Statistica Sinica Preprint No: SS-2021-0223	
Title	Factor-augmented Model for Functional Data
Manuscript ID	SS-2021-0223
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202021.0223
Complete List of Authors	Yuan Gao, Han Lin Shang and Yanrong Yang
Corresponding Authors	Yanrong Yang
E-mails	yanrong.yang@anu.edu.au
Notice: Accepted version subject to English editing.	

Statistica Sinica

FACTOR-AUGMENTED MODEL FOR FUNCTIONAL DATA

Yuan Gao†, Han Lin Shang* and Yanrong Yang†

†*Australian National University*

* *Macquarie University*

Abstract:

We propose modeling raw functional data as a mixture of a smooth function and a high-dimensional factor component. The conventional approach to retrieving the smooth function from the raw data is through various smoothing techniques. However, the smoothing model is inadequate to recover the smooth curve or capture the data variation in some situations. These include cases where there is a large amount of measurement error, the smoothing basis functions are incorrectly identified, or the step jumps in the functional mean levels are neglected. A factor-augmented smoothing model is proposed to address these challenges, and an iterative numerical estimation approach is implemented in practice. Including the factor model component in the proposed method solves the aforementioned problems since a few common factors often drive the variation that cannot be captured by the smoothing model. Asymptotic theorems are also established to demonstrate the effects of including factor structures on the smoothing results. Specifically, we show that the smoothing coefficients projected on the complement space of the factor loading matrix are asymptotically normal. As a byproduct of independent interest, an estimator for the population covariance matrix of the raw data is presented based on the proposed model. Extensive simulation studies illustrate that these factor adjustments are essential in improving estimation accuracy and avoiding the curse of dimensionality. The

The authors acknowlege Editor, Associate Editor and two reviewers for their constructive comments, that result in much improvement of this paper.

superiority of our model is also shown in modeling Australian temperature data.

Key words and phrases: Basis function misspecification, functional data smoothing, high-dimensional factor model, measurement error, statistical inference on covariance estimation

1. Introduction

Functional data analysis (FDA) has received growing attention with the increasing capability to store data over the last 20 years. Functional data are considered realizations of smooth random objects in graphical representations of curves, images, and shapes. The monographs of Ramsay and Silverman (2002, 2005) and Ramsay and Hooker (2017) provide a comprehensive account of the methodology and applications of the FDA; other relevant monographs include Ferraty and Vieu (2006) and Horváth and Kokoszka (2012). More recent advances in this field can be found in many survey papers (see, e.g., Cuevas, 2014; Febrero-Bande et al., 2017; Goia and Vieu, 2016; Reiss et al., 2017; Wang et al., 2016). One main challenge in the FDA lies in the fact that we cannot observe functional curves directly, but only discrete points, which are often contaminated by measurement errors. To model a mixture of functional data and high-dimensional measurement error, we introduce a factor-augmented smoothing model (FASM).

We denote a random sample of n functional data as $\mathcal{X}_i(u), i = 1, \ldots, n$, and $u \in \mathcal{I} \subset \mathbb{R}$, where $\mathcal{I}$ is a compact interval on the real line $\mathbb{R}$. In practice, the observed data are discrete points and are often contaminated by noise or measurement error. We use Y_{ij} to represent the jth observation on the ith subject; the observed data can then be expressed as a "signal plus noise" model:

$$Y_{ij} = \mathcal{X}_i(u_j) + \eta_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, p.$$

We use $\mathcal{X}_i(u_j)$ to denote the realization of the jth discrete point on the curve $\mathcal{X}_i(\cdot)$, and η_{ij} is the noise or measurement error. We assume that measurement error only occurs where the measurements are taken; thus, the error $\boldsymbol{\eta}_i = (\eta_{i1}, \ldots, \eta_{ip})$ is a multivariate term of dimension p. Though in practice, the signal function component $\boldsymbol{\mathcal{X}}_i = (\mathcal{X}_i(u_1), \ldots, \mathcal{X}_i(u_j))$ is of the same p dimension, it differs from $\boldsymbol{\eta}_i$ in nature. Although functions are potentially infinite-dimensional, we may impose smoothing assumptions on the functions, which usually implies functions possess one or more derivatives. This smoothness feature is used to separate the functions from measurement errors – a functional smoothing procedure.

When the variance of the noise level is a tiny fraction of the variance of the function, we say the signal-to-noise ratio is high. In this case, classic smoothing tools apply to functional data, including kernel methods (e.g. Wand and Jones, 1995), local polynomial smoothing (e.g. Fan and Gijbels, 1996), and spline smoothing (e.g. Wahba, 1990; Eubank, 1999; Green and Silverman, 1999). With pre-smoothed functions, estimates, such as mean and covariance functions, can be further obtained. More recent studies on functional smoothing approaches include Cai and Yuan (2011); Yao and Li (2013), and Zhang and Wang (2016). In this article, we apply basis smoothing to the functions $\mathcal{X}_i(u)$; that is, we represent $\mathcal{X}_i(u)$ as $\mathcal{X}_i(u) = \sum_{k=1}^{K} c_{ik}\phi_k(u)$, where $\{\phi_k(u),\ k = 1, \ldots, K\}$ are the basis functions and $\{c_{ik},\ i = 1, \ldots, n,\ k = 1, \ldots, K\}$ are the smoothing coefficients. The smoothing model then becomes

$$Y_{ij} = \sum_{k=1}^{K} c_{ik}\phi_k(u_j) + \eta_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, p.$$

When the signal-to-noise level is low, smoothing tools may not be adequate in removing

the measurement error and may cause an inefficient estimation of the smoothing coefficients. Let us take a further look at the measurement error η_{ij}. In the FDA, the number of discrete points p on each subject is often large compared with the sample size n. Hence the term $\boldsymbol{\eta}_i$ is a high-dimensional component. In this case, the observed data are, in fact, a mixture of functional data and high-dimensional data. The existence of the large measurement error η_{ij} raises the curse of dimensionality problem, which naturally calls for the application of dimension reduction models to η_{ij}. Many studies have been conducted on various dimension reduction techniques for high-dimensional data; among these, factor models are widely used (e.g. Fan et al., 2008; Lam et al., 2011).

We propose using a factor model for the measurement error term. Without further information on the measurement error, the factor model is appropriate since the estimation of latent factors does not require any observed variables. The high-dimensional measurement error is assumed to be driven by a small number of unobserved common factors.

$$\eta_{ij} = \boldsymbol{a}_j^\top \boldsymbol{f}_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, p,$$

where $\boldsymbol{f}_i \in \mathbb{R}^r$ are the unobserved factors, $\boldsymbol{a}_j \in \mathbb{R}^r$ are the unobserved factor loadings, r is the number of latent factors, and ϵ_{ij} are idiosyncratic errors with mean zero. Thus, the observed data Y_{ij} can be written as the sum of two components:

$$Y_{ij} = \sum_{k=1}^{K} c_{ik}\phi_k(u_j) + \boldsymbol{a}_j^\top \boldsymbol{f}_i + \epsilon_{ij}, \quad i = 1, \ldots, n, \ j = 1, \ldots, p.$$

This is a basis smoothing model with the factor-augmented form. This proposed model can be easily modified to adopt nonparametric smoothing methods. In Section S1, we illustrate the use of spline smoothing approaches. In Section S3.3, the nonparametric smoothing model is applied to simulated data.

This paper motivates the FASM in three considerations, as listed below. In these three cases, using the proposed model remedies the defects of the traditional smoothing model. Examples of the following three motivations are provided in Section 2.

1. In traditional smoothing models, the measurement error η_{ij} is assumed to be non-informative and independently and identically distributed (i.i.d.) in both directions. This is an unrealistic assumption when the measurement errors contain information. With the factor model applied, we assume that a small number of unobserved factors can capture the covariance in the measurement error. This is usually reasonable in practice because a few common factors often drive the occurrence of systematic measurement error.

2. When the smoothing basis functions are incorrectly identified, the smoothing model will lead to an erroneous coefficient estimate and large residuals. The proposed model deals with this problem since the unexplained variation resulting from the basis's misidentification can be modeled with a small number of unobserved common factors.

3. When there are step jumps in the mean level of the functions, neglecting the mean shift in smoothing models will result in large residuals at the point where the jumps occur. The changes in the mean levels of the functions come from a universal source and can be modeled by common factors.

Since the latent factors are unobserved, we propose an iterative approach to simultaneously estimate the smooth function and factors. Principal component analysis (PCA) is used as a tool in estimating the factor model, and penalized least squares estimation is applied to construct the estimator for the smoothing coefficient c_{ik}. We establish the asymptotic theories of the smoothing coefficient estimator, where the consistency of the estimator is proved. We also provide the asymptotic distribution of the projected estimator in the orthogonal complement of the space spanned by the factors $\boldsymbol{f}_i$. The interplay between the smooth component and the factor model component is manifested.

In the remainder of this article, we elaborate on the previously mentioned three motivations in detail, with examples given in Section 2. In Section 3, the model is formally stated, and the iterative estimation approach is provided. We discuss the asymptotic properties of the smoothing coefficients under various assumptions in Section 4. In Section 5, we conduct Monte-Carlo simulations on the proposed model under different settings. A real data example is given in Section 6, and conclusions are drawn in Section 7. Due to the page limit, we include some important model extensions and data analysis and the proofs for the theorems in the Supplement.

2. Motivation

We introduce three examples to motivate the proposed model. In these cases, the smoothing model is inadequate to capture the raw data's signal information. In the first example, when a large measurement error exists, the residuals after smoothing are large with extreme values. In the second example, when the basis functions are selected incorrectly, part of the functions'

variation cannot be captured by the smoothing model. In the third example, when there are step jumps in the functional data, the residuals after smoothing contain gaps. These examples demonstrate that further modeling of the smoothing residuals is needed.

2.1 Functional data with measurement error

Figure 1 shows the rainbow plots of the average daily temperature and log precipitation at 35 locations in Canada. Due to the nature of the two kinds of data, it is reasonable to assume that temperature and log precipitation are functions over time. The two graphs, however, display distinct features. Although there are some perturbations in the temperature plot, it is relatively easy to discern each curve's shape. In the precipitation plot, there is a tremendous amount of variability in the raw data, such that it is almost impossible to observe the underlying shape of the curves.

