Submitted to the Annals of Statistics
arXiv: arXiv:0000.0000

A NEW SCOPE OF PENALIZED EMPIRICAL LIKELIHOOD WITH HIGH-DIMENSIONAL ESTIMATING EQUATIONS

By Jinyuan Chang, Cheng Yong Tang and Tong Tong Wu

Southwestern University of Finance and Economics, Temple University, and University of Rochester

Statistical methods with empirical likelihood (EL) are appealing and effective especially in conjunction with estimating equations through which useful data information can be adaptively and flexibly incorporated. It is also known in the literature that EL approaches encounter difficulties when dealing with problems having high-dimensional model parameters and estimating equations. To overcome the challenges, we begin our study with a careful investigation on high-dimensional EL from a new scope targeting at estimating high-dimensional sparse model parameters. We show that the new scope provides an opportunity for relaxing the stringent requirement on the dimensionality of the model parameters. Motivated by the new scope, we then propose a new penalized EL by applying two penalty functions respectively regularizing the model parameters and the associated Lagrange multiplier in the optimizations of EL. By penalizing the Lagrange multiplier to encourage its sparsity, a drastic dimension reduction in the number of estimating equations can be effectively achieved without compromising the validity and consistency of the resulting estimators. Most attractively, such a reduction in dimensionality of estimating equations is actually equivalent to a selection among those high-dimensional estimating equations, resulting in a highly parsimonious and effective device for high-dimensional sparse model parameters. Allowing both the dimensionalities of model parameters and estimating equations growing exponentially with the sample size, our theory demonstrates that our new penalized EL estimator is sparse and consistent with asymptotically normally distributed nonzero components. Numerical simulations and a real data analysis show that the proposed penalized EL works promisingly.

[*] We are grateful to the Co-Editor, the Associate Editor and three referees for very constructive comments and suggestions that have greatly improved our paper. Chang was supported in part by the Fundamental Research Funds for the Central Universities (Grant No. JBK171121, JBK170161, JBK150501), NSFC (Grant No. 11501462), the Center of Statistical Research at SWUFE, and the Joint Lab of Data Science and Business Intelligence at SWUFE. Tang was support in part by NSF Grants IIS-1546087 and SES-1533956. Wu was supported in part by NIH Grants R01HL094572 and R01HL119058.

Primary 62G99, Secondary 62F40

Keywords and phrases: Empirical likelihood; Estimating equations; High-dimensional statistical methods; Moment selection; Penalized likelihood.

1. Introduction.

Statistical approaches using estimating equations are widely applicable to solve a broad class of practical problems. The most influential special cases of estimating equations include the fundamental maximum likelihood score equations and those from the popular generalized methods of moments (GMM, hereinafter) (Hansen, 1982). The approaches of using estimating equations are particularly appealing in practice with merits from requiring less stringent distributional assumptions on the data model, yet being adaptable to flexibly incorporate suitable information and conditions extracted from practical features in various scenarios of interests.

Empirical likelihood (EL, hereinafter) (Owen, 2001) coupled with estimating equations has been demonstrated successful since the seminal work of Qin and Lawless (1994). It is particularly appealing that the EL estimator asymptotically achieves the semiparametric efficiency bound (Qin and Lawless, 1994). The properties of EL are also desirable through some higher order analyses (Chen and Cui, 2006, 2007; Newey and Smith, 2004). Moreover, the Wilks' theorems (Owen, 1988, 1990; Qin and Lawless, 1994) for EL ensure that EL ratio is asymptotically χ^2-distributed when evaluated at the truth. Hence, EL provides an analogous device to the conventional fully parametric likelihood for statistical inferences, but without requiring more stringent distributional assumptions.

In recent years, high data dimensionality in practice has attracted increasing attention and brought unprecedented challenges to approaches based on estimating equations and EL. Studies in Chen et al. (2009), Hjort et al. (2009), Tang and Leng (2010), Leng and Tang (2012), and Chang et al. (2015) reveal that the conventional EL only works when both the dimensionality of the model parameters p and the number of the estimating equations r diverge at some rate slower than the sample size n. However, challenges due to high-dimensionality require a capacity to deal with cases where $p, r \gg n$. Tang and Leng (2010), Leng and Tang (2012), and Chang et al. (2015) attempt to utilize sparsity of the model parameters by applying penalty functions on those parameters, and show that sparse estimators with good properties are achievable. However, the restriction from the data dimensionality is not alleviated by using penalized EL in their works.

The challenges for EL from high data dimensionality are well documented in the literature. Tsao (2004) found that for fixed n with moderately large fixed p, the probability that the truth is contained in the EL based confidence region can be substantially smaller than the nominal level, resulting in the under-coverage problems. As remedies, Tsao and Wu (2013, 2014) propose extended EL to address the under-coverage problems due to the constraints on the parameter space. With a modification avoiding equality constraints, Bartolucci (2007) propose a penalized EL method via optimizing products of probability weights penalized by a loss function depending on the model parameters. Lahiri and Mukhopadhyay (2012) propose a different type of loss from that in Bartolucci (2007) and study its properties with high-dimensional model parameters and dependent data. To our best knowledge, no estimation problems have been investigated with the EL formulations of Bartolucci (2007) and Lahiri and Mukhopadhyay (2012).

In this paper, we study the properties of EL by carefully examining the impacts from the data dimensionality, and explore the opportunity from targeting at the sparse model parameters. We find that consistently estimating high-dimensional sparse model parameters by a penalized EL is feasible even $r < p$. This motivates us to propose a new penalized EL approach to tackle high-dimensional statistical problems where both p and r can grow at exponential rates of n. We solve the problem by employing two penalty functions when constructing the EL with high-dimensional estimating equations. Specifically, the first penalty function is on the magnitude of the model parameters to obtain sparse estimator. The second penalty function is imposed on the Lagrange multiplier to encourage its sparsity when optimizing the EL. We also observe that obtaining a sparse Lagrange multiplier in EL is equivalent to reducing the dimensionality r via an effective selection among those estimating equations, which itself is an interesting problem and a new scope; see our discussions in Sections 2 and 3.

Here we note that the effect of the penalty on the Lagrange multiplier relates to moment selection in the GMM, a problem that has been extensively studied in the econometrics literature; see, among others, Cheng and Liao (2015) and reference therein. Recently, Cheng and Liao (2015) and Shi (2016) study the problem with many moment conditions for estimating a fixed dimensional model parameter. Cheng and Liao (2015) propose to treat the sample averages of the moment conditions as additional parameters to be optimized, and to apply the L_1 penalty on them to encourage sparsity so that effective moment selection can be achieved. The role of the L_1 penalty in their approach is seen similar to ours on the Lagrange multiplier for the purpose of moment selection. In light of Dantzig selector (Candes and Tao, 2007), Shi (2016) propose a new EL formulation by relaxing the equality constraints to inequality ones involving some regularization parameter, so that effective moment selection is also achieved. Nevertheless, none of Cheng and Liao (2015) and Shi (2016) consider the impacts from diverging number of model parameters that potentially can be sparse.

Our investigation contributes to the area of EL with high-dimensional statistical problems from a new scope. Our approach successfully extends the EL approach with estimating functions to scenarios allowing both p and r growing exponentially with n. New results for high-dimensional penalized EL are established in Sections 2 and 3, and many of them are interesting in both areas of EL and estimating equations. Our analysis first reveals a result of its own interests that substantially broadens the understanding of the relationship between the number of estimating equations r and the number of model parameters p with penalized EL. Surprisingly, we find that with an appropriate penalization, a consistent and sparse estimator of the model parameters actually does not require $r \geq p$, thanks to the new scope from estimating sparse model parameters. In particular, we show that a sparse estimator with s nonzero components for the p-dimensional parameter technically may only require $r \geq s$. Such a result crucially supports the motivation in our new penalized EL approach for the second penalty function imposed on the Lagrange multiplier to reduce the effective number of estimating equations actually involved in the high-dimensional penalized EL. That is, the resulting sparse Lagrange multiplier from the penalization is equivalent to a selection among available estimating equations for the model parameters. Our theory shows that the new penalized EL estimator is consistent which estimates the zero components of the model parameters as zero with probability tending to one. Additionally, the nonzero components of the new penalized EL estimator is asymptotically normally distributed.

The rest of this paper is organized as follows. The new scope with high-dimensional sparse model parameters on EL and penalized EL is studied in Section 2. The new penalized EL with an additional penalty function on the Lagrange multiplier and its properties for estimating high-dimensional sparse model parameters are given in Section 3. An algorithm using coordinate descent is stated in Section 4. Numerical studies are shown in Section 5. Some discussions are given in Section 6. All technical details are provided in Section 7. The Supplementary Material contains more technical proofs.

2. Empirical likelihood and penalized empirical likelihood.

2.1. *An overview of EL with diverging dimensionality.* We define some notations first. For a matrix $\mathbf{B} = (b_{ij})_{s_1 \times s_2}$, let $|\mathbf{B}|_\infty = \max_{1\leq i\leq s_1, 1\leq j\leq s_2} |b_{ij}|$, $\|\mathbf{B}\|_1 = \max_{1\leq j\leq s_2} \sum_{i=1}^{s_1} |b_{ij}|$, $\|\mathbf{B}\|_\infty = \max_{1\leq i\leq s_1} \sum_{j=1}^{s_2} |b_{ij}|$ and $\|\mathbf{B}\|_2 = \lambda_{\max}^{1/2}(\mathbf{B}\mathbf{B}^{\mathrm{T}})$ where $\lambda_{\max}(\mathbf{B}\mathbf{B}^{\mathrm{T}})$ denotes the largest eigenvalue of $\mathbf{B}\mathbf{B}^{\mathrm{T}}$. Specifically, if $s_2 = 1$, we use $|\mathbf{B}|_1 = \sum_{i=1}^{s_1} |b_{i1}|$ and $|\mathbf{B}|_2 = (\sum_{i=1}^{s_1} b_{i1}^2)^{1/2}$ to denote the L_1-norm and L_2-norm of the s_1-dimensional vector $\mathbf{B}$, respectively.

Let $\mathbf{X}_1, \ldots, \mathbf{X}_n$ be d-dimensional i.i.d. observations and $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^{\mathrm{T}}$ be a p-dimensional parameter with support $\boldsymbol{\Theta}$. For an r-dimensional estimating function $\mathbf{g}(\mathbf{X}; \boldsymbol{\theta}) = \{g_1(\mathbf{X}; \boldsymbol{\theta}), \ldots, g_r(\mathbf{X}; \boldsymbol{\theta})\}^{\mathrm{T}}$, the information for the model parameter $\boldsymbol{\theta}$ is collected by the unbiased moment condition

$$\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta}_0)\} = \mathbf{0}, \tag{2.1}$$

where $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ is the unknown truth. When n grows, following Hjort et al. (2009) and Chang et al. (2015), the observations $\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}_{i=1}^n$ can be viewed as a triangular array where r, p, d, $\mathbf{X}_i$, $\boldsymbol{\theta}$ and $\mathbf{g}(\mathbf{X};\boldsymbol{\theta})$ may all depend on n. Qin and Lawless (1994) investigate an EL with estimating equations:

$$L(\boldsymbol{\theta}) = \sup\bigg\{\prod_{i=1}^n \pi_i : \pi_i > 0,\ \sum_{i=1}^n \pi_i = 1,\ \sum_{i=1}^n \pi_i \mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta}) = \mathbf{0}\bigg\}. \tag{2.2}$$

The so-called EL estimator is defined as $\widehat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$, which is equivalent to solve the corresponding dual problem:

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}, \tag{2.3}$$

where $\widehat{\Lambda}_n(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^r : \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta}) \in \mathcal{V},\ i = 1,\ldots,n\}$ for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\mathcal{V}$ is an open interval containing zero.

In a conventional setting where p and r are fixed as $n \to \infty$, $r \geq p$ is required to ensure all components of $\boldsymbol{\theta}$ are identifiable. In high-dimensional cases, however, it is documented in the literature that accommodating a diverging r is a key difficulty for EL; see, among others, Hjort et al. (2009), Chen et al. (2009), Leng and Tang (2012), and Chang et al. (2015). The reason is that the Lagrange multiplier $\boldsymbol{\lambda}$ in (2.3) is of the same high dimensionality r. Since $|\boldsymbol{\lambda}|_2$ is required to be $o_p(1)$ in theoretical analyses of EL, high-dimensional r is clearly cumbersome. A direct consequence is that dimensionality p and r for EL in (2.2) can only be accommodated at some polynomial rate of the sample size n. To explore EL with high-dimensional problems, we first present a general result for $\widehat{\boldsymbol{\theta}}$ with r estimating equations.

Proposition 1. *Write* $\widehat{\mathbf{V}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})^{\mathrm{T}}$ *and* $\bar{\mathbf{g}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})$. *Assume that there are uniform constants* $C_1 > 0$, $C_2 > 1$ *and* $\gamma > 2$ *such that*

$$\max_{1\leq j\leq r} \mathbb{E}\bigg\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |g_j(\mathbf{X}_i;\boldsymbol{\theta})|^{\gamma}\bigg\} \leq C_1, \tag{2.4}$$

and

$$\mathbb{P}\bigg[C_2^{-1} \leq \inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\min}\{\widehat{\mathbf{V}}(\boldsymbol{\theta})\} \leq \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \lambda_{\max}\{\widehat{\mathbf{V}}(\boldsymbol{\theta})\} \leq C_2\bigg] \to 1. \tag{2.5}$$

If $r = o(n^{1/2-1/\gamma})$, *then* $\widehat{\boldsymbol{\theta}}$ *defined in* (2.3) *satisfies* $|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2 = O_p(r^{1/2}n^{-1/2})$.

Conditions for Proposition 1 are standard. Condition (2.4) ensures that some moments with order larger than 2 exist for the estimating functions, and (2.5) says that the sample covariance matrices of the estimating functions should behave reasonably well. Consistent with the finding in Hjort et al. (2009) and Chen et al. (2009), the higher the order of the moment γ is, the more estimating functions can be accommodated. When the estimating functions are bounded, $\gamma = \infty$, r is allowed to be $o(n^{1/2})$. The key implication of Proposition 1 is that the sample mean of the estimating functions is well behaving, regardless the number of the model parameters p is. With r estimating functions, the optimum $|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2$ is $O_p(r^{1/2}n^{-1/2})$. Hence the impact on the behavior of the estimating function is the dimensionality r, which cannot grow faster than $n^{1/2}$ as $n \to \infty$.

Clearly, the impact from p on the EL estimator is on the identifiability of the model parameters. $\widehat{\boldsymbol{\theta}}$ is not uniquely defined when $r < p$ without further constraints, rendering ambiguity and inapplicability for estimating high-dimensional model parameters. An example of the situation is that the identifiability issue happens in the classic linear models if the design matrix is not of full column rank, so that the minimum of the least squares criterion function well exists but the ordinary least squares estimator is not uniquely defined in that case. To solve the problem, our next objective is to illustrate that identifying a sparse p-dimensional model parameter is still feasible.

2.2. *High-dimensional sparse model parameters.* The intuition here is that if one concerns a sparse $\boldsymbol{\theta}_0$ such that most of its components are zeros, then identification and estimation of such $\boldsymbol{\theta}_0$ are feasible with fewer estimating functions by EL with appropriate penalization. Specifically, write $\boldsymbol{\theta}_0 = (\theta_1^0, \ldots, \theta_p^0)^{\mathrm{T}}$ and let $\mathcal{S} = \{1 \le k \le p : \theta_k^0 \ne 0\}$ with $s = |\mathcal{S}|$. Here $\mathcal{S}$ is unknown, and $s \ll p$. Without loss of generality, let $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{0,\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}$ where $\boldsymbol{\theta}_{0,\mathcal{S}} \in \mathbb{R}^s$ is the nonzero components and $\boldsymbol{\theta}_{0,\mathcal{S}^c} = \mathbf{0} \in \mathbb{R}^{p-s}$. For identification of $\boldsymbol{\theta}_0$, we impose the following condition.

Condition 1. Assume that

$$\inf_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}=(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}},\boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}\in\boldsymbol{\Theta}:|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_\infty>\varepsilon,\boldsymbol{\theta}_{\mathcal{S}^c}=\mathbf{0}\}} |\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty \ge \Delta(\varepsilon) \tag{2.6}$$

for any $\varepsilon > 0$, where $\Delta(\cdot)$ is a function satisfying $\liminf_{\varepsilon\to 0^+} \varepsilon^{-\beta}\Delta(\varepsilon) \ge K_1$ for some uniform constants $K_1 > 0$ and $\beta > 0$.

The identification condition (2.6) can be viewed as a dedicated one for estimating sparse model parameters. Condition 1 is not stringent, and it ensures identifying the nonzero components of $\boldsymbol{\theta}_0$ locally. Studying local optimums in high-dimensional statistical problems is common in the literature with reasonable technical conditions; see, for example, Lv and Fan (2009) and Zhang (2010). Condition 1 means that the expected values of the estimating functions at the truth adequately differ from those outside a small neighborhood of the sparse support of $\boldsymbol{\theta}_0$. Here β is some generic constant related to the consistency result in Proposition 2. For estimating a high-dimensional mean parameter with $\mathbf{g}(\mathbf{X};\boldsymbol{\theta}) = \mathbf{X} - \boldsymbol{\theta}$, we can choose $\Delta(\varepsilon) = \varepsilon$. For linear models, $\mathbf{g}(\mathbf{X};\boldsymbol{\theta}) = \mathbf{Z}(Y - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\theta})$ with $\mathbf{Z}$ and Y being the covariates and response variable respectively, and $\mathbf{X} = (Y, \mathbf{Z}^{\mathrm{T}})^{\mathrm{T}}$, we can select $\Delta(\varepsilon) = \varepsilon\|\boldsymbol{\Sigma}_{\mathbf{Z},\mathcal{S}}^{-1}\|_\infty^{-1}$ with $\boldsymbol{\Sigma}_{\mathbf{Z},\mathcal{S}} = \mathbb{E}(\mathbf{Z}_\mathcal{S}\mathbf{Z}_\mathcal{S}^{\mathrm{T}})$. More generally, if there is a subset $\mathcal{E} \subset \{1,\ldots,r\}$ with $|\mathcal{E}| = s$ and $[\mathbb{E}\{\nabla_{\boldsymbol{\theta}_\mathcal{S}}\mathbf{g}_\mathcal{E}(\mathbf{X}_i;\boldsymbol{\theta})\}]^{-1}$ exists, where $\mathbf{g}_\mathcal{E}(\cdot)$ collects the set of estimating functions indexed by $\mathcal{E}$, then we can select $\Delta(\varepsilon) = \varepsilon\inf_{\boldsymbol{\theta}\in\{\boldsymbol{\theta}=(\boldsymbol{\theta}_\mathcal{S}^{\mathrm{T}},\boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}:\boldsymbol{\theta}_{\mathcal{S}^c}=\mathbf{0}\}}\|[\mathbb{E}\{\nabla_{\boldsymbol{\theta}_\mathcal{S}}\mathbf{g}_\mathcal{E}(\mathbf{X}_i;\boldsymbol{\theta})\}]^{-1}\|_\infty^{-1}$. Intuitively, Condition 1 ensures the identifiability of the s nonzero components of $\boldsymbol{\theta}_0$ so that a consistent sparse estimator is possible, provided that $r \geq s$ and conditions in Proposition 2. As a special case when $\mathcal{S}^c$ is empty, Condition 1 becomes a global identification for the dense model parameter $\boldsymbol{\theta}_0$. Similar global identification conditions can be found in Chen (2007) and Chen and Pouzo (2012) for some other models.

