Endogeneity in Ultrahigh Dimension

Jianqing Fan   Yuan Liao[*]

Department of Operations Research and Financial Engineering, Princeton University

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 exogeneous. Yet, endogeneity arises easily in high-dimensional regression due to a large pool of regressors and this causes the inconsistency of the penalized least-squares methods and possible false scientific discoveries. A necessary condition for model selection of a very general class of penalized regression methods is given, which allows us to prove formally the inconsistency claim. To cope with the possible endogeneity, we construct a novel penalized focussed generalized method of moments (FGMM) criterion function and offer a new optimization algorithm. The FGMM is not a smooth function. To establish its asymptotic properties, we first study the model selection consistency and an oracle property for a general class of penalized regression methods. These results are then used to show that the FGMM possesses an 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 recent years ultra-high dimensional models have gained considerable importance in many fields of science, engineering and humanities. In such models the overall number of regressors p grows extremely fast with the sample size n. In particular, $p = O(\exp(n^\alpha))$, for some $\alpha \in (0,1)$. Hence p can grow non-polynomially with n, as in the so-called NP-dimensional problem. Sparse modeling has been widely used to deal with high dimensionality and "Big Data". For example, in the regression model

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

it is assumed that most of the components in $\boldsymbol{\beta}_0$ are zero, and therefore only a few regressors are important that captures the main contributions to the regression. The goal of ultra high dimensional modeling is to achieve the oracle property, which aims at (1) achieving the variable selection consistency (identify the important regressors with high probability), and (2) making inference on the coefficients of the important regressors. There has been an extensive literature on addressing this problem (see for example, Fan and Li (2001), Donoho and Elad (2003), Donoho (2006), Zhao and Yu (2006), Candes and Tao (2007), Huang, Horowitz and Ma (2008), Lounici (2008), Zhang and Huang (2008), Wasserman and Roeder (2009), Lv and Fan (2009), Städler, Bühlmann and van de Geer (2010), Bühlmann, Kalisch and Maathuis (2010), Belloni and Chernozhukov (2011b) and Raskutti, Wainwright and Yu (2011)).

Has the goal of chasing the oracle been really achieved? While the majority of the papers in the literature have given various conditions under which the oracle property can be achieved, they assume that all the candidate regressors are uncorrelated with the regression error term, namely, $E(\varepsilon \mathbf{X}) = 0$. More stringently, they assume

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

This is a very restrictive and possibly unrealistic assumption, yet it is hard if not impossible to verify because of the high-dimensionality p. Without this assumption, all popular model selection techniques are inconsistent as to be shown in Theorems 2.1 and 2.2, which can lead to false scientific claims. Yet, violations to assumption (1.2) arise easily as a result of selection biases, measurement errors, autoregression with autocorrelated errors, omitted variables, and from many other sources (Engle, Hendry and Richard (1983)). In high dimensional models, this is even harder (if not impossible) to avoid due to a large collections of regressors. Indeed, regressors are collected because of their possible prediction powers to the response variable

Y. Yet, requesting equations (1.2) or even more specifically

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

to satisfy is indeed a scientific fiction and is an irresponsible assumption without any validations, particularly when p is large.

For example, in a wage equation, Y is the logarithm of an individual's wage, and the objects of interest in applications include the coefficients of $\mathbf{X}_S$ such as the years of education, years of labor-force experience, marital status and labor union membership. On the other hand, widely available data sets from CPS (Current Population Survey) can contain hundreds or even thousands of variables that are associated with wage but are unimportant predictors. But, some of these variables can be correlated with $y - \mathbf{X}^T \boldsymbol{\beta}_0$ (namely, ε) too, due to the large pool of predictors. The analogy also applies to genomic applications in which gene expression profiles can also be correlated with the regression errors, making false selection of irrelevant genes for scientific outcomes.

To solve the aforementioned issues, we borrow the terminology of *endogeneity* and *exogeneity* from the econometric literature. A regressor is said to be *endogenous* when there is a correlation between the regressor and the error term, and is said to be *exogenous* otherwise. Broadly, a loop of causality between the independent variable and regressor can lead to endogeneity (Verbeek (2008) and Hansen (2010)).

A more realistic and appealing model assumption should be:

$$Y = \mathbf{X}^T \boldsymbol{\beta}_0 + \varepsilon = \mathbf{X}_S^T \boldsymbol{\beta}_{0S} + \varepsilon, \quad E(Y - \mathbf{X}_S^T \boldsymbol{\beta}_{0S} | \mathbf{X}_S) = 0, \tag{1.4}$$

where $\mathbf{X}_S$ and $\boldsymbol{\beta}_{0S}$ denote the vector of important regressors and corresponding coefficients respectively, whose identities are, of course, unknown to us. This assumption is far easier to validate. One of the goals of this paper is to achieve the oracle property under model (1.4), in the presence of possible endogenous regressors.

What makes the model selection possible is the idea of over identification. Let S be the set of important variables in model (1.4) and $|S|$ be the size of the set. For the set S, there exists a solution to the *over-identified* equations (with respect to $\boldsymbol{\beta}_S$) such as

$$E(Y - \mathbf{X}_S^T \boldsymbol{\beta}_S)\mathbf{X}_S = 0 \quad \text{and} \quad E(Y - \mathbf{X}_S^T \boldsymbol{\beta}_S)\mathbf{X}_S^2 = 0, \tag{1.5}$$

where $\mathbf{X}_S^2$ is the vector consisting of squared elements of $\mathbf{X}_S$ and is used as an illustration. It can be replaced, for example, by $|\mathbf{X}_S|$ or many other functions of $\mathbf{X}_S$. In the above equations, we have only $|S|$ unknowns, but $2|S|$ linear equations. Yet, the solution exists and is given

3

by $\boldsymbol{\beta}_S = \boldsymbol{\beta}_{0S}$. On the other hand, for other sets $\tilde{S}$ of variables, the over-identified equations

$$E(Y - \mathbf{X}_{\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}})\mathbf{X}_{\tilde{S}} = 0 \quad \text{and} \quad E(Y - \mathbf{X}_{\tilde{S}}^T\boldsymbol{\beta}_{\tilde{S}})\mathbf{X}_{\tilde{S}}^2 = 0 \tag{1.6}$$

do not have a compatible solution unless $\tilde{S} \supset S$ and the support of $\boldsymbol{\beta}_{\tilde{S}}$ is S and

$$E\varepsilon\mathbf{X}_{\tilde{S}} = 0 \quad \text{and} \quad E\varepsilon\mathbf{X}_{\tilde{S}}^2 = 0, \tag{1.7}$$

where $\varepsilon = Y - \mathbf{X}_S^T\boldsymbol{\beta}_{0S}$.

We show that in the presence of endogenous regressors, the classical penalized least squares method is no longer consistent. Under model (1.4), we introduce a novel loss function, called *focussed generalized method of moments* (FGMM), which differs from the classical generalized method of moments (Hansen, 1982) in that the instrumental variables depend irregularly on unknown parameters. The new FGMM fully appreciates the information contained in the moment condition (1.4), and is powerful in detecting incorrectly specified moment condition of the form

$$E(Y - \mathbf{X}_S^T\boldsymbol{\beta}_{0S})X_l \neq 0 \tag{1.8}$$

if X_l is endogenous. It is also very different from the low-dimensional techniques of either moment selection (Andrews 1999, Andrews and Lu 2001) or shrinkage GMM (Liao 2010) in dealing with misspecifications of moment conditions; the latter introduces one unknown parameter to each possibly misspecified equation and is inappropriate in our high-dimensional endeavors. However, penalization is still needed in FGMM to avoid overfitting the model, since we allow some of unimportant predictors exogenous, satisfying (1.7). This results in a novel penalized FGMM. The proposed FGMM successfully achieves the oracle property in the presence of endogeneity. In particular, the estimator converges in probability to $\boldsymbol{\beta}_{0S}$ at the near *oracle rate* $O_p(\sqrt{(s \log s)/n})$ (Fan and Lv (2011)), and under certain over-identification condition, is a near global minimizer. In addition, it is shown that via a two-step procedure similar to ISIS (Fan and Lv, 2008) and *post-lasso* (Belloni and Chernozhukov, 2011a), we can achieve the semi-parametric efficiency in a more general nonlinear model.

In addition, we consider a more general framework of the ultra high dimensional variable selection problem, and derive both sufficient and necessary conditions for a penalized minimization procedure to achieve the oracle property, where both the loss function (the leading term of the criterion function) and the penalty function can take a very general form. Many results on the oracle property in the literature can be understood as applications of these general theorems.

We emphasize that the problem concerned in this paper is not a simple model misspecification, but rather a question about what kinds of model assumption are more realistic, and about with which assumptions the empirical researchers feel comfortable.

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 to solve the problem of endogeneity, and discusses the rationale of our construction as well as its numerical implementation. Section 4 gives sufficient conditions for establishing the oracle property for a general penalized regression. Section 5 applies these conditions to show the oracle property of FGMM. Section 6 discusses the global optimization. Section 7 is concerned about the semi-parametric efficient estimation of the non-vanishing parameters. Simulation results are demonstrated in Sections 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 $\|\mathbf{A}\|$, $\|\mathbf{A}\|_2$ and $\|\mathbf{A}\|_\infty$ as the Frobenius, operator and elementwise norms of a matrix $\mathbf{A}$ respectively, defined respectively as $\|\mathbf{A}\| = \mathrm{tr}^{1/2}(\mathbf{A}^T\mathbf{A})$, $\|\mathbf{A}\|_2 = \lambda_{\max}^{1/2}(\mathbf{A}^T\mathbf{A})$, and $\|\mathbf{A}\|_\infty = \max_{i,j}|\mathbf{A}_{ij}|$. When $\mathbf{A}$ is a vector, both $\|\mathbf{A}\|$ and $\|\mathbf{A}\|_2$ are equal to the Euclidean norm. For two sequences a_n and $b_n \neq 0$, write $a_n \ll b_n$ (equivalently, $b_n \gg a_n$) if $a_n = o(b_n)$. $|\boldsymbol{\beta}|_0$ denotes the number of nonzero components of a vector $\boldsymbol{\beta}$. In addition, $P_n'(t)$ and $P_n''(t)$ denote the first and second derivatives of a penalty function $P_n(t)$. Finally, we write w.p.a.1 as brevity for "with probability approaching one".

2 Necessary Condition for Variable Selection Consistency

2.1 Penalized regression and necessary condition

Let s denote the number of nonzero coefficients of $\boldsymbol{\beta}_0$. For notational simplicity without loss of generality, it is assumed throughout the paper that the coordinates are rearranged so that the non-vanishing coordinates of $\boldsymbol{\beta}_0$ are the first s coordinates, denoted by $\boldsymbol{\beta}_{0S}$. Therefore, the true structural parameter can be partitioned as $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S}^T, \boldsymbol{\beta}_{0N}^T)^T$, with $\boldsymbol{\beta}_{0N} = 0$. Accordingly, the regressors can be partitioned as $\mathbf{X} = (\mathbf{X}_S^T, \mathbf{X}_N^T)^T$, called *important regressors* and *unimportant regressors* respectively. The sparsity structure typically assumes that the number of important regressors $s = \dim(\mathbf{X}_S)$ grows slowly with the sample size:

$s = o(n)$.

A penalized regression problem in general takes a form of:

$$\min_{\boldsymbol{\beta} \in \mathbb{R}^p} L_n(\boldsymbol{\beta}) + \|P_n(\boldsymbol{\beta})\|_1,$$

where $P_n(\cdot)$ denotes a penalty function and $\|P_n(\boldsymbol{\beta})\|_1 = \sum_{j=1}^p P_n(|\beta_j|)$. While the current literature has been focusing on the sufficient conditions for the penalized estimator to achieve the oracle property, there is relatively much less attention to the necessary conditions. Zhao and Yu (2006) derived an *almost necessary* condition for the sign consistency. Zou (2006) provided a necessary condition for the variable selection consistency of the least squares estimator with Lasso penalty when $p/n \to 0$. To the authors' best knowledge, so far there has been no necessary condition on the loss function for the selection consistency in the ultra high dimensional framework. Such a necessary condition is important, because it provides us a way to justify whether a typical loss function can result in a consistent variable selection.

Theorem 2.1 (Necessary Condition). *Suppose:*
(i) $L_n(\boldsymbol{\beta})$ is twice differentiable, and

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

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

$$L_n(\boldsymbol{\beta}) + \|P_n(\boldsymbol{\beta})\|_1$$

such that $P(\widehat{\boldsymbol{\beta}}_N = 0) \to 1$, and $\sqrt{s}\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\| = o_p(1)$.
(iii) The penalty satisfies: $P_n(\cdot) \geq 0$, $P_n(0) = 0$, $P_n'(t)$ 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 such that $\beta_{0,l} = 0$,

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

Note that the conclusion (2.1) differs from the Karush-Kuhn-Tucker (KKT) condition 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 (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 2009), as long as the tuning parameter $\lambda_n \to 0$. Hence this theorem should be understood as a necessary condition imposed on the loss function instead of the

6

penalty.

2.2 Inconsistency of least squares with endogeneity

As an important application of Theorem 2.1, consider the simple linear model:

$$y \;=\; \mathbf{X}^T\boldsymbol{\beta}_0 + \varepsilon = \mathbf{X}_S^T\boldsymbol{\beta}_{0S} + \varepsilon, \tag{2.2}$$

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

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

$$\min_{\boldsymbol{\beta}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + \|P_n(\boldsymbol{\beta})\|_1.$$

In the simpler case when s, the number of non-vanishing components of $\boldsymbol{\beta}_0$, is bounded, it can be shown that if there exists some unimportant regressor correlated with the regression error ε, the PLS does not achieve the variable selection consistency. This is because the necessary condition in (2.1) does not hold for the least squares loss function. Hence without the ad-hoc exogeneity assumption, PLS would not work any more.

Theorem 2.2 (Inconsistency of PLS). *Suppose $s = O(1)$, and $\mathbf{X}_N$ has an endogenous component X_l, that is, $|E(X_l\varepsilon)| > c$ 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

$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2 + \|P_n(\boldsymbol{\beta})\|_1,$$

then either $\|\tilde{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \nrightarrow^p 0$, or

$$\limsup_{n\to\infty} P(\tilde{\boldsymbol{\beta}}_N = 0) < 1.$$

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

$$Y = \mathbf{X}^T\boldsymbol{\beta}_0 + \varepsilon, \quad \varepsilon \sim N(0,1),$$

$$\boldsymbol{\beta}_{0S} = (5, -4, 7, -1, 1.5); \quad \beta_{0j} = 0, \text{ for } 6 \leq j \leq p.$$
$$X_j = Z_j \text{ for } j \leq 5, \quad X_j = (Z_j + 5)(\varepsilon + 1), \text{ for } 6 \leq j \leq p.$$
$$Z \sim N_p(0, \Sigma), \text{ independent of } \varepsilon, \text{ with } (\Sigma)_{ij} = 0.5^{|i-j|},$$

Table 1: Performanceof PLS and FGMM over 100 replications. $p = 50$, $n = 300$

	PLS				FGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE_S	0.145	0.133	0.629	1.417	0.261	0.184	0.194	0.979
	(0.053)	(0.043)	(0.301)	(0.329)	(0.094)	(0.069)	(0.076)	(0.245)
MSE_N	0.126	0.068	0.072	0.095	0.001	0	0.001	0.003
	(0.035)	(0.016)	(0.016)	(0.019)	(0.010)	(0)	(0.009)	(0.014)
TP-Mean	5	5	4.82	3.63	5	5	5	4.5
	(0)	(0)	(0.385)	(0.504)	(0)	(0)	(0)	(0.503)
FP-Mean	37.68	35.36	8.84	2.58	0.08	0	0.02	0.14
	(2.902)	(3.045)	(3.334)	(1.557)	(0.337)	(0)	(0.141)	(0.569)

MSE_S is the average of $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|$ for non-vanishing coefficients. MSE_N is the average of $\|\widehat{\boldsymbol{\beta}}_N - \boldsymbol{\beta}_{0N}\|$ 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.

