Published in final edited form as:

Ann Stat. 2014 June 1; 42(3): 872–917. doi:10.1214/13-AOS1202.

Endogeneity in High Dimensions

Jianqing Fan and
Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544

Yuan Liao
Department of Mathematics, University of Maryland, College Park, MD 20742

Jianqing Fan: jqfan@princeton.edu; Yuan Liao: yuanliao@umd.edu

Abstract

Most papers on high-dimensional statistics are based on the assumption that none of the regressors are correlated with the regression error, namely, they are exogenous. Yet, endogeneity can arise incidentally from a large pool of regressors in a high-dimensional regression. This causes the inconsistency of the penalized least-squares method and possible false scientific discoveries. A necessary condition for model selection consistency of a general class of penalized regression methods is given, which allows us to prove formally the inconsistency claim. To cope with the incidental endogeneity, we construct a novel penalized focused generalized method of moments (FGMM) criterion function. The FGMM effectively achieves the dimension reduction and applies the instrumental variable methods. We show that it possesses the oracle property even in the presence of endogenous predictors, and that the solution is also near global minimum under the over-identification assumption. Finally, we also show how the semi-parametric efficiency of estimation can be achieved via a two-step approach.

Keywords

Focused GMM; Sparsity recovery; Endogenous variables; Oracle property; Conditional moment restriction; Estimating equation; Over identification; Global minimization; Semi-parametric efficiency

1. Introduction

In high-dimensional models, the overall number of regressors p grows extremely fast with the sample size n. It can be of order $\exp(n^{\alpha})$, for some $\alpha \in (0, 1)$. What makes statistical inference possible is the sparsity and *exogeneity* assumptions. For example, in the linear model

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0 + \varepsilon, \quad (1.1)$$

it is assumed that the number of elements in $S = \{j : \beta_{0j} \neq 0\}$ is small and $E\varepsilon\mathbf{X} = 0$, or more stringently

$$E(\varepsilon|\mathbf{X})=E(Y-\mathbf{X}^T\boldsymbol{\beta}_0|\mathbf{X})=0. \quad (1.2)$$

The latter is called "exogeneity". One of the important objectives of high-dimensional modeling is to achieve the variable selection consistency and make inference on the coefficients of important regressors. See, for example, Fan and Li (2001), Hunter and Li (2005), Zou (2006), Zhao and Yu (2006), Huang, Horowitz and Ma (2008), Zhang and Huang (2008), Wasserman and Roeder (2009), Lv and Fan (2009), Zou and Zhang (2009), Städler, Bühlmann and van de Geer (2010), and Bühlmann, Kalisch and Maathuis (2010). In these papers, (1.2) (or $E\varepsilon\mathbf{X} = 0$) has been assumed either explicitly or implicitly[1]. Condition of this kind is also required by the Dantzig selector of Candès and Tao (2007), which solves an optimization problem with constraint $\max_{j\le p}|\frac{1}{n}\sum_{i=1}^{n}X_{ij}(Y_i-\mathbf{X}_i^T\boldsymbol{\beta})|<C\sqrt{\frac{\log p}{n}}$ for some $C > 0$.

In high-dimensional models, requesting that ε and all the components of $\mathbf{X}$ be uncorrelated as (1.2), or even more specifically

$$E(Y-\mathbf{X}^T\boldsymbol{\beta}_0)X_j=0,\ \forall j=1,\dots,p, \quad (1.3)$$

can be restrictive particularly when p is large. Yet, (1.3) is a necessary condition for popular model selection techniques to be consistent. However, violations to assumption either (1.2) or (1.3) can arise as a result of selection biases, measurement errors, autoregression with autocorrelated errors, omitted variables, and from many other sources (Engle, Hendry and Richard 1983). They also arise from unknown causes due to a large pool of regressors, some of which are incidentally correlated with the random noise $Y-\mathbf{X}^T\boldsymbol{\beta}_0$. For example, in genomics studies, clinical or biological outcomes along with expressions of tens of thousands of genes are frequently collected. After applying variable selection techniques, scientists obtain a set of genes $\hat{S}$ that are responsible for the outcome. Whether (1.3) holds, however, is rarely validated. Because there are tens of thousands of restrictions in (1.3) to validate, it is likely that some of them are violated. Indeed, unlike low-dimensional least-squares, the sample correlations between residuals $\hat{\epsilon}$, based on the selected variables $\mathbf{X}_{\hat{S}}$, and predictors $\mathbf{X}$, are unlikely to be small, because all variables in the large set $\hat{S}^c$ are not even used in computing the residuals. When some of those are unusually large, endogeneity arises incidentally. In such cases, we will show that $\hat{S}$ can be inconsistent. In other words, violation of assumption (1.3) can lead to false scientific claims.

We aim to consistently estimate $\boldsymbol{\beta}_0$ and recover its sparsity under weaker conditions than (1.2) or (1.3) that are easier to validate. Let us assume that $\boldsymbol{\beta}_0=(\boldsymbol{\beta}_{0S}^T,0)^T$ and $\mathbf{X}$ can be partitioned as $\mathbf{X}=(\mathbf{X}_S^T,\mathbf{X}_N^T)^T$. Here $\mathbf{X}_S$ corresponds to the nonzero coefficients $\boldsymbol{\beta}_{0S}$, which we call *important regressors*, and $\mathbf{X}_N$ represents the *unimportant regressors* throughout the

[1] In fixed designs, e.g., Zhao and Yu (2006), it has been implicitly assumed that $n^{-1}\sum_{i=1}^{n}\varepsilon_iX_{ij}=o_p(1)$ for all $j < p$.

paper, whose coefficients are zero. We borrow the terminology of *endogeneity* from the econometric literature. A regressor is said to be *endogenous* when it is correlated with the error term, and *exogenous* otherwise. Motivated by the aforementioned issue, this paper aims to select $\mathbf{X}_S$ with probability approaching one and making inference about β_{0S}, allowing components of $\mathbf{X}$ to be endogenous. We propose a unified procedure that can address the problem of endogeneity to be present in either important or unimportant regressors, or both, and we do not require the knowledge of which case of endogeneity is present in the true model. The identities of $\mathbf{X}_S$ are unknown before the selection.

The main assumption we make is that, there is a vector of observable *instrumental variables* $\mathbf{W}$ such that

$$E\left[\varepsilon|\mathbf{W}\right]=0. \quad (1.4)$$

[2]

Briefly speaking, $\mathbf{W}$ is called an "instrumental variable" when it satisfies (1.4) and is correlated with the explanatory variable $\mathbf{X}$. In particular, as noted in the footnote, $\mathbf{W} = \mathbf{X}_S$ is allowed so that the instruments are unknown but no additional data are needed. Instrumental variables (IV) have been commonly used in the literature of both econometrics and statistics in the presence of endogenous regressors, to achieve identification and consistent estimations (e.g., Hall and Horowitz 2005). An advantage of such an assumption is that it can be validated more easily. For example, when $\mathbf{W} = \mathbf{X}_S$, one needs only to check whether the correlations between $\hat{\in}$ and $\mathbf{X}_{\hat{S}}$ are small or not, with $\mathbf{X}_{\hat{S}}$ being a relatively low-dimensional vector, or more generally, the moments that are actually used in the model fitting such as (1.5) below hold approximately In short, our setup weakens the assumption (1.2) to some verifiable moment conditions.

What makes the variable selection consistency (with endogeneity) possible is the idea of *over identification*. Briefly speaking, a parameter is called "over-identified" if there are more restrictions than those are needed to grant its identifiability (for linear models, for instance, when the parameter satisfies more equations than its dimension). Let $(f_1,\ldots, f_p)$ and $(h_1,\ldots, h_p)$ be two different sets of transformations, which can be taken as a large number of series terms, e.g., B-splines and polynomials. Here each f_j and h_j are scalar functions. Then (1.4) implies

$$E(\varepsilon f_j(\mathbf{W}))=E(\varepsilon h_j(\mathbf{W}))=0,\ j=1,\ldots,p.$$

Write $\mathbf{F} = (f_1(\mathbf{W}),\ldots, f_p(\mathbf{W}))^T$, and $\mathbf{H} = (h_1(\mathbf{W}),\ldots, h_p(\mathbf{W}))^T$. We then have $E\varepsilon\mathbf{F} = E\varepsilon\mathbf{H} = 0$. Let S be the set of indices of important variables, and let $\mathbf{F}_S$ and $\mathbf{H}_S$ be the subvectors of $\mathbf{F}$

[2] We thank the AE and referees for suggesting the use of a general vector of instrument $\mathbf{W}$, which extends to the more general endogeneity problem, allowing the presence of endogenous important regressors. In particular, $\mathbf{W}$ is allowed to be $\mathbf{X}_S$, which amounts to assume that $E(\varepsilon|\mathbf{X}_S) = 0$ by (1.4), but allow $E(\varepsilon|\mathbf{X}) \neq 0$. In this case, we can allow the instruments $\mathbf{W} = \mathbf{X}_S$ to be unknown, and $\mathbf{F}$ and $\mathbf{H}$ to be defined below can be transformations of $\mathbf{X}$. This is the setup of an earlier version of this paper, which is much weaker than (1.2) and allows some of $\mathbf{X}_N$ to be endogenous.

and $\mathbf{H}$ corresponding to the indices in S. Implied by $E\varepsilon\mathbf{F} = E\varepsilon\mathbf{H} = 0$, and $\varepsilon = Y - \mathbf{X}_S^T\boldsymbol{\beta}_{0S}$, there exists a solution $\beta_S = \beta_{0S}$ to the *over-identified* equations (with respect to β_S) such as

$$E(Y - \mathbf{X}_S^T\boldsymbol{\beta}_S)\boldsymbol{F}_S = 0 \text{ and } E(Y - \mathbf{X}_S^T\boldsymbol{\beta}_S)\boldsymbol{H}_S = 0. \quad (1.5)$$

In (1.5), we have twice as many linear equations as the number of unknowns, yet the solution exists and is given by $\beta_S = \beta_{0S}$. Because β_{0S} satisfies more equations than its dimension, we call β_{0S} to be *over-identified*. On the other hand, for any other set $\tilde{S}$ of variables, if $\tilde{S} \not\subset S$, then the following $2|\tilde{S}|$ equations (with $|\tilde{S}| = \dim(\beta_{\tilde{S}})$ unknowns)

$$E(Y - \mathbf{X}_{\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}})\mathbf{F}_{\tilde{S}} = 0 \text{ and } E(Y - \mathbf{X}_{\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}})\boldsymbol{H}_{\tilde{S}} = 0 \quad (1.6)$$

have no solution as long as the basis functions are chosen such that $\mathbf{F}_{\tilde{S}} \neq \mathbf{H}_{\tilde{S}}$.[3] The above setup includes $\mathbf{W} = \mathbf{X}_S$ with $\mathbf{F} = \mathbf{X}$ and $\mathbf{H} = \mathbf{X}^2$ as a specific example (or $\mathbf{H} = \cos(\mathbf{X}) + 1$ if $\mathbf{X}$ contain many binary variables).

We show that in the presence of endogenous regressors, the classical penalized least squares method is no longer consistent. Under model

$$Y = \mathbf{X}_S^T\boldsymbol{\beta}_{0S} + \varepsilon, \ E(\varepsilon|\mathbf{W}) = 0,$$

we introduce a novel penalized method, called *focused generalized method of moments* (FGMM), which differs from the classical GMM (Hansen 1982) in that the working instrument $\mathbf{V}(\beta)$ in the moment functions $n^{-1}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}(\boldsymbol{\beta})$ for FGMM also depends *irregularly* on the unknown parameter $\boldsymbol{\beta}$ (which also depends on ($\mathbf{F}$, $\mathbf{H}$), see Section 3 for details). With the help of *over identification,* the FGMM successfully eliminates those subset $\tilde{S}$ such that $\tilde{S} \not\subset S$. As we will see in Section 3, a penalization is still needed to avoid over-fitting. This results in a novel penalized FGMM.

We would like to comment that FGMM differs from the low-dimensional techniques of either moment selection (Andrews 1999, Andrews and Lu 2001) or shrinkage GMM (Liao 2013) in dealing with mis-specifications of moment conditions and dimension reductions. The existing methods in the literature on GMM moment selections cannot handle high-dimensional models. Recent literature on the instrumental variable method for high-dimensional models can be found in, e.g., Belloni et al. (2012), Caner and Fan (2012), García (2011). In these papers, the endogenous variables are in low dimensions. More closely related work is by Gautier and Tsybakov (2011), who solved a constrained minimization as an extension of Dantzig selector. Our paper, in contrast, achieves the oracle

[3] The compatibility of (1.6) requires very stringent conditions. If $E\mathbf{F}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T$ are invertible, then a necessary condition for (1.6) to have a common solution is that $(E\mathbf{F}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T)^{-1}E(Y\mathbf{F}_{\tilde{S}}) = (E\mathbf{H}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T)^{-1}E(Y\mathbf{H}_{\tilde{S}})$, which does not hold in general when $\mathbf{F} \neq \mathbf{H}$.

property via a penalized GMM. Also, we study a more general *conditional moment restricted model* that allows nonlinear models.

The remainder of this paper is as follows: Section 2 gives a necessary condition for a general penalized regression to achieve the oracle property. We also show that in the presence of endogenous regressors, the penalized least squares method is inconsistent. Sections 3 constructs a penalized FGMM, and discusses the rationale of our construction. Section 4 shows the oracle property of FGMM. Section 5 discusses the global optimization. Section 6 focuses on the semi-parametric efficient estimation after variable selection. Section 7 discusses numerical implementations. We present simulation results in Section 8. Finally, Section 9 concludes. Proofs are given in the appendix.

Notation

Throughout the paper, let $\lambda_{\min}(\mathbf{A})$ and $\lambda_{\max}(\mathbf{A})$ be the smallest and largest eigenvalues of a square matrix $\mathbf{A}$. We denote by $\|\mathbf{A}\|_F$, $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_\infty$ as the Frobenius, operator and element-wise norms of a matrix $\mathbf{A}$ respectively, defined respectively as $\|\mathbf{A}\|_F = \mathrm{tr}^{1/2}(\mathbf{A}^T\mathbf{A})$, $\|\mathbf{A}\|=\lambda_{\max}^{1/2}(\mathbf{A}^T\mathbf{A})$, and $\|\mathbf{A}\|_\infty = \max_{i,j} \|\mathbf{A}_{ij}\|$. For two sequences a_n and b_n, write $a_n \ll b_n$ (equivalently, $b_n \gg a_n$) if $a_n = o(b_n)$. Moreover, $|\beta|_0$ denotes the number of nonzero components of a vector β. Finally, $P_n^{'}(t)$ and $P_n^{''}(t)$ denote the first and second derivatives of a penalty function $P_n(t)$, if exist.

2. Necessary Condition for Variable Selection Consistency

2.1. Penalized regression and necessary condition

Let s denote the dimension of the true vector of nonzero coefficients β_{0S}. The sparse structure assumes that s is small compared to the sample size. A penalized regression problem, in general, takes a form of:

$$\min_{\beta\in\mathbb{R}^p} L_n(\boldsymbol{\beta})+\sum_{j=1}^{p} P_n(|\beta_j|),$$

where $P_n(\cdot)$ denotes a penalty function. There are relatively less attentions to the necessary conditions for the penalized estimator to achieve the oracle property. Zhao and Yu (2006) derived an *almost necessary* condition for the sign consistency, which is similar to that of Zou (2006) for the least squares loss with Lasso penalty. To the authors' best knowledge, so far there has been no necessary condition on the loss function for the selection consistency in the high-dimensional framework. Such a necessary condition is important, because it provides us a way to justify whether a specific loss function can result in a consistent variable selection.

Theorem 2.1 (Necessary Condition)—Suppose:

i. $L_n(\beta)$ is twice differentiable, and

$$\max_{1\le l,j\le p}\left|\frac{\partial^2 L_n(\boldsymbol{\beta}_0)}{\partial\beta_l\partial\beta_j}\right|=O_p(1).$$

ii. There is a local minimizer $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S,\hat{\boldsymbol{\beta}}_N)^T$ of

$$L_n(\boldsymbol{\beta})+\sum_{j=1}^{p}P_n(|\beta_j|)$$

such that $P(\hat{\boldsymbol{\beta}}_N=0)\to 1$, and $\sqrt{s}\,\|\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}_0\|=o_p(1)$.

iii. The penalty satisfies: $P_n(\cdot)\ge 0$, $P_n(0)=0$, $P_n'(t)$ is non-increasing when $t\in(0,u)$ for some $u>0$, and $\lim_{n\to\infty}\lim_{t\to 0^+}P_n'(t)=0$. Then for any $l\le p$,

$$\left|\frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial\beta_l}\right|\to^p 0. \quad (2.1)$$

The implication (2.1) is fundamentally different from the “irrepresentable condition” in Zhao and Yu (2006) and that of Zou (2006). It imposes a restriction on the loss function $L_n(\cdot)$, whereas the “irrepresentable condition” is derived under the least squares loss and $E(\varepsilon\mathbf{X})=0$. For the least squares, (2.1) reduces to either $n^{-1}\sum_{i=1}^{n}\varepsilon_i X_{il}=o_p(1)$ or $E\varepsilon\mathbf{X}_l=0$, which requires a exogenous relationship between ε and $\mathbf{X}$. In contrast, the irrepresentable condition requires a type of relationship between important and unimportant regressors and is specific to Lasso. It also differs from the Karush-Kuhn-Tucker (KKT) condition (e.g., Fan and Lv 2011) in that it is about the gradient vector evaluated at the true parameters rather than at the local minimizer.

The conditions on the penalty function in condition (iii) are very general, and are satisfied by a large class of popular penalties, such as Lasso (Tibshirani 1996), SCAD (Fan and Li 2001) and MCP (Zhang 2010), as long as their tuning parameter $\lambda_n\to 0$. Hence this theorem should be understood as a necessary condition imposed on the loss function instead of the penalty.

2.2. Inconsistency of least squares with endogeneity

As an application of Theorem 2.1, consider a linear model:

$$Y=\mathbf{X}^T\boldsymbol{\beta}_0+\varepsilon=\mathbf{X}_S^T\boldsymbol{\beta}_{0S}+\varepsilon, \quad (2.2)$$

where we may not have $E(\varepsilon\mathbf{X})=0$.

The conventional penalized least squares (PLS) problem is defined as:

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p} P_n(|\beta_j|).$$

In the simpler case when s, the number of nonzero components of β_0, is bounded, it can be shown that if there exist some regressors correlated with the regression error ε, the PLS does not achieve the variable selection consistency. This is because (2.1) does not hold for the least squares loss function. Hence without the possibly ad-hoc exogeneity assumption, PLS would not work any more, as more formally stated below.

Theorem 2.2 (Inconsistency of PLS)—Suppose the data are i.i.d., $s = O(1)$, and **X** has at least one endogenous component, that is, there is l such that $|E(\mathbf{X}_l\varepsilon)| > c$ for some $c > 0$. Assume that $EX_l^4 < \infty$, $E\varepsilon^4 < \infty$, and $P_n(t)$ satisfies the conditions in Theorem 2.1. If $\tilde{\boldsymbol{\beta}} = (\tilde{\boldsymbol{\beta}}_S^T, \tilde{\boldsymbol{\beta}}_N^T)^T$, corresponding to the coefficients of $(\mathbf{X}_S, \mathbf{X}_N)$, is a local minimizer of

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + \sum_{j=1}^{p} P_n(|\beta_j|),$$

then either $\|\tilde{\beta}_S - \beta_{0S}\| = o_p(1)$, or $\limsup_{n\to\infty} P(\tilde{\beta}_N = 0) < 1$.

The index l in the condition of the above theorem does not have to be an index of an important regressor. Hence the consistency for penalized least squares will fail even if the endogeneity is only present on the unimportant regressors.

