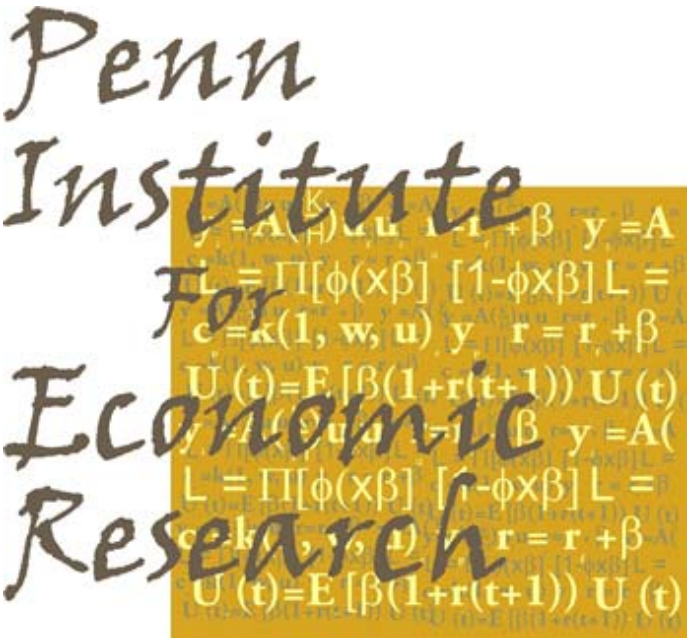

Penn Institute for Economic Research
Department of Economics
University of Pennsylvania
3718 Locust Walk
Philadelphia, PA 19104-6297
pier@econ.upenn.edu
http://economics.sas.upenn.edu/pier

PIER Working Paper 12-045

"Select the Valid and Relevant Moments:
A One-Step Procedure for GMM with Many Moments"

by

Xu Cheng and Zhipeng Liao

http://ssrn.com/abstract=2180906

Select the Valid and Relevant Moments: A One-Step Procedure for GMM with Many Moments

Xu Cheng*
University of Pennsylvania

Zhipeng Liao†
University of California – Los Angeles

First Version: June, 2011
This Version: November, 2012

Abstract

This paper considers the selection of valid and relevant moments for the generalized method of moments (GMM) estimation. For applications with many candidate moments, our asymptotic analysis accommodates a diverging number of moments as the sample size increases. The proposed procedure achieves three objectives in one-step: (i) the valid and relevant moments are selected simultaneously rather than sequentially; (ii) all desired moments are selected together instead of in a stepwise manner; (iii) the parameter of interest is automatically estimated with all selected moments as opposed to a post-selection estimation. The new moment selection method is achieved via an information-based adaptive GMM shrinkage estimation, where an appropriate penalty is attached to the standard GMM criterion to link moment selection to shrinkage estimation. The penalty is designed to signal both moment validity and relevance for consistent moment selection and efficient estimation. The asymptotic analysis allows for non-smooth sample moments and weakly dependent observations, making it generally applicable. For practical implementation, this one-step procedure is computationally attractive.

JEL Classification: C12, C13, C36

Keywords: Adaptive Penalty; GMM; Many Moments; Moment Selection; Oracle Properties; Shrinkage Estimation

*Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, USA. Email: xucheng@sas.upenn.edu

†Department of Economics, UC Los Angeles, 8379 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095. Email: zhipeng.liao@econ.ucla.edu

1 Introduction

In many applications of the generalized method of moments (GMM) estimation, the number of candidate moment conditions is much larger than that of the parameter of interest. However, one typically does not employ all candidate moment conditions due to two types of concerns. First, some moments may be invalid, which cause estimation bias if included. Second, some moment conditions may be redundant. A redundant moment condition does not contain additional information to improve efficiency and results in additional finite-sample bias. Therefore, it is important to identify the valid and relevant (non-redundant) moment conditions, especially when both concerns are elevated in the presence of many candidate moments. The large number of candidate moments raises the need as well as challenges for a consistent moment selection method.

This paper proposes a procedure that consistently selects all valid and relevant moments in an asymptotic framework where the number of candidate moments is allowed to increase with the sample size. This type of asymptotic framework reflects the complexity of the problem and the computation demand associated with a large number of candidate moments. The capacity of the proposed procedure to handle an increasing number of moments justifies its excellent finite-sample performance and mirrors its computational advantage. It only requires computation with the large number of candidate moments once. In contrast, all existing methods only allow for a fixed number of candidate moments in asymptotic analysis and typically require repeated estimations in practical implementation.

The procedure proposed in this paper takes into account validity and relevance simultaneously, whereas all existing procedures first select valid moments and then select the relevant ones out of the former set. A one-step procedure is not only computationally attractive, but also avoids the accumulation of model-selection errors.

The new moment selection method is achieved via an information-based adaptive GMM shrinkage estimation. The moment selection problem is transformed into a penalized GMM (P-GMM) estimation and a novel penalty is designed to incorporate information on both validity and relevance for adaptive estimation. The P-GMM estimation not only consistently select all valid and relevant moment conditions in one step, but also simultaneously estimate the parameter of interest by incorporating all valid and relevant moments and leaving out all invalid or redundant ones. Asymptotic results provide bounds on the penalty level to ensure consistent moment selection. We analyze these bounds as a function of the sample size and the number of moments and provide an algorithm for practical implementation of our procedure.

The moment selection and estimation results developed in the paper allow for (i) non-smooth sample moments, (ii) temporal dependence, and (iii) an increasing number of candidate moments.

High-level assumptions are first provided to capture the main characteristics of the problem and cover all application simultaneously, followed by primitive sufficient assumptions. In the framework of a high-dimensional P-GMM estimation, we develop results on consistency, rate of convergence, super efficiency, and asymptotic distribution, allowing the dimension of the unknown parameter to increase with the sample size. The paper focuses on GMM estimation, but the moment selection procedure works for minimum distance problems as well.

Next, we discuss alternative moment selection procedures available when the number of candidate moments is fixed. The standard J test detects the validity of a given set of moment conditions but it does not specify which ones are invalid and, hence, is not suitable for subset selection. In a seminal paper, Andrews (1999) proposes a moment selection criterion, based on a trade-off between the J statistic and the number of moment conditions, and downward and upward testing procedures. These procedures can consistently select the largest set of valid moments. Andrews and Lu (2001) generalize these methods and study applications to dynamic panel models. Hong, Preston, and Shum (2003) study moment selection based on the generalized empirical likelihood estimation using analogous approaches. On the selection of relevant moments, Hall, Inoue, Jana, and Shin (2007) propose a moment selection criterion that balance the information content and the number of moments. This procedure can be applied to select relevant moments, after all invalid moments are left out in the first step. For applications to DSGE models, Hall, Inoue, Nason, and Rossi (2010) propose two moment selection criteria of this sort to select all valid and relevant impulse response functions for matching estimation. Methods based on the moment selection criteria or sequential testing are stepwise, which requires intensive computation when the candidate set is large.

For the selection of valid moments, the shrinkage procedure proposed by Liao (2011) enjoys great computational advantage over the stepwise methods. If it is followed by a stepwise procedure to select the relevant moments, the computation advantage is diminished. This paper introduces an information-based penalty that enables a shrinkage method to select valid and relevant moments simultaneously rather than sequentially. Most importantly, the current paper allows the number of moments to increase with the sample size, whereas all previous papers select either the valid ones or the relevant ones over a fixed number of candidate moments. Liao (2011) demonstrates that incorporating additional valid moments through shrinkage estimation improves efficiency for strongly identified parameters and improves the rate of convergence for weakly identified parameters. This paper focuses on moment selection, assuming parameters are well identified.

The moment selection problem studied in this paper differs from selecting moments and instrumental variables (IVs) among those known to be valid for mean square error minimization, as in Donald and Newey (2001), Donald, Imbens, and Newey (2009), and Kuersteiner (2002), etc. These

papers focus on moments of similar qualities, but we consider invalid and redundant moments. Our problem is also different from that in Inoue (2006), where moment selection is based on confidence interval coverage. After our procedure having selected all valid and relevant moments, methods from these literatures can be applied subsequently.

Our paper contributes to the study of GMM moment validity and relevance and extends it to a high-dimensional framework. There is a long history on the study of instrumental variable (IV) and GMM moment validity, starting from Sargan (1958), Hansen (1982), Eichenbaum, Hansen, and Singleton (1988). More recent papers related to IV and GMM moment validity include Berkowitz, Caner, and Fang (2012), Conley, Hansen, and Rossi (2012), Doko Tchatoka and Dufour (2012), Guggenberger (2012), Nevo and Rosen (2012), and DiTraglia (2012), among others.

On the GMM moment relevance, Breusch, Qian, Schmidt, and Whyhowski (1999) discuss that, even though a moment is valid and useful by itself, it becomes redundant if its residual after projecting onto an existing set does not contain additional information. In a linear IV model, an IV is redundant if it does not improve the first-stage regression. Im, Ahn, Schmidt, and Wooldridge (1999) study efficient estimation in dynamic panel models in the presence of such redundant moments. Hall and Peixe (2003) study the selection of relevant IVs through canonical correlations and conduct simulations to demonstrate the importance of excluding redundant IVs in finite sample.

There is a large literature on many weak GMM moments and many weak IVs, see Chao and Swanson (2005), Stock and Yogo (2005), Han and Phillips (2006), Hansen, Hausman, and Newey (2005), Newey and Windmeijer (2005), and Andrews and Stock (2006). These papers assume all moments or IVs are valid, albeit weak. Although our paper also let the number of moments increase with the sample size, we allow the unknown invalid moments to mix with valid moments and our objective is moment selection rather than inference.

This paper also complements a growing literature on the application of high-dimensional methods to the linear IV and GMM estimation. Most papers in this literature investigate efficient estimation in the presence of many valid IVs. Belloni, Chernozhukov and Hansen (2010) and Belloni, Chen, Chernozhukov, and Hansen (2012) apply Lasso-type estimation to linear models with many IVs and show that the optimal IV is well approximated by the first stage shrinkage estimation. The boosting method is suggested for IV selection by Bai and Ng (2009). Carrasco (2012) studies efficient estimation with many IVs by regularization techniques. Shrinkage estimation for homoskedastic linear IV models is considered by Chamberlain and Imbens (2004) and Okui (2011). Gautier and Tsybakov (2011) propose a Danzig selector based IV estimator in high dimensional models. Kuersteiner and Okui (2010) recommend using the Mallows averaging methods to approximate the optimal IV in the first-stage regression. Caner (2009) and Liao (2011) study

P-GMM estimation with a fixed number of moments. Caner and Zhang (2012) study adaptive elastic net GMM estimation with an increasing number of parameters. Fan and Liao (2011) investigate P-GMM and penalized empirical likelihood estimation in ultra high dimensional models where the number of parameters increases faster than the sample size and provide a different type of asymptotic results. Our paper contributes to this literature by combining the selection of valid and relevant moments with efficient estimation instead of focusing on the latter, proposing a new information-based adaptive penalty, and considering a general nonlinear GMM estimation with possible non-smooth moment conditions and temporally dependent observations.

The rest of the paper is organized as follows. Section 2 describes the three categories of moment conditions, provides heuristic arguments on how shrinkage estimation distinguishes moments in different categories, and introduces the P-GMM estimator and its information-based penalty. Section 3 derives asymptotic results for the P-GMM estimator, including consistency, rate of convergence, super efficiency, and asymptotic distribution, and discusses their implications on consistent moment selection. Section 4 analyzes the asymptotic magnitudes of the information-based penalty and provides suggestions for practical implementation of the procedure. Section 5 provides finite-sample results through simulation. Section 6 concludes and discusses related topics under investigation. The Appendix includes the proofs and the a simple linear IV model to illustrate the verification of some assumptions.

Notation is standard. Throughout the paper, $\|\cdot\|$ denotes the Euclidean norm; $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues of a matrix A, respectively; $A \equiv B$ means that A is defined as B; the expression $a_n = o_p(b_n)$ means $\Pr(|a_n/b_n| \geq \epsilon) \to 0$ for all $\epsilon > 0$ as n go to infinity; $a_n = O_p(b_n)$ when $\Pr(|a_n/b_n| \geq M) \to 0$ as n and M go to infinity; $a_n \asymp b_n$ means that $(1 + o_p(1))b_n = a_n$ and vice versa; "$\to_p$" and "$\to_d$" denote convergence in probability and convergence in distribution, respectively; and w.p.a.1 abbreviates with probability approaching 1.

2 An Information-Based Penalized GMM Estimator

2.1 Three Categories of Moments Conditions

There exists a vector of moment conditions $n^{-1}\sum_{i=1}^{n} g(Z_i, \theta) : \Theta \to R^{k_n}$ for the estimation of $\theta \in R^{d_\theta}$, where $\{Z_i : i = 1, ..., n\}$ is stationary and ergodic and Z is used generically for Z_i. We allow the number of moments k_n to increase with the sample size. In particular, we are interested in applications where k_n is much larger than d_θ. In this case, it is not restrictive to assume that there exists a relatively small sub-vector of $g(Z, \theta)$, denoted by $g_C(Z, \theta) \in R^{k_0}$, for the identification of θ by $\mathbb{E}[g_C(Z, \theta_o)] = 0$, where θ_o is the true value of θ and $k_0 \geq d_\theta$. Typically, these are the moment

conditions one would use without further exploring the validity and relevance of other candidate moments. They are a "conservative" set of moment conditions to ensure identification, as indicated by the letter "C" in the subscript. Given the identification of θ_o, this paper proposes a moment selection procedure that explore all other candidate moments and yield the *largest* set of valid and relevant moment conditions.

Let $g_D(Z,\theta) \in R^{k_n-k_0}$ denote all of the moments not used for identification, where "D" indicates the "doubt" on the validity and/or relevance of these moments. Without loss of generality, write

$$g(Z,\theta) = \begin{bmatrix} g_C(Z,\theta) \\ g_D(Z,\theta) \end{bmatrix}. \tag{2.1}$$

We also use D to denote the indices of all moments in $g_D(Z,\theta)$. Let $g_\ell(Z,\theta)$ denote an element of $g(Z,\theta)$ indexed by ℓ. A moment is valid if $\mathbb{E}\left[g_\ell(Z_i,\theta_o)\right] = 0$ for $\ell \in D$. Given its validity, a moment is considered to be relevant if adding it yields a more efficient estimator than the one based on $\mathbb{E}\left[g_C(Z_i,\theta_o)\right] = 0$.

By the criteria of validity and relevance, the index set D is divided into three mutually disjoint sets

$$D = A \cup B1 \cup B0, \tag{2.2}$$

where A indexes the set of valid and relevant moments, $B1$ indexes the set of invalid moments, and $B0$ indexes the set of redundant moments. Our objective is to consistently estimate the set A, leaving out all moments indexed by the set $B = B1 \cup B0$.

2.2 Heuristic Arguments for Moment Selection from Shrinkage Estimation

For the purpose of moment selection, a slackness parameter β and its true value β_o are introduced:

$$\beta \equiv \mathbb{E}\left[g_D(Z,\theta)\right] \text{ and } \beta_o \equiv \mathbb{E}\left[g_D(Z,\theta_o)\right]. \tag{2.3}$$

With the introduction of β, all candidate moments, regardless of their validity, can be transformed to moment equalities and stacked into

$$\mathbb{E}\begin{bmatrix} g_C(Z,\theta_o) \\ g_D(Z,\theta_o) - \beta_o \end{bmatrix} = 0. \tag{2.4}$$

This set of moment conditions identifies both θ_o and β_o and enables their joint estimation. Our moment selection strategy is based on the estimation of β_o. Below we first list all desired properties

of the estimator for consistent moment selection, then propose an estimator of β_o that satisfy all of these properties.

Let $\widehat{\beta}_n$ denote an estimator of β_o with sample size n. Let $\widehat{\beta}_{n,\ell}$ and $\beta_{o,\ell}$ denote the estimator and true value of the slackness parameter associated with moment $\ell \in D$. We estimate the desired set A based on the zero elements of $\widehat{\beta}_n$, i.e.,

$$\widehat{A}_n \equiv \{\ell : \widehat{\beta}_{n,\ell} = 0\}. \tag{2.5}$$

For consistent selection of all valid and relevant moments in A, the estimator $\widehat{\beta}_n$ has to satisfy that (i) $\Pr(\widehat{\beta}_{n,\ell} = 0, \forall \ell \in A) \to 1$ and (ii) $\Pr(\widehat{\beta}_{n,\ell} = 0, \forall \ell \in B) \to 0$ for $\ell \in B \equiv B1 \cup B0$.

