Journal of Econometrics 186 (2015) 443–464

Contents lists available at ScienceDirect

Journal of Econometrics

journal homepage: www.elsevier.com/locate/jeconom

Select the valid and relevant moments: An information-based LASSO for GMM with many moments☆

Xu Cheng [a,*], Zhipeng Liao [b]

[a] Department of Economics, University of Pennsylvania, 3718 Locust Walk, Philadelphia, PA 19104, USA

[b] Department of Economics, UC Los Angeles, 8379 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095, USA

ARTICLE INFO

Article history:
Available online 25 March 2015

JEL classification:
C12
C13
C36

Keywords:
Adaptive penalty
GMM
Many moments
Moment selection
Oracle properties
Shrinkage estimation

ABSTRACT

This paper studies 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 distinguished from the invalid or irrelevant ones; (ii) all desired moments are selected in one step instead of in a stepwise manner; (iii) the parameters of interest are automatically estimated with all selected moments as opposed to a post-selection estimation. The new method performs moment selection and efficient estimation simultaneously 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. We develop asymptotic results for the high-dimensional GMM shrinkage estimator, allowing for non-smooth sample moments and weakly dependent observations. For practical implementation, this one-step procedure is computationally attractive.

1. Introduction

In many applications of the generalized method of moments (GMM), the number of candidate moment conditions is much larger than that of the parameters of interest. However, one typically does not employ all candidate moment conditions due to two concerns. First, some moments may be invalid, which cause inconsistent estimation if included. Second, some moment conditions may be redundant. A redundant moment condition does not contain additional information to improve estimation 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. This paper proposes a procedure that consistently selects all valid and relevant moments in econometric models 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.

Our method achieves consistent moment selection via an information-based adaptive GMM shrinkage estimation. Assuming there exists a conservative set of moment conditions that identifies the parameter of interest, the moment selection problem is embedded in a penalized GMM (P-GMM) estimation and a novel penalty is designed to incorporate information on both moment validity and relevance for adaptive estimation. This penalized GMM estimation not only consistently selects all valid and relevant moment conditions in one step, but also simultaneously and efficiently estimates the parameters of interest by incorporating all valid and relevant moments and leaving out all invalid or redundant ones automatically. Asymptotic results provide bounds on the penalty level to ensure consistent moment selection and efficient estimation. 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.

☆ We appreciate the insightful suggestions from the co-editor and three anonymous referees. We also thank Xiaohong Chen, Jinyong Hahn, Bruce Hansen, Michael Jansson, Frank Kleibergen, Adam McCloskey, Peter C.B. Phillips, Jack Porter, Eric Renault, and participants in the 2012 Great New York Area Econometrics Colloquium, 2013 North American Summer Meeting of the Econometric Society at USC, 2013 CEME Stanford/UCLA Conference at Stanford University, econometrics workshops at Brown University, University of Montreal, and the University of Wisconsin-Madison.

* Corresponding author.
E-mail addresses: xucheng@econ.upenn.edu (X. Cheng), zhipeng.liao@econ.ucla.edu (Z. Liao).

http://dx.doi.org/10.1016/j.jeconom.2015.02.019
0304-4076/

This paper develops asymptotic results for the high-dimensional GMM shrinkage estimator in a general framework, allowing for: (i) an increasing number of candidate moments; (ii) an increasing number of nuisance parameters; (iii) non-smooth moment functions; and (iv) weakly dependent observations. High-level assumptions are first provided to capture the main characteristics of the problem, followed by primitive sufficient assumptions. We develop results on consistency, rate of convergence, super efficiency, and the asymptotic distribution of the high-dimensional GMM shrinkage estimator. A linear instrumental variable (IV) model with independent and identically distributed (i.i.d.) observations is studied in detail to illustrate the general results.

Our paper contributes to the study of moment validity and relevance, and extends it to a high-dimensional framework. There is a long history on the study of moment validity, starting from Sargan (1958), Hansen (1982), Eichenbaum et al. (1988). More recent papers include Berkowitz et al. (2012), Conley et al. (2012), Doko Tchatoka and Dufour (2012), Guggenberger (2012), Nevo and Rosen (2012), and DiTraglia (2012), among others. There are moment selection methods in the literature which select moment conditions based on their validity. 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. Andrews and Lu (2001) generalize these methods and study applications to dynamic panel models. Hong et al. (2003) study moment selection based on the generalized empirical likelihood estimation. Liao (2013) proposes a GMM shrinkage procedure for selection of valid moment conditions. These moment selection methods only take the moment validity into account and they assume that the number of the candidate moments is fixed.

On the moment relevance, Breusch et al. (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 of moment conditions does not contain additional information. For example, in the linear IV model, an IV is redundant if it does not improve the first-stage regression. Im et al. (1999) study efficient estimation in dynamic panel models in the presence of 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 samples. Hall et al. (2007) propose a moment selection criterion that balances 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 et al. (2012) propose two moment selection criteria of this sort to select all valid and relevant impulse response functions for matching estimation. There are moment selection methods in the literature which select IVs or moments via different criteria, such as the mean square error or the coverage of confidence region of the estimators of structural parameters, see, e.g., Donald and Newey (2001), Donald et al. (2009), Kuersteiner (2002), and Inoue (2006). These selection procedures assume that all candidate moments are valid.

This paper also complements a growing literature on the application of high-dimensional methods to the IV and moment based econometric models. Most papers in this literature investigate efficient estimation in the presence of many valid IVs. Belloni et al. (2010, 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 model averaging methods to approximate the optimal IV in the first-stage regression. Lee and Zhou (2011) consider averaged instrumental variable estimator. Caner and Zhang (2012) study adaptive elastic net GMM estimation with an increasing number of parameters. Caner et al. (2013) study the valid IV selection and variable selection in linear IV models. 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 the literature by combining the selection of valid and relevant moments with efficient estimation, 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 applies the main theory of this paper to a linear IV model. Section 5 analyzes the asymptotic magnitudes of the information-based penalty and provides suggestions for practical implementation of the procedure. Section 6 provides finite-sample results through simulation. Section 7 concludes and discusses related topics under investigation. The Appendix contains all the technical proofs. A separate online Supplemental Appendix contains additional supporting materials and is available on the authors' websites.

The notations are standard. Throughout the paper, C denotes some generic finite positive constant; $\|\cdot\|$ denotes the Euclidean norm; A' denotes the transpose of a matrix A; $\rho_{\max}(A)$ and $\rho_{\min}(A)$ denote the largest and smallest eigenvalues of a matrix A respectively; for $d_1 \times 1$ vector function $f(x)$: $R^{d_2} \to R^{d_1}$ we use $\frac{\partial f(x)}{\partial x'}$ to denote the $d_1 \times d_2$ matrix whose ith row and jth column element is $\frac{\partial f_i(x)}{\partial x_j}$ where $f_i(\cdot)$ and x_j are the ith component in $f(\cdot)$ and jth component in x respectively; for any square matrix A, $A \geq 0$ means that A is a positive semi-definite matrix; for any positive integers k_1 and k_2, I_{k_1} denotes the $k_1 \times k_1$ identity matrix and $\mathbf{0}_{k_1 \times k_2}$ denotes the $k_1 \times k_2$ zero matrix; $A \equiv B$ means that A is defined as B; $a_n = o_p(b_n)$ means that for any constants $\epsilon_1, \epsilon_2 > 0$, there is $\Pr(|a_n/b_n| \geq \epsilon_1) < \epsilon_2$ eventually; $a_n = O_p(b_n)$ means that for any $\epsilon > 0$, there is a finite constant C_ϵ such that $\Pr(|a_n/b_n| \geq C_\epsilon) < \epsilon$ eventually; "$\to_p$" 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 GMM shrinkage estimator

2.1. Three categories of moment conditions

There exists a vector of moment functions $g(Z, \theta)$: $R^{d_z} \times \Theta \to R^{k_n}$ for the estimation of $\theta_o \in \Theta \subset R^{d_\theta}$, where $\{Z_i : i = 1, \ldots, n\}$ is stationary and ergodic, Z is used generically for Z_i and d_θ is a fixed positive integer. 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 subvector of $g(Z, \theta)$, denoted by $g_S(Z, \theta) \in R^{k_0}$, for the identification of θ_o. Let $\mathbb{E}[\cdot]$ denote the expectation operator taken with respect to the distribution of Z. The true value of θ is identified by $\mathbb{E}[g_S(Z, \theta_o)] = 0$. We assume that k_0 is a fixed positive integer

with $k_0 \geq d_\theta$. 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.[1] Given the identification of θ_o, this paper proposes a moment selection procedure that explores all other candidate moments and yields the *largest* set of valid and relevant moment conditions.

Let $g_D(Z,\theta)$ 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

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

We also use S and D to denote the sets of indices of all moments in $g_S(Z,\theta)$ and $g_D(Z,\theta)$ respectively. Let $g_\ell(Z,\theta)$ denote an element of $g(Z,\theta)$ indexed by ℓ. Given the order of the moment conditions in (2.1), we know that $S = \{1,\ldots,k_0\}$ and $D = \{k_0+1,\ldots,k_n\}$. A moment is valid if $\mathbb{E}[g_\ell(Z,\theta_o)] = 0$ 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 $\mathbb{E}[g_S(Z,\theta_o)] = 0$.

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

$$D = A \cup B_1 \cup B_0, \tag{2.2}$$

where A indexes the set of valid and relevant moments, B_1 indexes the set of invalid moments and B_0 indexes the set of valid but redundant moments. Moments in sets A and B_0 are both valid, but only those in A are relevant. Our objective is to consistently estimate the set A, leaving out all moments indexed by the set $B \equiv B_1 \cup B_0$. We use d_A, d_B, d_{B_1}, and d_{B_0} to denote the number of moment conditions indexed by the sets A, B, B_1 and B_0, respectively. Our general theory on consistent moment selection allows d_A and d_{B_1} to increase with the sample size n and d_{B_0} to be bounded from above by some large but fixed integer. The theory only restricts the relative rates of k_n and n and this condition is discussed as the theory progresses.

2.2. Heuristic arguments for shrinkage-based moment selection

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

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

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

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

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. For any $\ell = 1,\ldots,k_n$, we assume that $|\beta_{o,\ell}| \leq C$. Below we first list all desired properties of the estimator for consistent moment selection, then propose an estimator of β_o that satisfies 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 D, the estimator $\widehat{\beta}_n$ has to satisfy

$$\Pr(\widehat{\beta}_{n,\ell} = 0, \forall \ell \in A) \to 1 \quad \text{and} \quad \Pr(\widehat{\beta}_{n,\ell} = 0, \quad \forall \ell \in B) \to 0$$

as the sample size $n \to \infty$.

Table 2.1 summarizes the properties of the slackness parameters and their estimators for all three categories. First, for the valid and relevant moment (in A), $\beta_{o,\ell}$ is 0 and we need its estimator to be 0 w.p.a.1. This super efficiency type of property can be achieved by shrinking the estimator of $\beta_{o,\ell}$ to be 0 for $\ell \in A$. Second, for the invalid moment (in B_1), the estimator of $\beta_{o,\ell}$ differs from 0 w.p.a.1 provided that it is consistent, because $\beta_{o,\ell}$ is different from 0 in this case. Heavy shrinkage of $\beta_{o,\ell}$ toward 0 for $\ell \in B_1$ causes estimation bias not only to β_o but also to θ_o. To ensure consistent estimation of θ_o and β_o, the shrinkage effect on the estimator of $\beta_{o,\ell}$ has to be controlled for $\ell \in B_1$. Third, for the redundant moment (in B_0), $\beta_{o,\ell}$ is 0 because the moment is valid. However, its 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 $\ell \in B_0$, the shrinkage effect has to be controlled to prevent the estimator of $\beta_{o,\ell}$ from having point mass at 0.

To sum up, consistent moment selection requires a sparse estimation[2] of the slackness parameters, 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 propose a P-GMM estimation that incorporates the measure of validity and relevance for each moment. The resulting P-GMM estimator is shown to satisfy all the requirements above and yields consistent moment selection.

2.3. Information measure and penalized GMM estimation

For the ease of exposition, we define $\alpha' \equiv (\theta', \beta')$ and introduce the following notations

$$\begin{array}{ll}
m(Z,\theta) \equiv \begin{bmatrix} g_S(Z,\theta) \\ g_D(Z,\theta) \end{bmatrix} & g(Z,\alpha) \equiv \begin{bmatrix} g_S(Z,\theta) \\ g_D(Z,\theta) - \beta \end{bmatrix} \\
\overline{m}(\theta) \equiv \mathbb{E}[m(Z,\theta)] & \overline{g}(\alpha) \equiv \mathbb{E}[g(Z,\alpha)] \\
\overline{m}_n(\theta) \equiv \dfrac{1}{n}\sum_{i=1}^n m(Z_i,\theta) & \overline{g}_n(\alpha) \equiv \dfrac{1}{n}\sum_{i=1}^n g(Z_i,\alpha) \\
\overline{m}_S(\theta) \equiv \mathbb{E}[g_S(Z,\theta)] & \Gamma_S(\theta) \equiv \partial \overline{m}_S(\theta)/\partial\theta' \in R^{k_0\times d_\theta} \\
\overline{m}_D(\theta) \equiv \mathbb{E}[g_D(Z,\theta)] & \Gamma_D(\theta) \equiv \partial \overline{m}_D(\theta)/\partial\theta' \in R^{(k_n-k_0)\times d_\theta} \\
\overline{m}_{S,n}(\theta) \equiv \dfrac{1}{n}\sum_{i=1}^n g_S(Z_i,\theta) & \overline{m}_{D,n}(\theta) \equiv \dfrac{1}{n}\sum_{i=1}^n g_D(Z_i,\theta) \\
\Gamma(\theta) \equiv \begin{bmatrix} \Gamma_S(\theta) & \mathbf{0}_{k_0\times(k_n-k_0)} \\ \Gamma_D(\theta) & -I_{k_n-k_0} \end{bmatrix} & \overline{m}_{S+A}(\theta) \equiv \mathbb{E}\begin{bmatrix} g_S(Z,\theta) \\ g_A(Z,\theta) \end{bmatrix}.
\end{array} \tag{2.6}$$

By definition, the parameter space of α is $\Lambda_n \equiv \Theta \times \mathcal{B}_1 \times \cdots \times \mathcal{B}_{k_n-k_0}$, where $\mathcal{B}_j \equiv \{\beta_j : \beta_j = \overline{m}_{j+k_0}(\theta) \text{ and } \theta \in \Theta\}$ for $j = 1,\ldots,k_n - k_0$. The efficient estimation and moment selection are simultaneously achieved in the P-GMM estimation

$$\widehat{\alpha}_n = \arg\min_{\alpha\in\Lambda_n}\left[\overline{g}_n(\alpha)' W_n \overline{g}_n(\alpha) + \lambda_n \sum_{\ell\in D}\omega_{n,\ell}\,|\beta_\ell|\right], \tag{2.7}$$

[1] Because θ_o is unknown, the conservative set of moment conditions is needed not only for the identification and consistent estimation of θ_o but also for defining the valid and invalid moment conditions. For example, we may have $\mathbb{E}[g_\ell(Z,\theta_1)] = 0$ and $\mathbb{E}[g_\ell(Z,\theta_o)] \neq 0$ for some moment function $g_\ell(Z,\theta)$. In this case, $\mathbb{E}[g_\ell(Z,\theta)] = 0$ is a valid moment for θ_1 but invalid for θ_o. The conservative set of moment conditions uniquely identifies θ_o and hence define valid and invalid moment conditions in $g(Z,\theta)$. The consistent estimation of θ_o makes it possible to select the valid moments when the moment function is evaluated at θ_o.

[2] The sparse estimation means that the resulting estimator may have sparse solutions. That is, when the true parameter β_o has zero elements, the estimator of β_o may contain components which are identically zero in finite samples.

Table 2.1
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
B_1—invalid	$\beta_{o,\ell}\neq 0$	$\Pr(\widehat{\beta}_{n,\ell}=0)\to 0$	consistency
B_0—valid but redundant	$\beta_{o,\ell}=0$	$\Pr(\widehat{\beta}_{n,\ell}=0)\to 0$	no shrinkage effect

where W_n is a $k_n\times k_n$ symmetric weighting 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 $\ell\in D$. This is a LASSO type estimator that penalizes each individual slackness parameter β_ℓ using its ℓ_1-norm. The ℓ_1-penalty is particularly attractive in our framework because both the GMM criterion and the ℓ_1-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 in (2.7) 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 penalty for moment conditions in different categories, as listed in Table 2.1. 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 (with $r_1\geq r_2$) 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 $\beta_{o,\ell}$ toward 0. In contrast, redundant moments (B_0) are subject to small shrinkage because $\dot{\mu}_{n,\ell}$ is asymptotically 0 for $\ell\in B_0$. This information-based adjustment $\dot{\mu}_{n,\ell}$ differentiates the relevant moments from redundant ones.

Second, when data suggest the moment ℓ 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 $\beta_{o,\ell}$ toward 0. In contrast, invalid moments (B_1) 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 $\dot{\mu}_{n,\ell}$ and $\dot{\beta}_{n,\ell}$, $\omega_{n,\ell}$ provides a data-driven adjustment that separates the valid and relevant moments (A) from the rest. The constants r_1 and r_2 (with $r_1\geq r_2$) are introduced to ensure that $\omega_{n,\ell}$ is small when the moment ℓ is redundant. Roughly speaking, the individual adjustment $\omega_{n,\ell}$ is large only when the corresponding moment condition is valid and relevant. In consequence, $\beta_{o,\ell}$ 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 $\overline{m}_S(\theta_o)=0$ are used for GMM estimation of θ_o, the asymptotic variance of the optimal GMM estimator is

$$V_S^{-1}\equiv\Gamma_S(\theta_o)'\Omega_S^{-1}(\theta_o)\Gamma_S(\theta_o),\quad \text{where}$$
$$\Omega_S(\theta_o)\equiv\lim_{n\to\infty}\mathrm{Var}\left[n^{-\frac{1}{2}}\sum_{i=1}^n g_S(Z_i,\theta_o)\right]. \tag{2.9}$$

Let $\Gamma_\ell(\theta)=\frac{\partial\mathbb{E}[g_\ell(Z,\theta)]}{\partial\theta'}$ for any ℓ and any θ. When another moment $\ell\in D$ is added, we can define a new variance $V_{S+\ell}$ analogously to V_S but with $\mathbb{E}[g_S(Z,\theta)]$ replaced by $\mathbb{E}[g_{S+\ell}(Z,\theta)]$, where $g_{S+\ell}(Z,\theta)$ is a vector that stacks $g_S(Z,\theta)$ and $g_\ell(Z,\theta)$ together, i.e.,

$$V_{S+\ell}^{-1}\equiv\begin{bmatrix}\Gamma_S(\theta_o)\\ \Gamma_\ell(\theta_o)\end{bmatrix}'\Omega_{S+\ell}^{-1}(\theta_o)\begin{bmatrix}\Gamma_S(\theta_o)\\ \Gamma_\ell(\theta_o)\end{bmatrix},\quad \text{where}$$
$$\Omega_{S+\ell}(\theta_o)\equiv\lim_{n\to\infty}\mathrm{Var}\left[n^{-\frac{1}{2}}\sum_{i=1}^n g_{S+\ell}(Z_i,\theta_o)\right]. \tag{2.10}$$

It is well-known that adding a valid (but possibly irrelevant) moment condition will not decrease the efficiency of the GMM estimator. We next show that even if a moment condition is invalid, including this moment condition in calculating the asymptotic variance of the GMM estimator does not decrease the "efficiency" either.[3]

Lemma 2.1. *Suppose that $\Omega_S(\theta_o)$ is positive definite and $\Omega_{S+\ell}(\theta_o)$ is an invertible matrix for any $\ell\in D$. Then $V_S\geq V_{S+\ell}$ for all $\ell\in D$.*

Because the matrix $V_S-V_{S+\ell}$ 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

$$\mu_\ell\equiv\rho_{\max}(V_S-V_{S+\ell}) \tag{2.11}$$

as the measure of information in the moment condition indexed by $\ell\in D$. When $\mu_\ell>0$, the moment ℓ is considered to be relevant. A suitable consistent estimator $\dot{\mu}_{n,\ell}$ is

$$\dot{\mu}_{n,\ell}=\rho_{\max}(\dot{V}_{n,S}-\dot{V}_{n,S+\ell}), \tag{2.12}$$

where $\dot{V}_{n,S}$ and $\dot{V}_{n,S+\ell}$ are consistent estimators of V_S and $V_{S+\ell}$, respectively.

To obtain $\dot{\beta}_{n,\ell}$ and $\dot{\mu}_{n,\ell}$ (for $\ell\in D$), we construct an initial GMM estimator $\dot{\alpha}_n'=(\dot{\theta}_n',\dot{\beta}_n')$, defined as

$$\dot{\alpha}_n=\arg\min_{\alpha\in\Lambda_n}\left[\overline{g}_n'(\alpha)\widetilde{W}_n\overline{g}_n(\alpha)\right], \tag{2.13}$$

where $\widetilde{W}_n$ denotes a preliminary weighting matrix (e.g., $k_n\times k_n$ identity matrix) which satisfies Assumption 3.1(iii) in the next section. The preliminary estimator $\dot{\mu}_{n,\ell}$ can be constructed using the formula (2.12) and the preliminary estimator $\dot{\theta}_n$ (see Appendix C for more details). 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 in the next section, the properties of the P-GMM estimator, e.g., consistency and rate of convergence, also hold for the initial GMM estimator $\dot{\alpha}_n$.

In the online Supplemental Appendix, we provide primitive sufficient conditions under which the convergence rates of $\dot{\beta}_{n,\ell}$ and $\dot{\mu}_{n,\ell}$ are derived. These stochastic properties are listed in

[3] Note that we do not suggest estimating θ_o using possibly invalid moments here. In practice, the variance matrix $V_{S+\ell}^{-1}$ is calculated as if the moment ℓ was valid, but the estimator of θ_o used for this calculation is based on the conservative moment conditions S.

Assumption 5.1 below as high-level assumptions. Under these high-level assumptions, the stochastic properties of $\omega_{n,\ell}$ are studied in Lemma 5.1. For the theoretical analysis in Section 3, we derive general bounds on the tuning parameter λ_n as an implicit function of $\omega_{n,\ell}$. Given the stochastic order of $\omega_{n,\ell}$, these bounds for λ_n only depend on the sample size, the number of moments, and some constants. For readers who would like to see the practical choice of λ_n directly, the rate of λ_n is provided in (5.5) and a practical choice is suggested in (5.8).

3. Asymptotic theory

3.1. Consistency and rate of convergence

We first state and discuss the assumptions for the consistency of the P-GMM estimator $\widehat{\theta}_n$.

