Journal of Business & Economic Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/ubes20

# Culling the Herd of Moments with Penalized Empirical Likelihood

Jinyuan Chang, Zhentao Shi & Jia Zhang

**To cite this article:** Jinyuan Chang, Zhentao Shi & Jia Zhang (2022): Culling the Herd of Moments with Penalized Empirical Likelihood, Journal of Business & Economic Statistics, DOI: 10.1080/07350015.2022.2071903

**To link to this article:** https://doi.org/10.1080/07350015.2022.2071903

JOURNAL OF BUSINESS & ECONOMIC STATISTICS
2022, VOL. 00, NO. 0, 1–15
https://doi.org/10.1080/07350015.2022.2071903

Taylor & Francis
Taylor & Francis Group

Check for updates

# Culling the Herd of Moments with Penalized Empirical Likelihood

Jinyuan Chang[a,b], Zhentao Shi[c,d], and Jia Zhang[b]

[a]Guizhou Key Laboratory of Big Data Statistical Analysis, Guizhou University of Finance and Economics, Guiyang, China; [b]Joint Laboratory of Data Science and Business Intelligence, Southwestern University of Finance and Economics, Chengdu, China; [c]Department of Economics, the Chinese University of Hong Kong, Sha Tin, New Territories, Hong Kong Special Administrative Region of China; [d]School of Economics, Georgia Institute of Technology, Atlanta, GA

**ABSTRACT**

Models defined by moment conditions are at the center of structural econometric estimation, but economic theory is mostly agnostic about moment selection. While a large pool of valid moments can potentially improve estimation efficiency, in the meantime a few invalid ones may undermine consistency. This article investigates the empirical likelihood estimation of these moment-defined models in high-dimensional settings. We propose a penalized empirical likelihood (PEL) estimation and establish its oracle property with consistent detection of invalid moments. The PEL estimator is asymptotically normally distributed, and a projected PEL procedure further eliminates its asymptotic bias and provides more accurate normal approximation to the finite sample behavior. Simulation exercises demonstrate excellent numerical performance of these methods in estimation and inference.

**ARTICLE HISTORY**
Received August 2021
Accepted April 2022

**KEYWORDS**
Empirical likelihood; Estimating equations; High-dimensional statistical methods; Misspecification; Moment selection; Penalized likelihood

## 1. Introduction

Economists' perennial pursuit of structural mechanisms leads to models defined by moments. These models can be written in a semiparametric form $\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta}_0)\}=\mathbf{0}$, where $\mathbf{g}=(g_j)_{j\in\{1,\ldots,r\}}$ is a vector of $r$ estimating functions, $\boldsymbol{\theta}_0$ is a vector of unknown parameters, and $\mathbf{X}_i$ is observed data. To estimate these models, the most popular method is *generalized method of moments* (GMM) (Hansen 1982). *Empirical likelihood* (EL) (Qin and Lawless 1994) is a competitive alternative to GMM, thanks to its nice statistical properties. Both GMM and EL are essential building blocks of modern econometrics (Anatolyev and Gospodinov 2011).

Ideally, economists count on economic theory to guide the choice of variables and moments. However, the truth is that most economic theories are parsimonious abstractions and rarely pinpoint these choices in data-rich environments. The indeterminacy of moment selection brings about three related issues. The first is *weak moments* (Stock, Wright, and Yogo 2014), which threatens identification of the true value $\boldsymbol{\theta}_0$ when multiple $\boldsymbol{\theta}$'s satisfy $\mathbb{E}\{\mathbf{g}(\mathbf{X}_i;\boldsymbol{\theta})\}\approx\mathbf{0}$. Practitioners respond to the concerns of weak moments by adding more moment conditions in the hope to strengthen identification, causing the second issue of *many moments* (Roodman 2009). The hazard of many moments is the possible inclusion of *invalid moments*, meaning $\mathbb{E}\{g_j(\mathbf{X}_i;\boldsymbol{\theta}_0)\}\neq 0$ for some $j\in\{1,\ldots,r\}$ (Murray 2006), which is the third issue. These cited survey articles highlight the unease incurred by the three challenges and econometricians' efforts in coping with them.

When an underlying economic theory is ambivalent, empirical results based on it can be controversial and susceptible to cherry-picking. In such a circumstance, of great importance are data-driven methods to guide and discipline moment selection. In the low-dimensional settings, Andrews and Lu (2001) propose the GMM information criteria, and Hong, Preston, and Shum (2003) follow with the counterpart for EL. Information criteria are evaluated exhaustively at all combinations of moments, and the computation becomes infeasible when there are many potential moments. To overcome this challenge, Liao (2013) ushers the adaptive Lasso shrinkage into GMM to select among a finite number of moments, and Cheng and Liao (2015) further extend it to deal with a diverging number of moments and accommodate invalid ones.

Unprecedented progress in computation and information technology fuel an arms race between the sheer size of the data and the scale of empirical models. In the era of big data, on the one hand the cost of data collection and processing is tremendously lowered and rich datasets open new perspectives to inspect a myriad of problems; on the other hand economists attempt to build general models to capture various sources of heterogeneity in observational data. Empirical applications abound with models of many potential moments. For instance, Eaton, Kortum, and Kramarz (2011) create 1360 moments to estimate a structural trade model, Altonji, Smith, and Vidangos (2013) match 2429 moments implied by a model of earning dynamics, and early works of linear instrumental variables (IV) models produce thousands of instruments by interacting variables (Angrist and Krueger 1992). Although the sample sizes in these examples are nontrivial, the proliferation of moments calls for a moment selection procedure capable of handling high-dimensional moments at a magnitude unrestricted by the

**CONTACT** Zhentao Shi zhentao.shi@gatech.edu School of Economics, Georgia Institute of Technology, Atlanta, GA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/UBES.

sample size. In particular, invalid ones that jeopardize consistency must be identified and "culled" from the herd of moments.

The quadratic form of the usual GMM criterion function is incompatible with high-dimensional moments (Shi 2016a, 2016b), and thus Belloni et al. (2018) regularize GMM with the sup-norm. In this article, we contribute the most general and versatile procedure for high-dimensional nonlinear settings, to the best of our knowledge. We first develop a *penalized empirical likelihood* (PEL) solution (Chang, Tang, and Wu 2018) to deal with valid and invalid moments simultaneously. We neutralize the invalid moments by an auxiliary parameter, following Liao (2013) and Cheng and Liao (2015). We establish the rate of convergence and asymptotic normality of the PEL estimator, and show the efficiency gain from incorporating extra valid moments. Under suitable conditions, it transpires that PEL enjoys the oracle property of consistent moment selection and parameter selection.

The asymptotic normal distribution of the PEL estimator involves a bias term caused by the high-dimensional moments. To spare the estimation of the bias term, we can take further actions to project out the influence of the high-dimensional nuisance parameter in the PEL estimator, which is called *projected PEL* (PPEL) (Chang et al. 2021). The asymptotic normality of the PPEL estimator is free of bias, which facilitates statistical inference of the structural parameter as well as the validity of moments. Invoking statistical learning to assist our decisions, our method fits well in the recent trend of machine learning for the automatic selection of moments and variables.

Although this article follows Chang, Tang, and Wu (2018) for estimation and Chang et al. (2021) for inference procedures, the key insight lies in the observations that the high-dimensional auxiliary parameter, which signifies the magnitude of misspecification, can be incorporated in the EL method as an additional high-dimensional parameter. In this article, "high-dimensional" means that the numbers of parameters and/or moments are larger than the sample size, which goes beyond the scope of Liao (2013) and Cheng and Liao (2015). On the other hand, in order to adapt Chang, Tang, and Wu (2018) and Chang et al. (2021) to accommodate misspecified moments, we must deal with the distinctive roles of the main parameter of interest and the auxiliary parameter in identification and the technical challenges induced by them. Differences from Chang, Tang, and Wu (2018) are highlighted in Section 3 about the rates of converges of the two components of the parameters, and those from Chang et al. (2021) are elaborated in Section 5 about the ways of confidence region construction.

*Literature review*. Our article stands on strands of literature, which are too vast to survey exhaustively. The accumulation of moments started from the linear IV model (Angrist 1990; Angrist and Keueger 1991). The linear IV model motivates theoretical research on issues of many IV (Bekker 1994), weak IV (Stock and Yogo 2005; Andrews and Cheng 2012), many weak IV (Chao et al. 2011; Hansen and Kozbur 2014), invalid moments and many invalid IV (Kolesár et al. 2015; Windmeijer et al. 2019), to name a few. In high-dimensional contexts, Belloni et al. (2012) use Lasso method for IV selection in the first stage, and Belloni, Chernozhukov, and Hansen (2014) deal with post-selection inference. Using the linear structure, Gold, Lederer, and Tao (2020) and Caner and Kock (2018) provide inferential procedures for low-dimensional parameters in models with high-dimensional endogenous variables and high-dimensional IVs. Our method includes the linear IV model as a special case. In particular, the case of high-dimensional structural parameters is elaborated in Section 4.

The proliferation of moments spreads from linear IV models to nonlinear models. For example, in empirical industrial organization researchers bring in moments from various resources, some of which are guided by economic theory, to mitigate the concerns of weak identification and improve estimation efficiency (Ackerberg et al. 2007). In empirical macroeconomics, identification failure and moment misspecification are common issues (Mavroeidis 2005). Under the GMM framework, inference under weak moments (Stock and Wright 2000; Kleibergen 2005; Andrews and Mikusheva 2020), estimation under many weak moments (Han and Phillips 2006), and robust procedures for invalid moments (DiTraglia 2016; Caner, Han, and Lee 2018) have been developed.

EL's attractive theoretical properties are studied extensively (Kitamura 2001; Otsu 2010; Matsushita and Otsu 2013; Chang, Chen, and Chen 2015). Otsu (2006) and Newey and Windmeijer (2009) deal with its inference under weak IV, and Caner and Fan (2015) select instruments in linear IV models. Penalization schemes on EL have been introduced by Otsu (2007), Tang and Leng (2010) and Chang, Tang, and Wu (2018).

*Organization.* The rest of the article is organized as follows. Section 2 introduces the model and the analytic framework. We first derive the asymptotic properties of our estimation in the low-dimensional case in Section 3, extend them to the high-dimensional structural parameter in Section 4, and we further refine PEL with projection to eliminate its bias in Section 5. The theoretical results are supported in Section 6 by Monte Carlo simulations. The influential study of the determinants of economic outcomes after colonialism is revisited in Section 7. Section 8 concludes the article. Due to the limitations of space, the proofs and the technical details of secondary importance are relegated into the supplementary materials.

*Notations.* We conclude this Introduction with notations used throughout the article. "Low-dimensional" refers to the cases that the number of parameters or moments is much smaller than the sample size $n$, whereas "high-dimensional" refers to the opposite. For two sequences of positive numbers $\{a_n\}$ and $\{b_n\}$, we write $a_n \lesssim b_n$ or $b_n \gtrsim a_n$ if there exists a positive constant $c$ such that $\limsup_{n\to\infty} a_n/b_n \le c$, and write $a_n \ll b_n$ or $b_n \gg a_n$ if $\limsup_{n\to\infty} a_n/b_n = 0$.

Denote by $1(\cdot)$ the indicator function. For a positive integer $q$, we write $[q] = \{1, \ldots, q\}$. For a $q \times q$ symmetric matrix $\mathbf{M}$, denote by $\lambda_{\min}(\mathbf{M})$ and $\lambda_{\max}(\mathbf{M})$ the smallest and largest eigenvalues of $\mathbf{M}$, respectively. For a $q_1 \times q_2$ matrix $\mathbf{B} = (b_{i,j})_{q_1\times q_2}$, let $\mathbf{B}^{\mathrm{T}}$ be its transpose, $\mathbf{B}^{\otimes 2} = \mathbf{B}\mathbf{B}^{\mathrm{T}}$, $\mathbf{B}^{\circ\kappa} = (|b_{i,j}|^{\kappa})_{q_1\times q_2}$ for any $\kappa > 0$, $|\mathbf{B}|_\infty = \max_{i\in[q_1],j\in[q_2]} |b_{i,j}|$ be the sup-norm, and $\|\mathbf{B}\|_2 = \lambda_{\max}^{1/2}(\mathbf{B}^{\otimes 2})$ be the spectral norm. Specifically, if $q_2 = 1$, we use $|\mathbf{B}|_\infty = \max_{i\in[q_1]} |b_{i,1}|$, $|\mathbf{B}|_1 = \sum_{i=1}^{q_1} |b_{i,1}|$ and $|\mathbf{B}|_2 = (\sum_{i=1}^{q_1} b_{i,1}^2)^{1/2}$ to denote the $L_\infty$-norm, $L_1$-norm and $L_2$-norm of the $q_1$-dimensional vector $\mathbf{B}$, respectively. For two square matrices $\mathbf{M}_1$ and $\mathbf{M}_2$, we say $\mathbf{M}_1 \le \mathbf{M}_2$ if $(\mathbf{M}_2 - \mathbf{M}_1)$ is a positive semidefinite matrix.

The population mean is denoted by $\mathbb{E}(\cdot)$, and the sample mean by $\mathbb{E}_n(\cdot) = n^{-1}\sum_{i=1}^{n}\{\cdot\}$. For a given index set $\mathcal{L}$, let $|\mathcal{L}|$ be its cardinality. For a generic multivariate function $\mathbf{h}(\cdot;\cdot)$, 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 write $\mathbf{a}_{\mathcal{L}}$ as the corresponding subvector of $\mathbf{a}$. For simplicity and when no confusion arises, we use the generic notation $\mathbf{h}_i(\boldsymbol{\theta})$ as the equivalence to $\mathbf{h}(\mathbf{X}_i;\boldsymbol{\theta})$, and $\nabla_{\boldsymbol{\theta}}\mathbf{h}_i(\boldsymbol{\theta})$ for the first-order partial derivative of $\mathbf{h}_i(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Denote by $h_{i,k}(\boldsymbol{\theta})$ the $k$th component of $\mathbf{h}_i(\boldsymbol{\theta})$, and by $\nabla^2_{\boldsymbol{\theta}}h_{i,k}(\boldsymbol{\theta})$ the second derivative of $h_{i,k}(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. Let $\bar{\mathbf{h}}(\boldsymbol{\theta}) = \mathbb{E}_n\{\mathbf{h}_i(\boldsymbol{\theta})\}$, and write its $k$th component as $\bar{h}_k(\boldsymbol{\theta}) = \mathbb{E}_n\{h_{i,k}(\boldsymbol{\theta})\}$. Analogously, let $\mathbf{h}_{i,\mathcal{L}}(\boldsymbol{\theta}) = \mathbf{h}_{\mathcal{L}}(\mathbf{X}_i;\boldsymbol{\theta})$ and $\bar{\mathbf{h}}_{\mathcal{L}}(\boldsymbol{\theta}) = \mathbb{E}_n\{\mathbf{h}_{i,\mathcal{L}}(\boldsymbol{\theta})\}$.

## 2. Empirical Likelihood with a Herd of Moments

In this section we introduce the model and the EL estimation. Let $\mathbf{X}_1,\ldots,\mathbf{X}_n$ be $d$-dimensional independent and identically distributed generic observations, and $\boldsymbol{\theta} = (\theta_1,\ldots,\theta_p)^{\mathrm{T}}$ be a $p$-dimensional parameter taking values in $\boldsymbol{\Theta} \subset \mathbb{R}^p$. For a set of $r_1$ estimating functions $\mathbf{g}^{(\mathcal{I})}(\cdot;\cdot) = \{g_j^{(\mathcal{I})}(\cdot;\cdot)\}_{j\in\mathcal{I}}$, the information of the model parameter $\boldsymbol{\theta}$ is collected by the unbiased moment condition

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

at the unknown true parameter $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, where $r_1 \geq p$ is necessary for identifying $\boldsymbol{\theta}_0$. The superscript $(\mathcal{I})$ labels this *initial* set, to be distinguished from the other set $(\mathcal{D})$ in Section 2.2.

### 2.1. EL Estimation with Valid Moments

Motivated from empirical applications in asset pricing, the two-step GMM (Hansen 1982) was the default estimating method for the moment-defined model (2.1). Intensive theoretical studies and numerical evidence in 1980s and 1990s revealed some undesirable finite-sample properties of the two-step GMM (Altonji and Segal 1996). EL and the continuously updating GMM (CUE) (Hansen, Heaton, and Yaron 1996) emerged as competitive solutions, and they were later unified as members of *generalized empirical likelihood* (Newey and Smith 2003).

