SUPPLEMENTARY MATERIAL OF "ESTIMATION AND VARIABLE SELECTION UNDER THE FUNCTION-ON-SCALAR LINEAR MODEL WITH COVARIATE MEASUREMENT ERROR"

Yifan Sun[1], Grace Y. Yi[1,2,*]

[1]*Department of Statistical and Actuarial Sciences, University of Western Ontario, London, Ontario, Canada N6A 5B7*

[2]*Department of Computer Science, University of Western Ontario, London, Ontario, Canada N6A 5B7*

Regularity conditions, technical proofs, and simulation results are displayed in Section S1, Section S2 and Section S3, respectively. In Section S4 we provide additional notes on the real data analysis. In Section S5, we extend the development in the main text to accommodate the generalized least squares loss function, with both theoretical properties and numerical studies reported.

S1 Regularity Conditions

For a square matrix $\boldsymbol{A}$, let $\rho_{\min}(\boldsymbol{A})$ and $\rho_{\max}(\boldsymbol{A})$ denote the minimum and maximum eigenvalues of $\boldsymbol{A}$, respectively. The following regularity condi-

*Corresponding author: Grace Y. Yi, gyi5@uwo.ca

tions are imposed:

A. Assumptions about covariates and responses:

In addition to assuming $\{\{\boldsymbol{X}_i, Y_i(t)\} : i = 1, 2, \ldots, n\}$ to be i.i.d. for $t \in \mathcal{T}$, as in Section 2.1, we require that

A1. The dimension of $\boldsymbol{X}_i$, p, is fixed.

A2. $\mathbb{E}(X_{ij}^4) < \infty$ for $j = 1, \ldots, p$, and $\rho_{\min}\{\mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^\top)\} > 0$.

A3. $\mathcal{T}$ is a compact interval, as required in Section 2.1.

A4. $\boldsymbol{X}_i$ is independent of $\{\varepsilon_i(t) : t \in \mathcal{T}\}$, as required in Section 2.1.

B. Assumption about the coefficient functions in model (2.1):

There exist a positive integer $q' \leq d$ and a constant $C_1 > 0$, together with a constant $\nu > 0$, such that the q'th order derivative of functions $\beta_j(\cdot)$ satisfies

$$|\beta_j^{(q')}(t_1) - \beta_j^{(q')}(t_2)| \leq C_1 |t_1 - t_2|^\nu,$$

for all $t_1, t_2 \in \mathcal{T}$ and all $1 \leq j \leq p$. Define $q = q' + \nu$.

C. Assumptions about B-spline approximations:

C1. The number of knots satisfies

(i) $\frac{M^{2q}}{n} > C_2$ for some positive constant C_2;

(ii) $\frac{M}{n} \to 0$ as $n \to \infty$.

C2. For $j = 1, \ldots, M$, let $h_j = \tau_j - \tau_{j-1}$. Assume that

$$\max_{1 \le j \le M} |h_{j+1} - h_j| = o(M^{-1}) \text{ and } \max_{1 \le j \le M} h_j \le C_3 \cdot \min_{1 \le j \le M} h_j$$

for some positive constant C_3.

C3. There exists a positive constant $C_4 \ge 1$ such that

$$C_4^{-1} m/M \le \rho_{\min}(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})) \le \rho_{\max}(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})) \le C_4 m/M. \tag{S1.1}$$

These requirements imply the invertibility of $\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})$, yielding that $m \ge M + d$ since $\boldsymbol{\Phi}(\boldsymbol{t})$ is an $m \times (M + d)$ matrix.

D. Assumptions about measurement error $\boldsymbol{U}_i$ in model (3.7):

D1. For any $1 \le j \le p$, assume that $\mathbb{E}(U_{ij}^4) < \infty$.

D2. For $i = 1, \ldots, n$, $\boldsymbol{U}_i$ is independent of $\{Y_i(t) : t \in \mathcal{T}\}$ and $\boldsymbol{X}_i$, as required in Section 3.1.

E. Assumptions about the penalty function $P_\lambda(v)$ for $v \ge 0$:

E1. $P_\lambda(0) = 0$ and $P_\lambda(v) > 0$ for $v > 0$.

E2. The first derivative of $P_\lambda(v)$, denoted $P'_\lambda(v)$, exists on $(0, +\infty)$, and satisfies

(i) $P'_\lambda(v)$ is continuous and monotonically non-increasing;

(ii) $P'_\lambda(v) = 0$ for $v > a\lambda$ for some constant $a > 1$;

(iii) $\lim_{v\to 0^+} P'_\lambda(v) = \lambda$.

Assumption A1 states the setting we consider. Assumption A2 is imposed to avoid settings with the perfect covariate collinearity, which results in nonidentifiability issues; this assumption is widely used in the literature of function-on-scalar regression models (Wang et al., 2007; Parodi and Reimherr, 2018). Indeed, if $\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^\mathsf{T})$ is singular, there exists a non-zero vector $\boldsymbol{\gamma} \in \mathbb{R}^p$ such that $\boldsymbol{\gamma}^\mathsf{T}\boldsymbol{X}_i = 0$ almost surely. Then any function of the form $\boldsymbol{\beta}(t) + c\boldsymbol{f}_{\boldsymbol{\gamma}}(t)$ satisfies model (2.1), where $c \in \mathbb{R}$ is an arbitrary constant and $\boldsymbol{f}_{\boldsymbol{\gamma}}(t) \equiv \boldsymbol{\gamma}$.

Assumption B is commonly made for the function-on-scalar regression model, which allows the use of B-spline based methods to estimate the coefficient functions in the model. In fact, such $\beta_j(t)$ is smooth enough and can be approximated by some "best" function in the linear space spanned by all B-spline basis functions $\boldsymbol{\phi}(t)$, denoted $S_{d,M}$. To be specific, together with Condition C2, there always exist B-spline functions $\{\theta_j(t) \in S_{d,M} : j = 1, \ldots, p,\, t \in \mathcal{T}\}$, such that

$$\sup_{1\le j\le p} \|\theta_j - \beta_j\|_{L_\infty} \le C_1' M^{-q}, \tag{S1.2}$$

for any positive integer M, where C_1' is a positive constant functionally independent of M (De Boor, 1978, Theorem XII(6)), and thus yielding that $\|\theta_j - \beta_j\|_{L_\infty} \to 0$ as $M \to \infty$. For a B-spline function $\theta_j(t) \in S_{d,M}$

satisfying (S1.2), we write

$$\theta_j(t) = \boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{b}_{0,j}, \tag{S1.3}$$

where $\boldsymbol{b}_{0,j} \in \mathbb{R}^{M+d}$. Let $\boldsymbol{b}_0 = (\boldsymbol{b}_{0,1}^{\mathsf{T}}, \ldots, \boldsymbol{b}_{0,p}^{\mathsf{T}})^{\mathsf{T}}$ denote the $(M+d)p \times 1$ vector. In order to establish the estimation consistency of $\hat{\beta}_j$, we therefore examine $\|\theta_j - \hat{\beta}_j\|_{L_\infty}$, and it suffices to find an upper bound of $\|\theta_j - \hat{\beta}_j\|_{L_\infty}$ which converges to 0 as $n \to \infty$, as discussed in the following derivation of Theorem 1.

Generally speaking, the number of B-spline inner knots should be divergent to derive the function consistency, as shown in (S1.2). However, a too large M results in an estimator with an overly large variance. Imposing condition C1 is to constrain the divergence rate of M. The imposition of Condition C(i) is to control the B-spline approximation error, which makes the terms involving the B-spline error (i.e., I_{22} and I_{25} in (S2.28) of the proof of Theorem 1) negligible, relative to other terms. Without this condition, the convergence rate for $\hat{\boldsymbol{b}}$ in Theorem 1 is not $O_p(\sqrt{M/n})$ but becomes $O_p(\sqrt{M/n} + M^{1/2-q})$. Condition C2 is identical to Assumption 3 of Zhou et al. (1998), which is easy to be met. For instance, the equally spaced knots satisfies Condition C2.

While constraint (S1.1) in Condition C3 may look somewhat stringent, they are readily satisfied for many settings. In fact, to satisfy (S1.1), it

suffices to require $\boldsymbol{a}^{\mathsf{T}}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})/m\}\boldsymbol{a}$ to be of order $O(M^{-1})$ for any vector $\boldsymbol{a}$ with $\|\boldsymbol{a}\| = 1$. To this end, let $s(t) = \boldsymbol{a}^{\mathsf{T}}\boldsymbol{\phi}(t)$ for $t \in \mathcal{T}$. Then we write

$$\boldsymbol{a}^{\mathsf{T}}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})/m\}\boldsymbol{a} = \int_{\mathcal{T}} s(t)^2 \mathrm{d}F_m(t), \tag{S1.4}$$

where $F_m(t) \triangleq \frac{1}{m}\sum_{k=1}^{m} \mathbb{I}(t_k \leq t)$ is the step function of observations $\boldsymbol{t} = \{t_1, \ldots, t_m\}$. Suppose that there is a distribution function $F(t)$ with positive continuous density on the domain $\mathcal{T}$, which satisfies

$$\|F_m - F\|_{L_\infty} = o(M^{-1}). \tag{S1.5}$$

On one hand, it can be proved that $\int_{\mathcal{T}} s(t)^2 \mathrm{d}F(t) = O(M^{-1})$ (De Boor, 1978, Page 155). On the other hand, we have that $\int_{\mathcal{T}} s(t)^2 \mathrm{d}(F_m - F)(t) = o(M^{-1})$ due to (S1.5). Therefore, the right-hand-side of equation (S1.4) is of order $O(M^{-1})$. The details of this proof can be found in Zhou et al. (1998, Lemma 6.1).

Assumption D1 is useful for technical derivations, and it is commonly made in the literature of measurement error models in a tacit manner, for example, by assuming the normality of $\boldsymbol{U}_i$ in model (3.7) (e.g., Section 2.6 of Yi (2017)). Assumptions in E are widely used in the literature of variable selection. The well-known SCAD (Fan and Li, 2001) and MCP (Zhang, 2010) are two examples satisfying them.

Remark 1. Condition (S1.5) is easy to be met. For example, if the do-

main $\mathcal{T} \triangleq [0, 1]$ and the observations are equally spaced at $t_k = k/m$ for $k = 1, \ldots, m$, then we take F as the uniform distribution function on $[0, 1]$, which satisfies $\|F_m - F\|_{L_\infty} = m^{-1}$. Hence (S1.5) holds under the assumption that $M/m \to 0$ as $m \to \infty$.

Remark 2. The results in Section 5 of the main text are established by treating the observation points $\boldsymbol{t} = \{t_1, \ldots, t_m\}$ as deterministic or pre-specified. When the observation times are random variables, the results can still hold if proper modifications of the conditions are done; typically, the inequalities (S1.1) is now modified to hold with probability tending to 1. This new constraint can also be proved if $\|F_m - F\|_{L_\infty} = o_p(M^{-1})$, where the definitions of F_m and F are the same as before. In a special case where $\{t_1, \ldots, t_m\}$ are independently sampled from a common distribution function, say, $F_0(t)$, then we have that $\|F_m - F_0\|_{L_\infty} = O_p(m^{-1/2})$ according to Massart (1990). Therefore, the condition $\|F_m - F\|_{L_\infty} = o_p(M^{-1})$ is satisfied if we set $F = F_0$ and assume $M^2/m \to 0$ as $m \to \infty$.

S2 Technical Details

S2.1 Preliminary Preparations

For a vector $\boldsymbol{a} = (a_1, \ldots, a_k)^\mathsf{T}$, let $\|\boldsymbol{a}\|_\infty$ and $\|\boldsymbol{a}\|$ respectively denote the infinity norm and the Euclidean norm for $\boldsymbol{a}$, i.e., $\|\boldsymbol{a}\|_\infty = \max_{1\le j\le k} |a_j|$ and $\|\boldsymbol{a}\| = \sqrt{\sum_{j=1}^k a_j^2}$. For a $p \times q$ matrix $\boldsymbol{A} = [a_{jk}]_{p\times q}$ with the (j,k) element $a_{jk} \in \mathbb{R}$, let $\|\boldsymbol{A}\|_F$ and $\|\boldsymbol{A}\|_2$ respectively represent the Frobenius norm and the matrix 2-norm (i.e., the spectral norm) for $\boldsymbol{A}$, i.e., $\|\boldsymbol{A}\|_F = \sqrt{\sum_{j=1}^p \sum_{k=1}^q a_{jk}^2}$ and $\|\boldsymbol{A}\|_2 = \sqrt{\rho_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})}$. An equivalent definition for the matrix 2-norm is given by

$$\|\boldsymbol{A}\|_2 = \sup_{\|\boldsymbol{w}\|=1} \|\boldsymbol{A}\boldsymbol{w}\|, \tag{S2.6}$$

where $\boldsymbol{w}$ is a vector with suitable dimension.

In the subsequent derivations, we frequently use the following properties, together with Lemmas 1 and 2.

Property 1: For matrices $\boldsymbol{A}$, $\boldsymbol{B}$, $\boldsymbol{C}$ and $\boldsymbol{D}$ having suitable dimensions,

$$(\boldsymbol{A}\otimes\boldsymbol{B})^\mathsf{T} = \boldsymbol{A}^\mathsf{T}\otimes\boldsymbol{B}^\mathsf{T} \text{ and } (\boldsymbol{A}\otimes\boldsymbol{C})(\boldsymbol{B}\otimes\boldsymbol{D}) = (\boldsymbol{A}\boldsymbol{C})\otimes(\boldsymbol{B}\boldsymbol{D}).$$

Property 2: For any matrices $\boldsymbol{A}$, $\boldsymbol{B}$ and $\boldsymbol{C}$ with suitable dimensions,

$$(\boldsymbol{B}^\mathsf{T}\otimes\boldsymbol{A})\,\mathrm{vec}(\boldsymbol{C}) = \mathrm{vec}(\boldsymbol{A}\boldsymbol{C}\boldsymbol{B}).$$

Property 3: For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$,

$$\|\boldsymbol{A} \otimes \boldsymbol{B}\|_2 = \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_2.$$

Property 4: For any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ with suitable dimensions, we have that

(a) $\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_2 \|\boldsymbol{B}\|_F$ and $\|\boldsymbol{AB}\|_F \leq \|\boldsymbol{A}\|_F \|\boldsymbol{B}\|_2$.

(b) $\|\boldsymbol{A}\|_2 \leq \|\boldsymbol{A}\|_F$.

Property 5: For any symmetric positive semi-definite matrix $\boldsymbol{G}$,

$$\rho_{\max}(\boldsymbol{G}^2) = \{\rho_{\max}(\boldsymbol{G})\}^2.$$

Property 6: For positive semi-definite matrices $\boldsymbol{S}$ and $\boldsymbol{T}$,

$$\rho_{\min}(\boldsymbol{S} \otimes \boldsymbol{T}) = \rho_{\min}(\boldsymbol{S}) \rho_{\min}(\boldsymbol{T}).$$

Properties 1 and 2 can be found in Horn and Johnson (1991, Section 4.2) or proved by using the definition of the kronecker product. Properties 3 and 6 are directly resulted from Theorem 4.2.12 of Horn and Johnson (1991) and Properties 4 and 5 can be found in Horn and Johnson (2012, Section 5.6). We also present rigorous proofs for Properties 3-6 below.

Proof of Property 3, 4, 5 and 6:

For any rectangular matrices $\boldsymbol{S}_1$ and $\boldsymbol{T}_1$, let the nonzero singular values of $\boldsymbol{S}_1$ and $\boldsymbol{T}_1$ be $\{\lambda_1, \ldots, \lambda_s\}$ and $\{\mu_1, \ldots, \mu_t\}$, respectively. According to Theorem 4.2.12 of Horn and Johnson (1991), all nonzero singular values for $\boldsymbol{S}_1 \otimes \boldsymbol{T}_1$ is given by $\{\lambda_j \mu_k : 1 \leq j \leq s,\, 1 \leq k \leq t\}$. Hence for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, we have that

$$\begin{aligned} \|\boldsymbol{A} \otimes \boldsymbol{B}\|_2^2 &= \rho_{\max}\{(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) \otimes (\boldsymbol{B}^\mathsf{T}\boldsymbol{B})\} \\ &= \rho_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) \cdot \rho_{\max}(\boldsymbol{B}^\mathsf{T}\boldsymbol{B}) \\ &= \|\boldsymbol{A}\|_2^2 \|\boldsymbol{B}\|_2^2, \end{aligned}$$

where the first step is due to Property 1 and the definition of spectral norm, and the second step is because $\boldsymbol{A}^\mathsf{T}\boldsymbol{A}$ and $\boldsymbol{B}^\mathsf{T}\boldsymbol{B}$ are positive semi-definite, indicating that all eigenvalues are non-negative. Property 3 is then proved.

Similarly, for two positive semi-definite matrices $\boldsymbol{S}$ and $\boldsymbol{T}$,

$$\rho_{\min}(\boldsymbol{S} \otimes \boldsymbol{T}) = \rho_{\min}(\boldsymbol{S})\rho_{\min}(\boldsymbol{T}),$$

since their eigenvalues are all non-negative. Property 6 is proved.

Note that the Frobenius norm for matrix $\boldsymbol{A}$ can be rewrite as $\|\boldsymbol{A}\|_F = \sqrt{\mathrm{tr}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A})}$, where $\mathrm{tr}(\cdot)$ represents the trace operator. Property 4 (b) then directly follows since

$$\|\boldsymbol{A}\|_F^2 = \mathrm{tr}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) \geq \rho_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) = \|\boldsymbol{A}\|_2^2.$$

For Property 4 (a), let $\boldsymbol{B} = [\boldsymbol{B}_1, \ldots, \boldsymbol{B}_p]$, where $\boldsymbol{B}_j$ for $j = 1, \ldots, p$ are

column vectors of $\boldsymbol{A}$. Then we have that

$$\begin{aligned}\|\boldsymbol{AB}\|_F^2 &= \mathrm{tr}(\boldsymbol{B}^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{AB}) \\ &= \sum_{j=1}^{p} \boldsymbol{B}_j^\mathsf{T}\boldsymbol{A}^\mathsf{T}\boldsymbol{AB}_j \\ &\leq \rho_{\max}(\boldsymbol{A}^\mathsf{T}\boldsymbol{A}) \sum_{j=1}^{p} \boldsymbol{B}_j^\mathsf{T}\boldsymbol{B}_j \\ &= \|\boldsymbol{A}\|_2^2 \|\boldsymbol{B}\|_F^2.\end{aligned}$$

By the facts that $\|\boldsymbol{AB}\|_F$ also equals to $\mathrm{tr}(\boldsymbol{ABB}^\mathsf{T}\boldsymbol{A}^\mathsf{T})$ and $\rho_{\max}(\boldsymbol{BB}^\mathsf{T}) = \|\boldsymbol{B}\|_2^2$, we can prove the other inequality analogously.

