Proofs of Theorems

Let $f(x) \in \Re$ denote a twice-differentiable function of $x \in \Re^P$. $\nabla f(x^*)$ and $\nabla^2 f(x^*)$ are defined as the gradient and Hessian of $f(x)$ evaluated at x^*, respectively, i.e., $\nabla f(x^*) = \frac{\partial f(x^*)}{\partial x}$ and $\nabla^2 f(x^*) = \frac{\partial^2 f(x^*)}{\partial x \partial x^T}$. Given an index set $\mathcal{J} \subset \{1,2,\dots,P\}$, $\nabla_{\mathcal{J}} f(x^*)$ denotes the vector formed by $\left\{\frac{\partial f(x^*)}{\partial x_q}\right\}_{q \in \mathcal{J}}$, where x_q is the q^{th} element of x. In a similar manner, $\nabla_{\mathcal{J}}^2 f(x^*)$ is used to denote the $|\mathcal{J}| \times |\mathcal{J}|$ matrix formed by $\left\{\frac{\partial^2 f(x^*)}{\partial x_q \partial x_{q'}}\right\}_{q,q' \in \mathcal{J}}$, where $|\mathcal{J}|$ is the number of elements in $\mathcal{J}$. For a vector $x \in \Re^P$, $\|x\|_q = \left(\sum_{p=1}^P |x_p|^q\right)^{1/q}$ denotes the ℓ_q norm of x. In particular, $\|x\|$, $\|x\|_0$, and $\|x\|_\infty$ are defined as $\left(\sum_{p=1}^P x_p^2\right)^{1/2}$, $\sum_{p=1}^P 1\{x_p \neq 0\}$, and $\max\{|x_p|\}_{p=1}^P$, respectively. For a square matrix $A \in \Re^{P \times P}$, $\omega_{min}(A)$ and $\omega_{max}(A)$ are used to denote the smallest and largest eigenvalue of A.

To derive the asymptotic properties of PL estimator, the following regularity conditions are assumed.

Condition A. $\mathcal{Y}_N = \{Y_n\}_{n=1}^N$ is a random sample from some distribution F that satisfies (1) $\mathbb{E}(Y) = \mu^*$; (2) $\mathbb{V}\mathrm{ar}(Y) = \Sigma^* \succ 0$; i.e., Σ^* is positive definite; (3) there exists an $\varepsilon > 0$ such that $\mathbb{E}\left(|Y_p|^{4+\varepsilon}\right) < \infty$ for all p.

Condition B. For each $\theta \in \Theta$ and any combination of q, q', and q'' ($q, q', q'' = 1,2,\dots,Q$), $\frac{\partial^3 \tau(\theta)}{\partial \theta_q \partial \theta_{q'} \partial \theta_{q''}}$ exists.

Condition C. There exists a quasi-true parameter $\theta^* \in \Theta$ such that (1) $\theta^* \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathbb{E}(\mathcal{L}(\theta))$; (2) $\|\theta^*\|_0 < \|\theta\|_0$ for any $\theta \in \underset{\theta \in \Theta}{\operatorname{argmax}}\, \mathbb{E}(\mathcal{L}(\theta))$, but $\theta \neq \theta^*$; (3) θ^* is the unique maximizer of $\mathbb{E}(\mathcal{L}(\theta))$ on $\Theta_{\mathcal{A}^*}$, where $\mathcal{A}^* = \{q | \theta_q^* \neq 0\}$ is the support of θ^*; $\Theta_{\mathcal{A}^*} = \Theta \cap \left(\prod_{q=1}^Q \mathfrak{X}_q\right)$ is the restricted parameter space with $\mathfrak{X}_q = \Re$ if $q \in \mathcal{A}^*$, and $\mathfrak{X}_q = \{0\}$ otherwise; (4) there exists a neighborhood of θ^* on $\Theta_{\mathcal{A}^*}$, denoted by $\Omega_{\mathcal{A}^*}(\theta^*)$ and a constant $\kappa_1 > 0$ such that $\omega_{min}\left(\mathcal{F}_{\mathcal{A}^*}(\theta)\right) > \kappa_1$ for all $\theta \in \Omega_{\mathcal{A}^*}(\theta^*)$, where $\mathcal{F}_{\mathcal{A}^*}(\theta) = \mathbb{E}\left(-\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta_{\mathcal{A}^*} \partial \theta_{\mathcal{A}^*}^T}\right)$.

Condition D. For each combination of q, q', and q'', there exists an F-integrable random function $K_{qq'q''}(y)$ such that $\left|\frac{\partial^3 \log \varphi_\theta(y)}{\partial \theta_q \partial \theta_{q'} \partial \theta_{q''}}\right| < K_{qq'q''}(y)$ for all y and θ in the neighborhood of θ^*.

Condition E. The penalty term $\mathcal{R}(\theta, \gamma) = \sum_{q=1}^{Q} c_q \rho(|\theta_q|, \gamma)$ satisfies (1) $c_q = 1$ if $\theta_q^* = 0$; (2) $\rho(t, \gamma)$ is increasing and concave in $t > 0$; (3) $\frac{\partial \rho(t,\gamma)}{\partial t}$ is continuous in both t and γ; (4) $\frac{\partial \rho(0+,\gamma)}{\partial t} = \gamma$; (5) $\frac{\partial \rho(t,\gamma)}{\partial t} = 0$ if $t > \delta\gamma$.

Condition F. θ^* is the unique maximizer of $\mathbb{E}\left(\mathcal{L}(\theta)\right)$ on Θ, and there exists a neighborhood of θ^* on Θ, denoted by $\Omega(\theta^*)$, and a constant $\kappa_2 > 0$ such that $\omega_{min}\left(\mathcal{F}(\theta)\right) \geq \kappa_2$ for all $\theta \in \Omega(\theta^*)$, where $\mathcal{F}(\theta) = \mathbb{E}\left(-\frac{\partial^2 \mathcal{L}(\theta)}{\partial \theta \partial \theta^T}\right)$.

