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

Jinhan Xie\textsuperscript{a} Xiaodong Yan\textsuperscript{b} and Niansheng Tang\textsuperscript{a}

\textsuperscript{a}Yunnan Key Laboratory of Statistical Modeling and Data Analysis, Yunnan University
\textsuperscript{b}School of Economics, Shandong University

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

In this section, we investigate the consistency and oracle property of the penalized likelihood estimator $\hat{\gamma}$ of $\gamma$. Without loss of generality, we can write $\gamma = (\gamma_1^T, \gamma_2^T)^T$, where $\gamma_1 \in \mathbb{R}^{d_m}$ and $\gamma_2 \in \mathbb{R}^{q-d_m}$ correspond to the nonzero and zero components of $\gamma$, respectively. Thus, the true parameter vector $\gamma_0$ of $\gamma$ can be written as $\gamma_0 = (\gamma_{10}^T, 0^T)^T$, where $\gamma_{10}$ is the true value of $\gamma_1$. Also, the corresponding penalized likelihood estimator $\hat{\gamma}$ of $\gamma$ can be written as $\hat{\gamma} = (\hat{\gamma}_1^T, \hat{\gamma}_2^T)^T$. Let $\mathcal{A}_\gamma = \{j : \gamma_{0j} \neq 0\}$ be the index set of nonzero components of $\gamma_0$, where $\gamma_{0j}$ is the $j$th component of $\gamma_0$ for $j = 1, \ldots, q$. Denote the cardinality of $\mathcal{A}_\gamma$ as $d_m = |\mathcal{A}_\gamma|$, 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 local concavity of penalty function $f_{\lambda_n}(t)$ at \( v = (v_1, \ldots, v_q)^\top \in \mathbb{R}^q \) (i.e., \( \|v\|_0 = q \)) as

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

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

We use \( \text{tr}(A) \) to represent the trace of matrix \( A \). Denote

\[
g_i(\gamma) = \log[\pi(U_{i}; \gamma)/\{1 - \pi(U_{i}; \gamma)\}]
\]

for \( i = 1, \ldots, n \), and define Fisher information matrix as

\[
F_n(\gamma) = E\{-\partial^2 l_n(\gamma)/\partial \gamma \partial \gamma^\top\} = \partial_\gamma^\top g(\gamma) \Sigma(\gamma) \partial_\gamma g(\gamma),
\]

where \( g(\gamma) = (g_1(\gamma), \ldots, g_n(\gamma))^\top \), and \( \Sigma(\gamma) = \text{diag}(\pi_1(1 - \pi_1), \ldots, \pi_n(1 - \pi_n)) \) with \( \pi_i = \pi_i(\gamma) = \pi(U_{i}; \gamma) \) for \( i = 1, \ldots, n \). Let \( s_n = \frac{1}{2} \min_j \{|\gamma_{0j}|: \gamma_{0j} \neq 0\} \), and define \( \mathcal{N} = \{\tau = (\tau_1^\top, \tau_2^\top)^\top \in \mathbb{R}^q: \tau_2 = 0, \|\tau_1 - \gamma_{10}\|_\infty \leq s_n\} \). The following assumptions are required to ensure the consistency of \( \hat{\gamma} \).

**Assumption S1.** 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 some 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 \( \lambda_n \).

Assumption S1 holds for a class of penalty functions such as the SCAD and MCP penalty functions. By Assumption S1, \( f_{\lambda_n}(t) \) is a concave function with
Assumption S2. (i) \( \min_{\tau \in \mathcal{N}} \mathbb{E}_{\min} \{ F_n(\tau) \} \geq cn \) and \( \text{tr}\{ F_n(\gamma_0) \} = O(d_m n) \); 
(ii) \( \left\| \partial_{\gamma_1} g(\gamma_0) \Sigma(\gamma_0) \partial_{\gamma_1} g(\gamma_0) \right\|_{2,\infty} = O(n) \), where notation \( \|B\|_{2,\infty} \) represents \( \max\|v\|_2 = 1 \|Bv\|_\infty \); 
(iii) \( \max_{\tau \in \mathcal{N}, 1 \leq j \leq q} \mathbb{E}_{\max} \left[ \partial_{\gamma_1}^2 g(\tau) \text{diag}\{ |\partial_{\gamma_j} g(\tau)| \circ |\partial_{\gamma_1}^2 V(\tau)| \} \partial_{\gamma_1} g(\tau) \right] = O(n) \), where \( V(\gamma) = (\pi_1(\gamma), \ldots, \pi_n(\gamma))^\top \).

Assumption S2(i) has been used in Fan and Lv (2011), and ensures that information matrix \( F_n(\tau) \) is positive definite, and its eigenvalues are uniformly bounded. Assumption S2(ii) measures the correlation between each unimportant variable and important variable using the weighted matrix \( \Sigma(\gamma_0) \), 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 S2(iii) is used to control the order of the remainder term when taking the third-order expansion of the objective function.

Assumption S3. Suppose that \( s_n \gg \lambda_n \gg \sqrt{d_m/n} \), \( \sqrt{n} \partial f_{\lambda_n}(s_n) = O(1) \) and \( \rho_0 = o(1) \), where \( \rho_0 = \max_{\tau \in \mathcal{N}} \rho(\lambda_n; \tau) \).

Assumption S3 shows that the minimum signal \( s_n \) should be satisfied, and is used to obtain nice properties of the proposed penalized likelihood estimator (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 S3 is less restrictive for the SCAD penalty function.

**Theorem S1.1.** Suppose that Assumptions S1–S3 hold. There is a strict local maximizer $\hat{\gamma} = (\hat{\gamma}_1, \hat{\gamma}_2)^\top$ of the nonconcave penalized likelihood $Q_n(\gamma)$ with respect to $\gamma$ such that $\hat{\gamma}_2 = 0$ with probability tending to 1 as $n \to \infty$ and $\|\hat{\gamma}_1 - \gamma_0\|_2 = O_p(\sqrt{d_m/n})$.

Theorem S1.1 shows that the sparsity property of the above proposed PLE still holds in a high-dimensional parametric model. To wit, zero components in $\gamma_0$ are estimated as zero with probability tending to one. Also, Theorem S1.1 establishes the consistency of the above proposed PLE $\hat{\gamma}_1$ of $\gamma_1$, i.e., there is a root-$(n/d_m)$-consistent PLE of $\gamma_1$.

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

**Assumption S4.** Suppose that $\partial f_{\lambda_n}(s_n) = o(1/\sqrt{nd_m})$, max $E|\delta_i - \pi_i(\gamma_0)|^3 = O(1)$, and $\sum_{i=1}^{n} \{\partial_{\gamma_i} g_i(\gamma_0) F_{n11}(\gamma_0) \partial_{\gamma_i} g_i(\gamma_0) \}^{3/2} \to 0$ as $n \to \infty$, where $F_{n11}(\gamma_0) = \partial_{\gamma_i} g(\gamma_0) \Sigma(\gamma_0) \partial_{\gamma_i} g(\gamma_0)$.
Theorem S1.2. Suppose that Assumptions S1-S4 hold and $d_m = o(n^{1/4})$. Then, we have

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

(ii) (Asymptotic Normality) $U_n F_{n11}^{1/2} (\hat{\gamma}_1 - \gamma_{10}) \overset{\mathcal{L}}{\to} \mathcal{N}(0, G)$, where $U_n$ is a $m \times d$ matrix such that $U_n U_n^\top \to G$, $G$ is a $m \times m$ symmetric positive definite matrix with the fixed $m$, and $\overset{\mathcal{L}}{\to}$ represents convergence in distribution.

