Selecting the Number of Principal Components in Functional Data

Yehua Li
Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011, yehuali@iastate.edu

Naisyin Wang
Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, nwangaa@umich.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, carroll@stat.tamu.edu

Abstract

Functional principal component analysis (FPCA) has become the most widely used dimension reduction tool for functional data analysis. We consider functional data measured at random, subject-specific time points, contaminated with measurement error, allowing for both sparse and dense functional data, and propose novel information criteria to select the number of principal component in such data. We propose a Bayesian information criterion based on marginal modeling that can consistently select the number of principal components for both sparse and dense functional data. For dense functional data, we also developed an Akaike information criterion (AIC) based on the expected Kullback-Leibler information under a Gaussian assumption. In connecting with factor analysis in multivariate time series data, we also consider the information criteria by Bai & Ng (2002) and show that they are still consistent for dense functional data, if a prescribed undersmoothing scheme is undertaken in the FPCA algorithm. We perform intensive simulation studies and show that the proposed information criteria vastly outperform existing methods for this type of data. Surprisingly, our empirical evidence shows that our information criteria proposed for dense functional data also perform well for sparse functional data. An empirical example using colon carcinogenesis data is also provided to illustrate the results.

Key Words: Akaike information criterion; Bayesian information criterion; Functional data analysis; Kernel smoothing; Principal components;

Short title: Model Selection in Functional Data

1 Introduction

Advances in technology has made functional data (Ramsay and Silverman, 2005) increasingly available in many scientific fields, such as many longitudinal data in medical, biological research, electroencephalography (EEG) and functional magnetic resonance imaging (fMRI) data. There is tremendous research interest in functional data analysis (FDA) for the past decade. Among the newly developed methodology, functional principal component analysis (FPCA) has become the most widely used dimension reduction tool for functional data analysis. There is some existing work on selecting the number of functional principal components, but to the best of our knowledge, none of them were rigorously studied either theoretically or empirically. In this paper, we consider functional data that are observed at random, subject-specific observation times, allowing for both sparse and dense functional data. We propose novel information criteria to select the number of principal components, and investigate their theoretical and empirical performance.

There are two main streams of methods for FPCA, kernel based FPCA methods including Yao, Müller and Wang (2005a), Hall, Müller and Wang (2006), and spline based methods including Rice and Silverman (1991), James and Hastie (2001), and Zhou, Huang and Carroll (2008). Some applications of FPCA include Functional Generalized Linear Models, (Müller and Studtmüller, 2005; Yao, Müller and Wang, 2005b; Cai and Hall, 2005; Li, Wang and Carroll, 2010) and Functional Sliced Inverse Regression (Li and Hsing, 2010a).

At this point, the kernel based FPCA methods are better understood in terms of theoretical properties. This is due to the work of Hall and Hosseini-Nasab (2006), who proved various asymptotic expansions of the estimated eigenvalues and eigenfunction for dense functional data, and by Hall et al. (2006) who provided the optimal convergence rate of FPCA in sparse functional data. An important result of Hall et al. (2006) was that, although FPCA is applied to the covariance function estimated by a two dimensional smoother, when the bandwidths were properly tuned, estimating the eigenvalues is a semiparametric problem and enjoys a root n convergence rate, and estimating the eigenfunctions is a nonparametric problem with the convergence rate of a one dimensional smoother.

In the work on FDA mentioned above, functional data were classified as (a) dense functional data where the curves are densely sampled so that passing a smoother on each curve can effectively recover the true sample curves (Hall et al., 2006); and (b) sparse functional data where the number of observations per curve is bounded by a finite number and pooling all subjects together is required to obtain consistent estimates of the principal components (Yao et al., 2005a; Hall et al., 2006). There has been a gap in methodologies for dealing with these two types of data. Hall et al. (2006) showed that when the number of observations per curve diverges to ∞ with a rate of at least $n^{1/4}$, the pre-smoothing approach is justifiable and the errors in smoothing each individual curve are asymptotically negligible. However, in reality it is hard to decide when the observations are dense enough. In some longitudinal studies it is possible that we have dense observations on some subjects and sparse observations on the others. In view of these difficulties, Li and Hsing (2010b) studied all types of functional data in a unified framework, and derived a strong uniform convergence rate for FPCA, where the number of observations per curve can be of any rate relative to the sample size.

A common finding in the aforementioned work is that higher order principal components are much harder to estimate and harder to interpret. Because seeking sparse representation of the data is at the core of modern statistics, it is reasonable in many situations to model the high order principal components as noise. Therefore, selecting the number of principal components is an important model selection problem in almost all practical contexts of FDA. Yao et al. (2005a) proposed an AIC criterion for selecting the number of principal components in sparse functional data. However, so far there is no theoretical justification for this approach, and whether this criterion also works for dense functional data or the types of data in the grey zone between sparse and dense functional data remains unknown. Hall and

Vial (2006) included theoretical discussion about the difficulty of selecting the number of principal components using a hypothesis testing approach. The bootstrap approach proposed by Hall and Vial provides a confidence lower bound $\widehat{v}_q$ for the "unconfounded noise variance", and can provide some guidance in selecting the number of principal components. However, their approach is not a real model selection criterion, and one needs to watch the decreasing trend of $\widehat{v}_q$ and decide the cut point subjectively. The minimum description length (MDL) method by Poskitt and Sengarapillai (2011) is similar to Yao's AIC in that each principal component is counted as one parameter, although of course the criteria are numerically different. We emphasize that, in reality, each principal component consists of one variance parameter and one nonparametric function. A main point of our paper is to justify how much penalty is needed in a model selection criterion, when selecting the number of nonparametric components in the data.

We approach this problem from three directions, with all approaches built upon the foundation of information criteria. In the marginal modeling approach, we focus on the decay rate of the estimated eigenvalues and develop a Bayesian Information Criterion (BIC) based selection method. The advantages of this approach include that it only uses existing outcomes from FPCA, namely the estimated eigenvalues and the residual variance, and that it is consistent for all types of functional data. As an alternative, we find that, with some additional assumptions, a modified Akaike Information Criterion (AIC) based on conditional likelihood could produce superior numerical outcomes. A referee pointed out to us that when the data are observed densely on a regular grid, where no kernel smoothing is necessary, there is some existing work in the econometrics literature based on a factor analysis model (Bai and Ng, 2002) to select the number of principal components. We study this class of information criteria in our setting and find out that they are still consistent if a specific undersmoothing scheme is carried out in the FPCA method. In addition, we also provide some discussion for the case that the true number of principal components diverges to infinity.

The remainder of the paper is organized as follows. In Section 2, we describe the data structure and the FPCA algorithm. In Sections 3.1 and 3.2, we propose and study the new marginal BIC and conditional AIC criteria, and we investigate the information criteria by Bai and Ng in Section 3.3. The proposed information criteria are tested by simulation studies in Section 4, and applied to an empirical example in Section 5. Some concluding remarks are given in Section 6, where we also provide discussion for the case that the true number of principal components diverges. All proofs are provided in the **Supplementary Material**.

2 Functional principal component analysis

2.1 Data structure and model assumptions

Let $X(t)$ be functional data defined on a fixed interval $\mathcal{T} = [a, b]$, with mean function $\mu(t)$ and covariance function $R(s,t) = \text{cov}\{X(s), X(t)\}$. Suppose the covariance function has the eigen-decomposition $R(s,t) = \sum_{j=1}^{\infty} \omega_j \psi_j(s)\psi_j(t)$, where the ω_j are the nonnegative eigenvalues of $R(\cdot,\cdot)$, which, without loss of generality, satisfy $\omega_1 \geq \omega_2 \geq \cdots > 0$, and the ψ_j are the corresponding eigenfunctions.

Although, in theory, the spectral decomposition of the covariance function consists of infinite number of terms, to motivate practically useful information criteria, it is sensible to assume that there is a finite dimensional true model. Due to the nature of spectral decomposition, the higher order terms are less reliably assessed and their estimates tend to have high variation. Consequently, even though one could assume that there are an infinite number of components, unless the data size is very large, sensible variable selection criteria will still select a relatively small number of components – the first several that can be reasonably assessed. This phenomenon is reflected by the numerical outcomes reported in Table S.7 of the **Supplementary Material**, in which a much-improved performance of BIC is observed when the sample size increases to 2000. The performance of BIC is mostly determined by the accuracy of detecting non-zero eigenvalues and that this detection can be difficult for higher order terms. For the rest of the paper, except for Section 6.2, we assume that the spectral decomposition of R ends at a finite p terms, i.e. $\omega_j = 0$ for $j > p$. Then the Karhunen-Loève expansion of $X(t)$ is

$$X(t) - \mu(t) = \sum_{j=1}^{p} \xi_j \psi_j(t), \tag{1}$$

where $\xi_j = \int \psi_j(t)\{X(t) - \mu(t)\}dt$ has mean zero, with $\mathrm{cov}(\xi_j, \xi_{j'}) = I(j = j')\omega_j$. Let p_0 be the true value of p.

Suppose we sample from n independent sample trajectories, $X_i(\cdot)$, $i = 1, \cdots, n$. It often happens that the observations contain additional random errors and instead we observe

$$W_{ij} = X_i(t_{ij}) + U_{ij}, \quad j = 1, \cdots, m_i, \tag{2}$$

where U_{ij} are independent zero-mean errors, with $\mathrm{var}\{U_i(t)\} = \sigma_u^2$, and the U_{ij} are also independent of $X_i(\cdot)$. Here the (t_{ij}) are random, subject-specific observation times. Suppose t_{ij} has a continuous density $f_1(t)$ with support $\mathcal{T}$. We adopt the framework in Li and Hsing (2010b) so that m_i can be of any rate relative to n. The only assumption on m_i is that all $m_i \geq 2$, so that we can estimate the within-curve covariance matrix. In other words, we allow m_i to be bounded by a finite number as in sparse functional data, or diverging to ∞ as in dense functional data.

2.2 Functional principal component analysis

The functions $\mu(\cdot)$ and $R(\cdot, \cdot)$ can be estimated by local polynomial regression, and then $\psi_k(\cdot)$, ω_k and σ_u^2 can be estimated using the functional principal component analysis method proposed in Yao, et al. (2005a) and Hall, et al. (2006). We now briefly describe the method. We first estimate $\mu(\cdot)$ by a local linear regression, $\widehat{\mu}(t) = \widehat{a}_0$ where $(\widehat{a}_0, \widehat{a}_1) = \mathrm{argmin}_{a_0,a_1} n^{-1}\sum_{i=1}^{n} m_i^{-1} \sum_{j=1}^{m_i} \{W_{ij} - a_0 - a_1(t_{ij} - t)\}^2 K\{(t_{ij} - t)/h_\mu\}$,$K(\cdot)$ is a symmetric density function and h_μ is the bandwidth for estimating μ. Define $C_{XX}(s,t) = \mathrm{E}\{X(s)X(t)\}$ and $M_i = (m_i - 1)m_i$. We denote the bandwidth for estimating $C_{XX}(\cdot, \cdot)$ by h_C and let

$\widehat{C}_{XX}(s,t) = \widehat{b}_0$, where $(\widehat{b}_0, \widehat{b}_1, \widehat{b}_2)$ minimizes

$$n^{-1}\sum_{i=1}^{n} M_i^{-1}\sum_{j=1}^{m_i}\sum_{k\neq j}\{W_{ij}W_{ik} - b_0 - b_1(t_{ij}-s) - b_2(t_{ik}-t)\}^2 K\left(\frac{t_{ij}-s}{h_C}\right) K\left(\frac{t_{ik}-t}{h_C}\right).$$

Then $\widehat{R}(s,t) = \widehat{C}_{XX}(s,t) - \widehat{\mu}(s)\widehat{\mu}(t)$. In addition, (ω_k) and $\{\psi_k(\cdot)\}$ can be estimated from an eigenvalue decomposition of $\widehat{R}(\cdot,\cdot)$ by discretization of the smoothed covariance function, see Rice and Silverman (1991) and Capra and Müller (1997). Let $\sigma_w^2(t) = \mathrm{var}\{W(t)\} = R(t,t) + \sigma_u^2$, and $\widehat{\sigma}_w^2(t) = \widehat{c}_0 - \widehat{\mu}^2(t)$, where, with a given bandwidth, h_σ, $(\widehat{c}_0, \widehat{c}_1)$ minimizes $n^{-1}\sum_{i=1}^{n} m_i^{-1}\sum_{j=1}^{m_i}\{W_{ij}^2 - c_0 - c_1(t_{ij}-t)\}^2 K\{(t_{ij}-t)/h_\sigma\}$. One possible estimator if σ_u^2 is

$$\widetilde{\sigma}_{u,\mathrm{I}}^2 = (b-a)^{-1}\int_a^b \{\widehat{\sigma}_w^2(t) - \widehat{R}(t,t)\}dt. \tag{3}$$

Define $\widehat{\omega}_k$ and $\widehat{\psi}_k(\cdot)$ to be the k^{th} eigenvalue and eigenfunction of $\widehat{R}(s,t)$, respectively. Rates of convergence results for $\widehat{\mu}(\cdot)$, $\widehat{R}(\cdot)$, $\widehat{\sigma}_w^2(\cdot)$, $\widetilde{\sigma}_{u,\mathrm{I}}^2$ and $\widehat{\psi}_k(\cdot)$ are described in the **Supplementary Material**, Section S.1.

3 Methodology

3.1 Marginal Bayesian Information Criterion

In a traditional regression setting with sample size n, parameter size p, and normally distributed errors of mean zero and variance σ_u^2, BIC is commonly defined as

$$\log(\sigma^2) + p\log(n)/n.$$

Considering the model equations (1) and (2), linking the current setup for each subject and then marginalizing over all subjects, we consider a generalized BIC criterion of the structure of

$$\log(\widehat{\sigma}_u^2) + \mathcal{P}_n(p), \tag{4}$$

where $\widehat{\sigma}_u^2$ is an estimate of σ_u^2 by marginally pooling error information from all subjects and $\mathcal{P}_n(p)$ is a penalty term. Even though the concept behind our criterion has been motivated by the traditional BIC in regression setting, there are some marked differences. For example, the ξ_j in model (1) are random. As a result, marginally, there are not np parameters. Further, unlike the traditional regression problems, we do not need to estimate/predict ξ_j. Consequently, the number of parameters in a marginal analysis is not determined by the degrees of freedom of these unknown ξ_j. Inspired by standard BIC, we let the penalty be of the form $\mathcal{P}_n(p) = C_{n,p}p$ and then determine the rate of $C_{n,p}$.

Let $\widehat{\sigma}^2_{u,[p]}$ be the estimator of $\widehat{\sigma}^2_u$ based on the residuals after taking into account of the first p principal components. Define

$$R_{[p]}(s,t) = \textstyle\sum_{j=1}^{p} \omega_j \psi_j(s)\psi_j(t), \quad \widehat{R}_{[p]}(s,t) = \textstyle\sum_{j=1}^{p} \widehat{\omega}_j \widehat{\psi}_j(s)\widehat{\psi}_j(t).$$

If p is the true number of principal components, then $R_{[p]}(s,t) = R(s,t)$. Since $\int_a^b \widehat{\psi}_k^2(t)dt = 1$ for all k, we can estimate σ^2_u by

$$\widehat{\sigma}^2_{[p],\mathrm{marg}} = \frac{1}{b-a}\int \{\widehat{\sigma}^2_w(t) - \widehat{R}_{[p]}(t,t)\}dt = \frac{1}{b-a}\int \widehat{\sigma}^2_w(t)dt - \frac{1}{b-a}\sum_{k=1}^{p}\widehat{\omega}_k. \tag{5}$$

Replacing $\widehat{\sigma}^2_u$ by $\widehat{\sigma}^2_{[p],\mathrm{marg}}$ in (4), the new BIC criterion is given by

$$\mathrm{BIC}(p) = \log(\widehat{\sigma}^2_{[p],\mathrm{marg}}) + \mathcal{P}_n(p). \tag{6}$$

That is, instead of estimating $\widehat{\sigma}^2_{u,[p]}$ from the estimated residuals, we will estimate it from a 'marginal' approach by pooling all subjects together. This way, we avoid estimating the principal component scores and dealing with the estimation errors in them.

Denote $\|\cdot\|$ as the L^2 functional norm, and define $\gamma_{nk} = (n^{-1}\sum_{i=1}^{n} m_i^{-k})^{-1}$, which is the k^{th} harmonic mean of the m_i's. When $m_i = m$ for all i, we have that $\gamma_{n1} = m$ and $\gamma_{n2} = m^2$. For any bandwidth h, define

$$\begin{aligned} \delta_{n1}(h) &= [\{1 + (h\gamma_{n1})^{-1}\}/n]^{1/2}, \\ \delta_{n2}(h) &= [\{1 + (h\gamma_{n1})^{-1} + (h^2\gamma_{n2})^{-1}\}/n]^{1/2}. \end{aligned}$$

We make the following assumptions.

(C.1) The observations time $t_{ij} \sim f_1(t)$, $(t_{ij}, t_{ij'}) \sim f_2(t_1, t_2)$, where f_1 and f_2 are continuous density functions with bounds $0 < m_T \leq f_1(t_1), f_2(t_1, t_2) \leq M_T < \infty$ for all $t_1, t_2 \in \mathcal{T}$. Both f_1 and f_2 are differentiable with bounded (partial) derivatives.

(C.2) The kernel function $K(\cdot)$ is a symmetric probability density function on $[-1, 1]$, and is of bounded variation on $[-1, 1]$. Denote $\nu_2 = \int_{-1}^{1} t^2 K(t)dt$.

(C.3) $\mu(\cdot)$ is twice differentiable and its second derivative is bounded on $[a, b]$.

(C.4) All second-order partial derivatives of $R(s, t)$ exist and are bounded on $[a, b]^2$.

(C.5) There exists $C > 4$ such that $\mathrm{E}(|U_{ij}|^C) + \mathrm{E}\{\sup_{t\in[a,b]} |X(t)|^C\} < \infty$.

(C.6) $h_\mu, h_C, h_\sigma, \delta_{n1}(h_\mu), \delta_{n2}(h_C), \delta_{n1}(h_\sigma) \to 0$ as $n \to \infty$.

(C.7) We have $\omega_1 > \omega_2 > \cdots > \omega_{p_0} > 0$ and $\omega_k = 0$ for all $k > p_0$.

Let $\widehat{p}$ be the minimizer of BIC(p). The following theorem gives a sufficient condition for $\widehat{p}$ to be consistent to p_0.

Theorem 1 *Make assumptions (C.1)-(C.7). Recall that $\mathcal{P}_n(p)$ is the penalty defined in (6), and define $\delta_n^* = h_\mu^2 + \delta_{n1}(h_\mu) + h_C^2 + \delta_{n2}(h_C)$. Suppose the following conditions hold*

(i) for any $p < p_0$, $pr[\text{limsup}_{n\to\infty}\{\mathcal{P}_n(p_0) - \mathcal{P}_n(p)\} \leq 0] = 1$;

(ii) for any $p > p_0$, $pr[\mathcal{P}_n(p) > \mathcal{P}_n(p_0), \limsup_{n\to\infty} \delta_n^*/\{\mathcal{P}_n(p) - \mathcal{P}_n(p_0)\} = 0] = 1$.

Then $\lim_{n\to\infty} pr(\widehat{p} = p_0) = 1$.

By Theorem 1, there is a large range of penalties that can result in a consistent BIC criterion. For example, let $N = \sum_i m_i$ and recall that the penalty term $\mathcal{P}_n(p) = C_{n,p}p$. If we let $C_{n,p} \sim \log(N)\delta_n^*$, it is easy to verify that the conditions in Theorem 1 are satisfied.