Condition A requires each observation to be an independent realization from the same distribution satisfying some moment conditions. It is a standard assumption for minimum discrepancy function estimation in SEM (e.g., Browne, 1984; Shapiro, 1983). In SEM applications, the support of the manifest variable is often bounded, implying that Condition A holds. Condition B assumes that model $\tau(\theta)$ is smooth enough so that the quadratic approximation for $\mathcal{L}(\theta)$ is allowed. If the specified model is in the class of Equations (1) and (2) in the main text, Condition B is generally satisfied. The combination of Conditions A and B implies the existence of $\mathcal{F}(\theta)$ and $\mathcal{H}(\theta) = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^{N} \frac{\partial \log \varphi_\theta(Y_n)}{\partial \theta} \frac{\partial \log \varphi_\theta(Y_n)}{\partial \theta^T}\right)$. Both $\mathcal{F}(\theta)$ and $\mathcal{H}(\theta)$ play important roles for studying the asymptotic behavior of PL estimators. Condition C requires the existence and the uniqueness of a quasi-true parameter θ^* on the restricted parameter space $\Theta_{\mathcal{A}^*}$, even when $\tau(\theta)$ is not identifiable on the whole parameter space Θ. However, the positive-definiteness of $\mathcal{F}_{\mathcal{A}^*}(\theta)$ on $\Omega_{\mathcal{A}^*}(\theta^*)$ implies that $\tau(\theta)$ is at least locally identified on the restricted parameter space $\Theta_{\mathcal{A}^*}$. Condition D ensures that the remaining term of the quadratic approximation of $\mathcal{L}(\theta)$ around θ^* can be arbitrarily small in probability. Condition E makes several assumptions about the penalty term. The first assumption requires that the penalization weights must be one for all true-zero parameters. If such assumption is not satisfied for some $\theta_q^* = 0$, it is impossible to obtain a sparse PL estimate for θ_q^*. A simple way to fulfill this requirement is to set all the penalization indicators to be one except for the indicators for variance parameters. The remaining assumptions in Condition E restrict the shape of the penalty function. Both SCAD and MCP satisfy the all of the properties. However, the ℓ_1 penalty does not satisfy the last property and hence the established theorem cannot be applied to the ℓ_1-penalized estimator. Finally, Condition F is a more restricted version of Condition C and is required to establish a global theoretical result for the PL estimators.

Theorem 1 (local oracle property). If Conditions A-E are true, γ satisfies $\gamma \to 0$, and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$, then there exists a strictly local maximizer of $\mathcal{U}(\theta, \gamma)$, denoted by $\hat{\theta} = \hat{\theta}(\gamma)$, such that

(a) $\lim_{N\to\infty} \mathbb{P}\big(\hat{\mathcal{A}}(\gamma) = \mathcal{A}^*\big) = 1$, where $\hat{\mathcal{A}}(\gamma)$ is the estimated support of $\hat{\theta}(\gamma)$;

(b) $\sqrt{N}\big(\hat{\theta}_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\big) \to_{\mathcal{D}} \mathcal{N}\big(0, \mathcal{F}^{*\,-1}_{\mathcal{A}^*}\mathcal{H}^*_{\mathcal{A}^*}\mathcal{F}^{*\,-1}_{\mathcal{A}^*}\big)$, where $\mathcal{F}^*_{\mathcal{A}^*} = \mathbb{E}\left(-\frac{\partial^2 \mathcal{L}(\theta^*)}{\partial\theta_{\mathcal{A}^*}\partial\theta^T_{\mathcal{A}^*}}\right)$ and $\mathcal{H}^*_{\mathcal{A}^*} = \mathbb{E}\left(\frac{1}{N}\sum_{n=1}^N \frac{\partial \log\varphi_{\theta^*}(Y_n)}{\partial\theta_{\mathcal{A}^*}} \frac{\partial \log\varphi_{\theta^*}(Y_n)}{\partial\theta^T_{\mathcal{A}^*}}\right)$.

Theorem 1 can be established by proving the following three lemmas.

Lemma 1. Under Conditions A-E, there exists a sequence of maximizer of $\mathcal{L}(\theta)$ on the restricted parameter space $\Theta_{\mathcal{A}^*}$, denoted by $\tilde{\theta}^* = \tilde{\theta}^*_N$, such that

(a) $\lim_{N\to\infty} \mathbb{P}\left(\|\tilde{\theta}^* - \theta^*\| < \epsilon\right) = 1;$

(b) $\sqrt{N}\left(\tilde{\theta}^*_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\right) \longrightarrow_{\mathcal{D}} \mathcal{N}\left(0, {\mathcal{F}^*_{\mathcal{A}^*}}^{-1} \mathcal{H}^*_{\mathcal{A}^*} {\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\right).$

Proof: The technique in Section 6.5 of Lehmann and Casella (1998) is adopted to prove this lemma. For part (a), we want to show that for any sufficiently small $\varepsilon > 0$ with probability tending to 1 that

$$\mathcal{L}(\theta^*) > \mathcal{L}(\theta), \tag{1}$$

at all points θ on the surface of $\mathcal{S}_\varepsilon$, where $\mathcal{S}_\varepsilon$ is the sphere with center at θ^* and radius ε. Equation (1) implies that there exists a local maximum in the interior of $\mathcal{S}_\varepsilon$ and a consistent sequence of local maximum can be selected. By Taylor's theorem, we have

$$\begin{aligned}\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \le{}& \nabla_{\mathcal{A}^*}\mathcal{L}(\theta^*)^T(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}) + \frac{1}{2}(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*})^T \nabla^2_{\mathcal{A}^*}\mathcal{L}(\theta^*)(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}) \\ &+ \frac{1}{6}\sum_{q\in\mathcal{A}^*}\sum_{q'\in\mathcal{A}^*}\sum_{q''\in\mathcal{A}^*}\left(\theta_q - \theta^*_q\right)\left(\theta_{q'} - \theta^*_{q'}\right)\left(\theta_{q''} - \theta^*_{q''}\right)K_{qq'q''}(Y) \\ &= a_1 + a_2 + a_3.\end{aligned} \tag{2}$$

We know that $\left|\theta_q - \theta^*_q\right| = \varepsilon$, $\|\nabla_{\mathcal{A}^*}\mathcal{L}(\theta^*)\| \to_{\mathcal{P}} 0$, and $-\nabla^2_{\mathcal{A}^*}\mathcal{L}(\theta^*) \to_{\mathcal{P}} \mathcal{F}^*_{\mathcal{A}^*}$. Hence, for large N, with probability tending to 1 we have

$$|a_1| \le \varepsilon\|\nabla_{\mathcal{A}^*}\mathcal{L}(\theta^*)\| \le |\mathcal{A}^*|\varepsilon^3 = C_1\varepsilon^3, \tag{3}$$

$$\begin{aligned}|a_2| ={}& -\frac{1}{2}(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*})^T \mathcal{F}^*_{\mathcal{A}^*}(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}) + \frac{1}{2}(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*})^T\left(\nabla^2_{\mathcal{A}^*}\mathcal{L}(\theta^*) + \mathcal{F}^*_{\mathcal{A}^*}\right)(\theta_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}) \\ &\le \omega_{max}(-\mathcal{F}^*_{\mathcal{A}^*})\varepsilon^2 + |\mathcal{A}^*|\varepsilon^3 \le -C_2\varepsilon^2,\end{aligned} \tag{4}$$

and

$$|a_3| \le \frac{1}{6}\varepsilon^3|\mathcal{A}^*|^3 \sum\sum\sum \mathbb{E}\left(K_{qq'q''}(Y)\right) = C_3\varepsilon^3, \tag{5}$$

for some C_1, C_2, and $C_3 > 0$, indicating that

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \le C_1\varepsilon^3 - C_2\varepsilon^2 + C_3\varepsilon^3. \tag{6}$$

Therefore, we conclude that if $\varepsilon < C_2/(C_1 + C_3)$, we have $\mathcal{L}(\theta) - \mathcal{L}(\theta^*) < 0$ for all θ on the surface of $\mathcal{S}_\varepsilon$.