Smooth temperature data can be retrieved without much difficulty using basic smoothing techniques. The residuals are small, with constant variation. On the other hand, the residuals after smoothing exhibit a high level of variation for the precipitation data and even contain some extreme values. Our model endeavors to further explain the large residuals in similar cases to the precipitation data; we will show the fitting result in Section S4.

2.2 Misidentification of the basis function

It is important to choose the appropriate basis functions in the smoothing method. In this example, we show the inadequacy of the smoothing model when the basis functions are misidentified. We generate functional data using basis functions with changing frequencies. The raw data are

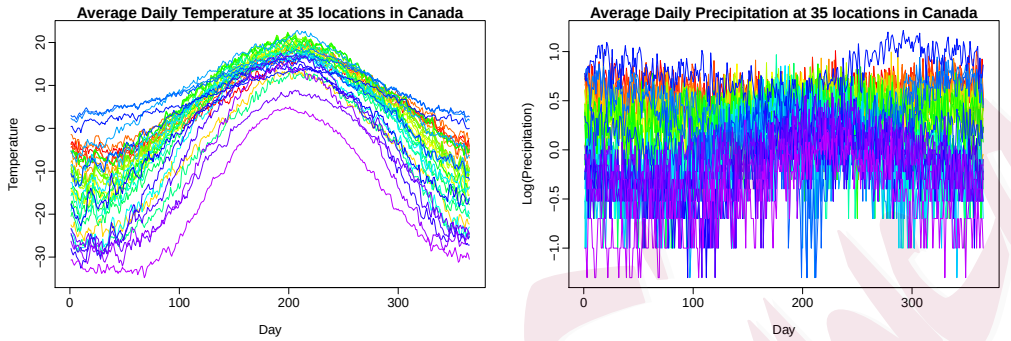
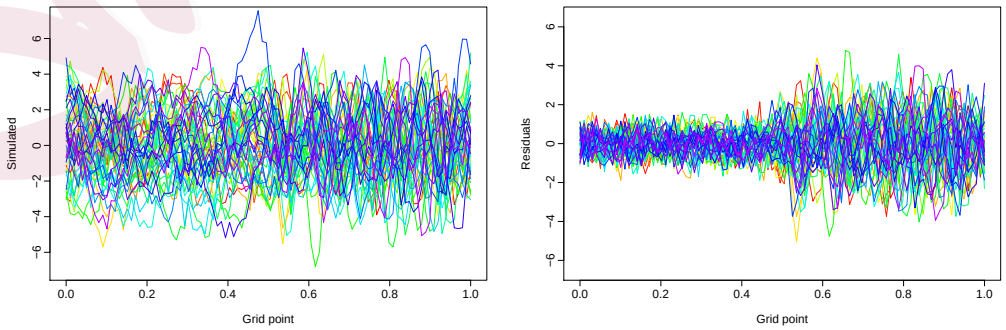


Figure 1: Average daily temperature and log precipitation in 35 Canadian weather stations averaged over the year 1960 to 1994.

shown in Figure 2a. Fourier basis functions are used. In the second half of the data, the frequency of the Fourier basis functions increases, so the data set exhibits more variation toward the right end. Suppose that we were unaware of the change in the frequencies in the basis functions, and still used the basis of the first half of the data for the whole curves. The consequence of misidentifying the basis functions when a smoothing model is applied can be observed in Figure 2b. The residuals are large in the second half. The smoothing model fails to reduce the residuals; a factor model can be used to further model the signal hidden in the large residuals. The data generating process and further analysis can be found in Section S3.4.

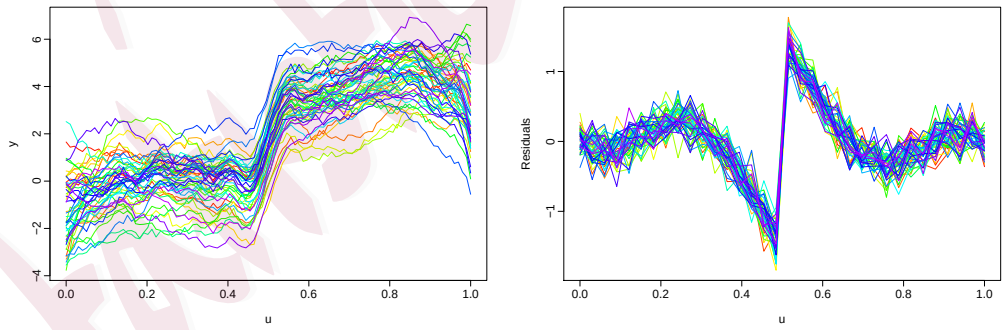


(a) Raw data (b) Residuals

Figure 2: A simulated sample of functional data with changing basis functions.

2.3 Functional data with step jumps in the mean level

We provide another example of functional data with step jumps to motivate our proposed model. Suppose we observed a sample of the raw functional data, as shown in Figure 3a. It can be seen that there is a jump at around $u = 0.5$. The jump applies to all the sample data, so this sudden shift is at the mean level. We will explain how the data are generated in Section S3.5. The residuals after smoothing are presented in Figure 3b. The large residuals around the jump clarify that without measures to deal with the step jumps, smoothing itself is not enough to model these kinds of data. We show in Section S3.5 that the proposed model applied to the same data generates smaller residuals and has less flexibility. This is indeed one of the main goals because of model selection.



(a) Raw data (b) Residuals

Figure 3: A simulated sample of functional data with step jump.

3. Model Specification and Estimation

This section formally states the proposed model in Section 3.1 and provides the estimation method in Section 3.2. We first show how the smoothing coefficient $\boldsymbol{c}_i$ and the latent factors $\boldsymbol{f}_i$

are estimated separately and then introduce an iterative approach to simultaneously find these estimates.

3.1 Factor-augmented smoothing model

We consider a sample of functional data $\mathcal{X}_i(u)$, which takes values in the space $H := L^2(\mathcal{I})$ of real-valued square integrable functions on $\mathcal{I}$. The space H is a Hilbert space, equipped with the inner product $\langle x, y\rangle := \int x(u)y(u)du$. The function norm is defined as $\|x\| := \langle x, x\rangle^{1/2}$. The functional nature of $\mathcal{X}_i(u)$ allows us to represent it as a linear expansion of a set of K smooth basis functions.

$$\mathcal{X}_i(u) = \sum_{k=1}^{K} c_{ik}\phi_k(u), \quad u \in \mathcal{I},$$

where $\{\phi_k(u), k = 1, \ldots, K\}$ is a set of common basis functions and c_{ik} is the kth coefficient for the ith curve. Therefore, we can express the full model as

$$Y_{ij} = \sum_{k=1}^{K} c_{ik}\phi_k(u_j) + \eta_{ij},$$

$$\eta_{ij} = \boldsymbol{a}_j^\top \boldsymbol{f}_i + \epsilon_{ij}, \quad i = 1, \ldots, n,\ j = 1, \ldots, p,$$

where $\boldsymbol{f}_i \in \mathbb{R}^r$ are the unobserved common factors, $\boldsymbol{a}_j \in \mathbb{R}^r$ are the unobserved factor loadings and r is the number of factors. We call this model the FASM. For the model to be identifiable, we require the following condition.

Identification Condition 1. We require

(i) $\{\mathcal{X}_i(u_j) : i = 1, \ldots, n; j = 1, \ldots, p\}$ are independent of $\{\eta_{ij} : i = 1, \ldots, n;\ j = 1, \ldots, p\}$;

(ii) $p^{-1}\sum_{j=1}^{p}\boldsymbol{a}_j\boldsymbol{a}_j^\top \xrightarrow{p} \Sigma_{\boldsymbol{a}} > 0$ for some $r \times r$ matrix $\Sigma_{\boldsymbol{a}}$, as $p \to \infty$;

$n^{-1}\sum_{i=1}^{n}\boldsymbol{f}_i\boldsymbol{f}_i^\top \xrightarrow{p} \Sigma_{\boldsymbol{f}} > 0$ for some $r \times r$ matrix $\Sigma_{\boldsymbol{f}}$, as $n \to \infty$.

The first part of the identification condition ensures the signal function component and the factor model component are independent. The second part ensures the existence of r factors, each of which makes a non-trivial contribution to the variance of η_{ij}, which in turn guarantees the identifiability between the factors and the error term ϵ_{ij}.

We treat the basis functions $\{\phi_k(u) : k = 1, 2, \ldots, K\}$ as known, and the number of basis functions K could be either fixed or going to infinity. There are various choices for the basis functions in empirical data analysis, and the decision can be subjective. For example, Fourier bases are preferred for periodic data, while spline basis systems are most commonly used for non-periodic data. Other bases include wavelet, polynomial, and some ad-hoc basis functions. The number of basis functions K controls the smoothness of the predicted functions. As K increases, the estimator variance increases but the bias decay. Cross-validation, for instance, can be used for decidng K (Wahba and Word, 1975). In this article, our model is estimated with a roughness penalty approach explained in the next section. The smoothness of the functional component is switched from being determined by K to being determined by the tuning parameter of the penalty term. In practice, we can use a relatively large K and select the tuning parameter carefully. In Remark 4, we discuss the common methods for selecting the tuning parameter.

3.2 Penalized estimation approach

We can write the model for the ith object as

$$\boldsymbol{Y}_i = \boldsymbol{\Phi}\boldsymbol{c}_i + \boldsymbol{A}\boldsymbol{f}_i + \boldsymbol{\epsilon}_i, \quad i = 1, \ldots, n \tag{3.1}$$

where

$$\boldsymbol{Y}_i = \begin{bmatrix} Y_{i1} \\ \vdots \\ Y_{ip} \end{bmatrix}, \ \boldsymbol{c}_i = \begin{bmatrix} c_{i1} \\ \vdots \\ c_{iK} \end{bmatrix}, \ \boldsymbol{\Phi} = \begin{bmatrix} \phi_1(u_1) & \ldots & \phi_K(u_1) \\ \vdots & & \vdots \\ \phi_1(u_p) & \ldots & \phi_K(u_p) \end{bmatrix}, \ \boldsymbol{A} = \begin{bmatrix} \boldsymbol{a}_1^\top \\ \vdots \\ \boldsymbol{a}_p^\top \end{bmatrix}, \ \boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_{i1} \\ \vdots \\ \epsilon_{ip} \end{bmatrix}.$$

Combining all the objects, we have in matrix form

$$\boldsymbol{Y} = \boldsymbol{\Phi}\boldsymbol{C} + \boldsymbol{A}\boldsymbol{F}^\top + \boldsymbol{E}, \tag{3.2}$$

where $\boldsymbol{Y}$ is $p \times n$ and $\boldsymbol{C} = (\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n)$ is a $K \times n$ matrix containing all the smoothing coefficients. The matrix $\boldsymbol{F} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n)^\top$ is $n \times r$ and $\boldsymbol{E} = (\boldsymbol{\epsilon}_1, \ldots, \boldsymbol{\epsilon}_n)$ is $p \times n$. Since $\boldsymbol{\Phi}$ is assumed to be known, we illustrate how the parameters $\boldsymbol{C}, \boldsymbol{A}$ and $\boldsymbol{f}$ are estimated in the following.