This article focuses on EL with estimating equations:

$$L(\boldsymbol{\theta}) = \max\left\{\prod_{i=1}^{n}\pi_i : \pi_i > 0\,,\ \sum_{i=1}^{n}\pi_i = 1\,,\ \sum_{i=1}^{n}\pi_i\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}) = \mathbf{0}\right\},$$

proposed by Qin and Lawless (1994) based on the seminal idea of EL (Owen 1988, 1990). Maximizing $L(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ delivers the EL estimator $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} = \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} L(\boldsymbol{\theta})$, which can be carried out equivalently by solving the corresponding dual problem

$$\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\ \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta})}\ \sum_{i=1}^{n}\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}\,, \tag{2.2}$$

where $\hat{\Lambda}_n^{(\mathcal{I})}(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^{r_1} : \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}) \in \mathcal{V}$ for any $i \in [n]\}$ and $\mathcal{V}$ is an open interval containing zero.

To fix ideas, we first study the asymptotic property of $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$ under regularity conditions. When the sample size $n$ grows, we adopt the asymptotic framework of Hjort, McKeague, and Van Keilegom (2009) and Chang, Chen, and Chen (2015) to take the observations $\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}_{i=1}^{n}$ as a multi-index array, where $r_1, p$ and $d$ may depend on $n$. Proposition A.1 in the supplementary materials shows that the standard asymptotic normality for $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$ holds under mild regularity conditions.

### 2.2. Oracle EL Estimation

In applied econometrics, there has been a tendency of assembling many IVs or creating many moments to identify the parameter of interest. In these applications, researchers often have some ideas about the relative importance of moments; in the meantime, when researchers work with more and more moments, some invalid ones may creep in.

*Example 1.* Eaton, Kortum, and Kramarz (2011) deem as the key moments $128(=2^7)$ combinations of the largest seven trade partners of France, which are more important than the other 1232 moments. Angrist (1990) treats the date of birth (DOB) as the key IV and it is complemented by the IVs generated by interactions; recently Kolesár et al. (2015) raise the potential invalidity among these interaction terms.[1]

To put into an analytic framework a herd of extra moments with unknown validity ex ante, suppose that $r_2$ estimating functions $\mathbf{g}^{(\mathcal{D})} = \{g_j^{(\mathcal{D})}\}_{j\in\mathcal{D}}$ are partitioned into two groups:

$$\mathcal{A} = \{j \in \mathcal{D} : \mathbb{E}\{g_{i,j}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\} = 0\} \quad \text{and}$$
$$\mathcal{A}^{\mathrm{c}} = \{j \in \mathcal{D} : \mathbb{E}\{g_{i,j}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\} \neq 0\}.$$

Here the estimating functions in the set $\mathcal{A}$ are correctly specified, which can help improve the efficiency in estimating $\boldsymbol{\theta}_0$. On the contrary, those in $\mathcal{A}^{\mathrm{c}}$ are misspecified, and they can only undermine the identification of $\boldsymbol{\theta}_0$.

If there is an "oracle" that reveals which components of $\mathbf{g}^{(\mathcal{D})}$ belong to $\mathcal{A}$, that is, the valid estimating functions, we can collect all valid estimating functions indexed by $\mathcal{H} := \mathcal{I}\cup\mathcal{A}$ to estimate $\boldsymbol{\theta}_0$. Denote $\mathbf{g}^{(\mathcal{H})} = \{\mathbf{g}^{(\mathcal{I}),\mathrm{T}}, \mathbf{g}^{(\mathcal{A}),\mathrm{T}}\}^{\mathrm{T}}$ and $h = |\mathcal{H}|$ as the total number of valid moments. The associated EL estimation for $\boldsymbol{\theta}_0$ based on the estimating functions $\mathbf{g}^{(\mathcal{H})}$ is given by

$$\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{H})} = \arg\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\ \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{H})}(\boldsymbol{\theta})}\ \sum_{i=1}^{n}\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{H})}(\boldsymbol{\theta})\}\,,$$

where $\hat{\Lambda}_n^{(\mathcal{H})}(\boldsymbol{\theta}) = \{\boldsymbol{\lambda} \in \mathbb{R}^h : \boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{H})}(\boldsymbol{\theta}) \in \mathcal{V}$ for any $i \in [n]\}$. By the same arguments as those in Proposition A.1 in the supplementary materials, the following Proposition 2.1 verifies the asymptotic normality for $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{H})}$.

*Proposition 2.1 (Oracle property).* Assume that $\mathbf{g}^{(\mathcal{H})}$ satisfies the conditions (A.1)–(A.5) in the supplementary materials associated with $\mathbf{g}^{(\mathcal{H})}$. If $h^3n^{-1+2/\gamma} = o(1)$ and $h^3p^2n^{-1} =$

[1] Identification of the simple linear IV model requires the IVs satisfying the orthogonality condition and the relevance condition. However, the more relevant an IV to the endogenous variables, the more likely it is that the so-called "IV" is correlated with the structural error, thereby violating orthogonality. There is a thin line between a valid IV and an invalid one.

$o(1)$, then $\sqrt{n}\boldsymbol{\alpha}^{\mathrm{T}}\{\mathbf{J}^{(\mathcal{H})}\}^{1/2}\{\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{H})}-\boldsymbol{\theta}_0\}\xrightarrow{d}\mathcal{N}(0,1)$ as $n\to\infty$ for any $\boldsymbol{\alpha}\in\mathbb{R}^p$ with $|\boldsymbol{\alpha}|_2=1$, where $\mathbf{J}^{(\mathcal{H})}=([\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_i^{(\mathcal{H})}(\boldsymbol{\theta}_0)\}]^{\mathrm{T}}\{\mathbf{V}^{(\mathcal{H})}(\boldsymbol{\theta}_0)\}^{-1/2})^{\otimes 2}$ with $\mathbf{V}^{(\mathcal{H})}(\boldsymbol{\theta}_0)=\mathbb{E}\{\mathbf{g}_i^{(\mathcal{H})}(\boldsymbol{\theta}_0)^{\otimes 2}\}$.

Obviously $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{H})}$ is more efficient than $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$ thanks for the additional moment restrictions in $\mathcal{A}$, which disclose more information about the parameter and thus tie down the estimation variability. This is parallel to Hall et al. (2007) in GMM for finite numbers of parameters and moments. In reality, however, we are oblivious to $\mathcal{A}$, and therefore we cannot blindly summon all $r_2$ estimating functions in $\mathbf{g}^{(\mathcal{D})}$ to estimate $\boldsymbol{\theta}_0$. Following Liao (2013), we introduce an auxiliary parameter $\boldsymbol{\xi}=\mathbb{E}\{\mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta})\}$ to neutralize the misspecified moments. Denote the augmented parameter $\boldsymbol{\psi}=(\boldsymbol{\theta}^{\mathrm{T}},\boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$, and the augmented parameter space $\boldsymbol{\Psi}=\boldsymbol{\Theta}\times\boldsymbol{\Upsilon}$. Let $r=r_1+r_2$ and stack these $r$ estimating functions as $\mathbf{g}^{(\mathcal{T})}(\mathbf{X};\boldsymbol{\psi})=\{\mathbf{g}^{(\mathcal{I})}(\mathbf{X};\boldsymbol{\theta})^{\mathrm{T}},\mathbf{g}^{(\mathcal{D})}(\mathbf{X};\boldsymbol{\theta})^{\mathrm{T}}-\boldsymbol{\xi}^{\mathrm{T}}\}^{\mathrm{T}}$. Then $\boldsymbol{\psi}_0=(\boldsymbol{\theta}_0^{\mathrm{T}},\boldsymbol{\xi}_0^{\mathrm{T}})^{\mathrm{T}}$ with $\boldsymbol{\xi}_0=\mathbb{E}\{\mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}$ can be identified by

$$\mathbb{E}\{\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}=\mathbb{E}\begin{Bmatrix}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\\ \mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta}_0)-\boldsymbol{\xi}_0\end{Bmatrix}=\mathbf{0}\,. \tag{2.3}$$

Can we directly include the auxiliary parameter $\boldsymbol{\xi}$ into the EL estimation? The answer is negative. If the auxiliary parameter is not regularized, the EL estimators with and without $\boldsymbol{\xi}$ are the same, up to numerical errors.

*Proposition 2.2.* If $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$ is the unique solution of (2.2), then $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}=\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$, where $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}$ is the associated subvector of $\hat{\boldsymbol{\psi}}_{\mathrm{EL}}^{(\mathcal{T})}=\arg\min_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})}\sum_{i=1}^n\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\}$ for the estimation of $\boldsymbol{\theta}_0$ with $\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})=\{\boldsymbol{\lambda}\in\mathbb{R}^r:\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\in\mathcal{V}$ for any $i\in[n]\}$.

Proposition 2.2 states the equivalence between $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{T})}$ and $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{I})}$. In the next section, we will show that desirable efficiency, as in the oracle estimator $\hat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(\mathcal{H})}$, can be achieved if we extend Liao (2013)'s idea of shrinking the auxiliary parameter[2] by further penalizing the Lagrange multipliers associated with the high-dimensional moments.

## 3. High-Dimensional Moments with Low-Dimensional $\boldsymbol{\theta}$

We consider in this section the model (2.3) with a low-dimensional parameter $\boldsymbol{\theta}$, low-dimensional estimating functions $\mathbf{g}^{(\mathcal{I})}$ and high-dimensional estimating functions $\mathbf{g}^{(\mathcal{D})}$. In this setting, $p$ and $r_1$ are either fixed or diverge at some slow polynomial rates of $n$, whereas $r_2$ can grow much larger than $n$. The components in the parameter $\boldsymbol{\theta}$ are indexed by $\mathcal{P}$. All moments together are indexed by $\mathcal{T}=\mathcal{I}\cup\mathcal{D}$, in which the researcher knows that those in $\mathcal{I}$ are correctly specified in the sense that $\mathbb{E}\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}_0)\}=\mathbf{0}$, whereas she is uncertain whether those in $\mathcal{D}=\mathcal{A}\cup\mathcal{A}^c$ satisfy $\mathbb{E}\{\mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}=\mathbf{0}$ or not. As there is a one-to-one relationship between $\boldsymbol{\xi}$ and the uncertain moments, we use $\mathcal{D}$ to index the components in $\boldsymbol{\xi}$ as well.

We propose the following PEL to simultaneously estimate the unknown parameter $\boldsymbol{\theta}_0$ and determine the validity of the estimating functions in $\mathbf{g}^{(\mathcal{D})}$:

$$(\hat{\boldsymbol{\theta}}_{\mathrm{PEL}}^{\mathrm{T}},\hat{\boldsymbol{\xi}}_{\mathrm{PEL}}^{\mathrm{T}})^{\mathrm{T}}=\arg\min_{\boldsymbol{\psi}\in\boldsymbol{\Psi}}\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})}\Bigg[\frac{1}{n}\sum_{i=1}^n\log\{1+\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\}-\sum_{j\in\mathcal{D}}P_{2,\nu}(|\lambda_j|)+\sum_{k\in\mathcal{D}}P_{1,\pi}(|\xi_k|)\Bigg], \tag{3.1}$$

where $\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})=\{\boldsymbol{\lambda}\in\mathbb{R}^r:\boldsymbol{\lambda}^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\in\mathcal{V}$ for any $i\in[n]\}$, and $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ are two penalty functions with tuning parameters $\pi$ and $\nu$, respectively. With the penalty function $P_{2,\nu}(\cdot)$ and appropriately selected tuning parameter $\nu$, the estimator $(\hat{\boldsymbol{\theta}}_{\mathrm{PEL}}^{\mathrm{T}},\hat{\boldsymbol{\xi}}_{\mathrm{PEL}}^{\mathrm{T}})^{\mathrm{T}}$ is associated with a sparse Lagrange multiplier $\boldsymbol{\lambda}$. Since the sparse $\boldsymbol{\lambda}$ invokes a subset of the estimating functions $\mathbf{g}^{(\mathcal{T})}(\cdot;\cdot)$, it digests the high-dimensional moments as long as the number of nonzero components in $\boldsymbol{\lambda}$ is small. On the other hand, the penalty $P_{1,\pi}(\cdot)$ is applied to identify which components in $\mathbf{g}^{(\mathcal{D})}(\cdot;\cdot)$ are correctly specified and can estimate $(\xi_{0,k})_{k\in\mathcal{A}}$ exactly as 0 with high probability.

These penalties originally appeared in Chang, Tang, and Wu (2018) though, there are several important differences. First, given the low-dimensional $\boldsymbol{\theta}$, it is unnecessary to assume sparsity on $\boldsymbol{\theta}_0$ for identification. As a result, our procedure here in (3.1) only penalizes $\boldsymbol{\xi}$ and $\boldsymbol{\lambda}_{\mathcal{D}}$, not the entire parameter $\boldsymbol{\psi}=(\boldsymbol{\theta}^{\mathrm{T}},\boldsymbol{\xi}^{\mathrm{T}})^{\mathrm{T}}$ and the associated Lagrange multiplier $\boldsymbol{\lambda}$. Were $\boldsymbol{\lambda}_{\mathcal{I}}$ penalized, we would rule out some components of $\mathbf{g}^{(\mathcal{I})}(\cdot;\cdot)$ and thus may suffer loss of estimation efficiency for $\boldsymbol{\theta}_0$. Second, the identification condition invoked in this article deviates from that in Chang, Tang, and Wu (2018). Our Condition 1 for identification is imposed on $\boldsymbol{\theta}_0$, rather than the whole parameter $\boldsymbol{\psi}_0$. In contrast, Chang, Tang, and Wu (2018) specify identification condition for the parameter entity. While estimators in Chang, Tang, and Wu (2018) share the same rate of convergence, we must deal with the disparate rates of convergence of the main component $\hat{\boldsymbol{\theta}}_{\mathrm{PEL}}$ and the auxiliary $\hat{\boldsymbol{\xi}}_{\mathrm{PEL}}$, which significantly complicates the theoretical analysis of (3.1), as detailed in Proposition 3.1.

For any penalty function $P_\tau(\cdot)$ with a tuning parameter $\tau$, let $\rho(t;\tau)=\tau^{-1}P_\tau(t)$ for any $t\in[0,\infty)$ and $\tau\in(0,\infty)$. We assume that the two penalty functions $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ involved in (3.1) belong to the following class:

$$\mathscr{P}=\{P_\tau(\cdot):\rho(t;\tau)\text{ is increasing in }t\in[0,\infty)\text{ and has continuous derivative }\rho'(t;\tau)\text{ for any }t\in(0,\infty)\text{ with }\rho'(0^+;\tau)\in(0,\infty),\text{ where }\rho'(0^+;\tau)\text{ is independent of }\tau\}\,. \tag{3.2}$$

This class $\mathscr{P}$, considered in Lv and Fan (2009), is broad and general. The commonly used $L_1$ penalty, SCAD penalty (Fan and Li 2001) and MCP penalty (Zhang 2010) are all included

[2] In GMM estimation under a fixed $r$, Liao (2013) shrinks $\boldsymbol{\xi}$ toward zero using the adaptive Lasso (Zou 2006).

in $\mathscr{P}$. When $P_\tau(\cdot) \in \mathscr{P}$, we write the associated $\rho'(0^+;\tau)$ as $\rho'(0^+)$ for simplification.

To establish the limiting distribution of $\hat{\boldsymbol\theta}_{\rm PEL}$ in (3.1), we assume the following regularity conditions.

*Condition 1.* There exists a universal constant $K_1 > 0$ such that

$$\inf_{\boldsymbol\theta\in\boldsymbol\Theta:\,|\boldsymbol\theta-\boldsymbol\theta_0|_\infty>\varepsilon} |\mathbb{E}\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol\theta)\}|_\infty \geq K_1\varepsilon$$

for any $\varepsilon > 0$.

