Ultra high dimensional variable selection with endogenous covariates

Yuan Liao
Rutgers, The State University of New Jersey

Ultra High Dimensional Variable Selection with Endogenous Covariates

Jianqing Fan Yuan Liao*

Department of Operations Research and Financial Engineering, Princeton University

This Version: July 2011

Abstract

In recent years high dimensional sparse models have gained considerable importance in several areas of economics and finance, which have emerged to deal with many new applications. In an ultra high dimensional sparse model, the number of regressors and candidate moment conditions can be possibly much larger than the sample size (growing in an exponential rate), but only a relatively small number of these regressors are important that interprets the main features of the regression function. The goal is to achieve the oracle property: identifying the important variables with high probability, when both the important and unimportant regressors are possibly endogenous. We derive sufficient conditions and necessary conditions for a general penalized minimization to achieve the oracle property, using a general form of penalty functions. We then show that the penalized GMM and penalized empirical likelihood are consistent in both estimation and selection when (i) the unimportant covariates are endogenous but the important ones are not, or (ii) the important covariates are also possibly endogenous and a set of valid instrumental variables are available. However, the penalized OLS is not. Finally, we develop new results for estimating the optimal instruments in the conditional moment restricted model with the number of instruments growing exponentially fast. This extends Belloni et al (2010) to the possibly nonlinear models as well as more general penalty that allows for SCAD, Lasso, and many other penalty functions.

*Address: Department of ORFE, Sherrerd Hall, Princeton University, Princeton, NJ 08544, USA, e-mail: *jqfan@princeton.edu*, *yuanliao@princeton.edu*. The authors are grateful to the participants at the 2011 Econometric Society Summer Meeting in St. Louis for helpful comments.

Keywords: Penalized GMM, Penalized empirical likelihood, Instrumental Variable, Endogenous variable, Sparsity recovery, Oracle property, Conditional moment restricted model

JEL Classification Numbers: C52, C30

1 Introduction

In recent years high dimensional models have gained considerable importance in several areas of economics and finance, which have emerged to deal with many new applications. In such models the overall number of regressors grows extremely fast with the sample size. For example, in housing price regression, the house price in one county may depend on several other counties, most likely its geographic neighbors. Since the correlation among the neighbor counties is unknown, initially the regression equation may include about one thousand counties in the country. Econometricians, however, may only observe a relatively small size of data series.

Example 1.1 (Housing price panel data)**.** The local housing price has cross-sectional correlation with the housing price in the surrounding counties, and probably in other states. To predict s period ahead housing price appreciation $y_{i,t+s}$ in county i, incorporating lagged variables of additional counties in other states may contribute additional predicting power:

$$y_{i,t+s} = \sum_{k=1}^{p} y_{kt}\beta_{ik} + \mathbf{x}_t^T\theta_i + \epsilon_{i,t+s},$$

where y_{kt} denotes the log-price of county k at period t, and $\mathbf{x}_t$ is a vector of national indices such as the per capita personal income, population, mortgage rate, the stock market index, etc. Since the cross-sectional correlation is unknown, p can be large (around 1000 counties in the US), while the sample size is typically less than two hundred for a data set containing the monthly repeated sales data in ten years (Fan, Lv and Lei (2011)).

Example 1.2 (Wage regression)**.** The expected wage function is approximated by a regression function with regressors being the transformations of education and experience:

$$y_i = \sum_{k=1}^{p} P_k(w_i)\beta_k + \epsilon_i,$$

where the transformation P_k are usually polynomials or B-splines. Empirical evidence shows that low-degree polynomials approximation, with relatively less smoothness flexibility, may fail to capture the information of the entire function provided by the data (Belloni and Chernozhukov (2010)). As the high-degree terms can also have large coefficients, the number of coefficients p can be large relative to the sample size.

