A Functional Information Criterion for Region Selection in Functional Linear Models

Yunxiang Huang and Qihua Wang

Academy of Mathematics and Systems Science, Chinese Academy of Sciences

University of Chinese Academy of Sciences

Zhejiang Gongshang University

Supplementary Material

This supplementary material contains proofs of Lemma 1 and Theorems 1–3, as well as some details about the simulation studies, additional simulations, and another application.

S1 Proof of Lemma 1

Let $\{\tau_1, \ldots, \tau_{d+P}\}$ be a sequence in $[0, T]$ such that $\tau_j = t_{j-(d+1)/2}$ when d is odd, and $\tau_j = (t_{j-(d+2)/2} + t_{j-d/2})/2$ when d is even. Here we set $t_j = t_0$ if $j < 0$, and $t_j = t_P$ if $j > P$. Let $b_{s,j} = \beta(\tau_j)$. By this construction, it immediately follows that $b_j = 0$ if $B_j(t)$ entirely lies inside the null region $N(\beta)$. Let $E = \{t : |t-u_i| \geq (d+1)P^{-1}, |t-v_i| \geq (d+1)P^{-1}, i = 1, \ldots, q\}$. An argument similar to the one used on pages 146 and 147 of de Boor (2001) shows that $|\beta(t) - \beta_S(t)| < cP^{-1}$ holds uniformly on E for some constant $c > 0$. By assumption (A1), β is bounded and has finite active intervals. Supposing that β has K active intervals, by direct calculation, it follows that

$$\begin{aligned}
\|\beta(t) - \beta_S(t)\|_2^2 &= \int_E [\beta(t) - \beta_S(t)]^2 \, dt + \int_{E^c} [\beta(t) - \beta_S(t)]^2 \, dt \\
&\leq T \times c^2 P^{-2} + \lambda^*(E^c) \times 4 \max\{[\beta(t)]^2 : t \in [0, T]\} \\
&\leq (c^2 T) P^{-2} + 8K(d+1)P^{-1} \max\{[\beta(t)]^2 : t \in [0, T]\},
\end{aligned}$$

which completes the proof.

S2 Proof of Theorem 1

We only give the proof of the case that M is the full model. The proof is almost identical for general M which includes all non-zero b_j. We first

introduce some useful properties of B-spline basis functions. Let $\langle \cdot, \cdot \rangle$ denote the L^2 inner product on $[0, T]$. Recall that J_m is an $(d+P) \times (d+P)$ matrix with entries $(J_m)_{ij} = \langle D^m B_i, D^m B_j \rangle$. Letting $\| \cdot \|$ denote the spectral norm of a matrix, we have the following result.

Lemma 1. *For $m \in [0, d]$, $\|J_m\| = \Theta(P^{2m-1})$.*

Proof. We first prove the result when $m = 0$. The B-spline basis functions form a partition of unity, that is, $\sum_{j=1}^{d+P} B_j(t) = 1$ for all $t \in [0, T]$. See page 89 of de Boor (2001) for details. From this, we have $\langle \sum_{j=1}^{d+P} B_j, \sum_{j=1}^{d+P} B_j \rangle = 1$, which implies $\|J_m\| = \Omega(P^{-1})$.

It remains to prove $\|J_m\| = O(P^{-1})$. Owning to the compact support property, it immediately follows that $\langle \sum_{j=1}^{d+P} B_j, B_i \rangle \leq (d+1)P^{-1}$ for each i, which implies that the row sums of J_0 are less than or equal to $(d+1)P^{-1}$. Then $\|J_m\| = O(P^{-1})$ follows from the fact that the largest eigenvalue of $\|J_m\|$ is less than or equal to its largest row sum. See Theorem 8.1.22 of Horn and Johnson (2013).

When $m > 0$, the mth derivative of $B_i(t)$ is a linear combination of B-spline basis functions of degree $d-m$. See page 116 of de Boor (2001) for details. The reminder of the proof is straightforward and so is omitted. □

For simplicity of notation, we first assume the intercept $a = 0$. In this

case, the spline estimator $\hat{\boldsymbol{b}}$ is given by

$$\hat{\boldsymbol{b}} = \underset{\boldsymbol{b}\in\mathcal{R}^{d+P}}{\arg\min} \left\{ \frac{1}{n}(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{Y} - \boldsymbol{Z}\boldsymbol{b}) + \lambda_n \boldsymbol{b}^T \boldsymbol{J}_m \boldsymbol{b} \right\}. \tag{S2.1}$$

We now prove $\|\hat{\boldsymbol{b}} - \boldsymbol{b}_S\| = O_p(1)$ where $\boldsymbol{b}_S$ is defined in Section S1.

Let $\Sigma(s,\, t) = \mathrm{Cov}[X(s),\, X(t)]$ be the covariance function of the functional predictor $X(t)$, and $\Sigma_n(s,\, t)$ be the sample version of $\Sigma(s,\, t)$ defined by

$$\Sigma_n(s,\, t) = \frac{1}{n}\sum_{i=1}^{n}(X_i(s) - \bar{X}(s))(X_i(t) - \bar{X}(t)).$$

As $\Sigma(s,\, t)$ is a positive (or semi-positive) definite operator on $L^2([0,\, T])$, with slight abuse of notation, for some $f \in L^2([0,\, T])$, write

$$(\Sigma f)(t) = \int_0^T \Sigma(s,\, t) f(s)\, dt\,.$$

The inner product and the norm induced by $\Sigma(s,\, t)$ are defined as $\langle f,\, g\rangle_\Sigma = \int_0^T (\Sigma f)(t) g(t)\, dt$ and $\|f\|_\Sigma = [\int_0^T (\Sigma f)(t) f(t)\, dt]^{1/2}$, respectively. Similarly, we define $(\hat{\Sigma}_n f)(t) = \int_0^T \Sigma_n(s,\, t) f(s)\, dt$, $\langle f,\, g\rangle_{\hat{\Sigma}_n} = \langle \hat{\Sigma}_n f,\, g\rangle$, and $\|f\|_{\hat{\Sigma}_n} = \langle \hat{\Sigma}_n f,\, f\rangle$. In addition, we denote the spectral norm of Σ and Σ_n by $\|\Sigma\|$ and $\|\Sigma_n\|$, respectively.

Let $Q(\boldsymbol{b})$ be the loss function in (S2.1), and write $Q(\boldsymbol{b}) = l(\boldsymbol{b}) + \lambda_n \boldsymbol{b}^T \boldsymbol{J}_2 \boldsymbol{b}$.

We have

$$\begin{aligned} l(\boldsymbol{b}) &= \frac{1}{n}(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{b})^T(\boldsymbol{Y}-\boldsymbol{Z}\boldsymbol{b}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^T X_i(t)(\beta(t)-\boldsymbol{B}^T(t)\boldsymbol{b})\,dt+\varepsilon_i\right]^2 \\ &= \|\beta-\boldsymbol{B}^T\boldsymbol{b}\|_{\Sigma_n}^2+\frac{2}{n}\sum_{i=1}^{n}\varepsilon_i\langle\beta-\boldsymbol{B}^T\boldsymbol{b},\,X_i(t)\rangle+\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i^2. \end{aligned}$$

The gradient of $Q(\boldsymbol{b})$ is

$$\nabla Q(\boldsymbol{b}) = 2\langle\boldsymbol{B}^T,\,\boldsymbol{B}^T\boldsymbol{b}-\beta\rangle_{\Sigma_n}-\frac{2}{n}\sum_{i=1}^{n}\langle\boldsymbol{B}^T,\,\varepsilon_i X_i\rangle+2\lambda_n\boldsymbol{J}_m\boldsymbol{b}. \tag{S2.2}$$

Let $\boldsymbol{H}$ be an $(d+P)\times(d+P)$ matrix with entries $(\boldsymbol{H})_{ij}=\langle B_i,\,B_j\rangle_{\Sigma_n}$. Then the Hessian matrix of $Q(\boldsymbol{b})$ is

$$\nabla^2 Q(\boldsymbol{b}) = 2\boldsymbol{H}+2\lambda_n\boldsymbol{J}_m,$$

which does not depend on $\boldsymbol{b}$. Noting that $\boldsymbol{H}=n^{-1}\boldsymbol{Z}^T\boldsymbol{Z}$, under assumptions (A2) and (A4), it is not hard to see that $\nabla^2 l(\boldsymbol{b})$ is strictly positive definite and

$$u^T[\nabla^2 l(\boldsymbol{b})]u = \|u\|^2\Omega(P^{-1}) \tag{S2.3}$$

holds for any non-zero $\boldsymbol{u}\in\mathbb{R}^{d+P}$. So $Q(\boldsymbol{b})$ is a strictly positive definite quadratic form with respect to $\boldsymbol{b}$, which ensures the existence and uniqueness of the minimizer.

We are now turning to the gradient of $Q(\boldsymbol{b})$. At the point $\boldsymbol{b}=\boldsymbol{b}_S$, we have the following result.

