Supplementary Material for Regularized Nonlinear Regression with Dependent Errors and its Application to a Biomechanical Model

Hojun You[1], Kyubaek Yoon[3], Wei-Ying Wu[4], Jongeun Choi[3] and Chae Young Lim[2]

[1] *University of Houston* [2] *Seoul National University* [3] *Yonsei University* [4] *National Dong Hwa University*

The proof of theoretical results and additional simulation studies are discussed in this supplementary material.

S1 Proof of Theorems

For the completeness of the proofs, we state assumptions and theorems again and the proofs are followed.

Assumption 1. *(1) The nonlinear function $f \in C^2$ on the compact set $\mathcal{D} \times \Theta$ where C^2 is the set of twice continuously differentiable functions.*

(2) As $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \to 0$, $\left(\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right)^{-1} \dot{\boldsymbol{F}}(\boldsymbol{\theta})^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}) \to I_p$, elementwisely and uniformly in $\boldsymbol{\theta}$.

(3) There exist symmetric positive definite matrices $\boldsymbol{\Gamma}$ and $\boldsymbol{\Gamma}_\epsilon$ such that

$$\frac{1}{n\lambda_w} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \to \boldsymbol{\Gamma}$$
$$\frac{1}{n\lambda_\epsilon \lambda_w^2} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \to \boldsymbol{\Gamma}_\epsilon.$$

(4) $\frac{\|\boldsymbol{W}\|_1 \cdot \|\boldsymbol{W}\|_\infty}{\|\boldsymbol{W}^T \boldsymbol{\Sigma}_n \boldsymbol{W}\|_2} = o(n^{1/2}\lambda_\epsilon^{1/2})$.

(5) $O(1) \le \lambda_\epsilon \le o(n)$ and $\lambda_w \ge O(1)$.

(6) $\{\epsilon_i^2\}$ is uniformly integrable.

(7) One of the following conditions is satisfied for ϵ_i.

- *(a) $\{\epsilon_i\}$ is a ϕ-mixing.*
- *(b) $\{\epsilon_i\}$ is a ρ-mixing and $\sum_{j\in\mathcal{N}} \rho(2^j) < \infty$.*
- *(c) For $\delta > 0$, $\{\epsilon_i\}$ is a α-mixing, $\{|\epsilon_i|^{2+\delta}\}$ is uniformly integrable, and $\sum_{j\in\mathcal{N}} n^{2/\delta}\alpha(n) < \infty$.*

Assumption 2. *The first derivative of a penalty function $p_{\tau_n}(\cdot)$ denoted by $q_{\tau_n}(\cdot)$, has the following properties:*

(1) $c_n = \max_{i\in\{1,\ldots,s\}}\{|q_{\tau_n}(|\theta_{0i}|)|\} = O\left((\lambda_\epsilon/n)^{1/2}\right)$

(2) $q_{\tau_n}(\cdot)$ *is Lipschitz continuous given* τ_n

(3) $n^{1/2}\lambda_\epsilon^{-1/2}\lambda_w^{-1}\tau_n \to \infty$

(4) For any $C > 0$, $\liminf_{n\to\infty} \inf_{\theta\in\left(0, C(\lambda_\epsilon/n)^{1/2}\right)} \tau_n^{-1} q_{\tau_n}(\theta) > 0$

Lemma 1. *For any $\varepsilon > 0$ and $a_n = (\lambda_\epsilon/n)^{1/2}$, under Assumption 1-(1), (2), (3), and (5) there exists a positive constant C such that*

$$P\left(\inf_{\|\boldsymbol{v}\|=C} S_n(\boldsymbol{\theta}_0 + a_n\boldsymbol{v}) - S_n(\boldsymbol{\theta}_0) > 0\right) > 1 - \varepsilon$$

for large enough n. Therefore, with probability tending to 1, there exists a local minimizer of $S_n(\boldsymbol{\theta})$, denoted by $\hat{\boldsymbol{\theta}}^{(s)}$, in the ball centered at $\boldsymbol{\theta}_0$ with the radius $a_n\boldsymbol{v}$. Since $a_n = o(1)$ by Assumption 1-(5), we have the consistency of $\hat{\boldsymbol{\theta}}^{(s)}$.

Proof. By Taylor's theorem,

$$\begin{aligned} S_n(\boldsymbol{\theta}_0 + a_n\boldsymbol{v}) - S_n(\boldsymbol{\theta}_0) &= a_n\boldsymbol{v}^T\nabla S_n(\boldsymbol{\theta}_0) + \frac{1}{2}a_n^2\boldsymbol{v}^T\nabla^2 S_n(\boldsymbol{\theta}_0 + a_n\boldsymbol{v}t)\boldsymbol{v} \\ &:= \mathbb{A} + \mathbb{B}, \qquad \text{where } t \in (0,1). \end{aligned}$$

For the term $\mathbb{A}$, since $\boldsymbol{\Sigma}_w\boldsymbol{1} = \boldsymbol{0}$

$$\begin{aligned} \mathbb{A} &= -2a_n\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\epsilon} \\ &= -2a_n\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\eta}, \end{aligned}$$

where $\boldsymbol{\eta} = \boldsymbol{\epsilon} - \mu\boldsymbol{1}$.

$$\begin{aligned} \mathrm{var}(\mathbb{A}) &= 4a_n^2\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v} \\ &\le 4a_n^2\lambda_\epsilon\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w^2\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v} \\ &\le 4a_n^2\lambda_\epsilon\lambda_w^2\|\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v}\|^2. \end{aligned}$$

By Assumption 1-(1) and the finiteness of p, $\|\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v}\|^2 = O(n)\|\boldsymbol{v}\|^2$, which implies $\mathrm{var}(\mathbb{A}) = O(na_n^2\lambda_\epsilon\lambda_w^2)\|\boldsymbol{v}\|^2$. Since $\mathrm{E}(\boldsymbol{\eta}) = 0$,

$$\mathbb{A} = O_p(n^{1/2}a_n\lambda_\epsilon^{1/2}\lambda_w)\|\boldsymbol{v}\|. \tag{S1.1}$$

Now, since $\nabla^2 S_n(\boldsymbol{\theta}) = 2\dot{\boldsymbol{F}}(\boldsymbol{\theta})^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}) + 2\ddot{\boldsymbol{F}}(\boldsymbol{\theta})^T(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{d}(\boldsymbol{\theta},\boldsymbol{\theta}_0)) - 2\ddot{\boldsymbol{F}}(\boldsymbol{\theta})^T(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{\epsilon})$, $\mathbb{B}$ is evaluated by the following four terms.

$$\begin{aligned} \mathbb{B} &= \frac{1}{2}a_n^2\boldsymbol{v}^T\nabla^2 S_n(\boldsymbol{\theta}_n)\boldsymbol{v} \\ &= a_n^2\boldsymbol{v}^T\left(\dot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_n) - \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right)\boldsymbol{v} \\ &\quad + a_n^2\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v} + a_n^2\boldsymbol{v}^T\ddot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T\left(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{d}(\boldsymbol{\theta}_n,\boldsymbol{\theta}_0)\right)\boldsymbol{v} \\ &\quad - a_n^2\boldsymbol{v}^T\ddot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T\left(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{\epsilon}\right)\boldsymbol{v} \\ &= \mathbb{B}_1 + \mathbb{B}_2 + \mathbb{B}_3 + \mathbb{B}_4, \qquad \text{where } \boldsymbol{\theta}_n = \boldsymbol{\theta}_0 + a_n\boldsymbol{v}t. \end{aligned}$$

$$\begin{aligned}
\mathbb{B}_1 &= a_n^2 \boldsymbol{v}^T \left(\dot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_n) - \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \right) \boldsymbol{v} \\
&= n a_n^2 \lambda_w \boldsymbol{v}^T \left(\frac{1}{n\lambda_w} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_n) - \frac{1}{n\lambda_w} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \right) \boldsymbol{v} \\
&= n a_n^2 \lambda_w \boldsymbol{v}^T \left(\boldsymbol{\Gamma}(1+o(1)) \left\{ \left(\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \right)^{-1} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_n) - I \right\} \right) \boldsymbol{v} \\
&= o(n a_n^2 \lambda_w) \|\boldsymbol{v}\|^2. \qquad \text{(S1.2)}
\end{aligned}$$