Table 1. Moment Selection Based on Shrinkage Estimation

Category	True Value	Estimator	Desired Property
$A-$ valid and relevant	$\beta_{o,\ell} = 0$	$\Pr(\widehat{\beta}_{n,\ell} = 0) \to 1$	super efficiency
$B1-$ invalid	$\beta_{o,\ell} \neq 0$	$\Pr(\widehat{\beta}_{n,\ell} = 0) \to 0$	consistency
$B0-$ redundant	$\beta_{o,\ell} = 0$	$\Pr(\widehat{\beta}_{n,\ell} = 0) \to 0$	no shrinkage effect

Table 1 summarizes the properties of the slackness parameter and its estimator for all three categories. First, for the valid and relevant moments (A), $\beta_{o,\ell}$ is 0 and we need its estimator to be 0 w.p.a.1. Having an estimator equal its true value w.p.a.1 is a much stronger result than consistency; the latter only requires the estimator to fall in any local neighborhood of the true value w.p.a.1. This type of super efficiency property can be achieved by shrinking the estimator of $\beta_{o,\ell}$ toward 0 for $\ell \in A$.

Second, for the invalid moments ($B1$), the estimator of $\beta_{o,\ell}$ differs from 0 w.p.a.1 provided it is consistent, because $\beta_{o,\ell}$ is different from 0 in this case. Heavy shrinkage of $\beta_{o,\ell}$ toward 0 for $\ell \in B1$ obviously causes estimation bias. To ensure consistency in this category, the shrinkage effect on the estimator of $\beta_{o,\ell}$ has to be controlled for $\ell \in B1$.

Third, for the redundant moments ($B0$), $\beta_{o,\ell}$ is 0 because the moments are valid. However, the estimator is required to be different from 0 in order to leave out redundant moments. This is completely opposite to the requirement for set A, although $\beta_{o,\ell} = 0$ in both cases. For $\ell \in B0$, the shrinkage effect has to be controlled to prevent the estimator of $\beta_{o,\ell}$ from having a point mass at 0.

To sum up, consistent moment selection requires a shrinkage estimator of the slackness parameter, however, the shrinkage effect has to be reduced when the moment is either invalid or redundant. Such requirements motivate the information-based adaptive shrinkage estimation proposed in this paper. We create a P-GMM estimator that incorporates the measure of validity and relevance for each moment. This estimator is shown to satisfy all the requirements above and yield consistent

moment selection.

2.3 Information Measure and GMM Shrinkage

For GMM estimation based on the transformed equalities in (2.4), define

$$\begin{aligned} \alpha' &\equiv (\theta', \beta'), \\ g(Z, \alpha) &\equiv \begin{bmatrix} g_C(Z, \theta) \\ g_D(Z, \theta) - \beta \end{bmatrix}, \\ \overline{g}_n(\alpha) &\equiv n^{-1} \sum_{i=1}^{n} g(Z_i, \alpha). \end{aligned} \tag{2.6}$$

Moment conditions in (2.4) can be written as $\mathbb{E}[g(Z, \alpha_o)] = 0$. The parameter space of α is $\mathcal{A} \equiv \{(\theta, \beta) : \beta = \mathbb{E}[g_D(Z, \theta)] \text{ and } \theta \in \Theta\}$. For any $\alpha = (\theta', \beta')' \in \mathcal{A}$, assume $|\beta_\ell| \leq C$ for some $C < \infty$ for any element of β.

The efficient estimation and moment selection are simultaneously achieved in the P-GMM estimation

$$\widehat{\alpha}_n = \underset{\alpha \in \mathcal{A}}{\arg\min} \left[\overline{g}_n'(\alpha) W_n \overline{g}_n(\alpha) + \lambda_n \sum_{\ell=1}^{k_n} \omega_{n,\ell} |\beta_\ell| \right], \tag{2.7}$$

where W_n is a $k_n \times k_n$ symmetric weight matrix, $\lambda_n \in R^+$ is a tuning parameter that controls the general penalty level, and $\omega_{n,\ell}$ is an information-based adaptive adjustment for each moment $\ell = 1, ..., k_n$. This is a LASSO type estimator that penalizes the slackness parameter β with respect to its L_1 norm. The L_1 penalty is particularly attractive in our framework because both the GMM criterion and the penalty function are convex in β, which makes the computation of the P-GMM estimator easy in practice.

The novelty of the P-GMM estimation lies in the individual adaptive adjustment $\omega_{n,\ell}$ which incorporates information on both validity and relevance. This individual adjustment is crucial because consistent moment selection requires different degrees of shrinkage for moment conditions in different categories, as listed in Table 1. To this end, define

$$\omega_{n,\ell} = \dot{\mu}_{n,\ell}^{r_1} \, |\dot{\beta}_{n,\ell}|^{-r_2}, \tag{2.8}$$

where $\dot{\mu}_{n,\ell} \geq 0$ is an empirical measure of the information in moment ℓ, $\dot{\beta}_{n,\ell}$ is a preliminary consistent estimator of $\beta_{o,\ell}$, and r_1, r_2 are user-selected positive constants. Before discussing the construction of $\dot{\mu}_{n,\ell}$ and $\dot{\beta}_{n,\ell}$, we first list the implications of this individual adjustment on consistent selection of valid and relevant moments.

First, when data suggest the moment ℓ is relevant, the empirical information measure will be large, which leads to a heavy shrinkage of $\widehat{\beta}_{o,\ell}$ toward 0. In contrast, redundant moments ($B0$) are subject to small shrinkage because $\dot{\mu}_{n,\ell}$ is asymptotically 0 for $\ell \in B0$. This information-based adjustment $\dot{\mu}_{n,\ell}$ differentiates the relevant moments from redundant ones.

Second, when data suggest the moment is likely to be valid, the magnitude of the preliminary estimator $|\dot{\beta}_{n,\ell}|$ will be small as $\dot{\beta}_{n,\ell}$ is consistent, which leads to a large penalty $\omega_{n,\ell}$ and hence, a heavy shrinkage of $\widehat{\beta}_{o,\ell}$ toward 0. In contrast, invalid moments ($B1$) are subject to small shrinkage toward 0, avoiding estimation bias. This validity-based adjustment $|\dot{\beta}_{n,\ell}|$ differentiates the valid moments from invalid ones. The application of $|\dot{\beta}_{n,\ell}|$ for adaptive shrinkage resembles the adaptive LASSO penalty proposed in Zou (2006).

Combining the two types of adaptive adjustment, $\omega_{n,\ell}$ provides a unique data-driven method that separates the valid and relevant moments (A) from the rest. Roughly speaking, the individual adjustment $\omega_{n,\ell}$ is large only when the corresponding moment condition is valid and relevant. In consequence, the estimator of $\beta_{o,\ell}$ is estimated as 0 w.p.a.1 only for $\ell \in A$, yielding a consistent moment selection procedure.

Next, we discuss the construction of the empirical information measure $\dot{\mu}_{n,\ell}$. For this purpose, we first define its population counterpart μ_ℓ, which is associated with the degree of efficiency improvement by adding the moment condition indexed by ℓ. When the moment conditions $\mathbb{E}\left[g_C(Z,\theta_o)\right] = 0$ are used for a GMM estimation of θ, the asymptotic variance of the optimal weighted GMM estimator, denoted by $\dot{\theta}_n$, is

$$\begin{aligned} V_c &\equiv \left[G_c'(\theta_o)\Omega_c^{-1}(\theta_o)G_c(\theta_o)\right]^{-1}, \text{ where} \\ G_c(\theta) &\equiv \frac{\partial \mathbb{E}\left[g_C(Z,\theta)\right]}{\partial \theta'} \text{ and} \\ \Omega_c(\theta) &\equiv \lim_{n\to\infty} \text{Var}\left[n^{-1/2}\sum_{i=1}^{n} g_C(Z_i,\theta)\right]. \end{aligned} \tag{2.9}$$

When another moment $\ell \in D$ is added, define a new variance $V_{c+\ell}$ analogously to V_c but with $\mathbb{E}\left[g_C(Z,\theta)\right]$ replaced by $\mathbb{E}\left[g_{C+\ell}(Z,\theta)\right]$, where $g_{C+\ell}(Z,\theta)$ is a vector that stacks $g_C(Z,\theta)$ and $g_\ell(Z,\theta)$ together. Because the matrix $V_c - V_{c+\ell}$ is positive semi-definite, its eigenvalues are always non-negative. Relevance requires that at least one of its eigenvalues is strictly larger than zero. Thus, we define $\mu_\ell \equiv \lambda_{\max}(V_c - V_{c+\ell})$ as the measure of information in the moment condition indexed by ℓ. When $\mu_\ell > 0$, the moment ℓ is considered to be relevant. A suitable consistent estimator $\dot{\mu}_{n,\ell}$ is

$$\dot{\mu}_{n,\ell} = \lambda_{\max}(\dot{V}_{n,c} - \dot{V}_{n,c+\ell}), \tag{2.10}$$

where $\dot{V}_{n,c}$ and $\dot{V}_{n,c+\ell}$ are consistent estimators of V_c and $V_{c+\ell}$, respectively.[1]

A suitable preliminary estimator $\dot{\beta}_{n,\ell}$ can be obtained from a first-step GMM estimator $\dot{\alpha}_n$, defined as

$$\dot{\alpha}_n = \arg\min_{\alpha \in \mathcal{A}} \left[n \overline{g}_n'(\alpha) W_n \overline{g}_n(\alpha) \right]. \tag{2.11}$$

It is clear that this initial estimator $\dot{\alpha}_n$ can be viewed as a special P-GMM estimator by setting $\lambda_n = 0$ in (2.7) for all n. Hence, as long as the tuning parameter $\lambda_n = 0$ satisfies the sufficient conditions provided below,[2] the properties of the P-GMM estimator, e.g., consistency and rate of convergence, also hold for the first-step GMM estimator $\dot{\alpha}_n$.

Now we return to the P-GMM estimation based on (2.7). To achieve consistent model selection, we first derive conditions on the general tuning parameter λ_n. Intuitively, there exist an upper bound and a lower bound on the convergence rate of λ_n. The upper bound ensures that the penalty is small enough such that $\widehat{\beta}_{o,\ell}$ is consistent with a continuous asymptotic distribution for $\ell \in B = B1 \cup B0$, whereas the lower bound ensures that the penalty is large enough such that $\widehat{\beta}_{o,\ell}$ is super efficient for $\ell \in A$. In Section 3 below, we treat $\omega_{n,\ell}$ as given and derive general bounds (which are functions of $\omega_{n,\ell}$) on λ_n. Section 4 provides explicit bounds for λ_n, following an analysis of the asymptotic orders of $\omega_{n,\ell}$ for moments in different categories, and suggests methods for choosing the tuning parameter in practice.

3 Asymptotic Theory

3.1 Consistency and Rate of Convergence

Define the sample moments

$$\overline{g}_n(\theta) \equiv n^{-1} \sum_{i=1}^{n} g(Z_i, \theta). \tag{3.1}$$

Note that $\overline{g}_n(\theta)$ does not involve centering with the slackness parameter β.

Throughout the paper, let C denote some generic finite positive constant.

Suppose $\mathbb{E}[g(Z,\theta)]$ is differentiable in θ. Define the partial derivative

$$\Gamma(\theta) \equiv \frac{\partial \mathbb{E}[g(Z,\theta)]}{\partial \theta'} \text{ and } \Gamma_o \equiv \Gamma(\theta_o). \tag{3.2}$$

Assumption 1. (i) For any $\varepsilon > 0$, $\inf_{\{\theta \in \Theta : \|\theta - \theta_o\| \geq \varepsilon\}} \|\mathbb{E}[g_C(Z,\theta)]\| > \delta_\varepsilon$ for some $\delta_\varepsilon > 0$.
(ii) $\sup_{\theta \in \Theta} \|\overline{g}_n(\theta) - \mathbb{E}[g(Z,\theta)]\| = o_p(1)$.

[1] When the moments are non-smooth, there are various ways of estimating $G_c(\theta_o)$. The estimation based on random perturbation, for example, is one of the attractive procedures (see, e.g., Chen, Hahn and Liao, 2012).

[2] See Assumptions P1, P2, and P4.

(iii) $\mathbb{E}[g(Z,\theta)]$ is differentiable in θ and $\sup_{||\theta-\theta_o||<\delta_n} ||\gamma_n'[\Gamma(\theta)-\Gamma_o]|| \to 0$ for any $\delta_n \to 0$ and $\gamma_n \in R^{k_n}$ with $||\gamma_n|| = 1$.
(iv) $C^{-1} \leq \lambda_{\min}(\Gamma_o'\Gamma_o) \leq \lambda_{\max}(\Gamma_o'\Gamma_o) \leq C$.
(v) W_n is symmetric and non-stochastic with $C^{-1} \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq C$ w.p.a.1.

Assumption 1(i) is a standard identifiable uniqueness condition for θ_o. Assumption 1(ii) is essentially a uniform law of large numbers (ULLN) and it requires uniform convergence of the sample moments to the population moments. Assumptions 1(iii) and 1(iv) impose standard regularity conditions on the first order derivative of the population moments. Assumptions 1(v) imposes regularity conditions on the weight matrix.

Assumption P1. The tuning parameter λ_n satisfies that $\lambda_n \sum_{\ell \in B1} \omega_{n,\ell} = o_p(1)$.

Assumption P1 imposes an upper bound on λ_n, which ensures that the penalty is small enough such that it does not cause inconsistency of the estimator. By construction, the P-GMM criterion has two parts, where the former is a quadratic form minimized by the true value of the parameter asymptotically and the latter is minimized by $\beta = 0$. When the penalty is too large, it shifts the estimator of $\beta_{o,\ell}$ towards 0 for all ℓ and causes estimation bias for $\beta_{o,\ell} \neq 0$. For this reason, the upper bound in Assumption P1 only involves the invalid moments in $B1$.

Lemma 1 *Suppose Assumptions* 1 *and* P1 *hold. Then,* $||\widehat{\alpha}_n - \alpha_o|| \to_p 0$.

Comment. Define

$$d_n = \min_{\ell \in B1} \left|\beta_{o,\ell}\right|.$$

If the slackness parameters $\beta_{o,\ell}$ for any $\ell \in B1$ satisfy $d_n \geq C > 0$, i.e., slackness parameters for invalid moments do not converge to 0, then using Lemma 1, we deduce that

$$\begin{aligned}
\Pr\left(\min_{\ell \in B1} |\widehat{\beta}_{n,\ell}| > 0\right) &\geq \Pr\left(\min_{\ell \in B1} \left[|\beta_{o,\ell}| - |\widehat{\beta}_{n,\ell} - \beta_{o,\ell}|\right] > 0\right) \\
&\geq \Pr\left(d_n - \max_{\ell \in B1} |\widehat{\beta}_{n,\ell} - \beta_{o,\ell}| > 0\right) \\
&\geq \Pr\left(C - ||\widehat{\alpha}_n - \alpha_o|| > 0\right) \to 1, \text{ as } n \to \infty \qquad (3.3)
\end{aligned}$$

which immediately implies that our method does not select the invalid moment conditions w.p.a.1. From the last inequality in (3.3), we see that the lower bound restriction $\min_{\ell \in B1} \left|\beta_{o,\ell}\right| \geq C$ can be relaxed by taking advantage of the convergence rate of $\widehat{\alpha}_n$.

Next, we derive the rate of convergence of the P-GMM estimator $\widehat{\alpha}_n$, whose dimension increases with the sample size.

Assumption 2. For a sequence of constants $\tau_n \to 0$,

$$\sup_{||\theta-\theta_o||\leq\delta_n} \|\overline{g}_n(\theta) - \mathbb{E}\left[g(Z_i,\theta)\right]\| = O_p(\tau_n)$$

for any $\delta_n \to 0$.

Assumption 2 is a high-level condition on the convergence rate of the empirical process indexed the moment functions. When the number of moment conditions is fixed, Assumption 2 holds with $\tau_n = n^{-1/2}$, following standard empirical process results; see Andrews (1994). Here, the sequence of constants τ_n is introduced to allow for an increasing number of moments $k_n \to \infty$. Assumption 2^* below provides sufficient conditions under which Assumption 2 holds with $\tau_n = \sqrt{k_n/n}$.

In Assumption 2^* below, let $g_\ell(Z,\theta)$ denote an element of $g(Z,\theta)$ indexed by $\ell = 1, ..., k_n$.

Assumption 2^* (i) The observations are i.i.d.
(ii) $g_\ell(Z,\theta)$ is differentiable in θ with the partial derivative denoted by $g_{\theta,\ell}(Z,\theta)$.
(iii) $\max_{\ell\leq k_n} \mathbb{E}(||\sup_{\theta\in\Theta} g_\ell(Z_i,\theta)||^2 + ||\sup_{\theta\in\Theta} g_{\theta,\ell}(Z_i,\theta)||^2) \leq C$.
(iv) Θ is compact.