Theorem S1.2 indicates that the sparsity and asymptotic normality of the above 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 $E(X_kY)$ in the presence of missing responses. Without loss generality, for each $k = 1, \ldots, p$, the $k$th column of covariates satisfies $E(X_k) = 0$, $E(X_k^2) = 1$. Then we have $E(X_kY) = cov(X_k, Y)$ and $\beta_k = cov(X_k, Y)$, which indicates that $\beta_k$ 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 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 $r_k = E(X_kY)$. To investigate the sure screening properties of the above presented screening criterion, we require the following assumptions.

**Assumption S5.** 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 S6.** 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 S7.** Variables $X_k$, $Y$ and $X_kY$ satisfy the sub-exponential tail probability uniformly in $p$. That is, there exists a positive constant $u_0$ such that for all $0 < \tilde{u} \leq 2u_0$, $\max_{1 \leq k \leq p} E\{\exp(2\tilde{u}X_k^2)\} < \infty$, $E\{\exp(2\tilde{u}Y^2)\} < \infty$, $\max_{1 \leq k \leq p} E\{\exp(2\tilde{u}X_kY)\} < \infty$.

**Assumption S8.** 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 S9.** $\lim_{p \to \infty} \inf \{\min_{k \in \mathcal{M}_*} |r_k| - \max_{k \not\in \mathcal{M}_*} |r_k|\} \geq m_0$ for some $m_0 > 0$.

Assumptions S5 and S6 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. Also, the assumption that $\sqrt{n}h^2 \to 0$ is to control the bias induced by the kernel smoothing. Assumption S7 has widely used in high dimensional data analysis (Fan and Lv, 2008; Li et al., 2012) and holds if $X, Y$ and $X_kY$ are bounded uniformly of $X, Y$ and $X_kY$ have multivariate normal distribution. Assumption S8 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 S9 ensures that the active and inactive predictors can be well separated in the population level. Assumption S9 is similar to Condition (C3) of Cui et al. (2015).

**Theorem S2.1. (Sure Screening Property)** Under Assumptions S1 and S4–S6, then for any constant $c_2 > 0$, there exists $c_7 > 0$ such that

$$
\Pr\left(\max_{1 \leq k \leq p} |\hat{r}_k - r_k| \geq c_2n^{-\varsigma}\right) \leq O\{p\exp(-c_7n^{(1-2\varsigma)/3} + \log(n))\}
$$

for sufficiently large $n$. Furthermore, under Assumption S8, by taking $\varrho_n = c_8n^{-\varsigma}$ with $c_8 \leq c_0/2$, there exists some positive constant $c_9$ such that

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

(Ranking Consistency Property) If Assumptions S1, S4–S9 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

$$
\lim_{n \to \infty} \inf \left\{ \min_{k \in \mathcal{M}_*} |\hat{r}_k| - \max_{k \in \mathcal{I}_*} |\hat{r}_k| \right\} > 0, \text{a.s.}
$$
Theorem S2.1 shows the sure screening and rank consistency properties of our proposed screening procedure, which indicates that our 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 \( \varrho_n \), which is smaller than the minimum signal.

S3. Proofs of all theorems

Let \( P_s^* = D_s(P_s - I) + I \), where \( P_s = X_s(X_s^\top X_s)^{-1}X_s \), \( D_s = \text{diag}(d_s^1, \ldots, d_s^n) \), \( d_s^i = 1/(1-h_{si}^s) \) and \( h_{si}^s \) is the \( i \)th diagonal element of \( P_s = X_s(X_s^\top X_s)^{-1}X_s^\top \). Let \( \|A\| \) be the Frobenius norm (e.g., \( \|A\| = \sqrt{\text{tr}(A^\top A)} \)), and \( \|A\|_2 \) be the Euclidean norm (e.g., \( \|A\|_2 = \sqrt{\text{E}_{\max}(A^\top A)} \)) of matrix \( A \), where \( A^* \) represents the conjugate transpose of matrix \( A \). Denote \( P^*(\omega) = \sum_{s=1}^S \omega_s P_s^* \) and \( P(\omega) = \sum_{s=1}^S \omega_s P_s \).

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

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

where \( \tau = (\tau_1, \ldots, \tau_{d_m})^\top \) and \( \tilde{l}_n(\tau) = \sum_{i=1}^n [\delta_i \log \{\pi_i(U_i^\tau; \tau)\} + (1 - \delta_i) \log \{1 - \pi(U_i^\tau; \tau)\}] \) in which \( U_i^\tau \) is the subvector of \( U_i \) corresponding to \( \tau \).
Let $\alpha_n = \sqrt{d_m/n}$, and define the closed set $\mathcal{N}_0 = \{ \tau \in \mathbb{R}^{d_m} : \| \tau - \gamma_{10} \|_2 \leq \alpha_n u \}$ for $u \in (0, \infty)$. Here, our purpose is to show that, for any $\kappa > 0$ and a sufficiently large $n$, we have

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

where $\partial \mathcal{N}_0$ denotes the boundary of the closed set $\mathcal{N}_0$, which implies that there exists a local maximizer $\hat{\gamma}_1$ in $\mathcal{N}_0$ such that $\| \hat{\gamma}_1 - \gamma_{10} \|_2 = O_p(\sqrt{d_m/n})$.

For a sufficiently large $n$, it follows from Assumption S3 that $\alpha_n u \leq d_m$. Taking Taylor expansion of $\tilde{Q}_n(\tau)$ at the true value $\gamma_{10}$ of $\gamma_1$ yields

$$\tilde{Q}_n(\tau) - \tilde{Q}_n(\gamma_{10}) = (\tau - \gamma_{10})^\top D_1 - \frac{1}{2}(\tau - \gamma_{10})^\top D_2 (\tau - \gamma_{10})$$

for any $\tau \in \mathcal{N}_0$, where $\mathbf{D}_1 = \partial \tilde{Q}_n(\gamma_{10})/\partial \gamma_1 = \partial \tilde{\gamma}_1 g(\gamma_{10})\{ \delta - V(\gamma_{10}) \}/n - \partial f_{\lambda_n}(\gamma_{10})$, $\mathbf{D}_2 = \partial^2 \tilde{Q}_n(\tilde{\gamma}_1)/\partial \gamma_1 \partial \gamma_1^\top$, and $\tilde{\gamma}_1$ lies on the line segment jointing $\tau$ and $\gamma_{10}$. Following the argument of Fan and Peng (2004), we can obtain