In the design, the unimportant regressors are endogenous. The penalized least squares (PLS) with SCAD-penalty was used for variable selection. From Table 1, PLS selects many unimportant regressors (FP-Mean). In contrast, using the proposed method penalized FGMM (to be introduced) we can do an excellent job in both selecting the important regressors and eliminating the unimportant regressors. Yet, the inefficiency of $\hat{\beta}_S$ by FGMM is due to the moment conditions used in the estimate. This can be improved further in Section 7.

3   Focussed GMM

3.1   Definition

Instead of the linear regression (1.1), in this paper we will consider a more general framework:

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

where Y stands for the dependent variable; $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is a known function. For simplicity, we require that g be one-dimensional, and should be thought of as a possibly nonlinear residual function. Our result can be naturally extended to multi-dimensional conditional moment restrictions.

Model (3.1) is called a *conditional moment restricted model*, which has been extensively studied in the literature: Newey (1993), Donald, Imbens and Newey (2003), Kitamura, Tripathi and Ahn (2004), etc. Some of the interesting examples in the generalized linear model that fit into (3.1) are:

- simple 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.

The conditional moment restriction (3.1) implies that

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{X}_S] = 0, \text{ and } E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{X}_S^2] = 0, \tag{3.2}$$

where $\mathbf{X}_S^2$ denotes a vector of squares of $\mathbf{X}_S$ taken coordinately and can be replaced by any other nonlinear functions such as $|\mathbf{X}_S|$ (assuming each variable has mean 0). 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 since the true identities of $\mathbf{X}_S$ are unknown to us. On the other hand, as explained in the introduction, the over-identified equations (3.2) do not have a solution for other sets that support $\boldsymbol{\beta}$.

To take advantage of the above intuition, let us introduce some additional notation. For any $\boldsymbol{\beta} \in \mathbb{R}^p / \{0\}$, and $i = 1, ..., n$, define $r = |\boldsymbol{\beta}|_0$-dimensional vectors

$$\mathbf{X}_i(\boldsymbol{\beta}) = (X_{i,l_1}, ..., X_{i,l_r})^T \text{ and } \mathbf{X}_i^2(\boldsymbol{\beta}) = (X_{i,l_1}^2, ..., X_{i,l_r}^2)^T,$$

where $(l_1, ..., l_r)$ denote the indices of the non-vanishing components of $\boldsymbol{\beta}$. For example, if $p = 3$ and $\boldsymbol{\beta} = (1, 0, 2)^T$, then $\mathbf{X}_i(\boldsymbol{\beta}) = (X_{i1}, X_{i3})^T$, and $\mathbf{X}_i^2(\boldsymbol{\beta}) = (X_{i1}^2, X_{i3}^2)^T$, $i \leq n$.

The FGMM weight matrix is specified as following: for each $j = 1, ..., p$, let $\overline{X}_j = \frac{1}{n} \sum_{i=1}^n X_{ij}$, $\overline{X_j^2} = \frac{1}{n} \sum_{i=1}^n X_{ij}^2$, and define

$$\widehat{\mathrm{var}}(X_j) = \frac{1}{n} \sum_{i=1}^n (X_{ij} - \overline{X}_j)^2, \quad \widehat{\mathrm{var}}(X_j^2) = \frac{1}{n} \sum_{i=1}^n (X_{ij}^2 - \overline{X_j^2})^2,$$

9

which are the sample variances of X_j and X_j^2 respectively. The $(2|\boldsymbol{\beta}|_0) \times (2|\boldsymbol{\beta}|_0)$ FGMM weight matrix is given by a diagonal matrix

$$\mathbf{W}(\boldsymbol{\beta}) = \text{diag}\{\widehat{\text{var}}(X_{l_1})^{-1}, ..., \widehat{\text{var}}(X_{l_r})^{-1}, \widehat{\text{var}}(X_{l_1}^2)^{-1}, ..., \widehat{\text{var}}(X_{l_r}^2)^{-1}\},$$

whereas again, $(l_1, ..., l_r)$ denote the indices of the non-vanishing components of $\boldsymbol{\beta}$.

Let

$$\mathbf{V}_i(\boldsymbol{\beta}) = \begin{pmatrix} \mathbf{X}_i(\boldsymbol{\beta}) \\ \mathbf{X}_i^2(\boldsymbol{\beta}) \end{pmatrix}.$$

Our Focussed Generalized Methods of Moments (FGMM) loss function is defined as

$$
\begin{aligned}
&L_{\text{FGMM}}(\boldsymbol{\beta}) \\
&= \sum_{j=1}^{p} I_{(\beta_j \neq 0)} \left[\frac{1}{\widehat{\text{var}}(X_j)} \left(\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) X_{ij} \right)^2 \right. \\
&\qquad \left. + \frac{1}{\widehat{\text{var}}(X_j^2)} \left(\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) X_{ij}^2 \right)^2 \right] \\
&= \left[\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{V}_i(\boldsymbol{\beta}) \right]^T \mathbf{W}(\boldsymbol{\beta}) \left[\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) \mathbf{V}_i(\boldsymbol{\beta}) \right]
\end{aligned}
$$

The loss function is a weighted average of two quadratic terms $\left(\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) X_{ij} \right)^2$ and $\left(\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}) X_{ij}^2 \right)^2$. As in the same spirit of the regular GMM's optimal weight matrix, the weights depend on the variance of the instrumental variables $\mathbf{X}(\boldsymbol{\beta})$ and $\mathbf{X}^2(\boldsymbol{\beta})$, and help to standardize the moment conditions.

The term $\mathbf{X}_i^2(\boldsymbol{\beta})$ is used here as an example. Other instrumental variables $\mathbf{V}_i(\boldsymbol{\beta})$ can also be used. An obvious example is to replace $\mathbf{X}^2(\boldsymbol{\beta})$ by $|\mathbf{X}(\boldsymbol{\beta}) - \bar{\mathbf{X}}(\boldsymbol{\beta})|$ in which $\bar{\mathbf{X}}(\boldsymbol{\beta})$ is the sample mean vector of $\mathbf{X}(\boldsymbol{\beta})$. Unlike the traditional GMM, the instrumental variables $\mathbf{V}_i(\boldsymbol{\beta})$ depend on the unknown $\boldsymbol{\beta}$ and is not continuous in $\boldsymbol{\beta}$. As to be further explained below, this allows to focus only on the equations with correct specifications and is therefore called the focussed GMM or FGMM for short. We then defined the FGMM estimator by minimizing the following criterion function:

$$Q_{\text{FGMM}}(\boldsymbol{\beta}) = L_{\text{FGMM}}(\boldsymbol{\beta}) + \|P_n(\boldsymbol{\beta})\|_1. \tag{3.3}$$

The penalty function $\|P_n(\boldsymbol{\beta})\|_1$ is also needed, because the indicator function in L_{FGMM} itself only plays a role of sure-screening, which is not enough to guarantee the variable selection consistency. Sufficient conditions on the penalty function for the oracle property will be

presented in Section 4.

3.2 Rationales behind the construction of FGMM

3.2.1 Inclusion of $\mathbf{V}(\boldsymbol{\beta})$

We construct the FGMM criterion function using

$$\mathbf{V}(\boldsymbol{\beta}) = (\mathbf{X}(\boldsymbol{\beta})^T, \mathbf{X}^2(\boldsymbol{\beta})^T)^T.$$

A natural question arises: including $\mathbf{X}^2(\boldsymbol{\beta})$ seems ad-hoc; why not just use $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{X}(\boldsymbol{\beta})$? We now explain the rationale behind the inclusion of the term such as $\mathbf{X}^2(\boldsymbol{\beta})$.

Let us consider a linear regression model (1.4) as an example. If $\mathbf{X}^2(\boldsymbol{\beta})$ were not included and $\mathbf{V}(\boldsymbol{\beta}) = \mathbf{X}(\boldsymbol{\beta})$ had been used, the GMM loss function would have been constructed as

$$L_v(\boldsymbol{\beta}) = \left[\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})\mathbf{X}_i(\boldsymbol{\beta})\right]^T \mathbf{W}(\boldsymbol{\beta}) \left[\frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})\mathbf{X}_i(\boldsymbol{\beta})\right].$$

For simplicity of illustration, we assume that $\mathbf{W}(\boldsymbol{\beta})$ is the identity matrix, and use the l_0 penalty $P_n(|\beta_j|) = \lambda_n I_{(|\beta_j|\neq 0)}$.

Suppose that the true $\boldsymbol{\beta}_0 = (\boldsymbol{\beta}_{0S}^T, 0, ..., 0)^T$ where only the first s components are non-vanishing and that $s > 1$. If we, however, restrict ourselves to $\boldsymbol{\beta}_p = (0, ..., 0, \beta_p)$, the criterion function now becomes

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

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

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

As a result, minimizing Q_{FGMM} is inconsistent for variable selection.

Including an additional term $\mathbf{X}^2(\boldsymbol{\beta})$ in $\mathbf{V}(\boldsymbol{\beta})$ can overcome this problem. Since the number of equations in

$$E[(Y - \mathbf{X}^T\boldsymbol{\beta})\mathbf{X}(\boldsymbol{\beta})] = 0 \quad \text{and} \quad E[(Y - \mathbf{X}^T\boldsymbol{\beta})\mathbf{X}^2(\boldsymbol{\beta})] = 0 \tag{3.4}$$

is twice as many as the number of unknowns (non-vanishing components in $\boldsymbol{\beta}$), it is very unlikely to have some $\boldsymbol{\beta}$ other than $\boldsymbol{\beta}_0$ to satisfy (3.4). As a result, if we define

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

the population version of L_{FGMM}, then as long as $\boldsymbol{\beta}$ is not close to $\boldsymbol{\beta}_0$, G should be bounded away from zero. Therefore, it is reasonable for us to assume that for any $\varepsilon > 0$,

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty > \varepsilon, \boldsymbol{\beta} \neq 0} G(\boldsymbol{\beta}) > \delta \tag{3.5}$$

for some $\delta > 0$. Due to condition (3.5) and that $G(\boldsymbol{\beta}_0) = 0$, implied by the model assumption $E(Y - \mathbf{X}_S^T\boldsymbol{\beta}_{0S}|\mathbf{X}_S) = 0$, minimizing L_{FGMM} forces the estimator to be close to $\boldsymbol{\beta}_0$.

It can be seen that instead of $\mathbf{X}^2(\boldsymbol{\beta})$, one can include other transformations of $\mathbf{X}(\boldsymbol{\beta})$ such as the trigonometric functions in $\mathbf{V}(\boldsymbol{\beta})$ to construction FGMM, as long as

$$\inf_{\|\boldsymbol{\beta}-\boldsymbol{\beta}_0\|_\infty > \varepsilon, \boldsymbol{\beta} \neq 0} \|Eg(Y, \mathbf{X}^T\boldsymbol{\beta})\mathbf{V}(\boldsymbol{\beta})\|^2 > \delta.$$

The specific choice of $\mathbf{V}(\boldsymbol{\beta})$ would not affect the oracle property, but only matters in the asymptotic variance of the estimator (see Sections 5 and 7 for details).

3.2.2 Indicator function

We handle the problems of ultra-high dimensionality and model mis-specification simultaneously by including an indicator function $I_{(\beta_j \neq 0)}$ in the loss function. As a result, the instrumental variables $\mathbf{V}(\beta)$ depend on the parameter $\boldsymbol{\beta}$, which leads to the novel focussed GMM. We now explain the rationale behind it.

Recently, there has been a growing literature on the shrinkage GMM, e.g., Caner (2009), Caner and Zhang (2009), etc, regarding estimation and variable selection based on a set of moment conditions like (3.2). The model considered by the authors above, besides restricted to specific penalty functions, significantly differs from ours, in that the moment conditions they considered are all correctly specified. More recently, Liao (2010) considered GMM with mis-specified moment conditions, but in a low dimensional parameter space, and use a very different idea.

In contrast, because we allow the presence of possibly endogenous regressors, the moment conditions of the form

$$E[g(Y, \mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}] = 0$$

are subject to mis-specification on some endogenous regressors. While only the important

12

regressors are assumed to satisfy

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{X}_S] = 0 \quad \text{and} \quad E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{X}_S^2] = 0,$$

the identities of the correct moment conditions are unknown to us. Without the indicator function in the definition of $L_{\text{FGMM}}(\boldsymbol{\beta})$, the oracle estimator can still have a large objective value due to the endogeneity of other predictors. Therefore the oracle estimator is not necessarily the minimizer.

Including the indicator function in $L_{\text{FGMM}}(\boldsymbol{\beta})$ eliminates the endogenous regressors. In addition, it automatically performs a *sure-screening* procedure that produces a sparse solution. Unless the support $S(\boldsymbol{\beta})$ of $\boldsymbol{\beta}$ contains the true variables in S, $L_{\text{FGMM}}(\boldsymbol{\beta})$ is large. Among those $S(\boldsymbol{\beta}) \supset S$, some variables can be exogenous, satisfying (1.7). The choice of zero or small coefficients are allowable when only $L_{\text{FGMM}}(\boldsymbol{\beta})$ is to be minimized without a penalty, whereas the penalty term in (3.3) makes this choice infeasible.

3.3 Implementation

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

3.3.1 Smoothed FGMM

As we discussed above, including an indicator function benefits us greatly in dimension reduction as well as in handling endogeneity. However, it also makes L_{FGMM} unsmooth. For each fixed subset $\tilde{S} \subset \{1, ..., p\}$, this criterion function is continuous in $\boldsymbol{\beta}$ on $\{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin \tilde{S}\}$, but is not continuous in $\boldsymbol{\beta}$ globally on $\mathbb{R}^p$. As there are 2^p subsets of $\{1, ..., p\}$, minimizing $Q_{\text{FGMM}}(\boldsymbol{\beta}) = L_{\text{FGMM}}(\boldsymbol{\beta}) + \text{Penalty}$ is generally NP-hard, that is, there are no algorithms to solve the problem in a polynomial time.

We overcome this discontinuity problem by applying the *smoothing* technique as in Horowitz (1992), 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
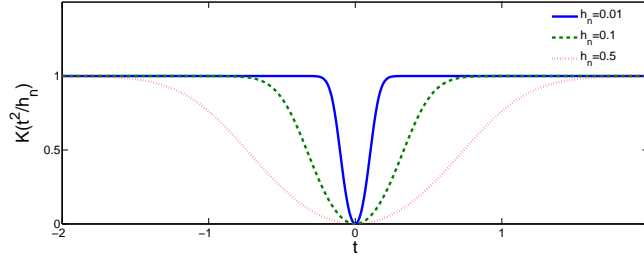
function in $\boldsymbol{\beta}$:

$$
\begin{aligned}
L_K(\boldsymbol{\beta}) \;=\; & \sum_{j=1}^{p} K\left(\frac{\beta_j^2}{h_n}\right)\left[\frac{1}{\widehat{\mathrm{var}}(X_j)}\left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}\right)^2\right. \\
& \left. +\frac{1}{\widehat{\mathrm{var}}(X_j^2)}\left(\frac{1}{n}\sum_{i=1}^{n} g(Y_i,\mathbf{X}_i^T\boldsymbol{\beta})X_{ij}^2\right)^2\right].
\end{aligned}
$$

Note that 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 $L_{\mathrm{FGMM}}(\boldsymbol{\beta})$ for finite sample. As an illustration, Figure 1 plots $K(t^2/h_n)$ as a function of t using the logistic cumulative distribution function, where

$$
K\left(\frac{t^2}{h_n}\right) = \frac{\exp(t^2/h_n)-1}{\exp(t^2/h_n)+1}.
$$

Figure 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 $I_{(t\neq 0)}$

3.3.2 Coordinate descent algorithm

After smoothing the indicator function by a kernel $K(\cdot)$, we employ the iterative coordinate algorithm for the 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 LLA (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 coordiate algorithm for the FGMM minimization. Suppose $\boldsymbol{\beta}^{(l)}$ is obtained at step l. For $k \in \{1,...,p\}$, denote by $\boldsymbol{\beta}^{(l)}_{(-k)}$ a $(p-1)$-dimensional vector consisting of all the components of $\boldsymbol{\beta}^{(l)}$ but $\beta_k^{(l)}$. Write $(\boldsymbol{\beta}^{(l)}_{(-k)},t)$ as the p-dimensional vector that replaces $\beta_k^{(l)}$ with t. The minimiztion with respect to t while keeping $\boldsymbol{\beta}^{(l)}_{(-k)}$ fixed is then

a univariate minimization problem, which can be carried out by a golden section search. To speed up the convergence, we can also use the second order approximation of $L_K(\boldsymbol{\beta}^{(l)}_{(-k)}, t)$ along the kth component:

$$L_K(\boldsymbol{\beta}^{(l)}_{(-k)}, t) \tag{3.6}$$
$$\approx L_K(\boldsymbol{\beta}^{(l)}) + \frac{\partial L_K(\boldsymbol{\beta}^{(l)})}{\partial \beta_k}(t - \beta_k^{(l)}) + \frac{1}{2}\frac{\partial^2 L_K(\boldsymbol{\beta}^{(l)})}{\partial \beta_k^2}(t - \beta_k^{(l)})^2$$
$$\equiv L_K(\boldsymbol{\beta}^{(l)}) + \hat{L}_K(\boldsymbol{\beta}^{(l)}_{(-k)}, t).$$

We solve for

$$t^* = \arg\min_t \hat{L}_K(\boldsymbol{\beta}^{(l)}_{(-k)}, t) + P_n'(|\beta_k^{(l)}|)|t|, \tag{3.7}$$

which admits an explicit analytical solution. We keep the remaining components at step l. We 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 algorithm runs as follows.