We conduct a simple simulated experiment to illustrate the impact of endogeneity on the variable selection. Consider

$$\begin{gathered}
Y = \mathbf{X}^T\boldsymbol{\beta}_0 + \varepsilon, \ \varepsilon \sim N(0,1), \\
\boldsymbol{\beta}_{0S} = (5, -4, 7, -2, 1.5); \ \beta_{0j} = 0, \ \text{for } 6 \le j \le p. \\
X_j = Z_j \text{ for } j \le 5, \ X_j = (Z_j + 5)(1 + \epsilon), \ \text{for } 6 \le j \le p. \\
Z \sim N_p(0, \textstyle\sum), \ \text{independent of } \varepsilon, \ \text{with } (\textstyle\sum)_{ij} = 0.5^{|i-j|}.
\end{gathered}$$

In the design, the unimportant regressors are endogenous. The penalized least squares (PLS) with SCAD-penalty was used for variable selection. The λ's in the table represent the tuning parameter used in the SCAD-penalty. The results are based on the estimated $(\hat{\boldsymbol{\beta}}_S^T, \hat{\boldsymbol{\beta}}_N^T)^T$, obtained from minimizing PLS and FGMM loss functions respectively (we shall discuss the construction of FGMM loss function and its numerical minimization in detail subsequently). Here $\tilde{\beta}_S$ and $\hat{\beta}_N$ represent the estimators for coefficients of important and unimportant regressors respectively.

From Table 1, PLS selects many unimportant regressors (FP). In contrast, the penalized FGMM performs well in both selecting the important regressors and eliminating the unimportant ones. Yet, the larger MSE$_S$ of $\hat{\beta}_S$ by FGMM is due to the moment conditions used in the estimate. This can be improved further in Section 6. Also, when endogeneity is present on the important regressors, PLS estimator will have larger bias (see additional simulation results in Section 8.)

3. Focused GMM

3.1. Definition

Because of the presence of endogenous regressors, we introduce an *instrumental variable* (IV) regression model. Consider a more general nonlinear model:

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) | \mathbf{W}] = 0, \quad (3.1)$$

where Y stands for the dependent variable; $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a known function. For simplicity, we require g be one-dimensional, and should be thought of as a possibly nonlinear residual function. Our result can be naturally extended to a multi-dimensional g function. Here $\mathbf{W}$ is a vector of observed random variables, known as instrumental variables.

Model (3.1) is called a *conditional moment restricted model*, which has been extensively studied in the literature, e.g., Newey (1993), Donald et al. (2009), Kitamura et al (2004). The high-dimensional model is also closely related to the semi/nonparametric model estimated by sieves with a growing sieve dimension, e.g., Ai and Chen (2003). Recently van de Geer (2008) and Fan and Lv (2011) considered generalized linear models without endogeneity. Some interesting examples of the generalized linear model that fit into (3.1) are:

- linear regression, $g(t_1, t_2) = t_1 - t_2$;
- logit model, $g(t_1, t_2) = t_1 - \exp(t_2)/(1 + \exp(t_2))$;
- probit model, $g(t_1, t_2) = t_1 - \Phi(t_2)$ where $\Phi(\cdot)$ denotes the standard normal cumulative distribution function.

Let $(f_1, \ldots, f_p)$ and $(h_1, \ldots, h_p)$ be two different sets of transformations of $\mathbf{W}$, which can be taken as a large number of series basis, e.g., B-splines, Fourier series, polynomials (see Chen 2007 for discussions of the choice of sieve functions). Here each f_j and h_j are scalar functions. Write $\mathbf{F} = (f_1(\mathbf{W}), \ldots, f_p(\mathbf{W}))^T$, and $\mathbf{H} = (h_1(\mathbf{W}), \ldots, h_p(\mathbf{W}))^T$. The conditional moment restriction (3.1) then implies that

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{F}_S] = 0, \text{ and} E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{H}_S] = 0, \quad (3.2)$$

where $\mathbf{F}_S$ and $\mathbf{H}_S$ are the subvectors of $\mathbf{F}$ and $\mathbf{H}$ whose supports are on the oracle set $S = \{j \leq p : \beta_{0j} \neq 0\}$. In particular, when all the components of $\mathbf{X}_S$ are known to be exogenous, we can take $\mathbf{F} = \mathbf{X}$ and $\mathbf{H} = \mathbf{X}^2$ (the vector of squares of $\mathbf{X}$ taken coordinately), or $\mathbf{H} = \cos(\mathbf{X}) + 1$ if $\mathbf{X}$ is a binary variable. A typical estimator based on moment conditions like (3.2) can be obtained via the generalized method of moments (GMM, Hansen 1982). However, in the problem considered here, (3.2) cannot be used directly to construct the GMM criterion function, because the identities of $\mathbf{X}_S$ are unknown.

Remark 3.1—One seemingly working solution is to define **V** as a vector of transformations of **W**, for instance **V** = **F**, and employ GMM to the moment condition $E[g(Y, \mathbf{X}^T \beta_0)\mathbf{V}] = 0$. However, one has to take dim(**V**) ≥ dim(*β*) = *p* to guarantee that the GMM criterion function has a unique minimizer (in the linear model for instance). Due to $p \gg n$, the dimension of V is too large, and the sample analogue of the GMM criterion function may not converge to its population version due to the accumulation of high-dimensional estimation errors.

Let us introduce some additional notation. For any $\beta \in \mathbb{R}^p/\{0\}$, and $i = 1, \ldots, n$, define $r = |\beta|_0$-dimensional vectors

$\mathbf{F}_i(\beta) = (f_{l_1}(\mathbf{W}_i), \ldots, f_{l_r}(\mathbf{W}_i))^T$ and $\mathbf{H}_{\boldsymbol{i}}(\beta) = (h_{l_1}(\mathbf{W}_i), \ldots, h_{l_r}(\mathbf{W}_i))^T$, where $(l_1, \ldots, l_r)$ are the indices of nonzero components of *β*. For example, if $p = 3$ and $\beta = (-1, 0, 2)^T$, then $\mathbf{F}_i(\beta) = (f_1(\mathbf{W}_i), f_3(\mathbf{W}_i))^T$, and $\mathbf{H}_i(\beta) = (h_1(\mathbf{W}_i), h_3(\mathbf{W}_i))^T$, $i \le$ n.

Our *focused GMM* (FGMM) loss function is defined as

$$L_{\text{FGMM}}(\boldsymbol{\beta}) = \sum_{j=1}^{p} I_{(\beta_j \neq 0)} \left\{ w_{j1} \left[\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) f_j(\mathbf{W}_i) \right]^2 + w_{j2} \left[\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) h_j(\mathbf{W}_i) \right]^2 \right\}, \quad (3.3)$$

where w_{j1} and w_{j2} are given weights. For example, we will take $w_{j1} = 1/\hat{\text{var}}(f_j(\mathbf{W}))$ and $w_{j2} = 1/\hat{\text{var}}(h_j(\mathbf{W}))$ to standardize the scale (here $\hat{\text{var}}$ represents the sample variance). Writing in the matrix form, for $\mathbf{V}_i(\beta) = (\mathbf{F}_i(\beta^T, \mathbf{H}_i(\beta)^T)^T$,

$$L_{\text{FGMM}}(\boldsymbol{\beta}) = \left[\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{V}_i(\boldsymbol{\beta}) \right]^T \mathbf{J}(\boldsymbol{\beta}) \left[\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{V}_i(\boldsymbol{\beta}) \right],$$

where $\mathbf{J}(\beta) = \text{diag}\{w_{l_1 1}, \ldots, w_{l_r 1}, w_{l_1 2}, \ldots, w_{l_r 2}\}$.[4]

Unlike the traditional GMM, the "working instrumental variables" **V**(*β*) depend irregularly on the unknown *β*. As to be further explained, this ensures the dimension reduction, and allows to focus only on the equations with the IV whose support is on the oracle space, and is therefore called the focused GMM or FGMM for short.

We then define the FGMM estimator by minimizing the following criterion function:

$$Q_{\text{FGMM}}(\boldsymbol{\beta}) = L_{\text{FGMM}}(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_n(|\beta_j|). \quad (3.4)$$

[4] For technical reasons we use a diagonal weight matrix and it is likely non-optimal. However, it does not affect the variable selection consistency in this step.

Sufficient conditions on the penalty function $P_n(|\beta_j|)$ for the oracle property will be presented in Section 4. Penalization is needed because otherwise small coefficients in front of unimportant variables would be still kept in minimizing $L_{\mathrm{FGMM}}(\beta)$. As to become clearer in Section 6, the FGMM focuses on the model selection and estimation consistency without paying much effort to the efficient estimation of β_{0S}.

3.2. Rationales behind the construction of FGMM

3.2.1. Inclusion of V(β)—

We construct the FGMM criterion function using

$$\mathbf{V}(\boldsymbol{\beta})=(\mathbf{F}(\boldsymbol{\beta})^T,\mathbf{H}(\boldsymbol{\beta})^T)^T.$$

A natural question arises: why not just use one set of IV's so that $\mathbf{V}(\beta) = \mathbf{F}(\beta)$? We now explain the rationale behind the inclusion of the second set of instruments $\mathbf{H}(\beta)$. To simplify notation, let $F_{ij} = f_j(\mathbf{W}_i)$ and $H_{ij} = h_j(\mathbf{W}_i)$ for $j \leq p$ and $i \leq n$. Then $\mathbf{F}_i = (F_{i1},\ldots, F_{ip})$ and $\mathbf{H}_i = (H_{i1},\ldots, H_{ip})$. Also write $F_j = f_j(\mathbf{W})$ and $H_j = h_j(\mathbf{W})$ for $j \leq p$.

Let us consider a linear regression model (2.2) as an example. If $\mathbf{H}(\beta)$ were not included and $\mathbf{V}(\beta) = \mathbf{F}(\beta)$ had been used, the GMM loss function would have been constructed as

$$L_v(\boldsymbol{\beta})=\|\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{F}_i(\boldsymbol{\beta})\|^2, \quad (3.5)$$

where for the simplicity of illustration, $\mathbf{J}(\beta)$ is taken as an identity matrix. We also use the L_0-penalty $P_n(|\beta_j|) = \lambda_n I_{(|\beta_j|\neq 0)}$ for illustration. Suppose that the true $\boldsymbol{\beta}_0=(\boldsymbol{\beta}_{0S}^T,0,\ldots,0)^T$ where only the first s components are nonzero and that $s > 1$. If we, however, restrict ourselves to $\beta_p = (0, \ldots, 0, \beta_p)$, the criterion function now becomes

$$Q_{\mathrm{FGMM}}(\boldsymbol{\beta}_p)=\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-X_{ip}\beta_p)F_{ip}\right]^2+\lambda_n.$$

It is easy to see its minimum is just λ_n. On the other hand, if we optimize Q_{FGMM} on the oracle space $\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,0)^T$, then

$$\min_{\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,0)^T,\boldsymbol{\beta}_{S,j}\neq 0} Q_{\mathrm{FGMM}}(\boldsymbol{\beta})\geq s\lambda_n.$$

As a result, it is inconsistent for variable selection.

The use of L_0-penalty is not essential in the above illustration. The problem is still present if the L_1-penalty is used, and is not merely due to the biasedness of L_1-penalty. For instance, recall that for the SCAD penalty with hyper parameter (a, λ_n), $P_n(\cdot)$ is non-decreasing, and $P_n(t)=\frac{(a+1)}{2}\lambda_n^2$ when $t \geq a\lambda_n$. Given that $\min_{j\in S}|\beta_{0j}| \gg \lambda_n$,

$$Q_{\text{FGMM}}(\boldsymbol{\beta}_0) \geq \sum_{j\in S} P_n(|\beta_{0j}|) \geq sP_n(\min_{j\in S}|\beta_{0j}|)=\frac{(a+1)}{2}\lambda_n^2 s.$$

On the other hand, $Q_{\text{FGMM}}(\boldsymbol{\beta}_p^*)=P_n(|\beta_p^*|) \leq \frac{(a+1)}{2}\lambda_n^2$ which is strictly less than $Q_{\text{FGMM}}(\boldsymbol{\beta}_0)$. So the problem is still present when an asymptotically unbiased penalty (e.g., SCAD, MCP) is used.

Including an additional term $\mathbf{H}(\boldsymbol{\beta})$ in $\mathbf{V}(\boldsymbol{\beta})$ can overcome this problem. For example, if we still restrict to $\boldsymbol{\beta}_p = (0,\ldots, \beta_p)$ but include an additional but different IV H_{ip}, the criterion function then becomes, for the L_0 penalty:

$$Q_{\text{FGMM}}(\boldsymbol{\beta}_p)=\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_{ip}\beta_p)F_{ip}\right]^2+\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - X_{ip}\beta_p)H_{ip}\right]^2+\lambda_n.$$

In general, the first two terms cannot achieve $o_p(1)$ simultaneously as long as the two sets of transformations $\{f_j(\cdot)\}$ and $\{h_j(\cdot)\}$ are fixed differently. so long as n is large and

$$(EX_pF_p)^{-1}E(YF_p) \neq (EX_pH_p)^{-1}E(YH_p). \quad (3.6)$$

As a result, $Q_{\text{FGMM}}(\boldsymbol{\beta}_p)$ is bounded away from zero with probability approaching one.

To better understand the behavior of $Q_{\text{FGMM}}(\boldsymbol{\beta})$, it is more convenient to look at the population analogues of the loss function. Because the number of equations in

$$E[(Y - \mathbf{X}^T\boldsymbol{\beta})\mathbf{F}(\boldsymbol{\beta})]=0 \text{ and } E[(Y - \mathbf{X}^T\boldsymbol{\beta})\mathbf{H}(\boldsymbol{\beta})]=0 \quad (3.7)$$

is twice as many as the number of unknowns (nonzero components in $\boldsymbol{\beta}$), if we denote $\tilde{S}$ as the support of $\boldsymbol{\beta}$, then (3.7) has a solution only when $(E\mathbf{F}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T)^{-1}E(Y\mathbf{F}_{\tilde{S}})=(E\mathbf{H}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T)^{-1}E(Y\mathbf{H}_{\tilde{S}})$, which does not hold in general unless $\tilde{S} = S$, the index set of the true nonzero coefficients. Hence it is natural for (3.7) to have a unique solution $\boldsymbol{\beta} = \boldsymbol{\beta}_0$. As a result, if we define

$$G(\boldsymbol{\beta})=\|E(Y - \mathbf{X}^T\boldsymbol{\beta})\mathbf{F}(\boldsymbol{\beta})\|^2+\|E(Y - \mathbf{X}^T\boldsymbol{\beta})\mathbf{H}(\boldsymbol{\beta})\|^2,$$

the population version of L_{FGMM}, then as long as β is not close to β_0, G should be bounded away from zero. Therefore, it is reasonable for us to assume that for any $\delta > 0$, there is $\gamma(\delta) > 0$ such that

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty>\delta,\boldsymbol{\beta}\neq 0} G(\boldsymbol{\beta})>\gamma(\delta). \quad (3.8)$$

On the other hand, $E(\varepsilon|\mathbf{W})=E(Y-\mathbf{X}_S^T\boldsymbol{\beta}_{0S}|\mathbf{W})=0$ implies $G(\beta_0) = 0$.

Our FGMM loss function is essentially a sample version of $G(\beta)$, so minimizing $L_{\text{FGMM}}(\beta)$ forces the estimator to be close to β_0, but small coefficients in front of unimportant but exogenous regressors may still be allowed. Hence a concave penalty function is added to L_{FGMM} to define Q_{FGMM}.

3.2.2. Indicator function

Another question readers may ask is that why not define $L_{\text{FGMM}}(\beta)$ to be, for some weight matrix $\mathbf{J}$,

$$\left[\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}_i\right]^T \mathbf{J}\left[\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}_i\right], \quad (3.9)$$

that is, why not replace the irregular β-dependent $\mathbf{V}(\beta)$ with $\mathbf{V}$, and use the entire $2p$-dimensional $\mathbf{V} = (\mathbf{F}^T, \mathbf{H}^T)^T$ as the IV? This is equivalent to the question why the indicator function in (3.3) cannot be dropped.

The indicator function is used to prevent the accumulation of estimation errors under the high dimensionality. To see this, rewrite (3.9) to be:

$$\sum_{j=1}^p \frac{1}{\hat{\text{var}}(F_j)}\left(\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})F_{ij}\right)^2+\frac{1}{\hat{\text{var}}(H_j)}\left(\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})H_{ij}\right)^2.$$

Since $\dim(\mathbf{V}_i) = 2p \gg n$, even if each individual term evaluated at $\beta = \beta_0$ is $O_p\left(\frac{1}{n}\right)$, the sum of p terms would become stochastically unbounded. In general, (3.9) does not converge to its population analogue when $p \gg n$ because the accumulation of high-dimensional estimation errors would have a non-negligible effect.

In contrast, the indicator function effectively reduces the dimension and prevents the accumulation of estimation errors. Once the indicator function is included, the proposed FGMM loss function evaluated at β_0 becomes:

$$\sum_{j\in S} \frac{1}{\hat{\text{var}}(F_j)}\left(\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta}_0)F_{ij}\right)^2+\frac{1}{\hat{\text{var}}(H_j)}\left(\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta}_0)H_{ij}\right)^2,$$

which is small because $E[g(Y, \mathbf{X}^T\beta_0)\mathbf{F}_S] = E[g(Y, \mathbf{X}^T\beta_0)\mathbf{H}_S] = 0$ and that there are only $s = |S|_0$ terms in the summation.

Recently, there has been growing literature on the shrinkage GMM, e.g., Caner (2009), Caner and Zhang (2013), Liao (2013), etc, regarding estimation and variable selection based on a set of moment conditions like (3.2). The model considered by these authors is restricted to either a low-dimensional parameter space or a low-dimensional vector of moment conditions, where there is no such a problem of error accumulations.

4. Oracle Property of FGMM

FGMM involves a non-smooth loss function. In the appendix, we develop a general asymptotic theory for high-dimensional models to accommodate the non-smooth loss function.

Our first assumption defines the penalty function we use. Consider a similar class of folded concave penalty functions as that in Fan and Li (2001).

For any $\boldsymbol{\beta} = (\beta_1,\ldots, \beta_s)^T \in \mathbb{R}^s$, and $|\beta_j| \neq 0$, $j = 1,\ldots, s$, define

$$\eta(\boldsymbol{\beta})=\lim_{\epsilon\to 0^+}\sup\max_{j\le s}\sup_{\substack{t_1<t_2\\(t_1,t_2)\in(|\beta_j|-\epsilon,|\beta_j|+\epsilon)}} -\frac{P_n'(t_2)-P_n'(t_1)}{t_2-t_1}, \quad (4.1)$$

which is $\max_{j\le s} - P_n''(|\beta_j|)$ if the second derivative of P_n is continuous. Let

$$d_n=\frac{1}{2}\min\{|\beta_{0j}|:\beta_{0j}\neq 0, j=1,\ldots,p\}$$

represent the strength of signals.

Assumption 4.1

The penalty function $P_n(t) : [0, \infty) \to \mathbb{R}$ satisfies:

i. $P_n(0) = 0$

ii. $P_n(t)$ is concave, non-decreasing on $[0, \infty)$, and has a continuous derivative $P_n'(t)$ when $t > 0$.

iii. $\sqrt{s}P_n'(d_n)=o(d_n)$.

iv. There exists $c > 0$ such that $\sup_{\beta\in B(\beta_{0S},cd_n)} \eta(\beta) = o(1)$.

These conditions are standard. The concavity of $P_n(\cdot)$ implies that $\eta(\beta) \geq 0$ for all $\beta \in \mathbb{R}^s$. It is straightforward to check that with properly chosen tuning parameters, the L_q penalty (for $q \leq 1$), hard-thresholding (Antoniadis 1996), SCAD (Fan and Li 2001), and MCP (Zhang 2010) all satisfy these conditions. As thoroughly discussed by Fan and Li (2001), a penalty function that is desirable for achieving the oracle properties should result in an estimator with three properties: unbiasedness, sparsity and continuity (see Fan and Li 2001 for details). These properties motivate the needs of using a folded concave penalty.

The following assumptions are further imposed. Recall that for $j \leq p$, $F_j = f_j(\mathbf{W})$ and $H_j = h_j(\mathbf{W})$.

Assumption 4.2

i. The true parameter β_0 is uniquely identified by $E(g(Y, \mathbf{X}^T \beta_0)|\mathbf{W}) = 0$.

ii. $(Y_1, \mathbf{X}_{1)},\ldots, (Y_n, \mathbf{X}_n)$ are independent and identically distributed.