To prove (b), according to Taylor's theorem,

$$\nabla_{\mathcal{A}^*}\mathcal{L}(\tilde{\theta}^*) = \nabla_{\mathcal{A}^*}\mathcal{L}(\theta^*) + \nabla^2_{\mathcal{A}^*}\mathcal{L}(\theta^*)\left(\tilde{\theta}^*_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\right) + o_p\left(N^{-\frac{1}{2}}\right). \tag{7}$$

Because $\nabla_{\mathcal{A}^*}\mathcal{L}(\tilde{\theta}^*) = 0$ and $-\nabla^2_{\mathcal{A}^*}\mathcal{L}(\theta^*) \to_{\mathcal{P}} \mathcal{F}^*_{\mathcal{A}^*}$, we have that $\sqrt{N}\left(\tilde{\theta}^*_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\right) = {\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\sqrt{N}\nabla_{\mathcal{A}^*}\mathcal{L}(\theta^*) + o_p(1)$. By the fact that $\sqrt{N}\nabla_{\mathcal{A}^*}\mathcal{L}(\theta^*) \longrightarrow_{\mathcal{D}} \mathcal{N}(0, \mathcal{H}^*_{\mathcal{A}^*})$ and Slutsky's theorem, we conclude that $\sqrt{N}\left(\tilde{\theta}^*_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\right) \longrightarrow_{\mathcal{D}} \mathcal{N}\left(0, {\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\mathcal{H}^*_{\mathcal{A}^*}{\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\right)$.

Lemma 2. Suppose $\hat{\theta} \in \Theta$ satisfies

$$\nabla_{\hat{\mathcal{A}}(\gamma)}\mathcal{L}(\hat{\theta}) = \nabla_{\hat{\mathcal{A}}(\gamma)}\mathcal{R}(\hat{\theta}, \gamma), \tag{8}$$

$$\left\|\nabla_{\hat{\mathcal{A}}(\gamma)^c}\mathcal{L}(\hat{\theta})\right\|_\infty < \gamma, \tag{9}$$

and

$$\omega_{min}\left(-\nabla^2_{\hat{\mathcal{A}}(\gamma)}\mathcal{L}(\hat{\theta}) + \nabla^2_{\hat{\mathcal{A}}(\gamma)}\mathcal{R}(\hat{\theta}, \gamma)\right) > 0, \tag{10}$$

then $\hat{\theta}$ is a local maximizer of $\mathcal{U}(\theta, \gamma)$, where $\hat{\mathcal{A}}(\gamma)^c$ is the complement of $\hat{\mathcal{A}}(\gamma)$.

Proof: Define $\Theta_{\hat{\mathcal{A}}(\gamma)} = \Theta \cap \left(\prod_{q=1}^{Q}\mathfrak{X}_q\right)$, where $\mathfrak{X}_q = \Re$ if $q \in \hat{\mathcal{A}}(\gamma)$ and $\mathfrak{X}_q = \{0\}$ otherwise. Let $\tilde{\mathcal{N}}$ denote a small neighborhood of $\hat{\theta}$ on $\Theta_{\hat{\mathcal{A}}(\gamma)}$. Equation (8) and (10) imply that $\hat{\theta}$ is the unique maximizer of $\mathcal{U}(\theta, \gamma)$ on $\tilde{\mathcal{N}}$ and hence a strictly local maximizer of $\mathcal{U}(\theta, \gamma)$ on $\Theta_{\hat{\mathcal{A}}(\gamma)}$. Now, we want to show that $\hat{\theta}$ is also a strictly local maximizer of $\mathcal{U}(\theta, \gamma)$ on Θ. Let $\mathcal{N}$ be a neighborhood of $\hat{\theta}$ on Θ such that $\mathcal{N} \cap \Theta_{\hat{\mathcal{A}}(\gamma)} \subset \tilde{\mathcal{N}}$. We claim that $\mathcal{U}(\hat{\theta}, \gamma) > \mathcal{U}(\vartheta, \gamma)$ for any $\vartheta \in \mathcal{N}\backslash\tilde{\mathcal{N}}$. Because $\hat{\theta}$ is the unique maximizer of $\mathcal{U}(\theta, \gamma)$ on $\tilde{\mathcal{N}}$, given any $\vartheta \in \mathcal{N}\backslash\tilde{\mathcal{N}}$, $\mathcal{U}(\hat{\theta}, \gamma) > \mathcal{U}(\tilde{\vartheta}, \gamma)$ holds, where $\tilde{\vartheta}$ is a projection of ϑ such that $\tilde{\vartheta}_q = \vartheta_q$ if $q \in \hat{\mathcal{A}}(\gamma)$ and $\tilde{\vartheta}_q = 0$ otherwise. Hence, it suffices to show that $\mathcal{U}(\tilde{\vartheta}, \gamma) > \mathcal{U}(\vartheta, \gamma)$ for any $\vartheta \in \mathcal{N}\backslash\hat{\theta}$. By the mean value theorem and the definition of $\tilde{\vartheta}$ and ϑ, we have

$$\mathcal{U}(\tilde{\vartheta}, \gamma) - \mathcal{U}(\vartheta, \gamma) = \nabla_{\hat{\mathcal{A}}(\gamma)^c}\mathcal{U}(\bar{\vartheta}, \gamma)^T\left(\tilde{\vartheta}_{\hat{\mathcal{A}}(\gamma)^c} - \vartheta_{\hat{\mathcal{A}}(\gamma)^c}\right)$$

$$= -\nabla_{\hat{\mathcal{A}}(\gamma)^c} \mathcal{U}(\bar{\vartheta},\gamma)^T \vartheta_{\hat{\mathcal{A}}(\gamma)^c}$$

$$= -\nabla_{\hat{\mathcal{A}}(\gamma)^c} \mathcal{L}(\bar{\vartheta},\gamma)^T \vartheta_{\hat{\mathcal{A}}(\gamma)^c} + \nabla_{\hat{\mathcal{A}}(\gamma)^c} \mathcal{R}(\bar{\vartheta},\gamma)^T \vartheta_{\hat{\mathcal{A}}(\gamma)^c}$$

$$= -\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \mathcal{L}(\bar{\vartheta})}{\partial|\vartheta_q|}\vartheta_q + \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \rho(|\bar{\vartheta}_q|,\gamma)}{\partial|\vartheta_q|} \operatorname{sign}(\bar{\vartheta}_q)\vartheta_q$$