We now derive a data-based version of $\mathcal{P}_n(p)$ that satisfies condition (i) and (ii). By Lemma S.1.1 in the **Supplementary Material**, δ_n^* is actually the L^2 convergence rate of

$\widehat{R}(\cdot,\cdot)$, which by Lemma S.1.3 in the **Supplementary Material** is also the bound for the null eigenvalues, $\{\widehat{\omega}_k; k > p_0\}$. In reality, $\|\widehat{R} - R\|$ not only depends on δ_n^* but also on unknown constants depending on the true function $R(\cdot,\cdot)$ and the distribution of W. To make the information criterion data-adaptive, we propose the following penalty

$$\mathcal{P}_{n,\text{adapt}}(p) = \log(N)p\|\widehat{R} - \widehat{R}_{[p]}\|/\widetilde{\sigma}^2_{u,\text{I}}. \tag{7}$$

Justification for (7) is given in the **Supplementary Material**, Section S.2.

3.2 Akaike Information Criterion based on conditional likelihood

The marginal BIC criterion can be computed by using outcomes from FPCA directly and it is consistent. However, its performances heavily rely on the precision in estimating ω_j, particularly when j is near the true number of principle components, p_0. It is known that the estimation of ω_j can deteriorate when j increases. In this subsection, we propose an alternative approach that, by having some additional conditions, allows us to take advantage of the use of likelihood. We consider the principal component scores as random effects, and proposed a new AIC criterion based on the conditional likelihood and estimated principal component scores. Such an approach is referred as conditional AIC in linear mixed models, see Claeskens and Hjort (2008). In an alternative context, Hurvich et al. (1998) proposed an AIC criterion for choosing the smoothing parameters in nonparametric smoothing. The FPCA method is to project the discrete longitudinal trajectories on some nonparametric functions (i.e. the eigenfunctions), and can thus be considered as simultaneously smoothing n curves. The AIC in the FPCA context is connected to that for the nonparametric smoothing problem, but the way of counting the effective number of parameters in the model will be different. Therefore, the penalty in our AIC will also be very different from that of the nonparametric smoothing problem.

Define $\boldsymbol{W}_i = (W_{i1}, \dots, W_{i,m_i})^{\text{T}}$, $\boldsymbol{\mu}_i = \{\mu(t_{i1}), \cdots, \mu(t_{i,m_i})\}^{\text{T}}$ and $\boldsymbol{\psi}_{ik} = \{\psi_k(t_{i1}), \cdots, \psi_k(t_{i,m_i})\}^{\text{T}}$. Under the assumption that there are p non-zero eigenvalues, denote $X_{i,[p]}(t) = \mu(t) +$

$\sum_{j=1}^{p} \xi_{ip}\psi_j(t)$, and $\boldsymbol{X}_{i,[p]} = \{X_{i,[p]}(t_{i1}), \ldots, X_{i,[p]}(t_{i,m_i})\}^{\mathrm{T}} = \boldsymbol{\mu}_i + \Psi_{i,[p]}\boldsymbol{\xi}_{i,[p]}$, where $\Psi_{i,[p]} = (\boldsymbol{\psi}_{i1}, \ldots, \boldsymbol{\psi}_{ip})$ and $\boldsymbol{\xi}_{i,[p]} = (\xi_{i1}, \ldots, \xi_{ip})^{\mathrm{T}}$. Under a Gaussian assumption, the conditional log likelihood of the observed data $\{\boldsymbol{W}_i\}$ given the principal component scores is

$$\begin{aligned} \mathcal{L}_{n,\mathrm{cond}}(p, \boldsymbol{X}_{[p]}, \sigma_u^2) &= \textstyle\sum_{i=1}^{n} \Big\{ -(m_i/2)\log(2\pi\sigma_u^2) - (2\sigma_u^2)^{-1}\|\boldsymbol{W}_i - \boldsymbol{X}_{i,[p]}\|^2 \Big\} \\ &= -(N/2)\log(2\pi\sigma_u^2) - (2\sigma_u^2)^{-1}\textstyle\sum_{i=1}^{n}\|\boldsymbol{W}_i - \boldsymbol{\mu}_i - \Psi_{i,[p]}\boldsymbol{\xi}_{i,[p]}\|^2, \end{aligned} \tag{8}$$

where $N = \sum_i m_i$ and $\boldsymbol{X}_{[p]} = (\boldsymbol{X}_{1,[p]}^{\mathrm{T}}, \ldots, \boldsymbol{X}_{n,[p]}^{\mathrm{T}})^{\mathrm{T}}$.

Following the method proposed by Yao et al. (2005a), we estimate the trajectories by

$$\widehat{X}_{i,[p]}(t) = \widehat{\mu}(t) + \textstyle\sum_{j=1}^{p} \widehat{\xi}_{ij}\widehat{\psi}_j(t), \tag{9}$$

where $\widehat{\mu}(\cdot)$ and $\widehat{\psi}_j(\cdot)$ are the estimators described in Section 2. The estimated principal component scores, $\widehat{\xi}_{ij}$, are given by the principal component analysis through the conditional expectation (PACE) estimator by Yao et al. (2005a). Under the Gaussian model, the best linear unbiased predictor (BLUP) for $\boldsymbol{\xi}_{i,[p]}$ is $\widetilde{\boldsymbol{\xi}}_{i,[p]} = \Lambda_{[p]}\Psi_{i,[p]}^{\mathrm{T}}\Sigma_{i,[p]}^{-1}(\boldsymbol{W}_i - \boldsymbol{\mu}_i)$, where $\Lambda_{[p]} = \mathrm{diag}(\omega_1, \ldots, \omega_p)$, $\Sigma_{i,[p]} = \Omega_{i,[p]} + \sigma_u^2 I_{m_i}$ and $\Omega_{i,[p]} = \Psi_{i,[p]}\Lambda_{[p]}\Psi_{i,[p]}^{\mathrm{T}}$. To estimate $\widetilde{\boldsymbol{\xi}}_{i,[p]}$, the PACE estimator requires a pilot estimator of σ_u^2, for which we can use the integral estimator $\widetilde{\sigma}_{u,\mathrm{I}}^2$ defined in (3). The PACE estimator is given by

$$\widehat{\boldsymbol{\xi}}_{i,[p]} = \widehat{\Lambda}_{[p]}\widehat{\Psi}_{i,[p]}^{\mathrm{T}}\widehat{\Sigma}_{i,[p]}^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i), \tag{10}$$

where $\widehat{\boldsymbol{\mu}}_i$, $\widehat{\Lambda}_{[p]}$ and $\widehat{\Psi}_{i,[p]}$ are the estimates using the FPCA method described in Section 2, and $\widehat{\Sigma}_{i,[p]} = \widehat{\Psi}_{i,[p]}\widehat{\Lambda}_{[p]}\widehat{\Psi}_{i,[p]}^{\mathrm{T}} + \widetilde{\sigma}_{u,\mathrm{I}}^2 I$.

To choose p, Yao et al. (2005a) proposed the pseudo AIC

$$\mathrm{AIC}_{\mathrm{Yao}}(p) = \mathcal{L}_{n,\mathrm{cond}}(p, \widehat{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}_{u,\mathrm{I}}^2) + p, \tag{11}$$

where $\widehat{\boldsymbol{X}}_{[p]}$ is the estimated value of $\boldsymbol{X}_{[p]}$ by interpolating the estimated trajectories defined in (9) on the subject-specific times. By adding a penalty p to the estimated conditional likelihood, Yao et al. essentially counted each principal component as one parameter.

To motivate our own AIC criterion, we consider dense functional data satisfying

$$m_i \asymp m \to \infty \text{ for all } i, \quad \sup_i |m_i - m|/m \to 0. \tag{12}$$

We follow the spirit of the derivation of Hurvich and Tsai (1989), and define the Kullback-Leibler information to be

$$\Delta(p, \widetilde{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}^2) = \mathrm{E}_F\{-2\mathcal{L}_{n,\mathrm{cond}}(p, \widetilde{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}^2)\}, \tag{13}$$

for any fixed $\widetilde{\boldsymbol{X}}_{[p]}$ and $\widetilde{\sigma}^2$, where F is the true normal distribution given the true curves $\{X_i(\cdot), i = 1, \ldots, n\}$. Using similar derivations as in Hurvich and Tsai (1989), for any fixed parameters $\widetilde{\boldsymbol{X}}_{[p]} = \{\widetilde{\boldsymbol{X}}_{i,[p]} = \widetilde{\boldsymbol{\mu}}_i + \widetilde{\Psi}_{i,[p]}\widetilde{\boldsymbol{\xi}}_{i,[p]}\}_{i=1}^n$ and $\widetilde{\sigma}^2$, we have

$$\begin{aligned}\Delta(p, \widetilde{\boldsymbol{X}}_{[p]}, \widetilde{\sigma}^2) &= N\log(2\pi\widetilde{\sigma}^2) + \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^n \mathrm{E}_F\|\boldsymbol{U}_i + \boldsymbol{X}_i - \widetilde{\boldsymbol{X}}_{i,[p]}\| \\ &= N\log(2\pi\widetilde{\sigma}^2) + N\frac{\sigma_u^2}{\widetilde{\sigma}^2} + \frac{1}{\widetilde{\sigma}^2}\sum_{i=1}^n \|(\boldsymbol{\mu}_i - \widetilde{\boldsymbol{\mu}}_i) + \Psi_{i,[p_0]}\boldsymbol{\xi}_{i,[p_0]} - \widetilde{\Psi}_{i,[p]}\widetilde{\boldsymbol{\xi}}_{i,[p]}\|^2. \end{aligned} \tag{14}$$

By substituting in the FPCA and PACE estimators, the estimated variance under the model with p principal components is given by

$$\begin{aligned}\widehat{\sigma}^2_{[p]} &= N^{-1}\textstyle\sum_{i=1}^n \|\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i - \widehat{\Psi}_{i,[p]}\widehat{\boldsymbol{\xi}}_{i,[p]}\|^2 = N^{-1}\sum_{i=1}^n \|(I - \widehat{\Omega}_{i,[p]}\widehat{\Sigma}^{-1}_{i,[p]})(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\|^2 \\ &= N^{-1}\textstyle\sum_{i=1}^n \|\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Sigma}^{-1}_{i,[p]}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\|^2.\end{aligned}$$

Then the Kullback-Leibler information for these estimators is

$$\Delta(p, \widehat{X}_{[p]}, \widehat{\sigma}^2_{[p]}) = N\log(\widehat{\sigma}^2_{[p]}) + \mathcal{A}_n(p), \tag{15}$$

where $\mathcal{A}_n(p) = N\sigma_u^2/\widehat{\sigma}^2_{[p]} + \widehat{\sigma}^{-2}_{[p]}\sum_{i=1}^n \|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i + \Psi_{i,[p_0]}\boldsymbol{\xi}_{i,[p_0]} - \widehat{\Psi}_{i,[p]}\widehat{\boldsymbol{\xi}}_{i,[p]}\|^2$.

To derive the new AIC criterion, we need the following theoretical results to evaluate the expected Kullback-Leibler information. As discussed in Hurvich et al. (1998, page 275), in derivation of AIC, one needs to assume that the true model is included in the family of candidate models, and any model bias is ignored. For example, Hurvich et al. (1998) ignored the smoothing bias when developing AIC for nonparametric regressions. Following the same argument, we will ignore all the biases in $\widehat{\mu}(\cdot)$ and $\widehat{\psi}_k(\cdot)$, and only take into account the variation in the estimators.

Proposition 1 *Under assumptions (C.1)-(C.7), condition (12) and the additional assumption that* $n(h_\mu + h_C) \to \infty$, $\widehat{\sigma}^2_{[p_0]}/\sigma^2_u = N^{-1}\sum_{i=1}^n \sum_{j=p_0+1}^{m_i} \mathcal{X}_{ij} + \mathcal{R}_n$, *where the* $\mathcal{X}_{ij}$ *are independent* χ^2_1 *random variables and* $\mathcal{R}_n = O_p\{\delta^2_{n1}(h_\mu) + \delta^2_{n1}(h_C)\} + o_p(nN^{-1})$. *As a result,* $\widehat{\sigma}_{[p_0]} \to \sigma^2_u$ *in probability as* $n \to \infty$

The next proposition gives the asymptotic expansion for $\mathrm{E}\{\mathcal{A}_n(p_0)\}$.

Proposition 2 *Under the same conditions as in Proposition 1,* $\mathrm{E}\{\mathcal{A}_n(p_0)\} = N + 2np_0 + o(n)$.

Thus, the expected Kullback-Leibler information is $\mathrm{E}_F\{\Delta(p_0, \widehat{X}_{[p_0]}, \widehat{\sigma}^2_{[p_0]})\} = \mathrm{E}_F\{N\log(\widetilde{\sigma}^2_{[p_0]})\} + N + 2np_0 + o(n)$. This justifies defining AIC as

$$\mathrm{AIC}(p) = N\log(\widehat{\sigma}^2_{[p]}) + N + 2np. \tag{16}$$

When $m_i \to \infty$ and p is fixed, an intuitive interpretation for the proposed AIC in (16) is to consider FPCA as a linear regression on the observed data $\boldsymbol{W}_i - \boldsymbol{\mu}_i$ against covariates $(\boldsymbol{\psi}_{i1}, \ldots, \boldsymbol{\psi}_{ip})$ for subject i, and consider the principal component scores as the subject-specific coefficients. By pooling n independent curves together and by adding up the individual AIC, we have a total of np regression parameters and the AIC in (16) coincides with that of a simple linear regression. The biggest difference between our AIC and that of Yao et al. in (11) is the way we count the number of parameters in the model.

3.3 Consistent information criteria

As pointed out by a referee, functional principal component analysis is closely related to factor models in econometrics, where there are some existing information criteria to choose the number of factors consistently (Bai and Ng, 2002). We stress that the data considered in the econometrics literature are multivariate time series data observed on regular time points, while we consider irregularly spaced functional data. The estimator and criteria proposed by Bai and Ng were based on matrix projections, while our FPCA method relies heavily on kernel smoothing and operator theory. As a result, deriving consistent model selection criteria for our problem is technically much more involved.

Inspired by Bai and Ng (2002), we consider two classes of information criteria:

$$PC(p) = \widehat{\sigma}^2_{[p]} + p g_n, \tag{17}$$

$$IC(p) = \log(\widehat{\sigma}^2_{[p]}) + p g_n, \tag{18}$$

where $\widehat{\sigma}^2_{[p]}$ is the error variance estimator used in our AIC (15) and g_n is a penalty. The estimator $\widehat{\sigma}^2_{[p]}$ in Bai and Ng (2002) was a mean squared error based on a simple regression, while our estimator is based on the PACE method involving kernel smoothing and BLUP.

For any $p \leq p_0$, denote $\boldsymbol{\psi}_{[p]}(t) = (\psi_1, \ldots, \psi_p)^{\mathrm{T}}(t)$, $\boldsymbol{\psi}_{[p+1:p_0]} = (\psi_{p+1}, \ldots, \psi_{p_0})^{\mathrm{T}}(t)$, and define the inner product matrices $\mathcal{J}_{1,p} = \int \boldsymbol{\psi}_{[p]}(t)\boldsymbol{\psi}^{\mathrm{T}}_{[p]}(t) f_1(t)dt$, $\mathcal{J}_{2,p} = \int \boldsymbol{\psi}_{[p+1:p_0]}(t)\boldsymbol{\psi}^{\mathrm{T}}_{[p+1:p_0]}(t)$ $f_1(t)dt$ and $\mathcal{J}_{12,p} = \int \boldsymbol{\psi}_{[p]}(t)\boldsymbol{\psi}^{\mathrm{T}}_{[p+1:p_0]}(t) f_1(t)dt$. Put $\Lambda_{[p+1:p_0]} = \mathrm{diag}(\omega_{p+1}, \ldots, \omega_{p_0})$, and

$$\tau_p = \mathrm{tr}\{(\mathcal{J}_{2,p} - \mathcal{J}^{\mathrm{T}}_{12,p}\mathcal{J}^{-1}_{1,p}\mathcal{J}_{12,p})\Lambda_{[p+1:p_0]}\}. \tag{19}$$

Theorem 2 *Suppose τ_p defined at (19) exists and is positive for all $0 \leq p < p_0$. Let $\widehat{p}$ be the minimizer of the information criteria defined in (17) or (18) among $0 \leq p \leq p_{\max}$ with $p_{\max} > p_0$ being a fixed search limit, and define $\varrho_n = h^2_\mu + h^2_C + h^2_\sigma + \delta_{n1}(h_\mu) + \delta_{n2}(h_C) + \delta^2_{n1}(h_\sigma)$. Under the assumptions (C.1) - (C.7) and condition (12), $\lim_{n\to\infty} pr(\widehat{p} = p_0) = 1$ if the penalty function g_n satisfies (i) $g_n \xrightarrow{p} 0$ and (ii) $g_n/(n/N + \varrho^2_n) \xrightarrow{p} \infty$.*

In the factor analysis context, the penalty term in the information criteria proposed by Bai and Ng (2002) converges to 0 with a rate slower than C_n^{-2}, where $C_n = \min(m^{1/2}, n^{1/2})$ translating to our notation. Their rate shows a sense of symmetry in the roles of m and

n. Indeed, when the curves are observed on a regular grid, the data can be arranged into a $n \times m$ matrix $\boldsymbol{W}$, the factor analysis can be carried out by a singular value decomposition of $\boldsymbol{W}$, and hence the roles of m and n are symmetric. For the random design that we consider, we apply nonparametric smoothing along t, not among the subjects. Therefore, m and n play different roles in our rate. Not only does the smoothing make our derivation much more involved, but the fact the within-subject covariance matrices are defined on subject specific time points poses many theoretical challenges. Our proof uses many techniques from perturbation theory of random operators and matrices.

The following corollary shows that when the bandwidths are chosen properly, penalties similar to those in Bai and Ng (2002) can still lead to consistent information criteria.

Corollary 1 *Suppose all conditions in Theorem 2 hold, and* $h_\mu \asymp \max(n,m)^{-c_1}$, $h_C \asymp \max(n,m)^{-c_2}$, $h_\sigma \asymp \max(n,m)^{-c_3}$, *where* $1/4 \le c_1, c_2 \le 1$, $1/4 \le c_3 \le 3/2$. *Then* $\widehat{p}$ *that minimizes* $PC(p)$ *or* $IC(p)$ *is consistent if (i)* $g_n \xrightarrow{p} 0$ *and (ii)* $C_n^2 g_n \xrightarrow{p} \infty$, *where* $C_n = \min(n^{1/2}, m^{1/2})$ *as defined in Bai and Ng (2002).*

Bai and Ng (2002) proposed the following information criteria that satisfy the conditions in Corollary 1,

$$
\begin{aligned}
PC_{p1}(p) &= \widehat{\sigma}^2_{[p]} + p\widehat{\sigma}^2_{\text{pilot}}\Big(\frac{n+m}{nm}\Big)\log\Big(\frac{nm}{n+m}\Big),\\
PC_{p2}(p) &= \widehat{\sigma}^2_{[p]} + p\widehat{\sigma}^2_{\text{pilot}}\Big(\frac{n+m}{nm}\Big)\log(C_n^2),\\
PC_{p3}(p) &= \widehat{\sigma}^2_{[p]} + p\widehat{\sigma}^2_{\text{pilot}}\Big\{\frac{\log(C_n^2)}{C_n^2}\Big\},\\
IC_{p1}(p) &= \log(\widehat{\sigma}^2_{[p]}) + p\Big(\frac{n+m}{nm}\Big)\log\Big(\frac{nm}{n+m}\Big),\\
IC_{p2}(p) &= \log(\widehat{\sigma}^2_{[p]}) + p\Big(\frac{n+m}{nm}\Big)\log(C_n^2),\\
IC_{p3}(p) &= \log(\widehat{\sigma}^2_{[p]}) + p\Big\{\frac{\log(C_n^2)}{C_n^2}\Big\}, \qquad (20)
\end{aligned}
$$

where $\widehat{\sigma}^2_{\text{pilot}}$ is a pilot estimator for σ_u^2. In our setting, we can use $\widetilde{\sigma}^2_{u,\text{I}}$ defined at (3) in place of $\widehat{\sigma}^2_{\text{pilot}}$, and replace m by either the arithmetic or the harmonic mean of m_i's. Under the undersmoothing choices of bandwidths described in Corollary 1, all information criteria in (20) are consistent. One can easily see the similarity between the IC_p criteria and the AIC proposed in (16). In general, the IC_p criteria impose greater penalties to over-fitting than AIC. By comparing AIC with the conditions in Theorem 2 and other consistent criteria we developed, we can see the penalty term in AIC is a little bit small and that explains the non-vanishing chance of overfitting witnessed in our simulation studies, see Section 4.

