Econometric Reviews, 35(8–10):1582–1608, 2016
Copyright © Taylor & Francis Group, LLC
ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474938.2015.1092805

Estimation of Sparse Structural Parameters with Many Endogenous Variables

Zhentao Shi
Department of Economics, Chinese University of Hong Kong, Hong Kong, China

We apply the generalized method of moments–least absolute shinkage and selection operator (GMM-Lasso) (Caner, 2009) to a linear structural model with many endogenous regressors. If the true parameter is sufficiently sparse, we can establish a new oracle inequality, which implies that GMM-Lasso performs almost as well as if we knew *a priori* the identities of the relevant variables. Sparsity, meaning that most of the true coefficients are too small to matter, naturally arises in econometric applications where the model can be derived from economic theory. In addition, we propose to use a modified version of AIC or BIC to select the tuning parameter in practical implementation. Simulations provide supportive evidence concerning the finite sample properties of the GMM-Lasso.

Keywords Big data; Endogeneity; GMM; High-dimensional; Sparsity.

JEL Classification C13; C26; C55.

1. INTRODUCTION

Endogenous variables are ubiquitous in empirical economic problems. When an empirical model is derived from economic theory, it is likely that the model includes several or many endogenous variables. For example, in demand theory a product on a market of K differentiated goods is associated with $K - 1$ cross-price elasticities, which leads to the well-known problem of "too many parameters"[1] (Ackerberg et al., 2007) when K is relatively large compared to the repeated observations across markets. For another example, many variables collected in a survey help explain expenditure, while a multitude of these explanatory variables are household decisions under the same budget constraints; Blundell et al. (1993) treat 14 variables as endogenous in their structural equation

Address correspondence to Zhentao Shi, 912 Esther Lee Building, the Chinese University of Hong Kong, Shatin, New Territories, Hong Kong, SAR; E-mail: zhentao.shi@cuhk.edu.hk

[1] In contrast to the recent paradigm (Berry et al., 1995), the classical approach results in "too many parameters" in price elasticity estimation, but it relies on fewer assumptions on economic agent behavior. See Ackerberg et al. (2007) for details.

comprising 93 explanatory variables. In both examples, the economic theories naturally endogenize many regressors.

Sparsity is another common phenomenon in sciences. Sparsity in this article roughly means that, in a model, many coefficients are either exactly zeros or close to zero. In economics, sparsity is buttressed by empirical observations. In the example of many differentiated products, we often observe that high-tier product prices are sensitive to those of similar quality, whereas lower-tier product prices have little influence on high-tier competitors. In the example of the expenditure survey, conventionally economists count on prior beliefs to select relevant variables, under which sparsity is implicitly assumed.

To allow asymptotics to achieve desirable accuracy in a model with many parameters, conventional estimation procedures would demand an enormous sample size that is sometimes too large to be realistic. However, if the problem is sparse, it is possible to modify a conventional procedure, usually via shrinkage, to take advantage of sparsity.

Shrinkage estimation dates back to Tikhonov (1943). Recently, our understanding of shrinkage methods has been extended to non-smooth penalties, in particular in high-dimensional regression models. Tibshirani (1996), Fan and Li (2001), Zou (2006), Bickel et al. (2009), and Belloni and Chernozhukov (2011) are among those important contributions.

In econometrics, Caner (2009) first introduces a least absolute shrinkage and selection operator (Lasso–type) penalty to the generalized method of moments (GMM) criterion function. Liao (2013) uses the Lasso-type shrinkage technique to select moment conditions in GMM. These two articles work with models of fixed dimensions. As large datasets become available, researchers are trying to extend econometric models to accommodate infinity-dimensional parameters. Gautier and Tsybakov (2013) establish non-asymptotic probabilistic bounds for linear instrumental variable (IV) models based on the Dantzig selector (Candes and Tao, 2007). Caner and Zhang (2014) explore the oracle property of their *elastic net estimator* in a GMM with the number of structural parameters diverging at a slower speed than the sample size. Moreover, Fan and Liao (2014) propose a novel *penalized focused generalized method of moments* criterion function to reduce the complexity of the high-dimension parameter, and they even allow the number of unknown parameters to be larger than the sample size. Under proper conditions, both Caner and Zhang (2014) and Fan and Liao (2014) achieve variable selection consistency—distinguishing with high probability the nonzero coefficients from those of zeros. If we call their setting the *exact* sparsity, our article allows the *approximately* sparse setting; that is, many components of the true parameter are small but do not have to be exactly zero.[2] Under approximate sparsity, it would be very challenging, if not impossible, for any method to correctly classify the coefficients at the margin. We therefore focus on the convergence rate instead.

Another line of research maintains a finite-dimensional parameter while allowing other components of the model to grow. Koenker and Machado (1999) derive the ratio of the

[2] The precise definition of sparsity will be given in Assumption 2.

number of moments to the sample size that can maintain the asymptotic normality of GMM. Chao and Swanson (2005) investigate the behavior of k-estimator under many weak instruments. Han and Phillips (2006) provide a comprehensive analysis of the consistency and the asymptotic distribution of a general nonlinear GMM with many moments. Belloni et al. (2012) propose a modified two-stage least square (2SLS) in which the reduced-form regression uses Lasso to select the optimal set of instruments. This is a highly relevant article, from which we actually acquire asymptotic techniques. However, the different settings lead to different procedures and results. Belloni et al. (2012) are concerned with a finite-dimensional structural parameter, so that they implement Lasso in the first stage reduced-form. GMM-Lasso penalizes the structural parameter while leaving the IV untouched. Belloni et al. (2012) develop the asymptotic distribution for the finite-dimensional parameter of interest, whereas our objective is the convergence rate of the high-dimensional parameter. Statistical inference theory for our setting has not been established, to the best of our knowledge.

In this article, we complement Caner's (2009) and Caner and Zhang's (2014) research by exploring the gain of shrinkage in a *moderately* high-dimensional linear IV model. "Moderately" here means that the unknown parameter dimension grows more slowly than the sample size—the same asymptotic framework taken by Caner and Zhang (2014). We establish a new oracle inequality to show the theoretical advantage of the *GMM-Lasso* estimator, a special case of Caner's (2009) Lasso-type GMM, in a sparse setting. In practice, we also provide a procedure to select the tuning parameter based on Akaike information criterion (AIC) or Bayesian information criterion (BIC). We evaluate the finite-sample performance in simulations, which support the theoretical properties of GMM-Lasso. Moreover, the simulations show interesting bias-variance trade-off of GMM-Lasso when both the shrinkage bias and the endogeneity bias are present.

The rest of the article is organized as follows. Section 2 discusses the setting and the estimator. Section 3 presents the key conditions and develops the theoretical results for GMM-Lasso. Section 4 shows its finite-sample performance in simulations.

2. SETTING AND ESTIMATION

We consider a linear structural equation

$$y_i = x_i'\beta_n + \epsilon_i,$$

where $x_i := \left(x_{i1}, \dots, x_{iK_n}\right)'$ is a K_n-vector of endogenous explanatory variables with zero mean and unit variance,[3] ϵ_i is a zero-mean structural error, y_i is a scalar dependent

[3] For notational conciseness in Section 3, we assume all components in x_i are endogenous with zero mean and unit variance. The analysis remains valid if some components in x_i are exogenous and they "instrument" themselves. The variance-normalization is desirable to make the effects of $\beta_k, k = 1, \dots, K_n$, comparable. In implementation, the zero-mean assumption has no effect when we add a constant in the explanatory variable; the sample standard error can be used to normalize the scale, which has no asymptotic effect.

variable, and β_n is a vector of coefficients. To address the endogeneity, we find an L_n-vector of instruments $z_i = \left(z_{i1}, \ldots, z_{iL_n}\right)'$. The sample consists of n independent identically distributed (i.i.d.) observations indexed by $i \in \{1, \ldots, n\}$.

2SLS is one of the most widely used estimators for linear models with endogenous regressors. It minimizes the criterion

$$(y - X\beta_n)' P_Z (y - X\beta_n), \tag{1}$$

where $y := (y_1, \ldots, y_n)'$ and $X := (x_1, \ldots, x_n)'$ stack the n observations into an n-vector and an $n \times K_n$ matrix, respectively; Z is defined similarly as an $n \times L_n$ matrix, and $P_Z := Z\left(Z'Z\right)^{-1} Z'$. When $Z'Z$ and $X'P_ZX$ are invertible, 2SLS has a closed-form

$$\widehat{\beta}_n^{(2SLS)} = \left(X'P_ZX\right)^{-1} X'P_Zy. \tag{2}$$

2SLS is a special case of GMM (Hansen, 1982), which minimizes, in this model, the criterion function

$$Q_n(\beta_n) := (y - X\beta_n)' ZW_nZ' (y - X\beta_n) \tag{3}$$

where W_n is a positive-definite weighting matrix.

The standard asymptotic framework passes n to infinity while fixing K_n and L_n. When the magnitude of K_n is comparable to n, however, an estimator's finite-sample behavior might be poorly approximated by asymptotic theory derived under such an embedding scheme. To acknowledge the influence of a large K_n, we allow K_n to diverge to infinity. Throughout this article, we work with a triangular array of models with the following assumption.

Assumptions 1. (a) $K_n \leq L_n$ for all n, and $K_n, L_n \to \infty$ as $n \to \infty$. (b) $n^{-1}L_n \log L_n \to 0$.

We call this embedding scheme in Assumption 1 the *diverging-K_n asymptotic framework.* Condition 1(a) allows the dimensions K_n and L_n to diverge while the order condition for identification holds—the number of the excluded instruments must be no fewer than the number of the endogenous variables. Condition (b) requires that n asymptotically dominates L_n. Were L_n of the same order as n, GMM would not converge to the true parameter in general.[4] The additional factor $\log L_n$ guarantees that those L_n sample moments converge uniformly in probability to their population moments.

[4] See Han and Phillips (2006) for the discussion of the source of potential inconsistency in the equally-weighted GMM. This restriction of the number of moments in GMM is overcome by Fan and Liao (2014) under exact sparsity via penalizing the focused GMM, while it remains unsolved under approximate sparsity.

In this setting, the standard GMM may be consistent under this diverging-K_n asymptotic framework along with proper assumptions. Nonetheless, if we know *a priori* that the true coefficient is sparse, we may modify GMM to achieve a faster rate of convergence. Inspired by Caner's (2009) Lasso-type GMM, we can attach a penalty term to the original GMM criterion so that we minimize

$$\frac{1}{n^2} Q_n(\beta_n) + \rho_n \|\beta_n\|_1, \tag{4}$$

where $\rho_n \in [0, \infty]$ is a tuning parameter, and $\|\beta_n\|_1 = \sum_{k=1}^{K_n} |\beta_k|$ is the l_1-norm of β_n. We denote $\widehat{\beta}_n$ as the minimizer of (4) and call it here the GMM-Lasso estimator, or simply GMM-Lasso.

