Supplementary Document for the Manuscript entitled

"Estimation of Sparse Functional Additive Models

with Adaptive Group LASSO"

Peijun Sang, Liangliang Wang and Jiguo Cao

Simon Fraser University

This document consists of procedures of estimating FPC scores introduced in Section 2.1, theoretical proofs of main results shown in Section 3, additional simulation results in Section 4 and one application example in Section 5.

S1. Estimation of FPC scores

When the functional predictor $X_i(t)$ has sparse observations, we employ the principal component analysis by conditional expectation (PACE) algorithm proposed by Yao *et al.* (2005) to obtain estimated FPC scores denoted by $\hat{\xi}_{ij}, \ j = 1, \ldots, d, \ i = 1, \ldots, n$. PACE first estimates the mean function of $X(t)$ via local linear regression and the corresponding estimator is denoted by $\hat{m}(t)$. Then the raw estimate of covariance evaluated at (t_{ij}, t_{il}),

$G_i(t_{ij}, t_{ij}) = (W_{ij} - \hat{m}(t_{ij}))(W_{il} - \hat{m}(t_{il}))$ is pooled. To account for the fact

that $\text{Cov}(W_{ij}, W_{il}) = \text{Cov}(X_i(t_{ij}), X_i(t_{il})) + \delta(t_{ij} = t_{il})$ where $\delta(t = s) = 1$

if $t = s$ and 0 otherwise, a local linear regression is employed to smooth the

non-diagonal estimates, $\{G_i(t_{ij}, t_{ij}), t_{ij} \neq t_{il}, i = 1, \ldots, n\}$. Let $\hat{G}(s,t)$ de-

note the smoothed covariance function and let $\hat{\lambda}_j$'s and $\hat{\phi}_j(t)$'s denote the

corresponding eigenvalues and eigenfunctions, respectively. They satisfy

$\int_{\mathcal{I}} \hat{G}(s,t)\hat{\phi}_j(t)dt = \hat{\lambda}_j \hat{\phi}_j(s)$. The estimated FPC scores, $\hat{\xi}_{ij}$'s, are estimated

via conditional expectation.

When the functional predictor $X_i(t)$ has dense observations, we smooth

each trajectory $X_i(t)$ via the local linear regression rather than smooth-

ing the mean function, and denote the smoothed trajectory by $\hat{X}_i(t)$, $i =$

$1, \ldots, n$. The estimated mean function and covariance function are then

given by $\hat{m}(t) = \frac{1}{n}\sum_{i=1}^{n}\hat{X}_i(t)$ and $\hat{G}(s,t) = \frac{1}{n}\sum_{i=1}^{n}(\hat{X}_i(s) - \hat{m}(s))(\hat{X}_i(t) -$

$\hat{m}(t))$, respectively. The estimated eigenvalues $\hat{\lambda}_j$ and eigenfunctions $\hat{\phi}_j(t)$

of $\hat{G}$ satisfy $\int_{\mathcal{I}} \hat{G}(s,t)\hat{\phi}_j(t)dt = \hat{\lambda}_j \hat{\phi}_j(s)$ as well. Unlike PACE, the FPC

scores in the scenario are calculated as $\hat{\xi}_{ij} = \int_{\mathcal{I}}(\hat{X}_i(s) - \hat{m}(s))\hat{\phi}_j(s)ds$,

$j = 1, \ldots, d$, $i = 1, \ldots, n$.

S2. Proofs

We follow the main ideas of Huang *et al.* (2010) to prove Theorems 1-3, but great effort needs to be taken to tackle the difficulties caused by using the estimated scaled FPC scores (rather than the true scaled FPC scores), to estimate the unknown parameters and functions. Before proving Proposition 1, we first present the following result which establishes consistency of the estimated scaled FPC scores.

Lemma 1. Suppose that Assumptions A and B1 hold. We have

$$\frac{1}{n}\sum_{i=1}^{n}\left(\sum_{j=1}^{d}|\hat{\zeta}_{ij}-\zeta_{ij}|\right)^{2}=O_{p}(n^{-1}).$$

Proof: See Lemma 2 in Zhu *et al.* (2014). □

B.1. Proof of Proposition 1:

For any function f on $[0, 1]$, let $||f||_{\infty}=\sup_{x\in[0,1]}|f(x)|$. For $f\in\mathscr{F}$ and $\mathrm{E}\left(f(\zeta_{j})\right)=0$, there exists $f_{j}'\in\mathscr{S}_{n}$, such that $||f-f_{j}'||_{\infty}=O(m_{n}^{-\rho})$; see De Boor (1968) or Lemma 5 of Stone (1985). Let $f_{nj}=f_{j}'-n^{-1}\sum_{i=1}^{n}f_{j}'(\hat{\zeta}_{ij})$. Then $f_{nj}\in\mathscr{S}_{nj}^{0}$ and we next prove that f_{nj} satisfies $\frac{1}{n}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij})-$

$f(\hat{\zeta}_{ij}))^2 = O_p(m_n^{-2\rho} + n^{-1})$. Since $(x+y)^2 \le 2(x^2 + y^2) \ \forall x, y \in \mathbb{R}$, we have

$$\frac{1}{n}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}))^2 = \frac{1}{n}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij}) - f'_j(\hat{\zeta}_{ij}) + f'_j(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}))^2$$

$$\le 2n^{-1}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij}) - f'_j(\hat{\zeta}_{ij}))^2 + (f'_j(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}))^2$$

$$= 2\left(\frac{1}{n}\sum_{i=1}^{n}f'_j(\hat{\zeta}_{ij})\right)^2 + O(m_n^{-2\rho})$$

$$= 2\left(\frac{1}{n}\sum_{i=1}^{n}(f'_j(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}) + f(\hat{\zeta}_{ij})\right.$$

$$\left. - f(\zeta_{ij}) + f(\zeta_{ij}))\right)^2 + O(m_n^{-2\rho})$$

$$\le 6E_1 + 6E_2 + 6E_3 + O(m_n^{-2\rho}), \qquad\qquad \text{(S2.1)}$$

where $E_1 = \left(\frac{1}{n}\sum_{i=1}^{n}(f'_j(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij})\right)^2$, $E_2 = \left(\frac{1}{n}\sum_{i=1}^{n}(f(\hat{\zeta}_{ij}) - f(\zeta_{ij}))\right)^2$

and $E_3 = \left(\frac{1}{n}\sum_{i=1}^{n}f(\zeta_{ij})\right)^2$. Obviously, $E_1 = O(m_n^{-2\rho})$ and $E_3 = O_p(n^{-1})$

by Chebyshev's inequality. As for E_2, there are two different cases. (i) If

$r = 0$, then

$$E_2 \le \left(n^{-1}\sum_{i=1}^{n}|f(\hat{\zeta}_{ij}) - f(\zeta_{ij})|\right)^2$$

$$\le \left(Ln^{-1}\sum_{i=1}^{n}|\hat{\zeta}_{ij} - \zeta_{ij}|^\nu\right)^2$$

$$\le \left(Ln^{-1}(\sum_{i=1}^{n}|\hat{\zeta}_{ij} - \zeta_{ij}|^2)^{\nu/2}n^{1-\nu/2}\right)^2$$

$$= O_p(n^{-\nu}).$$

The second inequality holds due to $f \in \mathscr{F}$ and the third inequality is

