econometrics

MDPI

Article

# On the Convergence Rate of the SCAD-Penalized Empirical Likelihood Estimator

**Tomohiro Ando [1] and Naoya Sueishi [2,*]**

[1] Melbourne Business School, University of Melbourne, 200 Leicester Street, Carlton, Victoria 3053, Australia; T.Ando@mbs.edu

[2] Graduate School of Economics, Kobe University, 2-1 Rokkodai-cho, Nada-ku, Kobe 657-8501, Japan

* Correspondence: sueishi@econ.kobe-u.ac.jp; Tel.: +81-78-803-6827

Received: 18 October 2018; Accepted: 18 March 2019; Published: 20 March 2019


**Abstract:** This paper investigates the asymptotic properties of a penalized empirical likelihood estimator for moment restriction models when the number of parameters ($p_n$) and/or the number of moment restrictions increases with the sample size. Our main result is that the SCAD-penalized empirical likelihood estimator is $\sqrt{n/p_n}$-consistent under a reasonable condition on the regularization parameter. Our consistency rate is better than the existing ones. This paper also provides sufficient conditions under which $\sqrt{n/p_n}$-consistency and an oracle property are satisfied simultaneously. As far as we know, this paper is the first to specify sufficient conditions for both $\sqrt{n/p_n}$-consistency and the oracle property of the penalized empirical likelihood estimator.

**Keywords:** diverging number of parameters; penalized empirical likelihood; sparse models

**JEL Classification:** C14; C52

## 1. Introduction

Recently, sparse regression models have received considerable attention in business, economics, genetics, and various other fields. In these models, the number of possible regressors can be potentially large; however, only a relatively small number of these regressors are relevant.

Penalization is an alternative to a classical subset selection. One of the drawbacks of subset selection is lack of stability due to its discrete nature, meaning that variables are either retained or are dropped from a model. As a result, a small perturbation in a sample may cause a drastic change in the post-selection results (Breiman 1996). Penalization addresses this issue by achieving variable selection and estimation simultaneously, through a continuous process.

Several penalization methods have been advocated for linear regression models. Examples include the bridge penalty (Frank and Friedman 1993), LASSO (Tibshirani 1996), the smoothly clipped absolute deviation (SCAD) penalty (Fan and Li 2001), and the elastic net penalty (Zou and Hastie 2005). However, penalized least squares methods are not applicable when endogeneity exists (Fan and Liao 2014). When endogeneity exists, parameters of interest are identified often by moment restrictions, using instrumental variables.

This study investigates the asymptotic properties of a penalized empirical likelihood (PEL) estimator for moment restriction models, when the number of parameters and/or the number of moment restrictions increases with the sample size. We extend the EL estimator of Qin and Lawless (1994) by employing the SCAD penalty, so that we can achieve estimation and variable selection simultaneously.

Some penalized estimators for moment restriction models have been proposed in the econometric literature. Caner (2009) and Shi (2016b) considered the GMM estimator with a LASSO-type penalty.

Caner and Zhang (2014) proposed the adaptive elastic net GMM estimator. Fan and Liao (2014) proposed the penalized focused GMM estimator. Leng and Tang (2012) and Chang et al. (2015) studied the asymptotic properties of the PEL estimator for independent and weakly dependent observations, respectively. Tang et al. (2018) considered a penalized exponential tilting estimator.

This paper shows that the SCAD-penalized EL estimator is $\sqrt{n/p_n}$-consistent, where $p_n$ is the number of parameters. Leng and Tang (2012) showed that the non-penalized EL estimator is $\sqrt{n/p_n}$-consistent under the assumption that $p_n/r_n \to c \in (0,1)$, where $r_n$ is the number of moment restrictions. Thus, essentially, they only proved $\sqrt{n/r_n}$-consistency. Chang et al. (2015) proved $\sqrt{n/p_n}$-consistency of the non-penalized EL estimator without imposing $p_n/r_n \to c \in (0,1)$, but they only obtained $\sqrt{n/r_n}$-consistency for the PEL estimator. We prove $\sqrt{n/p_n}$-consistency of the PEL estimator under a reasonable condition on the regularization parameter of the penalty function. Our result is important because it implies $\sqrt{n}$-consistency of the estimator when $p_n$ is fixed and only $r_n$ increases with the sample size. This is consistent with previous results in the EL literature such as Donald et al. (2003). In contrast, $\sqrt{n/r_n}$-consistency implies that only a slow rate of convergence can be achieved even when $p_n$ is finite and fixed.

This paper also shows that the PEL estimator satisfies the oracle property in the sense of Fan and Peng (2004) when the truth is sparse. That is, if the true parameter vector has some zero components, then they are estimated as zeros with probability approaching one, and the other nonzero components are estimated well, similar to the case when the zero components are known a priori. Although Leng and Tang (2012) and Chang et al. (2015) also discussed the oracle property of the PEL estimator, they obtained their results under high-level assumptions. As far as we know, this paper is the first to specify sufficient conditions for both $\sqrt{n/p_n}$-consistency and the oracle property of the PEL estimator.

Recently, Chang et al. (2018) proposed an alternative PEL estimator that regularizes both parameters and Lagrange multipliers. Their estimator allows the case where $r_n$ and $p_n$ increase at an exponential rate, while our PEL estimator allows a polynomial rate only. Their method is useful when the truth is actually sparse. In contrast, our estimator is valid even when the truth is not sparse because $\sqrt{n/p_n}$-consistency can be established without imposing sparsity.

There is also a large literature on instrument (moment) selection that addresses the problem of selecting/constructing optimal instruments when a large number of instruments are available (e.g., Donald and Newey 2001; Bai and Ng 2009; Kuersteiner and Okui 2010; Belloni et al. 2012; Caner and Fan 2015; Cheng and Liao 2015; Shi 2016a). In contrast to these papers, here we focus on variable selection in a structural model.

This paper is organized as follows. We first show $\sqrt{n/p_n}$-consistency of the SCAD-penalized EL estimator and compare our assumptions with those of Leng and Tang (2012) and Chang et al. (2015). Then, we obtain the asymptotic distribution. Our proofs are new in the EL literature. All the proofs are found in the Appendix A.

## 2. PEL Estimator and Asymptotic Results

Let $\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n\}$ be a random sample from an unknown distribution on $\mathbb{R}^{d_n}$. This study considers the moment restriction model

$$E\left[\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta}_0)\right] = \mathbf{0},$$

where $\boldsymbol{\theta}_0 = (\theta_{10}, \ldots, \theta_{p_n 0})' \in \Theta_n$ is a $p_n$-dimensional true parameter and

$$\boldsymbol{m}(\boldsymbol{y}, \boldsymbol{\theta}) = (m_1(\boldsymbol{y}, \boldsymbol{\theta}), \ldots, m_{r_n}(\boldsymbol{y}, \boldsymbol{\theta}))'$$

is an $r_n$-dimensional moment function. For instance, the model includes the linear instrumental variable model

$$E[\boldsymbol{z}_i(y_i - \boldsymbol{x}_i'\boldsymbol{\theta}_0)] = \mathbf{0},$$

where $\boldsymbol{z}_i$ is an $r_n \times 1$ vector of instrumental variables and $\boldsymbol{x}_i$ is a $p_n \times 1$ vector of explanatory variables. We consider the case where $r_n \geq p_n$. The subscript indicates that $d_n$, $p_n$, and $r_n$ may increase with the sample size.

The PEL estimator for $\boldsymbol{\theta}_0$ is

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta_n} \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})} \left\{ \frac{1}{n}\sum_{i=1}^{n} \log(1 - \boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta})) + \sum_{j=1}^{p_n} p_{\kappa_n}(\theta_j) \right\},$$

where $\hat{\Lambda}_n(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^{r_n} : \boldsymbol{\lambda}'\boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta}) < 1, i = 1, \ldots, n\}$ and $p_\kappa(\cdot)$ is a penalty function with a regularization parameter $\kappa$. Thus, the estimator is the same as that of Leng and Tang (2012).

