i2}) \times (m_{i1} + m_{i2})$ and can be partitioned as $\begin{pmatrix} \Sigma_{i1} & \Sigma_{i12} \\ \Sigma_{i12}^T & \Sigma_{i2} \end{pmatrix}$, where the

(j, j') element in Σ_{is} is $\Gamma_1(t_{ijs}, t_{ij's}) + \Gamma_2(t_{ijs}, t_{ij's}) + \sigma_\epsilon^2 I_{t_j = t_{j'}}$, and that in Σ_{i12} is $\Gamma_1(t_{ij1}, t_{ij'2})$.

We can follow Di and Crainiceanu (2010) to estimate $\Gamma_1(t, t')$ $\Gamma_2(t, t')$ and σ_ϵ^2. Let $N_s = \sum_{i=1}^n m_{is}$ and $N = N_1 + N_2$. Proposition 4.2.2 holds if we replace σ_k^2 by σ_{ks}^2, $\hat{\sigma}_k^2$ by $\hat{\sigma}_{ks}^2$, θ_k by $\theta_k^{(s)}$ and $\hat{\theta}_k$ by $\hat{\theta}_k^{(s)}$, for $s = 1, 2$, in condition (C6′). Therefore, we can apply the pseudo LRT approach to test whether $\mu_{21}(t) \equiv 0$.

4.3.3 Test for equality of more than 2 curves

When more than 2 visits are available, inference on the developmental trend of the visit-specific mean functions is a multiple testing problem. As an example, we will consider the dense setting with three visits per subject. But the idea can be similarly extended for settings with more visits.

Let $\mu_{21}(t) = \mu_2(t) - \mu_1(t)$ and $\mu_{32}(t) = \mu_3(t) - \mu_2(t)$. We are interested in testing whether $\mu_{21}(t) = \mu_{32}(t) = 0$. We still use the pth spline polynomials and model $\mu_{21}(t) = x_t\beta_{[1]} + z_t u_{[1]}$ and $\mu_{32}(t) = x_t\beta_{[2]} + z_t u_{[2]}$. Now we want to relate the problem with tests under the linear mixed model framework. In the dense setting, we can calculate $y_{i21}(t_j) = y_{i2}(t_j) - y_{i1}(t_j)$ and $y_{i32}(t_j) = y_{i3}(t_j) - y_{i2}(t_j)$. Let $Y_i = \{y_{i21}(t_1), \ldots, y_{i21}(t_m), y_{i32}(t_1), \ldots, y_{i32}(t_m)\}$, and let $Y_{\mathbf{b}} = n^{-1/2} \sum_{i=1}^n Y_i$. Similarly, we can define $e_{i21}(t_j)$'s as well as $e_{i32}(t_j)$'s, and construct e_i and $e_{\mathbf{b}}$. Let X and Z be the $m \times (p+1)$ and $m \times K$ matrices whose jth row are x_{t_j} and z_{t_j} respectively. Let $X_{\mathbf{b}} = I_2 \otimes X$ and $Z_{\mathbf{b}} = I_2 \otimes Z$, where I_2 is the 2×2 identity matrix. Then we can write the model as $Y_{\mathbf{b}} = X_{\mathbf{b}}\beta + Z_{\mathbf{b}} u + e$, where

$\beta = (\beta_{[1]}^T, \beta_{[2]}^T)^T$, and $u = (u_{[1]}^T, u_{[2]}^T)^T$ satisfying

$$\begin{pmatrix} u_{[1]} \\ u_{[2]} \\ e \end{pmatrix} \sim N \left\{ \mathbf{0}_{(2K+N)\times 1}, \begin{pmatrix} \lambda_{[1]} I_K & \mathbf{0}_{K\times K} & \mathbf{0}_{K\times N} \\ \mathbf{0}_{K\times K} & \lambda_{[2]} I_K & \mathbf{0}_{K\times N} \\ \mathbf{0}_{N\times K} & \mathbf{0}_{N\times K} & \Sigma \end{pmatrix} \right\},$$

for some $\lambda_{[1]} \geq 0$, $\lambda_{[2]} \geq 0$, and Σ positive definite. To test whether $\mu_{21}(t) = \mu_{32}(t) = 0$. We define the null hypothesis as

$$H_0 : \beta_q = 0 \quad \text{for } q = 0, \ldots, 2p+1 \qquad \text{and} \quad \lambda_{[1]} = 0 \quad \text{and } \lambda_{[2]} = 0, \tag{4.19}$$

and the composite alternative

$$H_A : \exists\, q_0 \quad \text{such that} \quad \beta_{q_0} \neq 0 \quad \text{or} \quad \lambda_{[1]} > 0 \quad \text{or } \lambda_{[2]} > 0. \tag{4.20}$$

Given Σ is known, twice the log-likelihood of $Y_{\mathbf{b}}$ equals

$$-\log\left|\Sigma + Z_{\mathbf{b}}\Sigma_u Z_{\mathbf{b}}^T\right| - (Y_{\mathbf{b}} - X_{\mathbf{b}}\beta)^T(\Sigma + Z_{\mathbf{b}}\Sigma_u Z_{\mathbf{b}}^T)^{-1}(Y_{\mathbf{b}} - X_{\mathbf{b}}\beta),$$

where $\Sigma_u = \begin{pmatrix} \lambda_{[1]} I_K & \mathbf{0}_{K\times K} \\ \mathbf{0}_{K\times K} & \lambda_{[2]} I_K \end{pmatrix}$. The pseudo log-likelihood is defined as

$$2\log \hat{L}_{\hat{Y}_{\mathbf{b}}}(\beta, \lambda_{[1]}, \lambda_{[2]}) \;=\; -\log|\hat{H}_{\lambda_{[1]},\lambda_{[2]}}| - (\hat{Y}_{\mathbf{b}} - \hat{X}_{\mathbf{b}}\beta)^T \hat{H}^{-1}_{\lambda_{[1]},\lambda_{[2]}}(\hat{Y}_{\mathbf{b}} - \hat{X}_{\mathbf{b}}\beta),$$

where $\hat{Y}_{\mathbf{b}}$, $\hat{X}_{\mathbf{b}}$ and $\hat{Z}_{\mathbf{b}}$ are defined as in (4.7), $\hat{H}_{\lambda_{[1]},\lambda_{[2]}} = I_N + \hat{Z}_{\mathbf{b}}\Sigma_u\hat{Z}_{\mathbf{b}}^T$, where N is the length of Y. The pseudo LRT statistic to test (4.19) versus (4.20) is defined as

$$pLRT_N = \sup_{H_0 \cup H_A} 2\log \hat{L}_{\hat{Y}_{\mathbf{b}}}(\beta, \lambda_{[1]}, \lambda_{[2]}) - \sup_{H_0} 2\log \hat{L}_{\hat{Y}_{\mathbf{b}}}(\beta, \lambda_{[1]}, \lambda_{[2]}).$$

Similar as Proposition 4.1.1, we have the following results.

PROPOSITION 4.3.1. *Assume the following conditions:*

(M1) The null hypothesis H_0 defined in (4.4) is true.

(M2) The covariance matrix $\Sigma = \mathrm{cov}(e_{\mathbf{b}})$ *in equation (4.3) is positive definite. Let* $\hat{\Sigma}$ *be its consistent estimate satisfying that*

$$a^T\hat{\Sigma}^{-1}a - a^T\Sigma^{-1}a = o_p(1) \qquad a^T\hat{\Sigma}^{-1}e_{\mathbf{b}} - a^T\Sigma^{-1}e_{\mathbf{b}} = o_p(1), \tag{4.21}$$

where a *is any* $(2m) \times 1$ *non random normalized vector.*

(M3) There exists positive constants ϱ' *and* ϱ *such that* $N^{-\varrho'}X_{\mathbf{b}}^T X_{\mathbf{b}}$ *and* $N^{-\varrho}Z_{\mathbf{b}}^T Z_{\mathbf{b}}$ *converge to nonzero matrices. Moreover,* $N^{-\varrho}\Sigma_u\tilde{Z}_{\mathbf{b}}^T\tilde{Z}_{\mathbf{b}}$ *and* $N^{-\varrho}\tilde{Z}_{\mathbf{b}}^T\tilde{Z}_{\mathbf{b}} - \tilde{Z}_{\mathbf{b}}^T\tilde{X}_{\mathbf{b}}(\tilde{X}_{\mathbf{b}}^T\tilde{X}_{\mathbf{b}})^{-1}\tilde{X}_{\mathbf{b}}^T\tilde{Z}_{\mathbf{b}}$ *converge to two nonzero matrices* A_1 *and* A_2 *resepctively. Let* $\xi_k(\lambda_{[1]}, \lambda_{[2]})$ *be the* k*th eigenvalues of* $\Sigma_u A_1$*, and* $\zeta_k(\lambda_{[1]}, \lambda_{[2]})$ *be the* k*th eigenvalues of* $\Sigma_u A_2$*.*

Denote τ *the cardinality of the set* Q *in the null hypothesis (4.4). Then*

$$pLRT_N \Rightarrow \sup_{\lambda \geq 0} LRT_\infty(\lambda) + \sum_{j=1}^{\tau} \nu_j^2, \tag{4.22}$$

$$LRT_\infty(\lambda_{[1]}, \lambda_{[2]}) = \sum_{k=1}^{2K} \frac{w_k^2(\lambda_{[1]}, \lambda_{[2]})}{1 + \zeta_k(\lambda_{[1]}, \lambda_{[2]})} - \sum_{k=1}^{2K} \log\{1 + \xi_k(\lambda_{[1]}, \lambda_{[2]})\},$$

$w_k(\lambda_{[1]}, \lambda_{[2]}) \sim N\{0, \zeta_k(\lambda_{[1]}, \lambda_{[2]})\}$ *for* $k = 1, \ldots, 2K$*,* $\nu_j \sim N(0,1)$ *for* $j = 1, \ldots, \tau$*, and all* w_k*'s* ν_j*'s are mutually independent.*

REMARK 4.3.1. *The eigenvalues* $\xi_k(\lambda_{[1]}, \lambda_{[2]})$*'s of* $\Sigma_u A_1$ *are real nonnegative. This is because they are also the eigenvalues of the semi-positive definite matrix* $\Sigma_u^{1/2} A_1 \Sigma_u^{1/2}$*. Similarly, all* $\zeta_k(\lambda_{[1]}, \lambda_{[2]})$*'s are real nonnegative eigenvalues.*

REMARK 4.3.2. *Similarly as Proposition 4.2.1, it can be shown that (M2) holds if condition (C4)–(C6) holds.*

When simulating its null distribution in (4.22), we need to search over $(\lambda_{[1]}, \lambda_{[2]})$. For each pair $(\lambda_{[1]}, \lambda_{[2]})$, we need to calculate the eigenvalues of

$\Sigma_u \hat{Z}_{\mathbf{b}}^T \hat{Z}_{\mathbf{b}}$ and $\Sigma_u \{\hat{Z}_{\mathbf{b}}^T \hat{Z}_{\mathbf{b}} - \hat{Z}_{\mathbf{b}}^T \hat{X}_{\mathbf{b}} (\hat{X}_{\mathbf{b}}^T \hat{X}_{\mathbf{b}})^{-1} \hat{X}_{\mathbf{b}}^T \hat{Z}_{\mathbf{b}}\}$ unless we further assume that $\lambda_{[1]} = \lambda_{[2]}$. It takes a longer time but the computation is not very expensive since we are only dealing with $(2K) \times (2K)$ dimensional matrix.

4.4 Simulations

We conduct simulation studies to evaluate finite sample performance of the pseudo LRT for functional data. Consider the single level functional data, which are generated from the model

$$y_i(t_{ij}) = \mu(t_{ij}) + \sum_{k=1}^{d} \gamma_{ik} \theta_k(t_{ij}) + \epsilon_{ij}. \tag{4.23}$$

We let $d = 3$, $\theta_1(t) = \sqrt{2}\cos(2\pi t)$, $\theta_2(t) = \sqrt{2}\sin(2\pi t)$, $\theta_3(t) = \sqrt{2}\cos(4\pi t)$. Assume that $\gamma_{ik} \sim N(0, \sigma_k^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$ and they are independent of each other. We set $\sigma_1^2 = 1$, $\sigma_2^2 = 0.5$ and $\sigma_3^2 = 0.25$. We consider two choices of the magnitude of noise: $\sigma_\epsilon^2 = 0.125$ (small) and $\sigma_\epsilon^2 = 2$ (large).

4.4.1 Simulation 1

We first consider the significance test of $\mu(t) \equiv 0$. We assume that the data are densely observed and the observations are recorded at the same time for all subjects. We set $n = 200$ subjects and each subject has $m = 80$ equally spaced observations $(j - 1/2)/m$, for $j = 1, \ldots, m$. For each subject, we create missing values by randomly throwing out 10 observations. We model $\mu(t)$ using pth spline polynomials with K equally spaced knots. We consider $p = 1, 3$ and $K = 40$. We consider three choices of the working covariance in the pseudo

LRT approach, where 'True' indicates the true covariance, 'Sam' stands for the sample covariance, and 'SMOOTH' is the estimation obtained by smoothing the sample covariance.

We use the Algorithm A.4 described below Proposition 4.1.1 to simulate the null distribution of the test statistic. To be specific, for one set of functional data $\{t_{ij}, y_i(t_{ij})\}$, $i = 1, \ldots, n$ and $j = 1, \ldots, m$, generated from model (4.23), we repeat Steps 2–4 100,000 times and then calculate the sample quantiles of the simulated pLRT. Similar as Crainiceanu and Ruppert (2004), we find that the maximizer of λ in (4.9) is 0 in more than 99% of the simulations. Correspondingly, the quantiles of the simulated test statistic is practically a χ^2_τ distribution, with τ being $p + 1$ because the coefficients $\beta_0 = \cdots = \beta_p = 0$ under the null.

Table 4.1 and 4.2 report the type I error of the pseudo LRT approach based on 1000 independent experiments with $\sigma^2_\epsilon = 2$ and $\sigma^2_\epsilon = 0.125$ respectively. It seems that the test statistic based on smooth covariance estimation performs similarly as that using the true covariance. In both cases, the type I errors are close to the nominal value regardless whether we model $\mu(t)$ using linear or cubic splines. However, the pseudo LRT approach using the sample covariance estimation failed. This finding is consistent with conclusions at the end of Section 4.2.1. Since the size of the test is seriously distorted when sample covariance is used, we will no longer consider using it as the working covariance in our simulations hereafter.

To calculate the power of the pseudo LRT test, we choose $\mu(t)$ in model (4.23) from two families of functions. For each family, a scalar $\tilde{\rho}$ controls the degree of departure from H_0 with $\tilde{\rho} = 0$ corresponding to H_0. The first family consists of

Table 4.1: Type I error of testing $\mu(t) \equiv 0$ based on 1000 experiments with $\sigma_\epsilon^2 = 2$ and linear/cubic polynomial splines.

Choices of Cov	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Sam	0.64 /0.726	0.55/0.592	0.432/0.518	0.278/ 0.310
Smooth	0.210 /0.189	0.119 /0.100	0.062 /0.050	0.014/ 0.016
True	0.214/ 0.196	0.116 /0.108	0.066 /0.052	0.011 /0.014

Table 4.2: Type I error of testing $\mu(t) \equiv 0$ based on 1000 experiments with $\sigma_\epsilon^2 = 0.125$ and linear/cubic polynomial splines.

Choices of Cov	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Sam	0.624/0.704	0.496/0.586	0.388/0.478	0.256/0.296
Smooth	0.206/0.175	0.113/0.097	0.066/0.048	0.011/0.018
True	0.206/0.202	0.118/0.116	0.066/0.052	0.008/0.017

increasing functions

$$\mu_{\tilde{\rho}}(x) = \tilde{\rho}/\{1 + e^{10(0.5-x)}\} - \tilde{\rho}/2 \tag{4.24}$$

and the second family consists of concave functions

$$\mu_{\tilde{\rho}}(x) = -\tilde{\rho}|0.5 - x|^{2.5}. \tag{4.25}$$

Figure 4.1 below reports the power plot of the pseudo LRT approach based on 200 simulations with $\sigma_\epsilon^2 = 2$. We model $\mu(t)$ by either the linear or cubic spline polynomials, i.e. $p = 1$ or $p = 3$. We compare the pseudo LRT approach with working covariance specified by the smooth covariance estimate and the true covariance. As γ increases, the deviation of $\mu_\gamma(t)$ from 0 is more obvious and we can see that the power of the pseudo LRT test also increases. However, we

do not find significant differences between using a smooth covariance estimate and the true covariance, or significant differences between using linear spline polynomials and cubic ones.