Lemma 2. *For non-zero* $\boldsymbol{u} \in \mathbb{R}^{d+P}$, $[\nabla l(\boldsymbol{b}_S)]^T \boldsymbol{u} = \|\boldsymbol{u}\| O_p(P^{-1})$.

Proof. We prove this lemma by deriving the asymptotic result for each term on the right side of (S2.2) at $\boldsymbol{b} = \boldsymbol{b}_S$. For the first term, by using Cauchy-Schwarz inequality, we have

$$\begin{aligned}\langle \boldsymbol{B}^T, \boldsymbol{B}^T\boldsymbol{b} - \beta\rangle_{\Sigma_n}\boldsymbol{u} &\le \|\boldsymbol{B}^T\boldsymbol{u}\|_2 \times \|\boldsymbol{B}^T\boldsymbol{b} - \beta\|_2 \|\Sigma_n\| \\ &\le \|\boldsymbol{u}\| \times O(P^{-1/2}) \times (c_1 P)^{-1/2} \times \|\Sigma_n\|,\end{aligned}$$

where the last inequality results from Lemma 1 and Lemma 1 in the main paper. By assumption (A3), $\|\Sigma\|$ is bounded. According to Proposition 1 of Dauxois et al. (1982), Σ_n converges uniformly almost surely to Σ under assumption (A3). So we have

$$\langle \boldsymbol{B}^T, \boldsymbol{B}^T\boldsymbol{b}_S - \beta\rangle_{\Sigma_n}\boldsymbol{u} = \|\boldsymbol{u}\| O_p(P^{-1}). \tag{S2.4}$$

For the second term, under assumption (A3), by the central limit theorem, it follows that $n^{-1/2}\sum_{i=1}^n \varepsilon_i X_i(t)$ converges in distribution to a zero-mean Gaussian random element in $L^2([0,\, T])$. Then using Cauchy-Schwarz inequality and Lemma 1 gives

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \langle \boldsymbol{B}^T, \varepsilon_i X_i\rangle \boldsymbol{u} &\le n^{-1/2}\|\boldsymbol{B}^T\boldsymbol{u}\|_2 \times \left\| n^{-1/2}\sum_{i=1}^n \varepsilon_i X_i \right\|_2 \\ &= \|\boldsymbol{u}\| O_p(n^{-1/2}P^{-1/2}).\end{aligned} \tag{S2.5}$$

For the last term, we have $\|\boldsymbol{b}_S\|_2 = O(P^{1/2})$ by the construction of $\boldsymbol{b}_S$

and $\|\boldsymbol{J}_m\|_2 = O(P^{2m-1})$ from Lemma 1. Then under assumption (A4), by using Cauchy-Schwarz inequality, it follows that

$$\begin{aligned}\|\lambda_n \boldsymbol{J}_m \boldsymbol{b}_S\|_2 &\leq \lambda_n \|\boldsymbol{J}_m\|_2 \times \|\boldsymbol{b}_S\|_2 \\ &= o(P^{-2m-1}) \times O(P^{2m-1}) \times O(P^{1/2}) = o(P^{-3/2}).\end{aligned} \tag{S2.6}$$

Combining this with (S2.4) and (S2.5) gives the desired result. □

By using Lemma 2 and (S2.3), we have

$$\begin{aligned}Q(\boldsymbol{b}_S + \boldsymbol{u}) - Q(\boldsymbol{b}_S) &= \nabla l(\boldsymbol{b}_S)\boldsymbol{u} + \frac{1}{2}\boldsymbol{u}^T \nabla^2 l(\boldsymbol{b}_S)\boldsymbol{u} \\ &= \|\boldsymbol{u}\| \times O_p(P^{-1}) + \|\boldsymbol{u}\|^2 \times \Omega(P^{-1}).\end{aligned}$$

So if $\|u\| = \omega(1)$ as $P, n \to \infty$, then $Q(\boldsymbol{b}_S + \boldsymbol{u}) - Q(\boldsymbol{b}_S) > 0$ holds in probability. This means the minimizer $\hat{\boldsymbol{b}}$ in (S2.1) must satisfies $\|\hat{\boldsymbol{b}} - \boldsymbol{b}_S\| = O_p(1)$. By the triangle inequality, it follows that

$$\begin{aligned}\|\hat{\beta} - \beta\|_2 &\leq \|[\boldsymbol{B}_{dP}^T \hat{\boldsymbol{b}} - \boldsymbol{B}_{dP}^T \boldsymbol{b}_S\|_2 + \|[\boldsymbol{B}_{dP}^T \boldsymbol{b}_S - \beta\|_2 \\ &= [(\hat{\boldsymbol{b}} - \boldsymbol{b}_S)^T \boldsymbol{J}_0 (\hat{\boldsymbol{b}} - \boldsymbol{b}_S)]^{1/2} + \|[\boldsymbol{B}_{dP}^T \boldsymbol{b}_S - \beta\|_2 \\ &= O_p(P^{-1/2}) + O(P^{-1/2}) \\ &= O_p(P^{-1/2}),\end{aligned}$$

where the third equality results from Lemma 1 and Lemma 1 in the main paper.

When $a \neq 0$, an argument similar to above can show that besides $\|\hat{\beta} - \beta\|_2^2 = O_p(P^{-1})$, $|\hat{a} - a| = O_p(P^{-1/2})$ also holds.

S3 Proof of Theorem 2

Obviously, Theorem 2 holds trivially when $\beta(t)$ is identical to zero on $[0, T]$. In what follows we assume that β is active on finite intervals.

Under assumptions (A1) and (A3), by the construction of $\beta_S(t)$ in Section S1, it follows that

$$\mu_i = \int_0^T X_i(t)\beta_S(t)\,dt + \int_0^T X_i(t)[\beta(t) - \beta_S(t)]\,dt =: \boldsymbol{Z}_i\boldsymbol{b}_S + \mu_{bias,\,i}, \quad \text{(S3.1)}$$

with $\mu_{bias,\,i} = O(P^{-1/2})$ for all i.

We first give a sufficient condition for the underfitted models in $\mathcal{M}$ such that the mean squared biases of the smoothing spline estimator for μ dominate P^{-1}.

Lemma 3. *If $\|\boldsymbol{b}_S(M^c)\|^2 = \Omega(P^{\delta})$ holds for some $\delta \geq 0$, then*

$$\Delta_n(M) =: \frac{1}{n}\|\boldsymbol{\mu} - \boldsymbol{H}(M)\boldsymbol{\mu}\|^2 = \Omega(P^{-1+\delta}). \quad \text{(S3.2)}$$

Proof. Observing $[I - \boldsymbol{H}(M)][\boldsymbol{Z}(M)\boldsymbol{b}_S(M)] = 0$ and that $[I - \boldsymbol{H}(M)]$ is

idempotent, we have

$$\begin{aligned}
&\frac{1}{n}\|\boldsymbol{\mu} - \boldsymbol{H}(M)\boldsymbol{\mu}\|^2 \qquad \text{(S3.3)}\\
&= \frac{1}{n}\|[I - \boldsymbol{H}(M)][\boldsymbol{Z}(M)\boldsymbol{b}_S(M) + \boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c) + \boldsymbol{\mu}_{bias}]\|^2\\
&\geq \frac{1}{n}\|[I - \boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|^2 + \frac{2}{n}\boldsymbol{\mu}_{bias}^T[I - \boldsymbol{H}(M)]\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c).
\end{aligned}$$

For the first term on the right side of the last inequality of (S3.3), we have

$$\begin{aligned}
&\frac{1}{n}\|[I - \boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|^2\\
&= \frac{1}{n}\boldsymbol{b}_S^T(M^c)\{\boldsymbol{Z}^T(M^c)\boldsymbol{Z}(M^c)\\
&\qquad - \boldsymbol{Z}^T(M^c)\boldsymbol{Z}(M)(\boldsymbol{Z}^T(M)\boldsymbol{Z}(M))^{-1}\boldsymbol{Z}(M)^T\boldsymbol{Z}(M^c)\}\boldsymbol{b}_S(M^c).
\end{aligned}$$

Let $A = \boldsymbol{Z}^T(M^c)\boldsymbol{Z}(M^c)$, $B = \boldsymbol{Z}^T(M^c)\boldsymbol{Z}(M)$, $C = \boldsymbol{Z}^T(M)\boldsymbol{Z}(M)$. According to the inverse of a partitioned matrix formula for Hermitian matrices on page 472 of Horn and Johnson (2013),

$$\begin{bmatrix} A & B \\ B^T & C \end{bmatrix}^{-1} = \begin{bmatrix} (A - BCB^T)^{-1} & A^{-1}B(B^TA^{-1}B - C)^{-1} \\ (B^TA^{-1}B - C)^{-1}B^TA^{-1} & (C - B^TAB)^{-1} \end{bmatrix},$$

it is not hard to see that the minimum eigenvalue of $(A - BC^{-1}B^T)$ is $\Theta_p(nM^{-1})$ under condition (A2). As a result, we have, for all M satisfying $\|\boldsymbol{b}_S(M^c)\|^2 = \Omega(P^\delta)$,

$$\frac{1}{n}\|[I - \boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|^2 = \frac{1}{n}\|\boldsymbol{b}_S(M^c)\|^2\Theta(nP^{-1}) = \Omega(P^{-1+\delta}). \qquad \text{(S3.4)}$$