Assumption 3.1. (i) $\overline{m}_S(\theta)$ is continuous in θ and for any $\varepsilon > 0$, there exists some $\delta_\varepsilon > 0$ such that

$$\inf_{\{\theta\in\Theta:\|\theta-\theta_o\|\geq\varepsilon\}} \|\overline{m}_S(\theta)\| > \delta_\varepsilon;$$

(ii) $\sup_{\theta\in\Theta} \|\overline{m}_n(\theta) - \overline{m}(\theta)\| = o_p(1)$;

(iii) W_n is a real matrix with $C^{-1} \leq \rho_{\min}(W_n) \leq \rho_{\max}(W_n) \leq C$ w.p.a.1;

(iv) the tuning parameter λ_n satisfies $\lambda_n \sum_{\ell\in B_1} \omega_{n,\ell} = o_p(1)$.

Assumption 3.1(i) is a standard identifiable uniqueness condition for θ_o. Assumption 3.1(ii) is essentially a uniform law of large numbers (ULLN) and it requires uniform convergence of the sample moments to the population moments. Assumption 3.1(iii) imposes regularity conditions on the weighting matrix. Assumption 3.1(iv) 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 $\widehat{\theta}_n$. 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 $\beta_{o,\ell}$ toward 0 for all ℓ and causes estimation bias for $\beta_{o,\ell} \neq 0$ and hence for θ_o. For this reason, the upper bound required by $\lambda_n \sum_{\ell\in B_1} \omega_{n,\ell} = o_p(1)$ only involves the invalid moments in B_1.

Lemma 3.1. *Under Assumption 3.1, we have* $\widehat{\theta}_n \to_p \theta_o$.

Next, we derive the rate of convergence of the P-GMM estimator $\widehat{\alpha}_n$, whose dimension increases with the sample size. Let ω_{n,B_1} denote a vector that collects $\omega_{n,\ell}$ for all $\ell \in B_1$ and

$$b_n \equiv \lambda_n \left\|\omega_{n,B_1}\right\|. \tag{3.1}$$

Assumption 3.2. (i) There exist a sequence of constants $\tau_n \to 0$ with $\tau_n^{-1} = O(n^{\frac{1}{2}})$ and a fixed constant $\delta_1 > 0$ such that

$$\sup_{\|\theta-\theta_o\|\leq\delta_1} \|\overline{m}_n(\theta) - \overline{m}(\theta)\| = O_p(\tau_n);$$

(ii) $\overline{m}(\theta)$ is continuously differentiable for any θ in the local neighborhood of θ_o;

(iii) $C^{-1} \leq \rho_{\min}\left[\Gamma_S(\theta_o)'\Gamma_S(\theta_o)\right]$ and $\rho_{\max}\left[\Gamma(\theta_o)'\Gamma(\theta_o)\right] \leq C$;

(iv) $\max_{\ell\leq k_n} \sup_{\|\theta-\theta_o\|\leq\delta_2} \|\Gamma_\ell(\theta) - \Gamma_\ell(\theta_o)\| \leq C\delta_2$ for some $\delta_2 > 0$;

(v) $\sqrt{k_n}b_n = o_p(1)$ and $\sqrt{k_n}\tau_n = o(1)$.

Assumption 3.2(i) is a high level condition on the convergence rate of the empirical process indexed by moment functions. When the number of moment conditions is fixed, Assumption 3.2(i) holds with $\tau_n = n^{-\frac{1}{2}}$, following standard empirical process results; see e.g., Andrews (1994). Here, the sequence of constants τ_n is introduced to allow for an increasing number of moments. Lemma D.1 in the Appendix D provides sufficient conditions under which Assumption 3.2(i) holds with $\tau_n = \sqrt{k_n/n}$. Assumption 3.2(ii)–(iv) impose standard regularity conditions on the first order derivative of the population moments. Assumption 3.2(v) imposes restrictions on the dimension of the moment functions k_n and the tuning parameter λ_n. Assumption 3.2(v) is a sufficient condition for Assumption 3.1(iv) because

$$\lambda_n \sum_{\ell\in B_1} \omega_{n,\ell} \left|\beta_{o,\ell}\right| \leq C\sqrt{k_n}\lambda_n \left\|\omega_{n,B_1}\right\| = C\sqrt{k_n}b_n, \tag{3.2}$$

by the Cauchy–Schwarz inequality and $|\beta_{o,\ell}| \leq C$ for any $\ell \in D$.

For some models, it is easier to verify Assumption 3.3, which is a high level assumption that can replace Assumptions 3.1 and 3.2, in conjunction with Assumption 3.1(iii).

Assumption 3.3. (i) There exists a sequence of constants $\tau_n \to 0$ with $\tau_n^{-1} = O(n^{\frac{1}{2}})$ such that

$$\|\overline{m}_n(\theta_o) - \overline{m}(\theta_o)\| = O_p(\tau_n);$$

(ii) for any $\alpha \in \Lambda_n$, $C^{-1}\|\alpha - \alpha_o\| \leq \|\overline{g}_n(\alpha) - \overline{g}_n(\alpha_o)\| \leq C\|\alpha - \alpha_o\|$ w.p.a.1.

If the data are i.i.d. and the second moment of $g_\ell(Z, \theta_o)$ is bounded from above by some finite constant uniformly over ℓ, Assumption 3.3(i) is satisfied with $\tau_n = \sqrt{k_n/n}$. When the number k_n of moment conditions is fixed, Assumption 3.3(i) holds by the central limit theorem with $\tau_n = n^{-\frac{1}{2}}$. Assumption 3.3(ii) essentially requires that the GMM criterion has a quadratic approximation. Assumption 3.3 is easy to verify in the linear IV model, as illustrated in Section 4.

Lemma 3.2. (a) *Suppose Assumptions 3.1 and 3.2 hold. Then,*

$$\|\widehat{\alpha}_n - \alpha_o\| = O_p(\tau_n + b_n);$$

(b) *Part* (a) *holds with Assumptions 3.1 and 3.2 replaced by Assumptions 3.1*(iii) *and 3.3.*

Remark 3.1. If $\lambda_n = 0$ for all n, then $\sqrt{k_n}b_n = 0$ for all n. Hence, if Assumptions 3.1(i)–(iii), 3.2 (i)–(iv) and $k_n\tau_n^2 = o(1)$ hold, Lemma 3.2(a) immediately implies that the initial GMM estimator $\dot{\alpha}_n$ defined in (2.13) satisfies that

$$\|\dot{\alpha}_n - \alpha_o\| = O_p(\tau_n). \tag{3.3}$$

If alternative Assumptions 3.1(iii) and 3.3 hold, Lemma 3.2(b) implies the same result. The convergence rate of the initial GMM estimator $\dot{\alpha}_n$ is useful to construct the adaptive penalty and the tuning parameter, as illustrated in Section 5.

Applying Lemma 3.2, we can show that the invalid moment conditions are not selected w.p.a.1 when the slackness parameters $\beta_{o,\ell}$ for any $\ell \in B_1$ are bounded away from 0 or converge to zero at a rate slower than τ_n. We consider the possibility that $\beta_{o,\ell}$ converges to 0 as the sample size increases because the number of moments in B_1 can diverge. To see this, first define

$$d_n \equiv \min_{\ell\in B_1} \left|\beta_{o,\ell}\right|. \tag{3.4}$$

We have $d_n > 0$ by definition. If $d_n \geq C > 0$, i.e., slackness parameters for invalid moments do not converge to 0, then using Lemma 3.2, we deduce that

$$\begin{aligned}\Pr\left(\min_{\ell\in B_1}|\widehat{\beta}_{n,\ell}| > 0\right) &\geq \Pr\left(\min_{\ell\in B_1}\left[|\beta_{o,\ell}| - |\widehat{\beta}_{n,\ell} - \beta_{o,\ell}|\right] > 0\right)\\ &\geq \Pr\left(d_n - \max_{\ell\in B_1}|\widehat{\beta}_{n,\ell} - \beta_{o,\ell}| > 0\right)\\ &\geq \Pr\left(C - \|\widehat{\alpha}_n - \alpha_o\| > 0\right) \to 1, \quad \text{as } n \to \infty \end{aligned} \tag{3.5}$$

which immediately implies that our method does not select the invalid moment conditions w.p.a.1. From the last inequality in (3.5), we see that the lower bound restriction $\min_{\ell\in B_1}|\beta_{o,\ell}| \geq C$ can be relaxed by applying the convergence rate of $\widehat{\alpha}_n$. Specifically, if $\|\widehat{\alpha}_n - \alpha_o\| = O_p(\tau_n)$ and the slackness parameters $\beta_{o,\ell}$ for any $\ell \in B_1$ satisfy $\tau_n = o(d_n)$, using the same arguments in (3.5), we have

$$\Pr\left(\min_{\ell\in B_1}|\widehat{\beta}_{n,\ell}| > 0\right) \geq \Pr\left(\frac{d_n}{\tau_n} > \frac{\|\widehat{\alpha}_n - \alpha_o\|}{\tau_n}\right) \to 1, \quad \text{as } n \to \infty. \tag{3.6}$$

Results in (3.5) and (3.6) immediately yield the following corollary.

Corollary 3.1 (*Invalid Moments*). (a) *Suppose Assumptions 3.1 and 3.2 hold. If we further have $d_n \geq C$ for all n, then*

$$\Pr\left(\cup_{\ell\in B_1}\left\{\widehat{\beta}_{n,\ell} = 0\right\}\right) \to 0 \quad \text{as } n \to \infty;$$

(b) *Part* (a) *holds under Assumptions 3.1, 3.2, $b_n = O_p(\tau_n)$ and $\tau_n = o(d_n)$;*

(c) *Parts* (a) *and* (b) *hold with Assumptions 3.1 and 3.2 replaced by Assumptions 3.1*(iii) *and 3.3.*

Remark 3.2. Corollary 3.1 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 $\ell \in B_1$ 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 $\ell \in B_1$ converge to zero at certain rate.

3.2. Super efficiency

In this section, we show that the valid and relevant moment conditions are selected w.p.a.1. The following restrictions on λ_n are needed.

Assumption 3.4. λ_n satisfies that (i) $b_n = O_p(\tau_n)$; (ii) $\lambda_n^{-1}\tau_n \max_{\ell\in A}\omega_{n,\ell}^{-1} = o_p(1)$.

Assumption 3.4(i) ensures $\|\widehat{\alpha}_n - \alpha_o\| = O_p(\tau_n)$. Assumption 3.4(ii) imposes a lower bound on λ_n. Assumption 3.4(ii) only involves the valid and relevant moment conditions because only β_ℓ for $\ell \in A$ is desired to be penalized heavily. This is a key condition to achieve the super efficiency on moment selection.

Theorem 3.2. (a) *Suppose Assumptions 3.1, 3.2 and 3.4 hold. Then,*

$$\Pr\left(\cap_{\ell\in A}\left\{\widehat{\beta}_{n,\ell} = 0\right\}\right) \to 1 \quad \text{as } n \to \infty;$$

(b) *Part* (a) *holds with Assumptions 3.1, 3.2 replaced by Assumptions 3.1*(iii) *and 3.3.*

Theorem 3.2 shows that all valid and relevant moments are simultaneously selected w.p.a.1, allowing for an increasing number of moments in A as $n \to \infty$. Corollary 3.1 and Theorem 3.2 are necessary but not sufficient to show that the set A is consistently estimated. For this purpose, it remains to show that the redundant moments in B_0 are not selected w.p.a.1.

3.3. Asymptotic normality

In this subsection, we establish the asymptotic distribution of the P-GMM estimator. Without loss of generality for the asymptotic results below, write $\beta' = (\beta_A', \beta_B')$, where β_A and β_B denote the sub-vector of β that collects β_ℓ for $\ell \in A$ and $\ell \in B$, respectively. Let $\widehat{\beta}_{A,n}$ and $\widehat{\beta}_{B,n}$ denote the P-GMM estimators of β_A and β_B, respectively. Theorem 3.2 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 $\alpha_B' \equiv (\theta', \beta_B')$ which is a $d_\theta + d_B$ dimensional vector. Now we stack all moment conditions and define

$$g(Z, \alpha_B) \equiv \begin{bmatrix} g_S(Z,\theta) \\ g_A(Z,\theta) \\ g_B(Z,\theta) - \beta_B \end{bmatrix}, \tag{3.7}$$

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 $\mathbb{E}[g(Z,\alpha_B)]$ with respect to α_B only depends on θ. Define

$$\Gamma_\alpha(\theta)' \equiv \begin{bmatrix} \left(\frac{\partial \mathbb{E}[g_S(Z,\theta)]}{\partial\theta'}\right)' & \left(\frac{\partial \mathbb{E}[g_A(Z,\theta)]}{\partial\theta'}\right)' & \left(\frac{\partial \mathbb{E}[g_B(Z,\theta)]}{\partial\theta'}\right)' \\ \mathbf{0}_{d_B\times k_0} & \mathbf{0}_{d_B\times d_A} & -I_{d_B} \end{bmatrix}$$

and $\Gamma_\alpha \equiv \Gamma_\alpha(\theta_o)$. (3.8)

Note that the link between $g(Z,\alpha_B)$ and $g(Z,\alpha)$ is that if we treat β_A in α to be zero, then we have $g(Z,\alpha) = g(Z,\alpha_B)$. 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 $\overline{g}_n(\alpha_o)$.

Assumption 3.5. Let $\upsilon_n(\theta) \equiv \overline{m}_n(\theta) - \overline{m}(\theta)$. There exists a sequence of constants $\varsigma_n \to 0$ such that

$$\sup_{\theta_1,\theta_2\in\{\theta\in\Theta:\|\theta-\theta_0\|\leq\delta_n\}} \frac{\|\upsilon_n(\theta_1) - \upsilon_n(\theta_2)\|}{n^{-\frac{1}{2}} + \|\theta_1 - \theta_2\|} = O_p(\varsigma_n) \tag{3.9}$$

for some sequence δ_n which converges to 0 slower than τ_n.

Assumption 3.5 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 et al. (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 3.5 holds with $o_p(1)$ on the right hand side.

A specific convergence rate ς_n associated with the empirical process $\upsilon_n(\theta)$ has to be derived in (3.9) to accommodate an increasing number of moments. Lemma D.2 in Appendix D provides primitive sufficient conditions under which Assumption 3.5 holds with $\varsigma_n = \sqrt{k_n/n}$.

Define the variance of the sample moments as

$$\Omega_n \equiv n\mathrm{Var}\left[\overline{g}_n(\alpha_o)\right]. \tag{3.10}$$

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

Assumption 3.6. (i) For any $\gamma_n \in R^{k_n}$ and $\|\gamma_n\| = 1$,

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

(ii) $C^{-1} \leq \rho_{\min}(\Omega_n) \leq \rho_{\max}(\Omega_n) \leq C$ for all n;

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

Assumption 3.6(i) assumes a triangular array central limit theorem (CLT) for scalar random sequences. Assumption 3.6(ii) requires that the variance matrix Ω_n is positive definite and bounded for all n. Assumption 3.6(iii) imposes similar regularity condition on $\Gamma_\alpha'\Gamma_\alpha$.

Assumption 3.7. The tuning parameter λ_n satisfies that $\lambda_n \left\|\omega_{n,B}\right\| = o_p(n^{-\frac{1}{2}})$.

Assumption 3.7 imposes an upper bound on λ_n, which ensures that a weighted linear combination of the P-GMM estimator has a mean-zero asymptotic normal distribution. Because $n^{-\frac{1}{2}} = O(\tau_n)$, we see that Assumption 3.7 implies Assumption 3.4(i). Moreover, when $k_n = o(n)$, Assumption 3.7 implies $\sqrt{k_n} b_n = o_p(1)$ in Assumption 3.2(v).

Assumptions 3.5–3.7 can be combined with Assumptions 3.1 and 3.2 to show the asymptotic normality of the P-GMM estimator. Alternatively, they can be used together with Assumptions 3.1(iii), 3.3, 3.8 for the same result.

Assumption 3.8. $\mathbb{E}\left[g(Z, \alpha_{1,B}) - g(Z, \alpha_{2,B})\right] = \Gamma_\alpha(\alpha_{1,B} - \alpha_{2,B})$ for any $\alpha_{1,B}$ and $\alpha_{2,B}$.

When the moment functions are linear in θ, Assumption 3.8 holds automatically.

Define

$$\Sigma_n \equiv \left(\Gamma_\alpha' W_n \Gamma_\alpha\right)^{-1} \left(\Gamma_\alpha' W_n \Omega_n W_n \Gamma_\alpha\right) \left(\Gamma_\alpha' W_n \Gamma_\alpha\right)^{-1}. \tag{3.11}$$

Theorem 3.3. (a) *Suppose Assumptions 3.1, 3.2, 3.4–3.7 hold. If we further have* $\sqrt{k_n}\tau_n^2 = o(n^{-\frac{1}{2}})$ *and* $\varsigma_n \tau_n = o(n^{-\frac{1}{2}})$, *then*

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

for any $\gamma_n \in R^{d_\theta + d_B}$ *with* $\|\gamma_n\| = 1$;

(b) *Part* (a) *holds under Assumptions* 3.1(iii), 3.3–3.8, *and* $\varsigma_n \tau_n = o(n^{-\frac{1}{2}})$.

Remark 3.3. The asymptotic distribution of the P-GMM estimator is derived by a perturbation technique on a local parameter space (see, e.g. Shen, 1997), which allows for non-smooth moment functions and an increasing number of parameters.

Remark 3.4. Theorem 3.3(a) requires $\sqrt{k_n}\tau_n^2 = o(n^{-\frac{1}{2}})$, which restricts the rate at which k_n diverges to infinity. When $\tau_n = \varsigma_n = \sqrt{k_n/n}$, which holds under the sufficient conditions in Lemmas D.1 and D.2 in the Appendix, this condition holds provided $k_n = o(n^{\frac{1}{3}})$, i.e., the number of moment conditions increases at a rate slower than $n^{\frac{1}{3}}$. On the other hand, Theorem 3.3(b) only needs $\varsigma_n \tau_n = o(n^{-\frac{1}{2}})$, which is satisfied if $k_n = o(n^{\frac{1}{2}})$, i.e., the number of moment conditions increases at a rate slower than $n^{\frac{1}{2}}$.

Remark 3.5. Theorem 3.3, in conjunction with the Cramér–Wold device, yields the asymptotic distribution of $\widehat{\theta}_n$. To see this, let $\gamma_{n,d_\theta}' = (\gamma_{d_\theta}', \mathbf{0}_{d_B}')$ where $\gamma_{d_\theta} \in R^{d_\theta}$. We have $\gamma_{n,d_\theta}'\alpha_B = \gamma_{d_\theta}'\theta$ because $\alpha_B' = (\theta', \beta_B')$. Let $\gamma_{n,d_\theta}^* = \Sigma_n^{1/2}\gamma_{n,d_\theta}\|\Sigma_n^{1/2}\gamma_{n,d_\theta}\|^{-1}$, which satisfies $\|\gamma_{n,d_\theta}^*\| = 1$. Applying Theorem 3.3, we get

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

which implies that

$$\|\Sigma_n^{1/2}\gamma_{n,d_\theta}\|^{-1}\sqrt{n}\gamma_{d_\theta}'(\widehat{\theta}_n - \theta_o) \to_d N(0, 1). \tag{3.12}$$

This asymptotic distribution can be applied to conduct inference for the parameter of interest θ_o. Let

$$\Omega_{S+A,n} \equiv \mathrm{Var}\left[n^{-\frac{1}{2}}\sum_{i=1}^n \begin{pmatrix} g_S(Z_i, \theta_o) \\ g_A(Z_i, \theta_o) \end{pmatrix}\right], \tag{3.13}$$

be a variance matrix that involves all valid and relevant moments. Suppose that W_n satisfies

$$\left\|W_n^{-1} - \Omega_n\right\| = o_p(1), \tag{3.14}$$

then we can use Assumptions 3.1(iii) and 3.6(iii) to show that[4]

$$\left|\gamma_{n,d_\theta}'\left[\Sigma_n - \left(\Gamma_\alpha'\Omega_n^{-1}\Gamma_\alpha\right)^{-1}\right]\gamma_{n,d_\theta}\right| = o_p(1), \tag{3.15}$$

where

$$\gamma_{n,d_\theta}'\left(\Gamma_\alpha'\Omega_n^{-1}\Gamma_\alpha\right)^{-1}\gamma_{n,d_\theta} = \gamma_{d_\theta}'\left[\left(\frac{\partial m_{S+A}(\theta_o)}{\partial\theta'}\right)'\Omega_{S+A,n}^{-1}\left(\frac{\partial m_{S+A}(\theta_o)}{\partial\theta'}\right)\right]^{-1}\gamma_{d_\theta}. \tag{3.16}$$

The matrix in the middle of the right hand side of (3.16) is the variance of the infeasible "oracle" estimator one would get with the complete knowledge of which moments are valid and relevant. Hence, the P-GMM estimator of θ_o is as efficient as the oracle estimator asymptotically.

Remark 3.6. When $\lambda_n = 0$, by the same arguments used to show Theorem 3.3, we obtain

$$\sqrt{n}\gamma_{\alpha,n}'\Sigma_{\alpha,n}^{-\frac{1}{2}}(\dot{\alpha}_n - \alpha_o) \to_d N(0, 1) \tag{3.17}$$

for any $\gamma_{\alpha,n} \in R^{d_\theta + k_n}$ with $\left\|\gamma_{\alpha,n}\right\| = 1$, where

$$\Sigma_{\alpha,n} \equiv \left[\Gamma(\theta_o)'W_n\Gamma(\theta_o)\right]^{-1}\left[\Gamma(\theta_o)'W_n\Omega_nW_n\Gamma(\theta_o)\right] \times \left[\Gamma(\theta_o)'W_n\Gamma(\theta_o)\right]^{-1}. \tag{3.18}$$

This together with Assumptions 3.1(iii), 3.2(iii) and $C^{-1} \leq \rho_{\min}\left[\Gamma(\theta_o)'\Gamma(\theta_o)\right]$ immediately implies the root-n normality of the preliminary estimator $\dot{\theta}_n$.

Let $\gamma_{n,\ell} \in R^{d_\theta + d_B}$ be a selection vector such that $\gamma_{n,\ell}'\alpha_B = \beta_{n,\ell}$ for any $\ell \in B_0$. Using arguments similar to those employed to derive (3.12), we can show

$$\|\Sigma_n^{1/2}\gamma_{n,\ell}\|^{-1}\sqrt{n}\widehat{\beta}_{n,\ell} \to_d N(0, 1) \quad \text{for any } \ell \in B_0. \tag{3.19}$$

Because $\widehat{\beta}_{n,\ell}$ has an asymptotic normal distribution for any individual $\ell \in B_0$, the probability that $\widehat{\beta}_{n,\ell} = 0$ approaches 0 for any individual $\ell \in B_0$. This result is particularly important for leaving out valid but redundant moments, which is not covered by Corollary 3.1. Corollary 3.4 states that all redundant moments are left out w.p.a.1 by the moment selection procedure.