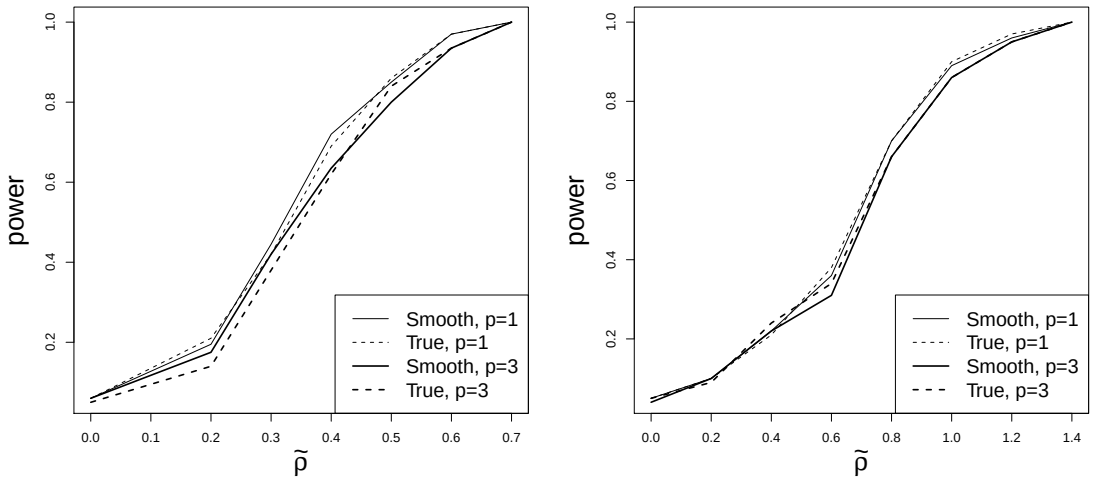

Figure 4.1: The power plot of significance test $\mu(t) \equiv 0$ based on 200 experiments.

4.4.2 Simulation 2

Next we consider testing whether $\mu(t)$ is a linear function or a general alternative. We let $\mu(t) = -t$ and use model (4.23) to generate sparsely observed data . Unlike the dense setting, we now assume that there are $n = 400$ subjects. Each subject has m_i observations, where m_i is randomly chosen from the discrete uniform distribution on $\{1, \ldots, 4\}$. The observation time points t_{ij} for $j = 1, \ldots, m_i$ are uniformly distributed.

We model $\mu(t)$ using cubic spline polynomials with K equally spaced knots.

We report the results based on $K = 40$. We consider two choices of the working covariance in the pseudo LRT approach, where 'True' indicates the true covariance and 'SMOOTH' is the estimation obtained by smoothing the sample covariance. Similar as the dense design, we find that the quantiles of simulated LRT using Algorithm A.4 is very close to that of the χ^2_τ distribution, where $\tau = 2$ because $\beta_2 = \beta_3 = 0$ under the null.

Table 4.3 reports the type I error after 1000 experiments. To calculate the power, we also consider two families of functions and let $\mu(t)$ be $-t + \mu_{\tilde{\rho}}(t)$, where $\mu_{\tilde{\rho}}(t)$ is specified from either the increasing family (4.24) or the concave family (4.25). Figure 4.2 provides the power plot based on 200 simulations. Under the sparse design, we have relatively small sample size and hence $\tilde{\rho}$ is larger than those in Figure 4.1. However, we do not find significant differences between using smooth covariance estimation and the true covariance.

Table 4.3: Type I error of testing a linear function versus a general alternative based on 1000 experiments with $\sigma^2_\epsilon = 2$ ($\sigma^2_\epsilon = 0.125$) and cubic spline polynomials.

Cov choice	$\alpha = 0.2$	$\alpha = 0.1$	$\alpha = 0.05$	$\alpha = 0.01$
Smooth	0.212 (0.195)	0.114 (0.094)	0.060 (0.049)	0.021 (0.011)
True	0.197 (0.194)	0.111 (0.091)	0.057 (0.041)	0.010 (0.009)

4.5 Sleep Heart Health Study

The sleep health heart study (SHHS) is a large-scale comprehensive multi-site study of sleep and its correlation with health outcomes. The principal aim of

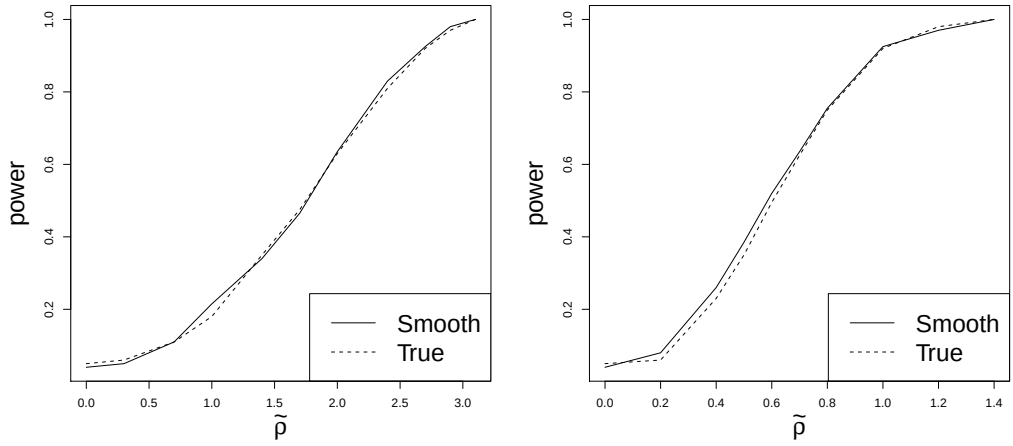

Figure 4.2: The power plot of testing a linear function versus a general alternative based on 200 experiments.

this study is to learn about the association between sleep and a variety of health-related outcomes such as hypertension and cardiovascular disease (CVD). The SHHS include the following: the Framingham Offspring and Omni Cohort Studies, the Atherosclerosis Risk in Communities Study (ARIC), the Cardiovascular Health Study (CHS), the Strong Heart Study, and the Tucson Epidemiologic Study of Respiratory Disease. Detailed descriptions of the SHHS can be found in Quan et al. (1997), Crainiceanu et al. (2009) and Di et al. (2009).

To acquire the sleep exposure variables, subjects underwent two in-home polysomnograms (PSGs), one at a baseline visit and one at a second visit, approximately five years later. A PSG is a quasi-continuous multi- channel recording of physiological signals acquired during sleep that include the following: two surface electroencephalograms (EEG), right and left electroocu- lograms for recording eye movements, leg and submentalis electromyograms, a precordial electrocardiogram, oxyhemoglobin saturation by pulse oximetry, and thoraco-

abdominal movement with plethysmography. In this section, we consider the electroencephalograms (EEG) of a sample of 500 patients.

To reduce the size of the data set, we follow the descriptions in Crainiceanu et al. (2009) and Di and Crainiceanu (2010) and transforming the EEG data to the frequency space. After Discrete Fourier Transform (DFT) and normalization, the raw EEG were summarized to the normalized power in some given frequency bands in each 30 second intervals (Crainiceanu et al., 2009). We are particularly interested in the $\delta-$power, which represents the low frequency (0.8–4.0 Hz) neuronal activity. The δ-power is a desired predictor for adverse health outcome and it reflects the homeostatic need for sleep (Borbely and Achermann, 1999). Lower values in δ-power might indicate difficulty in relaxing during sleep. We are interested in testing whether there are differences in the mean δ-power function between these two visits and thus answer how aging affects the mean δ-power function.

Let y_{ijs} denote the observed $\delta-$power for subject i at time t_j in visit s. We use the multilevel FPCA (4.15) and our model is

$$y_{ijs} = y_{is}(t_j) = \mu_s(t_j) + \eta_i(t_j) + \Omega_{is}(t_j) + \epsilon_{ijs}, \tag{4.26}$$

where $\mu_s(t)$ is the visit-specific mean function, $\eta_i(t)$ and $\Omega_{is}(t)$ are the subject-specific and subject/visit-specific deviations respectively, and $\epsilon_{is}(t)$ is the measurement error. Let $y_{ij} = y_{ij2} - y_{ij1}$ be the difference of delta power between two visits. Then

$$y_{ij} = \mu(t_j) + \Omega_i(t_j) + \epsilon_{ij}, \tag{4.27}$$

where $\mu(t) = \mu_2(t) - \mu_1(t)$ and $\Omega_i(t) = \Omega_{i2}(t) - \Omega_{i1}(t)$ and $\epsilon_{ij} = \epsilon_{ij2} - \epsilon_{ij1}$. We focus on the $\delta-$power associated with the first 4-hour sleep in each visit. Our

null hypothesis is $H_0 : \mu_1(t) \equiv \mu_2(t)$, or equivalently,

$$H_0 : \mu(t) \equiv 0, \text{ for } 0 < t < 4. \tag{4.28}$$

We conduct the pseudo LRT for this global inferential problem. We use linear splines to model the difference function $\mu(t)$. We follow the idea of the first smoothing then estimation methodology to estimate the within subject covariance caused by $\Omega_i(t)$; see also Fan and Zhang (2000), Hall et al. (2006). As suggested by AIC, we select 14 eigenfunctions. We also try using only 7 eigenfunctions, which is the minimum number of eigenfunctions that explain more than 95% of variations of the eigenvalues. But the results are quite similar and hence not reported here. According to subsection 4.2.1, we calculate $\bar{y}_j$'s, the average of y_{ij}'s across subjects for each j, and use their loglikelihood to derive the test statistics. Note that subjects might wake up at night, not all δ-power y_{ijs}'s are observed and we have y_{ij}'s that are undefined or not available. We introduce the indicator $\iota_i(t)$, which equals 1 if $y_{is}(t_j)$ are observed for both visits, and equals 0 otherwise.

We consider two approaches to calculate the average $\bar{y}_j$.

Approach 1–Missing at random (MAR): The idea is to treat missingness as uninformative and only consider available responses over all individuals and define the average as

$$\bar{y}_j^{[1]} = n_j^{-1} \sum_{i=1}^{n} y_{ij} \iota_i(t_j), \ j = 1, \ldots, m, \tag{4.29}$$

where $n_j = \sum_{i=1}^{n} \iota_i(t_j)$.

Approach 2–Nonignorable missing (NIM): The idea is to create simulated data to replace the missing data. The simulated data are obtained by the algorithm below.

1 Let $f_i(t) = \mu(t) + \Omega_i(t)$. We smooth each individual curve to obtain an estimate $\hat{f}_i(t)$.

2 Let $\sigma_\epsilon^2 = \mathrm{var}(\epsilon_{ij})$. We obtain its estimate $\hat{\sigma}_\epsilon^2$ using method of moments.

3 We simulate $\hat{y}_{ij} = \hat{f}_i(t) + \hat{\sigma}_\epsilon w_{ij}$, where w_{ij} is generated independently from $N(0, 1)$.

4 We define $\tilde{y}_{ij} = y_{ij}\iota_i(t_j) + \hat{y}_{ij}\{1 - \iota_i(t_j)\}$.

In contrast to the MAR approach, we define the average based on the augmented complete data set, i.e.,

$$\bar{y}_j^{[2]} = n^{-1} \sum_{i=1}^{n} \tilde{y}_{ij},\ j = 1, \ldots, m.$$

The test statistics for the null hypothesis in (4.28) are 16.8 and 35.6 for the MAR and NIM approaches respectively. Although the test statistics are different, they are both much greater than the critical value 9.4 at the level 0.01, which is obtained based on 100000 simulation using algorithm A.4. Hence we reject H_0 and conclude that there are differences between the mean δ-power function in both visits. Plots of the estimating mean difference based on the MAR and NIM approaches are provided in Figure 4.3. Following the idea of Chapter 6.4 in Ruppert et al. (2003) with modifications to account for the covariance structure induced by $\Omega_i(t)$'s, we construct a pointwise confidence interval band for $\hat{\mu}(t)$ in Figure 4.4. Note that penalized splines estimators has slower convergence rate at the boundary. We truncate the first and last 15 minutes. It seems that for the first three hours, both MAR and NIM approaches find that the mean δ-power function decreases in the second visit, roughly five years after the baseline visit. However, we see different patterns with respect to the time beyond 3 hours.

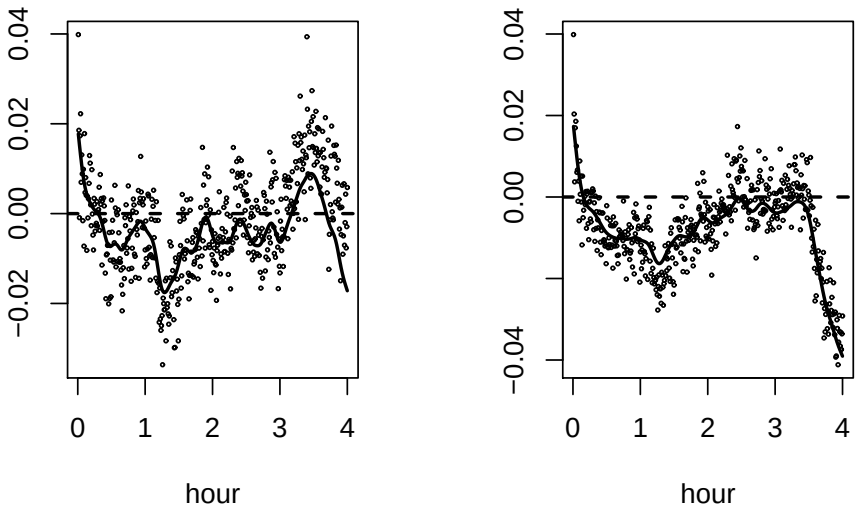

Figure 4.3: Plots of estimated difference function (the solid line) using the MAR (left) and NIM (right) approaches.

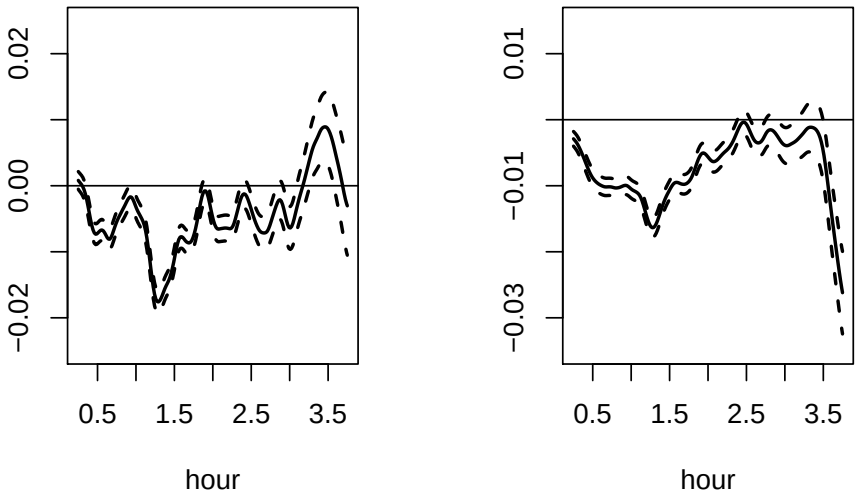

Figure 4.4: Pointwise Confidence Intervals for the difference function using the MAR (left) and NIM (right) approaches.

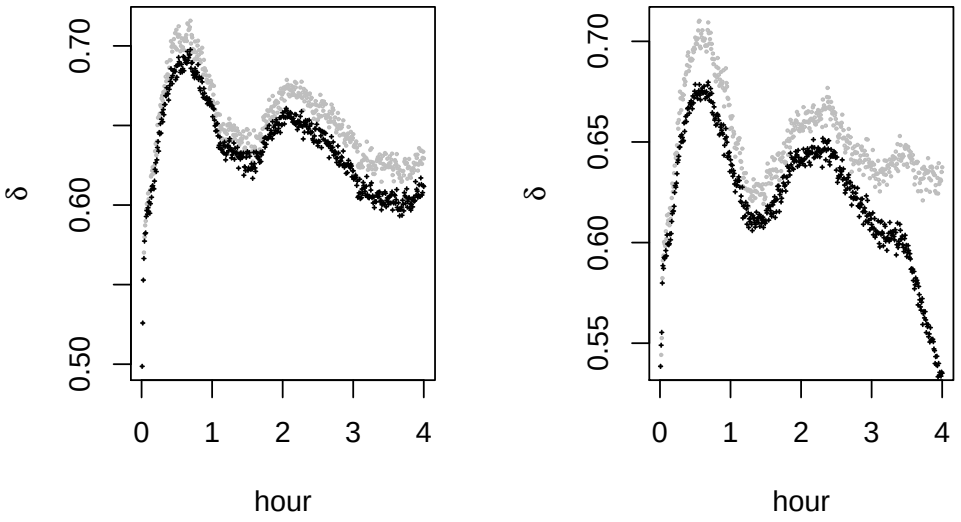

Figure 4.5: Comparisons of the MAR (grey points) and NIM (dark points) based on average responses from visit 1 and 2.

This motivates us to further compare the MAR and NIM approaches. We apply the idea of the NIM approach and simulate δ-power for unobserved y_{ijs}'s. Figure 4.5 δ-power using MAR and NIM approach. It seems that the simulated δ-power tends to be smaller and the NIM approach has a strong decrease near the end of the first 4 hours in the second plot, which might imply that more people woke up near the end of the first 4 hours compared with the first visit. We think that the data suggest $y_i(t)$ is not missing at random. The lower readings of δ-power are more likely to be censored than higher readings since the lower δ-power is associated with waking.

Since the MAR approach ignores this additional information, it might over estimate $\hat{\mu}_2(t)$ near the end of the first 4 hours. Therefore, the positivity of the difference in the mean function shows in Figure 4.4 might be spurious. In gen-

eral, the δ-power tends to go down in the second visit compared with the first one, indicating that as humans aging, they might have more difficulty in relaxing during sleep.