$$\mathbf{D}_2 = -\frac{1}{n} \partial^2 \tilde{I}_n(\tilde{\gamma}_1)/\partial \gamma_1 \partial \gamma_1^\top + \text{diag}\{ \partial^2 f_{\lambda_n}(\tilde{\gamma}_1) \}$$

$$= \frac{1}{n} \mathbf{F}_n(\tilde{\gamma}_1) - \frac{1}{n} \left\{ \partial^2 \tilde{I}_n(\tilde{\gamma}_1)/\partial \gamma_1 \partial \gamma_1^\top - E(\partial^2 \tilde{I}_n(\tilde{\gamma}_1)/\partial \gamma_1 \partial \gamma_1^\top) \right\} + \text{diag}\{ \partial^2 f_{\lambda_n}(\tilde{\gamma}_1) \}$$

$$= \frac{1}{n} \mathbf{F}_n(\tilde{\gamma}_1) + \text{diag}\{ \partial^2 f_{\lambda_n}(\tilde{\gamma}_1) \} + o_p(1).$$

Without loss of generality, when there is not the second-order derivative of the penalty function $f_{\lambda_n}(\cdot)$, it is easily shown that matrix $\mathbf{D}_2$ can be replaced by a diagonal matrix whose maximum absolute element is bounded by $\rho_0$. Thus, for a sufficiently large $n$, under Assumptions S1(i) and S6, we have $\mathbb{E}_{\min}(\mathbf{D}_2) \geq c - \rho_0 \geq \cdots$
where $c$ is some constant.

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

$$
Pr\left\{ \sup_{\tau \in \partial N_0} \tilde{Q}_n(\tau) < \tilde{Q}_n(\gamma_{10}) \right\} \geq Pr\left\{ u \alpha_n \left( \|D_1\|_2 - \frac{cu \alpha_n}{4} \right) < 0 \right\} \geq 1 - \frac{16E\|D_1\|_2^2}{c^2u^2 \alpha_n^2}.
$$

However, under Assumptions S5(i) and S6, we have

$$
E\|D_1\|_2^2 \leq \frac{1}{n^2} \text{tr}\{F_n(\gamma_{10})\} + d_m \partial f_{\lambda_n}(d_m)^2 = O(\alpha_n^2),
$$

which leads to

$$
Pr\left\{ \sup_{\tau \in \partial N_0} \tilde{Q}_n(\tau) < \tilde{Q}_n(\gamma_{10}) \right\} \geq 1 - O\left( \frac{16}{c^2u^2} \right).
$$

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

Next, to show the sparsity of the proposed estimator, it is necessary to prove that $\hat{\gamma} \in R^q$ is a strict local maximizer of $Q_n(\gamma)$ such that $\hat{\gamma}_{\neq,\gamma} = \hat{\gamma}_1 \in N_0 \subset N$ and $\hat{\gamma}_{\neq,\gamma} = \hat{\gamma}_2 = 0$. Similar to the proof of Theorem 1 of Fan and Lv (2011), we only require showing

$$
\left\| \frac{1}{n\lambda_n} \partial \gamma_2 l_n(\gamma) \right\|_\infty \leq \partial f_{\lambda_n}(0+). \quad (S3.3)
$$

Thus, we have

$$
\frac{1}{n} \partial \gamma_2 l_n(\gamma) = \frac{1}{n} \partial^T \gamma_2 g(\hat{\gamma}) \{ \delta - V(\hat{\gamma}) \}
= \frac{1}{n} \left[ \eta_{\neq}\hat{\gamma} + \{ \partial^T \gamma_2 g(\hat{\gamma}) \delta - \partial^T \gamma_2 g(\hat{\gamma}) V(\hat{\gamma}) \} \right. 
\left. - \{ \partial^T \gamma_2 g(\gamma_0) \delta - \partial^T \gamma_2 g(\gamma_0) V(\gamma_0) \} \right],
$$

where $\eta = (\eta_1, \ldots, \eta_q)^T = \partial^T \gamma g(\gamma_0) \{ \delta - V(\gamma_0) \}$. 

---

\[10\]
Denote $\mathcal{B}_\gamma = \{\|\eta_{\delta\gamma}\|_\infty \leq c_1 \sqrt{n}\}$, where $c_1$ is some constant. Under Assumption S1, it is easily known that the first- to third-order derivatives of $g_i(\gamma)$ with respect to $\gamma$ 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

$$
\Pr(\mathcal{B}_\gamma) = \Pr(\|\eta_{\delta\gamma}\|_\infty \leq c_1 \sqrt{n}) = 1 - \Pr(\|\eta_{\delta\gamma}\|_\infty > c_1 \sqrt{n}) \\
\geq 1 - 2(p - d) \exp(-c'n) \geq 1 - 2q \exp(-c'n) \to 1.
$$

Under Assumptions S1(ii), S5(iii) and S2, we consider Taylor expansion of $\partial_{\gamma^2} g(\hat{\gamma})\delta - \partial_{\gamma^2} g(\hat{\gamma})V(\hat{\gamma})$ at $\gamma_{10}$. Thus, we obtain

$$
\left\| \frac{1}{n\lambda_n} \partial_{\gamma^2} l_n(\hat{\gamma}) \right\|_\infty \leq \frac{1}{n\lambda_n} \left[ \|\eta_{\delta\gamma}\|_\infty \right. + \left. \left\| \left\{ \partial_{\gamma^2} g(\hat{\gamma})\delta - \partial_{\gamma^2} g(\hat{\gamma})V(\hat{\gamma}) \right\} \|ight]
$$

$$
= \frac{1}{n\lambda_n} \left\{ O(n)\|\hat{\gamma}_1 - \gamma_{10}\|_2 + O(n)\|\hat{\gamma}_1 - \gamma_{10}\|_2^2 \right\}
$$

$$
= o_p(1) + O_p(\lambda_n^{-1} \sqrt{d_m/n} ) = o_p(1).
$$

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 $\hat{\gamma}_1$. For the set $\mathcal{N}$, it is easily shown that $\hat{\gamma} = (\hat{\gamma}_1^T, 0^T)^T \in \mathcal{N}$ is a strict local maximizer of $\tilde{Q}_n(\tau)$, which leads to $\partial_{\gamma_1} \tilde{Q}_n(\hat{\gamma}) = 0$. 
Taking Taylor expansion of $\partial \gamma_1 \tilde{l}_n(\hat{\gamma})$ at $\gamma_{10}$ leads to

$$0 = \partial_1 \tilde{Q}_n(\hat{\gamma}) = \frac{1}{n} \partial_1 \tilde{l}_n(\hat{\gamma}) - \partial f_{\lambda_n}(\hat{\gamma})$$