Lemma 2 *Assumption* 2^* *implies that Assumption* 2(i) *holds with* $\tau_n = \sqrt{k_n/n}$.

For some models, it is easier to verify Assumption $2^\dagger$ below, which is a high-level assumption and can replace Assumptions 1 and 2, in conjecture with Assumption 1(v). Under Assumption $2^\dagger$ below, Assumption P1 can also be omitted when Assumption P2 holds.

Assumption $2^\dagger$ (i) $||\overline{g}_n(\theta_0) - \mathbb{E}[g(Z_i,\theta_0)]|| = O_p(\tau_n)$.
(ii) $||\overline{g}_n(\theta) - \overline{g}_n(\theta_0)|| \asymp ||\theta - \theta_0||$.

For a subset $S \subset D$, let $\omega_{n,S}$ denote a vector that collects $\omega_{n,\ell}$ for all $\ell \in S$.

Assumption P2. The tuning parameter λ_n satisfies that $\lambda_n \|\omega_{n,B1}\| = O_p(\tau_n)$.

Assumption P2 imposes an upper bound on λ_n, under which the penalization is small enough such that the rate of convergence of the P-GMM estimator is determined by the GMM sample moment rather than the penalization. Because $\omega_{n,\ell} > 0$ for any $\ell \in B1$, we see that $\|\omega_{n,B1}\| < \sum_{\ell\in B1} \omega_{n,\ell}$. Hence, Assumption P1 is not strictly weaker than Assumption P2, although the latter imposes a specific rate on $\lambda_n \|\omega_{n,B1}\|$. As in Assumption P1, this condition is only imposed on the invalid moments $B1$ because its purpose is to restrict estimation bias due to penalization.

Lemma 3 (a) *Suppose Assumptions* 1, 2, P1, *hold and* $\lambda_n \|\omega_{n,B1}\| = O_p(1)$. *Then,*

$$\|\widehat{\alpha}_n - \alpha_o\| = O_p(\tau_n + \lambda_n \|\omega_{n,B1}\|).$$

(b) *Suppose Assumptions* 1, 2, P1 *and* P2 *hold. Then,* $\|\widehat{\alpha}_n - \alpha_o\| = O_p(\tau_n)$.
(c) *Parts* (a) *and* (b) *hold with Assumptions* 1, 2, P1 *replaced by Assumptions* 1(v), $2^\dagger$, P2.

Comment. 1. The rate of convergence in Lemma 3 is employed to derive the super efficiency associated with set A (Theorem 1 below) and the asymptotic normality associated with Set B (Theorem 2 below).

2. It is clear that when $\lambda_n = 0$ for all n, Assumptions P1 and P2 are trivially satisfied. Hence if Assumptions 1, 2 or Assumptions 1(v), $2^\dagger$ hold, from Lemma 3 we immediately have

$$\|\dot{\alpha}_n - \alpha_o\| = O_p(\tau_n). \tag{3.4}$$

The convergence rate of the first-step GMM estimator $\dot{\alpha}_n$ is useful to construct the adaptive penalty and tuning parameter, as illustrated in Section 4.

If the slackness parameters $\beta_{o,\ell}$ for any $\ell \in B1$ satisfy $\tau_n = o(d_n)$, using the same arguments in (3.3), we have

$$\Pr\left(\min_{\ell \in B1} |\widehat{\beta}_{n,\ell}| > 0\right) \geq \Pr\left(\frac{d_n}{\tau_n} > \frac{||\widehat{\alpha}_n - \alpha_o||}{\tau_n}\right) \to 1, \text{ as } n \to \infty \tag{3.5}$$

which combined with the result in (3.3), immediately yields the following corollary.

Corollary 1 (Invalid Moments) (a) *Suppose Assumptions* 1, 2, *and* P1 *hold. If we further have* $d_n \geq C > 0$, *then*

$$\Pr\left(\widehat{\beta}_{n,\ell} = 0, \textit{ for any } \ell \in B1\right) \to 0 \textit{ as } n \to \infty. \tag{3.6}$$

(b) *Suppose Assumptions* 1, 2, P1 *and* P2 *hold. If we further have* $\tau_n = o(d_n)$, *then (3.6) holds.*
(c) *Parts* (a) *and* (b) *hold with Assumptions* 1, 2, P1 *replaced by Assumptions* 1(v), $2^\dagger$, P2.

Comment. Corollary 1 implies that the probability that the P-GMM estimation selects any invalid moment condition goes to zero. Part (a) is implied by the consistency of the P-GMM estimator when the magnitudes of the slackness parameters $\beta_{o,\ell}$ for any $\ell \in B1$ are uniformly bounded from below. Part (b) indicates that the invalid moment conditions will not be selected w.p.a.1 even when the magnitudes of the slackness parameters $\beta_{o,\ell}$ for any $\ell \in B1$ converge to zero at certain rate.

3.2 Super Efficiency

We select the valid and relevant moments in A based on shrinkage estimation. To this end, the shrinkage effect has to be large enough to ensure all slackness parameter $\beta_{o,\ell}$ for $\ell \in A$ are estimated as 0 w.p.a.1. Assumption P3 imposes a lower bound on λ_n for this purpose. Assumption P3 only involves the valid and relevant conditions in A because only β_ℓ for $\ell \in A$ are desired to be penalized heavily. This is a key condition to achieve the shrinkage result on moment selection.

Assumption P3. The tuning parameter λ_n satisfies that $\lambda_n^{-1}\tau_n \max_{\ell\in A}\omega_{n,\ell}^{-1} = o_p(1)$.

Theorem 1 (a) *Suppose Assumptions* 1, 2, P1-P3 *hold. Then,*

$$\Pr\left(\widehat{\beta}_{n,\ell} = 0, \textit{ for all } \ell \in A\right) \to 1 \textit{ as } n \to \infty.$$

(b) *Part* (a) *holds with Assumptions* 1, 2, P1 *replaced by Assumptions* 1(v), $2^\dagger$, P2.

Comments: 1. Theorem 1 shows that all valid and relevant moments are simultaneously selected w.p.a.1., allowing for an increasing number of moments in A as $n \to \infty$.

2. Corollary 1 and Theorem 1 are necessary but not sufficient to show that the set A is consistently estimated. For this purpose, it remains to show that any redundant moments in $B0$ are not selected w.p.a.1.

3.3 Asymptotic Normality

Next, we establish the asymptotic distribution of the P-GMM estimator. Following this asymptotic distribution, we conclude that all redundant moments are left out by the moment selection procedure, in addition to the invalid ones covered by Corollary 1. The following assumptions are needed to derive the asymptotic normal distribution.

Define

$$\upsilon_n(\theta) \equiv \overline{g}_n(\theta) - \mathbb{E}\left[g(Z_i,\theta)\right]. \tag{3.7}$$

Assumption 3. (i) For a sequence of constants $\varsigma_n \to 0$,

$$\sup_{||\theta_1-\theta_o||\le\delta_n, ||\theta_2-\theta_o||\le\delta_n} \frac{||\upsilon_n(\theta_1) - \upsilon_n(\theta_2)||}{n^{-1/2} + ||\theta_1 - \theta_2||} = O_p(\varsigma_n) \tag{3.8}$$

for any δ_n converges to 0 slower than τ_n. (ii) $\varsigma_n\tau_n = o(n^{-1/2})$.

Assumption 3(i) is a stochastic equicontinuity condition that accommodates non-smooth moment conditions. Similar stochastic equicontinuity conditions are employed in Pakes and Pollard

(1989), Andrews (2002), and Chen, Linton, van Keilegom (2003), among others. Empirical process results in Pollard (1984), Andrews (1994), and van der Vaart and Wellner (1996) can be used for the verification. When the number of moments is fixed, to ensure the root-n consistency of the GMM estimator, it is sufficient to show Assumption 3(i) holds with $o_p(1)$ on the right hand side. A specific convergence rate ς_n of the modulus of continuity of the empirical process $\upsilon_n(\theta)$ has to be derived in (3.8) to accommodate an increasing number of moments. Assumption 3* below, when applied together with Assumption 2*, provides primitive sufficient conditions under which Assumption 3(i) holds with $\varsigma_n = \sqrt{k_n/n}$.

Assumption 3(ii) restricts the rate at which k_n diverges to ∞. When $\tau_n = \varsigma_n = \sqrt{k_n/n}$, Assumption 3(ii) holds provided $k_n = o(n^{1/2})$, i.e., the number of moment conditions increases slower than $n^{1/2}$.

Assumption 3*. $g_\ell(Z,\theta)$ is twice differentiable in θ with the second partial derivative denoted by $g_{\theta\theta,\ell}(Z,\theta)$ and $\mathbb{E}\left[||\sup_{\theta\in\Theta} g_{\theta\theta,\ell}(Z_i,\theta)||^2\right] \leq C$ for any $\ell \geq 1$.

Lemma 4 (a) *Assumptions* 2* *and* 3* *imply that Assumption* 3(i) *hold with* $\varsigma_n = \sqrt{k_n/n}$.
(b) *Assumptions* 2* *and* 3* *and* $k_n = o(n^{1/2})$ *imply Assumption* 3(ii).

Without loss of generality for the asymptotic results below, write $\beta = (\beta_A, \beta_B)$, where β_A and β_B denote the subvector of β that collects β_ℓ for $\ell \in A$ and $\ell \in B$, respectively. The set $B = B1 \cup B0$ includes both the invalid moments $B1$ and the redundant moments $B0$. Let $\widehat{\beta}_{A,n}$ and $\widehat{\beta}_{B,n}$ denote the P-GMM estimators of β_A and β_B, respectively. Theorem 1 shows $\widehat{\beta}_{A,n} = 0$ w.p.a.1. It remains to develop the asymptotic distribution of $\widehat{\beta}_{B,n}$, together with the distribution of $\widehat{\theta}_n$. To this end, define

$$\alpha_B' \equiv (\theta', \beta_B'). \tag{3.9}$$

Now we stack all moment conditions and define

$$g(Z,\alpha_B) \equiv \begin{bmatrix} g_C(Z,\theta) \\ g_A(Z,\theta) \\ g_B(Z,\theta) - \beta_B \end{bmatrix}, \tag{3.10}$$

where $g_A(Z,\theta)$ denotes the valid and relevant moments and $g_B(Z,\theta)$ denotes the invalid or redundant moments. Because $g(Z,\alpha_B)$ is linear in β_B, the partial derivative of $\mathbb{E}[g(Z_i,\alpha_B)]$ w.r.t. α_B only depends on θ. Define

$$\Gamma_\alpha(\theta) \equiv \frac{\partial \mathbb{E}[g(Z,\alpha_B)]}{\partial \alpha_B'} \text{ and } \Gamma_\alpha \equiv \Gamma_\alpha(\theta_o). \tag{3.11}$$

The difference between $g(Z, \alpha_B)$ and $g(Z, \alpha)$ defined in (2.6) is that $\beta_A = 0$ in $g(Z, \alpha_B)$. Because the true value of β_A is 0, $g(Z, \alpha_o) = g(Z, \alpha_{B,o})$ by definition. Hence, the sample average of $g(Z, \alpha_{B,o})$ can be written as $\overline{g}_n(\alpha_o)$. Define the long-run variance of the sample moments

$$\Omega_n \equiv \text{Var}(n^{1/2}\overline{g}_n(\alpha_o)). \tag{3.12}$$

For i.i.d. observations, this variance matrix is simplified to $\Omega_n = \mathbb{E}[g(Z, \alpha_o)g'(Z, \alpha_o)]$ for all n.

Assumption 4. (i) For any $\gamma_n \in R^{k_n}$ and $||\gamma_n|| = 1$,

$$\gamma_n' \sqrt{n}\Omega_n^{-1/2}\overline{g}_n(\alpha_o) \to_d N(0, 1).$$

(ii) $C^{-1} \leq \lambda_{\min}(\Omega_n) \leq \lambda_{\max}(\Omega_n) \leq C$ for all n.
(iii) $C^{-1} \leq \lambda_{\min}(\Gamma_\alpha'\Gamma_\alpha) \leq \lambda_{\max}(\Gamma_\alpha'\Gamma_\alpha) \leq C$.

Assumption 4(i) assumes a triangular array central limit theorem for scalar random variables. Assumption 4(ii) requires that the long-run variance matrix Ω_n is positive definite and bounded for all n. Assumption 4(iii) imposes the same regularity condition to $\Gamma_\alpha'\Gamma_\alpha$.

Assumption P4. The tuning parameter λ_n satisfies that $\lambda_n \|\omega_{n,B}\| = o_p(n^{-1/2})$.

Assumption P4 imposes an upper bound on λ_n, which ensures that the P-GMM estimator of any finite-dimensional parameter has a mean-zero asymptotic normal distribution. Assumption P4 implies Assumption P2. It also implies Assumption P1 given that $k_n = o(n)$.

Define a covariance matrix

$$\Sigma_n \equiv \left(\Gamma_\alpha' W_n \Gamma_\alpha\right)^{-1} \left(\Gamma_\alpha' W_n \Omega_n W_n \Gamma_\alpha\right) \left(\Gamma_\alpha' W_n \Gamma_\alpha\right)^{-1}. \tag{3.13}$$

Theorem 2 (a) *Suppose Assumptions* 1-4 *and* P3-P4 *hold. Then,*

$$\gamma_n' \sqrt{n}\Sigma_n^{-1/2}(\widehat{\alpha}_{B,n} - \alpha_{B,0}) \to_d N(0, 1)$$

for any $\gamma_n \in R^{k_n}$ *with* $||\gamma_n|| = 1$.
(b) *Part* (a) *holds with Assumptions* 1 *and* 2 *replaced by Assumptions* 1(v) *and* $2^\dagger$.

Comments. 1. The asymptotic distribution of the P-GMM estimator is derived by a perturbation on a local parameter space (see, e.g., Shen (1997)), allowing for non-smooth sample moments and an increasing number of parameters.

2. Theorem 2, in conjuncture with the Cramer-Wold device, yields the asymptotic distribution of $\widehat{\theta}_n$. This asymptotic distribution can be applied to conduct inference for the parameter of interest θ_o. Although the primary purpose of the paper is moment selection, the P-GMM estimator automatically produces an estimator for θ_o, imposing all valid and relevant moment conditions in the estimation and leaving out all invalid or redundant moments. Therefore, the P-GMM estimator of θ_o is asymptotically equivalent to the ideal but infeasible "oracle" estimator one would get with the complete knowledge of which moments are valid and relevant. Simulation results in Section 5 demonstrate that the P-GMM estimator of θ_o is comparable to this oracle estimator in their finite-sample performances.

Because $\widehat{\beta}_{n,\ell}$ has an asymptotic normal distribution for $\ell \in B$, the probability that $\widehat{\beta}_{n,\ell} = 0$ approaches 0 for any $\ell \in B$. The set B includes both the invalid moments $B1$ and the redundant moments $B0$. This result is particularly important for the latter, which is not covered by Corollary 1. Corollary 2 states that any fixed subset of redundant moments are left out w.p.a.1 by the moment selection procedure.

Corollary 2 (Redundant Moments) (a) *Suppose Assumptions* 1-4 *and* P3-P4 *hold. Then,*

$$\Pr\left(\widehat{\beta}_{n,\ell} = 0, \text{ for any } \ell \in \dot{B}0\right) = 0 \text{ as } n \to \infty$$

where $\dot{B}0$ *is any fixed subset of* $B0$.
(b) *Part* (a) *holds with Assumptions* 1 *and* 2 *replaced by Assumptions* 1(v) *and* $2^\dagger$.

Comment. Combining Theorem 1 and Corollaries 1 and 2, we conclude that, by the P-GMM estimation, the invalid moment conditions are not selected with probability approaching 1, the valid and relevant moment conditions are selected with probability approaching 1 and any subset of the redundant moment conditions are not selected with probability approaching 1.