CHAPTER 5

CONCLUSIONS

Penalized splines approach has been a powerful tool in nonparametric and semiparametric regression. This methodology has been widely used in practice though its large-sample distribution theory, was less explored. In Chapter 2, we addressed this issue and established the asymptotic distribution of P-spline estimators, proposed by Eilers and Marx (1996), for the univariate mean regression. We then continued our study on the asymptotics of P-spline estimators for bivariate additive models in Chapter 3. We believe that our work can similarly be extended for multivariate additive models. It would also be interesting to apply P-spline approach for nonparametric robust regression. In Chapter 4, we considered the global inferential problems, where we utilized the mixed model based penalized splines and expressed the mean function by polynomial splines. We applied our approach to specification tests on the mean for functional data. There are several directions for future study. First, we assume that the covariance structure is sparse, i.e., the number of eigenvalues are a finite constant. It might be possible to relax this assumption by assuming the eigenvalues of the covariance function decay exponentially fast. Second, for densely observed functional data, we have assumed an upper bound on the number of observations per subject. It would be interesting to see whether this bound can be further relaxed. Third, due to technical difficulty, we use a fixed large number of knots to model the mean function in Chapter 4. One might consider allowing the number of knots to increase as the sample size does.

APPENDIX A

CHAPTER 2 OF APPENDIX

A.1 Proof of results in Chapter 2

Proof of Proposition 2.2.1 on page 14: Notice that if ρ is a root of (2.5), so is $1/\rho$. If we can prove (II) and (III), we can conclude that none of the roots of $P(\rho)$ has norm 1 and hence (I) is true.

Now we prove (II). Let $p(\rho) = P(\rho) - \lambda(-1)^m(\rho-1)^{2m}$, where $P(\rho)$ is defined as (2.5). Let

$$\eta_1(\rho) = (\rho-1)^{2m} + (-1)^m \frac{p(\rho)}{\lambda} \quad \text{and} \quad \eta_2(\rho) = (\rho-1)^{2m} + (-1)^m \frac{p(1)}{\lambda}.$$

Notice that $\eta_1(\rho)$ have the same roots as $P(\rho)$. We want to prove (2.7) by studying the roots in $\eta_1(\rho)$ and $\eta_2(\rho)$. Define

$$u_k = 1 + \lambda^{-\frac{1}{2m}}\psi_k, \tag{A.1}$$

where $\{\psi_k, k = 1, \ldots, 2m\}$ are the roots of $x^{2m} + (-1)^m = 0$. For each k, we define a circle γ_k with the center u_k and the radius $C_f/\lambda^{1/m}$, where the constant C_f will be given later. The key step is to prove that

(*) : $\eta_1(\rho)$ has one and only one root, say ρ_k inside each γ_k,

and hence we can conclude that $\rho_k = u_k + O(\lambda^{-1/m})$. The justification of (*) will be given later. Notice that $\psi_k, k = 1, \ldots, 2m$ can be expressed as $\pm\alpha_k - \beta_k\imath, k = 1, \ldots, m$ with $\alpha_k > 0$. Without loss of generality, we assume that $\psi_k = -\alpha_k - \beta_k\imath, k = 1, \ldots, m$. We assume that $|\rho_k| < 1$. Hence $\rho_k = u_k + O(\lambda^{1/m}) =$

$1-(\alpha_k+\beta_k\imath)\lambda^{-1/(2m)}+O(\lambda^{1/m})$ for $k=1,\ldots,m$. Equation (2.7) is true if we can prove (*).

Since $p(1)=P(1)\equiv 1$, we have $\eta_2(u_k)=0$ and each u_k is the root of $\eta_2(\rho)$. Notice that γ_k's are disjoint. There is one and only one root of $\eta_2(\rho)$ in γ_k. Hence we only need to show that $\eta_1(\rho)$ and $\eta_2(\rho)$ have the same number of roots in γ_k. We can use Rouché Theorem (stated at the end of this section) and it suffices to check the following condition,

$$|\eta_1(\rho)-\eta_2(\rho)|<|\eta_2(\rho)|,\ \text{for } \rho \text{ on } \gamma_k. \tag{A.2}$$

Let C_h be a positive constant such that $|p(\rho)-p(1)|\le C_h|\rho-1|$ for $|\rho-1|\le 1/2$. Then for any ρ on the circle γ_k,

$$|\eta_1(\rho)-\eta_2(\rho)|=\left|\frac{p(\rho)-p(1)}{\lambda}\right|\le\frac{2C_h|\rho-1|}{\lambda}=\frac{2C_h}{\lambda^{(2m+1)/(2m)}}.$$

Since u_i's are the roots of $\eta_2(\rho)=0$, $\eta_2(\rho)=\prod_{i=1}^{2m}(\rho-u_i)$. For any ρ on the circle γ_k,

$$|\eta_2(\rho)|=\frac{C_f}{\lambda^{1/m}}\Big|\prod_{i\neq k}(\rho-u_i)\Big|\ge\frac{C_f}{\lambda^{1/m}}|\prod_{i\neq k}\Big||u_k-u_i|-|\rho-u_k|\Big|.$$

Notice that $|u_k-u_i|$ dominates $|\rho-u_k|$ for any ρ on the circle γ_k. We can find a constant C_1 such that,

$$|\eta_2(\rho)|\ge\frac{C_fC_1C_2}{\lambda^{1/m}}\left(\frac{p(1)}{\lambda}\right)^{\frac{2m-1}{2m}},\ \text{for } \rho \text{ on } \gamma_k.$$

where $C_2=\prod_{i\neq k}|\psi_k-\psi_i|$ is a nonzero constant. Choose C_f such that $C_fC_1C_2>2C_h$. Then (A.2) is satisfied and (*) is true. Hence (II) is proved.

Now consider the case when $p>m$. Let $\tilde{p}(\rho)=P(\rho)-\lambda(-1)^m\rho^{p-m}(\rho-1)^{2m}$. Define

$$\eta_1'(\rho)=(\rho-1)^{2m}+(-1)^m\frac{\tilde{p}(\rho)}{\lambda\rho^{p-m}}\quad\text{and}\quad\eta_2'(\rho)=(\rho-1)^{2m}+(-1)^m\frac{\tilde{p}(1)}{\lambda},$$

$$\tilde{\eta}_1(\rho) = \rho^{p-m} + (-1)^m \frac{\tilde{p}(\rho)}{\lambda(\rho-1)^{2m}} \quad \text{and} \quad \tilde{\eta}_2(\rho) = \rho^{p-m} + (-1)^m \frac{\tilde{p}(0)}{\lambda}.$$

Notice that the roots of $\tilde{\eta}_1(\rho)$ contain the roots of $P(\rho)$ whose norms are less than $1/2$, while $\eta_1'(\rho)$ contains the rest. Similarly, we want to prove (III) by studying the roots between $\tilde{\eta}_1(\rho)$ and $\tilde{\eta}_2(\rho)$ or those between $\eta_1'(\rho)$ and $\eta_2'(\rho)$. The proof is very similar and we only take $\tilde{\eta}_1(\rho)$ and $\tilde{\eta}_2(\rho)$ as an example.

Let $\tilde{u}_k = (\frac{\tilde{p}(0)}{\lambda})^{\frac{1}{p-m}} \tilde{\psi}_k$, where $\tilde{\psi}_1, \ldots, \psi_{p-m}$, are defined in Proposition 2.2.1. For any k, define a circle $\tilde{\gamma}_k$, with center at $\tilde{u}_k$ and radius $\tilde{C}_f/\lambda^{2/(p-m)}$, where $\tilde{C}_f$ is given later. Similar as above, $\tilde{\gamma}_k$'s are disjoint and each of them contains one root of $\tilde{\eta}_2(\rho)$. To prove (2.8), we want to show that $\tilde{\eta}_1(\rho)$ and $\tilde{\eta}_2(\rho)$ have the same number of roots in each $\tilde{\gamma}_k$ and it suffices to check the following condition

$$|\tilde{\eta}_1(\rho) - \tilde{\eta}_2(\rho)| \leq \tilde{\eta}_2(\rho) \text{ for } \rho \text{ on } \gamma_k. \tag{A.3}$$

Let $\tilde{C}_h$ be a positive constant such that $|\tilde{p}(\rho) - \tilde{p}(0)| \leq \tilde{C}_h|\rho|$ for all $|\rho| \leq 1/2$. For ρ on the circle $\tilde{\gamma}_k$, $|\rho| \leq 2\{\tilde{p}(0)/\lambda\}^{1/(p-m)}$. Hence

$$|\tilde{\eta}_1(\rho) - \tilde{\eta}_2(\rho)| \leq \frac{\tilde{C}_h|\rho| + 4m\tilde{p}(0)|\rho|}{\lambda|(1-\rho)^{2m}|} \leq \frac{C}{\lambda^{(p-m+1)/(p-m)}}, \qquad \text{for} \quad \rho \quad \text{on} \quad \tilde{\gamma}_k,$$

where C is a finite positive constant. Notice that $\tilde{u}_i$'s are roots of $\tilde{\eta}_2(\rho) = 0$.

$$|\tilde{\eta}_2(\rho)| \geq \frac{\tilde{C}_f\tilde{C}_1}{\lambda^{2/(p-m)}} \left\{ \frac{\tilde{p}(0)}{\lambda}^{(p-m+1)/(p-m)} \right\} \Big| \prod_{i \neq k} (\tilde{\psi}_k - \tilde{\psi}_i) \Big| \geq \frac{\tilde{C}_1\tilde{C}_2\tilde{C}_f}{\lambda^{(p-m+1)/(p-m)}},$$

where $\tilde{C}_1$ and $\tilde{C}_2$ is a finite positive constant. We can choose $\tilde{C}_f$ such that $\tilde{C}_1\tilde{C}_2\tilde{C}_f > C$. Therefore, (A.3) is true and (2.8) is proved. □

In Proposition 2.2.1, we conclude that $P(x)$ has q roots and m of them satisfy equation (2.7). Now we further show that those m roots have the following property. This result will be used in the proof of Proposition 2.2.2.

LEMMA A.1.1. *Let φ_k be the roots of $x^m + (-1)^m = 0$. Let $\alpha_k + \beta_k \imath$ be defined as Proposition 2.2.1. As $\lambda \to \infty$, for $k = 1, \ldots, m$, we have*

$$\rho_k + 1/\rho_k - 2 \;=\; \varphi_k/\lambda^{1/m} + o(\lambda^{-1/m}) \sim (\alpha_k + \beta_k \imath)^2 \lambda^{-1/m}. \tag{A.4}$$

Proof of Lemma A.1.1: Recall that

$$P(\rho_k) = \lambda(-1)^m(\rho_k - 1)^{2m} + p(\rho_k) = 0. \tag{A.5}$$

By Proposition 2.7, $\rho_k \to 1$. Hence $\rho_k^j - \rho_k^{j'} = o(1)$ for $0 \le j \le m$ and $0 \le j' \le m$. Since $p(1) = 1$, we have $p(\rho_k) = 1 + o(1) = \rho_k^m\{1 + o(1)\}$. Therefore, equation (A.5) yields that

$$\begin{aligned} \lambda(-1)^m(\rho_k - 1)^{2m} &= -\rho_k^m\{1 + o(1)\} \\ \frac{(\rho_k - 1)^{2m}}{\rho_k^m} &= \frac{(-1)^{m+1}}{\lambda} + o(\lambda^{-1}). \end{aligned} \tag{A.6}$$

Notice that φ_k for $k = 1, \ldots, m$ are the m roots of $(-1)^{m+1}$. Hence we can associate ρ_k with φ_k such that

$$\frac{\rho_k^2 - 2\rho_k + 1}{\rho_k} = \frac{\varphi_k}{\lambda^{1/m}} + o(\lambda^{-1/m}).$$

Since $|\rho_k| > 0$, it is equivalent to consider the root of $\rho_k^2 - \{2 + \varphi_k/\lambda^{1/m} + o(\lambda^{-1/m})\}\rho_k + 1 = 0$. Because $|\rho_k| < 1$,

$$\rho_k = \frac{2 + \varphi_k/\lambda^{1/m} - 2\sqrt{\varphi_k/\lambda^{1/m}}}{2} + o(\lambda^{-1/(2m)}) = 1 - \frac{\sqrt{\varphi_k}}{\lambda^{1/(2m)}} + o(\lambda^{-1/(2m)}).$$

Comparing this with equation (2.7), we must have $\sqrt{\varphi_k} = \alpha_k + \beta_k \imath + o(1)$. So $\varphi_k = (\alpha_k + \beta_k \imath)^2 + o(1)$. Hence Lemma A.1.1 is proved. □

We also need the following lemmas for Proposition 2.2.2.

LEMMA A.1.2. *Let $z_1, \ldots, z_m$ be the roots of $x^m + (-1)^m = 0$. Then for any given i,*

$$\prod_{j \ne i}(z_i - z_j) = m(-1)^{m+1} z_i^{-1}. \tag{A.7}$$

Proof of Lemma A.1.2: The definition of z_j implies that $z^m + (-1)^m = \prod_{j=1}^m (z - z_j)$ for any z. Take its first derivative at z_i, we have

$$mz^{m-1}\Big|_{z=z_i} = \Big\{ \prod_{j=1}^m (z - z_j) \Big\}^{(1)} \Big|_{z=z_i} = \prod_{j \neq i} (z_i - z_j). \tag{A.8}$$

Since $z_i^m + (-1)^m = 0$, $mz_i^{m-1} = m(-1)^{m+1} z_i^{-1}$. Equation (A.7) holds. □

LEMMA A.1.3. *Define $s_j(\rho_k) = \mathbf{T}_i(\rho_k)(\Omega_{m,p})^T_{i-q+j}$ for $1 \leq j \leq q-1$. Then $s_j(\rho_k)$ does not depend on i. Moreover, for the first m roots, where equation (2.7) holds, we have*

$$s_j(\rho_k) = \sum_{\ell=0}^{j-1} \omega_{q-\ell} (\rho_k^{j-\ell} - \rho_k^{\ell-j}). \tag{A.9}$$

Proof of Lemma A.1.3: Notice that the ith element of $\mathbf{T}_i$ is 1. The $(i-q+j)$th element and the ith elements of $(\Omega_{m,p})_{i-q+1}$ are ω_0 and ω_{q-j} respectively. Hence

$$s_j(\rho_k) = \omega_q \rho_k^j + \cdots + \omega_{q-j} 1 + \omega_{q-j-1} \rho_k + \cdots + \omega_0 \rho_k^{q-j} + \cdots \omega_q \rho_k^{2q-j}, \tag{A.10}$$

and $s_j(\rho_k)$ in fact does not depend on i. Notice that ρ_k^{-1} exists for $k = 1, \ldots, m$. Since $P(\rho_k) = 0$,

$$\begin{aligned} & s_1(\rho_k) \\ = \; & \omega_q \rho_k - \left(\omega_q \rho_k^{-1} - \omega_q \rho_k^{-1} \right) + \omega_{q-1} 1 + \omega_{q-2} \rho_k + \cdots + \omega_0 \rho_k^{q-1} + \cdots \omega_q \rho_k^{2q-1} \\ = \; & \omega_q (\rho_k - \rho_k^{-1}) + P(\rho_k)/\rho_k = \omega_q (\rho_k - \rho_k^{-1}). \end{aligned}$$

In general, we can add and subtract terms involving $\rho_k^{-1}, \ldots, \rho_k^{-j}$ and conclude (A.9) holds. □

LEMMA A.1.4. *Let $s_j(\rho_k)$ be defined as in Lemma A.1.3. Let A be the $(q-1) \times q$ matrix whose (j,k)th element is $s_j(\rho_k)$. Let $A_{(i)}$ be the $(q-1) \times (q-1)$ matrix obtained by deleting the ith column of A. Then*

$$a_i / a_m = (-1)^{i-m} \det(A_{(i)}) / \det(A_{(m)}) \qquad \text{for} \quad 1 \leq i \leq q, \tag{A.11}$$

where det *is the determinant operator.*

Proof of Lemma A.1.4: Recall that $\mathbf{S}_i = \sum_{k=1}^{q} a_k T_i(\rho_k)$ and $\mathbf{S}_i(\Omega_{p,m})_{i-j}^T = 0$ by equation (2.6). Hence the vector $(a_1, \ldots, a_q)$ simultaneously satisfies $q-1$ equations:

$$\sum_{k=1}^{q} a_k s_j(\rho_k) = 0 \quad \text{for} \quad j = 1, \ldots, q-1. \tag{A.12}$$

Notice that the vector $\left\{(-1)^1 \det(A_{(1)}), \ldots, (-1)^q \det(A_{(q)})\right\}$ is also the solution to (A.12) because for $j = 1, \ldots, q-1$,

$$\sum_{k=1}^{q} (-1)^k \det(A_{(k)}) s_j(\rho_k) = -\det \begin{pmatrix} s_j(\rho_1) & \ldots & s_j(\rho_q) \\ s_1(\rho_1) & \ldots & s_1(\rho_q) \\ s_2(\rho_1) & \ldots & s_2(\rho_q) \\ \vdots & \ddots & \vdots \\ s_{q-1}(\rho_1) & \ldots & s_{q-1}(\rho_q) \end{pmatrix} = 0.$$