The third equality holds by Assumption 1-(3) and the last equality holds by Assumption 1-(2) since $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| = a_n \|\boldsymbol{v}\| t \to 0$. By Assumption 1-(3) again,

$$\begin{aligned}
\mathbb{B}_2 &= a_n^2 \boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v} \\
&= n a_n^2 \lambda_w \boldsymbol{v}^T \left(\frac{1}{n\lambda_w} \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \right) \boldsymbol{v} \\
&= n a_n^2 \lambda_w \boldsymbol{v}^T \boldsymbol{\Gamma} \boldsymbol{v} (1+o(1)). \qquad \text{(S1.3)}
\end{aligned}$$

By Assumption 1-(5), $\mathbb{B}_2$ does not vanish to zero.

$$\begin{aligned}
\mathbb{B}_3 &= a_n^2 \boldsymbol{v}^T \ddot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T \left(I \otimes \boldsymbol{\Sigma}_w \boldsymbol{d}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) \right) \boldsymbol{v} \\
&= a_n^2 \boldsymbol{v}^T \left[\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \boldsymbol{d}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) \right]_{k,l=\{1,\dots,p\}} \boldsymbol{v}. \qquad \text{(S1.4)}
\end{aligned}$$

The term with the bracket above is a matrix where the expression in the bracket equals to the element in the k-th row and the l-th column of the matrix.

$$\begin{aligned}
\left| \boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \boldsymbol{d}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0) \right| &\le \|\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)\| \lambda_w \|\boldsymbol{d}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0)\| \\
&= O(n^{1/2}\lambda_w) \|\boldsymbol{d}(\boldsymbol{\theta}_n, \boldsymbol{\theta}_0)\| \\
&= O(n^{1/2} a_n \lambda_w) \|\boldsymbol{v}\|.
\end{aligned}$$

The first equality follows from $\|\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)\|^2 = O(n)$ by Assumption 1-(1) and the second equality holds because $\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\| = O(a_n)\|\boldsymbol{v}\|$. Therefore,

$$\mathbb{B}_3 = O(n^{1/2} a_n^3 \lambda_w) \|\boldsymbol{v}\|^3 \qquad \text{(S1.5)}$$

since p is finite. To deal with $\mathbb{B}_4$,

$$\begin{aligned}
\mathbb{B}_4 &= -a_n^2 \boldsymbol{v}^T \ddot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T \left(I \otimes \boldsymbol{\Sigma}_w \boldsymbol{\epsilon} \right) \boldsymbol{v} \\
&= -a_n^2 \boldsymbol{v}^T \ddot{\boldsymbol{F}}(\boldsymbol{\theta}_n)^T \left(I \otimes \boldsymbol{\Sigma}_w \boldsymbol{\eta} \right) \boldsymbol{v} \\
&= -a_n^2 \boldsymbol{v}^T \left[\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n) \boldsymbol{\Sigma}_w \boldsymbol{\eta} \right]_{k,l=\{1,\dots,p\}} \boldsymbol{v}.
\end{aligned}$$

We first show $|\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \boldsymbol{\eta}| = O_p(n^{1/2}\lambda_\epsilon^{1/2}\lambda_w)$ by

$$\begin{aligned}\mathrm{var}(\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \boldsymbol{\eta}) &= \boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \boldsymbol{f}_{kl}(\boldsymbol{\theta}_n) \\ &\leq \lambda_\epsilon \lambda_w^2 \|\boldsymbol{f}_{kl}(\boldsymbol{\theta}_n)\|^2 \\ &= O(n\lambda_\epsilon\lambda_w^2).\end{aligned}$$

Thus,

$$\mathbb{B}_4 = O_p(n^{1/2}a_n^2\lambda_\epsilon^{1/2}\lambda_w)\|\boldsymbol{v}\|^2. \tag{S1.6}$$

By equations (S1.2), (S1.3), (S1.5), and (S1.6),

$$\begin{aligned}\mathbb{B} &= o(na_n^2\lambda_w)\|\boldsymbol{v}\|^2 + na_n^2\lambda_w \boldsymbol{v}^T\boldsymbol{\Gamma}\boldsymbol{v} + O(n^{1/2}a_n^3\lambda_w)\|\boldsymbol{v}\|^3 + O_p(n^{1/2}a_n^2\lambda_\epsilon^{1/2}\lambda_w)\|\boldsymbol{v}\|^2 \\ &= na_n^2\lambda_w\boldsymbol{v}\boldsymbol{\Gamma}\boldsymbol{v} + o_p(na_n^2\lambda_w)\|\boldsymbol{v}\|^2. \qquad (\text{S1.7})\end{aligned}$$

The second equality holds since $\lambda_\epsilon \leq o(n)$ by Assumption 1-(5). Finally, through (S1.1) and (S1.7),

$$\begin{aligned}S_n(\boldsymbol{\theta}_0 + a_n\boldsymbol{v}) - S_n(\boldsymbol{\theta}_0) &= O_p(n^{1/2}a_n\lambda_\epsilon^{1/2}\lambda_w)\|\boldsymbol{v}\| + na_n^2\lambda_w\boldsymbol{v}^T\boldsymbol{\Gamma}\boldsymbol{v}(1+o_p(1)) \\ &= O_p(\lambda_\epsilon\lambda_w)\|\boldsymbol{v}\| + \lambda_\epsilon\lambda_w\boldsymbol{v}^T\boldsymbol{\Gamma}\boldsymbol{v}(1+o_p(1)). \qquad (\text{S1.8})\end{aligned}$$

Therefore, with large enough $\|\boldsymbol{v}\|$, The desired result follows. □

Theorem 1. *For any $\varepsilon > 0$ and $b_n = (\lambda_\epsilon/n)^{1/2} + c_n$, under assumptions in Lemma 1 and 2-(1),(2), there exists a positive constant C such that*

$$P\left(\inf_{\|\boldsymbol{v}\|=C} Q_n(\boldsymbol{\theta}_0 + b_n\boldsymbol{v}) - Q_n(\boldsymbol{\theta}_0) > 0\right) > 1 - \varepsilon$$

for large enough n. Therefore, with probability tending to 1, there exists a local minimizer ($\hat{\boldsymbol{\theta}}$) of $Q_n(\boldsymbol{\theta})$ in the ball centered at $\boldsymbol{\theta}_0$ with the radius $b_n\boldsymbol{v}$. By Assumptions 1-(5) and 2-(1), $b_n = o(1)$, which leads to the consistency of $\hat{\boldsymbol{\theta}}$.