Finally, we consider the estimation of A by combining results in Theorem 1 and Corollaries 1 and 2. Theorem 1 implies that

$$\Pr(A \subseteq \widehat{A}_n) \to 1 \tag{3.14}$$

as $n \to \infty$, i.e., all valid and relevant moments are selected asymptotically. On the other hand,

$$\begin{aligned} \Pr(\widehat{A}_n \subseteq A) &= 1 - \Pr(\bigcup_{\ell \in B} \widehat{\beta}_{n,\ell} = 0) \\ &\geq 1 - \Pr(\bigcup_{\ell \in B1} \widehat{\beta}_{n,\ell} = 0) - \Pr(\bigcup_{\ell \in B0} \widehat{\beta}_{n,\ell} = 0). \end{aligned} \tag{3.15}$$

Note that

$$\Pr(\bigcup_{\ell\in B1}\widehat{\beta}_{n,\ell}=0)\to 0 \tag{3.16}$$

by Corollary 1. For the redundant moments,

$$\begin{aligned}\Pr(\bigcup_{\ell\in B0}\widehat{\beta}_{n,\ell}=0)\ &\leq\ \sum_{\ell\in B0}\Pr(\widehat{\beta}_{n,\ell}=0),\ \text{where}\\ \Pr(\widehat{\beta}_{n,\ell}=0)\ &\to\ 0\ \text{for any}\ \ell\in B0.\end{aligned} \tag{3.17}$$

By (3.17),

$$\Pr(\bigcup_{\ell\in B0}\widehat{\beta}_{n,\ell}=0)\to 0 \tag{3.18}$$

provided that the cardinality of $B0$ is bounded or it increases slowly such that $\Pr(\widehat{\beta}_{n,\ell}=0)\to 0$ for any $\ell\in B0$ implies that $\sum_{\ell\in B0}\Pr(\widehat{\beta}_{n,\ell}=0)\to 0$. Under this condition, (3.14)-(3.18) together yield

$$\lim_{n\to\infty}\Pr\left(\widehat{A}_n=A\right)=1. \tag{3.19}$$

4 Selection of the Tuning Parameter

The asymptotic results established in the previous section provide restrictions on the tuning parameter λ_n. These restrictions are implicit in the sense that they depend on the individual information-based adaptive penalties $\omega_{n,\ell}$ defined in (2.8), whose asymptotic magnitudes depend on the validity as well as relevance of the moment condition ℓ by construction. In this section, we analyze these individual penalties under general conditions and provide an explicit asymptotic bounds for the tuning parameter λ_n. These explicit bounds only depend on the sample size and the total number of moments. A practical choice is suggested. At the end of this session, an algorithm is listed for the practical implementation of the procedure.

To construct the adaptive penalty $\omega_{n,\ell}$ in (2.8), preliminary estimators $\dot{\mu}_{n,\ell}$ and $\dot{\beta}_{n,\ell}$ are employed, as defined in (2.10) and (2.11). From (3.4), we see that the first-step GMM estimators $\dot{\beta}_{n,\ell}$ have the joint τ_n rate of convergence. To analyze the asymptotic order of $\omega_{n,\ell}$, we assume that the preliminary estimators $\dot{\mu}_{n,\ell}$ are $\sqrt{n}$ consistent for their true values. Note that these preliminary estimators rely on standard GMM estimation with a fixed number of moments and a fixed number of well-identified unknown parameters.

Assumption 5. $\dot{\mu}_{n,\ell}=\mu_\ell+O_p(n^{-1/2})$ for any $\ell\in D$.

In Assumption 5, we do not specify the nature of the information measure μ_ℓ. It is clear that the definition of the index set $B0$ can be generalized to be $B0=\{\ell:\mu_\ell=O(n^{-1/2})\}$, because

in both cases ($\mu_\ell = 0$ and $\mu_\ell = O(n^{-1/2})$), the empirical information measure $\dot{\mu}_{n,\ell}$ goes to zero at the root-n rate. We call a moment condition ℓ nearly redundant, if $\mu_\ell = O(n^{-1/2})$. The near redundancy concept allows for not exact redundancy in finite-sample.

4.1 Practical Choice of the Tuning Parameter

Now we discuss practical choice of the penalty coefficient λ_n under the guidance of Assumptions P3 and P4. To investigate the bound suggested by Assumption P3, note that $\max_{\ell \in A} \omega_{n,\ell}^{-1} = O_p(\max_{\ell \in A} |\dot{\beta}_{n,\ell}|^{r_2})$. Because $\beta_{o,\ell} = 0$ for all $\ell \in A$, $\max_{\ell \in A} |\dot{\beta}_{n,\ell}| \leq (\sum_{\ell \in A} |\dot{\beta}_{n,\ell} - \beta_{o,l}|^2)^{1/2} \leq ||\dot{\beta}_n - \beta_o|| = O_p(\tau_n)$, where the $O_p(\tau_n)$ term follows from (3.4). Hence, Assumption P3 suggests $\lambda_n^{-1} = o(\tau_n^{-(r_2+1)})$. When $\tau_n = (k_n/n)^{1/2}$ as given in Lemma 4, the lower bound for λ_n is

$$\lambda_n^{-1} = o((k_n/n)^{-r_2/2-1/2}). \tag{4.1}$$

To investigate the bound suggested by Assumption P4, note that $\min_{\ell \in B1} |\dot{\beta}_{n,\ell}| > d_n - ||\dot{\alpha}_n - \alpha_o||$ using arguments analogous to that in (3.3). Note that $||\dot{\alpha}_n - \alpha_o|| = O_p(\tau_n)$ by (3.4). Hence, $d_n^{-1} \min_{\ell \in B1} |\dot{\beta}_{n,\ell}| > 1 - o_p(1)$ provided that $\tau_n = o(d_n)$. Given that we have shown $d_n^{-1} \min_{\ell \in B1} |\dot{\beta}_{n,\ell}|$ is bounded away from 0 for $\ell \in B1$, we know that $||\omega_{B1}|| \leq O_p(k_n^{1/2} d_n^{-r_2/2})$. Suppose $r_1 - r_2$ is large such that $\omega_{n,\ell} = o_p(1)$ for $\ell \in B0$, then Assumption P4 suggests

$$\lambda_n = o(d_n^{r_2/2} k_n^{-1/2} n^{-1/2}). \tag{4.2}$$

In practice, the choice of λ_n is a balance of the two conditions in (4.1) and (4.2). On the one hand, selecting valid and relevant moments require λ_n to converge to 0 slower than $k_n^{-r_2/2-1/2} n^{r_2/2+1/2}$ and as slow as possible. On the other hand, leaving out invalid or redundant moments requires λ_n to converge to 0 faster than $k_n^{-1/2} n^{-1/2}$ and as fast as possible.[3] We recommend balancing these two requirements by choosing

$$\lambda_n = c \; k_n^{r_2/4} \; n^{-1/2-r_2/4}, \tag{4.3}$$

where c is a loading coefficient. Asymptotic theories do not impose requirement on c. In practice, one common approach to choose level parameters of this sort is through cross validation.

4.2 Algorithm for Empirical Implementation

For practical implementation, our procedure is executed in the following steps.
(1). A preliminary estimator $\dot{\beta}_{n,\ell}$ follows from (2.11) for all $\ell \in D$.

[3] Because d_n is unknown, we use the scale coefficient c to accommodate its finite-sample effect.

(2). Estimate the information measure $\mu_{n,\ell}$ by (2.10) and construct the adaptive individual penalty $\omega_{n,\ell}$ by (2.8) for a given pair of (r_1, r_2), for all $\ell \in D$. The constants satisfy $r_1 > r_2 > 0$.
(3). Construct the general penalty λ_n by (4.3) with $c > 0$.
(4). Estimate $\widehat{\alpha}_n' = (\widehat{\theta}_n', \widehat{\beta}_n')$ by the P-GMM estimator defined in (2.7).
(5). The indices of the valid and relevant moments are estimated by $\widehat{A}_n = \{\ell : \widehat{\beta}_{n,\ell} = 0\}$.

5 Simulation

For finite-sample investigation, we consider a simple linear regression model

$$Y_1 = Y_2\theta_o + u, \tag{5.1}$$

where $Y_1, Y_2 \in R$ are endogenous and $\theta_o \in R$ is the parameter of interest. For applications with exogenous variables on the right hand side, Y_1 and Y_2 can be viewed as the residuals obtained after projections onto these exogenous variables. Valid and relevant IVs $Z_C \in R^2$ are available for the identification of θ_o. In addition, a vector of candidate IVs $Z_D \in R^{10}$ are considered, without knowing their validity or relevance. The candidate IVs comprise of $Z_D = (Z_A, Z_{B0}, Z_{B1})$, where $Z_A \in R^2$ is valid and relevant, $Z_{B0} \in R^4$ is redundant, and $Z_{B1} \in R^4$ is invalid. Specifically, the relationship between the endogenous variable Y_2 and the valid and relevant IVs Z_C and Z_A is

$$Y_2 = \pi_C' Z_C + \pi_A' Z_A + v. \tag{5.2}$$

We generate

$$(Z_C, Z_A, Z_{B0}, Z_{B1}^*, u, v) \sim N(0, \Sigma), \text{ where } \Sigma = diag(\Sigma_{AC}, \Sigma_B, \Sigma_{uv}). \tag{5.3}$$

By construction, $(Z_C, Z_A, Z_{B0}, Z_{B1}^*)$ are all valid, but only Z_C and Z_A are relevant based on (5.2). The invalid IVs Z_{B1} are obtained by contaminating Z_{B1}^* with the structural error u. Specifically,

$$Z_{B1,\ell} = Z_{B1,\ell}^* + c_\ell \cdot u, \tag{5.4}$$

where $Z_{B1,\ell}$ and $Z_{B1,\ell}^*$ are the ℓ-th element of Z_{B1} and Z_{B1}^*, respectively. The structure of (5.4) indicates that the degree of endogeneity of an invalid IV varies with the coefficient c_ℓ, which is given below.

Parameters in the data generating process are as follows. (i) $\theta_o = 0.5$, (ii) $\pi_C = (\pi_o, 0.1)'$, where the value $\pi_o = 0.1$ or 0.3 to experiment different identification strength, (iii) $\pi_A = (0.5, 0.5)$, (iv)

Σ_{AC} is a 4×4 matrix with the (i,j)-th element being $0.2^{|i-j|}$, (v) Σ_B is an 8×8 identity matrix, (vi) $\Sigma_{u,v}$ is a 2×2 matrix with diagonal elements $(0.5, 1)$ and off-diagonal elements $(0.6, 0.6)$, (vii) For $c_o = 0.2$ or 0.5 and $\ell = 1, ..., 4$, the coefficients in (5.4) are $c_1 = c_o$, $c_2 = c_o + (0.8 - c)/3$, $c_3 = c_o + 2(0.8 - c)/3$, $c_4 = 0.8$. A larger value of c_o is associated with stronger endogeneity of the invalid IVs.

For each specification of (π_o, c_o), we generate i.i.d. observations with sample size $n = 250$ and $n = 2500$. To construct the information-based penalty in (2.8), the user-selected constants are $r_1 = 3$ and $r_2 = 2$. The preliminary estimator $\dot{\mu}_{n,\ell}$ is constructed by sample analogs of the variance matrix and the preliminary estimator $\dot{\beta}_{n,\ell}$ follows from (2.11). The number of simulation repetition is 5000. The projected scaled sub-gradient method (active-set variant) method proposed in Schmidt (2010) is employed to solve the minimization problem in the GMM shrinkage estimation.

Table 2. Performance of Moment Selection by GMM Shrinkage Estimation

	$\pi_o = 0.3$							
	$n = 250$				$n = 2500$			
$c_o = 0.5$	.0000	.6888	.1878	.1234	.0000	.9606	.0284	.0011
$c_o = 0.2$	.0006	.6874	.1884	.1236	.0000	.9602	.0278	.0120
	$\pi_o = 0.1$							
	$n = 250$				$n = 2500$			
$c_o = 0.5$	.0016	.4944	.4932	.0108	.0000	.9028	.0946	.0026
$c_o = 0.2$	.0112	.4908	.4866	.0114	.0000	.9026	.0950	.0024

For each parameter combination, four numbers are reported. The first number is the probability of "selecting any invalid IVs". The second number is the probability of "selecting all valid and relevant IVs". The third number is the probability of "selecting all valid and relevant IVs plus some redundant IVs". The fourth column is the probability of all other events.

Table 2 presents the finite-sample performance of the moment selection procedure by the GMM shrinkage estimation. We first look at the case with strong identification ($\pi_o = 0.3$), strong endogeneity of invalid IVs ($c_o = 0.5$), and moderate sample size ($n = 250$). In this case, the probability of any invalid IVs being selected is about 0. Hence, the shrinkage procedure succeeds in selecting only the valid IVs. With a probability of 0.69, Z_A is the set of IVs selected. With a probability of 0.19, Z_A plus some elements in Z_{B0} are selected. This implies that with a probability of 0.88, the shrinkage procedure selects *all* of the valid and relevant IVs. When sample size is $n = 2500$, the probability of selecting all and only the valid and relevant IVs is 0.96, whereas the probability of selecting invalid IVs is 0 and the probability of selecting redundant IVs is as low as 0.03. Reducing the degree of identification and reducing the degree of endogeneity for the invalid IVs both make moment selection more challenging In the extreme case with relatively weak identification

($\pi_0 = 0.1$) and weak endogeneity ($c_o = 0.2$), the procedure is robust at not including any invalid IVs but tend to include some redundant ones. The probability of including redundant IVs is reduced significantly when sample size increases.

Table 3. Finite Sample Bias (BS), Standard Deviation (SD) and RMSE (RE) of Estimators of θ_o

	Automatic						Conservative					
(π_o, c_o)	$n = 250$			$n = 2500$			$n = 250$			$n = 2500$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.3 .5)	.0042	.0815	.0816	.0001	.0232	.0232	-.0013	.1614	.1614	.0003	.0501	.0501
(.3 .2)	.0042	.0816	.0817	.0001	.0232	.0232	-.0013	.1614	.1614	.0003	.0501	.0501
(.1 .5)	.0026	.0856	.0857	.0001	.0248	.0248	-.0035	.2613	.2613	.0006	.0786	.0786
(.1 .2)	.0030	.0847	.0848	.0001	.0248	.0248	-.0035	.2613	.2613	.0006	.0786	.0786

	Pooled (infeasible)						Aggressive					
(π_o, c_o)	$n = 250$			$n = 2500$			$n = 250$			$n = 2500$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.3 .5)	.0049	.0754	.0755	.0004	.0232	.0232	.1203	.1337	.1799	.1191	.0433	.1267
(.3 .2)	.0049	.0754	.0755	.0004	.0232	.0232	.0931	.1187	.1508	.0902	.0378	.0977
(.1 .5)	.0057	.0810	.0812	.0004	.0248	.0248	.1377	.1422	.1979	.1364	.0463	.1441
(.1 .2)	.0057	.0810	.0812	.0004	.0248	.0248	.1068	.1265	.1655	.1034	.0404	.1110

	Post-Shrinkage						Oracle (infeasible)					
(π_o, c_o)	$n = 250$			$n = 2500$			$n = 250$			$n = 2500$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.3 .5)	.0054	.0904	.0906	.0002	.0237	.0237	.0017	.0744	.0744	.0000	.0231	.0231
(.3 .2)	.0054	.0904	.0906	.0002	.0237	.0237	.0017	.0744	.0744	.0000	.0231	.0231
(.1 .5)	.0028	.0842	.0842	.0000	.0247	.0247	.0021	.0800	.0800	.0000	.0247	.0247
(.1 .2)	.0033	.0821	.0821	.0000	.0247	.0247	.0021	.0800	.0800	.0000	.0247	.0247

(i) The "automatic" estimation is obtained simultaneously with moment selection. (ii) The "conservative" estimation uses Z_C. (iii) The "pooled" estimation uses all valid IVs, including Z_C, Z_A, and Z_{B0}. (iv) The "aggressive" estimation uses all available IVs, including invalid ones. (v) The "post-shrinkage" estimation uses Z_C plus IVs selected by the shrinkage procedure. (vi) The "oracle" estimation uses Z_C and Z_A.

The P-GMM estimator proposed in this paper produces an automatic estimate of θ_o in the shrinkage estimation. Table 3 summaries finite-sample properties of this estimator, denoted by "automatic" in Table 3, and compares it with several alternative estimators. Some of the alternative estimators are infeasible, but serve as good benchmarks. To show the efficiency improvement by using more relevant and valid IVs, we compare the "automatic" estimator with a "conservative" estimator, which only uses Z_C without further exploring information in other candidate IVs. This comparison shows that the "automatic" estimator enjoys smaller standard deviation and root mean square error (RMSE) than the "conservative" estimator in all scenarios considered. To show the

finite-sample improvement by excluding redundant IVs, the "automatic" estimator is compared to a "pooled" estimator, which uses all valid IVs Z_C, Z_A, and Z_{B0}. This comparison indicates that the "automatic" estimator has smaller finite-sample bias. Note that this "pooled" estimator is actually infeasible because it excludes all invalid IVs and include all valid IVs. Table 2 suggests that there is a non-negligible probability that some valid and relevant IVs are not selected when the sample size is moderate, which is why the standard deviation of the "automatic" estimator is slighter larger than that of the "pooled" estimator for $n = 250$. This difference disappears for $n = 2500$. To show the importance of excluding invalid IVs, the "automatic" estimator is compared to an "aggressive" estimator, which uses all candidate IVs regardless of their validity. This comparison suggests that including invalid IVs increases finite-sample bias as expected. The "post-shrinkage" estimator is the GMM estimator uses all IVs selected by the shrinkage procedure. The difference between the "automatic" estimator and the "post-shrinkage" estimator is small, although the former tends to have smaller bias and the latter has smaller standard deviation in some cases. Finally, an important comparison is between the "automatic" estimator and the infeasible "oracle" estimator, which uses the desirable IVs Z_C and Z_A. This comparison indicates that the finite-sample properties of the "automatic" estimator are comparable to those of the "oracle" estimator, even for a moderate sample size, and the two are basically the same when the sample size is large.

