Functional Sparsity: Global versus Local

Haonan Wang and Bo Kai

Colorado State University and College of Charleston

Supplementary Material

This supplementary material includes proofs of the theoretical results and additional discussion of our proposed method. The proofs of Lemma 1, Theorem 1 and Theorem 2 are given in Sections S1-S3, respectively. In Section S4, we provide additional discussion on estimating globally and locally constant functions by modifying the proposed method.

Appendix

We first introduce some notation. For a real valued function f over an interval I, $\|f\|_\infty = \sup_x |f(x)|$. Given two sequences of positive numbers, say α_n and β_n, if both α_n/β_n and β_n/α_n are bounded, we denote it by $\alpha_n \sim \beta_n$.

In addition, we will use the following inequalities in the proofs:

(i) for any p-dimensional vector

$$\|\boldsymbol{\gamma}\|_2 \leq \|\boldsymbol{\gamma}\|_1 \leq p^{1/2}\|\boldsymbol{\gamma}\|_2; \tag{S.1}$$

(ii) for any $a > 0$, $b > 0$, and $\alpha \in (0, 1)$,

$$|b^\alpha - a^\alpha| \leq 2a^{\alpha-1}|b - a|; \tag{S.2}$$

(iii) for any $0 \leq a \leq b$ and $\alpha \in (0, 1)$,

$$\alpha b^{\alpha-1}(b - a) \leq b^\alpha - a^\alpha. \tag{S.3}$$

S1. Proof of Lemma 1

Proof of Lemma 1. Note that

$$\|f(x) - \boldsymbol{B}^T(x)\boldsymbol{\gamma}_0\|_\infty \leq \|f(x) - \boldsymbol{B}(x)^T\boldsymbol{\gamma}^*\|_\infty + \|\boldsymbol{B}(x)^T(\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0)\|_\infty.$$

By (3.1), the first term on the right hand side is of order M_n^{-r}. In addition, by the properties of B-splines, we have that

$$\|\boldsymbol{B}(x)^T(\boldsymbol{\gamma}^* - \boldsymbol{\gamma}_0)\|_\infty \leq \max_k |\gamma_k^* - \gamma_{0,k}|,$$

where $\boldsymbol{\gamma}_0 = (\gamma_{0,1}, \ldots, \gamma_{0,L_n})^T$. Since $\boldsymbol{\gamma}_0$ is a sparse modification of $\boldsymbol{\gamma}^*$, if $\gamma_k^* \neq \gamma_{0,k}$ for some k, $|\gamma_k^* - \gamma_{0,k}| = |\gamma_k^*|$, and there exists an index j, $k - d \leq j \leq k$, such that $k \in A_j$ and $|f(x)| \leq DM_n^{-r}$ for $x \in [\kappa_{j-1}, \kappa_j]$. As a direct consequence of (3.1), $\boldsymbol{B}(x)^T\boldsymbol{\gamma}^*$ is of order M_n^{-r} for $x \in [\kappa_{j-1}, \kappa_j]$. In addition, by the local support property of B-splines, at most $d+1$ $\gamma's$ are non-zero in $\boldsymbol{B}(x)^T\boldsymbol{\gamma}^*$ for $x \in [\kappa_{j-1}, \kappa_j]$ and γ_k^* is one of them. The desired result immediately follows. □

S2. Proof of Theorem 1

The following lemma will be used in the proof of Theorem 1. It is from Lemma A.3 of Huang et al. (2004).