Corollary 3.4 (*Redundant Moments*). (a) *Under the conditions of Theorem* 3.3(a),

$$\Pr\left(\cup_{\ell\in B_0}\left\{\widehat{\beta}_{n,\ell} = 0\right\}\right) \to 0 \quad \text{as } n \to \infty;$$

(b) *Part* (a) *holds under the conditions of Theorem* 3.3(b).

Remark 3.7. Combining Corollary 3.1, Theorem 3.2, and Corollary 3.4, we conclude that

$$\Pr(\widehat{A}_n = A) \to 1 \tag{3.20}$$

as $n \to \infty$, where the estimator $\widehat{A}_n$ is defined in (2.5). The P-GMM estimation achieves consistent moment selection under assumptions and conditions specified above.

[4] The formal proof is in the Supplemental Appendix of the paper.

Remark 3.8. The assumption that B_0 is a finite set is important for our argument to show Corollary 3.4. By the Bonferroni inequality, we have

$$\Pr\left(\cup_{\ell\in B_0}\left\{\widehat{\beta}_{n,\ell}=0\right\}\right)\leq\sum_{\ell\in B_0}\Pr\left(\widehat{\beta}_{n,\ell}=0\right). \tag{3.21}$$

As there are only finite many elements in B_0, to prove the result in Corollary 3.4, it is sufficient to have for any $\ell\in B_0$,

$$\Pr\left(\widehat{\beta}_{n,\ell}=0\right)\to 0 \quad \text{as } n\to\infty, \tag{3.22}$$

which follows from (3.19) because a random variable with an asymptotic normal distribution put zero mass at any given point w.p.a.1.[5]

Remark 3.9. Although the procedure can leave out moment conditions that are redundant to S, one potential limitation is that some moments in A might be redundant to the rest of A combined with S. That is, there could be a subset $A_0\subset A$ which contains the valid and relevant moments to S, while the rest of the moments in A, defined as A_0^c, are redundant to $A_0\cup S$.[6] To deal with this potential redundancy problem, we can use a two step procedure that selects A in the first step and selects A_0 out of A in the second step. The second step is similar to the first step but with a sequential information measure that takes into account the potential redundancy to $A_0\cup S$. This sequential information measure is defined as

$$\mu_\ell^*\equiv\rho_{\max}(V_{S+A_\ell}-V_{S+A_\ell-\ell}) \tag{3.23}$$

for $\ell=1,\ldots,d_A$, where

$$A_\ell\equiv\{j\in A: j<\ell \text{ and } \mu_j^*>0\}\cup\{j\in A: j\geq\ell\}, \tag{3.24}$$

V_{S+A_ℓ} is defined similarly to $V_{S+\ell}$ in (2.10) but with $g_\ell(Z,\theta)$ replaced by $g_{A_\ell}(Z,\theta)$, and $V_{S+A_\ell-\ell}$ is defined similarly to V_{S+A_ℓ} but with $g_\ell(Z,\theta)$ taken out of $g_{A_\ell}(Z,\theta)$. It is key that the set A_ℓ excludes those moments that are redundant to their predecessors. By definition, we set $\mu_\ell^*=0$ for any $\ell\in A_0^c$. Using the empirical analog of μ_ℓ^*, the P-GMM estimation can be used to consistently select A_0 provided that A_0^c is a finite set.[7]

4. Example: a linear IV model

In this section, we study a linear IV model to illustrate the general assumptions of the previous section. Consider the model

$$\begin{aligned}Y_i&=X_i\theta_o+u_i,\\ X_i&=\sum_{j=1}^{k_0}\pi_j Z_{1,i}(j)+\sum_{j=k_0+1}^{\infty}\pi_j Z_{1,i}(j)+v_i,\end{aligned} \tag{4.1}$$

where Y_i, X_i are scalar endogenous variables and $Z_{1,i}(j)$ for $j\in\mathbb{Z}_+\equiv\{1,2,\ldots\}$ are the excluded exogenous variables. For any vector Z, we use $Z(j)$ to denote the jth component of Z. We assume that

$$\mathbb{E}[u_i Z_{1,i}(j)]=0 \quad\text{and}\quad \mathbb{E}[v_i Z_{1,i}(j)]=0 \quad \text{for all } j. \tag{4.2}$$

For example, Eq. (4.1) can be obtained from the following conditional mean model

$$X_i=h(\omega_i)+v_i \quad \text{with } \mathbb{E}[(u_i,v_i)|\omega_i]=(0,0). \tag{4.3}$$

In this example, $Z_{1,i}(j)$ are the basis functions in the series expansion $h(\omega_i)=\sum_{j=1}^{\infty}\pi_j p_j(\omega_i)$, i.e., $Z_{1,i}(j)=p_j(\omega_i)$ for $j\in\mathbb{Z}_+$. Moment conditions in (4.2) are implied by the conditional moment restrictions in model (4.3).

We assume that an econometrician has the first k_0 IVs $Z_i^*\equiv(Z_{1,i}(1),\ldots,Z_{1,i}(k_0))'$ to construct moment conditions for identification of θ_o. The valid and relevant IVs $\underline{Z}_{1,i}\equiv(Z_{1,i}(k_0+1),\ldots,Z_{1,i}(k_0+d_A))'$ are mixed with invalid IVs $\underline{Z}_{2,i}\equiv(Z_{2,i}(1),\ldots,Z_{2,i}(d_{B_1}))'$ and irrelevant IVs $\underline{Z}_{3,i}\equiv(Z_{3,i}(1),\ldots,Z_{3,i}(d_{B_0}))'$. For this linear IV model, we have moment functions

$$g_S(Z_i,\theta)=(Y_i-X_i\theta_o)Z_i^* \tag{4.4}$$

for consistent estimation of θ_o and the following moment conditions for selection

$$\begin{aligned}g_A(Z_i,\theta)&=(Y_i-X_i\theta_o)\underline{Z}_{1,i},\\ g_{B_1}(Z_i,\theta)&=(Y_i-X_i\theta_o)\underline{Z}_{2,i},\\ g_{B_0}(Z_i,\theta)&=(Y_i-X_i\theta_o)\underline{Z}_{3,i}.\end{aligned} \tag{4.5}$$

Define $\underline{Z}_i'\equiv(\underline{Z}_{1,i}',\underline{Z}_{2,i}',\underline{Z}_{3,i}')$. Let $k_n=k_0+d_A+d_{B_1}+d_{B_0}$ denote the number of all available IVs when the sample size is n. In this example, we assume $k_n=o(n^{\frac{1}{2}})$. We next provide sufficient conditions for Assumptions 3.3, 3.5, 3.6 and 3.8, when the moment conditions are constructed from this linear IV model.

Condition 4.1. (i) $\{Y_i,X_i,Z_i^*,\underline{Z}_i\}_{i\leq n}$ *is a triangular array of i.i.d. process;*

(ii) $\mathbb{E}[u_i^2|Z_i^*]\leq C$ *and* $\mathbb{E}[u_i^2|\underline{Z}_i]\leq C$ *for all* n;

(iii) $\mathbb{E}[\|X_i\|^4]\leq C$, $\mathbb{E}[\|Z_i^*(j)\|^4]\leq C$, *and* $\mathbb{E}[\|\underline{Z}_i(j)\|^4]\leq C$ *for all* n;

(iv) $\mathbb{E}[Z_{1,i}(j)Z_{1,i}(k)]=\delta_{j,k}$ *where* $\delta_{j,k}$ *denotes the Kronecker's delta;*

(v) $\sum_{j=1}^{k_0}\pi_j^2>0$, $\mathbb{E}[\|X_i\underline{Z}_{2,i}\|]\leq C$, *and* $\mathbb{E}[X_i\underline{Z}_{3,i}]=0$ *for all* n.

The triangular array assumption is imposed because the number of IVs may increase with the sample size. Condition 4.1(ii) requires that the conditional second moments of the error term u_i given the IVs are uniformly bounded. Condition 4.1(iii) requires that the fourth moments of the endogenous variable X_i and the IVs are bounded from above uniformly. Condition 4.1(iv) implies that the valid and relevant IVs are orthonormalized, which is a normalization condition (see, e.g. Newey, 1997). Condition 4.1(v) contains three restrictions. The first restriction $\sum_{j=1}^{k_0}\pi_j^2>0$ ensures that the conservative IVs Z_i^* can be used to identify and consistently estimate θ_o. The second restriction implies that the aggregated information contained in the invalid IVs $\underline{Z}_{2,i}$ is bounded from above. The last restriction indicates that the irrelevant IVs $\underline{Z}_{3,i}$ contains no information about the endogenous variable X_i.

Lemma 4.1. *Under Condition 4.1,*

(a) $\|\overline{m}_n(\theta_o)-\overline{m}(\theta_o)\|^2=O_p(k_n/n)$;

(b) $\|\overline{g}_n(\alpha)-\overline{g}(\alpha_o)\|^2\leq C[1+O_p(k_n/n)+O_p(\sqrt{k_n/n})]\,\|\alpha-\alpha_o\|^2$;

(c) $\|\overline{g}_n(\alpha)-\overline{g}(\alpha_o)\|^2\geq C[1+O_p(k_n/n)+O_p(\sqrt{k_n/n})]\,\|\alpha-\alpha_o\|^2$;

(d) *Assumption 3.5 is satisfied with* $\varsigma_n=\sqrt{k_n/n}$.

[5] The proposed method can be extended to consistent moment selection with an increasing number of moments in B_0 if one can derive the rate at which $\Pr(\widehat{\beta}_{n,\ell}=0)\to 0$ uniformly over $\ell\in B_0$.

[6] It is clear that in the general scenario, the non-redundant set A_0 in A may not be unique. However, any non-redundant set in A can be uniquely determined by an order among the moments in A. That is, if we order the moments in A and delete the redundant moments by investigating the moments from the first to the last, we get a non-redundant set A_0 determined by the order. We implicitly impose an order on the moments in A and our calculation of the empirical information measure below follows this order.

[7] This is a heuristic argument based on proofs in this paper. A formal proof for this modified sequential procedure is beyond the scope of this paper and left for future research. Also note that the two step procedure is needed to ensure that moment conditions in A_0 which may be redundant to the moments in B_1 will not be ruled out.

Lemma 4.1(a) implies Assumption 3.3(i) holds with $\tau_n = \sqrt{k_n/n}$. Lemmas 4.1(b) and (c) and $k_n = o(n^{\frac{1}{2}})$ imply Assumption 3.3(ii). Because $\tau_n = \sqrt{k_n/n}$ and $k_n = o(n^{\frac{1}{2}})$, Lemma 4.1(d) implies that $\varsigma_n\tau_n = k_n n^{-1} = o(n^{-\frac{1}{2}})$, which is required by Theorem 3.3(b).

By definition, we write

$$\Omega_n = \mathbb{E}\left[\begin{pmatrix} u_i Z_i^* \\ u_i \underline{Z}_i - \beta_o \end{pmatrix}\begin{pmatrix} u_i Z_i^* \\ u_i \underline{Z}_i - \beta_o \end{pmatrix}'\right] \equiv \begin{pmatrix} \Omega_{0,n} & \Omega_{1,n} \\ \Omega_{1,n}' & \Omega_{2,n} \end{pmatrix} \tag{4.6}$$

where

$$\begin{aligned}
&\Omega_{0,n} \equiv \begin{pmatrix} \mathbb{E}\left[u_i^2 Z_i^* Z_i^{*\prime}\right] & \mathbb{E}\left[u_i^2 Z_i^* \underline{Z}_{1,i}'\right] \\ \mathbb{E}\left[u_i^2 \underline{Z}_{1,i} Z_i^{*\prime}\right] & \mathbb{E}\left[u_i^2 \underline{Z}_{1,i} \underline{Z}_{1,i}'\right] \end{pmatrix}, \\
&\Omega_{1,n}' \equiv \begin{pmatrix} \mathbb{E}\left[u_i^2 \underline{Z}_{-1,i} Z_i^{*\prime}\right] & \mathbb{E}\left[u_i^2 \underline{Z}_{-1,i} \underline{Z}_{1,i}'\right] \end{pmatrix}, \\
&\Omega_{2,n} \equiv \mathbb{E}\left[u_i^2 \underline{Z}_{-1,i} \underline{Z}_{-1,i}'\right] - \beta_{B,o}\beta_{B,o}', \\
&\text{and} \quad \underline{Z}_{-1,i}' \equiv (\underline{Z}_{2,i}', \underline{Z}_{3,i}').
\end{aligned} \tag{4.7}$$

Let $\Omega_n^{11} \equiv \Omega_{0,n} - \Omega_{1,n}\Omega_{2,n}^{-1}\Omega_{1,n}'$ and $\Omega_n^{22} \equiv \Omega_{2,n} - \Omega_{1,n}'\Omega_{0,n}^{-1}\Omega_{1,n}$.

Condition 4.2. (i) $\mathbb{E}[u_i^4 \,|\, Z_i^*] \le C$ *and* $\mathbb{E}[u_i^4 \,|\, \underline{Z}_i] \le C$ *for all* n;
(ii) $\rho_{\max}(\mathbb{E}[\underline{Z}_{-1,i}\underline{Z}_{-1,i}']) \le C$ *for all* n;
(iii) $\rho_{\min}(\Omega_n^{11}) \ge C^{-1}$ *and* $\rho_{\min}(\Omega_n^{22}) \ge C^{-1}$ *for all* n.

Condition 4.2(i) is stronger than Condition 4.1(i). The finite conditional fourth moment of the error term is a regularity condition for showing asymptotic normality of plug-in series estimator in conditional mean models, see, e.g., Newey (1997). Condition 4.2(ii) imposes an upper bound for the largest eigenvalue of $\mathbb{E}[\underline{Z}_{-1,i}\underline{Z}_{-1,i}']$ which is also a mild condition. To see the intuition behind Condition 4.2(iii), we write

$$\begin{aligned}
&\Omega_{0,n} - \Omega_{1,n}\Omega_{2,n}^{-1}\Omega_{1,n}' \\
&\quad = \mathbb{E}\left[u_i^2 (Z_{1,i} - \Omega_{1,n}\Omega_{2,n}^{-1}\underline{Z}_{-1,i})(Z_{1,i} - \Omega_{1,n}\Omega_{2,n}^{-1}\underline{Z}_{-1,i})'\right]
\end{aligned} \tag{4.8}$$

where $u_i Z_{1,i} - \Omega_{1,n}\Omega_{2,n}^{-1} u_i \underline{Z}_{-1,i}$ is the residual of the projection of $u_i Z_{1,i}$ on the space spanned by $u_i\underline{Z}_{-1,i}$. Hence, $\rho_{\min}(\Omega_n^{11}) \ge C^{-1}$ implies that for any random variable $\gamma' Z_{1,i} u_i$ generated by some linear combination of $Z_{1,i}u_i$, the optimal mean square linear prediction error based on the set of variables $u_i\underline{Z}_{-1,i}$ is bounded from below by some fixed constant for all n. The similar intuition applies to the restriction $\rho_{\min}(\Omega_n^{22}) \ge C^{-1}$. Hence, Condition 4.2(iii) requires that the distance between the Hilbert spaces generated by $u_i\underline{Z}_{1,i}$ and $u_i\underline{Z}_{-1,i}$ is bounded away from zero for all n.

Lemma 4.2. *Under Conditions 4.1 and 4.2, Assumptions 3.6 and 3.8 are satisfied.*

5. Selection of the tuning parameter

The asymptotic results established in previous sections 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 rely on the validity as well as relevance of the moment condition ℓ by construction. In this section, we analyze these individual penalties and provide an explicit formula for the tuning parameter λ_n.

Specifically, we consider the case where

$$\tau_n = \varsigma_n = k_n^{\frac{1}{2}} n^{-\frac{1}{2}}. \tag{5.1}$$

These specific rates apply under the sufficient conditions in Lemmas D.1 and D.2 in the Appendix A. Our goal is to choose λ_n that satisfies the upper bound

$$\lambda_n\left(n^{\frac{1}{2}}\left\|\omega_{n,B}\right\|\right) = o_p(1), \tag{5.2}$$

and the lower bound

$$\lambda_n^{-1}\left(k_n^{\frac{1}{2}} n^{-1/2} \max_{\ell\in A} \omega_{n,\ell}^{-1}\right) = o_p(1). \tag{5.3}$$

Under (5.1), conditions in (5.2) and (5.3) imply Assumptions 3.1(iv), 3.2(v), 3.4 and 3.7.

Assumption 5.1. (i) $\max_{\ell\in D}\left|\dot{\mu}_{n,\ell} - \mu_{o,\ell}\right| = O_p(\tau_n)$;
(ii) $\max_{\ell\in D}|\dot{\beta}_{n,\ell} - \beta_{o,\ell}| = O_p(\tau_n)$;
(iii) For any $\ell \in B_0$, $\left|\dot{\mu}_{n,\ell} - \mu_{o,\ell}\right| = O_p(n^{-\frac{1}{2}})$ and $\sqrt{n}(\dot{\beta}_{n,\ell} - \beta_{o,\ell}) \to_d N(0, \sigma_\ell^2)$ with $\sigma_\ell^2 > 0$;
(iv) $\max_{\ell\in A}|\mu_{o,\ell}^{-1}| \le C$, $\max_{\ell\in B_1}\left|\mu_{o,\ell}\right| \le C$, and $\max_{\ell\in B_1}|\beta_{o,\ell}^{-1}| \le C$.

Assumption 5.1(i) imposes a restriction on the convergence rate of the empirical information measure. Assumption 5.1(ii) is implied by (3.3), because

$$\max_{\ell\in D}\left|\dot{\beta}_{n,\ell} - \beta_{o,\ell}\right| \le \left\|\dot{\beta}_n - \beta_o\right\| \le \|\dot{\alpha}_n - \alpha_o\| = O_p(\tau_n). \tag{5.4}$$

Assumption 5.1(iii) is standard because the assumptions are on individual moments. The root-n normality of $\dot{\beta}_{n,\ell}$ is implied by (3.17) and the Cramer–Wold device. Assumption 5.1(iv) imposes bounds on $\left|\mu_{o,\ell}\right|$ for $\ell \in A$ and $\left|\beta_{o,\ell}\right|$ for $\ell \in B_1$.

Sufficient conditions for Assumptions 5.1(i) and 5.1(iii) are provided in the Supplemental Appendix.

Remark 5.1. In Appendix C, we show that the tuning parameter specified in (5.5) also allows $\min_{\ell\in A}\left|\mu_{o,\ell}\right|$ and $\min_{\ell\in B_1}\left|\beta_{o,\ell}\right|$ to go to zero at certain rate as $k_n \to \infty$, which relaxes Assumption 5.1(iv).

Lemma 5.1. *Suppose Assumption 5.1 holds.*

(a) *The upper bound in (5.2) is satisfied if* $\lambda_n k_n^{\frac{1}{2}} n^{\frac{1}{2}} = o(1)$ *and* $\lambda_n n^{\frac{1+r_2-r_1}{2}} = o(1)$;

(b) *The lower bound in (5.3) is satisfied if* $\lambda_n^{-1} k_n^{\frac{1+r_2}{2}} n^{-\frac{1+r_2}{2}} = o(1)$.

Remark 5.2. By choosing $r_1 \ge r_2$, Lemma 5.1(a) only requires that $\lambda_n k_n^{\frac{1}{2}} n^{\frac{1}{2}} = o(1)$. On the other hand, Lemma 5.1(b) requires that $\lambda_n^{-1} k_n^{\frac{1+r_2}{2}} n^{-\frac{1+r_2}{2}} = o(1)$. To balance these two rates, we set

$$\lambda_n = c k_n^{\frac{r_2}{4}} n^{-\frac{1}{2}-\frac{r_2}{4}} \tag{5.5}$$

where c is some finite positive constant.

Remark 5.3. Given r_1 and r_2 and $r_1 > r_2$, we propose a plug-in loading constant based on the argument in Liao (2013):

$$\widehat{c}_{\ell,n} = 2\|W_n^{\frac{1}{2}}(\ell)\widehat{\Pi}_n\| \quad \text{for any } \ell \in D \tag{5.6}$$

where $W_n^{\frac{1}{2}}(\ell)$ denotes the ℓth row of the matrix W_n, and $\widehat{\Pi}_n$ is an estimator of the following matrix

$$\Pi_n = I_{k_n} - W_n^{\frac{1}{2}}\Gamma_\alpha\left(\Gamma_\alpha' W_n \Gamma_\alpha\right)^{-1}\Gamma_\alpha' W_n^{\frac{1}{2}} \tag{5.7}$$

based on a preliminary P-GMM estimation with $\lambda_n = 2k_n^{\frac{r_2}{4}} n^{-\frac{1}{2}-\frac{r_2}{4}}$. To sum up, we propose using the tuning parameter

$$\widehat{\lambda}_{n,\ell} = 2\|W_n^{\frac{1}{2}}(\ell)\widehat{\Pi}_n\| k_n^{\frac{r_2}{4}} n^{-\frac{1}{2}-\frac{r_2}{4}} \tag{5.8}$$

for the ℓth moment condition in D.

6. Simulation

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

$$Y_1 = Y_2\theta_o + u, \tag{6.1}$$

where $Y_1, Y_2 \in R$ are endogenous and $\theta_o \in R$ is the parameter of interest. Valid and relevant IVs $Z_S \in R^2$ are available for the identification of θ_o. In addition, a vector of candidate IVs $Z_D = (Z_A, Z_{B_0}, Z_{B_1}) \in R^K$ are considered, where $Z_A \in R^{d_A}$ ($d_A = 2$) are valid and relevant, $Z_{B_0} \in R^{d_{B_0}}$ ($d_{B_0} = K/2 - 1$) are redundant, and $Z_{B_1} \in R^{d_{B_1}}$ ($d_{B_1} = K/2-1$) are invalid. We consider $K = 10, 30$, and 50 in the experiments. The relationship between Y_2 and (Z_S, Z_A) is

$$Y_2 = \pi_S' Z_S + \pi_A' Z_A + v \tag{6.2}$$

where π_S and π_A are $d_S \times 1$ and $d_A \times 1$ real vectors respectively.