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

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

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

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

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

   If

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

   update $\beta_k^{(l)}$ as t^*. Increase l by one when $k = p$.

3. Repeat Step 2 until convergence or l reaches a pre-determined maximum number of iterations.

When the second order approximation (3.6) 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 following form $\min_{\beta\in\mathbb{R}} \frac{1}{2}(z - \beta)^2 + \Lambda P_n(|\beta|)$ has an analytical solution.

15

4 Oracle Property of Penalized Regression for Ultra High Dimensional Models

FGMM involves a non-smooth loss function. We need to first develop a general asymptotic theory in ultra high dimensional models to accommodate this. Sufficient conditions of the oracle property are given when both the loss and penalty functions take general forms. Then in Section 5, the general theory will be applied to the newly proposed FGMM.

4.1 Penalty function

Fan and Li (2001) and Lv and Fan (2009) proposed a class of penalty functions that satisfy a set of general regularity conditions for the variable selection consistency. In this paper, we consider a similar class of penalty functions.

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

$$\eta(\boldsymbol{\beta}) = \limsup_{\varepsilon \to 0^+} \max_{j \leq s} \sup_{\substack{t_1 < t_2 \\ (t_1, t_2) \in (|\beta_j| - \varepsilon, |\beta_j| + \varepsilon)}} -\frac{P_n'(t_2) - P_n'(t_1)}{t_2 - t_1}, \tag{4.1}$$

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

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

represent the strength of signals.

We now define a class of penalty functions to be used throughout the paper:

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, increasing 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 $\sup_{\boldsymbol{\beta} \in B(\boldsymbol{\beta}_{0S}, cd_n)} \eta(\boldsymbol{\beta}) = o(1)$.

The concavity of $P_n(\cdot)$ implies that $\eta(\boldsymbol{\beta}) \geq 0$ for all $\boldsymbol{\beta} \in \mathbb{R}^s$. These conditions are standard, which are needed for establishing the oracle properties of the penalized optimization. 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.

4.2   Oracle property of general penalized regression

The following theorems provide sufficient conditions for the penalized regression (GMM, maximum likelihood, least squares, etc.) to have oracle properties in ultra high dimension.

Define $S = \{j \in \{1, ..., p\} : \beta_{0j} \neq 0\}$, and $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin S\}$. The variable selection aims to recover S with high probability. Our first theorem restricts the penalized optimization onto the s-dimensional subspace $\mathcal{B}$, which is the oracle parameter space. Though infeasible in practice, it gives us an idea of the oracle rate.

In the theorems below, write $L_n(\boldsymbol{\beta}_S, 0) = L_n(\boldsymbol{\beta})$ for $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, 0)^T \in \mathcal{B}$. Let $\boldsymbol{\beta}_S = (\beta_{S1}, ..., \beta_{Ss})$ and

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

Theorem 4.1 (Oracle Consistency). *Suppose $d_n = O(1)$, $s/\sqrt{n} = o(d_n)$ and Assumption 4.1 is satisfied. 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 $\mathcal{B}$, and there exists a positive sequence $\{a_n\}_{n=1}^\infty$ such that $a_n/d_n \to 0$, and a constant $c > 0$ such that:*
(i)

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

(ii) The Hessian matrix $\nabla_S^2 L_n(\boldsymbol{\beta}_S, 0)$ is element-wise continuous within a neighborhood of $\boldsymbol{\beta}_{0S}$, and with probability approaching one,

$$\lambda_{\min}(\nabla_S^2 L_n(\boldsymbol{\beta}_S, 0)) > c.$$

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

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

subject to $(\boldsymbol{\beta}_S^T, 0)^T \in \mathcal{B}$ such that

$$\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = O_p(a_n + \sqrt{s}P_n'(d_n)).$$

For a penalized regression estimator, the rate of convergence depends on both $\|\nabla_S L_n(\boldsymbol{\beta}_{0S}, 0)\|$ and the penalty P_n. Condition (i) requires that the score function should be asymptotically unbiased, whose rate is usually the leading term of the rate of convergence of the estimator. Condition (ii) ensures that asymptotically the Hessian matrix of $L_n(\boldsymbol{\beta}_S, 0)$ is positive definite in a neighborhood of $\boldsymbol{\beta}_{0S}$. Both conditions are satisfied by the likelihood-type loss function considered in Fan and Lv (2011) and Bradic, Fan and Wang (2011). It will

be shown in the next section that FGMM can achieve the near-oracle rate $O_p(\sqrt{(s \log s)/n})$.

The previous theorem assumes that the true support S were 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 $\mathcal{B}$ is also a local minimizer on $\mathbb{R}^p$. The following theorem establishes the sparsity recovery (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

$$\mathbb{T}\boldsymbol{\beta} = (\beta_1', \beta_2', ..., \beta_p')^T \in \mathcal{B}, \quad \beta_j' = \begin{cases} \beta_j & \text{if } j \in S \\ 0, & \text{if } j \notin S \end{cases}.$$

Theorem 4.2 (Sparsity recovery). *Suppose $L_n : \mathbb{R}^p \to \mathbb{R}$ satisfies the conditions in Theorem 4.1, and Assumption 4.1 holds. In addition, for $\widehat{\boldsymbol{\beta}}_S$ in Theorem 4.1, there exists a neighborhood $\mathcal{N}_1 \subset \mathbb{R}^p$ of $(\widehat{\boldsymbol{\beta}}_S^T, 0)^T$, such that for all $\gamma \in \mathcal{N}_1 \backslash \mathcal{B}$, with probability approaching one,*

$$L_n(\mathbb{T}\gamma) - L_n(\gamma) < \sum_{j \notin S} P_n(|\gamma_j|). \tag{4.2}$$

Then with probability approaching 1, $(\widehat{\boldsymbol{\beta}}_S^T, 0)^T$ is a strict local minimizer of

$$Q_n(\boldsymbol{\beta}) = L_n(\boldsymbol{\beta}) + \|P_n(|\boldsymbol{\beta}|)\|_1$$

in $\mathbb{R}^p$. In particular, if L_n is twice differentiable in a neighborhood of $\boldsymbol{\beta}_0$, then (4.2) holds with probability approaching one, if $\sqrt{s}(a_n + \sqrt{s}P'(d_n)) = o(P_n'(0^+))$,

$$\max_{l \notin S} \left| \frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \right| = o_p(P_n'(0^+)), \ and \ \max_{l \leq p, j \leq p} \left| \frac{\partial^2 L_n(\boldsymbol{\beta}_0)}{\partial \beta_l \partial \beta_j} \right| = O_p(1), \tag{4.3}$$

where we denote $P_n'(0^+) = \liminf_{t \to 0^+} P_n'(t)$.

Condition (4.2) is a high-level condition. Due to

$$\sum_{j=1}^{p} P_n(|\gamma_j|) - \sum_{j=1}^{p} P_n(|(\mathbb{T}\gamma)_j|) = \sum_{j \notin S} P_n(|\gamma_j|),$$

it almost is the proof of the theorem. It is imposed here because we want to allow $L_n(\boldsymbol{\beta})$ to be possibly nonsmooth, which is often seen in quantile regression (Belloni and Chernozhukov 2011b), and in our proposed FGMM. On the other hand, if $L_n(\boldsymbol{\beta})$ is assumed to be twice

18

differentiable, such a high level condition can be verified, and a sufficient condition (4.3) is provided.

For statistical inference, we have the following theorem on the asymptotic normality. Let $\text{sgn}(\cdot)$ denote the sign function.

Theorem 4.3 (Asymptotic normality)**.** *Suppose the assumptions in Theorem 4.1 hold, and there exists an $s \times s$ matrix $\boldsymbol{\Omega}_n$, such that:*
(i) For any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$, $\|\boldsymbol{\alpha}\| = 1$,

$$\boldsymbol{\alpha}^T \boldsymbol{\Omega}_n \nabla_S L_n(\boldsymbol{\beta}_{0S}, 0) \to^d N(0, 1);$$

(ii)

$$\left\| \boldsymbol{\Omega}_n \begin{pmatrix} P_n'(|\hat{\beta}_{S1}|)\text{sgn}(\hat{\beta}_{S1}) \\ \vdots \\ P_n'(|\hat{\beta}_{Ss}|)\text{sgn}(\hat{\beta}_{Ss}) \end{pmatrix} \right\| = o_p(1).$$

Then for any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$ with $\|\boldsymbol{\alpha}\| = 1$,

$$\boldsymbol{\alpha}^T \boldsymbol{\Omega}_n \nabla_S^2 L_n(\boldsymbol{\beta}_{0S}, 0)(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}) \to^d N(0, 1).$$

Therefore, the combination of the above theorems implies that, under the conditions of Theorems 4.1-4.3, $Q_n(\boldsymbol{\beta})$ has a strict local minimizer in $\mathbb{R}^p$ that can be partitioned as $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, \widehat{\boldsymbol{\beta}}_N^T)^T$, where the coordinates of $\widehat{\boldsymbol{\beta}}_S$ are inside S, such that

$$\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = O_p(a_n + \sqrt{s}P_n'(d_n)),$$

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

and in addition, $\widehat{\boldsymbol{\beta}}_S$ is asymptotically normal.

These sufficient conditions for the variable selection and parameter estimation are very general and not limited to any specific model. We will see in the next section that, with mild regularity conditions on the moments, all the conditions in Theorems 4.1, 4.2 and 4.3 are satisfied by the penalized FGMM in conditional moment restricted models.

5   Oracle Property of FGMM

With the help of general penalized regression theory, we are now ready to derive the oracle property of the penalized FGMM procedure. The following assumptions are imposed.

Assumption 5.1. *(i) The true parameter $\boldsymbol{\beta}_0$ is uniquely identified by $E(g(Y, \mathbf{X}^T \boldsymbol{\beta}_0)|\mathbf{X}_S) = 0$.*

(ii) $(Y_1, \mathbf{X}_1), ..., (Y_n, \mathbf{X}_n)$ are independent and identically distributed.

Assumption 5.2. *There exist $b_1, b_2 > 0$ and $r_1, r_2 > 0$ such that for any $t > 0$,*
(i) $P(|g(Y, \mathbf{X}^T \boldsymbol{\beta}_0)| > t) \leq \exp(-(t/b_1)^{r_1})$,
(ii) $\max_{l \leq p} P(|X_l| > t) \leq \exp(-(t/b_2)^{r_2})$.
(iii) $\min_{l \in S} \operatorname{var}(g(Y, \mathbf{X}^T \boldsymbol{\beta}_0) X_l)$ is bounded away from zero.
(iv) $\operatorname{var}(X_l)$ and $\operatorname{var}(X_l^2)$ are bounded away from both zero and infinity uniformly in $l = 1, ..., p$ and $p \geq 1$.

This assumption requires that both the regression residuals and the important regressors should have exponential tails, which enables us to apply the large deviation theory to show $\|n^{-1} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) \mathbf{V}_{iS}\| = O_p(\sqrt{s \log s / n})$. A simple example in which this assumption is satisfied is that $g(Y, \mathbf{X}^T \boldsymbol{\beta}_0)$ and $\mathbf{X}_s$ are Gaussian.

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

$$m(t_1, t_2) = \frac{\partial g(t_1, t_2)}{\partial t_2}, \quad q(t_1, t_2) = \frac{\partial^2 g(t_1, t_2)}{\partial t_2^2},$$

$$\mathbf{V}_S = \begin{pmatrix} \mathbf{X}_S \\ \mathbf{X}_S^2 \end{pmatrix}.$$

Assumption 5.3. $g(\cdot, \cdot)$ *is twice differentiable, $\sup_{t_1, t_2} |m(t_1, t_2)| < \infty$, and $\sup_{t_1, t_2} |q(t_1, t_2)| < \infty$.*

This assumption is satisfied by the simple linear regression, logistic regression, probit model, and most of the interesting examples in the generalized linear model.

Example 5.1. In linear regression, $m(t_1, t_2) = -1$. In logistic regression, $m(t_1, t_2) = \frac{\exp(t_2)}{(1 + \exp(t_2))^2} < \frac{1}{4}$, $|q(t_1, t_2)| = |\frac{\exp(t_2)(1 - \exp(t_2))}{(1 + \exp(t_2))^3}| < 1$. In probit regression, $m(t_1, t_2) = \phi(t_2) < (2\pi)^{-1/2}$, $|q(t_1, t_2)| = |t_2 \phi(t_2)| < (2\pi e)^{-1/2}$.

Assumption 5.4. *There exist $C_1 > 0$ and $C_2 > 0$ such that*

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

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

The first condition is needed for $\widehat{\boldsymbol{\beta}}_S$ to converge at a near oracle rate, that is, $a_n = O_p(\sqrt{(s \log s)/n})$ for a_n in Theorem 4.1. The second condition ensures that the Hessian matrix of $L_{\text{FGMM}}(\boldsymbol{\beta}_S, 0)$ is positive definite at $\boldsymbol{\beta}_{0S}$. In the generalized linear model, Assumption

5.4 is satisfied if proper conditions on the design matrices are imposed. For example, in the linear regression model, we assume

$$C_1 \leq \lambda_{\min}(E\mathbf{X}_S\mathbf{X}_S^T) \leq \lambda_{\max}(E\mathbf{X}_S\mathbf{X}_S^T) \leq C_2,$$

and

$$C_1 \leq \lambda_{\min}(E\mathbf{X}_S\mathbf{X}_S^{2T}E\mathbf{X}_S^2\mathbf{X}_S^T) \leq \lambda_{\max}(E\mathbf{X}_S\mathbf{X}_S^{2T}E\mathbf{X}_S^2\mathbf{X}_S^T) \leq C_2;$$

In the probit model, Assumption 5.4 holds if

$$C_1 \leq \lambda_{\min}(E\phi(\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{X}_S^T) \leq \lambda_{\max}(E\phi(\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{X}_S^T) \leq C_2,$$

and similar inequalities hold for $E\phi(\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{X}_S^{2T}$, where $\phi(\cdot)$ is the standard normal density function. Conditions in the same spirit are also assumed in Bradic, Fan and Wang (2011 Condition 4), and Fan and Lv (2011, Condition 4).