Lemma 2. *If* $\lim_{n\to\infty} M_n \log M_n/n = 0$*, there exists an interval* $[C_1, C_2]$ *(*$0 < C_1 < C_2 < \infty$*) such that all eigenvalues of* $(M_n/n)\boldsymbol{B}^T\boldsymbol{B}$ *fall in* $[C_1, C_2]$ *with probability approaching* 1 *as* $n \to \infty$*.*

Proof of Theorem 1. It can be seen that

$$\|\widehat{f} - f\|_2^2 \leq 2\|\boldsymbol{B}(x)^T\widehat{\boldsymbol{\gamma}} - \boldsymbol{B}(x)^T\boldsymbol{\gamma}_0\|_2^2 + 2\|f(x) - \boldsymbol{B}(x)^T\boldsymbol{\gamma}_0\|_2^2.$$

By Lemma 1, $\|f(x) - \boldsymbol{B}(x)^T\boldsymbol{\gamma}_0\|_2 = O(M_n^{-r}) = O(n^{-r/(2r+1)})$. By the properties of B-spline basis functions, we have

$$\|\boldsymbol{B}(x)^T\widehat{\boldsymbol{\gamma}} - \boldsymbol{B}(x)^T\boldsymbol{\gamma}_0\|_2^2 \sim M_n^{-1}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2.$$

Let $\widehat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}_0 + \eta_n\boldsymbol{v}$, where η_n is a scalar and $\boldsymbol{v}$ is a vector with $||\boldsymbol{v}||_2 = 1$. Therefore, it is sufficient to show that $\eta_n = O_p(M_n n^{-1/2})$.

Denote $D_n(\boldsymbol{v}) = Q_n(\boldsymbol{\gamma}_0 + \eta_n\boldsymbol{v}) - Q_n(\boldsymbol{\gamma}_0)$. In fact, $D_n(\boldsymbol{v})$ can be expressed as the sum of the following two quantities:

$$n^{-1}\sum_{i=1}^n \left(Y_i - \boldsymbol{B}(X_i)^T\widehat{\boldsymbol{\gamma}}\right)^2 - n^{-1}\sum_{i=1}^n \left(Y_i - \boldsymbol{B}(X_i)^T\boldsymbol{\gamma}_0\right)^2$$

and $\lambda_n \sum_{j=1}^{M_n+1} \|\widehat{\boldsymbol{\gamma}}_{A_j}\|_1^\alpha - \lambda_n \sum_{j=1}^{M_n+1} \|\boldsymbol{\gamma}_{0,A_j}\|_1^\alpha$, denoted by T_1 and T_2 respectively.

Let $r_i = \boldsymbol{B}(X_i)^T\boldsymbol{\gamma}_0 - f(X_i)$, $\boldsymbol{r} = (r_1, \ldots, r_n)^T$, $\varepsilon_i = Y_i - f(X_i)$ and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)^T$. Straightforward calculation shows that

$$\begin{aligned} T_1 &= \frac{1}{n}\sum_{i=1}^{n}\left(\varepsilon_i - r_i - \eta_n \boldsymbol{B}(X_i)^T\boldsymbol{v}\right)^2 - \frac{1}{n}\sum_{i=1}^{n}(\varepsilon_i - r_i)^2 \\ &= \frac{\eta_n^2}{n}\boldsymbol{v}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{v} - 2\frac{\eta_n}{n}(\boldsymbol{\varepsilon} - \boldsymbol{r})^T\boldsymbol{B}\boldsymbol{v}. \end{aligned}$$

Note that, as a direct consequence of Lemma 2, we have that, with probability approaching 1,

$$\frac{M_n}{n}\boldsymbol{v}^T\boldsymbol{B}^T\boldsymbol{B}\boldsymbol{v} \geq C_1.$$

Moreover, by Cauchy-Schwarz inequality, we have

$$\sup_{||\boldsymbol{v}||_2=1}\{(\boldsymbol{\varepsilon} - \boldsymbol{r})^T\boldsymbol{B}\boldsymbol{v}\}^2 \leq (\boldsymbol{\varepsilon} - \boldsymbol{r})^T\boldsymbol{B}\boldsymbol{B}^T(\boldsymbol{\varepsilon} - \boldsymbol{r}).$$