Remark 4.1

Condition (i) above is standard in the GMM literature (e.g., Newey 1993, Donald et al. 2009, Kitamura et al. 2004). This condition is closely related to the “over-identifying restriction”, and ensures that we can always find two sets of transformations **F** and **H** such that the equations in (3.2) are uniquely satisfied by $\beta_S = \beta_{0S}$. In linear models, this is a reasonable assumption, as discussed in Section 3.2. In nonlinear models, however, requiring the identifiability from either $E(g(Y, \mathbf{X}^T\beta_0)|\mathbf{W}) = 0$ or (3.2) may be restrictive. Indeed, Dominguez and Lobato 2004) showed that the identification condition in (i) may depend on the marginal distributions of **W**. Furthermore, in nonparametric regression problems as in Bickel et al. (2009) and Ai and Chen (2003), the sufficient condition of Condition (i) is even more complicated, which also depends on the conditional distribution of **X**|**W**, and is known to be statistically untestable (see Newey and Powell 2003, Canay et al 2013).

Assumption 4.3

There exist $b_1, b_2, b_3 > 0$ and $r_1, r_2, r_3 > 0$ such that for any $t > 0$,

i. $P(|g(Y, \mathbf{X}^T\beta_0)| > t) \leq \exp(-(t/b_1)^{r1})$,

ii. $\max_{l \leq p} P(|F_l| > t) \leq \exp(-(t/b_2)^{r2})$, $\max_{l \leq p} P(|H_l > t) \leq \exp(-(t/b_3)^{r3})$.

iii. $\min_{j \in S} \operatorname{var}(g(Y, \mathbf{X}^T \beta_0)F_j)$ and $\min_{j \in S} \operatorname{var}(g(Y, \mathbf{X}^T \beta_0)H_j)$ are bounded away from zero.

iv. $\operatorname{var}(F_j)$ and $\operatorname{var}(H_j)$ are bounded away from both zero and infinity uniformly in $j = 1,\ldots, p$ and $p \geq 1$.

We will assume $g(\cdot,\cdot)$ to be twice differentiable, and in the following assumptions, let

$$m(t_1, t_2) = \frac{\partial g(t_1, t_2)}{\partial t_2}, q(t_1, t_2) = \frac{\partial^2 g(t_1, t_2)}{\partial t_2^2}, \mathbf{V}_S = \begin{pmatrix} \mathbf{F}_S \\ \mathbf{H}_S \end{pmatrix}.$$

Assumption 4.4

i. $g(\cdot,\cdot)$ is twice differentiable.

ii. $\sup_{t_1, t_2} |m(t_1, t_2)| < \infty$, and $\sup_{t_1, t_2} |q(t_1, t_2)| < \infty$.

It is straightforward to verify Assumption 4.4 for linear, logistic and probit regression models.

Assumption 4.5

There exist $C_1 > 0$ and $C_2 > 0$ such that

$$\lambda_{\max}[(Em(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S^T)(Em(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S^T)^T] < C_1.$$
$$\lambda_{\min}[(Em(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S^T)(Em(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S^T)^T] < C_2;$$

These conditions require that the instrument $\mathbf{V}_S$ be not *weak*, that is, $\mathbf{V}_S$ should not be weakly correlated with the important regressors. In the generalized linear model, Assumption 4.5 is satisfied if proper conditions on the design matrices are imposed. For example, in the linear regression model and probit model, we assume the eigenvalues of $(E\mathbf{X}_S\mathbf{V}_S^T)(E\mathbf{X}_S\mathbf{V}_S^T)^T$ and $(E\phi(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}_S\mathbf{V}_S^T)(E\phi(\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}_S\mathbf{V}_S^T)^T$ are bounded away from both zero and infinity respectively, where $\varphi(\cdot)$ is the standard normal density function. Conditions in the same spirit are also assumed in, e.g., Bradic et al. (2011), and Fan and Lv (2011).

Define

$$\boldsymbol{\Upsilon} = \mathrm{var}(g(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{V}_S). \quad (4.2)$$

Assumption 4.6

- **i.** For some $c > 0$, $\lambda_{\min}(\Upsilon) > c$.
- **ii.** $sP_n'(d_n) + s\sqrt{(\log p)/n} + s^3(\log s)/n = o(P_n'(0^+))$, $P_n'(d_n)s^2 = O(1)$, and $s\sqrt{(\log p)/n} = o(d_n)$.
- **iii.** $P_n'(d_n) = o(1/\sqrt{ns})$ and $\sup_{\|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \le d_n/4} \eta(\boldsymbol{\beta}) = o((s \log p)^{-1/2})$.
- **iv.** $\max_{j \notin S} \|Em(y, \mathbf{X}^T\boldsymbol{\beta}_0)X_j\mathbf{V}_S\| \sqrt{(\log s)/n} = o(P_n(0^+))$.

This assumption imposes a further condition jointly on the penalty, the strength of the minimal signal and the number of important regressors. Condition (i) is needed for the asymptotic normality of the estimated nonzero coefficients. When either SCAD or MCP is used as the penalty function with a tuning parameter λ_n, $P_n'(d_n) = \sup_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_{0S}\| \le d_n/4} \eta(\boldsymbol{\beta}) = 0$ and $P_n'(0^+) = \lambda_n$ when $\lambda_n = o(d_n)$. Thus Conditions (ii)-(iv) in the assumption are satisfied as long as $s\sqrt{\log p/n} + s^3 \log s/n \ll \lambda_n \ll d_n$. This requires the signal d_n be strong and s be small compared to n. Such a condition is needed to achieve the variable selection consistency.

Under the foregoing regularity conditions, we can show the oracle property of a local minimizer of Q_{FGMM} (3.4).

Theorem 4.1

Suppose $s^3 \log p = o(n)$. Under Assumptions 4.1-4.6, there exists a local minimizer $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S^T, \hat{\boldsymbol{\beta}}_N^T)^T$ of $Q_{\mathrm{FGMM}}(\boldsymbol{\beta})$ with $\hat{\boldsymbol{\beta}}_S$ and $\hat{\boldsymbol{\beta}}_N$ being sub-vectors of $\hat{\boldsymbol{\beta}}$ whose coordinates are in S and S^c respectively, such that:

i.

$$\sqrt{n}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^{-1/2}\sum(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S})\to^d N(0,1),$$

for any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$, $\|\boldsymbol{\alpha}\| = 1$, where $\mathbf{A}=Em(Y,\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}_S\mathbf{V}_S^T$,

$$\boldsymbol{\Gamma}=4\mathbf{AJ}(\boldsymbol{\beta}_0)\boldsymbol{\Upsilon}\mathbf{J}(\boldsymbol{\beta}_0)\mathbf{A}^T, \mathit{and} \sum=\mathit{2}\mathbf{AJ}\boldsymbol{\beta}_{\mathit{0}})\mathbf{A}^T.$$

ii.

$$\lim_{n\to\infty}P(\hat{\boldsymbol{\beta}}_N=0)=1.$$

In addition, the local minimizer $\hat{\boldsymbol{\beta}}$ is strict with probability at least $1 - \delta$ for an arbitrarily small $\delta > 0$ and all large n.

iii. Let $\hat{S} = \{j \leq p : \hat{\beta}_j \neq 0\}$. Then

$$P(\hat{S}=S) \to 1.$$

Remark 4.2

As was shown in an earlier version of this paper Fan and Liao (2012), when it is known that $E[g(Y, \mathbf{X}^T\boldsymbol{\beta}_0)|\mathbf{X}_S] = 0$ but likely $E[g(Y, \mathbf{X}^T\boldsymbol{\beta}_0)|\mathbf{X}] \neq 0$, we can take $\mathbf{V} = (\mathbf{F}^T, \mathbf{H}^T)^T$ to be transformations of $\mathbf{X}$ that satisfy Assumptions 4.3-4.6. In this way, we do not need an extra instrumental variable $\mathbf{W}$, and Theorem 4.1 still goes through, while the traditional methods (e.g., penalized least squares in the linear model) can still fail as shown by Theorem 2.2. In the high-dimensional linear model, compared to the classical assumption: $E(\varepsilon|\mathbf{X}) = 0$, our condition $E(\varepsilon| \mathbf{X}_S) = 0$ is relatively easier to validate as $\mathbf{X}_S$ is a low-dimensional vector.

Remark 4.3

We now explain our required lower bound on the signal $s\sqrt{\log p/n}=o(d_n)$. When a penalized regression is used, which takes the form $\min_{\beta}L_n(\boldsymbol{\beta})+\sum_{j=1}^{p}P_n(|\beta_j|)$, it is required that if $L_n(\boldsymbol{\beta})$ is differentiable, $\max_{j\notin S}|\partial L_n(\boldsymbol{\beta}_0)/\partial\beta_j|=o(P_n'(0^+))$. This often leads to a requirement of the lower bound of d_n. Therefore, such a lower bound of d_n depends on the choice of both the loss function $L_n(\boldsymbol{\beta})$ and the penalty. For instance, in the linear model when least squares with a SCAD penalty is employed, this condition is equivalent to $\sqrt{\log p/n}=o(d_n)$. It is also known that the adaptive lasso penalty requires the minimal signal to be significantly larger than $\sqrt{\log p/n}$(Huang, Ma and Zhang 2008). In our framework, the requirement $s\sqrt{\log p}/n=o(d_n)$ arises from the use of the new FGMM loss function. Such a condition is stronger than that of the least squares loss function, which is the price paid to achieve variable selection consistency in the presence of endogeneity. This condition is still easy to satisfy as long as *s* grows slowly with *n*.

Remark 4.4

Similar to the "irrpresentable condition" for Lasso, the FGMM requires important and unimportant explanatory variables not be strongly correlated. This is fulfilled by Assumption 4.6(iv). For instance, in the linear model and $\mathbf{V}_S$ contains $\mathbf{X}_S$ as in our earlier version, this condition implies $\max_{j\notin S}\|EX_j\mathbf{X}_S\|\sqrt{\log s/n}=o(\lambda_n)$. Strong correlation between ($\mathbf{X}_S$, $\mathbf{X}_N$) is also ruled out by the identifiability condition Assumption 4.2. To illustrate the idea, consider a case of perfect linear correlation: $\mathbf{X}_S^T\boldsymbol{\alpha}-\mathbf{X}_N^T\boldsymbol{\delta}=0$ for some ($\boldsymbol{\alpha}$, $\boldsymbol{\delta}$) with $\boldsymbol{\delta}\neq 0$. Then, $\mathbf{X}^T\boldsymbol{\beta}_0=\mathbf{X}_S^T(\boldsymbol{\beta}_{0S}-\boldsymbol{\alpha})+\mathbf{X}_N^T\boldsymbol{\delta}$. As a result, the FGMM can be variable selection inconsistent because $\boldsymbol{\beta}_0$ and ($\boldsymbol{\beta}_{0S}-\boldsymbol{\alpha}$, $\boldsymbol{\delta}$) are observationally equivalent, violating Assumption 4.2.

5. Global minimization

With the over identification condition, we can show that the local minimizer in Theorem 4.1 is nearly global. To this end, define an l_∞ ball centered at $\boldsymbol{\beta}_0$ with radius δ:

$$\Theta_\delta=\{\boldsymbol{\beta}\in\mathbb{R}^p:|\beta_i-\beta_{0i}|<\delta, i=1,\dots,p\}.$$

Assumption 5.1 (over-identification)

For any $\delta > 0$, there is $\gamma > 0$ such that

$$\lim_{n\to\infty}P\left(\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}}\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}_i(\boldsymbol{\beta})\|^2>\gamma\right)=1.$$

This high-level assumption is hard to avoid in high-dimensional problems. It is the empirical counterpart of (3.8). In classical low-dimensional regression models, this assumption has often been imposed in the econometric literature, e.g., Andrews (1999), Chernozhukov and Hong (2003), among many others. Let us illustrate it by the following example.

Example 5.1

Consider a linear regression model of low dimensions: $E(Y-\mathbf{X}_S^T\boldsymbol{\beta}_{0S}|\mathbf{W})=0$, which implies $E[(Y-\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{F}_S]=0$ and $E[(Y-\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{H}_S]=0$ where *p* is either bounded or slowly diverging with *n*. Now consider the following problem:

$$\min_{\boldsymbol{\beta}\neq 0}G(\boldsymbol{\beta})\equiv\min_{\boldsymbol{\beta}\neq 0}\|E(Y-\mathbf{X}^T\boldsymbol{\beta})\mathbf{F}(\boldsymbol{\beta})\|^2+\|E(Y-\mathbf{X}^T\boldsymbol{\beta})\mathbf{H}(\boldsymbol{\beta})\|^2.$$

Once $[E\mathbf{F}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T]^{-1}E[\mathbf{F}_{\tilde{S}}Y]\neq[E\mathbf{H}_{\tilde{S}}\mathbf{X}_{\tilde{S}}^T]^{-1}E[\mathbf{H}_{\tilde{S}}Y]$ for all index set $\tilde{S}\neq S$, the objective function is then minimized to zero uniquely by $\boldsymbol{\beta}=\boldsymbol{\beta}_0$. Moreover, for any $\delta>0$ there is $\gamma>0$ such that when $\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}$, we have $G(\boldsymbol{\beta})>\gamma>0$. Assumption 5.1 then follows from the uniform weak law of large number: with probability approaching one, uniformly in $(\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}$,

$$\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{F}_i(\boldsymbol{\beta})(Y_i-\mathbf{X}_i^T\boldsymbol{\beta})\|^2+\|\frac{1}{n}\sum_{i=1}^{n}\mathbf{H}_i(\boldsymbol{\beta})(Y_i-\mathbf{X}_i^T\boldsymbol{\beta})\|^2>\gamma/2.$$

When *p* is much larger than *n*, the accumulation of the fluctuations from using the law of large number is no longer negligible. It is then challenging to show that $\|E[g(Y_i,\mathbf{X}^T\boldsymbol{\beta})\mathbf{V}(\boldsymbol{\beta})]\|$ is close to $\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}_i(\boldsymbol{\beta})\|$ uniformly for high-dimensional $\boldsymbol{\beta}$s, which is why we impose Assumption 5.1 on the empirical counterpart instead of the population.

Theorem 5.1

Assume $\max_{j\in S}P_n'(|\beta_{0j}|)=o(s^{-1})$. Under Assumption 5.1 and those of Theorem 4.1, the local minimizer $\hat{\boldsymbol{\beta}}$ in Theorem 4.1 satisfies: for any $\delta>0$, there exists $\gamma>0$,

$$\lim_{n\to\infty}P\left(Q_{\mathbf{FGMM}}(\hat{\boldsymbol{\beta}})+\gamma<\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}}Q_{FGMM}(\boldsymbol{\beta})\right)=1.$$

The above theorem demonstrates that $\hat{\boldsymbol{\beta}}$ is a *nearly global minimizer*. For SCAD and MCP penalties, the condition $\max_{j\in S}P_n'(|\beta_{0j}|)=o(s^{-1})$ holds when $\lambda_n=o(s^{-1})$, which is satisfied if *s* is not large.

Remark 5.1

We exclude the set {0} from the searching area in both Assumption 5.1 and Theorem 5.1 because we do not include the intercept in the model so $\mathbf{X}(0)=0$ by definition, and hence $Q_{\text{FGMM}}(0)=0$. It is reasonable to believe that zero is not close to the true parameter, since we assume there should be at least one important regressor in the model. On the other hand, if we always keep $X_1 = 1$ to allow for an intercept, there is no need to remove {0} in either Assumption 5.1 or the above theorem. Such a small change is not essential.

Remark 5.2

Assumption 5.1 can be slightly relaxed so that γ is allowed to decay slowly at a certain rate. The lower bound of such a rate is given by Lemma D.2 in the appendix. Moreover, Theorem 5.1 is based on an over-identification assumption, which is essentially different from the global minimization theory in the recent high-dimensional literature, e.g., Zhang (2010), Bühlmann and van de Geer (2011, ch 9), and Zhang and Zhang (2012).

6. Semi-parametric efficiency

The results in Section 5 demonstrate that the choice of the basis functions $\{f_j, h_j\}_{j\leq p}$ forming **F** and **H** influences the asymptotic variance of the estimator. The resulting estimator is in general not efficient. To obtain a semi-parametric efficient estimator, one can employ a second step post-FGMM procedure. In the linear regression, a similar idea has been used by Belloni and Chernozhukov (2013).

After achieving the oracle properties in Theorem 4.1, we have identified the important regressors with probability approaching one, that is,

$$\hat{S}=\{j:\hat{\beta}_j \neq 0\}, \hat{\mathbf{X}}_S=(X_j:j \in \hat{S}), P(\hat{S}=S) \rightarrow 1.$$

This reduces the problem to a low-dimensional problem. For simplicity, we restrict $s = O(1)$. The problem of constructing semi-parametric efficient estimator (in the sense of Newey (1990) and Bickel et al. (1998)) in a low-dimensional model

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S})|\mathbf{W}]=0$$

has been well studied in the literature (see, for example, Chamberlain (1987), Newey (1993)). The optimal instrument that leads to the semi-parametric efficient estimation of β_{0S} is given by $\mathbf{D}(\mathbf{W})\sigma(\mathbf{W})^{-2}$, where

$$\mathbf{D}(\mathbf{W})=E(\frac{\partial g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S})}{\partial \boldsymbol{\beta}_S}|\mathbf{W}), \sigma(\mathbf{W})^2=E(g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S})^2|\mathbf{W}).$$

Newey (1993) showed that the semi-parametric efficient estimator of β_{0S} can be obtained by GMM with the moment condition:

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S})\sigma(\mathbf{W})^{-2}\mathbf{D}(\mathbf{W})]=0. \quad (6.1)$$

In the post-FGMM procedure, we replace $\mathbf{X}_S$ with the selected $\mathbf{X}_{\hat{S}}$ obtained from the first-step penalized FGMM. Suppose there exist consistent estimators $\hat{\mathbf{D}}(\mathbf{W})$ and $\hat{\sigma}(\mathbf{W})^2$ of $\mathbf{D}(\mathbf{W})$ and $\sigma(\mathbf{W})^2$. Let us assume the true parameter $\|\beta_{0S}\|_\infty < M$ for a large constant $M > 0$. We then estimate β_{0S} by solving

$$\rho_n(\boldsymbol{\beta}_S)=\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\hat{\mathbf{X}}_{iS}^T\boldsymbol{\beta}_S)\hat{\sigma}(\mathbf{W}_i)^{-2}\hat{\mathbf{D}}(\mathbf{W}_i)=0, \quad (6.2)$$

on $\{\beta_S : \|\beta_{0S}\|_\infty \le M\}$, and the solution is assumed to be unique.

Assumption 6.1

i. There exist $C_1 > 0$ and $C_2 > 0$ so that

$$C_1<\inf_{\mathbf{w}\in\chi}\sigma(\mathbf{w})^2 \le \sup_{\mathbf{w}\in\chi}\sigma(\mathbf{w})^2<C_2.$$

In addition, there exist $\hat{\sigma}(\mathbf{w})^2$ and $\hat{\mathbf{D}}(\mathbf{w})$ such that

$$\sup_{\mathbf{w}\in\chi}|\hat{\sigma}(\mathbf{w})^2-\sigma(\mathbf{w})^2|=o_p(1), \, and \sup_{\mathbf{w}\in\chi}\|\hat{\mathbf{D}}(\mathbf{w})-\mathbf{D}(\mathbf{w})\|=o_p(\mathit{1})$$

where χ is the support of $\mathbf{W}$.

ii.

$$E(\sup_{\|\boldsymbol{\beta}\|_\infty\le M}g(Y,\mathbf{X}_S^T\boldsymbol{\beta}_S)^4)<\infty.$$

The consistent estimators for $\mathbf{D}(\mathbf{w})$ and $\sigma(\mathbf{w})^2$ can be obtained in many ways. We present a few examples below.

Example 6.1 (Homoskedasticity)

Suppose $Y=h(\mathbf{X}_S^T\boldsymbol{\beta}_{0S})+\varepsilon$ for some nonlinear function $h(\cdot)$. Then $\sigma(\mathbf{w})^2 = E(\varepsilon^2|\mathbf{W} = \mathbf{w}) = \sigma^2$, which does not depend on $\mathbf{w}$ under homoskedasticity. In this case, equations (6.1) and (6.2) do not depend on σ^2.