The simulated samples are generated in the following way. First we generate

$$(Z_S, Z_A, Z_{B_0}, Z^*_{B_1}, u, v) \sim N(0, \Sigma),$$
$$\text{where } \Sigma = diag(\Sigma_{AS}, \Sigma_B, \Sigma_{uv}) \tag{6.3}$$

where Σ_{AS}, Σ_B and Σ_{uv} are 4×4, $(K-2) \times (K-2)$ and 2×2 variance–covariance matrices respectively. By construction, $(Z_S, Z_A, Z_{B_0}, Z^*_{B_1})$ are all valid, but only Z_S and Z_A are relevant based on (6.2). The invalid IVs Z_{B_1} are obtained by contaminating $Z^*_{B_1}$ with the structural error u. Specifically,

$$Z_{B_1}(\ell) = Z^*_{B_1}(\ell) + c_\ell \times u, \tag{6.4}$$

where c_ℓ is a real constant, $Z_{B_1}(\ell)$ and $Z^*_{B_1}(\ell)$ are the ℓth element of Z_{B_1} and $Z^*_{B_1}$, respectively. The structure of (6.4) 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) $\pi_S = (\pi_o, 0.1)'$, where the value $\pi_o = 0.1$ or 0.3 to experiment different identification strength; (iii) $\pi_A = (0.5, 0.5)$; (iv) Σ_{AS} is a 4×4 matrix with the (i, j)th element being $0.2^{|i-j|}$; (v) Σ_B is an $(K-2) \times (K-2)$ 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 $\ell = 1, \ldots, (K/2 - 1)$, the coefficients in (6.4) are

$$c_\ell = c_o + \frac{(\ell - 1)(\bar{c} - c_o)}{K/2 - 1}, \tag{6.5}$$

where $\bar{c} = 2.4$ sets a large upper bound. A larger value of c_o is associated with stronger endogeneity of the invalid IVs.

For each specification of (π_o, c_o, K), we generate i.i.d. observations with sample size $n = 250$, $n = 2500$ and $n = 5000$. 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 (2.13). The weighting matrix W_n is defined as

$$W_n^{-1} = \frac{1}{n}\left[\sum_{i=1}^{n} g(Z, \dot{\alpha}_n) g(Z, \dot{\alpha}_n)'\right] \tag{6.6}$$

where $\dot{\alpha}_n$ is defined in (2.13) with identity weighting matrix, $g(Z, \alpha_n)$ is constructed using the IVs $(Z_S, Z_A, Z_{B_0}, Z_{B_1})$. 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 6.1 presents the finite-sample performances of the moment selection 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 small sample size ($n = 250$). In this case, the probabilities of any invalid IVs being selected are small for $K = 10, 30$, and 50. Hence, the shrinkage procedure succeeds in selecting only the valid IVs. The number of the moment conditions affect the probabilities of valid and/or relevant moment conditions to be selected. When $K = 10$, with a probability of 0.69, Z_A is the set of IVs selected and with a probability of 0.19, Z_A plus some elements in Z_{B_0} are selected. This implies that with a probability of 0.88, the shrinkage procedure selects *all* of the valid and relevant IVs. When K increases, the probability of selecting Z_A alone decreases and the probability of selecting Z_A plus some elements in Z_{B_0} increases. For example, when $K = 50$, the probability of selecting Z_A drops to 0.24, while the probability of selecting Z_A plus some elements in Z_{B_0} increases to 0.62. When sample size is $n = 2500$, the probabilities of selecting Z_A are 0.96 with $K = 10$, 0.90 with $K = 30$ and 0.85 with $K = 50$, whereas the probabilities of selecting invalid IVs are 0 and the probabilities of selecting redundant IVs are as low as 0.14 even with $K = 50$. When sample size is $n = 5000$, the probabilities of selecting Z_A are larger than 0.90 and the probabilities of selecting invalid or redundant IVs are close to zero. 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 ($\pi_o = 0.1$) 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.

The P-GMM estimator proposed in this paper produces an automatic estimate of θ_o in the shrinkage estimation. Table 6.2 summaries finite-sample properties of this estimator denoted by "automatic" in Table 6.2, and compares it with several alternative estimators.[8] 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_S 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_S, Z_A, and Z_{B_0}. 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 6.1 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. Finally, an important comparison is between the "automatic" estimator and the infeasible "oracle" estimator, which uses the desirable

[8] We only present finite sample properties of various GMM estimators with $K = 50$ here. More simulation results are available in the Supplemental Appendix of the paper.

Table 6.1
Performance of moment selection by GMM shrinkage estimation.

	$\pi_o = 0.1, K = 10$		
	n = 250	n = 2500	n = 5000
$c_o = 0.2$	.011 .495 .482 .012	.000 .906 .092 .002	.000 .951 .049 .000
$c_o = 0.5$	.002 .497 .489 .012	.000 .906 .092 .002	.000 .951 .049 .000
	$\pi_o = 0.3, K = 10$		
$c_o = 0.2$	.001 .687 .184 .128	.000 .961 .027 .012	.000 .982 .014 .004
$c_o = 0.5$	.000 .687 .184 .129	.000 .961 .027 .012	.000 .982 .014 .004
	$\pi_o = 0.1, K = 30$		
$c_o = 0.2$	.021 .132 .834 .013	.000 .701 .297 .002	.000 .828 .171 .001
$c_o = 0.5$	.007 .133 .847 .013	.000 .701 .297 .002	.000 .828 .171 .001
	$\pi_o = 0.3, K = 30$		
$c_o = 0.2$	.001 .398 .467 .135	.000 .902 .086 .011	.000 .945 .051 .004
$c_o = 0.5$	.000 .398 .468 .135	.000 .902 .086 .011	.000 .945 .051 .004
	$\pi_o = 0.1, K = 50$		
$c_o = 0.2$	.025 .052 .910 .013	.000 .558 .440 .002	.000 .731 .268 .001
$c_o = 0.5$	.009 .052 .927 .013	.000 .558 .440 .002	.000 .731 .268 .001
	$\pi_o = 0.3, K = 50$		
$c_o = 0.2$	.001 .241 .622 .136	.000 .849 .141 .010	.000 .918 .076 .006
$c_o = 0.5$	.000 .241 .623 .136	.000 .849 .141 .010	.000 .918 .076 .006

Note: 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 6.2
Finite sample bias (BS), Standard deviations (SD) and RMSEs (RE) with $K = 50$.

	Automatic estimate						Conservative GMM estimate					
	$n = 250$			$n = 2500$			$n = 250$			$n = 2500$		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.1.2)	.0067	.0855	.0858	.0004	.0254	.0254	.0044	.2581	.2581	-.0018	.0770	.0770
(.1.5)	.0063	.0858	.0860	.0004	.0254	.0254	.0044	.2581	.2581	-.0018	.0770	.0770
(.3.2)	.0045	.0811	.0812	.0003	.0237	.0237	.0004	.1608	.1608	-.0012	.0490	.0490
(.3.5)	.0045	.0811	.0812	.0003	.0237	.0237	.0004	.1608	.1608	-.0012	.0490	.0490
	Pooled GMM Estimate						Aggressive GMM Estimate					
(.1.2)	.0231	.0766	.0800	.0026	.0253	.0254	.1957	.0770	.2103	.1936	.0253	.1952
(.1.5)	.0231	.0766	.0800	.0026	.0253	.0254	.2034	.0770	.2175	.2030	.0253	.2046
(.3.2)	.0199	.0717	.0744	.0021	.0236	.0237	.1711	.0721	.1856	.1689	.0236	.1706
(.3.5)	.0199	.0717	.0744	.0021	.0236	.0237	.1778	.0721	.1918	.1773	.0236	.1788
	Post-Shrinkage GMM Estimate						Oracle GMM Estimate					
(.1.2)	.0048	.0828	.0830	.0003	.0253	.0254	.0012	.0773	.0773	.0004	.0253	.0253
(.1.5)	.0048	.0833	.0835	.0003	.0253	.0254	.0012	.0773	.0773	.0004	.0253	.0253
(.3.2)	.0061	.0900	.0902	.0003	.0240	.0240	.0008	.0722	.0723	.0002	.0236	.0236
(.3.5)	.0061	.0900	.0902	.0003	.0240	.0240	.0008	.0722	.0723	.0002	.0236	.0236

Note: (i) The "automatic" estimation is obtained simultaneously with moment selection. (ii) The "conservative" estimation uses Z_S. (iii) The "pooled" estimation uses all valid IVs, including Z_S, Z_A, and Z_{B_0}. (iv) The "aggressive" estimation uses all available IVs, including invalid ones. (v) The "post-shrinkage" estimation uses Z_S plus IVs selected by the shrinkage procedure. (vi) The "oracle" estimation uses Z_S and Z_A.

IVs Z_S 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 small 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 6.1, but also automatically estimate the parameter of interest. Table 6.2 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.

7. 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 et al. (2011), and McCloskey (2012), among others. Robust inference on the parameter of interest is beyond the scope of this paper and will be investigated in future research.

Appendix

In this appendix, for any two sequences a_n and b_n, we use $a_n \lesssim b_n$ to denote that $a_n \leq Cb_n$ where C is some fixed finite positive constant.

Appendix A. Proofs of main results in Sections 2 and 3

Proof of Lemma 2.1. For the ease of notation, we write

$$\Gamma_S \equiv \Gamma_S(\theta_o), \qquad \Gamma_\ell \equiv \Gamma_\ell(\theta_o), \qquad \Omega_S \equiv \Omega_S(\theta_o)$$
$$\text{and} \quad \Omega_{S+\ell} \equiv \Omega_{S+\ell}(\theta_o). \tag{A.1}$$

By definition, we write

$$\Omega_{S+\ell} \equiv \begin{pmatrix} \Omega_S & \Omega_{S,\ell} \\ \Omega_{\ell,S} & \Omega_\ell \end{pmatrix} \tag{A.2}$$

where Ω_S is the leading $k_0 \times k_0$ sub-matrix of $\Omega_{S+\ell}$, Ω_ℓ is the last diagonal element of $\Omega_{S+\ell}$ and $\Omega_{S,\ell} = \Omega'_{\ell,S}$ are corresponding sub-matrices of $\Omega_{S+\ell}$.

By the inverse formula of a block matrix, we have

$$\Omega_{S+\ell}^{-1} \begin{bmatrix} \Omega_S & \Omega_{S,\ell} \\ \Omega_{\ell,S} & \Omega_{\ell,S}\Omega_S^{-1}\Omega_{S,\ell} \end{bmatrix} \Omega_{S+\ell}^{-1} = \begin{bmatrix} \Omega_S^{-1} & \mathbf{0}_{k_o\times 1} \\ \mathbf{0}_{1\times k_o} & 0 \end{bmatrix}, \tag{A.3}$$

which further implies that

$$\begin{aligned} V_{S+\ell}^{-1} - V_S^{-1} &= \begin{bmatrix} \Gamma_S \\ \Gamma_\ell \end{bmatrix}' \left(\Omega_{S+\ell}^{-1} - \begin{bmatrix} \Omega_S^{-1} & \mathbf{0}_{k_o\times 1} \\ \mathbf{0}_{1\times k_o} & 0 \end{bmatrix} \right) \begin{bmatrix} \Gamma_S \\ \Gamma_\ell \end{bmatrix} \\ &= \begin{bmatrix} \Gamma_S \\ \Gamma_\ell \end{bmatrix}' \Omega_{S+\ell}^{-1} \left(\Omega_{S+\ell} - \begin{bmatrix} \Omega_S & \Omega_{S,\ell} \\ \Omega_{\ell,S} & \Omega_{\ell,S}\Omega_S^{-1}\Omega_{S,\ell} \end{bmatrix} \right) \Omega_{S+\ell}^{-1} \begin{bmatrix} \Gamma_S \\ \Gamma_\ell \end{bmatrix} \\ &= \begin{bmatrix} \Gamma_S \\ \Gamma_\ell \end{bmatrix}' \Omega_{S+\ell}^{-1} \begin{bmatrix} \mathbf{0}_{k_o\times k_o} & \mathbf{0}_{k_o\times 1} \\ \mathbf{0}_{1\times k_o} & \Omega_\ell - \Omega_{\ell,S}\Omega_S^{-1}\Omega_{S,\ell} \end{bmatrix} \Omega_{S+\ell}^{-1} \begin{bmatrix} \Gamma_S \\ \Gamma_\ell \end{bmatrix}, \end{aligned} \tag{A.4}$$

where

$$\begin{aligned} &\Omega_\ell - \Omega_{\ell,S}\Omega_S^{-1}\Omega_{S,\ell} \\ &= \lim_{n\to\infty} \operatorname{Var}\left[n^{-\frac{1}{2}} \sum_{i=1}^n \left[g_\ell(Z_i,\theta_o) - g_S'(Z_i,\theta_o)\Omega_S^{-1}\Omega_{S,\ell} \right] \right] \\ &\geq 0. \end{aligned} \tag{A.5}$$

This implies that $V_{S+\ell}^{-1} - V_S^{-1} \geq 0$ and hence $V_S \geq V_{S+\ell}$. The second result is an immediate implication of the first. ■

Proof of Lemma 3.1. Recall $\upsilon_n(\theta) = \overline{m}_n(\theta) - \overline{m}(\theta)$. Note that $\upsilon_n(\theta) = \overline{g}_n(\alpha) - \overline{g}(\alpha)$ for any $\alpha \in \mathcal{A}$, and $\upsilon_n(\theta_o) = \overline{g}_n(\alpha_o)$ because $\overline{g}(\alpha_o) = 0$. Hence, by Assumptions 3.1(ii) and (iii), we get

$$\overline{g}_n(\alpha_o)' W_n \overline{g}_n(\alpha_o) = \upsilon_n(\theta_o)' W_n \upsilon_n(\theta_o) = o_p(1). \tag{A.6}$$

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

$$\begin{aligned} &\overline{g}_n(\widehat{\alpha}_n)' W_n \overline{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell\in D} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \\ &\leq \overline{g}_n(\alpha_o)' W_n \overline{g}_n(\alpha_o) + \lambda_n \sum_{\ell\in D} \omega_{n,\ell} \left|\beta_{o,\ell}\right|. \end{aligned} \tag{A.7}$$

Let $\overline{m}_{S,n}(\widehat{\theta}_n)$ and $\overline{m}_{D,n}(\widehat{\theta}_n)$ denote the subvectors of $\overline{m}_n(\widehat{\theta}_n)$ associated with moments in S and D, respectively. The inequality in (A.7) implies

$$\left\|\overline{m}_{S,n}(\widehat{\theta}_n)\right\|^2 + \left\|\overline{m}_{D,n}(\widehat{\theta}_n) - \widehat{\beta}_n\right\|^2 = \|\overline{g}_n(\widehat{\alpha}_n)\|^2 = o_p(1), \tag{A.8}$$

because (i) $\lambda_n \sum_{\ell\in D} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \geq 0$; (ii) $\overline{g}_n(\alpha_o)' W_n \overline{g}_n(\alpha_o) = o_p(1)$ by (A.6); (iii) $\beta_{o,\ell} = 0$ for $\ell \notin B_1$; (iv) $\lambda_n \sum_{\ell\in B_1} \omega_{n,\ell} \left|\beta_{o,\ell}\right| = o_p(1)$ by Assumption 3.1(iv) and $|\beta_{o,\ell}| < C$ for all ℓ; and (iv) $\rho_{\min}(W_n) \geq C^{-1}$ w.p.a.1 by Assumption 3.1(iii).

Using the triangle inequality and result in (A.8), we have

$$o_p(1) = \left\|\overline{m}_{S,n}(\widehat{\theta}_n)\right\| \geq \left| \left\|\overline{m}_S(\widehat{\theta}_n)\right\| - \left\|\overline{m}_{S,n}(\widehat{\theta}_n) - \overline{m}_S(\widehat{\theta}_n)\right\| \right|, \tag{A.9}$$

which combined with Assumption 3.1(ii) implies that $\|\overline{m}_S(\widehat{\theta}_n)\| = o_p(1)$. Under Assumption 3.1(i), $\|\overline{m}_S(\widehat{\theta}_n)\| = o_p(1)$ implies that $\widehat{\theta}_n \to_p \theta_o$. ■

Proof of Lemma 3.2. We first prove part (a). Because $\lambda_n \omega_{n,\ell} \geq 0$ for all ℓ, the triangle inequality and the Cauchy–Schwarz inequality imply that,

$$\lambda_n \sum_{\ell\in B_1} \omega_{n,\ell} \left|\beta_{o,\ell}\right| - \lambda_n \sum_{\ell\in B_1} \omega_{n,\ell} \left|\widehat{\beta}_{n,\ell}\right| \leq b_n \left\|\widehat{\beta}_n - \beta_o\right\|. \tag{A.10}$$

Combining the inequalities in (A.7) and (A.10) and using the fact that $\lambda_n \omega_{n,\ell} \geq 0$ and $\beta_{o,\ell} = 0$ for $\ell \notin B_1$, we obtain

$$\overline{g}_n(\widehat{\alpha}_n)' W_n \overline{g}_n(\widehat{\alpha}_n) \leq b_n \left\|\widehat{\beta}_n - \beta_o\right\| + \overline{g}_n(\alpha_o)' W_n \overline{g}_n(\alpha_o). \tag{A.11}$$

By Assumption 3.1(iii) and $\overline{g}_n(\alpha_o)' W_n \overline{g}_n(\alpha_o) = O_p(\tau_n^2)$ under Assumption 3.2(i), we have

$$\begin{aligned} &\left\|\overline{m}_{S,n}(\widehat{\theta}_n)\right\|^2 + \left\|\overline{m}_{D,n}(\widehat{\theta}_n) - \widehat{\beta}_n\right\|^2 \\ &\lesssim b_n \left\|\widehat{\beta}_n - \beta_o\right\| + O_p(\tau_n^2) \end{aligned} \tag{A.12}$$

w.p.a.1.

To derive the rate of convergence of $\widehat{\theta}_n$, we next study the two terms in the left hand side of (A.12) and link them to $\|\widehat{\theta}_n - \theta_o\|$. First, note that

$$\begin{aligned} \left\|\overline{m}_{S,n}(\widehat{\theta}_n)\right\|^2 &= \left\|\overline{m}_{S,n}(\widehat{\theta}_n) - \overline{m}_S(\widehat{\theta}_n) + \overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o)\right\|^2 \\ &\geq \left\|\overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o)\right\|^2 - 2\left\|\overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o)\right\| \\ &\quad \times \left\|\overline{m}_{S,n}(\widehat{\theta}_n) - \overline{m}_S(\widehat{\theta}_n)\right\| \\ &= \left\|\overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o)\right\|^2 - O_p(\tau_n)\left\|\overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o)\right\|, \end{aligned} \tag{A.13}$$

where the first equality holds because $\overline{m}_S(\theta_o) = 0$, the inequality follows from an expansion of the quadratic term and the Cauchy–Schwarz inequality, the $O_p(\tau_n)$ term in the last equality follows from Assumption 3.2(i) and the consistency of $\widehat{\theta}_n$. By the mean value theorem,

$$\begin{aligned} \overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o) &= \Gamma_S(\theta_o)(\widehat{\theta}_n - \theta_o) \\ &\quad + \left[\Gamma_S(\widetilde{\theta}_n) - \Gamma_S(\theta_o)\right](\widehat{\theta}_n - \theta_o) \end{aligned} \tag{A.14}$$

where $\Gamma_S(\widetilde{\theta}_n)' = \left[\Gamma_1(\widetilde{\theta}_{1,n})', \ldots, \Gamma_{k_0}(\widetilde{\theta}_{k_0,n})'\right]$ and $\widetilde{\theta}_{\ell,n}$ is some value between $\widehat{\theta}_n$ and θ_o for any $\ell \in S$. By the Cauchy–Schwarz inequality, the consistency of $\widehat{\theta}_n$ and Assumption 3.2(iv),

$$\begin{aligned} &\left\|\left[\Gamma_S(\widetilde{\theta}_n) - \Gamma_S(\theta_o)\right](\widehat{\theta}_n - \theta_o)\right\| \\ &\leq \left\|\Gamma_S(\widetilde{\theta}_n) - \Gamma_S(\theta_o)\right\| \left\|\widehat{\theta}_n - \theta_o\right\| \lesssim \left\|\widehat{\theta}_n - \theta_o\right\|^2 \end{aligned} \tag{A.15}$$

w.p.a.1. Using Assumption 3.2(iii), we have

$$\left\|\Gamma_S(\theta_o)(\widehat{\theta}_n - \theta_o)\right\| \lesssim \left\|\widehat{\theta}_n - \theta_o\right\|, \tag{A.16}$$

which together with (A.14), (A.15), the consistency of $\widehat{\theta}_n$ and the Cauchy–Schwarz inequality implies that

$$\begin{aligned} \left\|\overline{m}_S(\widehat{\theta}_n) - \overline{m}_S(\theta_o)\right\|^2 &= (\widehat{\theta}_n - \theta_o)'\Gamma_S(\theta_o)'\Gamma_S(\theta_o)(\widehat{\theta}_n - \theta_o) \\ &\quad + o_p(1)\left\|\widehat{\theta}_n - \theta_o\right\|^2. \end{aligned} \tag{A.17}$$

The above equality combined with Assumption 3.2(iii) further implies that

$$\left\|\widehat{\theta}_n-\theta_o\right\| \lesssim \left\|\overline{m}_S(\widehat{\theta}_n)-\overline{m}_S(\theta_o)\right\| \lesssim \left\|\widehat{\theta}_n-\theta_o\right\| \tag{A.18}$$

w.p.a.1. Combining results in (A.13) and (A.18), we have w.p.a.1,

$$\left\|\overline{m}_{S,n}(\widehat{\theta}_n)\right\|^2 \gtrsim \left\|\widehat{\theta}_n-\theta_o\right\|^2 - O_p(\tau_n)\left\|\widehat{\theta}_n-\theta_o\right\|. \tag{A.19}$$

To study the second term on the left hand side of (A.12), we can write

$$\begin{aligned}\overline{m}_{D,n}(\widehat{\theta}_n)-\widehat{\beta}_n &= \left[\overline{m}_{D,n}(\widehat{\theta}_n)-\overline{m}_D(\widehat{\theta}_n)\right]+\left[\overline{m}_D(\widehat{\theta}_n)-\widehat{\beta}_n\right]\\ &= O_p(\tau_n)+\overline{m}_D(\widehat{\theta}_n)-\widehat{\beta}_n\\ &= Q_n-\left[\widehat{\beta}_n-\beta_o\right], \quad \text{where}\end{aligned}$$

$$Q_n \equiv \left[\overline{m}_D(\widehat{\theta}_n)-\beta_o\right]+O_p(\tau_n) \tag{A.20}$$

following the consistency of $\widehat{\theta}_n$ and Assumption 3.2(i). Then,

$$\begin{aligned}b_n\left\|\widehat{\beta}_n-\beta_o\right\|+O_p(\tau_n^2) &\geq \left\|\overline{m}_{D,n}(\widehat{\theta}_n)-\widehat{\beta}_n\right\|^2\\ &\geq \left\|\widehat{\beta}_n-\beta_o\right\|^2+\|Q_n\|^2\\ &\quad -2\|Q_n\|\left\|\widehat{\beta}_n-\beta_o\right\|\end{aligned} \tag{A.21}$$

w.p.a.1, where the first inequality follows from (A.12) and the second inequality follows from (A.20) and the Cauchy–Schwarz inequality. Reorganizing (A.21), we obtain

$$\begin{aligned}&\left\|\widehat{\beta}_n-\beta_o\right\|^2-(2\|Q_n\|+b_n)\left\|\widehat{\beta}_n-\beta_o\right\|\\ &+\|Q_n\|^2-O_p(\tau_n^2)\leq 0\end{aligned} \tag{A.22}$$

which implies

$$\left\|\widehat{\beta}_n-\beta_o\right\| \lesssim \left\|\overline{m}_D(\widehat{\theta}_n)-\overline{m}_D(\theta_o)\right\|+b_n+O_p(\tau_n) \tag{A.23}$$

using the definition of Q_n in (A.21) and $\beta_o=\overline{m}_D(\theta_o)$.