Now we verify Property 5. Since $\boldsymbol{G}$ is symmetric, there exists a a real orthogonal matrix $\boldsymbol{P}$ such that

$$\boldsymbol{G} = \boldsymbol{P}^\mathsf{T}\boldsymbol{\Lambda}\boldsymbol{P},$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ is the diagonal matrix with λ_j, $1 \leq j \leq m$, representing the ith eigenvalue of $\boldsymbol{G}$. Without loss of generality, we assume that $\lambda_1 = \rho_{\max}(\boldsymbol{G})$. It is then obvious that

$$\rho_{\max}(\boldsymbol{G}^2) = \rho_{\max}(\boldsymbol{P}^\mathsf{T}\boldsymbol{\Lambda}^2\boldsymbol{P}) = \lambda_1^2 = \{\rho_{\max}(\boldsymbol{G})\}^2,$$

where the first step is due to that $\boldsymbol{PP}^\mathsf{T}$ equals identity and the second step is because that all $\lambda_j \geq 0$, $j = 1, \ldots, m$. □

Lemma 1. *Let $\boldsymbol{X}$ and $\boldsymbol{Z}$ be two independent column vector variables with mean $\boldsymbol{0}$ and covariance matrices $\boldsymbol{\Sigma_X}$ and $\boldsymbol{\Sigma_Z}$ respectively. For any real-*

valued matrix $\boldsymbol{C}$ *with a suitable dimension, we have that*

$$\mathrm{Cov}\{(\boldsymbol{X}\otimes\boldsymbol{C})\boldsymbol{Z}\}=\boldsymbol{\Sigma}_{\boldsymbol{X}}\otimes(\boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{C}^{\mathsf{T}}).$$

Proof:

Let $\boldsymbol{X}=(X_1,\ldots,X_p)^{\mathsf{T}}$ and let $\boldsymbol{W}_j=X_j\boldsymbol{C}\boldsymbol{Z}$ for $j=1,\ldots,p$. Then $(\boldsymbol{X}\otimes\boldsymbol{C})\boldsymbol{Z}=[\boldsymbol{W}_1^{\mathsf{T}},\cdots,\boldsymbol{W}_p^{\mathsf{T}}]^{\mathsf{T}}$. It is direct to verify that

$$\mathrm{Var}(\boldsymbol{W}_j)=\mathbb{E}(X_j^2)\boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{C}^{\mathsf{T}}$$

and

$$\mathrm{Cov}(\boldsymbol{W}_j,\boldsymbol{W}_l)=\mathrm{Cov}(X_j,X_l)\boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{C}^{\mathsf{T}}.$$

It is then straightforward to combine these two results and obtain that

$$\mathrm{Cov}([\boldsymbol{W}_1^{\mathsf{T}},\cdots,\boldsymbol{W}_p^{\mathsf{T}}]^{\mathsf{T}})=\boldsymbol{\Sigma}_{\boldsymbol{X}}\otimes(\boldsymbol{C}\boldsymbol{\Sigma}_{\boldsymbol{Z}}\boldsymbol{C}^{\mathsf{T}}).$$

□

Lemma 2. *Let* $\boldsymbol{\Sigma}$ *be an* $m\times m$ *real-valued symmetric positive definite matrix and let* $\boldsymbol{\Psi}$ *be an* $m\times d$ *real-valued matrix with the full rank, where* $d\le m$. *Then*

$$\rho_{min}(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\Psi})\ge\rho_{min}(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi})\rho_{min}(\boldsymbol{\Sigma}),\tag{S2.7}$$

and

$$\rho_{max}(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Sigma}\boldsymbol{\Psi})\le\|\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}\|_2\rho_{max}(\boldsymbol{\Sigma}).\tag{S2.8}$$

Proof:

Since $\boldsymbol{\Sigma}$ is real-valued and symmetric, there exists a real-valued orthogonal matrix $\boldsymbol{P}$ such that

$$\boldsymbol{\Sigma} = \boldsymbol{P}^{\mathsf{T}}\boldsymbol{\Lambda}\boldsymbol{P},$$

where $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_m)$ is the diagonal matrix with λ_j representing an eigenvalue of $\boldsymbol{\Sigma}$ for $1 \leq j \leq m$.

Define $\tilde{\boldsymbol{\Psi}} = \boldsymbol{P}\boldsymbol{\Psi}$ and write $\tilde{\boldsymbol{\Psi}} = [\boldsymbol{\psi}_1, \cdots, \boldsymbol{\psi}_m]^{\mathsf{T}}$. Since $\boldsymbol{P}^{\mathsf{T}}\boldsymbol{P}$ equals the identity matrix, proving (S2.7) is equivalent to showing that

$$\rho_{\min}(\tilde{\boldsymbol{\Psi}}^{\mathsf{T}}\boldsymbol{\Lambda}\tilde{\boldsymbol{\Psi}}) \geq \rho_{\min}(\tilde{\boldsymbol{\Psi}}^{\mathsf{T}}\tilde{\boldsymbol{\Psi}})\rho_{\min}(\boldsymbol{\Lambda}), \tag{S2.9}$$

which is true because that

$$\begin{aligned}
\rho_{\min}(\tilde{\boldsymbol{\Psi}}^{\mathsf{T}}\boldsymbol{\Lambda}\tilde{\boldsymbol{\Psi}}) &= \rho_{\min}\left(\sum_{j=1}^{m}\lambda_j\boldsymbol{\psi}_j\boldsymbol{\psi}_j^{\mathsf{T}}\right) \\
&= \inf_{\|\boldsymbol{w}\|=1}\sum_{j=1}^{m}\lambda_j(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\psi}_j)^2 \\
&\geq \inf_{\|\boldsymbol{w}\|=1}\sum_{j=1}^{m}(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{\psi}_j)^2\rho_{\min}(\boldsymbol{\Lambda}) \\
&= \rho_{\min}(\boldsymbol{\Lambda}) \cdot \rho_{\min}\left(\sum_{j=1}^{m}\boldsymbol{\psi}_j\boldsymbol{\psi}_j^{\mathsf{T}}\right) \\
&= \rho_{\min}(\tilde{\boldsymbol{\Psi}}^{\mathsf{T}}\tilde{\boldsymbol{\Psi}})\rho_{\min}(\boldsymbol{\Lambda}),
\end{aligned}$$

where the second and the fourth steps are due to the fact that $\rho_{\min}(\boldsymbol{Q}) = \inf_{\|\boldsymbol{w}\|=1}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{Q}\boldsymbol{w}$ for any square matrix $\boldsymbol{Q}$. The proof of (S2.7) is then completed.

Expression (S2.8) can be proved using arguments analogous to those for (S2.7) and the identity $\rho_{\max}(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}) = \|\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}\|_2$. □

Lemma 3. *Consider the B-spline basis functions $\boldsymbol{\phi}(t)$ in Section 2.2 and the linear space $S_{d,M}$ spanned by them given before (S1.2). For $t \in \mathcal{T}$, let $f(t) = \boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{b} \in S_{d.M}$ denote a spline function over domain $\mathcal{T}$ with coefficients $\boldsymbol{b} \in \mathbb{R}^{M+d}$. Assume Assumption C2. Then there exists positive constants E_1 and E_2 depending only on d such that*

$$E_1\sqrt{M}\|f\|_{L_2} \leq \|\boldsymbol{b}\| \leq E_2\sqrt{M}\|f\|_{L_2}.$$

Proof:

In Section 2.2, we use $\tau_1 < \tau_2 < \cdots < \tau_{M+1}$ to denote knots on $\mathcal{T}$ for the setup of B-spline, where τ_1 and τ_{M+1} represent the two endpoints of $\mathcal{T}$. Define $\tau_k = \tau_1$ for all $k = 1-d, 1-d+1, \ldots, 0$ and $\tau_k = \tau_{M+1}$ for all $k = M+2, \ldots, M+2d+1$. Let $\boldsymbol{b} = (b_1, \ldots, b_{M+d})^{\mathsf{T}}$. According to equation (13) in Zhou et al. (1998), there exist positive constants $\tilde{E}_1$ and $\tilde{E}_2$ depending only on d, such that

$$\tilde{E}_1 \sum_{j=1}^{M+d} (\tau_{j+1} - \tau_{j-d}) b_j^2 \leq \|f\|_{L_2}^2 \leq \tilde{E}_2 \sum_{j=1}^{M+d} (\tau_{j+1} - \tau_{j-d}) b_j^2. \tag{S2.10}$$

Due to Assumption C2 and the fact that $\min_{1\leq j\leq M} h_j \leq 1/M$,

$$\max_{1\leq j\leq M} h_j \leq C_3 \min_{1\leq j\leq M} h_j \leq C_3/M.$$

Similarly, since $\max_{1\le j\le M} h_j \ge 1/M$,

$$\min_{1\le j\le M} h_j \ge C_3^{-1} \max_{1\le j\le M} h_j \ge 1/(C_3 M).$$

Hence, we have that for any $j = 1, \ldots, M$,

$$1/(C_3 M) \le \min_{1\le j\le M} h_j \le \tau_{j+1} - \tau_{j-d} \le (d+1) \max_{1\le j\le M} h_j \le C_3(d+1)/M. \tag{S2.11}$$

Then by (S2.10) and (S2.11),

$$\frac{\tilde{E}_1}{C_3 M}\|\boldsymbol{b}\|^2 \le \|f\|_{L_2}^2 \le \frac{C_3(d+1)\tilde{E}_2}{M}\|\boldsymbol{b}\|^2,$$

which completes the proof. □

S2.2 Proof of Theorem 1

The proof consists of two parts, each for proving (5.23) or (5.24).

Part 1: Prove (5.23).

Set $\delta_n = \sqrt{\frac{M}{n}}$. To show the existence of a local minimizer, say $\hat{\boldsymbol{b}}$, of $Q_n(\boldsymbol{b})$ satisfying $\|\hat{\boldsymbol{b}} - \boldsymbol{b}_0\| = O_p(\delta_n)$, it suffices to verify that, for any $\epsilon > 0$, there exists a positive constant K_ϵ such that

$$\mathbb{P}\left(\inf_{\|\boldsymbol{v}\|=K_\epsilon, \boldsymbol{v}\in\mathbb{R}^{p(M+d)}} Q_n(\boldsymbol{b}_0 + \delta_n \boldsymbol{v}) > Q_n(\boldsymbol{b}_0) \right) > 1 - \epsilon. \tag{S2.12}$$

The proof of (S2.12) consists of the following steps, where for any given constant $K > 0$, we consider any $\boldsymbol{v}$ satisfying $\|\boldsymbol{v}\| = K$.

Step 1: Identify a lower bound of $Q_n(\boldsymbol{b}_0 + \delta_n \boldsymbol{v}) - Q_n(\boldsymbol{b}_0)$.

For $i = 1, \ldots, n$, let

$$\boldsymbol{Z}_i = \boldsymbol{X}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}) \text{ and } \boldsymbol{Z}_i^* = \boldsymbol{X}_i^{*\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}). \tag{S2.13}$$

By (3.11),

$$Q_n(\boldsymbol{b}) = \frac{1}{2} \sum_{i=1}^{n} \|\boldsymbol{Y}_i(\boldsymbol{t}) - \boldsymbol{Z}_i^* \boldsymbol{b}\|^2 - \frac{n}{2} \boldsymbol{b}^{\mathsf{T}} \left[\boldsymbol{\Sigma} \otimes \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{b} + nm \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|).$$

Therefore,

$$
\begin{aligned}
Q_n(\boldsymbol{b}_0+\delta_n\boldsymbol{v})-Q_n(\boldsymbol{b}_0) &= \frac{1}{2}\sum_{i=1}^{n}\left\{\|\boldsymbol{Y}_i(\boldsymbol{t})-\boldsymbol{Z}_i^*(\boldsymbol{b}_0+\delta_n\boldsymbol{v})\|^2-\|\boldsymbol{Y}_i(\boldsymbol{t})-\boldsymbol{Z}_i^*\boldsymbol{b}_0\|^2\right\}\\
&\quad+nm\sum_{j=1}^{p}\left\{P_\lambda(\|\boldsymbol{b}_{0,j}+\delta_n\boldsymbol{v}_j\|)-P_\lambda(\|\boldsymbol{b}_{0,j}\|)\right\}\\
&\quad-\frac{n}{2}(\boldsymbol{b}_0+\delta_n\boldsymbol{v})^{\mathsf{T}}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right](\boldsymbol{b}_0+\delta_n\boldsymbol{v})\\
&\quad+\frac{n}{2}\boldsymbol{b}_0^{\mathsf{T}}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{b}_0\\
&= \frac{\delta_n^2}{2}\sum_{i=1}^{n}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{Z}_i^{*\mathsf{T}}\boldsymbol{Z}_i^*\boldsymbol{v}-\delta_n\sum_{i=1}^{n}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{Z}_i^{*\mathsf{T}}(\boldsymbol{Y}_i(\boldsymbol{t})-\boldsymbol{Z}_i^*\boldsymbol{b}_0)\\
&\quad+nm\sum_{j=1}^{p}\left\{P_\lambda(\|\boldsymbol{b}_{0,j}+\delta_n\boldsymbol{v}_j\|)-P_\lambda(\|\boldsymbol{b}_{0,j}\|)\right\}\\
&\quad-n\delta_n\boldsymbol{b}_0^{\mathsf{T}}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}-\frac{n\delta_n^2}{2}\boldsymbol{v}^{\mathsf{T}}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}\\
&\geq \frac{\delta_n^2}{2}\sum_{i=1}^{n}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{Z}_i^{*\mathsf{T}}\boldsymbol{Z}_i^*\boldsymbol{v}-\delta_n\sum_{i=1}^{n}\boldsymbol{v}^{\mathsf{T}}\boldsymbol{Z}_i^{*\mathsf{T}}(\boldsymbol{Y}_i(\boldsymbol{t})-\boldsymbol{Z}_i^*\boldsymbol{b}_0)\\
&\quad+nm\sum_{j\notin J_0}\left\{P_\lambda(\|\boldsymbol{b}_{0,j}+\delta_n\boldsymbol{v}_j\|)-P_\lambda(\|\boldsymbol{b}_{0,j}\|)\right\}\\
&\quad-n\delta_n\boldsymbol{b}_0^{\mathsf{T}}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}-\frac{n\delta_n^2}{2}\boldsymbol{v}^{\mathsf{T}}\left[\boldsymbol{\Sigma}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}\\
&\triangleq I_1+I_2+I_3+I_4+I_5,
\end{aligned}
\tag{S2.14}
$$

where $\boldsymbol{v}_j$ is the jth part of $\boldsymbol{v}$ such that $\boldsymbol{v}=(\boldsymbol{v}_1^{\mathsf{T}},\ldots,\boldsymbol{v}_p^{\mathsf{T}})^{\mathsf{T}}$. The inequality is due to Assumption E1 and the fact that $\boldsymbol{b}_{0,j}=\boldsymbol{0}$ for all $j\in J_0$.

Step 2: Examine I_1 and I_5 in (S2.14).

By definition and model (3.7),

$$\boldsymbol{Z}_i^* = \boldsymbol{Z}_i + \boldsymbol{U}_i^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t}).$$

Hence the term I_1 in (S2.14) can be further split as

$$\begin{aligned} I_1 &= \frac{\delta_n^2}{2} \sum_{i=1}^n \boldsymbol{v}^\mathsf{T} \boldsymbol{Z}_i^{*\mathsf{T}} \boldsymbol{Z}_i^* \boldsymbol{v} \\ &= \frac{\delta_n^2}{2} \sum_{i=1}^n \boldsymbol{v}^\mathsf{T} \mathbb{E}(\boldsymbol{Z}_i^\mathsf{T} \boldsymbol{Z}_i) \boldsymbol{v} + \delta_n^2 \sum_{i=1}^n \boldsymbol{v}^\mathsf{T} \boldsymbol{Z}_i^\mathsf{T} \{\boldsymbol{U}_i^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t})\} \boldsymbol{v} \\ &\quad + \frac{\delta_n^2}{2} \sum_{i=1}^n \boldsymbol{v}^\mathsf{T} \left[(\boldsymbol{U}_i \boldsymbol{U}_i^\mathsf{T}) \otimes \left\{ \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{v} \\ &\quad + \frac{\delta_n^2}{2} \sum_{i=1}^n \boldsymbol{v}^\mathsf{T} \left\{ \boldsymbol{Z}_i^\mathsf{T} \boldsymbol{Z}_i - \mathbb{E}(\boldsymbol{Z}_i^\mathsf{T} \boldsymbol{Z}_i) \right\} \boldsymbol{v} \\ &\triangleq I_{1,1} + I_{1,2} + I_{1,3} + I_{1,4}. \end{aligned} \tag{S2.15}$$

Now we examine (S2.15) term by term.

For $I_{1,1}$ in (S2.15), we have that

$$\begin{aligned} I_{1,1} &= \frac{\delta_n^2}{2} \sum_{i=1}^n \boldsymbol{v}^\mathsf{T} \mathbb{E}(\boldsymbol{Z}_i^\mathsf{T} \boldsymbol{Z}_i) \boldsymbol{v} \\ &= \frac{n\delta_n^2}{2} \boldsymbol{v}^\mathsf{T} \left[\mathbb{E}\left(\boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T} \right) \otimes \left\{ \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{v} \\ &\geq \frac{n\delta_n^2}{2} \|\boldsymbol{v}\|^2 \frac{m}{M} \rho_{\min} \left\{ \mathbb{E}(\boldsymbol{X}_i \boldsymbol{X}_i^\mathsf{T}) \right\} \rho_{\min} \left\{ \frac{M}{m} \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \\ &\geq \delta_n^2 K^2 \frac{nm}{M} C_0 \end{aligned} \tag{S2.16}$$

for some positive constant C_0, where the first equality is due to Property 1, the third step is due to Property 6 and the last inequality is due to Assumption A2 and Assumption C3 as well as $\|\boldsymbol{v}\| = K$.

For $I_{1,2}$ in (S2.15), we have that

$$
\begin{aligned}
|I_{1,2}| &= \delta_n^2 \left| \sum_{i=1}^{n} \boldsymbol{v}^\mathsf{T} \boldsymbol{Z}_i^\mathsf{T} \{ \boldsymbol{U}_i^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t}) \} \boldsymbol{v} \right| \\
&= \delta_n^2 \left| \boldsymbol{v}^\mathsf{T} \left[\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right) \otimes \left\{ \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{v} \right| \\
&\leq \delta_n^2 \|\boldsymbol{v}\| \left\| \left[\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right) \otimes \left\{ \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{v} \right\| \\
&\leq \delta_n^2 \|\boldsymbol{v}\|^2 \sup_{\|\boldsymbol{w}\|=1} \left\| \left[\left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right) \otimes \left\{ \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{w} \right\| \\
&= \delta_n^2 \|\boldsymbol{v}\|^2 \left\| \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right) \otimes \left\{ \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right\|_2 \\
&= \delta_n^2 \|\boldsymbol{v}\|^2 \left\| \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right\|_2 \left\| \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\|_2 , \qquad \text{(S2.17)}
\end{aligned}
$$

where the second equality is due to (S2.13) and Property 1, the third step is due to the Cauchy-Schwarz inequality, the fifth step is due to the definition (S2.6), and the last equality is due to Property 3.

Note that

$$
\left\| \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right\|_2 \leq \left\| \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^\mathsf{T} \right\|_F = \sqrt{\sum_{j,k=1}^{p} \left(\sum_{i=1}^{n} X_{ij} U_{ik} \right)^2} = O_p(\sqrt{n}) \qquad \text{(S2.18)}
$$

due to the central limit theorem, and by Assumption C3, we further know

that

$$\begin{aligned}\|\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\|_2 &= \sqrt{\rho_{\max}\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}} \\ &= \rho_{\max}\left(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right) \\ &\leq C_4 m/M, \end{aligned} \tag{S2.19}$$

where the first step is due to the definition of the matrix 2-norm, and the second step is due to Property 5. Therefore, (S2.17) is bounded:

$$|I_{1,2}| = O_p(\delta_n^2\sqrt{n}m/M)\cdot K^2 \tag{S2.20}$$

since $\|\boldsymbol{v}\| = K$.