For the latent factor estimation, there is an identification problem such that $\boldsymbol{A}\boldsymbol{F}^\top = \boldsymbol{A}\boldsymbol{U}\boldsymbol{U}^{-1}\boldsymbol{F}^\top$ for any $r \times r$ invertible matrix $\boldsymbol{U}$. Thus we impose the normalization restriction on the matrices $\boldsymbol{A}$ and $\boldsymbol{F}$

$$\boldsymbol{A}^\top\boldsymbol{A}/p = \boldsymbol{I}_r, \quad \text{and } \boldsymbol{F}^\top\boldsymbol{F} \text{ is a diagonal matrix.} \tag{3.3}$$

We propose to implement penalized least squares, where the objective function is defined as

$$\mathrm{SSR}(\boldsymbol{c}_i, \boldsymbol{A}, \boldsymbol{f}) = \frac{1}{np}\sum_{i=1}^{n}\left[(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i - \boldsymbol{A}\boldsymbol{f}_i)^\top(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i - \boldsymbol{A}\boldsymbol{f}_i) + \alpha\mathrm{PEN}_2(\mathcal{X}_i)\right],$$

where $\mathrm{PEN}_2(\mathcal{X}_i)$ is a penalty term used for regularization, and α is the tuning parameter controlling the degree of regularization. The same α is used for all the functional observations i. This is a simplified case, where we assume a similar degree of smoothness for all curves. The tuning parameter can be chosen by cross-validation or information criteria. We intend to penalize the "roughness" of the function term. To quantify the notion of "roughness" in a function, we use the square of the second derivative. Define the measure of roughness as

$$\mathrm{PEN}_2(\mathcal{X}_i) = \int_{\mathcal{I}}\left[D^2\mathcal{X}_i(s)\right]^2 ds,$$

where $D^2\mathcal{X}_i$ denotes taking the second derivative of the function $\mathcal{X}_i$, the larger the tuning parameter α, the smoother the estimated functions we obtain. Further, we denote

$$\boldsymbol{\Phi}(u) = [\phi_1(u), \ldots, \phi_K(u)]^\top. \tag{3.4}$$

Then

$$\mathcal{X}_i(u) = \boldsymbol{c}_i^\top\boldsymbol{\Phi}(u).$$

We can re-express the roughness penalty $\text{PEN}_2(\mathcal{X}_i)$ in a matrix form as follows:

$$\begin{aligned}
\text{PEN}_2(\mathcal{X}_i) &= \int_{\mathcal{I}} \left[D^2\mathcal{X}_i(s)\right]^2 ds \\
&= \int_{\mathcal{I}} \left[D^2\boldsymbol{c}_i^\top \boldsymbol{\Phi}(s)\right]^2 ds \\
&= \int_{\mathcal{I}} \boldsymbol{c}_i^\top D^2\boldsymbol{\Phi}(s) D^2\boldsymbol{\Phi}^\top(s)\boldsymbol{c}_i ds \\
&= \boldsymbol{c}_i^\top \left[\int_{\mathcal{I}} D^2\boldsymbol{\Phi}(s) D^2\boldsymbol{\Phi}^\top(s) ds\right] \boldsymbol{c}_i \\
&= \boldsymbol{c}_i^\top \boldsymbol{R}\boldsymbol{c}_i, \qquad i = 1, \ldots, n
\end{aligned}$$

where

$$\boldsymbol{R} \equiv \int_{\mathcal{I}} D^2\boldsymbol{\Phi}(s) D^2\boldsymbol{\Phi}^\top(s) ds. \tag{3.5}$$

The matrix $\boldsymbol{R}$ is the same for all subjects, and the penalty term $\text{PEN}_2(\mathcal{X}_i)$ differs for each subject only by the coefficient $\boldsymbol{c}_i$.

Remark 1. The number of smoothing coefficient $\boldsymbol{c}_i$ increases as the sample size increases. The inclusion of a penalty term penalizes the "roughness" of the smoothed function and mitigates the effect of the increasing number of parameters to control the model flexibility.

Thus, the objective function can be written as

$$\text{SSR}(\boldsymbol{c}_i, \boldsymbol{A}, \boldsymbol{f}) = \frac{1}{np}\sum_{i=1}^{n} \left[(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i - \boldsymbol{A}\boldsymbol{f}_i)^\top (\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i - \boldsymbol{A}\boldsymbol{f}_i) + \alpha \boldsymbol{c}_i^\top \boldsymbol{R}\boldsymbol{c}_i\right],$$

subject to the constraint $\boldsymbol{A}^\top\boldsymbol{A}/p = \boldsymbol{I}_r$.

We aim to estimate the smoothing coefficient $\boldsymbol{c}_i$. We left multiply a matrix to each term in (3.1) to project the factor model term onto a zero matrix. Define the projection matrix

$$\boldsymbol{M}_{\boldsymbol{A}} \equiv \boldsymbol{I}_p - \boldsymbol{A}(\boldsymbol{A}^\top \boldsymbol{A})^{-1}\boldsymbol{A}^\top = \boldsymbol{I}_p - \boldsymbol{A}\boldsymbol{A}^\top/p. \tag{3.6}$$

Then

$$\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{A}\boldsymbol{f}_i = \left(\boldsymbol{I}_p - \boldsymbol{A}\boldsymbol{A}^\top/p\right)\boldsymbol{A}\boldsymbol{f}_i = \left(\boldsymbol{A} - \boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{A}/p\right)\boldsymbol{f}_i = \boldsymbol{0}.$$

So we estimate $\boldsymbol{c}_i$ from the projected equation

$$\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{Y}_i = \boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{c}_i + \boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{\epsilon}_i.$$

The projected objective function becomes

$$\mathrm{SSR}(\boldsymbol{c}_i, \boldsymbol{A}) = \frac{1}{np}\sum_{i=1}^{n}\left[(\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{Y}_i - \boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{c}_i)^\top(\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{Y}_i - \boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{c}_i) + \alpha\boldsymbol{c}_i^\top\boldsymbol{R}\boldsymbol{c}_i\right]. \tag{3.7}$$

By taking the derivative of $\mathrm{SSR}(\boldsymbol{c}_i, \boldsymbol{A})$ with respective to each $\boldsymbol{c}_i$ and setting it to zero, we can solve for the estimator $\widehat{\boldsymbol{c}}_i$.

$$\frac{\partial \mathrm{SSR}(\boldsymbol{c}_i\boldsymbol{A})}{\partial \boldsymbol{c}_i} = -\frac{1}{np}(\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{Y}_i - \boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{\Phi}\boldsymbol{c}_i)^\top(\boldsymbol{M}_{\boldsymbol{A}}\boldsymbol{\Phi}) + \frac{1}{np}\alpha\boldsymbol{c}_i^\top\boldsymbol{R}.$$

Setting the derivative to zero and rearranging the terms, we have

$$\left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}}^\top \boldsymbol{M}_{\boldsymbol{A}} \boldsymbol{\Phi} + \alpha \boldsymbol{R}\right) \boldsymbol{c}_i = \boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}}^\top \boldsymbol{M}_{\boldsymbol{A}} \boldsymbol{Y}_i.$$

Using the fact that $\boldsymbol{M}_{\boldsymbol{A}}^\top \boldsymbol{M}_{\boldsymbol{A}} = \left(\boldsymbol{I}_p - \boldsymbol{A}\boldsymbol{A}^\top/p\right)^\top \left(\boldsymbol{I}_p - \boldsymbol{A}\boldsymbol{A}^\top/p\right) = \boldsymbol{M}_{\boldsymbol{A}}$, we obtain the least squares estimator for $\boldsymbol{c}_i$ given $\boldsymbol{A}$

$$\widehat{\boldsymbol{c}}_i = \left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}} \boldsymbol{\Phi} + \alpha \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}} \boldsymbol{Y}_i.$$

Next, to estimate $\boldsymbol{A}$ and $\boldsymbol{f}_i$, we focus on the factor model

$$\boldsymbol{\eta}_i = \boldsymbol{A}\boldsymbol{f}_i + \boldsymbol{\epsilon}_i,$$

and in matrix form

$$\boldsymbol{Z} = \boldsymbol{A}\boldsymbol{F}^\top + \boldsymbol{E},$$

where $\boldsymbol{Z} = (\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_n)$. In high-dimensional cases, the unknown factors and loadings are typically estimated by least squares (i.e., the principal component analysis; see, e.g., Fan et al. (2008); Onatski (2012). The least squares objective function is

$$\operatorname{tr}\left[(\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{F}^\top)(\boldsymbol{Z} - \boldsymbol{A}\boldsymbol{F}^\top)^\top\right]. \tag{3.8}$$