Example 6.2 (Simultaneous linear equations)

In the simultaneous linear equation model, $\mathbf{X}_S$ linearly depends on $\mathbf{W}$ as:

$$g(Y,\mathbf{X}_S^T\boldsymbol{\beta}_S)=Y-\mathbf{X}_S^T\boldsymbol{\beta}_S, \mathbf{X}_S=\mathbf{\Pi}\mathbf{W}+\mathbf{u}$$

for some coefficient matrix Π, where $\mathbf{u}$ is independent of $\mathbf{W}$. Then $\mathbf{D}(\mathbf{w}) = E(\mathbf{X}_S|\mathbf{W} = \mathbf{w}) = \Pi\mathbf{w}$. Let $\hat{\mathbf{X}} = (\mathbf{X}_{\hat{S}1}, \ldots, \mathbf{X}_{\hat{S}n})$, $\bar{\mathbf{W}} = (\mathbf{W}_1, \ldots, \mathbf{W}_n)$. We then estimate $\mathbf{D}(\mathbf{w})$ by $\hat{\Pi}\mathbf{w}$, where $\hat{\Pi} = (\hat{\mathbf{X}}\bar{\mathbf{W}}^T)(\bar{\mathbf{W}}\bar{\mathbf{W}})^{-1}$.

Example 6.3 (Semi-nonparametric estimation)

We can also assume a semi-parametric structure on the functional forms of $\mathbf{D}(\mathbf{w})$ and $\sigma(\mathbf{w})^2$:

$$\mathbf{D}(\mathbf{w})=\mathbf{D}(\mathbf{w};\theta_1),\, \sigma(\mathbf{w})^2=\sigma^2(\mathbf{w};\theta_2),$$

where $\mathbf{D}(\cdot;\theta_1)$ and $\sigma^2(\cdot;\theta_2)$ are semi-parametric functions parameterized by θ_1 and θ_2. Then $\mathbf{D}(\mathbf{w})$ and $\sigma(\mathbf{w})^2$ are estimated using a standard semi-parametric method. More generally, we can proceed by a pure nonparametric approach via respectively regressing $\partial g(Y, \hat{\mathbf{X}}_S^T\hat{\boldsymbol{\beta}}_S)/\partial\boldsymbol{\beta}_S$ and $g(Y, \hat{\mathbf{X}}_S^T\hat{\boldsymbol{\beta}}_S)^2$ on $\mathbf{W}$, provided that the dimension of $\mathbf{W}$ is either bounded or growing slowly with n (see Fan and Yao, 1998).

Theorem 6.1

Suppose $s = O(1)$, Assumption 6.1 and those of Theorem 4.1 hold. Then

$$\sqrt{n}(\hat{\boldsymbol{\beta}}_S^* - \boldsymbol{\beta}_{0S})\to^d N(0, [E(\sigma(\mathbf{W})^{-2}\mathbf{D}(\mathbf{W})\mathbf{D}(\mathbf{W})^T)]^{-1}),$$

and $[E(\sigma(\mathbf{W})^{-2}\mathbf{D}(\mathbf{W})\mathbf{D}(\mathbf{W})^T)]^{-1}$ is the semi-parametric efficiency bound in Chamberlain (1987).

7. Implementation

We now discuss the implementation for numerically minimizing the penalized FGMM criterion function.

7.1. Smoothed FGMM

As we previously discussed, including an indicator function benefits us in dimension reduction. However, it also makes L_{FGMM} unsmooth. Hence, minimizing $Q_{\mathrm{FGMM}}(\beta) = L_{\mathrm{FGMM}}(\beta)$+Penalty is generally NP-hard.

We overcome this discontinuity problem by applying the *smoothing* technique as in Horowitz (1992) and Bondell and Reich (2012), which approximates the indicator function by a smooth kernel $K : (-\infty,\infty) \to \mathbb{R}$ that satisfies

1. $0 \leq K(t) < M$ for some finite M and all $t \geq 0$.
2. $K(0) = 0$ and $\lim_{|t|\to\infty} K(t) = 1$.
3. $\limsup_{|t|\to\infty} |K'(t)t| = 0$, and $\limsup_{|t|\to\infty} |K''(t)t^2| < \infty$.

We can set $K(t)=\frac{F(t)-F(0)}{1-F(0)}$, where $F(t)$ is a twice differentiable cumulative distribution function. For a pre-determined small number h_n, L_{FGMM} is approximated by a continuous function $L_K(\beta)$ with the indicator replaced by $K(\beta_j^2/h_n)$. The objective function of the smoothed FGMM is given by

$$Q_K(\boldsymbol{\beta})=L_K(\boldsymbol{\beta})+\sum_{j=1}^{p}P_n(|\beta_j|).$$

As $h_n \to 0^+$, $K(\beta_j^2/h_n)$ converges to $I_{(\beta_j \neq 0)}$, and hence $L_K(\boldsymbol{\beta})$ is simply a smoothed version of $L_{\text{FGMM}}(\boldsymbol{\beta})$. As an illustration, Figure 1 plots such a function.

Smoothing the indicator function is often seen in the literature on high-dimensional variable selections. Recently, Bondell and Reich (2012) approximate $I_{(t \neq 0)}$ by $\frac{(h_n+1)t}{h_n+t}$ to obtain a tractable non-convex optimization problem. Intuitively, we expect that the smoothed FGMM should also achieve the variable selection consistency. Indeed, the following theorem formally proves this claim.

Theorem 7.1—Suppose $h_n^{1-\gamma}=o(d_n^2)$ for a small constant $\gamma \in (0, 1)$. Under the assumptions of Theorem 4.1 there exists a local minimizer $\hat{\boldsymbol{\beta}}'$ of the smoothed FGMM $Q_K(\boldsymbol{\beta})$ such that, for $\hat{S}'=\{j \leq p{:}\hat{\beta}_j' \neq 0\}$,

$$P(\hat{S}'=S) \to 1.$$

In addition, the local minimizer $\hat{\boldsymbol{\beta}}'$ is strict with probability at least $1 - \delta$ for an arbitrarily small $\delta > 0$ and all large *n*.

The asymptotic normality of the estimated nonzero coefficients can be established very similarly to that of Theorem 4.1, which is omitted for brevity.

7.2. Coordinate descent algorithm

We employ the iterative coordinate algorithm for the smoothed FGMM minimization, which was used by Fu (1998), Daubechies et al. (2004), Fan and Lv (2011), etc. The iterative coordinate algorithm minimizes one coordinate of $\boldsymbol{\beta}$ at a time, with other coordinates kept fixed at their values obtained from previous steps, and successively updates each coordinate. The penalty function can be approximated by local linear approximation as in Zou and Li (2008).

Specifically, we run the regular penalized least squares to obtain an initial value, from which we start the iterative coordinate algorithm for the smoothed FGMM. Suppose $\boldsymbol{\beta}^{(l)}$ is obtained at step *l*. For $k \in \{1, \ldots, p\}$, denote by $\boldsymbol{\beta}_{(-k)}^{(l)}$ a $(p - 1)$-dimensional vector consisting of all the components of ($\boldsymbol{\beta}^{(l)}$ but $\beta_k^{(l)}$. Write $(\boldsymbol{\beta}_{(-k)}^{(l)}, t)$ as the *p*-dimensional vector that replaces $\beta_k^{(l)}$ with *t*. The minimization with respect to *t* while keeping $\boldsymbol{\beta}_{(-k)}^{(l)}$ fixed is then a univariate minimization problem, which is not difficult to implement. To speed up the convergence, we can also use the second order approximation of $L_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t)$ along the *k*th component at $\beta_k^{(l)}$:

$$L_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t) \approx L_K(\boldsymbol{\beta}^{(l)}) + \frac{\partial L_K(\boldsymbol{\beta}^{(l)})}{\partial \beta_k}(t - \beta_k^{(l)}) + \frac{1}{2}\frac{\partial^2 L_K(\boldsymbol{\beta}^{(l)})}{\partial \beta_k^2}(t - \beta_k^{(l)})^2 \equiv L_K(\boldsymbol{\beta}^{(l)}) + \hat{L}_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t), \quad (7.1)$$

where $\hat{L}_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t)$ is a quadratic function of t. We solve for

$$t^* = \arg\min_t \hat{L}_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t) + P_n'(|\beta_k^{(l)}|)|t|, \quad (7.2)$$

which admits an explicit analytical solution, and keep the remaining components at step l.

Accept t^* as an updated kth component of $\boldsymbol{\beta}^{(l)}$ only if $L_K(\boldsymbol{\beta}^{(l)}) + \sum_{j=1}^p P_n(|\boldsymbol{\beta}_j^{(l)}|)$ strictly decreases. The coordinate descent algorithm runs as follows.

1. Set l = 1. Initialize $\boldsymbol{\beta}^{(1)} = \hat{\boldsymbol{\beta}}^*$, where $\hat{\boldsymbol{\beta}}^*$ solves

$$\min_{\beta \in \mathbb{R}^p} \frac{1}{n}\sum_{i=1}^n [g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta})]^2 + \sum_{j=1}^p P_n(|\beta_j|)$$

using the coordinate descent algorithm as in Fan and Lv (2011).

2. Successively for k = 1, …, p, let t^* be the minimizer of

$$\min_t \hat{L}_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t) + P_n'(|\beta_k^{(l)}|)|t|.$$

Update $\beta_k^{(l)}$ as t^* if

$$L_K(\boldsymbol{\beta}_{(-k)}^{(l)}, t^*) + P_n(|t^*|) < L_K(\boldsymbol{\beta}^{(l)}) + P_n(|\beta_k^{(l)}|).$$

Otherwise set $\beta_k^{(l)} = \beta_k^{(l-1)}$. Increase l by one when $k = p$.

3. Repeat Step 2 until $|Q_K(\boldsymbol{\beta}^{(l)}) - Q_K(\boldsymbol{\beta}^{(l+1)})| < \varepsilon$, for a pre-determined small ε.

When the second order approximation (7.1) is combined with SCAD in Step 2, the local linear approximation of SCAD is not needed. As demonstrated in Fan and Li (2001), when $P_n(t)$ is defined using SCAD, the penalized optimization of the form $\min_{\beta \in \mathbb{R}} \frac{1}{2}(z - \beta)^2 + \Lambda P_n(|\beta|)$ has an analytical solution.

We can show that the evaluated objective values $\{Q_K(\boldsymbol{\beta}^{(l)})\}_{l \geq 1}$ is a bounded Cauchy sequence. Hence for an arbitrarily small $\varepsilon > 0$, the above algorithm stops after finitely many steps. Let $M(\boldsymbol{\beta})$ denote the map defined by the algorithm from $\boldsymbol{\beta}^{(l)}$ to $\boldsymbol{\beta}^{(l+1)}$. We define a stationary point of the function $Q_K(\boldsymbol{\beta})$ to be any point $\boldsymbol{\beta}$ at which the gradient vector of $Q_K(\boldsymbol{\beta})$

is zero. Similar to the local linear approximation of Zou and Li (2008), we have the following result regarding the property of the algorithm.

Theorem 7.2—The sequence $\{Q_K(\beta^{(l)})\}_{l\geq 1}$ is a bounded non-increasing Cauchy sequence. Hence for any arbitrarily small $\varepsilon > 0$, the coordinate descent algorithm will stop after finitely many iterations. In addition, if $Q_K(\beta) = Q_K(M(\beta))$ only for stationary points of $Q_K(\cdot)$ and if β^* is a limit point of the sequence $\{(\beta^{(l)})_{l\geq 1}$, then β^* is a stationary point of $Q_K(\beta)$.

Theoretical analysis of non-convex regularization in the recent decade has focused on numerical procedures that can find local solutions (Hunter and Li 2005, Kim et al. 2008, Brehenry and Huang 2011). Proving that the algorithm achieves a solution that possesses the desired oracle properties is technically difficult. Our simulated results demonstrate that the proposed algorithm indeed reaches the desired sparse estimator. Further investigation along the lines of Zhang and Zhang (2012) and Loh and Wainwright (2013) is needed to investigate the statistical properties of the solution to non-convex optimization problems, which we leave as future research.

8. Monte Carlo Experiments

8.1. Endogeneity in both important and unimportant regressors

To test the performance of FGMM for variable selection, we simulate from a linear model:

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0 + \varepsilon, (\beta_{01}, \ldots, \beta_{05}) = (5, -4, 7, -2, 1.5); \beta_{0j} = 0, \text{for } 6 \leq j \leq p$$

with p = 50 or 200. Regressors are classified as being exogenous (independent of ε) and endogenous. For each component of **X**, we write $X_j = X_j^e$ if X_j is endogenous, and $X_j = X_j^x$ if X_j is exogenous, and X_j^e and X_j^x are generated according to:

$$X_j^e = (F_j + H_j + 1)(3\varepsilon + 1), X_j^x = F_j + H_j + u_j,$$

where $\{\varepsilon, u_1, \ldots, u_p\}$ are independent $N(0, 1)$. Here $\mathbf{F} = (F_1, \ldots, F_p)^T$ and $\mathbf{H} = (H_1, \ldots, H_p)^T$ are the transformations (to be specified later) of a three-dimensional instrumental variable $\mathbf{W} = (W_1, W_2, W_3)^T \sim N_3(0, I_3)$. There are m endogenous variables $(X_1, X_2, X_3, X_6, \ldots, X_{2+m})^T$, with m = 10 or 50. Hence three of the important regressors (X_1, X_2, X_3) are endogenous while two are exogenous (X_4, X_5).

We apply the Fourier basis as the working instruments:

$$\mathbf{F} = \sqrt{2}\{\sin(j\pi W_1) + \sin(j\pi W_2) + \sin(j\pi W_3) : j \leq p\},$$
$$\mathbf{H} = \sqrt{2}\{\cos(j\pi W_1) + \cos(j\pi W_2) + \cos(j\pi W_3) : j \leq p\}.$$

The data contain $n = 100$ i.i.d. copies of $(Y, \mathbf{X}, \mathbf{F}, \mathbf{H})$. PLS and FGMM are carried out separately for comparison. In our simulation we use SCAD with pre-determined tuning parameters of λ as the penalty function. The logistic cumulative distribution function with $h = 0.1$ is used for smoothing:

$$F(t)=\frac{\exp(t)}{1+\exp(t)}, K\left(\frac{\beta_j^2}{h}\right)=2F\left(\frac{\beta_j^2}{h}\right)-1.$$

There are 100 replications per experiment. Four performance measures are used to compare the methods. The first measure is the mean standard error (MSE_S) of the important regressors, determined by the average of $\|\hat{\beta}_S - \beta_{0S}\|$ over the 100 replications, where $S = \{1, \ldots, 5\}$. The second measure is the average of the MSE of unimportant regressors, denoted by MSE_N. The third measure is the number of correctly selected non-zero coefficients, that is, the true positive (TP), and finally, the fourth measure is the number of incorrectly selected coefficients, the false positive (FP). In addition, the standard error over the 100 replications of each measure is also reported. In each simulation, we initiate $\beta^{(0)} = (0,\ldots, 0)^T$, and run a penalized least squares (SCAD(λ)) for λ = 0.5 to obtain the initial value for the FGMM procedure. The results of the simulation are summarized in Table 2, which compares the performance measures of PLS and FGMM.

PLS has non-negligible false positives (FP). The average FP decreases as the magnitude of the penalty parameter increases, however, with a relatively large MSE_S for the estimated nonzero coefficients, and the FP rate is still large compared to that of FGMM. The PLS also misses some important regressors for larger λ. It is worth noting that the larger MSE_S for PLS is due to the bias of the least squares estimation in the presence of endogeneity. In contrast, FGMM performs well in both selecting the important regressors, and in correctly eliminating the unimportant regressors. The average MSE_S of FGMM is significantly less than that of PLS since the instrumental variable estimation is applied instead. In addition, after the regressors are selected by the FGMM, the post-FGMM further reduces the mean squared error of the estimators.

8.2. Endogeneity only in unimportant regressors

Consider a similar linear model but only the unimportant regressors are endogenous and all the important regressors are exogenous, as designed in Section 2.2, so the true model is as the usual case without endogeneity. In this case, we apply $(\mathbf{F}, \mathbf{H}) = (\mathbf{X}, \mathbf{X}^2)$ as the working instruments for FGMM with SCAD(λ) penalty, and need only data $\mathbf{X}$ and $\mathbf{Y} = (Y_1, \ldots, Y_n)$. We still compare the FGMM procedure with PLS. The results are reported in Table 3.

It is clearly seen that even though only the unimportant regressors are endogenous, however, the PLS still does not seem to select the true model correctly. This illustrates the variable selection inconsistency for PLS even when the true model has no endogeneity. In contrast, the penalized FGMM still performs relatively well.

8.3. Weak minimal signals

To study the effect on variable selection when the strength of the minimal signal is weak, we run another set of simulations with the same data generating process as in Design 1 but we change $\beta_4 = -0.5$ and $\beta_5 = 0.1$, and keep all the remaining parameters the same as before. The minimal nonzero signal becomes $|\beta_5| = 0.1$. Three of the important regressors are endogenous as in Design 1. Table 4 indicates that the minimal signal is so small that it is not easily distinguishable from the zero coefficients.

9. Conclusion and Discussion

Endogeneity can arise easily in the high-dimensional regression due to a large pool of regressors, which causes the inconsistency of the penalized least-squares methods and possible false scientific discoveries. Based on the over-identification assumption and valid instrumental variables, we propose to penalize an FGMM loss function. It is shown that FGMM possesses the oracle property, and the estimator is also a nearly global minimizer.

We would like to point out that this paper focuses on correctly specified sparse models, and the achieved results are “pointwise” for the true model. An important issue is the uniform inference where the sparse model may be locally misspecified. While the oracle property is of fundamental importance for high-dimensional methods in many scientific applications, it may not enable us to make valid inference about the coefficients uniformly across a large class of models (Leeb and Pötscher 2008, Belloni et al. 2013)[5]. Therefore, the “post-double-selection” method with imperfect model selection recently proposed by Belloni et al. (2013) is important for making uniform inference. Research along that line under high-dimensional endogeneity is important and we shall leave it for the future agenda.

Finally, as discussed in Bickel et al. (2009) and van de Geer (2008), high-dimensional regression problems can be thought of as an approximation to a nonparametric regression problem with a “dictionary” of functions or growing number of sieves. Then in the presence of endogenous regressors, model (3.1) is closely related to the non-parametric conditional moment restricted model considered by, e.g., Newey and Powell (2003), Ai and Chen (2003), and Chen and Pouzo (2008). While the penalization in the latter literature is similar to ours, it plays a different role and is introduced for different purposes. It will be interesting to find the underlying relationships between the two models.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We would like to thank the anonymous reviewers, Associate Editor and Editor for their helpful comments that helped to improve the paper.

This project was supported by the National Science Foundation grant DMS-1206464 and the National Institute of General Medical Sciences of the National Institutes of Health through Grant Numbers R01GM100474 and R01-

[5] We thank a referee for reminding us this important research direction.

GM072611. The bulk of the research was carried out while Yuan Liao was a postdoctoral fellow at Princeton University.

Appendix A: Proofs for Section 2

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses. Let sgn(·) denote the sign function.

A.1. Proof of Theorem 2.1

Proof. When $\hat{\beta}$ is a local minimizer of $Q_n(\beta)$, by the Karush-Kuhn-Tucker (KKT) condition, $\forall l \leq p$,

$$\frac{\partial L_n(\hat{\boldsymbol{\beta}})}{\partial \beta_l}+v_l=0,$$

where $v_l=P_n'(|\hat{\beta}_l|)\mathrm{sgn}(\hat{\beta}_l)$ if $\hat{\beta}_l \neq 0$; $v_l \in [-P_n'(0^+), P_n'(0^+)]$ if $\hat{\beta}_l = 0$, and we denote $P_n'(0^+)=\lim_{t\to 0^+} P_n'(t)$. By the monotonicity of $P_n'(t)$, we have $|\partial L_n(\hat{\boldsymbol{\beta}})/\partial\beta_l| \leq P_n'(0^+)$. By Taylor expansion and the Cauchy-Schwarz inequality, there is $\tilde{\beta}$ on the segment joining $\hat{\beta}$ and β_0 so that, on the event $\hat{\beta}_N = 0$, ($\hat{\beta}_j - \beta_{0j} = 0$ for all $j \notin S$)

$$\left|\frac{\partial L_n(\hat{\boldsymbol{\beta}})}{\partial \beta_l}-\frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l}|=|\sum_{j=1}^{p}\frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial\beta_l\partial\beta_j}(\hat{\beta}_j-\beta_{0j})|=|\sum_{j\in S}\frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial\beta_l\partial\beta_j}(\hat{\beta}_j-\beta_{0j})\right|.$$

The Cauchy-Schwarz inequality then implies that $\max_{l\leq p} |\partial L_n(\hat{\beta})/\partial\beta_l - \partial L_n(\beta_0)/\partial\beta_l|$ is bounded by

$$\max_{l,j\leq p}\left|\frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial\beta_l\partial\beta_j}\right|\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|_1 \leq \max_{l,j\leq p}\left|\frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial\beta_l\partial\beta_j}\right|\sqrt{s}\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|.$$

By our assumption, $\sqrt{s}\|\hat{\beta}_S - \beta_{0S}\| = o_p(1)$. Because $P(\hat{\beta}_N = 0) \to 1$,

$$\max_{l\leq p}\left|\frac{\partial L_n(\hat{\boldsymbol{\beta}})}{\partial \beta_l}-\frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l}\right|\to^p 0. \quad \text{(A.1)}$$

This yields that $\partial L_n(\beta_0)/\partial\beta_l = o_p(1)$.