Next, combining $I_{1,3}$ in (S2.15) and I_5 in (S2.14) gives that

$$I_{1,3} + I_5 = \frac{1}{2}\delta_n^2 \boldsymbol{v}^{\mathsf{T}}\left[\left\{\sum_{i=1}^{n}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma})\right\}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}. \tag{S2.21}$$

For any $1 \leq j, k \leq p$,

$$\operatorname{Var}\left\{\sum_{i=1}^{n}(U_{ij}U_{ik} - \Sigma_{jk})\right\} = n\operatorname{Var}(U_{ij}U_{ik}) = O(n), \tag{S2.22}$$

due to Assumption D1, where Σ_{jk} is the (j,k) element of $\boldsymbol{\Sigma}$. Therefore, by Property 4 (b) and the definition of the Frobenius norm,

$$\begin{aligned}\left\|\sum_{i=1}^{n}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma})\right\|_2^2 &\leq \left\|\sum_{i=1}^{n}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma})\right\|_F^2 \\ &= \sum_{j,k=1}^{p}\left\{\sum_{i=1}^{n}(U_{ij}U_{ik} - \Sigma_{jk})\right\}^2 \\ &= O_p(n),\end{aligned}$$

where the last step comes from the Markov inequality and (S2.22), so

$$\left\|\sum_{i=1}^{n}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}}-\boldsymbol{\Sigma})\right\|_2 = O_p(\sqrt{n}). \tag{S2.23}$$

Applying (S2.23) and (S2.19) to (S2.21) gives that

$$\begin{aligned}|I_{1,3}+I_5| &\le \frac{1}{2}\delta_n^2\|\boldsymbol{v}\|^2\left\|\sum_{i=1}^{n}(\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}}-\boldsymbol{\Sigma})\right\|_2 \left\|\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\|_2 \\ &= O_p(\delta_n^2\sqrt{n}m/M)\cdot K^2, \end{aligned}\tag{S2.24}$$

where the inequality is due to Property 3.

By (S2.13), (S2.19) and Properties 3 and 4,

$$\begin{aligned}|I_{1,4}| &= \frac{\delta_n^2}{2}\left|\sum_{i=1}^{n}\boldsymbol{v}^{\mathsf{T}}\left\{\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{Z}_i-\mathbb{E}(\boldsymbol{Z}_i^{\mathsf{T}}\boldsymbol{Z}_i)\right\}\boldsymbol{v}\right| \\ &= \frac{\delta_n^2}{2}\left|\boldsymbol{v}^{\mathsf{T}}\left[\left\{\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}-n\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}})\right\}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}\right| \\ &\le \frac{n\delta_n^2}{2}\|\boldsymbol{v}\|^2\left\|\frac{1}{n}\sum_{i=1}^{n}\{\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}-\mathbb{E}(\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}})\}\right\|_F \left\|\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\|_2 \\ &= O_p(\delta_n^2\sqrt{n}m/M)\cdot K^2, \end{aligned}\tag{S2.25}$$

where the last step comes from Assumption A2, the central limit theorem to bound the Frobenius norm term, and (S2.19).

Therefore, combining (S2.20), (S2.24) and (S2.25) gives that

$$|I_{1,2}+I_{1,3}+I_{1,4}+I_5| = O_p(\delta_n^2\sqrt{n}m/M)\cdot K^2. \tag{S2.26}$$

Step 3: Examine I_2 and I_4 in (S2.14).

To describe the difference between (2.3) and (S1.3), let $\Delta_j(t) \triangleq \beta_j(t) - \theta_j(t)$, which equals $\beta_j(t) - \boldsymbol{b}_{0,j}^{\mathsf{T}}\boldsymbol{\phi}(t)$, representing the B-spline approximation

error for $\beta_j(t)$. Then we write $\beta_j(t) = \boldsymbol{b}_{0,j}^{\mathsf{T}}\boldsymbol{\phi}(t) + \Delta_j(t)$. Inserting this into model (2.2) gives that

$$\begin{aligned}
\boldsymbol{Y}_i(\boldsymbol{t}) &= \sum_{j=1}^{p} X_{ij}\left(\beta_j(t_1), \cdots, \beta_j(t_m)\right)^{\mathsf{T}} + \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \\
&= \sum_{j=1}^{p} X_{ij}\{\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{b}_{0,j} + \boldsymbol{\Delta}_j(\boldsymbol{t})\} + \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \\
&= \boldsymbol{\Phi}(\boldsymbol{t})[\boldsymbol{b}_{0,1}, \cdots, \boldsymbol{b}_{0,p}]\boldsymbol{X}_i + \boldsymbol{\varepsilon}_i(\boldsymbol{t}) + \sum_{j=1}^{p} X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t}) \\
&= \{\boldsymbol{X}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t})\}\boldsymbol{b}_0 + \boldsymbol{\varepsilon}_i(\boldsymbol{t}) + \sum_{j=1}^{p} X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t}), \qquad \text{(S2.27)}
\end{aligned}$$

where $\boldsymbol{\Delta}_j(\boldsymbol{t}) = (\Delta_j(t_1), \ldots, \Delta_j(t_m))^{\mathsf{T}}$, and the last step is due to Property 2. Plugging (S2.27) into I_2 in (S2.14) yields that

$$\begin{aligned}
I_2 &= -\delta_n \sum_{i=1}^{n} \boldsymbol{v}^{\mathsf{T}} \boldsymbol{Z}_i^{*\mathsf{T}} (\boldsymbol{Y}_i(\boldsymbol{t}) - \boldsymbol{Z}_i^* \boldsymbol{b}_0) \\
&= -\delta_n \boldsymbol{v}^{\mathsf{T}} \left[\sum_{i=1}^{n} \left(\boldsymbol{Z}_i^{\mathsf{T}} + \boldsymbol{U}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \right) \left\{ \boldsymbol{\varepsilon}_i(\boldsymbol{t}) + \sum_{j=1}^{p} X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t}) - \left(\boldsymbol{U}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}) \right) \boldsymbol{b}_0 \right\} \right] \\
&= -\delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathsf{T}} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) - \delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathsf{T}} \left(\sum_{j=1}^{p} X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t}) \right) \\
&\quad + \delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathsf{T}} \left(\boldsymbol{U}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}) \right) \boldsymbol{b}_0 - \delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \left(\boldsymbol{U}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \right) \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \\
&\quad - \delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \left(\boldsymbol{U}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \right) \left(\sum_{j=1}^{p} X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t}) \right) + \delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \left(\boldsymbol{U}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \right) \left(\boldsymbol{U}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}) \right) \boldsymbol{b}_0 \\
&\triangleq I_{2,1} + I_{2,2} + I_{2,3} + I_{2,4} + I_{2,5} + I_{2,6}. \qquad \text{(S2.28)}
\end{aligned}$$

Note that

$$\begin{aligned}\left\|\sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathsf{T}} \boldsymbol{\varepsilon}_i(\boldsymbol{t})\right\| &= \left\|\sum_{i=1}^{n} \left(\boldsymbol{X}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\right) \boldsymbol{\varepsilon}_i(\boldsymbol{t})\right\| \\ &= \left\|\mathrm{vec}\left(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{X}_i^{\mathsf{T}}\right)\right\| \\ &= \left\|\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{X}_i^{\mathsf{T}}\right\|_F \\ &\leq \|\boldsymbol{\Phi}(\boldsymbol{t})\|_2 \left\|\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{X}_i^{\mathsf{T}}\right\|_F, \end{aligned} \tag{S2.29}$$

where the second equality of is due to Property 2 and the last inequality is due to Property 4 (a).

On one hand,

$$\|\boldsymbol{\Phi}(\boldsymbol{t})\|_2 = \sqrt{\rho_{\max}(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t}))} = O(\sqrt{m/M}), \tag{S2.30}$$

by Assumption C3 and the definition of the matrix 2-norm. On the other hand, by the Chebyshev inequality, we have that

$$\begin{aligned}\mathbb{P}\left(\left\|\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{X}_i^{\mathsf{T}}\right\|_F \geq \sqrt{nm}R\right) &\leq \frac{1}{nmR^2}\mathbb{E}\left(\left\|\sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{X}_i^{\mathsf{T}}\right\|_F^2\right) \\ &= \frac{n}{nmR^2} \sum_{j=1}^{p}\sum_{l=1}^{m} \mathbb{E}(X_{ij}^2)\mathbb{E}[\{\varepsilon_i(t_l)\}^2] \\ &= O(1)/R^2 \to 0 \text{ as } R \to \infty,\end{aligned}$$

where the second step is due to the independence assumption for $\boldsymbol{\varepsilon}_i(\boldsymbol{t})$ and $\boldsymbol{X}_i$, and the third step is due to Assumption A2 and the assumption that

$\sup_{t\in\mathcal{T}} \mathbb{E}[\{\varepsilon_i(t)\}^2] < \infty$. Hence,

$$\left\| \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{X}_i^{\mathsf{T}} \right\|_F = O_p(\sqrt{nm}). \tag{S2.31}$$

Combining this rate and (S2.30) with (S2.29) gives that

$$\left\| \sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathsf{T}} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \right\| = O_p\left(\frac{m\sqrt{n}}{\sqrt{M}} \right),$$

which leads to

$$|I_{2,1}| \leq \delta_n \|\boldsymbol{v}\| \left\| \sum_{i=1}^{n} \boldsymbol{Z}_i^{\mathsf{T}} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \right\| = O_p\left(\delta_n \frac{m\sqrt{n}}{\sqrt{M}} \right) \cdot K \tag{S2.32}$$

Similarly, following the steps of (S2.31) and using the independence assumption for $\boldsymbol{\varepsilon}_i(\boldsymbol{t})$ and $\boldsymbol{U}_i$ and Assumption D1, we obtain that $I_{2,4}$ in (S2.28) can be bounded as

$$\begin{aligned} |I_{2,4}| &\leq \delta_n \|\boldsymbol{v}\| \left\| \sum_{i=1}^{n} (\boldsymbol{U}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})) \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \right\| \\ &\leq \delta_n \|\boldsymbol{v}\| \|\boldsymbol{\Phi}(\boldsymbol{t})\|_2 \left\| \sum_{i=1}^{n} \boldsymbol{\varepsilon}_i(\boldsymbol{t}) \boldsymbol{U}_i^{\mathsf{T}} \right\|_F \\ &= O_p\left(\delta_n \frac{m\sqrt{n}}{\sqrt{M}} \right) K. \end{aligned} \tag{S2.33}$$

To examine $I_{2,2}$ in (S2.28), define the $m \times p$ matrix $\boldsymbol{\Delta}(\boldsymbol{t}) = [\boldsymbol{\Delta}_1(\boldsymbol{t}), \ldots, \boldsymbol{\Delta}_p(\boldsymbol{t})]$. Then by (S2.13),

$$I_{2,2} = -\delta_n \boldsymbol{v}^{\mathsf{T}} \sum_{i=1}^{n} \left\{ (\boldsymbol{X}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})) \boldsymbol{\Delta}(\boldsymbol{t}) \boldsymbol{X}_i \right\}.$$

Note that by Property 2,

$$\begin{aligned}\left\|\sum_{i=1}^{n}\left\{(\boldsymbol{X}_i\otimes\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}))\boldsymbol{\Delta}(\boldsymbol{t})\boldsymbol{X}_i\right\}\right\| &= \left\|\mathrm{vec}\left(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Delta}(\boldsymbol{t})\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\right)\right)\right\| \\ &= \left\|\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Delta}(\boldsymbol{t})\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\right)\right\|_F \\ &\le \|\boldsymbol{\Phi}(\boldsymbol{t})\|_2\times\|\boldsymbol{\Delta}(\boldsymbol{t})\|_F\times\left\|\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\right\|_2,\end{aligned} \tag{S2.34}$$

where the last step is due to Property 4. By (S1.2), each element of $\boldsymbol{\Delta}(\boldsymbol{t})$ is uniformly bounded by $C_1'M^{-q}$, and hence,

$$\|\boldsymbol{\Delta}(\boldsymbol{t})\|_F = O(\sqrt{m}M^{-q}). \tag{S2.35}$$

Furthermore, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\left\|\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\right\|_2 &= \rho_{\max}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\right) \\ &= \sup_{\|\boldsymbol{w}\|=1}\boldsymbol{w}^{\mathsf{T}}\left(\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i^{\mathsf{T}}\right)\boldsymbol{w} \\ &= \sup_{\|\boldsymbol{w}\|=1}\left\{\sum_{i=1}^{n}(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{X}_i)^2\right\} \\ &\le \sum_{i=1}^{n}\|\boldsymbol{X}_i\|^2 \\ &= O_p(n),\end{aligned} \tag{S2.36}$$

where the first step comes from Property 5, the second step is because that $\rho_{\max}(\boldsymbol{A}) = \sup_{\|\boldsymbol{w}\|=1}\boldsymbol{w}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{w}$ for any square matrix $\boldsymbol{A}$, and the last step is due to Assumption A2. Hence, applying (S2.30), (S2.35), (S2.36) to (S2.34)

gives

$$\left\| \sum_{i=1}^{n} \left\{ (\boldsymbol{X}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})) \boldsymbol{\Delta}(\boldsymbol{t}) \boldsymbol{X}_i \right\} \right\| = O_p(mM^{-q-1/2}n),$$

which leads to

$$|I_{2,2}| \leq \delta_n \|\boldsymbol{v}\| \left\| \sum_{i=1}^{n} \left\{ (\boldsymbol{X}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})) \boldsymbol{\Delta}(\boldsymbol{t}) \boldsymbol{X}_i \right\} \right\| = O_p(\delta_n mM^{-q-1/2}n)K \tag{S2.37}$$

since $\|\boldsymbol{v}\| = K$.

The term $I_{2,5}$ in (S2.28) can be treated analogously. By (S2.18), we can show that

$$\begin{aligned} |I_{2,5}| &\leq \delta_n \|\boldsymbol{v}\| \left\| \sum_{i=1}^{n} \left\{ (\boldsymbol{U}_i \otimes \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})) \boldsymbol{\Delta}(\boldsymbol{t}) \boldsymbol{X}_i \right\} \right\| \\ &\leq \delta_n \|\boldsymbol{v}\| \|\boldsymbol{\Phi}(\boldsymbol{t})\|_2 \|\boldsymbol{\Delta}(\boldsymbol{t})\|_F \left\| \sum_{i=1}^{n} \boldsymbol{U}_i \boldsymbol{X}_i^{\mathsf{T}} \right\|_2 \\ &= O_p(\delta_n mM^{-q-1/2}\sqrt{n})K. \end{aligned} \tag{S2.38}$$

Now examine $I_{2,3}$ in (S2.28) using a similar strategy. First, define a $p \times (M + d)$ matrix $\tilde{\boldsymbol{U}}$ whose transpose is $[\boldsymbol{b}_{0,1}, \ldots, \boldsymbol{b}_{0,p}]$. According to Assumption B, the true coefficient function $\beta_j(t)$ is uniformly bounded with respect of t, which implies that all elements of $\tilde{\boldsymbol{U}}$ is also uniformly bounded, i.e.,

$$\sup_{1 \leq j \leq p} \|\boldsymbol{b}_{0,j}\|_\infty = O(1). \tag{S2.39}$$

Therefore,

$$\|\tilde{\boldsymbol{U}}\|_F = O(\sqrt{M}). \tag{S2.40}$$

Then by (S2.13),

$$
\begin{aligned}
|I_{2,3}| &= \delta_n \boldsymbol{v}^{\mathsf{T}} \left[(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^{\mathsf{T}}) \otimes \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{b}_0 \\
&\leq \delta_n \|\boldsymbol{v}\| \left\| \left[(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^{\mathsf{T}}) \otimes \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{b}_0 \right\| \\
&\leq \delta_n \|\boldsymbol{v}\| \left\| \mathrm{vec} \left(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \tilde{\boldsymbol{U}} (\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^{\mathsf{T}}) \right) \right\| \\
&= \delta_n \|\boldsymbol{v}\| \left\| \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \tilde{\boldsymbol{U}} (\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^{\mathsf{T}}) \right\|_F \\
&\leq \delta_n \|\boldsymbol{v}\| \left\| \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\|_2 \|\tilde{\boldsymbol{U}}\|_F \left\| \sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{U}_i^{\mathsf{T}} \right\|_2 \\
&= O_p(\delta_n m \sqrt{n} / \sqrt{M}) K, \qquad \text{(S2.41)}
\end{aligned}
$$

where the third step is due to Property 2, the fifth step is due to Property 4 and the last step is due to (S2.18), (S2.19), and (S2.40).

Now combining $I_{2,6}$ in (S2.28) and I_4 in (S2.14), we obtain that

$$
I_{2,6} + I_4 = \delta_n \boldsymbol{v}^{\mathsf{T}} \left[\left\{ \sum_{i=1}^{n} (\boldsymbol{U}_i \boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}) \right\} \otimes \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{b}_0,
$$

which leads to

$$
\begin{aligned}
|I_{2,6}+I_4| &\le \delta_n\|\boldsymbol{v}\| \left\| \left[\left\{ \sum_{i=1}^{n} (\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}) \right\} \otimes \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t}) \right\} \right] \boldsymbol{b}_0 \right\| \\
&\le \delta_n\|\boldsymbol{v}\| \left\| \mathrm{vec} \left[\left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t}) \right\} \tilde{\boldsymbol{U}}^{\mathsf{T}} \left\{ \sum_{i=1}^{n} (\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}) \right\} \right] \right\| \\
&= \delta_n\|\boldsymbol{v}\| \left\| \left\{ \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t}) \right\} \tilde{\boldsymbol{U}}^{\mathsf{T}} \left\{ \sum_{i=1}^{n} (\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}) \right\} \right\|_F \\
&\le \delta_n\|\boldsymbol{v}\| \left\| \sum_{i=1}^{n} (\boldsymbol{U}_i\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}) \right\|_2 \|\boldsymbol{\Phi}(\boldsymbol{t})^{\mathsf{T}}\boldsymbol{\Phi}(\boldsymbol{t})\|_2 \|\tilde{\boldsymbol{U}}\|_F \\
&= O_p(\delta_n m\sqrt{n}/\sqrt{M})K, \qquad \text{(S2.42)}
\end{aligned}
$$

where the second and the fourth steps are due to Property 2 and Property 4 respectively, and the last step is due to (S2.23), (S2.19), and (S2.40). Therefore, combining (S2.32), (S2.33), (S2.37), (S2.38), (S2.41) and (S2.42) with (S2.28), we obtain that

$$
\begin{aligned}
|I_2+I_4| &= O_p(\delta_n m M^{-q-1/2} n)K + O_p(\delta_n m\sqrt{n}/\sqrt{M})K \\
&= O_p(\delta_n m\sqrt{n}/\sqrt{M})K, \qquad \text{(S2.43)}
\end{aligned}
$$

where the last step is due to Assumption C1 (i).

Step 4: Examine I_3 in (S2.14).

By the chain rule,

$$
\frac{\mathrm{d}P_\lambda(\|\boldsymbol{w}\|)}{\mathrm{d}\boldsymbol{w}} = P'_\lambda(\|\boldsymbol{w}\|)\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|},
$$

for any vector $\boldsymbol{w} \neq \boldsymbol{0}$. For $j \notin J_0$, the mean theorem gives that

$$
P_\lambda(\|\boldsymbol{b}_{0,j} + \delta_n\boldsymbol{v}_j\|) - P_\lambda(\|\boldsymbol{b}_{0,j}\|) = \delta_n\boldsymbol{v}_j^{\mathsf{T}} P'_\lambda(\|\tilde{\boldsymbol{b}}_j\|)\frac{\tilde{\boldsymbol{b}}_j}{\|\tilde{\boldsymbol{b}}_j\|},
$$

where $\tilde{\boldsymbol{b}}_j$ is a vector "between" $\boldsymbol{b}_{0,j}$ and $\boldsymbol{b}_{0,j} + \delta_n \boldsymbol{v}_j$. As $\|\boldsymbol{b}_{0,j}\|$ is bounded away from 0 for $j \notin J_0$, $\|\tilde{\boldsymbol{b}}_j\|$ is also bounded away from 0 for sufficiently large n. By Assumption E2 (ii), $P'_\lambda(t) = 0$ for any positive constant t as $\lambda \to 0$. Hence, when $n \to \infty$ and $\lambda \to 0$, we have that

$$P_\lambda(\|\boldsymbol{b}_{0,j} + \delta_n \boldsymbol{v}_j\|) - P_\lambda(\|\boldsymbol{b}_{0,j}\|) = 0,$$

which leads to

$$I_3 = 0. \tag{S2.44}$$

In conclusion, let $R \triangleq I_{1,2} + I_{1,3} + I_{1,4} + I_2 + I_3 + I_4 + I_5$. Then by (S2.26), (S2.43) and (S2.44), we obtain that

$$\begin{aligned} |R| &= O_p(\delta_n^2 \sqrt{n} m / M) K^2 + O_p(\delta_n m \sqrt{n} / \sqrt{M}) K \\ &= O_p(\delta_n m \sqrt{n} / \sqrt{M}) K, \end{aligned} \tag{S2.45}$$

where the second step is due to the definition of δ_n as $n \to \infty$.