Although we can establish the asymptotic theory as long as ρ_n satisfies some rate, we have to pick a number $\hat{\rho}_n$ for estimation in practice. We propose to use either the modified GMM-AIC

$$J_{B_n}^{\mathrm{AIC}}(\beta_n) = \frac{1}{n^2} Q_n(\beta_n) + \frac{2}{n} B_n |\beta_n|_0$$

or the modified GMM-BIC

$$J_{B_n}^{\mathrm{BIC}}(\beta_n) = \frac{1}{n^2} Q_n(\beta_n) + \frac{\log n}{n} B_n |\beta_n|_0,$$

where $|\cdot|_0$ is the number of nonzero components in a vector, and B_n is a slowly diverging deterministic sequence. They are Andrews and Lu's (2001) GMM-AIC and GMM-BIC adapted to a diverging number of models. This modification is suggested by Wang et al. (2009) for the linear regression, and adopted by Caner and Fan (2010). To make its dependence on ρ_n explicit, we can write $\widehat{\beta}_n = \widehat{\beta}_n(\rho_n)$. In the simulations, we will use $\widehat{\beta}_n\left(\widehat{\rho}_n^{(\mathrm{AIC})}\right)$ and $\widehat{\beta}_n\left(\widehat{\rho}_n^{(\mathrm{BIC})}\right)$, where

$$\widehat{\rho}_n^{(\mathrm{AIC})} = \arg\min_{\rho_n \in \mathbb{R}^+} J_{B_n}^{\mathrm{AIC}}\left(\widehat{\beta}_n(\rho_n)\right),$$

$$\widehat{\rho}_n^{(\mathrm{BIC})} = \arg\min_{\rho_n \in \mathbb{R}^+} J_{B_n}^{\mathrm{BIC}}\left(\widehat{\beta}_n(\rho_n)\right).$$

We call them GMM-Lasso-AIC and GMM-Lasso-BIC, respectively.

Next, we prepare the key conditions for GMM-Lasso and develop the main theoretical results in Section 3, and then we show the simulation performance of $\widehat{\beta}_n\left(\widehat{\rho}_n^{(\mathrm{AIC})}\right)$ and $\widehat{\beta}_n\left(\widehat{\rho}_n^{(\mathrm{BIC})}\right)$ in Section 4.

3. THEORETICAL ANALYSIS

3.1. Sparsity and Relevance of Instruments

Instruments are indispensable in addressing endogeneity in a structural econometric model. Besides being orthogonal to the structural error, valid instruments must also be *relevant* to the endogenous variables. Relevance here can be measured by the rank of the $L_n \times K_n$ population covariance matrix $\Sigma_n = (\sigma_{lk})_{l \leq L_n, k \leq K_n}$, where $\sigma_{lk} := \mathrm{cov}\,(z_{il}, x_{ik})$. The rank condition for identification demands a full-column-rank Σ_n. Under the diverging-K_n framework, we will assume a weakened counterpart of this rank condition.

In order to discuss the rank condition in this context, we need to introduce some notations. Let $\phi_{\Sigma_n 1}, \phi_{\Sigma_n 2}, \ldots, \phi_{\Sigma_n K_n}$ be the sequence of Σ_n's K_n singular values in descending order. Let $S_n \subset \{1, \ldots, K_n\}$ be a generic index set, and $\beta_{S_n} := \left(\beta_k \cdot 1\{k \in S_n\}\right)_{k=1}^{K_n}$ where $1\{\cdot\}$ is the indicator function. β_{S_n} is a K_n-vector that keeps the kth component of β_n if $k \in S_n$ but coerces the kth component to zero if $k \notin S_n$. Denote $S_n^c := \{1, \ldots, K_n\} \backslash S_n$ as S_n's complement set. By definition, $\beta_{S_n^c} = \left(\beta_k \cdot 1\{k \notin S_n\}\right)_{k=1}^{K_n}$ is a K_n-vector that keeps the components in S_n^c. Therefore, $\mathscr{B}(S_n) := \left\{\beta \in \mathbb{R}^{K_n} : \beta_{S_n^c} = \mathbf{0}_{K_n}\right\}$ is the set of the parameters with all the components in S_n^c being zeros; in other words, for any $\beta \in \mathscr{B}(S_n)$, the nonzero components reside only in S_n.

When all $\beta_n \in \mathbb{R}^{K_n}$ are considered, the rank condition for identification is equivalent to a nonzero smallest singular value Σ_n; that is, $\phi_2(K_n) := \min_{\beta_n \in \mathbb{R}^{K_n}} \|\beta_n\|_2^{-1} \sqrt{\beta_n' \Sigma_n' \Sigma_n \beta_n} > 0$, where $\|\beta_n\|_2 = \left(\sum_{k=1}^{K_n} \beta_k^2\right)^{1/2}$ is the l_2-norm of β_n. This rank condition can be weakened under sparsity, because we will consider not all possible $\beta_n \in \mathbb{R}^{K_n}$, but only on a restricted set instead. Define a *compatibility constant*[5] as

$$\phi_1(S_n) := \min_{\beta_n \in \mathscr{C}(S_n)} \frac{\sqrt{|S_n|}}{\left\|\beta_{S_n}\right\|_1} \sqrt{\beta_n' \Sigma_n' \Sigma_n \beta_n}.$$

where $|S_n|$ is the cardinality of S_n, and define

$$\mathscr{C}(S_n) := \left\{\beta_n \in \mathbb{R}^{K_n} : \left\|\beta_{S_n^c}\right\|_1 \leq 3 \left\|\beta_{S_n}\right\|_1\right\}.$$

The constant 3 in the definition of $\mathscr{C}(S_n)$ is simply a convenient valid integer. The set $\mathscr{C}(S_n)$ contains those parameters whose aggregate magnitude of the components on S_n^c is not too big relative to that on S_n. Unlike $\mathscr{B}(S_n)$, whose elements have at most $|S_n|$ nonzero components, the parameter component of the members in $\mathscr{C}(S_n)$ can include

[5] The compatibility constant here mimics Bühlmann and van de Geer's (2011, p. 109) definition in linear regressions. Bühlmann and van de Geer (2011, Chapter 6.13) give a comprehensive summary for various related quantities, amongst which the most well-known is Bickel et al. (2009)'s *restricted eigenvalue*, which minimizes an eigenvalue-type quantity over all sets of a certain cardinality. Discussion will follow in Remark 4 about the suitable choice of the minimal-eigenvalue-type quantities for our IV regression context.

those β_n such that $\beta_k \neq 0$ for all $k = 1, \ldots, K_n$. Here, $\phi_1(S_n) > 0$, instead of $\phi_2(K_n) > 0$, is the counterpart of the rank condition under the sparse setting. Note that the factor $\sqrt{|S_n| / \|\beta_{S_n}\|_1}$ in $\phi_1(S_n)$ replaces $\|\beta_n\|_2^{-1}$ in the definition of $\phi_2(K_n)$. It is obvious that $\phi_1(S_n) \geq \phi_2(K_n)$ for all $S_n \subset \{1, \ldots, K_n\}$ by the Cauchy–Schwarz inequality $\|\beta_{S_n}\|_1 \leq \sqrt{|S_n|}\,\|\beta_{S_n}\|_2$.

Now we come back to the model. To reasonably estimate the high-dimensional "true parameter" β_n^0 when we do not have an unrealistically big sample, some sense of sparsity of β_n^0 is a prerequisite. Sparsity roughly means only a few components in the parameter are large whereas the other components are small enough. To precisely define sparsity in Assumption 2 below, let ρ_n be a threshold chosen by the econometrician. We call β_k, the kth component of the vector β_n, a large coefficient if $|\beta_k| \geq \rho_n$; otherwise, we call it a small coefficient. Let $S_{0n} = S_{0n}(\beta_n^0, \rho_n) := \{k \in \{1, \ldots, K_n\} : |\beta_k^0| \geq \rho_n\}$ be the index of large components, and

$$t_{0n} = t_{0n}(\beta_n^0, \rho_n) := \sum_{k=1}^{K_n} |\beta_k^0| \, 1\{|\beta_k^0| \leq \rho_n\}$$

be the aggregate magnitude of the small components. It is difficult to accurately estimate too many small coefficients.

Besides β_n^0's small components, the difficulty of estimation arises when x_i and z_i are barely relevant. Under the standard asymptotic framework, GMM or 2SLS would encounter the "weak instruments" problem if $\phi_2(K_n)$ is too small. Similarly, in our diverging-K asymptotic framework under sparsity, we need a restriction on the relative sizes of K_n, L_n, n as well as $|S_n|$ and $\phi_1(S_n)$ to avoid weak instrumentation. To simplify notation, we denote

$$\lambda_n := \sqrt{n^{-1} L_n \log K_n L_n}$$
$$\psi(S_n) := |S_n| / \phi_1^2(S_n).$$

The definition of λ_n and Assumption 1 implies $\lambda_n \leq \sqrt{2n^{-1} L_n \log L_n} = o(1)$. Assumption 2 states in what sense we say a parameter is sparse.

Assumptions 2. There exists a sequence of (ρ_n) such that $\rho_n \to 0$, $\liminf_{n\to\infty} \rho_n/\lambda_n \geq C_\rho$,[6] and it satisfies (a) $t_{0n} \to 0$ and (b) $\rho_n \psi(S_{0n}) \to 0$.

Remark 3. Bühlmann and van de Geer (2011) present the Lasso theory with $\rho_n =$ constant × λ_n. Assumption 2 incorporates the case ρ_n/λ_n being a constant independent of n, while $\limsup_{n\to\infty} \rho_n/\lambda_n = \infty$ is also allowed. The terminology *compatibility constant*,

[6] The constant C_ρ's explicit form will be given in the proof of Theorem 10.

borrowed from the statistical literature, is somewhat misleading, since Condition (b) permits $\phi_1(S_{0n}) \to 0$ as $n \to \infty$ as long as $\rho_n \psi(S_{0n})$ vanishes in the limit.

Remark 4. Assumption 2(b) requires $\rho_n \psi(S_n) \to 0$ merely on the set S_{0n}. However, if we mimic Bickel et al.'s (2009, p. 1710) $\mathrm{RE}(s, c_0)$ and define a restrict eigenvalue as $\phi_1^{\mathrm{RE}}(s_n) := \min_{|S_n| \le s_n} \min_{\beta_n \in \mathscr{C}(S_n)} \frac{\sqrt{|S_n|}}{\|\beta_{S_n}\|_1} \sqrt{\beta_n' \Sigma_n' \Sigma_n \beta_n}$. Then the counterpart of Assumption 2(b) would be $\rho_n s_{0n} / \left[\phi_1^{\mathrm{RE}}(s_{0n})\right]^2 \to 0$, where $s_{0n} = |S_{0n}|$. Because $\phi_1(S_{0n})$ calculates the minimal-eigenvalue-type quantity only for $\beta \in \mathscr{C}(S_{0n})$ whereas $\phi_1^{\mathrm{RE}}(s_{0n})$ takes the minimum over $\beta \in \bigcup_{\{S_n : |S_n| \le s_{0n}\}} \mathscr{C}(S_n)$, the latter must be no greater than the former. We use $\phi_1(S_{0n})$ in our theory as it better fits the econometric purpose of parameter estimation. To keep $\psi(S_{0n})$ from growing too fast, we demand that the truly relevant explanatory variables are all well instrumented and sufficiently uncorrelated. In other words, Assumption 2(b) does not break down when (i) an unimportant x_{ik} is weakly instrumented, or even not instrumented at all; and (ii) a few unimportant x_{ik}'s are highly correlated, or even perfectly correlated. Since weak instrumentation has been a major concern in the IV theory and practice, Point (i) is particularly desirable in our context.