4 Simulation Studies

4.1 Empirical performance of the proposed criteria

To illustrate the finite sample performance of the proposed methods, we performed various simulation studies. Let $\mathcal{T} = [0, 1]$, and suppose that the data are generated from the model (1) and (2). Let the observation time points $T_{ij} \sim$ Uniform $[0, 1]$, $m_i = m$ for all i and $U_{ij} \sim \text{Normal}(0, \sigma_u^2)$.

We consider the following five scenarios.

Scenario I: Here the true mean function is $\mu(t) = 5(t - 0.6)^2$, the number of principal components is $p_0 = 3$, the true eigenvalues are $(\omega_1, \omega_2, \omega_3) = (0.6, 0.3, 0.1)$, the variance of the error is $\sigma_u^2 = 0.2$ and the eigenfunctions are $\psi_1(t) = 1$, $\psi_2(t) = \sqrt{2}\sin(2\pi t)$, $\psi_3(t) = \sqrt{2}\cos(2\pi t)$. The principal component scores are generated from independent normal distributions, i.e. $\xi_{ij} \sim \text{Normal}(0, \omega_j)$. Here $\omega_3 < \sigma_u^2$.

Scenario II: The data are generated in the same way as in Scenario I, except that we replace the third eigenfunction by a rougher function $\psi_3'(t) = \sqrt{2}\cos(4\pi t)$ so that the covariance function is less smooth, and we let the principal component scores follow a skewed Gaussian mixture model. Specifically, ξ_{ij} has 1/3 probability of following a $\text{Normal}(2\sqrt{\omega_j/3}, \omega_j/3)$ distribution, and 2/3 probability of following $\text{Normal}(-\sqrt{\omega_j/3}, \omega_j)$, for $j = 1, 2, 3$.

Scenario III: Set $\mu(t) = 12.5(t - 0.5)^2 - 1.25$, $\phi_1(t) = 1$, $\phi_2(t) = \sqrt{2}\cos(2\pi t)$, $\phi_3(t) = \sqrt{2}\sin(4\pi t)$, and $(\omega_1, \omega_2, \omega_3, \sigma^2) = (4.0, 2.0, 1.0, 0.5)$. The principal component scores are generated from a Gaussian distribution. Here $\omega_3 > \sigma_u^2$.

Scenario IV: The mean function, eigenvalues, eigenfunction and noise level are set to be the same as in Scenario Ⅲ, but the ξ_{ij}'s are generated from a Gaussian mixture model similar to that in Scenario Ⅱ.

Scenario V: In this simulation, we set $p_0 = 6$, the true eigenvalues are $(4.0, 3.5, 3.0, 2.5, 2.0, 1.5)$ and $\sigma_u^2 = 0.5$. We assume that the principal component scores are normal random variables and let the eigenfunctions be

$$\psi_1(t) = 1; \quad \psi_{2k}(t) = \sqrt{2}\sin(2k\pi t), \text{for } k = 1, 2, 3;$$
$$\psi_{2k+1}(t) = \sqrt{2}\cos(2k\pi t), \text{for } k = 1, 2.$$

In each simulation, we generated $n = 200$ trajectories from the models above, and compared the cases with $m = 5$, 10 and 50. The cases $m = 5$ and $m = 50$ may be viewed as representing sparse and dense functional data, respectively, whereas $m = 10$ represents scenarios between the two extremes. For each m, we apply the FPCA procedure to estimate $\{\mu(\cdot), R(\cdot,\cdot), \omega_k, \psi_k(\cdot), \sigma_w^2(t)\}$, then use the proposed information criteria to choose p. The simulation was then repeated 200 times for each scenario.

The performance of the estimators depends on the choice of bandwidths for $\mu(t)$, $C(s,t)$ and $\sigma_w^2(t)$, and the optimal bandwidths vary with n and m. We picked the bandwidths that are slightly smaller than those minimizing the integrated mean squared error (IMSE) of the corresponding functions, since undersmoothing in functional principal component analysis was also advocated by Hall et al. (2006) and Li and Hsing (2010b).

We consider Yao's AIC, MDL by Poskitt and Sengarapillai (2011), the proposed BIC and AIC in (6) and(16), and the criteria by Bai and Ng in (20). Yao's AIC is calculated using the publicly available PACE package (http://anson.ucdavis.edu/ mueller/data/pace.html), where all bandwidths are data-driven and selected by generalized cross-validation (GCV). The empirical distribution of $\widehat{p}$ under Scenarios I to IV are summarized in Tables 1-3. Since the true number of principal components p_0 is different in Scenario V, the distribution of $\widehat{p}$ is summarized in a separate Table 4.

The proposed BIC method is based on the convergence rate results on the eigenvalues, and does not rely much on the distributional assumptions for X and U. From Tables 1-3, we see that BIC picks the correct number of principal components with high percentage in almost all scenarios, except for the cases where the data are sparse, i.e. $m = 5$. This phenomena is as expected, because it is harder to pick up the correct number of signals from sparse and noisy data.

Compared to BIC, the performance of the proposed AIC method is even more impressive. Although BIC is designed to be a consistent model selector, the AIC method selects the right number of principal component with a higher percentage in most of the cases we considered. This is partially due to the fact that AIC makes more use of the information from the likelihood. Even though the data are non-Gaussian in Scenario II and IV, the AIC still performs better than the BIC, and it shows that both the PACE method and the AIC method are quite robust against mild violation of the Gaussian Assumption. Even though the motivation and theoretical development for the AIC method described in Section 3.2 are for dense functional data, it performs surprisingly well for sparse data, such as the case $m = 5$.

There are six criteria in (20), and we find that the PC_p's and the IC_p's tend to perform similarly. To save journal space, we only provide the results for PC_{p1} and IC_{p1}, and the results for the remaining criteria in (20) can be find in the expanded versions of Tables 1-4 in the **Supplementary Material**. As we can see, these criteria behave similar to the AIC, and they tend to do better only in a few occasions when AIC overestimates p.

For almost all scenarios considered, Yao's AIC hardly ever picks the correct model, with the exception of Scenario V, $m = 5$, which will be discussed in more detail below. When the true number of principal components is 3, Yao's AIC will normally chose a number greater than 5. This phenomenon becomes more severe when the data are dense. For example, when $m = 50$, Yao's AIC almost always pick the maximum order considered, which is 15 in our simulations. The behavior of the MDL by Poskitt and Sengarapillai (2011) is similar to Yao's AIC, and hence these results are only provided in Tables S.2 - S.5 in the **Supplementary Material**.

Scenario V, Table 4 is specially designed to check the performance of the proposed information criteria under the situations where we have a relatively large number of principal components. The proposed criteria worked reasonably well for $m = 10$ and 50, and performed much better than Yao's AIC. The case of $m = 5$ under Scenario V is the only case in all of our simulations that Yao's AIC picks the correct model more often than our criteria. With a closer look at the results, we find an explanation. The true covariance function under Scenario V is quite rough, and the GCV criterion in the PACE package chose a large bandwidth so that the local fluctuations on the true covariance surface are smoothed out. In other words, high frequency signals are smoothed out and treated as noise. In a typical run, the PACE estimates for the eigenvalues are $(4.1736, 2.1350, 1.6697, 1.0009, 0.3978, 0.0476)$ which are far from the truth, (4.0, 3.5, 3.0, 2.5, 2.0, 1.5), and the estimated error variance is 6.519 in contrast to the truth $\sigma_u^2 = 0.5$. It is the combination of seriously underestimating the high order eigenvalues and small penalty in AIC that makes Yao's criterion pick the correct number of principal components. Switching to our undersmoothing bandwidths, these estimates are improved but then Yao's AIC will choose much larger values for p. This case also highlights the difficulty of FPCA when p is large but the data are sparse. Unless we have a very large sample size, estimation of these principal components is very difficult, and comparing the model selection procedures in such a case would not be meaningful.

4.2 Further Simulations

The **Supplementary Material**, Section S.4 contains further simulations, including (a) Expanded results with other model selectors in Tables S.2-S.5; (a) an examination of the sensitivity of the results to the bandwidth (Supplementary Table S.6); (c) the behavior of BIC with much larger sample size (Supplementary Table S.7); and (c) results when the value of m is not constant, i.e., $m_i \neq m$ for all i (Supplementary Table S.8).

5 Data analysis

The colon carcinogenesis data in our study have been analyzed in Li, Wang et al. (2007, 2010) and Baladandayuthapani et al. (2008). The biomarker of interest in this experiment is *p27*, which is a protein that inhibits cell cycle. We have 12 rats injected with carcinogen and sacrificed 24 hours after the injection. Beneath the colon tissue of the rats, there are pore structures called 'colonic crypts'. A crypt typically contains 25 to 30 cells, lined up from the bottom to the top. The stem cells are at the bottom of the crypt, where daughter cells are generated. These daughter cells move towards the top as they mature. We sampled about 20 crypts from each of the 12 rats. The *p27* expression level was measured for each cell within the sampled crypts. As previously noted in the literature (Morris et al. 2001, 2003), the *p27* measurements, indexed by the relative cell location within the crypt, are natural functional data. We have $m =$ 25-30 observations (cells) on each function. As in the previous analyses, we consider *p27* in the logarithmic scale. By pooling data from the 12 rats, we have a total of $n = 249$ crypts (functions). In the literature, it has been noted that there is spatial correlation among the crypts within the same rat (Li et al., 2007, Baladandayuthapani et al., 2008). In this experiment, we sampled crypts sufficiently far apart so that the spatial correlations are negligible, and thus we can assume that the crypts are independent.

We perform the FPCA procedure as described in Section 2, with the bandwidths chosen by leave one curve out cross-validation. The estimated covariance function is given in the top panel of Figure 1. The estimated variance of measurement error by integration is $\widetilde{\sigma}_{u,\mathrm{I}} = 0.103$. In contrast, the top 3 eigenvalues are 0.8711, 0.0197 and 0.0053. Let $k_n = \max\{k;\ \widehat{\omega}_k > 0\}$, then the percentage of variation explained by the kth principal component is estimated by $\widehat{\omega}_k/(\sum_{j=1}^{k_n}\widehat{\omega}_j)$. The percentage of variation explained by the first 7 principal components are $(0.966, 0.022, 0.006, 0.003, 0.002, 0.001, 0.000)$.

We apply the proposed AIC, adaptive BIC, the Bai and Ng criteria (20) and Yao's AIC to the data. All of the proposed methods lead to $p = 3$ principal components, for which the corresponding eigenfunctions are shown in the middle panel Figure 1. As we can see, the first principal component is a constant over time, and the second and third eigenfunctions are essentially linear and quadratic functions. Eigenfunction 4 to 7 are shown in the bottom panel of Figure 1, and they are basically noises and are hard to interpret. We therefore can see that the variation among different crypts can be explained by random quadratic polynomials. Yao's AIC, on the other hand, picked a much large number of principal components, with $p = 9$. This is due to the fact that a much smaller penalty is used in Yao's AIC criterion. We have repeated the data analysis using other choices of bandwidths, and the results are the same.

6 Summary

6.1 Basic Summary

Choosing the number of principal components is a crucial step in functional data analysis. There have been some data-driven procedures proposed in the literature that can be used to choose the number of principal components, but these procedures have not been studied theoretically, nor were they tested numerically as extensively as in this paper.

To promote practically useful model selection criteria, we have assumed that there exists a finite dimensional true model. We found that the consistency of the model selection criteria depends on both the sample size n and the number of repeated measurements m on each curve. We proposed a marginal BIC criterion that is consistent for both dense and sparse functional data, which means m can be of any rate relative to n. In the framework of dense functional data, where both n and m diverge to infinity, we proposed a conditional Akaike information criterion, which is motivated by an asymptotic study of the expected Kullback-Leibler distance under Gaussian assumption.

Following the standard approach of Hurvich et al. (1998), we ignored smoothing biases in developing AIC. Our intensive simulation studies also confirm that bias plays a very small role in model selection. In our simulations in Section 4.2, we tried a wide range of bandwidths and thus increase or decrease the biases in the estimators, but the performance of AIC is almost the same. Intuitively, the models under different numbers of principal components are nested, for a fixed bandwidth the smoothing bias exists in all models that we compare, and therefore variation is a more decisive factor in model selection.

In view of the connection of FPCA with factor analysis in multivariate time series data, we revisited the information criteria proposed by Bai and Ng (2002). Even though our setting is fundamentally different, since we assumed that the observational times are random, and the FPCA estimators depend heavily on nonparametric smoothing and are much more complex than those in Bai and Ng, we show essentially similar information criteria can be constructed. Using perturbation theory of random operators and matrices, and under an under-smoothing scheme prescribed in Section 3.3, we showed that these information criteria are consistent when both n and m go to infinity.

6.2 Discussion of the case $p_0 \to \infty$

Some processes considered as functional data are intrinsically infinite dimensional. In those cases, the assumption of p_0 being finite is a finite sample approximation. As the sample size n increases, we can afford to include more principal components in the model and data analysis. It is helpful to consider that the true dimension p_{0n} increases to infinity as a function of n.

This setting was considered in the estimation of a functional linear model (Cai and Hall, 2006). To the best of our knowledge, no information criteria have been previously studied under this setting.

While allowing $p_{0n} \to \infty$, the convergence rates for $\widehat{\mu}(t)$ and $\widehat{R}(s,t)$ remain the same as those given in Lemma S.1.1 in the **Supplementary Material**, but the convergence rates for $\widehat{\psi}_j(t)$ are affected by the spacing of the true eigenvalues. Following condition (4.2) in Cai and Hall (2006), we assume that for some positive constants C and α,

$$C^{-1}j^{-\alpha} \le \omega_j \le Cj^{-\alpha}, \quad \omega_j - \omega_{j+1} \ge C^{-1}j^{-1-\alpha}, \quad \text{j=1,}\dots, p_{0n}. \tag{21}$$

To ensure that $\sum_j^{p_{0n}} \omega_j < \infty$, we assume that $\alpha > 1$. Define the distances between the eigenvalues, $\delta_j = \min_{k\le j}(\omega_k - \omega_{k+1})$, which is no less than $C^{-1}j^{-1-\alpha}$ under condition (21). By the asymptotic expansion of $\widehat{\psi}_j(t)$, see (2.8) in Hall and Hosseini-Nasab, 2006, one can show that the convergence rate of $\widehat{\psi}_j$ is δ_j^{-1} times those in Lemma S.1.2 in the **Supplementary Material**, i.e.

$$\widehat{\psi}_j(t) - \psi_j(t) = O_p[j^{\alpha+1} \times \{h_\mu^2 + \delta_{n1}(h_\mu) + h_C^2 + \delta_{n1}(h_C) + \delta_{n2}^2(h_C)\}], \quad \text{j=1,}\dots, p_{0n}.$$

Assume that $n, m, p_{0n} \to \infty$, $p_{0n}^{\alpha+1}\varrho_n \to 0$, and $p_{0n}^{\alpha+3}/\min(n,m) \to 0$. Following the proof of Theorem 2, while taking into account the increasing estimation error in $\widehat{\psi}_j(t)$ as j increases and the increasing dimensionality of the design matrix Ψ_i, we can show that

$$\widehat{\sigma}^2_{[p]} = \begin{cases} \sigma_u^2 + \tau_p + O_p(pm^{-1} + N^{-1/2}) + o_p(\tau_p + p^{\alpha+3}\varrho_n^2), & \text{for } p < p_{0n}; \\ \widehat{\sigma}^2_{[p_{0n}]} + O_p(m^{-1} + p_{0n}^{\alpha+3}\varrho_n^2), & \text{for } p \ge p_{0n}, \end{cases} \tag{22}$$

where $\tau_p \asymp \mathrm{tr}(\Lambda_{[p+1:p_{0n}]})$ is analogous to (19) and ϱ_n is as defined in Theorem 2. Since the eigenvalues are decaying to 0, the size of the signal $\tau_p \asymp p^{-\alpha}$ as p increases to p_{0n}. In order to have some hope of choosing p_{0n} correctly, we need τ_p to be greater than the size of the estimation error, which implies that $p_{0n}^{2\alpha+3}\varrho_n^2 \to 0$.

Now, consider the class of information criteria in Section 3.3. Suppose that p_{0n} increases slowly enough so that $p_{0n}^{2\alpha+3}/\min(n,m) \to 0$, and that the penalty term satisfies $\tau_p/(pg_n) \to$

∞ for $p < p_{0n}$ and $pg_n/(m^{-1} + p^{\alpha+3}\varrho_n^2) \to \infty$ for $p > p_{0n}$. Then we can show that the $\widehat{p}$ which minimizes $PC(p)$ or $IC(p)$ is consistent. These conditions translate to

$$p_{0n}^{\alpha+1} g_n \to 0, \quad g_n/(p_{0n}^{-1} m^{-1} + p_{0n}^{\alpha+2}\varrho_n^2) \to \infty. \tag{23}$$

If $p_{0n} = \{\min(m, n)\}^\beta$ where $0 < \beta < 1/(2\alpha + 3)$, one can see that the criteria in (20) do not satisfy the conditions in (23) automatically and hence are not guaranteed to be consistent. An information criterion satisfying condition (23) requires a priori knowledge of the decay rate of the eigenvalues. Developing a data-adaptive information criterion that does not require such a priori knowledge is a challenging topic for future research.

Acknowledgment

Li's research was supported by the National Science Foundation (DMS-1105634, DMS-1317118). Wang's research was supported by a grant from the National Cancer Institute (CA74552). Carroll's research was supported by a grant from the National Cancer Institute (R37-CA057030) and by Award Number KUS-CI-016-04, made by King Abdullah University of Science and Technology (KAUST). The authors thank the associate editor and two anonymous referees for their constructive comments that led to significant improvements in the paper.

References

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models, *Econometrika*, 70, 191-221.

Baladandayuthapani, V., Mallick, B., Hong, M., Lupton, J., Turner, N., and Carroll, R. J. (2008). Bayesian hierarchical spatially correlated functional data analysis with application to colon carcinogenesis. *Biometrics*, 64, 64-73.

Cai, T. and Hall, P. (2006). Prediction in functional linear regression, *Annals of Statistics*, 34, 2159-2179.

Capra, W. B. and Müller, H. G. (1997). An accelerated-time model for response curves. *Journal of the American Statistical Association*, 92, 72-83.

Claeskens, G. and Hjort, N. L. (2008). *Model Selection and Model Averaging*, Cambridge University Press, New York.

Hall, P. and Hosseini-Nasab, M. (2006). On properties of functional principal components analysis, *Journal of the Royal Statistical Society, Series B*, 68, 109-126.

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

Hall, P. and Vial, C. (2006). Assessing the finite dimensionality of functional data, *Journal of the Royal Statistical Society, Series B*, 68, 689-705.

Hurvich, C. M., Simonoff, J. S. and Tsai, C. L. (1998). Smoothing parameter selection in nonparametric regression using an improved Akaike information criterion, *Journal of the Royal Statistical Society, Series B*, 60, 271-293.

Hurvich, C. M. and Tsai, C. L. (1989). Regression of time series model selection in small samples, *Biometrika*, 76, 2, 297-307.

Horn, R. A. and Johnson, C. R. (1985). *Matrix Analysis*, Cambridge University Press, New York.

Kato, T. (1987). Variation of discrete spectra, *Communications in Mathematical Physics*, 111, 501-504.

Li, Y. and Hsing, T. (2010a). Deciding the dimension of effective dimension reduction space for functional and high-dimensional data, *Annals of Statistics*, 38, 3028-3062.