In sum, the GMM shrinkage estimator proposed in this paper not only produces consistent moment selection, as indicated in Table 2, but also automatically estimate the parameter of interest. Table 3 shows that this "automatic" estimator dominates all other feasible estimators and it is comparable to the ideal but infeasible "oracle" estimator in terms of finite-sample bias and variance.

6 Conclusion

This paper studies moment selection when the number of moments diverges with the sample size, allowing for both invalid and redundant moments in the candidate set. We show that the moment selection problem can be transformed to a P-GMM estimation problem, which consistently selects the subset of valid and relevant moments and automatically estimates the parameter of interest. In consequence, the P-GMM estimator is not only robust to the potential mis-specification introduced by invalid moments but also robust to the possible finite-sample bias introduced by redundant moments.

An interesting and challenging question related to this paper is inference on the parameter of interest θ_o when moment selection is necessary. Although the asymptotic distribution developed in this paper can be used to conduct inference on θ_o, this limiting distribution ignores the moment

selection error in finite sample. As a result, a robust inference procedure with correct asymptotic size is an important issue for the P-GMM estimator. This is related to the post model selection inference problem investigated by Leeb and Pötscher (2005, 2008), Andrews and Guggenberger (2009, 2010), Guggenberger (2010), Belloni, Chernozhukov, and Hansen (2011), and McCloskey (2012), among others. Robust inference on the parameter of interest is beyond the scope of this paper and investigated in future research.

7 Appendix

7.1 Proofs on Asymptotic Results

For notation simplicity, define

$$m(\theta) \equiv \mathbb{E}\begin{bmatrix} g_C(Z_i,\theta) \\ g_D(Z_i,\theta) \end{bmatrix} \text{ and } m(\alpha) \equiv \mathbb{E}\begin{bmatrix} g_C(Z_i,\theta) \\ g_D(Z_i,\theta) - \beta \end{bmatrix}. \tag{7.1}$$

Proof of Lemma 1. Define $\upsilon_n(\theta) \equiv \overline{g}_n(\theta) - m(\theta)$. Note that $\upsilon_n(\theta) = \overline{g}_n(\alpha) - m(\alpha)$ for any $\alpha \in \mathcal{A}$, and $\upsilon_n(\theta_o) = \overline{g}_n(\alpha_o)$ because $m(\alpha_o) = 0$. Hence,

$$\overline{g}_n'(\alpha_o) W_n \overline{g}_n(\alpha_o) = \upsilon_n(\theta_o)' W_n \upsilon_n(\theta_o) = o_p(1) \tag{7.2}$$

by Assumption 1(ii) and $\lambda_{\max}(W_n) \leq C$ for some $C < \infty$ w.p.a.1, where the latter holds by Assumptions 1(v).

The definition of $\widehat{\alpha}_n$ implies that

$$\overline{g}_n'(\widehat{\alpha}_n) W_n \overline{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell=1}^{k_n} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \leq \overline{g}_n'(\alpha_o) W_n \overline{g}_n(\alpha_o) + \lambda_n \sum_{\ell=1}^{k_n} \omega_{n,\ell} \left|\beta_{o,\ell}\right|. \tag{7.3}$$

This in turn yields

$$||\overline{g}_C(\widehat{\theta}_n)||^2 + ||\overline{g}_D(\widehat{\theta}_n) - \widehat{\beta}_n||^2 = ||\overline{g}(\widehat{\alpha}_n)||^2 = o_p(1) \tag{7.4}$$

because (i) $\lambda_n \sum_{\ell=1}^{k_n} \omega_{n,\ell} |\widehat{\beta}_{n,\ell}| > 0$, (ii) $\overline{g}_n'(\alpha_o) W_n \overline{g}_n(\alpha_o) = o_p(1)$ by (7.2), (iii) $\beta_{o,\ell} = 0$ for $\ell \notin B1$, (iv) $\lambda_n \sum_{\ell \in B1} \omega_{n,\ell} \left|\beta_{o,\ell}\right| = o_p(1)$ by Assumption P1 and that $|\beta_{o,\ell}|$ is bounded, and (iv) $\lambda_{\min}(W_n) \geq C$ for some $C > 0$ w.p.a.1.

Write

$$\overline{g}(\widehat{\alpha}_n) = \begin{bmatrix} \overline{g}_C(\widehat{\theta}_n) \\ \overline{g}_D(\widehat{\theta}_n) - \widehat{\beta}_n \end{bmatrix}. \tag{7.5}$$

Then, $||\overline{g}_C(\widehat{\theta}_n)|| = o_p(1)$ and $||\overline{g}_D(\widehat{\theta}_n) - \widehat{\beta}_n|| = o_p(1)$ by (7.4). Together with the triangle inequality and Assumptions 1(i), 1(ii), $||\overline{g}_C(\widehat{\theta}_n)|| = o_p(1)$ implies that $\widehat{\theta}_n \to_p \theta_o$.

To show the consistency of $\widehat{\beta}_n$, we first show that under Assumptions 1(iii) and 1(iv), there is

$$\sup_{||\theta - \theta_o|| \leq \delta_n} \|\mathbb{E}\left[g(Z_i,\theta) - g(Z_i,\theta_o)\right]\| \to 0 \text{ for any } \delta_n \to 0. \tag{7.6}$$

For this purpose, we notice that

$$\|\mathbb{E}\left[g(Z_i,\theta)-g(Z_i,\theta_o)\right]\| \leq \left[||\gamma'_{\theta,n}[\Gamma(\widetilde{\theta})-\Gamma_o]|| + ||\gamma'_{\theta,n}\Gamma_o||\right] \times ||\theta-\theta_o|| \tag{7.7}$$

where $\widetilde{\theta}$ lies between θ and θ_o and $\gamma_{\theta,n} = (\theta-\theta_o)/||\theta-\theta_o||$. From the inequality in (7.7) and Assumptions 1(iii), 1(iv), we deduce that

$$\sup_{||\theta-\theta_o||\leq\delta_n} \|\mathbb{E}\left[g(Z_i,\theta)-g(Z_i,\theta_o)\right]\| \leq \delta_n \sup_{||\theta-\theta_o||\leq\delta_n} \left[||\gamma'_{\theta,n}[\Gamma(\widetilde{\theta})-\Gamma_o]|| + C^{1/2}\right] \to 0 \tag{7.8}$$

for any $\delta_n \to 0$. This proves (7.6).

Let $\mathbb{E}_Z[\cdot]$ denote the expectation taking with respect to the distribution of Z. To show the consistency of $\widehat{\beta}_n$ note that

$$\begin{aligned}
\left\|\widehat{\beta}_n - \beta_o\right\| &\leq ||\overline{g}_D(\widehat{\theta}_n)-\beta_o|| + ||\widehat{\beta}_n - \overline{g}_D(\widehat{\theta}_n)|| \\
&\leq ||\overline{g}_D(\widehat{\theta}_n) - \mathbb{E}_Z[g_D(Z_i,\widehat{\theta}_n)]|| + ||\mathbb{E}_Z[g_D(Z_i,\widehat{\theta}_n)] - \mathbb{E}_Z[g_D(Z_i,\theta_o)]|| + o_p(1) \\
&= o_p(1),
\end{aligned} \tag{7.9}$$

where the first inequality follows from the triangle inequality, the second inequality holds by the triangle inequality, $\mathbb{E}[\overline{g}_D(\theta_o)] = \beta_o$, and (7.4), and the equality follows from Assumptions 1(ii), (7.6), and the consistency of $\widehat{\theta}_n$. This completes the proof. □

Proof of Lemma 3. Define $b_n \equiv \lambda_n \|\omega_{n,B1}\|$. We first prove part (a). Assumption 2, together with Assumption 1(v), implies that

$$\overline{g}'_n(\alpha_o) W_n \overline{g}_n(\alpha_o) = O_p(\tau_n^2). \tag{7.10}$$

The inequalities in (7.3) and the equation (7.10) imply that

$$\overline{g}'_n(\widehat{\alpha}_n) W_n \overline{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell\in B1} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \leq \lambda_n \sum_{\ell\in B1} \omega_{n,\ell} \left|\beta_{o,\ell}\right| + O_p(\tau_n^2). \tag{7.11}$$

By the Cauchy-Schwarz inequality,

$$\lambda_n \sum_{\ell\in B1} \omega_{n,\ell} \left|\beta_{o,\ell}\right| - \lambda_n \sum_{\ell\in B1} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \leq b_n \left\|\widehat{\alpha}_n - \alpha_o\right\|. \tag{7.12}$$

The inequalities in (7.11) and (7.12) imply that

$$\overline{g}_n'(\widehat{\alpha}_n)W_n\overline{g}_n(\widehat{\alpha}_n) \leq b_n \left\|\widehat{\alpha}_n - \alpha_o\right\| + O_p(\tau_n^2). \tag{7.13}$$

Applying $\overline{g}_n(\widehat{\alpha}_n) = m(\widehat{\alpha}_n) + \upsilon_n(\widehat{\theta}_n)$, we obtain

$$\begin{aligned}
\overline{g}_n'(\widehat{\alpha}_n)W_n\overline{g}_n(\widehat{\alpha}_n) &= \left[m(\widehat{\alpha}_n) + \upsilon_n(\widehat{\theta}_n)\right]' W_n \left[m(\widehat{\alpha}_n) + \upsilon_n(\widehat{\theta}_n)\right] \\
&= m(\widehat{\alpha}_n)'W_n m(\widehat{\alpha}_n) + \upsilon_n(\widehat{\theta}_n)'W_n\upsilon_n(\widehat{\theta}_n) + 2m(\widehat{\alpha}_n)'W_n\upsilon_n(\widehat{\theta}_n) \\
&\asymp ||m(\widehat{\alpha}_n)||^2 + ||m(\widehat{\alpha}_n)||O_p(\tau_n) + O_p(\tau_n^2),
\end{aligned} \tag{7.14}$$

w.p.a.1, using $\upsilon_n(\widehat{\theta}_n) = O_p(\tau_n)$ by Assumption 2 and $0 < C^{-1} \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq C$ for some $C < \infty$ w.p.a.1 by Assumption 1(v). Because $m(\alpha_o) = 0$,

$$||m(\widehat{\alpha}_n)||^2 = ||\mathbb{E}_Z[g(Z_i, \widehat{\theta}_n)] - \mathbb{E}_Z[g(Z_i, \theta_o)]||^2 + ||\widehat{\beta}_n - \beta_o||^2 \tag{7.15}$$

By a mean-value expansion,

$$\mathbb{E}_Z[g(Z_i, \widehat{\theta}_n)] - \mathbb{E}_Z[g(Z_i, \theta_o)] = \Gamma(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_o) \tag{7.16}$$

for some $\widetilde{\theta}_n$ between $\widehat{\theta}_n$ and θ_o. Assumptions 1(iii) and 1(iv) and the consistency of $\widehat{\theta}_n$ imply that

$$\begin{aligned}
||\Gamma(\widetilde{\theta}_n)(\widehat{\theta}_n - \theta_o)||^2 &= ||\widehat{\theta}_n - \theta_o||^2||\widehat{\gamma}_n'[\Gamma(\widetilde{\theta}_n) - \Gamma_o] + \widehat{\gamma}_n'\Gamma(\theta_o)||^2 \\
&= ||\widehat{\theta}_n - \theta_o||^2[||\widehat{\gamma}_n'\Gamma(\theta_o)||^2 + o_p(1)] \asymp ||\widehat{\theta}_n - \theta_o||
\end{aligned} \tag{7.17}$$

where $\widehat{\gamma}_n = (\widehat{\theta}_n - \theta_o)/||\widehat{\theta}_n - \theta_o||$. Combining (7.15)-(7.17) yields

$$||m(\widehat{\alpha}_n)|| \asymp ||\widehat{\alpha}_n - \alpha_o|| \tag{7.18}$$

w.p.a.1, which in turn gives

$$\overline{g}_n'(\widehat{\alpha}_n)W_n\overline{g}_n(\widehat{\alpha}_n) \asymp ||\widehat{\alpha}_n - \alpha_o||^2 + ||\widehat{\alpha}_n - \alpha_o||O_p(\tau_n) + O_p(\tau_n^2) \tag{7.19}$$

w.p.a.1, in conjuncture with (7.14).

By (7.13) and (7.19), we have

$$\left\|\widehat{\alpha}_n - \alpha_o\right\|^2 - O_p(\tau_n + b_n)\left\|\widehat{\alpha}_n - \alpha_o\right\| + O_p(\tau_n^2) \leq 0. \tag{7.20}$$

This implies that $\|\widehat{\alpha}_n - \alpha_o\| = O_p(\tau_n + b_n)$.

To verify part (b), note that Assumption $2^\dagger$ implies that (i) $||\overline{g}_n(\alpha_o)|| = O_p(\tau_n)$ and (ii) $||\overline{g}_n(\alpha) - \overline{g}_n(\alpha_o)|| \asymp ||\alpha - \alpha_o||$. Also note that

$$\begin{aligned}
&\overline{g}_n'(\widehat{\alpha}_n) W_n \overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n'(\alpha_o) W_n \overline{g}_n(\alpha_o) \\
&= [\overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n(\alpha_o)]' W_n [\overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n(\alpha_o)] \\
&\quad + 2 [\overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n(\alpha_o)]' W_n \overline{g}_n(\alpha_o) \\
&\geq C \|\overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n(\alpha_o)\|^2 - \|\overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n(\alpha_o)\| \, \|\overline{g}_n(\alpha_o)\| \\
&\geq [C + o_p(1)] \|\widehat{\alpha}_n - \alpha_o\|^2 - O_p(\tau_n) \|\widehat{\alpha}_n - \alpha_o\|, \qquad (7.21)
\end{aligned}$$

where the first inequality follows from Assumption 1(v) and the Cauchy-Schwarz inequality and the second inequality holds by Assumptions $2^\dagger$(i) and $2^\dagger$(ii). By the triangle inequality and the Cauchy-Schwarz inequality, we have

$$\begin{aligned}
&\lambda_n \sum_{j \in B_1} \omega_{n,j} \left|\widehat{\beta}_{n,j}\right| - \lambda_n \sum_{j \in B_1} \omega_{n,j} \left|\beta_{o,j}\right| \\
&\geq -\lambda_n \sum_{j \in B_1} \omega_{n,j} \left|\widehat{\beta}_{n,j} - \beta_{o,j}\right| \geq -\lambda_n \|\omega_{n,B_1}\| \, \|\widehat{\alpha}_n - \alpha_o\|. \qquad (7.22)
\end{aligned}$$

By Assumption P2 and the inequalities in (7.3), (7.21) and (7.22), we get

$$\|\widehat{\alpha}_n - \alpha_o\|^2 - [O_p(\tau_n) + \lambda_n \|\omega_{n,B_1}\|] \|\widehat{\alpha}_n - \alpha_o\| \leq 0, \qquad (7.23)$$

which implies that $||\widehat{\alpha}_n - \alpha_o|| = O_p(\tau_n + b_n)$. □

Proof of Theorem 1. We start with part (a). By the Karush–Kuhn–Tucker (KKT) optimality condition, $\widehat{\beta}_{n,\ell} = 0$ if

$$|W_n(k_o + \ell) \overline{g}_n(\widehat{\alpha}_n)| < \left|\frac{\lambda_n \omega_{n,\ell}}{2}\right|, \qquad (7.24)$$

where $W_n(k_o + \ell)$ is a row of W_n associated with $\omega_{n,\ell}$. Hence,

$$\Pr\left(\widehat{\beta}_{n,\ell} = 0, \, \ell \in A\right) \geq \Pr\left(\max_{\ell \in A} \left|\frac{W_n(k_o + \ell) \overline{g}_n(\widehat{\alpha}_n)}{\lambda_n \omega_{n,\ell}}\right| < \frac{1}{2}\right). \qquad (7.25)$$

To obtain the desired result, it remains to show

$$\max_{\ell \in A} \left|\frac{W_n(k_o + \ell) \overline{g}_n(\widehat{\alpha}_n)}{\lambda_n \omega_{n,\ell}}\right| = o_p(1). \qquad (7.26)$$

Following Assumptions 1(v),

$$0 < C_2 \leq W_n(k_o + \ell)W_n'(k_o + \ell) \leq C_1 < \infty \tag{7.27}$$

w.p.a.1 for some constants C_1 and C_2. By the Cauchy-Schwarz inequality and (7.27),

$$\max_{\ell \in A} \left| \frac{W_n(k_o + \ell)\overline{g}_n(\widehat{\alpha}_n)}{\lambda_n \omega_{n,\ell}} \right| \leq \max_{\ell \in A} \frac{\|W_n(k_o + \ell)\|}{\lambda_n \omega_{n,\ell}} \|\overline{g}_n(\widehat{\alpha}_n)\| \leq \frac{C \|\overline{g}_n(\widehat{\alpha}_n)\|}{\lambda_n} \max_{\ell \in A} \omega_{n,\ell}^{-1}. \tag{7.28}$$

for some constant $0 < C < \infty$. By the triangle inequality,

$$\|\overline{g}_n(\widehat{\alpha}_n)\| \leq \|m(\widehat{\alpha}_n)\| + ||\upsilon_n(\widehat{\theta}_n)|| = O_p(\tau_n), \tag{7.29}$$

where the equality follows from (7.18), Lemma 3(b) under Assumption P2, and Assumption 2.