Assumption 5.5. *There exist two nonnegative sequences $\kappa_n = O(\sqrt{s})$ and $\eta_n = O(\sqrt{s})$ such that*

$$\max_{l \notin S} \|Em(y, \mathbf{X}^T\boldsymbol{\beta}_0)X_l\mathbf{V}_S\|^2 = O(\kappa_n^2),$$

$$\max_{j \in S} \lambda_{\max}[Em(y, \mathbf{X}^T\boldsymbol{\beta}_0)^2X_j^2\mathbf{V}_S\mathbf{V}_S^T] = O(\eta_n^2),$$

and

$$s\kappa_n\eta_n(\sqrt{(\log s)/n} + P_n'(d_n)) = o(P_n'(0^+)).$$

This assumption is needed to satisfy condition (4.2) in Theorem 4.2. For the ordinary linear model, the above assumption is a statement on

$$\max_{l \notin S} \|EX_l\mathbf{V}_S\|, \text{ and } \max_{j \in S} \lambda_{\max}[EX_j^2\mathbf{V}_S\mathbf{V}_S^T]$$

which imposes some restrictions on the correlation between the important and unimportant regressors once the data are centered. In general, the above assumption imposes some restrictions on the order of the weighted covariance. By Assumptions 5.2 and 5.3, the first two equalities hold with $\kappa_n = \eta_n = \sqrt{s}$. Therefore, without the first two assumptions in Assumption 5.5, the oracle property in Theorem 5.1 below still holds if $s^2P_n'(d_n) + s^2\sqrt{\log s/n} = o(P_n'(0^+))$. This is satisfied by SCAD and MCP if the tuning parameter satisfies $s^2\sqrt{\log s/n} \ll \lambda_n \ll d_n$ and by l^q penalty $(q < 1)$ if $\lambda_n\sqrt{s} = o(d_n^{2-q})$.

On the other hand, when covariates are weakly correlated, we can take smaller order κ_n and η_n than the upper bound $\sqrt{s}$. This relaxes the third requirement in Assumption 5.5, and hence the restrictions on the number of important regressors s and the strength

of the minimal signal d_n. In particular, when $\kappa_n = \eta_n = 1$, our restriction reduces to $sP_n'(d_n) + s\sqrt{\log s/n} = o(P_n'(0^+))$.

Under the foregoing regularity conditions, we can show the oracle property of a local minimizer of the FGMM (3.3).

Theorem 5.1. *Suppose $s/\sqrt{n} = o(d_n)$, and $\log p = o(n)$. Under Assumptions 4.1, 5.1-5.5, there exists a strict local minimizer $\widehat{\boldsymbol\beta} = (\widehat{\boldsymbol\beta}_S^T, \widehat{\boldsymbol\beta}_N^T)^T$ of $Q_{FGMM}(\boldsymbol\beta)$ such that:*
(i)

$$\|\widehat{\boldsymbol\beta}_S - \boldsymbol\beta_{0S}\| = O_p(\sqrt{(s\log s)/n} + \sqrt{s}P_n'(d_n)),$$

where $\widehat{\boldsymbol\beta}_S$ is a subvector of $\widehat{\boldsymbol\beta}$ whose coordiates are in S, and
(ii)

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

Remark 5.1.   1. We only require $\mathbf{X}_S$ to be uncorrelated with the error term. In other words, even if some of the components in $\mathbf{X}_N$ are endogenous, penalized FGMM can still achieve the variable selection consistency.

2. The near oracle rate $\|\widehat{\boldsymbol\beta}_S - \boldsymbol\beta_{0S}\| = O_p(\sqrt{s\log s/n})$ is attained if $P_n'(d_n) = O(\sqrt{\log s/n})$. This is satisfied, for example, by SCAD and MCP if the tuning parameter $\lambda_n = o(d_n)$.

The asymptotic normality requires an additional assumption as follows. Define

$$\mathbf{V}_0 = \text{var}(g(Y, \mathbf{X}_S^T\boldsymbol\beta_{0S})\mathbf{V}_S). \tag{5.1}$$

Assumption 5.6. *(i) For some $c > 0$, $\lambda_{\min}(\mathbf{V}_0) > c$.*
(ii) $P_n'(d_n) = o(1/\sqrt{ns})$.
(iii) There exists $C > 0$, $\sup_{\|\boldsymbol\beta - \boldsymbol\beta_{0S}\| \le C\sqrt{(s\log s)/n}} \eta(\boldsymbol\beta) = o((s\log s)^{-1/2})$.

Conditions (ii) and (iii) are satisfied by the penalty functions SCAD, and MCP. For example, for SCAD, $\sup_{\|\boldsymbol\beta - \boldsymbol\beta_{0S}\| \le C\sqrt{(s\log s)/n}} \eta(\boldsymbol\beta) = 0$ when $\lambda_n + \sqrt{s\log s/n} = o(d_n)$. However, they are not satisfied by l_q-penalty ($q \in (0,2)$), or the elastic net (Zou and Hastie (2005)).

Theorem 5.2 (Asymptotic Normality). *Under the conditions in Theorem 5.1 and Assumption 5.6, the penalized FGMM estimator in Theorem 5.1 satisfies*

$$\sqrt{n}\boldsymbol\alpha^T\boldsymbol\Gamma_n^{-1/2}\boldsymbol\Sigma_n(\widehat{\boldsymbol\beta}_S - \boldsymbol\beta_{0S}) \to^d N(0,1),$$

for any unit vector $\boldsymbol\alpha \in \mathbb{R}^s$, $\|\boldsymbol\alpha\| = 1$, where

$$\boldsymbol\Gamma_n = 4\mathbf{A}_n\mathbf{W}(\boldsymbol\beta_0)\mathbf{V}_0\mathbf{W}(\boldsymbol\beta_0)\mathbf{A}_n^T, \quad \boldsymbol\Sigma_n = 2\mathbf{A}_n\mathbf{W}(\boldsymbol\beta_0)\mathbf{A}_n^T,$$

$$\mathbf{A}_n = Em(Y, \mathbf{X}^T\boldsymbol{\beta}_0)\mathbf{X}_S\mathbf{V}_S^T.$$

6 Global minimization

Theoretical analysis of minimizing a nonconvex criterion function for large p has so far focused on the properties of a specific local minimizer (e.g., Lv and Fan (2009), Bradic et al. (2011)). A natural question to ask is that how close is such a local minimizer to the global solution?

In the GMM literature, when the parameter satisfies a set of moment conditions whose dimension is larger than that of the parameter, the parameter is said to be *over-identified*. Relating the over-identification issue to the problem here, we can then show that the local minimizer in Theorems 5.1 and 5.2 can also be made nearly global.

For a fixed δ, define an l_∞ ball centered at $\boldsymbol{\beta}_0$ with radius δ:

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

Assumption 6.1 (over-identification). *For any $\delta > 0$, there exists $\varepsilon > 0$ such that*

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

This is a high-level assumption that is, however, hard to avoid in ultra-high dimensional problems. It is the empirical counterpart of (3.5). We now explain the rationale behind this assumption. As in the discussion of Section 3.2, the number of equations in

$$E[g(Y, \mathbf{X}^T\boldsymbol{\beta})\mathbf{X}(\boldsymbol{\beta})] = 0 \quad \text{and} \quad E[g(Y, \mathbf{X}^T\boldsymbol{\beta})\mathbf{X}^2(\boldsymbol{\beta})] = 0 \tag{6.1}$$

is twice as much as the number of unknowns (non-vanishing components in $\boldsymbol{\beta}$). As a result, the above simultaneous equations are in general incompatible (that is, have no solution) unless $\boldsymbol{\beta}$ is on the true parameter space $\boldsymbol{\beta} = (\boldsymbol{\beta}_S^T, 0)^T$. In other words, (6.1) has a unique solution $\boldsymbol{\beta} = \boldsymbol{\beta}_0$ and it is reasonable to assume that $\|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_i^T\boldsymbol{\beta})\mathbf{V}_i(\boldsymbol{\beta})\|$ is bounded away from zero whenever $\boldsymbol{\beta}$ is not close to $\boldsymbol{\beta}_0$.

We impose this assumption on the empirical counterpart instead of the population for technical reasons. Under ultra-high dimensionality, the accumulation of the approximation errors from using the law of large number is no longer negligible, and as a result, it is challenging to show that $\|E[g(Y, \mathbf{X}^T\boldsymbol{\beta})\mathbf{V}(\boldsymbol{\beta})]\|$ 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 $\boldsymbol{\beta}$.

Theorem 6.1. *Assume* $\max_{j \in S} P_n(|\beta_{0j}|) = o(s^{-1})$. *Under Assumption 6.1 and those of Theorem 5.1, the local minimizer* $\widehat{\boldsymbol{\beta}}$ *in Theorem 5.1 satisfies: for any* $\delta > 0$, *there exists* $\varepsilon > 0$,

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

Remark 6.1.   1. The result stated in this theorem is *near global*, in the sense that it excludes the set $\{0\}$ from the searching area because $Q_{\text{FGMM}}(0) = 0$ by definition. It is reasonable to believe that 0 is not close to the true parameter, since we assume there should be at least one important regressor in the model. In addition, our global minimization result 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), Zhang (2010), Bühlmann and van de Geer (2011, ch 9), and Zhang and Zhang (2012).

2. Assumption 6.1 can be relaxed a bit in that ε 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.

3. Including finitely many transformations of $\mathbf{X}$ in $\mathbf{V}$ also enables us to achieve the near global minimization if the over-identification assumption is satisfied.

7   Semi-parametric efficiency

The results in Sections 5-6 demonstrate that the choice of the instrumental variable $\mathbf{V}(\boldsymbol{\beta})$ only changes the asymptotic variance of the estimator, but does not affect the variable selection consistency or the rate of convergence. Therefore, the specific choice does not matter if our focus is just on these properties, but not on the semiparametric efficiency, that is, the minimum asymptotic variance of the estimator.

On the other hand, one can always follow a two-step post-FGMM procedure if the semiparametric efficiency is indeed one of the objectives. In linear regression, this has been achieved by Belloni and Chernozhukov (2011a).

After achieving the oracle properties in Theorem 5.1, we have exactly identified the important regressors with probability approaching one, that is,

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

Then the problem of achieving semiparametric efficiency (in the sense of Newey (1990) and

Bickel, Klaassen, Ritov, and Wellner (1998)) in a low dimensional model:

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

has been well studied in the literature (see, for example, Chamberlain (1987), Newey (1993)). In particular, Newey (1993) showed that the semiparametric efficient estimator of $\boldsymbol{\beta}_{0S}$ can be obtained using GMM with moment condition:

$$E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \sigma(\mathbf{X}_S)^{-2} \mathbf{D}(\mathbf{X}_S)] = 0 \tag{7.1}$$

where

$$\sigma(\mathbf{X}_S)^2 = E[g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S})^2 | \mathbf{X}_S], \text{ and } \mathbf{D}(\mathbf{X}_S) = E\left[\frac{\partial g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S})}{\partial \boldsymbol{\beta}_S} \bigg| \mathbf{X}_S \right].$$

For simplicity, we restrict $s = O(1)$, and only consider the nonlinear regression model:

$$g(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) = Y - h(\mathbf{X}_S^T \boldsymbol{\beta}_{0S})$$

for some known differentiable function $h(\cdot)$. Suppose there exists a consistent estimator $\widehat{\sigma}(\mathbf{X}_S)^2$ of $\sigma(\mathbf{X}_S)^2$, we then estimate $\boldsymbol{\beta}_{0S}$ by solving

$$\rho_n(\boldsymbol{\beta}_S) = \frac{1}{n} \sum_{i=1}^{n} (Y_i - h(\widehat{\mathbf{X}}_{iS}^T \boldsymbol{\beta}_S)) h'(\widehat{\mathbf{X}}_{iS}^T \widehat{\boldsymbol{\beta}}_S) \widehat{\sigma}(\mathbf{X}_i)^{-2} \widehat{\mathbf{X}}_{iS} = 0 \tag{7.2}$$

on a compact set $\Theta \subset \mathbb{R}^s$ in which $\boldsymbol{\beta}_{0S}$ is an interior point, where $h'(\cdot)$ denotes the first derivative of $h(\cdot)$.

Let χ be the support of $\mathbf{X}_S$.

Assumption 7.1. *(i) There exists $C_1 > 0$ and $C_2 > 0$ so that*

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

In addition, there exists $\widehat{\sigma}(\mathbf{x})$ such that

$$\sup_{\mathbf{x} \in \chi} |\widehat{\sigma}(\mathbf{x})^2 - \sigma(\mathbf{x})^2| = o_p(1).$$

(ii) Parameter space: $\boldsymbol{\beta}_{0S}$ lies in the interior of a compact set $\Theta \in \mathbb{R}^s$.
(iii) $E(\sup_{\boldsymbol{\beta}_S \in \Theta_S} h(\mathbf{X}_S^T \boldsymbol{\beta}_S)^4) < \infty$, $\sup_t |h'(t)| < \infty$, and $\sup_t |h''(t)| < \infty$.

The existence of a consistent estimator for $\sigma(\mathbf{x})^2$ can be obtained in many interesting examples.

Example 7.1 (Homoskedasticity)**.** Suppose $Y = h(\mathbf{X}_S^T \boldsymbol{\beta}_{0S}) + \varepsilon$, where ε and $\mathbf{X}_S$ are independent. Then

$$\sigma(\mathbf{X}_S)^2 = E(\varepsilon^2 | \mathbf{X}_S) = \sigma^2,$$

which does not depend on $\mathbf{X}_S$, and hence can be consistently estimated by $\widehat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - h(\widehat{\mathbf{X}}_{iS}^T \widehat{\boldsymbol{\beta}}_S))^2$. In this case, equations (7.1) and (7.2) do not depend on σ^2 and (7.2) is simply the normal equations of the ordinary least-squares.

Example 7.2 (Exponential family)**.** Consider a generalized linear model where the conditional density of Y given $\mathbf{X}_S$ belongs to the exponential family

$$f(Y; \mathbf{X}_S, \theta) = c(Y) \exp[Y \mathbf{X}_S^T \boldsymbol{\beta}_{0S} - b(\mathbf{X}_S^T \boldsymbol{\beta}_{0S})].$$

Then $\sigma(\mathbf{X}_S)^2 = b''(\mathbf{X}_S^T \boldsymbol{\beta}_{0S})$, and can be consistently estimated by $b''(\mathbf{X}_S^T \widehat{\boldsymbol{\beta}}_S)$.

Example 7.3 (Nonparametric approach)**.** One can also assume a semi-parametric structure on the functional form of $\sigma(\mathbf{X}_S)^2$:

$$\sigma(\mathbf{X}_S)^2 = f(\mathbf{X}_S; \theta),$$

where $f(\cdot; \theta)$ is a nonparametric function parameterized by θ. We can then estimate $\sigma(\mathbf{X}_S)^2$ using a standard semi-parametric method. More generally, we can proceed by a pure non-parametric approach via regressing $[Y - h(\widehat{\mathbf{X}}_S^T \widehat{\boldsymbol{\beta}}_S)]^2$ on $\widehat{\mathbf{X}}_S$ (see Fan and Yao, 1998).

Condition (iii) in Assumption 7.1 is a technical assumption. We need the fourth moment of $h(\cdot)$ to be uniformly bounded to apply the uniform weak law of large number:

$$\sup_{\boldsymbol{\beta}_S \in \Theta} |\frac{1}{n} \sum_{i=1}^n h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S)^4 - E h(\mathbf{X}_S^T \boldsymbol{\beta}_S)^4| = o_p(1).$$

For example in the linear regression, $h(\mathbf{X}_S^T \boldsymbol{\beta}_S) = \mathbf{X}_S^T \boldsymbol{\beta}_S$, then due to the compactness of Θ, $E(\sup_{\boldsymbol{\beta}_S \in \Theta_S} h(\mathbf{X}_S^T \boldsymbol{\beta}_S)^4) \leq CE\|\mathbf{X}_S\|^4 < \infty$. For other interesting models in GLM, this condition has been verified by Example 5.1 in Section 5.