To estimate sparse $\boldsymbol{\theta}_0$, we consider a penalized EL estimator as

$$\widetilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})}\left[\sum_{i=1}^n\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} + n\sum_{k=1}^p P_{1,\pi}(|\theta_k|)\right], \tag{2.7}$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^{\mathrm{T}}$, and $P_{1,\pi}(\cdot)$ is a penalty function with tuning parameter π. For any penalty function $P_\tau(\cdot)$ with tuning parameter τ, take $\rho(t;\tau) = \tau^{-1}P_\tau(t)$ for any $t \in [0,\infty)$ and $\tau \in (0,\infty)$. Let $P_{1,\pi}(\cdot)$ belong to the class as considered in Lv and Fan (2009):

$$\begin{aligned}\mathcal{P} = \{P_\tau(\cdot) : {}& \rho(t;\tau) \text{ is increasing in } t \in [0,\infty) \text{ and has continuous} \\ & \text{derivative } \rho'(t;\tau) \text{ for any } t \in (0,\infty) \text{ with } \rho'(0^+;\tau) \in \\ & (0,\infty), \text{ where } \rho'(0^+;\tau) \text{ is independent of } \tau\}.\end{aligned} \tag{2.8}$$

The class $\mathcal{P}$ is broad and general, which includes the commonly used L_1 penalty, SCAD penalty (Fan and Li, 2001) and MCP penalty (Zhang, 2010). To establish the consistency of $\widetilde{\boldsymbol{\theta}}_n$, we also assume the following condition.

CONDITION 2. For any $\mathbf{X}$ and $j = 1,\ldots,r$, $g_j(\mathbf{X};\boldsymbol{\theta})$ is continuously differentiable with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and satisfies the conditions

$$\max_{1\leq j\leq r}\max_{k\notin\mathcal{S}}\mathbb{E}\left\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{\partial g_j(\mathbf{X}_i;\boldsymbol{\theta})}{\partial\theta_k}\right|\right\} \leq K_2 \tag{2.9}$$

for some uniform constant $K_2 > 0$, and

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{1\le j\le r} \max_{k\notin\mathcal{S}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \left| \frac{\partial g_j(\mathbf{X}_i;\boldsymbol{\theta})}{\partial\theta_k} \right| \right\} = O_p(\varphi_n) \tag{2.10}$$

for some $\varphi_n > 0$, which may diverge with n.

Condition 2 is on the continuity of the estimating function with respect to $\boldsymbol{\theta}$. Typically, smooth estimating functions can be assumed to have bounded derivatives so that Condition 2 is easily satisfied. At the sample level, considering the high-dimensionality of the problem, we can accommodate diverging φ_n in (2.10) so that our results hold in broad situations. If there are functions $B_{n,jk}(\cdot)$ such that $|\partial g_j(\mathbf{X};\boldsymbol{\theta})/\partial\theta_k| \le B_{n,jk}(\mathbf{X})$ for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$, $j = 1,\ldots,r$ and $k\notin\mathcal{S}$, and $|\mathbb{E}\{B_{n,jk}^m(\mathbf{X}_i)\}| \le Km!H^{m-2}$ for any $m\ge 2$, $j = 1,\ldots,r$ and $k\notin\mathcal{S}$, where K and H are uniform positive constants independent of j and k, Theorem 2.8 of Petrov (1995) implies $\sup_{1\le j\le r}\sup_{k\notin\mathcal{S}} n^{-1}\sum_{i=1}^n B_{n,jk}(\mathbf{X}_i) = O_p(1)$ provided that $\max\{\log r, \log p\} = o(n)$. Thus, (2.10) holds with $\varphi_n = 1$, accommodating exponentially growing r and p. Since the identification condition (2.6) only provides a lower bound for $|\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty$ when $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}$ satisfies $|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty > \varepsilon$ and $\boldsymbol{\theta}_{\mathcal{S}^c} = \mathbf{0}$, we use (2.9) to derive a lower bound for $|\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty$ when $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}$ satisfies $\boldsymbol{\theta}_{\mathcal{S}^c} \ne \mathbf{0}$ but $|\boldsymbol{\theta}_{\mathcal{S}^c}|_1$ is small, and then $\boldsymbol{\theta}_0$ is a local minimizer for $|\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}|_\infty$. For special case with sparse linear regressions, Condition (2.9) becomes one similar to the well known crucial irrepresentable condition (Zhao and Yu, 2007) at the population level. We have the following proposition on the properties of the penalized EL estimator $\widetilde{\boldsymbol{\theta}}_n$ as in (2.7).

PROPOSITION 2. *Let $P_{1,\pi}(\cdot)\in\mathcal{P}$ for $\mathcal{P}$ defined in* (2.8). *Define $a_n = \sum_{k=1}^p P_{1,\pi}(|\theta_k^0|)$ and $b_n = \max\{rn^{-1}, a_n\}$. Assume* (2.4), (2.5), *Conditions* 1 *and* 2 *hold, and*

$$\max_{k\in\mathcal{S}} \sup_{0<t<|\theta_k^0|+c_n} P_{1,\pi}'(t) = O(\chi_n) \tag{2.11}$$

for some $\chi_n \to 0$ and $c_n \to 0$ with $b_n^{1/(2\beta)}c_n^{-1} \to 0$. If $r = o(n^{1/2-1/\gamma})$, $\max\{b_n, rs\chi_n b_n^{1/(2\beta)}\} = o(n^{-2/\gamma})$ and $r^{1/2}\varphi_n\max\{r^{1/2}n^{-1/2}, s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\} = o(\pi)$, then there is a local minimizer $\widetilde{\boldsymbol{\theta}}_n\in\boldsymbol{\Theta}$ for (2.7) *satisfying $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty = O_p\{b_n^{1/(2\beta)}\}$ and $\mathbb{P}(\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c} = \mathbf{0}) \to 1$ as $n\to\infty$.*

In Proposition 2, a_n depends on the true parameter $\boldsymbol{\theta}_0$ and the tuning parameter π in the penalty function. For a typical $P_{1,\pi}(\cdot)\in\mathcal{P}$ and $\boldsymbol{\theta}_0$ with

s nonzero components, it is the case that $a_n = O(s\pi)$. Condition (2.11) is used to control the bias introduced by $P_{1,\pi}(\cdot)$ on $\widetilde{\boldsymbol{\theta}}_n$. See (7.3) in Section 7.2 for details. With the assumption $b_n = o(\min_{k\in\mathcal{S}}|\theta_k^0|^{2\beta})$ that the signal strength of the nonzero components of $\boldsymbol{\theta}_0$ does not diminish to zero too fast, (2.11) can be replaced by

$$\max_{k\in\mathcal{S}} \sup_{c|\theta_k^0|<t<c^{-1}|\theta_k^0|} P'_{1,\pi}(t) = O(\chi_n) \tag{2.12}$$

for some constant $c \in (0,1)$. For those asymptotically unbiased penalties like SCAD and MCP, $\chi_n = 0$ in (2.12) for n sufficiently large if $b_n = o(\min_{k\in\mathcal{S}}|\theta_k^0|^{2\beta})$; see also Fan and Li (2001). Thus, with $\beta = 1$ in Condition 1, $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty = O_p(b_n^{1/2})$. Further, if π is chosen as $O\{(n^{-1}\log p)^{1/2}\}$, a common one in the literature, then $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty = O_p\{s^{1/2}(n^{-1}\log p)^{1/4}\}$, a conservative convergence rate of $\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}}$.

Let $F_n(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} n^{-1}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} + \sum_{k=1}^p P_{1,\pi}(|\theta_k|)$. The rationale of Proposition 2 is that for any $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}$ in a small neighborhood of $\boldsymbol{\theta}_0$ such that $|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty > \varepsilon_n$, where $\varepsilon_n \to 0$ at some slow enough rate, $F_n(\boldsymbol{\theta})$ will take value larger than $\xi_n F_n(\boldsymbol{\theta}_0)$ for some diverging ξ_n with probability tending to 1; see also Chang et al. (2013, 2016) for such a phenomenon of EL. Then with the penalty $P_{1,\pi}(\cdot)$ encouraging sparsity of $\widetilde{\boldsymbol{\theta}}_n$, we are able to establish the consistency of $\widetilde{\boldsymbol{\theta}}_n$ to the sparse $\boldsymbol{\theta}_0$.

Proposition 2 shows that the penalized EL can consistently estimate the high-dimensional sparse model parameter with p growing exponentially with n, though still requires r diverging at some slower rate than $n^{1/2}$. The development of Proposition 2 is fundamentally facilitated by our motivation: to estimate a high-dimensional sparse model parameter. With the new identification condition (2.6), sparse and consistent estimator can be obtained by using penalized EL. The intuition of our results is clear: to identify s nonzero components of a sparse p-dimensional model parameter, one essentially requires r ($r \geq s$) informative estimating functions for those s components. The practical interpretation is also clear: given fewer estimating functions than the model parameters, a reasonable direction is to identify and estimate a sparse model parameter. Such an observation is consistent with the ones found in Gautier and Tsybakov (2014) for high-dimensional instrumental variables regression with endogenity where the number of instrumental variables may be less than the model parameters in the regression problems.

3. A new penalized empirical likelihood.

With the penalized EL estimator $\widetilde{\boldsymbol{\theta}}_n$ in (2.7) capable of handling high-dimensional model parameter with fewer number of estimating functions, our next goal is to accommodate a more general situation: allowing both r and p to grow exponentially with n. For such a purpose, we propose to update the penalized EL estimator with an extra penalty encouraging sparsity in the Lagrange multiplier $\boldsymbol{\lambda}$:

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \Bigg[\sum_{i=1}^{n} \log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} - n\sum_{j=1}^{r} P_{2,\nu}(|\lambda_j|) \\
+ n\sum_{k=1}^{p} P_{1,\pi}(|\theta_k|)\Bigg],
\end{aligned}
\tag{3.1}
$$

where $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^{\mathrm{T}}$, $\boldsymbol{\lambda} = (\lambda_1,\ldots,\lambda_r)^{\mathrm{T}}$, and $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ are two penalty functions with tuning parameters π and ν, respectively. Our motivation is that with appropriately chosen penalty function $P_{2,\nu}(\cdot)$ and tuning parameter ν, the estimator $\widehat{\boldsymbol{\theta}}_n$ is associated with a sparse Lagrange multiplier $\boldsymbol{\lambda}$. Since sparse $\boldsymbol{\lambda}$ effectively uses a subset of the estimating functions $\mathbf{g}(\cdot;\cdot)$, r can be large as long as the number of nonzero components in $\boldsymbol{\lambda}$ is small, essentially satisfying the requirement in Proposition 2. Hence, one expects analogous properties of (3.1) to those in Proposition 2, but now being capable of accommodating high-dimensional p and r simultaneously.

Not surprisingly, involving $P_{2,\nu}(\cdot)$ makes the technical analysis much more challenging, especially when handling exponentially diverging p and r. For $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})$, let $f(\boldsymbol{\lambda};\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} - \sum_{j=1}^{r} P_{2,\nu}(|\lambda_j|)$ and $S_n(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} f(\boldsymbol{\lambda};\boldsymbol{\theta}) + \sum_{k=1}^{p} P_{1,\pi}(|\theta_k|)$. Here $f(\boldsymbol{\lambda};\boldsymbol{\theta})$ is a function of $\boldsymbol{\lambda}$ upon given $\boldsymbol{\theta}$. Let $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}) = \arg\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} f(\boldsymbol{\lambda};\boldsymbol{\theta})$ be the Lagrange multiplier defined at $\boldsymbol{\theta}$. For any subset $\mathcal{A}\subset\{1,\ldots,r\}$, denote by $\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i;\boldsymbol{\theta})$ the subvector of $\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})$ with elements indexed by $\mathcal{A}$. Write $\bar{\mathbf{g}}_{\mathcal{A}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i;\boldsymbol{\theta})$, $\widehat{\mathbf{V}}_{\mathcal{A}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i;\boldsymbol{\theta})\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i;\boldsymbol{\theta})^{\mathrm{T}}$ and $\mathbf{V}_{\mathcal{A}}(\boldsymbol{\theta}) = \mathbb{E}\{\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i;\boldsymbol{\theta})\mathbf{g}_{\mathcal{A}}(\mathbf{X}_i;\boldsymbol{\theta})^{\mathrm{T}}\}$. For any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$ and $j=1,\ldots,r$, define $\bar{g}_j(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n} g_j(\mathbf{X}_i;\boldsymbol{\theta})$. We first characterize the properties of $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ for $\boldsymbol{\theta}$ near the truth $\boldsymbol{\theta}_0$. To do this, we assume the following condition for the existence of higher order moments.

Condition 3. There exist some $K_3>0$ and $\gamma>4$ such that

$$
\max_{1\le j\le r}\mathbb{E}\Big\{\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |g_j(\mathbf{X}_i;\boldsymbol{\theta})|^{\gamma}\Big\} \le K_3.
$$

Let $\rho_2(t;\nu) = \nu^{-1}P_{2,\nu}(t)$. We also take $P_{2,\nu}(\cdot)\in\mathcal{P}$ for $\mathcal{P}$ as in (2.8), so that $\rho_2'(0^+;\nu)$ is independent of ν. Write it as $\rho_2'(0^+)$ for simplicity and define $\mathcal{M}_{\boldsymbol{\theta}} = \{1\le j\le r : |\bar{g}_j(\boldsymbol{\theta})| \ge \nu\rho_2'(0^+)\}$ for any $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Proposition 3 shows that for any $\boldsymbol{\theta}$ near the truth $\boldsymbol{\theta}_0$, the support of the Lagrange multiplier $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is a subset of $\mathcal{M}_{\boldsymbol{\theta}}$ with probability approaching one.

PROPOSITION 3. *Let $\{\boldsymbol{\theta}_n\}$ be a sequence in $\boldsymbol{\Theta}$ and $P_{2,\nu}(\cdot) \in \mathcal{P}$ be convex for $\mathcal{P}$ as in* (2.8)*. For some $C \in (0,1)$, take $\mathcal{M}^*_{\boldsymbol{\theta}_n} = \{1 \le j \le r : |\bar{g}_j(\boldsymbol{\theta}_n)| \ge C\nu\rho_2'(0^+)\}$. Assume Condition* 3 *hold. Further, we assume the eigenvalues of $\widehat{\mathbf{V}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)$ are uniformly bounded away from zero and infinity with probability approaching one, and $|\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}|_2 = O_p(u_n)$ for some $u_n \to 0$. Let $\max_{1\le j\le r} n^{-1}\sum_{i=1}^n |g_j(\mathbf{X}_i;\boldsymbol{\theta}_n)|^2 = O_p(\varsigma_n)$ for some $\varsigma_n > 0$ that may diverge with n. If $m_n^{1/2}u_n\varsigma_n = o(\nu)$ and $m_n^{1/2}u_n n^{1/\gamma} = o(1)$ with $m_n = |\mathcal{M}^*_{\boldsymbol{\theta}_n}|$, then with probability approaching one there is a sparse local maximizer $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_n) = (\widehat{\lambda}_{n,1},\ldots,\widehat{\lambda}_{n,r})^{\mathrm{T}}$ for $f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$ satisfying the three results:* (i) *$|\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_n)|_2 = O_p(u_n)$,* (ii) *$\mathrm{supp}\{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_n)\} \subset \mathcal{M}_{\boldsymbol{\theta}_n}$, and* (iii) *$\mathrm{sgn}(\widehat{\lambda}_{n,j}) = \mathrm{sgn}\{\bar{g}_j(\boldsymbol{\theta}_n)\}$ for any $j \in \mathcal{M}_{\boldsymbol{\theta}_n}$ with $\widehat{\lambda}_{n,j} \ne 0$.*

The sequence $\{\boldsymbol{\theta}_n\}$ can be taken as one that approaches the truth $\boldsymbol{\theta}_0$ as $n \to \infty$. Then $\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)$ will be small when n is large. As shown in Section 7.3, $\nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}$ is the asymptotically leading term of $\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)$. The reason is that the tuning parameter ν typically diminishes to 0 at some slower rate than $n^{-1/2}$, so that $\nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}$ leads to a non-negligible contribution, and our analysis shows that it leads to a correctable bias term in $\widehat{\boldsymbol{\theta}}_n$. Upon removing the leading order term, we assume $|\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}|_2 = O_p(u_n)$ with $u_n \to 0$, which can be easily satisfied. Requirement on the eigenvalues of $\widehat{\mathbf{V}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)$ is natural so that we can characterize the limiting behavior of the estimator $\widehat{\boldsymbol{\theta}}_n$. Furthermore, m_n is taken to be an upper bound of the size of $\mathcal{M}_{\boldsymbol{\theta}_n}$, the generic description such as $m_n^{1/2}u_n\varsigma_n = o(\nu)$ and $m_n^{1/2}u_n n^{1/\gamma} = o(1)$ can be viewed as characterizing the capacity of the penalized EL under which it is reliable for consistent estimators, depending on the behavior of the estimating function $\mathbf{g}(\cdot;\cdot)$ on its continuity and tail probabilistic properties.

Proposition 3 implies that when $\boldsymbol{\theta}$ is approaching $\boldsymbol{\theta}_0$, the sparse $\boldsymbol{\lambda}$ in (3.1) effectively conducts a moment selection by choosing the estimating functions such that $\bar{g}_j(\boldsymbol{\theta})$ has large absolute deviation from 0. Let $\mu_j(\boldsymbol{\theta}) = \mathbb{E}\{g_j(\mathbf{X}_i;\boldsymbol{\theta})\}$, then we know that $\mu_j(\boldsymbol{\theta}_0) = 0$ and $\bar{g}_j(\boldsymbol{\theta}) \to_p \mu_j(\boldsymbol{\theta})$ as $n \to \infty$. If $\boldsymbol{\theta}$ is in the neighborhood of $\boldsymbol{\theta}_0$, then Taylor expansion gives that $\mu_j(\boldsymbol{\theta}) = \mu_j(\boldsymbol{\theta}) - \mu_j(\boldsymbol{\theta}_0) = \{\nabla_{\boldsymbol{\theta}}\mu_j(\boldsymbol{\theta}^*)\}^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$ for some $\boldsymbol{\theta}^*$ between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Hence, those components of the estimating functions with large magnitude in the derivative of their expected value with respect to $\boldsymbol{\theta}$ will be selected. Since larger derivative indicates a steeper direction towards the truth $\boldsymbol{\theta}_0$, making it easier and more informative to find the optimum. Therefore, selecting components in $\mathcal{M}_{\boldsymbol{\theta}}$ is seen sensible. However, we note that without further strong and likely to be unrealistic conditions on the shape of the estimating functions, $\mathcal{M}_{\boldsymbol{\theta}}$ cannot be controlled as a fixed set even at the limiting case when $n \to \infty$, so that it will depend on the value of the parameter $\boldsymbol{\theta}$. Instead of requiring $\mathcal{M}_{\boldsymbol{\theta}}$ to be fixed, we show in the following that for any choice of its subset satisfying some reasonable conditions, the resulting penalized EL estimator is consistent and asymptotically normally distributed.

Let

$$\ell_n = \max_{\boldsymbol{\theta} \in \{\boldsymbol{\theta}=(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Theta} : |\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty \le c_n, \boldsymbol{\theta}_{\mathcal{S}^c} = \mathbf{0}\}} |\mathcal{M}_{\boldsymbol{\theta}}| \tag{3.2}$$

for some $c_n \to 0$ satisfying $b_n^{1/(2\beta)} c_n^{-1} \to 0$ where b_n is more clearly specified in Condition 6 below. Based on Proposition 3, we know the support of Lagrange multiplier $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is a subset of $\mathcal{M}_{\boldsymbol{\theta}}$ with probability approaching one when $\boldsymbol{\theta}$ is in a small neighborhood of $\boldsymbol{\theta}_0$. Here ℓ_n is a technical device controlling the maximum number of effective estimating functions when applying the new penalized EL, and it can be viewed as a cap of the r in Proposition 2. Though ℓ_n is a technical device, we remark that, practically, one can always achieve the control of the nonzero components of $\boldsymbol{\lambda}$ by appropriately choosing the tuning parameter ν.