Li, Y. and Hsing, T. (2010b). Uniform convergence rates for nonparametric regression and principal component analysis in functional/longitudinal data, *Annals of Statistics*, 38, 3321-3351.

Li, Y., Wang, N, Hong, M, Turner, N., Lupton, J. and Carroll. R. J. (2007). Nonparametric estimation of correlation functions in spatial and longitudinal data, with application to colon carcinogenesis experiments. *Annals of Statistics*, 35, 1608-1643.

Li, Y., Wang, N. and Carroll, R. J. (2010). Generalized functional linear models with semiparametric single-index interactions, *Journal of the American Statistical Association*, 105, 621-633.

Morris, J. S., Wang, N.,Lupton, J. R., Chapkin, R. S., Turner, N. D., Hong, M. Y. and Carroll, R. J. (2001). Parametric and nonparametric methods for understanding the relationship between carcinogen-induced DNA adduct levels in distal and proximal regions of the colon. *Journal of the American Statistical Association*, 96, 455, 816-826.

Morris, J. S., Vannucci, M., Brown, P. J. and Carroll, R. J. (2003). Wavelet-based nonparametric modeling of hierarchical functions in colon carcinogenesis, *Journal of the American Statistical Association*, 98, 463, 573-583.

Müller, H. -G. and Stadtmüller, U. (2005). Generalized functional linear models, *Annals of Statistics*, 33, 774-805.

Poskitt, D. S. and Sengarapillai, A. (2011). Description length and dimensionality reduction in functional data analysis. *Computational Statistics & Data Analysis*, in press.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis, 2nd Edition.* Springer-Verlag, New York.

Rice, J. and Silverman, B. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves, *Journal of the Royal Statistical Society, Series B*, 53,233-243.

Yao, F., Müller, H. G. and Wang, J. L. (2005a). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, 100, 577-590.

Yao, F., Müller, H. G., and Wang, J. L. (2005b) Functional linear regression analysis for longitudinal data, *Annals of Statistics*, 33, 2873-2903.

Zhou, L., Huang, J. Z. and Carroll, R. J. (2008). Joint modelling of paired sparse functional data using principal components. *Biometrika*, 95, 3, 601-619.

Scenario	Method	$\widehat{p} \leq 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \geq 5$
I	AIC_{PACE}	0.000	0.008	0.000	0.121	0.870
	AIC	0.000	0.405	0.580	0.010	0.005
	BIC	0.155	0.335	0.380	0.115	0.015
	PC_{p1}	0.005	0.565	0.410	0.010	0.010
	IC_{p1}	0.000	0.215	0.735	0.045	0.005
II	AIC_{PACE}	0.000	0.000	0.005	0.125	0.870
	AIC	0.000	0.205	0.630	0.155	0.010
	BIC	0.230	0.395	0.245	0.110	0.020
	PC_{p1}	0.000	0.000	0.375	0.440	0.185
	IC_{p1}	0.000	0.140	0.605	0.210	0.045
III	AIC_{PACE}	0.000	0.025	0.005	0.130	0.840
	AIC	0.000	0.035	0.720	0.170	0.075
	BIC	0.335	0.260	0.325	0.080	0.000
	PC_{p1}	0.000	0.220	0.640	0.075	0.065
	IC_{p1}	0.000	0.005	0.590	0.280	0.125
IV	AIC_{PACE}	0.000	0.015	0.015	0.145	0.825
	AIC	0.000	0.020	0.710	0.185	0.085
	BIC	0.315	0.180	0.410	0.070	0.025
	PC_{p1}	0.000	0.160	0.640	0.095	0.105
	IC_{p1}	0.000	0.015	0.560	0.260	0.165

Table 1: When $m = 5$, displayed are the distributions of the number of selected principal components $\widehat{p}$ for all methods and across Scenarios I-IV. The true number of principal components is 3.

Scenario	Method	$\widehat{p} \leq 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \geq 5$
I	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.005	0.980	0.015	0.000
	BIC	0.000	0.040	0.670	0.255	0.035
	PC_{p1}	0.000	0.040	0.955	0.000	0.005
	IC_{p1}	0.000	0.005	0.985	0.010	0.000
II	AIC_{PACE}	0.000	0.000	0.000	0.005	0.995
	AIC	0.000	0.000	0.710	0.260	0.030
	BIC	0.000	0.170	0.665	0.135	0.030
	PC_{p1}	0.000	0.000	0.570	0.355	0.075
	IC_{p1}	0.000	0.000	0.805	0.185	0.010
III	AIC_{PACE}	0.000	0.015	0.000	0.000	0.985
	AIC	0.000	0.000	0.580	0.400	0.020
	BIC	0.005	0.035	0.770	0.145	0.045
	PC_{p1}	0.000	0.000	0.965	0.030	0.005
	IC_{p1}	0.000	0.000	0.665	0.320	0.015
IV	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.830	0.150	0.020
	BIC	0.010	0.005	0.775	0.190	0.020
	PC_{p1}	0.000	0.000	0.920	0.045	0.035
	IC_{p1}	0.000	0.000	0.900	0.085	0.015

Table 2: When $m = 10$, displayed are the distributions of the number of selected principal components $\widehat{p}$ for all methods and across Scenarios I-IV. The true number of principal components is 3.

Scenario	Method	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	$\widehat{p}\geq 5$
I	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.830	0.150	0.020
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
II	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.630	0.320	0.050
	BIC	0.000	0.000	0.795	0.185	0.020
	PC_{p1}	0.000	0.000	0.955	0.045	0.000
	IC_{p1}	0.000	0.000	0.945	0.055	0.000
III	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.775	0.200	0.025
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
IV	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.945	0.055	0.000
	BIC	0.000	0.000	0.835	0.140	0.025
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000

Table 3: For $m = 50$, displayed are the distributions of the number of selected principal components $\widehat{p}$ for all methods and across Scenarios I-IV. The true number of principal components is 3.

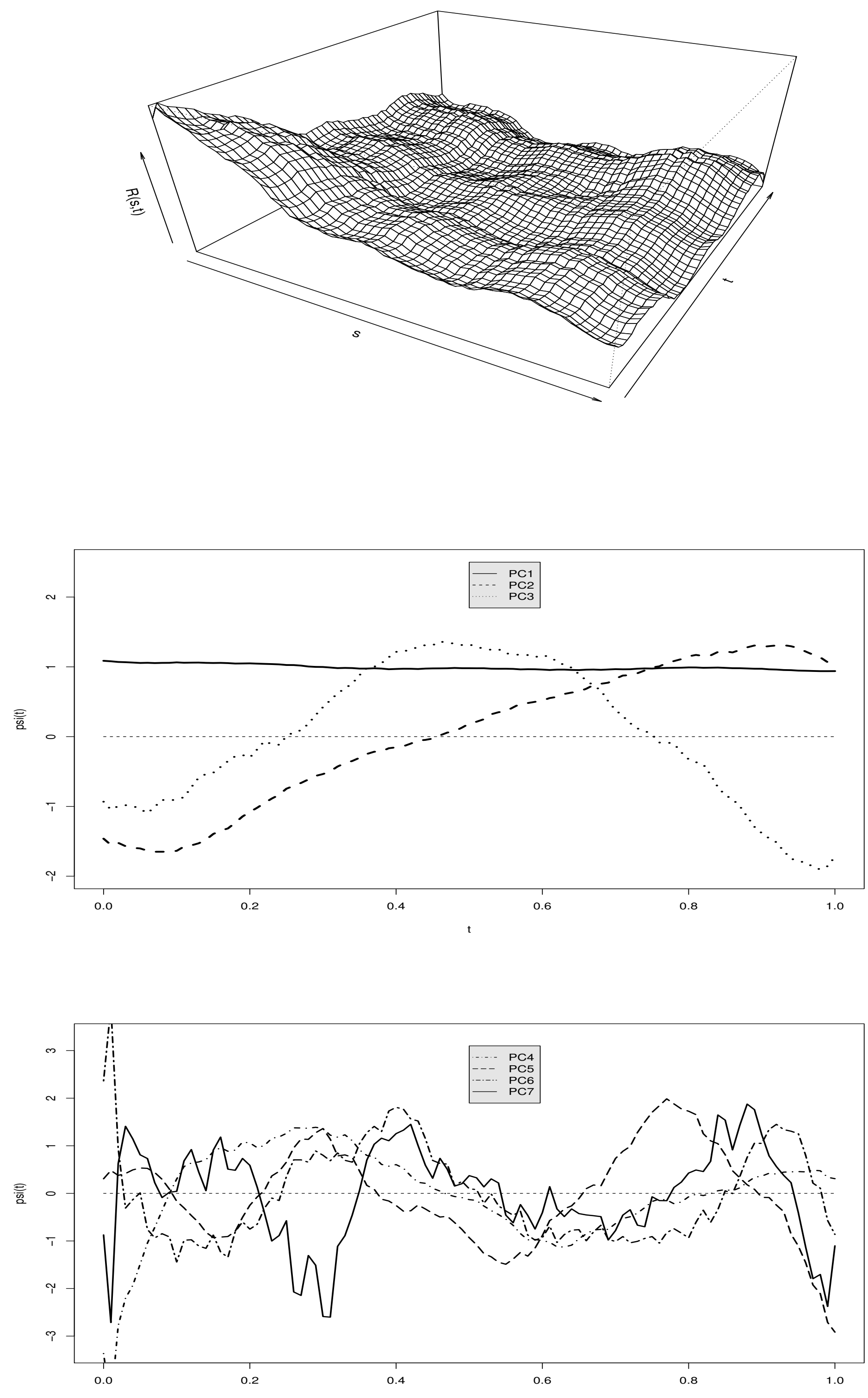

Figure 1: Functional principal component analysis for the colon carcinogenesis *p27* data. Top panel: estimated covariance function; middle panel: the first 3 eigenfunctions; lower panel: eigenfunctions 4-7.

Scenario	Method	$\widehat{p} \leq 4$	$\widehat{p} = 5$	$\widehat{p} = 6$	$\widehat{p} = 7$	$\widehat{p} \geq 8$
m=5	AIC_{PACE}	0.005	0.005	0.705	0.245	0.040
	AIC	0.165	0.330	0.470	0.035	0.000
	BIC	0.835	0.020	0.090	0.050	0.005
	PC_{p1}	0.580	0.345	0.070	0.005	0.000
	IC_{p1}	0.060	0.335	0.545	0.060	0.000
m=10	AIC_{PACE}	0.005	0.000	0.065	0.475	0.455
	AIC	0.000	0.000	0.570	0.280	0.15
	BIC	0.250	0.030	0.525	0.165	0.030
	PC_{p1}	0.000	0.145	0.775	0.020	0.060
	IC_{p1}	0.000	0.000	0.705	0.185	0.110
m=50	AIC_{PACE}	0.000	0.065	0.000	0.000	0.935
	AIC	0.000	0.000	0.260	0.405	0.335
	BIC	0.005	0.000	0.590	0.325	0.080
	PC_{p1}	0.000	0.000	0.980	0.010	0.010
	IC_{p1}	0.000	0.000	0.965	0.035	0.000

Table 4: Distributions of the number of selected principal components $\widehat{p}$ for Scenario V. The true number of principal components is 6.

Supplementary Material to *Selecting the Number of Principal Components in Functional Data*

Yehua Li
Department of Statistics & Statistical Laboratory, Iowa State University, Ames, IA 50011, yehuali@iastate.edu

Naisyin Wang
Department of Statistics, University of Michigan, Ann Arbor, MI 48109-1107, nwangaa@umich.edu

Raymond J. Carroll
Department of Statistics, Texas A&M University, TAMU 3143, College Station, TX 77843-3143, carroll@stat.tamu.edu

S.1 Asymptotic Results for Methods in Section 2.1 of the Main Paper

Lemma **S.1.1** *Under assumptions (C.1)-(C.6),*

$$\begin{aligned}
&\widehat{\mu}(t)-\mu(t)=O_p\{h_\mu^2+\delta_{n1}(h_\mu)\},\\
&\widehat{R}(s,t)-R(s,t)=O_p\{h_\mu^2+\delta_{n1}(h_\mu)+h_C^2+\delta_{n2}(h_C)\},\\
&\widehat{\sigma}_w^2(t)-\sigma_w^2(t)=O_p\{h_\sigma^2+\delta_{n1}(h_\sigma)+h_\mu^2+\delta_{n1}(h_\mu)\}.
\end{aligned}$$

Further, the integration estimator (3) has the convergence rate

$$\widetilde{\sigma}_{u,\mathrm{I}}^2-\sigma_u^2=O_p\{h_C^2+\delta_{n1}(h_C)+\delta_{n2}^2(h_C)+h_\sigma^2+\delta_{n1}^2(h_\sigma)\}.$$

With the same spirit as Bai and Ng (2002), we use the pointwise convergence rates in Lemma S.1.1 to develop the new information criteria, instead of uniform convergence rates. The convergence rates in Lemma S.1.1 are essentially the same as the strong uniform convergence rates proved by Li and Hsing (2010b), without the $\log(n)$ factor that controls the maximum absolute deviation.

Lemma **S.1.2** *Under the assumptions in the appendix, for $j\le p_0$,*

$$\begin{aligned}
&\widehat{\omega}_j-\omega_j=O_p\{n^{-1/2}+h_\mu^2+h_C^2+\delta_{n1}^2(h_\mu)+\delta_{n2}^2(h_C)\}\\
&\widehat{\psi}_j(t)-\psi_j(t)=O_p\{h_\mu^2+\delta_{n1}(h_\mu)+h_C^2+\delta_{n1}(h_C)+\delta_{n2}^2(h_C)\}.
\end{aligned}$$

The proof uses the asymptotic expansion of the eigenvalues and eigenfunctions proved in Hall and Hosseini-Nasab (2006). These expansions only exist for $j \le p_0$. For $j > p_0$, by the model assumption $\omega_j = 0$, and thus the ψ_j's are not even uniquely defined.

Lemma S.1.1 shows that $\widehat{R} - R$ defines a self-adjoint, Hilbert-Schmidt integral operator, which is also compact. The following inequality is a standard result in perturbation theory for compact self-adjoint operators, see Kato (1987).

<u>Lemma</u> **S.1.3** *Under the assumptions in the appendix,* $\sum_{j=1}^{\infty}(\widehat{\omega}_j - \omega_j)^2 \le \|\widehat{R} - R\|^2 = O_p\{h_\mu^4 + \delta_{n1}^2(h_\mu) + h_C^4 + \delta_{n2}^2(h_C)\}$.

Lemma S.1.3 implies that all the null eigenvalues of $\widehat{R}(s,t)$ are small, i.e. for any fixed $j > p_0$, $|\widehat{\omega}_j| \le \|\widehat{R} - R\| = O_p\{h_\mu^2 + \delta_{n1}(h_\mu) + h_C^2 + \delta_{n2}(h_C)\}$.

S.2 Justification of the Penalty (7)

We now provide some heuristic justification for (7). The basic idea is that when p is the correct number of principal components, $\widehat{R}_{[p]}$ is a better estimate of R, therefore $\|\widehat{R} - \widehat{R}_{[p]}\|$ gives us an estimate of $\|\widehat{R} - R\|$, which is the bound of the null eigenvalues. The factor $\widetilde{\sigma}_{u,\mathrm{I}}^2$ defined in (3) is used in the denominator of (7) to make the penalty scale invariant. Following the convention of classic BIC, we include the $\log(N)$ factor in (7) to ensure that the proposed penalty falls in the range of penalties defined in Theorem 1. Then

$$\|\widehat{R} - \widehat{R}_{[p]}\| = \Big[\int\int\Big\{\sum_{k=p+1}^{\infty}\widehat{\omega}_k\widehat{\psi}_k(s)\widehat{\psi}_k(t)\Big\}^2 dsdt\Big]^{1/2} = \Big(\sum_{k=p+1}^{\infty}\widehat{\omega}_k^2\Big)^{1/2}.$$

When $p \ge p_0$, the right hand side only includes the null eigenvalues, and therefore, by Lemma S.1.3, is of order $O_p(\delta_n^*)$. Interestingly, since $\widehat{R}(\cdot,\cdot)$ is not guaranteed to be positive semidefinite, some of the $\widehat{\omega}_k$'s may be negative, but these possible negative eigenvalues are still informative about the L^2 distance between $\widehat{R}$ and R. From our experience in simulation studies, the value of $\|\widehat{R} - \widehat{R}_{[p]}\|$ becomes quite stable when p is large. In other words, when $p > p_0$, further increasing p almost cause no changes in the value of $\|\widehat{R} - \widehat{R}_{[p]}\|$. As a result, for $p > p_0$, $\mathcal{P}_{n,\mathrm{adapt}}(p)$ becomes a monotone increasing function of p. Hence, one can verify that Condition (ii) in Theorem 1 is satisfied.

On the other hand, when $p < p_0$, $\|\widehat{R} - \widehat{R}_{[p]}\|$ includes some of the non-zero eigenvalues, therefore $\mathcal{P}_{n,\mathrm{adapt}}(p) = O_p\{\log(N)\}$. It is easy to verify that $\mathcal{P}_{n,\mathrm{adapt}}(p_0) - \mathcal{P}_{n,\mathrm{adapt}}(p) = O_p\{\log(N)\delta_n^*\} - O_p\{\log(N)\}$ is less or equal to 0 with probability tending to 1. Therefore, Condition (i) in Theorem 1 is also verified.