*Condition 2.* There exist universal constants $K_2 > 0$ and $\gamma > 4$ such that

$$\max_{j\in\mathcal{I}}\mathbb{E}\bigg\{\sup_{\boldsymbol\theta\in\boldsymbol\Theta}|g_{i,j}^{(\mathcal{I})}(\boldsymbol\theta)|^\gamma\bigg\} + \max_{j\in\mathcal{D}}\mathbb{E}\bigg\{\sup_{\boldsymbol\theta\in\boldsymbol\Theta}|g_{i,j}^{(\mathcal{D})}(\boldsymbol\theta)|^\gamma\bigg\} \leq K_2\,.$$

*Condition 3.* For each $j \in \mathcal{T}$, the function $g_j^{(\mathcal{T})}(\mathbf{X};\boldsymbol\psi)$ is twice continuously differentiable with respect to $\boldsymbol\psi \in \boldsymbol\Psi$ for any $\mathbf{X}$. For $\gamma$ specified in Condition 2,

$$\sup_{\boldsymbol\psi\in\boldsymbol\Psi}\bigg(|\mathbb{E}_n[\{\nabla_{\boldsymbol\psi}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol\psi)\}^{\circ 2}]|_\infty + \max_{j\in\mathcal{T}}|\mathbb{E}_n[\{\nabla^2_{\boldsymbol\psi}g_{i,j}^{(\mathcal{T})}(\boldsymbol\psi)\}^{\circ 2}]|_\infty + |\mathbb{E}_n[\{\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol\psi)\}^{\circ\gamma}]|_\infty\bigg) = O_{\rm p}(1)\,.$$

There is some universal constant $K_3 > 0$ such that $\sup_{\boldsymbol\theta\in\boldsymbol\Theta}|\mathbb{E}\{\nabla_{\boldsymbol\theta}\mathbf{g}_{i,\mathcal{A}^c}^{(\mathcal{D})}(\boldsymbol\theta)\}|_\infty \leq K_3$.

For any index set $\mathcal{F} \subset \mathcal{T}$ and $\boldsymbol\psi \in \boldsymbol\Psi$, define $\mathbf{V}_{\mathcal{F}}^{(\mathcal{T})}(\boldsymbol\psi) = \mathbb{E}\{\mathbf{g}_{i,\mathcal{F}}^{(\mathcal{T})}(\boldsymbol\psi)^{\otimes 2}\}$. When $\mathcal{F} = \mathcal{T}$, we write $\mathbf{V}^{(\mathcal{T})}(\boldsymbol\psi) = \mathbf{V}_{\mathcal{T}}^{(\mathcal{T})}(\boldsymbol\psi)$ for conciseness.

*Condition 4.* There exists a universal constant $K_4 > 1$ such that $K_4^{-1} < \lambda_{\min}\{\mathbf{V}^{(\mathcal{T})}(\boldsymbol\psi_0)\} \leq \lambda_{\max}\{\mathbf{V}^{(\mathcal{T})}(\boldsymbol\psi_0)\} < K_4$.

Conditions 1–4 are standard regularity assumptions in the literature. Condition 1 is an identification assumption of the estimating equations in the known set $\mathcal{I}$. Write $\boldsymbol\xi_0 = (\xi_{0,k})_{k\in\mathcal{D}}$, and we continue by defining

$$a_n = \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|) \text{ and } \phi_n = \max\{pa_n^{1/2}, pr_1^{1/2}\aleph_n, \nu\} \tag{3.3}$$

in this section, where $\aleph_n = (n^{-1}\log r)^{1/2}$. Suppose:

$$\text{There exist } \chi_n \to 0 \text{ and } c_n \to 0 \text{ with } \phi_n c_n^{-1} \to 0 \text{ such that } \max_{k\in\mathcal{A}^c}\sup_{0<t<|\xi_{0,k}|+c_n} P'_{1,\pi}(t) = O(\chi_n) \tag{3.4}$$

to control the bias induced by $P_{1,\pi}(\cdot)$ on $\hat{\boldsymbol\xi}_{\rm PEL}$. With the assumption $\phi_n = o(\min_{k\in\mathcal{A}^c}|\xi_{0,k}|)$ that the nonzero components of $\boldsymbol\xi_0$ do not diminish to zero too fast, (3.4) can be replaced by

$$\max_{k\in\mathcal{A}^c}\sup_{c|\xi_{0,k}|<t<c^{-1}|\xi_{0,k}|} P'_{1,\pi}(t) = O(\chi_n) \tag{3.5}$$

for some constant $c \in (0,1)$. If we select $P_{1,\pi}(\cdot)$ as an asymptotically unbiased penalty such as SCAD or MCP, we have $\chi_n = 0$ in (3.5) when

$$\min_{k\in\mathcal{A}^c}|\xi_{0,k}| \gg \max\{\phi_n,\pi\}\,. \tag{3.6}$$

To simplify the presentation, in this section we assume that (3.6) holds and $\chi_n = 0$ in (3.5).[3]

To allocate the parameter of interest $\boldsymbol\theta$ and those invalid moments in $\mathcal{A}^c$, we define an index set $\mathcal{S} = \mathcal{P}\cup\mathcal{A}^c$ and its cardinality $s = |\mathcal{S}|$. For any $\boldsymbol\psi \in \boldsymbol\Psi$, the set $\mathcal{S}$ picks out $\boldsymbol\psi_{\mathcal{S}} = (\boldsymbol\theta^{\rm T}, \boldsymbol\xi_{\mathcal{A}^c}^{\rm T})^{\rm T}$. Since $\mathcal{A} = \mathcal{S}^c = (\mathcal{P}\cup\mathcal{D})\backslash\mathcal{S}$, the corresponding auxiliary coefficients can be written as $\boldsymbol\psi_{\mathcal{S}^c} = \boldsymbol\xi_{\mathcal{A}}$ and furthermore $\boldsymbol\psi_{0,\mathcal{S}^c} = \mathbf{0}$ as $\mathcal{A}$ is the set of valid moments. Given some constant $C_* \in (0,1)$, define $\mathcal{M}^*_{\boldsymbol\psi} = \mathcal{I}\cup\mathcal{D}^*_{\boldsymbol\psi}$ with $\mathcal{D}^*_{\boldsymbol\psi} = \{j \in \mathcal{D} : |\bar{g}_j^{(\mathcal{T})}(\boldsymbol\psi)| \geq C_*\nu\rho_2'(0^+)\}$. We assume the existence of a sequence $\ell_n \to \infty$ such that

$$\mathbb{P}\bigg(\max_{\boldsymbol\psi\in\boldsymbol\Psi:\,|\boldsymbol\psi_{\mathcal{S}}-\boldsymbol\psi_{0,\mathcal{S}}|_\infty\leq c_n,\,|\boldsymbol\psi_{\mathcal{S}^c}|_1\leq\aleph_n} |\mathcal{M}^*_{\boldsymbol\psi}| \leq \ell_n\bigg) \to 1$$

with some $c_n \gg \phi_n$.

Let $b_{1,n} = \max\{a_n, r_1\aleph_n^2\}$ and $b_{2,n} = \max\{b_{1,n},\nu^2\}$. Then $\phi_n = \max\{pb_{1,n}^{1/2}, b_{2,n}^{1/2}\}$. Define $\boldsymbol\Psi_* = \{\boldsymbol\psi = (\boldsymbol\psi_{\mathcal{S}}^{\rm T}, \boldsymbol\psi_{\mathcal{S}^c}^{\rm T})^{\rm T} : |\boldsymbol\psi_{\mathcal{S}} - \boldsymbol\psi_{0,\mathcal{S}}|_\infty \leq \varepsilon, |\boldsymbol\psi_{\mathcal{S}^c}|_1 \leq \aleph_n\}$ for some fixed $\varepsilon > 0$. Consider

$$\hat{\boldsymbol\psi} = \arg\min_{\boldsymbol\psi\in\boldsymbol\Psi_*}\max_{\boldsymbol\lambda\in\hat\Lambda_n^{(\mathcal{T})}(\boldsymbol\psi)}\bigg[\frac{1}{n}\sum_{i=1}^n\log\{1+\boldsymbol\lambda^{\rm T}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol\psi)\} - \sum_{j\in\mathcal{D}}P_{2,\nu}(|\lambda_j|) + \sum_{k\in\mathcal{D}}P_{1,\pi}(|\xi_k|)\bigg]. \tag{3.7}$$

Proposition 3.1 shows that such a $\hat{\boldsymbol\psi}$ is a sparse local minimizer for (3.1).

*Proposition 3.1.* Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathscr{P}$ for $\mathscr{P}$ defined in (3.2), and $P_{2,\nu}(\cdot)$ be convex with bounded second derivative around 0. For $\hat{\boldsymbol\psi}$ defined as (3.7), assume there exists a constant $\tilde{c} \in (C_*,1)$ such that $\mathbb{P}[\cup_{j\in\mathcal{T}}\{|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol\psi})| \in [\tilde{c}\nu\rho_2'(0^+), \nu\rho_2'(0^+))\}] \to 0$. Under Conditions 1–4 and (3.6), if $\log r = o(n^{1/3})$, $s^2\ell_n\phi_n^2 = o(1)$, $b_{2,n} = o(n^{-2/\gamma})$, $\ell_n\aleph_n = o(\nu)$ and $\ell_n^{1/2}\aleph_n = o(\pi)$, then with probability approaching one this $\hat{\boldsymbol\psi} = (\hat{\boldsymbol\theta}^{\rm T}, \hat{\boldsymbol\xi}_{\mathcal{A}^c}^{\rm T}, \hat{\boldsymbol\xi}_{\mathcal{A}}^{\rm T})^{\rm T}$ provides a sparse local minimizer for the nonconvex optimization (3.1) such that (i) $|\hat{\boldsymbol\theta} - \boldsymbol\theta_0|_\infty = O_{\rm p}(b_{1,n}^{1/2})$, (ii) $|\hat{\boldsymbol\xi}_{\mathcal{A}^c} - \boldsymbol\xi_{0,\mathcal{A}^c}|_\infty = O_{\rm p}(\phi_n)$, and (iii) $\mathbb{P}(\hat{\boldsymbol\xi}_{\mathcal{A}} = \mathbf{0}) \to 1$ as $n \to \infty$.

In the rest of this section, we focus on the sparse local minimizer $\hat{\boldsymbol\psi}_{\rm PEL} = (\hat{\boldsymbol\theta}_{\rm PEL}^{\rm T}, \hat{\boldsymbol\xi}_{\rm PEL}^{\rm T})^{\rm T}$ specified in (3.7). We proceed with additional regularity conditions.

*Condition 5.* There exists a universal constant $K_5 > 1$ such that $K_5^{-1} < \lambda_{\min}\{\mathbf{Q}_{\mathcal{I}\cup\mathcal{B}_1,\mathcal{B}_2}^{(\mathcal{T})}\} \leq \lambda_{\max}\{\mathbf{Q}_{\mathcal{I}\cup\mathcal{B}_1,\mathcal{B}_2}^{(\mathcal{T})}\} < K_5$ for any $\mathcal{B}_2 \subset \mathcal{B}_1 \subset \mathcal{D}$ with $p + |\mathcal{B}_2| \leq r_1 + |\mathcal{B}_1| \leq \ell_n$, where $\mathbf{Q}_{\mathcal{I}\cup\mathcal{B}_1,\mathcal{B}_2}^{(\mathcal{T})} = ([\mathbb{E}\{\nabla_{\boldsymbol\psi_{\mathcal{P}\cup\mathcal{B}_2}}\mathbf{g}_{i,\mathcal{I}\cup\mathcal{B}_1}^{(\mathcal{T})}(\boldsymbol\psi_0)\}]^{\rm T})^{\otimes 2}$.

Condition 5 is the sparse Riesz condition (Zhang and Huang 2008; Chen and Chen 2008) in our setting to deal with the high-dimensional $\boldsymbol\psi$ when $p+r_2 > n$. Let $\hat{\boldsymbol\lambda}(\hat{\boldsymbol\psi}_{\rm PEL})$ be the $r\times 1$ vector

[3] If (3.6) is violated, the asymptotic normality of $\hat{\boldsymbol\theta}_{\rm PEL}$ in Theorem 3.1 will still hold under (3.4) along with more complicated notations to spell out the restrictions.

of Lagrange multiplier defined at $\hat{\boldsymbol{\psi}}_{\rm PEL}$:

$$\hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\rm PEL}) = \arg\max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})} \bigg[\frac{1}{n}\sum_{i=1}^n \log\{1+\boldsymbol{\lambda}^{\rm T}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})\} - \sum_{j\in\mathcal{D}} P_{2,\nu}(|\lambda_j|)\bigg].$$

Write $\hat{\boldsymbol{\lambda}} := \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\rm PEL}) = (\hat{\lambda}_1,\ldots,\hat{\lambda}_r)^{\rm T}$ and $\rho_2(t;\nu) = \nu^{-1}P_{2,\nu}(t)$. Let $\mathcal{R}_n = \mathcal{I}\cup\{j\in\mathcal{D}:\hat{\lambda}_j\neq 0\}$ be the set of the estimated binding moments. Since only those $\lambda_j$'s associated with $\mathcal{D}$ are penalized, it follows that

$$\frac{1}{n}\sum_{i=1}^n \frac{g_{i,j}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})}{1+\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}^{\rm T}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})} = \begin{cases} 0, & \text{if } j\in\mathcal{I},\\ \nu\rho_2'(|\hat{\lambda}_j|;\nu)\mathrm{sgn}(\hat{\lambda}_j), & \text{if } j\in\mathcal{D} \text{ with } \hat{\lambda}_j\neq 0.\end{cases}$$

For any unbinding moment $j\in\mathcal{R}_n^{\rm c}$, define

$$\hat{\eta}_j := \frac{1}{n}\sum_{i=1}^n \frac{g_{i,j}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})}{1+\hat{\boldsymbol{\lambda}}_{\mathcal{R}_n}^{\rm T}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})}. \tag{3.8}$$

If $\hat{\lambda}_j = 0$ for some $j\in\mathcal{D}$, the subdifferential at $\hat{\lambda}_j$ has to include the zero element (Bertsekas 1997). That is, $\hat{\eta}_j\in[-\nu\rho_2'(0^+),\nu\rho_2'(0^+)]$ for any $j\in\mathcal{R}_n^{\rm c}$. In our theoretical analysis, we impose the next condition.

*Condition 6.* $\mathbb{P}[\cup_{j\in\mathcal{R}_n^{\rm c}}\{|\hat{\eta}_j| = \nu\rho_2'(0^+)\}]\to 0$ as $n\to\infty$.

*Remark 1.* Condition 6 requires that $\hat{\eta}_j$ ($j\in\mathcal{R}_n^{\rm c}$) does not lie on the boundary with probability approaching one, which is realistic in practice. If the distribution function of $\hat{\eta}_j$ is continuous at $\pm\nu\rho_2'(0^+)$, we then have $\mathbb{P}\{|\hat{\eta}_j| = \nu\rho_2'(0^+)\} = 0$. Condition 6 makes sure that $\hat{\boldsymbol{\lambda}}(\boldsymbol{\psi})$ is continuously differentiable at $\hat{\boldsymbol{\psi}}_{\rm PEL}$ with probability approaching one. See Lemma A.5 in the supplementary materials.

Define $\mathcal{A}_* = \{j\in\mathcal{A}:\hat{\lambda}_j\neq 0\}$ and $\mathcal{A}_{*,{\rm c}} = \{j\in\mathcal{A}^{\rm c}:\hat{\lambda}_j\neq 0\}$. Let $\mathcal{I}^* = \mathcal{I}\cup\mathcal{A}_*$, and then $\mathcal{R}_n$ can be decomposed into two disjoint parts $\mathcal{I}^*$ and $\mathcal{A}_{*,{\rm c}}$. Furthermore, we define $\mathcal{S}_* = \mathcal{P}\cup\mathcal{A}_{*,{\rm c}}$. The fact $\mathcal{S}_*\subset\mathcal{S}$ implies that $|\mathcal{S}_*| = p+|\mathcal{A}_{*,{\rm c}}|\le|\mathcal{S}| = s$. For any $\boldsymbol{\psi}\in\boldsymbol{\Psi}$, we have $\boldsymbol{\psi}_{\mathcal{S}_*} = (\boldsymbol{\theta}^{\rm T},\boldsymbol{\xi}_{\mathcal{A}_{*,{\rm c}}}^{\rm T})^{\rm T}$. Define

$$\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})} = \big([\mathbb{E}\{\nabla_{\boldsymbol{\theta}}\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}]^{\rm T}\{\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}^{-1/2}\big)^{\otimes 2}$$

with $\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0) = \mathbb{E}\{\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)^{\otimes 2}\}$, and $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n} = \{\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}\}^{-1}[\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\rm T}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1}\hat{\boldsymbol{\eta}}_{\mathcal{R}_n}$, where $\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\psi}_{\mathcal{S}_*}}\mathbf{g}_{i,\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\rm T}\{\mathbf{V}_{\mathcal{R}_n}^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}^{-1/2})^{\otimes 2}$, $\hat{\boldsymbol{\eta}} = (\hat{\eta}_j)_{j\in\mathcal{T}}$ with $\hat{\eta}_j = 0$ for $j\in\mathcal{I}$, $\hat{\eta}_j = \nu\rho_2'(|\hat{\lambda}_j|;\nu)\mathrm{sgn}(\hat{\lambda}_j)$ for $j\in\mathcal{D}$ with $\hat{\lambda}_j\neq 0$, and $\hat{\eta}_j$ specified in (3.8) for $j\in\mathcal{R}_n^{\rm c}$. The limiting distribution of $\hat{\boldsymbol{\theta}}_{\rm PEL}$ is stated in Theorem 3.1.