For the last term on the right side of the last inequality of (S3.3), by using the Cauchy–Schwarz inequality, we have, for all M satisfying $\|\boldsymbol{b}_S(M^c)\|^2 = \Omega(P^\delta)$,

$$\begin{aligned} &\frac{|\boldsymbol{\mu}_{bias}^T[I-\boldsymbol{H}(M)]\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)|}{\|[I-\boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|^2} \\ \le\ &\frac{\|\boldsymbol{\mu}_{bias}\| \times \|[I-\boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|}{\|[I-\boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|^2} \\ =\ &\frac{n^{-1/2}\|\boldsymbol{\mu}_{bias}\|}{n^{-1/2}\|[I-\boldsymbol{H}(M)][\boldsymbol{Z}(M^c)\boldsymbol{b}_S(M^c)]\|} \\ =\ &\frac{O(P^{-1/2})}{\Omega(P^{-1/2+\delta/2})} = O(P^{-\delta/2}). \end{aligned} \tag{S3.5}$$

Combining (S3.3)–(S3.5) gives the desired result. □

Lemma 4. *Let $\mathcal{M}_1$ denote a subset of $\mathcal{M}$ such that $\dim(M_0 \setminus M) = \Theta(P)$ for each $M \in \mathcal{M}_1$. It follows that $\|\boldsymbol{b}(M^c)\|^2 = \Omega(P)$ for $M \in \mathcal{M}_1$ and hence $\Delta_n(M) = \Omega(1)$ by using Lemma 3.*

Proof. By the construction of $\boldsymbol{b}_S$, the behavior of $P^{-1}\sum_{k=j_0}^{j_1} b_k^2$ for $j_1 \ge j_0$ is analogy to a Riemann sum. Under assumption (A1), one can verify that $P^{-1}\sum_{k=j_0}^{j_1} b_k^2 = \int_{\tau(j_0)}^{\tau(j_1)} [\beta(t)]^2\,dt + O(P^{-1})$ for all $j_0 - (d+1)/2 \ge 1$, $j_1 - d/2 \le P$, and $(\tau(j_0), \tau(j_1)) \in A(\beta)$ where τ_j is defined in Section S1. Now $\|\boldsymbol{b}(M^c)\|^2 = \Omega(P)$ follows immediately from the fact that for a given $l > 0$, there exists a $C(l) > 0$ such that

$$\inf_{E \subset A(\beta),\, \lambda^*(E) \ge l} \left[\int_E (\beta(t))^2\,dt\right] \ge C(l). \tag{S3.6}$$

□

Next, we show that the within-sample mean squared error of the spline estimator and that of the least squares estimator are close as $(P,\, n) \to \infty$.

Lemma 5. *The following result holds uniformly for all $M \in \mathcal{M}$:*

$$\left| \frac{1}{n} \|(I - \boldsymbol{H}(M))\boldsymbol{Y}\|^2 - \frac{1}{n} \|(I - \boldsymbol{H}_{\lambda_n}(M))\boldsymbol{Y}\|^2 \right| = o_p(P^{-1}),$$

where $\boldsymbol{H}(M) = \boldsymbol{Z}(M)[\boldsymbol{Z}^T(M)\boldsymbol{Z}(M)]^{-1}\boldsymbol{Z}^T(M)$.

Proof. The following proof works uniformly over $M \in \mathcal{M}$. For brevity, we shall drop M in parentheses in this proof. By assumption (A2), it is apparent that all the eigenvalues of $n^{-1}Z^T(M)Z(M)$ lie in the interval $[c_2P^{-1},\, c_3P^{-1}]$ for any $M \in \mathcal{M}$. According to Section 5.8 in Horn and Johnson (2013), it follows that

$$\begin{aligned} &\|(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\| \leq \frac{\|(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\|^2 \|n\lambda_n\boldsymbol{J}_m\|}{1 - \|(\boldsymbol{Z}^T\boldsymbol{Z})^{-1}(n\lambda_n\boldsymbol{J}_m)\|} \\ &= \frac{O(P^2n^{-2}) \times O(n\lambda_nP^{2m-1})}{1 - O(Pn^{-1}) \times (n\lambda_nP^{2m-1})} = \frac{O(n^{-1}\lambda_nP^{2m+1})}{1 - O(\lambda_nP^{2m})} = o(n^{-1}), \end{aligned}$$

and hence

$$\begin{aligned} &\|(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1}(\boldsymbol{Z}^T\boldsymbol{Z})(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1}\| \\ &= \|(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1}[(\boldsymbol{Z}^T\boldsymbol{Z})(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1} - I]\| \\ &\leq \|(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1}\| \times \|(\boldsymbol{Z}^T\boldsymbol{Z})\| \times \|(\boldsymbol{Z}^T\boldsymbol{Z} + n\lambda_n\boldsymbol{J}_m)^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}]\| \\ &= o(n^{-1}). \end{aligned}$$

Observing that $\boldsymbol{H}^2 = \boldsymbol{H}$, we have

$$
\begin{aligned}
&\left|\frac{1}{n}\|(I-\boldsymbol{H})\boldsymbol{Y}\|^2 - \frac{1}{n}\|(I-\boldsymbol{H}_{\lambda_n})\boldsymbol{Y}\|^2\right| \\
&= \frac{1}{n}\left|-\boldsymbol{Y}^T\boldsymbol{H}\boldsymbol{Y} + 2\boldsymbol{Y}^T\boldsymbol{H}_{\lambda_n}\boldsymbol{Y} - \boldsymbol{Y}^T\boldsymbol{H}_{\lambda_n}^2\boldsymbol{Y}\right| \\
&\le \frac{1}{n}|\boldsymbol{Y}^T(\boldsymbol{H}-\boldsymbol{H}_{\lambda_n})\boldsymbol{Y}| + \frac{1}{n}|\boldsymbol{Y}^T(\boldsymbol{H}_{\lambda_n}-\boldsymbol{H}_{\lambda_n}^2)\boldsymbol{Y}| \\
&= \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Z}\left[(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right]\boldsymbol{Z}^T\boldsymbol{Y} \\
&\qquad + \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Z}\left[(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1}(\boldsymbol{Z}^T\boldsymbol{Z})(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1}\right. \\
&\qquad \left.-(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1}\right]\boldsymbol{Z}^T\boldsymbol{Y} \\
&\le \frac{1}{n}\|\boldsymbol{Z}^T\boldsymbol{Y}\|^2\left\|(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1} - (\boldsymbol{Z}^T\boldsymbol{Z})^{-1}\right\| \\
&\qquad + \frac{1}{n}\|\boldsymbol{Z}^T\boldsymbol{Y}\|^2\left\|(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1}(\boldsymbol{Z}^T\boldsymbol{Z})(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1}\right. \\
&\qquad \left.-(\boldsymbol{Z}^T\boldsymbol{Z}+n\lambda_n\boldsymbol{J}_m)^{-1}\right\| \\
&= \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Z}\boldsymbol{Z}^T\boldsymbol{Y} \times o(n^{-1}) \\
&\le \frac{1}{n}\boldsymbol{Y}^T\boldsymbol{Y} \times \|\boldsymbol{Z}^T\boldsymbol{Z}\| \times o(n^{-1}) \\
&= o_p(P^{-1}),
\end{aligned}
$$

where the last equality results from the law of large numbers. □

Write $L_n(M) = n^{-1}\|\boldsymbol{\mu} - \boldsymbol{H}(M)\boldsymbol{Y}\|$, and

$$R_n(M) = \mathrm{E}(L_n(M)) = \Delta_n(M) + \frac{1}{n}\sigma^2\dim(M).$$

The following lemma provides an asymptotic expression for FICf(M), which is crucial to our approach.

Lemma 6. *Under assumptions (A1) – (A8), the following result holds uniformly for $M \in \mathcal{M}$:*

$$FICf(M) = L_n(M) + \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 - \frac{2}{n}\sigma^2 \dim(M) + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) + o_p(L_n(M)) + o_p(P^{-1}). \tag{S3.7}$$

When ε is Gaussian, assumption (A6) can be replaced by (A6′).