Theorem 7.1. *Suppose* $s = O(1)$, *Assumption 7.1 and those of Theorem 5.1 hold. Then*

$$\sqrt{n}(\widehat{\boldsymbol{\beta}}_S^* - \boldsymbol{\beta}_{0S}) \to^d N(0, [E(\sigma(\mathbf{X}_S)^{-2} h'(\mathbf{X}_S^T \boldsymbol{\beta}_{0S})^2 \mathbf{X}_S \mathbf{X}_S^T)]^{-1}),$$

and $[E(\sigma(\mathbf{X}_S)^{-2}h'(\mathbf{X}_S^T\boldsymbol{\beta}_{0S})^2\mathbf{X}_S\mathbf{X}_S^T)]^{-1}$ *achieves the semi-parametric efficiency bound in Chamberlain (1987).*

8 Monte Carlo Experiments

8.1 Design 1

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

$$Y = \mathbf{X}^T\boldsymbol{\beta}_0 + \varepsilon, \quad \varepsilon \sim N(0,1).$$

$$(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}) = (5, -4, 7, -1, 1.5); \quad \beta_{0j} = 0, \text{ for } 6 \le j \le p.$$

The p-dimensional vector of regressors $\mathbf{X}$ is generated from the following process:

$$Z = (Z_1, ..., Z_p)^T \sim N_p(0, \Sigma), \quad (\Sigma)_{ij} = 0.5^{|i-j|},$$

$$(X_1, ..., X_5) = (Z_1, ..., Z_5), \quad X_j = (Z_j + 5)(\varepsilon + 1), \text{ for } 6 \le j \le p.$$

where Z is independent of ε. The unimportant regressors are correlated with both important regressors and the error term.

The data contains $n = 200$ i.i.d. copies of $(Y, \mathbf{X})$. 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.

We use the logistic cumulative distribution function with $h = 0.1$ for smoothing:

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

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 $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|$ over the 100 replications, where $S = \{1, ..., 5\}$. 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 $\boldsymbol{\beta}^{(0)} = (0, ..., 0)^T$, and run a penalized least squares (SCAD(λ)) for $\lambda = 0.01$ to obtain the initial value for the FGMM procedure. The results of the simulation are summarized in Tables 2-4, which

compare the performance measures of PLS and FGMM for three values of p.

Table 2: Performance Measures of PLS and FGMM when $p = 15$

	PLS				FGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE_S	0.147	0.138	0.626	1.452	0.193	0.177	0.203	0.953
	(0.055)	(0.052)	(0.306)	(0.320)	(0.066)	(0.067)	(0.061)	(0.241)
MSE_N	0.076	0.062	0.084	0.093	0.010	0.004	0.003	0.004
	(0.023)	(0.014)	(0.013)	(0.017)	(0.026)	(0.014)	(0.015)	(0.017)
TP-Mean	5	5	4.85	3.57	5	5	5	4.55
Median	5	5	5	4	5	5	5	5
	(0)	(0)	(0.357)	(0.497)	(0)	(0)	(0)	(0.5)
FP-Mean	9.356	8.84	2.7	1.34	0.099	0.090	0.02	0.04
Median	10	9	3	1	0	0	0	0
	(0.769)	(0.987)	(1.127)	(0.553)	(0.412)	(0.288)	(0.218)	(0.197)

PLS has non-negligible false positives (FP). The average FP decreases as the magnitude of the penalty parameter increases, however, with an increasing average MSE as well since larger penalties also incorrectly miss the important regressors. For $\lambda = 1$, the median of true positives is only 4. In contrast, FGMM performs quite well in both selecting the important regressors, and correctly eliminating the unimportant regressors. The average MSE of FGMM is only slightly larger than that of PLS when $\lambda = 0.05$ and 0.1. This is understandable since the FGMM as implemented does not intend to be efficient in estimating parameters. When the correct regressors are selected by the FGMM, since the error distribution is normal, adding an extra term $\mathbf{X}_S^2$ term in the square loss makes parameters inefficiently estimated. A solution to this efficient issue is the two-stage post-FGMM in which the ordinary least-squares are run again using the variables $\mathbf{X}_S$ (because the error is normal; see Section 7). Note that $\lambda = 0.4$ is a large tuning parameter that results in some incorrectly eliminated important regressors, and a larger MSE.

8.2 Design 2

Consider the same simple linear model with

$$(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}) = (5, -4, 7, -1, 1.5); \quad \beta_{0j} = 0, \text{ for } 6 \leq j \leq p.$$

Table 3: Performance Measures of PLS and FGMM when $p = 50$

	PLS				FGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE_S	0.145	0.133	0.629	1.417	0.261	0.184	0.194	0.979
	(0.053)	(0.043)	(0.301)	(0.329)	(0.094)	(0.069)	(0.076)	(0.245)
MSE_N	0.126	0.068	0.072	0.095	0.001	0	0.001	0.003
	(0.035)	(0.016)	(0.016)	(0.019)	(0.010)	(0)	(0.009)	(0.014)
TP-Mean	5	5	4.82	3.63	5	5	5	4.5
Median	5	5	5	4	5	5	5	4.5
	(0)	(0)	(0.385)	(0.504)	(0)	(0)	(0)	(0.503)
FP-Mean	37.68	35.36	8.84	2.58	0.08	0	0.02	0.14
Median	38	35	8	2	0	0	0	0
	(2.902)	(3.045)	(3.334)	(1.557)	(0.337)	(0)	(0.141)	(0.569)

Table 4: Performance Measures of PLS and FGMM when $p = 300$

	PLS				FGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE_S	0.186	0.159	0.650	1.430	0.274	0.187	0.193	1.009
	(0.073)	(0.054)	(0.304)	(0.310)	(0.086)	(0.102)	(0.123)	(0.276)
MSE_N	0.221	0.107	0.071	0.086	5×10^{-4}	0	5×10^{-4}	0.002
	(0.037)	(0.019)	(0.023)	(0.027)	(0.006)	(0)	(0.005)	(0.010)
TP-Mean	5	5	4.82	3.62	5	5	4.99	4.45
Median	5	5	5	4	5	5	5	4
	(0)	(0)	(0.384)	(0.487)	(0)	(0)	(0.100)	(0.557)
FP-Mean	227.96	210.47	42.78	7.94	0.11	0	0.01	0.05
Median	227	211	42	7	0	0	0	0
	(10.767)	(11.38)	(11.773)	(5.635)	(0.37)	(0)	(0.10)	(0.330)

The p-dimensional vector of regressors $\mathbf{X}$ is generated from the following process:

$$Z = (Z_1, ..., Z_p)^T \sim N_p(0, \Sigma), \quad (\Sigma)_{ij} = 0.5^{|i-j|},$$

$$(X_1, ..., X_{100}) = (Z_1, ..., Z_{100}), \quad X_j = (Z_j + 5)(\varepsilon + 1), \text{ for } 101 \le j \le p.$$

where Z is independent of ε. Now the first 95 unimportant regressors are exogenous while the rest are endogenous. We run the same FGMM procedure for $n = 200$ and $p = 300$, with an additional post-GMM step to improve the mean squared error of the estimates. The results are reported in Table 5. We can see that the penalized FGMM still performs quite well when there are both exogenous and endogenous unimportant regressors. In addition, after running the additional post-FGMM step, one achieves a better accuracy of estimation.

Table 5: Performance Measures of PLS, FGMM and post-FGMM when $p = 300$

	PLS		FGMM			
	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 0.1$	post-FGMM	$\lambda = 0.2$	post-FGMM
MSE$_S$	0.278	0.712	0.215	0.190	0.241	0.188
	(0.089)	(0.342)	(0.085)	(0.068)	(0.174)	(0.069)
MSE$_N$	0.541	0.118	0.018		0.006	
	(0.083)	(0.056)	(0.042)		(0.011)	
TP-Mean	5	4.733	5		4.97	
Median	5	5	5		5	
	(0)	(0.445)	(0)		(0.171)	
FP-Mean	206.26	31.14	3.56		3.58	
Median	207	31	3		3	
	(13.658)	(9.024)	(2.231)		(2.235)	

8.3 Design 3

To study the sensitivity of our procedure to the minimal non-vanishing signals, 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 non-vanishing signal becomes $|\beta_5| = 0.1$, and we run for $p = 50, 300$ and $n = 200$. All the unimportant regressors are endogenous as in Design 1. Table 6 indicates that the minimal signal is so small that it is not as easily distinguishable from the zero coefficients as before.

Table 6: Performance Measures of FGMM when $p = 50$, $\boldsymbol{\beta}_4 = -0.5$, $\boldsymbol{\beta}_5 = 0.1$

λ	0.001	0.005	0.01	0.05	0.1
MSE_S	0.160	0.155	0.150	0.199	0.277
	(0.050)	(0.047)	(0.055)	(0.051)	(0.163)
MSE_N	0.069	0.074	0.088	0.002	0.003
	(0.017)	(0.016)	(0.028)	(0.011)	(0.014)
TP-Mean	4.61	4.49	4.42	4	3.78
Median	5	4	4	4	4
	(0.492)	(0.502)	(0.496)	(0)	(0.416)
FP-Mean	15.94	3.96	1.48	0.07	0.07
Median	16	3	1	0	0
	(3.405)	(1.959)	(0.959)	(0.383)	(0.356)

Table 7: Performance Measures of FGMM when $p = 300$, $\boldsymbol{\beta}_4 = -0.5$, $\boldsymbol{\beta}_5 = 0.1$

λ	0.001	0.005	0.01	0.05	0.1
MSE_S	0.174	0.164	0.168	0.211	0.247
	(0.055)	(0.054)	(0.056)	(0.061)	(0.156)
MSE_N	0.107	0.097	0.083	5×10^{-4}	0.002
	(0.018)	(0.023)	(0.036)	(0.005)	(0.012)
TP-Mean	4.59	4.52	4.28	4.02	3.83
Median	5	5	4	4	4
	(0.494)	(0.502)	(0.451)	(0.141)	(0.378)
FP-Mean	76.43	7.83	1.4	0.01	0.06
Median	77	7	1	0	0
	(11.19)	(3.613)	(0.985)	(0.1)	(0.371)

9 Conclusion

Endogeneity arises easily in high-dimensional regression due to a large pool of regressors. This causes the inconsistency of the penalized least-squares methods and possible false scientific discoveries. When there exists an endogenous variable whose true regression coefficient is zero, the penalized LS does not satisfy the necessary condition of variable selection consistency regardless of the penalty function.

We propose to penalize an FGMM loss function. It is shown that FGMM possesses the oracle property. By the assumption of over-identification, one can also achieve the oracle property with near global minimization.

We give sufficient and necessary conditions for a general penalized optimization to achieve the consistency for both variable selection and estimation, and apply these results to the sparse conditional moment restricted model, which covers a broad range of applications.

In addition to FGMM, it is also possible to achieve the oracle property using the *penalized empirical likelihood*(PEL). The empirical likelihood was first proposed by Owen (1988). Since it is defined based on estimating equations and moment conditions, it has been an appealing alternative to GMM. The PEL criterion function can be constructed in a similar way, whose oracle properties can also be achieved. We will leave this for future research.

The current paper has assumed that the important regressors be exogenous. In some applications in social sciences, however, they are possibly endogenous as well. In this case, the oracle property should also be achieved with the help of instrumental variables. Recently Gautier and Tsybakov (2011) considered a high dimensional instrumental variable approach. We will explore this direction in depth in the future.

A   Proofs for Section 2

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses.

A.1   Proof of Theorem 2.1

Proof. When $\widehat{\boldsymbol{\beta}}$ is a local minimizer of $Q_n(\boldsymbol{\beta})$, by the Karush-Kuhn-Tucker (KKT) condition, $\forall l \notin S$,

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

where $v_l = P'_n(|\hat{\beta}_l|)\mathrm{sgn}(\hat{\beta}_l)$ if $\hat{\beta}_l = 0$; $v_l \in [-P'_n(0^+), P'_n(0^+)]$ 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

$$\left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} \right| \leq P'_n(0^+). \tag{A.1}$$

By Taylor expansion and the Cauchy-Schwarz inequality, there is $\tilde{\boldsymbol{\beta}}$ on the segment joining $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ so that

$$\max_{l \notin S} \left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \right| \leq \max_{l,j \leq p} \left| \frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial \beta_l \partial \beta_j} \right| \sqrt{s} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|.$$

Since $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = o_p(1)$, and due to the condition of the theorem, we have

$$\max_{l \notin S} \left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \right| \to^p 0. \tag{A.2}$$

Combining the last two labeled results, we conclude that

$$\frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \to^p 0.$$

Q.E.D.

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. Note that in penalized LS, $L_n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \mathbf{X}_i^T\boldsymbol{\beta})^2$. 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 the necessary condition of Theorem 2.1. Q.E.D.

B   Proofs for Section 4

B.1   Proof of Theorem 4.1

Lemma B.1. *Under Assumptions 4.1, and $s/\sqrt{n} = o(d_n)$, if $\boldsymbol{\beta} = (\beta_1, ..., \beta_s)^T$ is such that $\max_{i \leq s} |\beta_j - \beta_{0S,j}| \leq d_n$, then*

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

Proof. By Taylor's expansion, there exists $\boldsymbol{\beta}^*$ lying on the line segment joining $\boldsymbol{\beta}$ and $\boldsymbol{\beta}_{0S}$,

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

Then $\min\{|\beta_j^*| : j \leq s\}$

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

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

Proof of Theorem 4.1

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 $\mathcal{N}_\tau = \{\boldsymbol{\beta} \in \mathbb{R}^s : \|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \leq k_n\tau\}$ 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 $(\widehat{\boldsymbol{\beta}}_S^T, 0)^T$ of Q_n restricted on $\mathcal{B}$ inside $\{\boldsymbol{\beta} =$

$(\boldsymbol{\beta}_S^T, 0)^T : \boldsymbol{\beta}_S \in \mathcal{N}_\tau\}$. Therefore,

$$P(\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \le k_n\tau) \ge P(H_n(\tau)).$$

Hence it suffices to show that $\forall \varepsilon > 0$, there exists $\tau > 0$ so that $P(H_n(\tau)) > 1 - \varepsilon$, and that the local minimizer is strict.

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

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

By Condition (i) that $\|\nabla L_1(\boldsymbol{\beta}_{0S})\| = O_p(a_n)$, for any $\varepsilon > 0$, there exists $C_1 > 0$, so that

$$P((\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S})^T \nabla L_1(\boldsymbol{\beta}_{0S}) \ge -C_1\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|a_n) > 1 - \varepsilon. \tag{B.1}$$

In addition, Condition (ii) yields that there exists $C > 0$ such that w.p.a.1,

$$(\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S})^T \nabla^2 L_1(\boldsymbol{\beta}_{0S})(\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}) > C\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|^2.$$

Hence by the continuity of $\nabla^2 L_1(\cdot)$, and that $\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\| \to 0$,

$$(\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S})^T \nabla^2 L_1(\boldsymbol{\beta}^*)(\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}) > \frac{C}{2}\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|^2.$$

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 we can choose $\tau > 0$ large enough (for example, $\tau C/4 > \max\{1, C_1\}$) so that, on the event

$$(\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S})^T \nabla L_1(\boldsymbol{\beta}_{0S}) \ge -C_1\|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|a_n,$$

we have:

$$\min_{\boldsymbol{\beta} \in \partial\mathcal{N}_\tau} Q_1(\boldsymbol{\beta}) - Q_1(\boldsymbol{\beta}_{0S}) \ge \|\boldsymbol{\beta}_S - \boldsymbol{\beta}_{0S}\|(\frac{k_n\tau C}{4} - C_1 a_n - \sqrt{s}P_n'(d_n)) > 0.$$

By (B.1), $P(H_n(\tau)) > 1 - \varepsilon$.