According to (S2.14) and (S2.15), to prove (S2.12) it suffices to show that for any $\epsilon > 0$ there exists $K_\epsilon > 0$ such that

$$\mathbb{P}\left(\inf_{\|\boldsymbol{v}\| = K_\epsilon} I_{1,1} + R > 0 \right) > 1 - \epsilon. \tag{S2.46}$$

To this end, we prove a stronger version than (S2.46):

$$\mathbb{P}\left(\inf_{\|\boldsymbol{v}\| = K_\epsilon} I_{1,1} > \sup_{\|\boldsymbol{v}\| = K_\epsilon} |R| \right) > 1 - \epsilon. \tag{S2.47}$$

By the definition of δ_n, (S2.16) and (S2.45), we have that

$$I_{1,1} \geq m C_0 K^2 \text{ and } |R| = O_p(m) \cdot K, \tag{S2.48}$$

for any $\|\boldsymbol{v}\| = K$. Then for any given $\epsilon > 0$, by (S2.48) we take K_ϵ such that $K_\epsilon C_0 > O(1)$. Then such K_ϵ makes (S2.47) hold. As a result, the assertion (S2.12) is then proved.

Part 2: Prove (5.24).

Given (5.23), it is straightforward that

$$\begin{aligned}
\|\hat{\beta}_j - \beta_j\|_{L_\infty} &\leq \|\hat{\beta}_j - \theta_j\|_{L_\infty} + \|\theta_j - \beta_j\|_{L_\infty} \\
&= \|(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_{0,j})^\mathsf{T}\boldsymbol{\phi}\|_{L_\infty} + \|\theta_j - \beta_j\|_{L_\infty} \\
&\leq \|\hat{\boldsymbol{b}}_j - \boldsymbol{b}_{0,j}\|_\infty + \|\theta_j - \beta_j\|_{L_\infty} \\
&\leq \|\hat{\boldsymbol{b}}_j - \boldsymbol{b}_{0,j}\| + \|\theta_j - \beta_j\|_{L_\infty} \\
&= O_p(\delta_n),
\end{aligned}$$

where the first step is due to the triangle inequality, the second step comes from (3.13) and (S1.3), the third step is due to the fact that $\sum_{k=1}^{M+d} |\phi_k(t)| = 1$ for any $t \in \mathcal{T}$ (Farin et al., 2002, (6.6.7)), the fourth step is because of the relationship between the infinity norm and Euclidean norm for a vector, and the last step is due to Assumption C1(i), (S1.2) and (5.23).

According to Assumption A3, $\mathcal{T}$ is bounded, and thus, its measure, denoted $|\mathcal{T}|$, is finite. Therefore, we have that for any $j = 1, \ldots, p$,

$$\|\theta_j - \beta_j\|_{L_2} \leq \sqrt{|\mathcal{T}|} \cdot \|\theta_j - \beta_j\|_{L_\infty},$$

and thus, there exists a positive constant E_1 such that

$$\begin{aligned}\|\hat{\beta}_j - \beta_j\|_{L_2} &\leq \|\hat{\beta}_j - \theta_j\|_{L_2} + \|\theta_j - \beta_j\|_{L_2} \\ &= \|(\hat{\boldsymbol{u}}_j - \boldsymbol{u}_{0,j})^{\mathsf{T}}\boldsymbol{\phi}\|_{L_2} + \sqrt{|\mathcal{T}|}\|\theta_j - \beta_j\|_{L_\infty} \\ &\leq E_1^{-1}M^{-1/2}\|\hat{\boldsymbol{u}}_j - \boldsymbol{u}_{0,j}\| + \sqrt{|\mathcal{T}|}\|\theta_j - \beta_j\|_{L_\infty} \\ &= O_p(\delta_n/\sqrt{M}),\end{aligned}$$

where the third step follows from Lemma 3, and the last step is due to Assumption C1(i), (S1.2) and (5.23).

S2.3 Proof of Theorem 2

Proof of Theorem 2(i):

We show the result using proof by contradiction. Suppose that there exists some $k \in J_0$ such that $\hat{\boldsymbol{b}}_k \neq \mathbf{0}$. Then by (3.12), taking derivative of $Q_n(\boldsymbol{b})$ in (3.11) with respect to $\boldsymbol{b}_k$ and evaluating it at $\hat{\boldsymbol{b}}$ yields that

$$\begin{aligned}\mathbf{0} &= \frac{\partial Q_n(\hat{\boldsymbol{b}})}{\partial \boldsymbol{b}_k} \\ &= -\sum_{i=1}^{n} X_{ik}^{*}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})(\boldsymbol{Y}_i(\boldsymbol{t}) - \boldsymbol{Z}_i^{*}\hat{\boldsymbol{b}}) \\ &\quad + nmP_\lambda'(\|\hat{\boldsymbol{b}}_k\|)\frac{\hat{\boldsymbol{b}}_k}{\|\hat{\boldsymbol{b}}_k\|} - n\left[\boldsymbol{\Sigma}_k \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right]\hat{\boldsymbol{b}} \\ &= -\sum_{i=1}^{n}(X_{ik} + U_{ik})\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\left\{\boldsymbol{Y}_i(\boldsymbol{t}) - \boldsymbol{Z}_i\hat{\boldsymbol{b}} - (\boldsymbol{U}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}}\right\} \\ &\quad + nmP_\lambda'(\|\hat{\boldsymbol{b}}_k\|)\frac{\hat{\boldsymbol{b}}_k}{\|\hat{\boldsymbol{b}}_k\|} - n\left[\boldsymbol{\Sigma}_k \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right]\hat{\boldsymbol{b}}, \qquad \text{(S2.49)}\end{aligned}$$

where $\boldsymbol{\Sigma}_k$ represents the kth row of $\boldsymbol{\Sigma}$, and we use (3.7) and (S2.13).

By (3.13) and (S2.13), $\boldsymbol{Z}_i\hat{\boldsymbol{b}} = \hat{\boldsymbol{\beta}}(\boldsymbol{t})^\mathsf{T}\boldsymbol{X}_i$, where the $p \times m$ matrix $\hat{\boldsymbol{\beta}}(\boldsymbol{t}) = [\hat{\boldsymbol{\beta}}(t_1), \cdots, \hat{\boldsymbol{\beta}}(t_m)]$. Then further applying model (2.2) to the last expression in (S2.49) gives that

$$\begin{aligned}
\mathbf{0} = & -\sum_{i=1}^{n} X_{ik}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\varepsilon}_i(\boldsymbol{t}) - \sum_{i=1}^{n} X_{ik}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})(\boldsymbol{\beta}(\boldsymbol{t}) - \hat{\boldsymbol{\beta}}(\boldsymbol{t}))^\mathsf{T}\boldsymbol{X}_i \\
& + \sum_{i=1}^{n} X_{ik}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})(\boldsymbol{U}_i^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}} - \sum_{i=1}^{n} U_{ik}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\varepsilon}_i(\boldsymbol{t}) \\
& - \sum_{i=1}^{n} U_{ik}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})(\boldsymbol{\beta}(\boldsymbol{t}) - \hat{\boldsymbol{\beta}}(\boldsymbol{t}))^\mathsf{T}\boldsymbol{X}_i + \sum_{i=1}^{n} U_{ik}\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})(\boldsymbol{U}_i^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}} \\
& + nmP'_\lambda(\|\hat{\boldsymbol{b}}_k\|)\frac{\hat{\boldsymbol{b}}_k}{\|\hat{\boldsymbol{b}}_k\|} - n\left[\boldsymbol{\Sigma}_k \otimes \{\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right]\hat{\boldsymbol{b}} \\
\triangleq & \, T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_7 + T_8.
\end{aligned} \tag{S2.50}$$

We now show that (S2.50) does not hold because the sum $\sum_{j=1}^{8} T_j$ cannot equal zero. To this end, we examine the terms in (S2.50) individually with the following steps.

Step 1: Examination of T_7 in (S2.50):

According to Theorem 1, we have that $\|\hat{\boldsymbol{b}}_k\| = O_p(\delta_n)$, where $\delta_n = \sqrt{M/n}$ defined in Appendix A.1. By Assumption E2 (i) and the assumption that $\lambda/\delta_n \to \infty$, it is easily shown that $|P'_\lambda(\|\hat{\boldsymbol{b}}_k\|) - \lambda| \to 0$ as $n \to \infty$. Hence, for sufficiently large n, we claim that

$$\|T_7\| = nm\left|P'_\lambda(\|\hat{\boldsymbol{b}}_k\|)\right| > \frac{1}{2}nm\lambda. \tag{S2.51}$$

Step 2: Examination of $T_6 + T_8$ in (S2.50):

Simple calculations result in

$$\begin{aligned} T_6 + T_8 &= \sum_{i=1}^{n} (U_{ik}\boldsymbol{U}_i^{\mathsf{T}}) \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\hat{\boldsymbol{b}} - n\left[\boldsymbol{\Sigma}_k \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right]\hat{\boldsymbol{b}} \\ &= \left[\left\{\sum_{i=1}^{n}\left(U_{ik}\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}_k\right)\right\} \otimes \{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right]\hat{\boldsymbol{b}}, \end{aligned}$$

where the first equality is due to Property 1.

For $1 \leq j \leq p$, we have that

$$\|\hat{\boldsymbol{b}}_j\|_\infty \leq \|\hat{\boldsymbol{b}}_j - \boldsymbol{b}_{0,j}\|_\infty + \|\boldsymbol{b}_{0,j}\|_\infty = O_p(1),$$

where we use Theorem 1, Condition C1, and (S2.39). Therefore,

$$\|\hat{\boldsymbol{b}}_j\| \leq \sqrt{M+d}\|\hat{\boldsymbol{b}}_j\|_\infty = O_p(\sqrt{M}),$$

which implies that

$$\|\hat{\boldsymbol{b}}\| = O_p(\sqrt{M}) \tag{S2.52}$$

by Condition C1 that the dimension p is finite. Hence, we arrive at

$$\|T_6 + T_8\| \leq \left\|\sum_{i=1}^{n}\left(U_{ik}\boldsymbol{U}_i^{\mathsf{T}} - \boldsymbol{\Sigma}_k\right)\right\|_2 \left\|\boldsymbol{\Phi}(\boldsymbol{t})^{\mathsf{T}}\boldsymbol{\Phi}(\boldsymbol{t})\right\|_2 \|\hat{\boldsymbol{b}}\| = O_p(\sqrt{n}m/\sqrt{M}), \tag{S2.53}$$

where the first step is due to Property 3 and the second step is due to (S2.23), (S2.19) and (S2.52).

Step 3: Examination of T_1 and T_4 in (S2.50):

By Property 2, we can show that

$$T_1 = -\sum_{i=1}^{n}(X_{ik} \otimes \mathbf{\Phi}^{\mathsf{T}}(\boldsymbol{t}))\boldsymbol{\varepsilon}_i(\boldsymbol{t}) = \operatorname{vec}\left\{\mathbf{\Phi}^{\mathsf{T}}(\boldsymbol{t})\left(\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i(\boldsymbol{t})X_{ik}\right)\right\},$$

which leads to

$$\|T_1\| = \left\|\mathbf{\Phi}^{\mathsf{T}}(\boldsymbol{t})\left(\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i(\boldsymbol{t})X_{ik}\right)\right\|_F \leq \|\mathbf{\Phi}(\boldsymbol{t})\|_2 \left\|\sum_{i=1}^{n}\boldsymbol{\varepsilon}_i(\boldsymbol{t})X_{ik}\right\| = O_p(\sqrt{n}m/\sqrt{M}), \tag{S2.54}$$

where the inequality is due to Property 4 and the last step is due to (S2.30) and (S2.31).

Analogously, we can show that

$$\|T_4\| = \left\|-\sum_{i=1}^{n}U_{ik}\mathbf{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\varepsilon}_i(\boldsymbol{t})\right\| = O_p(\sqrt{n}m/\sqrt{M}), \tag{S2.55}$$

where the procedure is the same as (S2.54) except we replace X_{ik} in (S2.54) with U_{ik}.

Step 4: Examination of T_3 in (S2.50):

$$\begin{aligned}
\|T_3\| &= \left\|\sum_{i=1}^{n}X_{ik}\mathbf{\Phi}^{\mathsf{T}}(\boldsymbol{t})(\boldsymbol{U}_i^{\mathsf{T}} \otimes \mathbf{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}}\right\| \\
&= \left\|\left[\left(\sum_{i=1}^{n}X_{ik}\boldsymbol{U}_i^{\mathsf{T}}\right) \otimes \{\mathbf{\Phi}^{\mathsf{T}}(\boldsymbol{t})\mathbf{\Phi}(\boldsymbol{t})\}\right]\hat{\boldsymbol{b}}\right\| \\
&\leq \left\|\sum_{i=1}^{n}X_{ik}\boldsymbol{U}_i\right\|_2 \left\|\mathbf{\Phi}(\boldsymbol{t})^{\mathsf{T}}\mathbf{\Phi}(\boldsymbol{t})\right\|_2 \|\hat{\boldsymbol{b}}\| \\
&= O_p(\sqrt{n}m/\sqrt{M}),
\end{aligned} \tag{S2.56}$$

where the second and third steps are due to Property 1 and Property 3, respectively, and the last step is due to (S2.18), (S2.19) and (S2.52).

Step 5: Examination of T_2 and T_5 in (S2.50):

Write $\boldsymbol{\Theta} = \boldsymbol{\beta}(\boldsymbol{t}) - \hat{\boldsymbol{\beta}}(\boldsymbol{t})$. Then by Theorem 1, $\|\boldsymbol{\Theta}\|_F = O_p(\delta_n\sqrt{m})$. Then we have that

$$\begin{aligned}\|T_2\| &= \left\| \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Theta}^{\mathsf{T}} \left(\sum_{i=1}^{n} X_{ik}\boldsymbol{X}_i \right) \right\| \\ &\leq \left\| \sum_{i=1}^{n} X_{ik}\boldsymbol{X}_i \right\|_2 \|\boldsymbol{\Phi}(\boldsymbol{t})\|_2 \|\boldsymbol{\Theta}\|_F \\ &= O_p(nm\delta_n/\sqrt{M}), \end{aligned} \tag{S2.57}$$

where the inequality is due to Property 4 and the last step is due to (S2.30) and (S2.36).

Similarly it can be proved that

$$\|T_5\| = \left\| \boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Theta}^{\mathsf{T}} \left(\sum_{i=1}^{n} U_{ik}\boldsymbol{X}_i \right) \right\| = O_p(\sqrt{n}m\delta_n/\sqrt{M}) \tag{S2.58}$$

where (S2.18) is used to replace (S2.36).

In conclusion, combining results (S2.53)-(S2.58) leads to

$$\|T_1 + T_2 + T_3 + T_4 + T_5 + T_6 + T_8\| = O_p(nm\delta_n/\sqrt{M}),$$

which, by (S2.51), is asymptotically ignorable relative to T_7 if $\sqrt{n}\lambda \to \infty$. Therefore, with probability tending to 1, the sum $\sum_{j=1}^{8} T_j$ in (S2.50) is dominated by T_7, which can never be zero according to (S2.51). That

is, with probability tending to 1, (S2.50) is untrue, and thus, by proof of contradiction, there is no $k \in J_0$ such that $\hat{\boldsymbol{b}}_k \neq \mathbf{0}$.

Proof of Theorem 2(ii):

Corresponding to the active subvector $\boldsymbol{X}_{Ii}$ defined before Theorem 1, let $\boldsymbol{X}_{Ii}^*$ denote its contaminated version, and let $\boldsymbol{U}_{Ii}$ denote the corresponding subvector of $\boldsymbol{U}_i$ in (3.7). For the cardinality $s = |J_0|$, we write $\boldsymbol{v} = (\boldsymbol{b}_1^\mathsf{T}, \ldots, \boldsymbol{b}_s^\mathsf{T})^\mathsf{T}$, $\boldsymbol{v}_0 = (\boldsymbol{b}_{0,1}^\mathsf{T}, \ldots, \boldsymbol{b}_{0,s}^\mathsf{T})^\mathsf{T}$ and $\hat{\boldsymbol{v}} = (\hat{\boldsymbol{b}}_1^\mathsf{T}, \ldots, \hat{\boldsymbol{b}}_s^\mathsf{T})^\mathsf{T}$. Concentrating on the subvector $\boldsymbol{v}$, let $\check{Q}_n(\boldsymbol{v}) = Q_n(\boldsymbol{v}, \mathbf{0})$, where $Q_n(\boldsymbol{b})$ is defined in (3.11), and $\mathbf{0}$ here is understood to have the dimension so that $(\boldsymbol{v}^\mathsf{T}, \mathbf{0}^\mathsf{T})^\mathsf{T}$ is of dimension $p(M+d)$. Throughout the manuscript, we loosely use $\mathbf{0}$ to represent a zero vector or a zero matrix without differentiating them or saying its dimension.

Define

$$S(\boldsymbol{v}) = \partial \check{Q}_n(\boldsymbol{v}) / \partial \boldsymbol{v},$$

which, by (3.11), equals

$$\begin{aligned} S(\boldsymbol{v}) = &- \sum_{i=1}^{n} \left(\boldsymbol{X}_{Ii}^* \otimes \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\right) \left\{\boldsymbol{Y}_i(\boldsymbol{t}) - \left(\boldsymbol{X}_{Ii}^{*\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t})\right) \boldsymbol{v}\right\} \\ &- n \left[\boldsymbol{\Sigma}_I \otimes \{\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t}) \boldsymbol{\Phi}(\boldsymbol{t})\}\right] \boldsymbol{v} + nm \boldsymbol{P}(\boldsymbol{v}) \boldsymbol{v}, \end{aligned} \tag{S2.59}$$

where $\boldsymbol{P}(\boldsymbol{v}) \triangleq \operatorname{diag}\{P_\lambda'(\|\boldsymbol{b}_1\|)/\|\boldsymbol{b}_1\|, \ldots, P_\lambda'(\|\boldsymbol{b}_s\|)/\|\boldsymbol{b}_s\|\} \otimes \boldsymbol{I}_{M+d}$ and $\boldsymbol{I}_{M+d}$ is the $(M+d)$-dimensional identity matrix. Note that $S(\boldsymbol{v})$ is well-defined

if $\boldsymbol{b}_j \neq \mathbf{0}$ for all $j = 1, \ldots, s$.