Proof. By Lemma 5, we have

$$\begin{aligned}\mathrm{FICf}(M) &= \frac{1}{n}\|(I - \boldsymbol{H}(M))(\boldsymbol{\mu} + \boldsymbol{\varepsilon})\|^2 + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) + o_p(P^{-1}) \\ &= L_n(M) + \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 - \frac{2}{n}\boldsymbol{\varepsilon}^T \boldsymbol{H}(M)\boldsymbol{\varepsilon} + \frac{2}{n}\boldsymbol{\varepsilon}^T (I - \boldsymbol{H}(M))\boldsymbol{\mu} \\ &\quad + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) + o_p(P^{-1}),\end{aligned}$$

holds uniformly for all $M \in \mathcal{M}$. Now it suffices to prove

$$\max_{M \in \mathcal{M}} \left| \frac{R_n}{L_n} - 1 \right| \xrightarrow{p} 0, \tag{S3.8}$$

$$\max_{M \in \mathcal{M}} \left| \frac{\sigma^2 \dim(M) - \boldsymbol{\varepsilon}^T \boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nR_n} \right| \xrightarrow{p} 0, \tag{S3.9}$$

and

$$\max_{M \in \mathcal{M}} \left| \frac{\boldsymbol{\varepsilon}^T (I - \boldsymbol{H}(M))\boldsymbol{\mu}}{nR_n(M)} \right| \xrightarrow{p} 0. \tag{S3.10}$$

Note that (S3.8) is a direct consequence of (S3.9) and (S3.10). So it suffices to prove (S3.9) and (S3.10).

Recall that each $M \in \mathcal{M}$ satisfies assumption (A5), which implies that there are at most polynomial rate of candidate models in $\mathcal{M}$. As a consequence, under assumption (A7), it follows that

$$\sum_{M \in \mathcal{M}} \frac{1}{[nR_n(M)]^l} \xrightarrow{p} 0. \tag{S3.11}$$

Now under (S3.11), (S3.9) and (S3.10) can be shown in a similar way on page 970 of Li (1987).

In the case that ε is Gaussian and assumption (A6) is replace by (A6′), denoting the cardinality of candidate model set $\mathcal{M}$ by $\mathbf{card}(\mathcal{M})$, we have $\mathbf{card}(M) = o(2^M)$. Then it is not hard to verify that

$$\sum_{M \in \mathcal{M}_n} \delta^{nR_n(M)} \to 0, \tag{S3.12}$$

for any $0 < \delta < 1$ when $\beta(t)$ is not identical to zero. Now, (S3.9) can be shown by using Lemma 2.1 in Shibata (1981), and the derivation of (S3.10) is analogous that of equation (2.4) in Shibata (1981). □

We are now in a position to complete the proof. Let $M_0 \in \mathcal{M}$ be the largest model satisfying $\inf\{|b_S(j)| : j \in M_0\} > 0$. By the construction of $\beta_S(t)$, we have $\lambda^*\{N(\beta) \triangle N(M_0)\} \to 0$ as $(P, n) \to \infty$. Therefore, to develop the region selection consistency of $N(\hat{M})$, it suffices to prove

$$\lambda^*[N(M_0) \triangle N(\hat{M})] = o_p(1). \tag{S3.13}$$

Since $L_n(M) = \Delta_n(M) + n^{-1}\boldsymbol{\varepsilon}^T\boldsymbol{H}\boldsymbol{\varepsilon} = \Delta_n(M) + O_p(n^{-1}P)$, by using Theorem 1, it follows that $L_n(M_0) = O_p(P^{-1})$. Using Lemma 6 and assumption (A9.3), we have the following result:

$$
\begin{aligned}
&\min_{M\in\mathcal{M}_1}\left[\mathrm{FICf}(M) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] - \left[\mathrm{FICf}(M_0) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] \qquad \text{(S3.14)}\\
&= \min_{M\in\mathcal{M}_1}\left[L_n(M) - \frac{2}{n}\sigma^2\dim(M) + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) + o_p(L_n(M))\right]\\
&\quad - \left[L_n(M_0) - \frac{2}{n}\sigma^2\dim(M_0) + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M_0)) + o_p(L_n(M_0))\right]\\
&\quad + o_p(P^{-1})\\
&= \min_{M\in\mathcal{M}_1}\left\{[L_n(M) + o_p(L_n(M))] - \left[\frac{2}{n}\sigma^2(\dim(M) - \dim(M_0))\right]\right.\\
&\quad \left. + \left[\frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) - \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M_0))\right]\right\}\\
&\quad - [L_n(M_0) + o_p(L_n(M_0))] + o_p(P^{-1})\\
&= \min_{M\in\mathcal{M}_1}[L_n(M) + o_p(L_n(M))]\\
&\quad + O(Pn^{-1}) + o_p(1) + O_p(P^{-1}) + o_p(P^{-1}).
\end{aligned}
$$

By using (S3.9), we have

$$\begin{aligned}
&\max_{M\in\mathcal{M}}\left|\frac{\boldsymbol{\varepsilon}^T\boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nL_n}\right| \qquad\qquad (S3.15)\\
\leq\ &\max_{M\in\mathcal{M}}\left|\frac{\boldsymbol{\varepsilon}^T\boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nR_n}\right|\times\max_{M\in\mathcal{M}}\left|\frac{L_n}{R_n}\right|\\
\leq\ &\left[\max_{M\in\mathcal{M}}\left|\frac{\sigma^2\dim(M)-\boldsymbol{\varepsilon}^T\boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nR_n}\right|+\max_{M\in\mathcal{M}}\left|\frac{\sigma^2\dim(M)}{nR_n}\right|\right]\\
&\qquad\times\left[\max_{M\in\mathcal{M}}\left|\frac{L_n}{R_n}-1\right|+1\right]\\
=\ &\left[o_p(1)+\max_{M\in\mathcal{M}}\left|\frac{\sigma^2\dim(M)}{nR_n}\right|\right]\times[o_p(1)+1].
\end{aligned}$$

Hence on $\mathcal{M}_1$ we have

$$\max_{M\in\mathcal{M}_1}\left|\frac{\boldsymbol{\varepsilon}^T\boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nL_n}\right|=o_p(1). \qquad (S3.16)$$

Combining (S3.9) and (S3.10) gives

$$\max_{M\in\mathcal{M}}\left|\frac{\boldsymbol{\varepsilon}^T(I-\boldsymbol{H}(M))\boldsymbol{\mu}}{nL_n(M)}\right|\xrightarrow{p}0. \qquad (S3.17)$$

Then by using (S3.16) and (S3.17), we have

$$\max_{M\in\mathcal{M}_1}\left|\frac{\Delta_n(M)}{L_n(M)}-1\right|=\max_{M\in\mathcal{M}_1}\left|\frac{2\boldsymbol{\varepsilon}^T(I-\boldsymbol{H}(M))\boldsymbol{\mu}-\boldsymbol{\varepsilon}^T\boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nL_n(M)}\right|.=o_p(1). \qquad (S3.18)$$

Therefore for $M\in\mathcal{M}_1$, it follows that $L_n(M)=\Delta_n(M)+o_p(1)=\Theta_p(1)$ holds uniformly and hence $\mathrm{FICf}(M)-\mathrm{FICf}(M_0)>0$ also holds uniformly in probability as P, $n\to\infty$.

On the other hand, let $\mathcal{M}_2$ be a subset of $\mathcal{M}\setminus\mathcal{M}_1$ such that $\dim(M)-$

$\dim(M_0) = \Theta(P)$ for each $M \in \mathcal{M}_2$. Again, as P, $n \to \infty$, we have

$$\begin{aligned}
&\min_{M\in\mathcal{M}_2}\left[\mathrm{FICf}(M) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] - \left[\mathrm{FICf}(M_0) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] \qquad \text{(S3.19)}\\
=\ &\min_{M\in\mathcal{M}_2}\left[L_n(M) - \frac{2}{n}\sigma^2\dim(M) + \frac{1}{n}\hat{\sigma}^2 p_{P,\,n}(\dim(M))\right]\\
&\quad - \left[L_n(M_0) - \frac{2}{n}\sigma^2\dim(M_0) + \frac{1}{n}\hat{\sigma}^2 p_{P,\,n}(\dim(M_0))\right]\\
&\quad + o_p(L_n(M)) + o_p(L_n(M_0)) + o_p(P^{-1})\\
\geq\ &\min_{M\in\mathcal{M}_2}\left\{\frac{1}{n}\hat{\sigma}^2[p_{P,\,n}(\dim(M)) - p_{P,\,n}(\dim(M_0))]\right.\\
&\quad \left. -\frac{2}{n}[\dim(M) - \dim(M_0)]\right\} + o_p(1).
\end{aligned}$$

By assumption (A9.2), it is not hard to verify that $\mathrm{FICf}(M) - \mathrm{FICf}(M_0) > 0$ holds uniformly on $\mathcal{M}_2$ in probability as $(P,\, n) \to \infty$. From Lemma 4, we have $\dim(M_0 \setminus M) = o(P)$ on $\mathcal{M} \setminus \mathcal{M}_1$. As a result, for $M \in \mathcal{M}_2$,

$$\begin{aligned}
\dim(M \setminus M_0) =&\ \dim(M) - \dim(M_0) + \dim(M_0 \setminus M)\\
=&\ \dim(M) - \dim(M_0) + o(P),
\end{aligned}$$

which implies for $M \in \mathcal{M}_2$, $\dim(M) - \dim(M_0) = \Theta(P)$ also holds.

