FUNCTIONAL ADAPTIVE DOUBLE-SPARSITY ESTIMATOR FOR FUNCTIONAL LINEAR REGRESSION MODEL WITH MULTIPLE FUNCTIONAL COVARIATES

Cheng Cao[1], Jiguo Cao[2], Hailiang Wang[3], Kwok-Leung Tsui[4], Xinyue Li[1]

[1] *City University of Hong Kong*, [2] *Simon Fraser University*,

[3] *The Hong Kong Polytechnic University*, [4] *Virginia Polytechnic Institute and State University*

Supplementary Material

This supplementary material includes illustrations of the simulation studies, real data application, and theoretical results of our proposed method. We provide derivations of Alternating Direction Method of Multipliers (ADMM) subproblems in Section S1. The proofs of Theorem 1-Theorem 3 are given in Section S2. The figures to demonstrate simulation performance and Timed Up and Go test are shown in Section S3 and S4.

S1 Derivation of ADMM subproblems

S1.1 Solution to $\tilde{\boldsymbol{b}}$-update

Based on the Equation (8),

$$\begin{aligned}\tilde{\boldsymbol{b}}_l^{k+1} &= \underset{\tilde{\boldsymbol{b}}_l\in\mathbb{R}^{M_n+d}}{\arg\min}\ \frac{1}{2}||\boldsymbol{r}_{(-l)} - \widetilde{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l||_2^2 + \lambda_2||\tilde{\boldsymbol{b}}_l||_2 + (\boldsymbol{u}_l^k)^T((\boldsymbol{L}_l^T)^{-1}\tilde{\boldsymbol{b}}_l - \boldsymbol{z}_l^k) + \frac{\rho}{2}||(\boldsymbol{L}_l^T)^{-1}\tilde{\boldsymbol{b}}_l - \boldsymbol{z}_l^k||_2^2\\ &= \underset{\tilde{\boldsymbol{b}}_l\in\mathbb{R}^{M_n+d}}{\arg\min}\ \frac{1}{2}||\boldsymbol{r}_{(-l)} - \widetilde{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l||_2^2 + \lambda_2||\tilde{\boldsymbol{b}}_l||_2 + \frac{\rho}{2}||(\boldsymbol{L}_l^T)^{-1}\tilde{\boldsymbol{b}}_l - \boldsymbol{z}_l^k - \boldsymbol{u}_l^k/\rho||_2^2\\ &= \underset{\tilde{\boldsymbol{b}}_l\in\mathbb{R}^{M_n+d}}{\arg\min}\ \lambda_2||\tilde{\boldsymbol{b}}_l||_2 + \frac{1}{2}||\hat{\boldsymbol{r}}_{(-l)}^k - \widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l||_2^2,\end{aligned}$$

where $\widehat{\boldsymbol{U}}_l = \big(\widetilde{\boldsymbol{U}}_l^T, \sqrt{\rho}\boldsymbol{L}_l^{-1}\big)^T$ and $\hat{\boldsymbol{r}}_{(-l)}^k = \big(\boldsymbol{r}_{(-l)}^T, \sqrt{\rho}(\boldsymbol{z}_l^k + \boldsymbol{u}_l^k/\rho\big)^T)^T$. Since $\widehat{\boldsymbol{U}}_l$ is not identity, we cannot directly get the solution to the second operator. Fortunately, by following Wang and Yuan (2012) we can employ linearization technique to approximate the quadratic term efficiently. We linearize it by replacing $||\hat{\boldsymbol{r}}_{(-l)}^k - \widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l||_2^2/2$ with $\big(\widehat{\boldsymbol{U}}_l^T\big(\widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l^k - \hat{\boldsymbol{r}}_{(-l)}^k\big)\big)^T(\tilde{\boldsymbol{b}}_l - \tilde{\boldsymbol{b}}_l^k) + \frac{\nu_l}{2}||\tilde{\boldsymbol{b}}_l - \tilde{\boldsymbol{b}}_l^k||_2^2$, where the first term is apparently the gradient at $\tilde{\boldsymbol{b}}_l^k$. Hence,

$$\begin{aligned}\tilde{\boldsymbol{b}}_l^{k+1} &= \underset{\tilde{\boldsymbol{b}}_l\in\mathbb{R}^{M_n+d}}{\arg\min}\ \lambda_2||\tilde{\boldsymbol{b}}_l||_2 + \big(\widehat{\boldsymbol{U}}_l^T\big(\widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l^k - \hat{\boldsymbol{r}}_{(-l)}^k\big)\big)^T(\tilde{\boldsymbol{b}}_l - \tilde{\boldsymbol{b}}_l^k) + \frac{\nu_l}{2}||\tilde{\boldsymbol{b}}_l - \tilde{\boldsymbol{b}}_l^k||_2^2\\ &= \underset{\tilde{\boldsymbol{b}}_l\in\mathbb{R}^{M_n+d}}{\arg\min}\ \lambda_2||\tilde{\boldsymbol{b}}_l||_2 + \frac{\nu_l}{2}||\tilde{\boldsymbol{b}}_l - \tilde{\boldsymbol{b}}_l^k + \widehat{\boldsymbol{U}}_l^T\big(\widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l^k - \hat{\boldsymbol{r}}_{(-l)}^k\big)/\nu_l||_2^2\\ &= S_{2,\lambda_2/\nu_l}(\tilde{\boldsymbol{b}}_l^k - \widehat{\boldsymbol{U}}_l^T\big(\widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l^k - \hat{\boldsymbol{r}}_{(-l)}^k\big)/\nu_l).\end{aligned}$$

If ρ is fixed, $\tilde{\boldsymbol{b}}_l^{k+1}$ will tend to be zero when $\lambda_2 \geq \nu_l||\tilde{\boldsymbol{b}}_l^k - \widehat{\boldsymbol{U}}_l^T\big(\widehat{\boldsymbol{U}}_l\tilde{\boldsymbol{b}}_l^k - \hat{\boldsymbol{r}}_{(-l)}^k\big)/\nu_l||_2$.

S1.2 Solution to $\boldsymbol{z}$-update

Based on the Equation (9),

$$\begin{aligned}\boldsymbol{z}^{k+1} &= \underset{\boldsymbol{z}\in\mathbb{R}^{J\times(M_n+d)}}{\arg\min}\ \lambda_1||\boldsymbol{z}||_1 + (\boldsymbol{u}^k)^T(\boldsymbol{D}\tilde{\boldsymbol{b}}^{k+1} - \boldsymbol{z}) + \frac{\rho}{2}||\boldsymbol{D}\tilde{\boldsymbol{b}}^{k+1} - \boldsymbol{z}||_2^2 \\ &= \underset{\boldsymbol{z}\in\mathbb{R}^{J\times(M_n+d)}}{\arg\min}\ \lambda_1||\boldsymbol{z}||_1 + \frac{\rho}{2}||\boldsymbol{D}\tilde{\boldsymbol{b}}^{k+1} - \boldsymbol{z} + \boldsymbol{u}^k/\rho||_2^2 \\ &= S_{1,\lambda_1/\rho}(\boldsymbol{D}\tilde{\boldsymbol{b}}^{k+1} + \boldsymbol{u}^k/\rho).\end{aligned}$$

If ρ is fixed, z_r^{k+1} will tend to be zero when $\lambda_1 > \left|\rho\boldsymbol{D}_{r\cdot}^T\tilde{\boldsymbol{b}}^{k+1} + u_r^k\right|$, $r = 1,\ldots,J\times(M_n+d)$, where $\boldsymbol{D}_{r\cdot}$ is the rth row of $\boldsymbol{D}$.