Since the solutions to (A.12) span a 1-dimensional subspace, equation (A.11) must hold. □

LEMMA A.1.5. *Let A be defined as in Lemma A.1.4. Assume that $\lambda \to \infty$.*

Then when $i \leq m$,

$$\frac{\det(A_{(i)})}{\det(A_{(m)})} = (-1)^{i-m} \frac{\alpha_i + \beta_i \imath}{\alpha_m + \beta_m \imath} + O(\lambda^{-1/m}); \tag{A.13}$$

when $i > m$,

$$\frac{\det(A_{(i)})}{\det(A_{(m)})} = O(\lambda^{-p/(p-m)+1/(2m)}). \tag{A.14}$$

Proof of Lemma A.1.5: First consider the case when $p \leq m$. Notice that $s_j(\rho_k)$ can be simplified as in equation (A.9). By several applications of the fact that the determinant of a matrix is unchanged if a multiple of its kth row is added to its k'th row for any $k \neq k'$, we can relate $\det(A_{(m)})$ to the determinant of a

Vandermonde matrix and conclude that

$$\det(A_{(m)}) \;=\; \omega_q^{m-1}\Big\{\prod_{i=1}^{m-1}(\rho_i - 1/\rho_i)\Big\}\Big\{\prod_{1\le i<j\le m-1}(\rho_j + 1/\rho_j - \rho_i - 1/\rho_i)\Big\}.$$

The calculation of $\det(A_{(i)})$ is similar by replacing ρ_m with ρ_i. Hence

$$\frac{\det(A_{(i)})}{\det(A_{(m)})} = \left(\frac{\rho_m - 1/\rho_m}{\rho_i - 1/\rho_i}\right)\frac{\prod_{j\neq m}(\rho_m + 1/\rho_m - \rho_j - 1/\rho_j)}{\prod_{j\neq i}(\rho_i + 1/\rho_i - \rho_j - 1/\rho_j)}(-1)^{i-m}.$$

Note that $\rho_i - 1/\rho_i = -2(\alpha_i + \beta_i\imath)\lambda^{-1/(2m)} + O(\lambda^{-3/2m})$. Moreover,

$$\rho_i + 1/\rho_i - \rho_j - 1/\rho_j = (\alpha_i + \beta_i\imath)^2\lambda^{-1/m} - (\alpha_j + \beta_j\imath)^2\lambda^{-1/m} + O(\lambda^{-2/m}).$$

Therefore, as $\lambda \to \infty$,

$$(-1)^{i-m}\frac{\det(A_{(i)})}{\det(A_{(m)})} \;=\; \frac{\alpha_m + \beta_m\imath}{\alpha_i + \beta_i\imath}\frac{\prod_{j\neq m}\{(\alpha_m + \beta_m\imath)^2 - (\alpha_j + \beta_j\imath)^2\}}{\prod_{j\neq i}\{(\alpha_i + \beta_i\imath)^2 - (\alpha_j + \beta_j\imath)^2\}} + O(\lambda^{-1/m}).$$

Notice that $(\alpha_j + \beta_j\imath)^2$ for $j \le m$ form the m roots of $x^m + (-1)^m = 0$. Hence $\prod_{j\neq\ell}(\rho_\ell + 1/\rho_\ell - \rho_j - 1/\rho_j) = m(-1)^{m+1}(\alpha_\ell + \beta_\ell\imath)^{-2}$. By choosing $\ell = m$ or i, we can simplify the equations above and conclude that equation (A.13) holds when $p \le m$.

Now consider $p > m$. We still have,

$$\frac{\det(A_{(i)})}{\det(A_{(m)})} = \left(\frac{\rho_m - 1/\rho_m}{\rho_i - 1/\rho_i}\right)\frac{\prod_{j\neq m}(\rho_m + 1/\rho_m - \rho_j - 1/\rho_j)}{\prod_{j\neq i}(\rho_i + 1/\rho_i - \rho_j - 1/\rho_j)}(-1)^{i-m}.$$

When $i < m$, we can apply the same techniques as above and conclude that equation (A.13) still holds. By Proposition 2.2.1, if $i > m$ or $j > m$,

$$\rho_i + 1/\rho_i - \rho_j - 1/\rho_j = O(\lambda^{1/(p-m)}).$$

Hence when $i > m$,

$$(-1)^{i-m}\frac{\det\{A_{(i)}\}}{\det\{A_{(m)}\}} = O(\lambda^{-p/(p-m)+1/(2m)}).$$

Therefore, Lemma A.1.5 holds. □

Proof of Proposition 2.2.2 on page 15: By Proposition 2.2.1, for $1 \leq k \leq m$,

$$\begin{aligned} a_k \mathbf{T}_i(\rho_k)\mathbf{J} &= a_k \sum_{j=1}^{K+p} \rho_k^{|i-j|} \approx a_k \frac{1+\rho_k}{1-\rho_k} = \frac{a_k \lambda^{1/(2m)}}{\alpha_k + \beta_k \imath}[2 + O\{\lambda^{-1/(2m)}\}] \\ &= \frac{2a_k \lambda^{1/(2m)}}{\alpha_k + \beta_k \imath} + O(a_k). \end{aligned} \tag{A.15}$$

Lemma A.1.4 and Lemma A.1.5 imply that the dominant term in $(2a_k\lambda^{1/(2m)})/(\alpha_k + \beta_k\imath)$ remain the same for $k = 1, \ldots, m$. Recall that $\mathbf{S}_i\mathbf{J} = \sum_{k=1}^m a_k \mathbf{T}_i(\rho_k)\mathbf{J} = 1$. When $p \leq m$, each $a_k\mathbf{T}_i(\rho_k)\mathbf{J}$ must contribute $m^{-1} + o(1)$. Hence we have $a_k = (\alpha_k + \beta_k\imath)/(2m\lambda^{1/(2m)}) + O(\lambda^{-1/m})$. When $p > m$, $a_k = O(\lambda^{-p/(p-m)})$ for $k > m$. Hence Proposition 2.2.2 is true. □

Proof of Proposition 2.2.3 on page 16: Note that

$$\hat{\mu}(x) = \sum_k B_k^{[p]}(x)\hat{b}_k = M^{-1} \sum_{i=1}^n y_i \{\sum_k \sum_r B_k^{[p]}(x) B_r^{[p]}(x_i) S_{k,r}\},$$

where $S_{k,r} = \sum_{\nu=1}^p a_\nu \rho_\nu^{|k-r|}$. Note that if $\nu > m$, $a_\nu = O(\lambda^{-p/(p-m)})$. Hence $M^{-1}\sum_{i=1}^n y_i\{\sum_k \sum_r B_k^{[p]}(x) B_r^{[p]}(x_i) a_\nu \rho_\nu^{|k-r|}\} = O\{(Kh)^{-2}\}$. In other words, we only need to consider the dominant term in $S_{k,r}$, i.e., $\sum_{\nu=1}^m a_\nu \rho_\nu^{|k-r|} = H_m(\frac{l-j}{Kh})$.

First consider the case when $p > 0$. Let $\tilde{B}^{[p]}(x)$ be the cardinal B-spline. Then

$$\begin{aligned} \hat{b}_k &= M^{-1}K^{-1} \sum_{j=1}^K h^{-1} H_m(\frac{l-j}{Kh}) n^{-1} \sum_{i=1}^n K\tilde{B}^{[p]}[\frac{x_i - K\{j-(p+1)/2\}}{K}] y_i \\ &= n^{-1} \sum_{i=1}^n \sum_{j=1}^K h^{-1} H_m(\frac{l-j}{Kh}) \tilde{B}^{[p]}[\frac{x_i - K\{j-(p+1)/2\}}{K}] y_i \\ &= n^{-1} \sum_{i=1}^n h^{-1} H_m(\frac{l-j}{Kh}) y_i + n^{-1} \sum_{i=1}^n \sum_{j=1}^n h^{-1} * \\ &\quad \Big[H_m(\frac{l-j}{Kh}) - H_m\{\frac{l-(p+1)/2 - Kx_i}{Kh}\}\Big] \tilde{B}^{[p]}\Big[\frac{x_i - K\{j-(p+1)/2\}}{K}\Big] y_i \end{aligned}$$

Note that H_m has a bounded second derivative, hence the following term

$$\frac{1}{nh}\sum_{i=1}^{n}\sum_{j=1}^{n}\Big[H_m(\frac{l-j}{Kh}) - H_m\{\frac{l-(p+1)/2-Kx_i}{Kh}\}\Big]\tilde{B}^{[p]}\Big[\frac{x_i - K\{j-(p+1)/2\}}{K}\Big]$$

is $O\{(Kh)^{-2}\}$. Therefore, $\hat{b}_k = n^{-1}\sum_{i=1}^{n}[h^{-1}H_m(\frac{\{l-(p+1)/2\}-Kx_i}{Kh}) + O\{(Kh)^{-2}\}]y_i$ Applying the same technique again, we conclude that

$$\hat{\mu}(x) = \sum_k B_k^{[p]}(x)\hat{b}_k = n^{-1}\sum_{i=1}^{n}\Big[h^{-1}H_m(\frac{x-x_i}{h}) + O\{(Kh)^{-2}\}\Big]y_i.$$

Hence equation (2.12) holds when $p > 0$.

When $p = 0$, recall that $\tilde{y}_j$, which is defined as the jth element of $B_p^T\mathbf{Y}/M$, satisfies $E\tilde{y}_j = \mu(\bar{x}_j) + O(K^{-1})$. It is straight forward to show that $\hat{\mu}(x) = n^{-1}\sum_{i=1}^{n}h^{-1}H_m(\frac{x-x_i}{h})y_i + O_p(K^{-1})$. Therefore, Proposition 2.2.3 holds. □

THEOREM A.1.1. *(Rouché Theorem) Let h_1 and h_2 be analytic functions in a simply connected domain D containing a Jordan contour C. Suppose that $|h_1(z) - h_2(z)| < h_2(z)$ on C. Then h_1 and h_2 have the same number of zeros inside C.*

APPENDIX B

CHAPTER 3 OF APPENDIX

B.1 Proof of results in Chapter 3

Proof of Proposition 3.1.1 on page 23: Denote $S_d = (C_{dd} + \lambda_d P_m)^{-1}$. First we show by induction that for any j, $j = 1, 2, \ldots,$

$$\begin{aligned} \hat{b}_1^{(j)} &= \sum_{i=1}^{j} (S_1 C_{12} S_2 C_{21})^{i-1} (S_1 V_1 - S_1 C_{12} S_2 V_2), \\ \hat{b}_2^{(j)} &= \sum_{i=1}^{j} (S_2 C_{21} S_1 C_{12})^{i-1} (S_2 V_2 - S_2 C_{21} S_1 V_1), \end{aligned} \tag{B.1}$$

By definition, $\hat{b}_d^{(0*)} = S_d V_d$ and $\hat{b}_d^{(1)} = S_d V_d - S_d C_{dd'} S_{d'} V_{d'}$, where $d \neq d'$. Hence (B.1) is true for $j = 1$.

Suppose it is true for step j. Then

$$\begin{aligned} \hat{b}_1^{(j*)} &= S_1 V_1 - S_1 C_{12} \hat{b}_2^{(j)} \\ &= S_1 V_1 - S_1 C_{12} \Big\{ \sum_{i=1}^{j} (S_2 C_{21} S_1 C_{12})^{i-1} (S_2 V_2 - S_2 C_{21} S_1 V_1) \Big\} \\ &= S_1 V_1 + S_1 C_{12} \sum_{i=1}^{j} (S_2 C_{21} S_1 C_{12})^{i-1} S_2 C_{21} S_1 V_1 \\ &\quad - S_1 C_{12} \sum_{i=1}^{j} (S_2 C_{21} S_1 C_{12})^{i-1} S_2 V_2 \\ &= \left\{ \sum_{i=1}^{j+1} (S_1 C_{12} S_2 C_{21})^{i-1} \right\} S_1 V_1 - \left\{ \sum_{i=1}^{j} (S_1 C_{12} S_2 C_{21})^{i-1} \right\} S_1 C_{12} S_2 V_2. \end{aligned}$$

Similarly,

$$\hat{b}_2^{(j*)} = \left\{ \sum_{i=1}^{j+1} (S_2 C_{21} S_1 C_1)^{i-1} \right\} S_2 V_2 - \left\{ \sum_{i=1}^{j} (S_2 C_{21} S_1 C_{12})^{i-1} \right\} S_2 C_{21} S_1 V_1.$$

Hence

$$\begin{aligned}
\hat{b}_1^{(j+1)} &= S_1V_1 - S_1C_{12}\hat{b}_2^{(j*)} \\
&= S_1V_1 + S_1C_{12}\left\{\sum_{i=1}^{j}(S_2C_{21}S_1C_{12})^{i-1}\right\}S_2C_{21}S_1V_1 \\
&\quad -S_1C_{12}\left\{\sum_{i=1}^{j+1}(S_2C_{21}S_1C_{12})^{i-1}\right\}S_2V_2 \\
&= \sum_{i=1}^{j+1}(S_1C_{12}S_2C_{21})^{i-1}(S_1V_1 - S_1C_{12}S_2V_2),
\end{aligned}$$

and

$$\hat{b}_2^{(j+1)} = \sum_{i=1}^{j+1}(S_2C_{21}S_1C_{12})^{i-1}(S_2V_2 - S_2C_{21}S_1V_1).$$

Next we prove that $\lim_{j\to\infty} b_d^{(j)}$, $d = 1, 2$, exist. First consider $\hat{b}_1^{(j+1)}$. Perform an eigen decomposition on $S_1C_{12}S_2C_{21}$. Denote the eigen pairs as (u_k, ξ_k), where $u_1 \leq u_2 \leq \cdots \leq u_K$. Note that S_1 and $C_{12}S_2C_{21}$ are both positive definite matrices. From their definition, it is obvious that their eigenvalues are all between 0 and 1. Since $C_{12} = B_1^TB_2/M$, it has at most one eigenvalue that equals 1. Hence $u_1 \neq -1$ and $u_{K-1} \neq 1$. By Lemma B.1.1, we have

$$1_n^TB_1(S_1C_{12}S_2C_{21}) = 1_n^T(B_1S_1B_1^T/M)(B_2S_2B_2^T/M)B_1 = 1_n^TB_1.$$

Hence $u_K = 1$ and $\xi_k = B_1^T1_n$. Since

$$1_n^TB_1S_1V_1 = 1_n^TB_1S_1B_1^T\tilde{Y}/M = 1_n^T\tilde{Y} = 0,$$

we have

$$\xi_k^T(S_1V_1 - S_1C_{12}S_2V_1) = 1_n^TB_1(S_1V_1 - S_1C_{12}S_2V_1) = 0.$$

Hence we can express $S_1V_1 - S_1C_{12}S_2V_2$ as a linear combination $\sum_{k=1}^{K-1} e_k\xi_k$. Notice that

$$(S_1C_{12}S_2C_{21})^i(S_1V_1 - S_1C_{12}S_2V_2) = \sum_{k=1}^{K-1} u_k^ie_k\xi_k.$$

Correspondingly, we have

$$\lim_{j\to\infty} \hat{b}_1^{(j)} = \lim_{j\to\infty} \sum_{i=1}^{j} \sum_{k=1}^{K-1} u_k^{i-1} e_k \xi_k = \sum_{k=1}^{K-1} \frac{1}{1-u_k} e_k \xi_k.$$

The above argument can also be used to prove the existence of $\lim_{j\to\infty} \hat{b}_2^{(j)}$. Note that the limit both satisfy equation (3.6). Moreover, since $1_n^T B_d \hat{b}_d^{(j)} = 0$ by orthogonality, so does the limit. Hence $\lim_{j\to\infty} \hat{b}_d^{(j)}$ must be the same as the spline coefficients $\hat{b}_d$. Proposition 3.1.1 holds. □

LEMMA B.1.1. *Let $S_d = (C_{dd} + \lambda_d P_m)^{-1}$ and let 1_n be the $n \times 1$ vector whose elements are all 1. Then $M^{-1} B_d S_d B_d^T 1_n = 1_n$.*

Proof of Lemma B.1.1: Notice that $C_{dd} = M^{-1} B_d^T B_d$ and $P_m 1_{K+p_d} = 0_{K+p_d}$. Hence

$$M^{-1} B_d^T B_d 1_{K+p_d} = (M^{-1} B_d^T B_d + \lambda_d P_m) 1_{K+p_d}.$$

Multiply $B_d S_d$ on both sides. Then $M^{-1} B_d S_d B_d^T B_d 1_{K+p_d} = B_d 1_{K+p_d}$. Notice that $B_d 1_{K+p_d} = 1_n$. Hence Lemma B.1.1 is true. □