*Theorem 3.1.* Assume the conditions of Proposition 3.1 hold. Under Conditions 5 and 6, if $\ell_n^{3/2}\log r = o(n^{1/2-1/\gamma})$ and $\ell_n n^{1/2}s^{3/2}\phi_n\nu = o(1)$, then $n^{1/2}\boldsymbol{\alpha}^{\rm T}\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{1/2}\{\hat{\boldsymbol{\theta}}_{\rm PEL}-\boldsymbol{\theta}_0-\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}\}\xrightarrow{d}\mathcal{N}(0,1)$ as $n\to\infty$ for any $\boldsymbol{\alpha}\in\mathbb{R}^p$ with $|\boldsymbol{\alpha}|_2 = 1$, where $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}$ is the first $p$ components of $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n}$ and $(a_n,\phi_n)$ are given in (3.3).

*Remark 2.* Notice that $a_n\lesssim s\pi$. To satisfy the restrictions in Theorem 3.1, it is sufficient for $(n,r,p,\ell_n,s)$ to have: $r = o\{\exp(n^{1/3})\}$, $\ell_n = o\{n^{(\gamma-2)/(3\gamma)}(\log r)^{-2/3}\}$, $\ell_n^{5/2}s^{3/2}(\log r)\max\{p,\ell_n^{1/2}\} = o(n^{1/2})$, the tuning parameters $\nu$ and $\pi$ satisfying $\ell_n^{1/2}\aleph_n\ll\pi\ll\min\{s^{-1}n^{-2/\gamma},\ell_n^{-4}s^{-4}p^{-2}(\log r)^{-1}\}$, $\ell_n\aleph_n\ll\nu\ll\min\{(\ell_n^3s^3p^2\log r)^{-1/2},n^{-1/\gamma},(\ell_n^2ns^3)^{-1/4}\}$ and $\nu^2\pi\ll(\ell_n^2ns^4p^2)^{-1}$.

Thanks to the extra valid moments in $\mathbf{g}^{(\mathcal{D})}$, the asymptotic covariance of $\hat{\boldsymbol{\theta}}_{\rm PEL}$ is smaller than that of $\hat{\boldsymbol{\theta}}_{\rm EL}^{(\mathcal{I})}$ which uses the information in $\mathbf{g}^{(\mathcal{I})}$ only. On the other hand, given that $\hat{\boldsymbol{\theta}}_{\rm EL}^{(\mathcal{I})}$ provides asymptotic normality under the set of correctly specified moments $\mathcal{I}$, one may question whether it is worthwhile to consider all the available moments among which some may risk misspecification. This is a tradeoff between robustness and efficiency, a recurrent theme in modern econometrics. Although it inevitably depends on the data generating process (DGP), in big data environments the efficiency gain may be substantial. Benefits are demonstrated in the simulation exercises and the empirical application via very simple econometric models.[4]

*Remark 3.* The setting with the sets $\mathcal{I}$ and $\mathcal{D} = \mathcal{A}\cup\mathcal{A}^{\rm c}$ is the same as Liao (2013). When estimating (2.3) with low-dimensional moments based on GMM, Cheng and Liao (2015) require the number of estimating functions $r = o(n^{1/3})$, and the signal of invalid estimating functions $\min_{k\in\mathcal{A}^{\rm c}}|\mathbb{E}\{g_{i,k}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}|\gg(rn^{-1})^{1/2}$. Our conditions on the relative size of the dimensions are more general. A sufficient condition for (3.6) is $\min_{k\in\mathcal{A}^{\rm c}}|\mathbb{E}\{g_{i,k}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}|\gg\max\{\nu,p(s\pi)^{1/2},pr_1^{1/2}\aleph_n\}$, in view of $a_n\lesssim s\pi$. Under the restrictions discussed in Remark 2, if we select $\nu$ and $\pi$ sufficiently close to $\ell_n\aleph_n$ and $\ell_n^{1/2}\aleph_n$, respectively, then (3.6) holds provided that $\min_{k\in\mathcal{A}^{\rm c}}|\mathbb{E}\{g_{i,k}^{(\mathcal{D})}(\boldsymbol{\theta}_0)\}|\gg ps^{1/2}\ell_n^{1/4}\aleph_n^{1/2}$. This is similar to the "beta-min condition" which is necessary for consistent variable selection by shrinkage methods (Bühlmann 2013).

*Remark 4.* Under low-dimensional moments the asymptotic normality involves no bias; see Liao (2013, Theorem 3.5) and Cheng and Liao (2015, Theorem 3.3). In contrast, Theorem 3.1 here makes clear that the high-dimensional moments incur an additional asymptotic bias $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}$. To conduct hypothesis testing or construct confidence region about $\boldsymbol{\theta}$, in principle we can estimate and correct the bias $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}$. Such a direct bias correction approach, nevertheless, is undesirable in practice and in theory. The bias term involves multiplication and inverse of large matrices, which are difficult to estimate with precision in finite samples. Illustrated in our simulation experiments in Section 6, we are unsatisfied with the asymptotic normality approximation after estimating and correcting the bias. Thus, we view Theorem 3.1 as a characterization of the asymptotic behavior of $\hat{\boldsymbol{\theta}}_{\rm PEL}$, but we do not encourage carrying out statistical inference based on it. Instead, Section 5 recommends PPEL for inference.

[4] See the RMSEs in Table 2, the length of confidence intervals in Figure 2, and the standard deviations in Table 4.

While the augmented parameter $\boldsymbol{\xi}$ essentially validates all moments in $\mathcal{D}$, the efficiency gain comes from the penalty on $\boldsymbol{\xi}$ that shrinks some $\hat{\xi}_k$, $k \in \mathcal{A}$, to zero, thereby confirming the validity of these associated moments. To discuss moment selection, we write $\hat{\boldsymbol{\xi}}_{\rm PEL} = (\hat{\xi}_k)_{k\in\mathcal{D}}$. If all valid estimating functions in $\mathcal{A}$ are selected by the optimization (3.1), that is, $\hat{\xi}_k = 0$ for all $k \in \mathcal{A}$, then the asymptotic covariance of $\hat{\boldsymbol{\theta}}_{\rm PEL}$ is $\{\mathbf{J}^{(\mathcal{H})}\}^{-1}$.

Remind that $\{\mathbf{J}^{(\mathcal{H})}\}^{-1}$ in Proposition 2.1 is the semiparametric efficiency bound for the estimation of $\boldsymbol{\theta}_0$ under the oracle, and Proposition 3.1 shows that $\mathbb{P}(\hat{\boldsymbol{\xi}}_{{\rm PEL},\mathcal{A}} = \mathbf{0}) \to 1$ as $n \to \infty$, which provides a natural moment selection criterion

$$\hat{\mathcal{A}} = \{k \in \mathcal{D} : \hat{\xi}_k = 0\} \tag{3.9}$$

to identify the valid estimating functions in $\mathcal{A}$. Notice that Proposition 3.1 also indicates that $|\hat{\boldsymbol{\xi}}_{{\rm PEL},\mathcal{A}^{\rm c}} - \boldsymbol{\xi}_{0,\mathcal{A}^{\rm c}}|_\infty = O_{\rm p}(\phi_n)$. Together with (3.6), we have $\mathbb{P}[\cup_{k\in\mathcal{A}^{\rm c}}\{\hat{\xi}_k = 0\}] \to 0$ as $n \to \infty$. Based on these arguments, Theorem 3.2 supports our proposed moments selection criterion.

*Theorem 3.2.* Under the conditions of Proposition 3.1, it holds that $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ as $n \to \infty$.

Up to now, we have established the asymptotic properties of the PEL estimator for the moment-defined model (2.3) with high-dimensional moments and low-dimensional $\boldsymbol{\theta}$. The next section further extends the estimation to a high-dimensional structural parameter.

## 4. High-Dimensional Moments with High-Dimensional $\theta$

A high-dimensional parameter $\boldsymbol{\theta}$ is present when many control variables are included in the structural model. For example, the leading case of the linear IV model takes only one endogenous variable and it is accompanied by a few corresponding IVs. Nevertheless, such a simple setting may still involve many exogenous control variables in the main equation, and these control variables are natural instruments for themselves. An empirical example can be found in Blundell, Pashardes, and Weber (1993). Fan and Liao (2014) propose the *focused GMM* for such a linear IV model, which is a special case of the moment-defined model.

In this section, $p$ and $r_2$ are both allowed to be much larger than $n$, whereas $r_1$ is fixed or diverges at some slow polynomial rate of $n$. We assume $\boldsymbol{\theta}_0$ sparse in the sense that most of its components are zero. Write $\boldsymbol{\theta}_0 = (\theta_{0,l})_{l\in\mathcal{P}}$ and define the active set $\mathcal{P}_\sharp = \{l \in \mathcal{P} : \theta_{0,l} \neq 0\}$ with cardinality $p_\sharp = |\mathcal{P}_\sharp|$. To obtain a sparse estimate of $\boldsymbol{\theta}_0$, we add to (3.1) a penalty on $\boldsymbol{\theta}$ to produce the following optimization problem:

$$\begin{aligned}(\hat{\boldsymbol{\theta}}_{\rm PEL}^{\rm T}, \hat{\boldsymbol{\xi}}_{\rm PEL}^{\rm T})^{\rm T} = \arg\min_{\boldsymbol{\psi}\in\boldsymbol{\Psi}} \max_{\boldsymbol{\lambda}\in\hat{\Lambda}_n^{(\mathcal{T})}(\boldsymbol{\psi})} \Bigg[\frac{1}{n}\sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\rm T}\mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi})\} \\ - \sum_{j\in\mathcal{D}} P_{2,\nu}(|\lambda_j|) + \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_k|) \\ + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_l|)\Bigg].\end{aligned} \tag{4.1}$$

With slight abuse of notation, we keep using $(\hat{\boldsymbol{\theta}}_{\rm PEL}^{\rm T}, \hat{\boldsymbol{\xi}}_{\rm PEL}^{\rm T})^{\rm T}$ to denote the solution to (4.1).

As $p$ can be bigger than $r_1$ in high dimension, here we update Condition 1 for the identification of $\boldsymbol{\theta}_0$. In view of Chang, Tang, and Wu (2018), we impose Condition 1′ for the identification of the nonzero components of $\boldsymbol{\theta}_0$.

*Condition 1′.* (i) There exists a universal constant $K_1' > 0$ such that

$$\inf_{\boldsymbol{\theta}=(\boldsymbol{\theta}_{\mathcal{P}_\sharp}^{\rm T},\boldsymbol{\theta}_{\mathcal{P}_\sharp^{\rm c}}^{\rm T})^{\rm T}\in\boldsymbol{\Theta}:\,|\boldsymbol{\theta}_{\mathcal{P}_\sharp}-\boldsymbol{\theta}_{0,\mathcal{P}_\sharp}|_\infty>\varepsilon,\,\boldsymbol{\theta}_{\mathcal{P}_\sharp^{\rm c}}=\mathbf{0}} |\mathbb{E}\{\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}|_\infty \ge K_1'\varepsilon$$

for any $\varepsilon > 0$. (ii) There exists a universal constant $K_3' > 0$ such that $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} |\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{P}_\sharp^{\rm c}}}\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta})\}|_\infty \le K_3'$.

As we have shown in Section 3, the index set $\mathcal{S}$ and the quantities $a_n$ and $\phi_n$ play key roles in the theoretical analysis of the low-dimensional $\boldsymbol{\theta}$ in (3.1). Under the current high-dimensional setting, we define

$$\begin{aligned}a_n = \sum_{k\in\mathcal{D}} P_{1,\pi}(|\xi_{0,k}|) + \sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_{0,l}|) \ \text{ and} \\ \mathcal{S} = \mathcal{P}_\sharp \cup \mathcal{A}^{\rm c} \ \text{ with } s = |\mathcal{S}|\,.\end{aligned} \tag{4.2}$$

This $\mathcal{S}$ indexes all nonzero components in the true augmented parameter $\boldsymbol{\psi}_0$. Compared to its counterpart $a_n$ in Section 3, the newly defined $a_n$ here is amended with an extra term $\sum_{l\in\mathcal{P}} P_{1,\pi}(|\theta_{0,l}|)$ due to the penalty imposed on $\boldsymbol{\theta}$. In the current high-dimensional setting, we also redefine

$$\phi_n = \max\{p_\sharp a_n^{1/2}, p_\sharp r_1^{1/2}\aleph_n, \nu\} \tag{4.3}$$

with $a_n$ in (4.2). To control the bias introduced by $P_{1,\pi}(\cdot)$ on $\hat{\boldsymbol{\theta}}_{\rm PEL}$ and $\hat{\boldsymbol{\xi}}_{\rm PEL}$, similar to (3.6), we assume that among the elements in $\boldsymbol{\psi}_0 = (\psi_{0,1}, \ldots, \psi_{0,p+r_2})^{\rm T}$, the minimal active signal is

$$\min_{k\in\mathcal{S}} |\psi_{0,k}| \gg \max\{\phi_n, \pi\} \tag{4.4}$$

with $\phi_n$ in (4.3) and $\max_{k\in\mathcal{S}} \sup_{c|\psi_{0,k}|<t<c^{-1}|\psi_{0,k}|} P_{1,\pi}'(t) = 0$ for some constant $c \in (0, 1)$.[5] Given the newly defined $\mathcal{S}$ and $\phi_n$, we can further update $\ell_n$, $\mathcal{R}_n$, $\mathcal{A}_*$, $\mathcal{A}_{*,{\rm c}}$ and $\mathcal{I}^*$ in the same manner as their counterparts in Section 3. Moreover, we also define $\mathbf{J}_{\mathcal{R}_n}^{(\mathcal{T})}$ and $\hat{\boldsymbol{\xi}}_{\mathcal{R}_n}$ as specified in Section 3 with $\mathcal{S}_* = \mathcal{P}_\sharp \cup \mathcal{A}_{*,{\rm c}}$. For any $\boldsymbol{\psi} \in \boldsymbol{\Psi}$, we have $\boldsymbol{\psi}_{\mathcal{S}_*} = (\boldsymbol{\theta}_{\mathcal{P}_\sharp}^{\rm T}, \boldsymbol{\xi}_{\mathcal{A}_{*,{\rm c}}}^{\rm T})^{\rm T}$.

Proposition A.2 in the supplementary materials shows that there exists a sparse local minimizer $(\hat{\boldsymbol{\theta}}_{\rm PEL}^{\rm T}, \hat{\boldsymbol{\xi}}_{\rm PEL}^{\rm T})^{\rm T} \in \boldsymbol{\Psi}$ for the nonconvex optimization (4.1) such that $\mathbb{P}(\hat{\boldsymbol{\theta}}_{{\rm PEL},\mathcal{P}_\sharp^{\rm c}} = \mathbf{0}) \to 1$ as $n \to \infty$, which means all zero components of $\boldsymbol{\theta}_0$ can be estimated exactly as zero with high probability. The limiting distribution of such $\hat{\boldsymbol{\theta}}_{{\rm PEL},\mathcal{P}_\sharp}$ and the moment selection outcomes are stated in Theorem 4.1.

*Theorem 4.1.* Let $P_{1,\pi}(\cdot), P_{2,\nu}(\cdot) \in \mathscr{P}$ for $\mathscr{P}$ defined in (3.2), and $P_{2,\nu}(\cdot)$ be convex with bounded second derivative around 0. For the sparse local minimizer $\hat{\boldsymbol{\psi}}_{\rm PEL} = (\hat{\boldsymbol{\theta}}_{\rm PEL}^{\rm T}, \hat{\boldsymbol{\xi}}_{\rm PEL}^{\rm T})^{\rm T}$ for

[5] See the arguments below (3.5) for the validity of these assumptions.