4

obtained by Hölder's inequality. In this case $0.5 < \rho = \nu \leq 1$, we have $E_2 = O_p(n^{-\nu}) = O_p(n^{-\rho}) = O_p(m_n^{-2\rho})$. Therefore, $\frac{1}{n}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}))^2 = O_p(m_n^{-2\rho})$ according to (S2.1). (ii) In the second case, $r \geq 1$, so f is continuously differentiable over $[0, 1]$, which implies that the first derivative of f is bounded some positive constant M. By Lemma 1,

$$
\begin{aligned}
E_2 &\leq \left(n^{-1}\sum_{i=1}^{n}|f(\hat{\zeta}_{ij}) - f(\zeta_{ij})| \right)^2 \\
&\leq M^2 n^{-2} \left(\sum_{i=1}^{n}|\hat{\zeta}_{ij} - \zeta_{ij}| \right)^2 \\
&\leq M^2 n^{-1} \left(\sum_{i=1}^{n}|\hat{\zeta}_{ij} - \zeta_{ij}|^2 \right) \\
&= O_p(n^{-1}).
\end{aligned}
$$

It follows that $\frac{1}{n}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}))^2 = O_p(m_n^{-2\rho} + n^{-1})$ by (S2.1). Hence in both cases, $\frac{1}{n}\sum_{i=1}^{n}(f_{nj}(\hat{\zeta}_{ij}) - f(\hat{\zeta}_{ij}))^2 = O_p(m_n^{-2\rho} + n^{-1})$.

$\square$

B.2. Some useful lemmas:

To establish estimation and selection consistency of the proposed estimator, we need the following lemmas.

Lemma 2. Let $T_{jk} = n^{-\frac{1}{2}}m_n^{\frac{1}{2}}\sum_{i=1}^{n}\psi_k(\hat{\zeta}_{ij})\epsilon_i$, $j = 1,\ldots,d$, $k = 1,\ldots,m_n$ and $T_n = \max_{1\leq j\leq d, 1\leq k\leq m_n}|T_{jk}|$. Under Assumptions A, B and C2 in Section 3, then $T_n = O_p(\sqrt{\log m_n})$, if $m_n = o(n^{1/4})$.

5

Proof: Let $\hat{S}^2_{njk} = \sum_{i=1}^{n} \psi_k^2(\hat{\zeta}_{ij})$ and $S^2_{njk} = \sum_{i=1}^{n} \psi_k^2(\zeta_{ij})$. Conditional on $\hat{\zeta}_{ij}$'s, then T_{jk}'s are sub-gaussian. Let $\hat{S}^2_n = \max_{1 \leq j \leq d, 1 \leq k \leq m_n} \hat{S}^2_{njk}$ and $S^2_n = \max_{1 \leq j \leq d, 1 \leq k \leq m_n} S^2_{njk}$. Based on maximal inequalities on sub-gaussian random variables (Van Der Vaart and Wellner (1996)), we have

$$\mathrm{E}\left(T_n | \{\hat{\zeta}_{ij}, i = 1, \ldots, n, j = 1, \ldots, d\}\right) \leq C_1 n^{-\frac{1}{2}} m_n^{\frac{1}{2}} \sqrt{\log m_n} \hat{S}_n$$

for some constant $C_1 > 0$. It follows

$$T_n = O_p(n^{-\frac{1}{2}} m_n^{\frac{1}{2}} \sqrt{\log m_n} \hat{S}_n). \tag{S2.2}$$

Now we study the properties of $\hat{S}_n$, from which the order of T_n is immediately available. Since $\hat{S}^2_{njk} = (\hat{S}^2_{njk} - S^2_{njk}) + S^2_{njk}$, $\hat{S}^2_n \leq S^2_n + \max_{1 \leq j \leq d, 1 \leq k \leq m_n}(\hat{S}^2_{njk} - S^2_{njk})$. On the one hand,

$$\mathrm{E}\left(S^2_n\right) \leq \sqrt{2C_2 m_n^{-1} n \log m_n} + C_2 \log m_n + C_2 n m_n^{-1}$$

holds for a sufficiently large constant C_2 according to the proof of Lemma 2 in Huang *et al.* (2010). Since $m_n = O(n^\alpha)$ with $\alpha < 0.5$,

$$S^2_n = O_p(n m_n^{-1}). \tag{S2.3}$$

On the other hand, from (5) of Prochazkova (2005), we have $||\frac{dB_k(x)}{dx}||_\infty = O(m_n)$ when the order of the B-spline basis functions, l, satisfies $l \geq 2$ and

6

$||B_k||_\infty \leq 1$. Both of them hold uniformly for $1 \leq k \leq m_n$. It follows that

$$
\begin{aligned}
\hat{S}_{njk}^2 - S_{njk}^2 &= \sum_{i=1}^n \left(B_k(\hat{\zeta}_{ij}) - n^{-1}\sum_{i=1}^n B_k(\hat{\zeta}_{ij})\right)^2 - \left(B_k(\zeta_{ij}) - n^{-1}\sum_{i=1}^n B_k(\zeta_{ij})\right)^2 \\
&= \sum_{i=1}^n B_k^2(\hat{\zeta}_{ij}) - B_k^2(\zeta_{ij}) + n^{-1}\left(\sum_{i=1}^n B_k(\zeta_{ij})\right)^2 - n^{-1}\left(\sum_{i=1}^n B_k(\hat{\zeta}_{ij})\right)^2 \\
&= \sum_{i=1}^n (B_k(\hat{\zeta}_{ij}) + B_k(\zeta_{ij}))(B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})) \\
&\quad + n^{-1}\sum_{i=1}^n (B_k(\hat{\zeta}_{ij}) + B_k(\zeta_{ij}))\sum_{i=1}^n (B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})) \\
&\leq 4\sum_{i=1}^n |B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})| \\
&= O(m_n\sum_{i=1}^n |\hat{\zeta}_{ij} - \zeta_{ij}|) \\
&= O\left(m_n\sqrt{n}\sqrt{\sum_{i=1}^n |\hat{\zeta}_{ij} - \zeta_{ij}|^2}\right) \\
&= O_p(m_n\sqrt{n}) \qquad\qquad\qquad\qquad\qquad\qquad\qquad\text{(S2.4)}
\end{aligned}
$$

uniformly over $1 \leq j \leq d$ and $1 \leq k \leq m_n$. The first inequality is obtained using the Cauchy-Schwarz inequality and the last equation is a direct application of Lemma 1. Combining (S2.3) and (S2.4), we have $\hat{S}_n^2 = O_p(m_n\sqrt{n} + nm_n^{-1})$. $\hat{S}_n = O_p(n^{\frac{1}{2}}m_n^{-\frac{1}{2}})$ if $m_n = o(n^{1/4})$. As a result, $T_n = O_p(\sqrt{\log m_n})$. $\qquad\qquad\qquad\qquad\square$