Combining the inequalities in (A.12), (A.19) and (A.23), we get

$$\begin{aligned}&\left\|\widehat{\theta}_n-\theta_o\right\|^2-O_p(\tau_n)\left\|\widehat{\theta}_n-\theta_o\right\|\\ &\lesssim b_n\left\|\overline{m}_D(\widehat{\theta}_n)-\overline{m}_D(\theta_o)\right\|+O_p(\tau_n^2+b_n^2)\end{aligned} \tag{A.24}$$

w.p.a.1. By the mean value theorem,

$$\begin{aligned}\overline{m}_D(\widehat{\theta}_n)-\overline{m}_D(\theta_o) &= \left[\Gamma_D(\widetilde{\theta}_n)-\Gamma_D(\theta_o)\right](\widehat{\theta}_n-\theta_o)\\ &\quad +\Gamma_D(\theta_o)(\widehat{\theta}_n-\theta_o)\end{aligned} \tag{A.25}$$

where $\Gamma_D(\widetilde{\theta}_n)'=\left[\Gamma_{k_0+1}(\widetilde{\theta}_{k_0+1,n})',\ldots,\Gamma_{k_n}(\widetilde{\theta}_{k_n,n})'\right]$ and $\widetilde{\theta}_{\ell,n}$ is some value between $\widehat{\theta}_n$ and θ_o for any $\ell \in D$. Note that Assumption 3.2(iv) implies that

$$\left\|\left[\Gamma_D(\widetilde{\theta}_n)-\Gamma_D(\theta_o)\right](\widehat{\theta}_n-\theta_o)\right\| \lesssim \sqrt{k_n}\left\|\widehat{\theta}_n-\theta_o\right\|^2 \tag{A.26}$$

w.p.a.1. Under Assumption 3.2(iii), we have $\left\|\Gamma_D(\theta_o)(\widehat{\theta}_n-\theta_o)\right\|^2 \lesssim \|\widehat{\theta}_n-\theta_o\|^2$. Therefore,

$$\left\|\overline{m}_D(\widehat{\theta}_n)-\overline{m}_D(\theta_o)\right\| \lesssim \sqrt{k_n}\left\|\widehat{\theta}_n-\theta_o\right\|^2+\left\|\widehat{\theta}_n-\theta_o\right\| \tag{A.27}$$

w.p.a.1 by the triangle inequality and the Cauchy–Schwarz inequality. Combining (A.24) and (A.27) yields

$$\begin{aligned}\left[1-O_p(\sqrt{k_n}b_n)\right]\left\|\widehat{\theta}_n-\theta_o\right\|^2 &\leq O_p(b_n+\tau_n)\left\|\widehat{\theta}_n-\theta_o\right\|\\ &\quad +O_p(\tau_n^2+b_n^2)\end{aligned} \tag{A.28}$$

w.p.a.1. As $\sqrt{k_n}b_n=o_p(1)$, the inequality above implies

$$\left\|\widehat{\theta}_n-\theta_o\right\| = O_p(b_n+\tau_n). \tag{A.29}$$

Applying the results in (A.27) and (A.29) to (A.23), we obtain

$$\begin{aligned}\left\|\widehat{\beta}_n-\beta_o\right\| &\lesssim O_p(b_n+\tau_n)+\sqrt{k_n}O_p(b_n^2+\tau_n^2)\\ &= O_p(b_n+\tau_n),\end{aligned} \tag{A.30}$$

where the last equality follows from $\sqrt{k_n}(b_n+\tau_n)=o_p(1)$ under Assumption 3.2(v). Combining the results in (A.29) and (A.30), we get the result in part (a).

We next prove part (b). We first note that

$$\begin{aligned}&\overline{g}_n(\widehat{\alpha}_n)'W_n\overline{g}_n(\widehat{\alpha}_n)-\overline{g}_n(\alpha_o)'W_n\overline{g}_n(\alpha_o)\\ &=\left[\overline{g}_n(\widehat{\alpha}_n)-\overline{g}_n(\alpha_o)\right]'W_n\left[\overline{g}_n(\widehat{\alpha}_n)-\overline{g}_n(\alpha_o)\right]\\ &\quad+2\left[\overline{g}_n(\widehat{\alpha}_n)-\overline{g}_n(\alpha_o)\right]'W_n\overline{g}_n(\alpha_o)\\ &\gtrsim \|\overline{g}_n(\widehat{\alpha}_n)-\overline{g}_n(\alpha_o)\|^2-\|\overline{g}_n(\widehat{\alpha}_n)-\overline{g}_n(\alpha_o)\|\,\|\overline{g}_n(\alpha_o)\|\\ &\gtrsim \|\widehat{\alpha}_n-\alpha_o\|^2-\|\overline{g}_n(\alpha_o)\|\,\|\widehat{\alpha}_n-\alpha_o\|\end{aligned} \tag{A.31}$$

w.p.a.1, where the first inequality follows from Assumption 3.1(iii) and the Cauchy–Schwarz inequality and the second inequality holds by Assumption 3.3(ii). Combining the inequalities in (A.11) and (A.31), we obtain

$$\|\widehat{\alpha}_n-\alpha_o\|^2-\|\overline{g}_n(\alpha_o)\|\,\|\widehat{\alpha}_n-\alpha_o\| \lesssim b_n\|\widehat{\alpha}_n-\alpha_o\|, \tag{A.32}$$

which together with Assumption 3.3(i), implies $\|\widehat{\alpha}_n-\alpha_o\| = O_p(\tau_n+b_n)$. ■

Proof of Theorem 3.2. Let e_ℓ be a k_n-dimensional vector with the ℓth entry being 1 and others being 0. By the Karush–Kuhn–Tucker (KKT) optimality condition, $\widehat{\beta}_{n,\ell}=0$ if

$$\left|e_\ell' W_n\overline{g}_n(\widehat{\alpha}_n)\right| < \left|\frac{\lambda_n\omega_{n,\ell}}{2}\right|. \tag{A.33}$$

Hence,

$$\Pr\left(\widehat{\beta}_{n,\ell}=0,\forall\ell\in A\right) \geq \Pr\left(\max_{\ell\in A}\left|\frac{e_\ell' W_n\overline{g}_n(\widehat{\alpha}_n)}{\lambda_n\omega_{n,\ell}}\right|<\frac{1}{2}\right). \tag{A.34}$$

To obtain the desired result, it remains to show

$$\max_{\ell\in A}\left|\frac{e_\ell' W_n\overline{g}_n(\widehat{\alpha}_n)}{\lambda_n\omega_{n,\ell}}\right| = o_p(1). \tag{A.35}$$

Following Assumption 3.1(iii),

$$0<C^{-1}\leq e_\ell' W_n W_n e_\ell \leq C<\infty \tag{A.36}$$

for any ℓ w.p.a.1. By the Cauchy–Schwarz inequality and the inequalities in (A.36),

$$\begin{aligned}\max_{\ell\in A}\left|\frac{e_\ell' W_n\overline{g}_n(\widehat{\alpha}_n)}{\lambda_n\omega_{n,\ell}}\right| &\leq \max_{\ell\in A}\frac{\left\|e_\ell' W_n\right\|}{\lambda_n\omega_{n,\ell}}\|\overline{g}_n(\widehat{\alpha}_n)\|\\ &\lesssim \frac{\|\overline{g}_n(\widehat{\alpha}_n)\|}{\lambda_n}\max_{\ell\in A}\omega_{n,\ell}^{-1}\end{aligned} \tag{A.37}$$

w.p.a.1. By the triangle inequality,

$$\|\overline{g}_n(\widehat{\alpha}_n)\| \leq \|\overline{g}(\widehat{\alpha}_n)\|+\left\|\upsilon_n(\widehat{\theta}_n)\right\| = \|\overline{g}(\widehat{\alpha}_n)\|+O_p(\tau_n), \tag{A.38}$$

where the equality follows from Assumption 3.2(i). Note that

$$\begin{aligned}\|\overline{g}(\widehat{\alpha}_n)\|^2 &= \left\|\overline{m}_S(\widehat{\theta}_n)-\overline{m}_S(\theta_o)\right\|^2+\left\|\overline{m}_D(\widehat{\theta}_n)-\widehat{\beta}_n\right\|^2\\ &\lesssim \left\|\overline{m}_S(\widehat{\theta}_n)-\overline{m}_S(\theta_o)\right\|^2+\left\|\overline{m}_D(\widehat{\theta}_n)-\overline{m}_D(\theta_o)\right\|^2\\ &\quad+\left\|\widehat{\beta}_n-\beta_o\right\|^2\end{aligned} \tag{A.39}$$

which together with (A.18), (A.27), Lemma 3.2, $\sqrt{k_n}(\tau_n+b_n)=o(1)$ and $b_n = O_p(\tau_n)$ implies that $\|\overline{g}_n(\widehat{\alpha}_n)\| = O_p(\tau_n)$. This

combined with Assumption 3.4(ii) and (A.37) implies that (A.35) holds.

Next, we prove part (b). Under Assumption 3.3, we have

$$\begin{aligned}\|\bar{g}_n(\widehat{\alpha}_n)\| &\leq \|\bar{g}_n(\widehat{\alpha}_n) - \bar{g}_n(\alpha_o)\| + \|\bar{g}_n(\alpha_o)\| \\ &\lesssim \|\widehat{\alpha}_n - \alpha_o\| + \|\bar{g}_n(\alpha_o)\| = O_p(\tau_n), \end{aligned} \tag{A.40}$$

where the first inequality follows from the triangle inequality, the second inequality is by Assumption 3.3(ii) and it holds w.p.a.1, the last equality is by Assumptions 3.3(i), 3.4(i), and Lemma 3.2 (b). This combined with Assumption 3.4(ii) and (A.37) implies that (A.35) holds. ■

Proof of Theorem 3.3. Let ε_n be a sequence of constants such that (i) $\varepsilon_n = o(n^{-\frac{1}{2}})$, (ii) $\lambda_n \|\omega_{n,B}\| = O_p(\varepsilon_n)$, (iii) $\varsigma_n\tau_n = o(\varepsilon_n)$ (and (iv) $\sqrt{k_n}\tau_n^2 = O(\varepsilon_n)$ for the proof of part (a)). Such a sequence can be constructed because $\lambda_n \|\omega_{n,B}\| = o_p(n^{-1/2})$, $\varsigma_n\tau_n = o(n^{-1/2})$, (and $\sqrt{k_n}\tau_n^2 = o(n^{-1/2})$ for part (a)). Define

$$\widehat{\alpha}^*_{B,n} = \widehat{\alpha}_{B,n} + \varepsilon_n u^*_n, \tag{A.41}$$

where $u^*_n = (\Gamma'_\alpha W_n \Gamma_\alpha)^{-1}\gamma^*_n$, $\gamma^*_n \in R^{d_\theta + d_B}$ and $\|\gamma^*_n\| \leq C$. Because the smallest eigenvalues of W_n and $\Gamma'_\alpha\Gamma_\alpha$ are bounded from below by Assumptions 3.1(iii) and 3.6(iii), $\|u^*_n\| \leq C$. Hence,

$$\|\varepsilon_n u^*_n\|^2 = \varepsilon_n^2 \|u^*_n\|^2 = O_p(\varepsilon_n^2) = o_p(n^{-1}). \tag{A.42}$$

Write $\widehat{\alpha}^{*\prime}_{B,n} = (\widehat{\theta}^{*\prime}_n, \widehat{\beta}^{*\prime}_{B,n})$, then

$$\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\| = \|\varepsilon_n u^*_n\| = O_p(\varepsilon_n) \quad \text{and} \tag{A.43}$$

$$\|\widehat{\alpha}^*_{B,n} - \alpha_{B,o}\| \leq \|\widehat{\alpha}_{B,n} - \alpha_{B,o}\| + \|\varepsilon_n u^*_n\| = O_p(\tau_n). \tag{A.44}$$

For any $\alpha'_B = (\theta', \beta'_B)$, we define

$$\bar{g}_n(\alpha_B) = \frac{1}{n}\sum_{i=1}^n g(Z_i, \alpha_B) \quad \text{and} \quad \bar{g}(\alpha_B) = \mathbb{E}[g(Z, \alpha_B)], \tag{A.45}$$

where $g(Z, \alpha_B)$ is defined in (3.7). By the definition of $\widehat{\alpha}_n$,

$$\begin{aligned}&\bar{g}_n(\widehat{\alpha}_n)' W_n \bar{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell\in D} \omega_{n,\ell} |\widehat{\beta}_{n,\ell}| \\ &\leq \bar{g}_n(\widehat{\alpha}^*_{B,n})' W_n \bar{g}_n(\widehat{\alpha}^*_{B,n}) + \lambda_n \sum_{\ell\in B} \omega_{n,\ell} |\widehat{\beta}^*_{n,\ell}| \end{aligned} \tag{A.46}$$

where $\widehat{\beta}^*_{n,\ell}$ is the element of $\widehat{\alpha}^*_{B,n}$ corresponding to $\widehat{\beta}_{n,\ell}$ for any $\ell \in B$. By Theorem 3.2, the left hand side of (A.46) satisfies

$$\begin{aligned}&\bar{g}_n(\widehat{\alpha}_n)' W_n \bar{g}_n(\widehat{\alpha}_n) + \lambda_n \sum_{\ell\in D} \omega_{n,\ell} |\widehat{\beta}_{n,\ell}| \\ &= \bar{g}_n(\widehat{\alpha}_{B,n})' W_n \bar{g}_n(\widehat{\alpha}_{B,n}) + \lambda_n \sum_{\ell\in B} \omega_{n,\ell} |\widehat{\beta}_{n,\ell}| \end{aligned} \tag{A.47}$$

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

$$\begin{aligned}&\left|\lambda_n \sum_{\ell\in B} \omega_{n,\ell} \left(|\widehat{\beta}^*_{n,\ell}| - |\widehat{\beta}_{n,\ell}|\right)\right| \leq \lambda_n \sum_{\ell\in B} \omega_{n,\ell} |\widehat{\beta}^*_{n,\ell} - \widehat{\beta}_{n,\ell}| \\ &= \lambda_n \varepsilon_n \sum_{\ell\in B} \omega_{n,\ell} |u^*_{n,\ell}| \leq \lambda_n \varepsilon_n \|\omega_{n,B}\| \|u^*_{n,B}\| = O_p(\varepsilon_n^2), \end{aligned} \tag{A.48}$$

where $u^*_{n,B} \equiv (u^*_{n,d_\theta+1}, \ldots, u^*_{n,d_\theta+d_B})'$ is the vector of perturbation on β_B and the $O_p(\varepsilon_n^2)$ follows from $\|u^*_{n,B}\| \leq C$ and $\lambda_n \|\omega_{n,B}\| = O_p(\varepsilon_n)$. Combining (A.46)–(A.48) yields

$$\bar{g}_n(\widehat{\alpha}^*_{B,n})' W_n \bar{g}_n(\widehat{\alpha}^*_{B,n}) - \bar{g}_n(\widehat{\alpha}_{B,n})' W_n \bar{g}_n(\widehat{\alpha}_{B,n}) \geq O_p(\varepsilon_n^2). \tag{A.49}$$

We next prove part (a). Define

$$I_{1,n} = \upsilon_n(\widehat{\theta}^*_n) - \upsilon_n(\widehat{\theta}_n). \tag{A.50}$$

Because $g(Z, \alpha_B)$ is linear in β,

$$\bar{g}_n(\alpha_B) = \bar{g}(\alpha_B) + \upsilon_n(\theta) \tag{A.51}$$

for any $\alpha'_B = (\theta', \beta'_B)$. Applying the equality above, we obtain

$$\bar{g}_n(\widehat{\alpha}^*_{B,n}) - \bar{g}_n(\widehat{\alpha}_{B,n}) = \bar{g}(\widehat{\alpha}^*_{B,n}) - \bar{g}(\widehat{\alpha}_{B,n}) + I_{1,n}, \tag{A.52}$$

which implies that

$$\begin{aligned}&\left\|\bar{g}_n(\widehat{\alpha}^*_{B,n}) - \bar{g}_n(\widehat{\alpha}_{B,n})\right\|^2 \\ &\leq 2\left\|\bar{g}(\widehat{\alpha}^*_{B,n}) - \bar{g}(\widehat{\alpha}_{B,n})\right\|^2 + 2\left\|I_{1,n}\right\|^2. \end{aligned} \tag{A.53}$$

By the mean value theorem,

$$\begin{aligned}&\left\|\bar{g}(\widehat{\alpha}^*_{B,n}) - \bar{g}(\widehat{\alpha}_{B,n})\right\|^2 \\ &= \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \Gamma_\alpha(\widetilde{\theta}^*_n)' \Gamma_\alpha(\widetilde{\theta}^*_n) \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right) \\ &= \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \Gamma_\alpha(\theta_o)' \Gamma_\alpha(\theta_o) \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right) \\ &\quad + \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right]' \\ &\quad \times \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right] \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right) \\ &\quad + 2\left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right]' \\ &\quad \times \Gamma_\alpha(\theta_o) \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right) \end{aligned} \tag{A.54}$$

where the first equality is a mean value expansion with $\widetilde{\theta}^*_n$ between the first d_θ elements of $\widehat{\alpha}^*_{B,n}$ and $\widehat{\alpha}_{B,n}$, ($\widetilde{\theta}^*_n$ may vary element by element for the vector). Next, we show each term on the right hand side of the second equality of (A.54) is $O_p(\varepsilon_n^2)$. Under Assumption 3.2(iii), we see that

$$\begin{aligned}\gamma'\Gamma_\alpha\Gamma'_\alpha\gamma &= (\gamma'_{d_\theta}, \mathbf{0}'_{d_A}, \gamma'_{d_B}) \left[\Gamma(\theta_o)'\Gamma(\theta_o)\right] (\gamma'_{d_\theta}, \mathbf{0}'_{d_A}, \gamma'_{d_B})' \\ &\leq C\|\gamma\|^2 \end{aligned} \tag{A.55}$$

for any $\gamma = (\gamma'_{d_\theta}, \gamma'_{d_B})' \in R^{d_\theta + d_B}$, which implies that $\rho_{\max}(\Gamma'_\alpha\Gamma_\alpha) \leq C$. Under (A.55) and $\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\| = O_p(\varepsilon_n)$, we have

$$\begin{aligned}\left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \Gamma_\alpha(\theta_o)' \Gamma_\alpha(\theta_o) \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right) &\lesssim \left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\|^2 \\ &= O_p(\varepsilon_n^2). \end{aligned} \tag{A.56}$$

By the Cauchy–Schwarz inequality and $k_n\tau_n^2 = o(1)$, we have

$$\begin{aligned}&\left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right]' \\ &\quad \times \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right] \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right) \\ &\leq \left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\|^2 \left\|\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right\|^2 \\ &= k_n O_p(\tau_n^2) \left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\|^2 = o_p(\varepsilon_n^2) \end{aligned} \tag{A.57}$$

where the first equality is by $\|\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\| = O_p(\sqrt{k_n}\tau_n)$, which in turn holds by Assumption 3.2(iv) and $\|\widetilde{\theta}^*_n - \theta_o\| = O_p(\tau_n)$. Using (A.55), the Cauchy–Schwarz inequality, Assumption 3.2(iv), and $k_n\tau_n^2 = o(1)$, we have

$$\begin{aligned}&\left|\left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)' \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right]' \Gamma_\alpha(\theta_o) \left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)\right| \\ &\leq \left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\| \left\|\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right\| \left\|\Gamma_\alpha(\theta_o)(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})\right\| \\ &\lesssim \left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\|^2 \left\|\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right\| \\ &= \sqrt{k_n} O_p(\tau_n) \left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\|^2 = o_p(\varepsilon_n^2). \end{aligned} \tag{A.58}$$

Combining the results in (A.54) and (A.56)–(A.58), we deduce that

$$\left\|\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})\right\|^2 = O_p(\varepsilon_n^2). \tag{A.59}$$

Using Assumption 3.5, $\|\widehat{\theta}^*_n - \widehat{\theta}_n\| = O_p(\varepsilon_n)$, $\tau_n^{-1} = O(n^{\frac{1}{2}})$, and $\varsigma_n\tau_n = o(\varepsilon_n)$, we have

$$\left\|I_{1,n}\right\| = O_p(n^{-\frac{1}{2}}\varsigma_n + \varepsilon_n\varsigma_n) = O_p(n^{-\frac{1}{2}}\varsigma_n) = o_p(\varepsilon_n), \tag{A.60}$$

which together with (A.52), (A.59), and Assumption 3.1(iii) implies that

$$\begin{aligned}&\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]' W_n \left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]\\ &\lesssim \left\|\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right\|^2 = O_p(\varepsilon_n^2)\end{aligned} \tag{A.61}$$

where the inequality holds w.p.a.1. We can rewrite the inequality (A.49) to obtain

$$\begin{aligned}O_p(\varepsilon_n^2) &\le \overline{g}_n(\widehat{\alpha}^*_{B,n})'W_n\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})'W_n\overline{g}_n(\widehat{\alpha}_{B,n})\\ &= \left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]' W_n \left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]\\ &\quad + 2\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]' W_n\overline{g}_n(\widehat{\alpha}_{B,n}).\end{aligned} \tag{A.62}$$

Applying the results in (A.61) and (A.62), we obtain

$$\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]' W_n\overline{g}_n(\widehat{\alpha}_{B,n}) \ge O_p(\varepsilon_n^2). \tag{A.63}$$