(4.1) specified in Proposition A.2 in the supplementary materials, assume there exists a constant $\tilde{c} \in (C_*, 1)$ such that $\mathbb{P}[\cup_{j\in\mathcal{T}}\{|\bar{g}_j^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\rm PEL})| \in [\tilde{c}\nu\rho_2'(0^+), \nu\rho_2'(0^+))\}] \to 0$. Suppose Conditions 1′, 2–4 and (4.4) hold. Furthermore, assume Condition 5 holds with the newly defined $\ell_n$, replacing $\mathcal{P}$ by $\mathcal{P}_\sharp$ and replacing $p$ by $p_\sharp$, and Condition 6 holds with the newly defined $\mathcal{R}_n$. If $\log r = o(n^{1/3})$, $\max\{a_n, \nu^2\} = o(n^{-2/\gamma})$, $\ell_n^{3/2}\log r = o(n^{1/2-1/\gamma})$, $\ell_n n^{1/2}s^{3/2}\phi_n\nu = o(1)$ and $\ell_n\aleph_n = o(\min\{\nu,\pi\})$, then we have

$$n^{1/2}\boldsymbol{\alpha}^{\rm T}\{\mathbf{W}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{1/2}\{\hat{\boldsymbol{\theta}}_{{\rm PEL},\mathcal{P}_\sharp} - \boldsymbol{\theta}_{0,\mathcal{P}_\sharp} - \hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}\} \xrightarrow{d} \mathcal{N}(0,1)$$

for any $\boldsymbol{\alpha} \in \mathbb{R}^{p_\sharp}$ with $|\boldsymbol{\alpha}|_2 = 1$ as $n \to \infty$, where $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n,(1)}$ is the first $p_\sharp$ components of $\hat{\boldsymbol{\zeta}}_{\mathcal{R}_n}$, $(a_n, \phi_n)$ are given by (4.2) and (4.3), respectively, and $\mathbf{W}_{\mathcal{I}^*}^{(\mathcal{T})} = ([\mathbb{E}\{\nabla_{\boldsymbol{\theta}_{\mathcal{P}_\sharp}}\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}]^{\rm T}\{\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)\}^{-1/2})^{\otimes 2}$, with $\mathbf{V}_{\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0) = \mathbb{E}\{\mathbf{g}_{i,\mathcal{I}^*}^{(\mathcal{T})}(\boldsymbol{\theta}_0)^{\otimes 2}\}$. Moreover, under the conditions of Proposition A.2 in the supplementary materials, it holds that $\mathbb{P}(\hat{\mathcal{A}} = \mathcal{A}) \to 1$ as $n \to \infty$, where $\hat{\mathcal{A}}$ is specified in (3.9).

*Remark 5.* Instead of assuming $\ell_n^{1/2}\aleph_n = o(\pi)$ as in Theorem 3.1, here Theorem 4.1 strengthens $\ell_n\aleph_n = o(\pi)$. As shown in Section A.5 of the supplementary materials, this stronger condition guarantees that $\boldsymbol{\psi}_{0,\mathcal{S}^{\rm c}}$, the zero components of $\boldsymbol{\psi}_0$, can be shrunk to zero with probability approaching one in the high-dimensional $\boldsymbol{\theta}$ setting.

Compared to (3.1), a penalty on $\boldsymbol{\theta}$ is added to (4.1). To appreciate the benefit from this additional penalty on the sparse parameter, without loss of generality, we write $\boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{0,\mathcal{P}_\sharp}^{\rm T}, \mathbf{0}^{\rm T})^{\rm T}$ and the inverse of $\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}$ in Theorem 3.1 in the following compatible partitioned matrix

$$\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1} = \begin{pmatrix} [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{11} & [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{12} \\ [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{21} & [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{22} \end{pmatrix},$$

where $[\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{11}$ is a $p_\sharp \times p_\sharp$ matrix. The asymptotic covariance of $\hat{\boldsymbol{\theta}}_{{\rm PEL},\mathcal{P}_\sharp}$ by (3.1) is $[\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{11}$, and that in (4.1) is $\{\mathbf{W}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}$ according to Theorem 4.1. Notice that $\{\mathbf{W}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1} = [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{11} - [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{12}[\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{22}^{-1}[\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{21} \le [\{\mathbf{J}_{\mathcal{I}^*}^{(\mathcal{T})}\}^{-1}]_{11}$ by the inverse of a partitioned matrix. Implicitly restricting the parameter space, the penalty on $\boldsymbol{\theta}$ gains efficiency for the estimation of $\boldsymbol{\theta}_{0,\mathcal{P}_\sharp}$, the nonzero components of $\boldsymbol{\theta}_0$.

*Remark 6.* Theorem 4.1 spells out the limiting distribution and efficiency gain for the estimator of the nonzero components in $\boldsymbol{\theta}_0$. If one is interested in some coefficient in $\mathcal{P}_\sharp^{\rm c}$, we have consistency in that $\mathbb{P}(\hat{\boldsymbol{\theta}}_{{\rm PEL},\mathcal{P}_\sharp^{\rm c}} = \mathbf{0}) \to 1$ as $n \to \infty$, but the limiting distribution is irregular due to the shrinkage.[6] Furthermore, parallel to Theorem 3.1 the asymptotic bias is again present. Similar to Remark 4, we do not suggest conducting statistical inference predicated on this characterization of the asymptotic behavior of normality.

[6]This is a generic property shared by procedures of the oracle properties, for example SCAD (Fan and Li 2001) and the adaptive Lasso (Zou 2006).

Statistical inference is important in applied econometrics when researchers intend to assess whether the estimated result supports or rejects a hypothesized value of the parameter. The next section proposes an inferential procedure based on a projection of estimating functions, which is free of asymptotic biases.

## 5. Confidence Regions for a Subset of $\boldsymbol{\psi}$

Suppose we are interested in the inference for a subset of parameters $\boldsymbol{\psi}_{\mathcal{M}}$ for some generic small subset $\mathcal{M} \subset (\mathcal{P} \cup \mathcal{D})$ with $|\mathcal{M}| = m$. Here we allow $m$ to be fixed or diverge slowly with the sample size $n$. What is novel here is that $\boldsymbol{\psi}_{\mathcal{M}}$ is allowed to contain part of the auxiliary parameter $\boldsymbol{\xi}$, for which our method will provide a formal statistical inference for the validity of a subset of the high-dimensional moment restrictions. In contrast, Liao (2013) offers asymptotic normality for the structural parameter $\boldsymbol{\theta}$ under low-dimensional moments but does not characterize the asymptotic distribution of the auxiliary parameter $\boldsymbol{\xi}$.

A key technical issue is how to deal with the high-dimensional nuisance parameter $\boldsymbol{\psi}_{\mathcal{M}^{\rm c}}$, where $\mathcal{M}^{\rm c} = (\mathcal{P} \cup \mathcal{D})\backslash\mathcal{M}$. Our approach follows Chang et al. (2021) to project out the influence of the nuisance parameter. Such idea was also used in Ning and Liu (2017) and Neykov et al. (2018). Given an initial estimate $\boldsymbol{\psi}^*$ for $\boldsymbol{\psi}_0$, we first determine a linear transformation matrix $\mathbf{A}_n = (\mathbf{a}_k^n)_{k\in\mathcal{M}}^{\rm T} \in \mathbb{R}^{m\times r}$ with each row defined as

$$\mathbf{a}_k^n = \arg\min_{\mathbf{u}\in\mathbb{R}^r} |\mathbf{u}|_1 \quad \text{s.t.} \quad |\{\nabla_{\boldsymbol{\psi}}\bar{\mathbf{g}}^{(\mathcal{T})}(\boldsymbol{\psi}^*)\}^{\rm T}\mathbf{u} - \boldsymbol{\gamma}_k|_\infty \le \varsigma \tag{5.1}$$

for $k \in \mathcal{M}$, where $\varsigma \to 0$ as $n \to \infty$ is a tuning parameter, and $\{\boldsymbol{\gamma}_k\}_{k\in\mathcal{M}}$ is a basis of the linear space $\{\mathbf{b} = (b_j)_{j\in\mathcal{P}\cup\mathcal{D}} : \mathbf{b}_{\mathcal{M}^{\rm c}} = \mathbf{0}\}$. In practice, we can specify the $(p + r_2)$-dimensional vector $\boldsymbol{\gamma}_k = (1(j = k))_{j\in\mathcal{P}\cup\mathcal{D}}$ with all zero elements except one unit entry. As we will discuss in Remark 8, the solution from (3.1) or (4.1) can serve as the initial estimator $\boldsymbol{\psi}^*$.

Based on $\mathbf{A}_n$ as in (5.1), we then obtain the new $m$-dimensional estimating functions $\mathbf{f}^{\mathbf{A}_n}(\cdot;\cdot)$ by projecting $\mathbf{g}^{(\mathcal{T})}(\cdot;\cdot)$ on $\mathbf{A}_n$:

$$\mathbf{f}^{\mathbf{A}_n}(\mathbf{X};\boldsymbol{\psi}) = \mathbf{A}_n\mathbf{g}^{(\mathcal{T})}(\mathbf{X};\boldsymbol{\psi}).$$

Write $\boldsymbol{\gamma}_k = (\gamma_{k,j})_{j\in\mathcal{P}\cup\mathcal{D}}$ and define an $m \times (p + r_2)$ matrix $\boldsymbol{\Gamma} = (\gamma_{k,j})_{k\in\mathcal{M},j\in\mathcal{P}\cup\mathcal{D}}$. The definition of $\mathbf{A}_n$ implies that $|\nabla_{\boldsymbol{\psi}}\bar{\mathbf{f}}^{\mathbf{A}_n}(\boldsymbol{\psi}^*) - \boldsymbol{\Gamma}|_\infty \le \varsigma$. Since all the components in the $j$th column of $\boldsymbol{\Gamma}$ are zero for $j \in \mathcal{M}^{\rm c}$, the newly defined estimating functions $\mathbf{f}^{\mathbf{A}_n}$ is uninformative of the nuisance parameter $\boldsymbol{\psi}_{\mathcal{M}^{\rm c}}$. Informative is $\mathbf{f}^{\mathbf{A}_n}$ of the parameter $\boldsymbol{\psi}_{\mathcal{M}}$ of interest, as for $j \in \mathcal{M}$ some components in the $j$th column of $\boldsymbol{\Gamma}$ must be nonzero.

Given the projected estimating functions $\mathbf{f}^{\mathbf{A}_n}$, it is possible to directly borrow from Chang et al. (2021) to construct the confidence region of $\boldsymbol{\psi}_{\mathcal{M}}$ by the EL ratio

$$w_n(\boldsymbol{\psi}_{\mathcal{M}}) = 2\max_{\boldsymbol{\lambda}\in\tilde{\Lambda}_n(\boldsymbol{\psi}_{\mathcal{M}})}\sum_{i=1}^n \log\{1 + \boldsymbol{\lambda}^{\rm T}\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{\mathcal{M}}, \boldsymbol{\psi}_{\mathcal{M}^{\rm c}}^*)\}$$

with respect to $\boldsymbol{\psi}_{\mathcal{M}}$, where $\tilde{\Lambda}_n(\boldsymbol{\psi}_{\mathcal{M}}) = \{\boldsymbol{\lambda} \in \mathbb{R}^m : \boldsymbol{\lambda}^{\rm T}\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{\mathcal{M}}, \boldsymbol{\psi}_{\mathcal{M}^{\rm c}}^*) \in \mathcal{V}$ for any $i \in [n]\}$. Since $w_n(\boldsymbol{\psi}_{0,\mathcal{M}}) \xrightarrow{d} \chi_m^2$ as $n \to \infty$ for a fixed $m$, the set $\{\boldsymbol{\psi}_{\mathcal{M}} \in \mathbb{R}^m :$

$w_n(\boldsymbol{\psi}_{\mathcal{M}}) \leq \chi^2_{m,1-\alpha}\}$ provides a $100(1-\alpha)\%$ confidence region for $\boldsymbol{\psi}_{\mathcal{M}}$, where $\chi^2_{m,1-\alpha}$ is the $(1-\alpha)$-quantile of $\chi^2_m$ distribution. Nonetheless, the finite-sample performance of such an asymptotically valid confidence region depends crucially on the convexity of $w_n(\boldsymbol{\psi}_{\mathcal{M}})$ near $\boldsymbol{\psi}_{0,\mathcal{M}}$. If convexity fails in a finite sample, this EL-ratio-based confidence region will be voluminous in magnitude due to numerical instability. Verifying the convexity condition is onerous, in particular when $m$ is large, for $\mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})$ may be a nonlinear function of $\boldsymbol{\psi}_{\mathcal{M}}$.

To secure a stable confidence region for $\boldsymbol{\psi}_{\mathcal{M}}$, in this article we deviate from Chang et al. (2021) and instead recommend a re-estimation procedure for a confidence region based on the asymptotic normality of the estimator. Let

$$\tilde{\boldsymbol{\psi}}_{\mathcal{M}} = \arg\min_{\boldsymbol{\psi}_{\mathcal{M}} \in \boldsymbol{\Psi}^*_{\mathcal{M}}} \max_{\boldsymbol{\lambda} \in \tilde{\Lambda}_n(\boldsymbol{\psi}_{\mathcal{M}})} \sum_{i=1}^{n} \log\{1 + \boldsymbol{\lambda}^{\mathrm{T}} \mathbf{f}_i^{\mathbf{A}_n}(\boldsymbol{\psi}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})\},$$

where $\boldsymbol{\Psi}^*_{\mathcal{M}} = \{\boldsymbol{\psi}_{\mathcal{M}} : |\boldsymbol{\psi}_{\mathcal{M}} - \boldsymbol{\psi}^*_{\mathcal{M}}|_1 \leq O_{\mathrm{p}}(\varpi_{1,n})\}$ for some $\varpi_{1,n} \to 0$ such that $|\boldsymbol{\psi}^*_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}|_1 = O_{\mathrm{p}}(\varpi_{1,n})$. Since $\boldsymbol{\psi}^*_{\mathcal{M}}$ is an initial consistent estimator of $\boldsymbol{\psi}_{0,\mathcal{M}}$, it is reasonable to search for $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$ in a small neighborhood of $\boldsymbol{\psi}^*_{\mathcal{M}}$. We will specify $\varpi_{1,n}$ for some specific choices of $\boldsymbol{\psi}^*_{\mathcal{M}}$ in Remark 8.

To derive the limiting distribution of $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$, we assume the following condition.

*Condition 7.* For each $k \in \mathcal{M}$, there is a nonrandom $\mathbf{a}_k$ satisfying $[\mathbb{E}\{\nabla_{\boldsymbol{\psi}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}]^{\mathrm{T}} \mathbf{a}_k = \boldsymbol{\gamma}_k$, $|\mathbf{a}_k|_1 \leq K_6$ for some universal constant $K_6 > 0$, and $\max_{k \in \mathcal{M}} |\mathbf{a}_k^n - \mathbf{a}_k|_1 = O_{\mathrm{p}}(\omega_n)$ for some $\omega_n \to 0$. Let $\mathbf{A} = (\mathbf{a}_k)^{\mathrm{T}}_{k \in \mathcal{M}} \in \mathbb{R}^{m \times r}$. The eigenvalues of $\mathbf{A}^{\otimes 2}$ are uniformly bounded away from zero and infinity.

*Remark 7.* Let $\boldsymbol{\Xi} = \mathbb{E}\{\nabla_{\boldsymbol{\psi}} \mathbf{g}_i^{(\mathcal{T})}(\boldsymbol{\psi}_0)\}$ and $\widehat{\boldsymbol{\Xi}} = \nabla_{\boldsymbol{\psi}} \bar{\mathbf{g}}^{(\mathcal{T})}(\boldsymbol{\psi}^*)$. It follows from the existence of $\mathbf{a}_k$ that $\boldsymbol{\gamma}_k = \widehat{\boldsymbol{\Xi}}^{\mathrm{T}} \mathbf{a}_k + (\boldsymbol{\Xi} - \widehat{\boldsymbol{\Xi}})^{\mathrm{T}} \mathbf{a}_k = \widehat{\boldsymbol{\Xi}}^{\mathrm{T}} \mathbf{a}_k + \boldsymbol{\varepsilon}_k$, where $\boldsymbol{\varepsilon}_k = (\boldsymbol{\Xi} - \widehat{\boldsymbol{\Xi}})^{\mathrm{T}} \mathbf{a}_k$. Some mild conditions ensure $|\widehat{\boldsymbol{\Xi}} - \boldsymbol{\Xi}|_\infty = o_{\mathrm{p}}(1)$. This, together with the assumption $|\mathbf{a}_k|_1 \leq K_6$, implies that $\boldsymbol{\varepsilon}_k$ is stochastically small uniformly over all the components such that $|\boldsymbol{\varepsilon}_k|_\infty = o_{\mathrm{p}}(1)$, which can be viewed as an attempt to recover a nonrandom $\mathbf{a}_k$ with no noise asymptotically (Candes and Tao 2007; Bickel, Ritov, and Tsybakov 2009). It follows that $|\mathbf{a}_k^n - \mathbf{a}_k|_1 = o_{\mathrm{p}}(1)$ if $\widehat{\boldsymbol{\Xi}}$ satisfies the routine conditions for sparse signal recovering. Furthermore, the constant $K_6$ in Condition 7 may be replaced by some diverging $\varphi_n$ and our main results remain valid.