In addition, by the independence of $\boldsymbol{\varepsilon}$ and $\boldsymbol{r}$, we have

$$E(\boldsymbol{\varepsilon} - \boldsymbol{r})^T\boldsymbol{B}\boldsymbol{B}^T(\boldsymbol{\varepsilon} - \boldsymbol{r}) = E\boldsymbol{\varepsilon}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\varepsilon} + E\boldsymbol{r}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{r}. \tag{S.4}$$

For the first term in (S.4), we have

$$E\boldsymbol{\varepsilon}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{\varepsilon} = \sigma^2\sum_{k=1}^{L_n}\sum_{i=1}^{n}E\left\{B_k(X_i)^2\right\} \leq \sigma^2\sum_{k=1}^{L_n}\sum_{i=1}^{n}E\left\{B_k(X_i)\right\} = O(n).$$

Recall that, by Lemma 1, r_i are of the order $O(M_n^{-r})$. So, for the second term in (S.4), we have

$$\begin{aligned} E\boldsymbol{r}^T\boldsymbol{B}\boldsymbol{B}^T\boldsymbol{r} &= \sum_{k=1}^{L_n}\left[\sum_{i=1}^{n}E\left\{r_i^2B_k(X_i)^2\right\} + \sum_{i\neq i'}E\left\{r_ir_{i'}B_k(X_i)B_k(X_{i'})\right\}\right] \\ &= \sum_{k=1}^{L_n}\left[\sum_{i=1}^{n}E\left\{r_i^2B_k(X_i)^2\right\} + \sum_{i\neq i'}E\left\{r_iB_k(X_i)\right\}E\left\{r_{i'}B_k(X_{i'})\right\}\right] \\ &\leq \sum_{k=1}^{L_n}\left[\sum_{i=1}^{n}C_0^2M_n^{-2r}E\left\{B_k(X_i)^2\right\} + \sum_{i\neq i'}C_0^2M_n^{-2r}E\left\{B_k(X_i)\right\}E\left\{B_k(X_{i'})\right\}\right] \\ &\leq \sum_{k=1}^{L_n}\left[\sum_{i=1}^{n}C_0^2M_n^{-2r}E\left\{B_k(X_i)\right\} + \sum_{i\neq i'}C_0^2M_n^{-2r}E\left\{B_k(X_i)\right\}E\left\{B_k(X_{i'})\right\}\right] \\ &= O(M_nnM_n^{-2r}M_n^{-1}) + O(M_nn^2M_n^{-2r}M_n^{-2}) = O(n). \end{aligned}$$

Back to (S.4), we have

$$E(\boldsymbol{\varepsilon} - \boldsymbol{r})^T\boldsymbol{B}\boldsymbol{B}^T(\boldsymbol{\varepsilon} - \boldsymbol{r}) = O(n).$$

Therefore, applying Markov inequality yields that

$$(\boldsymbol{\varepsilon}-\boldsymbol{r})^T\boldsymbol{B}\boldsymbol{B}^T(\boldsymbol{\varepsilon}-\boldsymbol{r}) = O_p(n).$$

Thus, $T_1 \geq C_1 M_n^{-1}\eta_n^2 - \eta_n O_p(n^{-1/2})$.

Next, we will deal with the quantity T_2. Note the fact that $\|\boldsymbol{\gamma}_{0,A_j}\|_1 > 0$ for $j \in \mathcal{A}_3$ and $\|\boldsymbol{\gamma}_{0,A_j}\|_1 = 0$ for $j \in \mathcal{A}_1 \cup \mathcal{A}_2$. Moreover, by (S.1) and (S.2), we have