Minimizing the objective function with respect to $\boldsymbol{F}^\top$, we have $\boldsymbol{F}^\top = (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{Z} = \boldsymbol{A}^\top\boldsymbol{Z}/p$

using (3.3). Substituting in (3.8), we obtain the objective function

$$\begin{aligned}
&\mathrm{tr}\left[(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{Z}/p)(\boldsymbol{Z}-\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{Z}/p)^\top\right]\\
&=\mathrm{tr}\left(\boldsymbol{Z}\boldsymbol{Z}^\top-\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A}\boldsymbol{A}^\top/p-\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A}\boldsymbol{A}^\top/p+\boldsymbol{A}\boldsymbol{A}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A}\boldsymbol{A}^\top/p^2\right)\\
&=\mathrm{tr}(\boldsymbol{Z}\boldsymbol{Z}^\top)-\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A})/p,
\end{aligned}$$

where the last equality uses (3.3) and that $\mathrm{tr}(\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A}\boldsymbol{A}^\top)=\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A})$. Thus, minimizing the objective function is equivalent to maximizing $\mathrm{tr}(\boldsymbol{A}^\top\boldsymbol{Z}\boldsymbol{Z}^\top\boldsymbol{A})/p$. The estimator for $\boldsymbol{A}$ is obtained by finding the first r eigenvectors corresponding to the r largest eigenvalues of the matrix $\frac{1}{p}\boldsymbol{Z}\boldsymbol{Z}^\top$ in descending order, where

$$\frac{1}{p}\boldsymbol{Z}\boldsymbol{Z}^\top=\frac{1}{p}\sum_{i=1}^{n}\boldsymbol{\eta}_i\boldsymbol{\eta}_i^\top=\frac{1}{p}\sum_{i=1}^{n}(\boldsymbol{Y}_i-\boldsymbol{\Phi}\boldsymbol{c}_i)(\boldsymbol{Y}_i-\boldsymbol{\Phi}\boldsymbol{c}_i)^\top.$$

Therefore, knowing $\boldsymbol{c}_i$, we solve for $\widehat{\boldsymbol{A}}$ using

$$\left[\frac{1}{np}\sum_{i=1}^{n}(\boldsymbol{Y}_i-\boldsymbol{\Phi}\boldsymbol{c}_i)(\boldsymbol{Y}_i-\boldsymbol{\Phi}\boldsymbol{c}_i)^\top\right]\widehat{\boldsymbol{A}}=\widehat{\boldsymbol{A}}\boldsymbol{V}_{np}, \tag{3.9}$$

where $\boldsymbol{V}_{np}$ is an $r\times r$ diagonal matrix containing the r eigenvalues of the matrix in the square brackets in decreasing order. The additional coefficient $1/n$ is used for scaling.

Remark 2. The number of factors r is selected based on some criteria regarding the eigenvalues. There have been many studies on this topic. Examples include Bai and Ng (2002), where two model selection criteria functions were proposed; Onatski (2010), where the number of factors

was estimated using differenced adjacent eigenvalues; and Ahn and Horenstein (2013), where they select the number based on the ratio of two adjacent eigenvalues. A modified ratio criterion of Ahn and Horenstein (2013) is proposed in Section S3.1. It is worth noting that, simulations and empirical applications in this paper illustrate that our model has good performances under various common selection criteria provided in the literature above.

It can be seen that $\boldsymbol{A}$ is needed to find $\widehat{\boldsymbol{c}}_i$, and in turn $\boldsymbol{c}_i$ is needed to find $\widehat{\boldsymbol{A}}$. The final estimator $(\widehat{\boldsymbol{c}}_i, \widehat{\boldsymbol{A}})$ is the solution of the set of equations

$$\begin{cases} \widehat{\boldsymbol{c}}_i = \left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}} \boldsymbol{\Phi} + \alpha \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}} \boldsymbol{Y}_i, \qquad i = 1, \ldots, n \\ \left[\frac{1}{np} \sum_{i=1}^{n} \left(\boldsymbol{Y}_i - \boldsymbol{\Phi}\widehat{\boldsymbol{c}}_i\right)\left(\boldsymbol{Y}_i - \boldsymbol{\Phi}\widehat{\boldsymbol{c}}_i\right)^\top\right] \widehat{\boldsymbol{A}} = \widehat{\boldsymbol{A}} \boldsymbol{V}_{np}. \end{cases} \tag{3.10}$$

Since there is no closed-form expression of $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{c}}_i$, we propose using numerical iterations to find the estimates. The details of these iterations are as follows:

Algorithm 1: Iterations for estimating FASM

1. Denote the initial value as $\widehat{\boldsymbol{A}}^{(0)}$. Using (3.10), we obtain $\widehat{\boldsymbol{c}}_i^{(0)} = \left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}^{(0)}} \boldsymbol{\Phi} + \alpha \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}^{(0)}} \boldsymbol{Y}_i$.
2. With $\widehat{\boldsymbol{c}}_i^{(t)}$, we substitute into the second equation of (3.10), to obtain $\widehat{\boldsymbol{A}}^{(t+1)} = (\widehat{\boldsymbol{a}}_1^{(t+1)}, \ldots, \widehat{\boldsymbol{a}}_r^{(t+1)})^\top$, where $\widehat{\boldsymbol{a}}_j^{(t+1)}$ is the eigenvector of the matrix $(np)^{-1} \sum_{i=1}^{n} (\boldsymbol{Y}_i - \boldsymbol{\Phi}\widehat{\boldsymbol{c}}_i^{(t+1)})(\boldsymbol{Y}_i - \boldsymbol{\Phi}\widehat{\boldsymbol{c}}_i^{(t+1)})^\top$ corresponding to its jth largest eigenvalue.
3. With $\widehat{\boldsymbol{A}}^{(t+1)}$, we obtain $\widehat{\boldsymbol{c}}_i^{(t+1)} = \left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}^{(t+1)}} \boldsymbol{\Phi} + \alpha \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}^{(t+1)}} \boldsymbol{Y}_i$ using (3.10).
4. We then repeat steps 2 and 3 until $\|\widehat{\boldsymbol{c}}_i^{(t+1)} - \widehat{\boldsymbol{c}}_i^{(t)}\| < \delta$, where δ is a prescribed small positive value.

Remark 3. In this paper, we use $\widehat{\boldsymbol{A}}^{(0)} = \boldsymbol{0}$. This means we start by ignoring the factor model component so the initial value for the smoothing coefficient $\widehat{\boldsymbol{c}}_i^{(0)} = \left(\boldsymbol{\Phi}^\top \boldsymbol{\Phi} + \alpha \boldsymbol{R}\right)^{-1} \boldsymbol{\Phi}^\top \boldsymbol{Y}_i$, which

is simply the ridge estimator. The convergence of Newton's numeric iteration requires the convergence of this estimator, which in turn requires the factor model component η_{ij} to have an expectation of zero. The stopping criterion only focuses on $\widehat{\boldsymbol{c}}_i$ because we are interested in estimating η_{ij} as a whole.

Remark 4. Common methods for selecting the shrinkage parameter α include the Akaikes Information Criterion (AIC Akaike, 1974), the Bayesian Information Criterion (BIC Schwarz, 1978), and cross-validation. In this paper, we use the mean generalized cross-validation (mGCV) method (Golub et al., 1979). We define, at step t,

$$\text{mGCV}^{(t)} = \frac{1}{n}\sum_{i=1}^{n}\frac{\text{pSSE}_i^{(t)}}{[p - df^{(t)}(\alpha)]^2}, \tag{3.11}$$

where $\text{SSE}_i^{(t)}$ is the sum of squares residual for the ith object at step t and $df^{(t)}(\alpha)$ is the equivalent degrees of freedom measure, which can be calculated as

$$\text{df}^{(t)}(\alpha) = \text{trace}\left[\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}^{(t)}}\boldsymbol{\Phi} + \alpha\boldsymbol{R}\right)^{-1}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}^{(t)}}\right]. \tag{3.12}$$

At each step of the iteration, the tuning parameter α is chosen by minimizing the $mGCV^{(t)}$.

Remark 5. Algorithm 1 is an iteration procedure in which ridge regression and PCA are iterated. The convergence of this iterative algorithm is studied in Jiang et al. (2021). For instance, Theorem 2 of Jiang et al. (2021) provides some sufficient conditions under which the recursive algorithm converges to the true value or some other values. In particular, this algorithm will converge to

the true parameter when the regressors are independent of the common factors, or the factors involved in regressors are weaker than the common ones.

After we obtain the estimates $\widehat{\boldsymbol{A}}$ and $\widehat{\boldsymbol{c}}_i$, the estimated coefficient matrix $\widehat{\boldsymbol{C}}$ is constructed as $\widehat{\boldsymbol{C}} = (\widehat{\boldsymbol{c}}_1, \ldots, \widehat{\boldsymbol{c}}_n)$, and the estimated factor can be obtained by

$$\widehat{\boldsymbol{F}}^\top = \widehat{\boldsymbol{A}}^\top (\boldsymbol{Y} - \boldsymbol{\Phi}\widehat{\boldsymbol{C}}).$$

Finally, the functional component can be estimated by $\widehat{\mathcal{X}}_i(u) = \widehat{\boldsymbol{c}}_i^\top \boldsymbol{\Phi}(u)$, where $\boldsymbol{\Phi}(u)$ is defined in (3.4).

Remark 6. Although we have imposed the constraint in (3.3) and the identification condition 1, $\boldsymbol{A}$ and $\boldsymbol{f}_i$ are not uniquely determined, since the model (3.1) is unchanged if we replace $\boldsymbol{A}$ and $\boldsymbol{f}_i$ with $\boldsymbol{AU}$ and $\boldsymbol{U}^\top \boldsymbol{f}_i$ for any orthogonal $r \times r$ matrix $\boldsymbol{U}$. However, the linear space spanned by the columns of $\boldsymbol{A}$ is uniquely defined. Although we are not able to estimate $\boldsymbol{A}$, we can still estimate a rotation of $\boldsymbol{A}$, which spans the same space as $\boldsymbol{A}$ does. The matrix $\boldsymbol{M}_{\boldsymbol{A}}$ defined in (3.6) is a projecting matrix onto the orthogonal complement of the linear space spanned by the columns of $\boldsymbol{A}$. It is shown in the next section that the estimator $\boldsymbol{M}_{\widehat{\boldsymbol{A}}}$ for $\boldsymbol{M}_{\boldsymbol{A}}$ is consistent.

4. Asymptotic Theory

In this section, we study the asymptotic properties of the coefficient estimator $\widehat{\boldsymbol{c}}_i$ with growing sample size and dimension. We state the assumptions in Section 4.1 and provide the asymptotic results of $\widehat{\boldsymbol{c}}_i$ in Section 4.2.