A.2. Proof of Theorem 2.2

Proof. Let $\{X_{il}\}_{i=1}^n$ be the i.i.d. data of X_l where X_l is an endogenous regressor. For the penalized LS, $L_n(\beta)=\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\beta)^2$. Under the theorem assumptions, by the strong law of large number $\partial_{\beta_l} L_n(\boldsymbol{\beta}_0) = -\frac{2}{n}\sum_{i=1}^{n} X_{il}(Y_i - \mathbf{X}_i^T \boldsymbol{\beta}_0) \to -2E(X_l\varepsilon)$ almost surely, which does not satisfy (2.1) of Theorem 2.1.

Appendix B: General Penalized Regressions

We present some general results for the oracle properties of penalized regressions. These results will be employed to prove the oracle properties for the proposed FGMM. Consider a penalized regression of the form:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} L_n(\boldsymbol{\beta}) + \sum_{j=1}^{p} P_n(|\beta_j|),$$

Lemma B.1

Under Assumption 4.1, if ($\boldsymbol{\beta} = (\beta_1,\ldots,\beta_s)^T$ is such that $\max_{j\le s}|\beta_j - \beta_{0S,j}| \le d_n$, then

$$|\sum_{j=1}^{s} P_n(|\beta_j|) - P_n(|\beta_{0S,j}|)| \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \sqrt{s} P_n'(d_n).$$

Proof. By Taylor's expansion, there exists $\boldsymbol{\beta}^*$ ($\beta_j^* \neq 0$ for each j) lying on the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{0S}$, such that

$$\sum_{j=1}^{s}(P_n(|\beta_j|) - P_n(|\beta_{0S,j}|) = (P_n'(|\beta_1^*|)\mathrm{sgn}(\beta_1^*), \ldots, P_n'(|\beta_s^*|)\mathrm{sgn}(\beta_s^*))^T(\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}) \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \sqrt{s}\max_{j\le s} P_n'(|\beta_j^*|).$$

Then $\min\{|\beta_j^*|: j \le s\} \ge \min\{|\beta_{0S,j}|: j \le s\} - \max_{j\le s}|\beta_j^* - \beta_{0S,j}| \ge 2d_n - d_n = d_n$.

Since P_n' is non-increasing (as P_n is concave), $P_n'(|\beta_j^*|) \le P_n'(d_n)$ for all $j \le s$. Therefore $\sum_{j=1}^{s}(P_n(|\beta_j|) - P_n(|\beta_{0S,j}|) \le \|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \sqrt{s} P_n'(d_n)$.

In the theorems below, with $S = \{j : \beta_{0j} \neq 0\}$, define a so-called "oracle space" $\mathscr{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0$ if $j \notin S\}$. Write $L_n(\boldsymbol{\beta}_S, 0) = L_n(\boldsymbol{\beta})$ for $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, 0)^T \in \mathscr{B}$. Let $\boldsymbol{\beta}_S = (\beta_{S1},\ldots,\beta_{Ss})$ and

$$\nabla_S L_n(\boldsymbol{\beta}_S, 0) = \left(\frac{\partial L_n(\boldsymbol{\beta}_S, 0)}{\partial \beta_{S1}}, \ldots, \frac{\partial L_n(\boldsymbol{\beta}_S, 0)}{\partial \beta_{Ss}}\right)^T.$$

Theorem B.1 (Oracle Consistency)

Suppose Assumption 4.1 holds. In addition, suppose $L_n(\boldsymbol{\beta}_S, 0)$ is twice differentiable with respect to $\boldsymbol{\beta}_S$ in a neighborhood of $\boldsymbol{\beta}_{0S}$ restricted on the subspace $\mathscr{B}$, and there exists a positive sequence $a_n = o(d_n)$ such that:

i.

$$\|\nabla_S L_n(\boldsymbol{\beta}_{0S}, 0)\| = O_p(a_n).$$

ii. For any $\varepsilon > 0$, there is $C_\varepsilon > 0$ so that for all large n,

$$P(\lambda_{\min}(\nabla_S^2 L_n(\boldsymbol{\beta}_{0S}, 0)) > C_\varepsilon) > 1 - \varepsilon. \quad \text{(B.1)}$$

iii. For any $\varepsilon > 0$, $\delta > 0$, and any nonnegative sequence $\alpha_n = o(d_n)$, there is $N > 0$ such that when $n > N$,

$$P\left(\sup_{\|\beta_S - \beta_{0S}\| \le \alpha_n} \|\nabla_S^2 L_n(\boldsymbol{\beta}_S, 0) - \nabla_S^2 L_n(\boldsymbol{\beta}_{0S}, 0)\|_F \le \delta\right) > 1 - \varepsilon. \quad \text{(B.2)}$$

Then there exists a local minimizer $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S^T, 0)^T$ of

$$Q_n(\boldsymbol{\beta}_S, 0) = L_n(\boldsymbol{\beta}_S, 0) + \sum_{j \in S} P_n(|\beta_j|)$$

such that $\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = O_p(a_n + \sqrt{s} P_n'(d_n))$. In addition, for an arbitrarily small $\varepsilon > 0$, the local minimizer $\boldsymbol{\beta}$ is strict with probability at least $1 - \varepsilon$, for all large n.

Proof. The proof is a generalization of the proof of Theorem 3 in Fan and Lv (2011). Let $k_n = a_n + \sqrt{s} P_n'(d_n)$. It is our assumption that $k_n = o(1)$. Write $Q_1(\boldsymbol{\beta}_S) = Q_n(\boldsymbol{\beta}_S, 0)$, and $L_1(\boldsymbol{\beta}_S) = L_n(\boldsymbol{\beta}_S, 0)$. In addition, write

$$\nabla L_1(\boldsymbol{\beta}_S) = \frac{\partial L_n}{\partial \boldsymbol{\beta}_S}(\boldsymbol{\beta}_S, 0), \text{ and } \nabla^2 L_1(\boldsymbol{\beta}_S) = \frac{\partial^2 L_n}{\partial \boldsymbol{\beta}_S \boldsymbol{\beta}_S^T}(\boldsymbol{\beta}_S, 0).$$

Define $\mathcal{N}_\tau = \{\boldsymbol{\beta} \in \mathbb{R}^s : \|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \le k_n \tau\}$ for some $\tau > 0$. Let $\partial \mathcal{N}_\tau$ denote the boundary of $\mathcal{N}_\tau$. Now define an event

$$H_n(\tau) = \{Q_1(\boldsymbol{\beta}_{0S}) < \min_{\boldsymbol{\beta}_S \in \partial \mathcal{N}_\tau} Q_1(\boldsymbol{\beta}_S)\}.$$

On the event $H_n(\tau)$, by the continuity of Q_1, there exists a local minimizer of Q_1 inside $\mathcal{N}_\tau$. Equivalently, there exists a local minimizer $(\hat{\boldsymbol{\beta}}_S^T,0)^T$ of Q_n restricted on $\mathscr{B}=\{\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,0)^T\}$ inside $\{\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,0)^T:\boldsymbol{\beta}_S\in\mathcal{N}_\tau\}$. Therefore, it suffices to show that $\forall\varepsilon > 0$, there exists $\tau > 0$ so that $P(H_n(\tau)) > 1 - \varepsilon$ for all large n, and that the local minimizer is strict with probability arbitrarily close to one.

For any $\boldsymbol{\beta}_S \in \partial\mathcal{N}_\tau$, which is $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\| = k_n\tau$, there is $\boldsymbol{\beta}^*$ lying on the segment joining $\boldsymbol{\beta}_S$ and $\boldsymbol{\beta}_{0S}$ such that by the Taylor's expansion on $L_1(\boldsymbol{\beta}_S)$:

$$Q_1(\boldsymbol{\beta}_S)-Q_1(\boldsymbol{\beta}_{0S})=(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})^T\nabla L_1(\boldsymbol{\beta}_{0S})+\frac{1}{2}(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})^T\nabla^2L_1(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})+\sum_{j=1}^{s}[P_n(|\beta_{Sj}|)-P_n(|\beta_{0S,j}|)].$$

By Condition (i) $\|\nabla L_1(\boldsymbol{\beta}_{0S})\| = O_p(a_n)$, for any $\varepsilon > 0$, there exists $C_1 > 0$, so that the event H_1 satisfies $P(H_1) > 1 - \varepsilon/4$ for all large n, where

$$H_1=\{(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})^T\nabla L_1(\boldsymbol{\beta}_{0S})\ge -C_1\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|a_n\}. \quad \text{(B.3)}$$

In addition, Condition (ii) yields that there exists $C_\varepsilon > 0$ such that the following event H_2 satisfies $P(H_2) \ge 1 - \varepsilon/4$ for all large n, where

$$H_2=\{(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})^T\nabla^2L_1(\boldsymbol{\beta}_{0S})(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})>C_\varepsilon\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|^2\}. \quad \text{(B.4)}$$

Define another event $H_3 = \{\|\nabla^2L_1(\boldsymbol{\beta}_{0S}) - \nabla^2L_1(\boldsymbol{\beta}^*)\|_F < C_\varepsilon/4\}$. Since $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\| = k_n\tau$, by Condition (B.2) for any $\tau > 0$, $P(H_3) > 1$ — $\varepsilon/4$ for all large n. On the event $H_2 \cap H_3$, the following event H_4 holds:

$$H_4=\{(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})^T\nabla^2L_1(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S})>\frac{3C_\varepsilon}{4}\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|^2\}.$$

By Lemma B.1, $\sum_{j=1}^{s}[P_n(|\beta_{Sj}|)-P_n(|\beta_{0S,j}|)]\ge-\sqrt{s}P_n'(d_n)\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|$. Hence for any $\boldsymbol{\beta}_S \in \partial\mathcal{N}_\tau$, on $H_1 \cap H_4$,

$$Q_1(\boldsymbol{\beta}_S)-Q_1(\boldsymbol{\beta}_{0S})\ge k_n\tau\left(\frac{3k_n\tau C_\varepsilon}{8}-C_1a_n-\sqrt{s}P_n'(d_n)\right).$$

For $k_n=a_n+\sqrt{s}P_n'(d_n)$, we have $C_1a_n+\sqrt{s}P_n'(d_n)\le(C_1+1)k_n$. Therefore, we can choose $\tau > 8(C_1 + 1)/(3C_\varepsilon)$ so that $Q_1(\boldsymbol{\beta}_S)-Q_1(\boldsymbol{\beta}_{0S}) \ge 0$ uniformly for $\boldsymbol{\beta} \in \partial\mathcal{N}_\tau$. Thus for all large n, when $\tau > 8(C_1 + 1)/(3C_\varepsilon)$,

$$P(H_n(\tau)) \geq P(H_1 \cap H_4) \geq 1-\varepsilon.$$

It remains to show that the local minimizer in $\mathcal{N}_\tau$ (denoted by $\hat{\beta_S}$) is strict with a probability arbitrarily close to one. For each $h \in \mathbb{R}/\{0\}$, define

$$\psi(h)=\limsup_{\varepsilon\to 0^+} \sup_{\substack{t_1<t_2 \\ (t_1,t_2)\in(|h|-\varepsilon,|h|+\varepsilon)}} -\frac{P_n'(t_2)-P_n'(t_1)}{t_2-t_1}.$$

By the concavity of $P_n(\cdot)$, $\psi(\cdot) \geq 0$. We know that L_1 is twice differentiable on $\mathbb{R}^s$. For $\beta_S \in \mathcal{N}_\tau$ Let $\mathbf{A}(\beta_S) = \nabla^2 L_1(\beta_S) - \text{diag}\{\psi(\beta_{S1}), \ldots, \psi(\beta_{Ss})\}$. It suffices to show that $\mathbf{A}(\hat{\beta_S})$ is positive definite with probability arbitrarily close to one. On the event $H_5 = \{\eta(\hat{\beta_S}) \leq \sup_{\beta\in B(\beta_{0S},cd_n)} \eta(\beta)\}$ (where cd_n is as defined in Assumption 4.1(iv)),

$$\max_{j\leq s}\psi(\hat{\beta}_{S,j}) \leq \eta(\hat{\boldsymbol{\beta}}_S) \leq \sup_{\beta\in B(\beta_{0S},cd_n)} \eta(\boldsymbol{\beta})\}.$$

Also define events $H_6 = \{\|\nabla^2 L_1(\beta_S) - \nabla^2 L_1(\beta_{0S})\|_F < C_\varepsilon/4\}$ and $H_7 = \{\lambda_{\min}(\nabla^2 L_1(\beta_{0S})) > C_\varepsilon$. Then on $H_5 \cap H_6 \cap H7$, for any $\alpha \in \mathbb{R}^s$ satisfying $\|\alpha\| = 1$, by Assumption 4.1(iv),

$$\boldsymbol{\alpha}^T\mathbf{A}(\hat{\boldsymbol{\beta}}_S)\boldsymbol{\alpha} \geq \boldsymbol{\alpha}^T\nabla^2 L_1(\boldsymbol{\beta}_{0S})\boldsymbol{\alpha} - |\boldsymbol{\alpha}^T(\nabla^2 L_1(\hat{\boldsymbol{\beta}}_S) - \nabla^2 L_1(\boldsymbol{\beta}_{0S}))\boldsymbol{\alpha}| - \max_{j\leq s}\psi(\hat{\beta}_{S,j}) \geq 3C_\varepsilon/4 - \sup_{\beta\in B(\beta_{0S},d_n)} \eta(\boldsymbol{\beta}) \geq C_\varepsilon/4$$

for all large n. This then implies $\lambda_{\min}(\mathbf{A}(\hat{\beta_S})) \geq C_\varepsilon/4$ for all large n.

We know that $P(\lambda_{\min}[\nabla^2 L_1(\beta_{0S})] > C_\varepsilon) > 1 - \varepsilon$. It remains to show that $P(H_5 \cap H_6) > 1 — \varepsilon$ for arbitrarily small ε. Because $k_n = o(d_n)$, for an arbitrarily small $\varepsilon > 0$, $P(H_5) > P(\hat{\beta_S} \in B(\beta_{0S}, cd_n)) \geq 1 - \varepsilon/2$ for all large n. Finally,

$$\begin{aligned} P(H_6^c) &\leq P(H_6^c, \|\hat{\boldsymbol{\beta}}_S \\ &\quad - \boldsymbol{\beta}_{0S}\| \leq k_n) + P(\|\hat{\boldsymbol{\beta}}_S \\ &\quad - \boldsymbol{\beta}_{0S}\| > k_n) \leq P\left(\sup_{\|\beta_S-\beta_{0S}\|\leq k_n} \|\nabla^2 L_1(\boldsymbol{\beta}_S) - \nabla^2 L_1(\boldsymbol{\beta}_{0S})\|_F \geq C_\varepsilon/4\right) + \varepsilon/4 = \varepsilon/2. \end{aligned}$$

The previous theorem assumes that the true support S is known, which is not practical. We therefore need to derive the conditions under which S can be recovered from the data with probability approaching one. This can be done by demonstrating that the local minimizer of Q_n restricted on $\mathscr{B}$ is also a local minimizer on $\mathbb{R}^p$. The following theorem establishes the variable selection consistency of the estimator, defined as a local solution to a penalized regression problem on $\mathbb{R}^p$.

For any $\boldsymbol{\beta} \in \mathbb{R}^p$, define the projection function

$$\mathbb{T}\boldsymbol{\beta}=(\beta_1^{\prime},\beta_2^{\prime},\ldots,\beta_p^{\prime})^T \in \mathscr{B},\ \beta_j^{\prime}=\begin{cases}\beta_j & \text{if } j \in S\\ 0, & \text{if } j \notin S\end{cases}. \quad \text{(B.5)}$$

Theorem B.2(Variable selection)

Suppose $L_n : \mathbb{R}^p \to \mathbb{R}$ satisfies the conditions in Theorem B.1, and Assumption 4.1 holds. Assume the following Condition A holds:

Condition A: With probability approaching one, for $\hat{\boldsymbol{\beta}}_S$ in Theorem B.1, there exists a neighborhood $\mathscr{H} \subset \mathbb{R}^p$ of $(\hat{\boldsymbol{\beta}}_S^T,0)^T$, such that for all $\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,\boldsymbol{\beta}_N^T)^T \in \mathscr{H}$ but $\boldsymbol{\beta}_N \neq 0$,

$$L_n(\mathbb{T}\boldsymbol{\beta}) - L_n(\boldsymbol{\beta})<\sum_{j\notin S}P_n(|\beta_j|). \quad \text{(B.6)}$$

Then (i) with probability approaching one, $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S^T,0)^T$ is a local minimizer in $\mathbb{R}^p$ of

$$Q_n(\boldsymbol{\beta})=L_n(\boldsymbol{\beta})+\sum_{i=1}^{p}P_n(|\beta_i|)$$

(ii) For an arbitrarily small $\varepsilon > 0$, the local minimizer $\hat{\boldsymbol{\beta}}$ is strict with probability at least $1 - \varepsilon$, for all large n.

Proof. Let $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S^T,0)^T$ with $\hat{\boldsymbol{\beta}}_S$ being the local minimizer of $Q_1(\boldsymbol{\beta}_S)$ as in Theorem B.1. We now show: with probability approaching one, there is a random neighborhood of $\hat{\boldsymbol{\beta}}$, denoted by $\mathscr{H}$, so that $\forall \boldsymbol{\beta} = (\boldsymbol{\beta}_S, \boldsymbol{\beta}_N) \in \mathscr{H}$ with $\boldsymbol{\beta}_N \neq 0$, we have $Q_n(\hat{\boldsymbol{\beta}}) < Q_n(\boldsymbol{\beta})$. The last inequality is strict.

To show this, first note that we can take $\mathscr{H}$ sufficiently small so that $Q_1(\hat{\boldsymbol{\beta}}) \leq Q_1(\boldsymbol{\beta}_S)$ because $\hat{\boldsymbol{\beta}}_S$ is a local minimizer of $Q_1(\boldsymbol{\beta}_S)$ from Theorem B.1. Recall the projection defined to be $\mathbb{T}\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,0)^T$, and $Q_n(\mathbb{T}\boldsymbol{\beta}) = Q_1(\boldsymbol{\beta}_S)$ by the definition of Q_1. We have $Q_n(\hat{\boldsymbol{\beta}}) = Q_1(\hat{\boldsymbol{\beta}}_S) \leq Q_1(\boldsymbol{\beta}_S) = Q_n(\mathbb{T}\boldsymbol{\beta})$. Therefore, it suffices to show that with probability approaching one, there is a sufficiently small neighborhood of $\mathscr{H}$ of $\hat{\boldsymbol{\beta}}$, so that for any $\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T,\boldsymbol{\beta}_N^T)^T \in \mathscr{H}$ with $\boldsymbol{\beta}_N \neq 0$, $Q_n(\mathbb{T}\boldsymbol{\beta}) < Q_n(\boldsymbol{\beta})$.

In fact, this is implied by Condition (B.6):

$$Q_n(\mathbb{T}\boldsymbol{\beta}) - Q_n(\boldsymbol{\beta})=L_n(\mathbb{T}\boldsymbol{\beta}) - L_n(\boldsymbol{\beta}) - (\sum_{j=1}^{p}P_n(\beta_j) - \sum_{j=1}^{s}P_n(|(\mathbb{T}\boldsymbol{\beta})_j|))<0. \quad \text{(B.7)}$$

The above inequality, together with the last statement of Theorem B.1 implies part (ii) of the theorem.