S2 Proofs

B-splines are essential in the estimation of coefficient function for functional model. Before presenting the proofs of the proposed estimator, it is necessary to state some properties of B-splines. As mentioned in Section 2.3, B-splines have a local support property: at most $d+1$ consecutive subintervals are nonzero. Plus, for a collection of B-spline basis functions $\{B_k(t): k = 1,\ldots,M_n+d, t\in\mathcal{T}\}$, $B_k(t)\geq 0$ and $\sum_{k=1}^{M_n+d} B_k(t) = 1$ for all t. These properties imply that

$$\sup_{k,r}|\langle B_k, B_r\rangle| \leq 2(d+1)M_n^{-1}, \tag{S2.1}$$

and thus,

$$||B_k||_2^2 \leq \sup_{k,r} |\langle B_k, B_r \rangle| \leq 2(d+1)M_n^{-1}. \tag{S2.2}$$

In addition, three inequalities will be also used. For any $x \in \mathbb{R}^p$

$$||x||_2 \leq ||x||_1 \leq \sqrt{p}||x||_2;$$

$$||x||_\infty \leq ||x||_1 \leq p||x||_\infty;$$

$$||x||_\infty \leq ||x||_2 \leq \sqrt{p}||x||_\infty.$$

S2.1 Proof of Theorem 1

Proof. For simplicity of notation, we assume the model has no intercept, i.e., $\boldsymbol{\mu} = \mathbf{0}$ and rewrite the objective in Equation (2.5)

$$L_n(\boldsymbol{b}) = \frac{1}{n}||\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{b}||_2^2 + \Delta_n \lambda_1 \sum_{l=1}^{J} w_l^{(1)} ||\boldsymbol{b}_l||_1 + \lambda_2 \sum_{l=1}^{J} w_l^{(2)} (\boldsymbol{b}_l^T \boldsymbol{K}_{\varphi,l} \boldsymbol{b}_l)^{1/2}, \tag{S2.3}$$

where $\boldsymbol{K}_{\varphi,l} = \boldsymbol{\Phi}_l + \varphi \boldsymbol{\Omega}_l$ is a $(M_n + d) \times (M_n + d)$ matrix.

We first provide Lemma 2 and 3 to facilitate the proof. Lemma 2 follows Huang et al. (2004) Lemma A.3 and Zhou et al. (2013) A_8, while Lemma 3 refers to Lemma 6.2 of Cardot et al. (2003).

Lemma 2. *If* $\lim_{n\to\infty} M_n \log M_n/n = 0$*, there are positive constants* C_2 *and* C_3 *such that, all eigenvalues of* $(M_n/n)\boldsymbol{U}^T\boldsymbol{U}$ *are within the interval* $[C_2, C_3]$ *with probability tending to 1 as* $n \to \infty$.

Lemma 3. *There are positive constants C_4 and C_5 such that, all eigenvalues of $\boldsymbol{K}_{\varphi,l}$ are within the interval $[C_4\varphi M_n^{-1}, C_5 M_n^{-1}]$.*

Proof. We define $\boldsymbol{K}_{\varphi,l} = \boldsymbol{\Phi}_l + \varphi\boldsymbol{\Omega}_l$, where $(\boldsymbol{\Phi}_l)_{pq} = \int_t B_{lp}(t)B_{lq}(t)dt$ and $(\boldsymbol{\Omega}_l)_{pq} = \int_t B_{lp}^m(t)B_{lq}^m(t)dt$. The proof follows Lemma 6.2 (i) in Cardot et al. (2003).

Let $L_n(\boldsymbol{b} + \eta_n\boldsymbol{v}) - L_n(\boldsymbol{b})$, where η_n is a scalar and $\boldsymbol{v} \in \mathbb{R}^{J\times(M_n+d)}$. At the point $\boldsymbol{b} = \boldsymbol{\alpha}$, we let $\hat{\boldsymbol{b}} = \boldsymbol{\alpha} + \eta_n\boldsymbol{v}$. By the minimality of $\hat{\boldsymbol{b}}$, we have

$$
\begin{aligned}
&\frac{1}{n}\big(||\boldsymbol{Y} - \boldsymbol{U}(\boldsymbol{\alpha} + \eta_n\boldsymbol{v})||_2^2 - ||\boldsymbol{Y} - \boldsymbol{U}\boldsymbol{\alpha}||_2^2\big) \\
\leq &\Delta_n\lambda_1 \sum_{l\in\mathcal{A}} w_l^{(1)}\big(||\boldsymbol{\alpha}_l||_1 - ||\boldsymbol{\alpha}_l + \eta_n\boldsymbol{v}_l||_1\big) + \lambda_2 \sum_{l\in\mathcal{A}} w_l^{(2)}\big((\boldsymbol{\alpha}_l^T\boldsymbol{K}_{\varphi,l}\boldsymbol{\alpha}_l)^{1/2} - \\
&((\boldsymbol{\alpha}_l + \eta_n\boldsymbol{v}_l)^T\boldsymbol{K}_{\varphi,l}(\boldsymbol{\alpha}_l + \eta_n\boldsymbol{v}_l))^{1/2}\big),
\end{aligned}
\tag{S2.4}
$$

because of the fact that $\boldsymbol{\alpha}_l = \boldsymbol{0}$ if $l \in \mathcal{A}^c$.

Let $\epsilon_i = \boldsymbol{Y}_i - \langle\boldsymbol{X}_i, \boldsymbol{\beta}\rangle$, $\boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)$ and $e_i = \langle\boldsymbol{X}_i, \boldsymbol{B}^T\boldsymbol{\alpha}\rangle - \langle\boldsymbol{X}_i, \boldsymbol{\beta}\rangle = \langle\boldsymbol{X}_i, \boldsymbol{\beta}^\alpha - \boldsymbol{\beta}\rangle$, $\boldsymbol{e} = (e_1, \ldots, e_n)$, the LHS of (S2.4) gives

$$
\begin{aligned}
\text{LHS} =& \frac{1}{n}||\boldsymbol{\epsilon} - \boldsymbol{e} - \eta_n\boldsymbol{U}\boldsymbol{v}||_2^2 - \frac{1}{n}||\boldsymbol{\epsilon} - \boldsymbol{e}||_2^2 \\
=& \frac{\eta_n^2}{n}\boldsymbol{v}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{v} - \frac{2\eta_n}{n}(\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{v}.
\end{aligned}
$$

By Lemma 2, $(\eta_n^2/n)\boldsymbol{v}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{v} \geq (\eta_n^2/n)(C_2 n/M_n) = \eta_n^2 O_p(M_n^{-1})$.

Moreover, by Cauchy-Schwarz inequality, $(2\eta_n/n)(\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{v} \leq (2\eta_n/n)||\boldsymbol{v}||_2\big((\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{\epsilon} - \boldsymbol{e})\big)^{1/2}$. Since $\boldsymbol{\epsilon}$ and $\boldsymbol{e}$ are independent, $\mathbb{E}[(\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{\epsilon} -$

$\boldsymbol{e})^T] = \mathbb{E}[\boldsymbol{\epsilon}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}] + \mathbb{E}[\boldsymbol{e}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{e}]$. By A.1, properties of B-splines and independence of ϵ_i, we have

$$\begin{aligned}\mathbb{E}[\boldsymbol{\epsilon}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{\epsilon}] &= \mathbb{E}\left[\sum_{l=1}^{J}\sum_{k=1}^{M_n+d}\left[\sum_{i=1}^{n}\langle X_{li}, B_{lk}\rangle^2\epsilon_i^2 + \sum_{i'\neq i}\langle X_{li}, B_{lk}\rangle\langle X_{i'}, B_{lk}\rangle\epsilon_i\epsilon_{li'}\right]\right]\\ &\leq \sigma^2\sum_{l=1}^{J}\sup_k\sum_{k=1}^{M_n+d}\sum_{i=1}^{n}|\langle X_{li}, B_{lk}\rangle|^2\\ &\leq \sigma^2 Jn||X_{li}||_2^2\sup_k\sum_{r=1}^{M_n+d}|\langle B_{lr}, B_{lk}\rangle|\\ &= O(n).\end{aligned}$$