4.1 Assumptions

We use $(\boldsymbol{c}_i^0, \boldsymbol{A}^0)$ to denote the true parameters. In this paper, we use L^2 norm. The norm of a vector $\boldsymbol{x}$ is defined as $\|\boldsymbol{x}\| = \sqrt{\boldsymbol{x}^\top \boldsymbol{x}}$ and the norm of a matrix $\boldsymbol{U}$ is defined as $\|\boldsymbol{U}\| = \sqrt{\lambda_{\max}(\boldsymbol{U}^\top \boldsymbol{U})}$, where $\lambda_{\max}(\cdot)$ represents the largest eigenvalue of a matrix. We introduce the matrix

$$\boldsymbol{D}_i(\boldsymbol{A}) \equiv \frac{1}{p}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}} \boldsymbol{\Phi} - \frac{1}{p}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}} \boldsymbol{\Phi} \boldsymbol{f}_i^\top \left(\frac{\boldsymbol{F}^\top \boldsymbol{F}}{n}\right)^\top \boldsymbol{f}_i. \tag{4.13}$$

This matrix plays an important role in this article. It is used in the proof of the consistency of $\widehat{\boldsymbol{c}}_i$, as can be found in S5.1 Appendix A. The identifying condition for $\boldsymbol{c}_i^0$ is that $\boldsymbol{D}_i(\boldsymbol{A})$ is positive definite for all i, which is stated in Assumption 3.

First, we state the assumptions.

Assumption 1.

$$\frac{1}{p}\left\|\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right\| = O(1), \quad ||\mathbf{R}|| = O(1), \quad K = o(\min(n, p)), \quad p = o\left(n^2\right).$$

In this assumption, the number of basis functions K could be either fixed or tending to infinity.

Assumption 2.

$$\|\boldsymbol{c}_i^0\| = O_p(1), \text{ for all } i = 1, 2, \ldots, n.$$

This assumption is introduced to ensure the consistency of the estimated coefficients $\widehat{\boldsymbol{c}}_i^0$. Notice that the dimension of $\boldsymbol{c}_i^0$ is K, which may go to infinity, but its norm is restricted to be

of constant order. Thus this assumption meanwhile controls the flexibility of the model.

Assumption 3. Let $\mathcal{A} = \{\boldsymbol{A} : \boldsymbol{A}^\top \boldsymbol{A}/p = \boldsymbol{I}, \text{ and } \boldsymbol{A} \text{ independent of } \boldsymbol{\Phi}\}$. We assume

$$\inf_{\boldsymbol{A}\in\mathcal{A}} \boldsymbol{D}_i(\boldsymbol{A}) > 0.$$

This assumption is the identification condition for $\boldsymbol{c}_i^0$. The usual assumption for the least-squares estimator only contains the first term on the right-hand side of (4.13). The second term on the right-hand side of (4.13) arises because of the unobservable matrices $\boldsymbol{F}$ and $\boldsymbol{A}$.

Assumption 4. For some constant $M > 0$,

1. $\mathbb{E}\|\boldsymbol{a}_j^0\|^4 \leq M$, $j = 1, \ldots, p$ and $\frac{1}{p}\boldsymbol{A}^\top \boldsymbol{A} \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{a}} > 0$ for some $r \times r$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{a}}$ as $p \to \infty$.

2. $\mathbb{E}\|\boldsymbol{f}_i\|^4 \leq M$ and $\frac{1}{n}\boldsymbol{F}\boldsymbol{F}^\top \xrightarrow{p} \boldsymbol{\Sigma}_{\boldsymbol{f}} > 0$ for some $r \times r$ matrix $\boldsymbol{\Sigma}_{\boldsymbol{f}}$ as $n \to \infty$.

Assumption 5. For some constant $M > 0$, the error terms $\epsilon_{ji}, j = 1, \ldots, p, i = 1, \ldots, n$ are i.i.d. in both directions, with $\mathbb{E}(\epsilon_{ji}) = 0$, $\mathrm{Var}(\epsilon_{ji}) = \sigma^2$, and $\mathbb{E}|\epsilon_{ji}|^8 \leq M$.

Assumption 6. ϵ_{ji} is independent of ϕ_s, $\boldsymbol{f}_t$, and $\boldsymbol{a}_s^0$ for all j, i, s, t.

We require that the errors are independent in themselves and also of the functional term $\phi(u)$ and factor model terms $\boldsymbol{f}_i$ and $\boldsymbol{a}_j^0$. To not mask the main contribution of our method, we use a simplified setting on the error terms to exclude endogeneity. Nevertheless, with simple but tedious modifications, Assumption 5 can be relaxed, and our model can be extended to more complicated settings where correlations between the error term and the factor model term are allowed.

Assumption 7. The tuning parameter α satisfies $\alpha = o(p)$.

This is conventionally assumed in ridge regression (see, e.g., Knight and Fu, 2000) and assures that the estimator's asymptotic bias is zero.

Before stating the next assumption, we introduce some notations. Let $\boldsymbol{\omega}_j$, $j = 1, \ldots, p$ denote the jth column of the $K \times p$ matrix $\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0}$, and let ψ_{ik} denote the (i,k)th element of the matrix $\boldsymbol{M}_{\boldsymbol{F}}$, where

$$\boldsymbol{M}_{\boldsymbol{F}} \equiv \boldsymbol{I}_n - \boldsymbol{F}\left(\boldsymbol{F}^\top \boldsymbol{F}\right)\boldsymbol{F}^\top. \tag{4.14}$$

Then, for any vector $\boldsymbol{b} = (b_1, \ldots, b_n)^\top$, we can write

$$\frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0} \boldsymbol{E} \boldsymbol{M}_{\boldsymbol{F}} \boldsymbol{b} = \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \boldsymbol{\omega}_j \epsilon_{ji} \sum_k^n \psi_{ik} b_k \equiv \frac{1}{\sqrt{np}} \sum_i^n \sum_j^p \boldsymbol{x}_{ij}. \tag{4.15}$$

In (4.15), for notation simplicity, we define $\boldsymbol{x}_{ij}$ as $\boldsymbol{\omega}_j \epsilon_{ji} \sum_k^n \psi_{ik} b_k$. The matrix $\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0} \boldsymbol{E} \boldsymbol{M}_{\boldsymbol{F}}$ is of interest because it is the main component that contributes to the asymptotic distribution of the estimators, as shall be seen in the next section.

Let

$$\boldsymbol{L}_{np} \equiv \frac{\sigma^2}{np} \sum_i^n \sum_j^p \boldsymbol{\omega}_j \boldsymbol{\omega}_j^\top \left(\sum_k^n \psi_{ik} b_k \right)^2. \tag{4.16}$$

We make the following assumption.

Assumption 8. When K is fixed, we assume there exists a $K \times K$ matrix $\boldsymbol{L}$ such that

$$\boldsymbol{L} \equiv \lim_{n,p \to \infty} \boldsymbol{L}_{np}, \tag{4.17}$$

where $\boldsymbol{L}_{np}$ is defined in (4.16). Let ν^2 be the smallest eigenvalue of the matrix $\boldsymbol{L}$ defined in (4.17), then assume that $\nu^2 > 0$, and that, for all $\varepsilon > 0$,

$$\lim_{n,p\to\infty} \frac{1}{np\nu^2} \sum_{i=1}^{n}\sum_{j=1}^{p} \mathbb{E}\left[\|\boldsymbol{x}_{ij}\|^2 \mathbb{I}\left(\|\boldsymbol{x}_{ij}\|^2 \geq \varepsilon np\nu^2\right)\right] = 0, \tag{4.18}$$

where $\mathbb{I}(\cdot)$ is the indicator function, defined by $\mathbb{I}(A) = 1$ if A occurs and $\mathbb{I}(A) = 0$ otherwise.

This assumption is the multivariate Lindeberg condition, which is needed in constructing the central limit theorem when K is fixed. This is by no means a strong condition; for instance, when the factor model component is ignored, $\boldsymbol{\omega}_j$ is simply $\boldsymbol{\phi}_j$, and $\boldsymbol{x}_{ij} = \boldsymbol{\phi}_j b_i \epsilon_{ji}$. Since we assume $\boldsymbol{\phi}_j = O(1)$ in Assumption 1, the Lindeberg condition in (4.18) is met.

Next, we introduce a similar assumption to Assumption 8, used for the central limit theorem when K goes to infinity. Let $\widetilde{\boldsymbol{\omega}}_j$, $j = 1, \ldots, p$ denote the jth column of the $K \times p$ matrix $\left(\frac{1}{p}\Phi^\top \boldsymbol{M}_{\boldsymbol{A}^0}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0}$. Denote

$$\begin{aligned} &\boldsymbol{\gamma}^\top \left(\tfrac{1}{p}\Phi^\top \boldsymbol{M}_{\boldsymbol{A}^0}\boldsymbol{\Phi}\right)^{-1} \tfrac{1}{\sqrt{np}}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0}\boldsymbol{E}\boldsymbol{M}_{\boldsymbol{F}}\mathbf{b} \\ = \;& \tfrac{1}{\sqrt{np}}\textstyle\sum_i^n \sum_j^p \boldsymbol{\gamma}^\top \widetilde{\boldsymbol{\omega}}_j \epsilon_{ji} \sum_k^n \psi_{ik} b_k \equiv \tfrac{1}{\sqrt{np}}\sum_i^n \sum_j^p \widetilde{x}_{ij}. \end{aligned} \tag{4.19}$$

Assumption 9. Assume that there exists a constant $\widetilde{L} > 0$ such that

$$\widetilde{L} \equiv \lim_{n,p,K\to\infty} \widetilde{L}_{np}, \tag{4.20}$$

where $\widetilde{L}_{np}$ is defined as

$$\widetilde{L}_{np} \equiv \frac{\sigma^2}{np} \sum_{i}^{n} \sum_{j}^{p} \left(\boldsymbol{\gamma}^\top \widetilde{\boldsymbol{\omega}}_j\right)^2 \left(\sum_{k}^{n} \psi_{ik} b_k\right)^2. \tag{4.21}$$

Moreover, for any small value $\varepsilon > 0$, we assume

$$\lim_{n,p,K\to\infty} \frac{1}{np\nu^2} \sum_{i=1}^{n} \sum_{j=1}^{p} \mathbb{E}\left[|\widetilde{x}_{ij}|^2 \mathbb{I}\left(|\widetilde{x}_{ij}|^2 \geq \varepsilon np\nu^2\right)\right] = 0, \tag{4.22}$$

where $\mathbb{I}(\cdot)$ is the indicator function, defined by $\mathbb{I}(A) = 1$ if A occurs and $\mathbb{I}(A) = 0$ otherwise.