Define

$$I_{0,n} = \upsilon_n(\widehat{\theta}_n) - \upsilon_n(\theta_o). \tag{A.64}$$

Then

$$\overline{g}_n(\widehat{\alpha}_{B,n}) = \overline{g}_n(\alpha_o) + \overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o) + I_{0,n}. \tag{A.65}$$

Plugging (A.65) into (A.63) and using the definition of $I_{1,n}$ in (A.50) yields

$$\begin{aligned}O_p(\varepsilon_n^2) &\le \left[\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})\right]' W_n \left[\overline{g}_n(\alpha_o) + \overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)\right]\\ &\quad + A_n + B_n + C_n, \quad \text{where}\\ A_n &= I'_{1,n}W_n[\overline{g}_n(\alpha_o) + \overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)],\\ B_n &= [\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})]'W_nI_{0,n},\\ C_n &= I'_{1,n}W_nI_{0,n}.\end{aligned} \tag{A.66}$$

We next study the terms A_n, B_n and C_n one by one. By the triangle inequality and the Cauchy–Schwarz inequality,

$$\begin{aligned}|A_n| &\le \left|I'_{1,n}W_n\overline{g}_n(\alpha_o)\right| + \left|I'_{1,n}W_n[\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)]\right|\\ &\le \left|I'_{1,n}W_nI_{1,n}\right|^{\frac{1}{2}} \left|\overline{g}_n(\alpha_o)'W_n\overline{g}_n(\alpha_o)\right|^{\frac{1}{2}}\\ &\quad + \left|I'_{1,n}W_nI_{1,n}\right|^{\frac{1}{2}} \left|[\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)]'\right.\\ &\quad \left.\times\ W_n[\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)]\right|^{\frac{1}{2}}.\end{aligned} \tag{A.67}$$

By arguments similar to those used to show (A.59), we can show

$$\left\|\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)\right\|^2 = O_p(\tau_n^2). \tag{A.68}$$

Using Assumptions 3.1(iii) and 3.2(i), the results in (A.60), (A.67), (A.68), and $\varsigma_n\tau_n = o(\varepsilon_n)$, we have

$$|A_n| = O_p(n^{-\frac{1}{2}}\varsigma_n\tau_n) = o_p(n^{-\frac{1}{2}}\varepsilon_n). \tag{A.69}$$

To study B_n and C_n, first note that

$$\left\|I_{0,n}\right\| = O_p(n^{-\frac{1}{2}}\varsigma_n + \tau_n\varsigma_n) = o_p(\varepsilon_n), \tag{A.70}$$

where the first equality follows from Assumption 3.5 and $\|\widehat{\theta}_n - \theta_o\| = O_p(\tau_n)$ and the second equality follows from $\tau_n^{-1} = O(n^{\frac{1}{2}})$ and $\varsigma_n\tau_n = o(\varepsilon_n)$. By the Cauchy–Schwarz inequality, (A.59) and (A.70),

$$\begin{aligned}|B_n| &\le \left|\left[\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})\right]'W_n\left[\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})\right]\right|^{\frac{1}{2}} \left|I'_{0,n}W_nI_{0,n}\right|^{\frac{1}{2}}\\ &= o_p(\varepsilon_n^2).\end{aligned} \tag{A.71}$$

By the Cauchy–Schwarz inequality, (A.60) and (A.70),

$$|C_n| \le \left|I'_{1,n}W_nI_{1,n}\right|^{\frac{1}{2}} \left|I'_{0,n}W_nI_{0,n}\right|^{\frac{1}{2}} = o_p(\varepsilon_n^2). \tag{A.72}$$

Putting together (A.66), (A.69), (A.71), and (A.72) and using $\varepsilon_n = o(n^{-1/2})$, we obtain

$$\begin{aligned}&\left[\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})\right]' W_n \left[\overline{g}_n(\alpha_o) + \overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)\right]\\ &\ge o_p(n^{-\frac{1}{2}}\varepsilon_n)\end{aligned} \tag{A.73}$$

Next, we consider the mean-value expansions for $\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n})$ and $\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)$ in (A.73). First,

$$\begin{aligned}\overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n}) &= \Gamma_\alpha(\widetilde{\theta}^*_n)\left(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right)\\ &= \Gamma_\alpha(\theta_o)(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}) + \left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right](\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})\\ &= \Gamma_\alpha(\theta_o)(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}) + O_p(\sqrt{k_n}\tau_n\varepsilon_n),\end{aligned} \tag{A.74}$$

where the first equality is a mean value expansion with $\widetilde{\theta}^*_n$ between the first d_θ elements of $\widehat{\alpha}^*_{B,n}$ and $\widehat{\alpha}_{B,n}$, ($\widetilde{\theta}^*_n$ may vary element by element for the vector), the second equality is obvious, and the third equality follows from the Cauchy–Schwarz inequality, $\|\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\| = O_p(\sqrt{k_n}\tau_n)$, which in turn holds by Assumption 3.2(iv) and $\|\widetilde{\theta}^*_n - \theta_o\| = O_p(\tau_n)$, and $\left\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\right\| = O_p(\varepsilon_n)$. Similarly, we have

$$\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o) = \Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o}) + O_p(\sqrt{k_n}\tau_n^2). \tag{A.75}$$

Plugging (A.74) and (A.75) into (A.73), we obtain

$$\begin{aligned}o_p(n^{-1/2}\varepsilon_n) &\le (\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma_\alpha(\theta_o)'\\ &\quad \times W_n[\overline{g}_n(\alpha_o) + \Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})] + D_1 + D_2\\ &= (\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma_\alpha(\theta_o)'\\ &\quad \times W_n[\overline{g}_n(\alpha_o) + \Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})]\\ &\quad + o_p(n^{-\frac{1}{2}}\varepsilon_n),\end{aligned} \tag{A.76}$$

because

$$\begin{aligned}|D_1| &\equiv \left|(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\left[\Gamma_\alpha(\widetilde{\theta}^*_n) - \Gamma_\alpha(\theta_o)\right]'\right.\\ &\quad \left.\times\ W_n\left[\overline{g}_n(\alpha_o) + \overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)\right]\right|\\ &\le O_p(\sqrt{k_n}\tau_n\varepsilon_n)\left[\|\overline{g}_n(\alpha_o)\| + \left\|\overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_o)\right\|\right]\\ &= O_p(\sqrt{k_n}\tau_n^2\varepsilon_n) = o_p(n^{-\frac{1}{2}}\varepsilon_n),\end{aligned} \tag{A.77}$$

and

$$\begin{aligned}|D_2| &\equiv \left|(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma_\alpha(\theta_o)'\right.\\ &\quad \left.\times\ W_n\left[\Gamma_\alpha(\widetilde{\theta}_n) - \Gamma_\alpha(\theta_o)\right](\widehat{\alpha}_{B,n} - \alpha_{B,o})\right|\\ &= O_p(\varepsilon_n)O_p(\sqrt{k_n}\tau_n^2) = o_p(n^{-\frac{1}{2}}\varepsilon_n)\end{aligned} \tag{A.78}$$

for some $\widetilde{\theta}_n$ between the first d_θ elements of $\widehat{\alpha}^*_{B,n}$ and α_0 and $\widetilde{\theta}_n$ may vary element by element for the vector. The first inequality in (A.77) holds by the Cauchy–Schwarz inequality, the third equality in (A.74), and Assumption 3.1(iii), the first equality in (A.77) follows from Assumption 3.2(i) and (A.68), and the second equality in (A.77) holds because $\sqrt{k_n}\tau_n^2 = O(\varepsilon_n)$ and $\varepsilon_n = o(n^{-1/2})$, as stated

at the beginning of the proof. Results in (A.78) hold by similar arguments using $\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\| = O_p(\varepsilon_n)$, $\|\widehat{\alpha}_{B,n} - \alpha_{B,o}\| = O_p(\tau_n)$ and $\|\Gamma_\alpha(\widetilde{\theta}_n) - \Gamma_\alpha(\theta_o)\| = O_p(\sqrt{k_n}\tau_n)$.

The results in (A.76) immediately gives

$$o_p(\varepsilon_n n^{-\frac{1}{2}}) \leq (\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma_\alpha(\theta_o)' \times W_n[\overline{g}_n(\alpha_o) + \Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})] \quad (A.79)$$

which together with $\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n} = \varepsilon_n u^*_n = \varepsilon_n(\Gamma'_\alpha W_n \Gamma_\alpha)^{-1}\gamma^*_n$ implies that

$$\gamma^{*\prime}_n(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\Gamma'_\alpha W_n[\sqrt{n}\overline{g}_n(\alpha_o) + \sqrt{n}\Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})] \geq o_p(1) \quad (A.80)$$

where $\gamma^*_n \in R^{d_\theta + d_B}$ with $\|\gamma^*_n\| \leq C$. Next, define $\widehat{\alpha}^*_{B,n} = \widehat{\alpha}_{B,n} - \varepsilon_n u^*_n$ and using the same arguments in deriving (A.80), we deduce that

$$\gamma^{*\prime}_n(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\Gamma'_\alpha W_n[\sqrt{n}\overline{g}_n(\alpha_o) + \sqrt{n}\Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})] \leq o_p(1), \quad (A.81)$$

which combined with (A.80), implies that

$$\left|\gamma^{*\prime}_n(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\Gamma'_\alpha W_n[\sqrt{n}\overline{g}_n(\alpha_o) + \sqrt{n}\Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})]\right| = o_p(1). \quad (A.82)$$

The approximation in (A.82) can be rewritten as

$$\sqrt{n}\gamma^{*\prime}_n(\widehat{\alpha}_{B,n} - \alpha_{B,o}) = -\gamma^{*\prime}_n(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\Gamma'_\alpha W_n\Omega_n^{\frac{1}{2}}\left[\sqrt{n}\Omega_n^{-\frac{1}{2}}\overline{g}_n(\alpha_o)\right] + o_p(1). \quad (A.83)$$

Let $\gamma_n \in R^{d_\theta + d_B}$ be an arbitrary vector with $\|\gamma_n\| = 1$. Let

$$\gamma^{*\prime}_n \equiv \gamma'_n\Sigma_n^{-\frac{1}{2}} \quad \text{and} \quad \overline{\gamma}'_n \equiv \gamma'_n\Sigma_n^{-\frac{1}{2}}(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\Gamma'_\alpha W_n\Omega_n^{\frac{1}{2}}. \quad (A.84)$$

It is clear that

$$\|\overline{\gamma}_n\|^2 = \overline{\gamma}'_n\overline{\gamma}_n = \gamma'_n\Sigma_n^{-\frac{1}{2}}\Sigma_n\Sigma_n^{-\frac{1}{2}}\gamma_n = 1 \quad (A.85)$$

and

$$\|\gamma^*_n\|^2 = \gamma^{*\prime}_n\gamma^*_n = \gamma'_n\Sigma_n^{-1}\gamma_n \leq C \quad (A.86)$$

where the last inequality is by $\|\gamma_n\| = 1$, (A.55), and Assumptions 3.1(iii), 3.6(ii), and 3.6(iii). Hence, we deduce that

$$\sqrt{n}\gamma'_n\Sigma_n^{-\frac{1}{2}}(\widehat{\alpha}_{B,n} - \alpha_{B,o}) = -\overline{\gamma}'_n\left[\sqrt{n}\Omega_n^{-\frac{1}{2}}\overline{g}_n(\alpha_o)\right] + o_p(1) \to_d N(0,1), \quad (A.87)$$

where the weak convergence is by Assumption 3.6(i).

We next prove part (b). Using Assumption 3.3(ii) and (A.43), we get

$$\|\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\| = O_p\left(\|\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}\|\right) = O_p(\varepsilon_n) \quad (A.88)$$

which together with the expression in (A.62) implies that

$$\overline{g}_n(\widehat{\alpha}^*_{B,n})'W_n\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})'W_n\overline{g}_n(\widehat{\alpha}_{B,n}) = 2\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]'W_n\overline{g}_n(\widehat{\alpha}_{B,n}) + O_p(\varepsilon_n^2). \quad (A.89)$$

By Assumption 3.8,

$$\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n}) = \overline{g}(\widehat{\alpha}^*_{B,n}) - \overline{g}(\widehat{\alpha}_{B,n}) + I_{1,n} = \Gamma_\alpha(\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n}) + I_{1,n} \quad (A.90)$$

and similarly,

$$\overline{g}_n(\widehat{\alpha}_{B,n}) - \overline{g}_n(\alpha_o) = \overline{g}(\widehat{\alpha}_{B,n}) - \overline{g}(\alpha_{B,o}) + I_{0,n} = \Gamma_\alpha(\widehat{\alpha}_{B,n} - \alpha_{B,o}) + I_{0,n}. \quad (A.91)$$

By Assumption 3.5, $\|\widehat{\theta}^*_n - \widehat{\theta}_n\| = O_p(\varepsilon_n)$, $\|\widehat{\theta}_n - \theta_o\| = O_p(\tau_n)$, and $\varsigma_n\tau_n = o(\varepsilon_n)$, we have

$$\|I_{1,n}\| = O_p(n^{-\frac{1}{2}}\varsigma_n) \quad \text{and} \quad \|I_{0,n}\| = O_p(\varsigma_n\tau_n) = o_p(\varepsilon_n). \quad (A.92)$$

Applying (A.90)–(A.92), we have

$$\begin{aligned}&\left[\overline{g}_n(\widehat{\alpha}^*_{B,n}) - \overline{g}_n(\widehat{\alpha}_{B,n})\right]'W_n\overline{g}_n(\widehat{\alpha}_{B,n})\\ &= (\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma'_\alpha W_n[\overline{g}_n(\alpha_o) + \Gamma_\alpha(\widehat{\alpha}_{B,n} - \alpha_{B,o})]\\ &\quad + (\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma'_\alpha W_n I_{0,n}\\ &\quad + I'_{1,n}W_n[\overline{g}_n(\alpha_o) + \Gamma_\alpha(\widehat{\alpha}_{B,n} - \alpha_{B,o})] + I'_{1,n}W_nI_{0,n}\\ &= (\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n})'\Gamma'_\alpha W_n[\overline{g}_n(\alpha_o) + \Gamma_\alpha(\widehat{\alpha}_{B,n} - \alpha_{B,o})]\\ &\quad + o_p(n^{-\frac{1}{2}}\varepsilon_n), \end{aligned} \quad (A.93)$$

following (A.43), (A.55), Assumption 3.1(iii), $\varepsilon_n = o(n^{-\frac{1}{2}})$ and $O_p(n^{-\frac{1}{2}}\varsigma_n\tau_n) = o_p(n^{-\frac{1}{2}}\varepsilon_n)$. Applying (A.89) and (A.93) in (A.49) and using $\widehat{\alpha}^*_{B,n} - \widehat{\alpha}_{B,n} = \varepsilon_n u^*_n = \varepsilon_n(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\gamma^*_n$, we obtain

$$\gamma^{*\prime}_n(\Gamma'_\alpha W_n\Gamma_\alpha)^{-1}\Gamma'_\alpha W_n[\sqrt{n}\overline{g}_n(\alpha_o) + \sqrt{n}\Gamma_\alpha(\theta_o)(\widehat{\alpha}_{B,n} - \alpha_{B,o})] \geq o_p(1) \quad (A.94)$$

as in (A.80). The rest of the proof is the same as that of part (a) and hence is omitted. ■

Appendix B. Proofs on linear IV model

Proof of Lemma 4.1. By definition,

$$g_S(Z,\theta) = (Y_i - X_i\theta)Z^*_i \quad \text{and} \quad g_D(Z,\theta) = (Y_i - X_i\theta)\underline{Z}_i. \quad (B.1)$$

We write

$$\|\overline{m}_n(\theta_o) - \overline{m}(\theta_o)\|^2 = \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n u_iZ^*_i \\ \frac{1}{n}\sum_{i=1}^n u_i\underline{Z}_i - \beta_o \end{pmatrix}' \begin{pmatrix} \frac{1}{n}\sum_{i=1}^n u_iZ^*_i \\ \frac{1}{n}\sum_{i=1}^n u_i\underline{Z}_i - \beta_o \end{pmatrix} = \left\|\frac{1}{n}\sum_{i=1}^n u_iZ^*_i\right\|^2 + \left\|\frac{1}{n}\sum_{i=1}^n u_i\underline{Z}_i - \beta_o\right\|^2. \quad (B.2)$$

Using Conditions 4.1(i)–(iii), we get

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n u_iZ^*_i\right\|^2\right] = \frac{1}{n}\sum_{j=1}^{k_0}\mathbb{E}\left[\left|u_iZ^*_i(j)\right|^2\right] \leq \frac{k_0C}{n} \quad (B.3)$$

and

$$\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n u_i\underline{Z}_i - \beta_o\right\|^2\right] = \frac{1}{n}\sum_{j\in D}\mathbb{E}\left[\left|u_i\underline{Z}_i(j)\right|^2\right] \leq \frac{k_nC}{n}. \quad (B.4)$$

Using the upper bounds in (B.3) and (B.4), and the expression in (B.2), we can invoke the Markov inequality to deduce that

$$\|\overline{m}_n(\theta_o) - \overline{m}(\theta_o)\|^2 = O_p(k_n/n). \quad (B.5)$$

This proves part (a).

For any α, we can write

$$\overline{g}_n(\alpha) - \overline{g}_n(\alpha_o) = \begin{pmatrix} \frac{-1}{n}\sum_{i=1}^n X_iZ^*_i & \mathbf{0}_{k_0\times(k_n-k_0)} \\ \frac{-1}{n}\sum_{i=1}^n X_i\underline{Z}_i & -I_{k_n-k_0} \end{pmatrix} \times (\alpha - \alpha_o) \equiv \overline{\Gamma}_n(\alpha - \alpha_o). \quad (B.6)$$

Moreover, we get

$$\|\overline{g}_n(\alpha)-\overline{g}_n(\alpha_o)\|^2=(\alpha-\alpha_o)'\overline{\Gamma}_n'\overline{\Gamma}_n(\alpha-\alpha_o). \tag{B.7}$$

Define $\Gamma_n\equiv\mathbb{E}\left[\overline{\Gamma}_n\right]$. Then

$$\overline{\Gamma}_n-\Gamma_n=\begin{pmatrix}\frac{-1}{n}\sum_{i=1}^n X_iZ_i^*+\mathbb{E}\left[X_iZ_i^*\right] & \mathbf{0}_{k_0\times(k_n-k_0)}\\ \frac{-1}{n}\sum_{i=1}^n X_i\underline{Z}_i+\mathbb{E}\left[X_i\underline{Z}_i\right] & \mathbf{0}_{(k_n-k_0)\times(k_n-k_0)}\end{pmatrix}. \tag{B.8}$$

Using the Hölder inequality, the Cauchy–Schwarz inequality and Conditions 4.1(i) and (iii), we get

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n X_iZ_i^*-\mathbb{E}\left[X_iZ_i^*\right]\right\|^2\right]\\&=\frac{1}{n}\sum_{j=1}^{k_0}\mathbb{E}\left[\left|X_iZ_i^*(j)-\mathbb{E}[X_iZ_i^*(j)]\right|^2\right]\\&\le\frac{1}{n}\sum_{j=1}^{k_0}\mathbb{E}\left[\left|X_iZ_i^*(j)\right|^2\right]\le\frac{\sqrt{\mathbb{E}\left[X_i^4\right]}\sum_{j=1}^{k_0}\sqrt{\mathbb{E}\left[|Z_i^*(j)|^4\right]}}{n}\\&\le\frac{Ck_0}{n}.\end{aligned} \tag{B.9}$$

Similarly, we have

$$\begin{aligned}&\mathbb{E}\left[\left\|\frac{1}{n}\sum_{i=1}^n X_i\underline{Z}_i'-\mathbb{E}\left[X_i\underline{Z}_i'\right]\right\|^2\right]\\&=\frac{1}{n}\sum_{j\in D}\mathbb{E}\left[\left|X_i\underline{Z}_i(j)-\mathbb{E}[X_i\underline{Z}_i(j)]\right|^2\right]\\&\le\frac{1}{n}\sum_{j\in D}\mathbb{E}\left[\left|X_i\underline{Z}_i(j)\right|^2\right]\le\frac{\sqrt{\mathbb{E}\left[X_i^4\right]}\sum_{j\in D}\sqrt{\mathbb{E}\left[|\underline{Z}_i(j)|^4\right]}}{n}\\&\le\frac{Ck_n}{n}.\end{aligned} \tag{B.10}$$

Using the upper bounds in (B.9) and (B.10) and the expression in (B.8), we invoke the Markov inequality to deduce that

$$\left\|\overline{\Gamma}_n-\Gamma_n\right\|^2=O_p(k_n/n). \tag{B.11}$$

Define

$$\Gamma_{1,n}\equiv\mathbb{E}\left[X_iZ_i^*\right]\quad\text{and}\quad\Gamma_{2,n}=\mathbb{E}\left[X_i\underline{Z}_i\right]. \tag{B.12}$$

Then,

$$\begin{aligned}\Gamma_n'\Gamma_n&=\begin{pmatrix}\mathbb{E}\left[X_iZ_i^{*\prime}\right]\mathbb{E}\left[X_iZ_i^*\right]+\mathbb{E}\left[X_i\underline{Z}_i'\right]\mathbb{E}\left[X_i\underline{Z}_i\right] & \mathbb{E}\left[X_i\underline{Z}_i'\right]\\ \mathbb{E}\left[X_i\underline{Z}_i\right] & I_{k_n-k_0}\end{pmatrix}\\&=\begin{pmatrix}\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n} & \Gamma_{2,n}'\\ \Gamma_{2,n} & I_{k_n-k_0}\end{pmatrix}\\&=\begin{pmatrix}1 & \Gamma_{2,n}'\\ \mathbf{0}_{(k_n-k_0)\times1} & I_{k_n-k_0}\end{pmatrix}\begin{pmatrix}\Gamma_{1,n}'\Gamma_{1,n} & \mathbf{0}_{1\times(k_n-k_0)}\\ \mathbf{0}_{(k_n-k_0)\times1} & I_{k_n-k_0}\end{pmatrix}\\&\quad\times\begin{pmatrix}1 & \mathbf{0}_{1\times(k_n-k_0)}\\ \Gamma_{2,n} & I_{k_n-k_0}\end{pmatrix}.\end{aligned} \tag{B.13}$$