The inequalities in (7.28), (7.29), and Assumption P3 imply that

$$\max_{\ell \in A} \left| \frac{W_n(k_o + \ell)\overline{g}_n(\widehat{\alpha}_n)}{\lambda_n \omega_{n,\ell}} \right| = \lambda_n^{-1} \max_{\ell \in A} \omega_{n,\ell}^{-1} O_p(\tau_n) = o_p(1). \tag{7.30}$$

Next, we prove part (b). Under Assumption $2^\dagger$, we have

$$\begin{aligned} \|\overline{g}_n(\widehat{\alpha}_n)\| &\leq \|\overline{g}_n(\widehat{\alpha}_n) - \overline{g}_n(\alpha_o)\| + \|\overline{g}_n(\alpha_o)\| \\ &\leq [C + o_p(1)] \|\widehat{\alpha}_n - \alpha_o\| + O_p(\tau_n) = O_p(\tau_n) \end{aligned} \tag{7.31}$$

where the first inequality follows from the triangle inequality, the second inequality is by Assumptions $2^\dagger$(i), $2^\dagger$(ii), and Lemma 3(c). This completes the proof. □

Proof of Lemma 2. Let $\varepsilon_n = o(n^{-1/2})$ be a sequence of constants such that (i) $\lambda_n \|\omega_{n,B}\| = O_p(\varepsilon_n)$, (ii) $\varsigma_n \tau_n = O(\varepsilon_n)$ under Assumptions 3(ii) and P4. Define

$$\widehat{\alpha}_{B,n}^* = \widehat{\alpha}_{B,n} + \varepsilon_n u_n^*, \text{ where } u_n^* = (\Gamma_\alpha' W_n \Gamma_\alpha)^{-1} \gamma_n^*, \tag{7.32}$$

where $\gamma_n^* \in R^{k_n}$ and $||\gamma_n^*|| \leq 1$. Because both W_n and $\Gamma_\alpha' \Gamma_\alpha$ have bounded eigenvalues by Assumptions 1(v) and 4(iii), $||u_n^*|| \leq C$ for some $C < \infty$ w.p.a.1. Hence,

$$\|\varepsilon_n u_n^*\|^2 = \varepsilon_n^2 ||u_n^*||^2 = O(\varepsilon_n^2) = o(n^{-1}) \tag{7.33}$$

w.p.a.1. Write $\widehat{\alpha}_{B,n}^* = (\widehat{\theta}_n^*, \widehat{\beta}_{B,n}^*)$, then $\left\|\widehat{\alpha}_{B,n}^* - \widehat{\alpha}_{B,n}\right\| = O_p(\varepsilon_n^2)$.

By the definition of $\widehat{\alpha}_n$,

$$\overline{g}_n'(\widehat{\alpha}_n)W_n\overline{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell=1}^{k_n} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \leq \overline{g}_n'(\widehat{\alpha}^*_{B,n})W_n\overline{g}_n(\widehat{\alpha}^*_{B,n}) + \lambda_n \sum_{\ell\in B} \omega_{n,\ell} \left|\widehat{\beta}^*_{n,\ell}\right| \tag{7.34}$$

where $\widehat{\beta}^*_{n,\ell}$ is the element of $\widehat{\alpha}^*_{B,n}$ corresponding to $\widehat{\beta}_{n,\ell}$. By Theorem 1, the left hand side of (7.34) satisfies that

$$\overline{g}_n'(\widehat{\alpha}_n)W_n\overline{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell=1}^{k_n} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| = \overline{g}_n'(\widehat{\alpha}_{B,n})W_n\overline{g}_n(\widehat{\alpha}_{B,n}) + \lambda_n \sum_{\ell\in B} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \tag{7.35}$$

w.p.a.1. The triangle inequality and the Cauchy-Schwarz inequality imply that

$$\begin{aligned}
&\left|\lambda_n \sum_{\ell\in B} \omega_{n,\ell} \left(\left|\widehat{\beta}^*_{n,\ell}\right| - \left|\widehat{\beta}_{n,\ell}\right|\right)\right| \\
&\leq \lambda_n \sum_{\ell\in B} \omega_{n,\ell} \left|\widehat{\beta}^*_{n,\ell} - \widehat{\beta}_{n,\ell}\right| = \lambda_n\varepsilon_n \sum_{\ell\in B} \omega_{n,\ell} \left|u^*_{n,\ell}\right| \\
&\leq \varepsilon_n\lambda_n \left\|\omega_{n,B}\right\| \left\|u^*_{n,B}\right\| = O_p(\varepsilon_n^2),
\end{aligned} \tag{7.36}$$

where $u^*_{n,B} \equiv (u^*_{n,d_\theta+1}, ..., u^*_{n,d_\theta+d_B})'$ is the vector of perturbation on β and the $O_p(\varepsilon_n^2)$ follows from $||u^*_{n,B}|| \leq C$ for some $C < \infty$ and Assumption P4. Combining (7.34)-(7.36) yields

$$\overline{g}_n'(\widehat{\alpha}^*_{B,n})W_n\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n'(\widehat{\alpha}_{B,n})W_n\overline{g}_n(\widehat{\alpha}_{B,n}) \geq O_p(\varepsilon_n^2). \tag{7.37}$$

Define

$$I_{1,n} = \upsilon_n(\widehat{\theta}^*_{B,n}) - \upsilon_n(\widehat{\theta}_{B,n}). \tag{7.38}$$

Because $g(Z_i, \alpha)$ in linear in β,

$$\overline{g}_n(\alpha) = m(\alpha) + \upsilon_n(\theta). \tag{7.39}$$

Applying this equality, we obtain

$$\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n}) = m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n}) + I_{1,n}, \tag{7.40}$$

which implies that

$$||\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})||^2 \leq 2||m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n})||^2 + 2I_{1,n}^2 = O_p(\varepsilon_n^2), \tag{7.41}$$

where the $O_p(\varepsilon_n^2)$ term follows from (i) $||m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n})||^2 = O_p(\varepsilon_n^2)$ by Assumption 4(iii) and

$||\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}|| = O_p(\varepsilon_n^2)$ and (ii) $I_{1,n} = O_p(\varsigma_n n^{-1/2}) = O_p(\varsigma_n \tau_n) = O_p(\varepsilon_n)$ by Assumptions 3(i), 3(ii), and $||\widehat{\theta}^*_n - \widehat{\theta}_n|| = O_p(\varepsilon_n)$. Write the left hand side of (7.37) as

$$\begin{aligned} &\overline{g}'_n(\widehat{\alpha}^*_{B,n}) W_n \overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}'_n(\widehat{\alpha}_{B,n}) W_n \overline{g}_n(\widehat{\alpha}_{B,n}) \\ = &\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]' W_n \left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right] + 2\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]' W_n \overline{g}_n(\widehat{\alpha}_{B,n}). \end{aligned} \tag{7.42}$$

This and (7.41) imply that

$$\left[m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n}) + I_{1,n}\right]' W_n \overline{g}_n(\widehat{\alpha}_{B,n}) \geq O_p(\varepsilon_n^2). \tag{7.43}$$

Define

$$I_{0,n} = \upsilon_n(\widehat{\theta}_{B,n}) - \upsilon_n(\theta_o). \tag{7.44}$$

Then

$$\overline{g}_n(\widehat{\alpha}_{B,n}) = \overline{g}_n(\alpha_o) + m(\widehat{\alpha}_{B,n}) - m(\alpha_{B,0}) + I_{0,n}. \tag{7.45}$$

Plugging (7.45) into (7.43) yields

$$\begin{aligned} O_p(\varepsilon_n^2) \leq &\left[m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n})\right]' W_n \left(\overline{g}_n(\alpha_o) + m(\widehat{\alpha}_{B,n}) - m(\alpha_{B,0})\right) \\ &+ A + B + C, \text{ where} \\ A = &\, I'_{1,n} W_n [\overline{g}_n(\alpha_o) + m(\widehat{\alpha}_{B,n}) - m(\alpha_{B,0}) + I_{0,n}] \\ B = &\left[m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n})\right]' W_n I_{0,n} \\ C = &\, I'_{1,n} W_n I_{0,n}. \end{aligned} \tag{7.46}$$

The extra term $A = o_p(\varepsilon_n n^{-1/2})$ because $I_{1,n} = O_p(\varsigma_n n^{-1/2})$, $||\overline{g}_n(\alpha_o)|| = O_p(\tau_n)$, $m(\widehat{\alpha}_{B,n}) - m(\alpha_{B,0}) = O_p(\tau_n)$, $I_{0,n} = O_p(\varsigma_n \tau_n) = O(\varepsilon_n)$. The extra terms are $B = O_p(\varepsilon_n^2)$ and $C = O_p(\varepsilon_n^2)$ because $m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n}) = O_p(\varepsilon_n)$, $\|I_{0,n}\| = O(\varepsilon_n)$. Therefore, the inequality in (7.46) implies that

$$\left[m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n})\right]' W_n \left(\overline{g}_n(\alpha_o) + m(\widehat{\alpha}_{B,n}) - m(\alpha_{B,0})\right) \geq o_p(\varepsilon_n n^{-1/2}). \tag{7.47}$$

By mean-value expansions,

$$\begin{aligned} m(\widehat{\alpha}^*_{B,n}) - m(\widehat{\alpha}_{B,n}) &= \Gamma_\alpha(\widetilde{\alpha}^*_{B,n})(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}) \\ m(\widehat{\alpha}_{B,n}) - m(\alpha_{B,0}) &= \Gamma_\alpha(\widetilde{\alpha}_{B,n})(\widehat{\alpha}_{B,n} - \alpha_{B,0}) \end{aligned} \tag{7.48}$$

for $\widetilde{\alpha}^*_{B,n}$ between $\widehat{\alpha}^*_{B,n}$ and $\widehat{\alpha}_{B,n}$ and $\widetilde{\alpha}_{B,n}$ between $\widehat{\alpha}_{B,n}$ and $\alpha_{B,0}$. By (7.32), (7.47), and (7.48),

$$\overline{\gamma}_n'(\Gamma_\alpha' W_n \Gamma_\alpha)^{-1}\Gamma_\alpha(\widetilde{\alpha}^*_{B,n})' W_n \left[n^{1/2}\overline{g}_n(\alpha_o) + \Gamma_\alpha(\widetilde{\alpha}_{B,n}) n^{1/2}(\widehat{\alpha}_{B,n} - \alpha_{B,0})\right] \geq o_p(1). \tag{7.49}$$

Next, define $\widehat{\alpha}^*_{B,n} = \widehat{\alpha}_{B,n} - \varepsilon_n u^*_n$ and using the same arguments in deriving (7.49), we deduce that

$$\overline{\gamma}_n'(\Gamma_\alpha' W_n \Gamma_\alpha)^{-1}\Gamma_\alpha(\widetilde{\alpha}^*_{B,n})' W_n \left[n^{1/2}\overline{g}_n(\alpha_o) + \Gamma_\alpha(\widetilde{\alpha}_{B,n}) n^{1/2}(\widehat{\alpha}_{B,n} - \alpha_{B,0})\right] \leq o_p(1). \tag{7.50}$$

The inequalities in (7.49) and (7.50) and the consistency of $\widehat{\theta}_n$ yield

$$\left|\overline{\gamma}_n'(\Gamma_\alpha' W_n \Gamma_\alpha)^{-1}\Gamma_\alpha(\widetilde{\alpha}^*_{B,n})' W_n \left[n^{1/2}\overline{g}_n(\alpha_o) + \Gamma_\alpha(\widetilde{\alpha}_{B,n}) n^{1/2}(\widehat{\alpha}_{B,n} - \alpha_{B,0})\right]\right| = o_p(1). \tag{7.51}$$

Following Assumptions 1(iii)-i(v) and the consistency of $\widehat{\alpha}_n$,

$$\begin{aligned} &||(\Gamma_\alpha' W_n \Gamma_\alpha)^{-1}\Gamma_\alpha(\widetilde{\alpha}^*_{B,n}) W_n \Gamma_\alpha(\widetilde{\alpha}_{B,n}) - I_{d_\theta + d_B}|| \to_p 0, \\ &||\Gamma_\alpha(\widetilde{\alpha}^*_{B,n}) - \Gamma_\alpha|| = o_p(1). \end{aligned} \tag{7.52}$$

This and (7.51) together give

$$\overline{\gamma}_n' n^{1/2}(\widehat{\alpha}_{B,n} - \alpha_{B,0})(1 + o_p(1)) = -\overline{\gamma}_n'(\Gamma_\alpha' W_n \Gamma_\alpha)^{-1}\left(\Gamma_\alpha + o_p(1)\right)' W_n n^{1/2}\overline{g}_n(\alpha_o) + o_p(1). \tag{7.53}$$

Let $\gamma_n \in R^{k_n}$ be an arbitrary vector with $||\gamma_n|| = 1$. Take

$$\begin{aligned} \overline{\gamma}_n' &\equiv \gamma_n'(\Gamma_\alpha' W_n \Omega_n W_n \Gamma_\alpha)^{-1/2}\left(\Gamma_\alpha' W_n \Gamma_\alpha\right) = \gamma_n' \Sigma_n^{-1/2}, \\ \gamma^{*\prime}_n &\equiv \gamma_n'(\Gamma_\alpha' W_n \Omega_n W_n \Gamma_\alpha)^{-1/2}\Gamma_\alpha{}' W_n \Omega_n^{1/2}. \end{aligned} \tag{7.54}$$

Obviously, $||\gamma^*_n|| = 1$ and $||\overline{\gamma}_n|| \leq 1$. With this choice of γ^*_n and $\overline{\gamma}_n$, the right hand side of (7.53) satisfies

$$\begin{aligned} &\overline{\gamma}_n'(\Gamma_\alpha' W_n \Gamma_\alpha)^{-1}\left(\Gamma_\alpha + o_p(1)\right)' W_n n^{1/2}\overline{g}_n(\alpha_o) \\ =\ &\gamma_n'(\Gamma_\alpha' W_n \Omega_n W_n \Gamma_\alpha)^{-1/2}\left(\Gamma_\alpha + o_p(1)\right)' W_n \Omega_n^{1/2}\left[n^{1/2}\Omega_n^{-1/2}\overline{g}_n(\alpha_o)\right] \\ =\ &\left(\gamma^{*\prime}_n + o_p(1)\right)\left[n^{1/2}\Omega_n^{-1/2}\overline{g}_n(\alpha_o)\right] \\ \to_d\ &N(0,1), \end{aligned} \tag{7.55}$$

where the $o_p(1)$ term in the second equality follows from the bounds of W_n, Ω_n, and $\Gamma_\alpha'\Gamma_\alpha$ in Assumptions 1(v), 4(ii), and 4(iii) and the convergence in distribution follows from Assumption

4(i). By the Slutsky Theorem, (7.53), and (7.55), we obtain

$$\gamma_n' \Sigma_n^{-1/2} n^{1/2} (\widehat{\alpha}_{B,n} - \alpha_{B,0}) = \overline{\gamma}_n' n^{1/2} (\widehat{\alpha}_{B,n} - \alpha_{B,0}) \to_d N(0,1). \tag{7.56}$$

The same results hold in part (b) because the rate of convergence and super efficiency results hold under this set of conditions according to Lemma 3(c) and Theorem 1(b). This completes the proof. □