Appendix C: Proofs for Section 4

Throughout the proof, we write $\mathbf{F}_{iS} = \mathbf{F}_i(\beta_{0S})$, $\mathbf{H}_{iS} = \mathbf{H}_i(\beta_{0S})$ and $\mathbf{V}_{iS}=(\mathbf{F}_{iS}^T, \mathbf{H}_{iS}^T)^T$.

Lemma C.1

- **i.** $\max_{l\leq p}|\frac{1}{n}\sum_{i=1}^n (F_{ij}-\overline{F}_j)^2 - \mathrm{var}(F_j)|=o_p(1).$
- **ii.** $\max_{l\leq p}|\frac{1}{n}\sum_{i=1}^n (H_{ij}-\overline{H}_j)^2 - \mathrm{var}(H_j)|=o_p(1).$
- **iii.** $\sup_{\beta\in\mathbb{R}^p} \lambda_{\max}(\mathbf{J}(\beta)) = O_p(1)$, and $\lambda_{\min}(\mathbf{J}(\beta_0))$ is bounded away from zero with probability approaching one.

Proof. Parts (i)(ii) follow from an application of the standard large deviation theory by using Bernstein inequality and Bonferroni's method. Part (iii) follows from the assumption that $\mathrm{var}(F_j)$ and $\mathrm{var}(H_j)$ are bounded uniformly in $j \leq p$.

C.1. Verifying conditions in Theorems B.1, B.2

C.1.1. Verifying conditions in Theorem B.1

For any $\boldsymbol{\beta}\in\mathbb{R}^p$, we can write $\mathbb{T}\boldsymbol{\beta}=(\boldsymbol{\beta}_S^T, 0)^T$. Define

$$\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_S)=\left[\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_S)\mathbf{V}_{iS}\right]^T \mathbf{J}(\boldsymbol{\beta}_0)\left[\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_S)\mathbf{V}_{iS}\right].$$

Then $\tilde{L}_{\mathrm{FGMM}}(\beta_S) = L_{\mathrm{FGMM}}(\beta_S, 0)$.

Condition (i)— $\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})=2\mathbf{A}_n(\boldsymbol{\beta}_{0S})\mathbf{J}(\boldsymbol{\beta}_0)\left[\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}\right]$, where

$$\mathbf{A}_n(\boldsymbol{\beta}_S)\equiv\frac{1}{n}\sum_{i=1}^n m(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_S)\mathbf{X}_{iS}\mathbf{V}_{iS}^T. \quad \text{(C.1)}$$

By Assumption 4.5, $\|\mathbf{A}_n(\beta_0)\| = O_p(1)$. In addition, the elements in $\mathbf{J}(\beta_0)$ are uniformly bounded in probability due to Lemma C.1. Hence

$\|\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})\| \leq O_p(1)\|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}\|$. Due to $Eg(Y, \mathbf{X}_s^T\boldsymbol{\beta}_{0S})\mathbf{V}_S=0$, using the exponential-tail Bernstein inequality with Assumption 4.3 plus Bonferroni inequality, it can be shown that there is $C > 0$ such that for any $t > 0$,

$$P(\max_{l\le p}|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})F_{li}|>t)<P(\max_{l\le p}P(|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})F_{li}|>t)\le \exp(\log p-Ct^2/n),$$

which implies $\max_{l\le p}|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})F_{li}|=O_p(\sqrt{\frac{\log p}{n}})$. Similarly, $\max_{l\le p}|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})H_{li}|=O_p(\sqrt{\frac{\log p}{n}})$. Hence $\|\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})\|=O_p(\sqrt{(s\log p)/n})$.

Condition (ii)—Straightforward but tedious calculation yields

$$\nabla^2\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})=\sum(\boldsymbol{\beta}_{0S})+\mathbf{M}(\boldsymbol{\beta}_{0S}),$$

where $\Sigma(\beta_{0S}) = 2\mathbf{A}_n(\beta_{0S})\mathbf{J}(\beta_0)\mathbf{A}_n(\beta_{0S})^T$, and $\mathbf{M}(\beta_{0S}) = 2\mathbf{Z}(\beta_{0S})\mathbf{B}(\beta_{0S})$, with (suppose $\mathbf{X}_{iS} = (X_{il_1}, \ldots, X_{il_s})^T$)

$$\boldsymbol{Z}(\boldsymbol{\beta}_{0S})=\frac{1}{n}\sum_{i=1}^{n}q_i(Y_i,\mathbf{X}_{iS}\boldsymbol{\beta}_{0S})(X_{il_1}\mathbf{X}_{iS},\ldots,X_{il_s}\mathbf{X}_{iS})\mathbf{V}_{iS}^T,\mathbf{B}(\boldsymbol{\beta}_{0S})=\mathbf{J}(\boldsymbol{\beta}_0)\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}.$$

It is not hard to obtain $\|\mathbf{B}(\boldsymbol{\beta}_{0S})\|_F=O_p(\sqrt{s\log p/n})$, and $\|\mathbf{Z}(\beta_{0S})\|_F = O_p(s)$, and hence $\|\mathbf{M}(\boldsymbol{\beta}_{0S})\|_F=O_p(s\sqrt{s\log p/n})=o_p(1)$.

Moreover, there is a constant $C > 0$, $P(\min_{j\in S}\hat{\mathrm{var}}(X_j)^{-1}>C)>1-\varepsilon$ and $P(\min_{j\le p}\hat{\mathrm{var}}(X_j^2)^{-1}>C)>1-\varepsilon$ for all large n and any $\varepsilon > 0$. This then implies $P(\lambda_{\min}[\mathbf{J}(\beta_0)] > \mathrm{C}) > 1 - \varepsilon$. Recall Assumption 4.5 that $\lambda_{\min}(E\mathbf{A}_n(\beta_{0S})E\mathbf{A}_n(\beta_{0S})^T) > C_2$ for some $C_2 > 0$. Define events

$$G_1=\{\lambda_{\min}[\mathbf{J}(\boldsymbol{\beta}_0)]>C\},\ G_2=\{\|\mathbf{M}(\boldsymbol{\beta}_{0S})\|<C_2C/5\}\ G_3=\{\|\mathbf{A}_n(\boldsymbol{\beta}_{0S})\mathbf{A}_n(\boldsymbol{\beta}_{0S})^T-E\mathbf{A}_n(\boldsymbol{\beta}_{0S})E\mathbf{A}_n(\boldsymbol{\beta}_{0S})^T)\|<C_2/5\}.$$

Then on the event $\cap_{i=1}^{3}G_i$,

$$\begin{aligned}\lambda_{\min}[\nabla^2\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})] &\ge 2\lambda_{\min}(\mathbf{J}(\boldsymbol{\beta}_0))\lambda_{\min}(\mathbf{A}_n(\boldsymbol{\beta}_{0S})\mathbf{A}_n(\boldsymbol{\beta}_{0S})^T)-\|\mathbf{M}(\boldsymbol{\beta}_{0S})\|_F \ge 2C[\lambda_{\min}(E\mathbf{A}_n(\boldsymbol{\beta}_{0S})E\mathbf{A}_n(\boldsymbol{\beta}_{0S})^T)\\ &\quad -C_2/5]\\ &\quad -C_2C/5\ge 7CC_2/5.\end{aligned}$$

Note that $P(\cap_{i=1}^{3} G_i) \geq 1 - \sum_{i=1}^{3} P(G_i^c) \geq 1 - 3\varepsilon$. Hence Condition (B.1) is then satisfied.

Condition (iii)—It can be shown that for any nonnegative sequence $\alpha_n = o(d_n)$ where $d_n = \min_{k \in S} |\beta_{0k}|/2$, we have

$$P(\sup_{\|\beta_S - \beta_{0S}\| \leq \alpha_n} \|\mathbf{M}(\boldsymbol{\beta}_S) - \mathbf{M}(\boldsymbol{\beta}_{0S})\|_F \leq \delta) > 1 - \varepsilon. \quad \text{(C.2)}$$

holds for any ε and $\delta > 0$. As for $\Sigma(\beta_S)$, note that for all β_S such that $\|\beta_S - \beta_{0S}\| < d_n/2$, we have $\beta_{S,k} \neq 0$ for all $k \leq s$. Thus $\mathbf{J}(\beta_S) = \mathbf{J}(\beta_{0S})$. Then $P(\sup_{\|\beta_S - \beta_{0S}\| < \alpha_n} \|\Sigma(\beta_S) - \Sigma(\beta_{0S})\|_F \leq \delta) > 1 - \varepsilon$ holds since $P(\sup_{\|\beta_S - \beta_{0S}\| < \alpha_n} \|\mathbf{A}_n(\beta_S) - \mathbf{A}_n(\beta_{0S})\|_F \leq \delta) > 1 - \varepsilon$.

C.1.2. Verifying conditions in Theorem B.2

Proof. We verify Condition A of Theorem B.2, that is, with probability approaching one, there is a random neighborhood $\mathscr{H}$ of $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S^T, 0)^T$, such that for any $\boldsymbol{\beta} = (\beta_{\boldsymbol{S}}^T, \boldsymbol{\beta}_N^T)^T \in \mathscr{H}$ with $\beta_N \neq 0$, condition (B.6) holds.

Let $\mathbf{F}(\mathbb{T}\beta) = \{F_l : l \in S, \beta_l \neq 0\}$ and $\mathbf{H}(\mathbb{T}\beta) = \{H_l : l \in S, \beta_l \neq 0\}$ for any fixed $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, \boldsymbol{\beta}_N^T)^T$. Define

$$\Xi(\boldsymbol{\beta}) = \left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\right]^T \mathbf{J}_1(\mathbb{T}\boldsymbol{\beta}) \left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\right] + \left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta})\mathbf{H}_i(\mathbb{T}\boldsymbol{\beta})\right]^T \mathbf{J}_2(\mathbb{T}\boldsymbol{\beta}) \left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta})\right.$$

where $\mathbf{J}_1(\mathbb{T}\beta)$ and $\mathbf{J}_2(\mathbb{T}\beta)$ are the upper-$|S|_0$ and lower-$|S|_0$ sub matrices of $\mathbf{J}(\mathbb{T}\beta)$. Hence $L_{\mathrm{FGMM}}(\mathbb{T}(\beta)) = \Xi(\mathbb{T}\beta)$. Then $L_{\mathrm{FGMM}}(\beta) - \Xi(\beta)$ equals

$$\sum_{l \notin S, \beta_l \neq 0} \left[w_{l1} \left(\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{X}_i^T \boldsymbol{\beta}) F_{il}\right)^2 + w_{l2} \left(\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{X}_i^T \boldsymbol{\beta}) H_{il}\right)^2 \right],$$

where $w_{l1} = 1/\mathrm{var}(\hat{F}_l$ and $w_{l2} = 1/\mathrm{var}(\hat{H}_l$. So $L_{\mathrm{FGMM}}(\beta) \geq \Xi(\beta)$. This then implies $L_{\mathrm{FGMM}}(\mathbb{T}\beta) - L_{\mathrm{FGMM}}(\beta) \leq \Xi(\mathbb{T}\beta) - \Xi(\beta)$. By the mean value theorem, there exists $\lambda \in (0,1)$, for $\mathbf{h} = (\boldsymbol{\beta}_S^T, -\lambda\boldsymbol{\beta}_N^T)^T$,

$$\Xi(\mathbb{T}\boldsymbol{\beta}) - \Xi(\boldsymbol{\beta})$$
$$= \sum_{l \notin S, \beta_l \neq 0} \beta_l \left[\frac{1}{n}\sum_{i=1}^{n} X_{il} m(Y_i, \mathbf{X}_i^T \mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\right]^T \mathbf{J}_1(\mathbb{T}\boldsymbol{\beta}) \left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\right] + \sum_{l \notin S, \beta_l \neq 0} \beta_l \left[\frac{1}{n}\sum_{i=1}^{n} X_{il} m(Y_i, \mathbf{X}_i^T \mathbf{h})\mathbf{H}_i(\mathbb{T}\right.$$

Let $\mathscr{H}$ be a neighborhood of $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S^T, 0)^T$ (to be determined later). We have shown that $\Xi(\mathbb{T}\boldsymbol{\beta}) - \Xi(\boldsymbol{\beta}) = \Sigma_{l\notin S, \beta_l\neq 0}\, \beta_l(a_l(\boldsymbol{\beta}) + b_l(\boldsymbol{\beta}))$, for any $\boldsymbol{\beta} \in \mathscr{H}$,

$$a_l(\boldsymbol{\beta})=\left[\frac{1}{n}\sum_{i=1}^{n} X_{il} m(Y_i, \mathbf{X}_i^T\mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\right]^T \mathbf{J}_1(\mathbb{T}\boldsymbol{\beta}) \left[\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T\mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\right],$$

and $b_l(\boldsymbol{\beta})$ is defined similarly based on $\mathbf{H}$. Note that $\mathbf{h}$ lies in the segment joining $\boldsymbol{\beta}$ and $\mathbb{T}\boldsymbol{\beta}$, and is determined by $\boldsymbol{\beta}$, hence should be understood as a function of $\boldsymbol{\beta}$. By our assumption, there is a constant M, such that $|m(t_1, t_2)|$ and $|q(t_1, t_2)|$, the first and second partial derivatives of g, and $EX_l^2F_k^2$ are all bounded by M uniformly in t_1, t_2 and $l, k \le p$. Therefore the Cauchy-Schwarz and triangular inequalities imply

$$\|\frac{1}{n}\sum_{i=1}^{n} X_{il} m(Y_i, \mathbf{X}_i^T\mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\|^2 \le M^2 \max_{l\le p} |\frac{1}{n}\sum_{i=1}^{n} \|X_{il}\mathbf{F}_{iS}\|^2 - E\|X_l\mathbf{F}_S\|^2| + M^2 \max_{l\notin S} E\|X_l\mathbf{F}_S\|^2.$$

Hence there is a constant M_1 such that if we define the event (again, keep in mind that $\mathbf{h}$ is determined by $\boldsymbol{\beta}$)

$$B_n=\{\sup_{\beta\in\mathscr{H}} \|\frac{1}{n}\sum_{i=1}^{n} X_{il} m(Y_i, \mathbf{X}_i^T\mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\| < \sqrt{s}M_1,\ \sup_{\beta\in\mathscr{H}} \|\mathbf{J}_1(\mathbb{T}\boldsymbol{\beta})\| < M_1\},$$

then $P(B_n) \to 1$. In addition with probability one,

$$\|\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T\mathbf{h})\mathbf{F}_i(\mathbb{T}\boldsymbol{\beta})\| \le \sup_{\beta\in\mathscr{H}} \|\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T\boldsymbol{\beta})\mathbf{F}_{iS}\| \le \sup_{\beta\in\mathscr{H}} \|\frac{1}{n}\sum_{i=1}^{n} [g(Y_i, \mathbf{X}_i^T\boldsymbol{\beta}) - g(Y_i, \mathbf{X}_i^T\hat{\boldsymbol{\beta}})]\mathbf{F}_{iS}\|$$
$$+\|\frac{1}{n}\sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T\hat{\boldsymbol{\beta}})\mathbf{F}_{iS}\| \equiv Z_1 + Z_2,$$

where, $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S^T, 0)^T$. For some deterministic sequence r_n (to be determined later), we can define the above $\mathscr{H}$ to be dddd

$$\mathscr{H}=\{\boldsymbol{\beta}: \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\| < r_n/p\}$$

then $\sup_{\beta\in\mathscr{H}} \|\boldsymbol{\beta} - \hat{\boldsymbol{\beta}}\|_1 < r_n$. By the mean value theorem and Cauchy Schwarz inequality, there is $\tilde{\boldsymbol{\beta}}$:

$$Z_1=\sup_{\beta\in\mathscr{H}}\|\frac{1}{n}\sum_{i=1}^{n}m(Y_i,\mathbf{X}_i^T\tilde{\boldsymbol{\beta}})\mathbf{F}_{iS}\mathbf{X}_i^T(\boldsymbol{\beta}-\hat{\boldsymbol{\beta}})\| \leq \sqrt{s}\sup_{\beta\in\mathscr{H}}\|\frac{1}{n}\sum_{i=1}^{n}m(Y_i,\mathbf{X}_i^T\tilde{\boldsymbol{\beta}})\mathbf{F}_{iS}\mathbf{X}_i^T\|_{\infty} r_n \leq M\sqrt{s}\max_{k\in S,l\leq p}|\frac{1}{n}\sum_{i=1}^{n}(F_{ik}X_{il})^2|^{1/2}r_n.$$

Hence there is a constant M_2 such that $P(Z_1 < M_2\sqrt{s}r_n) \to 1$.

Let $\varepsilon_i=g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta}_0)$. By the triangular inequality and mean value theorem, there are $\tilde{\mathbf{h}}$ and $\tilde{\tilde{\mathbf{h}}}$ lying in the segment between $\hat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ such that

$$\begin{aligned}Z_2 \leq\ &\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i\mathbf{F}_{iS}\| +\|\frac{1}{n}\sum_{i=1}^{n}m(Y_i,\mathbf{X}_i^T\tilde{\mathbf{h}})\mathbf{F}_{iS}\mathbf{X}_{iS}^T(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S})\| \leq \sqrt{s}\max_{j\leq p}|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_iF_{ij}|\\ &+\|\frac{1}{n}\sum_{i=1}^{n}m(Y_i,\mathbf{X}_i^T\boldsymbol{\beta}_0)\mathbf{F}_{iS}\mathbf{X}_{iS}^T(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S})\|\\ &+\|\frac{1}{n}\sum_{i=1}^{n}q(Y_i,\mathbf{X}_i^T\tilde{\tilde{\mathbf{h}}})\mathbf{X}_{iS}^T(\boldsymbol{\beta}_{0S}-\tilde{\mathbf{h}}_S)\mathbf{F}_{iS}\mathbf{X}_{iS}^T(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S})\| \leq O_p(\sqrt{s\log p/n})\\ &+(o_p(1)+\|Em(Y,\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{F}_S\mathbf{X}_S^T\|)\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|(\frac{1}{n}\sum_{i=1}^{n}\|q(Y_i,\mathbf{X}_i^T\tilde{\tilde{\mathbf{h}}})\mathbf{X}_{iS}\|^2)^{1/2}(\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_{iS}\|^2\|\mathbf{F}_{iS}\|^2)^{1/2}\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2,\end{aligned}$$

where we used the assumption that $\|Em(Y,\mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}_S\mathbf{F}_S^T\|=O(1)$. We showed that $\|\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})\|=O_p(\sqrt{(s\log p)/n})$ in the proof of verifying conditions in Theorem B.1. Hence by Theorem B.1, $\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|=O_p(\sqrt{s\log p/n}+\sqrt{s}P_n'(d_n))$. Thus

$$Z_2=O_p(\sqrt{\frac{s\log p}{n}}+\sqrt{s}P_n'(d_n)+\frac{s^2\sqrt{s}\log s}{n}+s^2\sqrt{s}P_n'(d_n)^2)\equiv O_p(\xi_n).$$

By the assumption $\sqrt{s}\xi_n=o(P_n^{'}(0^+))$, hence $P(Z_2<P_n^{'}(0^+)/(8\sqrt{s}M_1^2))\to 1$, where M_1 is defined in the event B_n. Consequently, if we define an event $D_n=\{Z_1<M_2\sqrt{s}r_n, Z_2<P_n^{'}(0^+)/(8\sqrt{s}M_1^2)\}$, then $P(B_n \cap D_n \to 1$, and on the event $B_n \cap D_n$,

$$\sup_{\beta\in\mathscr{H}}|a_l(\boldsymbol{\beta})| \le M_1^2\sqrt{s}(M_2\sqrt{s}r_n+P_n^{'}(0^+))/(8\sqrt{s}M_1^2)=M_1^2M_2sr_n+P_n^{'}(0^+)/8.$$

We can $r_n<P_n^{'}(0^+)/(8M_1^2M_2s)$, and thus $\sup_{\beta\in\mathscr{H}}|a_l(\boldsymbol{\beta})| \le P_n^{'}(0^+)/4$.