It remains to show that the local minimizer in $\mathcal{N}_\tau$ (denoted by $\widehat{\boldsymbol{\beta}}_S$) is strict. 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 $\boldsymbol{\beta}_S \in \mathcal{N}_\tau$ Let

$$\mathbf{A}(\boldsymbol{\beta}_S) = \nabla^2 L_1(\boldsymbol{\beta}_S) - \mathrm{diag}\{\psi(\boldsymbol{\beta}_{S1}), ..., \psi(\boldsymbol{\beta}_{Ss})\}.$$

Since $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = o_p(1)$, by Condition (ii), there exists $C > 0$ such that for any non-vanishing $\boldsymbol{\alpha} \in \mathbb{R}^s$, with probability approaching one,

$$\boldsymbol{\alpha}^T \mathbf{A}(\widehat{\boldsymbol{\beta}}_S) \boldsymbol{\alpha} \geq C \boldsymbol{\alpha}^T \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \boldsymbol{\alpha} \max_{j \leq s} \psi(\hat{\beta}_{Sj}).$$

By assumption $k_n = o(d_n)$, hence $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \leq d_n$ w.p.a.1. By the definition of $\eta(\cdot)$, w.p.a.1,

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

Therefore,

$$P(\boldsymbol{\alpha}^T \mathbf{A}(\widehat{\boldsymbol{\beta}}_S) \boldsymbol{\alpha} \geq \|\boldsymbol{\alpha}\|(C - \sup_{\boldsymbol{\beta} \in B(\boldsymbol{\beta}_{0S}, d_n)} \eta(\boldsymbol{\beta}))) \to 1,$$

which implies $\boldsymbol{\alpha}^T \mathbf{A}(\widehat{\boldsymbol{\beta}}_S) \boldsymbol{\alpha} > C/2$ w.p.a.1 by Assumption 4.1. Therefore $\mathbf{A}(\widehat{\boldsymbol{\beta}}_S)$ is positive definite w.p.a.1. Q.E.D.

B.2   Proof of Theorem 4.2

Proof. Let $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, 0)^T$, with $\widehat{\boldsymbol{\beta}}_S \in \mathcal{N}_\tau$ being a strict local minimizer of $L_1(\boldsymbol{\beta}_S)$, as in the proof of Theorem 4.1. It remains to prove that $\widehat{\boldsymbol{\beta}}$ is indeed a strict local minimizer of $Q_n(\boldsymbol{\beta})$ on the space $\mathbb{R}^p$. To show this, take a sufficiently small ball $\mathcal{N}_1$ in $\mathbb{R}^p$ centered at $\widehat{\boldsymbol{\beta}}$ such that

$$\mathcal{N}_1 \cap \mathcal{B} \subset \{(\boldsymbol{\beta}_S^T, 0)^T : \boldsymbol{\beta}_S \in \mathcal{N}_\tau\}. \tag{B.2}$$

We recall the definition

$$\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } \beta_{0j} = 0\},$$

which is $\{\boldsymbol{\beta} = \mathbb{T}\boldsymbol{\beta}\}$. We then need to show that $\forall \gamma \in \mathcal{N}_1 \backslash \{\widehat{\boldsymbol{\beta}}\}$, $Q_n(\widehat{\boldsymbol{\beta}}) < Q_n(\gamma)$ w.p.a.1. Note that if $\gamma = (\gamma_S^T, \gamma_N^T)^T$ with $\gamma_N = 0$, then $\gamma \in \mathcal{B}$ and by Theorem 4.1, $Q_n(\widehat{\boldsymbol{\beta}}) < Q_n(\gamma)$. Therefore we consider the case when $\gamma_N \neq 0$. In addition, note that $Q_n(\widehat{\boldsymbol{\beta}}) \leq Q_n(\mathbb{T}\gamma)$, where $\mathbb{T}(\gamma) = (\gamma_S^T, 0)$, the projection of γ onto $\mathcal{B}$. Thus, it suffices to show:

Claim: There exists a sufficiently small $\mathcal{N}_1$ satisfying (B.2) such that $\forall \gamma \in \mathcal{N}_1$, with $\gamma_N \neq 0$, $Q_n(\mathbb{T}\gamma) < Q_n(\gamma)$ w.p.a.1.

In fact, this is implied by Condition (4.2):

$$Q_n(\mathbb{T}\gamma) - Q_n(\gamma) = L_n(\mathbb{T}\gamma) - L_n(\gamma) - (\sum_{j=1}^{p} P_n(\gamma_j) - \sum_{j=1}^{s} P_n(|(\mathbb{T}\gamma)_j|)) < 0.$$

If L_n is continuously differentiable in a neighborhood of $\boldsymbol{\beta}_0$, by the mean value theorem, there exists $\lambda \in (0,1)$ such that for $h = \lambda\gamma + (1-\lambda)\mathbb{T}\gamma$,

$$
\begin{aligned}
Q_n(\mathbb{T}\gamma) - Q(\gamma) &= \sum_{l \notin S} \frac{\partial L_n(h)}{\partial \beta_l}(-\gamma_l) - \sum_{l \notin S} P'_n(|h_l|)|\gamma_l| \\
&\leq \sum_{l \notin S} \left(\left| \frac{\partial L_n(h)}{\partial \beta_l} \right| - P'_n(|h_l|) \right) |\gamma_l|,
\end{aligned}
$$

where we used $dP_n(|t|)/dt = P'_n(|t|)\mathrm{sgn}(t)$, and the fact that $\mathrm{sgn}(h_l) = \mathrm{sgn}(\gamma_l)$ for $l \notin S$. It thus suffices to show, the following holds w.p.a.1:

$$\max_{l \notin S} \left| \frac{\partial L_n(h)}{\partial \beta_l} \right| - P'_n(|h_l|) < 0.$$

Suppose we have

$$\max_{l \notin S} \left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} \right| = o_p(P'_n(0^+)), \tag{B.3}$$

then by continuity, there is $\delta > 0$, for any $\boldsymbol{\beta}$ in a ball in $\mathbb{R}^p$ centered at $\widehat{\boldsymbol{\beta}}$ with radius δ,

$$\max_{l \notin S} \left| \frac{\partial L_n(\boldsymbol{\beta})}{\partial \beta_l} \right| - P'_n(\delta) < 0.$$

We further shrink the radius of the ball $\mathcal{N}_1$ to less than δ so that $|\gamma_j| < \delta$ for any $j \notin S$. Hence

$$
\begin{aligned}
\max_{l \notin S} \left| \frac{\partial L_n(h)}{\partial \beta_l} \right| - P'_n(|h_l|) &= \max_{l \notin S} \left| \frac{\partial L_n(h)}{\partial \beta_l} \right| - P'_n(\lambda|\gamma_l|) \\
&\leq \max_{l \notin S} \left| \frac{\partial L_n(h)}{\partial \beta_l} \right| - P'_n(\delta) < 0,
\end{aligned}
$$

where we used the monotonicity of $P'_n(\cdot)$. Hence it remains to prove (B.3). By the triangular inequality,

$$\max_{l \notin S} \left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} \right| \leq \max_{l \notin S} \left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \right| + \max_{l \notin S} \left| \frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \right|.$$

By assumption, $\max_{l \notin S} |\frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l}| = o_p(P_n'(0^+))$. For the first term on the right hand side, apply the mean value theorem (note that $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ only differ at the coordinates in S),

$$
\begin{aligned}
\max_{l \notin S} \left| \frac{\partial L_n(\widehat{\boldsymbol{\beta}})}{\partial \beta_l} - \frac{\partial L_n(\boldsymbol{\beta}_0)}{\partial \beta_l} \right| &\leq \max_{l \notin S} \left| \sum_{j \in S} \frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial \beta_l \partial \beta_j}(\hat{\beta}_j - \beta_{0j}) \right| \\
&\leq \max_{l,j \leq p} \left| \frac{\partial^2 L_n(\tilde{\boldsymbol{\beta}})}{\partial \beta_l \partial \beta_j} \right| \sqrt{s} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \\
&= o_p(P_n'(0^+)).
\end{aligned}
$$

where $\tilde{\boldsymbol{\beta}}$ lies on the line segment joining $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$, and we used the Cauchy-Schwarz inequality.

Q.E.D.

B.3 Proof of Theorem 4.3

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

$$
-P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \text{sgn}(\widehat{\boldsymbol{\beta}}_S) = \nabla_S L_n(\widehat{\boldsymbol{\beta}}_S, 0),
$$

where $\circ$ denotes the Hadamard product of two vectors. By the mean value theorem, there exists $\boldsymbol{\beta}^*$ lying on the segment joining $\boldsymbol{\beta}_{0S}$ and $\widehat{\boldsymbol{\beta}}_S$ such that

$$
\nabla_S L_n(\widehat{\boldsymbol{\beta}}_S, 0) = \nabla_S L_n(\boldsymbol{\beta}_{0S}, 0) + \nabla_S^2 L_n(\boldsymbol{\beta}^*, 0)(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}).
$$

Since $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = o_p(1)$, we have $\nabla_S^2 L_n(\boldsymbol{\beta}^*, 0) = \nabla_S^2 L_n((\boldsymbol{\beta}_{0S}, 0) + o_p(1)$, where $o_p(1)$ is in terms of the Frobenius norm. Therefore,

$$
(\nabla_S^2 L_n((\boldsymbol{\beta}_{0S}, 0) + o_p(1))(\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}) = -P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \text{sgn}(\widehat{\boldsymbol{\beta}}_S) - \nabla_S L_n(\boldsymbol{\beta}_{0S}, 0). \tag{B.4}
$$

For any unit vector $\boldsymbol{\alpha} \in \mathbb{R}^s$, by Condition (ii), $\|\boldsymbol{\alpha}^T \Omega_n[P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \text{sgn}(\widehat{\boldsymbol{\beta}}_S)]\| = o_p(1)$. Hence the result follows immediately from (B.4) and Condition (i). Q.E.D.

C Proofs for Section 5

According to Theorems 4.1 and 4.2, minimization of Q_{FGMM} can be first constrained on $\mathcal{B} = \{\boldsymbol{\beta} \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin S\}$, and consider $\tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_S) = L_{\text{FGMM}}(\boldsymbol{\beta}_S, 0)$ instead, which is assumed to be twice differentiable. We then proceed to show by using Theorem 4.1 that $\widehat{\boldsymbol{\beta}}_S$

is a local solution to

$$\min_{\boldsymbol{\beta}_S} \tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_S) + \sum_{j=1}^{s} P_n(|\beta_j|)$$

and that $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = o_p(1)$. After that, we use Theorem 4.2 to conclude that $(\widehat{\boldsymbol{\beta}}_S^T, 0)^T$ is also a local solution to $\min_{\boldsymbol{\beta} \in \mathbb{R}^p} Q_{\text{FGMM}}(\boldsymbol{\beta})$.

Throughout the proof, we write $\mathbf{X}_{iS}^2 = \mathbf{X}_i^2(\boldsymbol{\beta}_{0S})$ and $\mathbf{V}_{iS} = (\mathbf{X}_{iS}^T, \mathbf{X}_{iS}^{2T})^T$.

C.1 Lemmas

Lemma C.1. *(i)* $\max_{l \le p} |\frac{1}{n} \sum_{i=1}^{n} (X_{ij} - \overline{X_j})^2 - \text{var}(X_j)| = o_p(1)$.
(ii) $\max_{l \le p} |\frac{1}{n} \sum_{i=1}^{n} (X_{ij}^2 - \overline{X_j^2})^2 - \text{var}(X_j^2)| = o_p(1)$.
(iii) $\sup_{\boldsymbol{\beta} \in \mathbb{R}^p} \lambda_{\max}(\mathbf{W}(\boldsymbol{\beta})) = O_p(1)$, *and* $\lambda_{\min}(\mathbf{W}(\boldsymbol{\beta}_0))$ *is bounded away from zero w.p.a.1.*

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 by the assumption that $\text{var}(X_j)$ and $\text{var}(X_j^2)$ are bounded uniformly in $j \le p$.

Lemma C.2. *If* $\mathbf{A}$, $\mathbf{B}$ *and* $\mathbf{A} - \mathbf{B}$ *are all semi-positive definite, then* $\lambda_{\max}(\mathbf{A}) \ge \lambda_{\max}(\mathbf{B})$.

Proof. Let $\boldsymbol{\alpha}$ be the eigenvector of $\mathbf{B}$ corresponding to the largest eigenvalue, $\|\boldsymbol{\alpha}\| = 1$. Then

$$
\begin{aligned}
\lambda_{\max}(\mathbf{A}) - \lambda_{\max}(\mathbf{B}) &= \lambda_{\max}(\mathbf{A}) - \boldsymbol{\alpha}^T \mathbf{B} \boldsymbol{\alpha} \\
&= \lambda_{\max}(\mathbf{A}) + \boldsymbol{\alpha}^T (\mathbf{A} - \mathbf{B}) \boldsymbol{\alpha} - \boldsymbol{\alpha}^T \mathbf{A} \boldsymbol{\alpha} \\
&\ge \lambda_{\max}(\mathbf{A}) - \boldsymbol{\alpha}^T \mathbf{A} \boldsymbol{\alpha} \ge 0.
\end{aligned}
$$

Lemma C.3. $\max_{j \in S} \|\frac{1}{n} \sum_{i=1}^{n} m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{ij} \mathbf{V}_{iS}\|_2^2 = O_p(\eta_n^2 + \frac{s \log s}{n})$.

Proof. Note that the Bernstein inequality plus Bonferroni's method imply that

$$
\max_{j \in S} \|\frac{1}{n} \sum_{i=1}^{n} m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{ij} \mathbf{V}_{iS}\|_2
$$
$$
\le \max_{j \in S} \|E m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_j \mathbf{V}_S\|_2 + O_p(\sqrt{\frac{s \log s}{n}}).
$$

Since $E m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0)^2 X_j^2 \mathbf{V}_S \mathbf{V}_S^T - E m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_j \mathbf{V}_S E m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_j \mathbf{V}_S^T$ is semi-positive definite, by Lemma C.2 and Assumption 5.5,

$$\|E m(Y, \mathbf{X}^T \boldsymbol{\beta}_0) X_j \mathbf{V}_S\|_2^2 \le \lambda_{\max}(E m(Y, \mathbf{X}^T \boldsymbol{\beta}_0)^2 X_j^2 \mathbf{V}_S \mathbf{V}_S^T) = O(\eta_n^2).$$

C.2 Proof of Theorem 5.1

C.2.1 Consistency

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

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

Then $\tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_S) = L_{\text{FGMM}}(\boldsymbol{\beta}_S, 0)$. We proceed by verifying the conditions in Theorem 4.1.

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

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

By Assumption 5.4, $\|\mathbf{A}_n(\boldsymbol{\beta}_0)\|_2 = O_p(1)$. In addition, the elements in $\mathbf{W}(\boldsymbol{\beta}_0)$ are uniformly bounded in probablity due to Lemma C.1. Hence

$$\|\nabla \tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_{0S})\| \leq O_p(1) \|\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S}) \mathbf{V}_{iS}\|.$$

Due to $Eg(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{X}_S = Eg(Y, \mathbf{X}_S^T \boldsymbol{\beta}_{0S}) \mathbf{X}_S^2 = 0$, using the exponential-tail Bernstein inequality with Assumption 5.2 plus Bonferroni's method, it can be shown that for any $t > 0$,

$$P(\max_{l \in S} |\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S}) X_{li}| > t) < s \max_{l \in S} P(|\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S}) X_{li}| > t)$$

$$\leq \exp\left(\log s - \frac{Ct^2}{n} \right),$$

which implies that

$$\max_{l \in S} |\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S}) X_{li}| = O_p(\sqrt{\frac{\log s}{n}}). \tag{C.2}$$

Similarly,

$$\max_{l \in S} |\frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S}) X_{li}^2| = O_p(\sqrt{\frac{\log s}{n}}). \tag{C.3}$$

Hence $\|\nabla \tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_{0S})\| = O_p(\sqrt{(s \log s)/n})$.

