Supplementary Material for "A Model-averaging method for high-dimensional regression with missing responses at random"

Jinhan Xie[a] Xiaodong Yan[b] and Niansheng Tang[a]

[a]Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University

School of Economics, Shandong University

Supplementary Material

S1. Properties of penalized likelihood estimator $\hat{\boldsymbol{\gamma}}$

In this section, we investigate the consistency, and oracle property of the penalized likelihood estimator $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$. Without loss of generality, we can write $\boldsymbol{\gamma} = (\boldsymbol{\gamma}_1^\top, \boldsymbol{\gamma}_2^\top)^\top$, where $\boldsymbol{\gamma}_1 \in \boldsymbol{R}^{d_m}$ and $\boldsymbol{\gamma}_2 \in \boldsymbol{R}^{q-d_m}$ correspond to the nonzero and zero components of $\boldsymbol{\gamma}$, respectively. Thus, the true parameter vector $\boldsymbol{\gamma}_0$ of $\boldsymbol{\gamma}$ can be written as $\boldsymbol{\gamma}_0 = (\boldsymbol{\gamma}_{10}^\top, \mathbf{0}^\top)^\top$, where $\boldsymbol{\gamma}_{10}$ is the true value of $\boldsymbol{\gamma}_1$. In addition, the penalized likelihood estimator $\hat{\boldsymbol{\gamma}}$ of $\boldsymbol{\gamma}$ can be written as $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1^\top, \hat{\boldsymbol{\gamma}}_2^\top)^\top$. Let $\mathscr{A}_\gamma = \{j : \gamma_{0j} \neq 0\}$ be the index set of nonzero components of $\boldsymbol{\gamma}_0$, where γ_{0j} is the jth component of $\boldsymbol{\gamma}_0$ for $j = 1, \ldots, q$. Denote the cardinality of $\mathscr{A}_\gamma$ as $d_m = |\mathscr{A}_\gamma|$, which is usually unknown to be estimated in applications. Here, we assume that the non-

sparsity size $d_m \ll n$, and the dimensionality satisfies $\log(q) = O(n^\alpha)$ for some $\alpha \in (0, 1/2)$. Following Lv and Fan (2009), Zhang (2010), and Fan and Lv (2011), we define the local concavity of the penalty function $f_{\lambda_n}(t)$ at $\boldsymbol{v} = (v_1, \ldots, v_q)^\top \in \boldsymbol{R}^q$ (i.e., $\|\boldsymbol{v}\|_0 = q$) as

$$\rho(f_{\lambda_n}; \boldsymbol{v}) = \lim_{\epsilon \to 0^+} \max_{1 \le j \le q} \sup_{t_1 < t_2 \in (|v_j| - \epsilon, |v_j| + \epsilon)} -\frac{\partial_t f_{\lambda_n}(t_2) - \partial_t f_{\lambda_n}(t_1)}{t_2 - t_1},$$

where $\partial_t^k f_{\lambda_n}(t)$ represents the k-order derivation of $f_{\lambda_n}(t)$ with respect to t, and $\|\boldsymbol{A}\|_m$ denotes the L_m norm of a vector or matrix $\boldsymbol{A}$ for $m \in [0, \infty]$.

We use $\mathrm{tr}(\boldsymbol{A})$ to represent the trace of matrix $\boldsymbol{A}$. Denote $g_i(\boldsymbol{\gamma}) = \log[\pi(\boldsymbol{U}_i; \boldsymbol{\gamma})/\{1 - \pi(\boldsymbol{U}_i; \boldsymbol{\gamma})\}]$ for $i = 1, \ldots, n$, and define the Fisher information matrix as

$$\boldsymbol{F}_n(\boldsymbol{\gamma}) = E\{-\partial^2 l_n(\boldsymbol{\gamma})/\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}^\top\} = \partial_{\boldsymbol{\gamma}}^\top \boldsymbol{g}(\boldsymbol{\gamma}) \boldsymbol{\Sigma}(\boldsymbol{\gamma}) \partial_{\boldsymbol{\gamma}} \boldsymbol{g}(\boldsymbol{\gamma}),$$

where $\boldsymbol{g}(\boldsymbol{\gamma}) = (g_1(\boldsymbol{\gamma}), \ldots, g_n(\boldsymbol{\gamma}))^\top$, and $\boldsymbol{\Sigma}(\boldsymbol{\gamma}) = \mathrm{diag}(\pi_1(1 - \pi_1), \ldots, \pi_n(1 - \pi_n))$ with $\pi_i = \pi_i(\boldsymbol{\gamma}) = \pi(\boldsymbol{U}_i; \boldsymbol{\gamma})$, for $i = 1, \ldots, n$. Let $s_n = \frac{1}{2} \min_j \{|\gamma_{0j}| : \gamma_{0j} \neq 0\}$, and define $\mathscr{N} = \{\boldsymbol{\tau} = (\boldsymbol{\tau}_1^\top, \boldsymbol{\tau}_2^\top)^\top \in R^q : \boldsymbol{\tau}_2 = 0, \|\boldsymbol{\tau}_1 - \boldsymbol{\gamma}_{10}\|_\infty \le s_n\}$. The following assumptions are required to ensure the consistency of $\hat{\boldsymbol{\gamma}}$.

Assumption 4. The penalty function $f_{\lambda_n}(t)$ is increasing and concave with respect to $t \in [0, \infty)$, and has a continuous derivation $\partial f_{\lambda_n}(t)$ with $\partial f_{\lambda_n}(0+) = c_0$, where c_0 is a positive constant. Also, $\partial f_{\lambda_n}(t)$ is increasing with respect to $\lambda_n \in (0, \infty)$, and $\partial f_{\lambda_n}(0+)$ is independent of λ_n.

Assumption 4 holds for a class of penalty functions, such as the SCAD and MCP penalty functions. By Assumption 4, $f_{\lambda_n}(t)$ is a concave function with respect to $t \in [0, \infty)$ to ensure $\rho(f_{\lambda_n}; \boldsymbol{v}) \geq 0$.

Assumption 5. (i) $\min_{\boldsymbol{\tau} \in \mathscr{N}} \mathbb{E}_{\min}\{\boldsymbol{F}_n(\boldsymbol{\tau})\} \geq cn$ and $\mathrm{tr}\{\boldsymbol{F}_n(\boldsymbol{\gamma}_0)\} = O(d_m n)$;

(ii) $\left\| \partial^{\top}_{\boldsymbol{\gamma}_2} \boldsymbol{g}(\boldsymbol{\gamma}_0) \boldsymbol{\Sigma}(\boldsymbol{\gamma}_0) \partial_{\boldsymbol{\gamma}_1} \boldsymbol{g}(\boldsymbol{\gamma}_0) \right\|_{2,\infty} = O(n)$, where notation $\|B\|_{2,\infty}$ represents $\max_{\|\boldsymbol{v}\|_2=1} \|B\boldsymbol{v}\|_{\infty}$;

(iii) $\max_{\boldsymbol{\tau} \in \mathscr{N}, 1 \leq j \leq q} \mathbb{E}_{\max}\Big[\partial^{\top}_{\boldsymbol{\gamma}_1} \boldsymbol{g}(\boldsymbol{\tau}) \mathrm{diag}\big\{ |\partial_{\gamma_j} \boldsymbol{g}(\boldsymbol{\tau})| \circ |\partial^2_{\boldsymbol{\gamma}_1} \boldsymbol{V}(\boldsymbol{\tau})| \big\} \partial_{\boldsymbol{\gamma}_1} \boldsymbol{g}(\boldsymbol{\tau}) \Big] = O(n)$, where $\boldsymbol{V}(\boldsymbol{\gamma}) = (\pi_1(\boldsymbol{\gamma}), \ldots, \pi_n(\boldsymbol{\gamma}))^{\top}$.

Assumption 5(i) has been used in Fan and Lv (2011), and ensures that the information matrix $\boldsymbol{F}_n(\boldsymbol{\tau})$ is positive definite, and its eigenvalues are uniformly bounded. Assumption 5(ii) measures the correlation between each unimportant variable and important variable using the weighted matrix $\boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)$, and controls the uniformly growth rate of these regression coefficients. This assumption is similar to the strong irrepresentable condition of Zhao and Yu (2006) for the consistency of Lasso estimator. Assumption 5(iii) is used to control the order of the remainder term when taking the third-order expansion of the objective function.

Assumption 6. $s_n \gg \lambda_n \gg \sqrt{d_m/n}$, $\sqrt{n} \partial f_{\lambda_n}(s_n) = O(1)$ and $\rho_0 = o(1)$, where $\rho_0 = \max_{\boldsymbol{\tau} \in \mathscr{N}} \rho(f_{\lambda_n}; \boldsymbol{\tau})$.

Assumption 6 shows that the minimum signal s_n should be satisfied, and is used to obtain nice properties of the proposed PLE like other variable selection methods. However, for the L_1 penalty, $\lambda_n = \partial f_{\lambda_n}(s_n) = O(n^{-1/2})$ is in conflict with the assumption $\lambda_n \gg \sqrt{d_m/n}$, which implies that the L_1 penalized likelihood estimator can usually not achieve the consistency rate of $O_p(\sqrt{d_m/n})$ given in Theorem S1.1, and has not the oracle property like the SCAD penalty function. The assumption $s_n \gg \lambda_n$ holds automatically for the SCAD penalty function. Thus, Assumption 6 is less restrictive for the SCAD penalty function.

Theorem S1.1. *Suppose that Assumptions 4–6 hold. There is a strict local maximizer* $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1^\top, \hat{\boldsymbol{\gamma}}_2^\top)^\top$ *of the nonconcave penalized likelihood* $Q_n(\boldsymbol{\gamma})$ *with respect to* $\boldsymbol{\gamma}$ *such that* $\hat{\boldsymbol{\gamma}}_2 = 0$ *with probability tending to* 1 *as* $n \to \infty$ *and* $\|\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}\|_2 = O_p(\sqrt{d_m/n})$.

Theorem S1.1 shows that the sparsity property of the proposed PLE still holds in a high-dimensional parametric model. That is, zero components in $\boldsymbol{\gamma}_0$ are estimated as zero with probability tending to one. Also, Theorem S1.1 establishes the consistency of the proposed PLE $\hat{\boldsymbol{\gamma}}_1$ of $\boldsymbol{\gamma}_1$, i.e., there is a root-(n/d_m)-consistent PLE of $\boldsymbol{\gamma}_1$.

To establish the asymptotic normality of the proposed PLE, we need the following additional assumption, which is associated with the Lyapunov

assumptions.

Assumption 7. $\partial f_{\lambda_n}(s_n) = o(1/\sqrt{nd_m})$, $\max_i E|\delta_i - \pi_i(\boldsymbol{\gamma}_0)|^3 = O(1)$, and $\sum_{i=1}^n \{\partial_{\boldsymbol{\gamma}_1}^\top g_i(\boldsymbol{\gamma}_0)\boldsymbol{F}_{n11}^{-1}(\boldsymbol{\gamma}_0)\partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0)\}^{3/2} \to 0$ as $n \to \infty$, where $\boldsymbol{F}_{n11}(\boldsymbol{\gamma}_0) = \partial_{\boldsymbol{\gamma}_1}^\top \boldsymbol{g}(\boldsymbol{\gamma}_0)\ \boldsymbol{\Sigma}(\boldsymbol{\gamma}_0)\partial_{\boldsymbol{\gamma}_1}\boldsymbol{g}(\boldsymbol{\gamma}_0)$.