When the sequence (ρ_n) and the relevance of the instruments are given, whether Assumptions 2(a) and (b) are satisfied hinges on the true parameter β_n^0. We will see in Corollary 12 that t_{0n} and $\rho_n \psi(S_{0n})$ are critical in an explicit convergence rate of GMM-Lasso, and in Remark 13 an interpretation of these two terms follows.

The consistency of the standard GMM may break down when $\phi_2(K_n)$ shrinks to zero too fast, even if $L_n = o(n)$. By considering only the restricted set $\mathscr{B}(S_{0n})$, the concern about the smallest eigenvalue is alleviated.

3.2. Regularity Conditions

In the previous subsection, we discussed the key sparsity condition. Before we present the main results, we impose a few technical regularity conditions. Let $\underline{\phi}_n$ and $\bar{\phi}_n$ be the smallest and largest eigenvalues of the positive-definite symmetric weighting matrix W_n, respectively. W_n can either be random or nonrandom. Let $D_n := \max_{1 \le k \le K_n} \sum_{l=1}^{L_n} \sigma_{lk}^2$, where σ_{lk} is the (l, k)th element of Σ_n.

Assumptions 5. There exist two finite positive constants C_1 and C_2 such that the following statement follow:

(a) $1/C_1 \le \liminf_{n \to \infty} \underline{\phi}_n \le \limsup_{n \to \infty} \bar{\phi}_n \le C_1$ with probability approaching one (w.p.a.1.);
(b) $\limsup_{n \to \infty} D_n \le C_2$.

Remark 6. Assumptions 5(a) and 5(b) require a well-behaved weighting matrix W_n and covariance matrix Σ_n, respectively. If we decide to use a nonrandom W_n as in Han and Phillips (2006), the identity matrix trivially satisfies Condition (a). If W_n is random, for example $W_n = (Z'Z)^{-1}$, Condition (a) becomes a high-level assumption but is not restrictive in view of $L_n = o(n)$. Condition (b) gives an upper bound for the largest eigenvalue of $\Sigma_n'\Sigma_n$ according to the Gershgorin circle theorem, whereas the lower bound is controlled by the compatibility constant $\phi_1(s_n)$ together with Assumption 2.

To handle the randomness of x_i, z_i and ϵ_i, Assumption 7 imposes mild technical restrictions. Let $\xi_{iln} := \frac{1}{\sqrt{n}} z_{il}\epsilon_i$ and $\Delta_{ilkn} := \frac{1}{\sqrt{n}}(x_{ik}z_{il} - \sigma_{lk})$.

Assumptions 7. There exist finite constants C_3, C_4, C_5, and C_6 such that uniformly for all n, we have as follows:

(a) $\inf_{l\le L_n, k\le K_n} \mathbb{E}\left[\left|\sqrt{n}\Delta_{ikln}\right|^2\right] \ge 1/C_3$ and $\sup_{l\le L_n, k\le K_n} \mathbb{E}\left[\left|\sqrt{n}\Delta_{ikln}\right|^3\right] \le C_3$;
(b) $\inf_{l\le L_n} \mathbb{E}\left[\left|\sqrt{n}\xi_{iln}\right|^2\right] \ge 1/C_4$ and $\sup_{l\le L_n} \mathbb{E}\left[\left|\sqrt{n}\xi_{iln}\right|^3\right] \le C_4$;
(c) $\mathbb{P}\left(\bar{V}_{\Delta,n} > C_5\right) \to 0$, where $\bar{V}_{\Delta,n} := \max_{k\le K_n, l\le L_n}\left(\sum_{i=1}^n \Delta_{ikln}^2\right)^{1/2}$;
(d) $\mathbb{P}\left(\bar{V}_{\xi,n} > C_6\right) \to 0$, where $\bar{V}_{\xi,n} := \max_{k\le K_n, l\le L_n}\left(\sum_{i=1}^n \xi_{iln}^2\right)^{1/2}$.

Remark 8. Conditions (a) and (b) restrict the lower bounds of the second moments and the upper bounds of the third moments. Conditions (c) and (d) are high-level assumptions that give probability bounds for the sample variances. They can be established using the uniformly bounded 8th moments as sufficient low-level conditions (Belloni et al., 2012, Lemma 3).

With all the above assumptions, we are ready to proceed to the main theoretical results.

3.3. Main Results

In this section we derive an oracle inequality, which implies the rate of convergence of GMM-Lasso. Define $d_n : \mathbb{R}^{K_n} \mapsto \mathbb{R}^+$ as $d_n(\delta_n) = \delta_n' \frac{X'Z}{n} W_n \frac{Z'X}{n}\delta_n$, and $d : \mathbb{R}^{K_n} \mapsto \mathbb{R}^+$ as $d(\delta_n) = \delta_n'\Sigma_n' W_n \Sigma_n \delta_n$. Both $d_n(\cdot)$ and $d(\cdot)$ are measures of noncentrality; the only difference is the sample covariance matrix $Z'X/n$ versus its population counterpart Σ_n. Let two events be

$$E_1(A_1\lambda_n) := \left\{ \sup_{\delta_n \in \mathbb{R}^{K_n}} \frac{\left|d_n(\delta_n) - d(\delta_n)\right|}{\|\delta_n\|_1^2} \le A_1\lambda_n \right\},$$
$$E_2(A_2\lambda_n) := \left\{ \left\|n^{-2}\xi_n' W_n Z'X\right\|_\infty \le A_2\lambda_n \right\},$$

where A_1 and A_2 are two constants, $\|\cdot\|_\infty$ is the sup-norm of a vector, and $\xi_n := (\xi_{1n}, \ldots, \xi_{nn})'$ is an $n \times L_n$ matrix stacking the L_n-vectors $\xi_{in} = (\xi_{iln})_{i=1}^{L_n}$, $i = 1, \ldots, n$.

Under these two events, the randomness is controlled at the rate λ_n up to some constants. We will show in Lemma 9 that the two events happen with high probability as $n \to \infty$.

Lemma 9. *If Assumptions 1, 2,5, and 7 hold, then there exist finite constants A_1 and A_2 such that (a) $\mathbb{P}[E_1(A_1\lambda_n)] \to 1$, and (b) $\mathbb{P}[E_2(A_2\lambda_n)] \to 1$.*

We are familiar with the closed-form solution of the standard GMM $\widehat{\beta}_n^{(\mathrm{GMM})} = \beta_n^0 + (X'ZW_nZ'X)^{-1} X'ZW_nZ'\epsilon$ if the needed inverses exist. Under the diverging-K_n asymptotic framework, Lemma 9(a) and (b) are the counterparts of the convergence of $X'ZW_nZ'X$ and $X'ZW_nZ'\epsilon$, respectively, along with proper norms that accommodate the growing dimensionalities.

These events are prepared for the key theoretical result Theorem 10, which states the oracle inequality of GMM-Lasso. To construct a bound for the left-hand side of (5), an "oracle" is told about the identity of the large components of β_n^0. The oracle empowers us as if we could optimize the quantity in the squared brackets over the set $\mathscr{B}(S_{0n})$, although the constants there remind us that the oracle is unavailable in reality. The oracle shows that GMM-Lasso is optimal under a certain criterion.

Theorem 10. *If Assumptions 1, 2, 5, and 7 hold, then as n is sufficiently large, we have w.p.a.1.*

$$\begin{aligned} & d_n\left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \left\|\widehat{\beta}_n - \beta_n^0\right\|_1 \\ & \quad \le \min_{\tilde{\beta}_n \in \mathscr{B}(S_{0n})} \left[8 d_n\left(\tilde{\beta}_n - \beta_n^0\right) + \rho_n \left\|\tilde{\beta}_n - \beta_n^0\right\|_1 + 32\rho_n^2 \psi(S_{0n})\right]. \end{aligned} \tag{5}$$

Remark 11. On the left-hand side of (5) lies the discrepancy between $\widehat{\beta}_n$ and β_n^0 measured by $d_n(\cdot) + \rho_n \|\cdot\|_1$. The right-hand side, up to the multipliers, consists of the discrepancy between $\tilde{\beta}_n$ and β_n^0, plus an additional penalty term.

It is easy to verify the consistency of $\widehat{\beta}_n$ to β_n^0 under the l_1-norm. According to Theorem 10, any $\tilde{\beta}_n \in \mathscr{B}(S_{0n})$ delivers an upper bound for the left-hand side of (5). Denote $\beta_n^{(\mathrm{tr})} := \left(\beta_k^0 \cdot 1\left\{\left|\beta_k^0\right| \ge \rho_n\right\}\right)_{k=1}^{K_n}$, the truncated version of β_n^0 that maintains the big (relative to ρ_n) components and suppresses the small ones to zero. Indeed, as it depends purely on β_n^0 and ρ_n but no randomness is involved, $\beta_n^{(\mathrm{tr})}$ is not an estimator. However, $\beta^{(\mathrm{tr})}$ provides, in the proof of Corollary 12, an explicit upper bound of the rate of convergence. The resulting rate is governed by t_{0n} and $\rho_n\psi(S_{0n})$.

Corollary 12. *If Assumptions 1, 2, 5, and 7 are satisfied, then w.p.a.1.*

$$\left\|\widehat{\beta}_n - \beta_n^0\right\|_1 = O_p\left(t_{0n} \vee \rho_n \psi(S_{0n})\right).$$

Remark 13. Remember that $t_{0n} = \sum_{k=1}^{K_n} \left|\beta_k^0\right| 1\left\{\left|\beta_k^0\right| \le \rho_n\right\}$ is the aggregate magnitude of the small (relative to ρ_n) components. It corresponds to the estimation bias. For a given β_n^0, if we choose a large ρ_n, then t_{0n} becomes greater and the estimator is more biased. The second term $\rho_n \psi\left(S_{0n}\right) = \frac{\rho_n\left|S_{0n}\right|}{\phi_1^2\left(S_{0n}\right)}$ corresponds to the estimation variance. Its numerator $\rho_n\left|S_{0n}\right| = \rho_n \sum_{k=1}^{K_n} 1\left\{\left|\beta_k^0\right| > \rho_n\right\}$ can be interpreted as an effective magnitude of the big (relative to ρ_n) components, and the denominator $\phi_1^2\left(S_{0n}\right)$ is the square of the compatibility constant concerning the strength of the IVs. Given a β_n^0, if we increase ρ_n, then $\left|S_{0n}\right|$ drops and so does $1/\phi_1^2\left(S_{0n}\right)$. Furthermore, if $\left|S_{0n}\right|$ shrinks fast enough such that the effective magnitude $\rho_n\left|S_{0n}\right|$ falls, then the estimation variance certainly diminishes. The choice of ρ_n is critical in balancing the bias and variance.