4.2 Asymptotic properties

As we have mentioned previously, the identification problem of the latent factor implies that we use the estimator $\widehat{\boldsymbol{A}}$ to estimate a rotation of $\boldsymbol{A}^0$. Based on the objective function (3.7) in Section 3, we use a center-adjusted objective function, defined as below.

$$S_{np}(\boldsymbol{c}_i, \boldsymbol{A}) = \frac{1}{np} \sum_{i=1}^{n} \left[(\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i)^\top \boldsymbol{M}_{\boldsymbol{A}} (\boldsymbol{Y}_i - \boldsymbol{\Phi}\boldsymbol{c}_i) + \alpha \boldsymbol{c}_i^\top \boldsymbol{R} \boldsymbol{c}_i\right] - \frac{1}{np} \sum_{i=1}^{n} \epsilon_i^\top \boldsymbol{M}_{\boldsymbol{A}^0} \epsilon_i, \tag{4.23}$$

where $\boldsymbol{M}_{\boldsymbol{A}}$ is defined in (3.6), satisfying $\boldsymbol{A}^\top \boldsymbol{A}/p = \boldsymbol{I}_r$. The second term on the right-hand side of (4.23) does not contain the unknown $\boldsymbol{A}$ and $\boldsymbol{c}_i$, so the inclusion of this term does not affect the optimization result. This term is only used for center adjusting, so that the resulting objective

function has an expectation zero. We estimate $\boldsymbol{c}_i^0$ and $\boldsymbol{A}^0$ by

$$(\widehat{\boldsymbol{c}}_i, \widehat{\boldsymbol{A}}) = \underset{\boldsymbol{c}_i, \boldsymbol{A}}{\operatorname{argmin}} S_{np}(\boldsymbol{c}_i, \boldsymbol{A}). \tag{4.24}$$

In the following, we establish the asymptotic properties for the estimated coefficient matrix $\widehat{\boldsymbol{C}}$. In Theorem 1, the consistency of the matrix $\widehat{\boldsymbol{C}}$ is proved. In Theorem 2, we show the rate of convergence of $\widehat{\boldsymbol{C}}$. Theorem 3 provides the asymptotic distribution of $\widehat{\boldsymbol{C}}$.

Let $\boldsymbol{P}_{\boldsymbol{U}} = \boldsymbol{U}(\boldsymbol{U}^\top \boldsymbol{U})^{-1}\boldsymbol{U}^\top$ for a matrix $\boldsymbol{U}$.

Theorem 1. *Under Assumptions 1 - 6, as $n, p \to \infty$, we have the following statements*

(i) $\frac{1}{\sqrt{n}} \left\| \boldsymbol{C} - \widehat{\boldsymbol{C}} \right\| \xrightarrow{p} 0.$

(ii) $\left\| \boldsymbol{P}_{\widehat{\boldsymbol{A}}} - \boldsymbol{P}_{\boldsymbol{A}^0} \right\| \xrightarrow{p} 0.$

We start by proving consistency for the vector $\widehat{\boldsymbol{c}}_i$. This consistency is uniform for all $i = 1, \ldots, n$. Therefore, we could combine $\boldsymbol{c}_i$ for all $i = 1, \ldots, n$, and have the result for the coefficient matrix $\widehat{\boldsymbol{C}}$ in (i). The matrix $\widehat{\boldsymbol{C}}$ is of dimension $K \times n$, where K is fixed and the sample size n goes to infinity, so there is a $(\sqrt{n})^{-1}$ scale in the result of (i). In the second part of the theorem, note that $\boldsymbol{P}_{\boldsymbol{A}} = \boldsymbol{I}_p - \boldsymbol{M}_{\boldsymbol{A}}$, where $\boldsymbol{M}_{\boldsymbol{A}}$ is the projection matrix onto the orthogonal complement of the linear space spanned by the columns of $\boldsymbol{A}$. Thus, $\boldsymbol{P}_{\widehat{\boldsymbol{A}}}$ and $\boldsymbol{P}_{\boldsymbol{A}^0}$ represent the spaces spanned by $\widehat{\boldsymbol{A}}$ and $\boldsymbol{A}^0$, and we show that they are asymptotically the same in (ii).

Next, we obtain the rate of convergence.

Theorem 2. *Under Assumptions 1 - 6, if* $p/n \to \rho > 0$ *as* $n, p \to \infty$,

$$\left\| \frac{\left(\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}\right)}{\sqrt{n}} \boldsymbol{M}_{\boldsymbol{F}} \right\| = O_p\left(\frac{1}{\sqrt{p}}\right),$$

where $\boldsymbol{M}_{\boldsymbol{F}}$ *is defined in* (4.14).

We study the case when the dimension p and the sample size n are comparable. We achieve rate $\sqrt{p}$ convergence, considering $(\sqrt{n})^{-1} \left\| \boldsymbol{C}^0 - \widehat{\boldsymbol{C}} \right\|$ on the whole. It is expected that the rate of convergence for smoothing models depends on the number of discrete points p observed on each curve.

Remark 7. The asymptotic result in Theorem 2 contains a projection matrix $\boldsymbol{M}_{\boldsymbol{F}}$. This matrix projects $\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}$ onto the space orthogonal to the factor matrix $\boldsymbol{F}$. This theorem shows the interplay between $\boldsymbol{C}^0$ and $\boldsymbol{F}$. When $\boldsymbol{C}^0$ and $\boldsymbol{F}$ are orthogonal, $(\boldsymbol{C}^0 - \widehat{\boldsymbol{C}})\boldsymbol{M}_{\boldsymbol{F}} = \boldsymbol{C}^0 - \widehat{\boldsymbol{C}}$, and we obtain the rate of convergence of $\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}$. When $\boldsymbol{C}^0$ and $\boldsymbol{F}$ are not orthogonal, the inference on $\boldsymbol{C}^0$ will be affected by the existence of the factor model component.

We further begin to establish the limiting distribution. It is shown in S5.1 Appendix A that

$$\left\| \sqrt{p} \frac{\left(\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}\right)}{\sqrt{n}} \boldsymbol{M}_{\boldsymbol{F}} \right\| = \left\| \left(\frac{1}{p} \Phi^\top \boldsymbol{M}_{\boldsymbol{A}^0} \boldsymbol{\Phi}\right)^{-1} \frac{1}{\sqrt{np}} \boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0} \boldsymbol{E} \boldsymbol{M}_{\boldsymbol{F}} \right\| + o_p(1).$$

The limiting distribution is constructed based on the first term on the right-hand side.

Theorem 3. *In addition to Assumptions 1 - 8, we assume that* K *is fixed and, as* $n, p \to \infty$,

there exists a positive definite matrix $\widetilde{\boldsymbol{Q}}$ *such that*

$$\boldsymbol{Q}\left(\boldsymbol{A}^0\right) \equiv \frac{1}{p}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\boldsymbol{A}^0}\boldsymbol{\Phi} \xrightarrow{p} \widetilde{\boldsymbol{Q}}.$$

Then we have, for any vector $\boldsymbol{b} \in \mathbb{R}^n$

$$\sqrt{p}\left(\frac{\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}}{\sqrt{n}}\right)\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{b} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{L}\widetilde{\boldsymbol{Q}}^{-1}\right),$$

where $\boldsymbol{M}_{\boldsymbol{F}}$ *is defined in Theorem 2,* $\boldsymbol{L}$ *is defined in* (4.17).

Moreover if the limit of $\boldsymbol{\Phi}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{L}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Phi}^\top$ *exists, we have*

$$\sqrt{p}\boldsymbol{\Phi}\left(\frac{\boldsymbol{C}^0 - \widehat{\boldsymbol{C}}}{\sqrt{n}}\right)\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{b} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \lim_{n,p\to\infty}\boldsymbol{\Phi}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{L}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Phi}^\top\right),$$

$$\sqrt{\frac{p}{n}}\left(\boldsymbol{X} - \widehat{\boldsymbol{X}}\right)\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{b} \xrightarrow{d} \mathcal{N}\left(\boldsymbol{0}, \lim_{n,p\to\infty}\boldsymbol{\Phi}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{L}\widetilde{\boldsymbol{Q}}^{-1}\boldsymbol{\Phi}^\top\right).$$

The vector $\boldsymbol{b}$ comes from the same vector in Lemma 1. The asymptotic bias is zero since we assume no serial or cross-sectional correlation in the error terms. This simplified setting can be extended to allow for weak correlations in errors in both directions. In that case, the asymptotic distribution will include a non-zero bias term.

Remark 8. Theorem 3 shows that the asymptotic distribution of the coefficient matrix $\widehat{\boldsymbol{C}}$ relies on the unobserved factor loading matrix $\boldsymbol{A}^0$. Although we are unable to consistently estimate $\boldsymbol{A}^0$ with $\widehat{\boldsymbol{A}}$, what we need is in fact the projected matrix $\boldsymbol{M}_{\boldsymbol{A}^0}$, which can be estimated by $\boldsymbol{M}_{\widehat{\boldsymbol{A}}}$.

We are able to find estimators for $\boldsymbol{Q}$ and $\boldsymbol{L}$ based on $\boldsymbol{M}_{\widehat{\boldsymbol{A}}}$,

$$\widehat{\boldsymbol{Q}} = \frac{1}{p}\boldsymbol{\Phi}^\top \boldsymbol{M}_{\widehat{\boldsymbol{A}}}\boldsymbol{\Phi}$$
$$\widehat{\boldsymbol{L}} = \frac{\sigma^2}{np}\sum_i^n\sum_j^p \widehat{\boldsymbol{\omega}}_j^\top\widehat{\boldsymbol{\omega}}_j\left(\sum_k^n \widehat{\psi}_{ik}b_k\right)^2,$$

where $\widehat{\boldsymbol{\omega}}_j$ is the jth column of the $K\times p$ matrix $\boldsymbol{\Phi}^\top\boldsymbol{M}_{\widehat{\boldsymbol{A}}}$ and $\widehat{\psi}_{ik}$ is the (i,k)th element in the matrix $\boldsymbol{M}_{\widehat{\boldsymbol{F}}}$.

When K goes to infinity, the random vector $\left(\frac{\boldsymbol{C}^0-\widehat{\boldsymbol{C}}}{\sqrt{n}}\right)\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{b}$ considered in Theorem 3 is a high-dimensional random vector. Then we establish the asymptotic distribution for some linear combination of this random vector in the following theorem.