Theorem S1.2. *Suppose that Assumptions 4–7 hold and* $d_m = o(n^{1/4})$. *Then, we have*

(i) (Sparsity) $\hat{\boldsymbol{\gamma}}_2 = 0$ with probability tending to 1 as $n \to \infty$.

(ii) (Asymptotic Normality) $\boldsymbol{U}_n \boldsymbol{F}_{n11}^{1/2}(\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$, where $\boldsymbol{U}_n$ is an $m \times d$ matrix such that $\boldsymbol{U}_n\boldsymbol{U}_n^\top \to \boldsymbol{G}$, $\boldsymbol{G}$ is an $m \times m$ symmetric positive definite matrix with the fixed m, and $\xrightarrow{\mathcal{L}}$ represents convergence in distribution.

Theorem S1.2 indicates that the sparsity and asymptotic normality of the proposed PLE still hold even for dimensionality of nonpolynomial order of sample size.

S2. Properties of the proposed screening procedure

Under the assumption $Y \perp \delta | X_k$ for $k = 1, \ldots, p$, $\hat{r}_k$ can be regarded as the empirical estimator of $E(X_k Y)$ in the presence of responses MAR. Without of loss generality, for each $k = 1, \ldots, p$, the kth column of covariates satisfies $E(X_k) = 0$, $E(X_k^2) = 1$. Then we have $E(X_k Y) = cov(X_k, Y)$

and $\beta_k = cov(X_k, Y)$, which indicates that β_k is the covariance between X_k and Y. Hence, $\beta_k = 0$ is equivalent to the fact that X_k and Y are marginally uncorrelated. Thus, define the index set of the active predictors as $\mathcal{M}_* = \{k : \beta_k \neq 0 \text{ for } 1 \leq k \leq p\}$, which corresponds to the true sparse model with nonsparsity size $|\mathcal{M}_*|$, where $|\mathcal{M}_*|$ is the cardinality of $\mathcal{M}_*$, and denote $\mathcal{I}_* = \{1, \ldots, p\} \backslash \mathcal{M}_*$ as the index set of the inactive predictors, where $p \gg n$. Here, we assume that $p \gg |\mathcal{M}_*|$ in ultrahigh-dimensional data analysis and define $r_k = E(X_k Y)$. To investigate the sure screening properties of the presented screening criterion, we require the following assumptions.

Assumption 8. For $k = 1, \ldots, p$, the probability density function of X_k, say $f_k(x)$, has continuous and bounded second order derivatives over the support $\mathcal{X}_k$ of X_k, and is bounded away from zero and infinity uniformly over $\mathcal{X}_k$.

Assumption 9. The kernel function $K(\cdot)$ is a probability density function such that (i) it is bounded and has compact support; (ii) it is symmetric with $\int t^2 K(t) dt < 1$; (iii) $K(\cdot) \geq d_1$ for some positive constant d_1 in some closed interval centered at zero; (iv) $\sqrt{n} h^2 \to 0$ as $n \to \infty$.

Assumption 10. Variables X_k, Y and $X_k Y$ satisfy the sub-exponential tail probability uniformly in p. That is, there exists a positive constant u_0

such that for all $0 < \tilde{u} \leq 2u_0$,

$$\max_{1\leq k\leq p} E\{\exp(2\tilde{u}X_k^2)\} < \infty, \quad E\{\exp(2\tilde{u}Y^2)\} < \infty,$$

$$\max_{1\leq k\leq p} E\{\exp(2\tilde{u}X_kY)\} < \infty.$$

Assumption 11. There exists a positive constant $c_0 > 0$ and $0 < \varsigma < 1/2$ such that $\min_{k\in\mathcal{M}_*} |r_k| \geq c_0 n^{-\varsigma}$.

Assumption 12. $\lim_{p\to\infty} \inf\{\min_{k\in\mathcal{M}_*} |r_k| - \max_{k\notin\mathcal{M}_*} |r_k|\} \geq m_0$ for some $m_0 > 0$.

Assumptions 8 and 9 impose some regularity assumptions on the probability density functions $f_k(x)$ and kernel function $K(\cdot)$, respectively, which hold for the widely used distributions and kernel functions. The assumption that $\sqrt{n}h^2 \to 0$ is to control the bias induced by the kernel smoothing. Assumption 10 has widely used in high-dimensional data analysis (Fan and Lv, 2008, and Li et al., 2012), and holds if $\boldsymbol{X}$, Y and X_kY are bounded uniformly of $\boldsymbol{X}$, Y, and X_kY have multivariate normal distribution. Assumption 11 allows the minimal signal between active variables and response variable to be the order of $n^{-\varsigma}$, which is a widely used condition to guarantee the sure screening property. Assumption 12 ensures that the active and inactive predictors can be well separated in the population level. Assumption 12 is similar to Condition (C3) of Cui et al. (2015).

Theorem S2.1. *(Sure Screening Property) Under Assumptions 1 and 8–10, then for any constant $c_2 > 0$, there exists $c_7 > 0$ such that*

$$\Pr\Big(\max_{1\le k\le p} |\hat{r}_k - r_k| \ge c_2 n^{-\varsigma} \Big) \le O\{p \exp(-c_7 n^{(1-2\varsigma)/3} + \log(n))\}$$

for n sufficiently large. Furthermore, under Assumption 11, by taking $\varrho_n = c_8 n^{-\varsigma}$ with $c_8 \le c_0/2$, there exists some positive constant c_9 such that

$$\Pr(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \ge 1 - O\{|\mathcal{M}_*| \exp(-c_9 n^{(1-2\varsigma)/3} + \log(n))\}.$$

(Ranking Consistency Property) If Assumptions 1, 8–12 and additional assumptions $\log(p) = o(n^{1/3} m_0^{2/3})$ and $\log(n) = o(n^{1/3} m_0^{2/3})$ hold. Then, we have

$$\lim_{n\to\infty} \inf \left\{ \min_{k\in\mathcal{M}_*} |\hat{r}_k| - \max_{k\in\mathcal{I}_*} |\hat{r}_k| \right\} > 0, a.s..$$

Theorem S2.1 shows the sure screening and rank consistency properties of the proposed screening procedure, which indicates that the proposed MI-SIS method can handle the NP-dimensionality problem. Specifically, as $n \to \infty$, the maximum dimensional is $p = o\{\exp(n^{(1-2\varsigma)/3})\}$ for $\varsigma \in (0, 1/2)$. In addition, through the ranking consistency property, we can separate the active and inactive predictors by taking an ideal threshold value ϱ_n, which is smaller than the minimum signal.

S3. Proofs of all theorems

Let $\boldsymbol{P}_s^* = \boldsymbol{D}_s(\boldsymbol{P}_s - \boldsymbol{I}) + \boldsymbol{I}$, where $\boldsymbol{P}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top \boldsymbol{X}_s)^{-1}\boldsymbol{X}_s$, $\boldsymbol{D}_s = \mathrm{diag}(d_1^s, \ldots, d_n^s)$, $d_i^s = 1/(1-h_{ii}^s)$ and h_{ii}^s is the ith diagonal element of $\boldsymbol{P}_s = \boldsymbol{X}_s(\boldsymbol{X}_s^\top \boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top$. Let $\|\boldsymbol{A}\|$ be the Frobenius norm (e.g., $\|\boldsymbol{A}\| = \sqrt{\mathrm{tr}(\boldsymbol{A}^\top \boldsymbol{A})}$), and $\|\boldsymbol{A}\|_2$ be the Euclidean norm (e.g., $\|\boldsymbol{A}\|_2 = \sqrt{\mathbb{E}_{\max}(\boldsymbol{A}^* \boldsymbol{A})}$) of matrix $\boldsymbol{A}$, where $\boldsymbol{A}^*$ represents the conjugate transpose of matrix $\boldsymbol{A}$. Denote $\boldsymbol{P}^*(\boldsymbol{\omega}) = \sum_{s=1}^S \omega_s \boldsymbol{P}_s^*$ and $\boldsymbol{P}(\boldsymbol{\omega}) = \sum_{s=1}^S \omega_s \boldsymbol{P}_s$.

Proof of Theorem S1.1. First, we show the consistency of the proposed PLE in the d_m-dimensional subspace. To this end, we constrain the $Q_n(\boldsymbol{\gamma})$ on the d_m-dimensional subspace $\{\boldsymbol{\gamma} \in \mathbb{R}^q : \gamma_j = 0, j \in \mathscr{A}_\gamma^c\}$ of $\mathbb{R}^q$, and the corresponding constrained penalized likelihood function is given by

$$\tilde{Q}_n(\boldsymbol{\tau}) = \frac{1}{n}\tilde{l}_n(\boldsymbol{\tau}) - \sum_{j=1}^{d_m} f_{\lambda_n}(|\tau_j|), \tag{S3.1}$$

where $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_{d_m})^\top$ and $\tilde{l}_n(\boldsymbol{\tau}) = \sum_{i=1}^n [\delta_i \log\{\pi_i(\boldsymbol{U}_i^\tau; \boldsymbol{\tau})\} + (1-\delta_i)\log\{1 - \pi(\boldsymbol{U}_i^\tau; \boldsymbol{\tau})\}]$ in which $\boldsymbol{U}_i^\tau$ is the subvector of $\boldsymbol{U}_i$ corresponding to $\boldsymbol{\tau}$.

Let $\alpha_n = \sqrt{d_m/n}$, and define the closed set $\mathscr{N}_0 = \{\boldsymbol{\tau} \in \mathbb{R}^{d_m} : \|\boldsymbol{\tau} - \boldsymbol{\gamma}_{10}\|_2 \leq \alpha_n u\}$ for $u \in (0, \infty)$. Here, the purpose is to show that, for any $\kappa > 0$ and a sufficiently large n, we have

$$\Pr\left\{\sup_{\boldsymbol{\tau} \in \partial\mathscr{N}_0} \tilde{Q}_n(\boldsymbol{\tau}) < \tilde{Q}_n(\boldsymbol{\gamma}_{10})\right\} \geq 1 - \kappa,$$

where $\partial\mathscr{N}_0$ denotes the boundary of the closed set $\mathscr{N}_0$, which implies that

there exists a local maximizer $\hat{\boldsymbol{\gamma}}_1$ in $\mathscr{N}_0$ such that $\|\hat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_{10}\|_2 = O_p(\sqrt{d_m/n})$.