Given a matrix Σ, let $\Lambda_{\min}(\Sigma)$ and $\Lambda_{\max}(\Sigma)$ denote the smallest and largest eigenvalues of Σ, respectively. Let $\tilde{\boldsymbol{Z}}_{ij} = (\psi_1(\zeta_{ij}), \ldots, \psi_{m_n}(\zeta_{ij}))^T$

7

and $\tilde{\boldsymbol{Z}}_j = (\tilde{\boldsymbol{Z}}_{1j}, \ldots, \tilde{\boldsymbol{Z}}_{nj})^T$. For a subset of $\{1, \ldots, d\}$, A, $\boldsymbol{Q}_A = \frac{\boldsymbol{z}_A^T \boldsymbol{z}_A}{n} \in$

$\mathbb{R}^{|A|m_n \times |A|m_n}$ and $\tilde{\boldsymbol{Q}}_A = \frac{\tilde{\boldsymbol{z}}_A^T \tilde{\boldsymbol{z}}_A}{n} \in \mathbb{R}^{|A|m_n \times |A|m_n}$, where $\boldsymbol{Z}_A$ (or $\tilde{\boldsymbol{Z}}_A$) rep-

resents the matrix by stacking $\boldsymbol{Z}_j$ (or $\tilde{\boldsymbol{Z}}_j$), $j \in A$ by column. For any

$\boldsymbol{x} = (x_1, \ldots, x_p)^T \in \mathbb{R}^p$, $||\boldsymbol{x}||_2 = \sqrt{\sum_{i=1}^{p} x_i^2}$.

Lemma 3. *Suppose $m_n = o(n^{1/6})$. Under Assumptions A, B and C3 in*

Section 3, then given a nonempty subset of $\{1, \ldots, d\}$, A, with probability

converging to 1,

$$C_3 m_n^{-1} \leq \Lambda_{\min}(\boldsymbol{Q}_A) \leq \Lambda_{\max}(\boldsymbol{Q}_A) \leq C_4 m_n^{-1}$$

for some positive constants C_3 and C_4.

Proof: We first lay out some facts about matrix theory, which are useful in

determining the magnitude of eigenvalues. For an $m \times n$ matrix $\boldsymbol{G}$,

$$||\boldsymbol{G}||_2 = \sup_{\boldsymbol{x} \in \mathbb{R}^n} \frac{||\boldsymbol{G}x||_2}{||x||_2} \quad \text{and} \quad ||\boldsymbol{G}||_1 = \max_{1 \leq j \leq n} \sum_{i=1}^{n} |g_{ij}|,$$

where g_{ij} is the (i,j)th entry of $\boldsymbol{G}$. For a symmetric matrix $\boldsymbol{G}$, $\Lambda_{\max}(\boldsymbol{G}) \leq$

$||\boldsymbol{G}||_2$. The following inequalities will be used later to prove the lemma.

Let $\boldsymbol{G}_1$ and $\boldsymbol{G}_2$ be two $n \times n$ symmetric matrices.

1. Weyl's inequality (Weyl (1912)):

$$\Lambda_{\min}(\boldsymbol{G}_1) - \Lambda_{\max}(\boldsymbol{G}_2 - \boldsymbol{G}_1) \leq \Lambda_{\min}(\boldsymbol{G}_2),$$

$$\Lambda_{\max}(\boldsymbol{G}_2) \leq \Lambda_{\max}(\boldsymbol{G}_1) + \Lambda_{\max}(\boldsymbol{G}_2 - \boldsymbol{G}_1).$$

8

Thus $|\Lambda_{\min}(\boldsymbol{G}_1) - \Lambda_{\min}(\boldsymbol{G}_2)| \leq ||\boldsymbol{G}_1 - \boldsymbol{G}_2||_2$ and $|\Lambda_{\max}(\boldsymbol{G}_1) - \Lambda_{\max}(\boldsymbol{G}_2)|$

$\leq ||\boldsymbol{G}_1 - \boldsymbol{G}_2||_2$.

2. The Gershgorin circle theorem (Gershgorin (1931)):

$$||\boldsymbol{G}_1 - \boldsymbol{G}_2||_2 \leq ||\boldsymbol{G}_1 - \boldsymbol{G}_2||_1.$$

We are ready to prove Lemma 3 now. Denote $\gamma_{k_1 k_2}^{(j_1, j_2)} = \frac{1}{n}\sum_{i=1}^{n}(\psi_{k_1}(\hat{\zeta}_{ij_1})$

$\psi_{k_2}(\hat{\zeta}_{ij_2}) - \psi_{k_1}(\zeta_{ij_1})\psi_{k_2}(\zeta_{ij_2}))$, $1 \leq k_1, \ k_2 \leq m_n, \ 1 \leq j_1, \ j_2 \leq d$. Based

on the definition of ψ_k's, it's obvious that uniformly over $1 \leq k \leq m_n$,

$||\frac{d\psi_k(x)}{dx}||_\infty = O(m_n)$ and $||\psi_k||_\infty \leq ||B_k||_\infty + |n^{-1}\sum_{i=1}^{n}B_k(\hat{\zeta}_{ij})| \leq 2$. Then

according to Lemma 1, we have

$$
\begin{aligned}
|\gamma_{k_1 k_2}^{(j_1, j_2)}| &= \frac{1}{n}\Big|\sum_{i=1}^{n}(\psi_{k_1}(\hat{\zeta}_{ij_1})\psi_{k_2}(\hat{\zeta}_{ij_2}) - \psi_{k_1}(\hat{\zeta}_{ij_1})\psi_{k_2}(\zeta_{ij_2}) \\
&\qquad\qquad + \psi_{k_1}(\hat{\zeta}_{ij_1})\psi_{k_2}(\zeta_{ij_2}) - \psi_{k_1}(\zeta_{ij_1})\psi_{k_2}(\zeta_{ij_2}))\Big| \\
&= O(m_n n^{-1})\sum_{i=1}^{n}(|\hat{\zeta}_{ij_1} - \zeta_{ij_1}| + |\hat{\zeta}_{ij_2} - \zeta_{ij_2}|) \\
&= O(m_n n^{-\frac{1}{2}})\sqrt{\sum_{i=1}^{n}(|\hat{\zeta}_{ij_1} - \zeta_{ij_1}| + |\hat{\zeta}_{ij_2} - \zeta_{ij_2}|)^2} \\
&= O_p(m_n n^{-\frac{1}{2}}).
\end{aligned}
$$

The third equation follows directly from the Cauchy-Schwarz inequality.