$$= -\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \mathcal{L}(\bar{\vartheta})}{\partial|\vartheta_q|}\vartheta_q + \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \rho(|\bar{\vartheta}_q|,\gamma)}{\partial|\vartheta_q|} |\vartheta_q|, \tag{11}$$

where $\bar{\vartheta}$ lies on the line segment between $\tilde{\vartheta}$ and ϑ. Note that $\operatorname{sign}(\bar{\vartheta}_q)\vartheta_q = |\vartheta_q|$ because ϑ_q and $\bar{\vartheta}_q$ have the same sign. By $\|\nabla_{\hat{\mathcal{A}}(\gamma)^c}\mathcal{L}(\hat{\theta})\|_\infty < \gamma = \frac{\partial\rho(0+,\gamma)}{\partial t}$ in Equation (9), and the continuity of $\frac{\partial\rho(t,\gamma)}{\partial t}$ and $\tau(\theta)$ described in Condition E and B, there exists a $\varepsilon > 0$ such that for any θ in the neighborhood of $\hat{\theta}$ with radius ε we have

$$\|\nabla_{\hat{\mathcal{A}}(\gamma)^c}\mathcal{L}(\theta)\|_\infty < \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}. \tag{12}$$

Since the choice of $\mathcal{N}$ is arbitrary, we can choose $\mathcal{N}$ with radius smaller than ε so that $|\bar{\vartheta}_q| \le |\vartheta_q| < \varepsilon$ for $q \in \hat{\mathcal{A}}(\gamma)^c$. By the fact $\bar{\vartheta} \in \mathcal{N}$, Equation (12) implies that $\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \mathcal{L}(\bar{\vartheta})}{\partial|\vartheta_q|}\vartheta_q < \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}|\vartheta_q|$. Using the concavity of $\rho(t,\gamma)$ in t and the continuity of $\frac{\partial\rho(t,\gamma)}{\partial t}$, we further obtain $\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \rho(|\bar{\vartheta}_q|,\gamma)}{\partial|\vartheta_q|}|\vartheta_q| \ge \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}|\vartheta_q|$. Therefore, by $\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \mathcal{L}(\bar{\vartheta})}{\partial|\vartheta_q|}\vartheta_q < \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}|\vartheta_q|$ and $\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial \rho(|\bar{\vartheta}_q|,\gamma)}{\partial|\vartheta_q|}|\vartheta_q| \ge \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}|\vartheta_q|$, Equation (11) is strictly larger than

$$-\sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}|\vartheta_q| + \sum_{q\in\hat{\mathcal{A}}(\gamma)^c} \frac{\partial\rho(\varepsilon,\gamma)}{\partial t}|\vartheta_q| = 0. \tag{13}$$

which implies that $\mathcal{U}(\tilde{\vartheta},\gamma) - \mathcal{U}(\vartheta,\gamma) > 0$ for any $\vartheta \in \mathcal{N}\backslash\hat{\theta}$ such that $\|\vartheta - \hat{\theta}\| < \varepsilon$. We conclude that $\hat{\theta}$ is also a strictly local maximizer of $\mathcal{U}(\theta,\gamma)$ on Θ.

Lemma 3. Let $\hat{\mathcal{O}}$ denote the set containing all the strictly local maximizers of $\mathcal{U}(\theta,\gamma)$. If Conditions

A-E hold, γ satisfies $\gamma \to 0$ and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$, we have

$$\lim_{N\to\infty} \mathbb{P}\left(\tilde{\theta}^* \in \hat{\mathcal{O}}\right) = 1, \tag{14}$$

where $\tilde{\theta}^*$ is the ML estimator on the restricted parameter space $\Theta_{\mathcal{A}^*}$.

Proof: We want to show that $\tilde{\theta}^*$ satisfies Equations (8), (9), and (10) asymptotically, i.e., $\lim_{N\to\infty} \mathbb{P}(\mathcal{K}) = 1$, where $\mathcal{K} = \left\{\nabla_{\mathcal{A}^*}\mathcal{L}\left(\tilde{\theta}^*\right) = \nabla_{\mathcal{A}^*}\mathcal{R}\left(\tilde{\theta}^*\right)\right\} \cap \left\{\left\|\nabla_{\mathcal{A}^{*c}}\mathcal{L}\left(\tilde{\theta}^*\right)\right\|_\infty < \gamma\right\} \cap \left\{\omega_{min}\left(-\nabla^2_{\mathcal{A}^*}\mathcal{L}\left(\tilde{\theta}^*\right) + \nabla^2_{\mathcal{A}^*}\mathcal{R}\left(\tilde{\theta}^*, \gamma\right)\right) > 0\right\}$. Let $\mathcal{E} = \mathcal{E}_1 \cap \mathcal{E}_2 \cap \mathcal{E}_3$ with $\mathcal{E}_1$, $\mathcal{E}_2$, and $\mathcal{E}_3$ being

$$\mathcal{E}_1 = \left\{\min_{q\in\mathcal{A}^*}\left|\tilde{\theta}^*_q\right| > \delta\gamma\right\}, \tag{15}$$

$$\mathcal{E}_2 = \left\{\max_{q\in\mathcal{A}^{*c}}\left|\nabla_q\mathcal{L}\left(\tilde{\theta}^*\right)\right| < \gamma\right\}, \tag{16}$$

and

$$\mathcal{E}_3 = \left\{\omega_{min}\left(-\nabla^2_{\mathcal{A}^*}\mathcal{L}\left(\tilde{\theta}^*\right) + \nabla^2_{\mathcal{A}^*}\mathcal{R}\left(\tilde{\theta}^*\right)\right) > 0\right\}. \tag{17}$$

By $\frac{\partial\rho(t,\gamma)}{\partial t} = 0$ if $t > \delta\gamma$ described in Condition E, we have $\mathcal{E} \subseteq \mathcal{K}$. The de Morgan's law implies that the complement of $\mathcal{E}$, denoted by $\mathcal{E}^c$, is $\mathcal{E}_1^c \cup \mathcal{E}_2^c \cup \mathcal{E}_3^c$, where

$$\mathcal{E}_1^c = \cup_{q\in\mathcal{A}^*}\left\{\left|\tilde{\theta}^*_q\right| \le \delta\gamma\right\}, \tag{18}$$

$$\mathcal{E}_2^c = \cup_{q\in\mathcal{A}^{*c}}\left\{\left|\nabla_q\mathcal{L}\left(\tilde{\theta}^*\right)\right| \ge \gamma\right\}, \tag{19}$$

and

$$\mathcal{E}_3^c = \left\{\omega_{min}\left(-\nabla^2_{\mathcal{A}^*}\mathcal{L}\left(\tilde{\theta}^*\right) + \nabla^2_{\mathcal{A}^*}\mathcal{R}\left(\tilde{\theta}^*\right)\right) \le 0\right\}. \tag{20}$$

Because $\mathbb{P}(\mathcal{K}) \ge \mathbb{P}(\mathcal{E}) = 1 - \mathbb{P}(\mathcal{E}^c) > 1 - \sum_{k=1}^3 \mathbb{P}(\mathcal{E}_k^c)$, it suffices to show that $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_k^c) = 0$ for $k = 1, 2, 3$.