S.3 Sketch of Technical Arguments

S.3.1 Technical lemmas

LEMMA **S.3.1** *If the conditions above hold and we ignore all biases in nonparametric smoothing, the following asymptotic expansion holds uniformly for all $s, t \in \mathcal{T}$*

$$
\begin{aligned}
\widehat{\mu}(t) - \mu(t) &= \frac{1}{nf_1(t)} \sum_{i=1}^{n} m_i^{-1} \sum_{j=1}^{m_i} K_{h_\mu}(t_{ij} - t)\epsilon_{ij} + o_p\{\delta_{n1}(h_\mu)\};\\
\widehat{C}(s,t) - C(s,t) &= \frac{1}{nf_2(s,t)} \sum_{i=1}^{n} M_i^{-1} \sum_{j \neq j'} \epsilon^*_{i,jj'} K_{h_C}(t_{ij} - s) K_{h_C}(t_{ij'} - t) + o_p\{\delta_{n2}(h_C)\},
\end{aligned}
$$

where $\epsilon_{ij} = W_{ij} - \mu(t_{ij})$ and $\epsilon^*_{i,jj'} = W_{ij}W_{ij'} - C(t_{ij}, t_{ij'})$. Moreover, for $k = 1, \ldots, p_0$,

$$
\begin{aligned}
\widehat{\psi}_k(t) - \psi_k(t) &= \Big\{ \frac{1}{n} \sum_{i=1}^{n} \frac{1}{M_i} \sum_{j \neq j'} \epsilon^*_{i,jj'} \mathcal{G}_{2,k}(t_{ij}, t_{ij'}, t) + \frac{1}{n} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} \epsilon_{ij} \mathcal{G}_{1,k}(t_{ij}, t)\\
&\quad + \omega_k^{-1} \frac{1}{n} \sum_{i=1}^{n} \frac{1}{M_i} \sum_{j \neq j'} K_{h_C}(t_{ij'} - t) \epsilon^*_{i,jj'} \psi_k(t_{ij}) / f_2(t_{ij}, t)\\
&\quad - \omega_k^{-1} \langle \mu, \psi_k \rangle \frac{1}{nf_1(t)} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j=1}^{m_i} K_{h_\mu}(t_{ij} - t)\epsilon_{ij} \Big\}\\
&\quad + o_p\{\log(n) n^{-1/2} + \delta_{n1}(h_\mu) + \delta_{n1}(h_C)\}, \qquad \text{(S.1)}
\end{aligned}
$$

where

$$
\begin{aligned}
\mathcal{G}_{1,k}(t_1, t_2) &= - \sum_{\substack{k'=1 \\ k' \neq k}}^{p_0} \frac{\omega_{k'} \psi_{k'}(t_2)}{(\omega_k - \omega_{k'})\omega_k} \{\langle \mu, \psi_k \rangle \psi_{k'}(t_1) + \langle \mu, \psi_{k'} \rangle \psi_k(t_1)\} / f_1(t_1)\\
&\quad + 2\omega_k^{-1} \langle \mu, \psi_k \rangle \psi_k(t_2) \psi_k(t_1) / f_1(t_1) - \omega_k^{-1} \mu(t_2) \psi_k(t_1) / f_1(t_1),\\
\mathcal{G}_{2,k}(t_1, t_2, t_3) &= \Big\{ \sum_{\substack{k'=1 \\ k' \neq k}}^{p_0} \frac{\omega_{k'} \psi_{k'}(t_3)}{(\omega_k - \omega_{k'})\omega_k} \psi_k(t_1) \psi_{k'}(t_2) - \omega_k^{-1} \psi_k(t_3) \psi_k(t_1) \psi_k(t_2) \Big\} / f_2(t_1, t_2).
\end{aligned}
$$

Proof: The asymptotic expansions for $\widehat{\mu}$ and $\widehat{C}$ come directly from the derivations in Li and Hsing (2010b). Similar to Hall and Hosseini-Nasab (2006), we can show an asymptotic expansion for $\widehat{\psi}_k(t)$,

$$
\begin{aligned}
\widehat{\psi}_k(t) - \psi_k(t) &= \Big\{ \sum_{\substack{k'=1 \\ k' \neq k}}^{p_0} \frac{\omega_{k'} \psi_{k'}(t)}{(\omega_k - \omega_{k'})\omega_k} \int\int (\widehat{R} - R) \psi_k \psi_{k'} - \omega_k^{-1} \psi_j(t) \int\int (\widehat{R} - R) \psi_k \psi_k\\
&\quad + \omega_k^{-1} \int (\widehat{R} - R)(s,t) \psi_k(s) ds \Big\} \times \{1 + o_p(1)\}. \qquad \text{(S.2)}
\end{aligned}
$$

The expansion given in Hall and Hosseini-Nasab (2006) was for the case that $\{\psi_j(t)\}$ form a complete orthonormal basis for the L^2 space. In our case the higher order eigenfunctions are not uniquely defined, and the expansion in (S.2) holds for a finite eigensystem assumed in this paper. When $p_0 \to \infty$, (S.2) is equivalent to the expansion in Hall and Hosseini-Nasab (2006). Since

$$(\widehat{R} - R)(s,t) = (\widehat{C} - C)(s,t) - \mu(s)(\widehat{\mu} - \mu)(t) \times \{1 + o_p(1)\} - (\widehat{\mu} - \mu)(s)\mu(t),$$

(S.1) is obtained by plugging the expansion for $\widehat{\mu}$ and $\widehat{C}$ into (S.2).

S.3.2 Proof of Theorem 1

When $p < p_0$, $\mathrm{BIC}(p) - \mathrm{BIC}(p_0) = \{\log(\widehat{\sigma}^2_{[p],\mathrm{marg}}) - \log(\widehat{\sigma}^2_{[p_0],\mathrm{marg}})\} - \{\mathcal{P}_n(p_0) - \mathcal{P}_n(p)\}$. By Lemmas S.1.1 and S.1.2, $\widehat{\sigma}^2_{[p_0],\mathrm{marg}} = \sigma_u^2 + O_p\{\delta_n^* + h_\sigma^2 + \delta_{n1}(h_\sigma)\}$. By (5),

$$\log(\widehat{\sigma}^2_{[p],\mathrm{marg}}) - \log(\widehat{\sigma}^2_{[p_0],\mathrm{marg}}) = \log\left\{1 + (b-a)^{-1}\left(\sum_{k=p+1}^{p_0} \widehat{\omega}_k\right)/\widehat{\sigma}^2_{[p_0],\mathrm{marg}}\right\},$$

which converges to a positive number. Since $\limsup\{\mathcal{P}_n(p_0) - \mathcal{P}_n(p)\} \le 0$ with probability 1, $\mathrm{BIC}(p) - \mathrm{BIC}(p_0)$ is positive with probability approaching 1.

Next, for any fixed $p > p_0$, $\widehat{\sigma}^2_{[p],\mathrm{marg}} = \widehat{\sigma}^2_{[p_0],\mathrm{marg}} - (b-a)^{-1}\sum_{k=p_0+1}^{p} \widehat{\omega}_k$. By Lemma S.1.3, $\sum_{k=p_0+1}^{p} \widehat{\omega}_k = O_p(\delta_n^*)$. By Taylor expansion, $\log(1+x) = x - x^2 + \cdots$, so that

$$\log(\widehat{\sigma}^2_{[p],\mathrm{marg}}) - \log(\widehat{\sigma}^2_{[p_0],\mathrm{marg}}) = \log\{1 - (\textstyle\sum_{k=p_0+1}^{p} \widehat{\omega}_k)/\widehat{\sigma}^2_{[p_0],\mathrm{marg}}\} = -(\textstyle\sum_{k=p_0+1}^{p} \widehat{\omega}_k)/\sigma_u^2 + o_p(\delta_n^*).$$

By the condition that $\delta_n^*/\{\mathcal{P}_n(p) - \mathcal{P}_n(p_0)\} \to 0$, $\mathrm{BIC}(p) - \mathrm{BIC}(p_0) = \mathcal{P}_n(p) - \mathcal{P}_n(p_0) - O_p(\delta_n^*)$ is positive with probability approaching 1.

By combining the arguments above, we conclude $\widehat{p}$, the minimizer of $\mathrm{BIC}(p)$, converges to p_0 with probability tending to 1.

S.3.3 Proof of Proposition 1

We first introduce some notation. Define $\boldsymbol{\psi}_{ik} = \{\psi_k(t_{i1}), \ldots, \psi_k(t_{i,m_i})\}$ for $k = 1, \ldots, p$. Put

$$\boldsymbol{\xi}_{i,[p]} = (\xi_{i1}, \ldots, \xi_{ip})^{\mathrm{T}}, \quad \Psi_{i,[p]} = (\boldsymbol{\psi}_{i1}, \ldots, \boldsymbol{\psi}_{ip}), \quad \Lambda_{[p]} = \mathrm{diag}(\omega_1, \ldots, \omega_p), \quad \Omega_{i,[p]} = \Psi_{i,[p]}\Lambda_{[p]}\Psi_{i,[p]}^{\mathrm{T}},$$

then $\Sigma_{i,[p]} = \sigma_u^2 I + \Omega_{i,[p]}$ is the covariance matrix within the i^{th} curve under the assumption that there are p principal components. For ease of exposition, we shorten $\boldsymbol{\xi}_{i,[p_0]}$, $\Sigma_{i,[p_0]}$, $\Psi_{i,[p_0]}$, $\Lambda_{[p_0]}$ and $\Omega_{i,[p_0]}$ as $\boldsymbol{\xi}_i$, Σ_i, Ψ_i, Λ and Ω_i respectively. For the following derivation, we use the following algebraic facts: $I - \Omega_i\Sigma_i^{-1} = \sigma_u^2\Sigma_i^{-1} = (I + \Psi_i\Lambda\Psi_i^{\mathrm{T}}/\sigma_u^2)^{-1} = I - \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}$. Under assumption (12), we have $m_i^{-1}\boldsymbol{\psi}_{ik}^{\mathrm{T}}\boldsymbol{\psi}_{ik'} \to \int \psi_k(t)\psi_{k'}(t)f_1(t)dt$, for $k, k' = 1, \ldots, p_0$, and hence $\Psi_i^{\mathrm{T}}\Psi_i = O(m_i)$.

Define

$$\widetilde{\sigma}^2_{[p_0]} = N^{-1}\sum_{i=1}^{n}\|\sigma_u^2\Sigma_i^{-1}(\boldsymbol{W}_i - \boldsymbol{\mu}_i)\|^2 \quad \text{and} \quad \mathcal{R}_n = (\widehat{\sigma}^2_{[p_0]} - \widetilde{\sigma}^2_{[p_0]})/\sigma_u^2. \tag{S.3}$$

Then $\widehat{\sigma}^2_{[p_0]}/\sigma_u^2 = \widetilde{\sigma}^2_{[p_0]}/\sigma_u^2 + \mathcal{R}_n$. To show Proposition 1, we will first provide an asymptotic expansion for $\widetilde{\sigma}^2_{[p_0]}$ and then show that $\mathcal{R}_n = O_p\{\delta^2_{n1}(h_\mu) + \delta^2_{n1}(h_C)\} + o_p(nN^{-1})$.

Under the Gaussian assumption,

$$\|\sigma_u^2\Sigma_i^{-1}(\boldsymbol{W}_i - \boldsymbol{\mu}_i)\|^2 = \sigma_u^2\mathcal{Z}_i^{\mathrm{T}}(\Psi_i\Lambda\Psi_i^{\mathrm{T}}/\sigma_u^2 + I)^{-1}\mathcal{Z}_i,$$

where $\mathcal{Z}_i$ is an m_i-vector of independent Normal$(0,1)$ random variables. Define $\lambda_j(\cdot)$ to be the functional that computes the j^{th} eigenvalue of a matrix, and let the eigenvalues be in descending order. Denote

$$\theta_{ij} = \lambda_j(\Sigma_i/\sigma_u) = \lambda_j(\Psi_i\Lambda\Psi_i^{\mathrm{T}}/\sigma_u^2 + I) = \lambda_j(\Omega_i)/\sigma_u^2 + 1. \tag{S.4}$$

Since Ψ_i is of rank p_0, we see that $\theta_{ij} = 1$ for $j = p_0+1,\ldots,m_i$, and

$$\theta_{ij} = \lambda_j(\Omega_i)/\sigma_u^2 + 1 = \lambda_j(\Lambda\Psi_i^{\mathrm{T}}\Psi_i)/\sigma_u^2 + 1, \quad j = 1,\ldots,p_0.$$

Since $\Psi_i^{\mathrm{T}}\Psi_i = O(m_i)$, we conclude that $\theta_{ij} = O(m_i)$ for $j = 1,\ldots,p_0$. It is easy to see that

$$\sigma_u^2\mathcal{Z}_i^{\mathrm{T}}(\Psi_i\Lambda\Psi_i^{\mathrm{T}}/\sigma_u^2 + I)^{-1}\mathcal{Z}_i = \sigma_u^2\sum_{j=1}^{m_i}\theta_{ij}^{-1}\mathcal{X}_{ij},$$

where the $\mathcal{X}_{ij}$ are independent χ^2_1 random variable. Since $\min_i(m_i) \to \infty$,

$$\widetilde{\sigma}^2_{[p_0]} = \frac{\sigma_u^2}{N}\sum_{i=1}^{n}\sum_{j=p_0+1}^{m_i}\mathcal{X}_{ij} + \frac{\sigma_u^2}{N}\sum_{i=1}^{n}\sum_{j=1}^{p_0}\theta_{ij}^{-1}\mathcal{X}_{ij} = \frac{\sigma_u^2}{N}\sum_{i=1}^{n}\sum_{j=p_0+1}^{m_i}\mathcal{X}_{ij} + o_p(nN^{-1}). \tag{S.5}$$

By the Weak Law of Large Numbers, we have $\widetilde{\sigma}^2_{[p_0]} \to \sigma_u^2$ in probability.

Next, denote $\boldsymbol{\epsilon}_i = \boldsymbol{W}_i - \boldsymbol{\mu}_i$, $\widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i$, and by simple algebra, $\sigma^2(\Psi_i\Lambda\Psi_i + \sigma^2)^{-1} = I - \Psi_i(\sigma^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}$. Thus,

$$\begin{aligned}
\sigma_u^2\mathcal{R}_n &= N^{-1}\sum_{i=1}\widehat{\boldsymbol{\epsilon}}_i^{\mathrm{T}}\{I - \widehat{\Psi}_i(\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Lambda}^{-1} + \widehat{\Psi}_i^{\mathrm{T}}\widehat{\Psi}_i)^{-1}\widehat{\Psi}_i^{\mathrm{T}}\}^2\widehat{\boldsymbol{\epsilon}}_i - \boldsymbol{\epsilon}_i^{\mathrm{T}}\{I - \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\}^2\boldsymbol{\epsilon}_i \\
&= (\mathcal{R}_{1,n} + \mathcal{R}_{2,n} + \mathcal{R}_{3,n}) \times \{1 + o_p(1)\},
\end{aligned}$$

where

$$\mathcal{R}_{1,n} = -2\sigma_u^4N^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_i^{-2}\boldsymbol{\epsilon}_i,$$

$$\mathcal{R}_{2,n} = -2\frac{\sigma_u^2}{N}\sum_{i=1}^{n}\boldsymbol{\epsilon}_i^{\mathrm{T}}\{\widehat{\Psi}_i(\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Lambda}^{-1} + \widehat{\Psi}_i^{\mathrm{T}}\widehat{\Psi}_i)^{-1}\widehat{\Psi}_i^{\mathrm{T}} - \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\}\Sigma_i^{-1}\boldsymbol{\epsilon}_i,$$

$$\mathcal{R}_{3,n} = N^{-1}\sum_{i=1}^{n}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\{I - \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\}^2(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i).$$

Denote $\boldsymbol{g}_i = (g_{i1}, \ldots, g_{i,m_i})^{\mathrm{T}} = \sigma_u^4 \Sigma_i^{-2} \boldsymbol{\epsilon}_i$, then $\mathrm{E}(\boldsymbol{g}_i) = \mathbf{0}$, $\mathrm{cov}(\boldsymbol{g}_i, \boldsymbol{\epsilon}_i) = \sigma_u^4 \Sigma_i^{-1} = \sigma_u^2 (I - \Psi_i(\sigma^2 \Lambda^{-1} + \Psi_i^{\mathrm{T}} \Psi_i)^{-1} \Psi_i^{\mathrm{T}})$. Since $\Psi_i^{\mathrm{T}} \Psi_i = O(m_i)$, we have $\mathrm{E}(\epsilon_{ij} g_{ij'}) = O(m_i^{-1})$ if $j \neq j'$, and $= O(1)$ if $j = j'$. Similarly, since $\mathrm{cov}(\boldsymbol{g}_i, \boldsymbol{g}_i) = \sigma_u^8 \Sigma_i^{-3}$, we have $\mathrm{cov}(g_{ij}, g_{ij'}) = O(m_i^{-1})$ if $j \neq j'$, and $= O(1)$ if $j = j'$. By Lemma S.3.1,

$$\mathcal{R}_{1,n} = -\frac{2}{N} \sum_{i_1=1}^{n} \sum_{j_1=1}^{m_{i_1}} g_{i_1 j_1} \Big\{ \frac{1}{n} \sum_{i_2=1}^{n} \frac{1}{m_{i_2}} \sum_{j_2=1}^{m_{i_2}} \epsilon_{i_2 j_2} K_{h_\mu}(t_{i_1 j_1} - t_{i_2 j_2}) \Big\} \times \{1 + o(1)\}.$$

By straightforward calculations,

$$\begin{aligned}
\mathrm{E}(\mathcal{R}_{1,n}) &= \Big\{ -\frac{2}{nN} \sum_{i=1}^{n} \frac{1}{m_i} \sum_{j_1=1}^{m_i} \sum_{j_2=1}^{m_i} \mathrm{E}(g_{ij_1} \epsilon_{ij_2}) K_{h_\mu}(t_{ij_1} - t_{ij_2}) \Big\} \times \{1 + o(1)\} \\
&= \Big[-\frac{2}{nN} \sum_{i=1}^{n} \frac{1}{m_i} \Big\{ \sum_{j=1}^{m_i} \mathrm{E}(g_{ij} \epsilon_{ij}) h_\mu^{-1} K(0) + \sum_{j_1 \neq j_2} \mathrm{E}(g_{ij_1} \epsilon_{ij_2}) K_{h_\mu}(t_{ij_1} - t_{ij_2}) \Big\} \Big] \times \{1 + o(1)\}.
\end{aligned}$$

Since $\mathrm{E}(g_{ij_1} \epsilon_{ij_2}) = O(m_i^{-1})$ for $j_1 \neq j_2$, we can show $m_i^{-1} \sum_{j_1 \neq j_2} \mathrm{E}(g_{ij_1} \epsilon_{ij_2}) K_{h_\mu}(t_{ij_1} - t_{ij_2}) = O(1)$. Therefore, $\mathrm{E}(\mathcal{R}_{1,n}) = O(N^{-1} h_\mu^{-1})$.

Since $\mathrm{E}(g_{ij_1} g_{ij_2}) = O(m_i^{-1})$ if $j_1 \neq j_2$, and $= O(1)$ if $j_1 = j_2$, we have

$$\begin{aligned}
\mathrm{var}(\mathcal{R}_{1,n}) &= \frac{4}{n^2 N^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \frac{1}{m_{i_2}^2} \mathrm{var}\Big\{ \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} g_{i_1 j_1} \epsilon_{i_2 j_2} K_{h_\mu}(t_{i_2 j_2} - t_{i_1 j_1}) \Big\} \\
&\quad + \frac{4}{n^2 N^2} \sum_{i_1=1}^{n} \sum_{i_2 \neq i_1} \mathrm{cov}\Big\{ \frac{1}{m_{i_2}} \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} g_{i_1 j_1} \epsilon_{i_2 j_2} K_{h_\mu}(t_{i_2 j_2} - t_{i_1 j_1}), \\
&\qquad\qquad \frac{1}{m_{i_1}} \sum_{j_3=1}^{m_{i_2}} \sum_{j_4=1}^{m_{i_1}} g_{i_2 j_3} \epsilon_{i_1 j_4} K_{h_\mu}(t_{i_2 j_3} - t_{i_1 j_4}) \Big\} \\
&\leq \frac{8}{n^2 N^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \frac{1}{m_{i_2}^2} \mathrm{var}\Big\{ \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} g_{i_1 j_1} \epsilon_{i_2 j_2} K_{h_\mu}(t_{i_2 j_2} - t_{i_1 j_1}) \Big\} \\
&\leq \frac{8}{n^2 N^2} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \frac{1}{m_{i_2}^2} \mathrm{E}\Big\{ \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} g_{i_1 j_1} \epsilon_{i_2 j_2} K_{h_\mu}(t_{i_2 j_2} - t_{i_1 j_1}) \Big\}^2 \\
&= \frac{8}{n^2 N^2} \Big[\sum_{i=1}^{n} \frac{1}{m_i^2} \mathrm{E}\Big\{ \sum_{j_1=1}^{m_i} \sum_{j_2=1}^{m_i} g_{ij_1} \epsilon_{ij_2} K_{h_\mu}(t_{ij_2} - t_{ij_1}) \Big\}^2 \\
&\qquad + \sum_{i_1 \neq i_2} \frac{1}{m_{i_2}^2} \mathrm{E}\Big\{ \sum_{j_1=1}^{m_{i_1}} \sum_{j_2=1}^{m_{i_2}} g_{i_1 j_1} \epsilon_{i_2 j_2} K_{h_\mu}(t_{i_2 j_2} - t_{i_1 j_1}) \Big\}^2 \Big].
\end{aligned}$$

By similar arguments as above, one can show that $\mathrm{E}(g_{ij_1} \epsilon_{ij_2} g_{ij_3} \epsilon_{ij_4}) = O(m_i^{-1})$ if $j_1 \neq j_3$, and $= O(1)$ if $j_1 = j_3$. Then by more detailed calculations we have $\mathrm{var}(\mathcal{R}_{1,n}) = O\{(nN)^{-1} +$

$(nN^2h_\mu^2)^{-1}\}$, and therefore $\mathrm{E}(\mathcal{R}_{1,n}^2) = \mathrm{var}(\mathcal{R}_{1,n}) + \mathrm{E}^2(\mathcal{R}_{1,n}) = O(n^{-1}N^{-1} + h_\mu^{-2}N^{-2})$ and we conclude that $\mathcal{R}_{1,n} = o_p\{\delta_{n1}^2(h_\mu) + nN^{-1}\}$.