$$\begin{aligned} -T_2 = \lambda_n \sum_{j=1}^{M_n+1}\left(\|\boldsymbol{\gamma}_{0,A_j}\|_1^\alpha - \|\widehat{\boldsymbol{\gamma}}_{A_j}\|_1^\alpha\right) &\leq \lambda_n \sum_{j\in\mathcal{A}_3}\left(\|\boldsymbol{\gamma}_{0,A_j}\|_1^\alpha - \|\widehat{\boldsymbol{\gamma}}_{A_j}\|_1^\alpha\right) \\ &\leq 2\lambda_n \sum_{j\in\mathcal{A}_3}\|\boldsymbol{\gamma}_{0,A_j}\|_1^{\alpha-1}\|\widehat{\boldsymbol{\gamma}}_{A_j} - \boldsymbol{\gamma}_{0,A_j}\|_1 \\ &\leq 2\lambda_n \sum_{j\in\mathcal{A}_3}\|\boldsymbol{\gamma}_{0,A_j}\|_1^{\alpha-1}(d+1)^{1/2}\|\widehat{\boldsymbol{\gamma}}_{A_j} - \boldsymbol{\gamma}_{0,A_j}\|_2. \end{aligned}$$

In addition, by Cauchy-Schwarz inequality, we have

$$\begin{aligned} -T_2 &\leq 2\lambda_n(d+1)^{1/2}\left(\sum_{j\in\mathcal{A}_3}\|\boldsymbol{\gamma}_{0,A_j}\|_1^{2\alpha-2}\right)^{1/2}\left(\sum_{j\in\mathcal{A}_3}\|\widehat{\boldsymbol{\gamma}}_{A_j} - \boldsymbol{\gamma}_{0,A_j}\|_2^2\right)^{1/2} \\ &= O_p(\lambda_n\eta_n\alpha_n). \end{aligned}$$

By the minimality of $\widehat{\gamma}$, $D_n(\boldsymbol{v}) = Q_n(\widehat{\gamma}) - Q_n(\gamma_0) = T_1 + T_2 \leq 0$. Therefore,

$$0 \geq \frac{\eta_n^2}{M_n}C_1 - \eta_n O_p(n^{-1/2}) - O_p(\lambda_n\eta_n\alpha_n)$$

with probability approaching 1, which implies that $\eta_n = O_p\{M_n n^{-1/2} + \lambda_n M_n\alpha_n\}$. When $\lambda_n\alpha_n = O(n^{-1/2})$, we have $\eta_n = O_p(M_n n^{-1/2})$. This completes the proof. □

S3. Proof of Theorem 2

Proof of Theorem 2. Let $\widehat{\boldsymbol{\gamma}}^*$ be a vector defined by

$$\widehat{\gamma}_k^* = \begin{cases} 0, & k \in \mathcal{B}_1 \\ \widehat{\gamma}_k, & k \in \mathcal{B}_2 = \{1,\ldots,L_n\} - \mathcal{B}_1 \end{cases}$$

Hence, $\widehat{\boldsymbol{\gamma}}_{A_j}^* = \mathbf{0}$, for all $j \in \mathcal{A}_1 \cup \mathcal{A}_2$.

According to the Karush-Kuhn-Tucker conditions, if $\widehat{\gamma}_k \neq 0$, we have that

$$0 = -\frac{2}{n}\sum_{i=1}^n\left\{B_k(X_i)\left(Y_i - \sum_{j=1}^{L_n}\widehat{\gamma}_j B_j(X_i)\right)\right\} + \alpha\lambda_n\sum_{j:k\in A_j}\|\widehat{\boldsymbol{\gamma}}_{A_j}\|_1^{\alpha-1}\mathrm{sgn}(\widehat{\gamma}_k).$$

Multiplying $\widehat{\gamma}_k - \widehat{\gamma}_k^*$ on both sides yields that

$$\frac{2}{n}\sum_{i=1}^{n}\left\{B_k(X_i)\left(Y_i - \sum_{j=1}^{L_n}\widehat{\gamma}_j B_j(X_i)\right)\right\}(\widehat{\gamma}_k - \widehat{\gamma}_k^*) = \alpha\lambda_n \sum_{j:k\in A_j}\|\widehat{\boldsymbol{\gamma}}_{A_j}\|_1^{\alpha-1}|\widehat{\gamma}_k| I\{k \in \mathcal{B}_1\}.$$