$$= \frac{1}{n} \partial_1^\top g(\gamma_0) \{ \delta - V(\gamma_0) \} - \frac{1}{n} \left[ F_{n11}(\gamma_0) - \left\{ \frac{\partial^2 l_n(\gamma_0)}{\partial \gamma_1 \partial \gamma_1^\top} - E \left( \frac{\partial^2 l_n(\gamma_0)}{\partial \gamma_1 \partial \gamma_1^\top} \right) \right\} \right] (\hat{\gamma}_1 - \gamma_{10})$$

$$+ O(1) \| \hat{\gamma}_1 - \gamma_{10} \|_{2d_m} - \partial f_{\lambda_n}(\hat{\gamma})$$

$$= \frac{1}{n} \partial_1^\top g(\gamma_0) \{ \delta - V(\gamma_0) \} - \frac{1}{n} F_{n11}(\gamma_0)(\hat{\gamma}_1 - \gamma_{10}) - \partial f_{\lambda_n}(\hat{\gamma})$$

$$+ o_p(n^{-1/2}d_m^{-1/2}) + O_p(d_m^2 n^{-1}).$$

Under Assumption S4, it follows from $\min_j |\tau_j| \geq \min_{j \in \arg} |\gamma_{0j}| - s_n = s_n$ and the monotonicity of $\partial f_{\lambda_n}(t)$ that

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

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

$$F_{n11}(\gamma_0)(\hat{\gamma}_1 - \gamma_{10}) = \partial_1^\top g(\gamma_0) \{ \delta - V(\gamma_0) \} + o_p(\sqrt{n}).$$

By Assumption S2(i), we obtain

$$F_{n11}^{1/2}(\gamma_0)(\hat{\gamma}_1 - \gamma_{10}) = F_{n11}^{-1/2}(\gamma_0) \partial_1^\top g(\gamma_0) \{ \delta - V(\gamma_0) \} + o_p(1). \quad (S3.5)$$

Let $U_n U_n^\top = G$, where $U_n$ is a $m \times d_m$ matrix, and $G$ is a $m \times m$ symmetric positive define matrix. Then, we have $U_n F_{n11}^{1/2}(\gamma_0)(\hat{\gamma}_1 - \gamma_{10}) = \nu_n + o_p(1)$, where

$$\nu_n = U_n F_{n11}^{-1/2}(\gamma_0) \partial_1^\top g(\gamma_0) \{ \delta - V(\gamma_0) \}.$$
If we can prove $\nu_n \xrightarrow{L} \mathcal{N}(0, G)$, it follows from Slutsky’s theorem that $U_n F_{11}^{1/2}(\gamma_0)(\gamma_1 - \gamma_0) \xrightarrow{L} \mathcal{N}(0, G)$. For any unit vector $a \in \mathbb{R}^m$, we have

$$v_n = a^T \nu_n = a^T U_n F_{11}^{-1/2}(\gamma_0) \partial_{\gamma_1} g(\gamma_0) \{\delta - V(\gamma_0)\} = \sum_{i=1}^{n} z_i,$$

where $z_i = a^T U_n F_{11}^{-1/2}(\gamma_0) \partial_{\gamma_1} g_i(\gamma_0) \{\delta_i - V_i(\gamma_0)\}$. It is easily seen that $z_i$'s are independent. For $i = 1, \ldots, n$, we have $E(z_i) = 0$ and

$$\sum_{i=1}^{n} \text{var}(z_i) = a^T U_n F_{11}^{-1/2}(\gamma_0) F_{11}(\gamma_0) F_{11}^{-1/2}(\gamma_0) U_n^T a = a^T U_n U_n^T a \xrightarrow{p} a^T G a,$$

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

Under Assumption S4, it follows from Cauchy-Schwarz inequality that

$$\sum_{i=1}^{n} E|z_i|^3 = \sum_{i=1}^{n} |a^T U_n F_{11}^{-1/2}(\gamma_0) \partial_{\gamma_1} g_i(\gamma_0)|^3 E|\delta_i - V_i(\gamma_0)|^3$$

$$= O(1) \sum_{i=1}^{n} |a^T U_n F_{11}^{-1/2}(\gamma_0) \partial_{\gamma_1} g_i(\gamma_0)|^3$$

$$\leq O(1) \sum_{i=1}^{n} \|a^T U_n\|^3 \cdot \|F_{11}^{-1/2}(\gamma_0) \partial_{\gamma_1} g_i(\gamma_0)\|_2^3$$

$$= O(1) \sum_{i=1}^{n} \{\partial_{\gamma_1} g_i(\gamma_0) F_{11}^{-1}(\gamma_0) \partial_{\gamma_1} g_i(\gamma_0)\}^{3/2} = o(1).$$

Using the Lyapunov’s Theorem leads to $a^T \nu_n = \sum_{i=1}^{n} z_i \xrightarrow{L} \mathcal{N}(0, a^T G a)$, which holds for any unit vector $a \in \mathbb{R}^m$. 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) $\sup_{1 \leq s \leq S} \| \frac{1}{n} X_s^\top W X_s - \frac{1}{n} X_s^\top W X_s \| = O_p \left( p_m \sqrt{\frac{d_m}{n}} \right)$;

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

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

**Proof.** By the definition of \( \hat{W} \) and \( W \), we obtain

\[
\left\| \frac{1}{n} X_s X_s^\top \hat{W} X_s - \left( \frac{1}{n} X_s X_s^\top W X_s \right) \right\| = \left\| n^{-1} \sum_{i=1}^{n} \left( \frac{\delta_i}{\pi_i} - \frac{\delta_i}{\hat{\pi}_i} \right) X_{si} X_{si}^\top \right\|
\leq \sup_i \left\| \frac{\delta_i}{\pi_i} - \frac{\delta_i}{\hat{\pi}_i} \right\| n^{-1} \sum_{i=1}^{n} \left\| X_{si} X_{si}^\top \right\|.
\]

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

\[
\sup_i \left\| \frac{1}{\pi_i} - \frac{1}{\hat{\pi}_i} \right\| \leq \sup_i \left\| \frac{\partial \pi_i / \partial \gamma_1}{\pi_i^2} \right\| \cdot \| \gamma_1 - \gamma_0 \|_2 = O_p \left( \sqrt{\frac{d_m}{n}} \right).
\]

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 \( \hat{H}_s = X_s^\top \hat{W} X_s / n \) and \( H_s = X_s^\top W X_s / n \). Following Theorem 1 of Lewis and Reinsel (1985), we obtain

\[
\hat{H}_s^{-1} - H_s^{-1} = -\hat{H}_s^{-1} (\hat{H}_s - H_s) H_s^{-1} = -\{ H_s^{-1} + (\hat{H}_s^{-1} - H_s^{-1}) \} (\hat{H}_s - H_s) H_s^{-1}.
\]