Proof of Proposition 3.1.2 on page 24: We can use induction to prove (3.9). Note that $\hat{\mu}_d^{(0)}(x_d) = \bar{\mu}_d^{(0)}(x_d) = 0$. Hence (3.9) holds when $j = 0$. Suppose it also holds for j. Denote $\hat{\mu}_d^{(j*)}(x_d) = \sum_{k=1}^{K+p_d} \hat{b}_{dk}^{(j*)} B_k^{[p_d]}(x_d)$. By the definition of Approach A.1 and Proposition 2.2.3,

$$\hat{\mu}_d^{(j*)}(x_d) = \frac{1}{n h_d} \sum_{i=1}^{n} H_m(\frac{x_d - x_{di}}{h_d}) \{\tilde{y}_i - \hat{\mu}_{d'}^{(j)}(x_{d'i})\} + o_p(n^{-\frac{2m}{4m+1}}). \tag{B.2}$$

Since the marginal distribution is uniform, $\bar{f}(x_d) = 1$. By standard kernel techniques,

$$\int \bar{\mu}_{d'}^{(j*)}(x_{d'}) \bar{f}(x_d, x_{d'}) / \bar{f}_d(x_d) dx_{d'} = n^{-1} h_d^{-1} \sum_{i=1}^{n} H_m(\frac{x_d - x_{di}}{h_d}) \{\bar{\mu}_{d'}^{(j)}(x_{d'i}) + O_p(h_d^{2m})\}.$$

Therefore,

$$\begin{aligned}\bar{\mu}_d^{(j*)}(x_d) &= \tilde{u}_d(x_d) - \bar{y} - n^{-1}h_d^{-1}\sum_{i=1}^{n} H_m(\frac{x_d - x_{di}}{h_d})\{\bar{\mu}_{d'}^{(j)}(x_{d'i})\} + o_p(n^{-\frac{2m}{4m+1}}) \\ &= n^{-1}h_d^{-1}\sum_{i=1}^{n} H_m(\frac{x_d - x_{di}}{h_d})\{\tilde{y}_i - \bar{\mu}_{d'}^{(j)}(x_{d'i})\} + o_p(n^{-\frac{2m}{4m+1}}).\end{aligned}$$

Note that we assume equation (3.9) holds for j. Denote that $B_d(x_d) = \{B_{d1}(x_d), \ldots, B_{d,K+p_d}(x_d)\}$. By Proposition 2.2.3,

$$n^{2m/(4m+1)}\{B_d(x_d)S_dV_d - \tilde{\mu}_d(x_d)\} \xrightarrow{p} 0.$$

Together with equation (B.2), we have, $n^{2m/(4m+1)}\{\hat{\mu}_d^{(j*)}(x_d) - \bar{\mu}_d^{(j*)}(x_d)\} \xrightarrow{p} 0$. Similarly, we can conclude that $n^{2m/(4m+1)}\{\hat{\mu}_d^{(j+1)}(x_d) - \bar{\mu}_d^{(j+1)}(x_d)\} \xrightarrow{p} 0$. Therefore, equation (3.9) holds for $j+1$. Proposition 3.1.2 holds. □

Proof of Theorem 3.2.2 on page 29: Recall that $\hat{\mu}_d^s(x_d)$'s are the P-spline estimators using suboptimal penalty, $B_d(x_d) = \{B_{d1}(x_d), \ldots, B_{d,K+p_d}(x_d)\}$ and $\hat{\mu}_d(x_d) = B_d(x_d)\hat{b}_d$, where $\hat{b}_d$ is the coefficients from Approach A.3. By equation (3.15),

$$\begin{aligned}\hat{\mu}_d(x_d) &= B_d(x_d)S_dB_d^TW\{\tilde{Y} - \sum_{d'\neq d}\hat{\mu}_{d'}^s(X_d')\} \\ &= B_d(x_d)S_dB_d^TW\{\tilde{Y} - \sum_{d'\neq d}\mu_{d'}(X_{d'})\} \\ &\quad + B_d(x_d)S_dB_d^TW\sum_{d'\neq d}\{\mu_{d'}(X_{d'}) - \hat{\mu}_{d'}^s(X_{d'})\} \\ &=: J_1 + J_2.\end{aligned}$$

Note that the term J_1 is the univariate P-spline estimator. Hence

$$n^{2m/(4m+1)}\{J_1 - \mu_d(x_d)\} \Rightarrow N\{(h_d')^{2m}\mu_d^{(2m)}(x_d)\int t^{2m}H_m(t)dt, \Psi_d(x_d)\}.$$

It suffices to show that

$$n^{2m/(4m+1)}J_2 = o_p(1). \tag{B.3}$$

Using the connection between additive P-spline estimators and kernel estimators in backfitting, we have

$$\begin{aligned}
& \hat{\mu}^s_{d'}(x_{d'}) - \mu_{d'}(x_{d'}) \\
= \quad & \{\frac{1}{nh_s}\sum_{j=1}^{n} H_m\left(\frac{x_{d'} - x_{d'j}}{h_s}\right) w_{jj}\epsilon_j + h_s^{2m}\mu_{d'}^{(2m)}(x_{d'}) \int t^{2m} H_m(t)dt\}\{1 + o_p(1)\} \\
=: \quad & (J_3 + J_4)\{1 + o_p(1)\},
\end{aligned}$$

where $\epsilon_j = y_j - \sum_{d=1}^{2} \mu_d(x_{dj})$ and w_{jj} is the (j,j)th element of the diagonal matrix W. Notice that $B_d(x_d)S_dB_d^TWJ_4 = O(h_s^{2m}) = o(n^{-2m/(4m+1)})$ because of the undersmoothing in step 1. To justify equation (B.3), it remains show that

$$B_d(x_d)S_dB_d^TWJ_3 = o_p(n^{-2m/(4m+1)}).$$

We adopt the techniques in Linton (1997). By interchanging the order of summations and using the connection between P-splines and kernel approach, we have

$$\begin{aligned}
& B_d(x_d)S_dB_d^TWJ_3 \\
= \quad & \frac{1}{nh_d}\sum_{i=1}^{n} H_m\Big(\frac{x_d - x_{di}}{h_d}\Big)\frac{1}{nh_s}\sum_{j=1}^{n} H_m\Big(\frac{x_{d'i} - x_{d'j}}{h_s}\Big) w_{jj}\epsilon_j + o_p(n^{-\frac{4m}{4m+1}}) \\
= \quad & \frac{1}{n}\sum_{j=1}^{n}\Big\{\frac{1}{nh_sh_d}\sum_{i=1}^{n} H_m\Big(\frac{x_d - x_{di}}{h_d}\Big)H_m\Big(\frac{x_{d'i} - x_{d'j}}{h_s}\Big)\Big\} w_{jj}\epsilon_j + o_p(n^{-\frac{4m}{4m+1}}) \\
=: \quad & n^{-1}\sum_{j=1}^{n} s_j w_{jj}\epsilon_j + o_p(n^{-\frac{4m}{4m+1}}),
\end{aligned}$$

where s_j satisfies

$$s_j =: \frac{1}{nh_sh_d}\sum_{i=1}^{n} H_m\Big(\frac{x_d - x_{di}}{h_d}\Big)H_m\Big(\frac{x_{d'i} - x_{d'j}}{h_s}\Big) = O_p(1),$$

uniformly in j. Therefore, $B_d(x_d)S_dB_d^TWJ_3 = O_p(n^{-1/2}) = o_p(n^{-2m/(4m+1)}$ and Theorem 3.2.2 holds. □

APPENDIX C

CHAPTER 4 OF APPENDIX

C.1 Proof of results in Chapter 4

Proof of Proposition 4.1.1 on page 33: Recall that the pseudo log-likelihood $\log \hat{L}_{\hat{Y}}(\beta, \lambda)$ is defined in equation (4.6). Solving the first order condition for β, we get the maximum pseudo profile likelihood estimation $\hat{\beta}(\lambda) = (\hat{X}^T \hat{H}_\lambda^{-1} \hat{X})^{-1} \hat{X}^T \hat{H}_\lambda^{-1} \hat{Y}$. Let $\log \hat{L}_{\hat{Y}}(\lambda)$ be the pseudo profile likelihood when β is maximized out, i.e.,

$$2 \log \hat{L}_{\hat{Y}}(\lambda) = 2 \log \hat{L}_{\hat{Y}}\{\hat{\beta}(\lambda), \lambda\} = -[\log |\hat{H}_\lambda| + \{\hat{Y} - \hat{X}\hat{\beta}(\lambda)\}^T \hat{H}_\lambda^{-1} \{\hat{Y} - \hat{X}\hat{\beta}(\lambda)\}].$$

Let $\log \hat{L}^{0,N}$ be the maximum pseudo log-likelihood under the null hypothesis (4.4). Then we can decompose LRT_N into two parts, i.e.,

$$LRT_N = \sup_{\lambda \geq 0} \{2 \log \hat{L}_{\hat{Y}}(\lambda) - 2 \log \hat{L}_{\hat{Y}}(0)\} + 2\{\log \hat{L}_{\hat{Y}}(0) - \log \hat{L}^{0,N}\}, \qquad \text{(C.1)}$$

where the first part corresponds to testing for $\lambda = 0$ and the second part corresponds to testing for the fixed effects $\beta_q = 0$ for $q \in Q$.

Consider the first part in (C.1). Recall that $\hat{H}_\lambda = I_N + \lambda \hat{Z}\hat{Z}^T$. Define $\hat{\xi}_{K,N}$ be the kth eigenvalues of $N^{-\varrho}\hat{Z}^T\hat{Z}$ for $k = 1, \ldots, K$. Since $\hat{Z}\hat{Z}^T$ has the same nonzero eigenvalues as $\hat{Z}^T\hat{Z}$, we have,

$$\log |\hat{H}_\lambda| = \sum_{k=1}^{K} \log(1 + \lambda N^{\varrho} \hat{\xi}_{k,N}). \qquad \text{(C.2)}$$

According to Patterson and Thompson (1971), there exists an $N \times (N - p - 1)$ matrix $\hat{W}$ such that $\hat{W}\hat{W}^T = I_N - \hat{X}(\hat{X}^T\hat{X})\hat{X}^T$, $\hat{W}^T\hat{W} = I_{N-p-1}$, and

$$\{\hat{Y} - \hat{X}\hat{\beta}(\lambda)\}^T \hat{H}_\lambda^{-1} \{\hat{Y} - \hat{X}\hat{\beta}(\lambda)\} = \hat{Y}^T \hat{W}(\hat{W}^T \hat{H}_\lambda \hat{W})^{-1} \hat{W}^T \hat{Y}. \qquad \text{(C.3)}$$

We can further simplify (C.3) as

$$\begin{aligned}
&\hat{Y}^T\hat{W}(\hat{W}^T\hat{H}_\lambda\hat{W})^{-1}\hat{W}^T\hat{Y} = \hat{Y}^T\hat{W}(I_{N-p-1}+\lambda\hat{W}^T\hat{Z}\hat{Z}^T\hat{W})^{-1}\hat{W}^T\hat{Y} \\
= \;& \hat{Y}^T\hat{W}\hat{W}^T\hat{Y} - \lambda\hat{Y}^T\hat{W}\{\hat{W}^T\hat{Z}(I_K+\lambda\hat{Z}^T\hat{W}\hat{W}^T\hat{Z})^{-1}\hat{Z}^T\hat{W}\}\hat{W}^T\hat{Y},
\end{aligned}$$

where the second equation results from the Woodbury matrix identity (Woodbury, 1950). Note that $-2\log\hat{L}_{\hat{Y}}(0) = \hat{Y}^T\hat{W}\hat{W}^T\hat{Y}$. Hence

$$\begin{aligned}
2\log\hat{L}_{\hat{Y}}(\lambda) - 2\log\hat{L}_{\hat{Y}}(0) \;=\; & \lambda\hat{Y}^T\hat{W}\hat{W}^T\hat{Z}(I_K+\lambda\hat{Z}^T\hat{W}\hat{W}^T\hat{Z})^{-1}\hat{Z}^T\hat{W}\hat{W}^T\hat{Y} \\
& - \sum_{k=1}^{K}\log(1+\lambda\hat{\xi}_{k,N}). \qquad \text{(C.4)}
\end{aligned}$$

Define $\hat{\zeta}_{k,N}$ be the kth eigenvalue of $N^{-\varrho}\hat{Z}^T\hat{W}\hat{W}^T\hat{Z}$ and let $\hat{U}_{\hat{Z}\hat{W}}$ be the $K\times K$ matrix whose kth column is the eigenvectors asscociated with $\hat{\zeta}_{k,N}$. Note that $\hat{Z}^T\hat{W}\hat{W}^T\hat{Z} = \hat{U}_{\hat{Z}\hat{W}}\text{diag}(N^{\varrho}\hat{\zeta}_{1,N},\ldots,N^{\varrho}\hat{\zeta}_{K,N})\hat{U}_{\hat{Z}\hat{W}}$ and thus

$$(I_K+\lambda\hat{Z}^T\hat{W}\hat{W}^T\hat{Z})^{-1} = \hat{U}_{\hat{Z}\hat{W}}\text{diag}\left\{\frac{1}{1+\lambda N^{\varrho}\hat{\zeta}_{1,N}},\ldots,\frac{1}{1+\lambda N^{\varrho}\hat{\zeta}_{K,N}}\right\}\hat{U}_{\hat{Z}\hat{W}}.$$

Together with equation (C.3), we have

$$2\log\hat{L}_{\hat{Y}}(\lambda) - 2\log\hat{L}_{\hat{Y}}(0) = \sum_{k=1}^{K}\frac{\lambda N^{\varrho}\hat{w}_{k,N}^2}{1+\lambda N^{\varrho}\hat{\zeta}_{k,N}} - \sum_{k=1}^{K}\log(1+\lambda N^{\varrho}\hat{\xi}_{k,N}), \qquad \text{(C.5)}$$

where $\hat{w}_{k,N}$ is the kth component of the vector $\hat{w}_N = N^{-\varrho/2}\hat{U}_{\hat{Z}\hat{W}}^T\hat{Z}^T\hat{W}\hat{W}^T\hat{Y}$.

We want to show that the first part in (C.1) satisfies

$$2\sup_{\lambda\geq 0}\{\log\hat{L}_{\hat{Y}}(\lambda) - \log\hat{L}_{\hat{Y}}(0)\} \Rightarrow \sup_{\lambda\geq 0} LRT_\infty(\lambda), \qquad \text{(C.6)}$$

where $LRT_\infty(\lambda)$ is defined under equation (4.9). For this purpose, we define

$$\hat{f}_N(\lambda') =: \sum_{k=1}^{K}\frac{\lambda'\hat{w}_{k,N}^2}{1+\lambda'\hat{\zeta}_{k,N}} - \sum_{k=1}^{K}\log(1+\lambda'\hat{\xi}_{k,N}). \qquad \text{(C.7)}$$

According to equation (C.5), $2\sup_{\lambda\geq 0}\{\log\hat{L}_{\hat{Y}}(\lambda) - \log\hat{L}_{\hat{Y}}(0)\} = \sup_{\lambda'\geq 0}\hat{f}_N(\lambda')$. Hence it suffices to show that $\sup_{\lambda'\geq 0}\hat{f}_N(\lambda') \Rightarrow \sup_{\lambda'\geq 0} LRT_\infty(\lambda')$. This proof consists of two steps:

(S1) $\hat{f}_N(\lambda')$ converges weakly to $LRT_\infty(\lambda')$ on the space of $C[0,\infty]$;

(S2) a continuous Mapping Theorem type result holds for $\sup_{\lambda'\geq 0}\hat{f}_N(\lambda')$.

With respect to the first step (S1), it suffices to establish the finite dimensional convergence of $\hat{f}_N(\lambda')$ to $LRT_\infty(\lambda')$ and show that $\hat{f}_N(\lambda')$ is a tight sequence according to page 54 of Billingsley (1968). Note that the definition of $LRT_\infty(\lambda')$ is defined similarly to that of $\hat{f}_N(\lambda')$ except that it replaces $\hat{\zeta}_{k,N}$'s, $\hat{\xi}_{k,N}$'s and $\hat{w}_{k,N}$'s by ζ_k's, ξ_k's and w_k's. By conditions (C2), (C3) and Lemma C.1.1 below, we can conclude that for all k, $\hat{w}_{k,N}\to w_k$ in probability, $\hat{\zeta}_{k,N}\to\zeta_k$ in probability, and $\hat{\xi}_{k,N}\Rightarrow\xi_k$. With the continuous mapping theorem, we establish the finite dimensional convergence of $\{\hat{f}_N(\lambda'_1),\ldots,\hat{f}_N(\lambda'_L)\}$ to $\{LRT_\infty(\lambda'_1),\ldots,LRT_\infty(\lambda'_L)\}$.