Because $\boldsymbol{v}_0$ is bounded away from $\mathbf{0}$ and $\hat{\boldsymbol{v}}$ converges in probability to $\boldsymbol{v}_0$ as $n \to \infty$, $\hat{\boldsymbol{v}}$ is also bounded in probability away from $\mathbf{0}$ as $n \to \infty$. Furthermore, by Theorem 1 and Theorem 2 (i), with probability tending to 1, $(\hat{\boldsymbol{v}}^\mathsf{T}, \mathbf{0}^\mathsf{T})^\mathsf{T}$ is the minimizer of $Q_n(\boldsymbol{b})$, and thus, we have that

$$S(\hat{\boldsymbol{v}}) = \mathbf{0}. \tag{S2.60}$$

Combining (S2.59) and (S2.60) leads to

$$\begin{aligned}\mathbf{0} = &- \sum_{i=1}^{n} \left(\boldsymbol{X}_{Ii}^* \otimes \boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\right) \left\{\boldsymbol{Y}_i(\boldsymbol{t}) - \left(\boldsymbol{X}_{Ii}^{*\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t})\right) \hat{\boldsymbol{v}}\right\} \\ &- n \left[\boldsymbol{\Sigma}_I \otimes \{\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}\right] \hat{\boldsymbol{v}} + nm\boldsymbol{P}(\hat{\boldsymbol{v}})\hat{\boldsymbol{v}}. \end{aligned} \tag{S2.61}$$

Since $\boldsymbol{b}_{0,j'} = \mathbf{0}$ for all $j' > s$, (S2.27) yields that

$$\boldsymbol{Y}_i(\boldsymbol{t}) = (\boldsymbol{X}_{Ii}^\mathsf{T} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{v}_0 + \boldsymbol{\varepsilon}_i(\boldsymbol{t}) + \sum_{j=1}^{p} X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t}). \tag{S2.62}$$

Inserting (S2.62) into (S2.61) yields that

$$\begin{aligned}\boldsymbol{A}(\hat{\boldsymbol{v}}-\boldsymbol{v}_0) = & \frac{M}{nm}\sum_{i=1}^{n}\left(\boldsymbol{X}_{Ii}^{*}\otimes\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\right)\boldsymbol{\varepsilon}_i(\boldsymbol{t}) \\ & -\frac{M}{nm}\sum_{i=1}^{n}\left[\left(\boldsymbol{X}_{Ii}^{*}\boldsymbol{U}_{Ii}^{\mathsf{T}}-\boldsymbol{\Sigma}_I\right)\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}_0 \\ & -\frac{M}{nm}\sum_{i=1}^{n}\left[\left(\boldsymbol{X}_{Ii}^{*}\boldsymbol{X}_{Ii}^{*\,\mathsf{T}}-\boldsymbol{\Sigma}_I-\boldsymbol{\Sigma}_{\boldsymbol{X}_I}\right)\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right](\hat{\boldsymbol{v}}-\boldsymbol{v}_0) \\ & +\frac{M}{nm}\sum_{i=1}^{n}\left[\left(\boldsymbol{X}_{Ii}^{*}\otimes\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\right)\left\{\sum_{j=1}^{p}X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t})\right\}\right]-M\boldsymbol{P}(\hat{\boldsymbol{v}})\hat{\boldsymbol{v}} \\ \triangleq & \, I_1+I_2+I_3+I_4+I_5,\end{aligned} \tag{S2.63}$$

where $\boldsymbol{A}\triangleq\boldsymbol{\Sigma}_{\boldsymbol{X}_I}\otimes\left\{\frac{M}{m}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}$ is a matrix whose eigenvalues are bounded and bounded away from 0 due to Assumption C3. Multiplying the inverse of $\boldsymbol{A}$ on both sides of (S2.63), we can then investigate the limiting distribution of $\boldsymbol{\alpha}^{\mathsf{T}}(\hat{\boldsymbol{v}}-\boldsymbol{v}_0)$ for any vector $\boldsymbol{\alpha}$ satisfying $0<\|\boldsymbol{\alpha}\|<\infty$. To this end, we examine (S2.63) term by term.

Step 1: Examination of I_1 and I_2 in (S2.63):

As defined in Section 5, $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}$ is the covariance matrix of $\boldsymbol{X}_{Ii}$ and $\boldsymbol{\Sigma}_I$ represents the covariance matrix of $\boldsymbol{U}_{Ii}$. Then the covariance matrix of I_1

in (S2.63) is computed by

$$\begin{aligned}\mathrm{Cov}(I_1) &= \frac{M^2}{nm^2}\,\mathrm{Cov}\left\{\left(\boldsymbol{X}_{Ii}^{*}\otimes\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\right)\boldsymbol{\varepsilon}_i(\boldsymbol{t})\right\}\\ &= \frac{M^2}{nm^2}\,\mathrm{Cov}(\boldsymbol{X}_{Ii}^{*})\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\right\}\\ &= \frac{M^2}{nm^2}\left(\boldsymbol{\Sigma}_{\boldsymbol{X}_I}+\boldsymbol{\Sigma}_I\right)\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\right\},\end{aligned}$$

where the second step is due to Lemma 1.

Let $\boldsymbol{I}_s$ be the $s\times s$ identity matrix. Recall that $\boldsymbol{B}_{0s}=[\boldsymbol{b}_{0,1},\cdots,\boldsymbol{b}_{0,s}]$ is the $(M+d)\times s$ matrix formed by the blocks of $\boldsymbol{b}_0$, defined in Section 5. Then the covariance matrix of I_2 in (S2.63) is calculated as follows:

$$\begin{aligned}\mathrm{Cov}(I_2) &= \frac{M^2}{nm^2}\,\mathrm{Cov}\left(\left[\left(\boldsymbol{X}_{Ii}^{*}\boldsymbol{U}_{Ii}^{\mathsf{T}}-\boldsymbol{\Sigma}_I\right)\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{v}_0\right)\\ &= \frac{M^2}{nm^2}\,\mathrm{Cov}\left[\mathrm{vec}\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\left(\boldsymbol{U}_{Ii}\boldsymbol{X}_{Ii}^{*\,\mathsf{T}}\right)\right\}\right]\\ &= \frac{M^2}{nm^2}\,\mathrm{Cov}\left(\left[\boldsymbol{X}_{Ii}^{*}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\right\}\right]\boldsymbol{U}_{Ii}\right)\\ &= \frac{M^2}{nm^2}\,\mathrm{Cov}\left(\left[\boldsymbol{X}_{Ii}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\right\}\right]\boldsymbol{U}_{Ii}\right)\\ &\quad+\frac{M^2}{nm^2}\,\mathrm{Cov}\left(\left[\boldsymbol{U}_{Ii}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\right\}\right]\boldsymbol{U}_{Ii}\right)\\ &= \frac{M^2}{nm^2}\boldsymbol{\Sigma}_{\boldsymbol{X}_I}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\\ &\quad+\frac{M^2}{nm^2}\,\mathrm{Cov}\left\{\mathrm{vec}\left(\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\boldsymbol{U}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}}\right)\right\}\\ &= \frac{M^2}{nm^2}\boldsymbol{\Sigma}_{\boldsymbol{X}_I}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\\ &\quad+\frac{M^2}{nm^2}\,\mathrm{Cov}\left(\left[\boldsymbol{I}_s\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\right\}\right]\mathrm{vec}\left(\boldsymbol{U}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}}\right)\right)\\ &= \frac{M^2}{nm^2}\boldsymbol{\Sigma}_{\boldsymbol{X}_I}\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\\ &\quad+\frac{M^2}{nm^2}\left[\boldsymbol{I}_s\otimes\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\right\}\right]\mathrm{Cov}\left\{\mathrm{vec}\left(\boldsymbol{U}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}}\right)\right\}\left[\boldsymbol{I}_s\otimes\left\{\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right],\end{aligned}$$

where the second, third and sixth steps are due to Property 2, the fourth step is because that $\boldsymbol{X}_{Ii}$ has mean $\mathbf{0}$ and is independent from $\boldsymbol{U}_{Ii}$, and the fifth step is due to Lemma 1 and Property 2.

Define the $s^2 \times s^2$ matrix $\boldsymbol{\Gamma} = \mathrm{Cov}\left\{\mathrm{vec}\left(\boldsymbol{U}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}}\right)\right\}$. Since $\boldsymbol{\varepsilon}_i(\boldsymbol{t})$ is independent of $\boldsymbol{X}_{Ii}^*$ and $\boldsymbol{U}_{Ii}$, I_1 and I_2 are uncorrelated. Thus,

$$\begin{aligned}\mathrm{Cov}(I_1 + I_2) =& \frac{M^2}{nm^2}\left(\boldsymbol{\Sigma}_{\boldsymbol{X}_I} + \boldsymbol{\Sigma}_I\right) \otimes \left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\right\} \\ &+ \frac{M^2}{nm^2}\boldsymbol{\Sigma}_{\boldsymbol{X}_I} \otimes \left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\} \\ &+ \frac{M^2}{nm^2}\left[\boldsymbol{I}_s \otimes \left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\boldsymbol{B}_{0s}\right\}\right]\boldsymbol{\Gamma}\left[\boldsymbol{I}_s \otimes \left\{\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]. \end{aligned} \tag{S2.64}$$

By the assumption in Theorem 2 (ii) that $\rho_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \geq C_5\xi_m$ with positive constants C_5 and $\{\xi_m : m = 1, 2, \dots\}$, $\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}$ is always positive definite. Furthermore, $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}$ is also positive definite due to Assumption A2. Hence, $\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\,\mathrm{Cov}(I_1 + I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha} \neq 0$ for any $\boldsymbol{\alpha} \neq \mathbf{0}$. Therefore, by the central limit theorem, we have that as $n \to \infty$,

$$\frac{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}(I_1 + I_2)}{\sqrt{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\,\mathrm{Cov}(I_1 + I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}}} \xrightarrow{d} N(0, 1). \tag{S2.65}$$

We next show that all the scaled remaining terms in (S2.63), $\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}I_k$ for $k = 3, 4, 5$, are asymptotically negligible relative to $\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}(I_1 + I_2)$.

Step 2: Examination of I_3 in (S2.63):

Following similar argument in (S2.23) ,

$$\begin{aligned}
&\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{Ii}^{*}\boldsymbol{X}_{Ii}^{*\mathsf{T}} - \boldsymbol{\Sigma}_I - \boldsymbol{\Sigma}_{\boldsymbol{X}_I} \right\|_2 \\
\leq &\left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{Ii}\boldsymbol{X}_{Ii}^{\mathsf{T}} - \boldsymbol{\Sigma}_{\boldsymbol{X}_I} \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{U}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}} - \boldsymbol{\Sigma}_I \right\|_2 + \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{Ii}\boldsymbol{U}_{Ii}^{\mathsf{T}} \right\|_2 \\
= &O_p(n^{-1/2})
\end{aligned}$$

since $\mathbb{E}(X_{ij}^4) < \infty$ and $\mathbb{E}(U_{ij}^4) < \infty$ for $j = 1, \ldots, s$ by Assumptions A2 and D1, respectively. Hence, for I_3 in (S2.63), by (S2.19) and Theorem 1 we have that

$$\begin{aligned}
\left|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}I_3\right| &\leq \|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\| \left\| \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{X}_{Ii}^{*}\boldsymbol{X}_{Ii}^{*\mathsf{T}} - \boldsymbol{\Sigma}_I - \boldsymbol{\Sigma}_{\boldsymbol{X}_I} \right\|_2 \\
&\quad \cdot \left\| \frac{M}{m}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t}) \right\|_2 \|\hat{\boldsymbol{v}} - \boldsymbol{v}_0\| \\
&= \|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\| \cdot O_p(n^{-1/2}) \cdot O_p(\sqrt{M/n}). \qquad \text{(S2.66)}
\end{aligned}$$

Meanwhile, we have assumed in Theorem 2 (ii) that $\rho_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \geq C_5\xi_m$, and thus, by Assumption C3, we have that

$$\rho_{\min}\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\right\} \geq \rho_{\min}\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\right\}\rho_{\min}(\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}) \geq C_5'\frac{m\xi_m}{M},$$

where C_5' is a positive constant and the first inequality is due to Lemma 2.

Therefore, for some constant C_5'',

$$\begin{aligned}\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\operatorname{Cov}(I_1+I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha} &\geq \|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\|^2\rho_{\min}\{\operatorname{Cov}(I_1)\}\\ &= \frac{M^2}{nm^2}\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\|^2\rho_{\min}(\boldsymbol{\Sigma}_I+\boldsymbol{\Sigma}_{\boldsymbol{X}_I})\rho_{\min}\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\right\}\\ &\geq C_5''\frac{M}{nm}\xi_m\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\|^2,\end{aligned} \tag{S2.67}$$

where the equality is due to Property 6, and the last step is due to the facts that $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}$ is positive definite and $\boldsymbol{\Sigma}_I$ is semi-positive definite, which implies that the minimum eigenvalue for $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}+\boldsymbol{\Sigma}_I$ is bounded away from 0. As a result of (S2.66) and (S2.67),

$$\left|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}I_3\right| = o_p\left(\sqrt{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\operatorname{Cov}(I_1+I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}}\right),$$

since $\frac{n\xi_m}{m}\to\infty$ as required in Theorem 2 (ii) .

Step 3: Examination of I_4 in (S2.63):

Recall that $\boldsymbol{\Delta}(\boldsymbol{t})\triangleq[\boldsymbol{\Delta}_1(\boldsymbol{t}),\ldots,\boldsymbol{\Delta}_p(\boldsymbol{t})]$. By (S2.30), (S2.35) and (S2.36),

we have that

$$\begin{aligned}
|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}I_4| &\leq \frac{M}{nm}\left\|\sum_{i=1}^{n}\left[\left(\boldsymbol{X}_{Ii}^{*}\otimes\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\right)\left\{\sum_{j=1}^{p}X_{ij}\boldsymbol{\Delta}_j(\boldsymbol{t})\right\}\right]\right\|\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\| \\
&= \frac{M}{nm}\left\|\mathrm{vec}\left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Delta}(\boldsymbol{t})\left(\sum_{i=1}^{n}\boldsymbol{X}_{Ii}^{*}\boldsymbol{X}_i^{\mathsf{T}}\right)\right\}\right\|\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\| \\
&= \frac{M}{nm}\left\|\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Delta}(\boldsymbol{t})\left(\sum_{i=1}^{n}\boldsymbol{X}_{Ii}^{*}\boldsymbol{X}_i^{\mathsf{T}}\right)\right\|_F\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\| \\
&\leq \frac{M}{nm}\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\|\|\boldsymbol{\Phi}(\boldsymbol{t})\|_2\|\boldsymbol{\Delta}(\boldsymbol{t})\|_F\left\|\sum_{i=1}^{n}\boldsymbol{X}_{Ii}^{*}\boldsymbol{X}_i^{\mathsf{T}}\right\|_2 \\
&= \frac{M}{nm}\|\boldsymbol{A}^{-1}\boldsymbol{\alpha}\|\cdot O_p\left(\sqrt{\frac{m}{M}}\right)\cdot O\left(\frac{\sqrt{m}}{M^q}\right)\cdot O_p(n),
\end{aligned}$$

where the second step is due to Property 2, and the fourth is due to Property 4. Thus, we have that

$$\left|\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}I_4\right| = o_p\left(\sqrt{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\,\mathrm{Cov}(I_1+I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}}\right),$$

due to the assumption $\frac{M^{2q}\xi_m}{nm}\to\infty$ as required in Theorem 2 (ii) and (S2.67).

Step 4: Examination of I_5 in (S2.63):

Because $\boldsymbol{v}_0$ is bounded away from $\mathbf{0}$ and $\hat{\boldsymbol{v}}$ converges in probability to $\boldsymbol{v}_0$ as $n\to\infty$, $\hat{\boldsymbol{v}}$ is bounded away from $\mathbf{0}$ in probability as $n\to\infty$. Therefore, we have that $\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}I_5 = 0$ with probability tending to 1 since $P'_\lambda(t) = 0$ for any fixed $t > 0$ and sufficiently small λ by Assumption E2 (ii).

In conclusion,

$$\frac{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}(I_3+I_4+I_5)}{\sqrt{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\,\mathrm{Cov}(I_1+I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}}} = o_p(1).$$

Combining this with result (S2.65) and equation (S2.63) leads to

$$\frac{\boldsymbol{\alpha}^{\mathsf{T}}(\hat{\boldsymbol{v}} - \boldsymbol{v}_0)}{\sqrt{\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\operatorname{Cov}(I_1 + I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty. \tag{S2.68}$$

We next give an explicit expression for the asymptotic variance of $\boldsymbol{\alpha}^{\mathsf{T}}(\hat{\boldsymbol{v}} - \boldsymbol{v}_0)$. By the definition of $\boldsymbol{A}$ and inserting the expression (S2.64) into the denominator of (S2.68), we have that

$$\begin{aligned}
&\boldsymbol{\alpha}^{\mathsf{T}}\boldsymbol{A}^{-1}\operatorname{Cov}(I_1 + I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha} \\
=&\frac{1}{n}\boldsymbol{\alpha}^{\mathsf{T}}\Big(\left\{\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}\left(\boldsymbol{\Sigma}_{\boldsymbol{X}_I} + \boldsymbol{\Sigma}_I\right)\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}\right\} \\
&\qquad \otimes \left[\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\right] \\
&\qquad + \boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1} \otimes (\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}) + (\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1} \otimes \boldsymbol{B}_{0s})\boldsymbol{\Gamma}(\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1} \otimes \boldsymbol{B}_{0s}^{\mathsf{T}})\Big)\boldsymbol{\alpha}, \quad \text{(S2.69)}
\end{aligned}$$

where we use Property 1 and the fact that $(\boldsymbol{A}_1 \otimes \boldsymbol{B}_1)^{-1} = \boldsymbol{A}_1^{-1} \otimes \boldsymbol{B}_1^{-1}$ for any nonsingular matrices $\boldsymbol{A}_1$ and $\boldsymbol{B}_1$ (Horn and Johnson, 1991, Section 4.2).

For $j = 1, \ldots, s$, let $\boldsymbol{\alpha}_j = \boldsymbol{e}_j \otimes \boldsymbol{\phi}(t)$, where $\boldsymbol{e}_j$ is the s-dimensional vector with the jth element being 1 and the rest being 0. Then $\|\boldsymbol{\alpha}_j\| = \|\boldsymbol{\phi}(t)\|$ is bounded away from 0 for any t. Indeed, for any fixed $t \in \mathcal{T}$, at most $d + 1$ consecutive $\phi_j(t)$ are positive while the remaining are all zero due to the construction of the B-spline basis. Without loss of generality, suppose that $\phi_j(t) > 0$ for $j = 1, \ldots, d + 1$, and $\phi_j(t) = 0$ otherwise. Let $\boldsymbol{a} = (a_1, \ldots, a_{M+d})$ with $a_j = 1$ for $j = 1, \ldots, d + 1$ and $a_j = 0$ otherwise.

By Cauchy-Schwartz's inequality, we have that

$$(d+1)\|\boldsymbol{\phi}(t)\|^2 = \|\boldsymbol{a}\|^2\|\boldsymbol{\phi}(t)\|^2 \geq \left(\sum_{j=1}^{M+d} a_j\phi_j(t)\right)^2 = \left(\sum_{j=1}^{M+d}\phi_j(t)\right)^2 = 1,$$

where the last equality follows from Farin et al. (2002, (6.6.7)). Hence, $\|\boldsymbol{\alpha}_j\| \geq \sqrt{1/(d+1)}$.