Theorem 4. *Suppose Assumptions 1 - 7 and Assumption 9 are satisfied. As n, p, K tend to infinity, we have for any vectors $\boldsymbol{\gamma}\in\mathbb{R}^K$ and $\boldsymbol{b}\in\mathbb{R}^n$,*

$$\sqrt{p}\boldsymbol{\gamma}^\top\left(\frac{\boldsymbol{C}^0-\widehat{\boldsymbol{C}}}{\sqrt{n}}\right)\boldsymbol{M}_{\boldsymbol{F}}\boldsymbol{b}\xrightarrow{d}\mathcal{N}\left(\boldsymbol{0},\widetilde{L}\right), \tag{4.25}$$

where $\widetilde{L}$ is defined in (4.20).

5. Simulation Studies

In this section, we use simulated data to illustrate the superiority of the proposed model. We compare FASM with three other modes. The first one is the basis smoothing model with a penalty (Bsmooth). We use the same basis functions as FASM and the L2 penalty. So the

Bsmooth model is only different from FASM in that it does not consider the augmented factor component. The second one is the local linear smoothing model (Localin). Cross-validation is used to choose the bandwidth parameter. The third one is the principal component analysis with conditional expectation (PACE). This method is proposed in Yao et al. (2005) and is mainly used for sparse data, where very few data points are observed on each functional curve.

5.1 Data generating process

Setting 1

We generate simulated data Y_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$ from the following model:

$$\begin{aligned} Y_{ij} &= \mathcal{X}_i(u_j) + \eta_{ij} + \epsilon_{ji} \\ &= \sum_{k=1}^{13} c_{ik}\phi_k(u_j) + \sum_{k=1}^{4} f_{ik}a_{kj} + \epsilon_{ji}, \end{aligned}$$

where $\phi_k(u)$ are chosen as B-spline basis functions of order 4 and the smoothing coefficients c_{ik} are generated from $\mathcal{N}(0, 1.5^2)$. The factors loadings a_{kj} follow $\mathcal{N}(0, 0.6^2)$ and the factors $(f_{i1}, f_{i2}, f_{i3}, f_{i4})^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a 4 by 4 covariance matrix. We set the multivariate mean term $\boldsymbol{\mu} = \mathbf{0}$ and variance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_4$. We adjust the value of σ^2 to control the signal-to-noise ratio. When σ^2 is large, the signal-to-noise level is low, and when σ^2 is small, the signal-to-noise level is high. The random error terms ϵ_{ji} follow $\mathcal{N}(0, 0.5^2)$.

Setting 2

We generate simulated data Y_{ij}, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$ from the following model:

$$\begin{aligned} Y_{ij} &= \mathcal{X}_i(u_j) + \eta_{ij} + \epsilon_{ji} \\ &= \sum_{k=1}^{5} c_{ik}\phi_k(u_j) + \sum_{k=1}^{4} f_{ik}a_{kj} + \epsilon_{ji}, \end{aligned}$$

where $\phi_k(u)$ are chosen as B-spline basis functions of order 4 and the smoothing coefficients c_{ik} are generated from $\mathcal{N}(0, \gamma_k^2)$, where $\gamma_1 = 3, \gamma_2 = 2.5, \gamma_3 = 2, \gamma_4 = 1.5, \gamma_5 = 1$. The factors loadings a_{kj} follow $\mathcal{N}(0, 0.8^2)$ and the factors $(f_{i1}, f_{i2}, f_{i3}, f_{i4})^\top \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma}$ is a 4 by 4 covariance matrix. We set the multivariate mean term $\boldsymbol{\mu} = \mathbf{0}$ and variance $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}_4$. We adjust the value of σ^2 to control the signal-to-noise ratio. When σ^2 is large, the signal-to-noise level is low, and when σ^2 is small, the signal-to-noise level is high. The random error terms ϵ_{ji} follow $\mathcal{N}(0, 0.5^2)$.

Setting 3

We generate simulated data as in Setting 1, except we use $K = 21$ B-spline basis functions. We use $n = 40$ and $p = 50$. We show how the model works when the number of basis functions is large compared to n and p.

5.2 Results

We repeat the simulation setup 500 times and obtain the estimated smooth function $\widehat{\mathcal{X}}_i(u) = \widehat{\boldsymbol{c}}_i^{\top}\boldsymbol{\Phi}(u)$. The mean integrated squared error (MISE) for function estimation is calculated as

$$\text{MISE} = \frac{1}{n}\sum_{i=1}^{n}\int\left[\mathcal{X}_i(u) - \widehat{\mathcal{X}}_i(u)\right]^2 du. \tag{5.26}$$

The results for Setting 1 and 2 are reported in Table 1 and 2. With the same sample size n, increasing the number of points p on the curve decreases the estimation error. However, with the same value for p, increasing the sample size does not decrease the estimation error. This is consistent with the convergence rate stated in Section 4, where the estimator converges with the rate related to p.

When σ is large, i.e., the signal-to-noise ratio is high, the FASM performs better than the smoothing model. When $\sigma = 0$, the actual data are generated without the factor model component. The performance of the two methods is identical. This implies that the proposed model is robust in that it improves the estimation when there is measurement error, and it is no worse than the ordinary smoothing model when there is no measurement error. This should be credited to the selection criterion used in (S3.4). It always selects zero factors when there are no common factors.

In Table 3, we show the MISE result from Setting 3. Our proposed model also shows its supiority when K is large compared to n and p. This setting corresponds to allowing the number of basis functions K to go to infinity in Section 4.

Table 1: **Setting 1**: The MISE of the function estimates with different samples sizes and dimensions. The adjustment of σ^2 value is used to control the signal-to-noise ratio.

			MISE			
Sample Size	Dimension	Size of η	FASM	Bsmooth	Localin	PACE
$n = 20$	$p = 51$	$\sigma = 0$	**0.058**	**0.058**	0.075	0.238
		$\sigma = 0.5$	**0.131**	**0.131**	0.149	0.587
		$\sigma = 0.75$	**0.204**	0.208	0.226	1.026
		$\sigma = 1$	**0.297**	0.318	0.336	1.676
$n = 20$	$p = 101$	$\sigma = 0$	**0.031**	**0.031**	0.043	0.238
		$\sigma = 0.5$	**0.076**	**0.076**	0.089	0.595
		$\sigma = 0.75$	**0.115**	0.123	0.136	1.020
		$\sigma = 1$	**0.159**	0.187	0.201	1.686
$n = 50$	$p = 51$	$\sigma = 0$	**0.058**	**0.058**	0.076	0.195
		$\sigma = 0.5$	**0.131**	0.135	0.153	0.543
		$\sigma = 0.75$	**0.186**	0.214	0.233	1.013
		$\sigma = 1$	**0.248**	0.314	0.331	1.668
$n = 100$	$p = 101$	$\sigma = 0$	**0.031**	**0.031**	0.044	0.170
		$\sigma = 0.5$	**0.062**	0.074	0.089	0.463
		$\sigma = 0.75$	**0.090**	0.118	0.132	0.908
		$\sigma = 1$	**0.133**	0.181	0.194	1.570

6. Application to Climatology

We apply the aforementioned FASM, Bsmooth, Localin and PACE methods to some real data sets. This section looks at the Friday temperature data at Adelaide airport. We also provide the analysis of Canadian yearly temperature and precipitation data in the Supplement.

We choose Adelaide because it tends to have the hottest temperature among Australia's big cities. Data from other weekdays exhibit similar features and are not shown here. The data are measured every half an hour from 1997 to 2007. The sample size n is 508, and the number of discrete data points p from each curve is 48. The plot of the raw data can be found in Figure 4a.

Table 2: **Setting 2**: The MISE of the function estimates with different samples sizes and dimensions. The adjustment of σ^2 value is used to control the signal-to-noise ratio.

			MISE			
Sample Size	Dimension	Size of η	FASM	Bsmooth	Localin	PACE
$n = 20$	$p = 51$	$\sigma = 0$	**0.024**	**0.024**	0.070	0.116
		$\sigma = 0.5$	0.086	**0.085**	0.195	0.621
		$\sigma = 0.75$	**0.160**	**0.160**	0.340	1.431
		$\sigma = 1$	**0.272**	0.280	0.538	2.685
$n = 20$	$p = 101$	$\sigma = 0$	**0.012**	**0.012**	0.040	0.089
		$\sigma = 0.5$	0.047	**0.046**	0.117	0.660
		$\sigma = 0.75$	**0.079**	**0.079**	0.192	1.434
		$\sigma = 1$	**0.135**	0.138	0.301	2.549
$n = 50$	$p = 51$	$\sigma = 0$	**0.025**	**0.025**	0.070	0.076
		$\sigma = 0.5$	0.087	**0.086**	0.202	0.584
		$\sigma = 0.75$	**0.162**	0.169	0.348	1.435
		$\sigma = 1$	**0.250**	0.269	0.522	2.557
$n = 100$	$p = 101$	$\sigma = 0$	**0.012**	**0.012**	0.041	0.033
		$\sigma = 0.5$	**0.045**	**0.045**	0.114	0.518
		$\sigma = 0.75$	**0.085**	**0.085**	0.195	1.337
		$\sigma = 1$	**0.138**	0.139	0.299	2.618

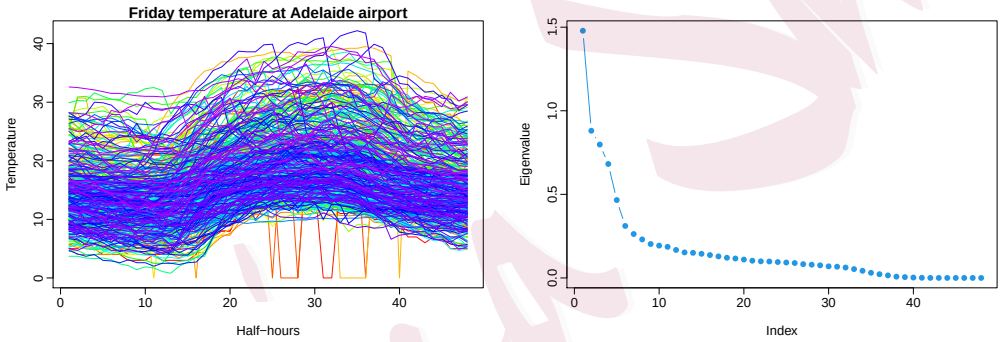
Table 3: **Setting 3**: The MISE of the function estimates when K is large compared to n and p. The adjustment of σ^2 value is used to control the signal-to-noise ratio.