For concreteness, we employ the SCAD penalty of Fan and Li (2001):

$$p_\kappa(u) = \begin{cases} \kappa|u| & |u| \leq \kappa \\ -(u^2 - 2a\kappa|u| + \kappa^2)/[2(a-1)] & \kappa < |u| \leq a\kappa \\ (a+1)\kappa^2/2 & |u| > a\kappa \end{cases}$$

for some $a > 2$. Similar asymptotic results are obtained also by using a different penalty function, such as the minimax concave penalty of Zhang (2010).

The true model may be sparse, that is, some elements of $\boldsymbol{\theta}_0$ may be zero. Let $q_n$ be the number of nonzero elements in $\boldsymbol{\theta}_0$. Without loss of generality, we can write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{10}', \boldsymbol{\theta}_{20}')' = (\boldsymbol{\theta}_{10}', \mathbf{0}')'$ with $\boldsymbol{\theta}_1 = (\theta_1, \ldots, \theta_{q_n})' \in \mathbb{R}^{q_n}$ and $\boldsymbol{\theta}_2 = (\theta_{q_n+1}, \ldots, \theta_{p_n})' \in \mathbb{R}^{p_n - q_n}$. For now, the sparsity assumption is not crucial. It is possible that $q_n = p_n$.

Let $\boldsymbol{m}_i(\boldsymbol{\theta}) = \boldsymbol{m}(\boldsymbol{y}_i, \boldsymbol{\theta})$ and $M_i(\boldsymbol{\theta}) = \partial \boldsymbol{m}_i(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$. Also, let $\boldsymbol{m}_i = \boldsymbol{m}_i(\boldsymbol{\theta}_0)$ and $M_i = M_i(\boldsymbol{\theta}_0)$. We define $Q_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = E[\log(1 - \boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))]$ and $\hat{Q}_n(\boldsymbol{\theta}, \boldsymbol{\lambda}) = n^{-1}\sum_{i=1}^{n} \log(1 - \boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))$. Moreover, we use $\boldsymbol{\lambda}(\boldsymbol{\theta})$ and $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ to denote $\arg\max_{\boldsymbol{\lambda}\in\Lambda_n(\boldsymbol{\theta})} Q_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$ and $\arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n(\boldsymbol{\theta})} \hat{Q}_n(\boldsymbol{\theta}, \boldsymbol{\lambda})$, respectively, where $\Lambda_n(\boldsymbol{\theta})$ is a subset in $\mathbb{R}^{r_n}$, such that $\mathbf{0} \in \text{int}(\Lambda_n(\boldsymbol{\theta}))$. Let $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ denote the minimum and maximum eigenvalues of a matrix $A$. Also, let $\|\cdot\|$ denote the Euclidean (Frobenius) norm.

We impose the following conditions for $\sqrt{n/p_n}$-consistency.

**Assumption 1.** *(i) The true parameter vector $\boldsymbol{\theta}_0$ is the unique minimizer of $Q_n(\boldsymbol{\theta}, \boldsymbol{\lambda}(\boldsymbol{\theta}))$ and belongs to the interior of $\Theta_n$; (ii) There are positive functions $\Delta_1(r, p)$ and $\Delta_2(\epsilon)$ such that for any $\epsilon > 0$*

$$\inf_{\{\boldsymbol{\theta}\in\Theta_n : \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| > \epsilon\}} Q_n(\boldsymbol{\theta}, \boldsymbol{\lambda}(\boldsymbol{\theta})) \geq \Delta_1(r_n, p_n)\Delta_2(\epsilon) > 0,$$

*where* $\liminf_{n\to\infty} \Delta_1(r_n, p_n) > 0$*; (iii)* $\sup_{\boldsymbol{\theta}\in\Theta_n} \left|\hat{Q}_n(\boldsymbol{\theta}, \boldsymbol{\lambda}(\boldsymbol{\theta})) - Q_n(\boldsymbol{\theta}, \boldsymbol{\lambda}(\boldsymbol{\theta}))\right| = o_p(\Delta_1(r_n, p_n))$.

**Assumption 2.** *(i)* $E[\sup_{\boldsymbol{\theta}\in\Theta_n}(\|\boldsymbol{m}_i(\boldsymbol{\theta})\| r_n^{-1/2})^\alpha] < \infty$ *for some* $\alpha > 4$*; (ii)* $\lim_{n\to\infty} r_n^4/n = 0$.

**Assumption 3.** *(i) There exists $C$ such that* $0 < 1/C \leq \lambda_{\min}(E[\boldsymbol{m}_i(\boldsymbol{\theta})\boldsymbol{m}_i(\boldsymbol{\theta})']) \leq \lambda_{\max}(E[\boldsymbol{m}_i(\boldsymbol{\theta})\boldsymbol{m}_i(\boldsymbol{\theta})']) < C < \infty$ *in a neighborhood of $\boldsymbol{\theta}_0$; (ii) There exists $C$ such that* $\lambda_{\max}(E[M_i]'E[M_i]) < C < \infty$*; (iii) There exists $C$ such that* $\lambda_{\max}(E[M_i(\boldsymbol{\theta})M_i(\boldsymbol{\theta})']) < C < \infty$ *in a neighborhood of $\boldsymbol{\theta}_0$.*

**Assumption 4.** *(i) The moment function $m(\boldsymbol{y}, \boldsymbol{\theta})$ is twice continuously differentiable in $\boldsymbol{\theta}$ for all $\boldsymbol{y}$ in a neighborhood of $\boldsymbol{\theta}_0$; (ii) There exists $C$ such that* $\lambda_{\min}\left(\frac{d^2\hat{Q}_n(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}))}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}\right) \geq C > 0$ *in a neighborhood of $\boldsymbol{\theta}_0$ with probability approaching one.*

**Assumption 5.** $\lim_{n\to\infty} \sqrt{q_n}\kappa_n / \min_{1\leq j\leq q_n} |\theta_{j0}| = 0$.

Assumption 1 is similar to condition 2.1 of Chang et al. (2015). Assumption 1 (iii) is an extension of the uniform convergence. If we restrict the parameter space such that $\Theta_n$ is compact and $E[\sup_{\boldsymbol{\theta}\in\Theta_n} \log(1-\boldsymbol{\lambda}(\boldsymbol{\theta})'\boldsymbol{m}_i(\boldsymbol{\theta}))] < \infty$, then Assumption 1 (iii) is satisfied with $\Delta_1(r,p) = 1$. Assumption 1 is used to show that $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$. Any condition that guarantees consistency of the estimator can replace 1.

Assumptions 2 (i) and (ii) are similar to Assumptions 2 and 4 in Leng and Tang (2012). However, we do not assume that $p_n/r_n \to c \in (0,1)$. Thus, $r_n$ can grow faster than $p_n$. We can allow the case where $p_n$ is fixed and only $r_n$ increases with the sample size.

Assumption 4 states that the objective function of the EL estimator is strictly convex in $\boldsymbol{\theta}$ in a neighborhood of $\boldsymbol{\theta}_0$. When $r_n$ and $p_n$ are fixed, this condition is satisfied under fairly weak conditions. We can also relax the condition so that $\lambda_{\min}\left(\frac{d^2\hat{Q}_n(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}))}{d\boldsymbol{\theta}d\boldsymbol{\theta}'}\right) \geq \rho_n$ with a positive sequence $\rho_n$ such that $\rho_n \to 0$. In that case, we obtain a different convergence rate of the estimator. Under certain conditions, we have $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n}/\rho_n)$.