To establish the consistency of the penalized EL estimator $\widehat{\boldsymbol{\theta}}_n$ as in (3.1), we need the following extra regularity conditions on the continuity and probabilistic behavior of the estimating functions.

Condition 4. There exist uniform constants $0 < K_4 < K_5$ such that $K_4 < \lambda_{\min}\{\mathbf{V}_{\mathcal{F}}(\boldsymbol{\theta}_0)\} \le \lambda_{\max}\{\mathbf{V}_{\mathcal{F}}(\boldsymbol{\theta}_0)\} < K_5$ for any $\mathcal{F} \subset \{1, \ldots, r\}$ with $|\mathcal{F}| \le \ell_n$, where ℓ_n is as in (3.2).

Condition 5. Assume that

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{1 \le j \le r} \max_{k \notin \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^n \left| \frac{\partial g_j(\mathbf{X}_i; \boldsymbol{\theta})}{\partial \theta_k} \right|^2 \right\} = O_p(\xi_n),$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{1 \le j \le r} \max_{k \in \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^n \left| \frac{\partial g_j(\mathbf{X}_i; \boldsymbol{\theta})}{\partial \theta_k} \right|^2 \right\} = O_p(\omega_n),$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{1 \le j \le r} \left\{ \frac{1}{n} \sum_{i=1}^n |g_j(\mathbf{X}_i; \boldsymbol{\theta})|^4 \right\} = O_p(\varrho_n)$$

for some $\xi_n > 0$, $\omega_n > 0$ and $\varrho_n > 0$ that may diverge with n.

Condition 6. Let $b_n = \max\{a_n, \nu^2\}$ with $a_n = \sum_{k=1}^p P_{1,\pi}(|\theta_k^0|)$. There exist $\chi_n \to 0$ and $c_n \to 0$ with $b_n^{1/(2\beta)} c_n^{-1} \to 0$ for β defined in Condition 1 such that $\max_{k \in \mathcal{S}} \sup_{0<t<|\theta_k^0|+c_n} P'_{1,\pi}(t) = O(\chi_n)$.

Here Condition 4 is actually a weaker one than that in (2.5) in the sense that it only requires the population covariance matrices of subsets of estimating functions to well behave at the truth $\boldsymbol{\theta}_0$. The first two bounds in Condition 5 are used to characterize the behavior of the eigenvalues of $\widehat{\mathbf{V}}_{\mathcal{F}}(\boldsymbol{\theta})$ when $\boldsymbol{\theta}$ in a small neighborhood of $\boldsymbol{\theta}_0$; see Lemma 1 in Section 7.4. We do not impose explicit rates on ξ_n, ω_n, and ϱ_n, so that the conditions are generally not restrictive. Similar to our earlier discussion for φ_n in (2.10) in Condition 2, we can actually choose $\xi_n = \omega_n = \varrho_n = 1$ under some additional mild conditions provided that $\max\{\log r, \log p\} = o(n)$. Condition 6 is similar to (2.11) in Proposition 2 with a differently defined b_n. Similar to that in Proposition 2, Condition 6 can be replaced by (2.12) if the minimal signal strength condition is satisfied for appropriately chosen tuning parameter π. Then $\chi_n = 0$ when n is large for those asymptotically unbiased penalties like SCAD and MCP. We now present the following theorem for the consistency of $\widehat{\boldsymbol{\theta}}_n$.

THEOREM 1. *Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathcal{P}$ for $\mathcal{P}$ defined in* (2.8)*, and $P_{2,\nu}(\cdot)$ be convex with bounded second derivative around* 0*. Assume Conditions* 1–6 *hold. Let $b_n = \max\{a_n, \nu^2\}$ with $a_n = \sum_{k=1}^{p} P_{1,\pi}(|\theta_k^0|)$, and $\kappa_n = \max\{\ell_n^{1/2} n^{-1/2}, s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\}$. If $\log r = o(n^{1/3})$, $\varrho_n = o(n^2)$, $s^2 \ell_n \omega_n b_n^{1/\beta} = o(1)$, $\ell_n^2 n^{-1} \varrho_n \log r = o(1)$, $\max\{b_n, \ell_n \kappa_n^2\} = o(n^{-2/\gamma})$, $\ell_n^{1/2} \varrho_n^{1/2} \kappa_n = o(\nu)$ and $\ell_n^{1/2} \xi_n^{1/2} \max\{\ell_n \nu, s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\} = o(\pi)$, then there is a local minimizer $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$ for* (3.1) *such that $|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty = O_p\{b_n^{1/(2\beta)}\}$ and $\mathbb{P}(\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c} = \mathbf{0}) \to 1$ as $n \to \infty$.*

Theorem 1 establishes the consistency of $\widehat{\boldsymbol{\theta}}_n$ under L_∞-norm with a conservative convergence rate $O_p\{b_n^{1/(2\beta)}\}$. Under some additional regularity conditions, the rate can be improved as $O_p(\nu)$. Theorem 1 holds for broad situations accommodating various cases of the estimating functions. In reasonable cases that we discussed earlier, $\chi_n = 0$ and $\xi_n = \omega_n = \varrho_n = 1$, Theorem 1 holds provided that $\log r = o(n^{1/3})$, $\ell_n = o(\min\{n^{1/2}(\log r)^{-1/2}, n^{1/2-1/\gamma}\})$, $a_n = o(\min\{s^{-2\beta}\ell_n^{-\beta}, n^{-2/\gamma}\})$, and the tuning parameters ν and π satisfy $\ell_n n^{-1/2} = o(\nu)$, $\nu = o(\min\{s^{-\beta}\ell_n^{-\beta/2}, n^{-1/\gamma}\})$ and $\ell_n^{3/2}\nu = o(\pi)$. Noticing that $a_n \lesssim s\pi$, choosing $\pi = o(\min\{s^{-2\beta-1}\ell_n^{-\beta}, s^{-1}n^{-2/\gamma}\})$ can ensure the result. Additionally, note that $s \le \ell_n$. Thus by letting $\log r \asymp n^\tau$ and $\ell_n \asymp n^\delta$ for some $\tau \in [0, 1/3)$ and $\delta \in [0, \min\{(\gamma-4)/(7\gamma), 1/(6\beta+7)\})$, $\widehat{\boldsymbol{\theta}}_n$ satisfies Theorem 1 if $\nu \asymp n^{-\phi_1}$ and $\pi \asymp n^{-\phi_2}$ with $\phi_1 \in (\max\{3\beta\delta/2, 1/\gamma\}, 1/2 - \delta)$ and $\phi_2 \in (\max\{(3\beta+1)\delta, 2/\gamma + \delta\}, \phi_1 - 3\delta/2)$, which are reasonable choices for the tuning parameters. To further establishing the limiting distribution of $\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}$, we need the following two additional conditions.

CONDITION 7. For any $\mathbf{X}$ and $j = 1, \dots, p$, $g_j(\mathbf{X}; \boldsymbol{\theta})$ is twice continuously differentiable with respect to $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \max_{1 \le j \le r} \max_{k_1, k_2 \in \mathcal{S}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left| \frac{\partial^2 g_j(\mathbf{X}_i; \boldsymbol{\theta})}{\partial \theta_{k_1} \partial \theta_{k_2}} \right|^2 \right\} = O_p(\varpi_n)$$

for some $\varpi_n \ge 0$ that may diverge with n.

CONDITION 8. Let $\mathbf{Q}_{\mathcal{F}} = [\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}} \mathbf{g}_{\mathcal{F}}(\mathbf{X}_i; \boldsymbol{\theta}_0)\}]^{\mathrm{T}} [\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}} \mathbf{g}_{\mathcal{F}}(\mathbf{X}_i; \boldsymbol{\theta}_0)\}]$ for any $\mathcal{F} \subset \{1, \dots, r\}$. There exist uniform constants $0 < K_6 < K_7$ such that $K_6 < \lambda_{\min}(\mathbf{Q}_{\mathcal{F}}) \le \lambda_{\max}(\mathbf{Q}_{\mathcal{F}}) < K_7$ for any $\mathcal{F}$ with $s \le |\mathcal{F}| \le \ell_n$.

Following similar discussion for Condition 5, $\varpi_n = 1$ in Condition 7 for reasonable models in practice. Write $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n) = (\widehat{\lambda}_1, \dots, \widehat{\lambda}_r)^{\mathrm{T}}$. Let $\mathcal{R}_n = \mathrm{supp}\{\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)\}$ and define

$$\begin{aligned} \widehat{\mathbf{J}}_{\mathcal{R}_n} &= \{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}} \widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n) \{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}, \\ \widehat{\boldsymbol{\psi}}_{\mathcal{R}_n} &= \widehat{\mathbf{J}}_{\mathcal{R}_n}^{-1} \{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}} \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}} \widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n) \widehat{\boldsymbol{\eta}}_{\mathcal{R}_n}, \end{aligned} \tag{3.3}$$

where $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \dots, \widehat{\eta}_r)^{\mathrm{T}}$ with $\widehat{\eta}_j = \nu \rho_2'(|\widehat{\lambda}_j|; \nu) \mathrm{sgn}(\widehat{\lambda}_j)$ for $\widehat{\lambda}_j \ne 0$ and $\widehat{\eta}_j \in [-\nu \rho_2'(0^+), \nu \rho_2'(0^+)]$ for $\widehat{\lambda}_j = 0$.

THEOREM 2. *Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathcal{P}$ for $\mathcal{P}$ defined in* (2.8), *and $P_{2,\nu}(\cdot)$ be convex with bounded second derivative around* 0. *Assume Conditions* 1–8 *hold. Let $b_n = \max\{a_n, \nu^2\}$ with $a_n = \sum_{k=1}^{p} P_{1,\pi}(|\theta_k^0|)$, and $\kappa_n = \max\{\ell_n^{1/2} n^{-1/2}, s^{1/2} \chi_n^{1/2} b_n^{1/(4\beta)}\}$. If $\log r = o(n^{1/3})$, $\varrho_n = o(n^2)$, $b_n = o(n^{-2/\gamma})$, $n s \chi_n^2 = o(1)$, $\ell_n^2 \varrho_n^{1/2} (\log r) \max\{s^2(\omega_n + s\varpi_n) b_n^{1/\beta}, n^{-1}(s\omega_n + \ell_n \varrho_n) \log r\} = o(1)$, $n \ell_n \kappa_n^4 \max\{s\omega_n, n^{2/\gamma}\} = o(1)$, $n \ell_n s^2 \varpi_n \max\{\ell_n^2 \nu^4, s^2 \chi_n^2 b_n^{1/\beta}\} = o(1)$, $\ell_n^{1/2} \varrho_n^{1/2} \kappa_n = o(\nu)$ and $\ell_n^{1/2} \xi_n^{1/2} \max\{\ell_n \nu, s^{1/2} \chi_n^{1/2} b_n^{1/(4\beta)}\} = o(\pi)$, then the local minimizer $\widehat{\boldsymbol{\theta}}_n \in \boldsymbol{\Theta}$ for* (3.1) *specified in Theorem* 1 *satisfies*

$$n^{1/2} \boldsymbol{\alpha}^{\mathrm{T}} \widehat{\mathbf{J}}_{\mathcal{R}_n}^{1/2} (\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}} - \widehat{\boldsymbol{\psi}}_{\mathcal{R}_n}) \xrightarrow{d} N(0,1) \tag{3.4}$$

for any $\boldsymbol{\alpha} \in \mathbb{R}^s$ with $|\boldsymbol{\alpha}|_2 = 1$, as $n \to \infty$, where $\widehat{\mathbf{J}}_{\mathcal{R}_n}$ and $\widehat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ are defined in (3.3).

Theorem 2 shows that subject to a bias correction, the new penalized EL estimator for nonzero components is asymptotically normal in the sense of (3.4). The bias term $\widehat{\boldsymbol{\psi}}_{\mathcal{R}_n}$ in (3.4) is due to the penalty function $P_{2,\nu}(\cdot)$ used in (3.1); see also our discussion after the Proposition 3. Similar to that in Theorem 1, with reasonable cases $\chi_n = 0$ and $\xi_n = \omega_n = \varrho_n = \varpi_n = 1$, descriptions on the dimensionality in Theorem 2 can be simplified. If $\ell_n \asymp s$, Theorem 2 holds provided that $\log r = o(n^{1/3})$, $s = o(\min\{n^{1/3}(\log r)^{-2/3}, n^{1/(10\beta+7)}(\log r)^{-2\beta/(10\beta+7)}, n^{(\gamma-4)/(7\gamma)}\})$, and ν and π satisfying $sn^{-1/2} = o(\nu)$, $\nu = o(\min\{n^{-1/\gamma}, s^{-5\beta/2}(\log r)^{-\beta/2}, n^{-1/4}s^{-5/4}\})$, $s^{3/2}\nu = o(\pi)$ and $\pi = o(\min\{n^{-2/\gamma}s^{-1}, s^{-5\beta-1}(\log r)^{-\beta}\})$. Generally speaking, conditions in Theorem 2 is stronger than those in Theorem 1, which can be viewed as the expense for the stronger asymptotic normality results. In summary, we have shown that the new penalized EL estimator $\widehat{\boldsymbol{\theta}}_n$ as in (3.1) has desirable properties including consistency in estimating nonzero components and identifying zero components of $\boldsymbol{\theta}_0$, and asymptotic normality for the estimator of the nonzero components of $\boldsymbol{\theta}_0$.

4. Algorithms for implementations.

For ease and stability in implementations, we calculate the new penalized EL estimator $\widehat{\boldsymbol{\theta}}_n$ by minimizing the following slightly modified objective function:

$$\begin{aligned}\widehat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} \Bigg[\sum_{i=1}^n \log_\star\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\} - n\sum_{j=1}^r P_{2,\nu}(|\lambda_j|) \\ + n\sum_{k=1}^p P_{1,\pi}(|\theta_k|)\Bigg],\end{aligned} \tag{4.1}$$

where $\log_\star(z)$ is a twice differentiable pseudo-logarithm function with bounded support adopted from Owen (2001):

$$\log_\star(z) = \begin{cases}\log(z), & \text{if } z \ge \epsilon;\\ \log(\epsilon) - 1.5 + 2z/\epsilon - z^2/(2\epsilon^2), & \text{if } z \le \epsilon;\end{cases}$$

where ϵ is chosen as n^{-1} in our implementations. In the optimization, we apply the quadratic approximation (Fan and Li, 2001) to the penalty functions $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$. More specifically, for a penalty function $P_\tau(\cdot)$, the quadratic approximation states $P_\tau(|t|) \approx P_\tau(|t_0|) + 2^{-1}P'_\tau(|t_0|)|t_0|^{-1}(t^2 - t_0^2)$ for t being in a small neighborhood of t_0. The first and second derivatives are approximated by $P'_\tau(|t|) \approx t|t_0|^{-1}P'_\tau(|t_0|)$ and $P''_\tau(|t|) \approx |t_0|^{-1}P'_\tau(|t_0|)$.

The computation of EL is challenging, especially with high-dimensional p and r. To compute $\widehat{\boldsymbol{\theta}}_n$, we propose to apply a modified two-layer coordinate decent algorithm extending the one in Tang and Wu (2014). The inner layer of the algorithm solves for $\boldsymbol{\lambda}$ with given $\boldsymbol{\theta}$ by maximizing $f(\boldsymbol{\lambda};\boldsymbol{\theta})$ as given in Section 3. This layer only involves maximizing a concave function, and hence is stable. The outer layer of the algorithm searches for the optimizer

$\widehat{\boldsymbol{\theta}}_n$. Both layers can be solved using coordinate descent by cycling through and updating each of the coordinates; see Tang and Wu (2014).

In the inner layer, $\boldsymbol{\lambda}$ is solved at a given $\boldsymbol{\theta}$, which can be done by optimizing (4.1) with respect to $\boldsymbol{\lambda}$ using coordinate descent. Let $\boldsymbol{\lambda}$ start at an initial value $\widehat{\boldsymbol{\lambda}}^{(0)}$. With the other coordinates fixed, the $(m+1)$th Newton's update for λ_j $(j = 1, \ldots, r)$, the jth component of $\boldsymbol{\lambda}$, is given by

$$\widehat{\lambda}_j^{(m+1)} = \widehat{\lambda}_j^{(m)} - \frac{\sum_{i=1}^n \log_\star'(t_{i,m}) g_j(\mathbf{X}_i; \boldsymbol{\theta}) - nP_{2,\nu}'\{|\widehat{\lambda}_j^{(m)}|\}}{\sum_{i=1}^n \log_\star''(t_{i,m})\{g_j(\mathbf{X}_i; \boldsymbol{\theta})\}^2 - nP_{2,\nu}''\{|\widehat{\lambda}_j^{(m)}|\}}, \tag{4.2}$$

where $t_{i,m} = 1 + \mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})^{\mathrm{T}} \widehat{\boldsymbol{\lambda}}^{(m)}$ with $\widehat{\boldsymbol{\lambda}}^{(m)} = \{\widehat{\lambda}_1^{(m)}, \ldots, \widehat{\lambda}_r^{(m)}\}^{\mathrm{T}}$. The procedure cycles through all the r components of $\boldsymbol{\lambda}$ and is repeated until convergence. During this process, the objective function needs to be checked to ensure it gets optimized in each step. If not, the step size continues to be halved until the objective function gets driven in the right direction. The iterative updating procedure (4.2) can be viewed as sequential univariate optimizations. The convergence rate and stability are studies in the optimization literature; see Friedman et al. (2007) and Wu and Lange (2008).

The outer layer of the algorithm is to optimize (4.1) with respect to $\boldsymbol{\theta}$, the main interest of the new penalized EL, using the coordinate descent algorithm. At a given $\boldsymbol{\lambda}$, the algorithm updates θ_k $(k = 1, \ldots, p)$, by minimizing $S_n(\boldsymbol{\theta})$ defined in Section 3 with respect to θ_k with other θ_l $(l \neq k)$ fixed. Let $\boldsymbol{\theta}$ start at an initial value $\widehat{\boldsymbol{\theta}}^{(0)}$. The $(m+1)$th update for θ_k is given by

$$\begin{aligned} &\widehat{\theta}_k^{(m+1)} \\ &\quad = \widehat{\theta}_k^{(m)} - \frac{\sum_{i=1}^n \log_\star'(s_{i,m}) w_{ik,m} + nP_{1,\pi}'\{|\widehat{\theta}_k^{(m)}|\}}{\sum_{i=1}^n \{\log_\star''(s_{i,m}) w_{ik,m}^2 + \log_\star'(s_{i,m}) z_{ik,m}\} + nP_{1,\pi}''\{|\widehat{\theta}_k^{(m)}|\}}, \end{aligned} \tag{4.3}$$

where $s_{i,m} = 1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{g}\{\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(m)}\}$, $w_{ik,m} = \boldsymbol{\lambda}^{\mathrm{T}} \partial \mathbf{g}\{\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(m)}\}/\partial \theta_k$ and $z_{ik,m} = \boldsymbol{\lambda}^{\mathrm{T}} \partial^2 \mathbf{g}\{\mathbf{X}_i; \widehat{\boldsymbol{\theta}}^{(m)}\}/\partial \theta_k^2$ with $\widehat{\boldsymbol{\theta}}^{(m)} = \{\widehat{\theta}_1^{(m)}, \ldots, \widehat{\theta}_p^{(m)}\}^{\mathrm{T}}$. Since quadratic approximations are applied in the algorithms, we follow Fan and Li (2001) and set a component $\widehat{\lambda}_j^{(m)}$ or $\widehat{\theta}_k^{(m)}$ as zero when it is less than a threshold level say 10^{-3} in an iteration.