Note that $\gamma_{k_1 k_2}^{(j_1, j_2)} = O_p(m_n n^{-\frac{1}{2}})$ holds uniformly over $1 \leq j_1, \ j_2 \leq d$ and

$1 \leq k_1, \ k_2 \leq m_n$. For a nonempty subset of $\{1, \ldots, d\}$, A, without loss of

generality, we can assume that $A = \{1, \ldots, q\}$, where $q \leq d$. Then we have

$$|\Lambda_{\min}(\boldsymbol{Q}_A) - \Lambda_{\min}(\tilde{\boldsymbol{Q}}_A)| \leq ||\boldsymbol{Q}_A - \tilde{\boldsymbol{Q}}_A||_2$$

$$\leq ||\boldsymbol{Q}_A - \tilde{\boldsymbol{Q}}_A||_1$$

$$= O_p\left(\sum_{j_1=1}^{q}\sum_{k_1=1}^{m_n}|\gamma_{k_1 k_2}^{(j_1,j_2)}|\right)$$

$$= O_p(m_n^2 n^{-\frac{1}{2}}).$$

This leads to $|\Lambda_{\min}(\boldsymbol{Q}_A) - \Lambda_{\min}(\tilde{\boldsymbol{Q}}_A)| = o_p(m_n^{-1})$ if $m_n = o(n^{1/6})$.

On the other hand, by Lemma 3 of Huang *et al.* (2010), there exists positive constants C_5 and C_6 such that $C_5 m_n^{-1} \leq \Lambda_{\min}(\tilde{\boldsymbol{Q}}_A) \leq \Lambda_{\max}(\tilde{\boldsymbol{Q}}_A) \leq C_6 m_n^{-1}$ with probability converging to 1. Therefore, we can find a positive constant C_3, satisfying $\Lambda_{\min}(\boldsymbol{Q}_A) \geq C_3 m_n^{-1}$ with probability converging to 1. Similarly, we can prove that, with probability tending to 1, there exists a constant C_4 such that $\Lambda_{\max}(\boldsymbol{Q}_A) \leq C_4 m_n^{-1}$. $\qquad\square$

B.3. Proof of Theorem 1:

Since $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_1^T, \ldots, \tilde{\boldsymbol{\beta}}_d^T)^T$ minimizes (2.5), it follows that

$$(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}})^T(\boldsymbol{y} - \boldsymbol{Z}\tilde{\boldsymbol{\beta}}) + \lambda_1\sum_{j=1}^{d}||\tilde{\boldsymbol{\beta}}_j||_2 \leq (\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}) + \lambda_1\sum_{j=1}^{d}||\boldsymbol{\beta}_j||_2.$$

$$(\text{S2.5})$$

For any set $A \subset \{1, \ldots, d\}$, $\boldsymbol{\beta}_A$ denotes the long vector obtained by stacking

vectors $\boldsymbol{\beta}_j$, $j \in A$. Let $A_2 = A_1 \cup \tilde{A}_1$. Then (S2.5) can be written as

$$||(\boldsymbol{y} - \boldsymbol{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{A_2})||_2^2 + \lambda_1 \sum_{j \in A_2} ||\tilde{\boldsymbol{\beta}}_j||_2 \leq ||(\boldsymbol{y} - \boldsymbol{Z}_{A_2}\boldsymbol{\beta}_{A_2})||_2^2 + \lambda_1 \sum_{j \in A_2} ||\boldsymbol{\beta}_j||_2. \quad \text{(S2.6)}$$

Let $\boldsymbol{u}_n = \boldsymbol{y} - \boldsymbol{Z}_{A_2}\boldsymbol{\beta}_{A_2}$ and $\boldsymbol{v}_n = \boldsymbol{Z}_{A_2}(\tilde{\boldsymbol{\beta}}_{A_2} - \boldsymbol{\beta}_{A_2})$. After some simple algebra,

we can write (S2.6) as $\boldsymbol{v}_n^T \boldsymbol{v}_n - 2\boldsymbol{u}_n \boldsymbol{v}_n \leq \lambda_1 \sum_{j \in A_2}(||\boldsymbol{\beta}_j||_2 - ||\tilde{\boldsymbol{\beta}}_j||)$. Note that

$$\sum_{j \in A_2}(||\boldsymbol{\beta}_j||_2 - ||\tilde{\boldsymbol{\beta}}_j||_2) \leq \sum_{j \in A_1}(||\boldsymbol{\beta}_j||_2 - ||\tilde{\boldsymbol{\beta}}_j||_2)$$

$$\leq \sum_{j \in A_1} ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2$$

$$\leq \sqrt{|A_1|} ||\boldsymbol{\beta}_{A_1} - \tilde{\boldsymbol{\beta}}_{A_1}||_2$$

$$\leq \sqrt{|A_1|} ||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2.$$

Hence (S2.6) can be simplified as

$$\boldsymbol{v}_n^T \boldsymbol{v}_n - 2\boldsymbol{u}_n^T \boldsymbol{v}_n \leq \lambda_1 \sqrt{|A_1|} ||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2. \quad \text{(S2.7)}$$

Let $\boldsymbol{u}_n^\star = \boldsymbol{Z}_{A_2}(\boldsymbol{Z}_{A_2}^T \boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}_{A_2}^T \boldsymbol{u}_n$. In other words, $\boldsymbol{u}_n^\star$ represents the projection of $\boldsymbol{u}_n$ onto the space spanned by the columns of $\boldsymbol{Z}_{A_2}$. The matrix inverse of $\boldsymbol{Z}_{A_2}^T \boldsymbol{Z}_{A_2}$ exists due to Lemma 3, which indicates that the $\Lambda_{\min}(\boldsymbol{Z}_{A_2}^T \boldsymbol{Z}_{A_2})$ is positive with probability approaching to 1. Applying the Cauchy-Schwarz inequality, we have

$$2|\boldsymbol{u}_n^T \boldsymbol{v}_n| = 2|(\boldsymbol{u}_n^\star)^T \boldsymbol{v}_n| \leq 2||\boldsymbol{u}_n^\star||_2 ||\boldsymbol{v}_n||_2 \leq 2||\boldsymbol{u}_n^\star||_2^2 + \frac{1}{2}||\boldsymbol{v}_n||_2^2, \quad \text{(S2.8)}$$

where the last inequality is from the fact that $ab \leq a^2 + b^2/4$ for $\forall a, b \in \mathbb{R}$.