Assumption 5 is similar to condition (B2) in Huang and Xie (2007), who obtained the convergence rate of the SCAD-penalized least squares estimator. Assumption 5 states that the minimum of nonzero elements in $\boldsymbol{\theta}_0$ may converge to 0, but the convergence rate must be sufficiently slow. If nonzero elements are too small compared to $\kappa_n$, then the PEL estimator cannot distinguish between zero and nonzero elements. Following Huang and Xie (2007), we prove $\sqrt{n/p_n}$-consistency of the PEL estimator in two steps. We first prove $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n} + \sqrt{q_n}\kappa_n)$ under Assumptions 1–4 and $q_n\kappa_n^2 \to 0$ (see Lemma A3 in the Appendix A). Then, we improve the convergence rate by using Assumption 5. Notice that if we assume $\sqrt{q_n}\kappa_n = O(\sqrt{p_n/n})$, then $\sqrt{n/p_n}$-consistency of the PEL estimator is obtained immediately from Lemma A3. However, as we will see later, this condition contradicts Assumption 6 (i), which is a key condition for the oracle property. Assumption 5 is imposed so that $\sqrt{n/p_n}$-consistency and the oracle property are satisfied simultaneously.

**Theorem 1.** *Suppose that Assumptions 1–5 hold. Then, we have* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n})$.

The sparsity assumption is not necessary for this theorem. The same result is obtained even if all elements in $\boldsymbol{\theta}_0$ are nonzero. Moreover, because Assumption 5 does not exclude $\kappa_n = 0$, the theorem also applies to the non-penalized EL estimator, whose $\sqrt{n/p_n}$-consistency has been established by Chang et al. (2015). As we will see in the next theorem, if the truth is sparse, then we obtain $\sqrt{n/q_n}$-consistency of the PEL estimator under certain additional assumptions.

Our convergence rate of the PEL estimator is better than that of Chang et al. (2015). Roughly speaking, different convergence rates are based on different equalities. The asymptotic analyses of Leng and Tang (2012) and Chang et al. (2015) are based on the moment equality $E[\boldsymbol{m}_i] = \mathbf{0}$, which implies $\|n^{-1}\sum_{i=1}^n \boldsymbol{m}_i\| = O_p(\sqrt{r_n/n})$. Leng and Tang (2012) obtained $\sqrt{n/p_n}$-consistency of the non-penalized EL estimator by assuming $r_n = O(p_n)$ and hence $\|n^{-1}\sum_{i=1}^n \boldsymbol{m}_i\| = O_p(\sqrt{p_n/n})$. On the other hand, our asymptotic analysis is based on the first-order condition $E\left[\frac{d\log(1-\boldsymbol{\lambda}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}}\right] = \mathbf{0}$, which implies $\left\|\frac{d\hat{Q}_n(\boldsymbol{\theta}_0,\hat{\boldsymbol{\lambda}}_n(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}}\right\| = O_p(\sqrt{p_n/n})$. Therefore, our proof is not a straightforward extension of that of Leng and Tang (2012) and Chang et al. (2015).

To obtain a convergence rate in line with the proof of Leng and Tang (2012) and Chang et al. (2015), we need a rather strong condition on the regularization parameter. For instance, Chang et al. (2015) assumed that $q_n\kappa_n r_n^{-1} n M^{-1} = O(1)$ to prove $\sqrt{n/r_n}$-consistency, where $M$ is the block length, which is equal to unity when the observations are independent. The condition of Chang et al. (2015) corresponds to the condition that $\sqrt{q_n}\kappa_n = o(\sqrt{p_n/n})$ in our case. As stated before, although this condition simplifies the proof of $\sqrt{n/p_n}$-consistency, it causes a problem for the oracle property of the estimator.

Next, we show sparsity and asymptotic normality of the PEL estimator. Let $\hat{\boldsymbol{\theta}}_{1n}$ and $\hat{\boldsymbol{\theta}}_{2n}$ be the corresponding estimators of $\boldsymbol{\theta}_{10}$ and $\boldsymbol{\theta}_{20}$, respectively. Furthermore, let $M_{1i} = \partial \boldsymbol{m}_i(\boldsymbol{\theta}_{10}, \mathbf{0})/\partial \boldsymbol{\theta}_1'$. We define $V_n = (E[M_i]' E[\boldsymbol{m}_i \boldsymbol{m}_i']^{-1} E[M_i])^{-1}$ and $V_{1n} = (E[M_{1i}]' E[\boldsymbol{m}_i \boldsymbol{m}_i']^{-1} E[M_{1i}])^{-1}$.

We impose additional conditions.

**Assumption 6.** *(i)* $\lim_{n\to\infty} \sqrt{n/p_n}\kappa_n = \infty$*; (ii)* $\lim_{n\to\infty} r_n p_n^{3/2}/\sqrt{n} = 0$

**Assumption 7.** *There exists* $B_{jkl}(\boldsymbol{y})$ *such that* $|\partial^2 m_l(\boldsymbol{y}, \boldsymbol{\theta})/\partial\theta_j \partial\theta_k| \leq B_{jkl}(\boldsymbol{y})$ *and* $E[B_{jkl}^2(\boldsymbol{y}_i)] < \infty$ *for all* $j, k = 1, \ldots, p_n$ *and* $l = 1, \ldots, r_n$ *in a neighborhood of* $\boldsymbol{\theta}_0$.

**Assumption 8.** *There exists* $C$ *such that* $0 < 1/C \leq \lambda_{\min}(V_n) \leq \lambda_{\max}(V_n) \leq C < \infty$.

Assumption 6 (i) is a key condition for sparsity of the PEL estimator. It requires that the regularization parameter is not too small so that zero elements in $\boldsymbol{\theta}_0$ are estimated as zero. The same condition is also employed by Leng and Tang (2012).

**Theorem 2.** *Suppose that Assumptions 1–8 hold. Let* $B_n$ *be an* $l \times q_n$ *matrix such that* $B_n B_n' \to G$*, where* $G$ *is an* $l \times l$ *matrix with fixed* $l$*. Then, the PEL estimator satisfies the following:*

1. *Sparsity:* $\hat{\boldsymbol{\theta}}_{2n} = \mathbf{0}$ *with probability approaching one.*
2. $\sqrt{n/q_n}$*-consistency:* $\|\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}\| = O_p(\sqrt{q_n/n})$.
3. *Asymptotic normality:* $\sqrt{n} B_n V_{1n}^{-1/2} (\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) \xrightarrow{d} N(0, G)$.

The selection of the matrix $B_n$ depends on the parameter of interest. For instance, suppose that the parameter of interest is the first element of $\boldsymbol{\theta}_{10}$. Let $\hat{\theta}_{1n,1}$ and $\theta_{10,1}$ denote first elements of $\hat{\boldsymbol{\theta}}_{1n}$ and $\boldsymbol{\theta}_{10}$, respectively. Then, we choose $B_n = (1, 0, \ldots, 0)$ and obtain $\sqrt{n}(\hat{\theta}_{1n,1} - \theta_{10,1}) \xrightarrow{d} N(0, v_{11})$, where $v_{11}$ is the limit of the first diagonal element of $V_{1n}$.

Although a detailed proof is given in the Appendix A, we give a sketch of the proof for asymptotic normality here. If $\boldsymbol{\lambda}(\boldsymbol{\theta})$ were known, then $\boldsymbol{\theta}_0$ can be estimated by

$$\tilde{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta} \in \Theta_n} \left\{ \frac{1}{n} \sum_{i=1}^{n} \log(1 - \boldsymbol{\lambda}(\boldsymbol{\theta})' \boldsymbol{m}_i(\boldsymbol{\theta})) + \sum_{j=1}^{p_n} p_{\kappa_n}(\theta_j) \right\},$$

which is a penalized maximum likelihood estimator using a least favorable submodel of the moment restriction model (see Sueishi 2016, for instance). Because $\tilde{\boldsymbol{\theta}}_n$ is the penalized maximum likelihood estimator, its distribution can be obtained in a manner similar to Fan and Peng (2004). We derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}_n$ by showing that $\hat{\boldsymbol{\theta}}_n$ is asymptotically equivalent to $\tilde{\boldsymbol{\theta}}_n$.