1. $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_1^c) = 0$.

By Lemma 1, we already know that for any $q \in \mathcal{A}^*$, $\tilde{\theta}^*_q$ is consistent to θ^*_q, which implies that $\mathbb{P}\left(\left|\tilde{\theta}^*_q\right| \le \delta\gamma\right) \to 0$ as $N \to \infty$ for $q \in \mathcal{A}^*$. Hence, we obtain that as $N \to \infty$

$$\mathbb{P}(\mathcal{E}_1^c) \leq \sum_{q\in\mathcal{A}^*} \mathbb{P}\big(\tilde{\theta}_q^* \leq \delta\gamma\big) \to 0. \tag{21}$$

2. $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_2^c) = 0.$

   We first observe that

$$\mathbb{P}\big(\big|\nabla_q \mathcal{L}(\tilde{\theta}^*)\big| \geq \gamma\big) \leq \mathbb{P}\big(\big|\nabla_q \mathcal{L}(\theta^*)\big| + \big|\nabla_q \mathcal{L}(\tilde{\theta}^*) - \nabla_q \mathcal{L}(\theta^*)\big| \geq \gamma\big)$$

$$\leq \mathbb{P}\big(\big|\nabla_q \mathcal{L}(\tilde{\theta}^*) - \nabla_q \mathcal{L}(\theta^*)\big| \geq \tfrac{\gamma}{2}\big) + \mathbb{P}\big(\big|\nabla_q \mathcal{L}(\theta^*)\big| \geq \tfrac{\gamma}{2}\big) = a_1 + a_2. \tag{22}$$

By Taylor's theorem and Cauchy-Schwarz inequality, it follows that

$$a_1 \leq \mathbb{P}\left(\left\|\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T}\right\| \left\|\tilde{\theta}_{\mathcal{A}^*}^* - \theta_{\mathcal{A}^*}^*\right\| > \tfrac{\gamma}{4}\right) + \mathbb{P}\big(O_P(N^{-1}) > \tfrac{\gamma}{4}\big) = a_{11} + a_{12}. \tag{23}$$

Note that

$$a_{11} \leq \mathbb{P}\big(\big\|\tilde{\theta}_{\mathcal{A}^*}^* - \theta_{\mathcal{A}^*}^*\big\| > \tfrac{1}{4}\big) + \mathbb{P}\left(\left\|\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T} - \mathbb{E}\left(\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T}\right)\right\| > \tfrac{\gamma}{2}\right)$$

$$+\mathbb{P}\left(\left\|\mathbb{E}\left(\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T}\right)\right\| > \tfrac{\gamma}{2}\right). \tag{24}$$

Because $\big\|\tilde{\theta}_{\mathcal{A}^*}^* - \theta_{\mathcal{A}^*}^*\big\|$ and $\left\|\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T} - \mathbb{E}\left(\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T}\right)\right\|$ are both $O_P\left(N^{-\frac{1}{2}}\right)$ and $\left\|\mathbb{E}\left(\frac{\partial \nabla_q \mathcal{L}(\theta^*)}{\partial \theta_{\mathcal{A}^*}^T}\right)\right\| > 0$, a_{11} converges to zero as $N \to \infty$. Clearly, a_{12} and a_2 also converge to zero by the fact $\big|\nabla_q \mathcal{L}(\theta^*)\big| = O_P\left(N^{-\frac{1}{2}}\right)$. Therefore, we conclude that $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_2^c) = 0$.

3. $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_3^c) = 0.$

   By Condition C, $\omega_{min}\left(-\nabla_{\mathcal{A}^*}^2 \mathcal{L}(\theta)\right) \geq \kappa_1$ on $\Omega_{\mathcal{A}^*}(\theta^*)$. Hence, for sufficiently large N and $\tilde{\theta}^* \in \Omega_{\mathcal{A}^*}(\theta^*)$, $\omega_{min}\left(-\nabla_{\mathcal{A}^*}^2 \mathcal{L}(\tilde{\theta}^*) + \nabla_{\mathcal{A}^*}^2 \mathcal{R}(\tilde{\theta}^*)\right) = \omega_{min}\left(-\nabla_{\mathcal{A}^*}^2 \mathcal{L}(\theta) + o(1)\right) > 0$ holds in probability, indicating $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_3^c) = 0$.

Lemma 1 shows that the ML estimator on the restricted parameter space, denoted by $\tilde{\theta}^*$, is consistent and asymptotically normal, which is just a standard result of ML estimator under misspecified likelihood (e.g., White, 1982). Lemma 2 gives the optimality condition for PL estimators (see also Fan & Lv, 2011). Lemma 3 indicates that asymptotically $\tilde{\theta}^*$ is also a local maximizer of the PL criterion (see also Kwon & Kim, 2012). Therefore, $\tilde{\theta}^*$ is of course an oracle estimator described in Theorem 1.

Theorem 2 (global oracle property). Under Conditions A-F and γ satisfies $\gamma \to 0$ and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$. Asymptotically, there exists a unique global maximizer of $\mathcal{U}(\theta, \gamma)$, denoted by $\hat{\theta}$, such that

(a) $\lim_{N\to\infty} \mathbb{P}\big(\hat{\mathcal{A}}(\gamma) = \mathcal{A}^*\big) = 1$;

(b) $\sqrt{N}\big(\hat{\theta}_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\big) \longrightarrow_{\mathcal{D}} \mathcal{N}\big(0, \mathcal{F}^{*}_{\mathcal{A}^*}{}^{-1}\mathcal{H}^*_{\mathcal{A}^*}\mathcal{F}^{*}_{\mathcal{A}^*}{}^{-1}\big)$.