Thus, summing up all k in $\mathcal{B}_1$ on both sides, we have

$$\begin{aligned} & \frac{2}{n}\sum_{k\in\mathcal{B}_1}\sum_{i=1}^{n}\left\{B_k(x_i)\left(Y_i - \sum_{j=1}^{L_n}\widehat{\gamma}_j B_j(x_i)\right)\right\}(\widehat{\gamma}_k - \widehat{\gamma}_k^*) \\ = \;& \alpha\lambda_n \sum_{k\in\mathcal{B}_1}|\widehat{\gamma}_k| \sum_{j:k\in A_j}\|\widehat{\gamma}_{A_j}\|_1^{\alpha-1} \\ = \;& \alpha\lambda_n \sum_{j=1}^{M_n+1}\|\widehat{\gamma}_{A_j}\|_1^{\alpha-1}\left(\|\widehat{\gamma}_{A_j}\|_1 - \|\widehat{\gamma}_{A_j}^*\|_1\right). \end{aligned}$$

That is,

$$\frac{2}{n}(\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}})^T \boldsymbol{B}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*) = \alpha\lambda_n \sum_{j=1}^{M_n+1}\|\widehat{\gamma}_{A_j}\|_1^{\alpha-1}\left(\|\widehat{\gamma}_{A_j}\|_1 - \|\widehat{\gamma}_{A_j}^*\|_1\right).$$

Using the inequality (S.3), we have

$$\frac{2}{n}\left|(\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}})^T \boldsymbol{B}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*)\right| \le \alpha\lambda_n \sum_{j\in\mathcal{A}_1\cup\mathcal{A}_2}\|\widehat{\gamma}_{A_j}\|_1^{\alpha} + \lambda_n \sum_{j\in\mathcal{A}_3}\left(\|\widehat{\gamma}_{A_j}\|_1^{\alpha} - \|\widehat{\gamma}_{A_j}^*\|_1^{\alpha}\right).$$

Moreover, by the minimality of $\widehat{\boldsymbol{\gamma}}$, we have

$$\frac{1}{n}\left(\|\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}}\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}}^*\|_2^2\right) \le \lambda_n \sum_{j=1}^{M_n+1}\|\widehat{\boldsymbol{\gamma}}^*\|_1^{\alpha} - \lambda_n \sum_{j=1}^{M_n+1}\|\widehat{\boldsymbol{\gamma}}\|_1^{\alpha}.$$

Combining the above two equations, we have

$$\begin{aligned} & \frac{2}{n}\left|(\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}})^T \boldsymbol{B}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*)\right| + (1-\alpha)\lambda_n \sum_{j\in\mathcal{A}_1\cup\mathcal{A}_2}\|\widehat{\gamma}_{A_j}\|_1^{\alpha} \\ \le \;& \lambda_n \sum_{j=1}^{M_n+1}\|\widehat{\boldsymbol{\gamma}}\|_1^{\alpha} - \lambda_n \sum_{j=1}^{M_n+1}\|\widehat{\boldsymbol{\gamma}}^*\|_1^{\alpha} \\ \le \;& \frac{1}{n}\left(\|\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}}^*\|_2^2 - \|\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}}\|_2^2\right) \\ = \;& \frac{1}{n}\|\boldsymbol{B}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*)\|_2^2 + \frac{2}{n}(\boldsymbol{Y} - \boldsymbol{B}\widehat{\boldsymbol{\gamma}})^T \boldsymbol{B}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*). \end{aligned}$$