To summarize, we have proved that

$$\min_{M\in\mathcal{M}_1\cup\mathcal{M}_2}[\mathrm{FICf}(M)] > \mathrm{FICf}(M_0)$$

holds in probability as P, $n \to \infty$, which implies that $\mathrm{P}\{\hat{M} \in [\mathcal{M} \setminus (\mathcal{M}_1 \cup \mathcal{M}_2)]\} \to 1$. Consequently, $\dim(M_0 \setminus \hat{M}) = o_p(P)$ and $\dim(\hat{M} \setminus M_0) = o_p(P)$

hold simultaneously and hence (S3.13) holds immediately, which completes the proof.

S4 Proof of Theorem 3

We first assume $L_n = \Theta(1)$. The following result is a sharp version of Lemma 4.

Lemma 7. *Let $\mathcal{M}_1(\delta_1)$ denote a subset of $\mathcal{M}$ such that for any for $M \in \mathcal{M}_1(\delta_1)$, $\dim(M_0 \setminus M) = \Omega(P^{(2p+\delta_1)/(2p+1)})$ holds for some $0 < \delta_1 < 1$. It follows that $\|\boldsymbol{b}(M^c)\|^2 = \Omega(P^{\delta_1})$ for $M \in \mathcal{M}_1(\delta_1)$ and hence $\Delta_n(M) = \Omega(P^{-1+\delta_1})$ by using Lemma 3.*

Proof. Recall that each $M \in \mathcal{M}$ satisfies assumption (A5). If M excludes some adjacent b_k's inside an active interval, it immediately follows that $\dim(M_0 \setminus M) = \Theta(P)$, and the desired result holds by using Lemma 4. Consequently, we only need to consider the case that M excludes some adjacent b_k's at the boundary between an active interval and a null interval. Suppose that there are $k = \Omega_p(P^{(2p+\delta_1)/(2p+1)})$ adjacent $(b_j, \ldots, b_{j+k})$ such that $b_{j-1} \in M_0^c$ and $b_{j+k+1} \in M_0$, or $b_{j-1} \in M_0$ and $b_{j+k+1} \in M_0^c$. When $b_{j-1} \in M_0^c$ and $b_{j+k+1} \in M_0$, by assumption (A10), it follows that

$$\sum_{l=j}^{j+k} b_l^2 = \Omega\left[\sum_{l=j}^{j+k}(P^{-1}(l-j+1))^{2p}\right] = \Omega(P^{-2p}k^{2p+1}) = \Omega(P^{\delta_1}).$$

The proof when $b_{j-1} \in M_0$ and $b_{j+k+1} \in M_0^c$ is identical. □

Now we are ready to prove Theorem 3. The proof is similar to that of Theorem 2. For any $0 < \delta_1 < 1$, as P, $n \to \infty$, by using (S3.14), we have

$$\begin{aligned}
&\min_{M \in \mathcal{M}_1(\delta_1)} \left[\mathrm{FICf}(M) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 \right] - \left[\mathrm{FICf}(M_0) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2 \right] \\
= &\min_{M \in \mathcal{M}_1(\delta_1)} \left\{ [L_n(M) + o_p(L_n(M))] - \left[\frac{2}{n}\sigma^2(\dim(M) - \dim(M_0)) \right] \right. \\
&\quad \left. + \left[\frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) - \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M_0)) \right] \right\} \\
&\quad - [L_n(M_0) + o_p(L_n(M_0))] + o_p(P^{-1}) \\
= &\min_{M \in \mathcal{M}_1(\delta_1)} [L_n(M) + o_p(L_n(M))] + O(P^{(2p+\delta_1)/(2p+1)} n^{-1}) \\
&\quad + o_p(P^{-1+\delta_1}) + O_p(P^{-1}) + o_p(P^{-1}) \\
= &\min_{M \in \mathcal{M}_1(\delta_1)} [L_n(M) + o_p(L_n(M))] + o_p(P^{-1+\delta_1}),
\end{aligned}$$

where the second inequality is due to assumption (A11.2). Using Lemma 7, we have $\Delta_n(M) = \Omega(P^{-1+\delta_1})$. An argument similar to the one used in the proof of (S3.18) shows that

$$\begin{aligned}
&\max_{M \in \mathcal{M}_1(\delta_1)} \left| \frac{\Delta_n(M)}{L_n(M)} - 1 \right| \\
= &\max_{M \in \mathcal{M}_1(\delta_1)} \left| \frac{2\boldsymbol{\varepsilon}^T(I - \boldsymbol{H}(M))\boldsymbol{\mu} - \boldsymbol{\varepsilon}^T \boldsymbol{H}(M)\boldsymbol{\varepsilon}}{nL_n(M)} \right| = o_p(1).
\end{aligned}$$

Then it follows that $\mathrm{P}(\hat{M} \in \mathcal{M}_1(\delta_1)) \xrightarrow{p} 0$ for any $0 < \delta_1 < 1$.

Otherwise, let $\mathcal{M}_2(\delta_1)$ be a subset of $\mathcal{M} \setminus \mathcal{M}_1(\delta_1)$ such that $\dim(M) - \dim(M_0) = \Omega(P^{(2p+\delta_1)/(2p+1)})$ for each $M \in \mathcal{M}_2$. By the definition of

$\mathcal{M}_1(\delta_1)$ in Lemma 7, it follows that $\dim(M_0 \setminus M) = o_p(P^{(2p+\delta_1)/(2p+1)})$ on $M \in \mathcal{M} \setminus \mathcal{M}_1(\delta_1)$. As a result, for $M \in \mathcal{M}_2(\delta_1)$,

$$\begin{aligned}\dim(M \setminus M_0) &= (\dim(M) - \dim(M_0)) + \dim(M_0 \setminus M) \\ &= (\dim(M) - \dim(M_0)) + o(P^{(2p+\delta_1)/(2p+1)}),\end{aligned}$$

which means $\dim(M) - \dim(M_0) = \Omega(P^{(2p+\delta_1)/(2p+1)})$ if and only if $\dim(M \setminus M_0) = \Omega(P^{(2p+\delta_1)/(2p+1)})$ for $M \in \mathcal{M}_2(\delta_1)$. Again, as P, $n \to \infty$, we have

$$\begin{aligned}&\min_{M\in\mathcal{M}_2(\delta_1)} \left[\mathrm{FICf}(M) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] - \left[\mathrm{FICf}(M_0) - \frac{1}{n}\|\boldsymbol{\varepsilon}\|^2\right] \\ =\ &\min_{M\in\mathcal{M}_2(\delta_1)} \left[L_n(M) + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)) + o_p(L_n(M))\right] \\ &\quad - \left[L_n(M_0) + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M_0))\right] + O_p(P^{-1}) \\ \geq\ &\min_{M\in\mathcal{M}_2(\delta_1)} \left[\frac{1}{n}\hat{\sigma}^2(p_{P,n}(\dim(M)) - p_{P,n}(\dim(M_0)))\right] - L_n(M_0) + O_p(P^{-1}) \\ \geq\ &\frac{1}{n}\hat{\sigma}^2\big[p_{P,n}\big(\dim(M_0) + \omega(P^{2p/(2p+1)})\big) - p_{P,n}(\dim(M_0))\big] + O_p(P^{-1}),\end{aligned}$$

where the second last inequality results from the monotonicity of $p_{P,n}(M)$. By using assumption (A11.1), it follows that $\mathrm{P}(\hat{M} \in \mathcal{M}_2(\delta_1)) \xrightarrow{p} 0$ for any $0 < \delta_1 < 1$.

To summarize, we have proved that

$$\mathrm{P}\{\hat{M} \in [\mathcal{M} \setminus (\mathcal{M}_1(\delta_1) \cup \mathcal{M}_2(\delta_1))]\} \xrightarrow{p} 1,$$

for any $0 < \delta_1 < 1$. Note that he case that $\delta_1 = 1$ has been proved in Theorem 2 and hence the desired result follows.

We now turn to the case that ε is Gaussian and $L_n = o(1)$. By using assumptions (A12.1) and (A12.2), a similar argument shows that the probability of that M excludes some adjacent b_k inside an active interval also goes to zero as $(P, n) \to \infty$. The rest of the proof is similar to the proof of Theorem 2 when ε is Gaussian. The details will not be reproduced here.

S5 Additional Simulations

We consider the following three additional models.

Case III No signal. $\beta(t) = 0$.

Case IV $\beta(t) = 7t^3 + 2\sin(4\pi t + 0.2)$. As β is active on [0, 1] and crosses zero twice, this is the case that does not favor the FICf procedure and one does not expect FICf to identify to any null region.