With $\boldsymbol{\alpha} = \boldsymbol{\alpha}_j$, the equation (S2.69) can be simplified as

$$\begin{aligned}
&\boldsymbol{\alpha}_j^{\mathsf{T}}\boldsymbol{A}^{-1}\operatorname{Cov}(I_1+I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}_j \\
=&\frac{1}{n}\Big((\Omega_{jj}+\tilde{\Omega}_{jj})\boldsymbol{\phi}^{\mathsf{T}}(t)\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{\Phi}(\boldsymbol{t})\}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\boldsymbol{\phi}(t) \\
&\quad + \Omega_{jj}\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\phi}(t) + \left[\boldsymbol{\Omega}_j^{\mathsf{T}}\otimes\{\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{B}_{0s}\}\right]\boldsymbol{\Gamma}\left[\boldsymbol{\Omega}_j\otimes\{\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\phi}(t)\}\right]\Big) \\
\triangleq&\sigma_j^2(t),
\end{aligned}$$

where Ω_{jj} and $\boldsymbol{\Omega}_j$ are the $(j,\, j)$ element and the jth column of $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}$, respectively; and $\tilde{\Omega}_{jj}$ is the $(j,\, j)$ element of $\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}\boldsymbol{\Sigma}_I\boldsymbol{\Sigma}_{\boldsymbol{X}_I}^{-1}$. Consequently, we have that

$$\frac{\boldsymbol{\alpha}_j^{\mathsf{T}}(\hat{\boldsymbol{v}}-\boldsymbol{v}_0)}{\sqrt{\sigma_j^2(t)}} \xrightarrow{d} N(0,1) \quad \text{as } n\to\infty. \tag{S2.70}$$

Finally, we derive the limiting distribution of $\hat{\beta}_j(t)$ for $j = 1, \ldots, s$. By definition,

$$\begin{aligned}
\hat{\beta}_j(t) - \beta_j(t) &= \boldsymbol{\phi}^{\mathsf{T}}(t)\hat{\boldsymbol{b}}_j - \beta_j(t) \\
&= \boldsymbol{\phi}^{\mathsf{T}}(t)(\hat{\boldsymbol{b}}_j - \boldsymbol{b}_{0,j}) + \boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{b}_{0,j} - \beta_j(t) \\
&= \boldsymbol{\alpha}_j^{\mathsf{T}}(\hat{\boldsymbol{v}}-\boldsymbol{v}_0) + \boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{b}_{0,j} - \beta_j(t).
\end{aligned} \tag{S2.71}$$

By (S1.2), $\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{b}_{0,j} - \beta_j(t) = O(M^{-q})$, which, by (S2.67), leads to

$$\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{b}_{0,j} - \beta_j(t) = o_p\left(\sqrt{\boldsymbol{\alpha}_j^{\mathsf{T}}\boldsymbol{A}^{-1}\operatorname{Cov}(I_1 + I_2)\boldsymbol{A}^{-1}\boldsymbol{\alpha}_j}\right) \tag{S2.72}$$

since $\frac{M^{2q}\xi_m}{nm} \to \infty$ as required in Theorem 2 (ii). Therefore, combining (S2.70), (S2.71) and (S2.72) yields that

$$\frac{\hat{\beta}_j(t) - \beta_j(t)}{\sqrt{\sigma_j^2(t)}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty.$$

S3 Simulation Studies

In this section, we conduct simulation studies to evaluate the finite-sample performance of the proposed method and also demonstrate the deleterious effects of ignoring measurement error in inferential procedures.

S3.1 Simulation Design

For $i = 1, \ldots, n$, we first generate covariate vector $\boldsymbol{X}_i$ independently from a centered multivariate normal distribution with covariance matrix whose (j, j') element is $0.5^{|j-j'|}$ for $1 \le j, j' \le p$, and then generate response $Y_i(t)$ from model (2.1), where the random error process $\varepsilon_i(t)$ is generated from a centered Gaussian process with covariance function $\Sigma_\varepsilon(t, t') = \sigma_\varepsilon^2 \rho_\varepsilon^{15|t-t'|^2}$ for $t, t' \in \mathcal{T}$. Here σ_ε^2, which is set to be 0.1, 0.25 or 0.5, representing the variance of $\varepsilon_i(t)$ for $t \in \mathcal{T}$, and ρ_ε is set to be 0.1 or 0.5 to reflect varying

auto-correlation. The common observation points are equally spaced on the domain $\mathcal{T} = [0, 1]$ and the number of observations, m, is set to be 30. Let the number of covariates $p = 6$ with the first three to be active. The functional coefficients are given by $\beta_1(t) = 2t^2$, $\beta_2(t) = \sqrt{2}\cos(3\pi t/2+\pi/2)$, $\beta_3(t) = \sqrt{2}\sin(\pi t/2) + 3\sqrt{2}\sin(3\pi t/2)$, and $\beta_4(t) = \beta_5(t) = \beta_6(t) = 0$ for $t \in \mathcal{T}$.

The measurement error $\boldsymbol{U}_i$ in model (3.7) follows a centered multivariate normal distribution with element (j, j') in the covariance matrix $\boldsymbol{\Sigma}$ to be $\Sigma_{j,j'} = \sigma_{\boldsymbol{U}}^2 \rho_{\boldsymbol{U}}^{|j-j'|}$ for $1 \leq j, j' \leq p$, where $\sigma_{\boldsymbol{U}}^2$ is the variance of each element of $\boldsymbol{U}_i$ and $\rho_{\boldsymbol{U}}$ reflects the auto-correlation of the elements in $\boldsymbol{U}_i$. To reflect different degrees of measurement error, we set $\sigma_{\boldsymbol{U}}^2 = 0.1, 0.2$ or 0.4, and $\rho_{\boldsymbol{U}} = 0.1$ or 0.5. Thus, the resulting reliability ratio for each variable, called a marginal reliability ratio, is $\mathrm{Var}(X_{ij})/\mathrm{Var}(X_{ij}^*) = 1/(1+\sigma_{\boldsymbol{U}}^2) = 0.91, 0.83$ or 0.71, respectively, where we consider a common marginal reliability ratio for all the covariates.

We consider the sample size $n = 50, 100, 200,$ or 500 and simulate $N \triangleq 300$ datasets for each parameter configuration. When implementing the proposed method, we determine tuning parameters M and λ by minimizing CV (4.21) or BIC (4.22) where we set the grid to be $\{M_k = 5k : k = 1, 2, 3\} \times \{\lambda_k = 10^{-1.5+2k/9} : k = 0, 1, \ldots, 9\}$, and take the threshold

parameter τ in (4.15) to be 10^{-4}. The impact on the choice of τ is studied in Section S3.5 of the supplementary material. The SCAD penalty (Fan and Li, 2001) is used as the penalty function.

S3.2 Analysis Methods and Performance Metrics

We analyze simulated data $\{\{\boldsymbol{Y}_i(\boldsymbol{t}), \boldsymbol{X}_i^*\} : i = 1, \ldots, n; t \in \mathcal{T}\}$ using two methods. The first method, called the naive method, disregards the difference between $\boldsymbol{X}_i^*$ and $\boldsymbol{X}_i$, which basically optimizes the objective function (3.11) with $\boldsymbol{\Sigma} = \boldsymbol{0}$. The second method, called the proposed method, accommodates the measurement error effects by implementing (3.12). As a comparison, we analyze the precisely measured values $\{\{\boldsymbol{Y}_i(\boldsymbol{t}), \boldsymbol{X}_i\} : i = 1, \ldots, n; t \in \mathcal{T}\}$ by implementing (3.12) with $L_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*)$ replaced by $\tilde{L}_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X})$ in (2.6), and call this procedure the oracle method. In implementing those methods, we consider both the CV and BIC.

We evaluate the performance of the resulting estimator using three metrics, the mean integrated squared error (MISE):

$$\text{MISE} = \frac{1}{N} \sum_{r=1}^{N} \sum_{j=1}^{p} \int_{\mathcal{T}} \left\{ \hat{\beta}_j^{[r]}(t) - \beta_j(t) \right\}^2 \mathrm{d}t,$$

the *integrated squared bias* (ISB):

$$\text{ISB} = \sum_{j=1}^{p} \int_{\mathcal{T}} \left\{ \frac{1}{N} \sum_{r=1}^{N} \hat{\beta}_j^{[r]}(t) - \beta_j(t) \right\}^2 \mathrm{d}t,$$

and the integrated variance (IV): IV = MISE − ISB, where $\hat{\beta}_j^{[r]}(t)$ is the estimated function of $\beta_j(t)$ for $j = 1, \ldots, p$ obtained from applying each of the aforementioned methods to the rth simulated dataset.

Further, regarding active covariates as "positive" and inactive covariates as "negative", we evaluate the finite sample performance of selection by the *positive selection ratio* (PSR), defined as the proportion of truly discovered active covariates among all the active covariates, and the *negative selection ratio* (NSR), defined as the proportion of truly discovered inactive covariates among all inactive covariates.

S3.3 Simulation Results - Impact of the Sample Size

First, we evaluate how the performance of the three methods may vary with the sample size, where we set $\sigma_{\boldsymbol{U}}^2 = 0.1$, $\rho_{\boldsymbol{U}} = 0.1$, $\sigma_\varepsilon^2 = 0.25$, and $\rho_\varepsilon = 0.1$, and we report in Table S3.1 the results corresponding to CV and BIC (in parentheses). Clearly, the naive method incurs notable biases in estimation, reflected by considerably larger values of ISB and IV than those of the oracle method. On the contrary, the proposed method greatly reduces the bias of the naive method, with the produced ISB values nearly identical to those output by the oracle method. While the proposed method yields larger IV values than the naive method, the overall metric MISE, combining both ISB

and IV, displays a lot smaller values for the proposed method than the naive method, especially when the sample size becomes large. The fact that the proposed method produces a larger IV value than the naive method reflects the price paid to reduce the bias in estimating the coefficient functions, and that a smaller value in the overall metric MISE for the proposed method than the naive method demonstrates the overall benefit of correcting for the measurement error effects in the inferential procedure.

In terms of variable selection, the impact of ignoring the measurement error effects may not be as profound as that for estimation; the naive method yields acceptable PSR values, especially when the sample size is not small. However, NSR values produced by the naive method noticeably deviate from those of the oracle method. As expected, the proposed method significantly improves the performance of the naive method, with the produced variable selection results fairly close to those of the oracle method.

Overall, the naive method yields unreliable results in terms of both estimation and variable selection, and biased results can exacerbate as the sample size increases. In contrast, the proposed method produces satisfactory results, which are fairly comparable to those output by the oracle method; and those results are further improved by increasing the sample

size.

While using CV and BIC yields different results, they reveal the same trend for the dependence on the sample size. In terms of estimation, using CV always results in smaller ISB and IV, and thus MISE, than BIC does, suggesting that using CV produces more accurate function estimates than using BIC. Regarding variable selection, however, using BIC seems preferred to using CV, because using BIC yields higher NSR than using CV.

S3.4 Simulation Results - Impact of the Measurement Error Degree

Here, we evaluate how different degrees of measurement error may affect the performance of the proposed method as well as the naive method. We consider the case with $n = 200$, $\sigma_\varepsilon^2 = 0.25$, and $\rho_\varepsilon = 0.1$. To reflect different magnitudes in measurement error, we consider $\rho_{\boldsymbol{U}} = 0.1$ or 0.5 and $\sigma_{\boldsymbol{U}}^2 = 0.2$ or 0.4, and report in Table S3.2 the results obtained from the three methods using both CV and BIC, with the results corresponding to BIC presented in parentheses.

Again, we observe the same patterns as in Section S3.3, except for cases with $\sigma_{\boldsymbol{U}}^2 = 0.4$ and BIC used. The proposed method outperforms the naive method in producing smaller values of MISE and larger values of NSR, while

Table S3.1: Simulated results in Section S3.3: Evaluation of the impact of the sample size n under setting with $\sigma^2_{\boldsymbol{U}} = 0.1$, $\rho_{\boldsymbol{U}} = 0.1$, $\sigma^2_{\varepsilon} = 0.25$, and $\rho_{\varepsilon} = 0.1$. The entries record the results obtained from using CV and BIC, with those corresponding to BIC included in parentheses.

n	Method	MISE	ISB	IV	PSR	NSR
50	naive	0.390(0.411)	0.249(0.261)	0.141(0.150)	0.998(0.999)	0.689(0.746)
	proposed	0.229(0.270)	0.004(0.009)	0.225(0.261)	1.000(0.996)	0.819(0.896)
	oracle	0.024(0.024)	0.000(0.000)	0.024(0.024)	1.000(1.000)	0.932(0.986)
100	naive	0.334(0.335)	0.263(0.264)	0.071(0.071)	1.000(1.000)	0.730(0.754)
	proposed	0.094(0.127)	0.000(0.007)	0.094(0.120)	1.000(1.000)	0.892(0.941)
	oracle	0.012(0.011)	0.000(0.000)	0.012(0.011)	1.000(1.000)	0.918(0.997)
200	naive	0.277(0.277)	0.239(0.239)	0.038(0.038)	1.000(1.000)	0.758(0.784)
	proposed	0.047(0.063)	0.001(0.003)	0.046(0.060)	1.000(1.000)	0.911(0.934)
	oracle	0.006(0.005)	0.000(0.000)	0.006(0.005)	1.000(1.000)	0.933(0.997)
500	naive	0.280(0.276)	0.263(0.260)	0.017(0.016)	1.000(1.000)	0.562(0.653)
	proposed	0.016(0.024)	0.000(0.001)	0.016(0.023)	1.000(1.000)	0.918(0.952)
	oracle	0.002(0.002)	0.000(0.000)	0.002(0.002)	1.000(1.000)	0.987(0.999)

both methods produce values of PSR close to 1. As the measurement error variance $\sigma^2_{\boldsymbol{U}}$ increases from 0.2 to 0.4, or equivalently, the marginal reliability ratio decreases from 0.83 to 0.71, both the naive and proposed methods

produce increased MISE values yet decreased NSR values. Interestingly, with a given $\sigma^2_{\boldsymbol{U}}$, when the autocorrelation $\rho_{\boldsymbol{U}}$ increases, an opposite pattern is observed for both methods. The impact of increasing $\rho_{\boldsymbol{U}}$ on the naive method seems more noticeable than on the proposed method.

Regarding the use of CV or BIC, again, as observed in Table S3.1, using CV tends to produce estimates with smaller MISE and NSR than using BIC. However, different from the patterns in Table S3.1, applying BIC to the proposed method does not yield satisfactory PSR values for the case with $\rho_{\boldsymbol{U}} = 0.1$ and $\sigma^2_{\boldsymbol{U}} = 0.4$, as the application of CV does.

Additionally, we evaluate how the performance of the proposed method may be influenced by other factors, including the association strength in the response model and the magnitudes of ρ_ε and σ^2_ε. Further, we assess the impact of different treatments of $\boldsymbol{\Sigma}$ on the performance of the proposed method. In particular, we conduct simulation studies for the case where $\boldsymbol{\Sigma}$ is estimated from repeated surrogate measurements of the covariates. We also consider the case where $\boldsymbol{\Sigma}$ is misspecified and conduct simulation studies accordingly. The details of these additional numerical studies are deferred to Sections S3.6- S3.8 of the supplementary material.

In summary, the simulation studies demonstrate that the naive method ignoring the measurement error effects produces biased results, and that

Table S3.2: Simulated results for Section S3.4: Evaluation of the impact of different measurement error degrees under the setting with $n = 200$, $\sigma_\varepsilon^2 = 0.25$, and $\rho_\varepsilon = 0.1$. The entries record the results obtained from using CV and BIC, with those corresponding to BIC included in parentheses.

$(\rho_{\boldsymbol{U}}, \sigma_{\boldsymbol{U}}^2)$	Method	MISE	ISB	IV	PSR	NSR
	oracle	0.006(0.006)	0.000(0.000)	0.006(0.006)	1.000(1.000)	0.918(0.997)
$(0.1, 0.2)$	naive	0.815(0.810)	0.752(0.750)	0.063(0.061)	1.000(1.000)	0.522(0.594)
	proposed	0.110(0.449)	0.004(0.306)	0.106(0.143)	1.000(0.999)	0.878(0.993)
$(0.1, 0.4)$	naive	1.984(1.976)	1.898(1.895)	0.085(0.081)	1.000(1.000)	0.359(0.448)
	proposed	0.362(2.241)	0.017(1.814)	0.345(0.427)	0.997(0.698)	0.776(0.999)
$(0.5, 0.2)$	naive	0.360(0.362)	0.319(0.320)	0.041(0.042)	1.000(1.000)	0.892(0.919)
	proposed	0.077(0.132)	0.001(0.026)	0.076(0.106)	1.000(1.000)	0.912(0.940)
$(0.5, 0.4)$	naive	1.003(1.007)	0.942(0.945)	0.061(0.062)	1.000(1.000)	0.822(0.857)
	proposed	0.207(0.720)	0.003(0.392)	0.204(0.328)	1.000(0.954)	0.827(0.983)

the proposed method greatly improves its performance and produces satisfactory results under a variety of settings. Using CV with the proposed method seems to be preferred over the use of BIC, especially for the case of parameter estimation where the reliability ratio is very low and the sample size is not big enough; in terms of inactive variable detection, the use of BIC may outperform the use of CV.

S3.5 Simulation Results - Effects of τ

As mentioned in Section 4 of the main text, the matrix $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^* - n\boldsymbol{\Sigma}$ may not be positive definite in some finite sample cases, and thus, (4.15) is essential since the unique Cholesky decomposition can only be done for positive definite matrices. To see how the nature of $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^* - n\boldsymbol{\Sigma}$ may depend on the sample size n, we consider the data generated as in Section S3.1 and report the frequencies that $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^* - n\boldsymbol{\Sigma}$ is not positive definite in Table S3.3. It shows that the frequency increases with an increasing $\sigma^2_{\boldsymbol{U}}$ and decreases with an increasing sample size n. Sufficiently large sample size almost always guarantees the matrix $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^* - n\boldsymbol{\Sigma}$ to be positive definite .

Further simulation studies are conducted to study the impact of the choice of the threshold parameter τ in (4.15). As opposed to the mean integrated squared error (MISE) considered in Section S3.2, we calculate the standard deviation of the integrated squared error $\sum_{j=1}^{p} \int_{\mathcal{T}} \left\{ \hat{\beta}_j^{[r]}(t) - \beta_j(t) \right\}^2 \mathrm{d}t$ for $r = 1, \ldots, N$, and let SDISE represents it, where $N = 300$ is the number of replicates and $\hat{\beta}_j^{[r]}(t)$ is the estimated function of $\beta_j(t)$ for $j = 1, \ldots, p$ by applying the proposed method to the rth simulated dataset. We display values of MISE, SDISE (in parentheses), mean of PSR, and NSR in Table S3.4 under the same setting in Section S3.1. As shown in the table, the performance of the proposed method does not seem to be very sensitive to

Table S3.3: Frequencies of non-positive definiteness of matrix $\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^{*} - n\boldsymbol{\Sigma}$, with $\rho_{\boldsymbol{U}}$=0.1.

n	$\sigma^2_{\boldsymbol{U}}$	marginal reliability ratio	Frequency
50	0.2	0.833	0.000
	0.4	0.714	0.063
	0.6	0.625	0.257
100	0.2	0.833	0.000
	0.4	0.714	0.000
	0.6	0.625	0.020
200	0.2	0.833	0.000
	0.4	0.714	0.000
	0.6	0.625	0.000

the choice of τ for a wide range of values. However, it is risky to choose relatively large τ for an extremely small sample size. For larger $\sigma^2_{\boldsymbol{U}}$ (i.e., smaller marginal reliability ratio), it is safer to choose a smaller τ to obtain accurate estimates. The results based on the BIC tuning parameter selection strategy are similar and thus omitted.

Table S3.4: Simulation results obtained from using CV under the setting with $\rho_{\boldsymbol{U}} = 0.1$, $\sigma_{\varepsilon}^2 = 0.25$ and $\rho_{\varepsilon} = 0.1$.