$\beta_n^{(\mathrm{tr})}$ focuses on the large coefficients but ignores all the small components. Nonetheless, (5) holds for any $\tilde{\beta}_n \in \mathscr{B}\left(S_{0n}\right)$, and we may find other candidates in $\mathscr{B}\left(S_{0n}\right)$ that work better than $\beta_n^{(\mathrm{tr})}$. Even though we cannot write in a closed-form the "oracle estimator"

$$\beta_n^* = \arg \min_{\tilde{\beta}_n \in \mathscr{B}(S_{0n})} \left[8 d_n\left(\tilde{\beta}_n - \beta_n^0\right) + \rho_n\left\|\tilde{\beta}_n - \beta_n^0\right\|_1 + 32\rho_n^2 \psi\left(S_{0n}\right)\right],$$

the convergence rate associated with β_n^* must be no slower than $O_p\left(t_{0n} \vee \rho_n \psi\left(S_{0n}\right)\right)$.

4. SIMULATIONS

The rate of convergence is of theoretical importance, whereas GMM-Lasso's finite sample performance depends on the choice of the tuning parameter ρ_n. In practice, when we choose $\rho_n = 0$, GMM-Lasso is exactly the same as the standard GMM. In contrast, when we choose a sufficiently large ρ_n, the penalty term will dominate the GMM quadratic criterion and the resulting estimate can be a vector of all zeros. At the end of Section 2, we introduced the modified GMM-AIC and GMM-BIC as information criteria to help determine the tuning parameter. Following Wang et al. (2009) and Caner and Fan (2010), in the simulation we will use $B_n = \log(\log n)$ to compute the GMM-Lasso estimators.

The least-angle regression algorithm (LARS) (Efron et al., 2004) efficiently calculates the solution path of a linear Lasso-type estimator for different values of ρ_n, starting from the all-zero solution towards the all-nonzero alternatives. A simple modification enables LARS to solve the path for GMM-Lasso. This modified algorithm makes our data-driven estimators computationally feasible.

In the rest of this section, we compare the finite-sample performance of GMM-Lasso-AIC and -BIC with the standard 2SLS[7] and an infeasible estimator. Experiment 1 adopts

[7] Besides 2SLS, the limited information maximum likelihood (LIML) is another popular estimator. We calculated LIML in our simulations. We witness that LIML enjoys the smallest bias compared to all the four other estimators; however its variance is much larger—too large to be displayed together with other estimators on the same graph. This is not surprising, as LIML's finite sample moments do not exist (Phillips, 1983, 1984). LIML's simulation results are available upon request.

the simplest design, with only a few large relevant explanatory variables. The bias-variance trade-off in this case is clear and easy to analyze. This design helps us understand the theoretical results, though the correlation structure is simplistic and unrealistic. To complement Experiment 1, we explore correlated x_{ik}, $k = 1, \ldots, K_n$, in Experiment 2.

4.1. Experiment 1

In our first simulation experiment, the value of the large coefficients is either $b_1 = 1$ or $b_1 = -1$, whereas the small coefficients are either $b_2 = 0$ (exactly sparse) or $b_2 = 0.01$ or -0.01 (approximately sparse). The true parameter in the linear structural equation is

$$\beta_n^0 = (\underbrace{b_1 \cdots b_1}_{K_0} \underbrace{b_2 \cdots b_2}_{K-K_0}).$$

with $K_0 = 10$. The explanatory x_{ki} is generated from

$$x_{ik} = 1 + \gamma_1 \sum_{l=1}^{L_n} 1\{l = k\} z_{il} + \gamma_2 1\{k \leq K_0\} \epsilon_i + \left(1 - |\gamma_1| - |\gamma_2|\right) u_{ik} \tag{6}$$

for $k = 1, \ldots, K_n$, where u_{ik} is a reduced-form idiosyncratic shock. All $\left(z_{ij}\right)_{j=1}^{L_n}$, $(u_{ik})_{k=1}^{K_n}$, and ϵ_i, are independent, and each follows the standard normal distribution. By construction, x_{ik} is correlated with a "supporting IV" z_{il} if $l = k$, and each x_{ik}, $k = 1, \ldots, K_0$ is endogenous due to its correlation with ϵ_i. All the other explanatory variables $\{x_{ik}\}_{k=K_0+1}^{K_n}$ are actually exogenous, but we suppose that the econometrician does not know that and he still uses the supporting IVs. γ_1 and γ_2 are two scalar constants such that $|\gamma_1| + |\gamma_2| < 1$. γ_1 indicates the strength of instrumentation; the greater γ_1 is, the higher the correlation between the endogenous variable and the corresponding instrument is. γ_2 indicates the magnitude of the endogeneity; the larger γ_2 is, the stronger the endogeneity is. We set $\gamma_1 = 0.3$ and $\gamma_2 = 0.3$, and specify $L = 1.1K$ as a slightly overidentified case.

We experiment with combinations of the parameter values (b_1, b_2) and different dimensions. K is either 20, 40, or 80, and n is either 200, 400, or 800. Let r be the index for the replication, and we use $R = 500$ as the total number of replications. For a generic estimate $\check{\beta}_{K_n}$, the aggregate empirical mean squared error (MSE) is defined as $\widehat{\text{MSE}}\left(\check{\beta}_{K_n}\right) = \sum_{k=1}^{K_n}\left[\frac{1}{R}\sum_{r=1}^{R}\left(\check{\beta}_{rk} - \beta_k^0\right)^2\right]$, the aggregate empirical squared-bias as $\sum_{k=1}^{K}\left(\frac{1}{R}\sum_{r=1}^{R}\check{\beta}_{rk} - \beta_k^0\right)^2$ and the aggregate empirical variance as $\sum_{k=1}^{K_n}\left[\frac{1}{R}\sum_{r=1}^{R}\left(\check{\beta}_{rk} - \frac{1}{R}\sum_{r=1}^{R}\check{\beta}_{rk}\right)^2\right]$.

In practice, we cannot easily compute the oracle estimator β_n^* because the set $\mathscr{C}(S_{0n})$ is comprised of too many elements so that we can hardly carry out the minimization in

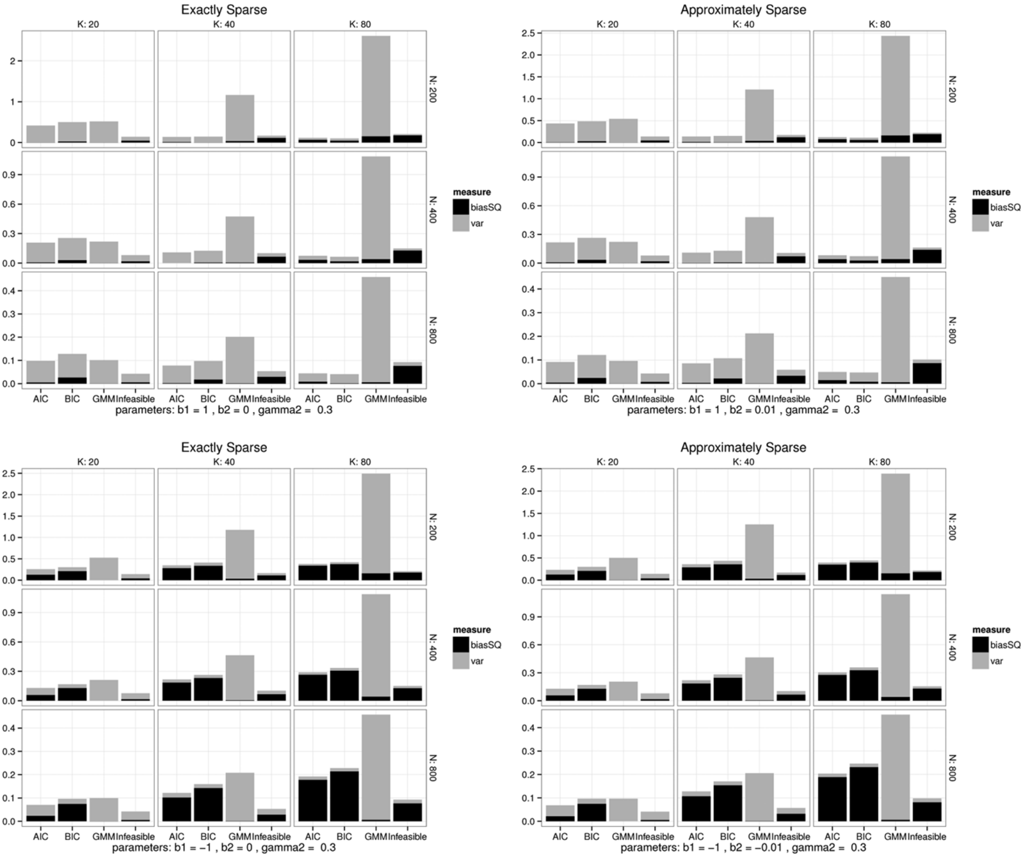

FIGURE 1 Experiment 1: uncorrelated endogenous regressors.

the definition of $\phi_1(S_{0n})$. Thus we introduce an implementable oracle estimator. Let X_{S^*} be the submatrix whose columns correspond to the coordinates of the large coefficients. The infeasible estimator

$$\widehat{\beta}_n = \left(X'_{S^*} P_Z X_{S^*}\right)^{-1} X'_{S^*} P_Z y,$$

is the 2SLS under the oracle $S^* = (1, \ldots, K_0)$. Here we use this implementable oracle estimator as a benchmark.

Figure 1 displays the estimation results. The graph consists of four panels and each panel includes nine subgraphs. The length of the black bar is the aggregate empirical squared-bias, and that of the gray bar is the aggregate empirical variance, so that the total length is the empirical MSE.

We discuss the top-left panel in detail. Along the vertical direction, the bias and variance both decrease as the sample size increases (note that the scales of the y-axes are different). The bias of the infeasible estimator is negligible when $K = 20$, but it deteriorates when $K = 80$, in which case as many as $80 \times 1.1 - 10 = 78$ redundant moment restrictions are involved. Among all the three feasible estimators, 2SLS exhibits small bias but large variance, which make it the worst feasible estimator in terms of MSE in all subgraphs. The automated GMM-Lasso-AIC and -BIC outperform 2SLS.

The other three panels show the results for different true coefficients. The top-right panel is the approximately sparse case with $b_2 = 0.01$ instead of $b_2 = 0$. This change worsens the bias of the infeasible estimator, as it only takes care of the large coefficients but ignores all the small coefficients. The general pattern of the estimators are close to that of the exactly sparse case.