Proof.

$$\begin{aligned}
&Q_n(\boldsymbol{\theta}_0 + b_n\boldsymbol{v}) - Q_n(\boldsymbol{\theta}_0)\\
&\quad= \mathbb{A}' + \mathbb{B}' + n\left(\sum_{i=1}^{p} p_{\tau_n}(|\theta_{0i} + b_n v_i|) - p_{\tau_n}(|\theta_{0i}|)\right)\\
&\quad\geq \mathbb{A}' + \mathbb{B}' + n\left(\sum_{i=1}^{s} p_{\tau_n}(|\theta_{0i} + b_n v_i|) - p_{\tau_n}(|\theta_{0i}|)\right)\\
&\quad= \mathbb{A}' + \mathbb{B}' + n\sum_{i=1}^{s} q_{\tau_n}(|\theta_{0i}^*|)sgn(\theta_{0i}^*)b_n v_i, \qquad \text{where } \theta_{0i}^* \text{ lies on a line segment } (\theta_{0i}, \theta_{0i} + b_n v_i)\\
&\quad= \mathbb{A}' + \mathbb{B}' + n\sum_{i=1}^{s} \left(q_{\tau_n}(|\theta_{0i}^*|) - q_{\tau_n}(|\theta_{0i}|) + q_{\tau_n}(|\theta_{0i}|)\right) sgn(\theta_{0i}^*)b_n v_i\\
&\quad= \mathbb{A}' + \mathbb{B}' + \mathbb{C} + \mathbb{D}, \qquad\qquad \text{(S1.9)}
\end{aligned}$$

where $\mathbb{A}'$ and $\mathbb{B}'$ are defined similarly to $\mathbb{A}$ and $\mathbb{B}$ in the proof of Lemma 1 with replacement of a_n to b_n. $\mathbb{C} = n\sum_i (q_{\tau_n}(|\theta_{0i}^*|) - q_{\tau_n}(|\theta_{0i}|))sgn(\theta_{0i^*})b_n v_i$ and $\mathbb{D} = n\sum_i q_{\tau_n}(|\theta_{0i}|)sgn(\theta_{0i^*})b_n v_i$. Referring to equation (S1.8)

$$\mathbb{A}' + \mathbb{B}' = O_p(n^{1/2}b_n\lambda_\epsilon^{1/2}\lambda_w)\|\boldsymbol{v}\| + nb_n^2\lambda_w\boldsymbol{v}^T\boldsymbol{\Gamma}\boldsymbol{v}.$$

It is enough to show that $\mathbb{C} = O(nb_n^2)\|\boldsymbol{v}\|$ and $\mathbb{D} = O(nb_n^2)\|\boldsymbol{v}\|$ since $\lambda_w \geq O(1)$. By Assumptions 2-(1) and (2),

$$|\mathbb{C}| \leq nb_n\sum_i |q_{\tau_n}(|\theta_{0i}^*|) - q_{\tau_n}(|\theta_{oi}|)||v_i| = O(nb_n^2)\|\boldsymbol{v}\|,$$

$$|\mathbb{D}| \leq nb_n\max_{1\leq i\leq s}|q_{\tau_n}(|\theta_{0i}|)|\sum_i v_i = O(nb_nc_n)\|\boldsymbol{v}\| \leq O(nb_n^2)\|\boldsymbol{v}\|.$$

Therefore, $nb_n^2\lambda_w\boldsymbol{v}^T\boldsymbol{\Gamma}\boldsymbol{v}$ dominates equation (S1.9) with large $\|\boldsymbol{v}\|$, which leads to the desired result.

□

Lemma 2. *[Theorem 2.2 from Peligrad and Utev (1997)] Let $\boldsymbol{\eta} = \{\eta_1, \ldots, \eta_n\}$ be a stochastic sequence and $\boldsymbol{h}_n = \{h_{n,1}, \ldots, h_{n,n}\}$ be a triangular weight vector, then*

$$\boldsymbol{h}_n^T \boldsymbol{\eta} \xrightarrow{d} N(0,1)$$

under the following conditions.

1. *$\boldsymbol{\eta}$ is a centered stochastic sequence*
2. $\sup_n \|\boldsymbol{h}_n\|_2^2 < \infty$ *and* $\|\boldsymbol{h}_n\|_\infty \to 0$ *as* $n \to \infty$.
3. $\{\eta_i^2\}$ *is uniformly integrable and* $\mathrm{var}(\boldsymbol{h}_n^T \boldsymbol{\eta}) = 1$
4. *For $\boldsymbol{\eta}$, one of the three following mixing conditions must be satisfied.*
   - *$\{\eta_i\}$ is a ϕ-mixing.*
   - *$\{\eta_i\}$ is a ρ-mixing and $\sum_{j \in \mathcal{N}} \rho(2^j) < \infty$.*
   - *$\{\eta_i\}$ is a α-mixing, for $\delta > 0$, $\{|\eta_i|^{2+\delta}\}$ is uniformly integrable, and $\sum_{j \in \mathcal{N}} n^{2/\delta} \alpha(n) < \infty$.*

Proof. We refer to Peligrad and Utev (1997) for the proof. □

Lemma 3 (Asymptotic normality)**.** *Under Assumption 1,*

$$\left(\frac{n}{\lambda_\epsilon}\right)^{1/2} \left(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\left(0, \boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\epsilon\boldsymbol{\Gamma}^{-1}\right),$$

where $\hat{\boldsymbol{\theta}}^{(s)}$ is a consistent estimator introduced in Lemma 1 with $S_n(\boldsymbol{\theta})$.

Proof. By the mean-value theorem of a vector-valued function (Feng et al., 2013), with $\zeta_n = (n\lambda_\epsilon\lambda_w^2)^{-1/2}$,

$$\begin{aligned}
&\zeta_n \nabla S_n(\boldsymbol{\theta}_0) \\
&\quad = \zeta_n \left(\nabla S_n(\hat{\boldsymbol{\theta}}^{(s)}) + \left(\int_0^1 \nabla^2 S_n \left(\hat{\boldsymbol{\theta}}^{(s)} + \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}^{(s)} \right) t \right) dt \right)^T \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}^{(s)} \right) \right) \\
&\quad = \zeta_n \left(\int_0^1 \nabla^2 S_n \left(\hat{\boldsymbol{\theta}}^{(s)} + \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}^{(s)} \right) t \right) dt \right)^T \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}^{(s)} \right)
\end{aligned}$$

since $\nabla S_n(\hat{\boldsymbol{\theta}}^{(s)}) = 0$. We show $\zeta_n \nabla S_n(\boldsymbol{\theta}_0)$ follows a normal distribution asymptotically and

$$\frac{1}{n\lambda_w} \int_0^1 \nabla^2 S_n \left(\hat{\boldsymbol{\theta}}^{(s)} + \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}^{(s)} \right) t \right) dt \xrightarrow{p} 2\boldsymbol{\Gamma}. \tag{S1.10}$$