Plugging (S2.8) into (S2.7), we have

$$\boldsymbol{v}_n^T \boldsymbol{v}_n \leq 4||\boldsymbol{u}_n^\star||_2^2 + 2\lambda_1 \sqrt{|A_1|}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2. \qquad (S2.9)$$

By Lemma 3, $\boldsymbol{v}_n^T \boldsymbol{v}_n \geq nC_3 m_n^{-1}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2$ for some positive constant C_3, with probability approaching 1. Since $2ab \leq a^2 + b^2 \ \forall a, b \in \mathbb{R}$, it can be derived from (S2.9) that with probability approaching 1,

$$nC_3 m_n^{-1}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2 \leq 4||\boldsymbol{u}_n^\star||_2^2 + 2\lambda_1 \sqrt{|A_1|}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2$$

$$\leq 4||\boldsymbol{u}_n^\star||_2^2 + \frac{(2\lambda_1 \sqrt{|A_1|})^2}{2nC_3 m_n^{-1}} + \frac{n}{2} C_3 m_n^{-1}||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2.$$

Therefore,

$$||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2 \leq \frac{8||\boldsymbol{u}_n^\star||_2^2}{nC_3 m_n^{-1}} + \frac{4\lambda_1^2 |A_1|}{n^2 C_3^2 m_n^{-2}}, \qquad (S2.10)$$

with probability approaching 1.

Next we study the magnitude of $||\boldsymbol{u}_n^\star||_2^2$. Denote the ith component in $\boldsymbol{u}_n$ as u_i, $i = 1, \ldots, n$. Then u_i can be expressed as

$$u_i = y_i - \bar{y} - \sum_{j \in A_2} \boldsymbol{Z}_{ij} \boldsymbol{\beta}_j$$

$$= y_i - a - \sum_{j=1}^d f_j(\zeta_{ij}) + a - \bar{y} + \sum_{j=1}^d f_j(\zeta_{ij}) - \sum_{j=1}^d f_j(\hat{\zeta}_{ij})$$

$$+ \sum_{j=1}^d f_j(\hat{\zeta}_{ij}) - \sum_{j \in A_2} \boldsymbol{Z}_{ij} \boldsymbol{\beta}_j$$

$$= E_{1i} + E_{2i} + E_{3i} + E_{4i},$$

where $E_{1i} = \epsilon_i$, $E_{2i} = a - \bar{y}$, $E_{3i} = \sum_{j=1}^d (f_j(\zeta_{ij}) - f_j(\hat{\zeta}_{ij}))$ and $E_{4i} =$

12

$\sum_{j \in A_2} (f_j(\hat{\zeta}_{ij}) - \boldsymbol{Z}_{ij}\boldsymbol{\beta}_j)$. Let $\boldsymbol{E}_j = (E_{j1}, \ldots, E_{jn})^T$, $1 \leq j \leq 4$ and $\boldsymbol{P}_{A_2} = \boldsymbol{Z}_{A_2}(\boldsymbol{Z}_{A_2}^T \boldsymbol{Z}_{A_2})^{-1}\boldsymbol{Z}_{A_2}^T$. Then we have

$$||\boldsymbol{u}_n^{\star}||_2 = ||\boldsymbol{P}_{A_2}\boldsymbol{u}_n||_2^2 \leq 4||\boldsymbol{\epsilon}^{\star}||_2^2 + 4||\boldsymbol{E}_2||_2^2 + 4||\boldsymbol{E}_3||_2^2 + 4||\boldsymbol{E}_4||_2^2, \quad \text{(S2.11)}$$

where $\boldsymbol{\epsilon}^{\star}$ is the projection of $\boldsymbol{E}_1$ onto the space spanned by columns in $\boldsymbol{Z}_{A_2}$.

Obviously $||\boldsymbol{E}_2||_2^2 = O_p(1)$ and $||\boldsymbol{E}_4||_2^2 = O_p(1 + nm_n^{-2\rho})$ by Proposition 1.

For $||\boldsymbol{E}_3||_2^2$, there are two different cases. (i) If $r = 0$, by Hölder's inequality, it follows that

$$
\begin{aligned}
||\boldsymbol{E}_3||_2^2 &= \sum_{i=1}^{n} \left(\sum_{j=1}^{d} (f_j(\zeta_{ij}) - f_j(\hat{\zeta}_{ij})) \right)^2 \\
&\leq L^2 \sum_{i=1}^{n} \left(\sum_{j=1}^{d} |\zeta_{ij} - \hat{\zeta}_{ij}|^{\rho} \right)^2 \\
&\leq L^2 d^{2-2\rho} \sum_{i=1}^{n} \left(\sum_{j=1}^{d} |\zeta_{ij} - \hat{\zeta}_{ij}| \right)^{2\rho} \\
&\leq L^2 d^{2-2\rho} n^{1-\rho} \left(\sum_{i=1}^{n} \left(\sum_{j=1}^{d} |\zeta_{ij} - \hat{\zeta}_{ij}| \right)^2 \right)^{\rho} \\
&= O_p(n^{1-\rho}). \quad \text{(S2.12)}
\end{aligned}
$$

The last equation is based on Lemma 1. (ii) If $r \geq 1$, then f_j's are continuously differentiable on [0,1]. Then there exists a positive constant M such

13

that $\max_{1 \leq j \leq d} |f_j(s) - f_j(t)| \leq M|s - t| \; \forall s, \; t \in [0, 1]$. Therefore,

$$||\boldsymbol{E}_3||_2^2 = \sum_{i=1}^{n} \left(\sum_{j=1}^{d} (f_j(\zeta_{ij}) - f_j(\hat{\zeta}_{ij})) \right)^2$$

$$\leq M^2 \sum_{i=1}^{n} \left(\sum_{j=1}^{d} |\zeta_{ij} - \hat{\zeta}_{ij}| \right)^2$$

$$= O_p(1), \tag{S2.13}$$

where the last equation is obtained from Lemma 1. For $||\boldsymbol{\epsilon}^\star||_2^2$, we have

$||\boldsymbol{\epsilon}^\star||_2^2 \leq ||(\boldsymbol{Z}_{A_2}^T \boldsymbol{Z}_{A_2})^{-\frac{1}{2}} \boldsymbol{Z}_{A_2}^T \boldsymbol{\epsilon}||_2^2 \leq \frac{m_n}{C_3 n} ||\boldsymbol{Z}_{A_2}^T \boldsymbol{\epsilon}||_2^2$ with probability approaching

1. By Lemma 2, we have

$$\max_{A:|A| \leq d} ||\boldsymbol{Z}_A^T \boldsymbol{\epsilon}||_2^2 = \max_{A:|A| \leq d} \sum_{j \in A} ||\boldsymbol{Z}_j^T \boldsymbol{\epsilon}||_2^2$$

$$\leq dm_n \max_{1 \leq j \leq d, 1 \leq k \leq m_n} |\sum_{i=1}^{n} \psi_k(\hat{\zeta}_{ij}) \epsilon_i|^2$$

$$= O_p(n \log m_n).$$

Hence,

$$||\boldsymbol{\epsilon}^\star||_2^2 = O_p(m_n \log m_n). \tag{S2.14}$$

In both cases ($r = 0$ or $r \geq 1$), combining (S2.10), (S2.11), (S2.12) (or (S2.13)) and (S2.14), we have

$$||\boldsymbol{\beta}_{A_2} - \tilde{\boldsymbol{\beta}}_{A_2}||_2^2 = O_p \left(\frac{m_n^2 \log m_n}{n} \right) + O_p \left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho-1}} \right) + O_p \left(\frac{\lambda_1^2 m_n^2}{n^2} \right).$$

$$\tag{S2.15}$$

This completes the proof of (ii) of Theorem 1.