On the other hand, Because ($\mathbb{T}\boldsymbol{\beta})_j = \boldsymbol{\beta}_j$ for either $j \in S$ or $\beta_j = 0$, there exists $\lambda_2 \in (0,1)$,

$$\sum_{j=1}^{p}(P_n(|\beta_j|)-P_n(|(\mathbb{T}\boldsymbol{\beta})_j|)=\sum_{j\notin S}P_n(|\beta_J|)=\sum_{l\notin S,\beta_l\neq 0}|\beta_l|P_n^{'}(\lambda_2|\beta_l|).$$

For all $l \notin S$, $|\beta_l| \le \|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_1 < r_n$. Due to the non-increasingness of $P_n^{'}(t)$, $\sum_{l\notin S}P_n(|\beta_l|)\ge\sum_{l\notin S,\beta_l\neq 0}|\beta_l|P_n^{'}(r_n)$. We can make r_n further smaller so that $P_n^{'}(r_n)\ge P_n^{'}(0^+)/2$ which is satisfied for example, when $r_n < \lambda_n$ if SCAD(λ_n) is used as the penalty. Hence

$$\sum_{l\notin S}\beta_l a_l(\boldsymbol{\beta}) \le \sum_{l\notin S}|\beta_l|\frac{P_n^{'}(0^+)}{4} \le \sum_{l\notin S}|\beta_l|\frac{P_n^{'}(r_n)}{2} \le \frac{1}{2}\sum_{l\notin S}P_n(|\beta_l|).$$

Using the same argument we can show $\sum_{l\notin S}\beta_l b_l(\boldsymbol{\beta}) \le \frac{1}{2}\sum_{l\notin S}P_n(|\beta_l|)$. Hence $L_{\text{FGMM}}(\mathbb{T}\boldsymbol{\beta}) - L_{\text{FGMM}}(\boldsymbol{\beta}) < \Sigma_{l\notin S,\beta_l\neq 0}\,\beta_l\,(a_l(\boldsymbol{\beta})+b_l(\boldsymbol{\beta})) \le \Sigma_{l\notin S}\,P_n(|\beta_l|)$ for all $\boldsymbol{\beta}\in\{\boldsymbol{\beta}: \|\boldsymbol{\beta}-\hat{\boldsymbol{\beta}}\|_1 < r_n\}$ under the event $B_n \cap D_n$. Here r_n is such that $r_n<P_n^{'}(0^+)/(8M_1^2M_2s)$ and $P_n^{'}(r_n)\ge P_n^{'}(0^+)/2$. This proves Condition A of Theorem B.2 due to $P(B_n \cap D_n) \to 1$.

C.2. Proof of Theorem 4.1: parts (ii) (iii)

We apply Theorem B.2 to infer that with probability approaching one, $\hat{\boldsymbol{\beta}}=(\hat{\boldsymbol{\beta}}_S^T,0)^T$ is a local minimizer of $Q_{\text{FGMM}}(\boldsymbol{\beta})$. Note that under the event that $(\hat{\boldsymbol{\beta}}_S^T,0)^T$ is a local minimizer of $Q_{\text{FGMM}}(\boldsymbol{\beta})$, we then infer that $Q_n(\boldsymbol{\beta})$ has a local minimizer $(\hat{\boldsymbol{\beta}}_S^T,\hat{\boldsymbol{\beta}}_N^T)^T$ such that $\hat{\boldsymbol{\beta}}_N = 0$. This reaches the conclusion of part (ii). This also implies $P(\hat{S} \subset S) \to 1$.

By Theorem B.1, and $\|\nabla\tilde{L}_{\text{FGMM}}(\boldsymbol{\beta}_{0S})\|=O_p(\sqrt{(s\log p)/n})$ as proved in verifying conditions in Theorem B.1, we have $\|\boldsymbol{\beta}_{0S}-\hat{\boldsymbol{\beta}}_S\| = o_p(d_n)$. So

$$P(S\not\subset\hat{S})=P(\exists j\in S,\hat{\beta}_j=0)\leq P(\exists j\in S,|\beta_{0j}-\hat{\beta}_j|\geq|\beta_{0j}|)\leq P(\max_{j\in S}|\beta_{0j}-\hat{\beta}_j|\geq d_n)\leq P(\|\boldsymbol{\beta}_{0S}-\hat{\boldsymbol{\beta}}_S\|\geq d_n)=o(1)$$

This implies $P(S \subset \hat{S}) \to 1$. Hence $P(\hat{S} = S) \to 1$.

C.3. Proof of Theorem 4.1: part (i)

Let $P_n'(|\hat{\boldsymbol{\beta}}_S|)=(P_n'(|\hat{\beta}_{S1}|),\ldots,P_n'(|\hat{\beta}_{Ss}|))^T$.

Lemma C.2

Under Assumption 4.1,

$$\|P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)\|=O_p(\max_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|\leq d_n/4}\eta(\boldsymbol{\beta})\sqrt{s\mathrm{log}p/n}+\sqrt{s}P_n'(d_n)),$$

where $\circ$ denotes the element-wise product.

Proof. Write $P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)=(v_1,\ldots,v_s)^T$, where $v_i=P_n'(|\hat{\beta}_{Si}|)\mathrm{sgn}(\hat{\beta}_{Si})$. By the triangular inequality and Taylor expansion,

$$|v_i|\leq|P_n'(|\hat{\beta}_{Si}|)-P_n'(|\beta_{0S,i}|)|+P_n'(|\beta_{0S,i}|)\leq\eta(\boldsymbol{\beta}^*)|\hat{\beta}_{Si}-\beta_{0S,i}|+P_n'(d_n)$$

where $\boldsymbol{\beta}^*$ lies on the segment joining $\hat{\boldsymbol{\beta}}_S$ and $\boldsymbol{\beta}_{0S}$. For any $\varepsilon > 0$ and all large n,

$$P(\eta(\boldsymbol{\beta}^*)>\max_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|\leq d_n/4}\eta(\boldsymbol{\beta}))\leq P(\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|>d_n/4)<\varepsilon.$$

This implies $\eta(\boldsymbol{\beta}^*) = O_p(\max_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|<d_n/4}\eta(\boldsymbol{\beta}))$. Therefore, $\|P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)\|^2=\sum_{i=1}^s v_j^2$ is upper-bounded by

$$2\max_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|\leq d_n/4}\eta(\boldsymbol{\beta})^2\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2+2sP_n'(d_n)^2,$$

which implies the result since $\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|=O_p(\sqrt{s\mathrm{log}p/n}+\sqrt{s}P_n'(d_n))$.

Lemma C.3

Let $\Omega_n = \sqrt{n}\Gamma^{-1/2}$. Then for any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$,

$$\boldsymbol{\alpha}^T\boldsymbol{\Omega}_n\nabla\tilde{L}_{FGMM}(\boldsymbol{\beta}_{0S})\rightarrow^d N(0,1).$$

Proof. We have $\nabla \tilde{L}_{\mathrm{FGMM}}(\beta_{0S}) = 2\mathbf{A}_n(\beta_{0S})\mathbf{J}(\beta_0)\mathbf{B}_n$, where $\mathbf{B}_n=\frac{1}{n}\sum_{i=1}^n g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}$. We write $\mathbf{A}=Em(Y,\mathbf{X}_s^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S^T$, $\boldsymbol{\Upsilon}=\mathrm{var}(\sqrt{n}\mathbf{B}_n)=\mathrm{var}(g(Y,\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{V}_S)$ and $\boldsymbol{\Gamma} = 4\mathbf{A}\mathbf{J}(\beta_0)\Upsilon\mathbf{J}(\beta_0)^T\mathbf{A}^T$.

By the weak law of large number and central limit theorem for iid data,

$$\|\mathbf{A}_n(\boldsymbol{\beta}_{0S})-\mathbf{A}\|=o_p(1),\ \ \sqrt{n}\tilde{\boldsymbol{\alpha}}^T\boldsymbol{\Upsilon}^{-1/2}\mathbf{B}_n\rightarrow^d N(0,1).$$

for any unit vector $\tilde{\alpha}\in\mathbb{R}^{2s}$. Hence by the Slutsky's theorem,

$$\sqrt{n}\boldsymbol{\alpha}^T\boldsymbol{\Gamma}^{-1/2}\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})\rightarrow^d N(0,1).$$

Proof of Theorem 4.1: part (i)

Proof. The KKT condition of $\hat{\beta}_S$ gives

$$-P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)=\nabla\tilde{L}_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_S), \quad \text{(C.3)}$$

By the mean value theorem, there exists β^* lying on the segment joining β_{0S} and $\hat{\beta}_S$ such that

$$\nabla\tilde{L}_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_S)=\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})+\nabla^2\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}^*)(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}).$$

Let $\mathbf{D} = (\nabla^2\tilde{L}_{\mathrm{FGMM}}(\beta^*) - \nabla^2\tilde{L}_{\mathrm{FGMM}}(\beta_{0S}))(\hat{\beta}_S - \beta_{0S})$. It then follows from (C.3) that for $\boldsymbol{\Omega}_n=\sqrt{n}\boldsymbol{\Gamma}_n^{-1/2}$, and any unit vector α,

$$\boldsymbol{\alpha}^T\boldsymbol{\Omega}_n\nabla^2\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})(\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S})=-\boldsymbol{\alpha}^T\boldsymbol{\Omega}_n[P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)+\nabla\tilde{L}_{\mathrm{FGMM}}(\boldsymbol{\beta}_{0S})+\mathbf{D}].$$

In the proof of Theorem 4.1, condition (ii), we showed that $\nabla^2\tilde{L}_{\mathrm{FGMM}}(\beta_{0S}) = \Sigma + O_p(1)$. Hence by Lemma C.3, it suffices to show $\boldsymbol{\alpha}^T\boldsymbol{\Omega}_n[P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)+\mathbf{D}]=o_p(1)$.

By Assumptions 4.5 and 4.6(i), $\lambda_{\min}(\Gamma_n)^{-1/2} = O_p(1)$.Thus $\|\alpha^T\Omega_n\| = O_p(\sqrt{n})$. Lemma C.2 then implies $\lambda_{\max}(\boldsymbol{\Omega}_n)\|P_n'(|\hat{\boldsymbol{\beta}}_S|)\circ\mathrm{sgn}(\hat{\boldsymbol{\beta}}_S)\|$ is bounded by $O_p(\sqrt{n})(\max_{\|\boldsymbol{\beta}_S-\boldsymbol{\beta}_{0S}\|\le d_n/4}\eta(\boldsymbol{\beta})\sqrt{s\log p/n}+\sqrt{s}P_n'(d_n))=o_p(1)$.

It remains to prove $\|\mathbf{D}\| = o_p(n^{-1/2})$, and it suffices to show that

$$\|\nabla^2 \tilde{L}_{\text{FGMM}}(\boldsymbol{\beta}^*) - \nabla^2 \tilde{L}_{\text{FGMM}}(\boldsymbol{\beta}_{0S})\| = o_p((s\log p)^{-1/2}) \quad \text{(C.4)}$$

due to $\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = O_p(\sqrt{s\log p/n} + \sqrt{s}P_n'(d_n))$, and Assumption 4.6 that $\sqrt{ns}P_n'(d_n) = o(1)$. Showing (C.4) is straightforward given the continuity of $\nabla^2 \tilde{L}_{\text{FGMM}}$.

Appendix D: Proofs for Sections 5 and 6

The local minimizer in Theorem 4.1 is denoted by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_S^T, \hat{\boldsymbol{\beta}}_N^T)^T$ and $P(\hat{\beta}_N) \to 1$. Let $\hat{\boldsymbol{\beta}}_G = (\hat{\boldsymbol{\beta}}_S^T, 0)^T$.

D.1. Proof of Theorem 5.1

Lemma D.1

$$L_{FGMM}(\hat{\boldsymbol{\beta}}_G) = O_p(s\log p/n + sP_n'(d_n)^2)$$

Proof. We have, $L_{\text{FGMM}}(\hat{\boldsymbol{\beta}}_G) \le \|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\hat{\boldsymbol{\beta}}_S)\mathbf{V}_{iS}\|^2 O_p(1)$. By Taylor expansion, with some $\hat{\beta}$ in the segment joining β_{0S} and $\hat{\beta}_S$,

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\hat{\boldsymbol{\beta}}_S)\mathbf{V}_{iS}\| \le \|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}\| \\
+\|\frac{1}{n}\sum_{i=1}^n m(Y_i, \mathbf{X}_{iS}^T\tilde{\boldsymbol{\beta}}_S)\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\hat{\boldsymbol{\beta}}_S \\
- \boldsymbol{\beta}_{0S}\| \le O_p(\sqrt{s\log p/n}) \\
+\|\frac{1}{n}\sum_{i=1}^n m(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\hat{\boldsymbol{\beta}}_S \\
- \boldsymbol{\beta}_{0S}\| + \frac{1}{n}\sum_{i=1}^n |m(Y_i, \mathbf{X}_{iS}^T\tilde{\boldsymbol{\beta}}_S) \\
- m(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})|\|\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|.
\end{aligned}$$

Note that $\|Em(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S\|$ is bounded due to Assumption 4.5. Apply Taylor expansion again, with some $\hat{\beta}^*$, the above term is bounded by

$$O_p(\sqrt{s\log p/n}) + O_p(1)\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| + \frac{1}{n}\sum_{i=1}^n |q(Y_i, \mathbf{X}_{iS}^T\tilde{\boldsymbol{\beta}}_S^*)|\|\mathbf{X}_{iS}\|\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|\|\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\hat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|.$$

Note that $\sup_{t_1,t_2}| q(t_1, t_2)| < \infty$ by Assumption 4.4. The second term in the above is bounded by $C\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_{iS}\|\|\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2$. Combining these terms, $\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\hat{\boldsymbol{\beta}}_S)\mathbf{V}_{iS}\|$ is bounded by $O_p(\sqrt{s\log p/n}+\sqrt{s}P_n'(d_n))+O_p(s\sqrt{s})\|\hat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2=O_p(\sqrt{s\log p/n}+\sqrt{s}P_n'(d_n))$.

Lemma D.2

Under the theorem's assumptions

$$Q_{FGMM}(\hat{\boldsymbol{\beta}}_G)=O_p\left(\frac{s\log p}{n}+sP_n'(d_n)^2+s\max_{j\in S}P_n(|\beta_{0j}|)+P_n'(d_n)s\sqrt{\frac{\log s}{n}}\right).$$

Proof. By the foregoing lemma, we have

$$Q_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_G)=O_p\left(\frac{s\log p}{n}+sP_n'(d_n)^2\right)+\sum_{j=1}^{s}P_n(|\beta_{Sj}|).$$

Now, for some $\tilde{\beta_{Sj}}$ in the segment joining $\hat{\beta_{Sj}}$ and β_{0j},

$$\begin{aligned}\sum_{j=1}^{s}P_n(|\beta_{Sj}|)&\le\sum_{j=1}^{s}P_n(|\beta_{0S,j}|)\\&+\sum_{j=1}^{s}P_n'(|\tilde{\boldsymbol{\beta}}_{Sj}|)|\boldsymbol{\beta}_{Sj}\\&-\boldsymbol{\beta}_{0S,j}|\le s\max_{j\in S}P_n(|\beta_{0j}|)\\&+\sum_{j=1}^{s}P_n'(d_n)|\beta_{Sj}\\&-\beta_{0S,j}|\le s\max_{j\in s}P_n(|\beta_{0j}|)\\&+P_n'(d_n)\|\hat{\beta}_S-\beta_{0S}\|\sqrt{s}.\end{aligned}$$

The result then follows.

Note that $\forall\delta>0$,

$$\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}}Q_{\mathrm{FGMM}}(\boldsymbol{\beta})\ge\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}}L_{\mathrm{FGMM}}(\boldsymbol{\beta})\ge\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}}\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}_i(\boldsymbol{\beta})\|^2\min_{j\le p}\left\{\hat{\mathrm{var}}(X_j),\hat{\mathrm{var}}(X_j^2)\right\}.$$

Hence by Assumption 5.1, there exists $\gamma>0$,

$$P(\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}} Q_{\mathrm{FGMM}}(\boldsymbol{\beta})>2\gamma)\to 1.$$

On the other hand, by Lemma D.2, $Q_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_G) = o_p(1)$. Therefore,

$$\begin{aligned}
& P(Q_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}) \\
& \quad +\gamma>\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}} Q_{\mathrm{FGMM}}(\boldsymbol{\beta}))P(Q_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_G) \\
& \quad\ +\gamma>\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}} Q_{\mathrm{FGMM}}(\boldsymbol{\beta})) \\
& \quad +o(1)\le P(Q_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_G) \\
& \quad\ +\gamma>2\gamma) \\
& \quad +P(\inf_{\boldsymbol{\beta}\notin\Theta_\delta\cup\{0\}} Q_{\mathrm{FGMM}}(\boldsymbol{\beta})<2\gamma) \\
& \quad +o(1)\le P(Q_{\mathrm{FGMM}}(\hat{\boldsymbol{\beta}}_G)>\gamma) \\
& \quad +o(1)=o(1).
\end{aligned}$$

Q.E.D.

D.2. Proof of Theorem 6.1

Lemma D.3

Define $\rho(\boldsymbol{\beta}_S)=E[g(Y,\mathbf{X}_S^T\boldsymbol{\beta}_S)\sigma(\mathbf{W})^{-2}\mathbf{D}(\mathbf{W})]$. Under the theorem assumptions, $\sup_{\boldsymbol{\beta}_S\in\Theta}\|\rho(\boldsymbol{\beta}_S)-\rho_n(\boldsymbol{\beta}_S)\|=o_p(1)$.

Proof. We first show three convergence results:

$$\sup_{\boldsymbol{\beta}_S\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\|g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_S)(\mathbf{D}(\mathbf{W}_i)-\hat{\mathbf{D}}(\mathbf{W}_i))\hat{\sigma}(\mathbf{W}_i)^{-2}\|=o_p(1), \quad \text{(D.1)}$$

$$\sup_{\boldsymbol{\beta}_S\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\|g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_S)\mathbf{D}(\mathbf{W}_i)(\hat{\sigma}(\mathbf{W}_i)^{-2}-\sigma(\mathbf{W}_i)^{-2})\|=o_p(1), \quad \text{(D.2)}$$

$$\sup_{\boldsymbol{\beta}_S\in\Theta}\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_S)\mathbf{D}(\mathbf{W}_i)\sigma(\mathbf{W}_i)^{-2}-Eg(Y,\mathbf{X}_S^T\boldsymbol{\beta}_S)\mathbf{D}(\mathbf{W})\sigma(\mathbf{W})^{-2}\|=o_p(1), \quad \text{(D.3)}$$

Because both $\sup_{\mathbf{w}}\|\hat{\mathbf{D}}(\mathbf{w})-\mathbf{D}(\mathbf{w})\|$ and $\sup_{\mathbf{w}}|\hat{\sigma}(\mathbf{w})^2-\sigma(\mathbf{w})^2|$ are $o_p(1)$, proving (D.1) and (D.2) is straightforward. In addition, given the assumption that $E(\sup_{\|\boldsymbol{\beta}\|_\infty\le M}g(Y,\mathbf{X}_S^T\boldsymbol{\beta}_S)^4)<\infty$, (D.3) follows from the uniform law of large number. Hence we have,

$$\sup_{\boldsymbol{\beta}_S\in\Theta}\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_S)\hat{\mathbf{D}}(\mathbf{W}_i)\hat{\sigma}(\mathbf{W}_i)^{-2}-Eg(Y,\mathbf{X}_S^T\boldsymbol{\beta}_S)\mathbf{D}(\mathbf{W})\sigma(\mathbf{W})^{-2}\|=o_p(1).$$

In addition, the event $\mathbf{X}_S = \mathbf{X}_{\hat{S}}$ occurs with probability approaching one, given the selection consistency $P(\hat{S} = S) \to 1$ achieved in Theorem 4.1.

The result then follows because $\rho_n(\boldsymbol{\beta}_S)=\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\hat{\mathbf{X}}_{iS}^T\boldsymbol{\beta}_S)\hat{\sigma}(\mathbf{W}_i)^{-2}\hat{\mathbf{D}}(\mathbf{W}_i).$

Given Lemma D.3, Theorem 6.1 follows from a standard argument for the asymptotic normality of GMM estimators as in Hansen (1982) and Newey and McFadden (1994, Theorem 3.4). The asymptotic variance achieves the semi-parametric efficiency bound derived by Chamberlain (1987) and Severini and Tripathi (2001). Therefore, $\hat{\beta}^*$ is semi-parametric efficient.