For an arbitrary vector $\boldsymbol{v} \in \mathcal{R}^p$,

$$\begin{aligned}\zeta_n \boldsymbol{v}^T \nabla S_n(\boldsymbol{\theta}_0) &= -2\zeta_n \boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\epsilon} \\ &= -2\zeta_n \boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\eta},\end{aligned}$$

where $\boldsymbol{\eta} = \boldsymbol{\epsilon} - \mu \mathbf{1}$. This converges to $N(0, 4\boldsymbol{v}^T \boldsymbol{\Gamma}_\epsilon \boldsymbol{v})$ since $\boldsymbol{h}_n^T \boldsymbol{\eta} \xrightarrow{d} N(0,1)$ with

$$\boldsymbol{h}_n = \left(\boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right)^{-1/2} \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v},$$

by Lemma 2. We first show that $\boldsymbol{h}_n^T \boldsymbol{\eta} \xrightarrow{d} N(0,1)$ by proving the conditions given in Lemma 2 are fulfilled and then we handle the remainder term. The first condition in Lemma 2 is trivial and the fourth condition is satisfied by Assumption 1-(7). The second condition is satisfied since

$$\begin{aligned}\|\boldsymbol{h}_n\|^2 &= \left(\boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right)^{-1} \boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w^2 \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v} \\ &= \lambda_\epsilon^{-1} \left(\frac{1}{n\lambda_\epsilon \lambda_w^2} \boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right)^{-1} \frac{1}{n\lambda_w^2} \boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w^2 \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v} \\ &\leq \lambda_\epsilon^{-1} \left(\boldsymbol{v}^T \boldsymbol{\Gamma}_\epsilon \boldsymbol{v}\right)^{-1} O(1) \|\boldsymbol{v}\|^2 \qquad \text{(Assumptions 1-(1) and (3))} \\ &= O(\lambda_\epsilon^{-1}) \\ &\leq O(1), \qquad \text{(Assumption 1-(5))} \\ \|\boldsymbol{h}_n\|_\infty &= \left(\boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right)^{-1/2} \left\|\boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right\|_\infty \\ &\leq \zeta_n \left(\boldsymbol{v}^T \boldsymbol{\Gamma}_\epsilon \boldsymbol{v}\right)^{-1/2} \|\boldsymbol{\Sigma}_w\|_\infty \left\|\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right\|_\infty \\ &\leq O(\zeta_n) \|\boldsymbol{W}^T\|_\infty \|\boldsymbol{\Sigma}_n\|_\infty \|\boldsymbol{W}\|_\infty \left\|\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right\|_\infty \\ &\leq O\left((n\lambda_\epsilon)^{-1/2}\right) \lambda_w^{-1} \|\boldsymbol{W}\|_1 \|\boldsymbol{W}\|_\infty \qquad (\|\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v}\|_\infty = O(1) \text{ and } \|\boldsymbol{\Sigma}_n\|_\infty \leq 2) \\ &= o(1). \qquad \text{(Assumption 1-(4))}\end{aligned}$$

$\{\eta_i^2\}$ is uniformly integrable by Assumption 1-(6). The remainder of the proof is to show $\text{var}(\boldsymbol{h}_n^T \boldsymbol{\eta}) = 1$.

$$\text{var}(\boldsymbol{h}_n^T \boldsymbol{\eta}) = \left(\boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right)^{-1} \left(\boldsymbol{v}^T \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T \boldsymbol{\Sigma}_w \boldsymbol{\Sigma}_\epsilon \boldsymbol{\Sigma}_w \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) \boldsymbol{v}\right) = 1.$$

Therefore, for arbitrary $\boldsymbol{v}$,

$$\begin{aligned}-2\zeta_n\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\eta} &= -2(n\lambda_\epsilon\lambda_w^2)^{-1/2}\left(\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v}\right)^{1/2}\boldsymbol{h}_n^T\boldsymbol{\eta}\\ &= -2\left(\frac{1}{n\lambda_\epsilon\lambda_w^2}\boldsymbol{v}^T\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\boldsymbol{v}\right)^{1/2}\boldsymbol{h}_n^T\boldsymbol{\eta}\\ &\xrightarrow{d} N(0,4\boldsymbol{v}^T\boldsymbol{\Gamma}_\epsilon\boldsymbol{v}).\end{aligned}$$

The asymptotic variance in the limiting distribution comes from Assumption 1-(3). By the Cramer-Wold device,

$$\zeta_n\nabla S_n(\boldsymbol{\theta}_0) = -2\zeta_n\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\boldsymbol{\eta} \xrightarrow{d} N(0,4\boldsymbol{\Gamma}_\epsilon). \tag{S1.11}$$

For equation (S1.10), we need to show

$$\frac{1}{n\lambda_w}\nabla^2 S_n(\boldsymbol{\theta}_0) \xrightarrow{p} 2\boldsymbol{\Gamma}, \tag{S1.12}$$

$$\frac{1}{n\lambda_w}\left(\int_0^1 \nabla^2 S_n\left(\hat{\boldsymbol{\theta}}^{(s)} + \left(\boldsymbol{\theta}_0 - \hat{\boldsymbol{\theta}}^{(s)}\right)t\right)dt - \nabla^2 S_n(\boldsymbol{\theta}_0)\right) \xrightarrow{p} 0. \tag{S1.13}$$

For (S1.12),

$$\frac{1}{n\lambda_w}\nabla^2 S_n(\boldsymbol{\theta}_0) = \frac{2}{n\lambda_w}\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0) - \frac{2}{n\lambda_w}\ddot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{\epsilon}).$$

Similar to equation (S1.3) and (S1.6), the first term converges to $2\boldsymbol{\Gamma}$ and the second term is $O_p((\lambda_\epsilon/n)^{1/2})$, which vanishes to $o_p(1)$ by Assumption 1-(5). (S1.13) is satisfied if

$$\max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Ca_n}\frac{1}{n\lambda_w}\left\|\nabla^2 S_n(\boldsymbol{\theta}) - \nabla^2 S_n(\boldsymbol{\theta}_0)\right\| \xrightarrow{p} 0. \tag{S1.14}$$

(S1.14) can be decomposed as three following terms.

$$\begin{aligned}&\max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Ca_n}\frac{1}{n\lambda_w}\left\|\nabla^2 S_n(\boldsymbol{\theta}) - \nabla^2 S_n(\boldsymbol{\theta}_0)\right\|\\ &\qquad\le \max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Ca_n}\frac{2}{n\lambda_w}\left\|\dot{\boldsymbol{F}}(\boldsymbol{\theta})^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}) - \dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)^T\boldsymbol{\Sigma}_w\dot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right\|\\ &\qquad\quad+ \max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Ca_n}\frac{2}{n\lambda_w}\left\|\ddot{\boldsymbol{F}}(\boldsymbol{\theta})^T\left(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{d}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\right)\right\|\\ &\qquad\quad+ \max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\le Ca_n}\frac{2}{n\lambda_w}\left\|\left(\ddot{\boldsymbol{F}}(\boldsymbol{\theta}) - \ddot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right)^T(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{\eta})\right\|.\end{aligned}$$

The first part converges to 0 by a similar procedure to equation (S1.2) and Assumption 1-(2). The second part converges to 0 because

$$\begin{aligned}\frac{2}{n\lambda_w}\left\|\ddot{\boldsymbol{F}}(\boldsymbol{\theta})^T\left(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{d}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\right)\right\| &= \frac{p}{n\lambda_w}\left\|\left[\boldsymbol{f}_{kl}(\boldsymbol{\theta})^T\boldsymbol{\Sigma}_w\boldsymbol{d}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\right]_{k,l=\{1,\ldots,p\}}\right\| \\ &\leq \frac{2p}{n\lambda_w}\cdot\max_{k,l}\left|\boldsymbol{f}_{kl}(\boldsymbol{\theta})^T\boldsymbol{\Sigma}_w\boldsymbol{d}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\right| \\ &\leq \frac{2p}{n}\max_{k,l}\|\boldsymbol{f}_{kl}(\boldsymbol{\theta})\|\cdot\|\boldsymbol{d}(\boldsymbol{\theta},\boldsymbol{\theta}_0)\| \\ &\leq \frac{2C_p}{n^{1/2}}\cdot\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|, \qquad \text{where } C_p \text{ is independent with } \boldsymbol{\theta} \\ &= O\left(n^{-1/2}a_n\right)=o(1).\end{aligned}$$