By Conditions 4.1(iv) and (v), we know that $\Gamma_{1,n}'\Gamma_{1,n}>0$, and hence $\Gamma_n'\Gamma_n$ is positive definite. Let $\lambda_{n,*}$ be an eigenvalue of $\Gamma_n'\Gamma_n$ with $\lambda_{n,*}\neq1$. Then

$$\begin{aligned}0&=\det\left[\begin{pmatrix}\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n}-\lambda_{n,*} & \Gamma_{2,n}'\\ \Gamma_{2,n} & (1-\lambda_{n,*})I_{k_n-k_0}\end{pmatrix}\right]\\&=\det\left[\begin{pmatrix}\frac{\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n}-\lambda_{n,*}}{1-\lambda_{n,*}} & \frac{\Gamma_{2,n}'}{1-\lambda_{n,*}}\\ \frac{\Gamma_{2,n}}{1-\lambda_{n,*}} & I_{k_n-k_0}\end{pmatrix}\right]\\&\quad\times(1-\lambda_{n,*})^{k_n-k_0+1}\\&=\left(\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n}-\lambda_{n,*}-\frac{\Gamma_{2,n}'\Gamma_{2,n}}{1-\lambda_{n,*}}\right)\\&\quad\times(1-\lambda_{n,*})^{k_n-k_0},\end{aligned} \tag{B.14}$$

where the last equality holds by $\det(ABC)=\det(A)\det(B)\det(C)$ with

$$\begin{aligned}A&=\begin{pmatrix}1 & \frac{\Gamma_{2,n}'}{1-\lambda_{n,*}}\\ \mathbf{0}_{(k_n-k_0)\times1} & I_{k_n-k_0}\end{pmatrix}=C'\quad\text{and}\\B&=\begin{pmatrix}\frac{\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n}-\lambda_{n,*}}{1-\lambda_{n,*}}-\frac{\Gamma_{2,n}'\Gamma_{2,n}}{(1-\lambda_{n,*})^2} & \mathbf{0}_{1\times(k_n-k_0)}\\ \mathbf{0}_{(k_n-k_0)\times1} & I_{k_n-k_0}\end{pmatrix}.\end{aligned} \tag{B.15}$$

From (B.14), we know that $\lambda_{n,*}$ satisfies

$$\lambda_{n,*}^2-(1+\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n})\lambda_{n,*}+\Gamma_{1,n}'\Gamma_{1,n}=0. \tag{B.16}$$

The above equation has the following two solutions

$$\lambda_{n,*}=\frac{\Pi_n\pm\sqrt{\Pi_n^2-4\Gamma_{1,n}\Gamma_{1,n}'}}{2}, \tag{B.17}$$

where $\Pi_n\equiv1+\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n}$. This implies that the eigenvalues of $\Gamma_n'\Gamma_n$ are bounded from below by $\min(\lambda_{n,*},1)$ and bounded from above by $\max(\lambda_{n,*},1)$. Under Conditions 4.1(iv) and (v),

$$\Gamma_{1,n}'\Gamma_{1,n}=\mathbb{E}\left[X_iZ_i^{*\prime}\right]\mathbb{E}\left[X_iZ_i^*\right]=\sum_{j=1}^{k_0}\pi_j^2 \tag{B.18}$$

and

$$\begin{aligned}\Gamma_{2,n}'\Gamma_{2,n}&=\mathbb{E}\left[X_i\underline{Z}_i'\right]\mathbb{E}\left[X_i\underline{Z}_i\right]=\sum_{j=k_0+1}^{k_0+d_A}\pi_j^2+\mathbb{E}\left[X_i\underline{Z}_{2,i}'\right]\mathbb{E}\left[X_i\underline{Z}_{2,i}\right]\\&\to\sum_{j=k_0+1}^{\infty}\pi_j^2+\lim_{n\to\infty}\left\|\mathbb{E}\left[X_i\underline{Z}_{2,i}\right]\right\|^2.\end{aligned} \tag{B.19}$$

Using the results in (B.18), (B.19) and the inequality

$$\Pi_n+\sqrt{\Pi_n^2-4\Gamma_{1,n}'\Gamma_{1,n}}\le\Pi_n+\sqrt{\Pi_n^2}=2\Pi_n, \tag{B.20}$$

we deduce that

$$\begin{aligned}\lambda_{n,*,-}&\equiv\frac{2\Gamma_{1,n}\Gamma_{1,n}'}{\Pi_n+\sqrt{\Pi_n^2-4\Gamma_{1,n}'\Gamma_{1,n}}}\ge\frac{\Gamma_{1,n}'\Gamma_{1,n}}{1+\Gamma_{1,n}'\Gamma_{1,n}+\Gamma_{2,n}'\Gamma_{2,n}}\\&\to\frac{\sum_{j=1}^{k_0}\pi_j^2}{1+\sum_{j=1}^{\infty}\pi_j^2+\lim_{n\to\infty}\left\|\mathbb{E}\left[X_i\underline{Z}_{2,i}\right]\right\|^2}\quad\text{as } n\to\infty.\end{aligned} \tag{B.21}$$

By $\mathbb{E}\left[X_i^2\right]\le C$ and Condition 4.1(iv), we see that $\sum_{j=1}^{\infty}\pi_j^2\le C$, which together with $\sum_{j=1}^{k_0}\pi_j^2>0$ and $\lim_{n\to\infty}\left\|\mathbb{E}\left[X_i\underline{Z}_{2,i}\right]\right\|^2$

$< C$ implies that the smallest eigenvalue of $\Gamma_n'\Gamma_n$ is bounded from below uniformly in n. Next, using the results in (B.18), (B.19) and the inequality in (B.20), we have

$$\lambda_{n,*,+} \equiv \frac{\Pi_n + \sqrt{\Pi_n^2 - 4\Gamma_{1,n}'\Gamma_{1,n}}}{2} \leq 1 + \Gamma_{1,n}'\Gamma_{1,n} + \Gamma_{2,n}'\Gamma_{2,n} \to 1 + \sum_{j=1}^{\infty}\pi_j^2 + \lim_{n\to\infty}\left\|\mathbb{E}\left[X_i\underline{Z}_{2,i}\right]\right\|^2, \tag{B.22}$$

which together with $\sum_{j=1}^{\infty}\pi_j^2 \leq C$ and $\lim_{n\to\infty}\left\|\mathbb{E}\left[X_i\underline{Z}_{2,i}\right]\right\|^2 < C$ implies that the largest eigenvalue of $\Gamma_n'\Gamma_n$ is bounded from above uniformly in n. Hence, for some constant C,

$$C^{-1} \leq \rho_{\min}(\Gamma_n'\Gamma_n) \leq \rho_{\max}(\Gamma_n'\Gamma_n) \leq C \quad \text{for all } n. \tag{B.23}$$

Using the inequalities in (B.23) and

$$\|\Gamma_n(\alpha-\alpha_o)\| = \left[(\alpha-\alpha_o)'\Gamma_n'\Gamma_n(\alpha-\alpha_o)\right]^{\frac{1}{2}}, \tag{B.24}$$

we deduce that

$$\|\alpha-\alpha_o\| \lesssim \|\Gamma_n(\alpha-\alpha_o)\| \lesssim \|\alpha-\alpha_o\|. \tag{B.25}$$

Note that

$$\begin{aligned}&(\alpha-\alpha_o)'\overline{\Gamma}_n'\overline{\Gamma}_n(\alpha-\alpha_o) - (\alpha-\alpha_o)'\Gamma_n'\Gamma_n(\alpha-\alpha_o)\\ &= (\alpha-\alpha_o)'(\overline{\Gamma}_n-\Gamma_n)'(\overline{\Gamma}_n-\Gamma_n)(\alpha-\alpha_o)\\ &\quad + 2(\alpha-\alpha_o)'(\overline{\Gamma}_n-\Gamma_n)'\Gamma_n(\alpha-\alpha_o).\end{aligned} \tag{B.26}$$

Using the triangle inequality, the Cauchy–Schwarz inequality, (B.11), (B.23) and (B.25), we obtain

$$\begin{aligned}&(\alpha-\alpha_o)'\overline{\Gamma}_n'\overline{\Gamma}_n(\alpha-\alpha_o)\\ &\leq (\alpha-\alpha_o)'\Gamma_n'\Gamma_n(\alpha-\alpha_o) + \left|(\alpha-\alpha_o)'\overline{\Gamma}_n'\overline{\Gamma}_n(\alpha-\alpha_o)\right.\\ &\quad \left. - (\alpha-\alpha_o)'\Gamma_n'\Gamma_n(\alpha-\alpha_o)\right|\\ &\lesssim \|\alpha-\alpha_o\|^2 + \left\|\overline{\Gamma}_n-\Gamma_n\right\|\|\alpha-\alpha_o\|^2\\ &\quad + \|\alpha-\alpha_o\|\left\|\overline{\Gamma}_n-\Gamma_n\right\|\|\Gamma_n(\alpha-\alpha_o)\|\\ &\lesssim \left[1 + O_p(k_n/n) + O_p(\sqrt{k_n/n})\right]\|\alpha-\alpha_o\|^2,\end{aligned} \tag{B.27}$$

which proves the result claimed in (b). Similarly,

$$\begin{aligned}&(\alpha-\alpha_o)'\overline{\Gamma}_n'\overline{\Gamma}_n(\alpha-\alpha_o)\\ &\geq (\alpha-\alpha_o)'\Gamma_n'\Gamma_n(\alpha-\alpha_o) - \left|(\alpha-\alpha_o)'\overline{\Gamma}_n'\overline{\Gamma}_n(\alpha-\alpha_o)\right.\\ &\quad \left. - (\alpha-\alpha_o)'\Gamma_n'\Gamma_n(\alpha-\alpha_o)\right|\\ &\gtrsim \|\alpha-\alpha_o\|^2 - \left\|\overline{\Gamma}_n-\Gamma_n\right\|\|\alpha-\alpha_o\|^2\\ &\quad - \|\alpha-\alpha_o\|\left\|\overline{\Gamma}_n-\Gamma_n\right\|\|\Gamma_n(\alpha-\alpha_o)\|\\ &\gtrsim \left[1 + O_p(k_n/n) + O_p(\sqrt{k_n/n})\right]\|\alpha-\alpha_o\|^2,\end{aligned} \tag{B.28}$$

which proves the result claimed in (c).

Finally, we verify the claim in part (d). For any θ_1, θ_2 with $\|\theta_1-\theta_2\| \leq \delta$, we can use the Cauchy–Schwarz inequality to deduce that

$$\begin{aligned}\|v_n(\theta_1) - v_n(\theta_2)\| &= \left\|\begin{pmatrix}\frac{1}{n}\sum_{i=1}^n X_iZ_i^* - \mathbb{E}[X_iZ_i^*]\\ \frac{1}{n}\sum_{i=1}^n X_i\underline{Z}_i - \mathbb{E}[X_i\underline{Z}_i]\end{pmatrix}(\theta_2-\theta_1)\right\|\\ &\leq \left\|\overline{\Gamma}_n-\Gamma_n\right\|\|\theta_1-\theta_2\|\end{aligned} \tag{B.29}$$

which together with (B.11) implies that

$$\begin{aligned}&\sup_{\theta_1,\theta_2\in\{\theta\in\Theta:\|\theta-\theta_o\|\leq\delta_n\}}\frac{\|v_n(\theta_1)-v_n(\theta_2)\|}{n^{-\frac{1}{2}}+\|\theta_1-\theta_2\|}\\ &\leq \sup_{\theta_1,\theta_2\in\{\theta\in\Theta:\|\theta-\theta_o\|\leq\delta_n\}}\frac{\left\|\overline{\Gamma}_n-\Gamma_n\right\|\|\theta_1-\theta_2\|}{\|\theta_1-\theta_2\|}\\ &= \left\|\overline{\Gamma}_n-\Gamma_n\right\| = O_p(\sqrt{k_n/n}).\end{aligned} \tag{B.30}$$

This finishes the proof. ■

Proof of Lemma 4.2. We first show that Assumption 3.6(ii) is satisfied. Under Conditions 4.1(ii) and 4.2(ii), for any $\gamma \in R^{d_{B_1}+d_{B_0}}$ we have

$$\begin{aligned}\gamma'\Omega_{2,n}\gamma &= \mathbb{E}\left[u_i^2\gamma'\underline{Z}_{-1,i}\underline{Z}_{-1,i}'\gamma\right] - (\gamma'\beta_{B,o})^2\\ &\leq C\gamma'\mathbb{E}\left[\underline{Z}_{-1,i}\underline{Z}_{-1,i}'\right]\gamma \leq C\gamma'\gamma\end{aligned}$$

which implies that $\rho_{\max}(\Omega_{2,n}) \leq C$. Similarly, using Condition 4.1(ii) and (iv), we have $\rho_{\max}(\Omega_{0,n}) \leq C$. Using Aronszajn's inequality (see, e.g., Theorem III.2.9 in Bhatia, 1997 or Theorem 1.2 in the Supplemental Appendix of this paper) and the fact that Ω_n is positive definite, we know that

$$\rho_{\max}(\Omega_n) \leq \rho_{\max}(\Omega_{0,n}) + \rho_{\max}(\Omega_{2,n}) \leq C. \tag{B.31}$$

By the inverse formula of partitioned matrix,

$$\Omega_n^{-1} = \begin{pmatrix}(\Omega_n^{11})^{-1} & -\Omega_{0,n}^{-1}\Omega_{1,n}(\Omega_n^{22})^{-1}\\ -(\Omega_n^{22})^{-1}\Omega_{1,n}'\Omega_{0,n}^{-1} & (\Omega_n^{22})^{-1}\end{pmatrix}. \tag{B.32}$$

Using the Aronszajn's inequality, Condition 4.2(iii), and the fact that Ω_n is positive definite, we have

$$\rho_{\max}(\Omega_n^{-1}) \leq \rho_{\max}((\Omega_n^{11})^{-1}) + \rho_{\max}((\Omega_n^{22})^{-1}) \leq C, \tag{B.33}$$

which implies that $\rho_{\min}(\Omega_n) \geq C^{-1}$. This together with (B.31) yields

$$C^{-1} \leq \rho_{\min}(\Omega_n) \leq \rho_{\max}(\Omega_n) \leq C \tag{B.34}$$

which shows Assumption 3.6(ii).

To verify Assumption 3.6(i), we only need to check the Lindeberg's condition of the triangular array CLT. For this purpose, we define

$$\phi_{i,n} \equiv n^{-\frac{1}{2}}\gamma_n'\Omega_n^{-\frac{1}{2}}g(Z_i,\alpha_o), \tag{B.35}$$

where $\gamma_n \in R^{k_n}$ with $\|\gamma_n\| = 1$, and

$$g(Z_i,\alpha_o) \equiv \begin{pmatrix}u_iZ_i^*\\ u_i\underline{Z}_i - \mathbb{E}[u_i\underline{Z}_i]\end{pmatrix}. \tag{B.36}$$

For any random variable U with finite fourth moment, we can use monotonicity of the expectation operator and the inequality $U^2I\{U>1\} \leq U^4$ to deduce that

$$\mathbb{E}\left[U^2I\{U>1\}\right] \leq \mathbb{E}\left[U^4\right]. \tag{B.37}$$

Using the inequality above and Condition 4.1(i)

$$\begin{aligned}\sum_{i=1}^n\mathbb{E}\left[\phi_{i,n}^2I\{\phi_{i,n}>\epsilon\}\right] &= n\epsilon^2\mathbb{E}\left[\left(\phi_{i,n}/\epsilon\right)^2I\{\phi_{i,n}/\epsilon>1\}\right]\\ &\leq \frac{1}{n\epsilon^2}\mathbb{E}\left[\left|\gamma_n'\Omega_n^{-\frac{1}{2}}g(Z_i,\alpha_o)g(Z_i,\alpha_o)'\Omega_n^{-\frac{1}{2}}\gamma_n\right|^2\right].\end{aligned} \tag{B.38}$$

As $\gamma_n\gamma_n'$ is semi-positive definite with eigenvalues 0 and 1, we know that for any symmetric matrix A, there is

$$(\gamma_n'A\gamma_n)^2 = (\gamma_n'A)\gamma_n\gamma_n'(\gamma_n'A)' \leq \gamma_n'A^2\gamma_n, \tag{B.39}$$

which can be applied to show that

$$\mathbb{E}\left[\left|\gamma_n'\Omega_n^{-\frac{1}{2}}g(Z_i,\alpha_o)g(Z_i,\alpha_o)'\Omega_n^{-\frac{1}{2}}\gamma_n\right|^2\right]$$
$$\leq \gamma_n'\Omega_n^{-\frac{1}{2}}\mathbb{E}\left[\left[g(Z_i,\alpha_o)g(Z_i,\alpha_o)'\right]^2\right]\Omega_n^{-\frac{1}{2}}\gamma_n\gamma_n'\Omega_n^{-1}\gamma_n. \quad \text{(B.40)}$$

Under Assumption 3.6(ii) (which has been verified above), we have

$$\mathbb{E}\left[\left|\gamma_n'\Omega_n^{-\frac{1}{2}}g(Z_i,\alpha_o)g(Z_i,\alpha_o)'\Omega_n^{-\frac{1}{2}}\gamma_n\right|^2\right]$$
$$\lesssim \left\|\mathbb{E}\left[\left[g(Z_i,\alpha_o)g(Z_i,\alpha_o)'\right]^2\right]\right\| \leq \mathbb{E}\left[\|g(Z_i,\alpha_o)\|^4\right], \quad \text{(B.41)}$$

where the first inequality is by (B.40), $\gamma_n'\Omega_n^{-1}\gamma_n \leq C$, and the Cauchy–Schwarz inequality, the second inequality is by the Jensen's inequality and the Cauchy–Schwarz inequality. Next note that

$$\begin{aligned}\mathbb{E}\left[\|g(Z_i,\alpha_o)\|^4\right] &= \mathbb{E}\left[\left[g(Z_i,\alpha_o)'g(Z_i,\alpha_o)\right]^2\right]\\ &= \mathbb{E}\left[\left|u_i^2Z_i^{*\prime}Z_i^* + \left(u_i\underline{Z}_i-\beta_o\right)'\left(u_i\underline{Z}_i-\beta_o\right)\right|^2\right]\\ &\lesssim \mathbb{E}\left[\left|u_i^2Z_i^{*\prime}Z_i^*\right|^2\right]\\ &\quad+\mathbb{E}\left[\left|\left(u_i\underline{Z}_i-\beta_o\right)'\left(u_i\underline{Z}_i-\beta_o\right)\right|^2\right], \quad \text{(B.42)}\end{aligned}$$

where

$$\begin{aligned}&\mathbb{E}\left[\left|\left(u_i\underline{Z}_i-\beta_o\right)'\left(u_i\underline{Z}_i-\beta_o\right)\right|^2\right]\\ &= \mathbb{E}\left[\left|u_i^2\left(Z_{1,i}'Z_{1,i}+Z_{3,i}'Z_{3,i}\right)\right.\right.\\ &\quad\left.\left.+\left(u_i\underline{Z}_{2,i}-\beta_{B_1,o}\right)'\left(u_i\underline{Z}_{2,i}-\beta_{B_1,o}\right)\right|^2\right]\\ &\lesssim \mathbb{E}\left[\left|u_i^2Z_{1,i}'Z_{1,i}\right|^2\right]+\mathbb{E}\left[\left|u_i^2Z_{3,i}'Z_{3,i}\right|^2\right]\\ &\quad+\mathbb{E}\left[\left|\left(u_i\underline{Z}_{2,i}-\beta_{B_1,o}\right)'\left(u_i\underline{Z}_{2,i}-\beta_{B_1,o}\right)\right|^2\right]. \quad \text{(B.43)}\end{aligned}$$

Combining the inequalities in (B.42) and (B.43) and applying the Cauchy–Schwarz inequality, we get

$$\begin{aligned}&\mathbb{E}\left[\|g(Z_i,\alpha_o)\|^4\right]\\ &\lesssim k_n\sum_{j=1}^{k_0+d_A}\mathbb{E}\left[u_i^4Z_{1,i}^4(j)\right]+k_n\sum_{j=1}^{d_{B_0}}\mathbb{E}\left[u_i^4Z_{3,i}^4(j)\right]\\ &\quad+k_n\sum_{j=1}^{d_{B_1}}\mathbb{E}\left[\left|u_i\underline{Z}_{2,i}(j)-\beta_{B_1,o}(j)\right|^4\right]\\ &\lesssim k_n(k_0+d_A+d_{B_0}+d_{B_1})=k_n^2 \quad \text{(B.44)}\end{aligned}$$

where the second inequality follows from Condition 4.2(i). The result in (B.44), together with (B.38) and (B.41), implies that

$$\sum_{i=1}^n\mathbb{E}\left[\phi_{i,n}^2I\{\phi_{i,n}>\epsilon\}\right]\lesssim\frac{k_n^2}{n\epsilon^2}=o(1). \quad \text{(B.45)}$$

Hence the Lindeberg's condition holds in the linear IV model, which verifies Assumption 3.6(i).

Using the eigenvalue bounds in (B.23) and the arguments used to show (A.55), we deduce that Assumption 3.6(iii) is also satisfied.

Finally, Assumption 3.8 holds automatically in this linear model. This finishes the proof. ■

Appendix C. Proof of results in Section 5

When the data are i.i.d. we use the preliminary estimator $\dot\theta_n$ in (2.13) to define

$$\begin{aligned}&\dot\Gamma_{S,n}\equiv\frac{1}{n}\sum_{i=1}^n\frac{\partial g_S(Z_i,\dot\theta_n)}{\partial\theta'}, \qquad \dot\Omega_{S,n}\equiv\frac{1}{n}\sum_{i=1}^n g_S(Z_i,\dot\theta_n)g_S(Z_i,\dot\theta_n)',\\ &\dot\Gamma_{S+\ell,n}\equiv\frac{1}{n}\sum_{i=1}^n\frac{\partial g_{S+\ell}(Z_i,\dot\theta_n)}{\partial\theta'},\\ &\dot\Omega_{S+\ell,n}\equiv\frac{1}{n}\sum_{i=1}^n g_{S+\ell}(Z_i,\dot\theta_n)g_{S+\ell}(Z_i,\dot\theta_n)'.\end{aligned} \quad \text{(C.1)}$$

The estimators of variance matrices V_S and $V_{S+\ell}$ are thus constructed as

$$\dot V_{n,S}\equiv\dot\Gamma_{S,n}'\dot\Omega_{S,n}^{-1}\dot\Gamma_{S,n} \quad\text{and}\quad \dot V_{n,S+\ell}\equiv\dot\Gamma_{S+\ell,n}'\dot\Omega_{S+\ell,n}^{-1}\dot\Gamma_{S+\ell,n}. \quad \text{(C.2)}$$

The empirical information measure is defined using the above variance estimators and formula (2.12).