14

Now we go back to prove part (i). For $j \in A_1$, $||f_j||_2 \geq c_f$ under Assumption C1. If $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})$, then from the proof of Proposition 1, we have $||f_j - f'_j||_2 = O(m_n^{-\rho})$, where $f'_j(x) = \sum_{k=1}^{m_n} B_k(x)\beta_{jk}$. Therefore, $||f'_j||_2 \geq ||f_j||_2 - ||f_j - f'_j||_2$ by the triangle inequality, which leads to $||f'_j||_2 \geq 0.5c_f$ when n is sufficiently large. On the other hand, based on (12) of Stone (1986), $C_7 m_n^{-1}||\boldsymbol{\beta}_j||_2^2 \leq ||f'_j||_2^2 \leq C_8 m_n^{-1}||\boldsymbol{\beta}_j||_2^2$ for some positive constants C_7 and C_8. Combining the above facts, we have $||\boldsymbol{\beta}_j||_2^2 \geq C_8^{-1} m_n ||f'_j||_2^2 \geq 0.25 C_8^{-1} m_n c_f^2$, $j \in A_1$. If $\tilde{\boldsymbol{\beta}}_j = 0$ for some $j \in A_1$, then for such j, we have $||\boldsymbol{\beta}_j||_2^2 = ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \geq 0.25 C_8^{-1} m_n c_f^2$. However, by part (ii), we have $||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 = o_p(1)$ since $m_n^2 \log(m_n)/n \to 0$ and $(\lambda_1^2 m_n^2)/n^2 \to 0$. Therefore, $||\tilde{\boldsymbol{\beta}}_j||_2 > 0$ $\forall j \in A_1$ with probability approaching 1. This completes the proof of part (i). $\qquad\square$

B.4. Proof of Theorem 2:

Part (i) is immediately available from the definition of $\tilde{f}_j$ and $\tilde{\boldsymbol{\beta}}_j$ and part (i) of Theorem 1. Based on properties of B-spline functions (De Boor (1968)), there exists $f'_j \in \mathscr{S}_n$ such that $||f_j - f'_j||_\infty = O(m_n^{-\rho})$. Let $f_{nj}(x) = f'_j(x) - n^{-1}\sum_{i=1}^n f'_j(\hat{\zeta}_{ij})$, As shown in Proposition 1, $f_{nj}(x)$ satisfies $\forall x \in [0,1]$

$$|f_{nj}(x) - f'_j(x)|^2 = \left|n^{-1}\sum_{i=1}^n f'_j(\hat{\zeta}_{ij})\right|^2 = O_p(m_n^{-2\rho} + n^{-1}).$$

15

Then it follows that

$$||f_j - f_{nj}||_2^2 = O_p(m_n^{-2\rho} + n^{-1}). \tag{S2.16}$$

Let $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm_n})$ and $\tilde{\boldsymbol{\beta}}_j = (\tilde{\beta}_{j1}, \ldots, \tilde{\beta}_{jm_n})$. Then $f_{nj}(x) = \sum_{k=1}^{m_n} \beta_{jk} B_k(x) - n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_n} \beta_{jk} B_k(\hat{\zeta}_{ij})$ and $\tilde{f}_j(x) = \sum_{k=1}^{m_n} \tilde{\beta}_{jk} B_k(x) - n^{-1} \sum_{i=1}^{n} \sum_{k=1}^{m_n} \tilde{\beta}_{jk} B_k(\hat{\zeta}_{ij})$. Therefore,

$$||f_{nj} - \tilde{f}_j||_2^2 = \int_0^1 (f_{nj}(x) - \tilde{f}_j(x))^2 dx$$

$$\leq 2E_1 + 2E_2, \tag{S2.17}$$

where $E_1 = \int_0^1 (\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) B_k(x))^2 dx$ and $E_2 = \left(\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^{n} B_k(\hat{\zeta}_{ij}) \right)^2$. By (12) of Stone (1986), we have

$$E_1 \leq C_9 m_n^{-1} ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \tag{S2.18}$$

for some positive constant C_9. The second term E_2 can be written as

$$E_2 = \Big(\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^{n} (B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij}))$$

$$+ \sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^{n} B_k(\zeta_{ij}) \Big)^2 \tag{S2.19}$$

$$\leq 2E_3 + 2E_4, \tag{S2.20}$$

where $E_3 = \left(\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^{n} (B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})) \right)^2$ and $E_4 = \left(\sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) n^{-1} \sum_{i=1}^{n} B_k(\zeta_{ij}) \right)^2$. Applying the Cauchy-Schwarz

16

inequality, we have

$$E_3 \leq ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2 \sum_{k=1}^{m_n} \left(n^{-1} \sum_{i=1}^{n} |B_k(\hat{\zeta}_{ij}) - B_k(\zeta_{ij})| \right)^2$$

$$\leq O(m_n^3) n^{-2} \left(\sum_{i=1}^{n} |\hat{\zeta}_{ij} - \zeta_{ij}| \right)^2 ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2$$

$$\leq O(m_n^3) n^{-1} \left(\sum_{i=1}^{n} |\hat{\zeta}_{ij} - \zeta_{ij}|^2 \right) ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2$$

$$= O_p(m_n^{-1}) ||\boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j||_2^2, \tag{S2.21}$$

where the second inequality holds since the derivative of B-spline basis functions satisfies $||B_k'||_\infty = O(m_n)$ uniformly over $1 \leq k \leq m_n$, and the third inequality follows from Lemma 1. Next we deal with E_4. Let $h(x) = \sum_{k=1}^{m_n} (\beta_{jk} - \tilde{\beta}_{jk}) B_k(x)$ for $x \in [0,1]$. Thus $h(x) \in \mathscr{S}_n$. As shown in Lemma 1 of Huang et $al.$ (2010), h satisfies $n^{-1} \sum_{i=1}^{n} h(\zeta_{ij}) - \mathrm{E}\, h(\zeta_{ij}) = O_p(n^{-\frac{1}{2}} m_n^{\frac{1}{2}})$.