By modifying the proof of Theorem 2, we can obtain easily the asymptotic distribution of the non-penalized EL estimator. Because the asymptotic distribution of the non-penalized EL estimator has already been derived by Leng and Tang (2012), we omit the derivation. We see that the efficiency of the PEL estimator for $\boldsymbol{\theta}_{10}$ is the same as that of the non-penalized EL estimator for which it is known a priori that $\boldsymbol{\theta}_{20} = \mathbf{0}$. Thus, our estimator satisfies the oracle property in the sense of Fan and Peng (2004).

Theorem 2 is similar to Theorem 3 of Leng and Tang (2012). However, they proved sparsity by assuming that the PEL estimator is $\sqrt{n/p_n}$-consistent. They did not state explicitly the conditions under which the non-penalized and penalized EL estimators have the same convergence rate.

Chang et al. (2015) showed a similar result to Theorem 2 for weakly dependent observations. They obtained $\sqrt{n/r_n}$-consistency and sparsity under two separate $\kappa_n$ rate conditions. Specifically, they assume: (i) $q_n \kappa_n r_n^{-1} n M^{-1} = O(1)$ for $\sqrt{n/r_n}$-consistency and (ii) $\kappa_n \sqrt{n/r_n} M^{-1} \to \infty$ for sparsity. If condition (ii) is satisfied, however, condition (i) requires that $q_n \sqrt{n/r_n} \to 0$, which is clearly impossible. This causes a trouble because their proof of sparsity requires $\sqrt{n/r_n}$-consistency of the estimator. We relaxed condition (i) and obtained sufficient conditions under which both $\sqrt{n/p_n}$-consistency and sparsity are satisfied.

## 3. Conclusions

We investigated the asymptotic properties of the PEL estimator when the number of parameters and/or the number of moment restrictions increases with the sample size. In particular, we showed that the PEL estimator is $\sqrt{n/p_n}$-consistent under a reasonable condition on the regularization parameter. Although we cannot compare our results directly to those of Chang et al. (2015) because they allow weakly dependent observations, our convergence rate is improved over the existing ones. In terms of converge rate, our result is even better than Tang et al. (2018) and Chang et al. (2018), because their convergence rates depend also on the number of moment restrictions.

A crucial issue with the PEL estimation concerns selecting the size of the regularization parameter. The asymptotic theory does not tell us how to select the regularization parameter in practice. Although some selection methods are considered by Leng and Tang (2012), Shi (2016b), and Ando and Sueishi (2019), this is still an underdeveloped area of research.

**Author Contributions:** Both authors contributed equally to this work.

**Funding:** This research was supported by JSPS KAKENHI Grant Number 15K03396.

**Acknowledgments:** The authors would like to thank anonymous reviewers for their comments.

**Conflicts of Interest:** The authors declare no conflict of interest.

## Appendix A

Throughout the Appendix, $C$ denotes a generic positive constant which may vary according to context. The qualifier "with probability approaching one" is abbreviated as w.p.a.1. We define

$$H_{11}(\boldsymbol{\theta},\boldsymbol{\lambda}) = E\left[\frac{\partial^2 \log(1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] = -E\left[\frac{\frac{\partial}{\partial\boldsymbol{\theta}'}(M_i(\boldsymbol{\theta})'\boldsymbol{\lambda})}{1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta})}\right] - E\left[\frac{M_i(\boldsymbol{\theta})'\boldsymbol{\lambda}\boldsymbol{\lambda}'M_i(\boldsymbol{\theta})}{(1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))^2}\right]$$
$$H_{12}(\boldsymbol{\theta},\boldsymbol{\lambda}) = E\left[\frac{\partial^2 \log(1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\lambda}'}\right] = -E\left[\frac{M_i(\boldsymbol{\theta})'}{1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta})}\right] - E\left[\frac{M_i(\boldsymbol{\theta})'\boldsymbol{\lambda}\boldsymbol{m}_i(\boldsymbol{\theta})'}{(1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))^2}\right]$$
$$H_{22}(\boldsymbol{\theta},\boldsymbol{\lambda}) = E\left[\frac{\partial^2 \log(1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))}{\partial\boldsymbol{\lambda}\partial\boldsymbol{\lambda}'}\right] = -E\left[\frac{\boldsymbol{m}_i(\boldsymbol{\theta})\boldsymbol{m}_i(\boldsymbol{\theta})'}{(1-\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta}))^2}\right].$$

We use $\hat{H}_{ij}(\boldsymbol{\theta},\boldsymbol{\lambda})$ to denote the sample analog of $H_{ij}(\boldsymbol{\theta},\boldsymbol{\lambda})$. Moreover, we define $\hat{Q}_n(\boldsymbol{\theta}) = \hat{Q}_n(\boldsymbol{\theta},\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}))$ and $Q_n(\boldsymbol{\theta}) = Q_n(\boldsymbol{\theta},\boldsymbol{\lambda}(\boldsymbol{\theta}))$.

We prepare some lemmas to prove Theorems 1 and 2.

**Lemma A1.** *Suppose that Assumptions 1, 2 and 3 (i) hold. Then, we have* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$ *if* $q_n\kappa_n^2 \to 0$.

**Proof of Lemma A1.** Let $\xi$ satisfy $1/\alpha + 1/8 \le \xi < 3/8$ and let $\bar{\Lambda}_n = \{\boldsymbol{\lambda} \in \mathbb{R}^{r_n} : \|\boldsymbol{\lambda}\| \le n^{-\xi}\}$. Then, by Assumption 2, we have

$$\max_{1\le i\le n}\sup_{\boldsymbol{\theta}\in\Theta_n}|\boldsymbol{\lambda}'\boldsymbol{m}_i(\boldsymbol{\theta})| \le n^{-\xi}\max_{1\le i\le n}\sup_{\boldsymbol{\theta}\in\Theta_n}\|\boldsymbol{m}_i(\boldsymbol{\theta})\| = o_p(n^{-\xi+1/\alpha}r_n^{1/2}) = o_p(1)$$

for all $\boldsymbol{\lambda} \in \bar{\Lambda}_n$. Let $\tilde{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}\in\bar{\Lambda}_n}\hat{Q}_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda})$. Because Assumptions 2 (ii) and 3 (i) imply $\lambda_{\min}(n^{-1}\sum_{i=1}^n \boldsymbol{m}_i\boldsymbol{m}_i') > C$ w.p.a.1, by expanding $\log(1-x)$ around $x=0$, we have

$$0 \le \hat{Q}_n(\boldsymbol{\theta}_0,\tilde{\boldsymbol{\lambda}}) \le -\tilde{\boldsymbol{\lambda}}'\bar{\boldsymbol{m}}_n - \frac{1}{2}\tilde{\boldsymbol{\lambda}}'\left\{\frac{1}{n}\sum_{i=1}^n\frac{\boldsymbol{m}_i\boldsymbol{m}_i'}{(1-\dot{\boldsymbol{\lambda}}'\boldsymbol{m}_i)^2}\right\}\tilde{\boldsymbol{\lambda}} \le \|\tilde{\boldsymbol{\lambda}}\|\|\bar{\boldsymbol{m}}_n\| - C\|\tilde{\boldsymbol{\lambda}}\|^2, \tag{A1}$$

where $\bar{\boldsymbol{m}}_n = n^{-1}\sum_{i=1}^n \boldsymbol{m}_i$ and $\dot{\boldsymbol{\lambda}}$ lies between $\mathbf{0}$ and $\tilde{\boldsymbol{\lambda}}$. Therefore, we obtain $\|\tilde{\boldsymbol{\lambda}}\| = O_p(\|\bar{\boldsymbol{m}}_n\|) = O_p(\sqrt{r_n/n}) = o_p(n^{-3/8})$ by Assumption 2 (ii), and hence $\tilde{\boldsymbol{\lambda}} \in \mathrm{int}(\bar{\Lambda}_n)$. Because $\bar{\Lambda}_n \subset \hat{\Lambda}_n(\boldsymbol{\theta}_0)$, the concavity of $\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda})$ implies $\tilde{\boldsymbol{\lambda}} = \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)$. Moreover, we obtain

$$\hat{Q}_n(\hat{\boldsymbol{\theta}}_n, \boldsymbol{\lambda}(\hat{\boldsymbol{\theta}}_n)) \leq \hat{Q}_n(\hat{\boldsymbol{\theta}}_n) \leq \hat{Q}_n(\boldsymbol{\theta}_0) + \sum_{j=1}^{p_n} p_{\kappa_n}(\theta_{j0}) = o_p(1). \tag{A2}$$