$\sigma_{\boldsymbol{U}}^2$	n	τ	MISE (SDISE)	PSR	NSR
0.4	50	0.0001	2.532 (2.726)	0.778	0.931
		0.001	2.532 (2.726)	0.778	0.931
		0.01	3240 (36227)	0.691	0.907
	100	0.0001	0.999 (1.181)	0.973	0.811
		0.001	0.999 (1.182)	0.973	0.811
		0.01	0.999 (1.183)	0.973	0.811
	200	0.0001	0.396 (0.570)	1.000	0.813
		0.001	0.396 (0.570)	1.000	0.813
		0.01	0.396 (0.570)	1.000	0.813
0.6	50	0.0001	4.156 (3.533)	0.574	0.948
		0.001	4.156 (3.533)	0.574	0.948
		0.01	25381 (135641)	0.480	0.810
	100	0.0001	2.080 (2.379)	0.827	0.911
		0.001	2.080 (2.380)	0.827	0.911
		0.01	2.652 (4.346)	0.803	0,901
	200	0.0001	0.896 (1.233)	0.981	0.797
		0.001	0.896 (1.233)	0.981	0.797
		0.01	0.896 (1.233)	0.981	0.797

S3.6 Simulation Results - Effect of ρ_ε and σ^2_ε

We evaluate how different association strength in the response model may influence the performance of the proposed estimator, and report the results in Table S3.5 for the case with $n = 100$, $\sigma^2_{\boldsymbol{U}} = 0.1$ and $\rho_{\boldsymbol{U}} = 0.1$, using CV or BIC (results are presented in parentheses), where we consider $\rho_\varepsilon = 0.1$ or 0.5 and $\sigma^2_\varepsilon = 0.1$ or 0.5. As expected, the oracle method always produces the best results among the three methods for all settings. Compared to the naive method, the proposed method produces consistently smaller values of MISE and larger values of NSR, yet both methods give satisfactory values of PSR. For all the three methods, an increase in σ^2_ε leads to an increase in MISE, but has relatively little effect on changing values of PSR and NSR. The impact of ρ_ε does not appear to be substantial. Under the setting with a small value of $\sigma^2_{\boldsymbol{U}}$ (i.e., $\sigma^2_{\boldsymbol{U}} = 0.1$), the performance of all the three methods appear stable, regardless of whether CV or BIC is used, though using CV tends to give smaller values for MISE and NSR than BIC does in all settings.

Table S3.5: Simulation results for Section S3.6: Evaluation of the impact of different association strengths in the response model, where $n = 100$, $\sigma_{\boldsymbol{U}}^2 = 0.1$, and $\rho_{\boldsymbol{U}} = 0.1$. The entries record the results obtained from using CV and BIC, with the results corresponding to BIC included in parentheses.

$(\rho_\varepsilon, \sigma_\varepsilon^2)$	Method	MISE	ISB	IV	PSR	NSR
$(0.1, 0.1)$	naive	0.325 (0.328)	0.260 (0.264)	0.065 (0.064)	1.000 (1.000)	0.761 (0.788)
	proposed	0.089 (0.112)	0.000 (0.004)	0.088 (0.108)	1.000 (1.000)	0.908 (0.927)
	oracle	0.005 (0.005)	0.000 (0.000)	0.005 (0.005)	1.000 (1.000)	0.943 (0.999)
$(0.1, 0.5)$	naive	0.343 (0.347)	0.262 (0.264)	0.081 (0.083)	1.000 (1.000)	0.733 (0.791)
	proposed	0.106 (0.137)	0.000 (0.007)	0.106 (0.130)	1.000 (1.000)	0.894 (0.951)
	oracle	0.024 (0.023)	0.000 (0.000)	0.023 (0.023)	1.000 (1.000)	0.910 (0.989)
$(0.5, 0.1)$	naive	0.329 (0.330)	0.264 (0.265)	0.065 (0.065)	1.000 (1.000)	0.767 (0.779)
	proposed	0.088 (0.119)	0.000 (0.006)	0.088 (0.112)	1.000 (1.000)	0.911 (0.932)
	oracle	0.005 (0.005)	0.000 (0.000)	0.005 (0.005)	1.000 (1.000)	0.947 (0.991)
$(0.5, 0.5)$	naive	0.345 (0.352)	0.262 (0.267)	0.083 (0.085)	1.000 (1.000)	0.749 (0.791)
	proposed	0.111 (0.141)	0.000 (0.008)	0.111 (0.133)	1.000 (1.000)	0.881 (0.939)
	oracle	0.025 (0.025)	0.000 (0.000)	0.025 (0.024)	1.000 (1.000)	0.928 (0.982)

S3.7 Simulation Study with the Measurement Error Covariance Matrix $\boldsymbol{\Sigma}$ Estimated

In this subsection, we investigate the performance of the proposed method for the case where $\boldsymbol{\Sigma}$ is unknown, but replicate surrogate measurement are available for certain study subjects, allowing for its estimation. In particular, we assume that half of n samples have additional K repeated surrogate measurements. For $k = 1, \ldots, K+1$ and $i = 1, \ldots, n/2$, with n assumed to be even, let $\boldsymbol{X}^*_{i[k]}$ denote the kth surrogate measurement of $\boldsymbol{X}_i$. Then using the method of moments, we estimate $\boldsymbol{\Sigma}$ by

$$\frac{2}{n}\sum_{i=1}^{n/2}\left\{\frac{1}{K}\sum_{k=1}^{K+1}(\boldsymbol{X}^*_{i[k]}-\bar{\boldsymbol{X}}^*_i)(\boldsymbol{X}^*_{i[k]}-\bar{\boldsymbol{X}}^*_i)^{\mathsf{T}}\right\}, \tag{S3.1}$$

where $\bar{\boldsymbol{X}}^*_i \triangleq \frac{1}{K+1}\sum_{k=1}^{K+1}\boldsymbol{X}^*_{i[k]}$. See equation (4.3) of Carroll et al. (2006) for a more general formula applicable to varying numbers of replicates for different subjects.

We generate $\{\{\boldsymbol{X}_i, Y_i(t)\} : t \in \mathcal{T}; i = 1\ldots, n\}$ and $\{\boldsymbol{X}^*_i : i = \frac{n}{2}+1, \ldots, n\}$ using the same manner of Section S3.1, and $\{\boldsymbol{X}^*_{i[k]} : k = 1, \ldots, K+1\}$ are generated independently from model (3.7) in the main text for $i = 1, \ldots, \frac{n}{2}$, where $\sigma_{\boldsymbol{U}} = 0.1$, $\rho_{\boldsymbol{U}} = 0.1$, $\sigma^2_\varepsilon = 0.25$ and $\rho_\varepsilon = 0.1$, together with $n = 50$ or 200. First, we apply (S3.1) to the repeated surrogate measurements $\{\boldsymbol{X}^*_{i[k]} : k = 1, \ldots, K+1; i = 1, \ldots, \frac{n}{2}\}$ to obtain an estimate

$\hat{\boldsymbol{\Sigma}}$ of $\boldsymbol{\Sigma}$. Next, having $\boldsymbol{\Sigma}$ replaced by $\hat{\boldsymbol{\Sigma}}$, we apply the proposed method to the data $\{\{\boldsymbol{X}_i^*, Y_i(t)\} : t \in \mathcal{T}, i = 1, \ldots, n\}$, with $\boldsymbol{X}_i^*$ taken as $\boldsymbol{X}_{i[1]}^*$ for $i = 1, \ldots, \frac{n}{2}$. where CV is used for the tuning parameter selection as in Section S3.1.

We consider $K = 0, 1$ or 2, where we let $K = 0$ represent the case for which no replicated surrogate measurements are available and $\boldsymbol{\Sigma}$ is taken as known when applying the proposed method. The results are reported in Table S3.6 in the same manner as for Table S3.1.

As expected, the performance of the proposed method improves as K increases due to the better estimation of $\boldsymbol{\Sigma}$ with a larger K. When $n = 200$, the proposed method with estimated $\boldsymbol{\Sigma}$ from $K = 4$ performs analogously to that for the case with $K = 0$.

S3.8 Effects of the Misspecified Measurement Error Covariance Matrix

In this subsection, we assess the sensitivity of the proposed method to the misspecification of the measurement error covariance matrix $\boldsymbol{\Sigma}$. In Section S3.1, the covariance matrix $\boldsymbol{\Sigma}$ used to simulate data is specified as the matrix, denoted as $\boldsymbol{\Sigma}_s$, with the (j, j') element set to be $\sigma_{\boldsymbol{U}}^2 \rho_{\boldsymbol{U}}^{|j-j'|}$ for $1 \leq j, j' \leq p$, where we consider $\sigma_{\boldsymbol{U}}^2 = \rho_{\boldsymbol{U}} = 0.1$. When implementing the

Table S3.6: Simulation results for Section S3.7: Evaluation of the impact of the number of additional repeated measurements K using CV, where $\sigma_{\boldsymbol{U}} = 0.1$, $\rho_{\boldsymbol{U}} = 0.1$, $\sigma_{\varepsilon}^2 = 0.25$, $\rho_{\varepsilon} = 0.1$, and $n = 50$ or 200.

		MISE	ISB	IV	PSR	NSR
n=50	$K = 0$	0.229	0.004	0.225	1.000	0.819
	$K = 1$	0.340	0.003	0.337	0.997	0.783
	$K = 4$	0.239	0.004	0.235	0.998	0.823
n=200	$K = 0$	0.047	0.001	0.046	1.000	0.911
	$K = 1$	0.054	0.001	0.053	1.000	0.878
	$K = 4$	0.047	0.001	0.046	1.000	0.903

proposed method, the covariance matrix $\boldsymbol{\Sigma}$ for the error term $\boldsymbol{U}_i$ is mistaken as a working matrix, denoted as $\boldsymbol{\Sigma}_w$, whose (j, j') element is set as $\tilde{\sigma}_{\boldsymbol{U}}^2 \tilde{\rho}_{\boldsymbol{U}}^{|j-j'|}$, where in case 1, we take $\tilde{\sigma}_{\boldsymbol{U}}^2 = \gamma_1 \sigma_{\boldsymbol{U}}^2$ and $\tilde{\rho}_{\boldsymbol{U}} = \rho_{\boldsymbol{U}}$ for a positive constant γ_1; and in case 2, we let $\tilde{\sigma}_{\boldsymbol{U}}^2 = \sigma_{\boldsymbol{U}}^2$ and $\tilde{\rho}_{\boldsymbol{U}} = \gamma_2 \rho_{\boldsymbol{U}}$ for a constant γ_2. When $\gamma_1 = 1$ or $\gamma_2 = 1$, the working matrix $\boldsymbol{\Sigma}_w$ is identical to the true matrix $\boldsymbol{\Sigma}_s$ used to generate data.

Tables S3.7 and S3.8 display the results for cases 1 and 2, respectively, in the same manner as for Table S3.1, where we take $\gamma_1 \in \{0.5, 0.8, 1.0, 1.2, 1.5, 2.0\}$, $\gamma_2 \in \{-1.0, -0.5, 0, 0.5, 1.0, 2.0\}$, and the results for the naive method are displayed in parentheses. Clearly, the values of MISE and ISB minimize

when true $\mathbf{\Sigma}$ is used in both cases. The values of IV tend to increase as γ_1 increases or γ_2 decreases; they decrease as n increases. The values of PSR are always 1 for all settings. When n is large, the best NSR is obtained when true $\mathbf{\Sigma}$ is used. However, best NSR values may be reached for $\gamma_1 > 1$ or $\gamma_2 < 1$ if n is small. Similar performance to the true $\mathbf{\Sigma}$-based method is achievable when the discrepancy between the working matrix $\mathbf{\Sigma}_w$ and $\mathbf{\Sigma}_s$ is moderate. However, a large discrepancy, such as $\gamma_1 = 2$, can lead to the proposed method performing even worse than the naive method, as shown in Table S3.7.

Table S3.7: Simulation results for Section S3.8: Misspecification of $\mathbf{\Sigma}$ for Case 1 using CV, where $\sigma_\varepsilon^2 = 0.25$, $\rho_\varepsilon = 0.1$, and $n = 100$ or 500. The results for the naive method is included in parentheses.

		$\gamma_1 = 0.5$	0.8	**1**	1.2	1.5	2
$n = 100$	MISE	0.148	0.096	**0.094** (0.334)	0.126	0.278	1.003
	ISB	0.074	0.012	**0.000** (0.263)	0.018	0.127	0.646
	IV	0.074	0.084	**0.094** (0.071)	0.108	0.151	0.357
	PSR	1.000	1.000	**1.000** (1.000)	1.000	1.000	1.000
	NSR	0.861	0.891	**0.892** (0.730)	0.916	0.903	0.837
$n = 500$	MISE	0.086	0.025	**0.016** (0.280)	0.037	0.151	0.742
	ISB	0.071	0.011	**0.000** (0.263)	0.017	0.119	0.666
	IV	0.015	0.014	**0.016** (0.017)	0.020	0.032	0.076
	PSR	1.000	1.000	**1.000** (1.000)	1.000	1.000	1.000
	NSR	0.786	0.904	**0.918** (0.562)	0.913	0.823	0.611

Table S3.8: Simulation results for Section S3.8: Misspecification of $\mathbf{\Sigma}$ for Case 2 using CV, where $\sigma_\varepsilon^2 = 0.25$, $\rho_\varepsilon = 0.1$, and $n = 100$ or 500. The results for the naive method is included in parentheses.

		$\gamma_2 = -1$	-0.5	0	0.5	**1**	2
$n = 100$	MISE	0.166	0.133	0.112	0.098	**0.094** (0.334)	0.097
	ISB	0.042	0.022	0.009	0.002	**0.000** (0.263)	0.006
	IV	0.124	0.111	0.103	0.096	**0.094** (0.071)	0.091
	PSR	1.000	1.000	1.000	1.000	**1.000** (1.000)	1.000
	NSR	0.900	0.912	0.911	0.923	**0.892** (0.730)	0.874
$n = 500$	MISE	0.076	0.047	0.030	0.020	**0.016** (0.280)	0.021
	ISB	0.044	0.022	0.009	0.002	**0.000** (0.263)	0.005
	IV	0.031	0.025	0.021	0.018	**0.016** (0.017)	0.016
	PSR	1.000	1.000	1.000	1.000	**1.000** (1.000)	1.000
	NSR	0.768	0.851	0.889	0.916	**0.918** (0.562)	0.881

S4 Analysis Results Additional to Section 6

This section records additional results for Section 6, which reports some results for $a = c = 1$ only. First, for $a = c = 1$, we display in Figure S4.1 the estimates of $\beta_j(t)$ corresponding to each of the twelve covariates obtained by using either CV or BIC to choose a suitable tuning parameter value. Estimates of $X_{i,9}$ and $X_{i,10}$ are quite different with the use of CV or BIC, for which the corresponding covariates $X_{i,9}$ and $X_{i,10}$ are excluded by CV but not by BIC. Additionally, all the covariates excluded by BIC are also excluded by CV.

Next, we report sensitivity in estimating $\beta_1(t)$, the coefficient function corresponding to the BMI Z-score ($X_{i,1}$), when different degrees of measurement error are present in the data, and display the estimates of $\beta_1(t)$ in Figure S4.2, where $a = 0, 1, 2$ and 3 and $c = 0, 5, 10$ and 15 are considered. Here $a = 0$ corresponds to the naive method which ignores the differences between $\boldsymbol{X}_i$ and $\boldsymbol{X}_i^*$. It seems that the value of a has more noticeable impact on estimation results than that of c. All other estimated functions share analogous trends and the variable selection results are consistent among all the chosen values for a and c, and are not reported here.

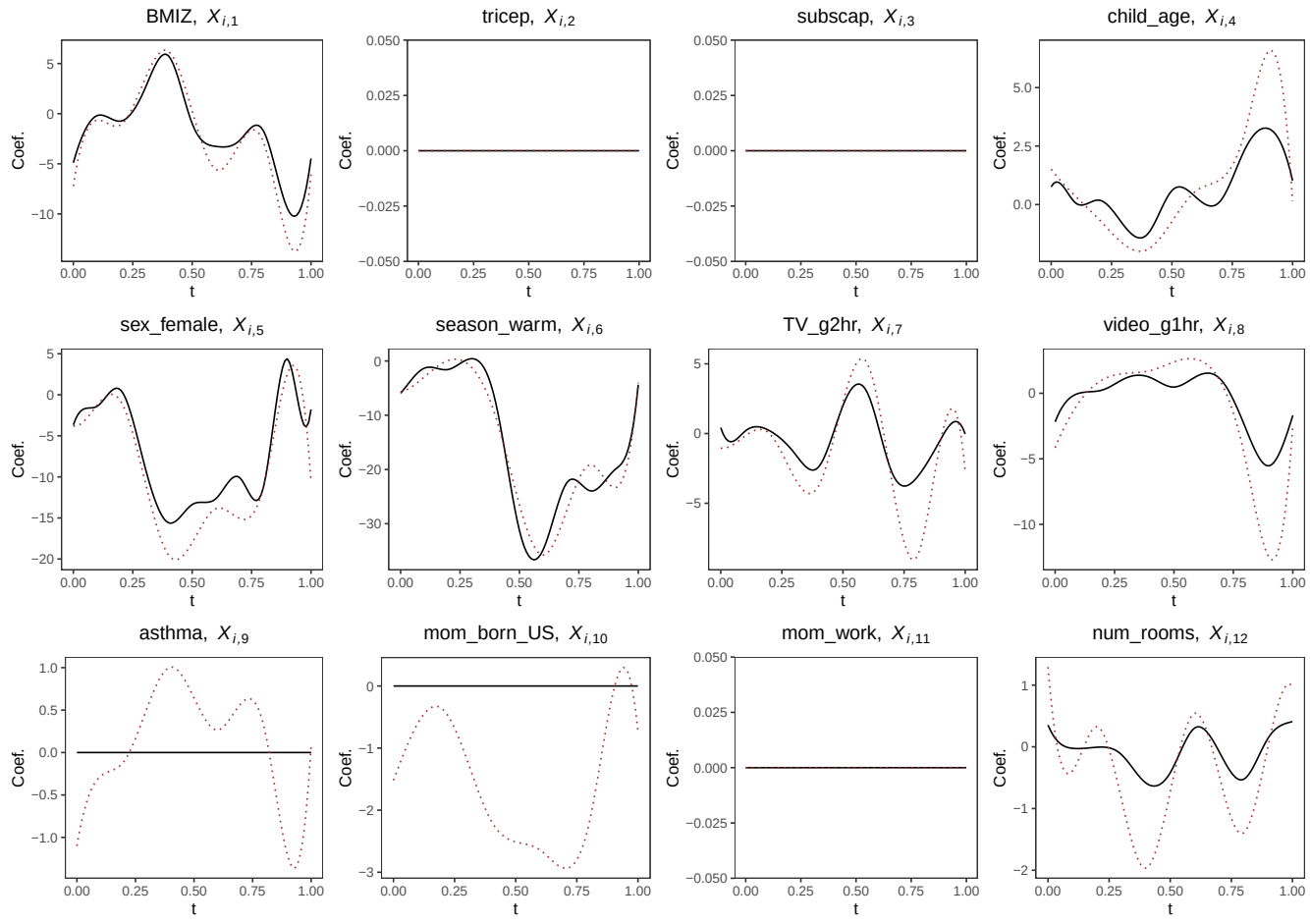

Figure S4.1: The estimates of $\beta_j(t)$ corresponding to each of the twelve covariates obtained by using CV (solid) or BIC (dotted) to choose a tuning parameter value with $a = c = 1$.