For a sufficiently large n, it follows from Assumption 6 that $\alpha_n u \leq d_m$. Taking Taylor expansion of $\tilde{Q}_n(\boldsymbol{\tau})$ at the true value $\boldsymbol{\gamma}_{10}$ of $\boldsymbol{\gamma}_1$ yields

$$\tilde{Q}_n(\boldsymbol{\tau}) - \tilde{Q}_n(\boldsymbol{\gamma}_{10}) = (\boldsymbol{\tau}-\boldsymbol{\gamma}_{10})^\top \boldsymbol{D}_1 - \frac{1}{2}(\boldsymbol{\tau}-\boldsymbol{\gamma}_{10})^\top \boldsymbol{D}_2 (\boldsymbol{\tau}-\boldsymbol{\gamma}_{10}) \qquad \text{(S3.2)}$$

for any $\boldsymbol{\tau} \in \mathscr{N}_0$, where $\boldsymbol{D}_1 = \partial \tilde{Q}_n(\boldsymbol{\gamma}_{10})/\partial \boldsymbol{\gamma}_1 = \partial^\top_{\boldsymbol{\gamma}_1} \boldsymbol{g}(\boldsymbol{\gamma}_{10})\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_{10})\}/n - \partial f_{\lambda_n}(\boldsymbol{\gamma}_{10})$, $\boldsymbol{D}_2 = \partial^2 \tilde{Q}_n(\tilde{\boldsymbol{\gamma}}_1)/\partial \boldsymbol{\gamma}_1 \partial \boldsymbol{\gamma}_1^\top$, and $\tilde{\boldsymbol{\gamma}}_1$ lies on the line segment jointing $\boldsymbol{\tau}$ and $\boldsymbol{\gamma}_{10}$. Following Fan and Peng (2004), we can obtain

$$\begin{aligned}
\boldsymbol{D}_2 &= -\tfrac{1}{n}\partial^2 \tilde{l}_n(\tilde{\boldsymbol{\gamma}}_1)/\partial \boldsymbol{\gamma}_1 \partial \boldsymbol{\gamma}_1^\top + \mathrm{diag}\{\partial^2 f_{\lambda_n}(|\tilde{\boldsymbol{\gamma}}_1|)\} \\
&= \tfrac{1}{n}\boldsymbol{F}_n(\tilde{\boldsymbol{\gamma}}_1) - \tfrac{1}{n}\left\{\partial^2 \tilde{l}_n(\tilde{\boldsymbol{\gamma}}_1)/\partial \boldsymbol{\gamma}_1 \partial \boldsymbol{\gamma}_1^\top - E(\partial^2 \tilde{l}_n(\tilde{\boldsymbol{\gamma}}_1)/\partial \boldsymbol{\gamma}_1 \partial \boldsymbol{\gamma}_1{}^\top)\right\} \\
&\quad + \mathrm{diag}\{\partial^2 f_{\lambda_n}(|\tilde{\boldsymbol{\gamma}}_1|)\} \\
&= \tfrac{1}{n}\boldsymbol{F}_n(\tilde{\boldsymbol{\gamma}}_1) + \mathrm{diag}\{\partial^2 f_{\lambda_n}(|\tilde{\boldsymbol{\gamma}}_1|)\} + o_p(1).
\end{aligned}$$

Without loss of generality, when there is no the second-order derivative of the penalty function $f_{\lambda_n}(\cdot)$, it is easily shown that matrix $\boldsymbol{D}_2$ can be replaced by a diagonal matrix whose maximum absolute element is bounded by ρ_0. Thus, for a sufficiently large n, under Assumptions 5(i) and 6, we have $\mathbb{E}_{\min}(\boldsymbol{D}_2) \geq c - \rho_0 \geq \frac{c}{2}$, where c is a constant.

Using the Markov's inequality and (S3.2) yields

$$\Pr\left\{\sup_{\boldsymbol{\tau}\in\partial\mathscr{N}_0} \tilde{Q}_n(\boldsymbol{\tau}) < \tilde{Q}_n(\boldsymbol{\gamma}_{10})\right\} \geq \Pr\left\{u\alpha_n\left(\|\boldsymbol{D}_1\|_2 - \frac{cu\alpha_n}{4}\right) < 0\right\} \geq 1 - \frac{16E\|\boldsymbol{D}_1\|_2^2}{c^2u^2\alpha_n^2}.$$

However, under Assumptions 5(i) and 6, we have

$$E\|\boldsymbol{D}_1\|_2^2 \leq \frac{1}{n^2}\mathrm{tr}\{\boldsymbol{F}_n(\boldsymbol{\gamma}_{10})\} + d_m \partial f_{\lambda_n}(d_m)^2 = O(\alpha_n^2),$$

which leads to

$$\Pr\left\{\sup_{\boldsymbol{\tau}\in\partial\mathscr{N}_0} \tilde{Q}_n(\boldsymbol{\tau}) < \tilde{Q}_n(\boldsymbol{\gamma}_{10})\right\} \geq 1 - O\left(\frac{16}{c^2u^2}\right).$$

Thus, we prove $\|\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}\|_2 = O_p(\sqrt{d_m/n})$.

Next, to show the sparsity of the proposed estimator, it is necessary to prove that $\hat{\boldsymbol{\gamma}} \in R^q$ is a strict local maximizer of $Q_n(\boldsymbol{\gamma})$ such that $\hat{\boldsymbol{\gamma}}_{\mathscr{A}_\gamma} = \hat{\boldsymbol{\gamma}}_1 \in \mathscr{N}_0 \subset \mathscr{N}$ and $\hat{\boldsymbol{\gamma}}_{\mathscr{A}_\gamma^c} = \hat{\boldsymbol{\gamma}}_2 = 0$. Similar to the proof of Theorem 1 of Fan and Lv (2011), we only require showing

$$\left\|\frac{1}{n\lambda_n}\partial_{\boldsymbol{\gamma}_2} l_n(\hat{\boldsymbol{\gamma}})\right\|_\infty \leq \partial f_{\lambda_n}(0+). \tag{S3.3}$$

Thus, we have

$$\begin{aligned}
\tfrac{1}{n}\partial_{\boldsymbol{\gamma}_2} l_n(\hat{\boldsymbol{\gamma}}) &= \tfrac{1}{n}\partial_{\boldsymbol{\gamma}_2}^\top \boldsymbol{g}(\hat{\boldsymbol{\gamma}})\{\boldsymbol{\delta} - \boldsymbol{V}(\hat{\boldsymbol{\gamma}})\} \\
&= \tfrac{1}{n}\big[\boldsymbol{\eta}_{\mathscr{A}_\gamma^c} + \big\{\partial_{\boldsymbol{\gamma}_2}^\top \boldsymbol{g}(\hat{\boldsymbol{\gamma}})\boldsymbol{\delta} - \partial_{\boldsymbol{\gamma}_2}^\top \boldsymbol{g}(\hat{\boldsymbol{\gamma}})\boldsymbol{V}(\hat{\boldsymbol{\gamma}})\big\} \\
&\quad -\big\{\partial_{\boldsymbol{\gamma}_2}^\top \boldsymbol{g}(\boldsymbol{\gamma}_0)\boldsymbol{\delta} - \partial_{\boldsymbol{\gamma}_2}^\top \boldsymbol{g}(\boldsymbol{\gamma}_0)\boldsymbol{V}(\boldsymbol{\gamma}_0)\big\}\big],
\end{aligned}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_q)^\top = \partial_{\boldsymbol{\gamma}}^\top \boldsymbol{g}(\boldsymbol{\gamma}_0)\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\}$.

Denote $\mathscr{B}_\gamma = \{\|\boldsymbol{\eta}_{\mathscr{A}_\gamma^c}\|_\infty \leq c_1\sqrt{n}\}$, where c_1 is some constant. Under Assumption 1, it is easily known that the first- to third-order derivatives of $g_i(\boldsymbol{\gamma})$ with respect to $\boldsymbol{\gamma}$ are bounded. By Bonferroni's inequality and

$\log(q) = O(n^{\alpha})$ for some $\alpha \in (0, 1/2)$, there exists a constant c' such that

$$\begin{aligned}\Pr(\mathscr{B}_{\gamma}) &= \Pr(\|\boldsymbol{\eta}_{\mathscr{A}_{\gamma}^{c}}\|_{\infty} \leq c_1\sqrt{n}) = 1 - \Pr(\|\boldsymbol{\eta}_{\mathscr{A}_{\gamma}^{c}}\|_{\infty} > c_1\sqrt{n}) \\ &\geq 1 - 2(p-d)\exp(-c'n) \geq 1 - 2q\exp(-c'n) \to 1.\end{aligned}$$

Under Assumptions 5(ii), 5(iii) and 6, we consider the Taylor expansion of $\partial_{\boldsymbol{\gamma}_2}^{\top}\boldsymbol{g}(\hat{\boldsymbol{\gamma}})\boldsymbol{\delta} - \partial_{\boldsymbol{\gamma}_2}^{\top}\boldsymbol{g}(\hat{\boldsymbol{\gamma}})\boldsymbol{V}(\hat{\boldsymbol{\gamma}})$ at $\boldsymbol{\gamma}_{10}$. Thus, we obtain

$$\begin{aligned}\big\|\tfrac{1}{n\lambda_n}\partial_{\boldsymbol{\gamma}_2} l_n(\hat{\boldsymbol{\gamma}})\big\|_{\infty} &\leq \tfrac{1}{n\lambda_n}\big[\|\boldsymbol{\eta}_{\mathscr{A}_{\gamma}^{c}}\|_{\infty} + \big\|\big\{\partial_{\boldsymbol{\gamma}_2}^{\top}\boldsymbol{g}(\hat{\boldsymbol{\gamma}})\boldsymbol{\delta} - \partial_{\boldsymbol{\gamma}_2}^{\top}\boldsymbol{g}(\hat{\boldsymbol{\gamma}})\boldsymbol{V}(\hat{\boldsymbol{\gamma}})\big\} \\ &\qquad -\big\{\partial_{\boldsymbol{\gamma}_2}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\boldsymbol{\delta} - \partial_{\boldsymbol{\gamma}_2}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\boldsymbol{V}(\boldsymbol{\gamma}_0)\big\}\big\|_{\infty}\big] \\ &\leq o_p(1) + \tfrac{1}{n\lambda_n}\Big\{O(n)\|\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}\|_2 + O(n)\|\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}\|_2^2\Big\} \\ &= o_p(1) + O_p(\lambda_n^{-1}\sqrt{d_m/n}) = o_p(1).\end{aligned}$$

Then, for a sufficiently large n, (S3.3) holds. Hence, we finish the proof of Theorem S1.1.

Proof of Theorem S1.2. From the proof of Theorem S1.1, we only require proving the asymptotic normality of $\hat{\boldsymbol{\gamma}}_1$. For the set $\mathscr{N}$, it is easily shown that $\hat{\boldsymbol{\gamma}} = (\hat{\boldsymbol{\gamma}}_1^{\top}, \mathbf{0}^{\top})^{\top} \in \mathscr{N}$ is a strict local maximizer of $\tilde{Q}_n(\boldsymbol{\tau})$, which leads to $\partial_{\boldsymbol{\gamma}_1}\tilde{Q}_n(\hat{\boldsymbol{\gamma}}) = 0$.