However, changing the value of b_1 and b_2 to be negative in the two bottom panels reveals a dramatic contrast of the bias-variance trade-off of the automated GMM-Lasso. In the top panels their biases are comparable to that of 2SLS, while in the bottom panels their bias dominates the variance. This phenomenon arises from the conflation of the endogeneity bias and the shrinkage bias. Thanks to the simple design of the data generating process (DGP), we can tell the sign of the bias. In the top panels the instruments and the endogenous variables are positively correlated; the bias of the standard 2SLS is positive. In the meantime, the Lasso-type estimators shrink the coefficients towards zero. When the true large coefficients are positive, the shrinkage offsets the positive endogeneity bias. That is, the endogeneity bias and the shrinkage bias are of the opposite direction, so they cancel each other. On the contrary, in the bottom panels the two biases both push the estimator towards zero, so the shrinkage effect amplifies the negative endogeneity bias. In terms of MSE, GMM-Lasso-AIC is still the winner amongst the feasible estimators. 2SLS and the infeasible estimator have no shrinkage effect; their bias-variance ratio is the same in the top and bottom panels.

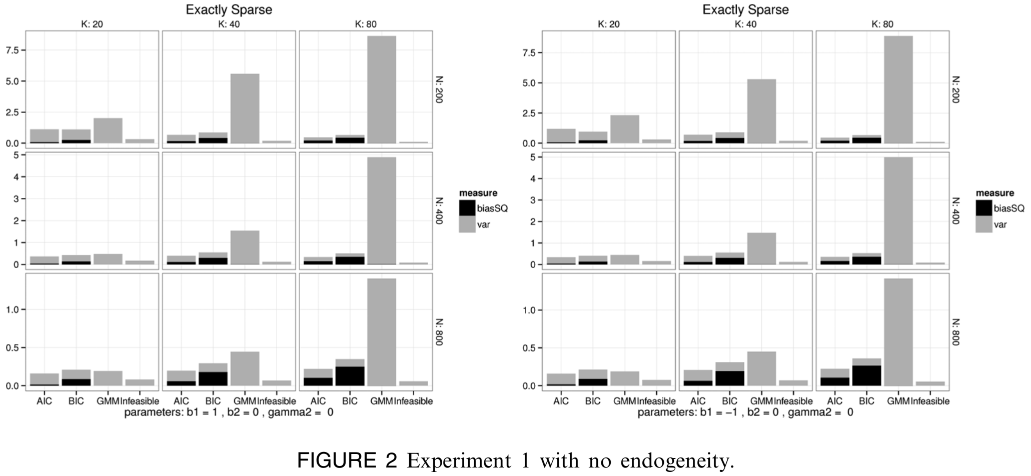

FIGURE 2 Experiment 1 with no endogeneity.

The above reasoning about the contrast of the biases in the top panels and those in the bottom panels predicts that the bias caused by the shrinkage should be symmetric for $b_1 = \pm 1$ if there is no endogeneity bias. To verify this prediction, we set $\gamma_2 = 0$ in (6). Figure 2 confirms the prediction that the infeasible estimator and the standard 2SLS should incur no bias and the bias of GMM-Lasso should be symmetric for $b_1 = \pm 1$.

4.2. Experiment 2

Experiment 2 differs from Experiment 1 in the generation of x_{ki}'s, where

$$x_{ik} = 1 + \gamma_1 \sum_{l=1}^{L_n} 0.5^{\left(|k-l|+1\right)1\{l \le K_n\}} z_{li} + \gamma_2 \epsilon_i + \left(1 - |\gamma_1| - |\gamma_2|\right) u_{ik}.$$

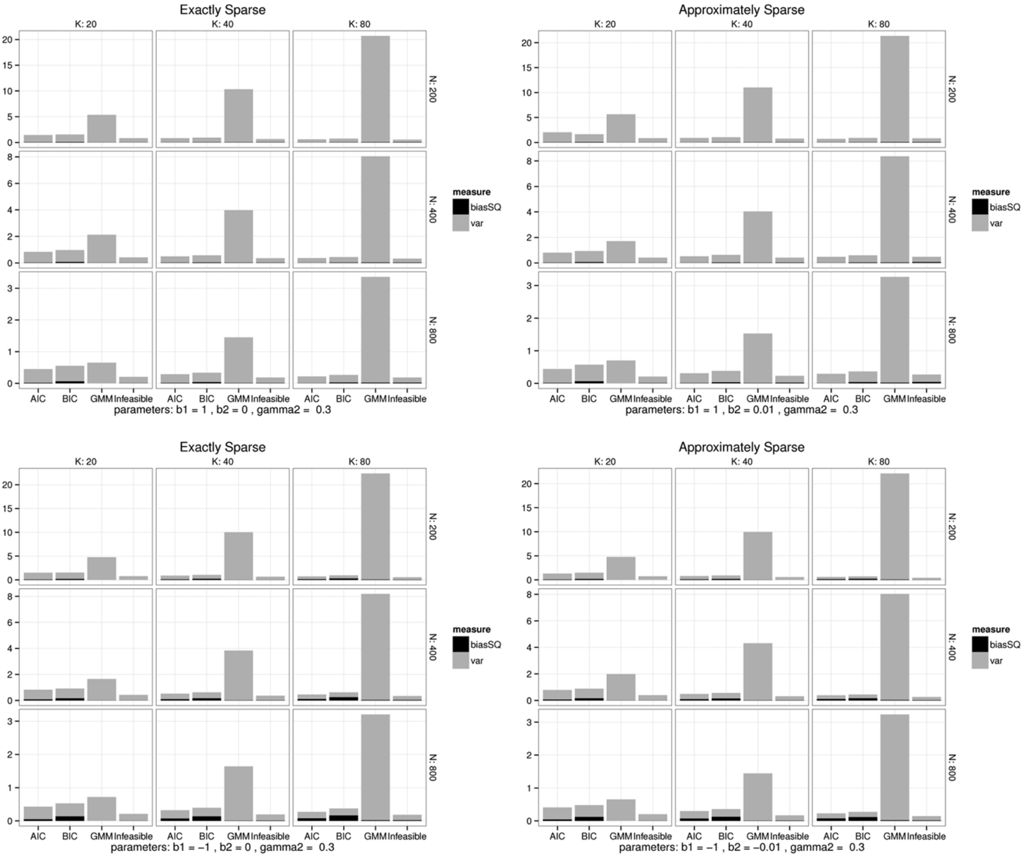

FIGURE 3 Experiment 2: correlated endogenous regressors.

All the x_{ik}'s are affected by the structural error, and x_{ik} is correlated with several or all instruments. The generation of x_{ik}'s induces strong correlation between x_{ik} and $x_{ik'}$ if k and k' are close. This design keeps the instrumentation strength to be γ_1 and the magnitude of endogeneity to be γ_2. The correlation pattern of the z_{ik}'s mimics a more realistic environment.

Figure 3 displays the empirical results. In view of the four panels, we do not see the drastic contrast of the bias-variance trade-off as in Fig. 1. Across different parameter values, the pattern is almost identical. The correlated design is much more complicated than that in Experiment 1, and it is difficult to predict the sign of the bias. The most salient feature is that the aggregate MSE of 2SLS is much larger than the automated GMM-Lasso. Again it confirms the importance of using shrinkage method when the true coefficients are sparse.

5. CONCLUSION

In this article, we analyze GMM-Lasso when the true structural parameter of a linear IV regression is sparse. We establish an oracle inequality, which implies consistency when the true coefficient is sparse enough. The Monte Carlo simulations show that GMM-Lasso-AIC and -BIC perform well in finite sample. Moreover, the simulations illustrate interesting bias-variance trade-off.

In view of GMM-Lasso's significant finite-sample bias in some DGP designs, a bias-correction procedure will be an important extension. Furthermore, it is possible to develop parallel results in more general nonlinear IV problems by invoking a different set of technical tools.

APPENDIX: PROOFS

Under Assumption 5(a), $E_3 = \{1/C_1 \leq W_n \leq C_1\}$ occurs w.p.a.1. To simplify notation, through the proofs we assume the event E_3 holds for all n to avoid writing w.p.a.1. whenever W_n appears. Conditioning on E_3 is innocuous in the asymptotic theory.

A.1. Probabilistic Tools

As discussed in the main text, we take the model as a triangular array of models. The moderate deviation theory of self-normalized sums gives the rate of tail probabilities (Peña et al., 2009, Theorem 7.4). The following Lemma 14 is taken from the proof of Lemma 5 of Belloni et al. (2012, p. 2410). We include it here for completeness.

Lemma 14 (Belloni et al., 2012). *Let $U_{1n}, \ldots, U_{nn}$ be a triangular array of independently non-identically distributed zero-mean random variables. Suppose that*

$$n^{1/6} M_n / \ell_n \geq 1, \text{ where } M_n = \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} U_{in}^2 \right)^{1/2} / \left(\frac{1}{n} \sum_{i=1}^{n} \mathbb{E} \left| U_{in} \right|^3 \right)^{1/3};$$

then uniformly on $0 \leq x \leq n^{1/6} M_n / \ell_n - 1$, the quantities $S_{nn} = \sum_{i=1}^{n} U_{in}$ and $V_{nn}^2 = \sum_{i=1}^{n} U_{in}^2$ obey

$$\left| \frac{\mathbb{P}\left(\left| S_{n,n} / V_{n,n} \right| \geq x \right)}{2\left(1 - \Phi(x)\right)} - 1 \right| \leq \frac{A}{\ell_n^3},$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution and A is a universal constant.

A.2. Proof of Lemma 9

Lemma 9 governs the behavior of the key stochastic terms. We put this proof of Part(b) prior to Part(a) since it illustrates the important technique that bounds the tail probability. This technique will reappear in the proof of Part(a).

Proof of Part(b). Decompose $Z'X/\sqrt{n}$ into its population mean and the deviation $Z'X/\sqrt{n} = \sqrt{n}\Sigma_n + \Delta_n$. Substitute the decomposition into $\xi_n' W_n \frac{Z'X}{\sqrt{n}}$ so that

$$\frac{1}{n} \left\| \xi_n' W_n \frac{Z'X}{\sqrt{n}} \right\|_\infty = \left\| \frac{1}{\sqrt{n}} \xi_n' W_n \Sigma_n + \frac{1}{n} \xi_n' W_n \Delta_n \right\|_\infty \leq \frac{1}{\sqrt{n}} \left\| \xi_n' W_n \Sigma_n \right\|_\infty + \frac{1}{n} \left\| \xi_n' W_n \Delta_n \right\|_\infty .$$

We will bound the two terms on the right-hand side, one by one. Let $\Sigma_{\cdot k}$ be the kth column of Σ_n. The first term

$$\begin{aligned} \frac{1}{\sqrt{n}} \left\| \xi_n' W_n \Sigma_n \right\|_\infty &= \max_{k \leq K_n} \frac{1}{\sqrt{n}} \left| \xi_n' W_n \Sigma_{\cdot k} \right| \\ &\leq \left(\max_{k \leq K_n} \sqrt{\Sigma_{\cdot k}' W_n \Sigma_{\cdot k}} \right) \sqrt{\frac{1}{n} \xi_n' W_n \xi_n} \\ &\leq \sqrt{\bar{\phi}_n D_n} \sqrt{\frac{1}{n} \xi_n' W_n \xi_n} \leq \sqrt{C_1 C_2} \sqrt{\frac{1}{n} \xi_n' W_n \xi_n}, \end{aligned} \tag{7}$$

where the first inequality follows by the Cauchy–Schwarz inequality and the last by Assumption 5. The second factor is stochastic, and it can be bounded by

$$\sqrt{\frac{1}{n}\xi_n' W_n \xi_n} \le \sqrt{\bar{\phi}_n}\sqrt{\frac{1}{n}\xi_n'\xi_n} \le \sqrt{C_1}\sqrt{\frac{1}{n}\sum_{l=1}^{L_n}\xi_{ln}^2} \le \sqrt{C_1}\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\left|\xi_{ln}\right|. \tag{8}$$