By simple algebra, $\mathcal{R}_{2,n} = (\mathcal{R}_{2,a} + \mathcal{R}_{2,b}) \times \{1 + o_p(1)\}$, where

$$\mathcal{R}_{2,a} = -2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i$$
$$-2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i,$$
$$\mathcal{R}_{2,b} = -2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i\{(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}^{-1} + \widehat{\Psi}_i^{\mathrm{T}}\widehat{\Psi}_i)^{-1} - (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i.$$

Since $\boldsymbol{\epsilon}_i = \Psi_i\boldsymbol{\xi}_i + \boldsymbol{U}_i$, and thus $(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{\epsilon}_i = \boldsymbol{\xi}_i + (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\sigma_u^2\Lambda^{-1}\boldsymbol{\xi}_i + (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{U}_i = \boldsymbol{\xi}_i + O_p(m_i^{-1/2})$. Further,

$$\begin{aligned}
\mathcal{R}_{2,b} &= \Big\{2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}^{-1} + \widehat{\Psi}_i^{\mathrm{T}}\widehat{\Psi}_i - \sigma_u^2\Lambda^{-1} - \Psi_i^{\mathrm{T}}\Psi_i) \\
&\qquad (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i\Big\} \times \{1 + o_p(1)\} \\
&= \Big\{2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i \\
&\quad +2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i \\
&\quad +2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i \\
&\quad +(\widetilde{\sigma}_{u,\mathrm{I}}^2 - \sigma_u^2) \times 2\frac{\sigma_u^2}{N}\sum_{i=1}^n \boldsymbol{\xi}_i^{\mathrm{T}}\Lambda^{-1}(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i \\
&\quad +2\frac{\sigma_u^4}{N}\sum_{i=1}^n \boldsymbol{\xi}_i^{\mathrm{T}}(\widehat{\Lambda}^{-1} - \Lambda^{-1})(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i\Big\} \times \{1 + o_p(1)\} \\
&= -\mathcal{R}_{2,a} - \Big[2\frac{\sigma_u^4}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Sigma_i^{-1}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i \\
&\qquad +2\frac{\sigma_u^4}{N}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}\Sigma_i^{-2}\boldsymbol{\epsilon}_i\Big] \\
&\qquad +2\frac{\sigma_u^4}{N}\sum_{i=1}^n \boldsymbol{\xi}_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i + o_p(nN^{-1}).
\end{aligned}$$

Denote $A_n = 2\sigma_u^4N^{-1}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Sigma_i^{-1}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i$ and $B_n = 2\sigma_u^4N^{-1}\sum_{i=1}^n \boldsymbol{\epsilon}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}\Sigma_i^{-2}\boldsymbol{\epsilon}_i$. Letting $\boldsymbol{g}_i = \sigma_u^2\Sigma_i^{-1}\boldsymbol{\epsilon}_i$ and $\boldsymbol{v}_i = \sigma_u^2(\sigma_u^2\Lambda^{-1} +$

$\Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\boldsymbol{\epsilon}_i$, we can easily see that $\mathrm{cov}(\boldsymbol{g}_i, \boldsymbol{g}_i) = \sigma_u^4\Sigma_i^{-1}$, $\mathrm{cov}(\boldsymbol{v}_i, \boldsymbol{v}_i) = \sigma_u^4(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Sigma_i^{-1}\Psi_i(\sigma_u^2\Lambda^{-1}+\Psi_i^{\mathrm{T}}\Psi_i)^{-1} = \sigma_u^2(\sigma_u^2\Lambda^{-1}+\Psi_i^{\mathrm{T}}\Psi_i)^{-1}\sigma_u^2\Lambda^{-1}(\sigma_u^2\Lambda^{-1}+\Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1}+\Psi_i^{\mathrm{T}}\Psi_i)^{-1} = O(m_i^{-2})$ and $\mathrm{cov}(\boldsymbol{g}_i, \boldsymbol{v}_i) = \sigma_u^2\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\sigma_u^2\Lambda^{-1}(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1} = O(m_i^{-2})$. These imply that $\mathrm{cov}(g_{ij}, g_{ij'}) = O(1)$ if $j = j'$, and $= O(m_i^{-1})$ if $j \neq j'$; $\mathrm{cov}(g_{ij}, v_{ik}) = O(m_i^{-2})$ for all j and k. By plugging in the asymptotic expansion of $\widehat{\Psi}_i$ given in Lemma S.3.1, and by similar calculations as for $\mathcal{R}_{1,n}$, we can show that $\mathrm{E}(A_n) = o(nN^{-1})$, $\mathrm{E}(A_n^2) = o(n^2N^{-2})$. Similar calculation shows that $B_n = o_p(nN^{-1})$. By combining $\mathcal{R}_{2,a}$ and $\mathcal{R}_{2,b}$ and by Lemma S.1.2, we conclude that

$$\mathcal{R}_{2,n} = 2N^{-1}\sum_{i=1}^{n}\boldsymbol{\xi}_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)\boldsymbol{\xi}_i + o_p(nN^{-1}) = O_p\{\delta_{n1}^2(h_\mu) + \delta_{n1}^2(h_C)\} + o_p(nN^{-1}).$$

It can be easily seen that $\mathcal{R}_{3,n} = N^{-1}\sum_{i=1}^n \|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 + \mathcal{R}_{3,a} + \mathcal{R}_{3,b}$, where $\mathcal{R}_{3,a} = -2N^{-1}\sum_{i=1}^n(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1}+\Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)$, $\mathcal{R}_{3,b} = N^{-1}\sum_{i=1}^n(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1}+\Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)$. By simple algebra, $\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}} \leq \Psi_i(\Psi_i^{\mathrm{T}}\Psi_i)^{-}\Psi_i^{\mathrm{T}}$, which is an idempotent matrix. By Lemma S.1.1,

$$\mathrm{E}|\mathcal{R}_{3,a}| \leq \mathrm{E}\Big\{\frac{2}{N}\sum_{i=1}^{n}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)^{\mathrm{T}}\Psi_i(\Psi_i^{\mathrm{T}}\Psi_i)^{-}\Psi_i^{\mathrm{T}}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\Big\} = O\{nN^{-1}\delta_{n1}^2(h_\mu)\} = o(nN^{-1}).$$

Similarly, we have $\mathcal{R}_{3,b} = o_p(nN^{-1})$, and therefore

$$\mathcal{R}_{3,n} = N^{-1}\sum_{i=1}^{n}\|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 + o_p(nN^{-1}) = O_p\{\delta_{n1}^2(h_\mu)\} + o_p(nN^{-1}).$$

Finally, by combining $\mathcal{R}_{1,n}$, $\mathcal{R}_{2,n}$ and $\mathcal{R}_{3,n}$, we conclude that

$$\sigma_u^2\mathcal{R}_n = O_p\{\delta_{n1}^2(h_\mu) + \delta_{n1}^2(h_C)\} + o_p(nN^{-1}). \tag{S.6}$$

Since $\widehat{\sigma}_{[p_0]}^2 = \widetilde{\sigma}_{[p_0]}^2 + \sigma_u^2\mathcal{R}_n$, the asymptotic expansion and consistency for $\widehat{\sigma}_{[p_0]}^2$ is obtained immediately from (S.5) and (S.6).

S.3.4 Proof of Proposition 2

Following the conventions in Proposition 1, we shorten $\boldsymbol{\xi}_{i,[p_0]}$, $\Sigma_{i,[p_0]}$, $\Psi_{i,[p_0]}$, $\Lambda_{[p_0]}$ and $\Omega_{i,[p_0]}$ as $\boldsymbol{\xi}_i$, Σ_i, Ψ_i, Λ and Ω_i respectively. We first prove the following lemma.

Lemma **S.3.2** *Suppose all assumptions for Proposition 2 hold and denote* $\mathcal{D}_i = \boldsymbol{U}_i^{\mathrm{T}}(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Sigma}_i^{-1} - \sigma_u^2\Sigma_i^{-1})\boldsymbol{\epsilon}_i$. *Then as* $m_i \to \infty$, $\mathrm{E}(\mathcal{D}_i) = o(1)$ *for* $i = 1, \ldots, n$.

Proof: We will study the asymptotic structure of $\mathcal{D}_i$ using Taylor series expansion. We will verify that the first order Taylor expansion of $\mathcal{D}_i$ has a mean of order $o(1)$. Similar conclusions can be verified for the higher order terms. Since $\sigma_u^2\Sigma_i^{-1} = I - \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}$, we have

$$\begin{aligned}\mathcal{D}_i &= \boldsymbol{U}_i^{\mathrm{T}}\{\widehat{\Psi}_i(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}^{-1} + \widehat{\Psi}_i^{\mathrm{T}}\widehat{\Psi}_i)^{-1}\widehat{\Psi}_i^{\mathrm{T}} - \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\}\boldsymbol{\epsilon}_i \\ &= (\mathcal{D}_{i1} + \mathcal{D}_{i2} + \mathcal{D}_{i3}) \times \{1 + o_p(1)\},\end{aligned}$$

where $\mathcal{D}_{i1}$-$\mathcal{D}_{i3}$ are the terms in the first order Taylor expansion of $\mathcal{D}_i$ given by

$$\begin{aligned}&\mathcal{D}_{i1} = \boldsymbol{U}_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{\epsilon}_i,\\ &\mathcal{D}_{i2} = \boldsymbol{U}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}\boldsymbol{\epsilon}_i,\\ &\mathcal{D}_{i3} = \boldsymbol{U}_i^{\mathrm{T}}\Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\{\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}^{-1} - \sigma_u^2\Lambda^{-1} + (\widehat{\Psi}_i^{\mathrm{T}} - \Psi_i)\Psi_i + \Psi_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)\}\\ &\qquad\qquad \times(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{\epsilon}_i.\end{aligned}$$

We first show that $\mathrm{E}(\mathcal{D}_{i1}) = o(1)$. Let $\boldsymbol{g}_i = (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{\epsilon}_i := (g_{i1}, \ldots, g_{ip_0})^{\mathrm{T}}$. Since $\Psi_i^{\mathrm{T}}\Psi_i = O(m_i)$, we have $\boldsymbol{g}_i = (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}(\Psi_i\boldsymbol{\xi} + \boldsymbol{U}_i) = \boldsymbol{\xi}_i + O_p(m_i^{-1/2})$. It is also easy to verify that $\mathrm{E}(\boldsymbol{g}_i) = \mathrm{E}(\boldsymbol{U}_i) = \mathbf{0}$, and $\mathrm{E}(\boldsymbol{U}_i\boldsymbol{g}_i^{\mathrm{T}}) = \Psi_i(\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1} = O(m_i^{-1})$, i.e. $\mathrm{E}(U_{ij}g_{ij'}) = O(m_i^{-1})$ for any j, j'. By the asymptotic expansion given in Lemma S.3.1, we have $\mathcal{D}_{i1} = (\mathcal{D}_{i11} + \mathcal{D}_{i12}) \times \{1 + o_p(1)\}$, where

$$\begin{aligned}\mathcal{D}_{i11} = &\frac{1}{n}\sum_{i'\neq i}\sum_{\ell=1}^{m_i}\sum_{k=1}^{p_0}\Big\{\frac{1}{M_{i'}}\sum_{j=1}^{m_{i'}}\sum_{j'\neq j}\epsilon^*_{i',jj'}u_{i\ell}g_{ik}\mathcal{G}_{2,k}(t_{i'j}, t_{i'j'}, t_{i\ell}) + \frac{1}{m_{i'}}\sum_{j=1}^{m_{i'}}\epsilon_{i'j}u_{i\ell}g_{ik}\mathcal{G}_{1,k}(t_{i'j}, t_{i\ell})\\ &+\frac{1}{M_{i'}}\sum_{j=1}^{m_{i'}}\sum_{j'\neq j}\epsilon^*_{i',jj'}u_{i\ell}g_{ik}\psi_k(t_{i'j})/\omega_k/f_2(t_{i'j}, t_{i\ell})K_{h_C}(t_{i'j'} - t_{i\ell})\\ &-\frac{\langle\mu, \psi_k\rangle}{\omega_k f_1(t_{i\ell})} \times \frac{1}{m_{i'}}\sum_{j=1}^{m_{i'}}\epsilon_{i'j}u_{i\ell}g_{ik}K_{h_\mu}(t_{i'j} - t_{i\ell})\Big\},\end{aligned}$$

and $\mathcal{D}_{i12}$ is similar to $\mathcal{D}_{i11}$ except one should replace the first summation by $i' = i$. It is easy to see that $\mathrm{E}(\mathcal{D}_{i11}) = 0$, and

$$\mathrm{E}(\mathcal{D}_{i12}) = \frac{1}{n}\sum_{\ell=1}^{m_i}\sum_{k=1}^{p_0}\frac{1}{M_i}\sum_{j'\neq j}\mathrm{E}(\epsilon^*_{i,jj'}u_{i\ell}g_{ik})\Big\{\mathcal{G}_{2,k}(t_{ij}, t_{ij'}, t_{i\ell}) + \frac{\psi_k(t_{ij})}{\omega_k f_2(t_{ij}, t_{i\ell})}K_{h_C}(t_{ij'} - t_{i\ell})\Big\}.$$

By definition, $\epsilon^*_{i,jj'} = W_{ij}W_{ij'} - C(t_{ij}, t_{ij'}) = \mu(t_{ij})\epsilon_{ij'} + \mu(t_{ij'})\epsilon_{ij} + \epsilon_{ij}\epsilon_{ij'} - R(t_{ij}, t_{ij'})$, then $\mathrm{E}(\epsilon^*_{i,jj'}u_{i\ell}g_{ik}) = \mathrm{E}(\epsilon_{ij}\epsilon_{ij'}u_{i\ell}g_{ik}) + O(m_i^{-1}) = O(1)$ if $\ell = j$ or j', $= O(m_i^{-1})$ otherwise. By detailed calculation, we have $\mathrm{E}(\mathcal{D}_{i12}) = O\{n^{-1} + (nh_C)^{-1}\}$. Hence we conclude $\mathrm{E}(\mathcal{D}_{i1}) = o(1)$. By similar calculation, we can show $\mathrm{E}(\mathcal{D}_{i2}) = o(1)$.

Finally, $\mathcal{D}_{i3} = \mathcal{D}_{i31} + \mathcal{D}_{i32} + \mathcal{D}_{i33}$, where

$$\mathcal{D}_{i31} = \boldsymbol{r}_i^{\mathrm{T}}(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}^{-1} - \sigma_u^2\Lambda)\boldsymbol{g}_i, \quad \mathcal{D}_{i32} = \boldsymbol{r}_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)^{\mathrm{T}}\Psi_i\boldsymbol{g}_i, \quad \mathcal{D}_{i33} = \boldsymbol{r}_i^{\mathrm{T}}\Psi_i^{\mathrm{T}}(\widehat{\Psi}_i - \Psi_i)\boldsymbol{g}_i,$$

$\boldsymbol{r}_i = (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{U}_i$, and $\boldsymbol{g}_i = (\sigma_u^2\Lambda^{-1} + \Psi_i^{\mathrm{T}}\Psi_i)^{-1}\Psi_i^{\mathrm{T}}\boldsymbol{\epsilon}_i$. By similar arguments as for $\mathcal{D}_{i1}$ we can show that $\mathrm{E}(\mathcal{D}_{i32}) = o(1)$ and $\mathrm{E}(\mathcal{D}_{i33}) = o(1)$. It remains to show that $\mathrm{E}(\mathcal{D}_{i31}) = o(1)$. It can be easily seen that $\boldsymbol{r}_i$ and $\boldsymbol{g}_i$ are p_0-dim vectors with $\boldsymbol{r}_i = O_p(m_i^{-1/2})$ and $\boldsymbol{g}_i = O_p(1)$. By Lemmas S.1.1 and S.1.2, we have $\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}^{-1} - \sigma_u^2\Lambda = o_p(1)$ and therefore $\mathcal{D}_{i31} = o_p(1)$ and $\mathrm{E}(\mathcal{D}_{i31}) = o(1)$. That completes the proof.

Proof of Proposition 2: We have

$$\begin{aligned}
\mathcal{A}_n(p_0) &= N + \widehat{\sigma}_{[p_0]}^{-2}\Big\{N(\sigma_u^2 - \widehat{\sigma}_{[p_0]}^2) + \sum_{i=1}^n \|\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i + \Psi_i\boldsymbol{\xi}_i - \widehat{\Omega}_i\widehat{\Sigma}_i^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\|^2\Big\} \\
&= N + N\frac{\sigma_u^2}{\widehat{\sigma}_{[p_0]}^2} + \frac{1}{\widehat{\sigma}_{[p_0]}^2}\Big\{\sum_{i=1}^n \|\boldsymbol{W}_i - \widehat{\mu}_i - \boldsymbol{U}_i - \widehat{\Omega}_i\widehat{\Sigma}_i^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\|^2 - N\widehat{\sigma}_{[p_0]}^2\Big\} \\
&= N + N\frac{\sigma_u^2}{\widehat{\sigma}_{[p_0]}^2} + \frac{1}{\widehat{\sigma}_{[p_0]}^2}\Big\{\sum_{i=1}^n \Big(\|\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Sigma}_i^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i) - \boldsymbol{U}_i\|^2 - \|\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Sigma}_i^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\|^2\Big)\Big\} \\
&= N + \frac{1}{\widehat{\sigma}_{[p_0]}^2}\Big\{N\sigma_u^2 + \sum_{i=1}^n \Big(\|\boldsymbol{U}_i\|^2 - 2\boldsymbol{U}_i^{\mathrm{T}}\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Sigma}_i^{-1}(\boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i)\Big)\Big\} \\
&:= N + \widehat{\sigma}_{[p_0]}^{-2}(\mathcal{A}_{1n} + \mathcal{A}_{2n} + \mathcal{A}_{3n} + \mathcal{A}_{4n}),
\end{aligned}$$

where

$$\mathcal{A}_{1n} = N\sigma_u^2 + \sum_{i=1}^n \|\boldsymbol{U}_i\|^2, \qquad \mathcal{A}_{2n} = -2\sum_{i=1}^n \boldsymbol{U}_i^{\mathrm{T}}\sigma_u^2\Sigma_i^{-1}\boldsymbol{\epsilon}_i,$$

$$\mathcal{A}_{3n} = 2\sum_{i=1}^n \boldsymbol{U}_i^{\mathrm{T}}\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Sigma}_i^{-1}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i), \quad \mathcal{A}_{4n} = -2\sum_{i=1}^n \boldsymbol{U}_i^{\mathrm{T}}(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Sigma}_i^{-1} - \sigma_u^2\Sigma_i^{-1})\boldsymbol{\epsilon}_i.$$

It is easy to show that $\mathrm{E}(\mathcal{A}_{1n}) = 2N\sigma_u^2$. Letting θ_{ij} be defined in (S.4), we have

$$\mathrm{E}(\mathcal{A}_{2n}) = -2\sigma_u^4\sum_{i=1}^n \mathrm{tr}(\Sigma_i^{-1}) = -2\sigma_u^2\sum_{i=1}^n\sum_{j=1}^{m_i}\theta_{ij}^{-1} = -2(N - np_0)\sigma_u^2 + o(n).$$

Following similar arguments as for $\mathcal{R}_{1n}$ in the proof of Proposition 1, we have

$$\mathcal{A}_{3n} = \Big\{2\sum_{i=1}^n \boldsymbol{U}_i^{\mathrm{T}}\sigma_u^2\Sigma_i^{-1}(\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i)\Big\} \times \{1 + o_p(1)\} = o_p(n).$$

By Lemma S.3.2, $\mathrm{E}(\mathcal{A}_{4n}) = o(n)$. Combining the results above, we have

$$\mathrm{E}(\mathcal{A}_{1n} + \mathcal{A}_{2n} + \mathcal{A}_{3n} + \mathcal{A}_{4n}) = 2np_0\sigma_u^2 + o(n).$$

By Proposition 1, $\widehat{\sigma}_{[p_0]}$ is consistent for σ_u^2, with $\mathrm{E}\widehat{\sigma}^2_{[p_0]} = \sigma_u^2 + o(1)$. Using the Delta method, one can show that $n^{-1}\widehat{\sigma}^{-2}_{[p_0]}(\mathcal{A}_{1n} + \mathcal{A}_{2n} + \mathcal{A}_{3n} + \mathcal{A}_{4n})$ is asymptotically normal with mean

$$\mathrm{E}\{n^{-1}\widehat{\sigma}^{-2}_{[p_0]}(\mathcal{A}_{1n} + \mathcal{A}_{2n} + \mathcal{A}_{3n} + \mathcal{A}_{4n})\} = \frac{1}{n\sigma_u^2}\mathrm{E}(\mathcal{A}_{1n} + \mathcal{A}_{2n} + \mathcal{A}_{3n} + \mathcal{A}_{4n}) \times \{1 + o(1)\}.$$

Therefore, we have $\mathrm{E}(\mathcal{A}_n) = N + 2np_0 + o(n)$, which completes the proof.