Theorem 5.1 gives the limiting distribution of $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$ with a generic initial estimator $\boldsymbol{\psi}^*$.

*Theorem 5.1.* Let $|\boldsymbol{\psi}^*_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}|_1 = O_{\mathrm{p}}(\varpi_{1,n})$, $|\boldsymbol{\psi}^*_{\mathcal{M}^c} - \boldsymbol{\psi}_{0,\mathcal{M}^c}|_1 = O_{\mathrm{p}}(\varpi_{2,n})$ for some $\varpi_{1,n} \to 0$ and $\varpi_{2,n} \to 0$. Under Conditions 2–4 and 7, if $m = o(n^{\min\{1/5,(\gamma-2)/(3\gamma)\}})$, $m\omega_n^2(m^2 + \log r) = o(1)$, $m\varpi_{1,n} = o(1)$ and $nm\varpi_{2,n}^2(\varsigma^2 + \varpi_{1,n}^2 + \varpi_{2,n}^2) = o(1)$, then $n^{1/2}\boldsymbol{\alpha}^{\mathrm{T}}(\hat{\mathbf{J}}^*)^{1/2}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}} - \boldsymbol{\psi}_{0,\mathcal{M}}) \xrightarrow{d} \mathcal{N}(0,1)$ as $n \to \infty$ for any $\boldsymbol{\alpha} \in \mathbb{R}^m$ with $|\boldsymbol{\alpha}|_2 = 1$, where $\hat{\mathbf{J}}^* = [\{\nabla_{\boldsymbol{\psi}_{\mathcal{M}}} \bar{\mathbf{f}}^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})\}^{\mathrm{T}} \{\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})\}^{-1/2}]^{\otimes 2}$ with $\widehat{\mathbf{V}}_{\mathbf{f}^{\mathbf{A}_n}}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c}) = \mathbb{E}_n\{\mathbf{f}_i^{\mathbf{A}_n}(\tilde{\boldsymbol{\psi}}_{\mathcal{M}}, \boldsymbol{\psi}^*_{\mathcal{M}^c})^{\otimes 2}\}$.

The above theorem is stated for $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$. It includes the estimation of $\boldsymbol{\theta}_{\mathcal{M}}$ as a special case if one's research interest falls on the main parameter for economic interpretation, and makes it possible to infer the validity of a subset of moments in view of selecting $\boldsymbol{\psi}_{\mathcal{M}} = \boldsymbol{\xi}_{\mathcal{M}}$.

*Remark 8.* We verify that the PEL estimator is qualified to serve as $\boldsymbol{\psi}^*$ in Theorem 5.1. Define $\bar{s} = |\mathcal{S} \cap \mathcal{M}|$, and thus $|\mathcal{S} \cap \mathcal{M}^c| = s - \bar{s}$. If $\boldsymbol{\theta}_0$ is low-dimensional and we choose $\boldsymbol{\psi}^* = \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ in (3.1), we have $\varpi_{1,n} = \bar{s}\phi_n$ and $\varpi_{2,n} = (s - \bar{s})\phi_n$ due to $|\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}} - \boldsymbol{\psi}_{0,\mathcal{S}}|_\infty = O_{\mathrm{p}}(\phi_n)$ and $\mathbb{P}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL},\mathcal{S}^c} = \mathbf{0}) \to 1$. Then $ms\phi_n = o(1)$ and $nms^2\phi_n^2(\varsigma^2 + s^2\phi_n^2) = o(1)$ are sufficient for the restrictions imposed on $\varpi_{1,n}$ and $\varpi_{2,n}$. Given $s$ and $\phi_n$ in Theorem 3.1, if $m$, $\omega_n$ and $\varsigma$ satisfy $m = o(n^{\min\{1/5,(\gamma-2)/(3\gamma)\}})$, $m\omega_n^2(m^2 + \log r) = o(1)$, $ms\phi_n = o(1)$ and $nms^2\phi_n^2(\varsigma^2 + s^2\phi_n^2) = o(1)$, the asymptotic normality of Theorem 5.1 holds under this choice of the initial value $\boldsymbol{\psi}^* = \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$. Analogously, if $\boldsymbol{\psi}^* = \hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$ in (4.1) when $\boldsymbol{\theta}_0$ is of high dimension, Theorem 5.1 also holds provided that $m$, $\omega_n$ and $\varsigma$ satisfy the same restrictions with the newly defined $s$ and $\phi_n$ in Section 4.

So far, we have established statistical inference results based on the PPEL estimator $\tilde{\boldsymbol{\psi}}_{\mathcal{M}}$ when we are interested in a small subset $\mathcal{M}$ of $\boldsymbol{\psi}$. In the next section, we check the finite sample performance via simulations.

## 6. Numerical Studies

One of the most important models in econometrics is the linear IV regression. We design a linear IV model here to mimic the empirical application in Section 7, whereas simulation results of a nonlinear panel regression with time-varying individual heterogeneity are presented in the supplementary materials.

Suppose that the researcher has at hand a dataset of $n$ independent observations of a vector $(y_i, x_i, \mathbf{z}_i, w_{1i}, \mathbf{w}_{2i}, \mathbf{w}_{3i})$, and is interested in estimating the main structural equation

$$y_i = \beta_x x_i + \boldsymbol{\beta}_{\mathbf{z}}^{\mathrm{T}} \mathbf{z}_i + \epsilon_i\,, \tag{6.1}$$

where $x_i$ is a scalar endogenous variable, and $\mathbf{z}_i$ is a $d_z \times 1$ vector of exogenous variables (including the intercept). Such an equation with a scalar endogenous variable is the leading case of IV regressions (Andrews, Stock, and Sun 2019). Due to space limitations, we report the results when we specify $\mathbf{z}_i = (1, z_{1,i}, z_{2,i})^{\mathrm{T}} \in \mathbb{R}^3$ with $(z_{1,i}, z_{2,i})^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_2)$, and the coefficients $(\beta_x, \boldsymbol{\beta}_{\mathbf{z}}^{\mathrm{T}})^{\mathrm{T}} = (0.5, 0.5, 0.5, 0.5)^{\mathrm{T}}$ in a plausible setting. The numerical performance are robust across our experiments when the parameters are varied.

The corresponding reduced-form equation is $x_i = \gamma_{w1} w_{1i} + \boldsymbol{\gamma}_{\mathbf{w}2}^{\mathrm{T}} \mathbf{w}_{2i} + \boldsymbol{\gamma}_{\mathbf{z}}^{\mathrm{T}} \mathbf{z}_i + u_i$, where $w_{1i} \in \mathbb{R}$ and $\mathbf{w}_{2i} \in \mathbb{R}^{d_{w2}}$ are excluded instruments. We specify $(w_{1i}, \mathbf{w}_{2i}^{\mathrm{T}})^{\mathrm{T}} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{d_{w2}+1})$, $\gamma_{w1} = 0.8$, $\boldsymbol{\gamma}_{\mathbf{w}2} = (\gamma_{\mathbf{w}2,j})_{j \in [d_{w2}]}$ with $\gamma_{\mathbf{w}2,j} = 0.4 - 0.3(j-1)/(d_{w2}-1)$, and $\boldsymbol{\gamma}_{\mathbf{z}} = (0.8, 0.8, 0.8)^{\mathrm{T}}$. The two error terms $(\epsilon_i, u_i)^{\mathrm{T}}$ are generated from the two-dimensional normal distribution with mean 0, covariance 1 and correlation 0.5 which are independent of $(\mathbf{z}_i, w_{1i}, \mathbf{w}_{2i})$. An additional vector $\mathbf{w}_{3i} = (w_{3i,j})_{j \in [s]} \in \mathbb{R}^s$, which serves as the invalid IV, follows $w_{3i,j} = \delta_j \epsilon_i + v_i$, where $v_i \sim \mathcal{N}(0,1)$ is independent of $(\epsilon_i, u_i)$, and $\delta_j \neq 0$ controls the strength of correlation. Write $\boldsymbol{\delta} = (\delta_j)_{j \in [s]}$. To emulate the scenarios of *weak*, *moderate*, and *strong* correlations between $\epsilon_i$ and $\mathbf{w}_{3i}$, we set $\delta_j = 0.3+0.2(j-1)/(s-1)$, $0.5+0.2(j-1)/(s-1)$ and $0.7+0.2(j-1)/(s-1)$, respectively. It is expected that the smaller is the coefficient, the less severe is misspecification, so estimation is more prone to selection mistakes in finite samples.

While the researcher is confident about the validity of $w_{1i}$ in that $\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}) = (w_{1i}, \mathbf{z}_i^{\mathrm{T}})^{\mathrm{T}} \times (y_i - \beta_x x_i - \boldsymbol{\beta}_{\mathbf{z}}^{\mathrm{T}}\mathbf{z}_i)$, where $\mathbf{z}_i$ is self-instrumented, she is uncertain about the validity of $(\mathbf{w}_{2i}, \mathbf{w}_{3i})$ and therefore must detect the invalid IVs. The two classes of undetermined (to the researcher) IVs consist of $\mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta}) = (\mathbf{w}_{2i}^{\mathrm{T}}, \mathbf{w}_{3i}^{\mathrm{T}})^{\mathrm{T}} \times (y_i - \beta_x x_i - \boldsymbol{\beta}_{\mathbf{z}}^{\mathrm{T}}\mathbf{z}_i)$. According to our DGP design, $\mathbf{w}_{2i}$ is the valid IV vector so the moments involving $\mathbf{w}_{2i}$ equal to zero under the true parameter; those moments involving $\mathbf{w}_{3i}$ are invalid.

Let $d_w = d_{\mathbf{w}2} + s + 1$ be the number of all the IVs, among which $s$ IVs are invalid. In the low-dimensional setting, we fix $s = 6$ and consider $(n, d_w) = (100, 50)$ or $(200, 100)$. In the high-dimensional setting, we vary the number of invalid IVs to be $s \in \{6, 8, 13\}$ for $(n, d_w) = (100, 120)$, and $s \in \{6, 12, 17\}$ for $(n, d_w) = (200, 240)$, where $s$ is specified by rounding, in addition, $2\log n$ and $3n^{1/3}$.

In terms of the numerical implementation, we compute the PEL estimates by the *modified two-layer coordinate descent algorithm* (Chang, Tang, and Wu 2018). The SCAD penalty is used for both $P_{1,\pi}(\cdot)$ and $P_{2,\nu}(\cdot)$ in (3.1) for all the numerical experiments in this article with local quadratic approximation (Fan and Li 2001), and the tuning parameters $\nu$ and $\pi$ are chosen by the Bayesian information criterion (BIC). Specifically, we use the following BIC type function:

$$\mathrm{BIC} = \ell(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) + (\log n) \cdot \mathrm{df}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}), \tag{6.2}$$

where $\mathrm{df}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})$ denotes the number of nonzero elements in $\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}$, and the log likelihood term is the EL ratio $\ell(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}}) = 2\sum_{i=1}^n \log\{1 + \hat{\boldsymbol{\lambda}}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})^{\mathrm{T}}\mathbf{g}_i^{(\mathcal{T})}(\hat{\boldsymbol{\psi}}_{\mathrm{PEL}})\}$.[7]

We first report results of moment selection by our selection criterion in (3.9). Let "FP" (false positive) denote the frequency that the valid moments being not selected, and "FN" (false negative) denote the frequency that the invalid moments being selected. In Table 1, PEL and DB-PEL denote the moment selection criterion based on the PEL estimator and its de-biased version, respectively. As expected, the strength of correlations between $\epsilon_i$ and $\mathbf{w}_{3i}$ does not affect FP, whereas FN quickly vanishes as the correlations get stronger. In all cases, larger sample sizes help reduce the chance of moment selection error.

The parameter of interest in the linear IV model is $\beta_x$ in (6.1) as it characterizes the "causal effect" of the endogenous variable, which bears economic interpretation. The root-mean-square error (RMSE), bias (BIAS) and standard deviation (STD) are calculated for PEL, DB-PEL and the classical two-stage least squares (2SLS) for (6.1), and an oracle estimator is added for comparison.[8] The results are summarized in Table 2. The performance of PEL and the oracle is comparable, and the gaps are narrowed when the sample size is increased from $n = 100$ to $n = 200$, suggesting the capacity for PEL to mimic the oracle by absorbing the signal from the valid moments and in the meantime keeping the invalid ones at bay. The RMSE of 2SLS is significantly larger than those of PEL and DB-PEL.

**Table 1.** PEL's performance in moment selection.

| | Correlation: | Weak | | Moderate | | Strong | |
|---|---|---|---|---|---|---|---|
| $(n, d_w, s)$ | method | FP | FN | FP | FN | FP | FN |
| | Panel A: low-dimensional setting | | | | | | |
| (100, 50, 6) | PEL | 0.1995 | 0.0267 | 0.1933 | 0.0003 | 0.2359 | 0.0000 |
| | DB-PEL | 0.1983 | 0.0273 | 0.1923 | 0.0003 | 0.2343 | 0.0000 |
| (200, 100, 6) | PEL | 0.0942 | 0.0033 | 0.1012 | 0.0000 | 0.0971 | 0.0000 |
| | DB-PEL | 0.0930 | 0.0033 | 0.1003 | 0.0000 | 0.0962 | 0.0000 |
| | Panel B: high-dimensional setting | | | | | | |
| (100, 120, 6) | PEL | 0.1067 | 0.0550 | 0.0850 | 0.0047 | 0.1141 | 0.0013 |
| | DB-PEL | 0.1061 | 0.0580 | 0.0845 | 0.0050 | 0.1131 | 0.0013 |
| (200, 240, 6) | PEL | 0.0478 | 0.0073 | 0.0476 | 0.0007 | 0.0444 | 0.0007 |
| | DB-PEL | 0.0468 | 0.0093 | 0.0464 | 0.0007 | 0.0435 | 0.0007 |
| (100, 120, 8) | PEL | 0.1009 | 0.0675 | 0.0872 | 0.0063 | 0.1087 | 0.0010 |
| | DB-PEL | 0.1004 | 0.0693 | 0.0863 | 0.0065 | 0.1074 | 0.0015 |
| (200, 240, 12) | PEL | 0.0498 | 0.0072 | 0.0497 | 0.0003 | 0.0494 | 0.0003 |
| | DB-PEL | 0.0486 | 0.0087 | 0.0488 | 0.0003 | 0.0485 | 0.0003 |
| (100, 120, 13) | PEL | 0.0709 | 0.0892 | 0.0558 | 0.0168 | 0.1288 | 0.0042 |
| | DB-PEL | 0.0706 | 0.0905 | 0.0554 | 0.0175 | 0.1268 | 0.0043 |
| (200, 240, 17) | PEL | 0.0402 | 0.0086 | 0.0469 | 0.0002 | 0.0456 | 0.0000 |
| | DB-PEL | 0.0394 | 0.0116 | 0.0458 | 0.0002 | 0.0447 | 0.0000 |

The left panel of Figure 1 plots the empirical cumulative distribution functions (ECDF) of the DB-PEL estimates. The dotted curve corresponds to the case of $(n, d_w, s) = (100, 120, 8)$, the dashed curve to that of $(n, d_w, s) = (200, 240, 12)$, and the solid curve is the cumulative distribution function (CDF) of $\mathcal{N}(0, 1)$ for comparison. A better normal approximation can be obtained by the PPEL estimate, as shown in the right panel of Figure 1, with its tuning parameter $\varsigma = 0.08(n^{-1}\log p)^{1/2}$. We also present the confidence intervals (CI) according to PPEL, DB-PEL and 2SLS, as reported in Table 3, for the 90%, 95% and 99% levels, where the CIs for DB-PEL and PPEL are predicated on the asymptotic normality from Theorems 3.1 and 5.1, respectively, and the CIs for 2SLS are based on its asymptotic normality as in standard textbooks. PPEL's coverage probability to the nominal counterpart is the best among the three estimators, and is much better than PEL. 2SLS's coverage probability seems too high in the high-dimensional setting, though it is acceptable in the low-dimensional case. Figure 2 shows that the width of 2SLS's CI is much wider than that of PPEL, which is caused by efficiency loss from abandoning the potentially valid estimating equations.