On the other hand, by Cauchy-Schwarz inequality, A.1, and Lemma 1, $e_i^2 = |\langle \boldsymbol{X}_i, \boldsymbol{\beta}^\alpha - \boldsymbol{\beta}\rangle|^2 \leq ||\boldsymbol{X}_i||_2^2\sum_l\int_{\mathcal{T}}\left[\sup_t|\beta_l^\alpha(t) - \beta_l(t)|\right]^2 dt \leq c_1^2 J|\mathcal{T}|(C_1M_n^{-\delta})^2$, where $|\mathcal{T}|$ represents the length of time domain. The inequality still holds for $e_{il}e_{i'}, \forall i \neq i'$. Hence,

$$\begin{aligned}\mathbb{E}[\boldsymbol{e}^T\boldsymbol{U}\boldsymbol{U}^T\boldsymbol{e}] &= \mathbb{E}\left[\sum_{l=1}^{J}\sum_{k=1}^{M_n+d}\left[\sum_{i=1}^{n}\langle X_{li}, B_{lk}\rangle^2 e_i^2 + \sum_{i'\neq i}\langle X_{li}, B_{lk}\rangle\langle X_{li'}, B_{lk}\rangle e_ie_{i'}\right]\right]\\ &\leq \sum_{k=1}^{M_n+d}\sum_{i=1}^{n} e_i^2\mathbb{E}[\langle X_{li}, B_{lk}\rangle^2] + \sum_{k=1}^{M_n+d}\sum_{i'\neq i} e_ie_{i'}\mathbb{E}[\langle X_{li}, B_{lk}\rangle\langle X_{li'}, B_{lk}\rangle]\\ &= O(M_n^{-2\delta}nM_n^{-1}) + O(M_n^{-2\delta}n(n-1)M_n^{-1})\\ &= O(n^2M_n^{-2\delta-1}).\end{aligned}$$

Therefore, by A.3, we have $\mathbb{E}[(\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{\epsilon} - \boldsymbol{e})^T] = O(n)$. By Markov inequality, $(\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{U}^T(\boldsymbol{\epsilon} - \boldsymbol{e})^T = O_p(n)$ and thus, $-(2\eta_n/n)(\boldsymbol{\epsilon} - \boldsymbol{e})^T\boldsymbol{U}\boldsymbol{v} \geq -2\eta_n||\boldsymbol{v}||_2 O_p(n^{-1/2})$. Therefore, the LHS gives

$$\text{LHS} \geq \eta_n^2 O_p(M_n^{-1}) - 2\eta_n O_p(n^{-1/2})||\boldsymbol{v}||_2. \tag{S2.5}$$

We denote the two terms of RHS of (S2.4) by T_1 and T_2. We first show the convergence rate of FDoS by assuming $w_l^{(1)} = w_l^{(2)} = 1$ for all l. With triangle inequality and $\Delta_n = |\mathcal{T}|/M_n$, and let $|\mathcal{A}|$ be the number of nonzero functions,

$$\begin{aligned} T_1 &\leq |\mathcal{T}| M_n^{-1} \lambda_1 \eta_n |\mathcal{A}| (J(M_n + d))^{1/2} ||\boldsymbol{v}||_2 \\ &= \eta_n O_p(\lambda_1 M_n^{-1/2}) ||\boldsymbol{v}||_2. \end{aligned} \tag{S2.6}$$

Suppose $\boldsymbol{x}$ is any vector and $\mathbf{A}$ is a symmetric matrix such that $g_{\mathbf{A}}(\boldsymbol{x}) = (\boldsymbol{x}^T \mathbf{A} \boldsymbol{x})^{1/2}$, g is a continuous function. By Taylor expansion, $g_{\boldsymbol{K}_{\varphi,l}}(\boldsymbol{\alpha}_l) - g_{\boldsymbol{K}_{\varphi,l}}(\boldsymbol{\alpha}_l + \eta_n \boldsymbol{v}) \leq -\eta_n \boldsymbol{v}^T \nabla g_{\boldsymbol{K}_{\varphi,l}}(\boldsymbol{\alpha}_l)$. By Lamma 3 and $\varphi = \lambda_2^2$, T_2 gives

$$\begin{aligned} T_2 &\leq -\lambda_2 |\mathcal{A}| \eta_n (\boldsymbol{\alpha}_l^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l)^{-1/2} ||\boldsymbol{v}||_2 ||\boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l||_2 \mathbb{I}(\boldsymbol{\alpha}_l \neq \mathbf{0}) \\ &= \eta_n O_p(\lambda_2^2) ||\boldsymbol{v}||_2. \end{aligned} \tag{S2.7}$$

Combining three inequalities (S2.5)-(S2.7), we have

$$\eta_n^2 O_p(M_n^{-1}) - 2\eta_n O_p(n^{-1/2}) ||\boldsymbol{v}||_2 \leq \eta_n O_p(\lambda_1 M_n^{-1/2}) ||\boldsymbol{v}||_2 + \eta_n O_p(\lambda_2^2) ||\boldsymbol{v}||_2.$$

For sufficient large constant C_6 such that $||\boldsymbol{v}||_2 = C_6$, we find $\eta_n = O_p(M_n n^{-1/2} + M_n^{1/2} \lambda_1 + M_n \lambda_2^2)$. When $\lambda_1 = O(M_n^{1/2} n^{-1/2})$ and $\lambda_2 = O(n^{-1/4})$, $\eta_n = O_p(M_n n^{-1/2})$. It means that for any given $\varepsilon > 0$, there always exists η_n such that

$$P\Big\{\exists \boldsymbol{v} \in \mathbb{R}^{M_n + d}, ||\boldsymbol{v}||_2 = C_6 : L_n(\boldsymbol{\alpha} + \eta_n \boldsymbol{v}) < L_n(\boldsymbol{\alpha})\Big\} \geq 1 - \varepsilon.$$

This further means that there is a local minimizer $\hat{\boldsymbol{b}} = \boldsymbol{\alpha} + \eta_n \boldsymbol{v}$, such that $||\hat{\boldsymbol{b}} - \boldsymbol{\alpha}||_2 = O_p(M_n n^{-1/2})$.

Therefore, by triangle inequality

$$
\begin{aligned}
||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}||_\infty &\leq ||\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\boldsymbol{\alpha}}||_\infty + ||\boldsymbol{\beta}^\alpha - \boldsymbol{\beta}||_\infty \\
&\leq \sup_t \sum_{k=1}^{M_n+d} |B_{lk}(t)| ||\hat{\boldsymbol{b}} - \boldsymbol{\alpha}||_\infty + ||\boldsymbol{\beta}^\alpha - \boldsymbol{\beta}||_\infty \\
&= O_p(M_n n^{-1/2}) + O(M_n^{-\delta}) = O_p(M_n n^{-1/2}).
\end{aligned}
$$

The last equation holds because of A.3.

The proof of convergence rate of FadDoS depends on the same reasoning before, except the step of T_1 and T_2. Let $\phi_1 = \sup_{l\in\mathcal{A}} ||\check{\beta}_l||_1^{-a}$ and $\phi_2 = \sup_{l\in\mathcal{A}} ||\check{\beta}_l||_2^{-a}$. When $\lambda_1\phi_1 = O(M_n^{1/2} n^{-1/2})$ and $\lambda_2^2\phi_2 = O(n^{-1/2})$, the results follow. □

S2.2 Proof of Theorem 2

Proof. Since the penalty terms of the objective function are separable, we assume other coefficient functions fixed and only consider the lth coefficient function here. The overall error can be divided into estimation error and approximation error as follows.

$$
(n/M_n)^{1/2}(\hat{\beta}_l - \beta_l) = (n/M_n)^{1/2}(\hat{\beta}_l - \beta_l^\alpha) + (n/M_n)^{1/2}(\beta_l^\alpha - \beta_l).
$$

Because the B-spline approximation error has been mentioned in Lemma 1, $(n/M_n)^{1/2}(\beta_l^\alpha - \beta_l) = O(n^{1/2} M_n^{-\delta-1/2})$, we only need to focus on the first term of RHS, $(n/M_n)^{1/2}(\hat{\beta}_l - \beta_l^\alpha) = (n/M_n)^{1/2}\boldsymbol{B}_l^T(\hat{\boldsymbol{b}}_l - \boldsymbol{\alpha}_l)$. Let $\boldsymbol{v} \in \mathbb{R}^{M_n+d}$

such that $\hat{\boldsymbol{b}}_l = \boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}$ and $\boldsymbol{r}_{(-l)} = \boldsymbol{Y} - \sum_{j\neq l}\boldsymbol{U}_j\boldsymbol{b}_j$, we define $Q_n(\boldsymbol{v})$ given (S2.3), assuming $w_l^{(1)} = w_l^{(2)} = 1$,

$$
\begin{aligned}
Q_n(\boldsymbol{v}) =& ||\boldsymbol{r}_{(-l)} - \boldsymbol{U}_l(\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v})||_2^2 + n\Delta_n\lambda_1||\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}||_1 + \\
& n\lambda_2((\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v})^T\boldsymbol{K}_{\varphi,l}(\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}))^{1/2}.
\end{aligned}
$$