Assumptions 1 and 2 imply that \( \sup_s \| H_s^{-1} \|_2 \leq \Lambda < \infty \) for some constant \( \Lambda \). According to Assumption 3(ii) and Lemma 1(i), and \( \sup_s \| \hat{H}_s - H_s \|_2 \leq \sup_s \| \hat{H}_s - H_s \| \to 0 \), there exists a constant \( \Lambda' \) such that \( \Lambda \| \hat{H}_s - H_s \|_2 < \Lambda' < 1 \) as \( n \to \infty \). Then, with the probability tending to one, we have

\[
\sup_{1 \leq s \leq S} \left\| \hat{H}_s^{-1} - H_s^{-1} \right\|_2 \leq \sup_{1 \leq s \leq S} \frac{\Lambda^2 \| \hat{H}_s - H_s \|_2}{1 - \Lambda \| \hat{H}_s - H_s \|_2},
\]
which leads to (ii). Using Triangle inequality and Assumption 3(ii) yields

\[
\sup_{1 \leq s \leq S} \| \hat{H}_s^{-1} \|_2 \leq \sup_{s} \| H_s^{-1} \|_2 + \sup_{1 \leq s \leq S} \| \hat{H}_s^{-1} - H_s^{-1} \|_2 = O_p(1),
\]

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

\[
\sup_{1 \leq s \leq S} \| \tilde{P}_s - P_s \| = \sup_{1 \leq s \leq S} \| \frac{1}{n} X_s \hat{H}_s^{-1} X_s^\top \hat{W} - \frac{1}{n} X_s (\frac{1}{n} X_s^\top X_s)^{-1} X_s^\top \| \\
\leq \sup_i \left| \frac{\delta_i}{\pi_i} - \frac{1}{\pi_i} \right| \cdot \sup_s \| \hat{H}_s^{-1} \|_2 \cdot \sup_s \frac{1}{n} \sum_{i=1}^n \| X_{si} X_{si}^\top \| \\
+ \sup_i \left| \frac{\delta_i}{\pi_i} - 1 \right| \cdot \sup_s \| \hat{H}_s^{-1} \|_2 \cdot \sup_s \frac{1}{n} \sum_{i=1}^n \| X_{si} X_{si}^\top \| \\
+ \sup_s \| \hat{H}_s^{-1} - (\frac{1}{n} X_s^\top X_s)^{-1} \|_2 \cdot \sup_s \frac{1}{n} \sum_{i=1}^n \| X_{si} X_{si}^\top \| \\
= O_p \left( p^2 m \sqrt{\frac{d_m}{n}} \right).
\]

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

\[
\sup_{1 \leq s \leq S} \| \tilde{P}_s - P^*_s \| = \sup_{1 \leq s \leq S} \| \tilde{D}_s (\tilde{P}_s - P_s) + (\tilde{D}_s - D_s) P_s + (D_s - \tilde{D}_s) \| \\
= O_p \left( p^2 m \sqrt{\frac{d_m}{n}} \right),
\]

which shows that (iv) holds.

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

**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 $1 \leq i \leq n$, where $a_i$ and $b_i$ are constants. Let $\bar{X} = \sum_{i=1}^{n} X_i/n$. Then, the following inequality holds

$$\Pr(|\bar{X} - E(\bar{X})| \geq t) \leq 2 \exp \left\{ - \frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2} \right\},$$

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 $L^*(\omega) = \{\mu - \tilde{\mu}(\omega)\}^\top \tilde{W} \{\mu - \tilde{\mu}(\omega)\}$ and $\tilde{\mu}(\omega) = \tilde{P}(\omega)Y$. Then, we have

\[
WCV(\omega) = \{Y - \tilde{\mu}(\omega)\}^\top \tilde{W} \{Y - \tilde{\mu}(\omega)\} = \{\mu + \varepsilon - \tilde{P}(\omega)Y\}^\top \tilde{W} \{\mu + \varepsilon - \tilde{P}(\omega)Y\} = \varepsilon^\top \tilde{W} \varepsilon + L^*(\omega) + 2 < \tilde{W}^{1/2} \varepsilon, \tilde{W}^{1/2} \{\mu - \tilde{P}(\omega)Y\} > = \varepsilon^\top \tilde{W} \varepsilon + L(\omega) \left\{ L^*(\omega) + \frac{2 < \tilde{W}^{1/2} \varepsilon, \tilde{W}^{1/2} \{\mu - \tilde{P}(\omega)Y\} >}{L(\omega)/R(\omega)} \right\}.
\]

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

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

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

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

we can obtain that $L(\omega)/\inf_{\omega \in W} L(\omega) \to 1$ is valid. Using Cauchy-Schwartz inequality-
ity leads to
\[
|L^*(\omega) - L(\omega)| = |\{\mu - \tilde{P}(\omega)Y\}^\top \hat{W} \{\mu - \tilde{P}(\omega)Y\} \nonumber \\
- \{\mu - \tilde{P}(\omega)Y\}^\top \hat{W} \{\mu - \tilde{P}(\omega)Y\}| 
\]
\[
= \|\hat{W}^{1/2} \{\tilde{P}(\omega) - \tilde{P}(\omega)\}Y\|^2 
- 2 < \hat{W}^{1/2} \{\mu - \tilde{P}(\omega)Y\}, \hat{W}^{1/2} \{\tilde{P}(\omega) - \tilde{P}(\omega)\}Y > | 
\leq \|\hat{W}^{1/2} \{\tilde{P}(\omega) - \tilde{P}(\omega)\}Y\|^2 
+ 2\sqrt{|L(\omega)|\|\hat{W}^{1/2} \{\tilde{P}(\omega) - \tilde{P}(\omega)\}Y\|}.
\]

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

\[
\sup_{\omega \in W} \|\hat{W}^{1/2} \{\tilde{P}(\omega) - \tilde{P}(\omega)\}Y\|^2 / L(\omega) \to 0. \tag{S3.9}
\]

From Lemma 1(iv), we have
\[
\tilde{P}(\omega) - \tilde{P}(\omega) = P^*(\omega) - P(\omega) + O_p(Sp_n^2 \sqrt{d_m/n}).
\]

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

\[
\sup_{\omega \in W} S^2 p_n^4 d_m n^{-1} / R(\omega) \to 0, \tag{S3.10}
\]
\[
\sup_{\omega \in W} \|\hat{W}^{1/2} \{P^*(\omega) - P(\omega)\}Y\|^2 / R(\omega) \to 0. \tag{S3.11}
\]

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

\[
\sup_{\omega \in W} S^2 p_n^4 d_m n^{-1} / R(\omega) \leq S^2 p_n^4 d_m n^{-1} / \xi_n 
= \left(\frac{S^{4G} \|\mu\|^{2G}}{\xi_n^{2G}}\right)^{1/2G} \sqrt{n} \left(\frac{p_n^{8/3} d_m}{n}\right)^{3/2} \cdot \frac{1}{\sqrt{d_m}} \to 0,
\]