The results above hold if $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|=O(a_n)$, which is still true for the maximum. For the last term,

$$\frac{2}{n\lambda_w}\left\|\left(\ddot{\boldsymbol{F}}(\boldsymbol{\theta})-\ddot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right)^T\left(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{\epsilon}\right)\right\| = \frac{2}{n\lambda_w}\left\|\left[\left(\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0)\right)^T\boldsymbol{\Sigma}_w\boldsymbol{\eta}\right]_{k,l=\{1,\ldots,p\}}\right\|.$$

We evaluate $\mathrm{var}((\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0))^T\boldsymbol{\Sigma}_w\boldsymbol{\eta})$ as follows.

$$\begin{aligned}\mathrm{var}\left((\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0))^T\boldsymbol{\Sigma}_w\boldsymbol{\eta}\right) &= (\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0))^T\boldsymbol{\Sigma}_w\boldsymbol{\Sigma}_\epsilon\boldsymbol{\Sigma}_w(\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0)) \\ &\leq \lambda_\epsilon\lambda_w^2\|\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0)\|^2 \\ &\leq \lambda_\epsilon\lambda_w^2 O\left(\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|^2\right) \qquad \text{(Assumption 1-(1))} \\ &= O\left(a_n^2\lambda_\epsilon\lambda_w^2\right).\end{aligned}$$

Therefore, $|\left(\boldsymbol{f}_{kl}(\boldsymbol{\theta})-\boldsymbol{f}_{kl}(\boldsymbol{\theta}_0)\right)^T\boldsymbol{\Sigma}_w\boldsymbol{\eta}|=O_p(a_n\lambda_\epsilon^{1/2}\lambda_w)$ and

$$\max_{\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|\leq Ca_n}\frac{2}{n\lambda_w}\left\|\left(\ddot{\boldsymbol{F}}(\boldsymbol{\theta})-\ddot{\boldsymbol{F}}(\boldsymbol{\theta}_0)\right)^T\left(I\otimes\boldsymbol{\Sigma}_w\boldsymbol{\eta}\right)\right\| = O_p\left(\frac{a_n\lambda_\epsilon^{1/2}}{n}\right)=o_p(1).$$

Thus, we prove equation (S1.13). Combining results of equations (S1.12) and (S1.13), we have equation (S1.10). Recall

$$\zeta_n\nabla S_n(\boldsymbol{\theta}_0)=\zeta_n\left(\int_0^1\nabla^2 S_n\left(\hat{\boldsymbol{\theta}}^{(s)}+\left(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}^{(s)}\right)t\right)dt\right)^T\left(\boldsymbol{\theta}_0-\hat{\boldsymbol{\theta}}^{(s)}\right),$$

with $\zeta_n=(n\lambda_\epsilon\lambda_w^2)^{-1/2}$. By equations (S1.10) and (S1.11) and Slutsky's theorem,

$$2n\lambda_w\zeta_n\boldsymbol{\Gamma}\left(\hat{\boldsymbol{\theta}}^{(s)}-\boldsymbol{\theta}_0\right)\xrightarrow{d}N(0,4\boldsymbol{\Gamma}_\epsilon).$$

Since $n\lambda_w\zeta_n = (n/\lambda_\epsilon)^{1/2}$,

$$\left(\frac{n}{\lambda_\epsilon}\right)^{1/2}\left(\hat{\boldsymbol{\theta}}^{(s)} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N(0, \boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\epsilon\boldsymbol{\Gamma}^{-1}).$$

□

Theorem 2 (Oracle property)**.** *With $\hat{\boldsymbol{\theta}}$, a consistent estimator introduced in Theorem 1 using $Q_n(\boldsymbol{\theta})$, if Assumptions 1 and 2 are satisfied,*

(i) $P\left(\hat{\theta}_i = 0\right) \to 1$, *for* $i \in \{s+1, \ldots, p\}$.

(ii) Also,

$$\left(\frac{n}{\lambda_\epsilon}\right)^{1/2} \left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01} + \left((2\lambda_w \boldsymbol{\Gamma})^{-1}\right)_{11} \boldsymbol{\beta}_{n,s}\right) \xrightarrow{d} N\left(0, \left(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\epsilon \boldsymbol{\Gamma}^{-1}\right)_{11}\right),$$

where $\hat{\boldsymbol{\theta}}_1 = (\hat{\theta}_1, \ldots, \hat{\theta}_s)^T$, $\boldsymbol{\theta}_{01} = (\theta_{01}, \ldots, \theta_{0s})^T$, $\boldsymbol{\beta}_{n,s} = (q_{\tau_n}(|\theta_{01}|)sgn(\theta_{01}), \ldots, q_{\tau_n}(|\theta_{0s}|)sgn(\theta_{0s}))^T$ *and* $\boldsymbol{A}_{11}$ *is the* $s \times s$ *upper-left matrix of* $\boldsymbol{A}$*.*

Proof. Proof of (i)
It is equivalent to show that $P\left(\hat{\theta}_i \neq 0\right) \to 0$ as $n \to \infty$ for $i \in \{s+1, \ldots, p\}$.

$$\begin{aligned} P\left(\hat{\theta}_i \neq 0\right) &= P\left(\hat{\theta}_i \neq 0,\ |\hat{\theta}_i| > Cb_n\right) + P\left(\hat{\theta}_i \neq 0,\ |\hat{\theta}_i| \leq Cb_n\right) \\ &:= P(\mathbb{E}) + P(\mathbb{F}). \end{aligned}$$

For any $\varepsilon > 0$ and large enough n, $P(\mathbb{E}) < \varepsilon/2$ by Theorem1. Now we show $P(\mathbb{F}) < \varepsilon/2$. By the vector-valued mean value theorem (Feng et al., 2013),

$$\zeta_n \nabla S_n(\boldsymbol{\theta}_0) = \zeta_n \left(\nabla S_n(\boldsymbol{\theta}) + \left(\int_0^1 \nabla^2 S_n\left(\boldsymbol{\theta} + (\boldsymbol{\theta}_0 - \boldsymbol{\theta})\, t\right) dt\right)^T (\boldsymbol{\theta}_0 - \boldsymbol{\theta})\right),$$

where $\zeta_n = (n\lambda_\epsilon \lambda_w^2)^{-1/2}$. From the proof of Lemma 3, $\zeta_n \nabla S_n(\boldsymbol{\theta}_0) = O_p(1)$ and with $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}\| = O_p(b_n)$, the similar results of (S1.10) inform that

$$n\zeta_n \lambda_w \left(\frac{1}{n\lambda_w} \int_0^1 \nabla^2 S_n\left(\boldsymbol{\theta} + (\boldsymbol{\theta}_0 - \boldsymbol{\theta})\, t\right) dt\right)^T (\boldsymbol{\theta}_0 - \boldsymbol{\theta}) = O_p(1).$$