Suppose the minimizer of $Q_n(\boldsymbol{v})$ is noted as $\hat{\boldsymbol{v}}_n$, then $\hat{\boldsymbol{v}}_n = (n/M_n)^{1/2}(\hat{\boldsymbol{b}}_l - \boldsymbol{\alpha}_l)$. We need to show the limiting distribution of $\hat{\boldsymbol{v}}_n$ by proving the finite distribution convergence of $V_{1(n)}^{(l)}$ to $V_1^{(l)}$. Note that $V_{1(n)}^{(l)}(\boldsymbol{v}) = Q_n(\boldsymbol{v}) - Q_n(\boldsymbol{0})$,

$$
\begin{aligned}
V_{1(n)}^{(l)}(\boldsymbol{v}) =& [\boldsymbol{v}^T(\frac{M_n}{n}\boldsymbol{U}_l^T\boldsymbol{U}_l)\boldsymbol{v} - 2(\frac{M_n}{n})^{1/2}(\boldsymbol{r}_{(-l)} - \boldsymbol{U}_l\boldsymbol{\alpha}_l)^T\boldsymbol{U}_l\boldsymbol{v}] + \\
& n\Delta_n\lambda_1\big(||\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}||_1 - ||\boldsymbol{\alpha}_l||_1\big) + \\
& n\lambda_2\big(((\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v})^T\boldsymbol{K}_{\varphi,l}(\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}))^{1/2} - (\boldsymbol{\alpha}_l^T\boldsymbol{K}_{\varphi,l}\boldsymbol{\alpha}_l)^{1/2}\big).
\end{aligned}
$$

Because $(M_n/n)\boldsymbol{U}_l^T\boldsymbol{U}_l \to \boldsymbol{C}_l$ and $\boldsymbol{W}_l \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{C}_l)$, the first term denoted by T_1 of RHS of $V_{1(n)}^{(l)}(\boldsymbol{v})$ is

$$
T_1 = \boldsymbol{v}^T\boldsymbol{C}_l\boldsymbol{v} - 2\boldsymbol{W}_l^T\boldsymbol{v},
$$

which is due to

$$
(M_n/n)^{1/2}(\boldsymbol{r}_{(-l)} - \boldsymbol{U}_l\boldsymbol{\alpha}_l)^T\boldsymbol{U}_l = (M_n/n)^{1/2}\boldsymbol{\epsilon}^T\boldsymbol{U}_l + (M_n/n)^{1/2}\boldsymbol{e}^T\boldsymbol{U}_l.
$$

We can see that $(M_n/n)^{1/2}\boldsymbol{\epsilon}^T\boldsymbol{U}_l = \boldsymbol{W}_l$ and $(M_n/n)^{1/2}\boldsymbol{e}^T\boldsymbol{U}_l = (M_n/n)^{1/2}O_p(M_n^{-\delta-1/2}n) = o_p(1)$. Furthermore, the second term denoted by T_2 of RHS is

$$
T_2 = \lambda_1(n/M_n)^{1/2}\sum_{k=1}^{M_n+d}\Big\{|v_k|\mathbb{I}(\alpha_{lk} = 0) + v_k\text{sgn}\big(\alpha_{lk}\big)\mathbb{I}(\alpha_{lk} \neq 0)\Big\}.
$$

By Lemma 3, the eigenvalues of $\boldsymbol{K}_{\varphi,l}$ is of the order $O(\varphi M_n^{-1})$, the third term denoted by T_3 of RHS gives

$$T_3 = \lambda_2 (M_n n)^{1/2} \left\{ (\boldsymbol{\alpha}_l^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l)^{-1/2} \boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l \mathbb{I}(\boldsymbol{\alpha}_l \neq \mathbf{0}) + (\boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{v})^{1/2} \mathbb{I}(\boldsymbol{\alpha}_l = \mathbf{0}) \right\} + o(M_n)\lambda_2$$

$$= \lambda_2^2 n^{1/2} \left\{ ||\boldsymbol{v}||_2 \mathbb{I}(\boldsymbol{\alpha}_l = \mathbf{0}) + (\boldsymbol{v}^T \boldsymbol{\alpha}_l / ||\boldsymbol{\alpha}_l||_2) \mathbb{I}(\boldsymbol{\alpha}_l \neq \mathbf{0}) \right\} + o(M_n)\lambda_2.$$

So, we have $V_{1(n)}^{(l)}(\boldsymbol{v}) \xrightarrow{p} V_1^{(l)}(\boldsymbol{v})$ for a fixed $\boldsymbol{v}$. $V_{1(n)}^{(l)}$ is a convex function and it follows the results of Geyer (1994) that $(n/M_n)^{1/2}(\hat{\boldsymbol{b}}_l - \boldsymbol{\alpha}_l) = \hat{\boldsymbol{v}}_n = \arg\min_{\boldsymbol{v}} V_{1(n)}^{(l)} \xrightarrow{p} \arg\min_{\boldsymbol{v}} V_1^{(l)}$ for $l = 1, \ldots, J$. Multipling $\boldsymbol{B}_l(t)$ on both sides obtains

$$(n/M_n)^{1/2}(\hat{\beta}_l(t) - \beta_l(t)) = (n/M_n)^{1/2}(\hat{\beta}_l(t) - \beta_l^{\alpha}(t)) + O(n^{1/2} M_n^{-\delta-1/2})$$
$$\xrightarrow{d} \boldsymbol{B}_l^T(t) \arg\min_{\boldsymbol{v}} V_1^{(l)}(\boldsymbol{v}).$$

□

S2.3 Proof of Proposition 1

Proof of 1. Since finding $\hat{\mathcal{A}}_n = \mathcal{A}$ is the intersection of correctly estimating nonzero values for nonzero coefficient function and correctly identifying zero coefficient function, $P(\hat{\mathcal{A}}_n = \mathcal{A}) \leq P(\hat{\boldsymbol{b}}_l = \mathbf{0}\ \forall l \notin \mathcal{A})$. For lth coefficient functions, let $\boldsymbol{v}_l^* = \arg\min_{\boldsymbol{v}} V_1^{(l)}(\boldsymbol{v}_l)$, Theorem 2 shows that $(n/M_n)^{1/2}(\hat{\boldsymbol{b}}_l - \boldsymbol{\alpha}_l) \xrightarrow{d} \boldsymbol{v}_l^*$. Therefore, we need to show that $c = P(\boldsymbol{v}_l^* = \mathbf{0}\ \forall l \notin \mathcal{A}) < 1$. There are two cases:

If $\gamma_1 = \gamma_2 = 0$, $\boldsymbol{v}_l^* = \boldsymbol{C}_l^{-1} \boldsymbol{W}_l \sim N(\mathbf{0}, \sigma^2 \boldsymbol{C}_l^{-1})$ and therefore $c = 0$.