Proof: Let $\tilde{\theta}^*$ denote the ML estimator on the restricted parameter space $\Theta_{\mathcal{A}^*}$. We only need to show that

$$\lim_{N\to\infty} \mathbb{P}\left(\mathcal{U}(\tilde{\theta}^*, \gamma) \geq \max_{\theta\in\Omega(\theta^*)} \mathcal{U}(\theta, \gamma)\right) = 1. \tag{25}$$

According to Taylor's theorem,

$$\mathcal{L}(\theta) - \mathcal{L}(\tilde{\theta}^*) = \nabla\mathcal{L}^T(\tilde{\theta}^*)(\theta - \tilde{\theta}^*) + \tfrac{1}{2}(\theta - \tilde{\theta}^*)^T\nabla^2\mathcal{L}(\bar{\theta}^*)(\theta - \tilde{\theta}^*). \tag{26}$$

By Lemma 3 and Condition F, for sufficiently large N, we have

$$\nabla\mathcal{L}^T(\tilde{\theta}^*)(\theta - \tilde{\theta}^*) \leq \gamma \textstyle\sum_{q\in\mathcal{A}^{*c}} |\theta_q|, \tag{27}$$

and

$$\tfrac{1}{2}(\theta - \tilde{\theta}^*)^T\nabla^2\mathcal{L}(\bar{\theta}^*)(\theta - \tilde{\theta}^*) \leq -\tfrac{1}{2}\kappa_2 \textstyle\sum_{q=1}^{Q}(\theta_q - \tilde{\theta}^*_q)^2. \tag{28}$$

Hence, for sufficiently large N, the following inequality holds

$$\mathcal{U}(\theta, \gamma) - \mathcal{U}(\tilde{\theta}^*, \gamma) \leq \textstyle\sum_{q=1}^{Q} a_q, \tag{29}$$

where

$$a_q = \begin{cases} -\frac{1}{2}\kappa_2\left(\theta_q - \tilde{\theta}_q^*\right)^2 + c_q\left[\rho\left(\left|\tilde{\theta}_q\right|, \gamma\right) - \rho\left(\left|\theta_q\right|, \gamma\right)\right] & \text{if } q \in \mathcal{A}^*, \\ \gamma\left|\theta_q\right| - \frac{1}{2}\kappa_2\left(\theta_q\right)^2 - c_q\rho\left(\left|\theta_q\right|, \gamma\right) & \text{if } q \in \mathcal{A}^{*c}. \end{cases} \tag{30}$$

For $q \in \mathcal{A}^*$, $-\frac{1}{2}\kappa_2\left(\theta_q - \tilde{\theta}_q^*\right)^2 < 0$ and $c_q\left[\rho\left(\left|\tilde{\theta}_q\right|, \gamma\right) - \rho\left(\left|\theta_q\right|, \gamma\right)\right] = 0$ hold asymptotically, which implies $a_q < 0$. For $q \in \mathcal{A}^{*c}$, by the fact that $\gamma \to 0$, the following inequality holds for sufficiently large N

$$a_q = \left|\theta_q\right|\left(\gamma - \frac{1}{2}\kappa_2\left|\theta_q\right|\right) - c_q\rho\left(\left|\theta_q\right|, \gamma\right) < 0. \tag{31}$$

Therefore, we conclude that $\mathbb{P}\left(\mathcal{U}\left(\tilde{\theta}^*, \gamma\right) \geq \max_{\theta \in \Omega(\theta^*)} \mathcal{U}(\theta, \gamma)\right) \to 1$.

Based on the result of Theorem 2, as long as we have a reliable algorithm to find the global maximizer, the global maximizer asymptotically performs as well as an oracle one. Note that the difference between Theorems 1 and 2 is that the latter requires the Fisher information matrix to be positive definite in the neighborhood of θ^* on the entire parameter space Θ, indicating that the specified model is at least locally in the neighborhood of θ^* on Θ. Therefore, if the specified model is not locally identified at θ^*, Theorem 2 would fail.

If Y is normally distributed and $\tau(\theta)$ is correctly specified, the information equality holds (i.e., ${\mathcal{F}_{\mathcal{A}^*}^*}^{-1} = \mathcal{H}_{\mathcal{A}^*}^*$) and Theorem 2 reduces to Corollary 1 below. The main implication of Corollary 1 is that under normality and correct model specification the PL estimator can achieve the Cramér-Rao lower bound, even when the true sparsity pattern is unknown beforehand. Furthermore, it also implies that $N \cdot \mathcal{D}_{ML}\left(\tau(\hat{\theta}), t\right)$ is asymptotically distributed as a chi-square random variable, where $\mathcal{D}_{ML}(\tau(\theta), t) = -\log\left|\Sigma(\theta)^{-1}S\right| + \mathrm{tr}(\Sigma(\theta)^{-1}S) - P + \left(\bar{Y} - \mu(\theta)\right)^T\Sigma(\theta)^{-1}\left(\bar{Y} - \mu(\theta)\right)$ and $t = (\mathrm{vech}(S)^T, \bar{Y}^T)^T$ with $S = \frac{1}{N}\sum_{n=1}^N (Y_n - \bar{Y})(Y_n - \bar{Y})^T$ and $\bar{Y} = \frac{1}{N}\sum_{n=1}^N Y_n$. Therefore, it is easy to construct an asymptotic $1 - \alpha$ level test for examining the null hypothesis $\tau = \tau(\theta)$ versus alternative hypothesis $\tau \neq \tau(\theta)$. Also, statistical tests for comparing several nested SEM models can be developed based on the result of sequential chi-square statistics (see Steiger, Shapiro, & Browne, 1985)

Corollary 1. Under Conditions A-F and γ satisfies $\gamma \to 0$ and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$. If the density of Y is actually $\varphi_\theta(y)$, then asymptotically, there exists a unique global maximizer of $\mathcal{U}(\theta, \gamma)$, denoted by $\hat{\theta}$, such that

(a) $\lim_{N\to\infty} \mathbb{P}\left(\hat{\mathcal{A}}(\gamma) = \mathcal{A}^*\right) = 1$;

(b) $\sqrt{N}\left(\hat{\theta}_{\mathcal{A}^*} - \theta^*_{\mathcal{A}^*}\right) \longrightarrow_{\mathcal{D}} \mathcal{N}\left(0, {\mathcal{F}^*_{\mathcal{A}^*}}^{-1}\right)$,

(c) $N \cdot \mathcal{D}_{ML}\left(\tau(\hat{\theta}), t\right) \longrightarrow_{\mathcal{D}} \chi^2_{df^*}$, where $df^* = P(P+3)/2 - |\mathcal{A}^*|$.