			MISE			
Sample Size	Dimension	Size of η	FASM	Bsmooth	Localin	PACE
$n = 50$	$p = 40$	$\sigma = 0$	**0.097**	**0.097**	0.121	0.225
		$\sigma = 0.5$	**0.206**	0.207	0.235	0.590
		$\sigma = 0.75$	**0.291**	0.309	0.335	1.044
		$\sigma = 1$	**0.390**	0.457	0.466	1.747

It can be seen that the data are quite noisy, with extreme values in some of the curves due to large measurement errors.

We use the B-spline basis functions of order 4 with knots at every data point. We show the residuals from the four models in Figure 5. Apart from FASM, all the other models produce large residuals with some extreme values. To further look at the residual struture from the Bsmooth

model, we conduct a principal component analysis on the residuals. In Figure 4b, the screeplot of the eigenvalues are presented. The residuals from Bsmooth model exhibit a "spike" structure, where the first few eigenvalues are significantly larger than the rest. This means the residuals contain information that can be captured by just a few factors, which calls for a further dimension reduction model on the residuals.

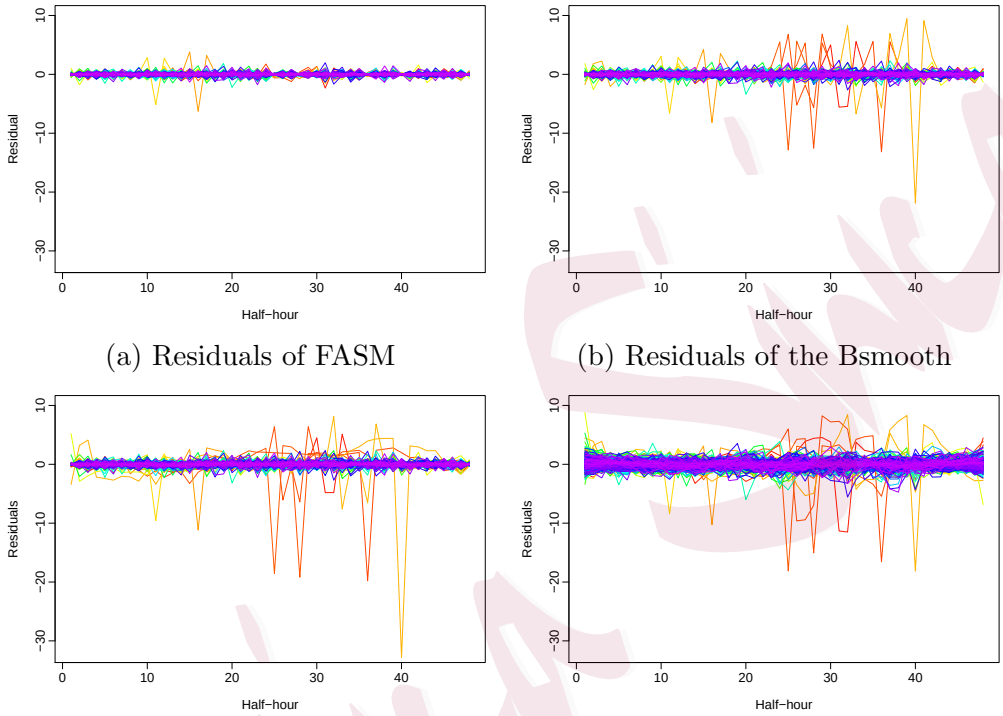


(a) Raw temperature data (b) Eigenvalues of the residuals in from Bsmooth

Figure 4: Half-hourly temperature data on Friday at Adelaide airport.

7. Conclusion and Future Work

In this paper, we propose a factor-augmented smoothing model for functional data. We study raw functional data, a mixture of functional curves and high-dimensional errors. When measurement error is informative, a smoothing model alone is inadequate to capture data variation and recover the signal functional component. The proposed model incorporates a factor structure into the smoothing model to further explain the large residuals. We propose a numerical iteration approach to obtain estimates in the smoothing and factor models simultaneously. The asymptotic distribution of the estimators is given with proof. Our model also serves as a dimension reduction



(a) Residuals of FASM (b) Residuals of the Bsmooth

(c) Residuals of Localin (d) Residuals of PACE

Figure 5: Half-hourly temperature data on Friday at Adelaide airport.

method on functional and high-dimensional mixture data, easing the path to making inferences. We provide an example of constructing a covariance estimator for the raw data. Further, we show that the model can be applied in situations where there is misidentification in the data structure, two examples of which are the misspecification of smoothing basis functions and the neglect of the step jumps in the mean level of the functions. The advantages of the proposed model are demonstrated in extensive simulation studies. We also show how our model performs via applications to Australian temperature data.

The proposed model is a good starting point for modeling complex data structures. We deal with a mixture of smooth functional curves and high-dimensional measurement error. The factor

model component can be regarded as a "boosting component that improves model accuracy. Extending from this idea, the model can be applied to other data structures. One example is the data that contain change points. The change point is a popular problem in many statistics and econometric topics and has been extensively studied in the multivariate setting. Previous literature on change points in functional data include Berkes et al. (2009); Hörmann and Kokoszka (2010) and Hörmann et al. (2015). It is shown in simulation examples that our model can be used for modeling functional data with change points in the cross-sectional direction. The model can be modified to account for change points also in the sample direction. Further research can be conducted along this line.

Supplementary Materials

The Supplement includes our extension of the model to non-parametric settings, introducing a covariance estimator and some extensive simulated and real data analysis. It also contains the proofs for the theorems in the main article. Due to the page limit, many important results are placed in the Supplement of this paper.

References

Ahn, S. C. and A. R. Horenstein (2013). Eigenvalue ratio test for the number of factors. *Econometrica 81*(3), 1203–1227.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models. *Econometrica 70*(1), 191–221.

REFERENCES

Berkes, I., R. Gabrys, L. Horváth, and P. Kokoszka (2009). Detecting changes in the mean of functional observations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 71*(5), 927–946.

Cai, T. T. and M. Yuan (2011). Optimal estimation of the mean function based on discretely sampled functional data: Phase transition. *The Annals of Statistics 39*(5), 2330–2355.

Cuevas, A. (2014). A partial overview of the theory of statistics with functional data. *Journal of Statistical Planning and Inference 147*, 1–23.

Eubank, R. L. (1999). *Nonparametric Regression and Spline Smoothing* (2nd ed.). New York: Marcel Dekker.

Fan, J., Y. Fan, and J. Lv (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics 147*(1), 186–197.

Fan, J. and I. Gijbels (1996). *Local Polynomial Modelling and Its Applications*. London: Chapman & Hall.

Febrero-Bande, M., P. Galeano, and W. González-Manteiga (2017). Functional principal component regression and functional partial least-squares regression: An overview and a comparative study. *International Statistical Review 85*(1), 61–83.

Ferraty, F. and P. Vieu (2006). *Nonparametric Functional Data Analysis: Theory and Practice*. New York: Springer Science & Business Media.

Goia, A. and P. Vieu (2016). An introduction to recent advances in high/infinite dimensional statistics. *Journal of Multivariate Analysis 146*, 1–6.

Golub, G. H., M. Heath, and G. Wahba (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics 21*(2), 215–223.

Green, P. J. and B. W. Silverman (1999). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. London: Chapman & Hall.

REFERENCES

Hörmann, S., L. Kidziński, and M. Hallin (2015). Dynamic functional principal components. *Journal of the Royal Statistical Society, Statistical Methodology, Series B 77*(2), 319–348.

Hörmann, S. and P. Kokoszka (2010). Weakly dependent functional data. *The Annals of Statistics 38*(3), 1845–1884.

Horváth, L. and P. Kokoszka (2012). *Inference for functional data with applications*, Volume 200. New York: Springer Science & Business Media.

Jiang, B., Y. Yang, J. Gao, and C. Hsiao (2021). Recursive estimation in large panel data models: Theory and practice. *Journal of Econometrics 224*(2), 439–465.

Knight, K. and W. Fu (2000). Asymptotics for lasso-type estimators. *The Annals of Statistics 28*(5), 1356–1378.

Lam, C., Q. Yao, and N. Bathia (2011). Estimation of latent factors for high-dimensional time series. *Biometrika 98*(4), 901–918.

Onatski, A. (2010). Determining the number of factors from empirical distribution of eigenvalues. *The Review of Economics and Statistics 92*(4), 1004–1016.

Onatski, A. (2012). Asymptotics of the principal components estimator of large factor models with weakly influential factors. *Journal of Econometrics 168*(2), 244–258.

Ramsay, J. O. and G. Hooker (2017). *Dynamic Data Analysis: Modeling Data with Differential Equations*. New York: Springer.

Ramsay, J. O. and B. W. Silverman (2002). *Applied Functional Data Analysis*. New York: Springer.

Ramsay, J. O. and B. W. Silverman (2005). *Functional Data Analysis*. New York: Springer.

Reiss, P. T., J. Goldsmith, H. L. Shang, and R. T. Ogden (2017). Methods for scalar-on-function regression. *International Statistical Review 85*(2), 228–249.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics 6*(2), 461–464.

REFERENCES

Wahba, G. (1990). *Spline models for observational data*, Volume 59. SIAM.

Wahba, G. and S. Word (1975). Periodic splines for spectral density estimation: The use of cross validation for determining the degree of smoothing. *Communications in Statistics:Theory and methods 4*(2), 125–141.

Wand, M. P. and C. M. Jones (1995). *Kernel Smoothing*. Boca Raton, FL: Chapman & Hall.

Wang, J.-L., J.-M. Chiou, and H.-G. Müller (2016). Functional data analysis. *Annual Review of Statistics and Its Application 3*, 257–295.

Yao, F., H. Müller, and J. Wang (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association 100*(470), 577–590.

Yao, W. and R. Li (2013). New local estimation procedure for a non-parametric regression function for longitudinal data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 75*(1), 123–138.

Zhang, X. and J. L. Wang (2016). From sparse to dense functional data and beyond. *The Annals of Statistics 44*(5), 2281–2321.

Yuan Gao

E-mail: (yuan.gao@anu.edu.au)

Phone: (+61 2 612 57290)

Han Lin Shang

E-mail: (hanlin.shang@mq.edu.au)

Yanrong Yang

E-mail: (yanrong.yang@anu.edu.au)