We summarize the computation procedure for $\boldsymbol{\theta}$ and $\boldsymbol{\lambda}$ in the following pseudo-code. Suppose ξ is a pre-defined small number, say, $\xi = 10^{-4}$.

```
1. Set the iteration counter $m = 0$, and initialize $\widehat{\boldsymbol{\theta}}^{(0)}$ and $\widehat{\boldsymbol{\lambda}}^{(0)}$;
2. Define the $\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta})$ function;
3. (Outer layer) For $k = 1, \dots, p$,
    (a) Calculate $\widehat{\theta}_k^{(m+1)}$ as in (4.3);
    (b) (Inner layer) For $j = 1, \dots, r$, update $\widehat{\lambda}_j^{(m)}$ as $\widehat{\lambda}_j^{(m+1)}$ defined in (4.2);
4. If $\max_{1 \le k \le p} |\widehat{\theta}_k^{(m+1)} - \widehat{\theta}_k^{(m)}| < \xi$, then stop;
5. Otherwise repeat steps 3 through 4.
```

5. Numerical examples.

The SCAD penalty (Fan and Li, 2001) is used for both $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ in (4.1) for all the numerical experiments in this paper. Since local quadratic approximation is applied in the algorithms, the convexity requirements of the results in Sections 2 and 3 are met. Three information criteria for choosing the tuning parameters π and ν in the penalty functions – BIC (Schwarz, 1978), BICC (Wang et al., 2009), and EBIC (Chen and Chen, 2008) – are used.

5.1. *Estimating high-dimensional mean parameter.*

The first simulation study is to calculate the mean of a multivariate normal distribution in $\mathbb{R}^p$. Let $\mathbf{X} = (X_1, \dots, X_p)^{\mathrm{T}} \sim N(\boldsymbol{\theta}_0, \boldsymbol{\Sigma})$. Suppose only three elements, X_1, X_2, and X_5, have nonzero means and the rest $p - 3$ elements have zero means, i.e., $\boldsymbol{\theta}_0 = (5, 4, 0, 0, 1, 0, \dots, 0)^{\mathrm{T}}$. The covariance matrix $\boldsymbol{\Sigma} = (\sigma_{kl})_{p \times p}$ is set as $\sigma_{kk} = 1$ for each $k = 1, \dots, p$ and $\sigma_{kl} = 0.9$ for any $k \neq l$. The estimating function is simply $\mathbf{g}(\mathbf{X}; \boldsymbol{\theta}) = \mathbf{X} - \boldsymbol{\theta}$. In this case, the number of parameters p is equal to the number of estimating equations r. We consider the underdetermined case where $p = r > n$.

Table 1 summarizes the results for $(n, p) = (50, 100)$, $(100, 200)$, and $(100, 500)$. The proposed penalized EL with two penalties (namely, PEL2) is compared to the single penalty approach (PEL) discussed in Tang and Leng (2010). In general, all the three BIC-type criteria work similarly, with BICC and EBIC yield slightly fewer nonzero parameters. The results from MLE for all p variables and the three true variables (i.e., MLE-Oracle) are also considered. We also report the model error (ME) defined by ME $= |\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}|_2^2$ for a given estimator $\widehat{\boldsymbol{\theta}}$. A smaller ME means a better estimation and prediction. Obviously, in the single penalty approach, all equating equations are used since no moment selection is performed. In each cell, standard error appears in the parentheses.

It is clear from the table that the double-penalty approach outperforms

(n, p, r)	Method	$\boldsymbol{\theta}_{\text{nonzeros}}$	$\boldsymbol{\theta}_{\text{true}}$	ME	No. EE's
$(50, 100, 200)$	MLE-Oracle	3 (0)	NA	0.062 (0.009)	NA
	MLE	100 (0)	3 (0)	2.096 (0.287)	NA
	PEL-BIC	24.06 (4.13)	0.72 (0.12)	33.276 (1.507)	100 (0)
	PEL-BICC	23.15 (4.08)	0.69 (0.12)	33.635 (1.483)	100 (0)
	PEL-EBIC	23.15 (4.08)	0.69 (0.12)	33.635 (1.483)	100 (0)
	PEL2-BIC	3.41 (0.17)	2.81 (0.04)	0.332 (0.041)	5.11 (0.34)
	PEL2-BICC	3.29 (0.15)	2.80 (0.04)	0.302 (0.041)	6.13 (0.33)
	PEL2-EBIC	3.15 (0.13)	2.76 (0.05)	0.341 (0.052)	8.20 (0.21)
$(100, 200, 400)$	MLE-Oracle	3 (0)	NA	0.024 (0.003)	NA
	MLE	200 (0)	3 (0)	1.743 (0.179)	NA
	PEL-BIC	22.02 (6.02)	0.33 (0.09)	38.078 (1.073)	199.98 (0.02)
	PEL-BICC	22.02 (6.02)	0.33 (0.09)	38.078 (1.073)	199.98 (0.02)
	PEL-EBIC	22.02 (6.02)	0.33 (0.09)	38.078 (1.073)	199.98 (0.02)
	PEL2-BIC	6.41 (1.84)	2.84 (0.04)	0.333 (0.091)	6.67 (0.23)
	PEL2-BICC	6.18 (1.84)	2.82 (0.04)	0.352 (0.092)	6.64 (0.23)
	PEL2-EBIC	5.82 (1.86)	2.80 (0.04)	0.372 (0.094)	6.69 (0.24)
$(100, 500, 1000)$	MLE-Oracle	3 (0)	NA	0.031 (0.005)	NA
	MLE	NA	NA	NA	NA
	PEL-BIC	85.71 (22.69)	0.51 (0.14)	37.585 (1.193)	500 (0)
	PEL-BICC	0 (0)	0 (0)	42 (0)	500 (0)
	PEL-EBIC	0 (0)	0 (0)	42 (0)	500 (0)
	PEL2-BIC	2.88 (0.11)	2.70 (0.06)	0.356 (0.057)	6.40 (0.36)
	PEL2-BICC	2.82 (0.09)	2.70 (0.06)	0.376 (0.058)	6.53 (0.35)
	PEL2-EBIC	2.83 (0.09)	2.71 (0.06)	0.369 (0.058)	6.97 (0.32)

TABLE 1

Simulation results for mean of a normal distribution based on 100 replicates. Here $\boldsymbol{\theta}_{\text{nonzeros}}$ *is the average number of selected nonzero components,* $\boldsymbol{\theta}_{\text{true}}$ *is the average number of true nonzero components that are selected, ME reports the model error, and No. EE's reports the number of estimating equations selected.*

the single-penalty approach, as expected. A much smaller subset of variables get selected with almost all the three true predictors identified by the double-penalty method. That says, the double-penalty approach yields lower false positives and higher true positives. While in the single-penalty approach, fewer true predictors are chosen in the larger set of selected variables or nothing can be picked out if $p > n$. What is the most interesting is that a small number (on average 5-8) of estimating equations are selected in the double-penalty approach. As a result, the double-penalty method yields a much smaller ME than the single-penalty method.

5.2. *Linear regression.* In this simulation study, we consider a linear regression model $Y_i = \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta}_0 + \varepsilon_i$, where $\boldsymbol{\theta}_0 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^{\mathrm{T}}$, $\mathbf{Z}_i \in \mathbb{R}^p$ are generated from $N(\mathbf{0}, \boldsymbol{\Sigma})$ with $\sigma_{kk} = 1$ for any $k = 1, \ldots, p$ and $\sigma_{kl} = 0.5$ for any $k \neq l$ in $\boldsymbol{\Sigma} = (\sigma_{kl})_{p\times p}$, and ε_i is a standard normal distributed random variable. Write $\mathbf{X}_i = (Y_i, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}$. The estimating function is $\mathbf{g}(\mathbf{X}; \boldsymbol{\theta}) = \mathbf{Z}(Y - \mathbf{Z}^{\mathrm{T}}\boldsymbol{\theta})$ with $p = r$.

The model error (ME) in the regression setting is defined by ME $= |\boldsymbol{\Sigma}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})|_2^2$ for a given estimator $\widehat{\boldsymbol{\theta}}$. Table 2 reports the results for $(n, p) = (50, 100), (100, 200)$, and $(100, 500)$. Similar to the previous example, the single-penalty approach (PEL) of Tang and Leng (2010) is compared with the double-penalty approach (PEL2). We also compare our method with LASSO. Since the number of parameters p doubles the number of sub-

(n, p, r)	Method	$\boldsymbol{\theta}_{\text{nonzeros}}$	$\boldsymbol{\theta}_{\text{true}}$	ME	No. EE's
$(50, 100, 100)$	MLE-Oracle	3 (0)	NA	0.069 (0.005)	NA
	LASSO	15.21 (0.88)	3 (0)	0.439 (0.034)	NA
	PEL-BIC	0 (0)	0 (0)	28.75 (0)	100 (0)
	PEL-BICC	0 (0)	0 (0)	28.75 (0)	100 (0)
	PEL-EBIC	0 (0)	0 (0)	28.75 (0)	100 (0)
	PEL2-BIC	6.39 (0.52)	2.98 (0.02)	0.497 (0.069)	10.46 (0.46)
	PEL2-BICC	6.33 (0.52)	2.98 (0.02)	0.498 (0.069)	10.49 (0.46)
	PEL2-EBIC	6.06 (0.52)	2.97 (0.02)	0.531 (0.070)	10.43 (0.47)
$(100, 200, 200)$	MLE-Oracle	3 (0)	NA	0.047 (0.005)	NA
	LASSO	17.79 (0.87)	3 (0)	0.374 (0.019)	NA
	PEL-BIC	0 (0)	0 (0)	28.75 (0)	200 (0)
	PEL-BICC	0 (0)	0 (0)	28.75 (0)	200 (0)
	PEL-EBIC	0 (0)	0 (0)	28.75 (0)	200 (0)
	PEL2-BIC	9.22 (1.27)	3 (0)	0.647 (0.118)	5.38 (0.17)
	PEL2-BICC	9.28 (1.28)	3 (0)	0.651 (0.119)	5.39 (0.17)
	PEL2-EBIC	8.38 (1.03)	3 (0)	0.632 (0.119)	5.34 (0.17)
$(100, 500, 500)$	MLE-Oracle	3 (0)	NA	0.039 (0.003)	NA
	LASSO	23.79 (1.23)	3 (0)	0.507 (0.028)	NA
	PEL-BIC	0 (0)	0 (0)	28.75 (0)	500 (0)
	PEL-BICC	0 (0)	0 (0)	28.75 (0)	500 (0)
	PEL-EBIC	0 (0)	0 (0)	28.75 (0)	500 (0)
	PEL2-BIC	6.28 (1.31)	3 (0)	0.601 (0.083)	5.48 (0.16)
	PEL2-BICC	5.96 (1.31)	3 (0)	0.593 (0.085)	5.38 (0.17)
	PEL2-EBIC	6.04 (1.32)	3 (0)	0.602 (0.086)	5.41 (0.16)

TABLE 2

Simulation results for linear regression based on 100 replicates. Here $\boldsymbol{\theta}_{\text{nonzeros}}$ is the average number of selected nonzero components, $\boldsymbol{\theta}_{\text{true}}$ is the average number of true nonzero components that are selected, ME reports the model error, and No. EE's reports the number of estimating equations selected.

jects n, the MLE method does not work in this example. We only report the results from MLE-Oracle (i.e., the MLE method using the true predictors), which gives the smallest model error. In all the three settings, the single-penalty method fails to select any predictor when using all r estimating equations. The double-penalty method identifies all true predictors from a handful of selected ones in most cases by using only a few estimating equations. With the default tuning parameter selection method in LASSO, we clearly see that the number of false inclusion of the predictors is high. Hence, compared with LASSO, we observe that our method has better performance in recovering a sparse model.

5.3. *Regression model with repeated measures.* This is an example with over-identification ($r > p$). Consider a repeated measures model such that $y_{ij} = \mathbf{z}_{ij}^{\mathrm{T}}\boldsymbol{\theta}_0 + \epsilon_{ij}$ ($i = 1, \ldots, n; j = 1, 2$), where $\boldsymbol{\theta}_0 = (3, 1.5, 0, 0, 2, 0, \ldots, 0)^{\mathrm{T}} \in \mathbb{R}^p$, $\mathbf{z}_{ij}$ are generated from $N(0, \boldsymbol{\Sigma})$ with $\sigma_{kl} = 0.5^{|k-l|}$ in $\boldsymbol{\Sigma} = (\sigma_{kl})_{p\times p}$. The random errors $(\epsilon_{i1}, \epsilon_{i2})^{\mathrm{T}}$ are generated from a two-dimensional normal distribution with mean zero and unit marginal compound symmetry covariance matrix with $\rho = 0.7$.

Let $\mathbf{Y}_i = (y_{i1}, y_{i2})^{\mathrm{T}}$ and $\mathbf{Z}_i = (\mathbf{z}_{i1}^{\mathrm{T}}, \mathbf{z}_{i2}^{\mathrm{T}})^{\mathrm{T}}$ respectively collect the response and predictor variables, and write $\mathbf{X}_i = (\mathbf{Y}_i^{\mathrm{T}}, \mathbf{Z}_i^{\mathrm{T}})^{\mathrm{T}}$. To incorporate the dependence among the repeated measures from the same subject

(n, p, r)	Method	$\boldsymbol{\theta}_{\text{nonzeros}}$	$\boldsymbol{\theta}_{\text{true}}$	ME	No. EE's
$(50, 100, 200)$	MLE-Oracle	3 (0)	NA	0.023 (0.002)	NA
	MLE	100 (0)	3 (0)	3.446 (0.106)	NA
	PEL-BIC	0 (0)	0 (0)	15.25 (0)	200 (0)
	PEL-BICC	0 (0)	0 (0)	15.25 (0)	200 (0)
	PEL-EBIC	0 (0)	0 (0)	15.25 (0)	200 (0)
	PEL2-BIC	27.92 (2.51)	2.95 (0.04)	5.252 (0.871)	5.29 (0.23)
	PEL2-BICC	27.00 (2.69)	2.95 (0.04)	4.532 (0.552)	5.21 (0.24)
	PEL2-EBIC	24.80 (2.87)	2.94 (0.04)	4.657 (0.625)	5.26 (0.25)
$(100, 200, 400)$	MLE-Oracle	3 (0)	NA	0.014 (0.001)	NA
	MLE	200 (0)	3 (0)	3.438 (0.068)	NA
	PEL-BIC	0 (0)	0 (0)	15.25 (0)	400 (0)
	PEL-BICC	0 (0)	0 (0)	15.25 (0)	400 (0)
	PEL-EBIC	0 (0)	0 (0)	15.25 (0)	400 (0)
	PEL2-BIC	45.46 (4.37)	3 (0)	5.241 (0.793)	5.51 (0.19)
	PEL2-BICC	43.00 (4.25)	2.99 (0.01)	4.736 (0.659)	5.50 (0.18)
	PEL2-EBIC	42.40 (4.33)	2.99 (0.01)	4.546 (0.649)	5.52 (0.19)
$(100, 500, 1000)$	MLE-Oracle	3 (0)	NA	0.011 (0.001)	NA
	MLE	NA	NA	NA	NA
	PEL-BIC	0 (0)	0 (0)	15.25 (0)	1000 (0)
	PEL-BICC	0 (0)	0 (0)	15.25 (0)	1000 (0)
	PEL-EBIC	0 (0)	0 (0)	15.25 (0)	1000 (0)
	PEL2-BIC	30.02 (6.11)	2.93 (0.03)	2.300 (0.359)	6.70 (0.16)
	PEL2-BICC	26.73 (6.02)	2.93 (0.03)	2.430 (0.377)	6.62 (0.16)
	PEL2-EBIC	25.09 (5.91)	2.93 (0.03)	2.415 (0.377)	6.59 (0.16)

TABLE 3

Simulation results for regression model for longitudinal data with repeated measures based on 100 replicates. Here $\boldsymbol{\theta}_{\text{nonzeros}}$ *is the average number of selected nonzero components,* $\boldsymbol{\theta}_{\text{true}}$ *is the average number of true nonzero components that are selected, ME reports the model error, and No. EE's reports the number of estimating equations selected.*

when estimating $\boldsymbol{\theta}_0$, we use the quadratic estimating equations proposed by Qu et al. (2000): $\mathbf{g}(\mathbf{X}_i; \boldsymbol{\theta}) = \{(\mathbf{Y}_i - \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{v}_i^{-1/2}\mathbf{M}_1\mathbf{v}_i^{-1/2}\mathbf{Z}_i, \ldots, (\mathbf{Y}_i - \mathbf{Z}_i^{\mathrm{T}}\boldsymbol{\theta})^{\mathrm{T}}\mathbf{v}_i^{-1/2}\mathbf{M}_m\mathbf{v}_i^{-1/2}\mathbf{Z}_i\}^{\mathrm{T}}$ where $\mathbf{v}_i$ is a diagonal matrix of the conditional variances of subject i, and $\mathbf{M}_j$ $(j = 1, \ldots, m)$ are working correlation matrices. Note that when $m = 1$, i.e., using only one working correlation matrix $\mathbf{M}_1$, the model becomes the one in Liang and Zeger (1986) and we have $r = p$. Here we choose two sets of basis matrices with $\mathbf{M}_1$ being the identity matrix of size n_i and $\mathbf{M}_2$ being the compound symmetry with the diagonal elements of 1 and off-diagonal elements of ρ. In our setting, $m = 2$ and therefore $r = 2p$ estimating equations to estimate p parameters.

We obtain the same quantities as those in the example of Section 5.2, and report them in Table 3. In comparison of the single-penalty method, we can conclude from Table 3, with the columns defined in the same way as those in Table 2, that the proposed double-penalty method has much better performance. This confirms the efficacy and efficiency of adding the additional penalty on the Lagrange multiplier $\boldsymbol{\lambda}$, which performs the selection of estimating equations by reducing the number of estimating equations to less than 10.

6. Discussion.

We study a new penalized EL approach with two penalties, with one encouraging sparsity of the estimator and the other encouraging sparsity of the Lagrange multiplier in the optimizations associated with the EL. Such an approach utilizes sparsity in the target parameters and effectively achieves a moment selection procedure for estimating the sparse parameter. Both theory and numerical examples confirm the merits of the new approach. One interesting extension is to explore inferences with estimating equations after the variable selection procedure. Such direction is a suitable stage for EL method with estimating equations who takes advantage of adaptivity to various moment conditions with less stringent distributional assumptions. The other interesting and challenging problem is to explore the optimality of the sparse estimator using estimating equations with high data dimensionality. Semiparametric efficiency of EL with estimating equations is shown in Qin and Lawless (1994). However, when the paradigm shifts to high-dimensional problems, the efficiency of the sparse estimator respecting its nonzero components remains open for further investigations. We plan to address the problems in future works.

7. Proofs.

In the sequel, we use the abbreviations "w.p.a.1" and "w.r.t" to denote, respectively, "with probability approaching one" and "with respect to", and C denotes a generic positive finite constant that may be different in different uses. For simplicity and when no confusion arises, we use notation $\mathbf{h}_i(\boldsymbol{\theta})$ as equivalent to $\mathbf{h}(\mathbf{X}_i;\boldsymbol{\theta})$ for a generic q-dimensional multivariate function $\mathbf{h}(\cdot;\cdot)$ and denote by $h_{i,k}(\boldsymbol{\theta})$ the kth component of $\mathbf{h}_i(\boldsymbol{\theta})$. Let $\bar{\mathbf{h}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathbf{h}_i(\boldsymbol{\theta})$, and $\bar{h}_k(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n h_{i,k}(\boldsymbol{\theta})$ be the kth component of $\bar{\mathbf{h}}(\boldsymbol{\theta})$. For a given set $\mathcal{L} \subset \{1,\ldots,q\}$, we denote by $\mathbf{h}_{\mathcal{L}}(\cdot;\cdot)$ the subvector of $\mathbf{h}(\cdot;\cdot)$ collecting the components indexed by $\mathcal{L}$. Analogously, we let $\mathbf{h}_{i,\mathcal{L}}(\boldsymbol{\theta}) = \mathbf{h}_{\mathcal{L}}(\mathbf{X}_i;\boldsymbol{\theta})$ and $\bar{\mathbf{h}}_{\mathcal{L}}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^n \mathbf{h}_{i,\mathcal{L}}(\boldsymbol{\theta})$.