Case V We set

$$\beta(t) = \begin{cases} 0, & 0 \le t \le 0.25, \\ 1600(t - 0.25)^2, & 0.25 < t \le 0.3, \\ -20t + 10, & 0.3 < t \le 0.7, \\ -1600(t - 0.75)^2, & 0.7 < t < 0.75, \\ 0, & 0.75 \le t \le 1. \end{cases}$$

In this case, β is not differentiable everywhere on the interior of the active region, and $\beta(t) = \beta'(t) = 0$ at the boundaries of the null region.

For case III, ε follows the standard normal distribution. Other settings are same as those in Section 4 in the main paper.

We use the following two indicators to respectively measure the estimation accuracy and the prediction accuracy: integrated squared estimating error of $\hat{\beta}$ for the selected model,

$$\text{ISE} = \int_0^T \{\hat{\beta}(t)\, \text{I}[t \in A(\hat{M})] - \beta_0(t)\}^2 \, dt\,, \tag{S5.1}$$

and average squared prediction errors of the selected model,

$$\text{ASPE} = \frac{1}{N} \sum_{i=1}^{N} \left\{ \int_0^T X_i'(t)(\hat{\beta}(t)\, \text{I}[t \in A(\hat{M})] - \beta_0(t))\, dt + (\hat{a} - a) \right\}^2, \tag{S5.2}$$

where X_i', $i = 1, \ldots, N$ are independent copies of X and we set $N = 1000$ in our simulations.

Tables 1–3 present summaries of $\Delta\lambda_\beta^*(\hat{M})$ in the main paper, ISE in (S5.1) and ASPE (S5.2), respectively.

(*Insert Table 1, Table 2 and Table 3 here.*)

S6 Effects of Different Model Complexity Penalty

We conduct a simulation study to investigate the finite sample performance of different model complexity penalties $p_{P,n}$ in

$$\mathrm{FICf}_{n,P}(M) = \frac{1}{n} S_n(M)^2 + \frac{1}{n}\hat{\sigma}^2 p_{P,n}(\dim(M)). \tag{S6.1}$$

We consider the penalties that have the form of

$$p_{P,n}(\dim(M)) = n^{a_1/9}\left(\frac{\dim(M)}{P}\right)^{a_2}, \tag{S6.2}$$

where $a_1 \in \{6, 6.5, 7, 7.5, 8\}$ and $a_2 \in \{1, 1.5\}$ are parameters. The settings are same as those in Section 4 in the main paper. We use the length of the symmetric difference $\Delta\lambda^*_\beta(\hat{M})$, ISE and ASPE of the selected model to measure the quality of the FICf method using different model complexity penalties. The results are summarized in Tables 4 – 6. It can be seen that the recommended $p_{P,n}$ in Section 2.3 of the main paper with $(a_1, a_2) = (7, 1)$ performs well in general.

(*Insert Table 4, Table 5 and Table 6 here.*)

S6.1 Application to Canadian Weather Data

The Canadian weather is consisted of mean daily temperature and precipitation data of 35 Canadian weather stations in a year. The data has been analysed in Ramsay and Silverman (2005); Ramsay et al. (2009); James

et al. (2009); Lin et al. (2017) for the purpose of predicting the logarithm of annual precipitation from daily mean temperature curve plotted in Figure 1(a).

(*Insert Figure 1 here.*)

In particular, we are interested in identifying the times of the year that have an effect on the logarithm of annual precipitation. We restrict the value of $\beta(t)$ at start of the year equal to that at end of the year and set $m = 3$ in equation (2.5) in the main paper as in James et al. (2009). Figure 1(b) displays the smoothing spline estimate of $\beta(t)$ on the active region selected by the FICf method (blue solid line), the smoothing estimate using Fourier basis with a roughness penalty for the full model (black dotted line), and the mean selecting frequency of the leave-one-out samples (red dashed line). The mean squared leave-one-out cross-validation error of the FICf method and the competing methods are reported in Table 7. In this application, the FICf method has a lower cross-validated error. The estimate of the null region is from late March to early October, which implies that the annual precipitation has no correlation with the daily temperature during this period. The smoothing spline estimate on the selected active region indicates a positive relationship from late October to early January in next year and a negative relationship form late January to late March. This

result interprets the positive relationship from December to January and the negative relationship in the spring months in Figure.5 (b) of James et al. (2009), which has been founded but has not been explored by James et al. (2009).

References

Dauxois, J., A. Pousse, and Y. Romain (1982). Asymptotic theory for the principal component analysis of a vector random function: some applications to statistical inference. *J. Multivariate Anal.* **12**(1), 136–154.

de Boor, C. (2001). *A practical guide to splines* (Revised ed.), Volume 27 of *Applied Mathematical Sciences*. Springer-Verlag, New York.

Grollemund, P.-M., C. Abraham, M. Baragatti, and P. Pudlo (2019). Bayesian functional linear regression with sparse step functions. *Bayesian Anal.* **14**(1), 111–135.

Horn, R. A. and C. R. Johnson (2013). *Matrix analysis* (Second ed.). Cambridge University Press, Cambridge.

James, G. M., J. Wang, and J. Zhu (2009). Functional linear regression that's interpretable. *Ann. Statist.* **37**(5A), 2083–2108.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: discrete index set. *Ann. Statist.* **15**(3), 958–975.

Lin, Y.-W., N. Xiao, L.-L. Wang, C.-Q. Li, and Q.-S. Xu (2017). Ordered homogeneity pursuit

lasso for group variable selection with applications to spectroscopic data. *Chemometr. Intell. Lab. Syst.* **168**, 62–71.

Lin, Z., J. Cao, L. Wang, and H. Wang (2017). Locally sparse estimator for functional linear regression models. *J. Comput. Graph. Statist.* **26**(2), 306–318.

Ramsay, J. O., G. Hooker, and S. Graves (2009). *Functional Data Analysis with R and MATLAB*. Use R! Springer, New York.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis* (Second ed.). Springer Series in Statistics. Springer, New York.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**(1), 45–54.

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

Table 1: Simulation results of the length of the symmetric difference $\Delta\lambda^*_\beta(\hat{M})$ in for Cases III – V. Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 100. FLiRTI: the method in James et al. (2009); Two-stage: the two-stage method in Zhou et al. (2013); SLoS: the smooth and locally Sparse method in Lin et al. (2017); Bliss: Bayesian functional Linear regression with Sparse Step functions in Grollemund et al. (2019); FICf: the proposed function information criterion method; FICf_0: similar to FICf but not using the smoothing penalty in region selection.

	FLiRTI	Two-stage	SLoS	Bliss*	FICf_0	FICf
Case III						
$n = 150$	13.3(16.1)	0.74(2.61)	4.99(15.8)	–	0.59(0.37)	0.41(2.65)
$n = 450$	12.3(16.5)	0.13(0.96)	3.76(13.2)	–	0.31(0.13)	0(0)
$n = 1000$	12.2(12.9)	0.11(0.79)	2.20(9.09)	–	0.20(0.68)	0(0)
Case IV						
$n = 150$	2.72(2.26)	3.37(8.94)	0.18(1.79)	1.91(4.15)	5.37(4.01)	2.34(5.52)
$n = 450$	1.88(1.65)	0.27(0.84)	0.00(0.00)	–	4.39(3.11)	0.79(1.53)
$n = 1000$	1.38(1.34)	0.17(0.45)	0.00(0.00)	–	4.26(4.10)	0.78(1.15)
Case V						
$n = 150$	41.7(9.24)	14.0(10.7)	9.21(5.62)	4.24(3.37)	38.4(10.4)	8.92(9.33)
$n = 450$	35.4(11.9)	10.1(8.86)	3.49(3.35)	–	11.7(8.37)	4.24(2.22)
$n = 1000$	35.2(11.5)	7.35(6.99)	1.89(1.27)	–	5.40(1.83)	4.57(1.04)

* The Bliss method does not work in Case III, $\beta(t) = 0$. We only report the results of the Bliss method with sample size of 150 due to computational cost.

Table 2: Simulation results of the integrated squared estimating errors for Cases I – V except for Case III. Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 100. Full: the smoothing spline estimate for the full model; FLiRTI: the method in James et al. (2009); Two-stage: the two-stage method in Zhou et al. (2013); SLoS: the smooth and locally Sparse method in Lin et al. (2017); Bliss-smooth: the smooth estimate of Bayesian functional Linear regression with Sparse Step functions in Grollemund et al. (2019); FICf: the smoothing spline estimate for the model selected by the proposed function information criterion method; FICf_0: similar to FICf but not using the smoothing penalty in region selection.