Now, suppose that $\hat{\boldsymbol{\theta}}_n$ is not consistent. Then, there exists a subsequence $\{n_k\}$ such that $\|\hat{\boldsymbol{\theta}}_{n_k} - \boldsymbol{\theta}_0\| > \epsilon$ for some $\epsilon > 0$ almost surely. By Assumption 1 (iii) and Equation (A2), we have $\|Q_{n_k}(\hat{\boldsymbol{\theta}}_{n_k})\| = o_p(\Delta_1(r_{n_k}, p_{n_k})) + o_p(1)$. In contrast, Assumption 1 (ii) implies $\|Q_{n_k}(\hat{\boldsymbol{\theta}}_{n_k})\| > \Delta_1(r_{n_k}, p_{n_k})\Delta_2(\epsilon)$. Because $\liminf_{n\to\infty} \Delta(r_n, p_n) > 0$, it is a contradiction. Therefore, we have $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = o_p(1)$. □

**Lemma A2.** *Suppose that Assumptions 1–3 hold. Then, we have*

$$\left\| \frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}} - \frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0)))}{d\boldsymbol{\theta}} \right\| = o_p\left(\frac{1}{\sqrt{n}}\right).$$

**Proof of Lemma A2.** Let $H_{ij}(\boldsymbol{\theta}) = H_{ij}(\boldsymbol{\theta}, \boldsymbol{\lambda}(\boldsymbol{\theta}))$ and $\hat{H}_{ij}(\boldsymbol{\theta}) = \hat{H}_{ij}(\boldsymbol{\theta}, \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}))$ for $i, j = 1, 2$. Also, let $H_{ij} = H_{ij}(\boldsymbol{\theta}_0)$ and $\hat{H}_{ij} = \hat{H}_{ij}(\boldsymbol{\theta}_0)$. Because $\boldsymbol{\lambda}(\boldsymbol{\theta}_0) = \mathbf{0}$, we have

$$\|\hat{H}_{12} - H_{12}\| \leq \left\| \frac{1}{n}\sum_{i=1}^n \frac{M_i' \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)\boldsymbol{m}_i'}{(1 - \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i)^2} \right\| + \left\| \frac{1}{n}\sum_{i=1}^n \frac{M_i}{1 - \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i} - E[M_i] \right\|.$$

From the proof of Lemma A1, we see that $\|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)\| = O_p(\sqrt{r_n/n})$. In addition, it follows from Assumptions 2 (ii) and 3 (iii) that $\lambda_{\max}(n^{-1}\sum_{i=1}^n M_i M_i') < C$ w.p.a.1. Because $n^{-1}\sum_{i=1}^n \|\boldsymbol{m}_i\|^2 = O_p(r_n)$ by Assumption 2 (i), we have

$$\begin{aligned} \left\| \frac{1}{n}\sum_{i=1}^n \frac{M_i' \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)\boldsymbol{m}_i'}{(1 - \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i)^2} \right\| &\leq C\sqrt{\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)' \left(\frac{1}{n}\sum_{i=1}^n M_i M_i'\right) \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}\sqrt{\frac{1}{n}\sum_{i=1}^n \|\boldsymbol{m}_i\|^2} \\ &= O_p\left(\frac{r_n}{\sqrt{n}}\right). \end{aligned}$$

Furthermore, because, $|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i| = o_p(1)$ for all $i$, we have $(1 - \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i)^{-1} = 1 + \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i + o_p(|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i|)$. Hence, we have

$$\begin{aligned} &\left\| \frac{1}{n}\sum_{i=1}^n \frac{M_i}{1 - \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i} - E[M_i] \right\| \\ &\quad \leq \left\| \frac{1}{n}\sum_{i=1}^n M_i - E[M_i] \right\| + C\left\| \frac{1}{n}\sum_{i=1}^n \hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i M_i \right\| = O_p\left(\frac{r_n}{\sqrt{n}}\right), \end{aligned}$$

which implies $\|\hat{H}_{12} - H_{12}\| = O_p(r_n/\sqrt{n})$. Similarly, we have

$$\begin{aligned}\left\|\hat{H}_{22}-H_{22}\right\| &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}_i\boldsymbol{m}_i'-E[\boldsymbol{m}_i\boldsymbol{m}_i']\right\|+C\left\|\frac{1}{n}\sum_{i=1}^{n}(\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\boldsymbol{m}_i)\boldsymbol{m}_i\boldsymbol{m}_i'\right\| \\ &\leq \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}_i\boldsymbol{m}_i'-E[\boldsymbol{m}_i\boldsymbol{m}_i']\right\| \\ &\quad +C\sqrt{\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)'\left(\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}_i\boldsymbol{m}_i'\right)\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\|\boldsymbol{m}_i\|^4} \\ &= O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right). \end{aligned} \tag{A3}$$

By the Taylor expansion,

$$\begin{aligned}&\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}-\frac{d\hat{Q}_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}} \\ &\quad=\frac{d}{d\boldsymbol{\theta}}\frac{\partial\hat{Q}_n(\boldsymbol{\theta},\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}))}{\partial\boldsymbol{\lambda}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{n0}}\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)+\left(\frac{\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}-\frac{\partial\boldsymbol{\lambda}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right)'\frac{\partial\hat{Q}_n(\boldsymbol{\theta}_0,\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0))}{\partial\boldsymbol{\lambda}},\end{aligned}$$

where $\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ locates between $\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})$ and $\boldsymbol{\lambda}(\boldsymbol{\theta})$. By applying the implicit function theorem to the first-order conditions, we obtain

$$\frac{\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}=-\hat{H}_{22}^{-1}\hat{H}_{21}\quad\text{and}\quad\frac{\partial\boldsymbol{\lambda}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}=-H_{22}^{-1}H_{21}.$$

Here we have $1/C\leq\lambda_{\min}(\hat{H}_{22})\leq\lambda_{\max}(\hat{H}_{22})<C$ by Assumptions 2 (ii) and 3 (i) and Equation (A3) w.p.a.1. Thus, by Assumption 3 (ii), we have

$$\left\|\frac{\partial\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}-\frac{\partial\boldsymbol{\lambda}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right\|\leq\left\|\hat{H}_{22}^{-1}(\hat{H}_{21}-H_{21})\right\|+\left\|(\hat{H}_{22}^{-1}-H_{22}^{-1})H_{21}\right\|=O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right).$$

Moreover, some calculation yields

$$\begin{aligned}&\left\|\frac{d}{d\boldsymbol{\theta}}\frac{\partial\hat{Q}_n(\boldsymbol{\theta},\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}))}{\partial\boldsymbol{\lambda}'}\bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}_{n0}}\right\| \\ &\quad=\left\|\hat{H}_{12}(\boldsymbol{\theta}_0,\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0))+\left(\frac{\partial\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right)'\hat{H}_{22}(\boldsymbol{\theta}_0,\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0))\right\| \\ &\quad\leq\left\|\hat{H}_{12}(\boldsymbol{\theta}_0,\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0))-H_{12}\right\|+\left\|\left(\frac{\partial\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}-\frac{\partial\boldsymbol{\lambda}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}'}\right)'\hat{H}_{22}(\boldsymbol{\theta}_0,\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0))\right\| \\ &\qquad+\left\|H_{12}H_{22}^{-1}\left(\hat{H}_{22}(\boldsymbol{\theta}_0,\dot{\boldsymbol{\lambda}}(\boldsymbol{\theta}_0))-H_{22}\right)\right\| \\ &\quad=O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right).\end{aligned}$$

Combining these results, we obtain

$$\left\|\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}-\frac{d\hat{Q}_n(\boldsymbol{\theta}_0,\boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}}\right\|=O_p\left(\frac{r_n^2}{n}\right),$$

which implies the desired result by Assumption 2 (ii). □

**Lemma A3.** *Suppose that Assumptions 1–4 hold. Then, we have* $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n} + \sqrt{q_n}\kappa_n)$.