To finish the proof of (S1), we need to show that $\hat{f}_N(\lambda')$ form a tight sequence. According to Theorem 8.3 of Billingsley (1968), it suffices to show that for every ε' and η' strictly positive, there exists δ' and N_0 such that for $N\geq N_0$,

$$\frac{1}{\delta'}P\{\sup_{t,t':t\leq t'\leq t+\delta'}|\hat{f}_N(t')-\hat{f}_N(t)|\geq\varepsilon'\}\leq\eta'. \tag{C.8}$$

Since $0\leq t\leq t'$ and all eigenvalues $\hat{\xi}_{k,N}$'s, $\hat{\zeta}_{k,N}$'s are nonnegative,

$$\begin{aligned}
& |\hat{f}_N(t)-\hat{f}_N(t')| \\
= & \left|\sum_{k=1}^K\frac{(t-t')\hat{w}_{k,N}^2}{(1+t\hat{\zeta}_{k,N})(1+t'\hat{\zeta}_{k,N})}+\sum_{k=1}^K\log\frac{1+t'\hat{\xi}_{k,N}}{1+t\hat{\xi}_{k,N}}\right| \\
\leq & \sum_{k=1}^K|t-t|\hat{w}_{k,N}^2+\sum_{k=1}^K\log\left\{1+\frac{(t'-t)\hat{\xi}_{k,N}}{1+t\hat{\xi}_{k,N}}\right\} \\
\leq & \sum_{k=1}^K\delta'\hat{w}_{k,N}^2+\sum_{k=1}^K\delta'\hat{\xi}_{k,N},
\end{aligned}$$

where the last inequality uses the facts that $\log(1+x)<x$ for $x>0$, and that $(t'-t)\hat{\xi}_{k,N}/(1+tN^{-\varrho})\leq(t'-t)\hat{\xi}_{k,N}$ for all k. For any $2K$ random variables

A_i's, if $\sum_{i=1}^{2K} A_i \geq a$, then there exists at least one $A_i \geq a/(2K)$. By Bonferroni inequality, $P(\sum_{i=1}^{2K} A_i \geq a) \leq \sum_{i=1}^{2K} P\{A_i \geq a/(2K)\}$. Hence

$$\begin{aligned} & P\{\sup_{t,t':t\leq t'\leq t+\delta'} |f_N(t') - f_N(t)| \geq \varepsilon'\} \\ \leq & \sum_{k=1}^{K} P\{\hat{w}_{k,N}^2 \geq \varepsilon'/(2K\delta')\} + \sum_{k=1}^{K} P\{\hat{\xi}_{k,N} \geq \varepsilon'/(2K\delta')\}. \end{aligned} \tag{C.9}$$

Consider the second term in (C.9). For any δ' and any N,

$$\begin{aligned} & P\{\hat{\xi}_{k,N} \geq \varepsilon'/(2K\delta')\} \\ \leq & P\{\hat{\xi}_{k,N} > \varepsilon'/(2K\delta'),\ |\hat{\xi}_{k,N} - \xi_k| \leq \varepsilon'\} + P(|\hat{\xi}_{k,N} - \xi_k| > \varepsilon') \\ \leq & P\{\varepsilon'/(2K\delta') < \hat{\xi}_{k,N} \leq \xi_k + \varepsilon'\} + P(|\hat{\xi}_{k,N} - \xi_k| > \varepsilon'). \end{aligned} \tag{C.10}$$

We can choose a sufficiently small δ' such that $\varepsilon'/(2K\delta') > \xi_k + \varepsilon'$, or equivalently, $\delta' < \varepsilon'/\{2K(\xi_k + \varepsilon')\}$, and hence the first term in equation (C.10) is 0. Let δ_1' satisfy inequality (C.14) below, which will be used to control the first term in (C.9). Later we will show that there exists such a δ_1'. Let δ_2' satisfy

$$\delta_2' < \min_{k=1,\ldots,K} [\varepsilon'/\{2K(\xi_k + \varepsilon')\}], \tag{C.11}$$

which is used to control the second term in (C.9). Let $\delta_0' = \min(\delta_1', \delta_2')$. Use the fact that $\hat{\xi}_{k,N} \Rightarrow \xi_k$. There exists N_1 such that when $N \geq N_1$,

$$P\{\hat{\xi}_{k,N} \geq \varepsilon'/(2K\delta')\} \leq P(|\hat{\xi}_{k,N} - \xi_k| > \epsilon') < \delta_0'\eta'/(2K), \quad k = 1, \ldots, K.$$

Hence we conclude that the second term in (C.9) satisfies

$$\sum_{k=1}^{K} P\{\hat{\xi}_{k,N} \geq \varepsilon'/(2K\delta_0')\} \leq K\delta_0'\eta'/(2K) = \delta_0'\eta'/2. \tag{C.12}$$

Next consider the first term in (C.9). Let $F_{k,N}(t)$ and $F_k(t)$ be the c.d.f. of $\hat{w}_{k,N}$ and w_k. Then

$$\begin{aligned} & P\{\hat{w}_{k,N}^2 \geq \frac{\varepsilon'}{2K\delta'}\} = 2\left\{1 - F_{k,N}\left(\sqrt{\frac{\varepsilon'}{2K\delta'}}\right)\right\} \\ \leq & 2\left\{1 - F_k\left(\sqrt{\frac{\varepsilon'}{2K\delta'}}\right)\right\} + 2\left|F_k\left(\sqrt{\frac{\varepsilon'}{2K\delta'}}\right) - F_{k,N}\left(\sqrt{\frac{\varepsilon'}{2K\delta'}}\right)\right|. \end{aligned} \tag{C.13}$$

Denote $x' = \varepsilon'/(2K\delta'\zeta_k)$. Since F_k is the c.d.f. of $N(0, \zeta_k)$, we have

$$\begin{aligned} & 1 - F_k\{\sqrt{\varepsilon'/(2K\delta_0')}\} = (\sqrt{2\pi})^{-1} \int_{x'}^{\infty} \exp(-x^2/2)dx \\ \leq \quad & (\sqrt{2\pi})^{-1} \int_{x'}^{\infty} \exp(-x/2)dx = \frac{2}{\sqrt{2\pi}} \exp(-x'/2), \end{aligned}$$

where the inequality holds whenever $x' \geq 1$. For any given positive constants a_1 and a_2, $a_2 x \exp(-x/a_1) \to 0$ as $x \to \infty$. In other words, for a sufficiently large x, we have $\exp(-x/a_1) \leq (a_2 x)^{-1}$. Hence there exists a sufficiently small δ_1' such that

$$1 - F_k\{\sqrt{\varepsilon'/(2K\delta_0')}\} \leq \delta_1'\eta/(8K). \tag{C.14}$$

Since $\delta_0' < \delta_1'$, we also have $1 - F_k\{\sqrt{\varepsilon'/(2K\delta_0')}\} \leq \delta_0'\eta/(8K)$. Since $\hat{w}_{k,N} \Rightarrow w_k$, we have $|F_{k,N}(t) - F_k(t)| \to 0$. Hence we can choose N_2 such that when $N \geq N_2$,

$$\left| F_k\left(\sqrt{\frac{\varepsilon'}{2K\delta_0'}}\right) - F_{k,N}\left(\sqrt{\frac{\varepsilon'}{2K\delta_0'}}\right) \right| < \delta_0'\eta'/(8K). \tag{C.15}$$

Combine equation (C.13), (C.14) and (C.15). We have

$$\sum_{k=1}^{K} P\Big\{\hat{w}_k^2 \geq \frac{\varepsilon'}{2K\delta_0'}\Big\} \leq \delta_0'\eta'/2. \tag{C.16}$$

Let $N_0 = \max(N_1, N_2)$. When $N \geq N_0$, equation (C.9), (C.12) and (C.16) imply that (C.8) holds. Therefore, $\hat{f}_N(\lambda')$ form a tight sequence.

Apply the same techniques as in the proof of Theorem 3 as Crainiceanu (2003), we can similarly show the continuous Mapping Theorem type result holds for $\sup_{\lambda' \geq 0} \hat{f}_N(\lambda')$ and hence $\sup_{\lambda' \geq 0} \hat{f}_N(\lambda') \Rightarrow \sup_{\lambda' \geq 0} LRT_\infty(\lambda')$. Therefore, we prove equation (C.6) and establish the weak convergence of the first term in (C.1). Next we will consider the second term in (C.1) and show that there exists i.i.d. $N(0,1)$ distributed $\nu_1, \ldots, \nu_\tau$ such that

$$2\log \hat{L}_{\hat{Y}}(0) - 2\log \hat{L}^{0,N} \Rightarrow \sum_{i=1}^{\tau} \nu_i^2, \tag{C.17}$$

where τ is the cardinality of the set Q in the null hypothesis (4.4). Note that equation (C.17) naturally holds if $\tau = 0$. We only need to discuss the case when $\tau > 0$.

Before we simplify the left hand side of equation (C.17), we introduce the following definition. Partition $\beta = (\beta_{(1)}^T | \beta_{(2)}^T)^T$, where $\beta_{(2)}$ contains all β_q^0 for $q \in Q$. Similarly, partition $X = (X_{(1)}|X_{(2)})$ according to the partition of β. We define $\hat{X}_{(i)} = \hat{\Sigma}^{-1/2} X_{(i)}$. For any full-rank matrix A, denote $S_A = A(A^T A)^{-1} A^T$. In the special case when $\tau = p + 1$, $X_{(2)} = X$ and $X_{(1)}$ does not exist, and we let $S_{\hat{X}_{(1)}}$ be the zero matrix $\mathbf{0}_{N \times N}$.

Consider the term $2 \log \hat{L}^{0,N}$ in equation (C.17). Note that all components in $\beta_{(2)}$ are 0. This is a linear model $\hat{Y} = \hat{X}_{(1)} \beta_{(1)} + \hat{e}$, and we have

$$2 \log \hat{L}^{0,N} = -\hat{Y}^T (I_N - S_{\hat{X}_1}) \hat{Y} = -\hat{e}^T (I_N - S_{\hat{X}_1}) \hat{e}, \tag{C.18}$$

where the second equation holds because $(I_N - S_{\hat{X}_1}) \hat{X}_{(1)} = \mathbf{0}_{N \times N}$. Similarly, we have

$$2 \log \hat{L}_{\hat{Y}}(0) = -\hat{e}^T (I_N - S_{\hat{X}}) \hat{e},$$

because $(I_N - S_{\hat{X}}) \hat{X} = (I_N - S_{\hat{X}})(\hat{X}_{(1)} | \hat{X}_{(2)}) = \mathbf{0}_{N \times N}$. Together with equation (C.18), we have,

$$2 \log \hat{L}_{\hat{Y}}(0) - 2 \log \hat{L}^{0,N} = \hat{e}^T (S_{\hat{X}} - S_{\hat{X}_{(1)}}) \hat{e}. \tag{C.19}$$

Note that $S_{\hat{X}} - S_{\hat{X}_{(1)}}$ is a projection matrix. There exists a $N \times \tau$ matrix such that $\hat{W}_0 \hat{W}_0^T = S_{\hat{X}} - S_{\hat{X}_{(1)}}$ and $\hat{W}_0^T \hat{W}_0 = I_\tau$. Denote $\hat{\varpi} = \hat{W}_0^T \hat{e}$. Following the proof of Lemma C.1.1, we can conclude that $\hat{\varpi}_i$'s are asymptotically i.i.d. $N(0, 1)$ under H_0. Therefore, equation (C.17) holds.

Together with equation (C.6), we have

$$LRT_{obs} \Rightarrow \sup LRT_{\infty}(\lambda) + \sum_{i=1}^{\tau} \nu_i^2.$$

Use the same technique in Theorem 3 as Crainiceanu (2003), we can further conclude that $\sup LRT_{\infty}(\lambda)$ and $\sum_{i=1}^{\tau} \nu_i^2$ are independent. Hence Proposition 4.1.1 holds. □

Before proving Lemma C.1.1, introduce a new notation ˜ as opposed to the notation ˆ. This notation indicates that the quantities is defined based on the true covariance Σ rather than the working covariance $\hat{\Sigma}$. To be specific, we mimic the definition of $\hat{Y}, \hat{X}, \hat{Z}$ in equation (4.7) and let $\tilde{Y} = \Sigma^{-1/2}Y$, $\tilde{X} = \Sigma^{-1/2}X$ and $\tilde{Z} = \Sigma^{-1/2}Z$. For any term $\hat{A}$, that is defined based on $\hat{Y}$, $\hat{X}$ and $\hat{Z}$, we define $\tilde{A}$ by replacing $\hat{Y}$, $\hat{X}$ and $\hat{Z}$ with $\tilde{Y}$, $\tilde{X}$ and $\tilde{Z}$. For example, we already defined $\hat{W}$ (see the line above equation (C.3)) such that it satisfies

$$\hat{W}\hat{W}^T = I_{N-p-1} - \hat{X}(\hat{X}^T\hat{X})^{-1}\hat{X}^T, \quad \text{and} \quad \hat{W}^T\hat{W} = I_{N-p-1}. \tag{C.20}$$

Similarly, $\tilde{W}$ satisfies

$$\tilde{W}\tilde{W}^T = I_{N-p-1} - \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T, \quad \text{and} \quad \tilde{W}^T\tilde{W} = I_{N-p-1}. \tag{C.21}$$

In Proposition 4.1.1, we already defined the eigenvalues $\tilde{\zeta}_k$'s and $\tilde{\xi}_k$'s. Note that their definitions do not contradict with the use of the notation ˜ here. For example, $\tilde{\xi}_k$ is the eigenvalue of $N^{-\varrho}Z^T\Sigma^{-1}Z = N^{-\varrho}\tilde{Z}^T\tilde{Z}$, while we know $\hat{\xi}_k$ is that of $N^{-\varrho}Z^T\hat{\Sigma}^{-1}Z = N^{-\varrho}\hat{Z}^T\hat{Z}$.

LEMMA C.1.1. *Let $\hat{\xi}_{k,N}$ be the kth eigenvalue of $N^{-\varrho}\hat{Z}^T\hat{Z}$ and $\hat{\zeta}_{k,N}$ be the kth eigenvalue of $N^{-\varrho}\{\hat{Z}^T\hat{Z} - \hat{Z}^T\hat{X}(\hat{X}^T\hat{X})^{-1}\hat{X}^T\hat{Z}\}$. Let $\hat{w}_N$ be defined as under equation (C.5) and $w = (w_1, \ldots, w_K)^T$ be defined as in Proposition 4.1.1. Assume condition (C2) and (C3) in Proposition 4.1.1. Then for $k = 1, \ldots, K$,*

$$\hat{\xi}_{k,N} \to \xi_k \text{ in probability}, \quad \hat{\zeta}_{k,N} \to \zeta_k, \text{ in probability}. \tag{C.22}$$

Further assume condition (C1). Then

$$\hat{w}_N \Rightarrow w. \tag{C.23}$$

Proof of Lemma C.1.1: Let $\tilde{\xi}_{k,N}$ and $\tilde{\zeta}_{k,N}$ be defined as similar to $\hat{\xi}_{k,N}$, $\hat{\zeta}_{k,N}$. Condition (C3) implies that $\tilde{\xi}_{k,N} \to \xi_k$ and $\tilde{\zeta}_{k,N} \to \zeta_k$ in probability for $k = 1, \ldots, K$. To prove (C.22), we need to show that $\hat{\xi}_k - \tilde{\xi}_k = o_p(1)$ and $\hat{\zeta}_k - \tilde{\zeta}_k = o_p(1)$. By Theorem 8.1-6 in Golub and Van Loan (1996), it suffices to show that

$$||N^{-\varrho}\hat{Z}^T\hat{Z} - N^{-\varrho}\tilde{Z}^T\tilde{Z}|| = o_p(1). \tag{C.24}$$

$$||N^{-\varrho}\hat{Z}^T\hat{W}\hat{W}^T\hat{Z} - N^{-\varrho}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{Z}|| = o_p(1), \tag{C.25}$$

where $||A|| = \sqrt{\sum_i \sum_j a_{ij}^2}$.

Note that $N^{-\varrho}\hat{Z}^T\hat{Z} = N^{-\varrho}Z^T\hat{\Sigma}^{-1}Z$ and $N^{-\varrho}\tilde{Z}^T\tilde{Z} = N^{-\varrho}Z^T\Sigma^{-1}Z$. Since $||N^{-\varrho}Z^TZ|| = O(1)$, condition (C2) implies that equation (C.24) holds. Similarly, we can conclude that

$$||N^{-\varrho'}\hat{X}^T\hat{X} - N^{-\varrho'}\tilde{X}^T\tilde{X}|| = o_p(1). \tag{C.26}$$

By Hölder's inequality, we have $||N^{-\varrho/2-\varrho'/2}\hat{X}^T\hat{Z} - N^{-\varrho/2-\varrho'/2}\tilde{X}^T\tilde{Z}|| = o_p(1)$. Note that

$$N^{-\varrho}\hat{Z}^T\hat{W}\hat{W}^T\hat{Z} = N^{-\varrho}\hat{Z}^T\hat{Z} - (N^{-\frac{\varrho+\varrho'}{2}}\hat{Z}^T\hat{X})(N^{-\varrho'}\hat{X}^T\hat{X})^{-1}(N^{-\frac{\varrho+\varrho'}{2}}\hat{X}^T\hat{Z}), \tag{C.27}$$

and $\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{Z}$ has the same structure except that we replace $\hat{}$ in equation (C.27) by $\tilde{}$. Therefore, equation (C.25) is true and thus equation (C.22) is proved.