The technique of the tail probability calculation in the next inequality will be used repeatedly in the proofs of this article. Thus we write down here every step in detail, and shorten the intermediate steps when we reuse it later. For a generic constant $a > 1$, by (7) and (8),

$$\mathbb{P}\left(\frac{1}{\sqrt{n}}\left\|\xi_n' W_n \Sigma_n\right\|_\infty \ge aC_1\sqrt{C_2}C_6\lambda_n\right) \le \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n} \ge aC_6\lambda_n\right).$$

Let $\xi_{ln} := \frac{1}{\sqrt{n}}\sum_{i=1}^n z_{il}\epsilon_i$. We self-normalize the term of interest, so that

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\left|\xi_{ln}\right| \ge aC_4\lambda_n\right)\\
&\quad\le \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^n \xi_{iln}^2}} \ge a\frac{C_6}{\bar{V}_{\xi,n}}\lambda_n\right)\\
&\quad= \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^n \xi_{iln}^2}} \ge a\frac{C_6}{\bar{V}_{\xi,n}}\lambda_n \middle| \frac{C_6}{\bar{V}_{\xi,n}} \ge 1\right)\mathbb{P}\left(\frac{C_6}{\bar{V}_{\xi,n}} \ge 1\right)\\
&\qquad+ \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^n \xi_{iln}^2}} \ge a\frac{C_6}{\bar{V}_{\xi,n}}\lambda_n \middle| \frac{C_6}{\bar{V}_{\xi,n}} < 1\right)\mathbb{P}\left(\frac{C_6}{\bar{V}_{\xi,n}} < 1\right)\\
&\quad\le \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^n \xi_{iln}^2}} \ge a\lambda_n\right)\mathbb{P}\left(\frac{C_6}{\bar{V}_{\xi,n}} \ge 1\right) + \mathbb{P}\left(\frac{C_6}{\bar{V}_{\xi,n}} < 1\right)\\
&\quad\le \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^n \xi_{iln}^2}} \ge a\lambda_n\right) + \mathbb{P}\left(\frac{C_6}{\bar{V}_{\xi,n}} < 1\right).
\end{aligned} \tag{9}$$

The next inequality is the key step in which we apply the moderate deviation theory for self-normalized sums. According to the Bonferroni inequality, the probability of the union of L_n events can be bounded by

$$
\begin{aligned}
&\mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\le L_n}\left(\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^{n}\xi_{iln}^{2}}}\right)\ge a\lambda_n\right)\\
&\le\sum_{l=1}^{L_n}\max_{l\le L_n}\mathbb{P}\left(\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^{n}\xi_{iln}^{2}}}\ge a\sqrt{\log K_nL_n}\right)\\
&\le 2L_n\Phi\left(-a\sqrt{\log K_nL_n}\right)(1+o(1))\\
&\le 4L_n\exp\left(-\frac{1}{2}a^2\log K_nL_n\right)(1+o(1)),
\end{aligned}
\tag{10}
$$

where the last inequality follows by the fact $1-\Phi(x)\le\frac{2}{1+x}\exp\left(-\frac{1}{2}x^2\right)$ for $x>0$. Now we explain how to use Lemma 14 to verify the second inequality of (1). For any l, we can replace $U_{i,n}$ in Lemma 14 by ξ_{iln}, so that $S_{n,n}=\xi_{ln}$ and $V_{n,n}=\left(\sum_{i=1}^{n}\xi_{iln}^{2}\right)^{1/2}$. Under Assumption 7, $M_{ln}:=\left(\mathbb{E}\left[\xi_{iln}^{2}\right]\right)^{1/2}/\left(\mathbb{E}\left[\xi_{iln}^{3}\right]\right)^{1/3}$ is a finite constant. When we replace ℓ_n in Lemma 14 by $a\log K_nL_n$, under Assumption 1 the condition $n^{1/6}M_{ln}/\ell_n\ge 1$ holds for a large n. This argument is valid uniformly for all $l\le L_n$, so that

$$
\max_{l\le L_n}\left|\mathbb{P}\left(\frac{\left|\xi_{ln}\right|}{\sqrt{\sum_{i=1}^{n}\xi_{iln}^{2}}}\ge a\sqrt{\log K_nL_n}\right)\Big/2\left(1-\Phi\left(a\sqrt{\log K_nL_n}\right)\right)-1\right|\le\frac{A}{(\log K_nL_n)}
$$

for n sufficiently large.

From the above derivation, for a generic constant a_1, by (9) and (10) we have

$$
\begin{aligned}
&\mathbb{P}\left(\frac{1}{\sqrt{n}}\left\|\xi_n'W_n\Sigma_n\right\|_\infty\ge a_1\lambda_n\right)\\
&=\mathbb{P}\left(\frac{1}{\sqrt{n}}\left\|\xi_n'W_n\Sigma_n\right\|_\infty\ge\left(\frac{a_1}{C_1\sqrt{C_2}C_6}\right)C_1\sqrt{C_2}C_6\lambda_n\right)\\
&\le 4L_n\exp\left(-\frac{1}{2}\left(\frac{a_1}{C_1\sqrt{C_2}C_6}\right)^2\log K_nL_n\right)(1+o(1))+\mathbb{P}\left(\frac{C_6}{\bar{V}_{\xi,n}}<1\right)\\
&\le 4\exp\left(-\left(\frac{a_1}{C_1\sqrt{C_2}C_6}\right)^2\right)+o(1),
\end{aligned}
\tag{11}
$$

where a in (10) is replaced by $a_1/\left(C_1\sqrt{C_2}C_6\right)$, and $\mathbb{P}\left(\bar{V}_{\xi,n} \geq C_6\right) = o\left(1\right)$ as assumed in Assumption 5. For any fixed small constant $\eta > 0$, there exists a finite a_1 such that $4\exp\left(-\left(\frac{a_1}{C_1\sqrt{C_2}C_6}\right)^2\right) < \eta$.

We move onto the second term $\frac{1}{n}\left\|\xi_n' W_n \Delta_n\right\|_\infty = \max_{k\leq K_n}\frac{1}{n}\left|\xi_n' W_n \Delta_{\cdot k}\right|$, where $\Delta_{\cdot k}$ is the kth column of Δ_n. The Cauchy–Schwarz inequality gives $\frac{1}{n}\xi_n' W_n \Delta_{\cdot k} \leq \sqrt{\frac{1}{n}\xi_n' W_n \xi_n}\sqrt{\frac{1}{n}\Delta_{\cdot k}' W_n \Delta_{\cdot k}}$. Let $\Delta_{lkn} := \frac{1}{\sqrt{n}}\sum_{i=1}^n \left(x_{ik}z_{il} - \sigma_{lk}\right)$:

$$\begin{aligned}
&\mathbb{P}\left(\sqrt{\Delta_{\cdot k}' W_n \Delta_{\cdot k}/n} \geq a\sqrt{C_1}C_5\lambda_n\right)\\
&\leq \mathbb{P}\left(\sqrt{\frac{L_n}{n}}\max_{l\leq L_n}\frac{\left|\Delta_{lkn}\right|}{\sqrt{\sum_{i=1}^n \Delta_{ikl}^2}} \geq a\lambda_n\frac{C_5}{\bar{V}_{\Delta,n}}\right)\\
&\leq 4L_n\exp\left(-\frac{1}{2}a^2\log K_nL_n\right) + \mathbb{P}\left(\frac{C_5}{\bar{V}_{\Delta,n}} < 1\right).
\end{aligned} \tag{12}$$

The reasoning is the same as that of (11), and the conditions needed to invoke Lemma 14 are assumed in Assumption 7. Thus for a generic a_2,

$$\begin{aligned}
&\mathbb{P}\left(\frac{1}{n}\left|\xi_n' W_n \Delta_{\cdot k}\right| \geq a_2\lambda_n^2\right)\\
&\leq \mathbb{P}\left(\sqrt{\xi_n' W_n \xi_n/n}\sqrt{\Delta_{\cdot k}' W_n \Delta_{\cdot k}/n} \geq a_2\lambda_n^2\right)\\
&\leq \mathbb{P}\left(\sqrt{\xi_n' W_n \xi_n/n} \vee \sqrt{\Delta_{\cdot k}' W \Delta_{\cdot k}/n} \geq \sqrt{a_2}\lambda_n\right)\\
&\leq \mathbb{P}\left(\sqrt{\xi_n' W_n \xi_n/n} \geq \sqrt{a_2}\lambda_n\right) + \mathbb{P}\left(\sqrt{\Delta_{\cdot k}' W_n \Delta_{\cdot k}/n} \geq \sqrt{a_2}\lambda_n\right)\\
&\leq \mathbb{P}\left(\sqrt{\frac{1}{n}\xi_n' W_n \xi_n} \geq \frac{\sqrt{a_2}}{C_6\sqrt{C_1}}C_6\sqrt{C_1}\lambda_n\right) + \mathbb{P}\left(\sqrt{\frac{1}{n}\Delta_{\cdot k}' W \Delta_{\cdot k}} \geq \frac{\sqrt{a_2}}{C_5\sqrt{C_1}}C_5\sqrt{C_1}\lambda_n\right)\\
&\leq 4L_n\left[\exp\left(-\frac{a_2}{2C_6^2C_1}\log K_nL_n\right) + \exp\left(-\frac{a_2}{2C_5^2C_1}\log K_nL_n\right)\right] + \mathbb{P}\left(\frac{C_5}{\bar{V}_{\Delta,n}} < 1\right)\\
&\leq 8L_n\exp\left(-\frac{a_2}{2C_1\left(C_5^2 \wedge C_6^2\right)}\log K_nL_n\right) + \mathbb{P}\left(\frac{C_5}{\bar{V}_{\Delta,n}} < 1\right),
\end{aligned}$$

where the fourth inequality follows by the same argument 12. The above inequality applies to each $\frac{1}{n}\left|\xi_n' W_n \Delta_{\cdot k}\right|$. To obtain the uniform bound over $k \leq K_n$, we invoke the

Bonferroni inequality once again, this time for the union of the K_n events,

$$\begin{aligned}\mathbb{P}\left(\frac{1}{n}\left\|\xi_n' W_n \Delta_n\right\|_\infty \geq a_2\lambda_0^2\right) &\leq \sum_{k=1}^{K_n}\mathbb{P}\left(\frac{1}{n}\left|\xi_n' W_n\Delta_{\cdot k}\right| \geq a_2\lambda_n^2\right) + \mathbb{P}\left(\frac{C_5}{\bar{V}_{\Delta,n}} < 1\right)\\ &\leq 8K_nL_n\exp\left(-\frac{a_2}{2C_1\left(C_5^2\wedge C_6^2\right)}\log K_nL_n\right) + \mathbb{P}\left(\frac{C_5}{\bar{V}_{\Delta,n}} < 1\right)\\ &= 8\exp\left(-\frac{a_2}{2C_1\left(C_5^2\wedge C_6^2\right)}\right) + o\left(1\right). \qquad (13)\end{aligned}$$

We conclude from (11) and (13) the statement of the lemma. □

The basic technique in the proof Part(a) is similar. We use the Bonferroni inequality to bound the random variables uniformly, while the exponential tail of the normal distribution is powerful enough to shrink to zero the probability of the undesirable events.