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

$$\| \hat{W}^{1/2} \{ P^*(\omega) - P(\omega) \} Y \|^2$$

$$= \left\{ \sum_{s=1}^{S} \omega_s \| \hat{W}^{1/2} (P_s^* - P_s) \mu \| + \sum_{s=1}^{S} \omega_s \| \hat{W}^{1/2} (P_s^* - P_s) \epsilon \| \right\}^2$$

$$\leq \left\{ \sum_{s=1}^{S} \| \hat{W}^{1/2} (P_s^* - P_s) \mu \| + \sum_{s=1}^{S} \| \hat{W}^{1/2} (P_s^* - P_s) \epsilon \| \right\}^2$$

$$\leq 2S^2 \left\{ \max_s \| \hat{W}^{1/2} (P_s^* - P_s) \mu \|^2 + \max_s \| \hat{W}^{1/2} (P_s^* - P_s) \epsilon \|^2 \right\}.$$ 

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

$$S^2 \max_s \| \hat{W}^{1/2} (P_s^* - P_s) \mu \|^2 / \xi_n \to 0,$$

(S3.12)

$$S^2 \max_s \| \hat{W}^{1/2} (P_s^* - P_s) \epsilon \|^2 / \xi_n \to 0.$$

(S3.13)

Note that

$$\frac{\delta_i}{\hat{\pi}_i} = \frac{\delta_i}{\pi_i} \left\{ 1 - \frac{\partial_i \hat{\pi}_i}{\pi_i} (\hat{\gamma}_1 - \gamma_{10}) + o_p(\sqrt{d_m/n}) \right\}.$$

By Theorem S1.1 and Assumption 1, we have

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

(S3.14)
Therefore, for (S3.12) and (S3.13), it is sufficient to show

\[ S^2 \max_s \| \hat{W}^{1/2} (P^*_s - P_s) \mu \|^2 / \xi_n \leq S^2 \max_s \sup_i |\delta_i / \hat{\pi}_i| \cdot \| (P^*_s - P_s) \mu \|^2 / \xi_n \]

\[ \leq C' S^2 \max_s \| (P^*_s - P_s) \mu \|^2 / \xi_n \to 0, \quad (S3.15) \]

\[ S^2 \max_s \| \hat{W}^{1/2} (P^*_s - P_s) \epsilon \|^2 / \xi_n \leq S^2 \max_s \sup_i |\delta_i / \hat{\pi}_i| \cdot \| (P^*_s - P_s) \epsilon \|^2 / \xi_n \]

\[ \leq C' S^2 \max_s \| (P^*_s - P_s) \epsilon \|^2 / \xi_n \to 0. \quad (S3.16) \]

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

\[ \Pr \left( S^2 \sum_s \| (P^*_s - P_s) \mu \|^2 / \xi_n > \kappa \right) \leq \sum_{s=1}^{S} \Pr \left( S^{4G} \| (P^*_s - P_s) \mu \|^{4G} / \xi_n^{2G} > \kappa^{2G} \right) \]

\[ \leq \frac{S^{4G}}{\kappa^{2G} \xi_n^{2G}} \sum_{s=1}^{S} \{ E \| (P^*_s - P_s) \mu \|^{4G} \} \]

\[ \leq C' \frac{S^{4G}}{\kappa^{2G} \xi_n^{2G}} \sum_{s=1}^{S} \| (P^*_s - P_s) \mu \|^{4G} \]

\[ \leq C' \frac{S^{4G}}{\kappa^{2G} \xi_n^{2G}} \sum_{s=1}^{S} \{ \max_s (P^*_s - P_s)^{4G} \| \mu \|^{4G} \}

\[ \leq C' \frac{S^{4G}}{\kappa^{2G} \xi_n^{2G}} \cdot \frac{\| \mu \|^2}{\xi_n} \cdot \left( \frac{p_{m/3}^4}{n} \right) \cdot \left( \frac{p_m}{n} \right)^{3G} \to 0, \]

which implies that (S3.15) holds.

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

\[ \sum_{s=1}^{S} \Pr \left\{ S^2 \| (P^*_s - P_s) \epsilon \|^2 - E \| (P^*_s - P_s) \epsilon \|^2 / \xi_n > \kappa \right\} \to 0, \quad (S3.17) \]

\[ S^2 \max_s E \| (P^*_s - P_s) \epsilon \|^2 / \xi_n \to 0. \quad (S3.18) \]
By Theorem 2 of Whittle (1960) and Lemma 1(iii), it follows from (S3.17) and Assumptions 3(i), (iii) and (vi) that

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

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

\[
S^2 \max_s E\| (P^*_s - P_s) \varepsilon \|^2 / \xi_n = S^2 \max_s \text{tr} \left\{ (P^*_s - P_s) \sigma \varepsilon (P^*_s - P_s)^\top \right\} / \xi_n^{-1} \\
\leq C^2 E_{\max}(\sigma \varepsilon) \cdot \frac{p_s^2 S^2}{n \xi_n} \\
= C^2 \cdot E_{\max}(\sigma \varepsilon) \cdot \frac{p_s^2}{n} \cdot \left( \frac{S^{4G}}{\xi_n^{2G}} \right)^{1/2G} \to 0,
\]

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

\[
| < \tilde{W}^{1/2} \varepsilon, \tilde{W}^{1/2} \{ \mu - \tilde{P}(\omega) Y \} > | \leq | < \tilde{W}^{1/2} \varepsilon, \tilde{W}^{1/2} \mu > | + | < \tilde{W}^{1/2} \varepsilon, \tilde{W}^{1/2} P^*(\omega) \mu > | \\
+ | < \tilde{W}^{1/2} \varepsilon, \tilde{W}^{1/2} P^*(\omega) \varepsilon > | + O_p \left( \frac{S p_m^2 \sqrt{d_m}}{n} \right).
\]
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

\[
\sup_{\omega \in W} | \langle \hat{W}^{1/2} \varepsilon, \hat{W}^{1/2} \mu \rangle | / R(\omega) \to 0, \tag{S3.19}
\]

\[
\sup_{\omega \in W} | \langle \hat{W}^{1/2} \varepsilon, \hat{W}^{1/2} P^*(\omega) \mu \rangle | / R(\omega) \to 0, \tag{S3.20}
\]

\[
\sup_{\omega \in W} | \langle \hat{W}^{1/2} \varepsilon, \hat{W}^{1/2} P^*(\omega) \varepsilon \rangle | / R(\omega) \to 0. \tag{S3.21}
\]

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

\[
\Pr \left\{ \sup_{\omega \in W} | \langle \hat{W}^{1/2} \varepsilon, \hat{W}^{1/2} \mu \rangle | / R(\omega) > \kappa \right\} 
\leq \frac{E|\varepsilon^T \hat{W} \mu|^{2G}}{\kappa^{2G} \xi^{2G}} 
\leq C' \kappa^{-2G} \frac{||\mu||^{2G}}{\xi^{2G}} \to 0.
\]