Condition (ii) Straightforward but tedious calculation yields $\nabla^2 \tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_{0S}) = \boldsymbol{\Sigma}(\boldsymbol{\beta}_{0S}) + \mathbf{M}(\boldsymbol{\beta}_{0S})$, where

$$\boldsymbol{\Sigma}(\boldsymbol{\beta}_{0S}) = 2\mathbf{A}_n(\boldsymbol{\beta}_{0S})\mathbf{W}(\boldsymbol{\beta}_0)\mathbf{A}_n(\boldsymbol{\beta}_{0S})^T,$$

and

$$\mathbf{M}(\boldsymbol{\beta}_{0S}) = 2\mathbf{H}(\boldsymbol{\beta}_{0S})\mathbf{B}(\boldsymbol{\beta}_{0S})$$

with (suppose $\mathbf{X}_{iS} = (X_{il_1}, ..., X_{il_s})^T$)

$$
\begin{aligned}
\mathbf{H}(\boldsymbol{\beta}_{0S}) &= \frac{1}{n}\sum_{i=1}^n q_i(Y_i, \mathbf{X}_{iS}\boldsymbol{\beta}_{0S})(X_{il_1}\mathbf{X}_{iS}, ..., X_{il_s}\mathbf{X}_{iS})\mathbf{V}_{iS}^T, \\
\mathbf{B}(\boldsymbol{\beta}_{0S}) &= \mathbf{W}(\boldsymbol{\beta}_0)\left[\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}\right].
\end{aligned}
$$

It is not hard to obtain $\|\mathbf{B}(\boldsymbol{\beta}_{0S})\| = O_p(\sqrt{s\log s/n})$, and $\|\mathbf{H}(\boldsymbol{\beta}_{0S})\| = O_p(s)$, and hence $\|\mathbf{M}(\boldsymbol{\beta}_{0S})\| = O_p(s\sqrt{s\log s/n}) = o_p(1)$. Therefore, the eigenvalues of $\nabla^2 \tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_{0S})$ are bounded away from zero w.p.a.1.

C.2.2 Sparsity

To show the sparsity, we check (4.2) in Theorem 4.2.
For some neighborhood $\mathcal{N}$ of $(\widehat{\boldsymbol{\beta}}_S^T, 0)^T$, and $\forall \gamma \in \mathcal{N}$, write

$$\gamma = (\gamma_S^T, \gamma_N^T)^T, \text{ and } \mathbb{T}\gamma = (\gamma_S^T, 0)^T.$$

In addition, we write $\mathbf{V}_i(\gamma_S) = \mathbf{V}_i(\mathbb{T}\gamma)$, $\mathbf{V}_i(\gamma_N) = \mathbf{V}_i(\gamma - \mathbb{T}\gamma)$, and $\mathbf{W}(\gamma_S) = \mathbf{W}(\mathbb{T}\gamma)$ for notational simplicity.

For all $\theta \in \mathbb{R}^p$, define

$$F(\theta) = \left[\frac{1}{n}\sum_{i=1}^n g(Y_i, X_i^T\theta)\mathbf{V}_i(\gamma_S)\right]^T \mathbf{W}(\gamma_S)\left[\frac{1}{n}\sum_{i=1}^n g(Y_i, X_i^T\theta)\mathbf{V}_i(\gamma_S)\right].$$

Hence $L_{\text{FGMM}}(\mathbb{T}\gamma) = F(\mathbb{T}\gamma)$, and $L_{\text{FGMM}}(\gamma) = F(\gamma) + \xi_2(\gamma)$, where

$$\xi_2(\gamma) = (\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_i^T\gamma)\mathbf{V}_i(\gamma_N))^T\mathbf{W}(\gamma_N)(\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_i^T\gamma)\mathbf{V}_i(\gamma_N)) \geq 0.$$

Hence

$$L_{\text{FGMM}}(\mathbb{T}\gamma) - L_{\text{FGMM}}(\gamma) \leq F(\mathbb{T}\gamma) - F(\gamma).$$

Note that $\mathbb{T}\gamma - \gamma = (0, -\gamma_N^T)^T$. By the mean value theorem, there exists $\lambda \in (0, 1)$, for $h = (\gamma_S^T, -\lambda\gamma_N^T)^T$,

$$
\begin{aligned}
F(\mathbb{T}\gamma) - F(\gamma) &- [\sum_{j=1}^p (P_n(|\gamma_j|) - P_n(|(\mathbb{T}\gamma)_j|)] \\
&= -\sum_{l \notin S, \gamma_l \neq 0} \gamma_l \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta_l} g(Y_i, \mathbf{X}_i^T h)\mathbf{V}_i(\gamma_S)\right]^T \mathbf{W}(\gamma_S)\left[\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_i^T h)\mathbf{V}_i(\gamma_S)\right] \\
&\quad - \sum_{l \notin S, \gamma_l \neq 0} |\gamma_l|P_n'(\lambda|\gamma_l|) \\
&\equiv \sum_{l \notin S, \gamma_l \neq 0} \gamma_l a_l(h) - |\gamma_l|P_n'(\lambda|\gamma_l|).
\end{aligned}
$$

Hence it suffices to show that there exists $\mathcal{N}$ so that for any $\gamma \in \mathcal{N}$,

$$
\max_{l \notin S, \gamma_l \neq 0} |\gamma_l a_l(h)| - |\gamma_l|P_n'(\lambda|\gamma_l|) < 0. \tag{C.4}
$$

Suppose we have, for $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, 0)^T$,

$$
\max_{l \notin S, \gamma_l \neq 0} |a_l(\widehat{\boldsymbol{\beta}})| = o_p(P_n'(0^+)), \tag{C.5}
$$

by continuity, there is $\delta > 0$, for any $\boldsymbol{\beta}$ in a ball in $\mathbb{R}^p$ centered at $\widehat{\boldsymbol{\beta}}$ with radius δ,

$$
\max_{l \notin S, \gamma_l \neq 0} |a_l(\boldsymbol{\beta})| - P_n'(\delta) < 0.
$$

We further shrink the radius of $\mathcal{N}$ to less than δ so that $|\gamma_l| < \delta$ for any $l \notin S$. By the monotonicity of $P_n'(\cdot)$,

$$
\max_{l \notin S, \gamma_l \neq 0} |a_l(h)| - P_n'(\lambda|\gamma_l|) \leq \max_{l \notin S, \gamma_l \neq 0} |a_l(h)| - P_n'(\delta) < 0.
$$

Hence it remains to prove (C.5). By the triangular inequality,

$$
\max_{l \notin S, \gamma_l \neq 0} |a_l(\widehat{\boldsymbol{\beta}})| \leq \max_{l \notin S} |a_l(\widehat{\boldsymbol{\beta}}) - a_l(\boldsymbol{\beta}_0)| + \max_{l \notin S} |a_l(\boldsymbol{\beta}_0)|.
$$

Since $E(g(Y, \mathbf{X}^T\boldsymbol{\beta}_0)|\mathbf{X}_S) = 0$, by Assumption 5.5, and (C.2)(C.3)

$$
\max_{l \notin S} |a_l(\boldsymbol{\beta}_0)| \leq \left\|\frac{1}{n}\sum_{i=1}^n m(Y_i, \mathbf{X}_i^T\boldsymbol{\beta}_0)X_{il}\mathbf{V}_i(\gamma_S)\right\|\left\|\mathbf{W}(\gamma_S)\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_i^T h)\mathbf{V}_i(\gamma_S)\right\|
$$

$$= O_p((\kappa_n + \sqrt{\frac{s \log p}{n}})\sqrt{(s \log s)/n}) = o_p(P'_n(0^+)),$$

where we used the triangular and Bernstein inequalities to obtain

$$\max_{l \notin S} \left\| \frac{1}{n} \sum_{i=1}^{n} m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{il} \mathbf{V}_i(\gamma_S) \right\| \leq \max_{l \notin S} \left\| Em(Y, \mathbf{X}^T \boldsymbol{\beta}_0) X_l \mathbf{V}_S \right\|$$

$$+ \max_{l \notin S} \left\| \frac{1}{n} \sum_{i=1}^{n} m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{il} \mathbf{V}_i(\gamma_S) - Em(Y, \mathbf{X}^T \boldsymbol{\beta}_0) X_l \mathbf{V}(\gamma_S) \right\|$$

$$= O(\kappa_n) + O_p(\sqrt{\frac{s \log p}{n}}).$$

On the other hand, applying the mean value theorem and Cauchy-Schwarz inequality gives (note that $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$ only differ at the coordinates in S),

$$\max_{l \notin S} |a_l(\widehat{\boldsymbol{\beta}}) - a_l(\boldsymbol{\beta}_0)| \leq \max_{l \notin S, j \in S} \left| \frac{\partial a_l(\tilde{\boldsymbol{\beta}})}{\partial \beta_j} \right| \sqrt{s} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = o_p(P'_n(0^+)).$$

where $\tilde{\boldsymbol{\beta}}$ lies on the line segment joining $\widehat{\boldsymbol{\beta}}$ and $\boldsymbol{\beta}_0$. Note that

$$\max_{l \notin S, j \in S} \left| \frac{\partial a_l(\boldsymbol{\beta}_0)}{\partial \beta_j} \right| \leq \left\| \frac{1}{n} \sum_{i=1}^{n} q(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{ij} X_{il} \mathbf{V}_i^T \mathbf{W}(\gamma_S) \frac{1}{n} \sum_{i=1}^{n} g(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) \mathbf{V}_i \right\|$$

$$+ \left\| \frac{1}{n} \sum_{i=1}^{n} m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{il} \mathbf{V}_i^T \mathbf{W}(\gamma_S) \frac{1}{n} \sum_{i=1}^{n} m(Y_i, \mathbf{X}_i^T \boldsymbol{\beta}_0) X_{ij} \mathbf{V}_i \right\|$$

$$= O_p(\sqrt{s \log s/n} + (\sqrt{s \log p/n} + \kappa_n)(\sqrt{s \log s/n} + \eta_n)),$$

where in the last equality, we used Lemma C.3 to bound the second term on the right. Therefore, (C.5) holds as long as $\kappa_n \eta_n s(P'_n(d_n) + \sqrt{\log s/n}) = o(P'_n(0^+))$. Q.E.D.

C.3 Proof of Theorem 5.2

Let $P'_n(|\widehat{\boldsymbol{\beta}}_S|) = (P'_n(|\hat{\beta}_{S1}|), ..., P'_n(|\hat{\beta}_{Ss}|))^T$. The asymptotic normality builds on the following lemmas.

Lemma C.4. *Under Assumption 4.1 and $s/\sqrt{n} = o(d_n)$, for $a_n, \widehat{\boldsymbol{\beta}}_S$ defined in Theorem 4.1,*

$$\|P'_n(|\widehat{\boldsymbol{\beta}}_S|) \circ \text{sgn}(\widehat{\boldsymbol{\beta}}_S)\| = O_p(\max_{\boldsymbol{\beta} \in \mathcal{N}_1} \eta(\boldsymbol{\beta}) a_n + \sqrt{s} P'_n(d_n)),$$

where $\mathcal{N}_1 = \{\boldsymbol{\beta} \in \mathbb{R}^s : \|\boldsymbol{\beta} - \boldsymbol{\beta}_{0S}\| \leq C\sqrt{(s \log s)/n}\}$, for some $C > 0$, and $\circ$ denotes the

element-wise product.

Proof. Write

$$P'_n(|\widehat{\boldsymbol{\beta}}_S|) \circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S) = (v_1, ..., v_s)^T, \text{ where } v_i = P'_n(|\hat{\beta}_{Si}|)\mathrm{sgn}(\hat{\beta}_{Si}).$$

By the triangular inequality and Taylor expansion,

$$|v_i| \leq |P'_n(|\hat{\beta}_{Si}|) - P'_n(|\beta_{0S,i}|)| + P'_n(|\beta_{0S,i}|) \leq \max_{\boldsymbol{\beta} \in \mathcal{N}_1} \eta(\boldsymbol{\beta})|\hat{\beta}_{Si} - \beta_{0S,i}| + P'_n(d_n).$$

Therefore,

$$
\begin{aligned}
\|P'_n(|\widehat{\boldsymbol{\beta}}_S|) \circ \mathrm{sgn}(\widehat{\boldsymbol{\beta}}_S)\|^2 &= \sum_{i=1}^s v_j^2 \leq 2\sum_{i=1}^s \max_{\mathcal{N}_1} \eta(\boldsymbol{\beta})^2 |\hat{\beta}_{Si} - \beta_{Si}|^2 + 2sP'_n(d_n)^2 \\
&\leq 2\max_{\boldsymbol{\beta} \in \mathcal{N}_1} \eta(\boldsymbol{\beta})^2 \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|^2 + 2sP'_n(d_n)^2,
\end{aligned}
$$

which implies the result since $\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = O_p(a_n + \sqrt{s}P'_n(d_n))$. Q.E.D.

Lemma C.5. *Let* $\boldsymbol{\Omega}_n = \sqrt{n}\boldsymbol{\Gamma}_n^{-1/2}$. *Then for any unit vector* $\boldsymbol{\alpha} \in \mathbb{R}^s$,

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

Proof. $\nabla \tilde{L}_{\mathrm{GMM}}(\boldsymbol{\beta}_{0S}) = 2\mathbf{A}_n(\boldsymbol{\beta}_{0S})\mathbf{W}(\boldsymbol{\beta}_0)\mathbf{B}_n$, 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

$$
\begin{aligned}
\boldsymbol{\Gamma}_n &= 4\mathbf{H}\mathbf{W}(\boldsymbol{\beta}_0)\mathbf{V}_0\mathbf{W}(\boldsymbol{\beta}_0)^T\mathbf{H}^T, \quad s \times s \\
\mathbf{V}_0 &= \mathrm{var}(\sqrt{n}\mathbf{B}_n) = \mathrm{var}(g(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{V}_S), \quad 2s \times 2s \\
\mathbf{H} &= Em(Y, \mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S^T, \quad s \times 2s.
\end{aligned}
$$

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

$$\|\mathbf{A}_n(\boldsymbol{\beta}_{0S}) - \mathbf{H}\| = o_p(1), \text{ and}$$

$$\sqrt{n}\tilde{\boldsymbol{\alpha}}^T\mathbf{V}_0^{-1/2}\mathbf{B}_n \to^d N(0,1).$$

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

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

Q.E.D.

Note that in the proof of Theorem 5.1, condition (ii), we showed that

$$\nabla^2\tilde{L}_{\text{GMM}}(\boldsymbol{\beta}_{0S}) = \boldsymbol{\Sigma}_n + o_p(1)$$

where $o_p(1)$ is in terms of the Frobenius norm. By Theorem 4.3. it remains to check that for $\Omega_n = \sqrt{n}\boldsymbol{\Gamma}_n^{-1/2}$, Condition (ii) in Theorem 4.3 holds. By Assumptions 5.4 and 5.6(i), $\lambda_{\min}(\boldsymbol{\Gamma}_n)^{-1/2} = O_p(1)$. Lemma C.4 then implies

$$\begin{aligned}
&\sqrt{n}\lambda_{\min}(\boldsymbol{\Gamma}_n)^{-1/2}\|P_n'(|\widehat{\boldsymbol{\beta}}_S|) \circ \text{sgn}(\widehat{\boldsymbol{\beta}}_S\| \\
\leq \ & C\sqrt{n}(\max\eta(\boldsymbol{\beta})\sqrt{s\log s/n} + \sqrt{s}P_n'(d_n)) \\
= \ & O_p(\sqrt{s\log s}\max\eta(\boldsymbol{\beta}) + \sqrt{ns}P_n'(d_n)) = o_p(1).
\end{aligned}$$

Q.E.D.

D   Proofs for Sections 6 and 7

The local minimizer in Theorem 5.1 is denoted by $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_S^T, \widehat{\boldsymbol{\beta}}_N^T)^T$, and $P(\widehat{\boldsymbol{\beta}}_N = 0) \to 1$. Let $\widehat{\boldsymbol{\beta}}_G = (\widehat{\boldsymbol{\beta}}_S^T, 0)^T$.