Taking Taylor expansion of $\partial_{\boldsymbol{\gamma}_1}\tilde{l}_n(\hat{\boldsymbol{\gamma}})$ at $\boldsymbol{\gamma}_{10}$ leads to

$$\begin{aligned}0 =& \partial_{\boldsymbol{\gamma}_1}\tilde{Q}_n(\hat{\boldsymbol{\gamma}}) = \frac{1}{n}\partial_{\boldsymbol{\gamma}_1}\tilde{l}_n(\hat{\boldsymbol{\gamma}}) - \partial f_{\lambda_n}(\hat{\boldsymbol{\gamma}}) \\ =& \frac{1}{n}\partial_{\boldsymbol{\gamma}_1}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\} \\ &- \frac{1}{n}\Big[\boldsymbol{F}_{n11}(\boldsymbol{\gamma}_0) - \Big\{\frac{\partial^2 l_n(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\gamma}_1^{\top}} - E\Big(\frac{\partial^2 l_n(\boldsymbol{\gamma}_0)}{\partial\boldsymbol{\gamma}_1\partial\boldsymbol{\gamma}_1^{\top}}\Big)\Big\}\Big](\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10})\end{aligned}$$

$$
\begin{aligned}
&+ O(1)\|\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}\|_2^2 d_m - \partial f_{\lambda_n}(\hat{\boldsymbol{\gamma}}) \\
=&\frac{1}{n}\partial_{\boldsymbol{\gamma}_1}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\} - \frac{1}{n}\boldsymbol{F}_{n11}(\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) - \partial f_{\lambda_n}(\hat{\boldsymbol{\gamma}}) \\
&+ o_p(n^{-1/2}d_m^{-1/2}) + O_p(d_m^2 n^{-1}).
\end{aligned}
$$

Under Assumption 7, it follows from $\min_j |\tau_j| \geq \min_{j\in\mathscr{A}_\gamma} |\gamma_{0j}| - s_n = s_n$ and the monotonicity of $\partial f_{\lambda_n}(t)$ that

$$
\|\partial f_{\lambda_n}(\hat{\boldsymbol{\gamma}})\|_2 \leq \sqrt{s_n}\partial f_{\lambda_n}(s_n) = o_p(n^{-1/2}). \tag{S3.4}
$$

Combining $d_m = o(n^{1/4})$, (S2.4) and (S3.4) leads to

$$
\boldsymbol{F}_{n11}(\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) = \partial_{\boldsymbol{\gamma}_1}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\} + o_p(\sqrt{n}).
$$

By Assumption 5(i), we obtain

$$
\boldsymbol{F}_{n11}^{1/2}(\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) = \boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0)\partial_{\boldsymbol{\gamma}_1}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\} + o_p(1). \tag{S3.5}
$$

Let $\boldsymbol{U}_n\boldsymbol{U}_n^{\top} = \boldsymbol{G}$, where $\boldsymbol{U}_n$ is an $m \times d_m$ matrix, and $\boldsymbol{G}$ is an $m \times m$ symmetric positive define matrix. Then, we have $\boldsymbol{U}_n\boldsymbol{F}_{n11}^{1/2}(\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) = \boldsymbol{\nu}_n + o_p(1)$, where $\boldsymbol{\nu}_n = \boldsymbol{U}_n\boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0)\partial_{\boldsymbol{\gamma}_1}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\ \{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\}$.

If we can prove $\boldsymbol{\nu}_n \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$, it follows from Slutsky's theorem that $\boldsymbol{U}_n\boldsymbol{F}_{n11}^{1/2}(\boldsymbol{\gamma}_0)(\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{G})$. For any unit vector $\boldsymbol{a} \in \boldsymbol{R}^m$, we have

$$
\boldsymbol{v}_n = \boldsymbol{a}^{\top}\boldsymbol{\nu}_n = \boldsymbol{a}^{\top}\boldsymbol{U}_n\boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0)\partial_{\boldsymbol{\gamma}_1}^{\top}\boldsymbol{g}(\boldsymbol{\gamma}_0)\{\boldsymbol{\delta} - \boldsymbol{V}(\boldsymbol{\gamma}_0)\} = \sum_{i=1}^{n} z_i,
$$

where $z_i = \boldsymbol{a}^{\top}\boldsymbol{U}_n\boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0)\partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0)\{\delta_i - V_i(\boldsymbol{\gamma}_0)\}$. It is easily seen that z_i's

are independent. For $i = 1, \ldots, n$, we have $E(z_i) = 0$ and

$$\sum_{i=1}^{n} \mathrm{var}(z_i) = \boldsymbol{a}^\top \boldsymbol{U}_n \boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0) \boldsymbol{F}_{n11}(\boldsymbol{\gamma}_0) \boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0) \boldsymbol{U}_n^\top \boldsymbol{a} = \boldsymbol{a}^\top \boldsymbol{U}_n \boldsymbol{U}_n^\top \boldsymbol{a} \xrightarrow{\mathcal{P}} \boldsymbol{a}^\top \boldsymbol{G} \boldsymbol{a},$$

where $\xrightarrow{\mathcal{P}}$ represents the convergence in probability.

Under Assumption 7, it follows from Cauchy-Schwarz inequality that

$$\begin{aligned}
\sum_{i=1}^{n} E|z_i|^3 &= \sum_{i=1}^{n} |\boldsymbol{a}^\top \boldsymbol{U}_n \boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0) \partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0)|^3 E|\delta_i - V_i(\boldsymbol{\gamma}_0)|^3 \\
&= O(1) \sum_{i=1}^{n} |\boldsymbol{a}^\top \boldsymbol{U}_n \boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0) \partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0)|^3 \\
&\leq O(1) \sum_{i=1}^{n} \|\boldsymbol{a}^\top \boldsymbol{U}_n\|_2^3 \cdot \|\boldsymbol{F}_{n11}^{-1/2}(\boldsymbol{\gamma}_0) \partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0)\|_2^3 \\
&= O(1) \sum_{i=1}^{n} \{\partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0) \boldsymbol{F}_{n11}^{-1}(\boldsymbol{\gamma}_0) \partial_{\boldsymbol{\gamma}_1} g_i(\boldsymbol{\gamma}_0)\}^{3/2} = o(1).
\end{aligned}$$

Using the Lyapunov's Theorem leads to $\boldsymbol{a}^\top \boldsymbol{\nu}_n = \sum_{i=1}^{n} z_i \xrightarrow{\mathcal{L}} \mathcal{N}(\boldsymbol{0}, \boldsymbol{a}^\top \boldsymbol{G} \boldsymbol{a})$, which holds for any unit vector $\boldsymbol{a} \in \mathbb{R}^m$. Thus, we finish the proof of Theorem S1.2(ii).

Lemma 1. *Suppose that Assumptions 1, 2 and 3(ii) hold. Let* $p_m = \sup_{1 \leq s \leq S} p_s$. *Then, as* $n \to \infty$, *we have*

(i) $\sup_{1 \leq s \leq S} \|\frac{1}{n} \boldsymbol{X_s}^\top \widehat{\boldsymbol{W}} \boldsymbol{X_s} - \frac{1}{n} \boldsymbol{X_s}^\top \boldsymbol{W} \boldsymbol{X_s}\| = O_p\left(p_m \sqrt{\frac{d_m}{n}}\right)$;

(ii) $\sup_{1 \leq s \leq S} \left\| \left(\frac{1}{n} \boldsymbol{X_s}^\top \widehat{\boldsymbol{W}} \boldsymbol{X_s}\right)^{-1} - \left(\frac{1}{n} \boldsymbol{X_s}^\top \boldsymbol{W} \boldsymbol{X_s}\right)^{-1} \right\|_2 = O_p\left(p_m \sqrt{\frac{d_m}{n}}\right)$;

(iii) $\sup_{1 \leq s \leq S} \left\| \left(\frac{1}{n} \boldsymbol{X_s}^\top \widehat{\boldsymbol{W}} \boldsymbol{X_s}\right)^{-1} \right\|_2 = O_p(1)$;

(iv) $\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) - \widehat{\boldsymbol{P}}(\boldsymbol{\omega}) = \boldsymbol{P}^*(\boldsymbol{\omega}) - \boldsymbol{P}(\boldsymbol{\omega}) + O_p\left(p_m^2 S \sqrt{\frac{d_m}{n}}\right)$.

Proof. By the definition of $\widehat{\boldsymbol{W}}$ and $\boldsymbol{W}$, we obtain

$$\begin{aligned}\big\|\frac{1}{n}\boldsymbol{X}_s^\top\widehat{\boldsymbol{W}}\boldsymbol{X}_s-\frac{1}{n}\boldsymbol{X}_s^\top\boldsymbol{W}\boldsymbol{X}_s\big\| &= \big\|n^{-1}\sum_{i=1}^{n}\left(\frac{\delta_i}{\hat{\pi}_i}-\frac{\delta_i}{\pi_i}\right)X_{si}X_{si}^\top\big\| \\ &\le \sup_i\big|\frac{\delta_i}{\hat{\pi}_i}-\frac{\delta_i}{\pi_i}\big|n^{-1}\sum_{i=1}^{n}\big\|X_{si}X_{si}^\top\big\|.\end{aligned}$$

Following Lemma 1 of Hirano et al. (2003), and by Theorem S1.1 and Assumption 1, we have

$$\sup_i\big|\frac{1}{\hat{\pi}_i}-\frac{1}{\pi_i}\big|\le\sup_i\big|\frac{\partial\pi_i/\partial\gamma_1}{\pi_i^2}\big|\cdot\|\hat{\boldsymbol{\gamma}}_1-\boldsymbol{\gamma}_{01}\|_2=O_p\left(\sqrt{\frac{d_m}{n}}\right).$$

From Assumption 2, we have $\sup_s\sum_{i=1}^{n}\|X_{si}X_{si}^\top\|/n=O_p(p_m)$. Thus, we have proved (i). For (ii), let $\widehat{\boldsymbol{H}}_s=\boldsymbol{X}_s^\top\widehat{\boldsymbol{W}}\boldsymbol{X}_s/n$ and $\boldsymbol{H}_s=\boldsymbol{X}_s^\top\boldsymbol{W}\boldsymbol{X}_s/n$. Following Theorem 1 of Lewis and Reinsel (1985), we obtain

$$\widehat{\boldsymbol{H}}_s^{-1}-\boldsymbol{H}_s^{-1}=-\widehat{\boldsymbol{H}}_s^{-1}(\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s)\boldsymbol{H}_s^{-1}=-\{\boldsymbol{H}_s^{-1}+(\widehat{\boldsymbol{H}}_s^{-1}-\boldsymbol{H}_s^{-1})\}(\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s)\boldsymbol{H}_s^{-1}.$$