For (S3.19), under (S3.14), we obtain \(|\varepsilon^T \hat{W} P^*(\omega) \mu| \leq C'|\varepsilon^T P^*(\omega) \mu|\). According to Lemma 2(iv) and Assumption 3(i) and (vi), we just require proving

\[
\Pr \left( \sup_{\omega \in W} | \varepsilon, P^*(\omega) \mu | / R(\omega) > \kappa \right) \leq \Pr \left( \max_s | \varepsilon, P_s^* \mu | > \kappa \xi_n \right)
\leq \sum_{s=1}^S \Pr \left( S^{2G} | \varepsilon, P_s^* \mu | > \kappa^{2G} \xi_n^{2G} \right)
\leq \frac{S^{2G}}{\xi_n^{2G} \kappa^{2G}} \sum_{s=1}^S E|\varepsilon P_s^* \mu|^{2G}
\leq C' \frac{S^{2G}}{\xi_n^{2G} \kappa^{2G}} \sum_{s=1}^S \mathbb{E}_{\max}(P_s^*)^{2G} ||\mu||^{2G}
\leq C' \cdot (1 + C)^{2G} \cdot \frac{S^{2G+1} ||\mu||^{2G}}{\kappa^{2G} \xi_n^{2G}} \to 0.
\]
Similarly, by Lemma 2(v), we only require proving

$$\Pr\left(\sup_{\omega \in W} |\epsilon, P^*(\omega)\epsilon > |/R(\omega) > \kappa\right) \leq \Pr\left(S_{\text{max}}|\epsilon, P'_s\epsilon > | > \kappa\xi_n\right)$$

$$\leq \frac{S^2G}{\kappa^2G\xi^2n} \sum_{s=1}^{S} E|\epsilon|^{2G}$$

$$\leq C'\sigma^2G \sum_{s=1}^{S} \left\{\text{tr}(P'_sP'_s\epsilon)\right\}^G$$

$$\leq C' \cdot C^G \cdot \sigma^2G \cdot \frac{S^2G + 1}{\kappa^2G\xi^2n} \rightarrow 0,$$

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

$$|L(\omega) - R(\omega)|$$

$$= |\{\mu - \hat{\mu}(\omega)\}^\top \tilde{W}\{\mu - \hat{\mu}(\omega)\} - E[\{\mu - \hat{\mu}(\omega)\}^\top \tilde{W}\{\mu - \hat{\mu}(\omega)\}]|$$

$$\leq \|\tilde{W}^{1/2}M(\omega)\mu\|^2 + \|\tilde{W}^{1/2}P(\omega)\epsilon\|^2 + 2|\tilde{W}^{1/2}M(\omega)\mu, \tilde{W}^{1/2}P(\omega)\epsilon > |$$

$$+ E\{\|\tilde{W}^{1/2}M(\omega)\mu\|^2|X\} + E\{\|\tilde{W}^{1/2}P(\omega)\epsilon\|^2|X\}$$

$$+ |E\{2 < \tilde{W}^{1/2}M(\omega)\mu, \tilde{W}^{1/2}P(\omega)\epsilon > |X\}$$

$$\leq C'\|\tilde{M}(\omega)\mu\|^2 + C'\|\tilde{P}(\omega)\epsilon\|^2 + 2C'\|\tilde{M}(\omega)\mu, \tilde{P}(\omega)\epsilon > |X\}$$

$$+ C'E\{\tilde{M}(\omega)\mu\|^2|X\} + C'E\{\tilde{P}(\omega)\epsilon\|^2|X\} + 2C'|E\{\tilde{M}(\omega)\mu, \tilde{P}(\omega)\epsilon > |X\}$$

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

\[ \sup_{\omega \in W} \left| \frac{\|M(\omega)\mu\|^2}{R(\omega)} \right| \rightarrow 0, \tag{S3.22} \]

\[ \sup_{\omega \in W} \left| \frac{\|P(\omega)\epsilon\|^2 - \text{tr}\{P(\omega)\sigma_\epsilon P^T(\omega)\}}{R(\omega)} \right| \rightarrow 0, \tag{S3.23} \]

\[ \sup_{\omega \in W} \left| \frac{< M(\omega)\mu, P(\omega)\epsilon >}{R(\omega)} \right| \rightarrow 0, \tag{S3.24} \]

where $M = I - P(\omega)$. 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

\[
\Pr \left\{ \sup_{\omega \in W} \left| \frac{\|P(\omega)\epsilon\|^2 - \text{tr}\{P(\omega)\sigma_\epsilon P^T(\omega)\}}{R(\omega)} \right| > \kappa \right\} \leq \Pr \left\{ \sup_{\omega \in W} \sum_{s=1}^{S} \sum_{k=1}^{S} \omega_s \omega_k |\epsilon^T P_s P_k \epsilon - \sigma_\epsilon \text{tr}(P_s P_k)| > \kappa \xi_n \right\} \leq \Pr \left\{ S^2 \max_{s,k} |\epsilon^T P_s P_k \epsilon - \sigma_\epsilon \text{tr}(P_s P_k)| > \kappa \xi_n \right\} \]

\[
\leq \sum_{s=1}^{S} \sum_{k=1}^{S} \Pr \left\{ |\epsilon^T P_s P_k \epsilon - \sigma_\epsilon \text{tr}(P_s P_k)| > \kappa \xi_n / S^2 \right\} \leq \sum_{s=1}^{S} \sum_{k=1}^{S} \frac{S^4 G}{\kappa^2 G \xi_n^2} \Pr \left\{ |\epsilon^T P_s P_k \epsilon - \sigma_\epsilon \text{tr}(P_s P_k)| \right\}^2 \leq C' \frac{S^4 G}{\kappa^2 G \xi_n^2} \sum_{s=1}^{S} \sum_{k=1}^{S} \left\{ \text{tr}(P_s^2 P_k^2) \right\}^G \leq C' \cdot \frac{S^4 G + 2 n G}{\xi_n^2} \cdot \kappa^{-2 G} \rightarrow 0,
\]

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

\[
\Pr \left\{ \sup_{\omega \in \mathcal{W}} \left| \frac{< M(\omega)\mu, P(\omega)\varepsilon >}{R(\omega)} \right| > \kappa \right\} \leq \Pr \left\{ \sup_{\omega \in \mathcal{W}} \sum_{s=1}^{S} \sum_{k=1}^{S} |\mu^T(I - P_s)P_k\varepsilon| > \kappa \zeta_n \right\}
\]

\[
\leq \frac{S^4 G}{\kappa^2 G \zeta_n} \sum_{s=1}^{S} \sum_{k=1}^{S} E|\mu^T(I - P_s)P_k\varepsilon|^{2G}
\]

\[
\leq C' \frac{S^{4G} + 2\|\mu\|^{2G}}{\zeta_n^{2G}} \cdot \kappa^{-2G} \to 0,
\]