Thus, we have

$$(1-\alpha)\lambda_n \sum_{j\in\mathcal{A}_1\cup\mathcal{A}_2}\|\widehat{\gamma}_{A_j}\|_1^{\alpha} \le \frac{1}{n}\|\boldsymbol{B}(\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*)\|_2^2 \le C_2 M_n^{-1}\|\widehat{\boldsymbol{\gamma}} - \widehat{\boldsymbol{\gamma}}^*\|_2^2 \le C_2 M_n^{-1}\|\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0\|_2^2.$$

It can be seen that

$$\sum_{j\in\mathcal{A}_1\cup\mathcal{A}_2}\|\widehat{\gamma}_{A_j}\|_1^\alpha \ge \left(\sum_{j\in\mathcal{A}_1\cup\mathcal{A}_2}\|\widehat{\gamma}_{A_j}\|_1\right)^\alpha \ge \|\widehat{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}^*\|_1^\alpha.$$

Finally, if $\|\widehat{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}^*\|_2 > 0$,

$$(1-\alpha)\lambda_n \le C_2 M_n^{-1}\|\widehat{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}^*\|_2^{2-\alpha} = O_p(1)(n^{-1+\alpha/2}M_n^{1-\alpha}).$$

Thus, we have

$$P(\|\widehat{\boldsymbol{\gamma}}-\widehat{\boldsymbol{\gamma}}^*\|_2 > 0) \le P\left\{\frac{\lambda_n}{n^{-1+\alpha/2}M_n^{1-\alpha}} \le O_p(1)\right\} \to 0.$$

□

S4. Additional Discussion on Estimating Globally and Locally Constant Functions

In the main paper, our research interest centers on detecting global sparsity and local sparsity in nonparametric regression models. Here, we take a closer look at the problem of nonparametric estimation of functions that are constant over the entire domain or part of the domain. In fact, this can be tackled in a similar fashion as our proposed method.

Continue to let $\mathbb{G}$ be the linear space of spline functions on $[0, 1]$ spanned by the B-spline basis functions $\{B_k(x) : k = 1, \ldots, L_n\}$. Here $L_n = M_n + d + 1$, where M_n is the number of equally-spaced interior knot points and d is the degree of polynomial pieces. For any $f \in \mathbb{G}$, we have $f(x) = \sum_{k=1}^{L_n}\gamma_k B_k(x)$. If $f(x) = c, x \in [\kappa_{j-1}, \kappa_j]$, for some constant c, as a consequence of local support property of B-splines, we have

$$\gamma_j B_j(x) + \cdots + \gamma_{j+d}B_{j+d}(x) = c.$$

By properties of B-spline basis functions and the fact

$$B_j(x) + \cdots + B_{j+d}(x) = 1, \quad \text{for } x \in [\kappa_{j-1}, \kappa_j],$$

we have $\gamma_j = \cdots = \gamma_{j+d} = c$.

From above discussion, it can be seen that, in order to identify the "flatness" on an interval $[\kappa_{j-1}, \kappa_j]$, it is essential to confirm the equality of the group of coefficients, $\{\gamma_j, \ldots, \gamma_{j+d}\}$. This observation motivates the following penalty function

$$\sum_{j=1}^{M_n+1} p(\boldsymbol{\gamma}_{A_j}), \tag{S.5}$$

where $A_j = \{j, j+1, \ldots j+d\}$ and

$$p(\boldsymbol{\gamma}_{\boldsymbol{A_j}}) = (|\gamma_{j+1} - \gamma_j| + \cdots + |\gamma_{j+d} - \gamma_{j+d-1}|)^{\alpha} \tag{S.6}$$

with $0 < \alpha < 1$. Moreover, the penalized least squares criterion can be written as

$$Q_n(\boldsymbol{\gamma}) = \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{B}\boldsymbol{\gamma}\|_2^2 + \lambda_n \sum_{j=1}^{M_n+1} p(\boldsymbol{\gamma}_{A_j}), \tag{S.7}$$

where $\boldsymbol{y}$, $\boldsymbol{B}$ and $\boldsymbol{\gamma}$ are defined in Equation (2.2) of the main paper.