Assumptions 1 and 2 imply that $\sup_s\|\boldsymbol{H}_s^{-1}\|_2\le\Lambda<\infty$ for some constant Λ. According to Assumption 3(ii) and Lemma 1(i), and $\sup_s\|\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s\|_2\le\sup_s\|\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s\|\to0$, there exists a constant Λ' such that $\Lambda\|\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s\|_2<\Lambda'<1$ as $n\to\infty$. Then, with the probability tending to one, we have

$$\sup_{1\le s\le S}\|\widehat{\boldsymbol{H}}_s^{-1}-\boldsymbol{H}_s^{-1}\|_2\le\sup_{1\le s\le S}\frac{\Lambda^2\|\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s\|_2}{1-\Lambda\|\widehat{\boldsymbol{H}}_s-\boldsymbol{H}_s\|_2},$$

which leads to (ii). Using the Triangle inequality and Assumption 3(ii) yields

$$\sup_{1\le s\le S}\|\widehat{\boldsymbol{H}}_s^{-1}\|_2\le\sup_s\|\boldsymbol{H}_s^{-1}\|_2+\sup_{1\le s\le S}\|\widehat{\boldsymbol{H}}_s^{-1}-\boldsymbol{H}_s^{-1}\|_2=O_p(1),$$

which shows that Lemma 1(iii) holds. It follows from Wiener and Masani (1958) that $\|\boldsymbol{AB}\| \leq \|\boldsymbol{A}\|_2\|\boldsymbol{B}\| + \|\boldsymbol{A}\|\|\boldsymbol{B}\|_2$ for any matrices $\boldsymbol{A}$ and $\boldsymbol{B}$. Combining Theorem S1.1 and Lemma 1(iii) leads to

$$\begin{aligned}
\sup_{1\leq s\leq S}\|\widehat{\boldsymbol{P}}_s - \boldsymbol{P}_s\| &= \sup_{1\leq s\leq S}\|\tfrac{1}{n}\boldsymbol{X}_s\widehat{\boldsymbol{H}}_s^{-1}\boldsymbol{X}_s^\top\widehat{\boldsymbol{W}} - \tfrac{1}{n}\boldsymbol{X}_s(\tfrac{1}{n}\boldsymbol{X}_s^\top\boldsymbol{X}_s)^{-1}\boldsymbol{X}_s^\top\| \\
&\leq \sup_i\left|\tfrac{\delta_i}{\hat{\pi}_i} - \tfrac{\delta_i}{\pi_i}\right| \cdot \sup_s\|\widehat{\boldsymbol{H}}_s^{-1}\|_2 \cdot \sup_s \tfrac{1}{n}\sum_{i=1}^n\|X_{si}X_{si}^\top\| \\
&\quad + \sup_i\left|\tfrac{\delta_i}{\pi_i} - 1\right| \cdot \sup_s\|\widehat{\boldsymbol{H}}_s^{-1}\|_2 \cdot \sup_s \tfrac{1}{n}\sum_{i=1}^n\|X_{si}X_{si}^\top\| \\
&\quad + \sup_s\|\widehat{\boldsymbol{H}}_s^{-1} - (\tfrac{1}{n}\boldsymbol{X}_s^\top\boldsymbol{X}_s)^{-1}\|_2 \cdot \sup_s \tfrac{1}{n}\sum_{i=1}^n\|X_{si}X_{si}^\top\| \\
&= O_p\left(p_m^2\sqrt{\tfrac{d_m}{n}}\right).
\end{aligned}$$

Similarly, we can obtain $\sup_s\|\widehat{\boldsymbol{D}}_s - \boldsymbol{D}_s\| = O_p\left(p_m^2\sqrt{\frac{d_m}{n}}\right)$. It follows from $\widetilde{\boldsymbol{P}}_s = \widehat{\boldsymbol{D}}_s(\widehat{\boldsymbol{P}}_s - \boldsymbol{I}) + \boldsymbol{I}$ and the definition of $\boldsymbol{P}_s^*$ that

$$\begin{aligned}
\sup_{1\leq s\leq S}\|\widetilde{\boldsymbol{P}}_s - \boldsymbol{P}_s^*\| &= \sup_{1\leq s\leq S}\|\widehat{\boldsymbol{D}}_s(\widehat{\boldsymbol{P}}_s - \boldsymbol{P}_s) + (\widehat{\boldsymbol{D}}_s - \boldsymbol{D}_s)\boldsymbol{P}_s + (\boldsymbol{D}_s - \widehat{\boldsymbol{D}}_s)\| \\
&= O_p\left(p_m^2\sqrt{\tfrac{d_m}{n}}\right),
\end{aligned}$$

which shows that (iv) holds.

Lemma 2. *Under Assumption 3(v), there exists a constant $C > 0$ such that* (i) $\mathbb{E}_{\max}(\boldsymbol{P}_s^* - \boldsymbol{P}_s) \leq Cp_s/n$; (ii) $\mathrm{tr}\{(\boldsymbol{P}_s^* - \boldsymbol{P}_s)^\top(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\} \leq C^2p_s^2/n$; (iii) $\mathrm{tr}\{(\boldsymbol{P}_s^* - \boldsymbol{P}_s)^\top(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\}^2 \leq Cp_s^4/n$; (iv) $\mathbb{E}_{\max}(\boldsymbol{P}_s^*) \leq 1 + Cp_s/n$; (v) $\mathrm{tr}(\boldsymbol{P}_s^*\boldsymbol{P}_s^{*\top}) \leq Cp_s$.

Proof. We can obtain the proof using Lemma 3.1 of Ando and Li (2014).

Lemma 3. *(Hoeffding's inequality) Let $X_1, \ldots, X_n$ be independent random*

variables. Assume that $\Pr(X_i \in [a_i, b_i]) = 1$ *for* $1 \le i \le n$*, where* a_i *and* b_i *are constants. Let* $\bar{X} = \sum_{i=1}^{n} X_i/n$*. Then the following inequality holds*

$$\Pr(|\bar{X} - E(\bar{X})| \ge t) \le 2\exp\Big\{ - \frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\Big\},$$

where t *is a positive constant and* $E(\bar{X})$ *is the expected value of* X*.*

Proof of Theorem 1. Let C' be a constant. Denote $L^*(\boldsymbol{\omega}) = \{\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}}\{\boldsymbol{\mu} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\}$ and $\tilde{\boldsymbol{\mu}}(\boldsymbol{\omega}) = \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}$. Then, we have

$$\begin{aligned}
\mathrm{WCV}(\boldsymbol{\omega}) &= \{\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top \widehat{\boldsymbol{W}}\{\boldsymbol{Y} - \tilde{\boldsymbol{\mu}}(\boldsymbol{\omega})\} \\
&= \{\boldsymbol{\mu} + \boldsymbol{\varepsilon} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\}^\top \widehat{\boldsymbol{W}}\{\boldsymbol{\mu} + \boldsymbol{\varepsilon} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\} \\
&= \boldsymbol{\varepsilon}^\top \widehat{\boldsymbol{W}}\boldsymbol{\varepsilon} + L^*(\boldsymbol{\omega}) + 2 < \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\{\boldsymbol{\mu} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\} > \\
&= \boldsymbol{\varepsilon}^\top \widehat{\boldsymbol{W}}\boldsymbol{\varepsilon} + L(\boldsymbol{\omega}) \left\{ \frac{L^*(\boldsymbol{\omega})}{L(\boldsymbol{\omega})} + \frac{2 < \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{\mu} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}) > /R(\boldsymbol{\omega})}{L(\boldsymbol{\omega})/R(\boldsymbol{\omega})} \right\}.
\end{aligned}$$

Thus, $\hat{\boldsymbol{\omega}}$ can be obtained by minimizing $\mathrm{WCV}^*(\boldsymbol{\omega}) = \mathrm{WCV}(\boldsymbol{\omega}) - \boldsymbol{\varepsilon}^\top \widehat{\boldsymbol{W}}\boldsymbol{\varepsilon}$ over $\boldsymbol{\omega} \in \mathcal{W}$. According to the definition of (2.4), if we can show

$$\sup_{\boldsymbol{\omega} \in \mathcal{W}} |L^*(\boldsymbol{\omega})/L(\boldsymbol{\omega}) - 1| \to 0, \tag{S3.6}$$

$$\sup_{\boldsymbol{\omega} \in \mathcal{W}} | < \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\{\boldsymbol{\mu} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\} > |/R(\boldsymbol{\omega}) \to 0, \tag{S3.7}$$

$$\sup_{\omega \in \mathcal{W}} |L(\boldsymbol{\omega})/R(\boldsymbol{\omega}) - 1| \to 0, \tag{S3.8}$$

we can obtain that $L(\hat{\boldsymbol{\omega}})/\inf_{\boldsymbol{\omega} \in \mathcal{W}} L(\boldsymbol{\omega}) \to 1$ is valid. Using the Cauchy-

Schwartz inequality leads to

$$\begin{aligned}
|L^*(\boldsymbol{\omega}) - L(\boldsymbol{\omega})| &= |\{\boldsymbol{\mu} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\}^\top \widehat{\boldsymbol{W}}\{\boldsymbol{\mu} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\} - \{\boldsymbol{\mu} - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\}^\top \widehat{\boldsymbol{W}}\{\boldsymbol{\mu} - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\}| \\
&= |\|\widehat{\boldsymbol{W}}^{1/2}\{\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\}\boldsymbol{Y}\|^2 \\
&\quad -2 < \widehat{\boldsymbol{W}}^{1/2}\{\boldsymbol{\mu} - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\}, \widehat{\boldsymbol{W}}^{1/2}\{\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\}\boldsymbol{Y} > | \\
&\leq \|\widehat{\boldsymbol{W}}^{1/2}\{\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\}\boldsymbol{Y}\|^2 + 2\sqrt{L(\boldsymbol{\omega})}\|\widehat{\boldsymbol{W}}^{1/2}\{\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\}\boldsymbol{Y}\|.
\end{aligned}$$

To show (S3.6), it is sufficient to show

$$\sup_{\boldsymbol{\omega}\in\mathcal{W}} \|\widehat{\boldsymbol{W}}^{1/2}\{\widetilde{\boldsymbol{P}}(\boldsymbol{\omega}) - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\}\boldsymbol{Y}\|^2/L(\boldsymbol{\omega}) \to 0. \tag{S3.9}$$

From Lemma 1(iv), we have $\widetilde{\boldsymbol{P}}(\boldsymbol{\omega})-\widehat{\boldsymbol{P}}(\boldsymbol{\omega}) = \boldsymbol{P}^*(\boldsymbol{\omega})-\boldsymbol{P}(\boldsymbol{\omega})+O_p(Sp_m^2\sqrt{d_m/n})$.