Proof of Lemma 5.1. (a) By the definition of $\|\omega_{n,B_1}\|$, the triangle inequality and the C_r-inequality,

$$\begin{aligned}n\lambda_n^2\|\omega_{n,B_1}\|^2 &= n\lambda_n^2\sum_{\ell\in B_1}\dot\mu_{n,\ell}^{2r_1}|\dot\beta_{n,\ell}|^{-2r_2}\\ &\lesssim n\lambda_n^2\sum_{\ell\in B_1}\frac{\mu_{o,\ell}^{2r_1}+\left|\dot\mu_{n,\ell}-\mu_{o,\ell}\right|^{2r_1}}{\left[|\beta_{o,\ell}|-|\dot\beta_{n,\ell}-\beta_{o,\ell}|\right]^{2r_2}}\end{aligned} \quad \text{(C.3)}$$

which together with $\max_{\ell\in B_1}|\dot\beta_{n,\ell}-\beta_{o,\ell}|=O_p(\tau_n)$ and $|\beta_{o,\ell}|\geq C^{-1}$ for all $\ell\in B_1$, implies that

$$n\lambda_n^2\|\omega_{n,B_1}\|^2\lesssim n\lambda_n^2\sum_{\ell\in B_1}|\beta_{o,\ell}|^{-2r_2}\left[\mu_{o,\ell}^{2r_1}+\left|\dot\mu_{n,\ell}-\mu_{o,\ell}\right|^{2r_1}\right] \quad \text{(C.4)}$$

w.p.a.1. Using $\max_{\ell\in B_1}\left|\dot\mu_{n,\ell}-\mu_{o,\ell}\right|=O_p(\tau_n)$, and the inequality above, we deduce that

$$\begin{aligned}n\lambda_n^2\|\omega_{n,B_1}\|^2 &\lesssim n\lambda_n^2\sum_{\ell\in B_1}|\beta_{o,\ell}|^{-2r_2}\mu_{o,\ell}^{2r_1}\\ &\quad+n\lambda_n^2\sum_{\ell\in B_1}|\beta_{o,\ell}|^{-2r_2}\left|\dot\mu_{n,\ell}-\mu_{o,\ell}\right|^{2r_1}\\ &\lesssim n\lambda_n^2k_n+n\lambda_n^2k_nO_p(\tau_n^{2r_1})=o_p(1),\end{aligned} \quad \text{(C.5)}$$

where the last equality follows from $r_1>0$ and $n^{\frac{1}{2}}k_n^{\frac{1}{2}}\lambda_n=o(1)$. By the definition of $\|\omega_{n,B_0}\|$,

$$\begin{aligned}n\lambda_n^2\|\omega_{n,B_0}\|^2 &= n\lambda_n^2\sum_{\ell\in B_0}\dot\mu_{n,\ell}^{2r_1}|\dot\beta_{n,\ell}|^{-2r_2}\\ &= n^{1+r_2-r_1}\lambda_n^2\sum_{\ell\in B_0}\frac{\left|\sqrt{n}\left(\dot\mu_{n,\ell}-\mu_{o,\ell}\right)\right|^{2r_1}}{\left|\sqrt{n}\left(\dot\beta_{n,\ell}-\beta_{o,\ell}\right)\right|^{2r_2}},\end{aligned} \quad \text{(C.6)}$$

which together with $\max_{\ell\in B_0}\left|\dot\mu_{n,\ell}-\mu_{o,\ell}\right|=O_p(n^{-\frac{1}{2}})$ and $\sqrt{n}\left(\dot\beta_{n,\ell}-\beta_{o,\ell}\right)\to_d N(0,\sigma_\ell^2)$ implies that

$$n\lambda_n^2\|\omega_{n,B_0}\|^2=n^{1+r_2-r_1}\lambda_n^2O_p(1)=o_p(1), \quad \text{(C.7)}$$

where the last equality is by $n^{\frac{1+r_2-r_1}{2}}\lambda_n=o(1)$. The desired result holds because $\|\omega_{n,B}\|^2=\|\omega_{n,B_0}\|^2+\|\omega_{n,B_1}\|^2$.

(b). By the definition of $\omega_{n,\ell}$ and the triangle inequality,

$$\begin{aligned}\frac{k_n \max_{\ell\in A}\omega_{n,\ell}^{-2}}{n\lambda_n^2} &= \frac{k_n}{n\lambda_n^2}\max_{\ell\in A}\frac{\dot\beta_{n,\ell}^{2r_2}}{\dot\mu_{n,\ell}^{2r_1}}\\ &\le \frac{k_n}{n\lambda_n^2}\max_{\ell\in A}\frac{\left|\dot\beta_{n,\ell}-\beta_{o,\ell}\right|^{2r_2}}{\left|\mu_{o,\ell}-\left|\dot\mu_{n,\ell}-\mu_{o,\ell}\right|\right|^{2r_1}}\\ &\lesssim \frac{k_n}{n\lambda_n^2}\max_{\ell\in A}|\mu_{o,\ell}|^{-2r_1}\left|\dot\beta_{n,\ell}-\beta_{o,\ell}\right|^{2r_2} \qquad \text{(C.8)}\end{aligned}$$

w.p.a.1, where the last inequality follows from $\max_{\ell\in A}|\dot\mu_{n,\ell}-\mu_{o,\ell}| = O_p(\tau_n)$ and $\max_{\ell\in A}|\mu_{o,\ell}^{-1}| \le C$. Using the inequality above, $\max_{\ell\in A}|\mu_{o,\ell}^{-1}| \le C$ and $\|\dot\beta_n-\beta_o\| = O_p(\tau_n)$, we have

$$\begin{aligned}\frac{k_n \max_{\ell\in A}\omega_{n,\ell}^{-2}}{n\lambda_n^2} &\lesssim \frac{k_n}{n\lambda_n^2}\left\|\dot\beta_n-\beta_o\right\|^{2r_2} = \frac{k_n}{n\lambda_n^2}O_p(\tau_n^{2r_2})\\ &= O_p(n^{-1}\lambda_n^{-2}k_n\tau_n^{2r_2}), \qquad \text{(C.9)}\end{aligned}$$

which together with $\tau_n = \sqrt{k_n/n}$, $k_n^{1+r_2}n^{-1-r_2}\lambda_n^{-2} = o(1)$ implies that $\frac{k_n\max_{\ell\in A}\omega_{n,\ell}^{-2}}{n\lambda_n^2} = o_p(1)$. ■

As discussed in Remark 5.1, the tuning parameter specified in (5.5) allows for the case that $\min_{\ell\in A}\left|\mu_{o,\ell}\right|$ and $\min_{\ell\in B_1}\left|\beta_{o,\ell}\right|$ go to zero as $k_n\to\infty$. To see it, suppose Assumption 5.1(i)–(iii) hold and the following restrictions

$$\begin{aligned}&\max_{\ell\in A}\left|\mu_{o,\ell}^{-1}\right| = o(n^{\frac{r_2}{4r_1}}k_n^{-\frac{1}{2r_1}-\frac{r_2}{4r_1}}) \quad \text{and}\\ &\max_{\ell\in B_1}\left|\beta_{o,\ell}^{-1}\right| = o(n^{\frac14}k_n^{-\frac{1}{2r_2}-\frac14}),\end{aligned} \qquad \text{(C.10)}$$

are satisfied. Using the restrictions above and the arguments in the proof of Lemma 5.1, we deduce that

$$n\lambda_n^2\left\|\omega_{n,B_1}\right\|^2 \lesssim n^{-\frac{r_2}{2}}k_n^{1+\frac{r_2}{2}}\max_{\ell\in B_1}\left|\beta_{o,\ell}\right|^{-2r_2} = o(1), \qquad \text{(C.11)}$$

when λ_n is as specified in (5.5) and moreover,

$$\begin{aligned}\frac{k_n \max_{\ell\in A}\omega_{n,\ell}^{-2}}{n\lambda_n^2} &\lesssim \frac{k_n}{n\lambda_n^2}\max_{\ell\in A}\frac{\left|\dot\beta_{n,\ell}-\beta_{o,\ell}\right|^{2r_2}}{|\mu_{o,\ell}|^{2r_1}}\\ &= O_p(n^{-\frac{r_2}{2}}k_n^{1+\frac{r_2}{2}})\max_{\ell\in A}|\mu_{o,\ell}|^{-2r_1} = o_p(1), \qquad \text{(C.12)}\end{aligned}$$

which implies that the tuning parameter in (5.5) also ensures that the lower bound is satisfied.

Appendix D. Some sufficient conditions

Let a.e. abbreviates almost everywhere. For any moment function $g_\ell(Z,\theta)$ and any $\theta\in\Theta$, we define $g_{\theta,\ell}(Z,\theta) = \frac{\partial g_\ell(Z,\theta)}{\partial\theta'}$. By definition, $g_{\theta,\ell}(Z,\theta)$ is a $1\times d_\theta$ vector with the ith element being $\frac{\partial g_\ell(Z,\theta)}{\partial\theta(i)}$, where $\theta(i)$ denotes the ith element of θ for any $i = 1,\dots,d_\theta$. We use $g_{\theta\theta,\ell}(Z,\theta)$ to denote the $d_\theta\times d_\theta$ matrix whose (i,j)th element being $\frac{\partial^2 g_\ell(Z,\theta)}{\partial\theta(i)\partial\theta(j)}$, where $\theta(i)$ and $\theta(j)$ $(i,j = 1,\dots,d_\theta)$ denote the ith and jth elements of θ respectively.

Lemma D.1. *Suppose* (i) *the observations are i.i.d.;* (ii) $g_\ell(Z,\theta)$ *is differentiable in* θ *a.e. for* $\ell = 1,\dots,k_n$*;* (iii)

$$\max_{\ell\le k_n}\mathbb{E}\left[\sup_{\theta\in\Theta}\|g_\ell(Z,\theta)\|^2 + \sup_{\theta\in\Theta}\left\|\partial g_{\theta,\ell}(Z,\theta)\right\|^2\right] \le C; \qquad \text{(D.1)}$$

and (iv) Θ *is compact. Then Assumption 3.2*(i) *holds with* $\tau_n = \sqrt{k_n/n}$.

Proof of Lemma D.1. Define $\mathcal{F} = \{g_\ell(Z,\theta) : \theta\in\Theta\}$. By Lemma 2.13 of Pakes and Pollard (1989), $\mathcal{F}$ is a Euclidean class with the envelope $F = \sup_{\theta\in\Theta}|g_\ell(Z,\theta)| + C\sup_{\theta\in\Theta}\|g_{\theta,\ell}(Z,\theta)\|$ for some positive constant C under assumptions (ii)–(iv) of Lemma D.1. For the definition of Euclidean class, see (2.7) of Pakes and Pollard (1989). By the maximal inequality (Section 4.3 of Pollard, 1989), for any ℓ and any n

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^n g_\ell(Z_i,\theta) - \mathbb{E}\left[g_\ell(Z,\theta)\right]\right|^2\right] \le Cn^{-1}. \qquad \text{(D.2)}$$

Hence,

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|n^{-1}\sum_{i=1}^n g(Z_i,\theta) - \mathbb{E}\left[g(Z,\theta)\right]\right\|^2\right] \le \frac{Ck_n}{n}, \qquad \text{(D.3)}$$

which implies

$$\sup_{\theta\in\Theta}\left\|n^{-1}\sum_{i=1}^n g(Z_i,\theta) - \mathbb{E}\left[g(Z,\theta)\right]\right\| = O_p(\sqrt{k_n/n}) \qquad \text{(D.4)}$$

by the Markov inequality. ■

Lemma D.2. *Suppose* (i) *conditions of Lemma D.1 hold;* (ii) $g_\ell(Z,\theta)$ *is twice differentiable in* θ *a.e. for* $\ell = 1,\dots,k_n$*;* (iii)

$$\max_{\ell\le k_n}\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|g_{\theta\theta,\ell}(Z_i,\theta)\right\|^2\right] \le C. \qquad \text{(D.5)}$$

Then, (a) *Assumption 3.5 holds with* $\varsigma_n = \sqrt{k_n/n}$*;* (b) $\varsigma_n\tau_n = o(n^{-\frac12})$ *holds if* $k_n = o(n^{\frac12})$.

Proof of Lemma D.2. Define

$$\upsilon_{\ell,n}(\theta) = n^{-1}\sum_{i=1}^n g_\ell(Z_i,\theta) - \mathbb{E}\left[g_\ell(Z,\theta)\right], \qquad \text{(D.6)}$$

which is continuously differentiable with respect to θ a.e. Let $\theta_1,\theta_2\in\Theta$ denote two different points in Θ. By the mean value expansion,

$$\begin{aligned}&\upsilon_{\ell,n}(\theta_1) - \upsilon_{\ell,n}(\theta_2)\\ &= \left[n^{-1}\sum_{i=1}^n\frac{\partial g_\ell(Z_i,\widetilde\theta_{n,\ell})}{\partial\theta'} - \frac{\partial\mathbb{E}\left[g_\ell(Z,\widetilde\theta_{n,\ell})\right]}{\partial\theta'}\right](\theta_1-\theta_2) \qquad \text{(D.7)}\end{aligned}$$

where $\widetilde\theta_{n,\ell}$ is some value between θ_1 and θ_2. We have $\max_{\ell\le k_n}\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{\partial g_\ell(Z,\theta)}{\partial\theta(j)}\right|\right] \le C$ for $j = 1,\dots,d_\theta$ by assumption (iii) of Lemma D.1. Hence, by the dominated convergence theorem, we can exchange "$\mathbb{E}$" and "∂" to obtain

$$\frac{\partial\mathbb{E}\left[g_\ell(Z,\theta)\right]}{\partial\theta'} = \mathbb{E}\left[\frac{\partial g_\ell(Z,\theta)}{\partial\theta'}\right] \quad \text{for any } \theta\in\Theta, \qquad \text{(D.8)}$$

which together with (D.7) implies that

$$\begin{aligned}&\upsilon_{\ell,n}(\theta_1) - \upsilon_{\ell,n}(\theta_2)\\ &= \left[n^{-1}\sum_{i=1}^n g_{\theta,\ell}(Z_i,\widetilde\theta_{n,\ell}) - \mathbb{E}[g_{\theta,\ell}(Z_i,\widetilde\theta_{n,\ell})]\right](\theta_1-\theta_2). \qquad \text{(D.9)}\end{aligned}$$

It follows that

$$\|\upsilon_n(\theta_1)-\upsilon_n(\theta_2)\|^2=\sum_{\ell\le k_n}\left\|\upsilon_{\ell,n}(\theta_1)-\upsilon_{\ell,n}(\theta_2)\right\|^2$$
$$\le\left(\sum_{\ell\le k_n}\left\|n^{-1}\sum_{i=1}^n g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})-\mathbb{E}[g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})]\right\|^2\right)$$
$$\times\ \|\theta_1-\theta_2\|^2 \tag{D.10}$$

by the Cauchy–Schwarz inequality.

Applying the proof of Lemma D.1 with $g_\ell(Z,\theta)$ replaced by $g_{\theta,\ell}(Z,\theta)$, under assumptions (i), (iii), (iv) of Lemma D.1 and assumption (iii) of Lemma D.2, we obtain

$$\mathbb{E}\left\|n^{-1}\sum_{i=1}^n g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})-\mathbb{E}[g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})]\right\|^2$$
$$\le\mathbb{E}\left[\sup_{\theta\in\Theta}\left\|n^{-1}\sum_{i=1}^n g_{\theta,\ell}(Z_i,\theta)-\mathbb{E}\left[g_{\theta,\ell}(Z,\theta)\right]\right\|^2\right]\le\frac{C}{n} \tag{D.11}$$

for all n and ℓ. It follows that

$$\mathbb{E}\left[\sum_{\ell\le k_n}\left\|n^{-1}\sum_{i=1}^n g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})-\mathbb{E}[g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})]\right\|^2\right]$$
$$\le\frac{Ck_n}{n}, \tag{D.12}$$

which in turn implies

$$\sum_{\ell\le k_n}\left\|n^{-1}\sum_{i=1}^n g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})-\mathbb{E}[g_{\theta,\ell}(Z_i,\widetilde{\theta}_{n,\ell})]\right\|^2=O_p(k_n/n) \tag{D.13}$$

by the Markov inequality.

Combining (D.10) and (D.13), we obtain

$$\sup_{\theta_1,\theta_2\in\{\theta\in\Theta:\|\theta-\theta_0\|\le\delta_n\}}\frac{\|\upsilon_n(\theta_1)-\upsilon_n(\theta_2)\|}{n^{-\frac{1}{2}}+\|\theta_1-\theta_2\|}$$
$$\le\sup_{\theta_1,\theta_2\in\Theta}\frac{\|\upsilon_n(\theta_1)-\upsilon_n(\theta_2)\|}{\|\theta_1-\theta_2\|}=O_p(\sqrt{k_n/n}). \tag{D.14}$$

This verifies Assumption 3.5(i) with $\varsigma_n=\sqrt{k_n/n}$. Part (b) holds because $\tau_n\varsigma_n=k_n/n=o(n^{-\frac{1}{2}})$ when $k_n=o(n^{\frac{1}{2}})$. ■

Appendix E. Supplementary data

Supplementary material related to this article can be found online at http://dx.doi.org/10.1016/j.jeconom.2015.02.019.

References

Andrews, D.W.K., 1994. Empirical process methods in econometrics. In: Engle, R.F., McFadden, D. (Eds.), Handbook of Econometrics, vol. IV. North-Holland, Amsterdam.

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. J. Bus. Econom. Statist. 20 (4), 530–544.

Andrews, D.W.K., Guggenberger, P., 2009. Hybrid and size-corrected subsampling methods. Econometrica 77 (3), 721–762.

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

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

Bai, J., Ng, S., 2009. Selecting instrumental variables in a data rich environment. J. Time Ser. Econom. 1 (1).

Bhatia, R., 1997. Matrix Analysis. Springer, New York.

Belloni, A., Chen, D., Chernozhukov, V., Hansen, C., 2012. Sparse models and methods for optimal instruments with an application to eminent domain. Econometrica 80, 2369–2429.

Belloni, A., Chernozhukov, V., Hansen, C., 2010. LASSO Methods for Gaussian Instrumental Variables Models. Preprint, arXiv:1012.1297.

Belloni, A., Chernozhukov, V., Hansen, C., 2011. Inference on Treatment Effects After Selection amongst High-dimensional Controls. Preprint, arXiv:1201.0224.

Berkowitz, D., Caner, M., Fang, Y., 2012. The validity of instruments revisited. J. Econometrics 166 (2), 255–266.

Breusch, T., Qian, H., Schmidt, P., Wyhowski, D., 1999. Redundancy of moment conditions. J. Econometrics 91, 89–111.

Caner, M., Han, X., Lee, Y., 2013. Bias-Corrected Semiparametrically Efficient High Dimensional GMM Estimator with Many Invalid Moment Conditions: An Application to Dynamic Panel Data Models, MIMEO.

Caner, M., Zhang, H., 2012. Adaptive Elastic Net for Generalized Methods of Moments. Unpublished Manuscript.

Carrasco, M., 2012. A regularization approach to the many instruments problem. J. Econometrics 170 (2), 383–398.

Chamberlain, G., Imbens, G.W., 2004. Random effects estimators with many instrumental variables. Econometrica 72, 295–306.

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

Conley, T.G., Hansen, C.B., Rossi, P.E., 2012. Plausibly exogenous. Rev. Econ. Statist. 94 (1), 260–272.

DiTraglia, F., 2012. Using Invalid Instruments on Purpose: Focused Moment Selection and Averaging for GMM, University of Pennsylvania, Working Paper.

Doko Tchatoka, F., Dufour, J.-M., 2012. Identification-robust Inference for Endogeneity Parameters in Linear Structural Models, MPRA Paper 40695, University Library of Munich, Germany.

Donald, S.G., Imbens, G.W., Newey, W.K., 2009. Choosing instrumental variables in conditional moment restriction models. J. Econometrics 152 (1), 28–36.

Donald, S.G., Newey, W.K., 2001. Choosing the number of instruments. Econometrica 69 (5), 1161–1191.

Eichenbaum, M.S., Hansen, L.P., Singleton, K.J., 1988. A time series analysis of representative agent models of consumption and leisure choice under uncertainty. Quart. J. Econ. 103 (1), 51–78.

Fan, J., Liao, Y., 2011. Ultra High dimensional Variable Selection with Endogenous Covariates, Princeton University and University of Maryland, Working Paper.

Gautier, E., Tsybakov, A.B., 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., Inoue, A., Jana, K., Shin, C., 2007. Information in generalized method of moments estimation and entropy based moment selection. J. Econometrics 138 (2), 488–512.

Hall, A.R., Inoue, A., Nason, J.M., Rossi, B., 2012. Information criteria for impulse response function matching estimation of DSGE models. J. Econometrics 170 (2), 499–518.

Hall, A.R., Peixe, F.P.M., 2003. A consistent method for the selection of relevant instruments. Econometric Rev. 22 (3), 269–288.

Hansen, L.P., 1982. Large sample properties of generalized method of moments estimators. Econometrica 50 (4), 1029–1054.

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

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

Inoue, A., 2006. A bootstrap approach to moment selection. Econom. J. 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., Okui, R., 2010. Constructing optimal instruments by first-stage prediction averaging. Econometrica 78 (2), 697–718.

Lee, Y., Zhou, Y., 2011. Averaged Instrumental Variable Estimator. University of Michigan, Working Paper.

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

Leeb, H., Pötscher, B.M., 2008. Sparse estimators and the oracle property, or the return of the hodges estimator. J. Econometrics 142 (1), 201–211.

Liao, Z., 2013. Adaptive GMM shrinkage estimation with consistent moment selection. Econometric Theory 29, 1–48.

McCloskey, A., 2012. Bonferroni-based Size-correction for Nonstandard Testing Problems. Brown University, Working Paper.

Newey, W.K., 1997. Convergence rates and asymptotic normality for series estimators. J. Econometrics 79, 147–168.

Nevo, A., Rosen, A., 2012. Identification with imperfect instruments. Rev. Econ. Stat. 93 (3), 659–671.

Okui, R., 2011. Instrumental variable estimation in the presence of many moment conditions. J. Econometrics 165 (1), 70–86.

Pakes, A., Pollard, D., 1989. Simulation and the asymptotics of optimization estimators. Econometrica 57, 1027–1057.

Pollard, D., 1984. Convergence of Stochastic Processes. Springer–Verlag, New York.

Pollard, D., 1989. Asymptotics via empirical processes. Statist. Sci. 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. Ann. Statist. 25, 2555–2591.
van der Vaart, A.W., Wellner, J.A., 1996. Weak Convergence and Empirical Processes. Springer, New York.
Zou, H., 2006. The adaptive lasso and its oracle properties. J. Amer. Statist. Assoc. 101 (476), 1418–1429.