**Proof of Lemma A3.** We denote $\nabla^2\hat{Q}_n(\boldsymbol{\theta}) = d^2\hat{Q}_n(\boldsymbol{\theta})/d\boldsymbol{\theta}d\boldsymbol{\theta}'$. By Assumption 4, $\nabla^2\hat{Q}_n(\boldsymbol{\theta})$ is positive definite in a neighborhood of $\boldsymbol{\theta}_0$ w.p.a.1. By the definition of the PEL estimator, we have

$$\hat{Q}_n(\boldsymbol{\theta}_0) + \sum_{j=1}^{p_n} p_{\kappa_n}(\theta_{j0}) \geq \hat{Q}_n(\hat{\boldsymbol{\theta}}_n). \tag{A4}$$

Because $p_{\kappa_n}(\theta_{j0}) \leq (a+1)\kappa^2/2$ for $j = 1, \ldots, q_n$ and $p_{\kappa_n}(\theta_{j0}) = 0$ for $j = q_n+1, \ldots, p_n$, expanding Equation (A4) yields

$$\begin{aligned} 0 &\geq 2\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)'\nabla^2\hat{Q}_n(\dot{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) - (a+1)q_n\kappa_n^2 \\ &= \left\|\nabla^2\hat{Q}_n^{1/2}(\dot{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \nabla^2\hat{Q}_n^{-1/2}(\dot{\boldsymbol{\theta}}_n)\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right\|^2 - \frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'}\nabla^2\hat{Q}_n^{-1}(\dot{\boldsymbol{\theta}}_n)\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}} \\ &\quad -(a+1)q_n\kappa_n^2 \end{aligned}$$

for some $\dot{\boldsymbol{\theta}}_n$ located between $\hat{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_0$. Therefore, by the Loève's $C_2$-inequality, we obtain

$$\begin{aligned} &\left\|\nabla^2\hat{Q}_n^{1/2}(\dot{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right\|^2 \\ &\leq 2\left\|\nabla^2\hat{Q}_n^{1/2}(\dot{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + \nabla^2\hat{Q}_n^{-1/2}(\dot{\boldsymbol{\theta}}_n)\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right\|^2 + 2\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'}\nabla^2\hat{Q}_n^{-1}(\dot{\boldsymbol{\theta}}_n)\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}} \\ &\leq 4\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'}\nabla^2\hat{Q}_n^{-1}(\dot{\boldsymbol{\theta}}_n)\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}} + 2(a+1)q_n\kappa_n^2. \end{aligned}$$

By Lemma A2, we obtain $\left\|\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}\right\| = O_p(\sqrt{p_n/n})$, and hence

$$C\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\|^2 \leq \left\|\nabla^2\hat{Q}_n^{1/2}(\dot{\boldsymbol{\theta}}_n)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)\right\|^2 = O_p\left(\frac{p_n}{n} + q_n\kappa_n^2\right)$$

by Assumption 4 (ii). □

**Proof of Theorem 1.** If $\sqrt{q_n}\kappa_n = O(\sqrt{p_n/n})$, then we trivially have $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n})$ by Lemma A3. Thus, we only consider the case where $\sqrt{q_n}\kappa_n/\sqrt{p_n/n} \to \infty$.

By Lemma A3, we have

$$\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(u_n) \quad \text{with} \quad u_n = \sqrt{\frac{p_n}{n}} + \sqrt{q_n}\kappa_n.$$

Furthermore, for any $M$ and for any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^M u_n$, we have

$$\min_{1\leq j\leq q_n}|\theta_j| \geq \min_{1\leq j\leq q_n}|\theta_{j0}| - 2^M u_n.$$

By Assumption 5, we have $u_n/\min_{1\leq j\leq q_n}|\theta_{j0}| < 2^{-M-1}$ for sufficiently large $n$, and hence

$$\min_{1\leq j\leq q_n}|\theta_j| \geq \frac{1}{2}\min_{1\leq j\leq q_n}|\theta_{j0}|.$$

This implies that $\min_{1\leq j\leq q_n}|\theta_j| > a\kappa_n$ for sufficiently large $n$.

Let $\{h_n\}$ be a positive sequence that converges to 0 as $n \to \infty$. Following Huang and Xie (2007), we decompose $\Theta_n \setminus \{\mathbf{0}\}$ into shells $S_{n,k} = \{\boldsymbol{\theta} : 2^{k-1}h_n \leq \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| \leq 2^k h_n\}$ for $k = 1, 2, \ldots$. For $\boldsymbol{\theta} \in S_{n,k}$ such that $2^k h_n \leq 2^M u_n$, we obtain

$$\hat{Q}_n(\boldsymbol{\theta}) - \hat{Q}_n(\boldsymbol{\theta}_0) = \frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\nabla^2\hat{Q}_n(\dot{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \boldsymbol{\theta}_0)$$

and

$$\frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)'\nabla^2\hat{Q}_n(\dot{\boldsymbol{\theta}}_n)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \geq 2^{2k-3}Ch_n^2 \tag{A5}$$

w.p.a.1. Let $E_n$ be the event such that Equation (A5) is satisfied. Because Lemma A2 implies that the difference between $\frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\boldsymbol{\theta}}$ and $\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}}$ is asymptotically negligible, we have

$$\begin{aligned}
&P\left(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > 2^L h_n\right) \\
&\leq P\left(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > 2^M u_n\right) + P\left(\left\{2^L h_n < \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq 2^M u_n\right\} \cap E_n\right) \\
&= o(1) + \sum_k P\left(\{\hat{\boldsymbol{\theta}}_n \in S_{n,k}\} \cap E_n\right) \\
&\leq o(1) + \sum_k P\left(\left\{\inf_{\boldsymbol{\theta} \in S_{n,k}} \hat{Q}_n(\boldsymbol{\theta}) + \sum_{j=1}^{p_n} p_{\kappa_n}(\theta_j) \leq \hat{Q}_n(\boldsymbol{\theta}_0) + \sum_{j=1}^{p_n} p_{\kappa_n}(\theta_{j0})\right\} \cap E_n\right) \\
&\leq o(1) + \sum_k P\left(\sup_{\boldsymbol{\theta} \in S_{n,k}} -\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \geq 2^{2k-3}Ch_n^2\right),
\end{aligned}$$

where $\sum_k$ stands for $\sum_{k:k>L, 2^k h_n \leq 2^M u_n}$. Moreover, some calculation yields that

$$\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^n E[M_i]'E[\boldsymbol{m}_i\boldsymbol{m}_i]^{-1}\boldsymbol{m}_i.$$

Thus, it follows from the Markov and Cauchy-Schwarz inequalities that

$$\begin{aligned}
&\sum_k P\left(\left\{\sup_{\boldsymbol{\theta} \in S_{n,k}} -\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \geq 2^{2k-3}Ch_n^2\right\}\right) \\
&\leq C\sum_k \frac{E\left[\sup_{\boldsymbol{\theta} \in S_{n,k}} \left|\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}'}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)\right|\right]}{2^{2k-3}h_n^2} \\
&\leq C\sum_{k:k>L} \frac{2^k h_n (\mathrm{tr}\{E[M_i]'E[\boldsymbol{m}_i\boldsymbol{m}_i']^{-1}E[M_i]\}/n)^{1/2}}{2^{2k-3}h_n^2} \\
&\leq C\sum_{k:k>L} \frac{\sqrt{p_n/n}}{2^{k-3}h_n}.
\end{aligned}$$

Notice that $\sum_k$ is changed to $\sum_{k:k>L}$ in the second inequality. By choosing $h_n = \sqrt{p_n/n}$, we obtain the desired result. □