If $\gamma_1 \neq 0$ or $\gamma_2 \neq 0$,

$$V_1^{(l)}(\boldsymbol{v}_l) = \begin{cases} \boldsymbol{v}_l^T \boldsymbol{C}_l \boldsymbol{v}_l - 2\boldsymbol{W}_l^T \boldsymbol{v}_l + \gamma_1 \Gamma_1^{(l)}(\boldsymbol{v}_l) + \gamma_2(\boldsymbol{v}_l^T \boldsymbol{\alpha}_l/||\boldsymbol{\alpha}_l||_2) & \text{if } l \in \mathcal{A} \\ \boldsymbol{v}_l^T \boldsymbol{C}_l \boldsymbol{v}_l - 2\boldsymbol{W}_l^T \boldsymbol{v}_l + \gamma_1 ||\boldsymbol{v}_l||_1 + \gamma_2 ||\boldsymbol{v}_l||_2 & \text{if } l \notin \mathcal{A}, \end{cases}$$

It can be seen that $V_1^{(l)}(\boldsymbol{v}_l)$ is not differentiable at $v_{lk} = 0$ for any k. By KKT conditions, $\boldsymbol{v}_l^*$ should satisfies

$$\begin{cases} 2\boldsymbol{C}_l \boldsymbol{v}_l^* - 2\boldsymbol{W}_l + \gamma_1 \boldsymbol{p}_l + \gamma_2 \boldsymbol{\alpha}_l/||\boldsymbol{\alpha}_l||_2 = 0 & \text{if } l \in \mathcal{A} \\ 2\boldsymbol{C}_l \boldsymbol{v}_l^* - 2\boldsymbol{W}_l + \gamma_1 \boldsymbol{q}_l + \gamma_2 \boldsymbol{z}_l = 0 & \text{if } l \notin \mathcal{A}, \end{cases} \tag{S2.8}$$

where $p_{lk} = \partial_{v_{lk}^*}\big\{|v_{lk}^*|\mathbb{I}(\alpha_{lk} = 0) + v_{lk}^* \text{sgn}\big(\alpha_{lk}\big)\mathbb{I}(\alpha_{lk} \neq 0)\big\}$,

$$q_{lk} = \begin{cases} \text{sgn}(v_{lk}^*) & \text{if } v_{lk}^* \neq 0 \\ \in \{q_{lk} : |q_{lk}| \leq 1\} & \text{if } v_{lk}^* = 0, \end{cases} \quad \boldsymbol{z}_l = \begin{cases} \boldsymbol{v}_l^*/||\boldsymbol{v}_l^*||_2 & \text{if } \boldsymbol{v}_l^* \neq \boldsymbol{0} \\ \in \{\boldsymbol{z}_l : ||\boldsymbol{z}_l||_2 \leq 1\} & \text{if } \boldsymbol{v}_l^* = \boldsymbol{0}. \end{cases}$$

To examine the variable selection consistency, we have to introduce new denotations by combining all coefficient functions. We let $\boldsymbol{C} = (M_n/n)\boldsymbol{U}^T\boldsymbol{U}$, $\boldsymbol{U} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_J)^T$, $\boldsymbol{W} = (\boldsymbol{W}_1, \ldots, \boldsymbol{W}_J)^T$, $\boldsymbol{p} = (\boldsymbol{p}_1, \ldots, \boldsymbol{p}_J)^T$, $\boldsymbol{q} = (\boldsymbol{q}_1, \ldots, \boldsymbol{q}_J)^T$, $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_J)^T$, and $\boldsymbol{v}^* = (\boldsymbol{v}_1^*, \ldots, \boldsymbol{v}_J^*)^T$. Without loss of generality, rewrite matrix $\boldsymbol{C}$ in a block-wise form involving either $l \in \mathcal{A}$ or $\mathcal{A}^c$ such that

$$\boldsymbol{C} = \begin{bmatrix} \boldsymbol{C}_{\mathcal{A}\mathcal{A}} & \boldsymbol{C}_{\mathcal{A}\mathcal{A}^c} \\ \boldsymbol{C}_{\mathcal{A}^c\mathcal{A}} & \boldsymbol{C}_{\mathcal{A}^c\mathcal{A}^c} \end{bmatrix},$$

and rewrite vectors $\boldsymbol{W}$, $\boldsymbol{p}$, $\boldsymbol{q}$, $\boldsymbol{\alpha}$, and $\boldsymbol{v}_l^*$ likewise. If $\boldsymbol{v}_l^* = \boldsymbol{0}$ for any $l \notin \mathcal{A}$, the optimality conditions (S2.8) become

$$\begin{cases} 2\boldsymbol{C}_{\mathcal{A}\mathcal{A}}\boldsymbol{v}_{\mathcal{A}}^* - 2\boldsymbol{W}_{\mathcal{A}} + \gamma_1\boldsymbol{p}_{\mathcal{A}} + \gamma_2\boldsymbol{\alpha}_{\mathcal{A}}/||\boldsymbol{\alpha}_{\mathcal{A}}||_2 = 0 \\ ||2\boldsymbol{C}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{v}_{\mathcal{A}}^* - 2\boldsymbol{W}_{\mathcal{A}^c} + \gamma_1\boldsymbol{q}_{\mathcal{A}^c}||_2 \leq \gamma_2. \end{cases} \tag{S2.9}$$

Combining these optimality conditions (S2.9) above,

$$\left|\left|\boldsymbol{C}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{C}_{\mathcal{A}\mathcal{A}}^{-1}(2\boldsymbol{W}_{\mathcal{A}} - \gamma_1\boldsymbol{p}_{\mathcal{A}} - \gamma_2\boldsymbol{\alpha}_{\mathcal{A}}/||\boldsymbol{\alpha}_{\mathcal{A}}||_2) - 2\boldsymbol{W}_{\mathcal{A}^c} + \gamma_1\boldsymbol{q}_{\mathcal{A}^c}\right|\right|_2 \leq \gamma_2.$$

Thus, we obtain,

$$c \leq P\left\{\left|\left|\boldsymbol{C}_{\mathcal{A}^c\mathcal{A}}\boldsymbol{C}_{\mathcal{A}\mathcal{A}}^{-1}(2\boldsymbol{W}_{\mathcal{A}} - \gamma_1\boldsymbol{p}_{\mathcal{A}} - \gamma_2\boldsymbol{\alpha}_{\mathcal{A}}/||\boldsymbol{\alpha}_{\mathcal{A}}||_2) - 2\boldsymbol{W}_{\mathcal{A}^c} + \gamma_1\boldsymbol{q}_{\mathcal{A}^c}\right|\right|_2 \leq \gamma_2\right\} < 1.$$

□

Proof of 2. Given the same reasoning of 1, $P(\hat{\beta}_l(t) = 0) \leq P(\hat{b}_{lk} = \boldsymbol{0}\ \forall k \notin \mathcal{B})$. We need to show that $c = P(v_{lk}^* = \boldsymbol{0}\ \forall k \notin \mathcal{B}) < 1$. There are also two cases:

If $\gamma_1 = \gamma_2 = 0$, $\boldsymbol{v}_l^* = \boldsymbol{C}_l^{-1}\boldsymbol{W}_l \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{C}_l^{-1})$ and therefore $c = 0$.

If $\gamma_1 = \gamma_2 = 0$, for $l \in \mathcal{A}$,

$$\begin{aligned} V_1^{(l)}(\boldsymbol{v}_l) =& \boldsymbol{v}_l^T\boldsymbol{C}_l\boldsymbol{v}_l - 2\boldsymbol{W}_l\boldsymbol{v}_l + \gamma_1\Gamma_1^{(l)}(\boldsymbol{v}_l) + \gamma_2\boldsymbol{v}_l^T\boldsymbol{\alpha}_{\mathcal{A}}/||\boldsymbol{\alpha}_{\mathcal{A}}||_2 \\ =& \begin{cases} \boldsymbol{v}_l^T\boldsymbol{C}_l\boldsymbol{v}_l - 2\boldsymbol{W}_l\boldsymbol{v}_l + \gamma_1 v_{lk}\text{sgn}\big(\alpha_{lk}\big) + \gamma_2\boldsymbol{v}_l^T\boldsymbol{\alpha}_{\mathcal{A}}/||\boldsymbol{\alpha}_{\mathcal{A}}||_2 & \text{if } k \in \mathcal{B} \\ \boldsymbol{v}_l^T\boldsymbol{C}_l\boldsymbol{v}_l - 2\boldsymbol{W}_l\boldsymbol{v}_l + \gamma_1|v_{lk}| + \gamma_2\boldsymbol{v}_l^T\boldsymbol{\alpha}_{\mathcal{A}}/||\boldsymbol{\alpha}_{\mathcal{A}}||_2 & \text{if } k \notin \mathcal{B}. \end{cases} \end{aligned}$$

We can follow the KKT conditions with respect to individual v_{lk}, and then depend on the similar reasoning of the *Proof of 1.* to obtain the results. □

S2.4 Proof of Theorem 3

Proof. The proof of Theorem 3 requires Lemma 4, which shows the rate of convergence of the initial estimator which is derived from the penalized B-splines estimator in Cardot et al. (2003). Let λ be the tuning parameter for the functional generalization of ridge regularization.