D.1   Proof of Theorem 6.1

Lemma D.1.

$$L_{FGMM}(\widehat{\boldsymbol{\beta}}_G) = O_p\left(\frac{s\log s}{n} + sP_n'(d_n)^2\right).$$

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

$$\begin{aligned}
\|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\widehat{\boldsymbol{\beta}}_S)\mathbf{V}_{iS}\| &\leq \|\frac{1}{n}\sum_{i=1}^n g(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{V}_{iS}\| \\
&+ \|\frac{1}{n}\sum_{i=1}^n m(Y_i, \mathbf{X}_{iS}^T\tilde{\boldsymbol{\beta}}_S)\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|_2\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \\
&\leq O_p(\sqrt{s\log s/n}) + \|\frac{1}{n}\sum_{i=1}^n m(Y_i, \mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|_2\|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|
\end{aligned}$$

$$+\frac{1}{n}\sum_{i=1}^{n}|m(Y_i,\mathbf{X}_{iS}^T\tilde{\boldsymbol{\beta}}_S)-m(Y_i,\mathbf{X}_{iS}^T\boldsymbol{\beta}_{0S})|\|\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|.$$

Note that $\|Em(Y,\mathbf{X}_S^T\boldsymbol{\beta}_{0S})\mathbf{X}_S\mathbf{V}_S\|_2$ is bounded due to Assumption 5.4. Apply Taylor expansion again, with some $\tilde{\boldsymbol{\beta}}^*$, the above term is bounded by

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

Note that $\sup_{t_1,t_2}|q(t_1,t_2)|<\infty$ by Assumption 5.3. We thus have,

$$\frac{1}{n}\sum_{i=1}^{n}|q(Y_i,\mathbf{X}_{iS}^T\tilde{\boldsymbol{\beta}}_S^*)|\|\mathbf{X}_{iS}\|\|\tilde{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|\|\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|$$
$$\leq C\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{X}_{iS}\|\|\mathbf{X}_{iS}\mathbf{V}_{iS}^T\|\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2$$
$$\leq CE\|\mathbf{X}_S\|\|\mathbf{X}_S\mathbf{V}_S^T\|(1+o_p(1))\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2.$$

Combining these terms, we obtain

$$\|\frac{1}{n}\sum_{i=1}^{n}g(Y_i,\mathbf{X}_{iS}^T\widehat{\boldsymbol{\beta}}_S)\mathbf{V}_{iS}\| = O_p(\sqrt{s\log s/n}+\sqrt{s}P_n'(d_n))+O_p(s\sqrt{s})\|\widehat{\boldsymbol{\beta}}_S-\boldsymbol{\beta}_{0S}\|^2$$
$$= O_p(\sqrt{s\log s/n}+\sqrt{s}P_n'(d_n)).$$

Lemma D.2.

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

Proof. By the foregoing lemma, we have

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

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

$$\sum_{j=1}^{s}P_n(|\hat{\beta}_{Sj}|) \leq \sum_{j=1}^{s}P_n(|\beta_{0S,j}|)+\sum_{j=1}^{s}P_n'(|\tilde{\beta}_{Sj}|)|\hat{\beta}_{Sj}-\beta_{0S,j}|$$

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

The result then follows. Q.E.D.

Note that $\forall \delta > 0$,

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

Hence by Assumption 6.1, there exists $\varepsilon > 0$,

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

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

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

Q.E.D.

D.2 Proof of Theorem 7.1

Lemma D.3. *Define* $\rho(\boldsymbol{\beta}_S) = E(Y - h(\mathbf{X}_S^T \boldsymbol{\beta}_S))h'(\mathbf{X}_S^T \boldsymbol{\beta}_{0S})\mathbf{X}_S \sigma(\mathbf{X}_S)^{-2}$. *Under the theorem assumptions,*

$$
\sup_{\boldsymbol{\beta}_S \in \Theta} \|\rho(\boldsymbol{\beta}_S) - \rho_n(\boldsymbol{\beta}_S)\| = o_p(1).
$$

Proof. Given $E(\sup_{\boldsymbol{\beta} \in \Theta} h(\mathbf{X}_S^T \boldsymbol{\beta})^4) < \infty$ and $\sup_t |h''(t)| < \infty$, we have the uniform law of large number (Newey and McFadden 1994, Lemma 2.4)

$$
\sup_{\boldsymbol{\beta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} h''(\mathbf{X}_{iS}^T \boldsymbol{\beta})^2 - E h''(\mathbf{X}_S^T \boldsymbol{\beta})^2 = o_p(1),
$$

$$\sup_{\boldsymbol{\beta} \in \Theta} \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{X}_{iS}^T \boldsymbol{\beta})^4 - E h(\mathbf{X}_S^T \boldsymbol{\beta})^4 = o_p(1).$$

Using these, we show three convergence results:

$$\frac{1}{n} \sum_{i=1}^{n} \|Y_i \mathbf{X}_{iS} (h'(\mathbf{X}_{iS}^T \widehat{\boldsymbol{\beta}}_S) - h'(\mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S})) \widehat{\sigma}(\mathbf{X}_{iS})^{-2}\| = o_p(1), \tag{D.1}$$

$$\sup_{\boldsymbol{\beta}_S \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \|h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S) \mathbf{X}_{iS} (h'(\mathbf{X}_{iS}^T \widehat{\boldsymbol{\beta}}_S) - h'(\mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S})) \widehat{\sigma}(\mathbf{X}_{iS})^{-2}\| = o_p(1), \tag{D.2}$$

$$\sup_{\boldsymbol{\beta}_S \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \|(Y_i - h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S)) h'(\mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S}) \mathbf{X}_{iS} (\widehat{\sigma}(\mathbf{X}_{iS})^{-2} - \sigma(\mathbf{X}_{iS})^{-2})\| = o_p(1). \tag{D.3}$$

For (D.1), the left hand side is upper bounded by (for some $\tilde{\boldsymbol{\beta}}$ in the segment joining $\boldsymbol{\beta}_{0S}$ and $\widehat{\boldsymbol{\beta}}_S$, and apply Cauchy-Schwarz inequality)

$$\frac{1}{n} \sum_{i=1}^{n} \|Y_i \mathbf{X}_{iS} \mathbf{X}_{iS}^T h''(\mathbf{X}_{iS}^T \tilde{\boldsymbol{\beta}})\| \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \widehat{\sigma}(\mathbf{X}_{iS})^{-2}$$

$$\leq O_p(1) \sqrt{\frac{1}{n} \sum_{i=1}^{n} \|Y_i \mathbf{X}_{iS} \mathbf{X}_{iS}^T\|^2} \sqrt{\frac{1}{n} \sum_{i=1}^{n} h''(\mathbf{X}_{iS}^T \tilde{\boldsymbol{\beta}})^2} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|$$

$$\leq O_p(1) \sqrt{o_p(1) + \sup_{\boldsymbol{\beta} \in \Theta} E h''(\mathbf{X}_S^T \boldsymbol{\beta})^2} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| = o_p(1),$$

where in the second inequality, we used the uniform weak law of large number. Similarly, the left hand side of (D.2) is upper bounded by

$$\sup_{\boldsymbol{\beta}_S \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \|h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S) \mathbf{X}_{iS} \mathbf{X}_{iS}^T h''(\mathbf{X}_{iS}^T \tilde{\boldsymbol{\beta}})\| \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\| \widehat{\sigma}(\mathbf{X}_{iS})^{-2}$$

$$\leq O_p(1) \left(\sup_{\boldsymbol{\beta}_S \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \|h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S) \mathbf{X}_{iS} \mathbf{X}_{iS}^T\|^2 \right)^{1/2} \left(\frac{1}{n} \sum_{i=1}^{n} h''(\mathbf{X}_{iS}^T \tilde{\boldsymbol{\beta}})^2 \right)^{1/2} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|$$

$$\leq O_p(1) \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_{iS} \mathbf{X}_{iS}^T\|^4 \sup_{\boldsymbol{\beta}_S \in \Theta} \frac{1}{n} \sum_{i=1}^{n} h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S)^4 \right)^{1/4} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|$$

$$\leq O_p(1) \left(\frac{1}{n} \sum_{i=1}^{n} \|\mathbf{X}_{iS} \mathbf{X}_{iS}^T\|^4 (o_p(1) + \sup_{\boldsymbol{\beta}_S \in \Theta} E h(\mathbf{X}_S^T \boldsymbol{\beta}_S)^4) \right)^{1/4} \|\widehat{\boldsymbol{\beta}}_S - \boldsymbol{\beta}_{0S}\|$$

$$= o_p(1),$$

where both the first and second inequalities follow from the Cauchy-Schwarz inequality, and the third inequality follows from the uniform law of large number. (D.3) can be established

in a similar way since $\widehat{\sigma}(\mathbf{X}_S)^2$ uniformly converges to $\sigma(\mathbf{X}_S)^2$.

Due to the previous convergences and that the event $\mathbf{X}_S = \widehat{\mathbf{X}}_S$ occurs with probability approachong one, it remains to show that $\sup_{\boldsymbol{\beta}_S \in \Theta} \|\rho(\boldsymbol{\beta}_S)\| < \infty$ and

$$\sup_{\boldsymbol{\beta}_S \in \Theta} \|\frac{1}{n} \sum_{i=1}^{n} \mathbf{X}_{iS} h'(\mathbf{X}_{iS}^T \boldsymbol{\beta}_{0S})(Y_i - h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S))\sigma(\mathbf{X}_{iS})^{-2}$$

$$-E\mathbf{X}_S h'(\mathbf{X}_S^T \boldsymbol{\beta}_{0S})(Y - h(\mathbf{X}_S^T \boldsymbol{\beta}_S))\sigma(\mathbf{X}_S)^{-2}\| = o_p(1).$$

The above result follows from the uniform law of large number to $\frac{1}{n} \sum_{i=1}^{n} h(\mathbf{X}_{iS}^T \boldsymbol{\beta}_S)^2 - Eh(\mathbf{X}_S^T \boldsymbol{\beta}_S)^2$, given that $E \sup_{\boldsymbol{\beta}_S \in \Theta} h(\mathbf{X}_S^T \boldsymbol{\beta}_S)^4 < \infty$. The fact that $\sup_{\boldsymbol{\beta}_S \in \Theta} \|\rho(\boldsymbol{\beta}_S)\| < \infty$ follows from repeatedly using Cauchy-Schwarz inequality.

Q.E.D.

Given the foregoing Lemma D.3, Theorem 7.1 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 asysmptotic variance achieves the semiparametric efficiency bound derived by Chamberlain (1987) and Severini and Tripathi (2001). Therefore, $\widehat{\boldsymbol{\beta}}^*$ is semiparametric efficient.

Q.E.D.

References

ANDREWS, D. (1999). Consistent moment selection procedures for generalized method of moments estimation. *Econometrica*, **67** 543-564

ANDREWS, D. and LU, B. (2001). Consistent model and moment selection procedures for GMM estimation with application to dynamic panel data models. *J. Econometrics*, **101** 123-164

ANTONIADIS, A. (1996). Smoothing noisy data with tapered coiflets series. *Scand. J. Stat.*, **23**, 313-330

BELLONI, A. and CHERNOZHUKOV, V. (2011a). Least squares after model selection in high-dimensional sparse models. Forthcoming in *Bernoulli. Manuscript.* MIT.

BELLONI, A. and CHERNOZHUKOV, V. (2011b). l_1-penalized quantile regression in high-dimensional sparse models. *Ann. Statist.*, **39**, 82-130.

BICKEL, P., KLAASSEN, C., RITOV, Y. and WELLNER, J. (1998). *Efficient and adaptive estimation for semiparametric models.* Springer, New York.

BRADIC, J., FAN, J. and WANG, W. (2011). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. *J. R. Stat. Soc. Ser. B*, **73**, 325-349.

BÜHLMANN, P., KALISCH, M. and MAATHUIS, M. (2010). Variable selection in high-dimensional models: partially faithful distributions and the PC-simple algorithm. *Biometrika*, **97**, 261-278

BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications.* Springer, New York.

CANER, M. (2009). Lasso-type GMM estimator. *Econometric Theory,* **25** 270-290

CANER, M. and ZHANG,H. (2009). General estimating equations: model selection and estimation with diverging number of parameters. *Manuscript*, North Carolina State University

CANDES, E. and TAO, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.*, **35** 2313-2404

CHAMBERLAIN, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *J. Econometrics,* **34** 305-334

DAUBECHIES, I., DEFRISE, M. and DE MOL, C. (2004). An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Comm. Pure Appl. Math.*, **57**, 1413-1457.

DONALD, S., IMBENS, G. and NEWEY, W. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *J. Econometrics,***117** 55-93

Donoho, D. L. (2006). Compressed sensing. *IEEE Trans. Inform. Theory* **52**, 1289–1306.

Donoho, D. L. and Elad, E. (2003). Maximal sparsity representation via l_1 Minimization, *Proc. Nat. Aca. Sci.*, **100**, 2197-2202.

ENGLE, R., HENDRY, D. and RICHARD, J. (1983). Exogeneity. *Econometrica.* **51**, 277-304.

FAN, J. and LI, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Amer. Statist. Assoc.*, **96** 1348-1360

FAN, J. and LV, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B*, **70**, 849-911.

FAN, J. and LV, J. (2011). Non-concave penalized likelihood with NP-dimensionality. *IEEE Trans. Inform. Theory*, **57**,5467-5484.

FAN, J. and YAO, Q. (1998). Efficient estimation of conditional variance functions in stochastic regression. *Biometrika*, **85**, 645-660.

FU, W. (1998). Penalized regression: The bridge versus the LASSO. *J. Comput. Graph. Statist.*, **7**, 397-416.

GAUTIER, E. and TSYBAKOV, A. (2011). High dimensional instrumental variables regression and confidence sets. *Manuscript*.

HANSEN, B. (2010). *Econometrics*, Unpublished manuscript. University of Wisconsin.

HANSEN, L. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, **50** 1029-1054

HOROWITZ, J. (1992). A smoothed maximum score estimator for the binary response model. *Econometrica* **60** 505-531

HUANG, J., HOROWITZ, J. and MA, S. (2008). Asymptotic properties of bridge estimators in sparse high-dimensional regression models. *Ann. Statist.* **36** 587-613

KITAMURA, Y., TRIPATHI, G. and AHN, H. (2004). Empirical likelihood-based inference in conditional moment restriction models. *Econometrica*, **72** 1667-1714

LIAO, Z. (2010). Adaptive GMM shrinkage estimation with consistent moment selection. *Manuscript*. Yale University.

LOUNICI, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electron. J. Stat.*, **2**, 90-102.

LV. J. and FAN, Y. (2009). A unified approach to model selection and sparse recovery using regularized least squares. *Ann. Statist.* **37** 3498-3528

NEWEY, W. (1990). Semiparametric efficiency bound *J. Appl. Econometrics*, **5** 99-125

NEWEY, W. (1993). Efficient estimation of models with conditional moment restrictions, in *Handbook of Statistics, Volume 11: Econometrics,* ed. by G. S. Maddala, C. R. Rao, and H. D. Vinod. Amsterdam: North-Holland.

NEWEY, W. and MCFADDEN, D. (1994). Large sample estimation and hypothesis testing, in *Handbook of Econometrics, Chapter 36*, ed. by R. Engle and D. McFadden.

OWEN, A. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, **75**, 237-249.

RASKUTTI, G., WAINWRIGHT, M. and YU, B. (2011). Minimax rates of estimation for high-dimensional linear regression over l_q-balls. *IEEE Trans. Inform. Theory*, **57**,6976-6994.

STÄDLER, N., BÜHLMANN, P. and VAN DE GEER, S. (2010). l1-penalization for mixture regression models (with discussion). *Test*, **19**, 209-256

SEVERINI, T. and TRIPATHI, G. (2001). A simplified approach to computing efficiency bounds in semiparametric models. *J. Econometrics*, **102**, 23-66.

TIBSHIRANI, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B*, **58** 267-288

VERBEEK, M. (2008). *A guide to modern econometrics.* 3rd edition. John Wiley and Sons, England.

WASSERMAN L. and ROEDER, K.(2009). High-dimensional variable selection. *Ann. Statist.*, **37** 2178-2201.

ZHANG, C. (2010). Nearly unbiased variable selection under minimax concave penalty, *Ann. Statist.*, **38** 894-942

ZHANG, C. and HUANG, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear models. *Ann. Statist.*, **36** 1567-1594.

ZHANG, C. and ZHANG, T. (2012). A general theory of concave regularization for high dimensional sparse estimation problems/ *Manuscript*, Rutgers University.

ZHANG, T. (2010). Analysis of multi-stage convex relaxation for sparse regularization. *J. Mach. Learn. Res.*, **11** 1087-1107.

ZHAO, P. and YU, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, **7** 2541-2563

ZOU, H. (2006). The adaptive Lasso and its oracle properties. *J. Amer. Statist. Assoc.*, **101** 1418-1429

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B* **67** 301-320

Zou, H. and Li, R. (2008). One-step sparse estimates in nonconcave penalized likelihood models. *Ann. Statist.*, **36** 1509-1533