Hence

$$
\begin{aligned}
E_4 &= \left(n^{-1} \sum_{i=1}^{n} h(\zeta_{ij}) \right)^2 \\
&= \left(n^{-1} \sum_{i=1}^{n} (h(\zeta_{ij}) - \mathrm{E}\, h(\zeta_{ij})) + \mathrm{E}\, h(\zeta_{ij}) \right)^2 \\
&\leq O_p(n^{-1} m_n) + 2(\mathrm{E}\, h(\zeta_{ij}))^2 \\
&= O_p(n^{-1} m_n) + 2 \left(\int_0^1 h(x) g_j(x) dx \right)^2 \\
&\leq O_p(n^{-1} m_n) + 2 C_2^2 \int_0^1 h^2(x) dx \\
&\leq O_p(n^{-1} m_n) + 2 C_2^2 C_9 m_n^{-1} \| \boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j \|_2^2, \qquad \text{(S2.22)}
\end{aligned}
$$

where the second inequality is by the Cauchy-Schwarz inequality and the last one is based on (S2.18). Combining (S2.17)- (S2.22), we have

$$
\| f_{nj} - \tilde{f}_j \|_2^2 = O_p(m_n^{-1} \| \boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j \|_2^2 + n^{-1} m_n). \qquad \text{(S2.23)}
$$

Since $\| \tilde{f}_j - f_j \|_2^2 \leq 2 \| f_j - f_{nj} \|_2^2 + 2 \| \tilde{f}_j - f_{nj} \|_2^2$, (S2.16) and (S2.23) lead to

$$
\| \tilde{f}_j - f_j \|_2^2 = O_p(m_n^{-2\rho} + n^{-1}) + O_p(m_n^{-1} \| \boldsymbol{\beta}_j - \tilde{\boldsymbol{\beta}}_j \|_2^2 + n^{-1} m_n).
$$

Additionally, by part (ii) of Theorem 1, it follows that for $j \in A_1 \cup \tilde{A}_1$,

$$
\| \tilde{f}_j - f_j \|_2^2 = O_p \left(\frac{m_n \log m_n}{n} \right) + O_p \left(\frac{m_n}{n} + \frac{1}{m_n^{2\rho}} \right) + O_p \left(\frac{m_n \lambda_1^2}{n^2} + \frac{m_n}{n} \right).
$$

This completes the proof of part (ii). $\qquad \square$

18

The proof of Theorem 3 follows the proof of Corollary 2 in Huang *et al.* (2010) with a similar change made to prove the part (ii) of Theorem 1, and is omitted here.

S3. Additional simulation results

Simulation results for $n = 200$ or 300 curves are summarized in Table S1.

Statistics	n	Methods					
		MARS	FAM	S-FAM	CSE-FAM	AGL-FAM	CSS-FAM
MSPE	200	.92 (.14)	.79 (.11)	.93 (.15)	.81 (.11)	.97 (.14)	.83 (.12)
	300	.87 (.10)	.77 (.08)	.84 (.10)	.78 (.09)	.88 (.10)	.77 (.09)
$\text{RSS}(\hat{f}_1)$	200	-	1.1 (1.8)	1.8 (2.0)	1.0 (1.3)	6.2 (3.4)	1.4 (1.4)
$(\times 10^{-2})$	300	-	0.6 (0.9)	1.0 (0.9)	0.8 (0.7)	3.9 (1.5)	1.0 (0.8)
$\text{RSS}(\hat{f}_2)$	200	-	1.8 (2.3)	5.0 (3.7)	6.2 (4.0)	9.0 (4.4)	4.8 (3.3)
$(\times 10^{-2})$	300	-	1.2 (1.5)	3.0 (2.0)	4.0 (2.2)	5.2 (2.5)	2.9 (2.2)
$\text{RSS}(\hat{f}_4)$	200	-	2.4 (3.4)	1.8 (2.2)	1.7 (1.3)	9.2 (6.2)	1.6 (1.3)
$(\times 10^{-2})$	300	-	1.4 (1.8)	1.2 (1.0)	0.9 (0.7)	4.0 (1.8)	0.9 (0.7)
TP%	200	100 (.0)	-	100 (.0)	99.7 (.03)	96.7 (.10)	96.7 (.10)
	300	100 (.0)	-	100 (.0)	100 (.0)	100 (.0)	100 (.0)
FP%	200	25.5 (.13)	-	11.8 (.10)	2.1 (.05)	0.1 (.01)	0.1 (.01)
	300	27.5 (12.43)	-	9.5 (.08)	1.5 (.04)	0.1 (.01)	0.1 (.01)

Table S1: Summary statistics for evaluating six methods. MSPE refers to the mean squared prediction error on the test data; the residual sum of squares (RSS) for each estimated component $\hat{f}_j$ is defined as: $\text{RSS}(\hat{f}_j) = \int_0^1 (\hat{f}_j(x) - f_j(x))^2 dx$; TP% and FP% stand for the true positive and false positive rates in percentage, respectively. The point estimate for each measure is averaged over 100 simulation replicates, and the corresponding estimated standard error is given in parenthesis.

S4. Application to air pollution data

It is of great interest to study the association between air pollution and respiratory system diseases (Xing *et al.* 2016). In this example we investigate this association using the sparse functional additive model (2.3). More specifically, in this article we study the effect of fine particulate matter (PM 2.5) on the rate of death caused by respiratory diseases (RESP).

The data consists of the median of the daily observations of PM 2.5 measured in micrograms per cubic meter (μg/m^3) from 1987 to 2000 across 108 cities of the United States. Some negative values appear due to detrending the time series in order to make them centered around 0 (see Page 42 in Peng and Dominici 2008). This data is obtained from the **NMMAPS-data** package (Peng and Welty 2004), which was originally assembled for the national morbidity, mortality, and air pollution study. In this article we focus on exploring the relationship between the daily observations of PM 2.5 and the RESP death rate in the year 2000, since considerably fewer missing values are present during that year. There are 7 cities that have no data for PM 2.5 in that year, so we remove these 7 cities from the data set. Among the 101 remaining cites, we randomly select 80 cities as the training set. The test set consists of the trajectories of PM 2.5 in the remaining 21 cities.

Six methods are used to predict the RESP death rate using the trajectories of PM 2.5. Table S2 displays the mean squared prediction errors (MSPEs) of these six methods. Our proposed CSS-FAM method has the lowest prediction error among the six methods. Compared with MARS and S-FAM, LAF achieves better performance in prediction. A possible reason is that even though MARS and S-FAM models the relationship between the functional predictor PM 2.5 and the response in a more flexible way than the linear model, variabilities in these models are not well-controlled. LAF, however, reduces the variability to a great extent by variable selection. Furthermore, since CSS-FAM and CSE-FAM combine nonparametric modelling and selection of components, they can not only provide a more adequate characterization of the relationship between the functional predictor and the response than FLR, but also result in a model with less variability and better prediction. In addition, the obvious distinction between CSS-FAM and AGL-FAM suggests that the further smoothing of the estimate from the adaptive group LASSO method via smoothing splines can improve the prediction accuracy.

We use 15 cubic B-spline basis functions to represent the additive components in Model (2.3) when applying our proposed CSS-FAM method. In the group LASSO step, the optimal value of the tuning parameter λ_1 is

	Methods					
	MARS	LAF	S-FAM	CSE-FAM	AGL-FAM	CSS-FAM
Air Pollution ($\times 10^{-8}$)	1.68	1.68	1.77	1.61	1.68	1.49

Table S2: Mean squared prediction errors on the test data for six methods in the air pollution data.