By (S3.8) and triangle inequality, the proof of (S3.9) is equivalent to proving

$$\sup_{\boldsymbol{\omega}\in\mathcal{W}} S^2 p_m^4 d_m n^{-1}/R(\boldsymbol{\omega}) \to 0, \tag{S3.10}$$

$$\sup_{\boldsymbol{\omega}\in\mathcal{W}} \|\widehat{\boldsymbol{W}}^{1/2}\{\boldsymbol{P}^*(\boldsymbol{\omega}) - \boldsymbol{P}(\boldsymbol{\omega})\}\boldsymbol{Y}\|^2/R(\boldsymbol{\omega}) \to 0. \tag{S3.11}$$

Under Assumptions 3(iii), (iv) and (vi), we obtain

$$\begin{aligned}
\sup_{\boldsymbol{\omega}\in\mathcal{W}} S^2 p_m^4 d_m n^{-1}/R(\boldsymbol{\omega}) &\leq S^2 p_m^4 d_m n^{-1}/\xi_n \\
&= \left(\frac{S^{4G}\|\mu\|^{2G}}{\xi_n^{2G}}\right)^{1/2G} \cdot \frac{\sqrt{n}}{\|\mu\|} \cdot \left(\frac{p_m^{8/3} d_m}{n}\right)^{3/2} \cdot \frac{1}{\sqrt{d_m}} \to 0,
\end{aligned}$$

which indicates that (S3.10) holds. Applying the triangle inequality to (S3.11) yields

$$\begin{aligned}
&||\widehat{\boldsymbol{W}}^{1/2}\{\boldsymbol{P}^*(\boldsymbol{\omega}) - \boldsymbol{P}(\boldsymbol{\omega})\}\boldsymbol{Y}||^2 \\
&= \left\{\sum_{s=1}^{S} \omega_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\| + \sum_{s=1}^{S} \omega_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|\right\}^2
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\{ \sum_{s=1}^{S} \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\| + \sum_{s=1}^{S} \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\| \right\}^2 \\
&\leq S^2 \left\{ \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\| + \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\| \right\}^2 \\
&\leq 2S^2 \left\{ \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^2 + \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2 \right\}.
\end{aligned}
$$

To prove (S3.11), it suffices to show that as $n \to \infty$, we have

$$
S^2 \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^2 / \xi_n \to 0, \tag{S3.12}
$$

$$
S^2 \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2 / \xi_n \to 0. \tag{S3.13}
$$

Note that

$$
\frac{\delta_i}{\hat{\pi}_i} = \frac{\delta_i}{\pi_i} \Big\{ 1 - \frac{\partial_{\boldsymbol{\gamma}_1}^\top \pi_i}{\pi_i} (\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}) + o_p(\sqrt{d_m/n}) \Big\}.
$$

By Theorem S1.1 and Assumption 1, we have

$$
\sup_i \left| \frac{\delta_i}{\hat{\pi}_i} \right| \leq \sup_i \left| \frac{\delta_i}{\pi_i} \right| + \sup_i \left| \frac{\partial_{\boldsymbol{\gamma}_1}^\top \pi_i}{\pi_i^2} \right| \cdot \|\hat{\boldsymbol{\gamma}}_1 - \boldsymbol{\gamma}_{10}\|_2 \leq \frac{1}{C_0} + O_p(\sqrt{d_m/n}) \leq C'. \tag{S3.14}
$$

Therefore, for (S3.12) and (S3.13), it is sufficient to show

$$
\begin{aligned}
S^2 \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^2 / \xi_n &\leq S^2 \max_s \sup_i |\delta_i / \hat{\pi}_i| \cdot \|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^2 / \xi_n \\
&\leq C' S^2 \max_s \|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^2 / \xi_n \to 0,
\end{aligned} \tag{S3.15}
$$

$$S^2 \max_s \|\widehat{\boldsymbol{W}}^{1/2}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2/\xi_n \leq S^2 \max_s \sup_i |\delta_i/\hat{\pi}_i| \cdot \|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2/\xi_n$$
$$\leq C' S^2 \max_s \|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2/\xi_n \to 0. \tag{S3.16}$$

By Lemma 2(i), Assumptions 1 and 3(iii), (iv) and (vi), for any $\kappa > 0$, we only require showing

$$\begin{aligned}
&\Pr\left(S^2 \max_s \|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^2/\xi_n > \kappa\right) \\
\leq &\sum_{s=1}^{S} \Pr\left(S^{4G}\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^{4G}/\xi_n^{2G} > \kappa^{2G}\right) \\
\leq &\frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S} \{E\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^{4G}\} \\
\leq &C' \frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S} \|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\mu}\|^{4G} \\
\leq &C' \frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S} \{\mathbb{E}_{\max}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\}^{4G}\|\boldsymbol{\mu}\|^{4G} \\
\leq &C' C^{4G} \cdot \frac{S^{4G+1}\|\mu\|^{2G}}{\xi_n^{2G}\kappa^{2G}} \cdot \left(\frac{\|\mu\|^2}{n}\right)^G \cdot \left(\frac{p_m^{4/3}}{n}\right)^{3G} \to 0,
\end{aligned}$$

which implies that (S3.15) holds.

To prove (S3.16), it is sufficient to show that for any $\kappa > 0$, we have

$$\sum_{s=1}^{S} \Pr\left\{S^2 |\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2 - E\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2|/\xi_n > \kappa\right\} \to 0, \tag{S3.17}$$

$$S^2 \max_s E\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2/\xi_n \to 0. \tag{S3.18}$$

By Theorem 2 of Whittle (1960) and Lemma 1(iii), it follows from (S3.17)

and Assumptions 3(i), (iii) and (vi) that

$$\begin{aligned}
&\sum_{s=1}^{S} \Pr\left\{S^2|\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2 - E\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2|/\xi_n > \kappa\right\} \\
\leq &\sum_{s=1}^{S} S^{4G} E\left\{\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2 - E\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2\right\}^{2G} /(\xi_n^{2G}\kappa^{2G}) \\
\leq &C' \sum_{s=1}^{S} S^{4G}\left[\mathrm{tr}\{(\boldsymbol{P}_s^* - \boldsymbol{P}_s)^\top(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\}^2\right]^G /(\xi_n^{2G}\kappa^{2G}) \\
\leq &C' C^G \cdot \frac{S^{4G}}{\xi_n^{2G}\kappa^{2G}} \sum_{s=1}^{S} \left(\frac{p_s^4}{n^3}\right)^G \\
\leq &C' C^G \cdot \frac{S^{4G+1}}{\xi_n^{2G}\kappa^{2G}} \cdot \left(\frac{p_m^{4/3}}{n}\right)^{3G} \to 0,
\end{aligned}$$

which implies that (S3.17) holds. Also, it follows from Assumptions 3(iii), (vi) and Lemma 2(ii) that

$$\begin{aligned}
S^2 \max_s E\|(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\boldsymbol{\varepsilon}\|^2/\xi_n =&S^2 \max_s \mathrm{tr}[(\boldsymbol{P}_s^* - \boldsymbol{P}_s)\sigma_{\boldsymbol{\varepsilon}}(\boldsymbol{P}_s^* - \boldsymbol{P}_s)^\top]\xi_n^{-1} \\
\leq &C^2 \mathbb{E}_{\max}(\sigma_{\boldsymbol{\varepsilon}}) \frac{p_s^2 S^2}{n\xi_n} \\
= &C^2 \cdot \mathbb{E}_{\max}(\sigma_{\boldsymbol{\varepsilon}}) \cdot \frac{p_s^2}{n} \cdot \left(\frac{S^{4G}}{\xi_n^{2G}}\right)^{1/2G} \to 0,
\end{aligned}$$

which implies that (S3.18) holds. Combining the proof of Lemma 1(iv) and triangle inequality, we obtain the following decomposition of (S3.7):

$$\begin{aligned}
&|< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\{\boldsymbol{\mu} - \widetilde{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{Y}\} > | \\
\leq &|< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\mu} > | + |< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\mu} > | \\
&+ |< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\varepsilon} > | + O_p\left(Sp_m^2\sqrt{\frac{d_m}{n}}\right).
\end{aligned}$$

Similar to the proof of (S3.10), under Assumptions 3(iii), (iv) and (vi), we delete the term $O_p(Sp_m^2\sqrt{d_m/n})$. Thus, the proof of (S3.7) is equivalent to proving

$$\sup_{\boldsymbol{\omega}\in\mathcal{W}} |< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\mu} > |/R(\boldsymbol{\omega}) \to 0, \tag{S3.19}$$

$$\sup_{\boldsymbol{\omega}\in\mathcal{W}} |< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\mu} > |/R(\boldsymbol{\omega}) \to 0, \tag{S3.20}$$

$$\sup_{\boldsymbol{\omega}\in\mathcal{W}} |< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\varepsilon} > |/R(\boldsymbol{\omega}) \to 0. \tag{S3.21}$$

Using the similar arguments of (S3.14) and (S3.15), by Markov's inequality, under Assumptions 3(i), (vi) and (S3.14), given any $\kappa > 0$, we have

$$\begin{aligned}
&\Pr\left\{\sup_{\boldsymbol{\omega}\in\mathcal{W}} |< \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\varepsilon}, \widehat{\boldsymbol{W}}^{1/2}\boldsymbol{\mu} > |/R(\boldsymbol{\omega}) > \kappa\right\} \\
\leq &\frac{E|\boldsymbol{\varepsilon}^\top\widehat{\boldsymbol{W}}\boldsymbol{\mu}|^{2G}}{\kappa^{2G}\xi_n^{2G}} \leq C'\kappa^{-2G}\frac{\|\boldsymbol{\mu}\|^{2G}}{\xi_n^{2G}} \to 0.
\end{aligned}$$

For (S3.19), under (S3.14), we obtain $|\boldsymbol{\varepsilon}^\top\widehat{\boldsymbol{W}}\boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\mu}| \leq C'|\boldsymbol{\varepsilon}^\top\boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\mu}|$. According to Lemma 2(iv) and Assumption 3(i) and (vi), we just require proving

$$\begin{aligned}
&\Pr\left(\sup_{\boldsymbol{\omega}\in\mathcal{W}} |< \boldsymbol{\varepsilon}, \boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\mu} > |/R(\boldsymbol{\omega}) > \kappa\right) \\
\leq &\Pr\left(S\max_s |< \boldsymbol{\varepsilon}, \boldsymbol{P}_s^*\boldsymbol{\mu} > | > \kappa\xi_n\right) \\
\leq &\sum_{s=1}^{S} \Pr\left(S^{2G}|< \boldsymbol{\varepsilon}, \boldsymbol{P}_s^*\boldsymbol{\mu} > |^{2G} > \kappa^{2G}\xi_n^{2G}\right)
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{S^{2G}}{\xi_n^{2G}\kappa^{2G}} \sum_{s=1}^{S} E|\boldsymbol{\varepsilon}\boldsymbol{P}_s^*\boldsymbol{\mu}|^{2G} \\
&\leq C' \frac{S^{2G}}{\xi_n^{2G}\kappa^{2G}} \sum_{s=1}^{S} \mathbb{E}_{\max}(\boldsymbol{P}_s^*)^{2G}\|\boldsymbol{\mu}\|^{2G} \\
&\leq C' \cdot (1+C)^{2G} \cdot \frac{S^{2G+1}\|\boldsymbol{\mu}\|^{2G}}{\kappa^{2G}\xi_n^{2G}} \to 0.
\end{aligned}
$$

Similarly, by Lemma 2(v), we only require proving

$$
\begin{aligned}
&\Pr\left(\sup_{\boldsymbol{\omega}\in\mathcal{W}} |<\boldsymbol{\varepsilon}, \boldsymbol{P}^*(\boldsymbol{\omega})\boldsymbol{\varepsilon}>|/R(\boldsymbol{\omega}) > \kappa\right) \\
\leq &\Pr\left(S\max_s |<\boldsymbol{\varepsilon}, \boldsymbol{P}_s^*\boldsymbol{\varepsilon}>| > \kappa\xi_n\right) \\
\leq &\frac{S^{2G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S} E|\boldsymbol{\varepsilon}^\top\boldsymbol{P}_s^*\boldsymbol{\varepsilon}|^{2G} \\
\leq &C'\sigma_{\boldsymbol{\varepsilon}}^{2G}\frac{S^{2G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S}\{\mathrm{tr}(\boldsymbol{P}_s^*\boldsymbol{P}_s^{*\top})\}^G \\
\leq &C' \cdot C^G \cdot \sigma_{\boldsymbol{\varepsilon}}^{2G} \cdot \frac{S^{2G+1}n^G}{\kappa^{2G}\xi_n^{2G}} \to 0,
\end{aligned}
$$

where the third inequality holds because of Assumption 1 and Assumption 3(i), the last term converges to zero because of Assumption 3(vi).