7.1. *Proof of Proposition* 1. Define $A_n(\boldsymbol{\theta},\boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta})\}$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta})$. Let $\widetilde{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda})$. Pick $\delta_n = o(r^{-1/2}n^{-1/\gamma})$ and $r^{1/2}n^{-1/2} = o(\delta_n)$. Let $\bar{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\Lambda_n} A_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda})$ where $\Lambda_n = \{\boldsymbol{\lambda} \in \mathbb{R}^r : |\boldsymbol{\lambda}|_2 \le \delta_n\}$. By Markov inequality, $\max_{1\le i\le n} |\mathbf{g}_i(\boldsymbol{\theta}_0)|_2 = O_p(r^{1/2}n^{1/\gamma})$. Then $\max_{1\le i\le n,\boldsymbol{\lambda}\in\Lambda_n} |\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_0)| = o_p(1)$. By Taylor expansion,

$$
\begin{aligned}
0 = A_n(\boldsymbol{\theta}_0,\mathbf{0}) \le A_n(\boldsymbol{\theta}_0,\bar{\boldsymbol{\lambda}}) &= \bar{\boldsymbol{\lambda}}^{\mathrm{T}}\bar{\mathbf{g}}(\boldsymbol{\theta}_0) - \frac{1}{2n}\sum_{i=1}^n \frac{\bar{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_0)\mathbf{g}_i(\boldsymbol{\theta}_0)^{\mathrm{T}}\bar{\boldsymbol{\lambda}}}{\{1+c\bar{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_0)\}^2} \\
&\le |\bar{\boldsymbol{\lambda}}|_2|\bar{\mathbf{g}}(\boldsymbol{\theta}_0)|_2 - C|\bar{\boldsymbol{\lambda}}|_2^2\{1+o_p(1)\}.
\end{aligned}
\tag{7.1}
$$

Notice that $|\bar{\mathbf{g}}(\boldsymbol{\theta}_0)|_2 = O_p(r^{1/2}n^{-1/2})$, (7.1) yields $|\bar{\boldsymbol{\lambda}}|_2 = O_p(r^{1/2}n^{-1/2}) = o_p(\delta_n)$. Thus $\bar{\boldsymbol{\lambda}} \in \mathrm{int}(\Lambda_n)$ w.p.a.1. Since $\Lambda_n \subset \widehat{\Lambda}_n(\boldsymbol{\theta}_0)$ w.p.a.1, $\widetilde{\boldsymbol{\lambda}} = \bar{\boldsymbol{\lambda}}$ w.p.a.1 by the concavity of $A_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda})$ and $\widehat{\Lambda}_n(\boldsymbol{\theta}_0)$. Hence, $\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda}) =$

$O_p(rn^{-1})$. For δ_n specified above, let $\boldsymbol{\lambda}^* = \delta_n\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})/|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2$, then $\boldsymbol{\lambda}^* \in \Lambda_n$. By Taylor expansion,

$$\begin{aligned} A_n(\widehat{\boldsymbol{\theta}},\boldsymbol{\lambda}^*) &= \boldsymbol{\lambda}^{*,\mathrm{T}}\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) - \frac{1}{2n}\sum_{i=1}^{n}\frac{\boldsymbol{\lambda}^{*,\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol{\theta}})\mathbf{g}_i(\widehat{\boldsymbol{\theta}})^{\mathrm{T}}\boldsymbol{\lambda}^*}{\{1+c\boldsymbol{\lambda}^{*,\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol{\theta}})\}^2} \\ &\geq \delta_n|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2 - C\delta_n^2\{1+o_p(1)\}. \end{aligned} \tag{7.2}$$

Since $A_n(\widehat{\boldsymbol{\theta}},\boldsymbol{\lambda}^*) \leq \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda}) = O_p(rn^{-1})$, then $|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2 = O_p(\delta_n)$. Consider any $\epsilon_n \to 0$ and let $\boldsymbol{\lambda}^{**} = \epsilon_n\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})$, then $|\boldsymbol{\lambda}^{**}|_2 = o_p(\delta_n)$. We have $\epsilon_n|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2^2 - C\epsilon_n^2|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2^2\{1+o_p(1)\} = O_p(rn^{-1})$. Then $\epsilon_n|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2^2 = O_p(rn^{-1})$. Notice that we can select arbitrary slow $\epsilon_n \to 0$, following a standard result from probability theory, we have $|\bar{\mathbf{g}}(\widehat{\boldsymbol{\theta}})|_2^2 = O_p(rn^{-1})$. □

7.2. *Proof of Proposition* 2. With $A_n(\boldsymbol{\theta},\boldsymbol{\lambda})$ defined in Section 7.1, define $F_n(\boldsymbol{\theta}) = \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})} A_n(\boldsymbol{\theta},\boldsymbol{\lambda}) + \sum_{k=1}^{p} P_{1,\pi}(|\theta_k|)$. Recall $a_n = \sum_{k=1}^{p} P_{1,\pi}(|\theta_k^0|)$ and $b_n = \max\{rn^{-1}, a_n\}$. Since $\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda}) = O_p(rn^{-1})$, then $F_n(\boldsymbol{\theta}_0) = O_p(b_n)$. Define $\boldsymbol{\Theta}_* = \{\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}} : |\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty \leq \varepsilon, |\boldsymbol{\theta}_{\mathcal{S}^c}|_1 \leq n^{-1/2}\varphi_n^{-1}\}$ for some fixed $\varepsilon > 0$. Let $\widetilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_*} F_n(\boldsymbol{\theta})$. As $F_n(\widetilde{\boldsymbol{\theta}}_n) \leq F_n(\boldsymbol{\theta}_0)$, then $F_n(\widetilde{\boldsymbol{\theta}}_n) \leq O_p(b_n)$. We will first show $\widetilde{\boldsymbol{\theta}}_n \in \mathrm{int}(\boldsymbol{\Theta}_*)$ w.p.a.1. Our proof includes two steps: (i) to show that for any $\epsilon_n \to \infty$ satisfying $b_n\epsilon_n^{2\beta}n^{2/\gamma} = o(1)$, there exists a uniform constant $K > 0$ independent of $\boldsymbol{\theta}$ such that $\mathbb{P}\{F_n(\boldsymbol{\theta}) > Kb_n\epsilon_n^{2\beta}\} \to 1$ as $n \to \infty$ for any $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Theta}_*$ satisfying $|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty > \epsilon_n b_n^{1/(2\beta)}$. Thus $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty = O_p\{\epsilon_n b_n^{1/(2\beta)}\}$. Notice that we can select arbitrary slow diverging ϵ_n, then $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty = O_p\{b_n^{1/(2\beta)}\}$, (ii) to show that $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1 < n^{-1/2}\varphi_n^{-1}$.

For (i), we will use the technique developed for the proof of Theorem 1 in Chang et al. (2013). For any $\boldsymbol{\theta} = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}} \in \boldsymbol{\Theta}_*$ satisfying $|\boldsymbol{\theta}_{\mathcal{S}} - \boldsymbol{\theta}_{0,\mathcal{S}}|_\infty > \epsilon_n b_n^{1/(2\beta)}$, take $\boldsymbol{\theta}^* = (\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ and $j_0 = \arg\max_{1\leq j\leq r}|\mathbb{E}\{g_{i,j}(\boldsymbol{\theta}^*)\}|$. Let $\mu_{j_0} = \mathbb{E}\{g_{i,j_0}(\boldsymbol{\theta})\}$, $\mu_{j_0}^* = \mathbb{E}\{g_{i,j_0}(\boldsymbol{\theta}^*)\}$, and $\widetilde{\boldsymbol{\lambda}} = \delta b_n^{1/2}\epsilon_n^{\beta}\mathbf{e}_{j_0}$ where $\delta > 0$ is a constant to be determined later, and $\mathbf{e}_{j_0}$ is an r-dimensional vector with the j_0-th component being 1 and other components being 0. Without lose of generality, we assume $\mu_{j_0}^* > 0$. (2.4) and Markov inequality yield $\max_{1\leq i\leq n}|g_{i,j_0}(\boldsymbol{\theta})| = O_p(n^{1/\gamma})$ and $\max_{1\leq i\leq n}|\widetilde{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta})| = O_p(b_n^{1/2}\epsilon_n^{\beta}n^{1/\gamma}) = o_p(1)$. Then $\widetilde{\boldsymbol{\lambda}} \in \widehat{\Lambda}_n(\boldsymbol{\theta})$ w.p.a.1. Write $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\lambda}} = (\widetilde{\lambda}_1,\ldots,\widetilde{\lambda}_r)^{\mathrm{T}}$. It holds w.p.a.1 that $F_n(\boldsymbol{\theta}) \geq n^{-1}\sum_{i=1}^{n}\log\{1+\widetilde{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta})\} + \sum_{k=1}^{p}P_{1,\pi}(|\theta_k|) \geq n^{-1}\sum_{i=1}^{n}\widetilde{\lambda}_{j_0}g_{i,j_0}(\boldsymbol{\theta}) - n^{-1}\sum_{i=1}^{n}\{\widetilde{\lambda}_{j_0}g_{i,j_0}(\boldsymbol{\theta})\}^2$. Thus, $\mathbb{P}\{F_n(\boldsymbol{\theta}) \leq Kb_n\epsilon_n^{2\beta}\} \leq \mathbb{P}[\bar{g}_{j_0}(\boldsymbol{\theta}) - \mu_{j_0} \leq b_n^{1/2}\epsilon_n^{\beta}\{K\delta^{-1} + \delta n^{-1}\sum_{i=1}^{n}g_{i,j_0}^2(\boldsymbol{\theta})\} - \mu_{j_0}] + o(1)$. From (2.4) and Markov inequality, there is a uniform constant $L > 0$ such that

$\mathbb{P}\{n^{-1}\sum_{i=1}^{n} g_{i,j_0}^2(\boldsymbol{\theta}) > L\} \to 0$ for any $\boldsymbol{\theta}$. With $\delta = (K/L)^{1/2}$, we have $\mathbb{P}\{F_n(\boldsymbol{\theta}) \le K b_n \epsilon_n^{2\beta}\} \le \mathbb{P}\{\bar{g}_{j_0}(\boldsymbol{\theta}) - \mu_{j_0} \le 2b_n^{1/2}\epsilon_n^{\beta}(KL)^{1/2} - \mu_{j_0}\} + o(1)$. From (2.6) and (2.9), $\mu_{j_0}^* \ge \Delta(\epsilon_n b_n^{1/(2\beta)}) \ge K_1 \epsilon_n^{\beta} b_n^{1/2}/2$ with K_1 specified in (2.6) for sufficiently large n, and $|\mu_{j_0} - \mu_{j_0}^*| \le \sum_{k \notin \mathcal{S}} \mathbb{E}\{\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_*} |\partial g_{i,j_0}(\boldsymbol{\theta})/\partial \theta_k|\}|\theta_k| \le K_2 |\boldsymbol{\theta}_{\mathcal{S}^c}|_1 = o(b_n^{1/2})$ for K_2 specified in (2.9). Therefore, $\mu_{j_0} \ge K_1 \epsilon_n^{\beta} b_n^{1/2}/3$ for sufficiently large n. For sufficiently small K independent of $\boldsymbol{\theta}$, we have $2b_n^{1/2}\epsilon_n^{\beta}(KL)^{1/2} - \mu_{j_0} \le -c\mu_{j_0}$ for some $c \in (0,1)$. Then $n^{1/2}\{2b_n^{1/2}\epsilon_n^{\beta}(KL)^{1/2} - \mu_{j_0}\} \lesssim -\epsilon_n^{\beta} b_n^{1/2} n^{1/2} \to -\infty$. As $n^{1/2}\{\bar{g}_{j_0}(\boldsymbol{\theta}) - \mu_{j_0}\} \to_d N(0, \sigma^2)$ for some $\sigma > 0$, then $\mathbb{P}\{F_n(\boldsymbol{\theta}) \le K b_n \epsilon_n^{2\beta}\} \to 0$. We complete the proof for (i).

For (ii), if $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1 = n^{-1/2}\varphi_n^{-1}$, we will take $\widetilde{\boldsymbol{\theta}}_n^* = (\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}}^{\mathrm{T}}, \tau \widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}$ for some $\tau \in (0,1)$ and show $F_n(\widetilde{\boldsymbol{\theta}}_n^*) < F_n(\widetilde{\boldsymbol{\theta}}_n)$ w.p.a.1. Since $\widetilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}_*} F_n(\boldsymbol{\theta})$, it is a contradiction. Thus $|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1 < n^{-1/2}\varphi_n^{-1}$. Write $\widetilde{\boldsymbol{\theta}}_n = (\widetilde{\theta}_{n,1}, \ldots, \widetilde{\theta}_{n,p})^{\mathrm{T}}$ and $\widetilde{\boldsymbol{\theta}}_n^* = (\widetilde{\theta}_{n,1}^*, \ldots, \widetilde{\theta}_{n,p}^*)^{\mathrm{T}}$. By $F_n(\widetilde{\boldsymbol{\theta}}_n) \le F_n(\boldsymbol{\theta}_0)$, $\max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\widetilde{\boldsymbol{\theta}}_n)} A_n(\widetilde{\boldsymbol{\theta}}_n, \boldsymbol{\lambda}) \le \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) + \sum_{k=1}^{p} P_{1,\pi}(|\theta_k^0|) - \sum_{k=1}^{p} P_{1,\pi}(|\widetilde{\theta}_{n,k}|)$. Notice that

$$\begin{aligned} &\sum_{k=1}^{p} P_{1,\pi}(|\theta_k^0|) - \sum_{k=1}^{p} P_{1,\pi}(|\widetilde{\theta}_{n,k}|) \\ &\quad \le \sum_{k=1}^{s} P_{1,\pi}'\{c_k|\widetilde{\theta}_{n,k}| + (1-c_k)|\theta_k^0|\}|\widetilde{\theta}_{n,k} - \theta_k^0| = O_p\{s\chi_n b_n^{1/(2\beta)}\} \end{aligned} \tag{7.3}$$

for some $c_k \in (0,1)$. Since $\max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\boldsymbol{\theta}_0)} A_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}) = O_p(rn^{-1})$, it holds that $\max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\widetilde{\boldsymbol{\theta}}_n)} A_n(\widetilde{\boldsymbol{\theta}}_n, \boldsymbol{\lambda}) = O_p(rn^{-1}) + O_p\{s\chi_n b_n^{1/(2\beta)}\}$. Pick $\delta_n = o(r^{-1/2}n^{-1/\gamma})$ and $\max\{rn^{-1}, s\chi_n b_n^{1/(2\beta)}\} = o(\delta_n^2)$. Same as (7.2), we know $o_p(\delta_n^2) = \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\widetilde{\boldsymbol{\theta}}_n)} A_n(\widetilde{\boldsymbol{\theta}}_n, \boldsymbol{\lambda}) \ge \delta_n |\bar{\mathbf{g}}(\widetilde{\boldsymbol{\theta}}_n)|_2 - C\delta_n^2\{1 + o_p(1)\}$, which implies $|\bar{\mathbf{g}}(\widetilde{\boldsymbol{\theta}}_n)|_2 = O_p(\delta_n)$. Following the same arguments below (7.2), we have $|\bar{\mathbf{g}}(\widetilde{\boldsymbol{\theta}}_n)|_2 = O_p(r^{1/2}n^{-1/2}) + O_p\{s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\}$. Notice that $|\bar{\mathbf{g}}(\widetilde{\boldsymbol{\theta}}_n^*)|_2 \le |\bar{\mathbf{g}}(\widetilde{\boldsymbol{\theta}}_n)|_2 + |\{\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}(\bar{\boldsymbol{\theta}})\}(\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n)|_2$ for some $\bar{\boldsymbol{\theta}}$ between $\widetilde{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n^*$. Since $\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}} = \widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}}^*$, by (2.10), $|\{\nabla_{\boldsymbol{\theta}} \bar{\mathbf{g}}(\bar{\boldsymbol{\theta}})\}(\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n)|_2 = O_p(r^{1/2}n^{-1/2})$. Hence, $|\bar{\mathbf{g}}(\widetilde{\boldsymbol{\theta}}_n^*)|_2 = O_p(r^{1/2}n^{-1/2}) + O_p\{s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\}$. Let $\boldsymbol{\lambda}^* = \arg\max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\widetilde{\boldsymbol{\theta}}_n^*)} A_n(\widetilde{\boldsymbol{\theta}}_n^*, \boldsymbol{\lambda})$. With the arguments for (7.1), $|\boldsymbol{\lambda}^*|_2 = O_p(r^{1/2}n^{-1/2}) + O_p\{s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\}$. Since $F_n(\widetilde{\boldsymbol{\theta}}_n) \ge A_n(\widetilde{\boldsymbol{\theta}}_n, \boldsymbol{\lambda}^*) + \sum_{k=1}^{p} P_{1,\pi}(|\widetilde{\theta}_{n,k}|)$, then $F_n(\widetilde{\boldsymbol{\theta}}_n^*) \le F_n(\widetilde{\boldsymbol{\theta}}_n) + [n^{-1}\sum_{i=1}^{n} \boldsymbol{\lambda}^{*,\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathbf{g}_i(\check{\boldsymbol{\theta}})\{1 + \boldsymbol{\lambda}^{*,\mathrm{T}} \mathbf{g}_i(\check{\boldsymbol{\theta}})\}^{-1}](\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n) + \sum_{k=s+1}^{p} P_{1,\pi}(\tau|\widetilde{\theta}_{n,k}|) - \sum_{k=s+1}^{p} P_{1,\pi}(|\widetilde{\theta}_{n,k}|)$ for some $\check{\boldsymbol{\theta}}$ between $\widetilde{\boldsymbol{\theta}}_n$ and $\widetilde{\boldsymbol{\theta}}_n^*$. Meanwhile, notice that $\max_{1 \le i \le n} |\boldsymbol{\lambda}^{*,\mathrm{T}} \mathbf{g}_i(\check{\boldsymbol{\theta}})| = o_p(1)$, then we have $|[n^{-1}\sum_{i=1}^{n} \boldsymbol{\lambda}^{*,\mathrm{T}} \nabla_{\boldsymbol{\theta}} \mathbf{g}_i(\check{\boldsymbol{\theta}})\{1 + \boldsymbol{\lambda}^{*,\mathrm{T}} \mathbf{g}_i(\check{\boldsymbol{\theta}})\}^{-1}](\widetilde{\boldsymbol{\theta}}_n^* - \widetilde{\boldsymbol{\theta}}_n)| \le |\boldsymbol{\lambda}^*|_2 |[n^{-1}\sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \mathbf{g}_i(\check{\boldsymbol{\theta}})\{1 + \boldsymbol{\lambda}^{*,\mathrm{T}} \mathbf{g}_i(\check{\boldsymbol{\theta}})\}^{-1}](\widetilde{\boldsymbol{\theta}}_n^* -$

$\widetilde{\boldsymbol{\theta}}_n)|_2 \le |\boldsymbol{\lambda}^*|_2|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1 O_p(r^{1/2}\varphi_n)$. On the other hand, $\sum_{k=s+1}^p P_{1,\pi}(\tau|\widetilde{\theta}_{n,k}|) - \sum_{k=s+1}^p P_{1,\pi}(|\widetilde{\theta}_{n,k}|) = -(1-\tau)\cdot\sum_{k=s+1}^p P_{1,\pi}'\{(c_k\tau+1-c_k)|\widetilde{\theta}_{n,k}|\}|\widetilde{\theta}_{n,k}| \le -(1-\tau)C\pi\sum_{k=s+1}^p|\widetilde{\theta}_{n,k}| = -(1-\tau)C\pi|\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1$ for some $c_k\in(0,1)$. If $r^{1/2}\varphi_n\max\{r^{1/2}n^{-1/2}, s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\} = o(\pi)$, we have $F_n(\widetilde{\boldsymbol{\theta}}_n^*) < F_n(\widetilde{\boldsymbol{\theta}}_n)$ w.p.a.1. We complete the proof of (ii).