7.2 Proofs on Sufficient Conditions

Proof of Lemma 2. Define $\mathcal{F} = \{g_\ell(Z,\theta) : \theta \in \Theta\}$. By Lemma (2.13) of Pakes and Pollard (1989), $\mathcal{F}$ is Euclidean for the envelope $F = \sup_{\theta\in\Theta} |g_\ell(Z,\theta)| + \sup_{\theta\in\Theta} |g_{\theta,\ell}(Z,\theta)|$ under Assumption 2*. For the definition of a Euclidean class of functions, see (2.7) of Pakes and Pollard (1989). By the maximal inequality (Section 4.3 of Pollard (1989)), for any n,

$$\mathbb{E} \sup_{\theta\in\Theta} (g_\ell(Z_i,\theta) - \mathbb{E}g_\ell(Z_i,\theta))^2 \leq Cn^{-1}. \tag{7.57}$$

Hence,

$$\mathbb{E} \sup_{\theta\in\Theta} ||g(Z_i,\theta) - \mathbb{E}g_\ell(Z_i,\theta)||^2 \leq Ck_n n^{-1}, \tag{7.58}$$

which implies

$$\sup_{\theta\in\Theta} ||g(Z_i,\theta) - \mathbb{E}g_\ell(Z_i,\theta)|| = O_p(\sqrt{k_n/n}) \tag{7.59}$$

by the Markov's inequality. □

Proof of Lemma 4. When the sample moments are differentiable, let $g_\theta(Z,\theta)$ denote the partial derivative wrt θ. By a mean-value expansion and an exchange of "$\mathbb{E}$" and "∂",

$$\upsilon_n(\theta_1) - \upsilon_n(\theta_2) = \left[n^{-1} \sum_{i=1}^n g_\theta(Z_i, \widetilde{\theta}) - \mathbb{E}g_\theta(Z_i, \widetilde{\theta}) \right] (\theta_1 - \theta_2). \tag{7.60}$$

for some $\widetilde{\theta}$ between θ_1 and θ_2, where $\widetilde{\theta}$ can be different for different rows. Applying the proof of Lemma 2 with $g_\ell(Z,\theta)$ replaced by $g_{\theta,\ell}(Z_i,\widetilde{\theta})$ under Assumptions 2* and 3*, we obtain

$$\mathbb{E} \sup_{\theta\in\Theta} ||n^{-1} \sum_{i=1}^n g_{\theta,\ell}(Z_i,\theta) - \mathbb{E}g_{\theta,\ell}(Z_i,\theta)||^2 \leq Cn^{-1} \tag{7.61}$$

for all n and ℓ. Hence,

$$\mathbb{E}\sup_{\theta\in\Theta}||n^{-1}\sum_{i=1}^{n}g_\theta(Z_i,\theta)-\mathbb{E}g_\theta(Z_i,\theta)||^2\leq Ck_nn^{-1}. \tag{7.62}$$

Combining (7.60) and (7.62), we have

$$\begin{aligned}\sup_{\theta_1,\theta_2\in\Theta}\frac{||\upsilon_n(\theta_1)-\upsilon_n(\theta_2)||}{||\theta_1-\theta_2||} &\leq \sup_{\theta\in\Theta}||n^{-1}\sum_{i=1}^{n}g_{\theta,\ell}(Z_i,\theta)-\mathbb{E}g_{\theta,\ell}(Z_i,\theta)|| \\ &= O_p(\sqrt{k_n/n}),\end{aligned} \tag{7.63}$$

where the inequality follows from the Cauchy-Schwarz inequality and the $O_p(\sqrt{k_n/n})$ term follows from the Markov's inequality. This verifies Assumption 3(i) with $\varsigma_n=\sqrt{k_n/n}$. Assumption 3(ii) holds because $\tau_n\varsigma_n=k_n/n=o(n^{-1/2})$ when $k_n=o(n^{1/2})$. □

7.3 Example: A Linear Model with Instrumental Variables

In this example, we consider a simple linear IV model to illustrate the verification of some general assumptions. The model

$$Y_i = X_i\theta_o+u_i, \tag{7.64}$$

$$X_i = W_i^*+v_i=\sum_{j=1}^{q}\pi_{j,o}Z_{1,j,i}+\sum_{j=q+1}^{\infty}\pi_{j,o}Z_{1,j,i}+v_i, \tag{7.65}$$

where Y_i, X_i are scaler endogenous variables and $Z_{1,j}$, $j\in\mathbb{Z}\equiv\{1,2,...\}$, are the excluded exogenous variable. We assume that

$$\mathbb{E}[u_i|\,Z_{1,j,i}]=0 \text{ for all } j \tag{7.66}$$

and the empirical researcher has the first q instrumental variables to construct the moment conditions for identification.

The rest of the IVs are mixed with invalid IVs $W_{1,j}$ in the sense that $\mathbb{E}[u_iW_{1,j,i}]\neq 0$ for $j\in B1$, and redundant IVs in the sense that $\mathbb{E}[X_iW_{2,j,i}]=0$ for $j\in B0$. In this example, we have

$$\mathbb{E}\left[(Y_i-X_i\theta_o)\,Z_{1,j,i}\right]=0 \text{ with } j\in Q\equiv\{1,...,q\} \tag{7.67}$$

for identification and consistent estimation of θ_o; and the following moment conditions

$$\mathbb{E}\left[\left(Y_i - X_i\theta_o\right) Z_{1,j,i}\right] \stackrel{?}{=} 0 \text{ with } j \in A \subset \mathbb{Z}\backslash Q, \tag{7.68}$$

$$\mathbb{E}\left[\left(Y_i - X_i\theta_o\right) W_{1,j,i}\right] \stackrel{?}{=} 0 \text{ with } j \in B1, \tag{7.69}$$

$$\text{and } \mathbb{E}\left[\left(Y_i - X_i\theta_o\right) W_{2,j,i}\right] \stackrel{?}{=} 0 \text{ with } j \in B0. \tag{7.70}$$

For the ease of notation, we use $Z_{2n,i}$ to denote the instrumental variables in the second set for selection and $Z_{1,i} = (Z_{1,1,i}, ..., Z_{1,q,i})'$ to denote the instrumental variables in the first set that are known to be valid and relevant for identification.

We next provide sufficient conditions for Assumptions $2^\dagger$, 3, and 4, when the moment conditions are constructed from this linear IV model.

Define $Z'_{n,i} \equiv (Z'_{1,i}, Z'_{2n,i})$ and $\check{Z}'_{n,i} \equiv (Y_i, X_i, Z'_{n,i})$. Let $Z_{n,i}(j)$ denote the j-th component in $Z_{n,i}$

Condition 1 *Suppose* (i). $\{\check{Z}_{n,i}\}_{i\leq n}$ *is a triangle array of* i.i.d. *process;* (ii). $\mathbb{E}[Z^4_{n,i}(j)] < C$, $\mathbb{E}[W_i^{*2}] < \infty$ *and* $\mathbb{E}[Z_{n,i}Z'_{n,i}] = I_{q+k_n}$ *for all* n *and* j*;* (iii). $\mathbb{E}[u_i^4 \big| Z_{n,i}] < C$ *and* $\mathbb{E}[v_i^4 \big| Z_{n,i}] < C$*;* (iv). *there are finite constants* Π_1 *and* Π_2 *such that* $\sum_{j=1}^q \pi^2_{j,o} = \Pi_1^2 > 0$ *and* $\lim_{n\to\infty} \mathbb{E}\left[v_i Z'_{2n,i}\right] \mathbb{E}\left[v_i Z_{2n,i}\right] = \Pi_2^2$.

For the linear IV model, the following results hold. We assume $k_n = o(n^{1/2})$.

Lemma 5 (a) *Under Condition* 1*,*

$$\left\|\overline{g}_n(\theta_o)\right\|^2 = O_p\left(k_n/n\right) \tag{7.71}$$

$$\text{and } \left\|\overline{g}_n(\theta) - \overline{g}_n(\theta_o)\right\|^2 \asymp \left[1 + O_p(k_n/n) + O_p(\sqrt{k_n/n})\right] \left\|\theta - \theta_o\right\|^2. \tag{7.72}$$

Hence, Assumption $2^\dagger$ *holds with* $\tau_n = \sqrt{k_n/n}$.

(b) *Under Condition* 1*, Assumption* 3 *holds with* $\varsigma_n = \sqrt{k_n/n}$.

(c) *Condition* 1 *implies Assumption* 4*.*

Proof of Lemma 5. We first show part (a). First note that by definition,

$$g_C(Z,\theta) = \left(Y_i - X_i\theta\right) Z_{1,i},$$
$$g_D(Z,\theta) = \left(Y_i - X_i\theta\right) Z_{2n,i}.$$

and we can rewrite

$$\begin{aligned}\|\overline{g}_n(\theta_o)\|^2 &= \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n u_i Z_{1,i} \\ \frac{1}{n}\sum_{i=1}^n (u_i Z_{2n,i}) \end{pmatrix}' \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n u_i Z_{1,i} \\ \frac{1}{n}\sum_{i=1}^n (u_i Z_{2n,i}) \end{pmatrix} \\ &= \left\| \frac{1}{n}\sum_{i=1}^n u_i Z_{1,i} \right\|^2 + \left\| \frac{1}{n}\sum_{i=1}^n (u_i Z_{2n,i} - \mathbb{E}[u_i Z_{2n,i}]) \right\|^2. \end{aligned} \tag{7.73}$$

Using the Markov's inequality, Condition 1(i)-(iii), we have

$$\left\| \frac{1}{n}\sum_{i=1}^n u_i Z_{1,i} \right\|^2 = O_p(n^{-1}) \text{ and } \left\| \frac{1}{n}\sum_{i=1}^n (u_i Z_{2n,i} - \mathbb{E}[u_i Z_{2n,i}]) \right\|^2 = O_p(k_n n^{-1}). \tag{7.74}$$

By definition, we have

$$\overline{g}_n(\theta) - \overline{g}_n(\theta_o) = \begin{pmatrix} \frac{-1}{n}\sum_{i=1}^n X_i Z_{1,i} \\ \frac{-1}{n}\sum_{i=1}^n X_i Z_{2n,i} \end{pmatrix} (\theta - \theta_o) \equiv -\overline{G}_n(\theta - \theta_o). \tag{7.75}$$

As a result, we have

$$\|\overline{g}_n(\theta) - \overline{g}_n(\theta_o)\|^2 = (\theta - \theta_o)'\overline{G}_n'\overline{G}_n(\theta - \theta_o). \tag{7.76}$$

By definition, we can rewrite

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n X_i Z_{n,i} - \mathbb{E}\left[X_i Z_{n,i}\right] \\ =\ &\frac{1}{n}\sum_{i=1}^n \left[W_i^* Z_{n,i} - \mathbb{E}\left[W_i^* Z_{n,i}\right]\right] + \frac{1}{n}\sum_{i=1}^n \left[v_i Z_{n,i} - \mathbb{E}\left[v_i Z_{n,i}\right]\right]. \end{aligned} \tag{7.77}$$

Using Condition 1(i)-(iii) and the Hölder's inequality, we get

$$\mathbb{E}\left[\left\| \frac{1}{n}\sum_{i=1}^n \left[W_i^* Z_{n,i} - \mathbb{E}\left[W_i^* Z_{n,i}\right]\right] \right\|^2 \right] \leq \frac{1}{n}\sum_{j=1}^{q+k_n} \mathbb{E}\left[W_i^{*2} Z_{n,i}^2\right] = O(k_n/n) \tag{7.78}$$

which, together with the Markov's inequality, implies that

$$\left\| \frac{1}{n}\sum_{i=1}^n \left[W_i^* Z_{n,i} - \mathbb{E}\left[W_i^* Z_{n,i}\right]\right] \right\| = O_p(\sqrt{k_n/n}). \tag{7.79}$$

Similarly, we can show that

$$\left\| \frac{1}{n} \sum_{i=1}^{n} \left[v_i Z_{n,i} - \mathbb{E}\left[v_i Z_{n,i} \right] \right] \right\| = O_p(\sqrt{k_n/n}). \tag{7.80}$$

From the results in (7.79) and (7.80), we get

$$\left\| \overline{G}_n - \mathbb{E}\left[\overline{G}_n \right] \right\|^2 = \left\| \frac{1}{n} \sum_{i=1}^{n} X_i Z_{n,i} - \mathbb{E}\left[X_i Z_{n,i} \right] \right\|^2 = O_p(k_n/n) \tag{7.81}$$

which implies that

$$\begin{aligned} & \left\| (\theta - \theta_o)' \left[\overline{G}_n' \overline{G}_n - \mathbb{E}\left[\overline{G}_n \right]' \mathbb{E}\left[\overline{G}_n \right] \right] (\theta - \theta_o) \right\| \\ \leq & \left\| \overline{G}_n - \mathbb{E}\left[\overline{G}_n \right] \right\|^2 \left\| \theta - \theta_o \right\|^2 + 2 \left\| \theta - \theta_o \right\| \left\| \mathbb{E}\left[\overline{G}_n \right] (\theta - \theta_o) \right\| \left\| \overline{G}_n - \mathbb{E}\left[\overline{G}_n \right] \right\| \\ = & O_p(k_n/n) \left\| \theta - \theta_o \right\|^2 + O_p(\sqrt{k_n/n}) \left\| \theta - \theta_o \right\| \left\| \mathbb{E}\left[\overline{G}_n \right] (\theta - \theta_o) \right\|. \end{aligned} \tag{7.82}$$

Note that the eigenvalues of $\mathbb{E}\left[\overline{G}_n \right]' \mathbb{E}\left[\overline{G}_n \right]$ are bounded by some general constants, which together with the inequality in (7.82) implies that

$$\begin{aligned} & \left\| (\theta - \theta_o)' \left[\overline{G}_n' \overline{G}_n - \mathbb{E}\left[\overline{G}_n \right]' \mathbb{E}\left[\overline{G}_n \right] \right] (\theta - \theta_o) \right\| \\ = & \left[O_p(k_n/n) + O_p(\sqrt{k_n/n}) \right] \left\| \theta - \theta_o \right\|^2. \end{aligned} \tag{7.83}$$

Now, combining the results in (7.76) and (7.83), we deduce that

$$\begin{aligned} \left\| \overline{g}_n(\theta) - \overline{g}_n(\theta_o) \right\|^2 = & (\theta - \theta_o)' \mathbb{E}\left[\overline{G}_n \right]' \mathbb{E}\left[\overline{G}_n \right] (\theta - \theta_o) \\ & + (\theta - \theta_o)' \left[\overline{G}_n' \overline{G}_n - \mathbb{E}\left[\overline{G}_n \right]' \mathbb{E}\left[\overline{G}_n \right] \right] (\theta - \theta_o) \\ \asymp & \left[1 + O_p(k_n/n) + O_p(\sqrt{k_n/n}) \right] \left\| \theta - \theta_o \right\|^2 \end{aligned} \tag{7.84}$$

which finishes part (a).

Next, we verify Assumption 3 in part (b). For any θ_1 and θ_2 with $\|\theta_1 - \theta_2\| \leq \delta$, we have

$$\begin{aligned} & \overline{g}_n(\theta_1) - \overline{g}_n(\theta_2) - [m(\theta_1) - m(\theta_2)] \\ = & \begin{pmatrix} \frac{-1}{n} \sum_{i=1}^{n} \left[X_i Z_{1,i} - \mathbb{E}\left(X_i Z_{1,i} \right) \right] \\ \frac{-1}{n} \sum_{i=1}^{n} \left[X_i Z_{2n,i} - \mathbb{E}\left(X_i Z_{2n,i} \right) \right] \end{pmatrix} (\theta_1 - \theta_2) \\ \equiv & - \left[\overline{G}_n - \mathbb{E}\left(\overline{G}_n \right) \right] (\theta - \theta_o) \end{aligned} \tag{7.85}$$

which together with (7.81) implies that

$$\sup_{||\theta_1-\theta_2||\leq\delta} \|\overline{g}_n(\theta_1) - \overline{g}_n(\theta_2) - [m(\theta_1) - m(\theta_2)]\| = O(\sqrt{k_n/n}\delta). \tag{7.86}$$

Hence, we have $\varsigma_n = \sqrt{k_n/n}$ and $\tau_n = \sqrt{k_n/n}$ in the linear IV example. This ensures that $\varsigma_n\tau_n = O(\varepsilon_n)$ is satisfied when $k_n/\sqrt{n} = o(1)$.