To show (S3.8), using the similar arguments of (S3.14) and (S3.15), by Triangle inequality, we have

$$
\begin{aligned}
&|L(\boldsymbol{\omega}) - R(\boldsymbol{\omega})| \\
=&|\{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top\widehat{\boldsymbol{W}}\{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\} - E[\{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}^\top\widehat{\boldsymbol{W}}\{\boldsymbol{\mu} - \hat{\boldsymbol{\mu}}(\boldsymbol{\omega})\}|\boldsymbol{X}]| \\
\leq &\|\widehat{\boldsymbol{W}}^{1/2}\widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}\|^2 + \|\widehat{\boldsymbol{W}}^{1/2}\widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon}\|^2 + 2|<\widehat{\boldsymbol{W}}^{1/2}\widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}, \widehat{\boldsymbol{W}}^{1/2}\widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon}>|
\end{aligned}
$$

$$
\begin{aligned}
&+ E\{\|\widehat{\boldsymbol{W}}^{1/2}\widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}\|^2|\boldsymbol{X}\} + E\{\|\widehat{\boldsymbol{W}}^{1/2}\widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon}\|^2|\boldsymbol{X}\} \\
&+ |E\{2 < \widehat{\boldsymbol{W}}^{1/2}\widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}, \widehat{\boldsymbol{W}}^{1/2}\widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon} > |\boldsymbol{X}\}| \\
\leq & C'\|\widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}\|^2 + C'\|\widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon}\|^2 + 2C'| < \widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}, \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon} > | \\
&+ C'E\{\|\widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}\|^2|\boldsymbol{X}\} + C'E\{\widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon}\|^2|\boldsymbol{X}\} + 2C'|E\{< \widetilde{\boldsymbol{M}}(\boldsymbol{\omega})\boldsymbol{\mu}, \widehat{\boldsymbol{P}}(\boldsymbol{\omega})\boldsymbol{\varepsilon} > |\boldsymbol{X}\}|
\end{aligned}
$$

where $\widetilde{\boldsymbol{M}}(\boldsymbol{\omega}) = \boldsymbol{I} - \widehat{\boldsymbol{P}}(\boldsymbol{\omega})$. From the proof of Lemma 1(iv), by (S3.10) and Assumption 3(vi), if we delete the remainder term $O_p(S^2p_m^4d_m/n)$, the proof of (S3.8) is equivalent to show,

$$
\sup_{\boldsymbol{\omega}\in\mathcal{W}} \Big| \frac{\|\boldsymbol{M}(\boldsymbol{\omega})\boldsymbol{\mu}\|^2}{R(\boldsymbol{\omega})} \Big| \to 0, \tag{S3.22}
$$

$$
\sup_{\boldsymbol{\omega}\in\mathcal{W}} \Big| \frac{\|\boldsymbol{P}(\boldsymbol{\omega})\boldsymbol{\varepsilon}\|^2 - \mathrm{tr}\{\boldsymbol{P}(\boldsymbol{\omega})\sigma_{\boldsymbol{\mathcal{E}}}\boldsymbol{P}^{\top}(\boldsymbol{\omega})\}}{R(\boldsymbol{\omega})} \Big| \to 0, \tag{S3.23}
$$

$$
\sup_{\boldsymbol{\omega}\in\mathcal{W}} \Big| \frac{< \boldsymbol{M}(\boldsymbol{\omega})\boldsymbol{\mu}, \boldsymbol{P}(\boldsymbol{\omega})\boldsymbol{\varepsilon} >}{R(\boldsymbol{\omega})} \Big| \to 0, \tag{S3.24}
$$

where $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\omega})$. Based on Assumption 3(vi), we can obtain (S3.22) using the method given in the proof of (S3.12). Next, we prove (S3.23) under Assumptions 3(i), (iv) and (vi). For any $\kappa > 0$, we have

$$
\begin{aligned}
&\Pr\left\{ \sup_{\boldsymbol{\omega}\in\mathcal{W}} \Big| \frac{\|\boldsymbol{P}(\boldsymbol{\omega})\boldsymbol{\varepsilon}\|^2 - \mathrm{tr}\left\{\boldsymbol{P}(\boldsymbol{\omega})\sigma_{\boldsymbol{\mathcal{E}}}\boldsymbol{P}^{\top}(\boldsymbol{\omega})\right\}}{R(\boldsymbol{\omega})} \Big| > \kappa \right\} \\
\leq & \Pr\left\{ \sup_{\boldsymbol{\omega}\in\mathcal{W}} \sum_{s=1}^{S}\sum_{k=1}^{S} \omega_s\omega_k |\boldsymbol{\varepsilon}^{\top}\boldsymbol{P}_s\boldsymbol{P}_k\boldsymbol{\varepsilon} - \sigma_{\boldsymbol{\mathcal{E}}}\mathrm{tr}(\boldsymbol{P}_s\boldsymbol{P}_k)| > \kappa\xi_n \right\} \\
\leq & \Pr\left\{ S^2 \max_s \max_k |\boldsymbol{\varepsilon}^{\top}\boldsymbol{P}_s\boldsymbol{P}_k\boldsymbol{\varepsilon} - \sigma_{\boldsymbol{\mathcal{E}}}\mathrm{tr}(\boldsymbol{P}_s\boldsymbol{P}_k)| > \kappa\xi_n \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\leq \sum_{s=1}^{S}\sum_{k=1}^{S} \Pr\left\{|\boldsymbol{\varepsilon}^\top \boldsymbol{P}_s \boldsymbol{P}_k \boldsymbol{\varepsilon} - \sigma_{\boldsymbol{\varepsilon}} \mathrm{tr} \boldsymbol{P}_s \boldsymbol{P}_k| > \kappa \xi_n / S^2\right\} \\
&\leq \sum_{s=1}^{S}\sum_{k=1}^{S} \frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} E|\boldsymbol{\varepsilon}^\top \boldsymbol{P}_s \boldsymbol{P}_k \boldsymbol{\varepsilon} - \sigma_{\boldsymbol{\varepsilon}} tr \boldsymbol{P}_s \boldsymbol{P}_k|^{2G} \\
&\leq C' \frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S}\sum_{k=1}^{S} \{\mathrm{tr}(\boldsymbol{P}_s^2 \boldsymbol{P}_k^2)\}^G \\
&\leq C' \cdot \frac{S^{4G+2} n^G}{\xi_n^{2G}} \cdot \kappa^{-2G} \to 0,
\end{aligned}
$$

where the last inequality holds since $\boldsymbol{P}(\boldsymbol{\omega})$ is the idempotent matrix. Similarly, $\boldsymbol{I} - \boldsymbol{P}(\boldsymbol{\omega})$ is also the idempotent matrix. Then, under Assumptions 3(i) and (vi), for any $\kappa > 0$, we have

$$
\begin{aligned}
&\Pr\left\{\sup_{\boldsymbol{\omega}\in\mathcal{W}} \left| \frac{< \boldsymbol{M}(\boldsymbol{\omega})\boldsymbol{\mu}, \boldsymbol{P}(\boldsymbol{\omega})\boldsymbol{\varepsilon} >}{R(\boldsymbol{\omega})} \right| > \kappa \right\} \\
\leq &\Pr\left\{\sup_{\boldsymbol{\omega}\in\mathcal{W}} \sum_{s=1}^{S}\sum_{k=1}^{S} |\boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{P}_s)\boldsymbol{P}_k \boldsymbol{\varepsilon}| > \kappa \xi_n \right\} \\
\leq &\frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S}\sum_{k=1}^{S} E|\boldsymbol{\mu}^\top (\boldsymbol{I} - \boldsymbol{P}_s)\boldsymbol{P}_k \boldsymbol{\varepsilon}|^{2G} \\
\leq &C' \frac{S^{4G}}{\kappa^{2G}\xi_n^{2G}} \sum_{s=1}^{S}\sum_{k=1}^{S} \|\boldsymbol{P}_s (\boldsymbol{I} - \boldsymbol{P}_k)\boldsymbol{\mu}\|^{2G} \\
\leq &C' \cdot \frac{S^{4G+2} \|\boldsymbol{\mu}\|^{2G}}{\xi_n^{2G}} \cdot \kappa^{-2G} \to 0,
\end{aligned}
$$

which indicates that we have proved (S3.21). Thus, we finish the proof of Theorem 1.

Proof of Theorem S2.1. We first show the sure screening property. For

each $k = 1, \ldots, p$, define

$$m(x) = E(X_k Y | X_k = x)$$

and

$$\hat{m}(x) = \frac{\sum_{j=1}^{n} \delta_j K_h(x - X_{jk}) Y_j}{\sum_{j=1}^{n} \delta_j K_h(x - X_{jk})}.$$

By the definition of $\hat{r}_k$ and r_k, we have

$$\begin{aligned}
\hat{r}_k - r_k =& \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i X_{ik} Y_i + (1 - \delta_i) \frac{1}{m} \sum_{v=1}^{m} X_{ik} \tilde{Y}_{iv}^k \right\} - E(X_k Y) \\
=& \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \{ \hat{m}(X_{ik}) - m(X_{ik}) \} + \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \Big\{ \frac{1}{m} \sum_{v=1}^{m} X_{ik} \tilde{Y}_{iv}^k - \hat{m}(X_{ik}) \Big\} \\
& + \frac{1}{n} \sum_{i=1}^{n} \delta_i \{ X_{ik} Y_i - m(X_{ik}) \} + \frac{1}{n} \sum_{i=1}^{n} \{ m(X_{ik}) - E(X_k Y) \} \\
=& J_{k1} + J_{k2} + J_{k3} + J_{k4}.
\end{aligned}$$

For J_{k2}, Wang and Chen (2009) have proved $J_{k3} = o_p(1/\sqrt{n})$ as n and $m \to \infty$. Thus, as n is sufficiently large, for any $\varsigma \in (0, 1/2)$ and $c_2 > 0$, we have

$$\begin{aligned}
\Pr\{ |\hat{r}_k - r_k| \geq c_2 n^{-\varsigma} \} &= \Pr(|J_{k1} + J_{k2} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma}) \\
&\leq \Pr(|J_{k1} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma} - |J_{k2}|) \\
&\leq \Pr(|J_{k1} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma} / 2).
\end{aligned}$$