Lemma 4. *Under* (A.1)-(A.3), *if* $\lambda = O(M_n n^{-1/2})$, $||\hat{\hat{\beta}}_l - \beta_l||_\infty = O_p(M_n n^{-1/2})$.

Proof. The proof depends on the identical procedure as Theorem 1.

Now we are ready to give the proof of Theorem 3. Based on (S2.3), we define $Q_n(\boldsymbol{v})$, $\boldsymbol{v} \in \mathbb{R}^{M_n+d}$ by adding adaptive weights as following

$$Q_n(\boldsymbol{v}) = ||\boldsymbol{r}_{(-l)} - \boldsymbol{U}_l(\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v})||_2^2 + n\Delta_n\lambda_1\hat{w}_l^{(1)}||\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}||_1 + n\lambda_2\hat{w}_l^{(2)}((\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v})^T\boldsymbol{K}_{\varphi,l}(\boldsymbol{\alpha}_l + (M_n/n)^{1/2}\boldsymbol{v}))^{1/2}.$$

Suppose the minimizer of $Q_n(\boldsymbol{v})$ is noted as $\hat{\boldsymbol{v}}_n$, then $\hat{\boldsymbol{v}}_n = (n/M_n)^{1/2}(\hat{\boldsymbol{b}}_l - \boldsymbol{\alpha}_l)$. We need to show the limiting distribution of $\hat{\boldsymbol{v}}_n$. Similarly, $V_{2(n)}^{(l)}(\boldsymbol{v}) =$

$Q_n(\boldsymbol{v}) - Q_n(\boldsymbol{0}) = \boldsymbol{v}^T \boldsymbol{C}_l \boldsymbol{v} - 2\boldsymbol{W}_l^T \boldsymbol{v} + \Gamma_{1,2}(\boldsymbol{v})$, where

$$\Gamma_{1,2}(\boldsymbol{v}) = \begin{cases} \lambda_1 (n/M_n)^{1/2} \hat{w}_l^{(1)} \sum_{k=1}^{M_n+d} \Big\{ |v_k| \mathbb{I}(\alpha_k = 0) + v_k \text{sgn}\big(\alpha_{lk}\big) \mathbb{I}(\alpha_{lk} \neq 0) \Big\} + & \\ \lambda_2 (M_n n)^{1/2} \hat{w}_l^{(2)} (\boldsymbol{\alpha}_l^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l)^{-1/2} \boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l & \text{if } l \in \mathcal{A} \\ \lambda_1 (n/M_n)^{1/2} \hat{w}_l^{(1)} ||\boldsymbol{v}||_1 + \lambda_2 (M_n n)^{1/2} \hat{w}_l^{(2)} (\boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{v})^{1/2} & \text{if } l \notin \mathcal{A}, \end{cases}$$

and $(M_n/n)\boldsymbol{U}_l^T \boldsymbol{U}_l \to \boldsymbol{C}_l$, $\boldsymbol{W}_l = N(\boldsymbol{0}, \sigma^2 \boldsymbol{C}_l)$ as before.

If $l \notin \mathcal{A}$, $\hat{w}_l^{(1)} = ||\hat{\tilde{\beta}}||_1^{-a} = O_p(M_n^{-a} n^{a/2})$ and $\hat{w}_l^{(2)} = ||\hat{\tilde{\beta}}||_2^{-a} = O_p(M_n^{-a} n^{a/2})$,

$$\lambda_1 (\frac{n}{M_n})^{1/2} \hat{w}_l^{(1)} ||\boldsymbol{v}||_1 \propto \lambda_1 (\frac{n}{M_n})^{1/2} \frac{n^{a/2}}{M_n^a} \hat{w}_l^{(1)} \frac{M_n^a}{n^{a/2}} ||\boldsymbol{v}||_1 = \frac{\lambda_1 n^{(a+1)/2}}{M_n^{a+1/2}} \hat{w}_l^{(1)} \frac{M_n^a}{n^{a/2}} ||\boldsymbol{v}||_1 \xrightarrow{p} \infty,$$

since $\lambda_1 n^{(a+1)/2} / M_n^{a+1/2} \to \infty$ and $(M_n^a / n^{a/2}) \hat{w}_l^{(1)} = O_p(1)$; on the other hand, given Lemma 3 and $\varphi = \lambda_2^2$,

$$\lambda_2 (M_n n)^{1/2} \hat{w}_l^{(2)} (\boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{v})^{1/2} = \lambda_2^2 n^{1/2} \frac{n^{a/2}}{M_n^a} \hat{w}_l^{(2)} \frac{M_n^a}{n^{a/2}} = \frac{\lambda_2^2 n^{(a+1)/2}}{M_n^a} \hat{w}_l^{(2)} \frac{M_n^a}{n^{a/2}} \xrightarrow{p} \infty,$$

since $\lambda_2^2 n^{(a+1)/2} / M_n^a \to \infty$ and $(M_n^a / n^{a/2}) \hat{w}_l^{(2)} = O_p(1)$. Hence, for $l \notin \mathcal{A}$

$$V_{2(n)}^{(l)}(\boldsymbol{v}) = \infty. \tag{S2.10}$$

If $l \in \mathcal{A}$ and $k \notin \mathcal{B}$, we have

$$V_{2(n)}^{(l)}(\boldsymbol{v}) = \boldsymbol{v}^T \boldsymbol{C}_l \boldsymbol{v} - 2\boldsymbol{W}_l^T \boldsymbol{v} + \lambda_1 (\frac{n}{M_n})^{1/2} \hat{w}_l^{(1)} |v_k| + \lambda_2 (M_n n)^{1/2} \hat{w}_l^{(2)} (\boldsymbol{\alpha}_l^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l)^{-1/2} \boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l.$$

Under the same conditions before, we know that $\lambda_1 (\frac{n}{M_n})^{1/2} \hat{w}_l^{(1)} |v_k| \xrightarrow{p} \infty$ and $\lambda_2 (M_n n)^{1/2} \hat{w}_l^{(2)} (\boldsymbol{\alpha}_l^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l)^{-1/2} \boldsymbol{v}^T \boldsymbol{K}_{\varphi,l} \boldsymbol{\alpha}_l \xrightarrow{p} \infty$. We re-expressed $\boldsymbol{v}$ in a blockwise form mentioned before such that $\boldsymbol{v} = (\boldsymbol{v}_{\mathcal{B}}, \boldsymbol{v}_{\mathcal{B}^c})^T$. Hence,

$$V_{2(n)}^{(l)}(\boldsymbol{v}_{\mathcal{B}^c}) = \infty. \tag{S2.11}$$

If $l \in \mathcal{A}$ and $k \in \mathcal{B}$, due to $t \in I_0(\beta_l)$, $\hat{\tilde{\beta}}_l(t) \xrightarrow{p} \beta_l(t) \neq 0$, the last two terms converge to zero in probability under the conditions $\lambda_1(n/M_n)^{1/2} \to 0$ and $\lambda_2^2 n^{1/4} \to 0$. Thus,

$$V_{2(n)}^{(l)}(\boldsymbol{v}_{\mathcal{B}}) = \boldsymbol{v}_{\mathcal{B}}^T(\boldsymbol{C}_l)_{\mathcal{B}\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} - 2(\boldsymbol{W}_l)_{\mathcal{B}}^T\boldsymbol{v}_{\mathcal{B}}. \tag{S2.12}$$