S.3.5 Proof of Theorem 2

Let $\boldsymbol{\xi}_{i,[p]}$, $\Psi_{i,[p]}$, $\lambda_{[p]}$, $\Omega_{i,[p]}$ and $\Sigma_{i,[p]}$ be defined as at the beginning of Section S.3.3, and let $\widehat{\boldsymbol{\xi}}_{i,[p]}$, $\widehat{\Psi}_{i,[p]}$, $\widehat{\lambda}_{[p]}$, $\widehat{\Omega}_{i,[p]}$ and $\widehat{\Sigma}_{i,[p]}$ be the estimators of these quantities using the estimation procedure in Section 2. For any $p_1 \le p_2$, we also define $\Psi_{i,[p_1:p_2]} = (\boldsymbol{\psi}_{i,p_1}, \dots, \boldsymbol{\psi}_{i,p_2})$, $\Lambda_{[p_1:p_2]} = \mathrm{diag}(\omega_{p_1}, \dots, \omega_{p_2})$, and let $\widehat{\Psi}_{i,[p_1:p_2]}$ and $\widehat{\Lambda}_{[p_1:p_2]}$ be their estimators. For convenience, $\Psi_{i,[p_1:p_2]}$ and $\Lambda_{[p_1:p_2]}$ equal to $\mathbf{0}$ matrices for $p1 > p_2$.

<u>Lemma</u> **S.3.3** *Consider the cases that $p \le p_0$. Under the conditions in Theorem 2, $\widehat{\sigma}^2_{[p]} - \sigma_u^2 \to \tau_p$ in probability, whereτ_p is defined in (19) for $p < p_0$ and $\tau_p = 0$ for $p = p_0$.*

Proof: Similar to the proof of Proposition 1, we find that

$$\widehat{\sigma}^2_{[p]} = N^{-1}\sum_{i=1}^n \|\widetilde{\sigma}^2_{u,I}\widehat{\Sigma}^{-1}_{i,[p]}\widehat{\boldsymbol{\epsilon}}_i\|^2, \quad \text{where } \widehat{\boldsymbol{\epsilon}}_i = \boldsymbol{W}_i - \widehat{\boldsymbol{\mu}}_i.$$

Define $\widetilde{\sigma}^2_{[p]} = N^{-1}\sum_{i=1}^n \|\sigma_u^2\Sigma^{-1}_{i,[p]}\boldsymbol{\epsilon}_i\|^2$, $\boldsymbol{\epsilon}_i = \boldsymbol{W}_i - \boldsymbol{\mu}_i$ and $\mathcal{R}_{n,p} = (\widehat{\sigma}^2_{[p]} - \widetilde{\sigma}^2_{[p]})/\sigma_u^2$. By simple algebra,

$$\sigma_u^2\Sigma^{-1}_{i,[p]} = \sigma_u^2(\sigma_u^2 I + \Psi_{i,[p]}\Lambda_{[p]}\Psi^{\mathrm{T}}_{i,[p]})^{-1} = I - \Psi_{i,[p]}(\sigma_u^2\Lambda^{-1}_{[p]} + \Psi^{\mathrm{T}}_{i,[p]}\Psi_{i,[p]})^{-1}\Psi^{\mathrm{T}}_{i,[p]}.$$

Recall that $\boldsymbol{\epsilon}_i = \Psi_{i,[p]}\boldsymbol{\xi}_{i,[p]} + \Psi_{i,[p+1:p_0]}\boldsymbol{\xi}_{i,[p+1:p_0]} + \boldsymbol{U}_i$, we have

$$\widetilde{\sigma}^2_{[p]} = \frac{1}{N}\sum_{i=1}^n \|\boldsymbol{a}_i + \boldsymbol{b}_i + \boldsymbol{c}_i\|^2,$$

where

$$\begin{aligned}
\boldsymbol{a}_i &= \sigma_u^2\Psi_{i,[p]}(\sigma_u^2\Lambda^{-1}_{[p]} + \Psi^{\mathrm{T}}_{i,[p]}\Psi_{i,[p]})^{-1}\Lambda^{-1}_{[p]}\boldsymbol{\xi}_{i,[p]},\\
\boldsymbol{b}_i &= \{I - \Psi_{i,[p]}(\sigma_u^2\Lambda^{-1}_{[p]} + \Psi^{\mathrm{T}}_{i,[p]}\Psi_{i,[p]})^{-1}\Psi^{\mathrm{T}}_{i,[p]}\}\Psi_{i,[p+1:p_0]}\boldsymbol{\xi}_{i,[p+1:p_0]},\\
\boldsymbol{c}_i &= \{I - \Psi_{i,[p]}(\sigma_u^2\Lambda^{-1}_{[p]} + \Psi^{\mathrm{T}}_{i,[p]}\Psi_{i,[p]})^{-1}\Psi^{\mathrm{T}}_{i,[p]}\}\boldsymbol{U}_i.
\end{aligned}$$

It is easy to see that $m_i^{-1}\Psi^{\mathrm{T}}_{i,[p]}\Psi_{i,[p]} \xrightarrow{p} \mathcal{J}_{1,p}$, $m_i^{-1}\Psi^{\mathrm{T}}_{i,[p+1:p_0]}\Psi_{i,[p+1:p_0]} \xrightarrow{p} \mathcal{J}_{2,p}$ and $m_i^{-1}\Psi^{\mathrm{T}}_{i,[p]}$ $\Psi_{i,[p+1:p_0]} \xrightarrow{p} \mathcal{J}_{12,p}$. Therefore,

$$\|\boldsymbol{a}_i\|^2 = O_p(m_i^{-1}), \quad \|\boldsymbol{b}_i\|^2 = m_i\boldsymbol{\xi}^{\mathrm{T}}_{i,[p+1:p_0]}(\mathcal{J}_{2,p} - \mathcal{J}^{\mathrm{T}}_{12,p}\mathcal{J}^{-1}_{1,p}\mathcal{J}_{12,p})\boldsymbol{\xi}_{i,[p+1:p_0]} + O_p(1).$$

On the other hand, $\Psi_{i,[p]}(\sigma_u^2\Lambda_{[p]}^{-1}+\Psi_{i,[p]}^{\mathrm{T}}\Psi_{i,[p]})^{-1}\Psi_{i,[p]}^{\mathrm{T}} \le \Psi_{i,[p]}(\Psi_{i,[p]}^{\mathrm{T}}\Psi_{i,[p]})^{-1}\Psi_{i,[p]}^{\mathrm{T}}$, which is an idempotent matrix of rank p. Hence, $\|\boldsymbol{c}_i\|^2 = \|\boldsymbol{U}_i\|^2 + O_p(1)$. By Cauchy-Schwarz inequality, $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{b}_i$ and $\boldsymbol{a}_i^{\mathrm{T}}\boldsymbol{c}_i$ are of order $O_p(1)$. By the independence between $\boldsymbol{U}_i$ and $\boldsymbol{\xi}_i$, we have $\mathrm{E}(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{c}_i) = 0$, $\mathrm{E}(\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{c}_i)^2 = \sigma_u^2\mathrm{tr}[\{I-\Psi_{i,[p]}(\sigma_u^2\Lambda_{[p]}^{-1}+\Psi_{i,[p]}^{\mathrm{T}}\Psi_{i,[p]})^{-1}\Psi_{i,[p]}^{\mathrm{T}}\}^4\Psi_{i,[p+1:p_0]}\Lambda_{[p+1:p_0]}\Psi_{i,[p+1:p_0]}^{\mathrm{T}}] = O(m_i)$, and hence $\boldsymbol{b}_i^{\mathrm{T}}\boldsymbol{c}_i = O_p(m_i^{1/2})$. Combining the calculations above we find that

$$
\begin{aligned}
\widetilde{\sigma}_{[p]}^2 &= N^{-1}\sum_{i=1}^n\{\|\boldsymbol{U}_i\|^2 + m_i\boldsymbol{\xi}_{i,[p+1:p_0]}^{\mathrm{T}}(\mathcal{J}_{2,p}-\mathcal{J}_{12,p}^{\mathrm{T}}\mathcal{J}_{1,p}^{-1}\mathcal{J}_{12,p})\boldsymbol{\xi}_{i,[p+1:p_0]}\} + O(m^{-1/2}) \\
&\xrightarrow{p} \sigma_u^2 + \tau_p \quad \text{by Laws of Large Numbers.}
\end{aligned}
$$

It remains to show that $\mathcal{R}_{n,p} \xrightarrow{p} 0$. Following similar calculations as in Proposition 1,

$$
\mathcal{R}_{n,p} = \{\mathcal{R}_{1n,p}+\mathcal{R}_{2n,p}+\mathcal{R}_{3n,p}\}\times\{1+o_p(1)\},
$$

where

$$
\begin{aligned}
\mathcal{R}_{1n,p} &= -2\sigma_u^4 N^{-1}\sum_{i=1}^n(\widehat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i)^{\mathrm{T}}\Sigma_{i,[p]}^{-2}\boldsymbol{\epsilon}_i, \\
\mathcal{R}_{2n,p} &= -2N^{-1}\sum_{i=1}^n\boldsymbol{\epsilon}_i^{\mathrm{T}}\{\widehat{\Psi}_{i,[p]}(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}_{[p]}^{-1}+\widehat{\Psi}_{i,[p]}^{\mathrm{T}}\widehat{\Psi}_{i,[p]})^{-1}\widehat{\Psi}_{i,[p]}^{\mathrm{T}} \\
&\qquad\qquad -\Psi_{i,[p]}(\sigma_u^2\Lambda_{[p]}^{-1}+\Psi_{i,[p]}^{\mathrm{T}}\Psi_{i,[p]})^{-1}\Psi_{i,[p]}^{\mathrm{T}}\}\Sigma_{i,[p]}^{-1}\boldsymbol{\epsilon}_i, \\
\mathcal{R}_{3n,p} &= N^{-1}\sum_{i=1}^n(\widehat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i)^{\mathrm{T}}\{I-\Psi_{i,[p]}(\sigma_u^2\Lambda_{[p]}^{-1}+\Psi_{i,[p]}^{\mathrm{T}}\Psi_{i,[p]})^{-1}\Psi_{i,[p]}^{\mathrm{T}}\}^2(\widehat{\boldsymbol{\mu}}_i-\boldsymbol{\mu}_i).
\end{aligned}
$$

By the convergence results for $\widehat{\mu}(\cdot)$, $\widehat{\omega}_j$, $\widehat{\psi}_j(\cdot)$ and $\widetilde{\sigma}_{u,\mathrm{I}}$ in Lemmas S.1.1 and S.1.2, it is easy to check that all the terms above converge to 0 in probability.

Lemma **S.3.4** *When $p > p_0$, under the conditions in Theorem 2, $\widehat{\sigma}_{[p]}^2 - \widehat{\sigma}_{[p_0]}^2 = O_p(n/N + \varrho_n^2)$.*

Proof: For $p > p_0$, $\widehat{\Sigma}_{i,[p]} - \widehat{\Sigma}_{i,[p_0]} = \widehat{\Psi}_{i,[p_0+1:p]}\widehat{\Lambda}_{[p_0+1:p]}\widehat{\Psi}_{i,[p_0+1:p]}^{\mathrm{T}}$. By simply algebra, $\widehat{\Sigma}_{i,[p]}^{-1} = \widehat{\Sigma}_{i,[p_0]}^{-1} - \widehat{\Pi}_i\widehat{\Sigma}_{i,[p_0]}^{-1}$ where $\widehat{\Pi}_i = \widehat{\Sigma}_{i,[p_0]}^{-1}\widehat{\Psi}_{i,[p_0+1:p]}(\widehat{\Lambda}_{[p_0+1:p]}^{-1} + \widehat{\Psi}_{i,[p_0+1:p]}^{\mathrm{T}}\widehat{\Sigma}_{i,[p_0]}^{-1}\widehat{\Psi}_{i,[p_0+1:p]})^{-1}\widehat{\Psi}_{i,[p_0+1:p]}^{\mathrm{T}}$. Put $\widehat{\boldsymbol{U}}_i = \widetilde{\sigma}_{u,I}^2\widehat{\Sigma}_{i,[p_0]}^{-1}\widehat{\boldsymbol{\epsilon}}_i$, then

$$
\widehat{\sigma}_{[p]}^2 = \frac{1}{N}\sum_{i=1}^n\|(I-\widehat{\Pi}_i)\widehat{\boldsymbol{U}}_i\|^2 = \widehat{\sigma}_{[p_0]}^2 - 2\frac{1}{N}\sum_{i=1}^n\widehat{\boldsymbol{U}}_i^{\mathrm{T}}\widehat{\Pi}_i\widehat{\boldsymbol{U}}_i + \frac{1}{N}\sum_{i=1}^n\widehat{\boldsymbol{U}}_i^{\mathrm{T}}\widehat{\Pi}_i^{\mathrm{T}}\widehat{\Pi}_i\widehat{\boldsymbol{U}}_i.
$$

Using the same technique as above,

$$
\begin{aligned}
\widehat{\boldsymbol{U}}_i &= \{I-\widehat{\Psi}_{i,[p_0]}(\widetilde{\sigma}_{u,\mathrm{I}}^2\widehat{\Lambda}_{[p_0]}^{-1}+\widehat{\Psi}_{i,[p_0]}^{\mathrm{T}}\widehat{\Psi}_{i,[p_0]})^{-1}\widehat{\Psi}_{i,[p_0]}^{\mathrm{T}}\}(\Psi_{i,[p_0]}\boldsymbol{\xi}_{i,[p_0]}+\boldsymbol{U}_i+\boldsymbol{\mu}_i-\widehat{\boldsymbol{\mu}}_i) \\
&:= \boldsymbol{U}_i + \widehat{\boldsymbol{r}}_i,
\end{aligned}
$$

where $\widehat{\boldsymbol{r}}_i = \widehat{\boldsymbol{r}}_{1i} - \widehat{\boldsymbol{r}}_{2i} + \widehat{\boldsymbol{r}}_{3i}$, $\widehat{\boldsymbol{r}}_{1i} = \{I - \widehat{\Psi}_{i,[p_0]}(\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Lambda}^{-1}_{[p_0]} + \widehat{\Psi}^{\mathrm{T}}_{i,[p_0]}\widehat{\Psi}_{i,[p_0]})^{-1}\widehat{\Psi}^{\mathrm{T}}_{i,[p_0]}\}(\boldsymbol{\mu}_i - \widehat{\boldsymbol{\mu}}_i)$, $\widehat{\boldsymbol{r}}_{2i} = \widehat{\Psi}_{i,[p_0]}(\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Lambda}^{-1}_{[p_0]} + \widehat{\Psi}^{\mathrm{T}}_{i,[p_0]}\widehat{\Psi}_{i,[p_0]})^{-1}\widehat{\Psi}^{\mathrm{T}}_{i,[p_0]}\boldsymbol{U}_i$, $\widehat{\boldsymbol{r}}_{3i} = \{I - \widehat{\Psi}_{i,[p_0]}(\widetilde{\sigma}^2_{u,\mathrm{I}}\widehat{\Lambda}^{-1}_{[p_0]} + \widehat{\Psi}^{\mathrm{T}}_{i,[p_0]}\widehat{\Psi}_{i,[p_0]})^{-1}\widehat{\Psi}^{\mathrm{T}}_{i,[p_0]}\}\Psi_{i,[p_0]}\boldsymbol{\xi}_{i,[p_0]}$. Using rate calculations as before, we can see that

$$\|\widehat{\boldsymbol{r}}_{1i}\|^2 \le \|\widehat{\boldsymbol{\mu}}_i - \boldsymbol{\mu}_i\|^2 = O_p[m_i \times \{h_\mu^4 + \delta^2_{n1}(h_\mu)\}], \quad \|\widehat{\boldsymbol{r}}_{2i}\|^2 = O_p(1),$$
$$\|\widehat{\boldsymbol{r}}_{3i}\|^2 \le \|(\widehat{\Psi}_{i,[p_0]} - \Psi_{i,[p_0]})\boldsymbol{\xi}_{i,[p_0]}\|^2 + O_p(1) = O_p(m_i\varrho_n^2)$$

Therefore, $m_i^{-1}\|\widehat{\boldsymbol{r}}_i\|^2 \le 3(\|\widehat{\boldsymbol{r}}_{1i}\|^2 + \|\widehat{\boldsymbol{r}}_{2i}\|^2 + \|\widehat{\boldsymbol{r}}_{3i}\|^2)/m_i = O_p(\varrho_n^2)$.

Next, it is easy to see that $\widehat{\Pi}$ is of rank $p - p_0$, and suppose it yields a singular value decomposition $\widehat{\Pi}_i = \widehat{\boldsymbol{P}}_i \mathrm{diag}(\widehat{\pi}_{i1}, \ldots, \widehat{\pi}_{i,p-p_0})\widehat{\boldsymbol{Q}}_i^{\mathrm{T}}$, where $\widehat{\boldsymbol{P}}_i = (\widehat{\boldsymbol{p}}_{i,1}, \ldots \widehat{\boldsymbol{p}}_{i,p-p_0})$ and $\widehat{\boldsymbol{Q}}_i = (\widehat{\boldsymbol{q}}_{i,1}, \ldots, \widehat{\boldsymbol{q}}_{i,p-p_0})$ are $m_i \times (p-p_0)$ matrices with the ℓth columns, $\widehat{\boldsymbol{p}}_{i,\ell} = (\widehat{p}_{i,1\ell}, \ldots, \widehat{p}_{i,m_i\ell})^{\mathrm{T}}$ and $\widehat{\boldsymbol{q}}_{i,\ell} = (\widehat{q}_{i,1\ell}, \ldots, \widehat{q}_{i,m_i\ell})^{\mathrm{T}}$, being the left and right singular vectors of $\widehat{\Pi}_i$. One can easily show (e.g. by Theorem 7.7.6 in Hort and Johnson, 1985), that $0 \le \widehat{\pi}_{ij} \le 1$ for $j = 1, \ldots, p - p_0$.