## 7. Empirical Application

Colonialism was widespread over the globe prior to the Second World War. While European institutions which protected private properties and checked government powers were successfully replicated in a few colonies, many others suffered expropriation of natural and human resources. In an influential study, Acemoglu, Johnson, and Robinson (2001) (AJR, henceforth) systematically explored the relationship between Europeans' mortality rates and the modes of institutions.

[7] Leng and Tang (2012) employ this BIC to choose the tuning parameter for the high-dimensional parameter $\boldsymbol{\psi}$, and Chang, Tang, and Wu (2018) apply it to select multiple tuning parameters. We follow their practice as (6.2) remains valid in our setting where two tuning parameters are included.

[8] The oracle EL in Section 2.2 is for low-dimensional parameters and moments. To handle the large pool of orthogonal IVs, the oracle estimator here is a 2SLS estimator taking advantage of a few most relevant IVs—those with the top $0.1n$ big coefficients $\gamma_{\mathbf{w}2,j}$ in the reduced-form equation.

**Table 2.** Point estimations for $\beta_x$.

| | Correlation: | Weak | | | Moderate | | | Strong | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n, d_W, s)$ | method | RMSE | BIAS | STD | RMSE | BIAS | STD | RMSE | BIAS | STD |
| | | | | | Panel A: low-dimensional setting | | | | | |
| (100, 50, 6) | Oracle | 0.074 | 0.020 | 0.071 | 0.074 | 0.020 | 0.071 | 0.074 | 0.020 | 0.071 |
| | PEL | 0.088 | 0.030 | 0.082 | 0.092 | 0.032 | 0.086 | 0.081 | 0.019 | 0.079 |
| | DB-PEL | 0.080 | 0.032 | 0.073 | 0.082 | 0.035 | 0.075 | 0.077 | 0.026 | 0.073 |
| | 2SLS | 0.145 | −0.004 | 0.145 | 0.145 | −0.004 | 0.145 | 0.145 | −0.004 | 0.145 |
| (200, 100, 6) | Oracle | 0.039 | 0.010 | 0.038 | 0.039 | 0.010 | 0.038 | 0.039 | 0.010 | 0.038 |
| | PEL | 0.044 | 0.013 | 0.042 | 0.047 | 0.013 | 0.045 | 0.045 | 0.013 | 0.044 |
| | DB-PEL | 0.042 | 0.021 | 0.036 | 0.042 | 0.020 | 0.037 | 0.042 | 0.020 | 0.037 |
| | 2SLS | 0.101 | −0.003 | 0.101 | 0.101 | −0.003 | 0.101 | 0.101 | −0.003 | 0.101 |
| | | | | | Panel B: high-dimensional setting | | | | | |
| (100, 120, 6) | Oracle | 0.067 | 0.018 | 0.064 | 0.067 | 0.018 | 0.064 | 0.067 | 0.018 | 0.064 |
| | PEL | 0.090 | 0.031 | 0.084 | 0.098 | 0.040 | 0.090 | 0.092 | 0.033 | 0.087 |
| | DB-PEL | 0.083 | 0.029 | 0.078 | 0.082 | 0.031 | 0.076 | 0.077 | 0.026 | 0.073 |
| | 2SLS | 0.181 | −0.016 | 0.180 | 0.181 | −0.016 | 0.180 | 0.181 | −0.016 | 0.180 |
| (200, 240, 6) | Oracle | 0.035 | 0.012 | 0.033 | 0.035 | 0.012 | 0.033 | 0.035 | 0.012 | 0.033 |
| | PEL | 0.045 | 0.014 | 0.043 | 0.044 | 0.014 | 0.042 | 0.037 | 0.014 | 0.034 |
| | DB-PEL | 0.043 | 0.017 | 0.040 | 0.042 | 0.017 | 0.039 | 0.035 | 0.016 | 0.031 |
| | 2SLS | 0.128 | 0.002 | 0.128 | 0.128 | 0.002 | 0.128 | 0.128 | 0.002 | 0.128 |
| (100, 120, 8) | Oracle | 0.063 | 0.015 | 0.061 | 0.063 | 0.015 | 0.061 | 0.063 | 0.015 | 0.061 |
| | PEL | 0.110 | 0.036 | 0.104 | 0.116 | 0.045 | 0.107 | 0.112 | 0.039 | 0.105 |
| | DB-PEL | 0.100 | 0.033 | 0.094 | 0.098 | 0.038 | 0.090 | 0.092 | 0.032 | 0.087 |
| | 2SLS | 0.211 | −0.011 | 0.211 | 0.211 | −0.011 | 0.211 | 0.211 | −0.011 | 0.211 |
| (200, 240, 12) | Oracle | 0.034 | 0.010 | 0.033 | 0.034 | 0.010 | 0.033 | 0.034 | 0.010 | 0.033 |
| | PEL | 0.038 | 0.013 | 0.036 | 0.036 | 0.013 | 0.033 | 0.036 | 0.013 | 0.033 |
| | DB-PEL | 0.039 | 0.017 | 0.035 | 0.038 | 0.016 | 0.034 | 0.038 | 0.016 | 0.034 |
| | 2SLS | 0.126 | −0.005 | 0.126 | 0.126 | −0.005 | 0.126 | 0.126 | −0.005 | 0.126 |
| (100, 120, 13) | Oracle | 0.065 | 0.020 | 0.062 | 0.065 | 0.020 | 0.062 | 0.065 | 0.020 | 0.062 |
| | PEL | 0.117 | 0.051 | 0.105 | 0.139 | 0.069 | 0.121 | 0.110 | 0.036 | 0.104 |
| | DB-PEL | 0.103 | 0.045 | 0.092 | 0.113 | 0.054 | 0.099 | 0.093 | 0.033 | 0.087 |
| | 2SLS | 0.200 | −0.008 | 0.200 | 0.200 | −0.008 | 0.200 | 0.200 | −0.008 | 0.200 |
| (200, 240, 17) | Oracle | 0.036 | 0.012 | 0.034 | 0.036 | 0.012 | 0.034 | 0.036 | 0.012 | 0.034 |
| | PEL | 0.047 | 0.013 | 0.045 | 0.052 | 0.011 | 0.051 | 0.041 | 0.011 | 0.039 |
| | DB-PEL | 0.044 | 0.016 | 0.041 | 0.053 | 0.015 | 0.051 | 0.039 | 0.015 | 0.036 |
| | 2SLS | 0.133 | −0.014 | 0.132 | 0.133 | −0.014 | 0.132 | 0.133 | −0.014 | 0.132 |

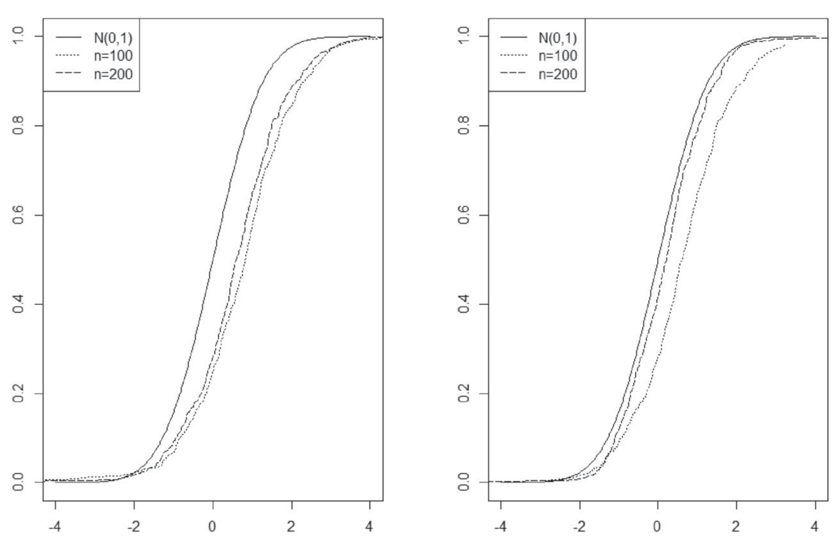


**Figure 1.** ECDF of DB-PEL (left) and PPEL (right) of $\beta_x$ with moderate correlation between $\epsilon_i$ and $\mathbf{w}_{3i}$.

We revisit AJR's open-access dataset. AJR's main structural equation is $y_i = \gamma_y + \beta_x x_i + \beta_z z_i + \epsilon_i$, where the dependent variable $y_i$ is `logarithm GDP per capita in 1995`, and the key explanatory variable of interest $x_i$ is `average protection against expropriation risk`, a continuous variable of a scale between 0 and 10. The `latitude` of a colony, denoted as $z_i$, is an additional control variable. For substantial measurement errors in the institutional index $x_i$ can bias the OLS estimator, the credibility of the empirical evidence counts on IVs.

AJR compiled from historical documents a seminal variable `logarithm of European settler mortality`, and argued forcefully that it qualifies as a valid IV for the endogenous variable $x_i$. Their empirical evidence was based on the 2SLS regression in cross-sectional data of countries. The baseline result in AJR's Column (2) of Table 4 (p.1386)

**Table 3.** Coverage probabilities for the CIs of $\beta_x$.

| | Correlation: | Weak | | | Moderate | | | Strong | | |
|---|---|---|---|---|---|---|---|---|---|---|
| $(n, d_w, s)$ | method | 90 | 95 | 99 | 90 | 95 | 99 | 90 | 95 | 99 |
| | | | | | Panel A: low-dimensional setting | | | | | |
| (100, 50, 6) | PPEL | 0.894 | 0.940 | 0.988 | 0.890 | 0.940 | 0.986 | 0.888 | 0.942 | 0.986 |
| | DB-PEL | 0.736 | 0.830 | 0.926 | 0.726 | 0.794 | 0.916 | 0.744 | 0.818 | 0.902 |
| | 2SLS | 0.912 | 0.964 | 0.994 | 0.912 | 0.964 | 0.994 | 0.912 | 0.964 | 0.994 |
| (200, 100, 6) | PPEL | 0.902 | 0.950 | 0.994 | 0.902 | 0.948 | 0.994 | 0.902 | 0.946 | 0.994 |
| | DB-PEL | 0.762 | 0.838 | 0.934 | 0.734 | 0.842 | 0.940 | 0.752 | 0.852 | 0.928 |
| | 2SLS | 0.920 | 0.956 | 1.000 | 0.920 | 0.956 | 1.000 | 0.920 | 0.956 | 1.000 |
| | | | | | Panel B: high-dimensional setting | | | | | |
| (100, 120, 6) | PPEL | 0.908 | 0.956 | 0.996 | 0.892 | 0.948 | 0.994 | 0.888 | 0.940 | 0.992 |
| | DB-PEL | 0.746 | 0.832 | 0.922 | 0.784 | 0.850 | 0.926 | 0.768 | 0.836 | 0.936 |
| | 2SLS | 0.956 | 0.982 | 1.000 | 0.956 | 0.982 | 1.000 | 0.956 | 0.982 | 1.000 |
| (200, 240, 6) | PPEL | 0.914 | 0.964 | 0.988 | 0.916 | 0.960 | 0.988 | 0.906 | 0.954 | 0.988 |
| | DB-PEL | 0.754 | 0.846 | 0.966 | 0.778 | 0.854 | 0.958 | 0.776 | 0.866 | 0.966 |
| | 2SLS | 0.966 | 0.992 | 1.000 | 0.966 | 0.992 | 1.000 | 0.966 | 0.992 | 1.000 |
| (100, 120, 8) | PPEL | 0.888 | 0.942 | 0.992 | 0.870 | 0.930 | 0.986 | 0.852 | 0.924 | 0.980 |
| | DB-PEL | 0.778 | 0.836 | 0.904 | 0.738 | 0.818 | 0.920 | 0.730 | 0.806 | 0.910 |
| | 2SLS | 0.960 | 0.988 | 1.000 | 0.960 | 0.988 | 1.000 | 0.960 | 0.988 | 1.000 |
| (200, 240, 12) | PPEL | 0.924 | 0.962 | 0.990 | 0.914 | 0.958 | 0.986 | 0.892 | 0.950 | 0.982 |
| | DB-PEL | 0.794 | 0.872 | 0.954 | 0.788 | 0.864 | 0.946 | 0.764 | 0.860 | 0.938 |
| | 2SLS | 0.972 | 0.998 | 1.000 | 0.972 | 0.998 | 1.000 | 0.972 | 0.998 | 1.000 |
| (100, 120, 13) | PPEL | 0.884 | 0.940 | 0.992 | 0.852 | 0.914 | 0.976 | 0.862 | 0.912 | 0.974 |
| | DB-PEL | 0.788 | 0.862 | 0.938 | 0.744 | 0.798 | 0.904 | 0.688 | 0.768 | 0.882 |
| | 2SLS | 0.942 | 0.986 | 0.996 | 0.942 | 0.986 | 0.996 | 0.942 | 0.986 | 0.996 |
| (200, 240, 17) | PPEL | 0.916 | 0.956 | 0.984 | 0.914 | 0.958 | 0.982 | 0.886 | 0.938 | 0.982 |
| | DB-PEL | 0.792 | 0.884 | 0.968 | 0.780 | 0.866 | 0.954 | 0.798 | 0.864 | 0.952 |
| | 2SLS | 0.974 | 0.988 | 1.000 | 0.974 | 0.988 | 1.000 | 0.974 | 0.988 | 1.000 |

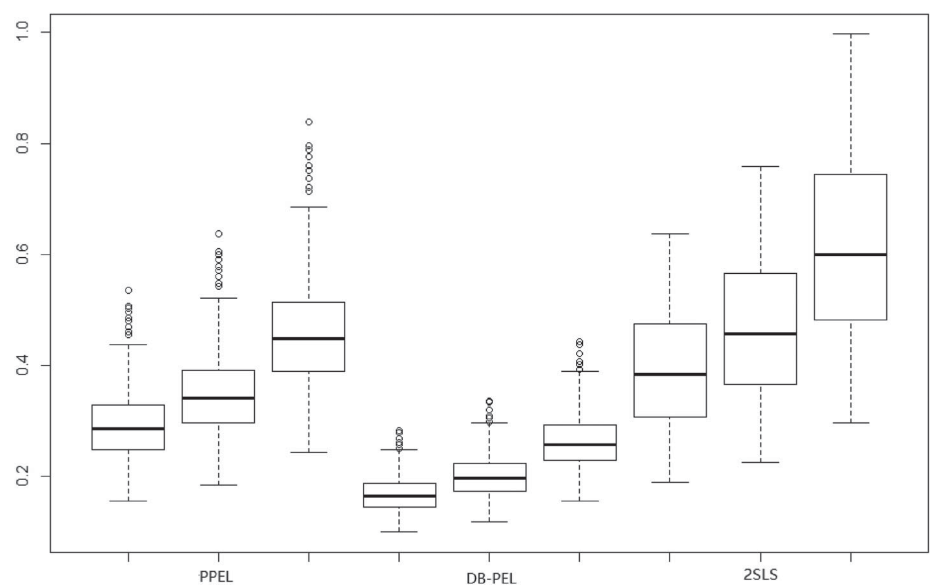


**Figure 2.** Width of 90% (left), 95% (middle), and 99% (right) CI for $\beta_x$ estimated by PPEL, DB-PEL, and 2SLS under moderate correlation between $\epsilon_i$ and $\mathbf{w}_{3i}$ with $(n, d_w, s) = (100, 120, 8)$.

corresponds to the three estimating functions $\mathbf{g}_i^{(\mathcal{I})}(\boldsymbol{\theta}) = (w_{1i}, z_i, 1)^{\mathrm{T}} \times (y_i - \gamma_y - \beta_x x_i - \beta_z z_i)$ by the notations in our simulation, and the 2SLS reports $\hat{\beta}_x = 1.00$ with STD 0.22.

There are another 11 potential health variables and institutional variables, which serve as potential IVs. AJR were uncertain about their validity, and they experimented in their Table 7 (p.1392) and Table 8 (p.1394) the empirical results under various IV configurations. These variables, denoted as $\mathbf{w}_2$, are (i) `Malaria in 1994`, (ii) `Yellow fever`, (iii) `Life expectancy`, (iv) `Infant mortality`, (v) `Mean temperature`, (vi) `Distance from coast`, (vii) `European settlements in 1990`, (viii) `Democracy (1st year of independence)`, (ix) `Constraint on executive (1st year of independence)`, (x) `Democracy in 1900`, and (xi) `Constraint on executive in 1900`. (i)–(iv) are health variables, (v)–(vi) are geographic variables, and (vii)–(xi) are institutional variables. All these extra IVs are associated with the estimating functions $\mathbf{g}_i^{(\mathcal{D})}(\boldsymbol{\theta}) = \mathbf{w}_{2i} \times (y_i - \gamma_y - \beta_x x_i - \beta_z z_i)$. Given the moderate sample size of 56 countries, a total of 12 IVs is nontrivial.