	Full	FLiRTI	Two-stage	SLoS	Bliss-smooth*	FICf_0	FICf
Case I							
$n = 150$	2.11(1.03)	9.62(4.65)	4.49(2.19)	1.85(0.94)	3.28(1.25)	9.54(8.95)	1.83(1.39)
$n = 450$	0.75(0.32)	3.23(1.32)	1.92(0.78)	0.76(0.40)	–	8.13(5.74)	0.64(0.43)
$n = 1000$	0.40(0.17)	1.58(0.61)	0.89(0.36)	0.35(0.18)	–	8.19(6.10)	0.33(0.26)
Case II							
$n = 150$	3.81(1.71)	9.85(4.30)	5.70(2.22)	3.20(1.26)	4.15(1.35)	7.29(5.12)	3.31(1.99)
$n = 450$	2.10(0.77)	4.45(4.77)	2.75(0.93)	1.81(0.37)	–	7.03(7.95)	1.88(1.06)
$n = 1000$	1.35(0.37)	2.10(0.70)	1.53(0.40)	1.46(0.21)	–	5.48(4.13)	1.26(0.53)
Case IV							
$n = 150$	11.9(6.00)	67.8(32.4)	34.4(24.1)	13.8(7.36)	46.3(13.3)	53.7(40.1)	13.6(9.82)
$n = 450$	4.73(2.74)	25.5(24.3)	10.6(4.64)	6.12(2.77)	–	43.9(31.1)	4.96(2.96)
$n = 1000$	2.42(1.46)	12.0(3.81)	5.30(1.70)	2.84(1.12)	–	42.6(41.0)	2.65(1.70)
Case V							
$n = 150$	14.6(8.21)	15.6(7.98)	14.3(6.76)	10.4(4.75)	11.4(3.58)	19.4(14.1)	12.3(9.55)
$n = 450$	5.66(1.38)	5.88(5.28)	6.07(2.44)	3.94(1.56)	–	13.3(14.4)	5.68(4.40)
$n = 1000$	3.23(0.69)	3.31(0.96)	3.54(1.59)	2.34(0.56)	–	7.56(7.39)	5.08(3.56)

* We only report the results of the Bliss method with sample size of 150 due to computational cost.

Table 3: Simulation results of the average squared prediction errors for Cases I – V except for Case III. Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 10^4. Full: the smoothing spline estimator for the full model; FLiRTI: the method in James et al. (2009); Two-stage: the two-stage method in Zhou et al. (2013); SLoS: the smooth and locally Sparse method in Lin et al. (2017); Bliss-smooth: the smooth estimate of Bayesian functional Linear regression with Sparse Step functions in Grollemund et al. (2019); FICf: the smoothing spline estimator for the model selected by the proposed function information criterion method; FICf_0: similar to FICf but not using the smoothing penalty in region selection.

	Full	FLiRTI	Two-stage	SLoS	Bliss-smooth*	FICf_0	FICf
Case I							
$n = 150$	2.88(1.30)	9.71(3.53)	4.85(2.07)	2.60(1.34)	5.36(2.48)	5.91(3.74)	2.36(1.52)
$n = 450$	1.04(0.42)	3.12(0.96)	2.00(0.71)	0.97(0.49)	–	3.07(1.34)	0.81(0.42)
$n = 1000$	0.51(0.21)	1.44(0.36)	0.84(0.30)	0.42(0.20)	–	2.00(6.75)	0.37(0.19)
Case II							
$n = 150$	4.49(1.90)	9.82(3.72)	5.95(2.34)	3.85(1.73)	6.16(2.61)	6.15(2.91)	3.56(2.02)
$n = 450$	2.08(1.72)	3.61(1.12)	2.45(0.84)	1.72(0.46)	–	3.14(1.53)	1.50(0.69)
$n = 1000$	1.14(0.33)	1.62(0.46)	1.15(0.35)	1.32(0.29)	–	1.75(0.62)	0.76(0.29)
Case IV							
$n = 150$	16.8(7.76)	66.9(32.3)	39.9(28.1)	19.5(9.84)	66.6(19.5)	37.7(22.0)	18.8(12.6)
$n = 450$	6.37(3.15)	22.6(6.71)	12.3(4.94)	7.95(3.42)	–	22.3(11.4)	6.46(3.31)
$n = 1000$	3.13(1.57)	10.8(2.59)	5.87(1.68)	3.59(1.41)	–	16.0(8.04)	3.21(1.74)
Case V							
$n = 150$	15.8(6.36)	17.3(7.64)	14.9(6.18)	12.1(5.75)	15.1(5.81)	17.4(8.92)	11.8(6.91)
$n = 450$	5.64(1.57)	5.22(2.14)	5.52(2.12)	4.16(1.92)	–	5.35(3.02)	3.33(1.60)
$n = 1000$	2.92(0.71)	2.58(0.94)	2.78(1.16)	2.08(0.60)	–	2.03(0.99)	1.83(0.54)

* We only report the results of the Bliss method with sample size of 150 due to computational cost.

Table 4: Simulation results of the length of the symmetric difference $\Delta\Delta\lambda^*_\beta(\hat{M})$ for Cases I–V with different (a_1, a_2) defined in (S6.2). Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 100.

(a_1, a_2)	(6, 1)	(6.5, 1)	(7, 1)	(7.5, 1)	(8, 1)	(6, 1.5)	(6.5, 1.5)	(7, 1.5)	(7.5, 1.5)	(8, 1.5)
Case I										
$n = 150$	7.75(7.40)	6.01(5.88)	4.93(4.25)	4.33(2.74)	4.24(2.43)	29.9(11.7)	28.3(12.0)	25.7(12.3)	23.2(12.5)	20.5(12.6)
$n = 450$	3.39(3.51)	2.78(2.15)	2.86(1.45)	3.41(1.67)	4.22(1.84)	21.0(12.0)	16.9(11.6)	13.4(10.5)	9.64(8.86)	7.16(7.11)
$n = 1000$	2.00(1.01)	2.10(1.10)	2.66(1.32)	3.47(1.44)	4.56(1.33)	16.4(11.9)	11.8(10.4)	7.78(8.22)	4.91(5.73)	3.15(3.79)
Case II										
$n = 150$	11.4(8.25)	9.07(6.58)	7.07(4.73)	5.85(3.32)	5.16(2.54)	29.5(10.9)	28.8(11.0)	26.4(11.4)	24.4(11.8)	21.6(12.0)
$n = 450$	4.91(4.65)	3.43(2.93)	3.81(1.93)	2.51(1.39)	2.27(0.90)	27.0(10.8)	23.9(11.0)	19.0(10.8)	14.8(9.69)	11.4(8.55)
$n = 1000$	3.15(2.03)	2.67(1.43)	2.35(1.11)	2.19(0.92)	2.31(0.90)	26.3(9.79)	21.3(10.2)	15.9(9.92)	10.6(7.97)	6.49(5.47)
Case III										
$n = 150$	1.46(6.60)	0.68(3.79)	0.41(2.65)	0.20(2.00)	0.14(1.43)	45.4(42.3)	33.0(39.3)	25.0(36.1)	17.2(31.4)	12.5(27.2)
$n = 450$	0.25(2.62)	0.09(1.21)	0.00(0.00)	0.00(0.00)	0.00(0.00)	19.3(32.5)	12.6(26.9)	6.91(19.2)	4.74(14.7)	2.13(7.95)
$n = 1000$	0.06(0.91)	0.06(0.91)	0.00(0.00)	0.00(0.00)	0.00(0.00)	11.1(21.7)	6.90(17.4)	3.26(11.3)	1.68(7.97)	0.88(4.96)
Case IV										
$n = 150$	0.88(2.58)	1.58(4.39)	2.34(5.52)	3.56(7.33)	7.32(10.5)	0.35(0.44)	0.36(0.44)	0.36(0.44)	0.44(1.05)	0.45(1.05)
$n = 450$	0.44(0.48)	0.54(0.61)	0.79(1.53)	1.37(2.91)	3.00(5.14)	0.26(0.35)	0.28(0.36)	0.29(0.38)	0.32(0.40)	0.34(0.44)
$n = 1000$	0.50(0.48)	0.64(1.13)	0.78(1.15)	1.02(1.56)	1.83(2.62)	0.25(0.36)	0.26(0.37)	0.29(0.38)	0.33(0.39)	0.39(0.43)
Case V										
$n = 150$	17.7(11.6)	12.7(11.2)	8.92(9.33)	6.73(7.09)	6.43(4.25)	43.0(8.25)	41.4(9.49)	40.0(10.1)	37.9(11.0)	34.5(12.1)
$n = 450$	8.27(7.37)	5.51(4.88)	4.24(2.22)	4.27(1.70)	6.14(4.78)	41.1(8.17)	37.7(9.41)	33.3(10.1)	27.7(10.8)	21.7(9.74)
$n = 1000$	5.78(3.92)	4.37(1.17)	4.57(1.04)	4.88(1.01)	6.24(3.59)	38.6(8.97)	33.5(9.48)	26.6(9.09)	18.7(9.27)	11.9(8.36)

Table 5: Simulation results of the integrated squared estimating errors for Cases I–V except Case III with different (a_1, a_2) defined in (S6.2). Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 100.