Next, we show part (c). To verify Assumption 4(i), we only need to check the Lindeberg condition of the triangle array CLT. For this purpose, we define

$$\phi_{i,n} = n^{-1/2}\gamma_n'\Omega_n^{-1/2}g(\check{Z}_i,\theta_o) \equiv n^{-1/2}\gamma_n'\Omega_n^{-1/2}\begin{pmatrix} u_iZ_{1,i} \\ u_iZ_{2n,i} - \beta_o \end{pmatrix}, \tag{7.87}$$

then we have

$$\begin{aligned}\sum_{i=1}^n \mathbb{E}\left[\phi_{i,n}^2 I\{\phi_{i,n} > \epsilon\}\right] &= n\mathbb{E}\left[\left(\phi_{i,n}/\epsilon\right)^2 I\{\phi_{i,n}/\epsilon > 1\}\right] \\ &\leq \frac{1}{n\epsilon^4}\mathbb{E}\left[\left(\gamma_n'\Omega_n^{-1/2}g(\check{Z}_i,\theta_o)g'(\check{Z}_i,\theta_o)\Omega_n^{-1/2}\gamma_n\right)^2\right]. \end{aligned}\tag{7.88}$$

As $(\gamma_n'A\gamma_n)^2 \leq \gamma_n'A^2\gamma_n\gamma_n'\gamma_n$ for any symmetric matrix, we deduce that

$$\begin{aligned}&\mathbb{E}\left[\left(\gamma_n'\Omega_n^{-1/2}g(\check{Z}_i,\theta_o)g'(\check{Z}_i,\theta_o)\Omega_n^{-1/2}\gamma_n\right)^2\right] \\ &\leq \gamma_n'\Omega_n^{-1/2}\mathbb{E}\left[\left[g(\check{Z}_i,\theta_o)g'(\check{Z}_i,\theta_o)\right]^2\right]\Omega_n^{-1/2}\gamma_n\gamma_n'\Omega_n^{-1}\gamma_n. \end{aligned}\tag{7.89}$$

If Assumption 4(ii) holds (which is verified independently below), we have

$$\begin{aligned}&\mathbb{E}\left[\left(\gamma_n'\Omega_n^{-1/2}g(\check{Z}_i,\theta_o)g'(\check{Z}_i,\theta_o)\Omega_n^{-1/2}\gamma_n\right)^2\right] \\ &\leq \mathbb{E}\left[\left[\gamma_n'g(\check{Z}_i,\theta_o)\right]^2\left\|g(\check{Z}_i,\theta_o)\right\|^2\right] \\ &\leq \sqrt{\mathbb{E}\left[\left[\gamma_n'g(\check{Z}_i,\theta_o)\right]^4\right]}\sqrt{\mathbb{E}\left[\left\|g(\check{Z}_i,\theta_o)\right\|^4\right]} \\ &\leq \mathbb{E}\left[\left\|g(\check{Z}_i,\theta_o)\right\|^4\right] \end{aligned}\tag{7.90}$$

where the second inequality is the Hölder's inequality and the last inequality is by the Cauchy-

Schwarz inequality. Next by the Cauchy-Schwarz inequality, we get

$$\begin{aligned}\left\|g(\check{Z}_i,\theta_o)\right\|^4 &= \left[g'(\check{Z}_i,\theta_o)g(\check{Z}_i,\theta_o)\right]^2 = \left[g'(\check{Z}_i,\theta_o)g(\check{Z}_i,\theta_o)\right]^2 \\ &\leq (q+k_n)\sum_{j=1}^{q}\mathbb{E}\left[u_i^4 Z_{1,j,i}^4\right] + 4(q+k_n)\sum_{i=1}^{k_n}\mathbb{E}\left[\left(u_i^4 Z_{2n,j,i}^4 + \beta_{o,j}^4\right)\right] \\ &\leq q(q+k_n)C + 8(q+k_n)\sum_{i=1}^{k_n}\mathbb{E}\left[u_i^4 Z_{2n,j,i}^4\right] \leq C(q+k_n)^2 \end{aligned} \tag{7.91}$$

which, together with (7.88), (7.89), (7.90) and (7.91), implies that

$$\sum_{i=1}^{n}\mathbb{E}\left[\phi_{i,n}^2 I\{\phi_{i,n}>\epsilon\}\right] \leq \frac{Ck_n^2}{n\epsilon^4} = o(1)$$

where the last equality is by $k_n^2/n = o(1)$. Hence the the Lindeberg condition holds in the linear IV model, which verifies Assumption 4(i).

Assumption 4(ii) holds with

$$\Omega_n = \mathbb{E}\begin{pmatrix} u_i Z_{1,i} \\ u_i Z_{2n,i} \end{pmatrix}\begin{pmatrix} u_i Z_{1,i} \\ u_i Z_{2n,i} \end{pmatrix}' \tag{7.92}$$

and is implied by Condition 1(ii) and (iii) automatically.

Next, we verify Assumption 4(iii). Define $\Gamma_1 = \mathbb{E}\left[X_i Z_{1,i}\right]$ and $\Gamma_{2,n} = \mathbb{E}\left[X_i Z_{2n,i}\right]$, then

$$\Gamma_\alpha'\Gamma_\alpha = \begin{pmatrix} \Gamma_1'\Gamma_1 + \Gamma_{2,n}'\Gamma_{2,n} & \Gamma_{2,n}' \\ \Gamma_{2,n} & I_{k_n} \end{pmatrix}. \tag{7.93}$$

It is clear that Γ_α has full column rank, hence $\Gamma_\alpha'\Gamma_\alpha$ is strictly positive definite. Let $\lambda_{n,*}$ be the eigenvalues of $\Gamma_\alpha'\Gamma_\alpha$ such that $\lambda_{n,*} \neq 1$, then

$$\begin{aligned} 0 &= \det\left[\begin{pmatrix} \Gamma_1'\Gamma_1 + \Gamma_{2,n}'\Gamma_{2,n} - \lambda_{n,*} & \Gamma_{2,n}' \\ \Gamma_{2,n} & I_{k_n} - \lambda_{n,*} I_{k_n} \end{pmatrix}\right] \\ &= \left(\Gamma_1'\Gamma_1 + \Gamma_{2,n}'\Gamma_{2,n} - \lambda_{n,*} - \frac{\Gamma_{2,n}'\Gamma_{2,n}}{1-\lambda_{n,*}}\right)(1-\lambda_{n,*})^{k_n}, \end{aligned} \tag{7.94}$$

which means that $\lambda_{n,*}$ satisfies

$$\lambda_{n,*}^2 - (1 + \Gamma_1'\Gamma_1 + \Gamma_{2,n}'\Gamma_{2,n})\lambda_{n,*} + \Gamma_1'\Gamma_1 = 0. \tag{7.95}$$

The above equation has the following two solutions

$$\begin{aligned}\lambda_{n,*,-} &= \frac{\Gamma_n'\Gamma_n - \sqrt{(\Gamma_n'\Gamma_n)^2 - 4\Gamma_{1,n}'\Gamma_{1,n}}}{2} \text{ and} \\ \lambda_{n*,+} &= \frac{\Gamma_n'\Gamma_n + \sqrt{(\Gamma_n'\Gamma_n)^2 - 4\Gamma_{1,n}'\Gamma_{1,n}}}{2},\end{aligned} \tag{7.96}$$

where $\Gamma_n \equiv 1+\Gamma_1'\Gamma_1+\Gamma_{2,n}'\Gamma_{2,n}$. This implies that the eigenvalues of $\Gamma_\alpha'\Gamma_\alpha$ are bounded from below by $\min(\lambda_{n,*,-}, 1)$ and bounded from above by $\max(\lambda_{n*,+}, 1)$. Under Condition 1(ii) and (iv), we have

$$\begin{aligned}\lambda_{n,*,-} &= \frac{2\Gamma_1'\Gamma_1}{\Gamma_n'\Gamma_n + \sqrt{(\Gamma_n'\Gamma_n)^2 - 4\Gamma_1'\Gamma_1}} \\ &\geq \frac{\Gamma_1'\Gamma_1}{1+\Gamma_1'\Gamma_1+\Gamma_{2,n}'\Gamma_{2,n}} \to \frac{\Pi_1^2}{1+\mathbb{E}\left[W_i^{*2}\right]+\Pi_2^2} > 0\end{aligned} \tag{7.97}$$

and

$$\begin{aligned}\lambda_{n,*,+} &= \frac{\Gamma_n'\Gamma_n + \sqrt{(\Gamma_n'\Gamma_n)^2 - 4\Gamma_{1,n}'\Gamma_{1,n}}}{2} \\ &\leq 1+\Gamma_1'\Gamma_1+\Gamma_{2,n}'\Gamma_{2,n} \to 1+\mathbb{E}\left[W_i^{*2}\right]+\Pi_2^2 < \infty.\end{aligned} \tag{7.98}$$

From the results in (7.97) and (7.98), we deduce that the eigenvalues of $\Gamma_\alpha'\Gamma_\alpha$ are bounded by some general constants. This completes the proof. □

References

Andrews, D. W. K.(1994): "Empirical Process Methods in Econometrics," in *Handbook of Econometrics*, Vol. IV, ed. by R. F. Engle and D. McFadden. Amsterdam: North-Holland.

Andrews, D. W. K. (1999): "Consistent moment selection procedures for generalized method of moments estimation," *Econometrica*, 67(3), 543-563.

Andrews, D. W. K. (2002): "Generalized method of moments estimation when a parameter is on a boundary," *Journal of Business and Economic Statistics*, 20(4), 530–544.

Andrews, D. W. K. and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods, " *Econometrica*, 77(3), 721–762.

Andrews, D. W. K. and P. Guggenberger (2010): "Asymptotic size and a problem with subsampling and with the m out of n bootstrap, "*Econometric Theory,* 26, 426–468.

Andrews, D. W. K. and B. Lu (2001): "Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models," *Journal of Econometrics*, 101(1), 123-164.

Andrews, D. W. K. and J. H. Stock (2007): "Testing with many weak instruments," *Journal of Econometrics*, 138(1), 24-46.

Bai, J., and S. Ng (2009): "Selecting instrumental variables in a data rich environment," *Journal of Time Series Econometrics*, 1(1).

Belloni, A., D. Chen, V. Chernozhukov and C. Hansen (2012): "Sparse models and methods for optimal instruments with an application to eminent domain," *Econometrica,* forthcoming.

Belloni, A., V. Chernozhukov and C. Hansen (2010): "LASSO methods for Gaussian instrumental variables models," *Preprint,* arXiv:1012.1297.

Belloni, A., V. Chernozhukov and C. Hansen (2011): "Inference on treatment effects after selection amongst high-dimensional controls," *Preprint,* arXiv:1201.0224.

Berkowitz, D. M. Caner and Y. Fang (2012) :"The validity of instruments revisited," *Journal of Econometrics*, 166(2), 255-266.

Breusch, T., H. Qian, P. Schmidt and D. Wyhowski (1999): "Redundancy of moment conditions," *Journal of Econometrics*, 91 89-111.

Caner, M. (2009): "Lasso-type GMM estimator," *Econometric Theory*, 25(1), 270-290.

Caner, M. and H. Zhang (2012): "Adaptive elastic net for generalized methods of moments," *Unpublished Manuscript.*

Carrasco, M. (2012): "A regularization approach to the many instruments problem," *Journal of Econometrics*, 170(2), 383-398.

Chamberlain, G., and G. W. Imbens (2004): "Random effects estimators with many instrumental variables," *Econometrica*, 72, 295-306.

Chao J. C. and N. R. Swanson (2005): "Consistent estimation with a large number of weak instruments," *Econometrica*, 73(5), 1673-1692.

Chen, X. J. Hahn and Z. Liao (2012): "Semiparametric Two-Step GMM with Weakly Dependent Data," UCLA and Yale, *Working Paper.*

Chen, X., O. Linton, and I. van Keilegom (2003): "Estimation of semiparametric models when the criterion function is not smooth," *Econometrica*, 71(5), 1591–1608.

Conley T. G., C. B. Hansen, and P. E. Rossi (2012): "Plausibly exogenous," *Review of Economics and Statistics*, 94(1), 260-272.

Doko Tchatoka, F. and J.-M. Dufour (2012): "Identification-robust inference for endogeneity parameters in linear structural models," *MPRA Paper 40695*, University Library of Munich, Germany.

Donald, S. G., G. W. Imbens and W. K. Newey (2009) "Choosing instrumental variables in conditional moment restriction models," *Journal of Econometrics*, 152(1), 28-36.

Donald, S. G. and W. K. Newey (2001): "Choosing the number of instruments," *Econometrica*, 69(5), 1161-91.

DiTraglia, F. (2012): "Using invalid instruments on purpose: focused moment selection and averaging for GMM," University of Pennsylvania, *Working Paper.*

Eichenbaum, M. S., L. P. Hansen and K. J. Singleton (1988): "A time series analysis of representative agent models of consumption and leisure choice under uncertainty," *Quarterly Journal of Economics*, 103(1), 51-78.

Fan, J. and Y. Liao (2011): "Ultra high dimensional variable selection with endogenous covariates", Princeton University and University of Maryland, *Working Paper.*

Gautier, E., and A.B. Tsybakov (2011): "High-dimensional instrumental variables regression and confidence sets," *Preprint,* arXiv: 1105.2454v2.

Guggenberger, P. (2010): "The Impact of a Hausman pretest on the asymptotic size of a hypothesis test," *Econometric Theory*, 26, 369-382.

Guggenberger, P. (2012): "On the asymptotic size distortion of tests when instruments locally violate the exogeneity assumption," *Econometric Theory*, 28(2), 387-421.

Hall, A. R., A. Inoue, K. Jana, and C. Shin (2007): "Information in generalized method of moments estimation and entropy based moment selection," *Journal of Econometrics*, 138(2), 488–512.

Hall, A. R., A. Inoue, J. M. Nason, and B. Rossi (2010): "Information criteria for impulse response function matching estimation of DSGE Models," *Journal of Econometrics*, forthcoming.

Hall, A. R. and F. P. M. Peixe (2003): "A consistent method for the selection of relevant instruments", *Econometric Reviews*, 22(3), 269-288.

Han, C. and P. C. B. Phillips (2006): "GMM with many moment conditions," *Econometrica*, 74(1), 147-192.

Hansen, C. B. J. A. Hausman, and W. K. Newey (2008): "Estimation with many instrumental variables," *Journal of Business & Economic Statistics*, 26(4), 398-422.

Hansen, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica*, 50(4), 1029-54.

Hong, H., B. Preston and M. Shum (2003): "Generalized empirical likelihood-based model selection criteria for moment condition models," *Econometric Theory*, 19(6), 923-943.

Im, K. S., S. C. Ahn, P. Schmidt and J. M. Wooldridge (1999): "Efficient estimation of panel data models with strictly exogenous explanatory variables," *Journal of Econometrics*, 93(1), 177-201.

Inoue, A. (2006): "A bootstrap approach to moment selection," *Econometrics Journal*, 9(1), 48-75.

Kuersteiner, G. M. (2002): "Mean square error reduction for GMM estimators of linear time series models," UC Davis, *Working Paper.*

Kuersteiner, G. M. and R. Okui (2010): "Constructing optimal instruments by first-stage prediction averaging," *Econometrica*, 78(2), 697-718.

Leeb, H. and B. M. Pötscher (2005): "Model selection and inference: facts and fiction," *Econometric Theory*, 21(1), 21–59.

Leeb, H. and B. M. Pötscher (2008): "Sparse estimators and the oracle property, or the return of the Hodges estimator," *Journal of Econometrics*, 142(1), 201-211.

Liao, Z. (2011): "Adaptive GMM shrinkage estimation with consistent moment selection," UCLA, *Working Paper*.

McCloskey, A. (2012): "Bonferroni-based size-correction for nonstandard testing problems," Brown University, *Working Paper*.

Newey, W. K. and F. Windmeijer (2009): "Generalized method of moments with many weak moment conditions," *Econometrica*, 77(3), 687–719.

Nevo, A. and A. Rosen (2012): "Identification with imperfect instruments," *Review of Economics and Statistics*, 93(3), 659-671.

Okui, R. (2011): "Instrumental variable estimation in the presence of many moment conditions," *Journal of Econometrics,* 165(1), 70-86.

Pakes, A. and Pollard, D. (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57, 1027-1057.

Pollard, D. (1984): *Convergence of Stochastic Processes*. New York: Springer–Verlag.

Pollard, D. (1989): "Asymptotics via empirical processes, " *Statistical Science*, 4(4), 341-354.

Sargan, J. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26(3), 393-415.

Schmidt, M. (2010): "Graphical model structure learning with L1-regularization," *Thesis*, University of British Columbia.

Shen, X. (1997): "On Methods of sieves and penalization," *Annals of Statistics*, 25, 2555-2591.

Stock, J. H. and M. Yogo (2005): "Asymptotic distributions of instrumental variables statistics with many instruments, " in *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, D. W. K. Andrews and J. H. Stock, eds., Cambridge: Cambridge University Press, 109–120.

van der Vaart, A. W. and J. A. Wellner (1996): *Weak Convergence and Empirical Processes.* New York: Springer

Zou, H. (2006): "The adaptive lasso and its oracle properties," *Journal of the American Statistical Association*, 101(476), 1418-1429.