Therefore,

$$\widehat{\sigma}^2_{[p_0]} - \widehat{\sigma}^2_{[p]} = \frac{1}{N}\sum_{i=1}^{n}\sum_{\ell=1}^{p-p_0}\{2\widehat{\pi}_{i,\ell}(\widehat{\boldsymbol{p}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i) - \widehat{\pi}^2_{i,\ell}(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)^2\}.$$

It can be seen that $\widehat{q}_{i,j\ell}$ is a functional of the estimated covariance function $\widehat{R}(\cdot,\cdot)$ and the variance estimator $\widetilde{\sigma}^2_{u,\mathrm{I}}$. We can define its counterpart $\widehat{q}^{(-i)}_{i,j\ell}$ by plugging in the estimators $\widehat{R}^{(-i)}(\cdot,\cdot)$ and $(\widetilde{\sigma}^{(-i)}_{i,\mathrm{I}})^2$ excluding data from the ith curve. By the asymptotic convergence rates and expansions in Lemmas S.1.1 and S.3.1, considering the influence of the ith curve on the $\widehat{R}(\cdot)$ and $\widetilde{\sigma}^2_{u,\mathrm{I}}$, we find that $\widehat{q}_{i,j\ell} - \widehat{q}^{(-i)}_{i,j\ell} = O_p(n^{-1/2}\varrho_n)$. Combining the results above, one can see that

$$\begin{aligned}
m_i^{-1}(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)^2 &\le 2m_i^{-1}(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\boldsymbol{U}_i)^2 + 2m_i^{-1}(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{r}}_i)^2 \\
&\le 4m_i^{-1}(\boldsymbol{U}_i^{\mathrm{T}}\widehat{\boldsymbol{q}}^{(-i)}_{i,\ell})^2 + 4m_i^{-1}\{\boldsymbol{U}_i^{\mathrm{T}}(\widehat{\boldsymbol{q}}_{i,\ell} - \widehat{\boldsymbol{q}}^{(-i)}_{i,\ell})\}^2 + 2m_i^{-1}\|\widehat{\boldsymbol{r}}_i\|^2 \\
&= 4m_i^{-1}(\boldsymbol{U}_i^{\mathrm{T}}\widehat{\boldsymbol{q}}^{(-i)}_{i,\ell})^2 + O_p(n^{-1}\varrho_n^2) + O_p(\varrho_n^2).
\end{aligned}$$

It is easy to see that $\widehat{\boldsymbol{q}}^{(-i)}_{i,\ell}$ is independent with $\boldsymbol{U}_i$, and $\mathrm{E}\{(\boldsymbol{U}_i^{\mathrm{T}}\widehat{\boldsymbol{q}}^{(-i)}_{i,\ell})^2\} = \sigma_u^2\mathrm{E}(\|\widehat{\boldsymbol{q}}^{-i}_{i,\ell}\|^2) = \sigma_u^2$.

Therefore, $(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)^2 = O_p(1 + m_i\varrho_n^2)$, and similarly $(\widehat{\boldsymbol{p}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)^2$ has the same rate. By straight forward rate calculations, we have

$$|\widehat{\sigma}^2_{[p_0]} - \widehat{\sigma}^2_{[p]}| \le \frac{1}{N}\sum_{i=1}^{n}\sum_{\ell=1}^{p-p_0}\{(\widehat{\boldsymbol{p}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)^2 + 2(\widehat{\boldsymbol{q}}^{\mathrm{T}}_{i,\ell}\widehat{\boldsymbol{U}}_i)^2\} = O_p(n/N + \varrho_n^2).$$

Proof of Theorem 2: For the interest of space, we only show the consistency of IC. The consistency of PC follows the similar arguments.

For $p < p_0$, by Lemma S.3.3, we have $IC(p) - IC(p_0) = (\widehat{\sigma}^2_{[p]} - \widehat{\sigma}^2_{[p_0]})/\widehat{\sigma}^2_{[p_0]} \times \{1 + o_p(1)\} + (p - p_0)g_n \xrightarrow{p} \tau_p/\sigma^2_u > 0$. Therefore $IC(p) > IC(p_0)$ with probability tending to 1.

When $p > p_0$, by Lemma S.3.4, $IC(p) - IC(p_0) = (\widehat{\sigma}^2_{[p]} - \widehat{\sigma}^2_{[p_0]})/\widehat{\sigma}^2_{[p_0]} \times \{1 + o_p(1)\} + (p - p_0)g_n = (p - p_0)g_n + O_p(n/N + \varrho^2_n)$. By condition (ii) of the theorem, again, we have $IC(p) > IC(p_0)$ with probability tending to 1.

Therefore, $\widehat{p}$ that minimizes $IC(p)$ converge to p_0 with probability tending to 1.

Proof of Corollary 1: Again, we only show the consistency of $IC(p)$. Following the proof of Theorem 2, condition (i) guarantees $IC(p) > IC(p_0)$ with probability tending to 1, for $p < p_0$.

When $p > p_0$, under the choice of bandwidths in the Corollary, $\widehat{\sigma}^2_{[p]} - \widehat{\sigma}^2_{[p_0]} = O_p(C_n^{-2})$. Using similar arguments as for Theorem 2, condition (ii) ensures $IC(p) > IC(p_0)$ with probability tending to 1 for $p > p_0$.

S.4 Additional Simulations

S.4.1 Expanded tables

Tables S.2 - S.5 are expanded versions of Tables 1 - 4 in the paper. We provide additional results on the minimum description length methods (criteria named DL_2 and DL_N) by Poskitt and Sengarapillai (2011) and the PC_p and IC_p criteria defined in (20).

S.4.2 Sensitivity of the proposed criteria to the choice of bandwidths

To test the sensitivity of the proposed information criteria to the choice of the bandwidths, we repeat the simulation for Scenario I and for the case $m = 10$ using some different bandwidths. We multiply our original choice of bandwidths by a common factor $\varrho = 0.5, 0.9, 1.1$ or 1.5. In other words, we either increase or decrease all bandwidths by 50% or 10%. The new results are shown in Table S.6. By comparing the results above with those in Table S.3, we find that all of the proposed procedures are valid for a relative wide range of bandwidths and are not sensitive to these choices. Despite the changes in the bandwidths, Yao's AIC and the MDL methods by Poskitt and Sengarapillai (2011) consistently pick much larger orders than the truth.

S.4.3 Performance of the proposed criteria under large sample size

For the limited sample sizes considered above, the proposed BIC performs not so well for the sparse data case, e.g. the case of $m = 5$. To verify its consistency, we repeat the simulations in Scenario I and increase the sample size to $n = 2000$. We present the case when the data are relatively sparse, i.e. $m = 5$ and 10. For such a large sample size, the automatic bandwidth selection algorithm (i.e. GCV) in the PACE package broke down because the computer ran out of memory (our simulations were run on a Dell PowerEdge 1950 server with two dual core processors at 3.73 GHz and 4 GB RAM). Therefore, for Yao's AIC we use our own choice of bandwidths.

The empirical distributions of $\widehat{p}$ for various criteria under this large sample scenario are presented in Table S.7. By comparing to the results in Table S.2-S.3, we find that the empirical probability of the proposed BIC picking the correct order has increased significantly by increasing the sample size. Especially for the case $m = 5$, this empirical probability has increased from 38% to 93.5%. The proposed AIC and the IC_p criteria in (20) perform consistently well, picking the correct order 100% of the time. The PC_p criteria perform less well for the sparse case where $m = 5$, but pick the right model 100% of the time when $m = 10$. In contrast, Yao's AIC and the MDL methods continue to pick much larger numbers than the true value.

S.4.4 Performance of the information criteria when m is random

We adopt the setting in Scenario I, allowing m_i to be subject specific. We let m_i's follow a discrete uniform distribution from 5 to 15, such that $\mathrm{E}(m_i) = 10$. The performance of the considered information criteria is show in Table S.8. The results for AIC, PC_p and IC_p are slightly worse than when m is held fixed at 10. Compared with the results in Table S.3, these methods seem to have a slightly higher tendency of selecting an over-fitted model when m_i's vary. On the other hand, the proposed BIC seems to be rather robust under the random m setting. The pseudo-AIC by Yao et al. and the description length by Poskitt and Sengarapillai (2011) continue to fail.

Scenario	Method	$\widehat{p} \leq 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \geq 5$
I	AIC_{PACE}	0.000	0.008	0.000	0.121	0.870
	AIC	0.000	0.405	0.580	0.010	0.005
	BIC	0.155	0.335	0.380	0.115	0.015
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.005	0.565	0.410	0.010	0.010
	PC_{p2}	0.005	0.570	0.405	0.010	0.010
	PC_{p3}	0.005	0.555	0.420	0.010	0.010
	IC_{p1}	0.000	0.215	0.735	0.045	0.005
	IC_{p2}	0.000	0.220	0.730	0.045	0.005
	IC_{p3}	0.000	0.210	0.740	0.045	0.005
II	AIC_{PACE}	0.000	0.000	0.005	0.125	0.870
	AIC	0.000	0.205	0.630	0.155	0.010
	BIC	0.230	0.395	0.245	0.110	0.020
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	0.375	0.440	0.185
	PC_{p2}	0.000	0.000	0.380	0.445	0.175
	PC_{p3}	0.000	0.000	0.365	0.450	0.185
	IC_{p1}	0.000	0.140	0.605	0.210	0.045
	IC_{p2}	0.000	0.140	0.620	0.200	0.040
	IC_{p3}	0.000	0.135	0.605	0.215	0.045
III	AIC_{PACE}	0.000	0.025	0.005	0.130	0.840
	AIC	0.000	0.035	0.720	0.170	0.075
	BIC	0.335	0.260	0.325	0.080	0.000
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.220	0.640	0.075	0.065
	PC_{p2}	0.000	0.230	0.630	0.075	0.065
	PC_{p3}	0.000	0.215	0.640	0.080	0.065
	IC_{p1}	0.000	0.005	0.590	0.280	0.125
	IC_{p2}	0.000	0.005	0.600	0.275	0.120
	IC_{p3}	0.000	0.005	0.585	0.285	0.125
IV	AIC_{PACE}	0.000	0.015	0.015	0.145	0.825
	AIC	0.000	0.020	0.710	0.185	0.085
	BIC	0.315	0.180	0.410	0.070	0.025
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.160	0.640	0.095	0.105
	PC_{p2}	0.000	0.165	0.640	0.090	0.105
	PC_{p3}	0.000	0.150	0.645	0.100	0.105
	IC_{p1}	0.000	0.015	0.560	0.260	0.165
	IC_{p2}	0.000	0.015	0.570	0.260	0.155
	IC_{p3}	0.000	0.015	0.545	0.275	0.165

Table S.2: Expanded version of Table 1.

Scenario	Method	$\widehat{p} \leq 1$	$\widehat{p} = 2$	$\widehat{p} = 3$	$\widehat{p} = 4$	$\widehat{p} \geq 5$
I	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.005	0.980	0.015	0.000
	BIC	0.000	0.040	0.670	0.255	0.035
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_n	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.040	0.955	0.000	0.005
	PC_{p2}	0.000	0.040	0.955	0.000	0.005
	PC_{p3}	0.000	0.030	0.965	0.000	0.005
	IC_{p1}	0.000	0.005	0.985	0.010	0.000
	IC_{p2}	0.000	0.005	0.985	0.010	0.000
	IC_{p3}	0.000	0.005	0.985	0.010	0.000
II	AIC_{PACE}	0.000	0.000	0.000	0.005	0.995
	AIC	0.000	0.000	0.710	0.260	0.030
	BIC	0.000	0.170	0.665	0.135	0.030
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	0.570	0.355	0.075
	PC_{p2}	0.000	0.000	0.575	0.355	0.070
	PC_{p3}	0.000	0.000	0.545	0.380	0.075
	IC_{p1}	0.000	0.000	0.805	0.185	0.010
	IC_{p2}	0.000	0.000	0.805	0.185	0.010
	IC_{p3}	0.000	0.000	0.785	0.200	0.015
III	AIC_{PACE}	0.000	0.015	0.000	0.000	0.985
	AIC	0.000	0.000	0.580	0.400	0.020
	BIC	0.005	0.035	0.770	0.145	0.045
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	0.965	0.030	0.005
	PC_{p2}	0.000	0.000	0.970	0.025	0.005
	PC_{p3}	0.000	0.000	0.965	0.030	0.005
	IC_{p1}	0.000	0.000	0.665	0.320	0.015
	IC_{p2}	0.000	0.000	0.670	0.320	0.010
	IC_{p3}	0.000	0.000	0.665	0.320	0.015
IV	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.830	0.150	0.020
	BIC	0.010	0.005	0.775	0.190	0.020
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	0.920	0.045	0.035
	PC_{p2}	0.000	0.000	0.930	0.040	0.030
	PC_{p3}	0.000	0.000	0.920	0.040	0.040
	IC_{p1}	0.000	0.000	0.900	0.085	0.015
	IC_{p2}	0.000	0.000	0.920	0.070	0.010
	IC_{p3}	0.000	0.000	0.895	0.090	0.015

Table S.3: Expanded version of Table 2.

Scenario	Method	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	$\widehat{p}\geq 5$
I	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.830	0.150	0.020
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_n	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	PC_{p2}	0.000	0.000	1.000	0.000	0.000
	PC_{p3}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p2}	0.000	0.000	1.000	0.000	0.000
	IC_{p3}	0.000	0.000	1.000	0.000	0.000
II	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.630	0.320	0.050
	BIC	0.000	0.000	0.795	0.185	0.020
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	0.955	0.045	0.000
	PC_{p2}	0.000	0.000	0.965	0.035	0.000
	PC_{p3}	0.000	0.000	0.915	0.085	0.000
	IC_{p1}	0.000	0.000	0.945	0.055	0.000
	IC_{p2}	0.000	0.000	0.955	0.045	0.000
	IC_{p3}	0.000	0.000	0.910	0.090	0.000
III	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.775	0.200	0.025
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	PC_{p2}	0.000	0.000	1.000	0.000	0.000
	PC_{p3}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p2}	0.000	0.000	1.000	0.000	0.000
	IC_{p3}	0.000	0.000	1.000	0.000	0.000
IV	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.945	0.055	0.000
	BIC	0.000	0.000	0.835	0.140	0.025
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	PC_{p2}	0.000	0.000	1.000	0.000	0.000
	PC_{p3}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p2}	0.000	0.000	1.000	0.000	0.000
	IC_{p3}	0.000	0.000	0.995	0.005	0.000

Table S.4: Expanded version of Table 3.

Scenario	Method	$\widehat{p} \leq 4$	$\widehat{p} = 5$	$\widehat{p} = 6$	$\widehat{p} = 7$	$\widehat{p} \geq 8$
m=5	AIC_{PACE}	0.005	0.005	0.705	0.245	0.040
	AIC	0.165	0.330	0.470	0.035	0.000
	BIC	0.835	0.020	0.090	0.050	0.005
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.580	0.345	0.070	0.005	0.000
	PC_{p2}	0.590	0.345	0.060	0.005	0.000
	PC_{p3}	0.570	0.355	0.070	0.005	0.000
	IC_{p1}	0.060	0.335	0.545	0.060	0.000
	IC_{p2}	0.070	0.325	0.545	0.060	0.000
	IC_{p3}	0.060	0.325	0.550	0.065	0.000
m=10	AIC_{PACE}	0.005	0.000	0.065	0.475	0.455
	AIC	0.000	0.000	0.570	0.280	0.15
	BIC	0.250	0.030	0.525	0.165	0.030
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.145	0.775	0.020	0.060
	PC_{p2}	0.000	0.170	0.750	0.025	0.055
	PC_{p3}	0.000	0.130	0.790	0.020	0.060
	IC_{p1}	0.000	0.000	0.705	0.185	0.110
	IC_{p2}	0.000	0.000	0.720	0.190	0.090
	IC_{p3}	0.000	0.000	0.700	0.190	0.110
m=50	AIC_{PACE}	0.000	0.065	0.000	0.000	0.935
	AIC	0.000	0.000	0.260	0.405	0.335
	BIC	0.005	0.000	0.590	0.325	0.080
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	0.980	0.010	0.010
	PC_{p2}	0.000	0.000	0.985	0.005	0.010
	PC_{p3}	0.000	0.000	0.980	0.010	0.010
	IC_{p1}	0.000	0.000	0.965	0.035	0.000
	IC_{p2}	0.000	0.000	0.975	0.025	0.000
	IC_{p3}	0.000	0.000	0.930	0.070	0.000

Table S.5: Expanded version of Table 4.

ϱ	Method	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	$\widehat{p}\geq 5$
0.5	$\mathrm{AIC}_{\mathrm{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.005	0.935	0.040	0.020
	BIC	0.285	0.535	0.150	0.010	0.020
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.045	0.850	0.040	0.065
	PC_{p2}	0.000	0.055	0.845	0.035	0.065
	PC_{p3}	0.000	0.040	0.855	0.040	0.065
	IC_{p1}	0.000	0.010	0.970	0.020	0.000
	IC_{p2}	0.000	0.010	0.975	0.015	0.000
	IC_{p3}	0.000	0.010	0.965	0.025	0.000
0.9	$\mathrm{AIC}_{\mathrm{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	0.995	0.005	0.000
	BIC	0.000	0.035	0.770	0.155	0.040
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.010	0.980	0.010	0.000
	PC_{p2}	0.000	0.010	0.980	0.010	0.000
	PC_{p3}	0.000	0.010	0.980	0.010	0.000
	IC_{p1}	0.000	0.005	0.995	0.000	0.000
	IC_{p2}	0.000	0.005	0.995	0.000	0.000
	IC_{p3}	0.000	0.005	0.995	0.000	0.000
1.1	$\mathrm{AIC}_{\mathrm{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.015	0.730	0.200	0.055
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.010	0.990	0.000	0.000
	PC_{p2}	0.000	0.015	0.985	0.000	0.000
	PC_{p3}	0.000	0.010	0.990	0.000	0.000
	IC_{p1}	0.000	0.005	0.995	0.000	0.000
	IC_{p2}	0.000	0.005	0.995	0.000	0.000
	IC_{p3}	0.000	0.005	0.995	0.000	0.000
1.5	$\mathrm{AIC}_{\mathrm{PACE}}$	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.730	0.230	0.040
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.040	0.960	0.000	0.000
	PC_{p2}	0.000	0.055	0.945	0.000	0.000
	PC_{p3}	0.000	0.035	0.965	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p2}	0.000	0.000	1.000	0.000	0.000
	IC_{p3}	0.000	0.000	1.000	0.000	0.000

Table S.6: Sensitivity of the criteria to the choice of bandwidths, based on Scenario I, $m = 10$. All bandwidths (h_μ, h_C and h_σ) are multiplied by a common factor ϱ, and the table shows the empirical distribution of $\widehat{p}$ for various information criteria considered.

m	Method	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	$\widehat{p}\geq 5$
5	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.010	0.935	0.045	0.010
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.180	0.820	0.000	0.000
	PC_{p2}	0.000	0.185	0.815	0.000	0.000
	PC_{p3}	0.000	0.180	0.820	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p2}	0.000	0.000	1.000	0.000	0.000
	IC_{p3}	0.000	0.000	1.000	0.000	0.000
10	AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
	AIC	0.000	0.000	1.000	0.000	0.000
	BIC	0.000	0.000	0.925	0.075	0.000
	DL_2	0.000	0.000	0.000	0.000	1.000
	DL_N	0.000	0.000	0.000	0.000	1.000
	PC_{p1}	0.000	0.000	1.000	0.000	0.000
	PC_{p2}	0.000	0.000	1.000	0.000	0.000
	PC_{p3}	0.000	0.000	1.000	0.000	0.000
	IC_{p1}	0.000	0.000	1.000	0.000	0.000
	IC_{p2}	0.000	0.000	1.000	0.000	0.000
	IC_{p3}	0.000	0.000	1.000	0.000	0.000

Table S.7: Performance of the considered criteria under large samples. The simulations are based on Scenario I, with the sample size increased to $n = 2000$.

Method	$\widehat{p}=1$	$\widehat{p}=2$	$\widehat{p}=3$	$\widehat{p}=4$	$\widehat{p}\geq 5$
AIC_{PACE}	0.000	0.000	0.000	0.000	1.000
AIC	0.000	0.000	0.680	0.210	0.110
BIC	0.020	0.145	0.730	0.095	0.010
DL_2	0.000	0.000	0.000	0.000	1.000
DL_N	0.000	0.000	0.000	0.000	1.000
PC_{p1}	0.000	0.000	0.775	0.090	0.135
PC_{p2}	0.000	0.000	0.780	0.090	0.130
PC_{p3}	0.000	0.000	0.770	0.080	0.150
IC_{p1}	0.000	0.000	0.805	0.150	0.045
IC_{p2}	0.000	0.000	0.810	0.150	0.040
IC_{p3}	0.000	0.000	0.795	0.150	0.055

Table S.8: Performance of the considered criteria under Scenario I, when m_i's are random with the mean value equals to 10.