**Lemma A4.** *Suppose that Assumptions 2, 3, 4 (i) and 7 hold. Then, for any $\boldsymbol{\theta}$ such that $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n})$, we have*

$$\left\|\nabla^2\hat{Q}_n(\boldsymbol{\theta}) - \nabla^2 Q_n(\boldsymbol{\theta}_0)\right\| = O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right) + O_p\left(\frac{r_n p_n}{\sqrt{n}}\right).$$

**Proof of Lemma A4.** Let $\boldsymbol{\theta}$ satisfy $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| = O_p(\sqrt{p_n/n})$. By a simple calculation, we obtain

$$\nabla^2 \hat{Q}_n(\boldsymbol{\theta}) = \hat{H}_{11}(\boldsymbol{\theta}) - \hat{H}_{12}(\boldsymbol{\theta})\hat{H}_{22}^{-1}(\boldsymbol{\theta})\hat{H}_{21}(\boldsymbol{\theta})$$

and

$$\nabla^2 Q_n(\boldsymbol{\theta}_0) = H_{11} - H_{12}H_{22}^{-1}H_{21} = E[M_i]'E[\boldsymbol{m}_i\boldsymbol{m}_i']^{-1}E[M_i].$$

Thus, it is sufficient to show that

$$\|\hat{H}_{11}(\boldsymbol{\theta})\| + \left\|-\hat{H}_{12}(\boldsymbol{\theta})\hat{H}_{22}(\boldsymbol{\theta})^{-1}\hat{H}_{21}(\boldsymbol{\theta}) - E[M_i]'E[\boldsymbol{m}_i\boldsymbol{m}_i']^{-1}E[M_i]\right\| = O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right) + O_p\left(\frac{r_n p_n}{\sqrt{n}}\right).$$

By using a similar argument as in Equation (A1), we have $\|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\| = O_p(\sqrt{r_n/n})$. Also, the $(j,k)$ element of $\frac{\partial}{\partial\boldsymbol{\theta}'}(M_i(\boldsymbol{\theta})'\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}))$ is given by $\sum_{l=1}^{r_n}\partial^2 m_l(\boldsymbol{y}_i,\boldsymbol{\theta})/\partial\theta_j\partial\theta_k\hat{\lambda}_l(\boldsymbol{\theta})$ and

$$\left|\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{r_n}\frac{\partial^2 m_l(\boldsymbol{y}_i,\boldsymbol{\theta})}{\partial\theta_j\partial\theta_k}\hat{\lambda}_l(\boldsymbol{\theta})\right| \le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{l=1}^{r_n}B_{jkl}^2(\boldsymbol{y}_i)}\|\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\| = O_p\left(\frac{r_n}{\sqrt{n}}\right)$$

by Assumption 7. Therefore, we have

$$\begin{aligned}\|\hat{H}_{11}(\boldsymbol{\theta})\| &\le C\left\|\frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}'}(M_i(\boldsymbol{\theta})'\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta}))\right\| + C\left\|\frac{1}{n}\sum_{i=1}^{n}M_i(\boldsymbol{\theta})'\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})\hat{\boldsymbol{\lambda}}(\boldsymbol{\theta})'M_i(\boldsymbol{\theta})\right\| \\ &= O_p\left(\frac{r_n p_n}{\sqrt{n}}\right).\end{aligned}$$

Moreover, by doing similar calculations as in the proof of Lemma A2, we obtain

$$\begin{aligned}\|-\hat{H}_{12}(\boldsymbol{\theta}) - E[M_i]\| &\le \left\|\frac{1}{n}\sum_{i=1}^{n}M_i(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n}M_i\right\| + O_p\left(\frac{r_n}{\sqrt{n}}\right) \\ &\le \sqrt{\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{p_n}\left\|\frac{\partial M_i(\dot{\boldsymbol{\theta}})}{\partial\theta_j}\right\|^2}\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\| + O_p\left(\frac{r_n}{\sqrt{n}}\right) \\ &= O_p\left(\frac{r_n^{1/2}p_n^{3/2}}{\sqrt{n}}\right) + O_p\left(\frac{r_n}{\sqrt{n}}\right)\end{aligned}$$

and

$$\begin{aligned}&\|-\hat{H}_{22}(\boldsymbol{\theta}) - E[\boldsymbol{m}_i\boldsymbol{m}_i']\| \\ &\le \left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}_i(\boldsymbol{\theta})\boldsymbol{m}_i(\boldsymbol{\theta})' - \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}_i\boldsymbol{m}_i'\right\| + O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right) \\ &\le 2\left\|\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{m}_i'M_i(\dot{\boldsymbol{\theta}})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\right\| + (\boldsymbol{\theta}-\boldsymbol{\theta}_0)'\left(\frac{1}{n}\sum_{i=1}^{n}M_i(\dot{\boldsymbol{\theta}})M_i'(\dot{\boldsymbol{\theta}})\right)(\boldsymbol{\theta}-\boldsymbol{\theta}_0) + O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right) \\ &= O_p\left(\frac{r_n^{3/2}}{\sqrt{n}}\right)\end{aligned}$$

for some $\dot{\boldsymbol{\theta}}$ that is located between $\boldsymbol{\theta}$ and $\boldsymbol{\theta}_0$. Hence, we obtain the result. □

**Proof of Theorem 2.** We first prove sparsity. Theorem 1 and Assumption 6 (i) imply that $\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \leq \kappa_n$ w.p.a.1. Thus, it is sufficient to show that w.p.a.1,

$$\frac{d\hat{Q}_n(\boldsymbol{\theta}_0 + \boldsymbol{v})}{d\theta_j} + p'_{\kappa_n}(v_j) > 0 \quad (0 < v_j < \kappa_n)$$
$$\frac{d\hat{Q}_n(\boldsymbol{\theta}_0 + \boldsymbol{v})}{d\theta_j} + p'_{\kappa_n}(v_j) < 0 \quad (-\kappa_n < v_j < 0)$$

for any $\boldsymbol{v} = (v_1, \ldots, v_{p_n})'$ such that $\|\boldsymbol{v}\| = O(\sqrt{p_n/n})$ and for $j = q_n + 1, \ldots, p_n$. Because $p'_{\kappa_n}(u) = \kappa_n \text{sgn}(u)$ for $|u| \leq \kappa_n$, we have

$$\begin{aligned} \frac{d\hat{Q}_n(\boldsymbol{\theta}_0 + \boldsymbol{v})}{d\theta_j} + p'_{\kappa_n}(v_j) &= \frac{d\hat{Q}_n(\boldsymbol{\theta}_0)}{d\theta_j} + \frac{d^2\hat{Q}_n(\boldsymbol{\theta}_0 + \dot{\boldsymbol{v}})}{d\theta_j d\boldsymbol{\theta}'}\boldsymbol{v} + \kappa_n \text{sgn}(v_j) \\ &\equiv I_1 + I_2 + I_3 \end{aligned}$$

for $j = q_n + 1, \ldots, p_n$ and for some $\dot{\boldsymbol{v}}$ such that $\|\dot{\boldsymbol{v}}\| = O_p(\sqrt{p_n/n})$. By Lemma A2, we have $|I_1| = O_p(\sqrt{p_n/n})$. Moreover, by Assumption 8 and Lemma A4, we have

$$\left\| \frac{d^2\hat{Q}_n(\boldsymbol{\theta}_0 + \dot{\boldsymbol{v}})}{d\theta_j d\boldsymbol{\theta}'} \right\| = O_p(1),$$

and thus $|I_2| = O_p(\sqrt{p_n/n})$. Therefore, $I_1$ and $I_2$ are asymptotically dominated by $I_3$. The sign of $d\hat{Q}_n(\boldsymbol{\theta}_0 + \boldsymbol{v})/d\theta_j + p'_{\kappa_n}(v_j)$ is determined by the sign of $v_j$.