Now, the asymptotic properties of AIC and BIC are derived under the framework of the proposed PL method. Given a model $\tau(\theta)$, for any index set $\mathcal{A} \subset \{1,2, \ldots, Q\}$, the MDF value of $\tau(\theta)$ on $\Theta_{\mathcal{A}}$ is defined as

$$\mathcal{D}^*_{\mathcal{A}} = \min_{\theta \in \Theta_{\mathcal{A}}} \mathcal{D}_{ML}(\tau(\theta), \tau^*). \tag{32}$$

where $\tau^* = \left(\text{vech}(\Sigma^*)^T, \mu^{*T}\right)^T$ is the true moment vector. Hence, by examining the values of $\mathcal{D}^*_{\mathcal{A}}$ and $\mathcal{D}^*_{\mathcal{A}'}$, the correctness of $\tau(\theta)$ restricted on $\Theta_{\mathcal{A}}$ and $\Theta_{\mathcal{A}'}$ can be compared. According to the definition of $\mathcal{A}^*$, $\mathcal{D}^*_{\mathcal{A}^*} \leq \mathcal{D}^*_{\mathcal{A}}$ for any $\mathcal{A} \subset \{1,2, \ldots, Q\}$. If some $\mathcal{A}$ satisfies $\mathcal{D}^*_{\mathcal{A}^*} = \mathcal{D}^*_{\mathcal{A}}$, Condition D indicates that $\mathcal{A}^*$ is still more parsimonious than $\mathcal{A}$, i.e., $|\mathcal{A}^*| < |\mathcal{A}|$. Given a random sample $\mathcal{Y}_N$, the set of regularization parameters is partitioned into three subsets

$$\Gamma^* = \left\{\gamma | \mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)} = \mathcal{D}^*_{\mathcal{A}^*}, \left|\hat{\mathcal{A}}(\gamma)\right| = |\mathcal{A}^*|\right\}, \tag{33}$$

$$\Gamma^+ = \left\{\gamma | \mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)} = \mathcal{D}^*_{\mathcal{A}^*}, \left|\hat{\mathcal{A}}(\gamma)\right| > |\mathcal{A}^*|\right\}, \tag{34}$$

and

$$\Gamma^{-} = \left\{\gamma | \mathcal{D}^{*}_{\hat{\mathcal{A}}(\gamma)} > \mathcal{D}^{*}_{\mathcal{A}^{*}}\right\}. \tag{35}$$

The subset Γ^{*} contains all the values of γ where the optimal model $\mathcal{A}^{*}$ is attained. On the other hand, Γ^{+} and Γ^{-} are formed by γ such that the corresponding models are overfitted and underfitted, respectively. Note that $\hat{\mathcal{A}}(\gamma)$ with $\gamma \in \Gamma^{+}$ may not be really "overfitting" in the usual sense. An overfitting model is generally used to refer a model that explains the phenomenon perfectly but contains unnecessary parameters. However, "overfitting" here is merely used to emphasize that $\hat{\mathcal{A}}(\gamma)$ contains unnecessary parameters because it is possible that $\mathcal{D}^{*}_{\hat{\mathcal{A}}(\gamma)} > 0$. Given any estimated support $\hat{\mathcal{A}}(\gamma)$, $\tilde{\theta}(\gamma)$ is used to denote a global maximizer of $\mathcal{L}(\theta)$ on $\hat{\mathcal{A}}(\gamma)$.

Theorem 3. Let $\hat{\gamma}^{AIC}$ and $\hat{\gamma}^{BIC}$ denote the selection results based on AIC and BIC respectively. Under Conditions A-F, we have

(a) $\lim_{N\to\infty} \mathbb{P}(\hat{\gamma}^{AIC} \in \Gamma^{-}) = 0$ and $\lim_{N\to\infty} \mathbb{P}(\hat{\gamma}^{AIC} \in \Gamma^{+}) > 0$;

(b) $\lim_{N\to\infty} \mathbb{P}(\hat{\gamma}^{BIC} \in \Gamma^{*}) = 1$.

Proof: To prove first part of (a), we want to show that the probability of $\mathcal{E}_1 = \bigcup_{\gamma' \in \Gamma^{*} \cup \Gamma^{+}} \left\{\inf_{\gamma \in \Gamma^{-}} AIC(\gamma) - AIC(\gamma') > 0\right\}$ converges to one. Let $t = (\text{vech}(S)^{T}, \bar{Y}^{T})^{T}$ denote a vector of sample moment, where $S = \frac{1}{N}\sum_{n=1}^{N}(Y_n - \bar{Y})(Y_n - \bar{Y})^{T}$ and $\bar{Y} = \frac{1}{N}\sum_{n=1}^{N} Y_n$. We use $\mathcal{D}(\theta) = \mathcal{D}_{ML}(\tau(\theta), t)$ to represent the sample discrepancy evaluated at θ. By the fact that $\mathcal{D}\left(\tilde{\theta}(\gamma)\right) \leq \mathcal{D}\left(\hat{\theta}(\gamma)\right)$ and $\left\{\inf_{\gamma \in \Gamma^{-}} AIC(\gamma) - AIC(0) > 0\right\} \subset \mathcal{E}_1$, the following inequality holds

$$\begin{aligned}\mathbb{P}(\mathcal{E}_1) &\geq \mathbb{P}\left(\inf_{\gamma \in \Gamma^{-}} AIC(\gamma) - AIC(0) > 0\right) \\ &\geq \mathbb{P}\left(\min_{\mathcal{A}(\gamma) \in \{\mathcal{A} | \mathcal{A}^{*} \not\subset \mathcal{A}\}} \mathcal{D}\left(\tilde{\theta}(\gamma)\right) - \mathcal{D}(\tilde{\theta}) - \frac{2}{N}Q > 0\right). \end{aligned} \tag{36}$$

Note that $\lim_{N\to\infty} \min_{\mathcal{A}(\gamma)\in\{\mathcal{A}|\mathcal{A}^*\not\subset\mathcal{A}\}} \mathcal{D}\left(\tilde{\theta}(\gamma)\right) \geq \min_{\mathcal{A}\in\{\mathcal{A}|\mathcal{A}^*\not\subset\mathcal{A}\}} \mathcal{D}_{\mathcal{A}}^* > \mathcal{D}_{\mathcal{A}^*}^*$ and $\lim_{N\to\infty} \mathcal{D}(\tilde{\theta}) = \mathcal{D}_{\mathcal{A}^*}^*$. Hence,

$$\mathbb{P}(\mathcal{E}_1) \geq \mathbb{P}\left(\min_{\mathcal{A}\in\{\mathcal{A}|\mathcal{A}^*\not\subset\mathcal{A}\}} \mathcal{D}_{\mathcal{A}}^* - \mathcal{D}_{\mathcal{A}^*}^* - o_p(1) > 0\right) \to 1. \tag{37}$$

For proving the second part of (a), we need to show that the probability of $\mathcal{E}_2 = \cup_{\gamma'\in\Gamma^+}\left\{\inf_{\gamma\in\Gamma^*} AIC(\gamma) - AIC(\gamma') > 0\right\}$ is larger than some nonzero constant. Again, by the fact $\left\{\inf_{\gamma\in\Gamma^*} AIC(\gamma) - AIC(0) > 0\right\} \subset \mathcal{E}_2$ and $\inf_{\gamma\in\Gamma^*} AIC(\gamma) > \mathcal{D}(\tilde{\theta}^*) + \frac{2}{N}|\mathcal{A}^*|$, we have that

$$\mathbb{P}(\mathcal{E}_2) \geq \mathbb{P}\left(\inf_{\gamma\in\Gamma^*} AIC(\gamma) - AIC(0) > 0\right)$$

$$\geq \mathbb{P}\left(\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}(\tilde{\theta}) + \frac{2}{N}(|\mathcal{A}^*| - Q) > 0\right). \tag{38}$$