Nextly, we will show $\mathbb{P}(\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c} = \mathbf{0}) \to 1$. Let $\widehat{G}_n(\boldsymbol{\theta},\boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta})\} + \sum_{k=1}^p P_{1,\pi}(|\theta_k|)$. Then $\widetilde{\boldsymbol{\theta}}_n$ and its Lagrange multiplier $\widehat{\boldsymbol{\lambda}}$ satisfy $\nabla_{\boldsymbol{\lambda}}\widehat{G}_n(\widetilde{\boldsymbol{\theta}}_n,\widehat{\boldsymbol{\lambda}}) = \mathbf{0}$. By the implicit theorem [Theorem 9.28 of Rudin (1976)], for all $\boldsymbol{\theta}$ in a $|\cdot|_2$-neighborhood of $\widetilde{\boldsymbol{\theta}}_n$, there is a $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ such that $\nabla_{\boldsymbol{\lambda}}\widehat{G}_n\{\boldsymbol{\theta},\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \mathbf{0}$ and $\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$. By the concavity of $\widehat{G}_n(\boldsymbol{\theta},\boldsymbol{\lambda})$ w.r.t $\boldsymbol{\lambda}$, $\widehat{G}_n\{\boldsymbol{\theta},\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\} = \max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})}\widehat{G}_n(\boldsymbol{\theta},\boldsymbol{\lambda})$. Write $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1,\ldots,\widehat{\lambda}_r)^{\mathrm{T}}$. From the envelope theorem, $\mathbf{0} = \nabla_{\boldsymbol{\theta}}\widehat{G}_n\{\boldsymbol{\theta},\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_n}$. Write $\widehat{\mathbf{h}} = (\widehat{h}_1,\ldots,\widehat{h}_p)^{\mathrm{T}} = \nabla_{\boldsymbol{\theta}}\widehat{G}_n\{\boldsymbol{\theta},\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\}|_{\boldsymbol{\theta}=\widetilde{\boldsymbol{\theta}}_n}$. Then $\widehat{h}_k = n^{-1}\sum_{i=1}^n\sum_{j=1}^r\widehat{\lambda}_j\{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\widetilde{\boldsymbol{\theta}}_n)\}^{-1}\partial g_{i,j}(\widetilde{\boldsymbol{\theta}}_n)/\partial\theta_k + \widehat{\kappa}_k$ with $\widehat{\kappa}_k = \pi\rho_1'(|\widetilde{\theta}_k|;\pi)\mathrm{sgn}(\widetilde{\theta}_k)$ for $\widetilde{\theta}_k\neq 0$ and $\widehat{\kappa}_k\in[-\pi\rho_1'(0^+),\pi\rho_1'(0^+)]$ otherwise. Since $\sup_{k\notin\mathcal{S}}|n^{-1}\sum_{i=1}^n\sum_{j=1}^r\widehat{\lambda}_j\{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\widetilde{\boldsymbol{\theta}}_n)\}^{-1}\partial g_{i,j}(\widetilde{\boldsymbol{\theta}}_n)/\partial\theta_k| \le [\sum_{j=1}^r|\widehat{\lambda}_j|\sup_{k\notin\mathcal{S}}\{n^{-1}\sum_{i=1}^n|\partial g_{i,j}(\widetilde{\boldsymbol{\theta}}_n)/\partial\theta_k|\}]$ $\cdot\{1+o_p(1)\} \le O_p(\varphi_n)\cdot\sum_{j=1}^r|\widehat{\lambda}_j| = o(\pi)$, if $\widetilde{\theta}_k\neq 0$ for some $k\notin\mathcal{S}$, then $\pi\rho_1'(|\widetilde{\theta}_k|;\pi)\mathrm{sgn}(\widetilde{\theta}_k)$ will dominates the sign of $\widehat{h}_k$. By the arguments for the proof of Lemma 1 in Fan and Li (2001), we know $\widetilde{\boldsymbol{\theta}}_{n,\mathcal{S}^c} = \mathbf{0}$ w.p.a.1. □

7.3. *Proof of Proposition* 3. Recall $\mathcal{M}_{\boldsymbol{\theta}_n} = \{1\le j\le r : |\bar{g}_j(\boldsymbol{\theta}_n)| \ge \nu\rho_2'(0^+)\}$ and $\mathcal{M}_{\boldsymbol{\theta}_n}^* = \{1\le j\le r : |\bar{g}_j(\boldsymbol{\theta}_n)| \ge C\nu\rho_2'(0^+)\}$ for some $C\in(0,1)$. Clearly, $\mathcal{M}_{\boldsymbol{\theta}_n}\subset\mathcal{M}_{\boldsymbol{\theta}_n}^*$. Recall $m_n = |\mathcal{M}_{\boldsymbol{\theta}_n}^*|$. Given $\mathcal{M}_{\boldsymbol{\theta}_n}$, we select δ_n satisfying $\delta_n = o(m_n^{-1/2}n^{-1/\gamma})$ and $u_n = o(\delta_n)$. Let $\bar{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda}\in\Lambda_n} f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$ where $\Lambda_n = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}}^{\mathrm{T}},\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^c}^{\mathrm{T}})^{\mathrm{T}}\in\mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}}|_2\le\delta_n \text{ and } \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^c} = \mathbf{0}\}$. For given $\mathcal{M}_{\boldsymbol{\theta}_n}$, by Condition 3 and Markov inequality, $\max_{1\le i\le n}|\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)|_2 = O_p(m_n^{1/2}n^{1/\gamma})$, which leads to $\max_{1\le i\le n}|\bar{\boldsymbol{\lambda}}_n^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n)| = o_p(1)$. Write $\bar{\boldsymbol{\lambda}}_n = (\bar{\lambda}_{n,1},\ldots,\bar{\lambda}_{n,r})^{\mathrm{T}}$. Notice that $\mathbb{P}[\lambda_{\min}\{\widehat{\mathbf{V}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}\ge C]\to 1$. By the definition of $\bar{\boldsymbol{\lambda}}_n$ and Taylor expansion, noting $P_{2,\nu}(t) = \nu\rho_2(t;\nu)$ and $\rho_2'(t;\nu)\ge\rho_2'(0^+)$ for any $t>0$, we have

$$\begin{aligned}
0 &= f(\mathbf{0};\boldsymbol{\theta}_n) \le f(\bar{\boldsymbol{\lambda}}_n;\boldsymbol{\theta}_n)\\
&= \frac{1}{n}\sum_{i=1}^n\bar{\boldsymbol{\lambda}}_n^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n) - \frac{1}{2n}\sum_{i=1}^n\frac{\bar{\boldsymbol{\lambda}}_n^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_0)\mathbf{g}_i(\boldsymbol{\theta}_n)^{\mathrm{T}}\bar{\boldsymbol{\lambda}}_n}{\{1+c\bar{\boldsymbol{\lambda}}_n^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n)\}^2} - \sum_{j=1}^r P_{2,\nu}(|\bar{\lambda}_{n,j}|)\\
&\le \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}}^{\mathrm{T}}[\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}] - C|\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}}|_2^2\{1+o_p(1)\}
\end{aligned}$$

Since $|\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)\}|_2 = O_p(u_n)$, then $|\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}}|_2 = O_p(u_n) = o_p(\delta_n)$. Write $\bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}} = (\bar{\lambda}_1, \ldots, \bar{\lambda}_{|\mathcal{M}_{\boldsymbol{\theta}_n}|})^{\mathrm{T}}$. We have w.p.a.1 that

$$\mathbf{0} = \frac{1}{n}\sum_{i=1}^{n} \frac{\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)}{1 + \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}}^{\mathrm{T}} \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n)} - \widehat{\boldsymbol{\eta}} \tag{7.4}$$

where $\widehat{\boldsymbol{\eta}} = (\widehat{\eta}_1, \ldots, \widehat{\eta}_{|\mathcal{M}_{\boldsymbol{\theta}_n}|})^{\mathrm{T}}$ with $\widehat{\eta}_j = \nu\rho_2'(|\bar{\lambda}_j|;\nu)\mathrm{sgn}(\bar{\lambda}_j)$ for $\bar{\lambda}_j \neq 0$ and $\widehat{\eta}_j \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $\bar{\lambda}_j = 0$. (7.4) implies that $\widehat{\boldsymbol{\eta}} = \bar{\mathbf{g}}_{\mathcal{M}_{\boldsymbol{\theta}_n}}(\boldsymbol{\theta}_n) + \mathbf{R}$ with $|\mathbf{R}|_\infty = O_p(\varsigma_n^{1/2}u_n)$. Since $\varsigma_n^{1/2}u_n = o(\nu)$, then w.p.a.1 $\mathrm{sgn}(\bar{\lambda}_j) = \mathrm{sgn}\{\bar{g}_j(\boldsymbol{\theta}_n)\}$ for any $\bar{\lambda}_j \neq 0$.

We will show that $\bar{\boldsymbol{\lambda}}_n$ is a local maximizer for $f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$ w.p.a.1. We first show that $\bar{\boldsymbol{\lambda}}_n = \arg\max_{\boldsymbol{\lambda}\in\Lambda_n^*(\boldsymbol{\theta}_n)} f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$ w.p.a.1, where $\Lambda_n^*(\boldsymbol{\theta}_n) = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^*}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^*}|_2 \leq \epsilon, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}} = \mathbf{0}\}$ for some $\epsilon > 0$. Since $f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$ is concave w.r.t $\boldsymbol{\lambda}$, it suffices to show that $\mathbf{w} = \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}^*}^{\mathrm{T}} =: (w_1, \ldots, w_{m_n})^{\mathrm{T}}$ satisfies $\mathbf{0} = n^{-1}\sum_{i=1}^n \mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}^*}(\boldsymbol{\theta}_n)\{1 + \mathbf{w}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}^*}(\boldsymbol{\theta}_n)\}^{-1} - \widehat{\boldsymbol{\eta}}^*$ w.p.a.1, where $\widehat{\boldsymbol{\eta}}^* = (\widehat{\eta}_1^*, \ldots, \widehat{\eta}_{m_n}^*)^{\mathrm{T}}$ with $\widehat{\eta}_j^* = \nu\rho_2'(|w_j|;\nu)\mathrm{sgn}(w_j)$ for $w_j \neq 0$ and $\widehat{\eta}_j^* \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $w_j = 0$. From (7.4), we know $0 = n^{-1}\sum_{i=1}^n g_{i,j}(\boldsymbol{\theta}_n)\{1 + \mathbf{w}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}^*}(\boldsymbol{\theta}_n)\}^{-1} - \widehat{\eta}_j^*$ holds for any $j \in \mathcal{M}_{\boldsymbol{\theta}_n}$. For each $j \in \mathcal{M}_{\boldsymbol{\theta}_n}^* \backslash \mathcal{M}_{\boldsymbol{\theta}_n}$, it holds that $n^{-1}\sum_{i=1}^n g_{i,j}(\boldsymbol{\theta}_n)\{1 + \mathbf{w}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}^*}(\boldsymbol{\theta}_n)\}^{-1} = \bar{g}_j(\boldsymbol{\theta}_n) + O_p(\varsigma_n^{1/2}u_n)$ where $O_p(\varsigma_n^{1/2}u_n)$ is uniform for any $j \in \mathcal{M}_{\boldsymbol{\theta}_n}^* \backslash \mathcal{M}_{\boldsymbol{\theta}_n}$. Since $C\nu\rho_2'(0^+) \leq |\bar{g}_j(\boldsymbol{\theta}_n)| < \nu\rho_2'(0^+)$ for $j \in \mathcal{M}_{\boldsymbol{\theta}_n}^* \backslash \mathcal{M}_{\boldsymbol{\theta}_n}$, if $\varsigma_n^{1/2}u_n = o(\nu)$, then $|n^{-1}\sum_{i=1}^n g_{i,j}(\boldsymbol{\theta}_n)\{1 + \mathbf{w}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}^*}(\boldsymbol{\theta}_n)\}^{-1}| < \nu\rho_2'(0^+)$ w.p.a.1 for any $j \in \mathcal{M}_{\boldsymbol{\theta}_n}^* \backslash \mathcal{M}_{\boldsymbol{\theta}_n}$. Then there exists $\widehat{\eta}_j^*$ such that $0 = n^{-1}\sum_{i=1}^n g_{i,j}(\boldsymbol{\theta}_n)\{1 + \mathbf{w}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{M}_{\boldsymbol{\theta}_n}^*}(\boldsymbol{\theta}_n)\}^{-1} - \widehat{\eta}_j^*$ holds for any $j \in \mathcal{M}_{\boldsymbol{\theta}_n}^* \backslash \mathcal{M}_{\boldsymbol{\theta}_n}$.

Secondly, we prove $\bar{\boldsymbol{\lambda}}_n$ is a local maximizer for $f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$ over $\boldsymbol{\lambda} \in \widetilde{\Lambda}_n(\boldsymbol{\theta}_n)$ w.p.a.1, where $\widetilde{\Lambda}_n(\boldsymbol{\theta}_n) = \{\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^*}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}^{\mathrm{T}})^{\mathrm{T}} \in \mathbb{R}^r : |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^*} - \bar{\boldsymbol{\lambda}}_{n,\mathcal{M}_{\boldsymbol{\theta}_n}^*}|_2 \leq o(u_n), |\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}|_1 = o(r^{-1/\gamma}n^{-1/\gamma})\}$. Note that $\max_{1\leq i\leq n, \boldsymbol{\lambda}\in\widetilde{\Lambda}_n(\boldsymbol{\theta}_n)} |\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n)| = o_p(1)$. For any $\boldsymbol{\lambda} \in \widetilde{\Lambda}_n(\boldsymbol{\theta}_n)$, we write $\boldsymbol{\lambda} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^*}^{\mathrm{T}}, \boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}^{\mathrm{T}})^{\mathrm{T}}$ and denote by $\widetilde{\boldsymbol{\lambda}} = (\boldsymbol{\lambda}_{\mathcal{M}_{\boldsymbol{\theta}_n}^*}^{\mathrm{T}}, \mathbf{0}^{\mathrm{T}})^{\mathrm{T}}$ the projection of $\boldsymbol{\lambda}$ onto the subspace $\Lambda_n^*(\boldsymbol{\theta}_n)$. We only need to show $\mathbb{P}[\sup_{\boldsymbol{\lambda}\in\widetilde{\Lambda}_n(\boldsymbol{\theta}_n)}\{f(\boldsymbol{\lambda};\boldsymbol{\theta}_n) - f(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\theta}_n)\} \leq 0] \to 1$. By Taylor expansion, $\sup_{\boldsymbol{\lambda}\in\widetilde{\Lambda}_n(\boldsymbol{\theta}_n)}\{f(\boldsymbol{\lambda};\boldsymbol{\theta}_n) - f(\widetilde{\boldsymbol{\lambda}};\boldsymbol{\theta}_n)\} = \sup_{\boldsymbol{\lambda}\in\widetilde{\Lambda}_n(\boldsymbol{\theta}_n)}[n^{-1}\sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}_n)^{\mathrm{T}}(\boldsymbol{\lambda} - \widetilde{\boldsymbol{\lambda}})\{1 + \boldsymbol{\lambda}_*^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n)\}^{-1} - \sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}} P_{2,\nu}(|\lambda_j|)]$ for some $\boldsymbol{\lambda}_*$ between $\boldsymbol{\lambda}$ and $\widetilde{\boldsymbol{\lambda}}$. We have $|n^{-1}\sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}_n)^{\mathrm{T}}(\boldsymbol{\lambda} - \widetilde{\boldsymbol{\lambda}})\{1 + \boldsymbol{\lambda}_*^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n)\}^{-1}| \leq C\nu\rho_2'(0^+)\sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}|\lambda_j| + O_p(m_n^{1/2}u_n\varsigma_n) \cdot \sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}|\lambda_j|$, where the term $O_p(m_n^{1/2}u_n\varsigma_n)$ is uniformly for any $\boldsymbol{\lambda} \in \widetilde{\Lambda}_n(\boldsymbol{\theta}_n)$. On the other hand, we have $\sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}} P_{2,\nu}(|\lambda_j|) \geq$

$\nu\rho_2'(0^+)\sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}|\lambda_j|$. Hence, $n^{-1}\sum_{i=1}^n \mathbf{g}_i(\boldsymbol{\theta}_n)^{\mathrm{T}}(\boldsymbol{\lambda}-\widetilde{\boldsymbol{\lambda}})\{1+\boldsymbol{\lambda}_*^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta}_n)\}^{-1}-\sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}P_{2,\nu}(|\lambda_j|)\le\{-(1-C)\nu\rho_2'(0^+)+O_p(m_n^{1/2}u_n\varsigma_n)\}\sum_{j\in\mathcal{M}_{\boldsymbol{\theta}_n}^{*,c}}|\lambda_j|$. Notice that $m_n^{1/2}u_n\varsigma_n/\nu\to 0$, then $-(1-C)\nu\rho_2'(0^+)+O_p(m_n^{1/2}u_n\varsigma_n)\le 0$ w.p.a.1. Hence, $\bar{\boldsymbol{\lambda}}_n$ w.p.a.1 is a local maximizer of $f(\boldsymbol{\lambda};\boldsymbol{\theta}_n)$. □

7.4. *Proof of Theorem* 1. Let $\mathcal{G}_0=\operatorname{supp}\{\widehat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)\}$. Then it holds that $\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)}f(\boldsymbol{\lambda};\boldsymbol{\theta}_0)\le\max_{\boldsymbol{\eta}\in\widehat{\Lambda}_n^{\dagger}(\boldsymbol{\theta}_0)}n^{-1}\sum_{i=1}^n\log\{1+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}_0}(\boldsymbol{\theta}_0)\}$, where $\widehat{\Lambda}_n^{\dagger}(\boldsymbol{\theta}_0)=\{\boldsymbol{\eta}\in\mathbb{R}^{|\mathcal{G}_0|}:\boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}_0}(\boldsymbol{\theta}_0)\in\mathcal{V},i=1,\ldots,n\}$ for some open interval $\mathcal{V}$ containing zero. Given $\mathcal{G}_0$, since $|\mathcal{G}_0|\le\ell_n$, following the proof of Proposition 1, $\max_{\boldsymbol{\eta}\in\widehat{\Lambda}_n^{\dagger}(\boldsymbol{\theta}_0)}n^{-1}\sum_{i=1}^n\log\{1+\boldsymbol{\eta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{G}_0}(\boldsymbol{\theta}_0)\}=O_p(\ell_n n^{-1})$ which implies $\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)}f(\boldsymbol{\lambda};\boldsymbol{\theta}_0)=O_p(\ell_n n^{-1})$. Recall $a_n=\sum_{k=1}^p P_{1,\pi}(|\theta_k^0|)$, $b_n=\max\{\ell_n n^{-1},a_n,\nu^2\}$ and $S_n(\boldsymbol{\theta})=\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta})}f(\boldsymbol{\lambda};\boldsymbol{\theta})+\sum_{k=1}^p P_{1,\pi}(|\theta_k|)$. Define $\boldsymbol{\Theta}_*=\{\boldsymbol{\theta}=(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}},\boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}:|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_\infty\le\varepsilon,|\boldsymbol{\theta}_{\mathcal{S}^c}|_1\le\aleph_n\}$ for some fixed $\varepsilon>0$ and $\aleph_n=\min\{s\omega_n^{1/2}b_n^{1/(2\beta)}\xi_n^{-1/2},o(b_n^{1/2}),o(\nu\varrho_n^{-1/2}\ell_n^{-3/2}\xi_n^{-1/2})\}$. Let $\widehat{\boldsymbol{\theta}}_n=\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_*}S_n(\boldsymbol{\theta})$. As we have shown above, $\mathbb{P}\{S_n(\boldsymbol{\theta}_0)\le a_n+O_p(\ell_n n^{-1})\}\to 1$ as $n\to\infty$. As $S_n(\widehat{\boldsymbol{\theta}}_n)\le S_n(\boldsymbol{\theta}_0)$, we have $\mathbb{P}\{S_n(\widehat{\boldsymbol{\theta}}_n)\le a_n+O_p(\ell_n n^{-1})\}\to 1$ as $n\to\infty$. We will show that $\widehat{\boldsymbol{\theta}}_n\in\operatorname{int}(\boldsymbol{\Theta}_*)$ w.p.a.1. Same as the proof of Proposition 2, our proof includes two steps: (i) to show that for any $\epsilon_n\to\infty$ satisfying $b_n\epsilon_n^{2\beta}n^{2/\gamma}=o(1)$, there exists a uniform constant $K>0$ independent of $\boldsymbol{\theta}$ such that $\mathbb{P}\{S_n(\boldsymbol{\theta})>Kb_n\epsilon_n^{2\beta}\}\to 1$ as $n\to\infty$ for any $\boldsymbol{\theta}=(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}},\boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}\in\boldsymbol{\Theta}_*$ satisfying $|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_\infty>\epsilon_n b_n^{1/(2\beta)}$, which leads to $|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_\infty=O_p\{b_n^{1/(2\beta)}\}$. (ii) to show that $|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1<\aleph_n$. The proof of (i) is the same as that stated in Section 7.2, thus we omit its proof and only show (ii) here. We need the following lemma whose proof is given in the supplementary material.