Table 4. Estimation of the effect of institution ($\beta_X$).

| | $\zeta_c$ | PE | STD | 95% CI |
|---|---|---|---|---|
| PEL | NA | 0.937 | 0.078 | (0.786, 1.090) |
| DB-PEL | NA | 0.938 | 0.078 | (0.786, 1.090) |
| 2SLS | NA | 0.945 | 0.200 | (0.553, 1.338) |
| PPEL | 0.04 | 0.942 | 0.159 | (0.631, 1.254) |
| | 0.06 | 0.941 | 0.136 | (0.675, 1.207) |
| | **0.08** | **0.945** | **0.126** | **(0.698, 1.193)** |
| | 0.12 | 0.964 | 0.150 | (0.669, 1.259) |
| | 0.16 | 0.967 | 0.152 | (0.669, 1.266) |

NOTE: Our sample size is 56, after removing countries with missing variables from the original sample of 64 countries, so the 2SLS point estimate is 0.945, slightly different from AJR's 1.00. The line with bold font follows the simulation exercise which sets the constant of the tuning parameter as 0.08.

Our PPEL estimates the main equation as

$$\log \widehat{\text{GDP per capita}} = \underset{(1.194)}{2.048} + \underset{(0.126)}{0.945} \times \text{institution} - \underset{(0.977)}{0.785} \times \text{lattitude}.$$

For the estimation and inference of the key parameter of interest $\beta_x$, we compare the results of PEL, DB-PEL, PPEL and the conventional 2SLS. Its point estimation (PE), STD, and CIs are presented in Table 4. Enjoying the efficiency gain from the extra IVs, the STD of PPEL is 0.126 when we select the tuning parameter $\varsigma = \zeta_c(n^{-1}\log p)^{1/2}$ with $\zeta_c = 0.08$ (the boldface row), as suggested by our simulations. This STD achieves a 37% reduction relative to that of the plain 2SLS with the sole valid IV. Moreover, when we vary the constant $\zeta_c = 0.08$ in $\varsigma$ as 0.04, 0.06, 0.12 and 0.16, PPEL is rather robust over the wide range of tuning parameters.

Our method is particularly important in unifying the IV selection in AJR's Tables 7 and 8 into a single set of automatically selected IVs. Among the 11 potential IVs, our moment selection criterion (3.9) invalidates all institutional variables. Five IVs survive the testing: the climate variable `mean temperature`, the geographic variable `distance from coast`, and three health variables `dummy of yellow fever`, `infant mortality` and `life expectancy`. All the other health and institutional variables are assessed as endogenous and are unsuitable for IVs in this study. The justification for `mean temperature` and `distance from coast` is straightforward because humans were unable to interfere with these natural conditions in the era of colonialism. Furthermore, AJR argued for the validity of `dummy of yellow fever`, though due to concerns of lack of variation they did not employ it as the main IV (AJR's p.1393, Paragraph 2). Our variable selection result provides supportive evidence to AJR's heuristics.

## 8. Conclusion

This article considers a general setting of an economic structural model with many potential moments, some of which may be invalid. These invalid moments must be disciplined in order to estimate the structural parameter consistently. We propose a PEL approach to estimate the parameter of interest while coping with the invalid moments. We show that the PEL estimator is normally distributed asymptotically, and invalid moments can be consistently detected thanks to the oracle property. To overcome the difficulty of estimating the bias in the limiting distribution of the PEL estimator, we further devise the PPEL approach for statistical inference of a low-dimensional object of interest, which is useful for hypothesis testing and confidence region construction. Simulation exercises are carried out to demonstrate excellent finite sample performance of our methods. We revisit an empirical application concerning economic development and shed new insight about its candidate instruments.

## Supplementary Materials

The supplementary materials consists of three parts. Part A provides the proofs and technical details about the methods developed in the present article. Part B reports additional simulation results concerning the liner IV model in the main text and an additional dynamic panel data model, respectively. Part C checks the robustness of PEL in the empirical application.

## Acknowledgments

The three coauthors contributed to this article equally. We thank the editor, the associate editor, and two anonymous referees for their constructive comments and suggestions.

## Funding

Chang and Zhang were supported in part by the National Natural Science Foundation of China (grant nos. 72125008, 71991472, 11871401 and 72003150). Chang was also supported by the Center of Statistical Research at Southwestern University of Finance and Economics.

## ORCID

Jinyuan Chang https://orcid.org/0000-0001-7933-4449
Zhentao Shi https://orcid.org/0000-0002-7289-2840

## References

Acemoglu, D., Johnson, S., and Robinson, J. A. (2001), "The Colonial Origins of Comparative Development: An Empirical Investigation," *The American Economic Review*, 91, 1369–1401. [10]

Ackerberg, D., Benkard, C. L., Berry, S. and Pakes, A. (2007), "Econometric Tools for Analyzing Market Outcomes," *Handbook of Econometrics*, 6, 4171–4276. [2]

Altonji, J. G., and Segal, L. (1996), "Small-Sample Bias in GMM Estimation of Covariance Structures," *Journal of Business & Economic Statistics*, 14, 353–366. [3]

Altonji, J. G., Smith, A. A. J., and Vidangos, I. (2013), "Modeling Earnings Dynamics," *Econometrica*, 81, 1395–1454. [1]

Anatolyev, S., and Gospodinov, N. (2011), *Methods for Estimation and Inference in Modern Econometrics*, Boca Raton, FL: CRC Press. [1]

Andrews, D., and Cheng, X. (2012), "Estimation and Inference with Weak, Semi-strong, and Strong Identification," *Econometrica*, 80, 2153–2211. [2]

Andrews, D., and Lu, B. (2001), "Consistent Model and Moment Selection Procedures for GMM Estimation with Application to Dynamic Panel Data Models," *Journal of Econometrics*, 101, 123–164. [1]

Andrews, I., and Mikusheva, A. (2020), "Optimal Decision Rules for Weak GMM," arXiv preprint arXiv:2007.04050. [2]

Andrews, I., Stock, J. H., and Sun, L. (2019), "Weak Instruments in Instrumental Variables Regression: Theory and Practice," *Annual Review of Economics*, 11, 727–753. [9]

Angrist, J. D., (1990), "Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records," *The American Economic Review*, 80, 313–336. [2,3]

Angrist, J. D., and Keueger, A. B. (1991), "Does Compulsory School Attendance Affect Schooling and Earnings?," *The Quarterly Journal of Economics*, 106, 979–1014. [2]

——— (1992), "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments from Two Samples," *Journal of the American Statistical Association*, 87, 328–336. [1]

Bekker, P. A. (1994), "Alternative Approximations to the Distributions of Instrumental Variable Estimators," *Econometrica*, 62, 657–681. [2]

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012), "Sparse Models and Methods for Optimal Instruments with an Application to Eminent Domain," *Econometrica*, 80, 2369–2429. [2]

Belloni, A., Chernozhukov, V., Chetverikov, D., Hansen, C., and Kato, K. (2018), "High-Dimensional Econometrics and Regularized GMM," arXiv preprint arXiv:1806.01888. [2]

Belloni, A., Chernozhukov, V., and Hansen, C. (2014), "Inference on Treatment Effects After Selection Among High-Dimensional Controls," *The Review of Economic Studies*, 81, 608–650. [2]

Bertsekas, D. P. (1997), "Nonlinear Programming," *Journal of the Operational Research Society*, 48, 334–334. [6]

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009), "Simultaneous Analysis of Lasso and Dantzig Selector," *The Annals of Statistics*, 37, 1705–1732. [9]

Blundell, R., Pashardes, P., and Weber, G. (1993), "What do we learn about Consumer Demand Patterns from Micro Data?" *The American Economic Review*, 83, 570–597. [7]

Bühlmann, P. (2013), "Statistical Significance in High-Dimensional Linear Models," *Bernoulli*, 19, 1212–1242. [6]

Candes, E., and Tao, T. (2007), "The Dantzig Selector: Statistical Estimation When p is much Larger than n," *The Annals of Statistics*, 35, 2313–2351. [9]

Caner, M., and Fan, Q. (2015), "Hybrid Generalized Empirical Likelihood Estimators: Instrument Selection with Adaptive Lasso," *Journal of Econometrics*, 187, 256–274. [2]

Caner, M., Han, X., and Lee, Y. (2018), "Adaptive Elastic Net GMM Estimation with Many Invalid Moment Conditions: Simultaneous Model and Moment Selection," *Journal of Business & Economic Statistics*, 36, 24–46. [2]

Caner, M., and Kock, A. B. (2018), "High Dimensional Linear GMM," arXiv preprint arXiv:1811.08779. [2]

Chang, J., Chen, S. X. and Chen, X. (2015), "High Dimensional Generalized Empirical Likelihood for Moment Restrictions with Dependent Data," *Journal of Econometrics*, 185, 283–304. [2,3]

Chang, J., Chen, S. X., Tang, C. Y. and Wu, T. T. (2021), "High-Dimensional Empirical Likelihood Inference," *Biometrika*, 108, 127–147. [2,8,9]

Chang, J., Tang, C. Y., and Wu, T. T. (2018), "A New Scope of Penalized Empirical Likelihood with High-Dimensional Estimating Equations," *The Annals of Statistics*, 46, 3185–3216. [2,4,7,10]

Chao, J. C., Swanson, N. R., Hausman, J. A., Newey, W. K., and Woutersen, T. (2011), "Asymptotic Distribution of JIVE in a Heteroskedastic IV Regression with Many Instruments," *Econometric Theory*, 28, 42–86. [2]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection with Large Model Spaces," *Biometrika*, 95, 759–771. [5]

Cheng, X., and Liao, Z. (2015), "Select the Valid and Relevant Moments: An Information-Based Lasso for GMM with Many Moments," *Journal of Econometrics*, 186, 443–464. [1,2,6]

DiTraglia, F. J. (2016), "Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM," *Journal of Econometrics*, 195, 187–208. [2]

Eaton, J., Kortum, S., and Kramarz, F. (2011), "An Anatomy of International Trade: Evidence from French Firms," *Econometrica*, 79, 1453–1498. [1,3]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [4,8,10]

Fan, J., and Liao, Y. (2014), "Endogeneity in High Dimensions," *The Annals of Statistics*, 42, 872–917. [7]

Gold, D., Lederer, J., and Tao, J. (2020), "Inference for High-Dimensional Instrumental Variables Regression," *Journal of Econometrics*, 217, 79–111. [2]

Hall, A. R., Inoue, A., Jana, K., and Shin, C. (2007), "Information in Generalized Method of Moments Estimation and Entropy-Based Moment Selection," *Journal of Econometrics*, 138, 488–512. [4]

Han, C., and Phillips, P. (2006), "GMM with Many Moment Conditions," *Econometrica*, 74, 147–192. [2]

Hansen, C., and Kozbur, D. (2014), "Instrumental Variables Estimation with Many Weak Instruments Using Regularized JIVE," *Journal of Econometrics*, 182, 290–308. [2]

Hansen, L. (1982), "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50, 1029–1054. [1,3]

Hansen, L., Heaton, J., and Yaron, A. (1996), "Finite-Sample Properties of Some Alternative GMM Estimators," *Journal of Business & Economic Statistics*, 14, 262–280. [3]

Hjort, N. L., McKeague, I. W., and Van Keilegom, I. (2009), "Extending the Scope of Empirical Likelihood," *The Annals of Statistics*, 37, 1079–1111. [3]

Hong, H., Preston, B., and Shum, M. (2003), "Generalized Empirical Likelihood-Based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 19, 923–943. [1]

Kitamura, Y., (2001), "Asymptotic Optimality of Empirical Likelihood for Testing Moment Restrictions," *Econometrica*, 69, 1661–1672. [2]

Kleibergen, F. (2005), "Testing Parameters in GMM Without Assuming that they are Identified," *Econometrica*, 73, 1103–1123. [2]

Kolesár, M., Chetty, R., Friedman, J., Glaeser, E., and Imbens, G. W. (2015), "Identification and Inference with Many Invalid Instruments," *Journal of Business & Economic Statistics*, 33, 474–484. [2,3]

Leng, C., and Tang, C. Y. (2012), "Penalized Empirical Likelihood and Growing Dimensional General Estimating Equations," *Biometrika*, 99, 703–716. [10]

Liao, Z. (2013), "Adaptive GMM Shrinkage Estimation with Consistent Moment Selection," *Econometric Theory*, 29, 857–904. [1,2,4,6,8]

Lv, J., and Fan, Y. (2009), "A Unified Approach to Model Selection and Sparse Recovery Using Regularized Least Squares," *The Annals of Statistics*, 37, 3498–3528. [4]

Matsushita, Y., and Otsu, T. (2013), "Second-Order Refinement of Empirical Likelihood for Testing Overidentifying Restrictions," *Econometric Theory*, 29, 324–353. [2]

Mavroeidis, S. (2005), "Identification Issues in Forward-Looking Models Estimated by GMM, with an Application to the Phillips Curve," *Journal of Money, Credit and Banking*, 37, 421–448. [2]

Murray, M. P. (2006), "Avoiding Invalid Instruments and Coping with Weak Instruments," *Journal of Economic Perspectives*, 20, 111–132. [1]

Newey, W. K., and Smith, R. J. (2003), "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72, 219–255. [3]

Newey, W. K., and Windmeijer, F. (2009), "Generalized Method of Moments with Many Weak Moment Conditions," *Econometrica*, 77, 687–719. [2]

Neykov, M., Ning, Y., Liu, J. S. and Liu, H. (2018), "A Unified Theory of Confidence Regions and Testing for High-Dimensional Estimating Equations," *Statistical Science*, 33, 427–443. [8]

Ning, Y., and Liu, H. (2017), "A General Theory of Hypothesis Tests and Confidence Regions for Sparse High Dimensional Models," *The Annals of Statistics*, 45, 158–195. [8]

Otsu, T. (2006), "Generalized Empirical Likelihood Inference for Nonlinear and Time Series Models Under Weak Identification," *Econometric Theory*, 22, 513–527. [2]

——— (2007), "Penalized Empirical Likelihood Estimation of Semiparametric Models," *Journal of Multivariate Analysis*, 98, 1923–1954. [2]

——— (2010), "On Bahadur Efficiency of Empirical Likelihood," *Journal of Econometrics*, 157, 248–256. [2]

Owen, A. B. (1988), "Empirical Likelihood Ratio Confidence Intervals for a Single Functional," *Biometrika*, 75, 237–249. [3]

——— (1990), "Empirical Likelihood Ratio Confidence Regions," *The Annals of Statistics*, 18, 90–120. [3]

Qin, J., and Lawless, J. (1994), "Empirical Likelihood and General Estimating Equations," *The Annals of Statistics*, 22, 300–325. [1,3]

Roodman, D. (2009), "A Note on the Theme of Too Many Instruments," *Oxford Bulletin of Economics and Statistics*, 71, 135–158. [1]

Shi, Z. (2016a), "Econometric Estimation with High-Dimensional Moment Equalities," *Journal of Econometrics*, 195, 104–119. [2]

Shi, Z. (2016b), "Estimation of Sparse Structural Parameters with Many Endogenous Variables," *Econometric Reviews*, 35, 1582–1608. [2]

Stock, J. H., and Wright, J. H. (2000), "GMM with Weak Identification," *Econometrica*, 68, 1055–1096. [2]

Stock, J. H., Wright, J. H., and Yogo, M. (2014), "A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments," *Journal of Business & Economic Statistics*, 20, 518–529. [1]

Stock, J. H., and Yogo, M. (2005), "Testing for Weak Instruments in Linear IV Regression," in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, pp. 109–120. [2]

Tang, C. Y., and Leng, C. (2010), "Penalized High-Dimensional Empirical Likelihood," *Biometrika*, 97, 905–920. [2]

Windmeijer, F., Farbmacher, H., Davies, N., and Smith, G. D. (2019), "On the Use of the Lasso for Instrumental Variables Estimation with Some Invalid Instruments," *Journal of the American Statistical Association*, 114, 1339–1350. [2]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [4]

Zhang, C.-H., and Huang, J. (2008), "The Sparsity and Bias of the Lasso Selection in High-Dimensional Linear Regression," *The Annals of Statistics*, 36, 1567–1594. [5]

Zou, H. (2006), "The Adaptive Lasso and its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [4,8]