Proof of Part(a). For any nonzero $K_n \times 1$ vector $\delta_n \in \mathbb{R}^{K_n}$, the discrepancy between the empirical noncentrality measure d_n and its population counterpart d is

$$\begin{aligned}\left|d_n\left(\delta_n\right) - d\left(\delta_n\right)\right| &= \delta_n'\left[\left(\frac{X'Z}{n}W_n\frac{Z'X}{n}\right) - \Sigma_n' W_n\Sigma_n\right]\delta_n\\ &= \delta_n'\left[\left(\Sigma_n + \Delta_n/\sqrt{n}\right)' W_n\left(\Sigma_n + \Delta_n/\sqrt{n}\right) - \Sigma_n' W_n\Sigma_n\right]\delta_n\\ &= \delta_n'\left(\Delta_n' W_n\Sigma_n/\sqrt{n} + \Sigma_n' W_n\Delta_n/\sqrt{n} + \Delta_n' W_n\Delta_n/n\right)\delta_n\\ &\leq \left\|\delta_n\right\|_1^2\left(2\left\|\Sigma_n' W_n\Delta_n/\sqrt{n}\right\|_\infty + \left\|\Delta_n' W_n\Delta_n/n\right\|_\infty\right).\end{aligned}$$

$\left|d_n\left(\delta_n\right) - d\left(\delta_n\right)\right| / \left\|\delta_n\right\|_1^2$ is bounded by a zero-mean term and a quadratic term. According to (12), for each $k \leq K_n$,

$$\mathbb{P}\left(\sqrt{\Delta_{\cdot k}' W_n\Delta_{\cdot k}/n} \geq a\sqrt{C_1}C_5\lambda_n\right) \leq 4L_n\exp\left(-\frac{1}{2}a^2\log K_nL_n\right) + o\left(1\right)$$

and then

$$\begin{aligned}&\mathbb{P}\left(\left\|\Delta_n' W_n\Delta_n/n\right\|_\infty \geq a_1C_5^2\lambda_n^2\right)\\ &\quad\leq \sum_{k=1}^{K_n}\mathbb{P}\left(\sqrt{\Delta_{\cdot k}' W_n\Delta_{\cdot k}/n} \geq \sqrt{a_1}C_5\lambda_n\right)\end{aligned}$$

$$
\begin{aligned}
&\le \sum_{k=1}^{K_n} \mathbb{P}\left(\sqrt{\Delta'_{\cdot k} W_n \Delta_{\cdot k}/n} \ge \frac{\sqrt{a_1}}{\sqrt{C_1}}\sqrt{C_1}C_5\lambda_n\right)\\
&\le 4K_nL_n \exp\left(-\frac{a_1}{2C_1}\log K_nL_n\right) = 4\exp\left(-\frac{a_1}{2C_1}\right),
\end{aligned}
\tag{14}
$$

where we only need to take care of the diagonal elements of $\Delta'_n W_n \Delta_n/n$ because it is of the quadratic form so that the maximal element must be on the diagonal, according to the Cauchy–Schwarz inequality.

Next we move back to the first term. The Cauchy–Schwarz inequality gives

$$
\begin{aligned}
\left\|\Sigma'_n W_n \Delta_n/\sqrt{n}\right\|_\infty &= \max_{k,k'\le K_n}\left|\Sigma'_{\cdot k} W_n \Delta_{\cdot k'}\right|/\sqrt{n}\\
&\le \max_{k'\le K_n}\sqrt{\Sigma'_{\cdot k} W_n \Sigma_{\cdot k}}\max_{k'\le K_n}\sqrt{\Delta'_{\cdot k'} W_n \Delta_{\cdot k'}/n}\\
&\le \sqrt{C_1C_2}\max_k \sqrt{\Delta'_{\cdot k} W_n \Delta_{\cdot k}/n},
\end{aligned}
\tag{15}
$$

Therefore, we bound the probability as

$$
\begin{aligned}
&\mathbb{P}\left(\left\|\Sigma'_n W_n \Delta_n\right\|_\infty \ge \sqrt{a_2}\lambda_n\right)\\
&\quad\le \mathbb{P}\left(\max_{k\le K_n}\Delta'_{\cdot k} W_n \Delta_{\cdot k}/n \ge \frac{\sqrt{a_2}}{\sqrt{C_1C_2}}\sqrt{C_1}C_2\lambda_n\right) + o(1)\\
&\quad\le 4K_nL_n\exp\left(-\frac{a_2}{C_1C_2^2}\log K_nL_n\right) + o(1) = 4\exp\left(-\frac{a_2}{C_1C_2^2}\right) + o(1).
\end{aligned}
\tag{16}
$$

Combining (14) and (16), we establish Part (a). □

A.3. Proof of Theorem 10

Under the assumptions in the statement, we have w.p.a.1. the events $E_1(A_1\lambda_n)$ and $E_2(A_2\lambda_n)$. Conditional on these events, we conduct deterministic calculation. Though the essence of the proof is similar to that of a linear regression under a fixed design (Bühlmann and van de Geer, 2011, Theorem 6.2), there are much more to handle as we work with a random design in the GMM context. Let $\rho_n = 2r_nA_2\lambda_n$ where $(r_n \ge 4)_{n\in\mathbb{N}}$ is a deterministic sequence.[8]

Proof. As $\widehat{\beta}_n$ minimizes the penalized GMM criterion,

$$
Q_n\left(\widehat{\beta}_n\right) + \rho_n\left\|\widehat{\beta}_n\right\|_1 \le Q_n\left(\tilde{\beta}_n\right) + \rho_n\left\|\tilde{\beta}_n\right\|_1 \tag{17}
$$

[8] Here it is clear that $C_\rho = 8A_2$ in Assumption 2.

for any $\tilde{\beta}_n$. The quadratic form of the GMM criterion admits a decomposition

$$Q_n\left(\widehat{\beta}_n\right) = n^{-2}\xi_n' W_n \xi_n + 2n^{-2}\xi_n' W_n Z_n' X_n\left(\beta_n^0 - \widehat{\beta}_n\right) + d_n\left(\widehat{\beta}_n - \beta_n^0\right),$$
$$Q_n\left(\tilde{\beta}_n\right) = n^{-2}\xi_n' W_n \xi_n + 2n^{-2}\xi_n' W_n Z_n' X_n\left(\beta_n^0 - \tilde{\beta}_n\right) + d_n\left(\tilde{\beta}_n - \beta_n^0\right).$$

Substituting these two equalities into (17) and rearranging give

$$d_n\left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \left\|\widehat{\beta}_n\right\|_1 \leq 2n^{-2}\xi_n' W_n Z_n' X_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right) + \rho_n \left\|\tilde{\beta}_n\right\|_1 + d_n\left(\tilde{\beta}_n - \beta_n^0\right).$$

In the above equation, the first term on the right-hand side can be bounded by

$$\begin{aligned} 2n^{-2}\left|\xi_n' W_n Z_n' X_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right)\right| &\leq 2\left\|n^{-2}\xi_n' W_n Z_n' X_n\right\|_\infty \left\|\widehat{\beta}_n - \tilde{\beta}_n\right\|_1 \\ &\leq 2A_2\lambda_n \left\|\widehat{\beta}_n - \tilde{\beta}_n\right\|_1 = \frac{\rho_n}{r_n}\left\|\widehat{\beta}_n - \tilde{\beta}_n\right\|_1, \end{aligned}$$

where the first and the second inequality follow by the definitions of $E_2\left(A_2\lambda_n\right)$ and ρ_n, respectively. Combine the previous two inequalities,

$$d_n\left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \left\|\widehat{\beta}_n\right\|_1 \leq \frac{\rho_n}{r_n}\left\|\widehat{\beta}_n - \tilde{\beta}_n\right\|_1 + \rho_n\left\|\tilde{\beta}_n\right\|_1 + d_n\left(\tilde{\beta}_n - \beta_n^0\right). \tag{18}$$

Let S be a generic set such that $\tilde{\beta}_{S^c} = \mathbf{0}_{K_n}$ and $\left|S\right| \leq s_{0n}$. Elementary calculation gives

$$\left\|\widehat{\beta}_n - \tilde{\beta}_n\right\|_1 = \left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 + \left\|\widehat{\beta}_{S^c} - \tilde{\beta}_{S^c}\right\|_1 = \left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 + \left\|\widehat{\beta}_{S^c}\right\|_1$$

and

$$\left\|\widehat{\beta}_n\right\|_1 = \left\|\widehat{\beta}_S\right\|_1 + \left\|\widehat{\beta}_{S^c}\right\|_1 \geq \left\|\tilde{\beta}_S\right\|_1 - \left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 + \left\|\widehat{\beta}_{S^c}\right\|_1.$$

Substituting them into (18) and rearranging gives

$$d_n\left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n\frac{r_n - 1}{r_n}\left\|\widehat{\beta}_{S^c}\right\|_1 \leq \rho_n\frac{r_n + 1}{r_n}\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 + d_n\left(\tilde{\beta}_n - \beta_n^0\right). \tag{19}$$

We proceed to two possible cases as follows:

$$\text{Case (i) } \rho_n\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 \geq d_n\left(\tilde{\beta}_n - \beta_n^0\right), \tag{20}$$

$$\text{Case (ii) } \rho_n\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 \quad < d_n\left(\tilde{\beta}_n - \beta_n^0\right). \tag{21}$$