Thus, we see that $V_{2(n)}^{(l)}(\boldsymbol{v}) \xrightarrow{p} V_2^{(l)}(\boldsymbol{v})$ for a fixed $\boldsymbol{v}$ and $V_{2(n)}^{(l)}$ is a convex function. Given (S2.10), (S2.11), and (S2.12), it follows the results of Geyer (1994) that $\hat{\boldsymbol{v}}_n^{(l)} = \arg\min_{\boldsymbol{v}} V_{2(n)}^{(l)} \xrightarrow{p} \arg\min_{\boldsymbol{v}} V_2^{(l)}$, where

$$V_2^{(l)}(\boldsymbol{v}) = \begin{cases} \boldsymbol{v}_{\mathcal{B}}(\boldsymbol{C}_l)_{\mathcal{B}\mathcal{B}}\boldsymbol{v}_{\mathcal{B}} - 2(\boldsymbol{W}_l)_{\mathcal{B}}^T\boldsymbol{v}_{\mathcal{B}} & \text{if } l \in \mathcal{A} \text{ and } k \in \mathcal{B} \\ \infty & \text{if } l \in \mathcal{A} \text{ and } k \notin \mathcal{B} \\ \infty & \text{if } l \notin \mathcal{A}, \end{cases}$$

and $(\boldsymbol{C}_l)_{\mathcal{B}\mathcal{B}}$, $(\boldsymbol{W}_l)_{\mathcal{B}}$ are similarly defined. Therefore, we have $\hat{\boldsymbol{v}}_n \xrightarrow{d} (\boldsymbol{C}_l)_{\mathcal{B}\mathcal{B}}^{-1}(\boldsymbol{W}_l)_{\mathcal{B}}$ if $l \in \mathcal{A}$ and $k \in \mathcal{B}$; $\hat{\boldsymbol{v}}_n \xrightarrow{d} \mathbf{0}$ if $l \in \mathcal{A}$ and $k \notin \mathcal{B}$; $\hat{\boldsymbol{v}}_n \xrightarrow{d} \mathbf{0}$ if $l \notin \mathcal{A}$. Since $(\boldsymbol{W}_l)_{\mathcal{B}} \sim N(\mathbf{0}, \sigma^2(\boldsymbol{C}_l)_{\mathcal{B}\mathcal{B}})$, it can be seen that for $t \in I_1(\beta_l)$,

$$\begin{aligned} (n/M_n)^{1/2}(\hat{\beta}_l(t) - \beta_l(t)) &= (n/M_n)^{1/2}(\hat{\beta}_l(t) - \beta_l^{\alpha}(t)) + (n/M_n)^{1/2}(\beta_l^{\alpha}(t) - \beta_l(t)) \\ &= (n/M_n)^{1/2}\boldsymbol{B}_l^T(\hat{\boldsymbol{b}}_{l\mathcal{B}} - \boldsymbol{\alpha}_{l\mathcal{B}}) + O(n^{1/2}M_n^{-\delta-1/2}) \\ &\xrightarrow{d} \boldsymbol{B}_l^T(\boldsymbol{C}_l)_{\mathcal{B}\mathcal{B}}^{-1}(\boldsymbol{W}_l)_{\mathcal{B}} \\ &\xrightarrow{d} N(0, \Sigma_{lt}), \end{aligned}$$

where $\Sigma_{lt} = \sigma^2 \boldsymbol{B}_l^T(t)(\boldsymbol{C}_l)^{-1}_{\mathcal{BB}}\boldsymbol{B}_l(t)$.

We now want to prove the consistency of global and local selection. We start with the global variable selection. The asymptotic normality indicates that $P(l \in \hat{\mathcal{A}}) \to 1$. Then it suffices to show that for any $l' \notin \mathcal{A}$, $P(l' \in \hat{\mathcal{A}}) \to 0$. We rewrite the objective function for $\boldsymbol{b}_{l'}$, resulting in $L_n(\boldsymbol{b}_{l'}) = n^{-1}||\boldsymbol{r}_{(-l')} - \boldsymbol{U}_{l'}\boldsymbol{b}_{l'}||_2^2 + \Delta_n\lambda_1\hat{w}_{l'}^{(1)}||\boldsymbol{b}_{l'}||_1 + \lambda_2\hat{w}_{l'}^{(2)}(\boldsymbol{b}_{l'}^T\boldsymbol{K}_{\varphi,l'}\boldsymbol{b}_{l'})^{1/2}$. By Lemma 3 that eigenvalues of $\boldsymbol{K}_{\varphi,l}$ is of the order $O(\varphi M_n^{-1})$, we can derive the KKT conditions for global and local selection respectively.

If $l' \in \hat{\mathcal{A}}$, under KKT conditions, we have

$$2\boldsymbol{U}_{l'}^T(\boldsymbol{r}_{(-l')} - \boldsymbol{U}_{l'}\boldsymbol{b}_{l'}) = n\Delta_n\lambda_1\hat{w}_{l'}^{(1)}\boldsymbol{q}_{l'} + nM_n^{-1/2}\lambda_2^2\hat{w}_{l'}^{(2)}\boldsymbol{b}_{l'}/||\boldsymbol{b}_{l'}||_2 \quad \text{(S2.13)}$$

where $q_{l'k}$ is defined similary as above with respect to $b_{l'k}$ here. Multiplying $(M_n/n)^{1/2}$ on both sides, the LHS gives

$$2(\frac{M_n}{n})^{1/2}\boldsymbol{U}_{l'}^T(\boldsymbol{r}_{(-l')} - \boldsymbol{U}_{l'}\boldsymbol{b}_{l'}) = 2\boldsymbol{W}_{l'} + 2(\frac{M_n}{n})\boldsymbol{U}_{l'}^T\boldsymbol{U}_{l'}(\frac{n}{M_n})^{1/2}(\boldsymbol{\alpha}_{l'} - \boldsymbol{b}_{l'}).$$

Because $\boldsymbol{W}_{l'} \sim N(\boldsymbol{0}, \sigma^2\boldsymbol{C}_{l'})$ and the second term asymptotically converges to normal distribution, $2(M_n/n)^{1/2}\boldsymbol{U}_{l'}^T(\boldsymbol{r}_{(-l')} - \boldsymbol{U}_{l'}\boldsymbol{b}_{l'}) \xrightarrow{d}$ some normal distribution by Slutsky's theorem. Given $(n/M_n)^{1/2}\lambda_1\hat{w}_l^{(1)} \xrightarrow{p} \infty$ and $n^{1/2}\lambda_2^2\hat{w}_l^{(2)} \xrightarrow{p} \infty$ under the conditions, we know the probability that (S2.13) holds tends to be 0, and thus, $P(l' \in \hat{\mathcal{A}}) \to 0$.

Now we want to prove the consistency of local selection only considering

$l \in \mathcal{A}$ and $k \in \mathcal{B}$. Similarly, it suffices to show that for any $k' \notin \mathcal{B}$, $P(k' \in \hat{\mathcal{B}} | l \in \mathcal{A}) \to 0$.

If $k' \in \hat{\mathcal{B}}$, under KKT conditions, we have

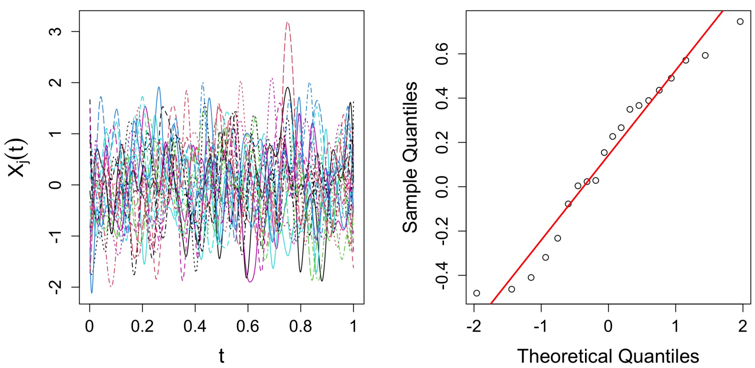
$$2\boldsymbol{U}_l^T(\boldsymbol{r}_{(-l)} - \boldsymbol{U}_l\boldsymbol{b}_l)I_{k'} = n\Delta_n\lambda_1\hat{w}_l^{(1)}\text{sgn}(b_{lk'}) + nM_n^{-1/2}\lambda_2^2\hat{w}_l^{(2)}b_{lk'}/||\boldsymbol{b}_l||_2, \tag{S2.14}$$

where $I_{k'}$ denotes (M_n+d)-dimensional unit vector with 1 at the k'th entry. Following a similar reasoning as above, we show the probability that (S2.14) holds tends to be 0 as well. Hence, $P(k' \in \hat{\mathcal{B}} | l \in \hat{\mathcal{A}}) \to 0$. $\square$