This leads to $\zeta_n \nabla S_n(\boldsymbol{\theta}) = O_p(1)$ for $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O_p(b_n)$. Since $\hat{\boldsymbol{\theta}}$ is the local minimizer of $Q_n(\boldsymbol{\theta})$ with $\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| = O_p(b_n)$, we attain, for $i \in \{s+1, \ldots, p\}$,

$$n\zeta_n q_{\tau_n}(|\hat{\theta}_i|) = O_p(1)$$

from

$$\zeta_n \left.\frac{\partial Q_n(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} = \zeta_n \left.\frac{\partial S_n(\boldsymbol{\theta})}{\partial \theta_i}\right|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} + n\zeta_n q_{\tau_n}(|\hat{\theta}_i|) sgn(\hat{\theta}_i).$$

Therefore, there exists a $M > 0$ such that $P\left(\left|n\zeta_n q_{\tau_n}(|\hat{\theta}_i|)\right| > M\right) < \varepsilon/2$ for large enough n, which implies

$$P\left(\hat{\theta}_i \neq 0,\ |\hat{\theta}_i| \leq Cb_n,\ n\zeta_n q_{\tau_n}(|\hat{\theta}_i|) > M\right) < \frac{\varepsilon}{2}.$$

By Assumptions 2-(3) and (4),

$$P\left(\hat{\theta}_i \neq 0,\ |\hat{\theta}_i| \leq Cb_n,\ n\zeta_n q_{\tau_n}(|\hat{\theta}_i|) > M\right) = P\left(\hat{\theta}_i \neq 0,\ |\hat{\theta}_i| \leq Cb_n\right)$$

for large enough n. At last, we have $P(\mathbb{F}) < \varepsilon/2$. Together with $P(\mathbb{E}) < \varepsilon/2$, this implies $P(\hat{\theta}_i \neq 0) \to 0$.

Proof of (ii)

Note that

$$\zeta_n \nabla Q_n(\hat{\boldsymbol{\theta}}) = \zeta_n \nabla S_n(\hat{\boldsymbol{\theta}}) + n\zeta_n \boldsymbol{q}_{\tau_n}(|\hat{\boldsymbol{\theta}}|)sgn(\hat{\boldsymbol{\theta}}),$$

where $\boldsymbol{q}_{\tau_n}(|\hat{\boldsymbol{\theta}}|)sgn(\hat{\boldsymbol{\theta}}) = (q_{\tau_n}(|\hat{\theta}_1|)sgn(\hat{\theta}_1), \ldots, q_{\tau_n}(|\hat{\theta}_p|)sgn(\hat{\theta}_p))^T$. Since $\hat{\boldsymbol{\theta}}$ is a local minimizer of $Q_n(\boldsymbol{\theta})$, $\nabla Q_n(\hat{\boldsymbol{\theta}}) = 0$, which implies

$$\begin{aligned} -\zeta_n \nabla S_n(\boldsymbol{\theta}_0) = &\left(\frac{1}{n\lambda_w}\int_0^1 \nabla^2 S_n\left(\boldsymbol{\theta}_0 + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)t\right) dt\right)^T \left(n\zeta_n\lambda_w(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)\right) \\ &+ n\zeta_n \boldsymbol{q}_{\tau_n}(|\hat{\boldsymbol{\theta}}|)sgn(\hat{\boldsymbol{\theta}}). \end{aligned}$$

The left-hand side converges to $N(0, 4\boldsymbol{\Gamma}_\epsilon)$ and, similarly to (S1.10),

$$\frac{1}{n\lambda_w}\int_0^1 \nabla^2 S_n\left(\boldsymbol{\theta}_0 + (\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)t\right) dt \xrightarrow{p} 2\boldsymbol{\Gamma}.$$

Thus, by the Slutsky's theorem,

$$\left(\frac{n}{\lambda_\epsilon}\right)^{1/2}\left(2\boldsymbol{\Gamma}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) + \lambda_w^{-1}\boldsymbol{q}_{\tau_n}(|\hat{\boldsymbol{\theta}}|)sgn(\hat{\boldsymbol{\theta}})\right) \xrightarrow{d} N\left(0, 4\boldsymbol{\Gamma}_\epsilon\right).$$

Slicing the first s components of $\hat{\boldsymbol{\theta}}$, we obtain

$$\left(\frac{n}{\lambda_\epsilon}\right)^{1/2}\left(\hat{\boldsymbol{\theta}}_1 - \boldsymbol{\theta}_{01} + \left((2\lambda_w\boldsymbol{\Gamma})^{-1}\right)_{11}\boldsymbol{\beta}_{n,s}\right) \xrightarrow{d} N\left(0, \left(\boldsymbol{\Gamma}^{-1}\boldsymbol{\Gamma}_\epsilon\boldsymbol{\Gamma}^{-1}\right)_{11}\right).$$

□

Table S1: Estimation results with LASSO for the equation (S2.15) when the error process is AR(1) with $\rho = 0.5$. Mean squared error values are presented with standard deviation in the parenthesis. The rows without ρ or (ρ, ϕ) indicate that no weight matrix is used.

AR(1) with $\rho = 0.5$				
(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
$(0.1, 0.5)$	PMWLS	11.09 (0.51)	6.86 (0.54)	4.09 (0.47)
	PMWLS ($\rho = 0.5$)	11.22 (0.48)	6.73 (0.53)	3.65 (0.41)
	PMWLS ($\rho = 0.9$)	11.67 (0.45)	6.77 (0.53)	3.39 (0.41)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	11.96 (0.44)	7.35 (0.54)	3.45 (0.40)
	PWLS	12.11 (0.46)	7.09 (0.64)	2.51 (0.47)
	PWLS ($\rho = 0.5$)	12.38 (0.45)	6.70 (0.59)	2.81 (0.43)
	PWLS ($\rho = 0.9$)	11.68 (0.51)	6.72 (0.54)	3.47 (0.39)
	PWLS ($\rho = 0.8, \phi = 0.4$)	12.23 (0.40)	6.93 (0.53)	3.76 (0.40)
$(0.5, 0.5)$	PMWLS	10.26 (0.55)	8.02 (0.54)	4.17 (0.46)
	PMWLS ($\rho = 0.5$)	11.04 (0.52)	7.56 (0.51)	3.79 (0.44)
	PMWLS ($\rho = 0.9$)	11.22 (0.55)	7.34 (0.51)	3.77 (0.45)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	11.60 (0.48)	8.23 (0.50)	3.86 (0.44)
	PWLS	9.41 (0.58)	6.35 (0.64)	2.07 (0.43)
	PWLS ($\rho = 0.5$)	10.92 (0.57)	7.87 (0.61)	2.19 (0.39)
	PWLS ($\rho = 0.9$)	11.36 (0.51)	8.06 (0.49)	4.34 (0.40)
	PWLS ($\rho = 0.8, \phi = 0.4$)	12.22 (0.45)	8.32 (0.49)	4.56 (0.40)

$\ast$ The actual MSE values are 0.01× the reported values.

S2 More simulation results

Here, we provide tables from the simulation study conducted in Section 3 of the main article.