LEMMA 1. *Let* $\mathscr{F}=\{\mathcal{F}\subset\{1,\ldots,r\}:|\mathcal{F}|\le\ell_n\}$ *and* $\boldsymbol{\Theta}_n=\{\boldsymbol{\theta}=(\boldsymbol{\theta}_{\mathcal{S}}^{\mathrm{T}},\boldsymbol{\theta}_{\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}:|\boldsymbol{\theta}_{\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_\infty=O_p\{b_n^{1/(2\beta)}\},|\boldsymbol{\theta}_{\mathcal{S}^c}|_1\le\aleph_n\}$. *Under Conditions* 4 *and* 5, *then* $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_n}\sup_{\mathcal{F}\in\mathscr{F}}\|\widehat{\mathbf{V}}_{\mathcal{F}}(\boldsymbol{\theta})-\mathbf{V}_{\mathcal{F}}(\boldsymbol{\theta}_0)\|_2=O_p\{s(\ell_n\omega_n b_n^{1/\beta})^{1/2}\}+O_p\{\ell_n(n^{-1}\varrho_n\log r)^{1/2}\}$ *provided that* $\log r=o(n^{1/3})$, $s^2\ell_n\omega_n b_n^{1/\beta}=o(1)$ *and* $\ell_n^2 n^{-1}\varrho_n\log r=o(1)$.

We begin to prove (ii) now. If $|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1=\aleph_n$, we define $\widehat{\boldsymbol{\theta}}_n^*=(\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}^{\mathrm{T}},\tau\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}^{\mathrm{T}})^{\mathrm{T}}$ for some $\tau\in(0,1)$ and will show $S_n(\widehat{\boldsymbol{\theta}}_n^*)<S_n(\widehat{\boldsymbol{\theta}}_n)$ w.p.a.1. Notice that $\widehat{\boldsymbol{\theta}}_n=\arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_*}S_n(\boldsymbol{\theta})$. This will be a contradiction. Thus $|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1<\aleph_n$. Write $\widehat{\boldsymbol{\theta}}_n=(\widehat{\theta}_{n,1},\ldots,\widehat{\theta}_{n,p})^{\mathrm{T}}$. Since $\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\widehat{\boldsymbol{\theta}}_n)}f(\boldsymbol{\lambda};\widehat{\boldsymbol{\theta}}_n)\le\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\boldsymbol{\theta}_0)}f(\boldsymbol{\lambda};\boldsymbol{\theta}_0)+\sum_{k=1}^p P_{1,\pi}(|\theta_k^0|)-\sum_{k=1}^p P_{1,\pi}(|\widehat{\theta}_{n,k}|)$, by (7.3), we have $\max_{\boldsymbol{\lambda}\in\widehat{\Lambda}_n(\widehat{\boldsymbol{\theta}}_n)}f(\boldsymbol{\lambda};\widehat{\boldsymbol{\theta}}_n)=$

$O_p(\ell_n n^{-1}) + O_p\{s\chi_n b_n^{1/(2\beta)}\}$. Pick δ_n satisfying $\delta_n = o(\ell_n^{-1/2} n^{-1/\gamma})$ and $\max\{\ell_n n^{-1}, s\chi_n b_n^{1/(2\beta)}\} = o(\delta_n^2)$. Select $\boldsymbol{\lambda}^*$ such that $\boldsymbol{\lambda}^*_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}} = \delta_n[\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}]/|\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}|_2$ and $\boldsymbol{\lambda}^*_{\mathcal{M}^c_{\widehat{\boldsymbol{\theta}}_n}} = \mathbf{0}$. Write $\boldsymbol{\lambda}^* = (\lambda_1^*, \ldots, \lambda_r^*)^{\mathrm{T}}$. Then

$$\begin{aligned} o_p(\delta_n^2) = {} & \max_{\boldsymbol{\lambda} \in \widehat{\Lambda}_n(\widehat{\boldsymbol{\theta}}_n)} f(\boldsymbol{\lambda}; \widehat{\boldsymbol{\theta}}_n) \\ \geq {} & \boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}} \bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) - \frac{1}{2n}\sum_{i=1}^n \frac{\boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}} \mathbf{g}_{i,\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) \mathbf{g}_{i,\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)^{\mathrm{T}} \boldsymbol{\lambda}^*_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}}{\{1 + c\boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}} \mathbf{g}_{i,\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}^2} \\ & - \sum_{j \in \mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}} P_{2,\nu}(|\lambda_j^*|) \\ \geq {} & \boldsymbol{\lambda}^{*,\mathrm{T}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}} \{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) - \nu\rho_2'(0^+)\mathrm{sgn}(\boldsymbol{\lambda}^*_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}})\} - C\delta_n^2\{1 + o_p(1)\} \end{aligned}$$

for some $c, c_j \in (0,1)$. Notice that $\mathrm{sgn}(\boldsymbol{\lambda}^*_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}) = \mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}$. Thus $|\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}|_2 = O_p(\delta_n)$. Using the technique developed in Section 7.1, we have $|\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n) - \nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}|_2 = O_p(\ell_n^{1/2} n^{-1/2}) + O_p\{s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\}$.

By Lemma 1 and Condition 4, we know $\lambda_{\min}\{\widehat{\mathbf{V}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\} \geq C$ w.p.a.1. Thus Proposition 3 leads to $|\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)|_2 = O_p(\ell_n^{1/2} n^{-1/2}) + O_p\{s^{1/2}\chi_n^{1/2} b_n^{1/(4\beta)}\}$. Based on this property of the Lagrange multiplier $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)$, we can follow the same arguments stated in Section 7.2 to construct (ii). Specifically, write $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)$ and $\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n^*)$ as $\widehat{\boldsymbol{\lambda}} = (\widehat{\lambda}_1, \ldots, \widehat{\lambda}_r)^{\mathrm{T}}$ and $\widehat{\boldsymbol{\lambda}}^* = (\widehat{\lambda}_1^*, \ldots, \widehat{\lambda}_r^*)^{\mathrm{T}}$, respectively. In the sequel, we use $\check{\boldsymbol{\theta}}$ to denote a generic vector lying on the jointing line between $\widehat{\boldsymbol{\theta}}_n$ and $\widehat{\boldsymbol{\theta}}_n^*$ that may be different in different uses. Write $\widehat{\boldsymbol{\theta}}_n^* = (\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,p}^*)^{\mathrm{T}}$. By Taylor expansion, it holds that

$$\begin{aligned} S_n(\widehat{\boldsymbol{\theta}}_n^*) \leq S_n(\widehat{\boldsymbol{\theta}}_n) & + \underbrace{\sum_{j=1}^r P_{2,\nu}(|\widehat{\lambda}_j|) - \sum_{j=1}^r P_{2,\nu}(|\widehat{\lambda}_j^*|)}_{\mathrm{I}} \\ & + \underbrace{\frac{1}{n}\sum_{i=1}^n \frac{\widehat{\boldsymbol{\lambda}}^{*,\mathrm{T}} \nabla_{\boldsymbol{\theta}_{\mathcal{S}^c}} \mathbf{g}_i(\check{\boldsymbol{\theta}})}{1 + \widehat{\boldsymbol{\lambda}}^{*,\mathrm{T}} \mathbf{g}_i(\check{\boldsymbol{\theta}})} (\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}^* - \widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c})}_{\mathrm{II}} \\ & + \underbrace{\sum_{k=s+1}^p P_{1,\pi}(\tau|\widehat{\theta}_{n,k}|) - \sum_{k=s+1}^p P_{1,\pi}(|\widehat{\theta}_{n,k}|)}_{\mathrm{III}}. \end{aligned} \tag{7.5}$$

We will show $\mathrm{I}+\mathrm{II}+\mathrm{III}<0$ w.p.a.1 as follows.

For I, we will first specify the convergence rate of $|\widehat{\boldsymbol{\lambda}}^{*}-\widehat{\boldsymbol{\lambda}}|_1$. Define $\widehat{H}_n(\boldsymbol{\theta},\boldsymbol{\lambda})=n^{-1}\sum_{i=1}^{n}\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i(\boldsymbol{\theta})\}+\sum_{k=1}^{p}P_{1,\pi}(|\theta_k|)-\sum_{j=1}^{r}P_{2,\nu}(|\lambda_j|)$ for any $\boldsymbol{\theta}=(\theta_1,\ldots,\theta_p)^{\mathrm{T}}$ and $\boldsymbol{\lambda}=(\lambda_1,\ldots,\lambda_r)^{\mathrm{T}}$. Then $\widehat{\boldsymbol{\theta}}_n$ and its Lagrange multiplier $\widehat{\boldsymbol{\lambda}}$ satisfy the score equation $\nabla_{\boldsymbol{\lambda}}\widehat{H}_n(\widehat{\boldsymbol{\theta}}_n,\widehat{\boldsymbol{\lambda}})=\mathbf{0}$, i.e.

$$\mathbf{0}=\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}_i(\widehat{\boldsymbol{\theta}}_n)}{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol{\theta}}_n)}-\widehat{\boldsymbol{\eta}}, \tag{7.6}$$

where $\widehat{\boldsymbol{\eta}}=(\widehat{\eta}_1,\ldots,\widehat{\eta}_r)^{\mathrm{T}}$ with $\widehat{\eta}_j=\nu\rho_2'(|\widehat{\lambda}_j|;\nu)\mathrm{sgn}(\widehat{\lambda}_j)$ for $\widehat{\lambda}_j\neq 0$ and $\widehat{\eta}_j\in[-\nu\rho_2'(0^{+}),\nu\rho_2'(0^{+})]$ for $\widehat{\lambda}_j=0$. Let $\mathcal{R}_n=\mathrm{supp}\{\widehat{\boldsymbol{\lambda}}(\widehat{\boldsymbol{\theta}}_n)\}$. Restricted on $\mathcal{R}_n$, for any $\boldsymbol{\theta}\in\mathbb{R}^p$ and $\boldsymbol{\zeta}=(\zeta_1,\ldots,\zeta_{|\mathcal{R}_n|})^{\mathrm{T}}\in\mathbb{R}^{|\mathcal{R}_n|}$ with each $\zeta_j\neq 0$, define $\mathbf{m}(\boldsymbol{\zeta},\boldsymbol{\theta})=n^{-1}\sum_{i=1}^{n}\mathbf{g}_{i,\mathcal{R}_n}(\boldsymbol{\theta})\{1+\boldsymbol{\zeta}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\boldsymbol{\theta})\}^{-1}-\mathbf{w}$, where $\mathbf{w}=(w_1,\ldots,w_{|\mathcal{R}_n|})^{\mathrm{T}}$ with $w_j=\nu\rho_2'(|\zeta_j|;\nu)\mathrm{sgn}(\zeta_j)$. Then, $\widehat{\boldsymbol{\lambda}}_{\mathcal{R}_n}$ and $\widehat{\boldsymbol{\theta}}_n$ satisfy $\mathbf{m}(\widehat{\boldsymbol{\lambda}}_{\mathcal{R}_n},\widehat{\boldsymbol{\theta}}_n)=\mathbf{0}$. By the implicit theorem [Theorem 9.28 of Rudin (1976)], for all $\boldsymbol{\theta}$ in a $|\cdot|_2$-neighborhood of $\widehat{\boldsymbol{\theta}}_n$, there is a $\boldsymbol{\zeta}(\boldsymbol{\theta})$ such that $\mathbf{m}\{\boldsymbol{\zeta}(\boldsymbol{\theta}),\boldsymbol{\theta}\}=\mathbf{0}$ and $\boldsymbol{\zeta}(\boldsymbol{\theta})$ is continuously differentiable in $\boldsymbol{\theta}$. Since $\widehat{\boldsymbol{\theta}}^{*}_{n,\mathcal{S}}=\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}$, we have $|\boldsymbol{\zeta}(\widehat{\boldsymbol{\theta}}^{*}_n)-\widehat{\boldsymbol{\lambda}}_{\mathcal{R}_n}|_1=|\{\nabla_{\boldsymbol{\theta}}\boldsymbol{\zeta}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}}\}(\widehat{\boldsymbol{\theta}}^{*}_n-\widehat{\boldsymbol{\theta}}_n)|_1\le\|\nabla_{\boldsymbol{\theta}_{\mathcal{S}^c}}\boldsymbol{\zeta}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}}\|_1|\widehat{\boldsymbol{\theta}}^{*}_{n,\mathcal{S}^c}-\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1$. Notice that

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}_{\mathcal{S}^c}}\boldsymbol{\zeta}(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}}\\
&=\Bigg[\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})^{\mathrm{T}}}{\{1+\boldsymbol{\zeta}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})\}^2}+\nu\mathrm{diag}[\rho_2''\{|\zeta_1(\check{\boldsymbol{\theta}})|;\nu\},...,\rho_2''\{|\zeta_{|\mathcal{R}_n|}(\check{\boldsymbol{\theta}})|;\nu\}]\Bigg]^{-1}\\
&\qquad\times\left\{\frac{1}{n}\sum_{i=1}^{n}\frac{\nabla_{\boldsymbol{\theta}_{\mathcal{S}^c}}\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})}{1+\boldsymbol{\zeta}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})}-\frac{1}{n}\sum_{i=1}^{n}\frac{\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})\boldsymbol{\zeta}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\nabla_{\boldsymbol{\theta}_{\mathcal{S}^c}}\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})}{\{1+\boldsymbol{\zeta}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})\}^2}\right\}\\
&=:\mathbf{A}(\check{\boldsymbol{\theta}})\times\mathbf{B}(\check{\boldsymbol{\theta}}).
\end{aligned}$$

Since $\max_{1\le i\le n}|\boldsymbol{\zeta}(\check{\boldsymbol{\theta}})^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\check{\boldsymbol{\theta}})|=o_p(1)$, from Lemma 1, we know $\|\mathbf{A}(\check{\boldsymbol{\theta}})\|_1\le|\mathcal{R}_n|^{1/2}\|\mathbf{A}(\check{\boldsymbol{\theta}})\|_2=O_p(\ell_n^{1/2})$. Meanwhile, we have $|\mathbf{B}(\check{\boldsymbol{\theta}})|_\infty=O_p(\xi_n^{1/2})$ which implies $\|\mathbf{B}(\check{\boldsymbol{\theta}})\|_1=O_p(\xi_n^{1/2}\ell_n)$. Then $\|\nabla_{\boldsymbol{\theta}_{\mathcal{S}^c}}\boldsymbol{\zeta}(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\check{\boldsymbol{\theta}}}\|_1\le\|\mathbf{A}(\check{\boldsymbol{\theta}})\|_1\|\mathbf{B}(\check{\boldsymbol{\theta}})\|_1=O_p(\ell_n^{3/2}\xi_n^{1/2})$, which implies $|\boldsymbol{\zeta}(\widehat{\boldsymbol{\theta}}^{*}_n)-\widehat{\boldsymbol{\lambda}}_{\mathcal{R}_n}|_1=O_p(\ell_n^{3/2}\xi_n^{1/2})|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1$. Let $\widetilde{\boldsymbol{\lambda}}$ satisfy $\widetilde{\boldsymbol{\lambda}}_{\mathcal{R}_n}=\boldsymbol{\zeta}(\widehat{\boldsymbol{\theta}}^{*}_n)$ and $\widetilde{\boldsymbol{\lambda}}_{\mathcal{R}_n^c}=\mathbf{0}$. For any $j\in\mathcal{R}_n^c$, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\frac{g_{i,j}(\widehat{\boldsymbol{\theta}}^{*}_n)}{1+\widetilde{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol{\theta}}^{*}_n)}&=\frac{1}{n}\sum_{i=1}^{n}\frac{g_{i,j}(\widehat{\boldsymbol{\theta}}_n)}{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol{\theta}}_n)}+O_p(\varrho_n^{1/2})|\widetilde{\boldsymbol{\lambda}}-\widehat{\boldsymbol{\lambda}}|_1\\
&\qquad+O_p(\xi_n^{1/2})|\widehat{\boldsymbol{\theta}}^{*}_{n,\mathcal{S}^c}-\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}^c}|_1\\
&=\frac{1}{n}\sum_{i=1}^{n}\frac{g_{i,j}(\widehat{\boldsymbol{\theta}}_n)}{1+\widehat{\boldsymbol{\lambda}}^{\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol{\theta}}_n)}+o_p(\nu),
\end{aligned}$$

where the term $o_p(\nu)$ holds uniformly for any $j \in \mathcal{R}_n^c$. Write $\widetilde{\boldsymbol\lambda} = (\widetilde\lambda_1, \ldots, \widetilde\lambda_r)^{\mathrm{T}}$. Recall that $\mathbf{m}\{\boldsymbol\zeta(\widehat{\boldsymbol\theta}_n^*), \widehat{\boldsymbol\theta}_n^*\} = \mathbf{0}$ and (7.6) holds, then it holds w.p.a.1 that $\mathbf{0} = n^{-1}\sum_{i=1}^n \mathbf{g}_i(\widehat{\boldsymbol\theta}_n^*)\{1 + \widetilde{\boldsymbol\lambda}^{\mathrm{T}}\mathbf{g}_i(\widehat{\boldsymbol\theta}_n^*)\}^{-1} - \widehat{\boldsymbol\eta}^*$ for $\widehat{\boldsymbol\eta}^* = (\widehat\eta_1^*, \ldots, \widehat\eta_r^*)^{\mathrm{T}}$ with $\widehat\eta_j^* = \nu\rho_2'(|\widetilde\lambda_j|; \nu)\mathrm{sgn}(\widetilde\lambda_j)$ for $\widetilde\lambda_j \neq 0$ and $\widehat\eta_j^* \in [-\nu\rho_2'(0^+), \nu\rho_2'(0^+)]$ for $\widetilde\lambda_j = 0$. By the concavity of $f(\boldsymbol\lambda; \boldsymbol\theta) = n^{-1}\sum_{i=1}^n \log\{1 + \boldsymbol\lambda^{\mathrm{T}}\mathbf{g}_i(\boldsymbol\theta)\} - \sum_{j=1}^r P_{2,\nu}(|\lambda_j|)$, we know $\widehat{\boldsymbol\lambda}^* = \widetilde{\boldsymbol\lambda}$ w.p.a.1. Hence, $|\widehat{\boldsymbol\lambda}^* - \widehat{\boldsymbol\lambda}|_1 = O_p(\ell_n^{3/2}\xi_n^{1/2})|\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c}|_1$. This implies $\mathrm{I} = O_p(\ell_n^{3/2}\xi_n^{1/2}\nu)|\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c}|_1$. Let $\mathcal{J}_* = \mathrm{supp}(\widehat{\boldsymbol\lambda}^*)$. Since $\max_{1\le i\le n}|\widehat{\boldsymbol\lambda}^{*,\mathrm{T}}\mathbf{g}_i(\check{\boldsymbol\theta})| = o_p(1)$, then $|\mathrm{II}| \le |\widehat{\boldsymbol\lambda}^*|_2|[n^{-1}\sum_{i=1}^n \nabla_{\boldsymbol\theta_{\mathcal{S}^c}}\mathbf{g}_{i,\mathcal{J}_*}(\check{\boldsymbol\theta})\{1 + \widehat{\boldsymbol\lambda}^{*,\mathrm{T}}\mathbf{g}_i(\check{\boldsymbol\theta})\}^{-1}](\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c}^* - \widehat{\boldsymbol\theta}_{n,\mathcal{S}^c})|_2 \le |\widehat{\boldsymbol\lambda}^*|_2|\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c}|_1 O_p(\ell_n^{1/2}\xi_n^{1/2}) = \max\{\ell_n^{1/2}n^{-1/2}, s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\}|\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c}|_1 \cdot O_p(\ell_n^{1/2}\xi_n^{1/2})$. Notice that $\mathrm{III} = -(1-\tau)\sum_{k=s+1}^p P_{1,\pi}'\{(c_k\tau+1-c_k)|\widehat\theta_{n,k}|\}|\widehat\theta_{n,k}| \le -(1-\tau)C\pi|\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c}|_1$ for some $c_k \in (0,1)$. Since $\max\{\ell_n^{3/2}\xi_n^{1/2}\nu, \ell_n\xi_n^{1/2}n^{-1/2}, \ell_n^{1/2}\xi_n^{1/2}s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\} = o(\pi)$, then (7.5) implies $S_n(\widehat{\boldsymbol\theta}_n^*) < S_n(\widehat{\boldsymbol\theta}_n)$ w.p.a.1. Hence, we complete the proof of (ii). Together with (i), we know such defined $\widehat{\boldsymbol\theta}_n$ is a local minimizer of $S_n(\boldsymbol\theta)$. Following the same arguments stated in Section 7.2, we can prove $\mathbb{P}(\widehat{\boldsymbol\theta}_{n,\mathcal{S}^c} = \mathbf{0}) \to 1$. □