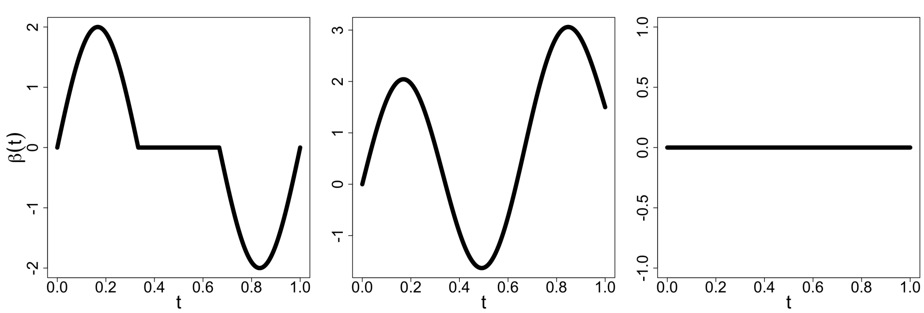
S3 Simulation Studies

S3.1 Data Generation

Figure S1 illustrates the simulated data and three different types of coefficient functions. The functional predictors are generated to be standardized



(a)



(b)

Figure S1: Generating functional predictor, response, and coefficient functions in simulations study: (a) The left panel is generated $X_{ij}(t), j = 1, \ldots, 10$ for all samples. Each curve represents the functional covariate for a subject over the time domain. The right panel is QQ-plot of generated Y_i. Both figures are drew under the sample size $n = 20$; (b) True coefficient functions, from left to right panel, are $\beta_1(t)$, $\beta_2(t)$, and $\beta_j(t), j = 3, \ldots, 10$.

and independent, which ensures a fair assessment of method performance, without confounding factors of differing scales or inter-predictor dependencies. Additionally, the three types of coefficient functions demonstrate the global and local sparsity structure in the functional linear regression model.

S3.2 Effects of the Smoothness Parameters

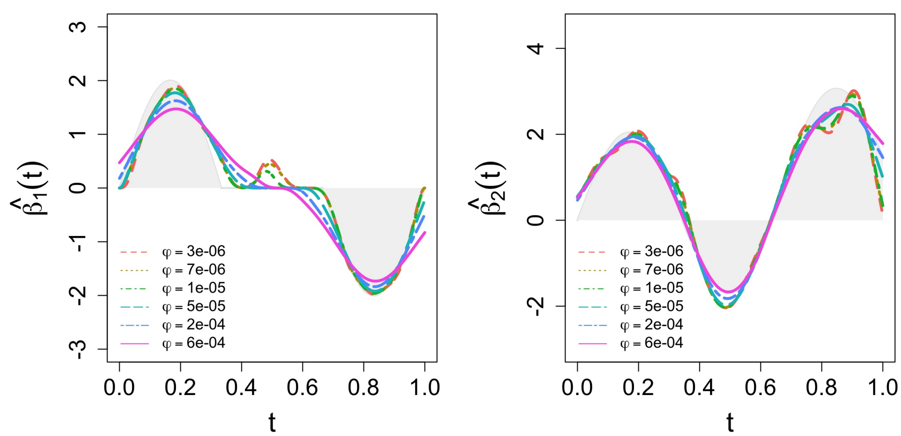


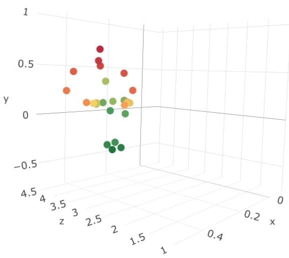
Figure S2: Estimated coefficient functions for $\beta_1(t)$ and $\beta_2(t)$ of the proposed estimator FadDoS with varying φ.

While the functional $\ell_{1,2}$ penalty induces global sparsity, the smoothing parameter φ controls overall curvature to ensure estimate smoothness. Figure S2 illustrates the effect of varying φ with λ_1 and λ_2 fixed. Insufficiently small φ leads to excessively wiggly functional estimates compared to the true functions. This impedes accurate identification of zero subregions and decreases the TNR due to inadequate shrinkage. Conversely, excessively large φ overly linearizes the estimates. As shown in Table S1, moderate values of φ around 5e-5 yield superior PMSE and ISE. The optimal level of smoothing avoids both under- and over-smoothing, thereby enabling precise estimation of zero subregions while maintaining estimate accuracy. Through controlling total curvature, the smoothing parameter provides localized regularization that complements the sparsity induced by the functional $\ell_{1,2}$ penalty.

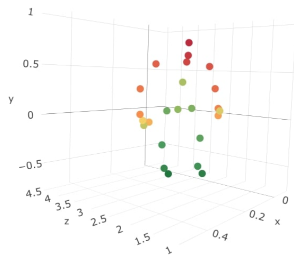
φ	PMSE ($\times 10^{-2}$)	$\text{ISE}_0(\hat{\beta}_1)$	$\text{ISE}_1(\hat{\beta}_1)$	$\text{ISE}(\hat{\beta}_2)$	$\sum_{j=3}^{10}\text{ISE}(\hat{\beta}_j)$	avgTNR
3e-6	2.59(0.18)	52.87(53.87)	103.83(48.22)	99.52(36.11)	0.37(3.29)	0.99(0.0176)
7e-6	2.54(0.17)	32.56(41.13)	85.29(44.36)	82.69(33.41)	0.27(2.68)	0.99(0.13)
1e-5	2.52(0.17)	26.38(37.95)	78.68(41.63)	76.37(32.58)	0.25(2.49)	0.99(0.13)
5e-5	2.46(0.17)	12.41(18.06)	64.67(40.28)	56.78(28.58)	0.00(0.00)	1.00(0.00)
2e-4	2.50(0.18)	28.48(24.42)	95.83(49.68)	53.38(29.61)	0.00(0.00)	1.00(0.00)
6e-4	2.64(0.20)	69.21(38.84)	173.73(60.80)	75.7(40.58)	0.00(0.00)	1.00(0.00)

Table S1: PMSE ($\times 10^{-2}$), ISE, and average TNR (avgTNR) of the proposed estimator FadDoS with varying φ. The test sample size is 1000. The entry in the parenthesis corresponds to the standard deviation among 100 simulation replicates.

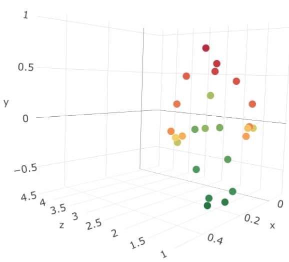
S4 Illustration of Timed Up and Go Test



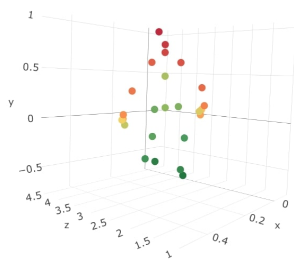
(a)



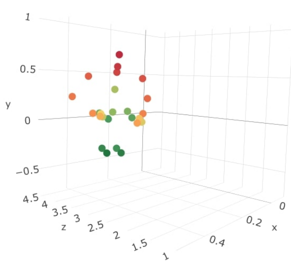
(b)



(c)



(d)



(e)

Figure S3: Instruction of Timed Up and Go (TUG) Test : (a) Stand up from the chair; (b) Walk forward at a normal pace; (c) Turn; (d) Walk backward to the chair at a normal pace; (e) Sit down. Each joint is color coded as Figure 1(b).

Bibliography

Cardot, H., F. Ferraty, and P. Sarda (2003). Spline estimators for the functional linear model. *Statistica Sinica 13*, 571–791.

Geyer, C. (1994). On the asymptotics of constrained m-estimation. *The Annals of Statistics 22*, 1993–2010.

Huang, J. Z., C. O. Wu, and L. Zhou (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica 14*, 763–788.

Wang, X. and X. Yuan (2012). The linearized alternating direction method for dantzig selector. *SIAM Journal on Scientific Computing 34*, A2792–A2811.

Zhou, J., N. Wang, and N. Wang (2013). Functional linear model with zero-value coefficient function at sub-regions. *Statistica Sinica 23*(1), 25–50.