We now prove equation (C.23). Let $\tilde{w}_N$ be defined similarly as $\hat{w}_N$ under (C.5). Since $\tilde{W}$ satisfies (C.21),

$$\tilde{W}^T\tilde{X} = \tilde{W}^T\tilde{W}\tilde{W}^T\tilde{X} = \tilde{W}^T\{I_{N-p-1} - \tilde{X}(\tilde{X}^T\tilde{X})^{-1}\tilde{X}^T\}\tilde{X} = \tilde{W}^T 0_{(N-p-1)\times(p+1)}.$$

Under H_0, $\tilde{W}^T\tilde{Y} = \tilde{W}^T(\tilde{X}\beta + \tilde{e}) = \tilde{W}^T\tilde{e}$ and hence

$$\tilde{w}_N = N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{Y} = N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e}. \tag{C.28}$$

Note that $\tilde{e} = \Sigma^{-1/2}e \sim N(\mathbf{0}_{N\times 1}, I_N)$, and

$$\mathrm{var}(\tilde{w}_N) \;=\; N^{-\varrho}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{W}\tilde{W}^T\tilde{Z}\tilde{U}_{\tilde{Z}\tilde{W}} = \tilde{U}^T_{\tilde{Z}\tilde{W}}(N^{-\varrho}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{Z})\tilde{U}_{\tilde{Z}\tilde{W}},$$

where the second equation uses the fact that $\tilde{W}^T\tilde{W} = I_N$. Recall that the kth column of $\tilde{U}_{\tilde{Z}\tilde{W}}$ is the kth eigenvector of $N^{-\varrho}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{Z}$ associated with $\tilde{\zeta}_{k,N}$. Hence $\mathrm{var}(\tilde{w}_N)$ is a diagonal matrix whose (k,k)th element is $\tilde{\zeta}_{k,N}$. By condition (C3), $\tilde{\zeta}_{k,N} \to \zeta_k$ in probability and thus $\tilde{w}_N \Rightarrow w$. Therefore, it suffices to show that $||\hat{w}_N - \tilde{w}_N|| = o_p(1)$.

Recall that $\hat{w} - \tilde{w} = N^{-\varrho/2}\hat{U}^T_{\hat{Z}\hat{W}}\hat{Z}^T\hat{W}\hat{W}^T\hat{e} - N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e}$. Hence we can first show that

$$||N^{-\varrho/2}\hat{Z}^T\hat{W}\hat{W}^T\hat{e} - N^{-\varrho/2}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e}|| = o_p(1). \tag{C.29}$$

Note that

$$\begin{aligned}
\hat{Z}^T\hat{W}\hat{W}^T\hat{e} &= Z^T\hat{\Sigma}^{-1}e - (N^{-\frac{\varrho'}{2}}Z^T\hat{\Sigma}^{-1}X)(N^{-\varrho'}X^T\hat{\Sigma}^{-1}X)^{-1}N^{-\frac{\varrho'}{2}}X^T\hat{\Sigma}^{-1}e,\\
\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e} &= Z^T\Sigma^{-1}e - (N^{-\frac{\varrho'}{2}}Z^T\Sigma^{-1}X)(N^{-\varrho'}X^T\Sigma^{-1}X)^{-1}N^{-\frac{\varrho'}{2}}X^T\Sigma^{-1}e.
\end{aligned}$$

In equation (C.26) and the line below, we have shown that

$$||N^{-\varrho'}\hat{X}^T\hat{X} - N^{-\varrho'}\tilde{X}^T\tilde{X}|| = o_p(1), \quad ||N^{-\frac{\varrho+\varrho'}{2}}\hat{X}^T\hat{Z} - N^{-\frac{\varrho+\varrho'}{2}}\tilde{X}^T\tilde{Z}|| = o_p(1).$$

Condition (C2) implies that

$$||N^{-\varrho/2}Z^T\hat{\Sigma}^{-1}e - N^{-\varrho/2}Z^T\Sigma^{-1}e|| = o_p(1) \tag{C.30}$$

$$||N^{-\varrho'/2}X^T\hat{\Sigma}^{-1}e - N^{-\varrho'/2}X^T\Sigma^{-1}e|| = o_p(1). \tag{C.31}$$

Hence equation (C.29) holds.

Recall that $\hat{U}_{\hat{Z}\hat{W}}$ and $\tilde{U}_{\tilde{Z}\tilde{W}}$ contain all eigenvectors of $N^{-\varrho}\hat{Z}^T\hat{W}\hat{W}\hat{Z}$ and $N^{-\varrho}\tilde{Z}^T\tilde{W}\tilde{W}\tilde{Z}$ respectively. By equation (C.25) and Theorem 8.3-6 of Golub and Van Loan (1996), $||\hat{U}_{\hat{Z}\hat{W}} - \tilde{U}_{\tilde{Z}\tilde{W}}|| = o_p(1)$. Since $\tilde{U}_{\tilde{Z}\tilde{W}}$ is an orthogonal matrix, we have

$$||N^{-\varrho/2}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e}|| = ||N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e}|| = ||w|| = O(1). \tag{C.32}$$

Together with Equation (C.30), $||N^{-\varrho/2}\hat{Z}^T\hat{W}\hat{W}^T\hat{e}|| = O_p(1)$. Hence

$$\begin{aligned} ||\hat{w} - \tilde{w}|| &= ||N^{-\varrho/2}\hat{U}^T_{\hat{Z}\hat{W}}\hat{Z}^T\hat{W}\hat{W}^T\hat{e} - N^{-\varrho/2}\tilde{U}^T_{\tilde{Z}\tilde{W}}\tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e}|| \\ &\leq ||\hat{U}^T_{\hat{Z}\hat{W}} - \tilde{U}^T_{\tilde{Z}\tilde{W}}||||N^{-\varrho/2}\hat{Z}^T\hat{W}\hat{W}^T\hat{e}|| \\ &\quad +||\tilde{U}^T_{\tilde{Z}\tilde{W}}||||N^{-\varrho/2}(\hat{Z}^T\hat{W}\hat{W}^T\hat{e} - \tilde{Z}^T\tilde{W}\tilde{W}^T\tilde{e})|| \\ &= o_p(1). \end{aligned}$$

Therefore, equation (C.23) is true and thus Lemma C.1.1 holds. □

Proof of Proposition 4.2.1 on page 38: Under the dense design, X_i and Z_i are identical respectively, and $\Sigma = I_n \otimes \Sigma_0$, where the (j, j')th element of Σ_0 is $\Gamma(t_j, t_{j'}) + \sigma^2_\epsilon I(t_j = t_{j'})$. It suffices to prove condition (C2′) holds, i.e.

$$a^T\hat{\Sigma}_0^{-1}a - a^T\Sigma_0^{-1}a = o_p(1) \tag{C.33}$$

$$a^T\hat{\Sigma}_0^{-1}\Sigma_0^{1/2}\epsilon_0 - a^T\Sigma_0^{-1}\Sigma_0^{1/2}\epsilon_0 = o_p(1), \tag{C.34}$$

where a is any $m \times 1$ non random vector whose L_2 norm satisfies $||a|| = 1$, $\epsilon_0 = \Sigma_0^{-1/2}e_0$ and $e_0 = n^{-1/2}\sum_{i=1}^n e_i$.

Consider equation (C.33). Let $\hat{G}$ be the $m \times M$ matrix that contain the first M eigenvectors of $\hat{\Sigma}_0$. Note that $\hat{\Sigma}_0 = \hat{\sigma}^2_\epsilon(I_m + \hat{G}\hat{\Lambda}\hat{G}^T)$. By Woodbury matrix identity (Woodbury, 1950),

$$a^T\hat{\Sigma}_0 a = \hat{\sigma}_\epsilon^{-2}a^Ta - \hat{\sigma}_\epsilon^{-2}a^T\hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^Ta, \tag{C.35}$$

$$a^T\Sigma_0 a = \sigma_\epsilon^{-2}a^Ta - \sigma_\epsilon^{-2}a^TG(\Lambda^{-1} + G^TG)^{-1}G^Ta. \tag{C.36}$$

Since $a^T a = O(1)$ and $\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2 = o_p(1)$, we have $(\hat{\sigma}_\epsilon^{-2} - \sigma_\epsilon^{-2})a^T a = o_p(1)$. To prove equation (C.33), it remains to show that for $||a|| = 1$,

$$a^T \hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^T a - a^T G(\Lambda^{-1} + G^T G)^{-1}G^T a = o_p(1). \tag{C.37}$$

Note that the (i,k)th element of $\hat{G}$ and G are $\hat{\theta}_k(t_i)/\sqrt{m}$ and $\theta_k(t_i)/\sqrt{m}$ respectively. By Hölder's inequality and condition (C6), we have

$$||a^T(\hat{G} - G)|| \leq ||a|| \cdot ||\hat{G} - G|| = O_p(n^{-\alpha}). \tag{C.38}$$

Moreover, $\hat{\Lambda}$ and Λ are both diagonal matrices with $\hat{\Lambda}_{ii} = m\hat{\sigma}_k^2/\hat{\sigma}_\epsilon^2$ and $\Lambda_{ii} = m\sigma_k^2/\sigma_\epsilon^2$ respectively. Since $\hat{G}^T\hat{G} = G^T G = I_M$,

$$||(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1} - (\Lambda^{-1} + G^T G)^{-1}|| = o_p(m^{-1}).$$

Together with equation (C.38), we conclude that (C.37) as well as (C.33) holds.

Now consider equation (C.34). Since $G^T G = I_M$, it suffices to show that equation (C.34) holds when a is any column of G or when a is orthogonal to G. First consider the case when $a = G_i$, i.e. the ith column G. By Woodbury matrix identity again,

$$a^T\hat{\Sigma}_0^{-1}\Sigma_0^{1/2}\epsilon_0 = \hat{\sigma}_\epsilon^{-2}a^T\Sigma_0^{1/2}\epsilon_0 - \hat{\sigma}_\epsilon^{-2}a^T\hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^T\Sigma_0^{1/2}\epsilon_0. \tag{C.39}$$

Let $\hat{G}_c\hat{G}_c^T = I_m - \hat{G}\hat{G}^T$. Then

$$\begin{aligned}
G_i^T\hat{\Sigma}_0^{-1}\Sigma_0^{1/2}\epsilon_0 &= \hat{\sigma}_\epsilon^{-2}G_i^T\Sigma_0^{1/2}\epsilon_0 - \hat{\sigma}_\epsilon^{-2}G_i^T\hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^T\Sigma_0^{1/2}\epsilon_0 \\
&= \hat{\sigma}_\epsilon^{-2}G_i^T\hat{G}\hat{G}^T\Sigma_0^{1/2}\epsilon_0 - \hat{\sigma}_\epsilon^{-2}G_i^T\hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^T\Sigma_0^{1/2}\epsilon_0 \\
&\quad +\hat{\sigma}_\epsilon^{-2}G_i^T\hat{G}_c\hat{G}_c^T\Sigma_0^{1/2}\epsilon_0 \\
&= \hat{\sigma}_\epsilon^{-2}G_i^T(\sum_{j=1}^{M}\frac{\hat{\Lambda}_{jj}^{-1}}{1+\hat{\Lambda}_{jj}^{-1}}\hat{G}_j\hat{G}_j^T)\Sigma_0^{1/2}\epsilon_0 + \hat{\sigma}_\epsilon^{-2}G_i^T\hat{G}_c\hat{G}_c^T\Sigma_0^{1/2}\epsilon_0.
\end{aligned}$$

Recall that $\Sigma_0 = \sigma_\epsilon^2(I + G\Lambda G^T)$ and Λ is a diagonal matrix with $\Lambda_{kk} = m\sigma_k^2/\sigma_\epsilon^2$. Hence the first M eigenvalues of $\Sigma_0^{1/2}$ is of order $\sqrt{m}$ and $\hat{\Lambda}_{jj}$ is of order m^{-1}. By

condition (C6), $||G_i^T\hat{G}_c\hat{G}_c^T|| = ||(\hat{G}_i - G_i)^T\hat{G}_c\hat{G}_c^T|| \leq ||\hat{G}_i - G_i|| = O_p(n^{-\alpha})$. Hence

$$G_i^T\hat{\Sigma}_0^{-1}\Sigma_0^{1/2}\epsilon_0 = o_p(m^{-1/2}) + O_p(n^{-\alpha}\sqrt{m}) = o_p(1). \tag{C.40}$$

Note that $G_i^T\Sigma_0^{-1/2}\epsilon_0 = O(m^{-1/2})G_i^T\epsilon_0 = o_p(1)$, equation (C.34) holds when $a = G_i$ for $i = 1, \ldots, M$.

Next consider the case when a is orthogonal to G. Since $G^Ta = \mathbf{0}_{M\times 1}$, we have $a^T\Sigma_0^{1/2} = a^T$ and $||a^T\hat{G}|| = ||a^T(\hat{G} - G)|| = O_p(n^{-\alpha})$. Since equation (C.39) still holds if we replace all estimators with the true parameters, we have,

$$\begin{aligned}
& a^T\hat{\Sigma}_0^{-1}\Sigma_0^{1/2}\epsilon_0 - a^T\Sigma_0^{-1}\Sigma_0^{1/2}\epsilon_0 \\
= \;& \hat{\sigma}_\epsilon^{-2}a^T\Sigma_0^{1/2}\epsilon_0 - \hat{\sigma}_\epsilon^{-2}a^T\hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^T\Sigma_0^{1/2}\epsilon_0 \\
& -\sigma_\epsilon^{-2}a^T\Sigma_0^{1/2}\epsilon_0 + \sigma_\epsilon^{-2}a^TG(\Lambda^{-1} + G^TG)^{-1}G^T\Sigma_0^{1/2}\epsilon_0 \\
= \;& (\hat{\sigma}_\epsilon^{-2} - \sigma_\epsilon^{-2})a^T\Sigma_0^{1/2}\epsilon_0 - \hat{\sigma}_\epsilon^{-2}a^T\hat{G}(\hat{\Lambda}^{-1} + \hat{G}^T\hat{G})^{-1}\hat{G}^T\Sigma_0^{1/2}\epsilon_0 + 0 \\
= \;& (\hat{\sigma}_\epsilon^{-2} - \sigma_\epsilon^{-2})a^T\epsilon_0 - O_p(n^{-\alpha})||\hat{G}^T\Sigma_0^{1/2}\epsilon_0|| \\
= \;& O_p(n^{-\alpha}\sqrt{m}) = o_p(1).
\end{aligned}$$

Therefore, condition (C2') holds. Following the proof of Lemma C.1.1, we can similarly conclude that $\hat{\xi}_{k,N} \to \xi_k$ in probability, $\hat{\zeta}_{k,N} \to \zeta_k$ in probability and $\hat{w}_N \Rightarrow w$. Hence, the asymptotic null distribution of the pseudo LRT statistic is the same as that in Proposition 4.1.1. Proposition 4.2.1 is true. □

Proof of Proposition 4.2.2 on page 41: Under the sparse design, Σ is a block diagonal matrix, whose ith block is the $m_i \times m_i$ dimensional matrix Σ_i and can be written as $\Sigma_i = \sigma_\epsilon^2(I_{m_i} + G_i\Lambda_iG_i)^T$, where the (j,k)th element of G_i is $\theta_j(t_{ik})$ and Λ_i is a diagonal matrix whose (k,k)th component is $\sigma_k^2/\sigma_\epsilon^2$. Note that G_i and Λ_i are defined differently from those in the proof of Porposition 4.2.1. Similarly, we can define $\hat{\Sigma}_i = \hat{\sigma}_\epsilon^2(I_{m_i} + \hat{G}_i\hat{\Lambda}_i\hat{G}_i)^T$. Partition a and e in condition (C2) into n

vectors, with a_i and e_i of length m_i. We want to prove condition (C2),

$$\sum_{i=1}^{n} a_i^T \hat{\Sigma}_i^{-1} a_i - \sum_{i=1}^{n} a_i^T \Sigma_i^{-1} a_i = o_p(1), \tag{C.41}$$

$$\sum_{i=1}^{n} a_i^T \hat{\Sigma}_i^{-1} e_i - \sum_{i=1}^{n} a_i^T \Sigma_i^{-1} e_i = o_p(1). \tag{C.42}$$

First consider equation (C.41). Let $\rho(A)$ be the spectral radius of any matrix A. By Corollary 6.1.5 of Horn and Johnson (1985),

$$\rho(A) \leq \min\{\sup_i \sum_j |A_{ij}|, \sup_j \sum_i |A_{ij}|\}.$$

Since $||a|| = 1$, $a^T A a \leq \rho(A)$ for any symmetric matrix A. Using condition (C6') and the fact that m_i is bounded, we conclude that

$$|a^T(\hat{\Sigma}^{-1} - \Sigma^{-1})a| \leq \rho(\hat{\Sigma}^{-1} - \Sigma^{-1}) = \rho\{\hat{\Sigma}^{-1}(\Sigma - \hat{\Sigma})\Sigma^{-1}\} = O_p(n^{-\alpha}) = o_p(1).$$

Hence equation (C.41) holds. Next consider equation (C.42). By Woodbury matrix identity again,

$$\begin{aligned}
\sum_{i=1}^{n} a_i^T \hat{\Sigma}_i^{-1} e_i &= \hat{\sigma}_\epsilon^2 \sum_{i=1}^{n} a_i^T e_i - \hat{\sigma}_\epsilon^2 \sum_{i=1}^{n} a_i^T \hat{G}_i (\hat{\Lambda}_i^{-1} + \hat{G}_i^T \hat{G}_i)^{-1} \hat{G}_i^T e_i. \\
\sum_{i=1}^{n} a_i^T \Sigma_i^{-1} e_i &= \sigma_\epsilon^2 \sum_{i=1}^{n} a_i^T e_i - \sigma_\epsilon^2 \sum_{i=1}^{n} a_i^T G_i (\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T e_i.
\end{aligned}$$