3×10^{-6}, which is determined by a 5-fold cross validation. Another 5-fold cross-validation suggests that $\lambda_2 = 1.5 \times 10^{-10}$ is an optimal choice of the tuning parameter in the adaptive LASSO step. As a result, 3 non-vanishing components, $\{\hat{f}_5, \hat{f}_8, \hat{f}_{11}\}$, are selected. These raw estimates turn out to be excessively wiggly. We therefore refine these raw estimates via smoothing splines, in which the optimal choice of the smoothing parameter λ_3 determined by GCV is 1.2×10^{-5}.

References

De Boor, C. (1968). On uniform approximation by splines. *Journal of Approximation Theory*, **1**(2), 219–235.

Gershgorin, S. A. (1931). Uber die abgrenzung der eigenwerte einer matrix. *Izv. Akad. Nauk. USSR Otd. Fiz.-Mat. Nauk (in German)*, (6), 749–754.

Huang, J., Horowitz, J. L., and Wei, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics*, **38**(4), 2282.

Peng, R. D. and Dominici, F. (2008). *Statistical methods for environmental epidemiology with R*. New York: Springer.

Peng, R. D. and Welty, L. J. (2004). The NMMAPSdata package. *R news*, **4**(2), 10–14.

Prochazkova, J. (2005). Derivative of B-spline function. In *Proceedings of the 25th Conference on Geometry and Computer Graphics*. Prague, Czech Republic.

Stone, C. J. (1985). Additive regression and other nonparametric models. *The Annals of Statistics*, **13**(2), 689–705.

Stone, C. J. (1986). The dimensionality reduction principle for generalized additive models. *The Annals of Statistics*, **14**(2), 590–606.

Van Der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes*. New York: Springer.

Weyl, H. (1912). Das asymptotische verteilungsgesetz der eigenwerte linearer partieller differentialgleichungen (mit einer anwendung auf die theorie der hohlraumstrahlung). *Mathematische Annalen*, **71**(4), 441–479.

Xing, Y.-F., Xu, Y.-H., Shi, M.-H., and Lian, Y.-X. (2016). The impact of PM2. 5 on the human respiratory system. *Journal of thoracic disease*, **8**(1), 69–74.

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

Zhu, H., Yao, F., and Zhang, H. H. (2014). Structured functional additive regression in repro-

25

ducing kernel Hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **76**(3), 581–603.
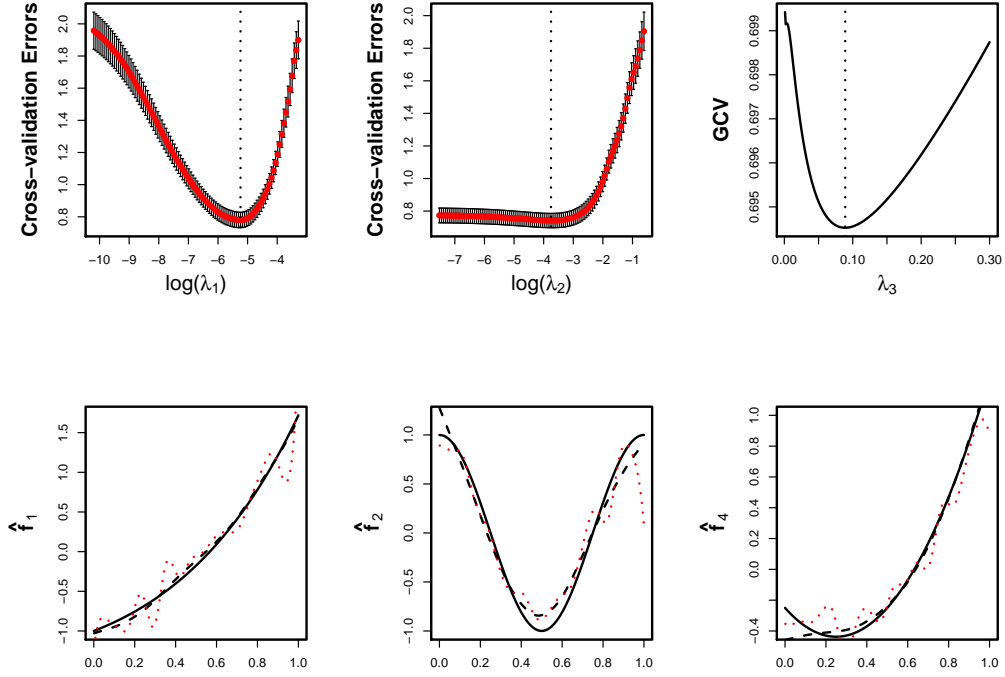
Figure S1: The top three panels illustrate how the cross-validation errors and GCV change with the tuning parameter λ_1, λ_2 and λ_3 in group LASSO, adaptive group LASSO and smoothing spline, respectively. The bottom three panels compare the estimated nonparametric components and the true underlying nonparametric components (f_k, $k = 1, 2, 4$). Dashed lines and dotted lines represent the estimated nonparametric components with and without smoothing spline after the adaptive group LASSO fitting, respectively; while solid lines represent the true underlying nonparametric components.
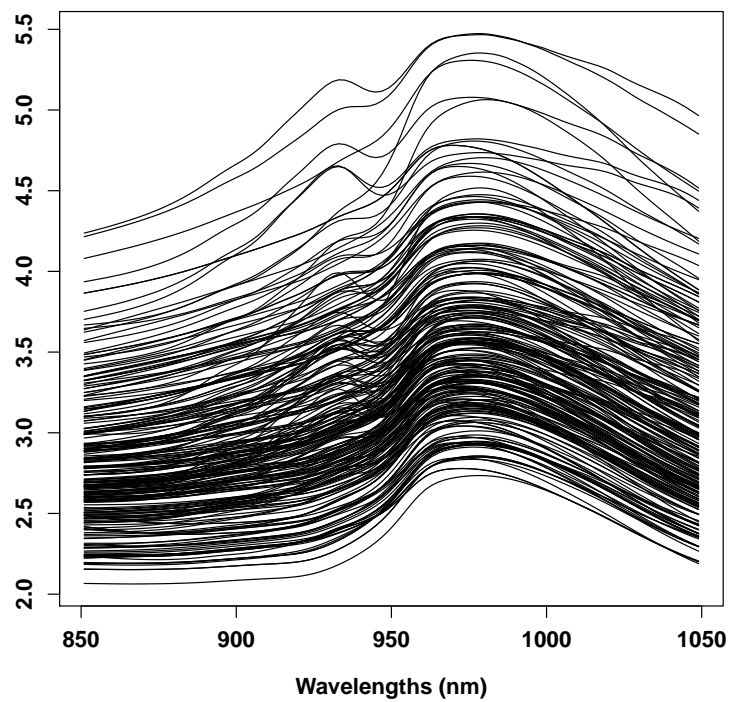
Figure S2: The profiles of the 100-channel spectrum of absorbance of the 240 meat samples.