Tables S1- S3 report the values of mean squared error (MSE) with standard deviation of squared error (SD) in parenthesis for the estimates from 100 repetitions of data generated from the model

$$y_t = \frac{1}{1 + \exp(-\boldsymbol{x}_t^T \boldsymbol{\theta}_0)} + \epsilon_t, \tag{S2.15}$$

where $\boldsymbol{\theta}_0 = (\theta_{01}, \theta_{02}, \ldots, \theta_{0,20})^T$ with $\theta_{01} = 1, \theta_{02} = 1.2, \theta_{03} = 0.6$, and the others being zero. The first component of the covariate $\boldsymbol{x}$ comes from $U[-1, 1]$, a uniform distribution on $[-1, 1]$, and the other components of $\boldsymbol{x}$ are simulated from a joint normal distribution with the zero mean, the variance being 0.6 and pairwise covariance being 0.1. For ϵ_t, the AR(1) and ARMA(1,1) with the non-zero mean are considered since these processes not only represent typical time series processes but also possess the strong mixing property. The choices of the AR(1) coefficient, ρ, are 0.5 and 0.9 and for the ARMA process, the parameters for the AR and MA parts are fixed as 0.8 (ρ) and 0.4 (ϕ), respectively. For the non-zero mean, μ, the choices are 0.1 and 0.5. For the standard deviation, $\sigma = 0.5$. The formulae to calculate MSE and SD are given in Section 3 of the main article.

Table S2: Estimation results with LASSO for the equation (S2.15) when the error process is AR(1) with $\rho = 0.9$. The other configurations are identical to Table S1.

AR(1) with $\rho = 0.9$				
(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
$(0.1, 0.5)$	PMWLS	8.50 (0.60)	6.71 (0.51)	3.64 (0.45)
	PMWLS ($\rho = 0.5$)	6.61 (0.56)	4.66 (0.46)	2.03 (0.27)
	PMWLS ($\rho = 0.9$)	7.41 (0.55)	4.79 (0.46)	1.75 (0.26)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	7.19 (0.56)	4.98 (0.51)	1.87 (0.28)
	PWLS	8.01 (0.62)	6.34 (0.61)	2.72 (0.47)
	PWLS ($\rho = 0.5$)	7.40 (0.61)	5.28 (0.54)	1.52 (0.31)
	PWLS ($\rho = 0.9$)	6.80 (0.53)	4.67 (0.45)	1.64 (0.24)
	PWLS ($\rho = 0.8, \phi = 0.4$)	7.13 (0.53)	4.90 (0.49)	1.90 (0.26)
$(0.5, 0.5)$	PMWLS	7.63 (0.58)	6.79 (0.58)	3.64 (0.46)
	PMWLS ($\rho = 0.5$)	6.60 (0.57)	3.93 (0.45)	2.10 (0.32)
	PMWLS ($\rho = 0.9$)	7.01 (0.57)	3.83 (0.44)	1.90 (0.28)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	7.18 (0.55)	3.95 (0.46)	2.03 (0.29)
	PWLS	7.23 (0.63)	5.41 (0.59)	2.07 (0.43)
	PWLS ($\rho = 0.5$)	6.66 (0.60)	4.58 (0.59)	1.36 (0.32)
	PWLS ($\rho = 0.9$)	7.34 (0.54)	4.28 (0.42)	2.26 (0.26)
	PWLS ($\rho = 0.8, \phi = 0.4$)	7.53 (0.54)	5.07 (0.44)	2.66 (0.29)

$*$ The actual MSE values are 0.01× the reported values.

Table S3: Estimation results with LASSO for the equation (S2.15) when the error process is ARMA(1,1) with $(\rho, \phi) = (0.8, 0.4)$. The other configurations are identical to Table S1.

ARMA(1,1) with $\rho = 0.8, \phi = 0.4$				
(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
$(0.1, 0.5)$	PMWLS	6.55 (0.54)	4.11 (0.48)	2.47 (0.40)
	PMWLS ($\rho = 0.5$)	5.13 (0.52)	2.64 (0.34)	1.52 (0.26)
	PMWLS ($\rho = 0.9$)	5.12 (0.51)	2.61 (0.34)	1.42 (0.24)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	5.45 (0.51)	2.58 (0.35)	1.36 (0.24)
	PWLS	6.20 (0.63)	3.05 (0.50)	1.62 (0.38)
	PWLS ($\rho = 0.5$)	5.52 (0.57)	2.12 (0.37)	0.96 (0.23)
	PWLS ($\rho = 0.9$)	5.14 (0.52)	2.68 (0.32)	1.37 (0.20)
	PWLS ($\rho = 0.8, \phi = 0.4$)	5.35 (0.54)	2.73 (0.36)	1.61 (0.24)
$(0.5, 0.5)$	PMWLS	7.37 (0.59)	3.95 (0.50)	2.66 (0.39)
	PMWLS ($\rho = 0.5$)	5.35 (0.51)	2.57 (0.36)	1.29 (0.24)
	PMWLS ($\rho = 0.9$)	5.86 (0.50)	2.49 (0.37)	1.40 (0.23)
	PMWLS ($\rho = 0.8, \phi = 0.4$)	6.04 (0.52)	2.47 (0.36)	1.74 (0.27)
	PWLS	6.13 (0.61)	1.67 (0.41)	1.35 (0.34)
	PWLS ($\rho = 0.5$)	5.73 (0.57)	1.87 (0.38)	0.80 (0.22)
	PWLS ($\rho = 0.9$)	6.41 (0.49)	2.96 (0.35)	1.94 (0.24)
	PWLS ($\rho = 0.8, \phi = 0.4$)	6.81 (0.50)	3.18 (0.33)	2.03 (0.27)

$*$ The actual MSE values are 0.01× the reported values.

Table S4: Selection results with LASSO for the equation (S2.15) when the error process is AR(1) with $\rho = 0.5$.

AR(1) with $\rho = 0.5$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
$(0.1, 0.5)$	PMWLS	0.75	1.85	2.61	16.87	16.65	16.38
	PMWLS ($\rho = 0.5$)	0.69	1.89	2.76	16.85	16.72	16.70
	PMWLS ($\rho = 0.9$)	0.59	1.90	2.79	16.86	16.74	16.55
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.52	1.71	2.75	16.85	16.80	16.54
	PWLS	0.42	1.66	2.71	16.95	16.82	16.66
	PWLS ($\rho = 0.5$)	0.36	1.78	2.74	16.91	16.79	16.84
	PWLS ($\rho = 0.9$)	0.58	1.86	2.75	16.79	16.73	16.62
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.41	1.75	2.69	16.89	16.76	16.56
$(0.5, 0.5)$	PMWLS	0.94	1.53	2.55	16.78	16.72	16.51
	PMWLS ($\rho = 0.5$)	0.75	1.71	2.58	16.83	16.77	16.59
	PMWLS ($\rho = 0.9$)	0.70	1.73	2.58	16.78	16.76	16.61
	PMWLS ($\rho = 0.8, \phi = 0.4$)	0.64	1.50	2.60	16.92	16.84	16.51
	PWLS	1.06	1.81	2.77	16.87	16.75	16.76
	PWLS ($\rho = 0.5$)	0.77	1.42	2.77	16.78	16.88	16.78
	PWLS ($\rho = 0.9$)	0.66	1.58	2.56	16.89	16.89	16.74
	PWLS ($\rho = 0.8, \phi = 0.4$)	0.46	1.52	2.51	16.91	16.88	16.75

Tables S4-S6 demonstrate selection results of PMWLS and PWLS methods with the LASSO penalty. True positive (TP) counts the number of significant estimates among the significant true parameters and true negative (TN) counts the number of insignificant estimates among the insignificant true parameters.

Tables S7 and S8 provides the estimation and selection results using nonlinear multiplicative model:

$$y_t = \frac{1}{1 + \exp(-\boldsymbol{x}_t^T \boldsymbol{\theta}_0)} \times \epsilon_t. \tag{S2.16}$$

For ϵ_t, the exponentiated AR processes or an ARMA process are considered since the ϵ_t's in the equation (S2.16) are allowed to have only positive values. The AR and ARMA coefficients and the parameter setting of $\boldsymbol{\theta}$ are the same as the one in the model (S2.15). We transformed the model in the log scale and apply our approach. For comparison, an 'Additive' method is considered, where the estimator is calculated as if the data are from a nonlinear additive model without log transformation. For this simulation, we provide the results using the SCAD penalty.