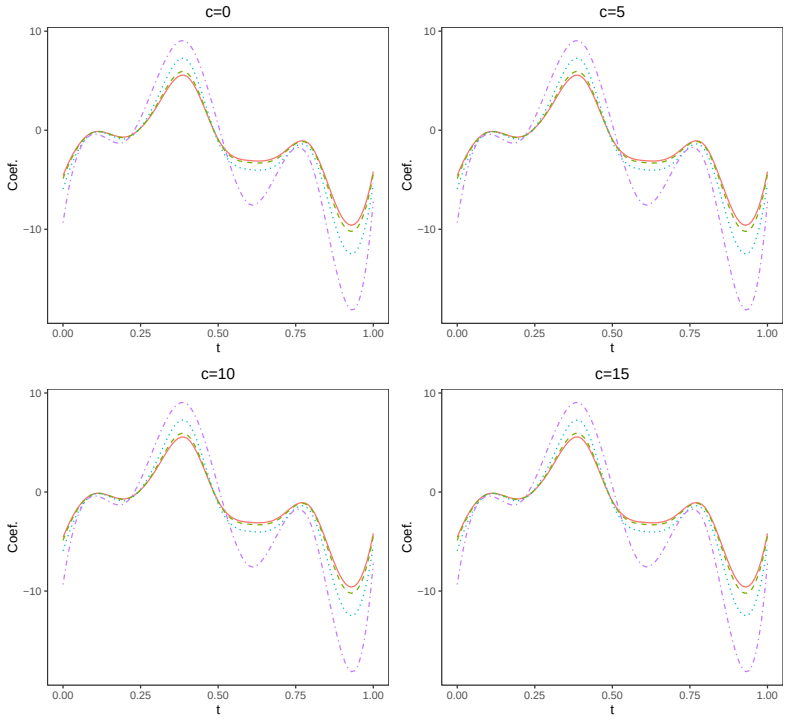

Figure S4.2: BMI Z-score coefficient function estimates for different values of a and c: the estimates corresponding to $a = 0, 1, 2$ and 3 are expressed by curves marked with solid red, dash green, dotted blue and dash-dotted purple, respectively.

S5 Extension to Accommodating the Generalized Least Squares Loss Function

S5.1 Method and Properties

The simple least squares loss function (2.6) focuses on expressing the differences between the responses $\boldsymbol{Y}(\boldsymbol{t})$ and their approximate mean $(\boldsymbol{X}\otimes\boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b}$, without accounting for the correlation within the error process $\boldsymbol{\varepsilon}(\boldsymbol{t})$ over $\boldsymbol{t}$. In this section, we extend the method in the main text to accommodate generalized least squares loss functions. Let $\boldsymbol{G}$ denote a symmetric positive-definite $m \times m$ matrix. Define

$$\tilde{L}_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}) \triangleq \frac{1}{2}\sum_{i=1}^{n} \left\|\boldsymbol{Y}_i(\boldsymbol{t}) - (\boldsymbol{X}_i^{\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\boldsymbol{b}\right\|_G^2, \tag{S5.2}$$

where $\|\boldsymbol{a}\|_G^2$ represents $\boldsymbol{a}^{\mathsf{T}}\boldsymbol{G}\boldsymbol{a}$ for any m-dimensional vector $\boldsymbol{a}$. The new loss function (S5.2) differs from (2.6) in the inclusion of a weight matrix $\boldsymbol{G}$.

Under the model (3.7), (3.9) can be modified as

$$\mathbb{E}\left\{\tilde{L}_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*)\right\} = \mathbb{E}\left\{\tilde{L}_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X})\right\} + \frac{n}{2}\boldsymbol{b}^{\mathsf{T}}\left[\boldsymbol{\Sigma} \otimes \left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{b},$$

which motivates the generalized version of (3.10):

$$L_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*) \triangleq \tilde{L}_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*) - \frac{n}{2}\boldsymbol{b}^{\mathsf{T}}\left[\boldsymbol{\Sigma} \otimes \left\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})\right\}\right]\boldsymbol{b}. \tag{S5.3}$$

Define

$$Q_n^G(\boldsymbol{b}) = L_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*) + nm \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|),$$

which is identical to (3.11) except replacing the loss function $L_n(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*)$ with $L_n^G(\boldsymbol{b}; \boldsymbol{Y}(\boldsymbol{t}), \boldsymbol{X}^*)$.

Replacing $Q_n(\boldsymbol{b})$ in (3.12) with $Q_n^G(\boldsymbol{b})$, we let $\hat{\boldsymbol{b}}^G$ denote the resulting estimators, and let $\hat{\beta}_j^G$ for $j = 1, \ldots, p$ denote the estimators determined by (3.13) with $\hat{\boldsymbol{b}}_j$ replaced by the jth element of $\hat{\boldsymbol{b}}^G$. Analogous to Theorems 1 and 2 in the main text, we establish asymptotic properties for $\hat{\boldsymbol{b}}^G$ and $\hat{\beta}_j^G$ with $j = 1, \ldots, p$.

Corollary 1. *Assume that the weight matrix $\boldsymbol{G}$ satisfies*

$$C_4'^{-1}\zeta_m' \leq \rho_{min}(\boldsymbol{G}) \leq \rho_{max}(\boldsymbol{G}) \leq C_4'\zeta_m, \tag{S5.4}$$

where C_4' is a positive constant, ζ_m and ζ_m' are sequences of m satisfying $\zeta_m/\zeta_m' = o(\sqrt{n/M})$. Then under the conditions of Theorem 1, there exists a local minimizer $\hat{\boldsymbol{b}}^G$ of $Q_n^G(\boldsymbol{b})$ such that

$$\|\hat{\boldsymbol{b}}^G - \boldsymbol{b}_0\| = O_p(\zeta_m\sqrt{M/n}/\zeta_m'),$$

and hence, for $j = 1, \ldots, p$,

$$\|\hat{\beta}_j^G - \beta_j\|_{L_\infty} = O_p(\zeta_m\sqrt{M/n}/\zeta_m') \text{ and } \|\hat{\beta}_j^G - \beta_j\|_{L_2} = O_p(\zeta_m\sqrt{1/n}/\zeta_m').$$

Remark 3. In Corollary 1, ζ_m is allowed to be divergent and ζ_m' is allowed to converge to 0 as $m \to \infty$. For example, if we choose $\boldsymbol{G}$ to be the inverse

of $\mathbf{\Sigma}_\varepsilon$, then ζ_m can be ξ_m^{-1}, as discussed in Remark 2, and thus, can be divergent for some settings of $\varepsilon_i(t)$. Also, the largest eigenvalue of $\mathbf{\Sigma}_\varepsilon$ may be divergent as $m \to \infty$, and thus, the minimum eigenvalue of the inverse of $\mathbf{\Sigma}_\varepsilon$, $\boldsymbol{G}$, converges to 0, showing that $\zeta'_m \to 0$. Corollary 1 implies that incorporating $\boldsymbol{G}$ may result in a lower convergence rate than Theorem 1.

The following corollary encompasses Theorem 2 as a special case by setting $\zeta'_m = \zeta_m = 1$ and $\boldsymbol{G}$ to be the identity matrix.

Corollary 2. *Suppose the conditions in Corollary 1 hold. Assume further that* $\lambda\sqrt{n/M} \cdot \zeta'_m/\zeta_m \to \infty$ *and* $\lambda\sqrt{n} \cdot \zeta'_m/\zeta_m^2 \to \infty$ *as* $n \to \infty$.

(i) Then with probability tending to 1, the minimizer $\hat{\boldsymbol{b}}$ *satisfies* $\hat{\boldsymbol{b}}_j = \mathbf{0}$ *for all* $j \in J_0$, *and thus,* $\hat{J}_0 = J_0$.

(ii) Assume $\rho_{min}(\mathbf{\Sigma}_{\boldsymbol{\varepsilon}}) \geq C_5\xi_m$, *where* C_5 *is a positive constant and* $\{\xi_m : m = 1, 2, \dots\}$ *is a sequence of constants satisfying* $\frac{n\xi_m\zeta_m'^4}{m\zeta_m^4} \to \infty$ *and* $\frac{M^{2q}\xi_m\zeta_m'^2}{nm\zeta_m^2} \to \infty$ *as* $n \to \infty$ *and* $m \to \infty$. *Then for any* $t \in \mathcal{T}$ *and* $j \notin J_0$, *we have that*

$$\frac{\hat{\beta}_j(t) - \beta_j(t)}{\sqrt{\sigma_j^2(t)}} \xrightarrow{d} N(0,1) \quad \text{as } n \to \infty,$$

where

$$\begin{aligned}&\sigma_j^2(t)\\ =&\frac{1}{n}\Big((\Omega_{jj}+\tilde{\Omega}_{jj})\boldsymbol{\phi}^{\mathsf{T}}(t)\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})\}\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\boldsymbol{\phi}(t)\\ &+\Omega_{jj}\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{B}_{0s}\boldsymbol{\Sigma}_I\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\phi}(t)+\left[\boldsymbol{\Omega}_j^{\mathsf{T}}\otimes\{\boldsymbol{\phi}^{\mathsf{T}}(t)\boldsymbol{B}_{0s}\}\right]\boldsymbol{\Gamma}\left[\boldsymbol{\Omega}_j\otimes\{\boldsymbol{B}_{0s}^{\mathsf{T}}\boldsymbol{\phi}(t)\}\right]\Big)\end{aligned}$$

with Ω_{jj}, $\boldsymbol{\Omega}_j$ *and* $\tilde{\Omega}_{jj}$ *given in Theorem 2.*

Comparing the variance $\sigma_j^2(t)$ in Corollary 2 and the counterpart in Theorem 2, the only difference is the inclusion of $\boldsymbol{G}$ in the first term. Specifically, if we set $\boldsymbol{G}=\boldsymbol{\Sigma}_{\varepsilon}^{-1}$, the first term can be simplified as

$$\frac{1}{n}(\Omega_{jj}+\tilde{\Omega}_{jj})\boldsymbol{\phi}^{\mathsf{T}}(t)\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{\Sigma}_{\boldsymbol{\varepsilon}}^{-1}\boldsymbol{\Phi}(\boldsymbol{t})\}^{-1}\boldsymbol{\phi}(t).$$

The proofs of Corollaries 1 and 2 are similar to those for Theorems 1 and 2 with slight modifications, and thus, are omitted.

S5.2 Implementation

The algorithm of minimizing $Q_n^G(\boldsymbol{b})$ is analogous to Section 4 in the main text, except the inclusion of the matrix $\boldsymbol{G}$. To be specific, let $\boldsymbol{W}^G \triangleq (\boldsymbol{X}^{*\mathsf{T}}\boldsymbol{X}^*-n\boldsymbol{\Sigma})\otimes\{\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})\}$ be the counterpart of $\boldsymbol{W}$ in (4.14). We then obtain $\overline{\boldsymbol{W}}^G$ by (4.15) and the Cholesky decomposition $\overline{\boldsymbol{W}}^G=(\boldsymbol{V}^G)^{\mathsf{T}}\boldsymbol{V}^G$.

Define $\tilde{\boldsymbol{b}}^G \triangleq (\overline{\boldsymbol{W}}^G)^{-1}(\boldsymbol{X}^*\otimes\boldsymbol{\Phi}(\boldsymbol{t}))^{\mathsf{T}}\mathbb{G}\boldsymbol{Y}(\boldsymbol{t})$, where $\mathbb{G}\triangleq\mathrm{diag}(\boldsymbol{G},\cdots,\boldsymbol{G})$

with n blocks of $\boldsymbol{G}$. We then consider minimizing the objective function

$$\begin{aligned}\tilde{Q}_n^G(\boldsymbol{b}) &= \frac{1}{2nm}\left\|\boldsymbol{V}^G\tilde{\boldsymbol{b}}^G - \boldsymbol{V}^G\boldsymbol{b}\right\|^2 + \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|) \\ &= \frac{1}{2(M+d)p}\left\|\sqrt{\frac{(M+d)p}{nm}}\boldsymbol{V}^G\tilde{\boldsymbol{b}}^G - \sqrt{\frac{(M+d)p}{nm}}\boldsymbol{V}^G\boldsymbol{b}\right\|^2 + \sum_{j=1}^{p} P_\lambda(\|\boldsymbol{b}_j\|).\end{aligned}$$

The remaining procedure is identical to the discussion for (4.19) in the main text, except that the term $\boldsymbol{V}^G\tilde{\boldsymbol{b}}^G$ is now computed by solving the equation

$$(\boldsymbol{V}^G)^\mathsf{T}\tilde{\boldsymbol{Y}} = (\boldsymbol{X}^* \otimes \boldsymbol{\Phi}(\boldsymbol{t}))^\mathsf{T}\,\mathbb{G}\boldsymbol{Y}(\boldsymbol{t})$$

for unknown $\tilde{\boldsymbol{Y}}$.

For the tuning parameters selection, the CV method is similar to (4.21) except the ordinary $\|\cdot\|$ term is replaced by $\|\cdot\|_G$ and the $\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{\Phi}(\boldsymbol{t})$ is now $\boldsymbol{\Phi}^\mathsf{T}(\boldsymbol{t})\boldsymbol{G}\boldsymbol{\Phi}(\boldsymbol{t})$. Analogously, BIC is the same as (4.22) with $\|\cdot\|$ replaced by $\|\cdot\|_\mathbb{G}$, where $\|\boldsymbol{a}\|_\mathbb{G}^2$ represents $\boldsymbol{a}^\mathsf{T}\mathbb{G}\boldsymbol{a}$ for any mn-dimensional vector $\boldsymbol{a}$.

S5.3 A Simulation Study

In the literature of generalized least squares (GLS) methods, the matrix $\boldsymbol{G}$ is commonly set as the inverse matrix $\boldsymbol{\Sigma}_\varepsilon^{-1}$. In this subsection, we compare the performance of the original method based on least squares (LS) which minimizes (3.11), to the extended approach with a generalized least squares function, where we set $\boldsymbol{G}$ to be $\boldsymbol{\Sigma}_\varepsilon^{-1}$ or $\hat{\boldsymbol{\Sigma}}_\varepsilon^{-1}$, and let GLS0 and GLS refer to those methods, respectively. Here $\hat{\boldsymbol{\Sigma}}_\varepsilon^{-1}$ is simply obtained using the

residuals, formed as follows: the residual for unit i is given by

$$\hat{\boldsymbol{e}}_i \triangleq \boldsymbol{Y}_i(\boldsymbol{t}) - (\boldsymbol{X}_i^{*\mathsf{T}} \otimes \boldsymbol{\Phi}(\boldsymbol{t}))\hat{\boldsymbol{b}} \text{ for } i = 1, \ldots, n,$$

where $\hat{\boldsymbol{b}}$ denotes the estimate obtained from the LS method (3.12). To account for the use of the contaminated $\boldsymbol{X}_i^*$, we estimate $\boldsymbol{\Sigma}_\varepsilon$ by adding a debias term as follows:

$$\hat{\boldsymbol{\Sigma}}_\varepsilon = \frac{1}{n}\sum_{i=1}^{n} \hat{\boldsymbol{e}}_i \hat{\boldsymbol{e}}_i^{\mathsf{T}} - \boldsymbol{\Phi}(\boldsymbol{t})\hat{\boldsymbol{B}}\boldsymbol{\Sigma}\hat{\boldsymbol{B}}^{\mathsf{T}}\boldsymbol{\Phi}^{\mathsf{T}}(\boldsymbol{t}), \tag{S5.5}$$

where $\hat{\boldsymbol{B}}$ is the $(M+d) \times p$ matrix version of $\hat{\boldsymbol{b}}$ satisfying $\text{vec}(\hat{\boldsymbol{B}}) = \hat{\boldsymbol{b}}$. We can then compute the inverse $\hat{\boldsymbol{\Sigma}}_\varepsilon^{-1}$ for GLS. If there is no measurement error, the first part of $\hat{\boldsymbol{\Sigma}}_\varepsilon$ in (S5.5) is often considered in multivariate responses models. See, for example, Sofer et al. (2014) and the reference therein. In functional response models, Chen et al. (2016) proposed a similar procedure.

The setting is analogous to those in Section S3.1 except for the random error. Indeed, direct calculation of $\boldsymbol{\Sigma}_\varepsilon^{-1}$ based on the definition, i.e., $\Sigma_\varepsilon(t, t') = \sigma_\varepsilon^2 \rho_\varepsilon^{15|t-t'|^2}$ for $t, t' \in \mathcal{T}$, is unstable of infeasible since those $\varepsilon_i(t_1), \ldots, \varepsilon_i(t_m)$ are highly correlated even when ρ_ε is very small. Alternatively, we consider enhancing the diagonal elements of $\hat{\boldsymbol{\Sigma}}_\varepsilon$ by defining $\Sigma_\varepsilon(t, t') = \sigma_\varepsilon^2 \rho_\varepsilon^{15|t-t'|^2} + \sigma_\varepsilon'^2 \mathbf{1}(t = t')$ for $t, t' \in \mathcal{T}$ now, where σ_ε' is a positive constant.

Repeating the experiments for three methods 300 times, Table S5.1 re-

ports MISE and SDISE (displayed in parentheses), the definition of which is given in Section S3.5) and means of PSR anf NSR under different sample sizes n using the CV method. Clearly, GLS0 is consistently the best, though it only slightly outperforms LS . When n is relatively small, GLS can perform worse than LS because of inaccurate estimation of $\mathbf{\Sigma}_\varepsilon$.

Table S5.1: Simulation results for comparing three methods under the setting with $\sigma_{\boldsymbol{U}}^2 = 0.1$, $\rho_{\boldsymbol{U}} = 0.1$, $\sigma_\varepsilon^2 = \sigma_\varepsilon'^2 = 0.25$, and $\rho_\varepsilon = 0.1$

n	Method	MISE (SDISE)	PSR	NSR
100	LS	0.097 (0.090)	1.000	0.903
	GLS	0.124 (0.142)	0.998	0.848
	GLS0	0.093 (0.088)	1.000	0.913
200	LS	0.047 (0.041)	1.000	0.894
	GLS	0.046 (0.036)	1.000	0.896
	GLS0	0.047 (0.040)	1.000	0.912
500	LS	0.017 (0.016)	1.000	0.922
	GLS	0.016 (0.014)	1.000	0.911
	GLS0	0.015 (0.012)	1.000	0.953

References

Carroll, R. J., D. Ruppert, L. A. Stefanski, and C. M. Crainiceanu (2006). *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC.

Chen, Y., J. Goldsmith, and R. T. Ogden (2016). Variable selection in function-on-scalar regression. *Stat 5*(1), 88–101.

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

Fan, J. and R. Li (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association 96*(456), 1348–1360.

Farin, G., J. Hoschek, and M. S. Kim (2002). *Handbook of Computer Aided Geometric Design*. Elsevier.

Horn, R. A. and C. R. Johnson (1991). *Topics in Matrix Analysis*. Cambridge University Press.

Horn, R. A. and C. R. Johnson (2012). *Matrix Analysis*. Cambridge University Press.

Massart, P. (1990). The tight constant in the dvoretzky-kiefer-wolfowitz inequality. *The Annals of Probability*, 1269–1283.

Parodi, A. and M. Reimherr (2018). Simultaneous variable selection and smoothing for high-dimensional function-on-scalar regression. *Electronic Journal of Statistics 12*(2), 4602–4639.

Sofer, T., L. Dicker, and X. Lin (2014). Variable selection for high dimensional multivariate outcomes. *Statistica Sinica 24*(4), 1633.

REFERENCES

Wang, L., G. Chen, and H. Li (2007). Group scad regression analysis for microarray time course gene expression data. *Bioinformatics 23*(12), 1486–1494.

Yi, G. Y. (2017). *Statistical Analysis with Measurement Error or Misclassification: Strategy, Method and Application*. Springer.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics 38*(2), 894–942.

Zhou, S., X. Shen, and D. A. Wolfe (1998). Local asymptotics for regression splines and confidence regions. *The Annals of Statistics 26*(5), 1760–1782.