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

\[
\hat{r}_k - r_k = \frac{1}{n} \sum_{i=1}^{n} \left\{ \delta_i X_{ik} Y_i + (1 - \delta_i) \frac{1}{m} \sum_{v=1}^{m} X_{ik} \tilde{Y}_{iv}^k \right\} - E(X_k Y)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \{ \hat{m}(X_{ik}) - m(X_{ik}) \} + \frac{1}{n} \sum_{i=1}^{n} (1 - \delta_i) \left\{ \frac{1}{m} \sum_{v=1}^{m} X_{ik} \tilde{Y}_{iv}^k - \hat{m}(X_{ik}) \right\}
\]

\[
+ \frac{1}{n} \sum_{i=1}^{n} \delta_i \{ X_{ik} Y_i - m(X_{ik}) \} + \frac{1}{n} \sum_{i=1}^{n} \{ m(X_{ik}) - E(X_k Y) \}
\]

\[
= J_{k1} + J_{k2} + J_{k3} + J_{k4}.
\]
For $J_{k2}$, Wang and Chen (2009) have proved $J_{k3} = o_p(1/\sqrt{n})$ as $n, m \to \infty$. Thus, as $n$ is sufficiently large, for any $\varsigma \in (0, 1/2)$ and $c_2 > 0$, we have

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

Define $\eta(x) = \pi(x)f_k(x)$ 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

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

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, S5 and S6, by the similar certification of Wang and Chen (2009),
\[ J_{k1} - \tilde{J}_{k1} = o_p(1/\sqrt{n}), \ J_{k12} = o_p(1/\sqrt{n}) \text{ and } J_{k13} = o_p(1/\sqrt{n}). \] Then, we have

\[
\Pr(|J_{k1} + J_{k3} + J_{k4}| \geq c_2n^{-c}/2)
\leq \Pr(|\tilde{J}_{k11} + (J_{k11} - \tilde{J}_{k11}) + J_{k12} + J_{k13} + J_{k3} + J_{k4}| \geq c_2n^{-c}/2)
\leq \Pr(|\tilde{J}_{k11} + J_{k3} + J_{k4}| \geq c_2n^{-c}/2 - |J_{k11} - \tilde{J}_{k11}| - |J_{k12}| - |J_{k13}|)
\leq \Pr(|\tilde{J}_{k11} + J_{k3} + J_{k4}| \geq c_2n^{-c}/16).
\]

Note that

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

Under Assumption 1, for any \( M > 0 \), we have

\[
\Pr(|I_1| \geq c_2n^{-c}/32) = \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \delta_i \{X_{ik}Y_i - E(X_kY)\} \frac{1}{\pi(X_{ik})} \right| \geq c_2n^{-c}/32 \right\}
\leq \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \delta_i \{X_{ik}Y_i - E(X_kY)\} \right| \geq C_0c_2n^{-c}/32 \right\}
\leq \Pr \left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \delta_i \{X_{ik}Y_i - E(X_kY)\} \right| \geq C_0c_2n^{-c}/32, \ \max_i |\delta_i X_{ik}Y_i| < M \right\}
+ \Pr (\max_i |X_{ik}Y_i| \geq M).
\]
According to Assumption S7 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 $\Pr(|X_kY| \geq M) \leq c_3 \exp(-c_4 M)$. 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

$$\Pr\left\{ \left| \frac{1}{n} \sum_{i=1}^{n} \delta_i \{X_{ik}Y_i - E(X_kY)\} \right| \geq C_0 c_2 n^{-\varsigma}/32, \ \max_i |\delta_i X_{ik}Y_i| < M \right\}$$

$$\leq 2 \exp(-n^{1-2\varsigma}/M^2) + \sum_{i=1}^{n} \Pr(|X_{ik}Y_i| \geq M)$$

$$\leq O(n) \exp\{-c_5 n^{(1-2\varsigma)/3}\}.$$

According to the Assumptions 1 and S7, the above argument can be used to $I_2$, we have $\Pr(|I_2| \geq c_2 n^{-\varsigma}/32) \leq O(n) \exp\{-c_6 n^{(1-2\varsigma)/3}\}$ for some constant $c_6$. Thus, there exists a constant $c_7$ such that

$$\Pr\left( \max_{1 \leq k \leq p} |\hat{r}_k - r_k| \geq c_2 n^{-\varsigma} \right) \leq p \Pr(|I_1| \geq c_2 n^{-\varsigma}/32) + p \Pr(|I_2| \geq c_2 n^{-\varsigma}/32)$$

$$\leq O(n) p \exp\{-c_7 n^{(1-2\varsigma)/3}\}$$

$$= O\{p \exp(-c_7 n^{(1-2\varsigma)/3} + \log(n))\}.$$

In fact, by $S_n = \{ \max_{k \in \mathcal{M}_*} |\hat{r}_k - r_k| \leq c_0 n^{-\varsigma}/2 \}$ and Assumption S8, we have

$$\min_{k \in \mathcal{M}_*} |\hat{r}_k| \geq \min_{k \in \mathcal{M}_*} (|r_k| - |\hat{r}_k - r_k|) \geq \min_{k \in \mathcal{M}_*} |r_k| - \max_{k \in \mathcal{M}_*} |\hat{r}_k - r_k| \geq c_0 n^{-\varsigma}/2.$$

Thus, by taking $\varsigma_n = c_8 n^{-\varsigma}$ with $c_8 \leq c_0/2$, there exists some positive constant $c_9$ such that $\Pr(\mathcal{M}_* \subset \hat{\mathcal{M}}) \geq \Pr(S_n) \geq 1 - O\{|\mathcal{M}_*| \exp(-c_9 n^{(1-2\varsigma)/3} + \log(n))\}$. Now
we show the ranking consistency property. It is easily shown that

\[
\Pr \left\{ \left( \min_{k \in M} \hat{r}_k - \max_{k \in I} \hat{r}_k \right) < m_0/2 \right\} \\
\leq \Pr \left\{ \left( \max_{k \in M} \left| \hat{r}_k - \max_{k \in I} \hat{r}_k \right| \right) < -m_0/2 \right\} \\
\leq \Pr \left\{ \left( \min_{k \in M} \left| \hat{r}_k - \max_{k \in I} \hat{r}_k \right| \right) - \left( \min_{k \in M} \left| r_k - \max_{k \in I} r_k \right| \right) \geq m_0/2 \right\} \\
\leq \Pr \left( \max_{1 \leq k \leq p} \left| \hat{r}_k - r_k \right| \geq m_0/2 \right) \\
\leq p \Pr \left( \left| \hat{r}_k - r_k \right| \geq m_0/4 \right) \\
\leq O(n) p \exp(-c_{10}n^{1/3}m_0^{2/3})
\]

for some constants $c_{10}$, where the first inequality holds because of Assumption S9.

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

\[
\sum_{n=n_0}^{\infty} pn \exp(-c_{10}n^{1/3}m_0^{2/3}) \leq 2 \sum_{n=n_0}^{\infty} \exp\left\{ c_{10}n^{1/3}m_0^{2/3}/2 - c_{10}n^{1/3}m_0^{2/3} + \log(n) \right\} \\
\leq 2 \sum_{n=n_0}^{\infty} \exp\left\{ -3\log(n) \right\} \leq 2 \sum_{n=n_0}^{\infty} n^{-3} < +\infty.
\]

Hence, by Borel Contelli Lemma, we have

\[
\lim \inf_{n \to \infty} \left\{ \min_{k \in M} \left| \hat{r}_k - \max_{k \in I} \hat{r}_k \right| \right\} > 0 \ a.s.
\]

References


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