Table S5: Selection results with LASSO for the equation (S2.15) when the error process is AR(1) with $\rho = 0.9$.

AR(1) with $\rho = 0.9$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
$(0.1, 0.5)$	PMWLS	1.38	1.89	2.62	16.52	16.54	16.28
	PMWLS ($\rho = 0.5$)	1.85	2.36	2.96	16.85	16.94	16.82
	PMWLS ($\rho = 0.9$)	1.67	2.31	2.95	16.88	16.94	16.80
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.75	2.21	2.94	16.85	16.89	16.78
	PWLS	1.36	1.84	2.73	16.89	16.80	16.58
	PWLS ($\rho = 0.5$)	1.59	2.15	2.90	16.85	16.98	16.93
	PWLS ($\rho = 0.9$)	1.84	2.32	2.95	16.90	16.90	16.83
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.79	2.24	2.94	16.86	16.91	16.79
$(0.5, 0.5)$	PMWLS	1.67	1.88	2.63	16.45	16.53	16.30
	PMWLS ($\rho = 0.5$)	1.83	2.47	2.88	16.73	16.83	16.80
	PMWLS ($\rho = 0.9$)	1.73	2.48	2.92	16.80	16.90	16.94
	PMWLS ($\rho = 0.8, \phi = 0.4$)	1.71	2.44	2.92	16.78	16.81	16.85
	PWLS	1.59	2.02	2.73	16.76	16.83	16.65
	PWLS ($\rho = 0.5$)	1.79	2.17	2.89	16.79	16.88	16.94
	PWLS ($\rho = 0.9$)	1.69	2.44	2.94	16.88	16.97	16.95
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.68	2.26	2.89	16.85	16.95	16.90

Table S6: Selection results with LASSO for the equation (S2.15) when the error process is ARMA(1,1) with $(\rho, \phi) = (0.8, 0.4)$.

ARMA(1,1) with $\rho = 0.8, \phi = 0.4$							
(μ, σ)	Methods	TP			TN		
		50	100	200	50	100	200
$(0.1, 0.5)$	PMWLS	1.96	2.50	2.89	16.63	16.36	15.99
	PMWLS ($\rho = 0.5$)	2.29	2.82	3.00	16.74	16.74	16.59
	PMWLS ($\rho = 0.9$)	2.34	2.79	3.00	16.81	16.76	16.60
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.22	2.82	3.00	16.77	16.69	16.48
	PWLS	1.85	2.59	2.93	16.80	16.55	16.45
	PWLS ($\rho = 0.5$)	2.11	2.83	3.00	16.77	16.85	16.87
	PWLS ($\rho = 0.9$)	2.32	2.83	3.00	16.78	16.86	16.77
	PWLS ($\rho = 0.8, \phi = 0.4$)	2.20	2.78	3.00	16.72	16.76	16.64
$(0.5, 0.5)$	PMWLS	1.75	2.58	2.95	16.45	16.16	16.03
	PMWLS ($\rho = 0.5$)	2.22	2.82	3.00	16.67	16.72	16.47
	PMWLS ($\rho = 0.9$)	2.10	2.81	3.00	16.84	16.68	16.72
	PMWLS ($\rho = 0.8, \phi = 0.4$)	2.04	2.81	2.98	16.75	16.68	16.57
	PWLS	1.88	2.81	2.96	16.71	16.63	16.58
	PWLS ($\rho = 0.5$)	2.02	2.81	3.00	16.80	16.88	16.87
	PWLS ($\rho = 0.9$)	2.00	2.77	2.99	16.88	16.87	16.89
	PWLS ($\rho = 0.8, \phi = 0.4$)	1.95	2.77	2.97	16.88	16.86	16.77

Table S7: Estimation results with SCAD for the data from the equation (S2.16). The Model column refers to the exponentiated error process for the data generation. The other configurations are identical to Table S1.

Model	(μ, σ)	Methods	$n = 50$	$n = 100$	$n = 200$
AR(1) $\rho = 0.5$	$(0.1, 0.5)$	PMWLS	0.88 (0.42)	0.32 (0.26)	0.15 (0.17)
		Additive	6.59 (1.07)	4.25 (0.72)	2.67 (0.49)
	$(0.5, 0.5)$	PMWLS	0.68 (0.37)	0.30 (0.25)	0.14 (0.16)
		Additive	79.04 (3.22)	85.64 (2.64)	48.82 (1.46)
AR(1) $\rho = 0.9$	$(0.1, 0.5)$	PMWLS	0.61 (0.35)	0.24 (0.22)	0.13 (0.16)
		Additive	16.30 (1.68)	10.87 (1.30)	3.98 (0.66)
	$(0.5, 0.5)$	PMWLS	0.55 (0.33)	0.28 (0.24)	0.13 (0.16)
		Additive	92.78 (3.47)	68.63 (2.54)	47.81 (1.93)
ARMA(1,1) $(\rho, \phi) = (0.8, 0.4)$	$(0.1, 0.5)$	PMWLS	0.37 (0.27)	0.14 (0.17)	0.09 (0.14)
		Additive	8.27 (1.16)	3.44 (0.68)	1.51 (0.37)
	$(0.5, 0.5)$	PMWLS	0.44 (0.30)	0.15 (0.18)	0.07 (0.12)
		Additive	66.88 (2.77)	41.68 (1.91)	31.44 (1.04)

Table S8: Selection results with SCAD for the data from the equation (S2.16). The other configurations refer to Table S7.

Model	(μ, σ)	Methods	TP			TN		
			n=50	n=100	200	50	100	200
AR(1) $\rho = 0.5$	$(0.1, 0.5)$	PMWLS	2.88	2.99	3.00	16.83	16.92	16.94
		Additive	2.54	2.72	2.99	16.99	17.00	17.00
	$(0.5, 0.5)$	PMWLS	2.94	2.99	3.00	16.84	16.94	17.00
		Additive	2.26	2.61	2.89	16.99	17.00	16.99
AR(1) $\rho = 0.9$	$(0.1, 0.5)$	PMWLS	2.94	3.00	3.00	16.85	16.92	16.96
		Additive	2.62	2.90	2.96	16.99	17.00	17.00
	$(0.5, 0.5)$	PMWLS	2.97	2.99	3.00	16.74	16.90	16.97
		Additive	2.40	2.68	2.82	16.99	17.00	16.99
ARMA(1,1) $(\rho, \phi) = (0.8, 0.4)$	$(0.1, 0.5)$	PMWLS	2.99	3.00	3.00	16.81	16.96	16.92
		Additive	2.87	2.98	3.00	16.93	16.99	16.99
	$(0.5, 0.5)$	PMWLS	2.98	3.00	3.00	16.70	16.94	16.95
		Additive	2.64	2.93	2.97	16.95	16.99	17.00

Bibliography

Feng, C., H. Wang, Y. Han, Y. Xia, and X. M. Tu (2013). The mean value theorem and taylor's expansion in statistics. *The American Statistician 67*(4), 245–248.

Peligrad, M. and S. Utev (1997). Central limit theorem for linear processes. *The Annals of Probability 25*(1), 443–456.