Note that, if we define $\xi_0 = \gamma_1$ and $\xi_k = \gamma_{k+1} - \gamma_k$ for $k = 1, \ldots, L_n - 1$, we have

$$\begin{pmatrix} \xi_0 \\ \xi_1 \\ \vdots \\ \xi_{L_n-1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ -1 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{L_n} \end{pmatrix} \triangleq \boldsymbol{U}\boldsymbol{\gamma}.$$

For our convenience, denote $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_{L_n-1})^T$. Applying this transformation to (S.7) yields

$$Q_n(\boldsymbol{\xi}) = \frac{1}{n}\left\| \boldsymbol{y} - \boldsymbol{B}\boldsymbol{U}^{-1} \begin{pmatrix} \xi_0 \\ \boldsymbol{\xi} \end{pmatrix} \right\|_2^2 + \lambda_n \sum_{j=1}^{M_n+1} ||\boldsymbol{\xi}_{A_j^*}||_1^{\alpha}, \tag{S.8}$$

where $A_j^* = \{j, j+1, \ldots, j+d-1\}$ and $\boldsymbol{\xi}_{A_j^*}$ denotes the sub-vector of coefficients $(\xi_j, \ldots, \xi_{j+d-1})^T$.

In addition, we can consider the following transformed basis functions,

$$\boldsymbol{C}(x) \equiv \begin{pmatrix} C_1(x) \\ C_2(x) \\ \vdots \\ C_{L_n}(x) \end{pmatrix} = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ 0 & 1 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \begin{pmatrix} B_1(x) \\ B_2(x) \\ \vdots \\ B_{L_n}(x) \end{pmatrix}.$$

We have $\boldsymbol{C}(x) = (\boldsymbol{U}^T)^{-1}\boldsymbol{B}(x)$. Let $\boldsymbol{C} = (\boldsymbol{C}(X_1), \ldots, \boldsymbol{C}(X_n))^T$, and we have $\boldsymbol{C} = \boldsymbol{B}\boldsymbol{U}^{-1}$. We further note that $C_1(x) = 1$, and thus, the entries of the first column of $\boldsymbol{C}$ are 1. Moreover, $\boldsymbol{C}$ can be partitioned as

$$\boldsymbol{C} = (\boldsymbol{1}_n, \boldsymbol{C}_1).$$

Using the transformed basis functions, the penalized criterion (S.8) can be rewritten as

$$Q_n(\boldsymbol{\xi}) = \frac{1}{n}\|\boldsymbol{y} - \xi_0 \boldsymbol{1}_n - \boldsymbol{C}_1\boldsymbol{\xi}\|_2^2 + \lambda_n \sum_{j=1}^{M_n+1} ||\boldsymbol{\xi}_{A_j^*}||_1^{\alpha}, \tag{S.9}$$

which is similar to our proposed penalized criterion as expressed in Equation (2.3) of the main paper. Note that, replacing ξ_0 by its unpenalized least squares estimate, minimization of (S.9) can be carried out by our proposed algorithm in Section 2.3.

Finally, we provide an alternative view of the penalty function in (S.6). Note that, detecting the "flatness" of a smooth function $f(x)$ can be viewed as the problem of detecting global and local sparsities of its first derivative $f'(x)$. Consider a linear combination of B-spline basis functions, say $f(x) = \sum_{k=1}^{L_n} \gamma_k B_k(x)$. Its first derivative can be written as a linear combination, with coefficients $\gamma_i - \gamma_{i-1}$, of rescaled B-spline basis functions of degree $d-1$; see de Boor (1978) for more details. As discussed in our main paper, the functional sparsity of $f'(x)$ can be inferred by those coefficeints through the group bridge penalty, i.e., Equation (S.6).

References

de Boor, C. (1978), *A Practical Guide to Splines*, Springer, New York.

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