Through defining $\eta(x) = \pi(x) f_k(x)$ and $\hat{\eta}(x) = \sum_{j=1}^{n} \delta_j K_h(X_{ik} - x)/n$, where $\pi(\cdot)$ is the selection probability function and $f_k(\cdot)$ is the probability density function of X_k. Thus we can decompose J_{k1} as

$$
\begin{aligned}
J_{k1} &= \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}(X_{ik}) - m(X_{ik})\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\sum_{j=1}^{n}\delta_j K_h(X_{jk}-X_{ik})\{X_{ik}Y_i - m(X_{jk})\}/n}{\eta(X_{ik})} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\{\hat{m}(X_{ik}) - m(X_{ik})\}\frac{\eta(X_{ik}) - \hat{\eta}(X_{ik})}{\eta(X_{ik})} \\
&\quad + \frac{1}{n}\sum_{i=1}^{n}(1-\delta_i)\frac{\sum_{j=1}^{n}\delta_j K_h(X_{jk}-X_{ik})\{m(X_{jk}) - m(X_{ik})\}/n}{\eta(X_{ik})} \\
&= J_{k11} + J_{k12} + J_{k13} \\
&= \tilde{J}_{k11} + (J_{k11} - \tilde{J}_{k11}) + J_{k12} + J_{k13},
\end{aligned}
$$

where $\tilde{J}_{k11} = (1/n)\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - m(X_{ik})\}\{1-\pi(X_{ik})\}/\pi(X_{ik})$. Under the Assumptions 1, 8, and 9, by the similar certification of Wang and Chen (2009), $J_{k11} - \tilde{J}_{k11} = o_p(1/\sqrt{n})$, $J_{k12} = o_p(1/\sqrt{n})$ and $J_{k13} = o_p(1/\sqrt{n})$. Then we can write

$$
\begin{aligned}
&\Pr(|J_{k1} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma}/2) \\
\leq &\Pr(|\tilde{J}_{k11} + (J_{k11} - \tilde{J}_{k11}) + J_{k12} + J_{k13} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma}/2) \\
\leq &\Pr(|\tilde{J}_{k11} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma}/2 - |J_{k11} - \tilde{J}_{k11}| - |J_{k12}| - |J_{k13}|) \\
\leq &\Pr(|\tilde{J}_{k11} + J_{k3} + J_{k4}| \geq c_2 n^{-\varsigma}/16).
\end{aligned}
$$

Note that

$$
\begin{aligned}
&\tilde{J}_{k11} + J_{k3} + J_{k4} \\
=&(1/n)\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - m(X_{ik})\}\{1-\pi(X_{ik})\}/\pi(X_{ik}) \\
&+\frac{1}{n}\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - m(X_{ik})\} + \frac{1}{n}\sum_{i=1}^{n}\{m(X_{ik}) - E(X_kY)\} \\
=&\frac{1}{n}\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - E(X_kY)\}\frac{1}{\pi(X_{ik})} \\
&+\frac{1}{n}\sum_{i=1}^{n}\frac{\{\pi(X_{ik})-\delta_i\}}{\pi(X_{ik})}[m(X_{ik}) - E\{m(X_{ik})\}] \\
=&I_1 + I_2.
\end{aligned}
$$

Under Assumption 1, for any $M > 0$, we have

$$
\begin{aligned}
\Pr(|I_1| \geq c_2 n^{-\varsigma}/32) =&\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - E(X_kY)\}\frac{1}{\pi(X_{ik})}\right| \geq c_2 n^{-\varsigma}/32\right\} \\
\leq&\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - E(X_kY)\}\right| \geq C_0 c_2 n^{-\varsigma}/32\right\} \\
\leq&\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - E(X_kY)\}\right| \geq C_0 c_2 n^{-\varsigma}/32, \quad \max_i|\delta_i X_{ik}Y_i| < M\right\} \\
&+\Pr(\max_i|X_{ik}Y_i| \geq M).
\end{aligned}
$$

According to Assumption 10 and Lemma S3 of Liu et al. (2014), there are some positive constants c_3 and c_4 such that for any $M > 0$, we have $\Pr(|X_kY| \geq M) \leq c_3\exp(-c_4M)$. Thus, by taking $M = c_3^{-1}n^{(1-2\varsigma)/3}$, applying the Hoeffdings inequality in Lemma 3 and yields that there exists

a positive constant c_5 such that

$$\begin{aligned}
&\Pr\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\delta_i\{X_{ik}Y_i - E(X_kY)\}\right| \geq C_0c_2n^{-\varsigma}/32, \quad \max_i|\delta_iX_{ik}Y_i| < M\right\} \\
&\quad + \Pr(\max_i|X_{ik}Y_i| \geq M) \\
\leq & 2\exp(-n^{1-2\varsigma}/M^2) + \sum_{i=1}^{n}\Pr(|X_{ik}Y_i| \geq M) \\
\leq & O(n)\exp\{-c_5n^{(1-2\varsigma)/3}\}.
\end{aligned}$$

According to the Assumptions 1 and 10, the above argument can be used to I_2, we have $\Pr(|I_2| \geq c_2n^{-\varsigma}/32) \leq O(n)\exp\{-c_6n^{(1-2\varsigma)/3}\}$ for some constant c_6. Thus, there exists a constant c_7 such that

$$\begin{aligned}
&\Pr\left(\max_{1\leq k\leq p}|\hat{r}_k - r_k| \geq c_2n^{-\varsigma}\right) \\
\leq & p\Pr(|I_1| \geq c_2n^{-\varsigma}/32) + p\Pr(|I_2| \geq c_2n^{-\varsigma}/32) \\
\leq & O(n)p\exp\{-c_7n^{(1-2\varsigma)/3}\} = O\{p\exp(-c_7n^{(1-2\varsigma)/3} + \log(n))\}.
\end{aligned}$$

In fact, by $S_n = \{\max_{k\in\mathcal{M}_*}|\hat{r}_k - r_k| \leq c_0n^{-\varsigma}/2\}$ and Assumption 11, we have

$$\min_{k\in\mathcal{M}_*}|\hat{r}_k| \geq \min_{k\in\mathcal{M}_*}(|r_k| - |\hat{r}_k - r_k|) \geq \min_{k\in\mathcal{M}_*}|r_k| - \max_{k\in\mathcal{M}_*}|\hat{r}_k - r_k| \geq c_0n^{-\varsigma}/2.$$

Thus, by taking $\varrho_n = c_8n^{-\varsigma}$ with $c_8 \leq c_0/2$, there exists some positive constant c_9 such that

$$\Pr(\mathcal{M}_* \subset \widehat{\mathcal{M}}) \geq \Pr(S_n) \geq 1 - O\{|\mathcal{M}_*|\exp(-c_9n^{(1-2\varsigma)/3} + \log(n))\}.$$

Now we show the ranking consistency property. It is easily shown that

$$\begin{aligned}
& \Pr\left\{\left(\min_{k\in\mathcal{M}_*}|\hat{r}_k| - \max_{k\in\mathcal{I}_*}|\hat{r}_k|\right) < m_0/2\right\} \\
\leq\ & \Pr\left\{\left(\min_{k\in\mathcal{M}_*}|\hat{r}_k| - \max_{k\in\mathcal{I}_*}|\hat{r}_k|\right) - \left(\min_{k\in\mathcal{M}_*}|r_k| - \max_{k\in\mathcal{I}_*}|r_k|\right) < -m_0/2\right\} \\
\leq\ & \Pr\left\{\left|\left(\min_{k\in\mathcal{M}_*}|\hat{r}_k| - \max_{k\in\mathcal{I}_*}\left|\hat{r}_k\right|\right) - \left(\min_{k\in\mathcal{M}_*}|r_k| - \max_{k\in\mathcal{I}_*}|r_k|\right)\right| \geq m_0/2\right\} \\
\leq\ & \Pr\left(2\max_{1\leq k\leq p}|\hat{r}_k - r_k| \geq m_0/2\right) \\
\leq\ & p\Pr(|\hat{r}_k - r_k| \geq m_0/4) \\
\leq\ & O(n)p\exp(-c_{10}n^{1/3}m_0^{2/3})
\end{aligned}$$

for some constants c_{10}, where the first inequality holds because of Assumption 12.

Note that $\log(n) = o(n^{1/3}m_0^{2/3})$ and $\log(p) = o(n^{1/3}m_0^{2/3})$ imply that $p \leq \exp(c_{10}n^{1/3}m_0^{2/3}/2)$, $c_{10}n^{1/3}m_0^{2/3}/2 \geq 4\log(n)$ for sufficiently large n. Thus, for some n_0, we have

$$\begin{aligned}
\sum_{n=n_0}^{\infty} pn\exp(-c_{10}n^{1/3}m_0^{2/3}) &\leq 2\sum_{n=n_0}^{\infty}\exp\{c_{10}n^{1/3}m_0^{2/3}/2 - c_{10}n^{1/3}m_0^{2/3} + \log(n)\} \\
&\leq 2\sum_{n=n_0}^{\infty}\exp\{-3\log(n)\} \leq 2\sum_{n=n_0}^{\infty} n^{-3} < +\infty.
\end{aligned}$$

Hence, by Borel Contelli Lemma, we have

$$\liminf_{n\to\infty}\left\{\min_{k\in\mathcal{M}_*}|\hat{r}_k| - \max_{k\in\mathcal{I}_*}|\hat{r}_k|\right\} > 0 \quad a.s.$$

References

Cui, H., Li, R. and Zhong, W. (2015). Model-free feature screening for ultrahigh dimensional discriminant analysis. *Journal of the American Statistical Association* **110**, 630-641.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society, Series B* **70**, 849-911.

Fan, J. and Lv, J. (2011). Nonconcave penalized likelihood with NP-dimensionality. *IEEE Transactions on Information Theory* **57**, 5467-5484.

Fan, J. and Peng, H. (2004). Nonconcave penalized likelihood with a diverging number of parameters. *The Annals of Statistics* **32**, 928-961.

Hirano, K., Imbens, G. and Ridder, G. (2003). Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica* **71**, 1161-1189.

Lewis, R. and Reinsel, G. C. (1985). Prediction of multivariate time series by autoregressive model fitting. *Journal of Multivariate Analysis* **16**, 393-411.

Li, R., Zhong, W. and Zhu, L. (2012). Feature screening via distance correlation learning. *Journal of the American Statistical Association* **107**, 1129-1139.

Liu, J., Li, R. and Wu, R. (2014). Feature selection for varying coefficient models with ultrahigh dimensional covariates. *Journal of the American Statistical Association* **109**, 266-274.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *The Annals of Statistics* **37**, 3498-3528.

Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory of Probability and its Applications* **5**, 302-305.

Wiener, N. and Masani, P. (1958). The prediction theory of multivariate stochastic processes,

II. The linear predictor. *Acta Mathematica* **99**, 93-137.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research* **7**, 2541-2567.

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