In Case (i), (19) becomes

$$d_n\left(\widehat{\beta}_n - \beta_n\right) + \rho_n \frac{r_n - 1}{r_n}\left\|\widehat{\beta}_{S^c}\right\|_1 \leq \rho_n \frac{2r_n + 1}{r_n}\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1, \tag{22}$$

which implies

$$\left\|\widehat{\beta}_{S^c}\right\|_1 \leq \frac{2r_n + 1}{r_n - 1}\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 \leq 3\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1$$

where the second inequality holds as $r_n \geq 4$. The fact $\tilde{\beta}_{S^c} = \mathbf{0}$ indicates

$$\left\|\widehat{\beta}_{S^c} - \tilde{\beta}_{S^c}\right\|_1 = \left\|\widehat{\beta}_{S^c}\right\|_1 \leq 3\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1; \tag{23}$$

therefore, $\left(\widehat{\beta}_n - \tilde{\beta}_n\right) \in \mathscr{C}\left(s_{0n}\right)$. Using the definition of the compatibility constant $\phi_1\left(s_{0n}\right)$ gives

$$\begin{aligned}
\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 &\leq \sqrt{\psi\left(s_{0n}\right) d\left(\widehat{\beta}_n - \tilde{\beta}_n\right)} \\
&\leq \sqrt{\psi\left(s_{0n}\right)\left[d_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right) + \left|d\left(\widehat{\beta}_n - \tilde{\beta}_n\right) - d_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right)\right|\right]} \\
&\leq \sqrt{\psi\left(s_{0n}\right) d_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right)} + \sqrt{\psi\left(s_{0n}\right)\left|d\left(\widehat{\beta}_n - \tilde{\beta}_n\right) - d_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right)\right|} \\
&\leq \sqrt{\psi\left(s_{0n}\right) d_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right)} + \sqrt{A_1 \lambda_n \psi\left(s_{0n}\right)}\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1,
\end{aligned}$$

where the second inequality is a triangle inequality, the third follows by $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and the last one holds by $E_1\left(A_1\lambda_n\right)$. Under Assumption 2, when n is sufficiently large we have $1 - \sqrt{A_1\lambda_n\psi\left(s_{0n}\right)} \geq \sqrt{3}/2$. As a result,

$$\begin{aligned}
\rho_n\left\|\widehat{\beta}_S - \tilde{\beta}_S\right\|_1 &\leq \frac{2}{\sqrt{3}}\rho_n\sqrt{\psi\left(s_{0n}\right) d_n\left(\widehat{\beta}_n - \tilde{\beta}_n\right)} \\
&\leq \frac{2}{\sqrt{3}}\rho_n\sqrt{\psi\left(s_{0n}\right) d_n\left(\widehat{\beta}_n - \beta_n^0\right)} + \frac{2}{\sqrt{3}}\rho_n\sqrt{\psi\left(s_{0n}\right) d_n\left(\tilde{\beta}_n - \beta_n^0\right)} \\
&\leq 2\rho_n^2\psi\left(s_{0n}\right) + \frac{2}{9}d_n\left(\widehat{\beta}_n - \beta_n^0\right) + \frac{2}{3}\rho_n^2\psi\left(s_{0n}\right) + \frac{2}{3}d_n\left(\tilde{\beta}_n - \beta_n^0\right) \\
&= \frac{8}{3}\rho_n^2\psi\left(s_{0n}\right) + \frac{2}{9}d_n\left(\widehat{\beta}_n - \beta_n^0\right) + \frac{2}{3}d_n\left(\tilde{\beta}_n - \beta_n^0\right),
\end{aligned} \tag{24}$$

where the second inequality is a triangle inequality, and the third one holds in view of the trivial inequalities $ab \le \frac{3}{2}a^2 + \frac{1}{6}b^2$ and $ab \le \frac{1}{2}a^2 + \frac{1}{2}b^2$. Add $\rho_n \frac{r_n - 1}{r_n} \left\| \widehat{\beta}_S - \tilde{\beta}_S \right\|_1$ to both sides of (22), so we have

$$\begin{aligned} & d_n \left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \frac{r_n - 1}{r_n} \left\| \widehat{\beta}_n - \tilde{\beta}_n \right\|_1 \\ & \quad \le 3\rho_n \left\| \widehat{\beta}_S - \tilde{\beta}_S \right\|_1 \le 8\rho_n^2 \psi\left(s_{0n}\right) + \frac{2}{3} d_n \left(\widehat{\beta}_n - \beta_n^0\right) + 2d_n \left(\tilde{\beta}_n - \beta_n^0\right), \end{aligned} \tag{25}$$

where the second inequality holds by (24). Multiplying 3 and subtract $2d_n\left(\widehat{\beta}_n - \beta_n^0\right)$ on both sides of the above inequality, we have

$$24\rho_n^2 \psi\left(s_{0n}\right) + 6d_n \left(\tilde{\beta}_n - \beta_n^0\right) \ge d_n \left(\widehat{\beta}_n - \beta_n^0\right) + 3\rho_n \frac{r_n - 1}{r_n} \left\| \widehat{\beta}_n - \tilde{\beta}_n \right\|_1 . \tag{26}$$

The calculation in Case (ii) is more straightforward. Again, we add $\rho_n \frac{r_n - 1}{r_n} \left\| \widehat{\beta}_S - \tilde{\beta}_S \right\|_1$ to both sides of (19), so that the left-hand side is the same as that in (25) but the right-hand side is $2\rho_n \left\| \widehat{\beta}_S - \tilde{\beta}_S \right\|_1 + d_n \left(\tilde{\beta}_n - \beta_n^0\right)$, which is bounded by $3d_n\left(\tilde{\beta}_n - \beta_n^0\right)$ according to (21). Thus we obtain

$$3d_n \left(\tilde{\beta}_n - \beta_n^0\right) \ge d_n \left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \frac{r_n - 1}{r_n} \left\| \widehat{\beta}_n - \tilde{\beta}_n \right\|_1 . \tag{27}$$

Combine (26) and (27), we have

$$\begin{aligned} d_n \left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \left\| \widehat{\beta}_n - \tilde{\beta}_n \right\|_1 & \le \frac{r_n}{r_n - 1} \left(24\rho_n^2 \psi\left(s_{0n}\right) + 6d_n \left(\tilde{\beta}_n - \beta_n^0\right)\right) \\ & \le 32\rho_n^2 \psi\left(s_{0n}\right) + 8d_n \left(\tilde{\beta}_n - \beta_n^0\right). \end{aligned}$$

Furthermore, we use the triangle inequality to conclude

$$d_n \left(\widehat{\beta}_n - \beta_n^0\right) + \rho_n \left\| \widehat{\beta}_n - \beta_n^0 \right\|_1 \le 32\rho_n^2 \psi\left(s_{0n}\right) + 8d_n \left(\tilde{\beta}_n - \beta_n^0\right) + \rho_n \left\| \tilde{\beta}_n - \beta_n^0 \right\|_1 .$$

This completes the proof. □

A.4. Proof of Corollary 12

Proof. Because $\beta_n^{(\mathrm{tr})} \in \mathscr{B}\left(S_{0n}\right)$, according to the oracle inequality (5), we have

$$\rho_n \left\| \widehat{\beta}_n - \beta_n^0 \right\|_1 \le 8d_n \left(\beta_n^{(\mathrm{tr})} - \beta_n^0\right) + \rho_n \left\| \beta_n^{(\mathrm{tr})} - \beta_n^0 \right\|_1 + 32\rho_n^2 \psi\left(S_{0n}\right) \tag{28}$$

in which the first term on the left-hand side of (5) is removed. By the definition of $\beta_n^{(\mathrm{tr})}$, it differs from β_n only on the small coordinates. Note that under $E_1\left(A_1\lambda_n\right)$

$$\begin{aligned} d_n\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right) &\leq d\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right)+\left|d_n\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right)-d\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right)\right| \\ &\leq d\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right)+A_1\lambda_n\left\|\beta_n^{(\mathrm{tr})}-\beta_n^0\right\|_1 \\ &\leq d\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right)+\rho_n\left\|\beta_n^{(\mathrm{tr})}-\beta_n^0\right\|_1. \end{aligned} \tag{29}$$

Moreover, given Assumption 5, we obtain

$$\begin{aligned} d\left(\beta_n^{(\mathrm{tr})}-\beta_n^0\right) &\leq C_1C_2\left\|\beta_n^{(\mathrm{tr})}-\beta_n^0\right\|_2^2=C_1C_2\sum_{k=1}^{K_n}\left|\beta_k^0\right|^2 1\left\{\beta_k^0\leq\rho_n\right\} \\ &\leq C_1C_2\rho_n\sum_{k=1}^{K_n}\left|\beta_k^0\right| 1\left\{\beta_k^0\leq\rho_n\right\}=O\left(\rho_n t_{0n}\right). \end{aligned} \tag{30}$$

The conclusion follows by inserting (29) and (30) into (28) and dividing ρ_n on both sides. □

ACKNOWLEDGMENTS

I thank my Ph.D. advisor Peter Phillips for his guidance and encouragement. I thank Timothy Armstrong, Donald Andrews, Richard Blundell, Xiaohong Chen, Hiroaki Kaido, Yuichi Kitamura, and Taisuke Otsu for their helpful comments. I thank two anonymous referees, whose points and suggestions greatly improve this article. All remaining errors are mine.

REFERENCES

Ackerberg, D., Benkard, C. L., Berry, S., Pakes, A. (2007). Econometric tools for analyzing market outcomes. *Handbook of Econometrics* 6:4171–4276.

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

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. (2011). l1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics* 39(1):82–130.

Berry, S., Levinsohn, J., Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica* 63(4): 841–890.

Bickel, P., Ritov, Y., Tsybakov, A. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics* 37(4):1705–1732.

Blundell, R., Pashardes, P., Weber, G. (1993). What do we learn about consumer demand patterns from micro data? *The American Economic Review* 83:570–597.

Bühlmann, P., van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer.

Candes, E., Tao, T. (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics* 35(6):2313–2351.

Caner, M. (2009). Lasso-type GMM estimator. *Econometric Theory* 25(01):270–290.

Caner, M., Fan, Q. (2015). Hybrid generalized empirical likelihood estimators: Instrument selection with adaptive lasso. *Journal of Econometrics* 187(1):256–274.

Caner, M., Zhang, H. (2014). Adaptive elastic net gmm with diverging number of moments. *Journal of Business and Economics Statistics* 32:30–47.

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

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics* 32(2):407–499.

Fan, J., Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96(456):1348–1360.

Fan, J., Liao, Y. (2014). Endogeneity in high dimensions. *The Annals of Statistics* 42(3):872–917.

Gautier, E., Tsybakov, A. (2013). Pivotal uniform inference in high-dimensional regression with random design in wide classes of models via linear programming. Technical report, CREST, ENSAE.

Han, C., Phillips, P. (2006). Gmm with many moment conditions. *Econometrica* 74(1):147–192.

Hansen, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society* 50:1029–1054.

Koenker, R., Machado, J. (1999). Gmm inference when the number of moment conditions is large. *Journal of Econometrics* 93(2):327–344.

Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory* 29(5):857–904.

Peña, V., Lai, T., Shao, Q. (2009). *Self-normalized Processes: Limit Theory and Statistical Applications.* Berlin Heidelberg: Springer-Verlag.

Phillips, P. (1983). Exact small sample theory in the simultaneous equations model. *Handbook of Econometrics* 1:449–516.

Phillips, P. (1984). The exact distribution of exogenous variable coefficient estimators. *Journal of Econometrics* 26(3):387–398.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* 58:267–288.

Tikhonov, A. N. (1943). On the stability of inverse problems. In: *Doklady Akademil Nauk SSSR*, volume 39, pp. 195–198.

Wang, H., Li, B., Leng, C. (2009). Shrinkage tuning parameter selection with a diverging number of parameters. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71(3):671–683.

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