Since $||a|| = 1$, $\sum_{i=1}^{n} a_i^T e_i = O_p(1)$. Moreover, $\hat{\sigma}_\epsilon^2 \to \sigma_\epsilon^2$ in probability. Therefore, $(\hat{\sigma}_\epsilon^2 - \sigma_\epsilon^2) \sum_{i=1}^{n} a_i^T e_i = o_p(1)$. It remains to show that

$$\sum_{i=1}^{n} a_i^T \{\hat{G}_i (\hat{\Lambda}_i^{-1} + \hat{G}_i^T \hat{G}_i)^{-1} \hat{G}_i^T - G_i (\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T\} e_i = o_p(1).$$

By equation (4.12), $e_i = G_i \gamma_i + \epsilon_i = \sum_{j=1}^{M} G_{ij} \gamma_{ij} + \epsilon_i$, where G_{ij} is the jth column of G_i and $\gamma_i = (\gamma_{i1}, \ldots, \gamma_{iM})^T$ contains all principal scores for subject i. It suffices to show that for $j = 1, \ldots, M$,

$$\sum_{i=1}^{n} a_i^T \{\hat{G}_i (\hat{\Lambda}_i^{-1} + \hat{G}_i^T \hat{G}_i)^{-1} \hat{G}_i^T - G_i (\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T\} G_{ij} \gamma_{ij} = o_p(1), \tag{C.43}$$

$$\sum_{i=1}^{n} a_i^T \{\hat{G}_i (\hat{\Lambda}_i^{-1} + \hat{G}_i^T \hat{G}_i)^{-1} \hat{G}_i^T - G_i (\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T\} \epsilon_i = o_p(1). \tag{C.44}$$

The proof of (C.43) and (C.44) are similar. We will only demonstrate how to prove equation (C.44) here. Since all eigenvalues and eigenfunctions are consistently estimated, we only need to prove that the dominant terms on the left hand side of equation (C.44) are all $o_p(1)$, i.e.

$$\sum_{i=1}^{n} a_i^T(\hat{G}_i - G_i)(\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T \epsilon_i = o_p(1) \tag{C.45}$$

$$\sum_{i=1}^{n} a_i^T G_i(\Lambda_i^{-1} + G_i^T G_i)^{-1}(\hat{G}_i - G_i)^T \epsilon_i = o_p(1) \tag{C.46}$$

$$\sum_{i=1}^{n} a_i^T G_i\{(\hat{\Lambda}_i^{-1} + \hat{G}_i^T \hat{G}_i)^{-1} - (\Lambda_i^{-1} + G_i^T G_i)^{-1}\} G_i^T \epsilon_i = o_p(1). \tag{C.47}$$

The key idea is to use the fact that the observation time points are chosen from $(t_1, \ldots, t_m)$. If $t_{il} = t_{i'l'} = t_j$, then $\hat{G}_{ik,l} = \hat{G}_{i'k,l'} = \hat{\theta}_k(t_j)$, where $\hat{G}_{ik,l}$ and $\hat{G}_{i'k,l'}$ are the lth and l'th elements of $\hat{G}_{ik}$ and $\hat{G}_{i'k}$. Hence we can rearrange the summation over i into summation over j to complete the proof.

First consider equation (C.45). Let $\epsilon_i' = (\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T \epsilon_i$. Then

$$\begin{aligned}
& \sum_{i=1}^{n} a_i^T(\hat{G}_i - G_i)(\Lambda_i^{-1} + G_i^T G_i)^{-1} G_i^T \epsilon_i \\
= & \sum_{i=1}^{n}\sum_{l=1}^{m_i}\sum_{k=1}^{M} a_{il}(\hat{G}_{ik,l} - G_{ik,l})\epsilon_{ik}' \\
= & \sum_{k=1}^{M}\sum_{j=1}^{m}\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}\epsilon_{ik}'(\hat{G}_{ik,l} - G_{ik,l}) I_{t_{il}=t_j} \\
= & \sum_{k=1}^{M}\sum_{j=1}^{m}\{\hat{\theta}_k(t_j) - \theta_k(t_j)\}\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}\epsilon_{ik}' I_{t_{il}=t_j}.
\end{aligned} \tag{C.48}$$

Define $S_j = \sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}\epsilon_{ik}' I_{t_{il}=t_j}$. Since M is finite, it suffices to show that

$$E[\sum_{j=1}^{m}\{\hat{\theta}_k(t_j) - \theta_k(t_j)\}S_j]^2 = o(1). \tag{C.49}$$

By condition (C6′), $\sup_{t\in[0,1]} |\hat{\theta}_k(t) - \theta_k(t)| = O(n^{-\alpha})$. Hence

$$E[\sum_{j=1}^{m}\{\hat{\theta}_k(t_j) - \theta_k(t_j)\}S_j]^2 \le n^{-2\alpha} E(\sum_{j=1}^{m}|S_j|)^2 \le n^{-2\alpha} m \sum_{j=1}^{m} E(S_j^2). \tag{C.50}$$

Since the observation time points are sampled uniformly without replacement, we have, $t_{il} \neq t_{il'}$ if $l \neq l'$ and $E(I_{t_{il}=t_j}) = m^{-1}$. Moreover, ϵ'_{ik}'s are independent across i and they are independent of t_{ik}'s. Hence

$$E(S_j^2) = E\{\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}^2(\epsilon'_{ik})^2 I_{t_{il}=t_j}\} = m^{-1}E\{\sum_{i=1}^{n}\sum_{l=1}^{m_i} a_{il}^2(\epsilon'_{ik})^2\} = O(m^{-1}). \quad \text{(C.51)}$$

Note that equation (C.51) holds for all j. Together with equation (C.50) and condition (C6'),

$$E[\sum_{j=1}^{m}\{\hat{\theta}_k(t_j) - \theta_k(t_j)\}S_j]^2 = O(n^{-2\alpha}m) = o(1),$$

and equation (C.45) is proved.

Now we want to prove equation (C.47). Direct calculations show that

$$\begin{aligned}
&\sum_{i=1}^{n} a_i^T G_i\{(\hat{\Lambda}_i^{-1} + \hat{G}_i^T\hat{G}_i)^{-1} - (\Lambda_i^{-1} + G_i^T G_i)^{-1}\}G_i^T\epsilon_i \\
= &\sum_{i=1}^{n} a_i^T G_i(\hat{\Lambda}_i^{-1} + \hat{G}_i^T\hat{G}_i)^{-1}(\Lambda_i^{-1} + G_i^T G_i - \hat{\Lambda}_i^{-1} - \hat{G}_i^T\hat{G}_i)(\Lambda_i^{-1} + G_i^T G_i)^{-1}G_i^T\epsilon_i.
\end{aligned}$$

Note that its dominant term is

$$\begin{aligned}
&\sum_{i=1}^{n} a_i^T G_i(\Lambda_i^{-1} + G_i^T G_i)^{-1}(\hat{\Lambda}_i^{-1} + \hat{G}_i^T\hat{G}_i - \Lambda_i^{-1} - G_i^T G_i)(\Lambda_i^{-1} + G_i^T G_i)^{-1}G_i^T\epsilon_i \\
= &\sum_{i=1}^{n} a'_i(\hat{\Lambda}_i^{-1} + \hat{G}_i^T\hat{G}_i - \Lambda_i^{-1} - G_i^T G_i)\epsilon'_i \qquad \text{(C.52)}
\end{aligned}$$

Applying the same techniques as above, we conclude that

$$\sum_{i=1}^{n} a'_i(\hat{G}_i - G_i)^T G_i\epsilon'_i = o_p(1), \qquad \sum_{i=1}^{n} a'_i G_i^T(\hat{G}_i - G_i)\epsilon'_i = o_p(1).$$

Therefore, $\sum_{i=1}^{n} a'_i(\hat{G}_i^T\hat{G}_i - G_i^T G_i)\epsilon'_i = o_p(1)$. Moreover, $\hat{\Lambda}_i$ and Λ_i remain the same for all i. They are both $M \times M$ diagonal matrices whose (i,i)th diagonal elements are $\hat{\sigma}_k^2/\hat{\sigma}_\epsilon^2$ and $\sigma_k^2/\sigma_\epsilon^2$ respectively. Therefore,

$$\sum_{i=1}^{n} a'_i(\hat{\Lambda}_i^{-1} - \Lambda_i^{-1})\epsilon'_i = o_p(1).$$

Hence the term in (C.52) is $o_p(1)$ and equation (C.47) is true. Together with (C.45)–(C.47), we conclude that equation (C.44) holds. Therefore we conclude that condition (C2′) holds and Proposition 4.2.2 is true. □

BIBLIOGRAPHY

Antoniadis, A. and Sapatinas, T. (2007). Estimation and inference in functional mixed-effects models. *Comput. Statist. Data Anal.*, 51(10):4793–4813.

Apanasovich, T., Li, Y., and Ruppert, D. (2010). Asymptotics of penalized splines for general degree splines and general order penalties. *Submitted*.

Bellman, R. (1961). *Adaptive control processes: A guided tour*. Princeton University Press, Princeton, N.J.

Billingsley, P. (1968). *Convergence of probability measures*. John Wiley & Sons Inc., New York.

Borbely, A. A. and Achermann, P. (1999). Sleep homeostasis and models of sleep regulation. *Journal of the Biological Rhythms*, 14:557.

Buja, A., Hastie, T., and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.*, 17(2):453–555.

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009). Asymptotic properties of penalized spline estimators. *Biometrika*, 96(3):529–544.

Crainiceanu, C. M. (2003). *Nonparametric Likelihood Ratio Testing*. Cornell University.

Crainiceanu, C. M., Caffo, B., Di, C., and Punjabi, N. M. (2009). Nonparametric signal extraction and measurement error in the analysis of electroencephalographic activity during sleep. *J. Amer. Statist. Assoc.*

Crainiceanu, C. M. and Ruppert, D. (2004). Likelihood ratio tests in linear mixed models with one variance component. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 66(1):165–185.

de Boor, C. (1978). *A practical guide to splines*, volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

Di, C. and Crainiceanu, C. M. (2010). Multilevel sparse functional principal component analysis. *Submitted*.

Di, C., Crainiceanu, C. M., Caffo, B., and Punjabi, N. M. (2009). Multilevel functional principal component analysis. *Ann. Appl. Statist.*, 3:458–488.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statist. Sci.*, 11(2):89–121. With comments and a rejoinder by the authors.

Fan, J. (1992). Design-adaptive nonparametric regression. *Journal of the American Statistical Association*, 87(420):998–1004.

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized likelihood ratio statistics and Wilks phenomenon. *Ann. Statist.*, 29(1):153–193.

Fan, J. and Zhang, J.-T. (2000). Two-step estimation of functional linear models with applications to longitudinal data. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 62(2):pp. 303–322.

Fan, Y. and Li, Q. (1996). Consistent model specification tests: omitted variables and semiparametric functional forms. *Econometrica*, 64(4):865–890.

Friedman, J. H. (1991). Multivariate adaptive regression splines. *Ann. Statist.*, 19(1):1–141. With discussion and a rejoinder by the author.

Friedman, J. H. and Silverman, B. W. (1989). Flexible parsimonious smoothing and additive modeling. *Technometrics*, 31(1):3–39. With discussions by Trevor Hastie and Douglas M. Hawkins and a reply by the authors.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix computations*. Johns Hopkins Studies in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, third edition.

Greven, S., Crainiceanu, C. M., Caffo, B., and Reich, D. (2010). Longitudinal functional principal component. *Electronic Journal of Statistics*.

Greven, S., Crainiceanu, C. M., Ku"chenhoff, H., and Peters, A. (2008). Restricted likelihood ratio testing for zero variance components in linear mixed models. *Journal of Computational and Graphical Statistics*, 17:870–891.

Guo, W. (2002). Functional mixed effects models. *Biometrics*, 58(1):pp. 121–128.

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

Hall, P. and Opsomer, J. D. (2005). Theory for penalised spline regression. *Biometrika*, 92(1):105–118.

Hall, P. and Wand, M. P. (1996). On the accuracy of binned kernel density estimators. *J. Multivariate Anal.*, 56(2):165–184.

Härdle, W. (1990). *Applied nonparametric regression*, volume 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge.

Härdle, W., Hall, P., and Ichimura, H. (1993). Optimal smoothing in single-index models. *Ann. Statist.*, 21(1):157–178.

Härdle, W. and Mammen, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.*, 21(4):1926–1947.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized additive models*, volume 43 of *Monographs on Statistics and Applied Probability*. Chapman and Hall Ltd., London.

Hong, Y. and White, H. (1995). Consistent specification testing via nonparametric series regression. *Econometrica*, 63(5):1133–1159.

Horn, R. A. and Johnson, C. R. (1985). *Matrix analysis / Roger A. Horn, Charles R. Johnson*. Cambridge University Press, Cambridge [Cambridgeshire] ; New York :.

Ichimura, H. (1993). Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. *J. Econometrics*, 58(1-2):71–120.

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

Jank, W. and Shmueli, G. (2006). Functional data analysis in electronic commerce research. *Statist. Sci.*, 21(2):155–166.

Jiang, J., Fan, Y., and Fan, J. (2010). Estimation in additive models with highly or nonhighly correlated covariates. *Ann. Statist.*, 38(3):1403–1432.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 71(2):pp. 487–503.

Kulasekera, K. B. (1995). Comparison of regression curves using quasi-residuals. *J. Amer. Statist. Assoc.*, 90(431):1085–1093.

Li, Y. and Ruppert, D. (2008). On the asymptotics of penalized splines. *Biometrika*, 95(2):415–436.

Linton, O. and Nielsen, J. P. (1995). A kernel method of estimating structured nonparametric regression based on marginal integration. *Biometrika*, 82(1):93–100.

Linton, O. B. (1997). Efficient estimation of additive nonparametric regression models. *Biometrika*, 84(2):469–473.

Mammen, E., Linton, O., and Nielsen, J. (1999). The existence and asymptotic properties of a backfitting projection algorithm under weak conditions. *Ann. Statist.*, 27(5):1443–1490.

Marx, B. D. and Eilers, P. H. C. (1998). Direct generalized additive modeling with penalized likelihood. *Computational Statistics & Data Analysis*, 28(2):193–209.

Morris, J.S. Brown, P. J., Herrick, R., Baggerly, K. A., and Coombes, K. (2008). Bayesian analysis of mass spectrometry proteomic data using wavelet-based functional mixed models. *Biometrics*, 64:479–489.

Nadaraja, È. A. (1964). Some new estimates for distribution functions. *Teor. Verojatnost. i Primenen.*, 9:550–554.

Newey, W. K. (1994). Kernel estimation of partial means and a general variance estimator. *Econometric Theory*, 10(2):233–253.

Nielsen, J. P. and Linton, O. B. (1998). An optimization interpretation of integration and back-fitting estimators for separable nonparametric models. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 60(1):217–222.

Opsomer, J. D. and Ruppert, D. (1997). Fitting a bivariate additive model by local polynomial regression. *Ann. Statist.*, 25(1):186–211.

O'Sullivan, F. (1986). A statistical perspective on ill-posed inverse problems. *Statist. Sci.*, 1(4):502–527. With comments and a rejoinder by the author.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are unequal. *Biometrika*, 58:545–554.

Quan, S. F., Howard, B. V., Iber, C., Kiley, J. P., Nieto, F. J., O'Connor, G. T., Rapoport, D. M., Redline, S., Robbins, J., Samet, J. M., and Wahl, P. W. (1997). The sleep heart health study: Design, rationale, and methods.

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

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

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003). *Semiparametric regression*, volume 12 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Silverman, B. W. (1984). Spline smoothing: the equivalent variable kernel method. *Ann. Statist.*, 12(3):898–916.

Simonoff, J. S. (1998). *Smoothing Methods in Statistics (Springer Series in Statistics)*. Springer.

Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010). Fast methods for spatially correlated multilevel functional data. *Biostatistics*, 11:177–194.

Stone, C. J. (1985). Additive regression and other nonparametric models. *Ann. Statist.*, 13(2):689–705.

Tjøstheim, D. and Auestad, B. H. (1994). Nonparametric identification of nonlinear time series: projections. *J. Amer. Statist. Assoc.*, 89(428):1398–1409.

Wahba, G. (1990). *Spline models for observational data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.

Wand, M. P. (1999). On the optimal amount of smoothing in penalised spline regression. *Biometrika*, 86(4):936–940.

Wand, M. P. and Ormerod, J. T. (2008). On semiparametric regression with O'Sullivan penalized splines. *Aust. N. Z. J. Stat.*, 50(2):179–198.

Wang, X., Shen, J., and Ruppert, D. (2010). Local asymptotics of p-spline smoothing. *Electronic Journal of Statistics, to appear*.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyā Ser. A*, 26:359–372.

Woodbury, M. A. (1950). Inverting modified matrices. *Memorandum Report Princeton University*.

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

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *J. Amer. Statist. Assoc.*, 97(460):1042–1054.