Next, we show asymptotic normality. Let $\hat{Q}_{1n}(\boldsymbol{\theta}_1) = \hat{Q}_n(\boldsymbol{\theta}_1, \mathbf{0})$. Lemma A3 and Assumption 5 imply that $\min_{1 \leq j \leq q_n} |\hat{\theta}_j| > a\kappa_n$ w.p.a.1. Moreover, we have $P(\hat{\boldsymbol{\theta}}_{2n} = \mathbf{0}) \to 1$. Thus, expanding the first-order condition for $\hat{\boldsymbol{\theta}}_{1n}$ yields

$$\mathbf{0} = \frac{d\hat{Q}_{1n}(\boldsymbol{\theta}_{10})}{d\boldsymbol{\theta}_1} + \frac{d^2\hat{Q}_{1n}(\dot{\boldsymbol{\theta}}_{1n})}{d\boldsymbol{\theta}_1 d\boldsymbol{\theta}_1'}(\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10})$$

for some $\dot{\boldsymbol{\theta}}_{1n}$ that is located between $\hat{\boldsymbol{\theta}}_{1n}$ and $\boldsymbol{\theta}_{10}$. Combining this with Lemmas A2 and A4 and Assumptions 2 (ii) and 6 (ii), we have

$$V_{1n}^{-1}(\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) = -\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}_1} + o_p\left(\frac{1}{\sqrt{n}}\right),$$

which immediately implies that $\|\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}\| = O_p(\sqrt{q_n/n})$. Moreover, because $\text{tr}(B_n V_{1n} B_n') < C\text{tr}(B_n B_n') < C$ by the assumption of Theorem 2 and Assumption 8, we have

$$\begin{aligned} \sqrt{n}B_n V_{1n}^{-1/2}(\hat{\boldsymbol{\theta}}_{1n} - \boldsymbol{\theta}_{10}) &= -\sqrt{n}B_n V_{1n}^{1/2}\frac{d\hat{Q}_n(\boldsymbol{\theta}_0, \boldsymbol{\lambda}(\boldsymbol{\theta}_0))}{d\boldsymbol{\theta}_1} + o_p(\|B_n V_{1n}^{1/2}\|) \\ &= \sum_{i=1}^{n} \boldsymbol{z}_{ni} + o_p(1), \end{aligned}$$

where

$$\boldsymbol{z}_{ni} = -\frac{1}{\sqrt{n}}B_n V_{1n}^{1/2} E[M_{1i}]' E[\boldsymbol{m}_i \boldsymbol{m}_i']^{-1} \boldsymbol{m}_i.$$

Here, by Assumptions 2 (i) and 8, we have

$$\begin{aligned} E\left[\|z_{ni}\|^4\right] &= \frac{1}{n^2}E\left[\left\{m_i' E[m_i m_i']^{-1} E[M_{1i}] V_{1n}^{1/2} B_n' B_n V_{1n}^{1/2} E[M_{1i}]' E[m_i m_i']^{-1} m_i\right\}^2\right] \\ &\leq \frac{C}{n^2}E\left[\{m_i' m_i\}^2\right] \\ &= O\left(\frac{r_n^2}{n^2}\right). \end{aligned}$$

Furthermore, because $B_n B_n' \to G$, we have $\sum_{i=1}^n E[z_{ni} z_{ni}'] \to G$ and

$$P(\|z_{ni}\| > \epsilon) \leq \frac{E[z_{ni}' z_{ni}]}{\epsilon^2} = O\left(\frac{1}{n}\right).$$

Therefore, we obtain

$$\sum_{i=1}^n E\left[\|z_{ni}\|^2 1\{\|z_{ni}\|^2 > \epsilon\}\right] \leq nE\left[\|z_{ni}\|^4\right]^{1/2} P(\|z_{ni}\| > \epsilon)^{1/2} = o(1),$$

and thus $\sum_{i=1}^n z_{ni} \xrightarrow{d} N(0, G)$ by the Lindeberg-Feller central limit theorem. □

## References


Ando, Tomohiro, and Naoya Sueishi. 2019. Regularization parameter selection for penalized empirical likelihood estimator. *Economics Letters* 178: 1–4. [CrossRef]

Bai, Jushan, and Serena Ng. 2009. Selecting instrumental variables in a data rich environment. *Journal of Time Series Econometrics* 1: 4.

Belloni, Alexandre, Daniel Chen, Victor Chernozhukov, and Christian Hansen. 2012. Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica* 80: 2369–429.

Breiman, Leo. 1996. Heuristics of instability and stabilization in model selection. *Annals of Statistics* 24: 2350–83. [CrossRef]

Caner, Mehmet, and Qingliang Fan. 2015. Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics* 187: 256–74. [CrossRef]

Caner, Mehmet, and Hao Helen Zhang. 2014. Adaptive elastic net for generalized methods of moments. *Journal of Business & Economic Statistics* 32: 30–47. [CrossRef]

Caner, Mehmet. 2009. Lasso-type gmm estimator. *Econometric Theory* 25: 270–90.

Chang, Jinyuan, Song Xi Chen, and Xiaohong Chen. 2015. High dimensional generalized empirical likelihood for moment restrictions with dependent data. *Journal of Econometrics* 185: 283–304. [CrossRef]

Chang, Jinyuan, Cheng Yong Tang, and Tong Tong Wu. 2018. A new scope of penalized empirical likelihood with high-dimensional estimating equations. *Annals of Statistics* 46: 3185–216. [CrossRef]

Cheng, Xu, and Zhipeng Liao. 2015. Select the valid and relevant moments: An information-based lasso for gmm with many moments. *Journal of Econometrics* 186: 443–64. [CrossRef]

Donald, Stephen G., and Whitney K. Newey. 2001. Choosing the number of instruments. *Econometrica* 69: 1161–91. [CrossRef]

Donald, Stephen G., Guido W. Imbens, and Whitney K. Newey. 2003. Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics* 117: 55–93. [CrossRef]

Fan, Jianqing, and Yuan Liao. 2014. Endogeneity in high dimensions. *Annals of Statistics* 42: 872–917. [CrossRef]

Fan, Jianqing, and Runze Li. 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96: 1348–60.

Fan, Jianqing, and Heng Peng. 2004. Nonconcave penalized likelihood with a diverging number of parameters. *Annals of Statistics* 32: 928–61. [CrossRef]

Frank, Ildiko E., and Jerome H. Friedman. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35: 109–35. [CrossRef]

Huang, Jian, and Huiliang Xie. 2007. Asymptotic oracle properties of scad-penalized least squares estimators. *IMS Lecture Notes–Monograph Series* 55: 149–66.

Kuersteiner, Guido, and Ryo Okui. 2010. Constructing optimal instruments by first-stage prediction averaging. *Econometrica* 78: 697–718.

Leng, Chenlei, and Cheng Yong Tang. 2012. Penalized empirical likelihood and growing dimensional general estimating equations. *Biometrika* 99: 703–16. [CrossRef]

Qin, Jin, and Jerry Lawless. 1994. Empirical likelihood and general estimating equations. *Annals of Statistics* 22: 300–25. [CrossRef]

Shi, Zhentao. 2016a. Econometric estimation with high-dimensional moment equalities. *Journal of Econometrics* 195: 104–19. [CrossRef]

Shi, Zhentao. 2016b. Estimation of sparse structral parameter with many endogenous variables. *Econometric Reviews* 35: 1582–608. [CrossRef]

Sueishi, Naoya. 2016. A simple derivation of the efficiency bound for conditional moment restriction models. *Economics Letters* 138: 57–59. [CrossRef]

Tang, Niansheng, Xiaodong Yan, and Puying Zhao. 2018. Exponentially tilted likelihood inference on growing dimensional unconditional moment models. *Journal of Econometrics* 202: 57–74. [CrossRef]

Tibshirani, Robert. 1996. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B* 58: 267–88. [CrossRef]

Zhang, Cun-Hui. 2010. Nearly unbiased variable selection under minimax concave penalty. *Annals of Statistics* 38: 894–942. [CrossRef]

Zou, Hui, and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67: 301–20.