Appendix E: Proofs for Section 7

The proof of Theorem 7.1 is very similar to that of Theorem 4.1, which we leave to the online supplementary material, downloadable from http://terpconnect.umd.edu/~yuanliao/high/supp.pdf

Proof of Theorem 7.2

Proof. Define $Q_{l,k}=L_K(\boldsymbol{\beta}_{(-k)}^{(l)},\boldsymbol{\beta}_k^{(l)})+\sum_{j\le k}P_n(|\boldsymbol{\beta}_j^{(l)}|)+\sum_{j>k}P_n(|\boldsymbol{\beta}_j^{(l-1)}|)$. We first show $Q_{l,k} \le Q_{l,k-1}$ for $1 < k \le p$ and $Q_{l+1,1} \le Q_{l,p}$. For $1 < k \le p$, $Q_{l,k} - Q_{l,k-1}$ equals

$$L_K(\boldsymbol{\beta}_{(-k)}^{(l)},\boldsymbol{\beta}_k^{(l)})+P_n(|\boldsymbol{\beta}_k^{(l)}|)-[L_K(\boldsymbol{\beta}_{(-(k-1))}^{(l)},\boldsymbol{\beta}_{k-1}^{(l)})+P_n(|\boldsymbol{\beta}_k^{(l-1)}|)].$$

Note that the difference between $(\boldsymbol{\beta}_{(-k)}^{(l)},\boldsymbol{\beta}_k^{(l)})$ and $(\boldsymbol{\beta}_{(-(k-1))}^{(l)},\boldsymbol{\beta}_{k-1}^{(l)})$ only lies on the *k*th position. The *k*th position of $(\boldsymbol{\beta}_{(-k)}^{(l)},\boldsymbol{\beta}_k^{(l)})$ is $\beta_k^{(l)}$ while that of $(\boldsymbol{\beta}_{(-(k-1))}^{(l)},\beta_{k-1}^{(l)})$ is $\beta_k^{(l-1)}$. Hence by the updating criterion, $Q_{l,k} \le Q_{l,k-1}$ for $k \le p$.

Because $(\boldsymbol{\beta}_{(-1)}^{(l+1)},\beta_1^{(l+1)})$ is the first update in the $l + 1$th iteration, $(\boldsymbol{\beta}_{(-1)}^{(l+1)},\beta_1^{(l+1)})=(\boldsymbol{\beta}_{(-1)}^{(l)},\beta_1^{(l+1)})$. Hence

$$Q_{l+1,1}=L_K(\boldsymbol{\beta}_{(-1)}^{(l)},\beta_1^{(l+1)})+P_n(|\beta_1^{(l+1)}|)+\sum_{j>1}P_n(|\beta_j^{(l)}|).$$

On the other hand, for $\boldsymbol{\beta}^{(l)}=(\boldsymbol{\beta}_{(-p)}^{(l)},\beta_p^{(l)})$,

$$Q_{l,p}=L_K(\boldsymbol{\beta}^{(l)})+\sum_{j>1}P_n(|\beta_j^{(l)}|)+P_n(|\boldsymbol{\beta}_1^{(l)}|).$$

Hence $Q_{l+1,1}-Q_{l,p}=L_K(\boldsymbol{\beta}_{(-1)}^{(l)},\beta_1^{(l+1)})+P_n(|\beta_1^{(l+1)}|)-[L_K(\boldsymbol{\beta}^{(l)})+P_n(|\boldsymbol{\beta}_1^{(l)}|)]$. Note that $(\boldsymbol{\beta}_{(-1)}^{(l)},\beta_1^{(l+1)})$ differs $\boldsymbol{\beta}^{(l)}$ only on the first position. By the updating criterion, $\mathrm{Q}_{l+1,1}-Q_{l,p}\le 0$.

Therefore, if we define $\{L_m\}_{m\ge1}=\{Q_{1,1},\ldots,Q_{1,p},Q_{2,1},\ldots,Q_{2,p},\ldots\}$, then we have shown that $\{L_m\}_{m\ge1}$ is a non-increasing sequence. In addition, $L_m\ge0$ for all $m\ge1$. Hence L_m is a bounded convergent sequence, which also implies that it is Cauchy. By the definition of $Q_K(\boldsymbol{\beta}^{(l)})$, we have $Q_K(\boldsymbol{\beta}^{(l)})=Q_{l,p}$, and thus $\{Q_K(\boldsymbol{\beta}^{(l)})\}_{l\ge1}$ is a sub-sequence of $\{L_m\}$. Hence it is also bounded Cauchy. Therefore, for any $\varepsilon>0$, there is $N>0$, when $l_1,l_2>N$, $|Q_K(\boldsymbol{\beta}^{(l1)})-Q_K(\boldsymbol{\beta}^{(l2)})|<\varepsilon$, which implies that the iterations will stop after finite steps.

The rest of the proof is similar to that of the Lyapunov's theorem of Lange (1995, Prop. 4). Consider a limit point $\boldsymbol{\beta}^*$ of $\{\boldsymbol{\beta}^{(l)}\}_{l\ge1}$ such that there is a subsequence $\lim_{k\to\infty}\boldsymbol{\beta}^{(lk)}=\boldsymbol{\beta}^*$. Because both $Q_K(\cdot)$ and $M(\cdot)$ are continuous, and $Q_K(\boldsymbol{\beta}^{(l)})$ is a Cauchy sequence, taking limits yields

$$Q_K(M(\boldsymbol{\beta}^*))=\lim_{k\to\infty}Q_K(M(\boldsymbol{\beta}^{(l_k)}))=\lim_{k\to\infty}Q_K(\boldsymbol{\beta}^{(l_k)})=Q_K(\boldsymbol{\beta}^*).$$

Hence $\boldsymbol{\beta}^*$ is a stationary point of $Q_K(\boldsymbol{\beta})$.

References

Ai C, Chen X. Efficient estimation of models with conditional moment restrictions containing unknown functions. Econometrica. 2003; 71:1795–1843.

Andrews D. Consistent moment selection procedures for generalized method of moments estimation. Econometrica. 1999; 67:543–564.

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

Antoniadis A. Smoothing noisy data with tapered coiflets series. Scand J Stat. 1996; 23:313–330.

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

Belloni A, Chernozhukov V. Least squares after model selection in high-dimensional sparse models. Bernoulli. 2013; 19:521–547.

Belloni A, Chernozhukov V, Hansen C. Inference on treatment effects after selection amongst high-dimensional controls. Review of Economic Studies. 2013 Forthcoming.

Bickel, P.; Klaassen, C.; Ritov, Y.; Wellner, J. Efficient and adaptive estimation for semiparametric models. Springer; New York: 1998.

Bickel P, Ritov Y, Tsybakov R. Simultaneous analysis of Lasso and Dantzig selector. Ann Statist. 2009; 37:1705–1732.

Bondell H, Reich B. Consistent high-dimensional Bayesian variable selection via penalized credible regions. J Amer Statist Assoc. 2012; 107:1610–1624.

Bradic J, Fan J, Wang W. Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. J R Stat Soc Ser B. 2011; 73:325–349.

Breheny P, Huang J. Coordinate descent algorithms for non convex penalized regression, with applications to biological feature selection. Ann Appl Statist. 2011; 5:232–253.

Bühlmann P, Kalisch M, Maathuis M. Variable selection in high-dimensional models: partially faithful distributions and the PC-simple algorithm. Biometrika. 2010; 97:261–278.

Bühlmann, P.; van de Geer, S. Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer; New York: 2011.

Canay I, Santos A, Shaikh A. On the testability of identification in some nonparametric odes with endogeneity. Econometrica. 2013; 81:2535–2559.

Caner M. Lasso-type GMM estimator. Econometric Theory. 2009; 25:270–290.

Caner M, Fan Q. Hybrid generalized empirical likelihood estimators: instrument selection with adaptive lasso. Manuscript. 2012

Caner M, Zhang H. Adaptive elastic net GMM with diverging number of moments. Journal of Business and Economic Statistics. 2013 forthcoming.

Candès E, Tao T. The Dantzig selector: statistical estimation when p is much larger than *n*. Ann Statist. 2007; 35:2313–2404.

Chamberlain G. Asymptotic efficiency in estimation with conditional moment restrictions. J Econometrics. 1987; 34:305–334.

Chen, X. Large sample sieve estimation of semi-nonparametric models. In: Heckman, JJ.; Leamer, EE., editors. Handbook of Econometrics. Vol. VI ch 76. 2007.

Chen X, Pouzo D. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. Econometrica. 2012; 80:277–321.

Chernozhukov V, Hong H. An MCMC approach to classical estimation. J Econometrics. 2003; 115:293–346.

Daubechies I, Defrise M, De Mol C. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. Comm Pure Appl Math. 2004; 57:1413–1457.

Dominguez M, Lobato I. Consistent estimation of models defined by conditional moment restrictions. Econometrica. 2004; 72:1601–1615.

Donald S, Imbens G, Newey W. Choosing instrumental variables in conditional moment restriction models. J Econometrics. 2009; 153:28–36.

Engle R, Hendry D, Richard J. Exogeneity. Econometrica. 1983; 51:277–304.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Amer Statist Assoc. 2001; 96:1348–1360.

Fan J, Liao Y. Endogeity in ultra high dimensions. Manuscript. 2012

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. J R Stat Soc Ser B. 2008; 70:849–911.

Fan J, Lv J. Non-concave penalized likelihood with NP-dimensionality. IEEE Trans Inform Theory. 2011; 57:5467–5484.

Fan J, Yao Q. Efficient estimation of conditional variance functions in stochastic regression. Biometrika. 1998; 85:645–660.

Fu W. Penalized regression: The bridge versus the LASSO. J Comput Graph Statist. 1998; 7:397–416.

García E. Linear regression with a large number of weak instruments using a post-l_1-penalized estimator. Manuscript. 2011

Gautier E, Tsybakov A. High dimensional instrumental variables regression and confidence sets. Manuscript. 2011

van de Geer S. High-dimensional generalized linear models and the lasso. Annals of Statistics. 2008; 36:614–645.

Hall P, Horowitz J. Nonparametric methods for inference in the presence of instrumental variables. Ann Statist. 2005; 33:2904–2929.

Hansen L. Large sample properties of generalized method of moments estimators. Econometrica. 1982; 50:1029–1054.

Horowitz J. A smoothed maximum score estimator for the binary response model. Econometrica. 1992; 60:505–531.

Huang J, Horowitz J, Ma S. Asymptotic properties of bridge estimators in sparse high-dimensional regression models. Ann Statist. 2008; 36:587–613.

Huang J, Ma S, Zhang C. Adaptive lasso for sparse high-dimensional regression models. Statistica Sinica. 2008; 18:1603–1618.

Hunter D, Li R. Variable selection using MM algorithms. Ann Statist. 2005; 33:1617–1642.

Kim Y, Choi H, Oh H. Smoothly Clipped Absolute Deviation on High Dimensions. J Amer Statist Assoc. 2008; 103:1665–1673.

Lange K. A gradient algorithm locally equivalent to the EM algorithm. J Roy Statist Soc Ser B. 1995; 57:425–437.

Kitamura Y, Tripathi G, Ahn H. Empirical likelihood-based inference in conditional moment restriction models. Econometrica. 2004; 72:1667–1714.

Leeb H, Pötscher B. Sparse estimators and the oracle property, or the return of Hodges' estimator. J Econometrics. 2008; 142:201–211.

Liao Z. Adaptive GMM shrinkage estimation with consistent moment selection. Econometric Theory. 2013; 29:857–904.

Loh P, Wainwright M. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. Manuscript. 2013

Lv J, Fan Y. A unified approach to model selection and sparse recovery using regularized least squares. Ann Statist. 2009; 37:3498–3528.

Newey W. Semiparametric efficiency bound. J Appl Econometrics. 1990; 5:99–125.

Newey, W. Efficient estimation of models with conditional moment restrictions. In: Maddala, GS.; Rao, CR.; Vinod, HD., editors. Handbook of Statistics, Volume 11: Econometrics. Amsterdam; North-Holland: 1993.

Newey, W.; McFadden, D. Large sample estimation and hypothesis testing. In: Engle, R.; McFadden, D., editors. Handbook of Econometrics. Vol. Chapter 36. 1994.

Newey W, Powell J. Instrumental variable estimation of nonpara-metric models. Econometrica. 2003; 71:1565–1578.

Städler N, Bühlmann P, van de Geer S. l1-penalization for mixture regression models with discussion. Test. 2010; 19:209–256.

Severini T, Tripathi G. A simplified approach to computing efficiency bounds in semiparametric models. J Econometrics. 2001; 102:23–66.

Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B. 1996; 58:267–288.

Wasserman L, Roeder K. High-dimensional variable selection. Ann Statist. 2009; 37:2178–2201.

Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Statist. 2010; 38:894–942.

Zhang C, Huang J. The sparsity and bias of the Lasso selection in high-dimensional linear models. Ann Statist. 2008; 36:1567–1594.

Zhang C, Zhang T. A general theory of concave regularization for high dimensional sparse estimation problems. Statistical Science. 2012; 27:576–593.

Zhao P, Yu B. On model selection consistency of Lasso. J Mach Learn Res. 2006; 7:2541–2563.

Zou H. The adaptive Lasso and its oracle properties. J Amer Statist Assoc. 2006; 101:1418–1429.

Zou H, Li R. One-step sparse estimates in nonconcave penalized likelihood models. Ann Statist. 2008; 36:1509–1533.

Zou H, Zhang H. On the adaptive elastic-net with a diverging number of parameters. Ann Statist. 2009; 37:1733–1751.

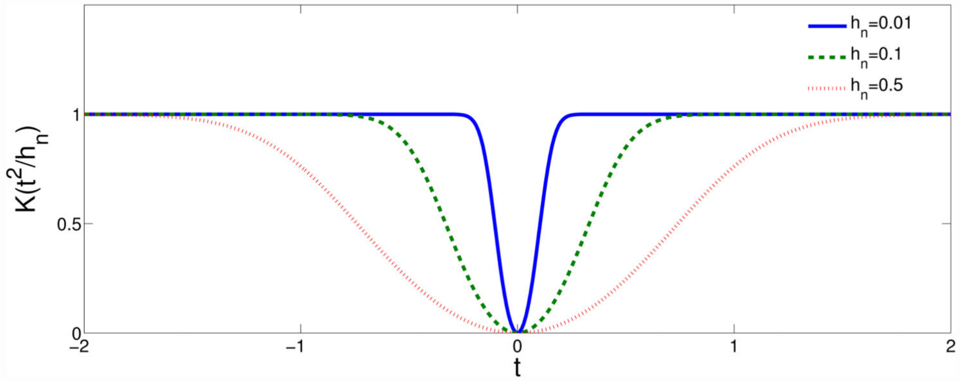

Fig 1. $K\left(\frac{t^2}{h_n}\right)=\frac{\exp(t^2/h_n)-1}{\exp(t^2/h_n)+1}$ **as an approximation to** $\boldsymbol{I_{(t\neq 0)}}$

Table 1

Performance of PLS and FGMM over 100 replications. $p = 50$, $n = 200$

	PLS				FGMM		
	λ = **0.05**	λ = **0.1**	λ = **0.5**	λ = **1**	λ = **0.05**	λ = **0.1**	λ = **0.2**
MSE_S	0.145 (0.053)	0.133 (0.043)	0.629 (0.301)	1.417 (0.329)	0.261 (0.094)	0.184 (0.069)	0.194 (0.076)
MSE_N	0.126 (0.035)	0.068 (0.016)	0.072 (0.016)	0.095 (0.019)	0.001 (0.010)	0 (0)	0.001 (0.009)
TP	5 (0)	5 (0)	4.82 (0.385)	3.63 (0.504)	5 (0)	5 (0)	5 (0)
FP	37.68 (2.902)	35.36 (3.045)	8.84 (3.334)	2.58 (1.557)	0.08 (0.337)	0 (0)	0.02 (0.141)

MSE_S is the average of $\|\hat{\beta_S} - \beta_{0S}\|$ for nonzero coefficients. MSE_N is the average of $\|\hat{\beta_N} - \beta_{0N}\|$ for zero coefficients. *TP* is the number of correctly selected variables, and *FP* is the number of incorrectly selected variables. The standard error of each measure is also reported.

Table 2

Endogeneity in both important and unimportant regressors, *n* = 100

	PLS			FGMM			
	$\lambda = 1$	**$\lambda =3$**	**$\lambda = 4$**	**$\lambda = 0.08$**	**$\lambda = 0.1$**	**$\lambda = 0.3$**	**post-FGMM**
				$p = 50$ $m = 10$			
MSE_S	0.190 (0.102)	0.525 (0.283)	0.491 (0.328)	0.106 (0.051)	0.097 (0.043)	0.102 (0.037)	0.088 (0.026)
MSE_N	0.171 (0.059)	0.240 (0.149)	0.183 (0.149)	0.090 (0.030)	0.085 (0.035)	0.048 (0.034)	
TP	5 (0)	5 (0)	4.97 (0.171)	5 (0)	5 (0)	5 (0)	
FP	27.69 (6.260)	14.63 (5.251)	10.37 (4.539)	3.76 (1.093)	3.5 (1.193)	1.63 (1.070)	
				$p = 200$ $m = 50$			
MSE_S	0.831 (0.787)	0.966 (0.595)	1.107 (0.678)	0.111 (0.048)	0.104 (0.041)	0.231 (0.431)	0.092 (0.032)
MSE_N	1.286 (1.333)	0.936 (0.799)	0.828 (0.656)	0.062 (0.018)	0.063 (0.021)	0.053 (0.075)	
TP	5 (0)	4.9 (0.333)	4.73 (0.468)	5 (0)	5 (0)	4.94 (0.246)	
FP	86.760 (27.41)	42.440 (15.08)	35.070 (13.84)	4.726 (1.358)	4.276 (1.251)	2.897 (2.093)	

m is the number of endogenous regressors. MSE_S is the average of $\|\hat{\beta_S} - \beta_{0S}\|$ for nonzero coefficients. MSE_N is the average of $\|\hat{\beta_N} - \beta_{0N}\|$ for zero coefficients. *TP* is the number of correctly selected variables; *FP* is the number of incorrectly selected variables, and *m* is the total number of endogenous regressors. The standard error of each measure is also reported.

Table 3

Endogeneity only in unimportant regressors, *n* = 200

	PLS			FGMM		
	$\lambda = 0.1$	**$\lambda = 0.5$**	**$\lambda = 1$**	**$\lambda = 0.05$**	**$\lambda = 0.1$**	**$\lambda = 0.2$**
			$p = 50$			
MSE_S	0.133 (0.043)	0.629 (0.301)	1.417 (0.329)	0.261 (0.094)	0.184 (0.069)	0.194 (0.076)
MSE_N	0.068 (0.016)	0.072 (0.016)	0.095 (0.019)	0.001 (0.010)	0 (0)	0.001 (0.009)
TP	5 (0)	4.82 (0.385)	3.63 (0.504)	5 (0)	5 (0)	5 (0)
FP	35.36 (3.045)	8.84 (3.334)	2.58 (1.557)	0.08 (0.337)	0 (0)	0.02 (0.141)
			$p = 300$			
MSE_S	0.159 (0.054)	0.650 (0.304)	1.430 (0.310)	0.274 (0.086)	0.187 (0.102)	0.193 (0.123)
MSE_N	0.107 (0.019)	0.071 (0.023)	0.086 (0.027)	5×10^{-4}(0.006)	0 (0)	5×10^{-4}(0.005)
TP	5 (0)	4.82 (0.384)	3.62 (0.487)	5 (0)	5 (0)	4.99 (0.100)
FP	210.47 (11.38)	42.78 (11.773)	7.94 (5.635)	0.11 (0.37)	0 (0)	0.01 (0.10)

Table 4

FGMM for weak minimal signal $\beta_4 = -0.5$, $\beta_5 = 0.1$

	$p = 50$	$m = 10$		$p = 200$	$m = 50$	
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$
MSE_S	0.128 (0.020)	0.107 (0.000)	0.118 (0.056)	0.138 (0.061)	0.125 (0.074)	0.238 (0.154)
MSE_N	0.155 (0.054)	0.097 (0.000)	0.021 (0.033)	0.134 (0.052)	0.108 (0.043)	0.084 (0.062)
TP	4.12 (0.327)	4 (0)	4 (0)	4.04 (0.281)	3.98 (0.141)	3.8 (0.402)
FP	4.93 (1.578)	5 (0)	2.08 (0.367)	4.72 (1.198)	4.3 (0.948)	1.95 (1.351)