7.5. *Proof of Theorem* 2. Recall $\mathcal{R}_n = \mathrm{supp}\{\widehat{\boldsymbol\lambda}(\widehat{\boldsymbol\theta}_n)\}$. We still write $\widehat{\boldsymbol\lambda} = \widehat{\boldsymbol\lambda}(\widehat{\boldsymbol\theta}_n) = (\widehat\lambda_1, \ldots, \widehat\lambda_r)^{\mathrm{T}}$. From (7.6), we have

$$\begin{aligned}\mathbf{0} &= \frac{1}{n}\sum_{i=1}^n \mathbf{g}_{i,\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n) - \frac{1}{n}\sum_{i=1}^n \frac{\mathbf{g}_{i,\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n)\mathbf{g}_{i,\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n)^{\mathrm{T}}\widehat{\boldsymbol\lambda}_{\mathcal{R}_n}}{\{1 + c\widehat{\boldsymbol\lambda}_{\mathcal{R}_n}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n)\}^2} - \widehat{\boldsymbol\eta}_{\mathcal{R}_n}\\ &=: \bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n) - \mathbf{C}(\widehat{\boldsymbol\theta}_n)\widehat{\boldsymbol\lambda}_{\mathcal{R}_n} - \widehat{\boldsymbol\eta}_{\mathcal{R}_n}\end{aligned}$$

for some $|c| < 1$, which implies $\widehat{\boldsymbol\lambda}_{\mathcal{R}_n} = \{\mathbf{C}(\widehat{\boldsymbol\theta}_n)\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n) - \widehat{\boldsymbol\eta}_{\mathcal{R}_n}\}$. On the other hand, together with $\mathbf{0} = \nabla_{\boldsymbol\theta}\widehat{H}_n\{\boldsymbol\theta, \widehat{\boldsymbol\lambda}(\boldsymbol\theta)\}|_{\boldsymbol\theta = \widehat{\boldsymbol\theta}_n}$, it holds that

$$\begin{aligned}\mathbf{0} &= \left\{\frac{1}{n}\sum_{i=1}^n \frac{\nabla_{\boldsymbol\theta_{\mathcal{S}}}\mathbf{g}_{i,\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n)}{1 + \widehat{\boldsymbol\lambda}_{\mathcal{R}_n}^{\mathrm{T}}\mathbf{g}_{i,\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n)}\right\}^{\mathrm{T}}\{\mathbf{C}(\widehat{\boldsymbol\theta}_n)\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n) - \widehat{\boldsymbol\eta}_{\mathcal{R}_n}\} + \widehat{\boldsymbol\kappa}_{\mathcal{S}}\\ &=: \{\mathbf{D}(\widehat{\boldsymbol\theta}_n)\}^{\mathrm{T}}\{\mathbf{C}(\widehat{\boldsymbol\theta}_n)\}^{-1}\{\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n) - \widehat{\boldsymbol\eta}_{\mathcal{R}_n}\} + \widehat{\boldsymbol\kappa}_{\mathcal{S}}\end{aligned} \tag{7.7}$$

where $\widehat{\boldsymbol\kappa}_{\mathcal{S}} = \{\sum_{k=1}^p \nabla_{\boldsymbol\theta_{\mathcal{S}}}P_{1,\pi}(|\theta_k|)\}|_{\boldsymbol\theta_{\mathcal{S}} = \widehat{\boldsymbol\theta}_{n,\mathcal{S}}}$. By Condition 6, $|\widehat{\boldsymbol\kappa}_{\mathcal{S}}|_\infty = O_p(\chi_n)$. We will use (7.7) to derive the limiting distribution of $\widehat{\boldsymbol\theta}_{n,\mathcal{S}}$ where the next lemmas are needed whose proofs are given in the supplementary material.

LEMMA 2. *Assume the conditions of Theorem* 1 *hold. Then* $\|\mathbf{C}(\widehat{\boldsymbol\theta}_n) - \widehat{\mathbf{V}}_{\mathcal{R}_n}(\boldsymbol\theta_0)\|_2 = O_p(\ell_n n^{-1/2+1/\gamma}) + O_p\{\ell_n^{1/2}s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}n^{1/\gamma}\}$, *and* $|\{\mathbf{D}(\widehat{\boldsymbol\theta}_n) - \nabla_{\boldsymbol\theta_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol\theta}_n)\}\mathbf{z}|_2 = |\mathbf{z}|_2[O_p(\ell_n s^{1/2}\omega_n^{1/2}n^{-1/2}) + O_p\{\ell_n^{1/2}s\omega_n^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\}]$ *holds uniformly for any* $\mathbf{z} \in \mathbb{R}^s$.

LEMMA 3. *Assume the conditions of Theorem* 1 *and Condition* 7 *hold. For* $\mathscr{F}$ *defined in Lemma* 1, $\sup_{\mathcal{F}\in\mathscr{F}}|[\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{F}}(\widehat{\boldsymbol{\theta}}_n)-\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\mathbf{g}_{i,\mathcal{F}}(\boldsymbol{\theta}_0)\}]\mathbf{z}|_2=|\mathbf{z}|_2[O_p\{s^{3/2}\ell_n^{1/2}\varpi_n^{1/2}b_n^{1/(2\beta)}\}+O_p\{(n^{-1}s\ell_n\omega_n\log r)^{1/2}\}]$ *holds uniformly for any* $\mathbf{z}\in\mathbb{R}^s$.

LEMMA 4. *Let* $\widehat{\mathbf{J}}_{\mathcal{F}}=\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{F}}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{F}}^{-1}(\widehat{\boldsymbol{\theta}}_n)\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{F}}(\widehat{\boldsymbol{\theta}}_n)\}$ *for any* $\mathcal{F}\in\mathscr{F}$, *where* $\mathscr{F}$ *is defined in Lemma* 1. *Under conditions for Lemma* 3 *and Condition* 8, *if* $s^2\ell_n^2b_n^{1/\beta}\varrho_n^{1/2}\max\{\omega_n,s\varpi_n\}\log r=o(1)$, $n^{-1}\ell_n^2s\omega_n\varrho_n^{1/2}(\log r)^2=o(1)$ *and* $n^{-1}\ell_n^3\varrho_n^{3/2}(\log r)^2=o(1)$, *we have that for any* $u\in\mathbb{R}$ *and* $\boldsymbol{\alpha}\in\mathbb{R}^s$ *with* $|\boldsymbol{\alpha}|_2=1$, $\sup_{\mathcal{F}\in\mathscr{F}}|\mathbb{P}[n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\mathbf{J}}_{\mathcal{F}}^{-1/2}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{F}}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{F}}^{-1}(\widehat{\boldsymbol{\theta}}_n)\bar{\mathbf{g}}_{\mathcal{F}}(\boldsymbol{\theta}_0)\le u]-\Phi(u)|\to0$, *where* $\Phi(\cdot)$ *is the cumulative distribution function of the standard normal distribution.*

Recall $\widehat{\mathbf{J}}_{\mathcal{R}_n}=\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}$. For any $\boldsymbol{\alpha}\in\mathbb{R}^s$ with unit L_2-norm, let $\boldsymbol{\delta}=\widehat{\mathbf{J}}_{\mathcal{R}_n}^{-1/2}\boldsymbol{\alpha}$, then it holds that $|\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}\boldsymbol{\delta}|_2^2\le\lambda_{\max}\{\widehat{\mathbf{V}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}|\mathbf{U}(\mathbf{U}^{\mathrm{T}}\mathbf{U})^{-1/2}\boldsymbol{\alpha}|_2^2=\lambda_{\max}\{\widehat{\mathbf{V}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}$ with $\mathbf{U}=\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1/2}(\widehat{\boldsymbol{\theta}}_n)\cdot\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}$. Thus, by Lemma 1, $|\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}\boldsymbol{\delta}|_2=O_p(1)$. Since $|\boldsymbol{\delta}|_2=O_p(1)$, by Lemma 2, $|\mathbf{D}(\widehat{\boldsymbol{\theta}}_n)\boldsymbol{\delta}|_2=O_p(1)$. As shown in Section 7.4, $|\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)-\nu\rho_2'(0^+)\mathrm{sgn}\{\bar{\mathbf{g}}_{\mathcal{M}_{\widehat{\boldsymbol{\theta}}_n}}(\widehat{\boldsymbol{\theta}}_n)\}|_2=O_p(\ell_n^{1/2}n^{-1/2})+O_p\{s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\}$. From Proposition 3, $|\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)-\widehat{\boldsymbol{\eta}}_{\mathcal{R}_n}|_2=O_p(\ell_n^{1/2}n^{-1/2})+O_p\{s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\}$. From Lemmas 2 and 3, (7.7) leads to $\boldsymbol{\delta}^{\mathrm{T}}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)\{\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)-\widehat{\boldsymbol{\eta}}_{\mathcal{R}_n}\}=O_p\big(\ell_n^{1/2}\max\{\ell_nn^{-1},s\chi_nb_n^{1/(2\beta)}\}\max\{s^{1/2}\omega_n^{1/2},n^{1/\gamma}\}\big)+O_p(s^{1/2}\chi_n)$. Expanding $\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)$ around $\boldsymbol{\theta}=\boldsymbol{\theta}_0$, it holds w.p.a.1 that

$$\begin{aligned}&\boldsymbol{\delta}^{\mathrm{T}}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)[\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widetilde{\boldsymbol{\theta}})\}(\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}})-\widehat{\boldsymbol{\eta}}_{\mathcal{R}_n}]\\&\quad=-\boldsymbol{\delta}^{\mathrm{T}}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)\bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta}_0)+O_p(s^{1/2}\chi_n)\\&\qquad+O_p\big(\ell_n^{1/2}\max\{\ell_nn^{-1},s\chi_nb_n^{1/(2\beta)}\}\max\{s^{1/2}\omega_n^{1/2},n^{1/\gamma}\}\big),\end{aligned}\tag{7.8}$$

where $\widetilde{\boldsymbol{\theta}}$ is on the line joining $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$. Notice that $|\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)-\bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta}_0)|_2\le|\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)|_2+|\bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta}_0)|_2=O_p(\ell_n^{1/2}\nu)+O_p\{s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\}$. By Taylor expansion, $|\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)-\bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta}_0)|_2\ge\lambda_{\min}([\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\dot{\boldsymbol{\theta}})]^{\mathrm{T}}[\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\dot{\boldsymbol{\theta}})])|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_2$ for some $\dot{\boldsymbol{\theta}}$ lying on the line jointing $\boldsymbol{\theta}_0$ and $\widehat{\boldsymbol{\theta}}_n$. Same as Lemma 3, $\lambda_{\min}([\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\dot{\boldsymbol{\theta}})]^{\mathrm{T}}[\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\dot{\boldsymbol{\theta}})])$ is bounded away from zero w.p.a.1, which implies $|\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}|_2=O_p(\ell_n^{1/2}\nu)+O_p\{s^{1/2}\chi_n^{1/2}b_n^{1/(4\beta)}\}$. Together with Condition 7, it holds that $|\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widetilde{\boldsymbol{\theta}})-\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}(\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}})|_2=$

$O_p(\ell_n^{3/2}s\varpi_n^{1/2}\nu^2)+O_p\{\ell_n^{1/2}s^2\varpi_n^{1/2}\chi_n b_n^{1/(2\beta)}\}$. Therefore, (7.8) leads to

$$\begin{aligned}
&\boldsymbol{\delta}^{\mathrm{T}}\widehat{\mathbf{J}}_{\mathcal{R}_n}\big[\widehat{\boldsymbol{\theta}}_{n,\mathcal{S}}-\boldsymbol{\theta}_{0,\mathcal{S}}-\widehat{\mathbf{J}}_{\mathcal{R}_n}^{-1}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)\widehat{\boldsymbol{\eta}}_{\mathcal{R}_n}\big]\\
={}&-\boldsymbol{\alpha}^{\mathrm{T}}\widehat{\mathbf{J}}_{\mathcal{R}_n}^{-1/2}\{\nabla_{\boldsymbol{\theta}_{\mathcal{S}}}\bar{\mathbf{g}}_{\mathcal{R}_n}(\widehat{\boldsymbol{\theta}}_n)\}^{\mathrm{T}}\widehat{\mathbf{V}}_{\mathcal{R}_n}^{-1}(\widehat{\boldsymbol{\theta}}_n)\bar{\mathbf{g}}_{\mathcal{R}_n}(\boldsymbol{\theta}_0)\\
&+O_p(\ell_n^{3/2}s\varpi_n^{1/2}\nu^2)+O_p\{\ell_n^{1/2}s^2\varpi_n^{1/2}\chi_n b_n^{1/(2\beta)}\}+O_p(s^{1/2}\chi_n)\\
&+O_p\big(\ell_n^{1/2}\max\{\ell_n n^{-1},s\chi_n b_n^{1/(2\beta)}\}\max\{s^{1/2}\omega_n^{1/2},n^{1/\gamma}\}\big).
\end{aligned}$$

By Lemma 4, we complete the proof of Theorem 2. □

References.

Bartolucci, F. (2007). A penalized version of the empirical likelihood ratio for the population mean. *Statistics and Probability Letters*, **77**, 104–110.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Ann. Statist.*, **35**, 2313–2351.

Chang, J., Chen, S. X. and Chen, X. (2015). High dimensional generalized empirical likelihood for moment restrictions with dependent data. *J. Econometrics* **185** 283–304.

Chang, J., Tang, C. Y. and Wu, Y. (2013). Marginal empirical likelihood and sure independence feature screening. *Ann. Statist.*, **41**, 2123–2148.

Chang, J., Tang, C. Y. and Wu, Y. (2016). Local independence feature screening for nonparametric and semiparametric models by marginal empirical likelihood. *Ann. Statist.*, **44**, 515–539.

Chen, J. and Chen, Z. (2008). Extended Bayesian information criterion for model selection with large model space. *Biometrika*, **95**, 759–771.

Chen, S. X. and Cui, H. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika*, **93**, 215–220.

Chen, S. X. and Cui, H. (2007). On the second properties of empirical likelihood with moment restrictions. *J. Econometrics*, **141**, 492–516.

Chen, S. X., Peng, L. and Qin, Y. L. (2009). Effects of data dimension on empirical likelihood. *Biometrika*, **96**, 711–722.

Chen, X. (2007). Large sample sieve estimation of semi-nonparametric models. In: Heckman, J.J., Leamer., E. (Eds.), The Handbook of Econometrics, 6B. North- Holland, Amsterdam.

Chen, X. and Pouzo (2012). Sieve quasi likelihood ratio inference on semi/nonparametric conditional moment models. *Econometrica*, **80**, 277–321.

Cheng, X. and Liao, Z. (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many moments. *J. Econometrics*, **186**, 443–464.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96**, 1348–1360.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *The Annals of Applied Statistics*, **2**, 302–332.

Gautier, E. and Tsybakov, A. B. (2014). High-dimensional instrumental variables regression and confidence sets. Manuscript. arXiv: 1105.2454v4.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50**, 1029–1054.

Hjort, N. L., McKeague, I. and Van Keilegom, I. (2009). Extending the scope of empirical likelihood. *Ann. Statist.*, **37**, 1079–1111.

Lahiri, S. N. and Mukhopadhyay, S. (1986). A penalized empirical likelihood method in high dimensions. *Ann. Statist.*, **40**, 2511–2540.

Liang, K. Y. and Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, **73**, 13–22.

Leng, C. and Tang, C. Y. (2012). Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika*, **99**, 703–716.

Lv, J. and Fan, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.*, **37**, 3498–3528.

Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, **72**, 219–255.

Owen, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237–249.

Owen, A. (1990). Empirical likelihood ratio confidence regions. *Ann. Statist.*, **18**, 90–120.

Owen, A. (2001). *Empirical Likelihood.* Chapman and Hall-CRC, New York.

Petrov, V. V. (1995). *Limit Theorems of Probability Theory: Sequences of Independent Random Variables.* Oxford University Press.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Ann. Statist.*, **22**, 300–325.

Qu, A., Lindsay, B. G. and Li, B. (2000). Improving estimating equations using quadratic inference functions. *Biometrika*, **87**, 823–836.

Rudin, W. (1976). *Principles of Mathematical Analysis.* McGraw-Hill, New York.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, **6**, 461–464.

Shi, Z. (2016). Econometric estimation with high-dimensional moment equalities. *J. Econometrics*, **195**, 104–119.

Tang, C. Y. and Leng, C. (2010). Penalized high dimensional empirical likelihood. *Biometrika*, **97**, 905–920.

Tang, C. Y. and Wu, T. T. (2014). Nested coordinate descent algorithms for empirical likelihood. *Journal of Statistical Computation and Simulation*, **84**, 1917-1930.

Tsao, M. (2004). Bounds on coverage probabilities of the empirical likelihood ratio confidence regions. *Ann. Statist.*, **32**, 1215–1221.

Tsao, M. and Wu, F. (2013). Empirical likelihood on the full parameter space. *Ann. Statist.*, **41**, 2176–2196.

Tsao, M. and Wu, F. (2014). Extended empirical likelihood for estimating equations. *Biometrika*, **101**, 703–710.

Wang, H., Li, B. and Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, **71**, 671–683.

Wu, T. T. and Lange, K. (2008). Coordinate descent algorithms for lasso penalized regression. *The Annals of Applied Statistics*, **2**, 224–244.

Zhang, C. H. (2010). Nearly unbiased variable selection under minimax concave penalty. *Ann. Statist.*, **38**, 894–942.

Zhao, P. and Yu, B. (2007). On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

J. Chang
E-mail: changjinyuan@swufe.edu.cn

C. Y. Tang
E-mail: yongtang@temple.edu

T. T. Wu
E-mail: Tongtong_Wu@urmc.rochester.edu