According to the result of White (1982), we have that $N\left(\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}(\tilde{\theta})\right)$ is distributed as a mixture of chi-square random variables asymptotically. Hence, we conclude that

$$\mathbb{P}(\mathcal{E}_2) \geq \mathbb{P}\left(N\left(\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}(\tilde{\theta})\right) > 2(Q - |\mathcal{A}^*|)\right) > 0. \tag{39}$$

To prove (b), by Theorem 2 we already derived that if γ satisfies $\gamma \to 0$ and $\sqrt{N}\gamma \to \infty$ as $N \to \infty$, then the $\lim_{N\to\infty} \mathbb{P}\left(BIC(\gamma) = \mathcal{D}(\tilde{\theta}^*) + \frac{\log N}{N}|\mathcal{A}^*|\right) = 1$, which implies that asymptotically Γ^* is not empty if we set $\Gamma = [0, L]$ for a sufficiently large L. The question remains whether BIC can select a $\gamma^* \in \Gamma^*$. Now, we want to show that for any $\gamma^* \in \Gamma^*$, the probability of $\mathcal{E}_1 = \cup_{\gamma\in\Gamma^-\cup\Gamma^+}\{BIC(\gamma^*) - BIC(\gamma) > 0\}$ converge to zero. By the fact that $\left\{\hat{\mathcal{A}}(\gamma)\right\}_{\gamma\in\Gamma} \subset \{\mathcal{A}\}$ and $BIC(\gamma) \geq \mathcal{D}\left(\tilde{\theta}(\gamma)\right) + \frac{\log N}{N} e(\gamma)$, $\mathcal{E}_1$ is contained in a set $\mathcal{E}_2$, a union of finite sets,

$$\mathcal{E}_2 = \cup_{\mathcal{A}(\gamma')\neq\mathcal{A}^*}\left\{BIC(\gamma^*) - \left(\mathcal{D}\left(\tilde{\theta}(\gamma)\right) + \frac{\log N}{N} e(\gamma)\right) > 0\right\}. \tag{40}$$

If for any $\hat{\mathcal{A}}(\gamma) \neq \mathcal{A}^*$, $\lim_{N\to\infty} \mathbb{P}\left(BIC(\gamma^*) - \left(\mathcal{D}\left(\tilde{\theta}(\gamma)\right) + \frac{\log N}{N} e(\gamma)\right) > 0\right) = 0$ holds, then by the fact $\mathcal{E}_1 \subset \mathcal{E}_2$, $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_1) = 0$ must be true. If $\hat{\mathcal{A}}(\gamma) \neq \mathcal{A}^*$ but $\hat{\mathcal{A}}(\gamma) \supset \mathcal{A}^*$, by the fact that

$BIC(\gamma^*) = \mathcal{D}(\tilde{\theta}^*) + \frac{\log N}{N}|\mathcal{A}^*|$ with probability tending to one, it suffices to show that the probability of $\left\{\left(\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}\left(\tilde{\theta}(\gamma)\right)\right) + \frac{\log N}{N}(|\mathcal{A}^*| - e(\gamma)) > 0\right\}$ can be arbitrarily small. By the fact $\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}\left(\tilde{\theta}(\gamma)\right) = O_p(N^{-1})$ and $|\mathcal{A}^*| < e(\gamma)$, we have

$$\mathbb{P}\left(\left(\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}\left(\tilde{\theta}(\gamma)\right)\right) + \frac{\log N}{N}(|\mathcal{A}^*| - e(\gamma)) > 0\right)$$

$$= \mathbb{P}\left(O_p(N) > \frac{\log N}{N}(e(\gamma) - |\mathcal{A}^*|)\right) \to 0, \quad (41)$$

as $N \to \infty$. For $\hat{\mathcal{A}}(\gamma) \neq \mathcal{A}^*$ but $\hat{\mathcal{A}}(\gamma) \not\supset \mathcal{A}^*$,

$$\mathbb{P}\left(\left(\mathcal{D}(\tilde{\theta}^*) - \mathcal{D}\left(\tilde{\theta}(\gamma)\right)\right) + \frac{\log N}{N}(|\mathcal{A}^*| - e(\gamma)) > 0\right)$$

$$= \mathbb{P}\left(\mathcal{D}^*_{\mathcal{A}^*} - \mathcal{D}^*_{\hat{\mathcal{A}}(\gamma)} + o_p(1) > 0\right) \to 0. \quad (42)$$

as $N \to \infty$. Therefore, we conclude that $\lim_{N\to\infty} \mathbb{P}(\mathcal{E}_1) = 0$ and hence $\lim_{N\to\infty} \mathbb{P}(\hat{\gamma}^{BIC} \in \Gamma^*) = 1$.

Theorems 3 shows that asymptotically both AIC and BIC select a model that attains the smallest MDF value $\mathcal{D}^*_{\mathcal{A}^*}$. However, only BIC yields a consistent selection result with respect to $\mathcal{A}^*$. AIC may suffer from the problem of overfitting. Of course, if Γ^+ is empty, AIC can also select the quasi-true model with probability one. The derived results are consistent with the typical behaviors of AIC and BIC in parametric regression models (e.g., Zhang, Li, & Tsai, 2012; Shao, 1997).

References

Browne, M. W. (1984). Asymptotic distribution-free methods for the analysis of covariance structures. *British Journal of Mathematical and Statistical Psychology, 37,* 62–83.

Fan, J.-Q., & Lv, J.-C. (2011). Non-concave penalized likelihood with NP-Dimensionality. *IEEE - Information Theory, 57,* 5467–5484.

Kwon, S., & Kim, Y. (2012). Large sample properties of the SCAD-penalized maximum likelihood estimation on high dimensions. *Statistica Sinica, 22,* 629–653.

Lehmann, E., & Casella, G. (1998), *Theory of point estimation* (2nd ed.). New York: Springer-Verlag.

Shao, J. (1997). n asymptotic theory for linear model selection. *Statistica Sinica, 7,* 221–264.

Shapiro, A. (1983). Asymptotic distribution theory in the analysis of covariance structures (a unified theory). *South African Statistical Journal, 17,* 33–81.

Steiger, J., Shapiro, A., & Browne, M. (1985). On the multivariate asymptotic distribution of sequential Chi-square statistics. *Psychometrika, 50,* 253–263.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica, 50,* 1–25.

Zhang, Y.-Y., Li, R.-Z., & Tsai, C.-L. (2010). Regularization parameter selections via generalized information criterion. Journal of the American Statistical Association, 105, 312–323.