(a_1, a_2)	(6, 1)	(6.5, 1)	(7, 1)	(7.5, 1)	(8, 1)	(6, 1.5)	(6.5, 1.5)	(7, 1.5)	(7.5, 1.5)	(8, 1.5)
Case I										
$n = 150$	1.98(1.39)	1.92(1.47)	1.83(1.39)	1.76(1.28)	1.71(1.15)	2.13(1.09)	2.14(1.11)	2.16(1.13)	2.18(1.23)	2.17(1.23)
$n = 450$	0.66(0.45)	0.64(0.45)	0.64(0.43)	0.67(0.46)	0.72(0.47)	0.76(0.35)	0.75(0.34)	0.76(0.39)	0.74(0.39)	0.76(0.60)
$n = 1000$	0.33(0.45)	0.34(0.45)	0.33(0.26)	0.35(0.25)	0.40(0.28)	0.39(0.21)	0.38(0.22)	0.36(0.22)	0.35(0.23)	0.34(0.30)
Case II										
$n = 150$	3.54(2.05)	3.42(2.02)	3.31(1.99)	3.27(2.07)	3.39(2.36)	3.84(1.80)	3.83(1.81)	3.83(1.83)	3.83(1.85)	3.79(1.90)
$n = 450$	2.03(1.41)	1.96(1.06)	1.88(1.06)	2.01(1.07)	2.10(1.03)	2.06(0.79)	2.06(0.79)	2.07(0.89)	2.06(0.91)	2.04(0.92)
$n = 1000$	1.17(0.40)	1.20(0.43)	1.26(0.53)	1.36(0.65)	1.67(0.99)	1.35(0.39)	1.37(0.46)	1.35(0.48)	1.30(0.41)	1.20(0.42)
Case IV										
$n = 150$	12.4(7.16)	12.9(8.17)	13.6(9.82)	15.0(11.8)	19.3(16.0)	12.0(6.17)	12.0(6.17)	12.0(6.17)	12.1(6.22)	12.1(6.22)
$n = 450$	4.80(2.71)	4.84(2.75)	4.96(2.96)	5.35(4.00)	6.73(6.01)	4.74(2.66)	4.75(2.67)	4.75(2.68)	4.76(2.68)	4.77(2.70)
$n = 1000$	2.53(1.49)	2.58(1.65)	2.65(1.70)	2.79(2.08)	3.42(3.81)	2.48(1.46)	2.49(1.48)	2.48(1.46)	2.49(1.46)	2.51(1.49)
Case V										
$n = 150$	13.4(6.91)	12.8(8.30)	12.3(9.55)	11.9(9.56)	12.1(9.31)	14.6(7.74)	14.6(7.76)	14.6(7.79)	14.6(7.70)	14.6(7.83)
$n = 450$	5.67(3.42)	5.63(4.40)	5.68(4.40)	6.24(4.77)	7.83(6.06)	6.03(1.57)	6.12(1.63)	6.18(1.72)	6.25(1.87)	6.21(2.05)
$n = 1000$	4.32(3.25)	4.62(3.53)	5.08(3.56)	5.64(3.75)	7.14(5.64)	3.73(1.10)	3.84(1.15)	4.16(1.88)	4.26(2.62)	4.29(2.70)

Table 6: Simulation results of the average squared prediction errors in (S5.2) for Cases I–V except Case III with different (a_1, a_2) defined in (S6.2). Each entry is the Monte Carlo average of 200 simulation replicates. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 10^4.

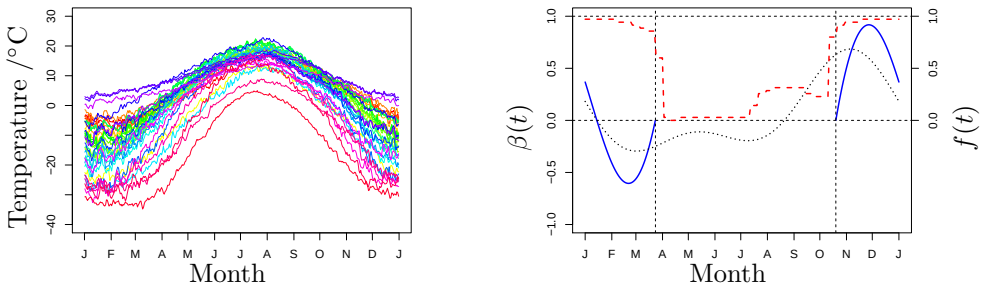
(a_1, a_2)	(6, 1)	(6.5, 1)	(7, 1)	(7.5, 1)	(8, 1)	(6, 1.5)	(6.5, 1.5)	(7, 1.5)	(7.5, 1.5)	(8, 1.5)
Case I										
$n = 150$	2.58(1.57)	2.48(1.63)	2.36(1.52)	2.27(1.35)	2.20(1.18)	2.87(1.31)	2.88(1.34)	2.91(1.37)	2.91(1.43)	2.90(1.43)
$n = 450$	0.85(0.47)	0.81(0.44)	0.81(0.42)	0.83(0.43)	0.80(0.44)	1.03(0.44)	1.01(0.44)	1.01(0.47)	0.98(0.47)	0.97(0.55)
$n = 1000$	0.37(0.21)	0.37(0.20)	0.37(0.19)	0.40(0.20)	0.42(0.22)	0.49(0.24)	0.47(0.24)	0.45(0.25)	0.42(0.24)	0.40(0.25)
Case II										
$n = 150$	3.91(2.12)	3.73(2.07)	3.56(2.02)	3.44(1.91)	3.42(1.92)	4.42(1.91)	4.42(1.92)	4.38(1.95)	4.36(1.97)	4.30(2.02)
$n = 450$	1.64(0.82)	1.54(0.72)	1.50(0.69)	1.48(0.70)	1.49(0.69)	2.00(0.76)	1.99(0.77)	1.96(0.79)	1.90(0.77)	1.83(0.78)
$n = 1000$	0.78(0.39)	0.76(0.29)	0.76(0.29)	0.76(0.31)	0.83(0.40)	1.10(0.34)	1.08(0.35)	1.05(0.36)	0.98(0.36)	0.90(0.31)
Case IV										
$n = 150$	17.3(9.15)	18.0(10.8)	18.8(12.6)	20.7(15.5)	26.1(22.0)	16.9(7.95)	16.9(7.96)	16.9(7.96)	16.9(8.03)	16.9(8.03)
$n = 450$	8.06(3.11)	6.36(3.14)	6.46(3.31)	6.77(4.27)	8.19(7.03)	6.32(3.08)	6.32(3.09)	6.33(3.10)	6.33(3.10)	6.34(3.11)
$n = 1000$	3.15(1.59)	3.18(1.70)	3.21(1.74)	3.30(2.07)	3.67(2.92)	3.14(1.57)	3.15(1.58)	3.14(1.57)	3.15(1.58)	3.15(1.59)
Case V										
$n = 150$	13.6(5.99)	12.5(6.34)	11.8(6.91)	11.1(6.72)	10.7(6.27)	14.6(7.74)	15.6(6.34)	15.6(6.38)	15.6(6.34)	15.4(6.51)
$n = 450$	4.03(1.85)	3.60(1.69)	3.33(1.60)	3.29(1.60)	4.20(1.89)	6.03(1.57)	5.59(1.67)	5.51(1.70)	5.40(1.72)	5.18(1.81)
$n = 1000$	2.00(0.74)	1.84(0.58)	1.83(0.54)	1.85(0.55)	2.39(1.95)	2.92(0.72)	2.90(0.73)	2.85(0.82)	2.67(0.88)	2.42(0.91)

Table 7: The mean squared leave-one-out cross-validated errors of different methods for the Canadian weather data. The corresponding standard deviation is reported in parentheses. All the values are multiplied by 100. OHPL: the ordered homogeneity pursuit LASSO method in Lin et al. (2017); Full: the smoothing estimate for the full model; FLiRTI: the method in James et al. (2009); Two-stage: the two-stage method in Zhou et al. (2013); SLoS: the smooth and locally Sparse method in Lin et al. (2017); Bliss-smooth: the smooth estimate of Bayesian functional Linear regression with Sparse Step functions in Grollemund et al. (2019); FICf: the smoothing spline estimate for the model selected by the proposed function information criterion method.

Full	OHPL	FLiRTI	Two-stage	SLoS	BLISS-smooth*	FICf
16.3(31.2)	15.9(31.6)	13.8(28.0)	–†	16.0(36.1)	53.6(24.8)	12.9(29.0)

* The within-sample mean squared errors are reported instead for the BLISS-smooth method.

† The full model is selected.



(a) (b)

Figure 1: (a) Daily temperature curves of 35 Canadian weather stations. (b) The smoothing spline estimate of the slope function β on the region selected by FICf (blue solid line), the smoothing estimate using the Fourier basis with a roughness penalty for the full model (black dotted line), and the mean selecting frequency of the leave-one-out samples (red dashed line).