Example 1.3 (Instrumental selection)**.** Consider a linear instrumental variable model as in

Angrist and Krueger (1991) and Belloni and Chernozhukov (2011):

$$\begin{aligned} y_i &= \theta_0 + \theta_1 x_i + \mathbf{w}_i^T \gamma + \epsilon_i, \\ x_i &= \mathbf{z}_i^T \beta + \mathbf{w}_i^T \delta + u_i, \end{aligned}$$

with $E(\epsilon_i|\mathbf{w}_i, \mathbf{z}_i) = E(u_i|\mathbf{w}_i, \mathbf{z}_i) = 0$. Here y_i, x_i, and $\mathbf{w}_i$ denote wage, education, and a vector of control variables respectively, and $\mathbf{z}_i$ denotes a vector of instrumental variables that have direct effect on education and indirectly on the wage. The data set in Angrist and Krueger (1991) contains a total of 180 instruments in $\mathbf{z}_i$. It is well known that using only a few instruments results in an estimator of the schooling coefficient θ_1 with a large variance, while using all the 180 instruments results in a large bias. On the other hand, Belloni and Chernozhukov (2011) showed that using just 37 instruments selected by the Lasso technique (Tibshirani 1996) can produce a nearly efficient estimator with a small bias at the same time.

Besides, high dimensional data have also emerged in many other fields of sciences, engineering and humanities. Examples include marketing, microarray data in genomics, signal processing, among others.

We assume that the parameters enter the model as the coefficients of a linear combination of the covariates as in $\mathbf{x}^T\beta_0$, where $\dim(x) = p$ grows with the sample size n. we consider an *ultra high* dimensional pool of covariates, meaning that, $p = O(\exp(n^\alpha))$, 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, which assumes that many components of β_0 are either exactly or near zero. As a result, the true structural parameter can be partitioned as $\beta_0 = (\beta_{0S}^T, \beta_{0N}^T)^T$, with $\beta_{0N} = (\approx)0$. Accordingly, the covariates can be partitioned as $\mathbf{x} = (\mathbf{x}_S^T, \mathbf{x}_N^T)^T$, called *important regressors* and *unimportant regressors*. The sparsity structure typically assumes that, although p, the number of total candidates, is NP-dimensional, the number of important regressors $s = \dim(\mathbf{x}_S)$ grows slowly with the sample size: $s = o(n)$. As in Example 1.1, conditional on $\mathbf{x}_t$, only a small number of counties are useful for predicting the housing price (Fan, Lv and Qi (2010)).

The goal of ultra high dimensional modeling is to (1) achieve the variable selection consistency (identify the nonzero components β_{0S} and therefore the important regressors with probability approaching one), and (2) make inference on β_{0S}. This is called the *oracle property* of high dimensional model selection and estimation (Fan and Li (2001), Zou (2006), and Belloni and Chernozhukov (2009)). While the majority of the literature has achieved this property by minimization a penalized objective function taken the form

$$\text{Loss function} + \text{Penalization},$$

they all assume that all the candidate covariates in $\mathbf{x}$ are exogenous, i.e., uncorrelated with the regression error term. The most popular choices of the loss function in ultra high dimensional modeling are least squares, negative log-likelihood and their various modifications (see, e.g., Bradic, Fan and Wang (2010) and Fan and Lv (2011)).

Has the goal of recovering the oracle been really achieved? Note that the unimportant regressors are artificially added into the model, while oracle knows what the true important regressors are, which does not depend on what kind of unimportant regressors are present. Consequently, a real "mimicking oracle" procedure should be the one under which the oracle property is achieved regardless of whether the unimportant regressors are correlated with the residual or not. Therefore the exogeneity assumption on the unimportant regressors is stronger than that is enough ideally to recover the true sparsity structure, and none of the proposed methods in the literature developed a simultaneous test to check this assumption. Besides, in the application of econometrics, there are also many examples in which the important covariates are also endogenous such as the education in studying the effect of schooling on the wage. We will show in this paper that in the presence of endogenous covariates, the penalized OLS is inconsistent in variable selection.

We consider a more general framework of the ultra high dimensional variable selection problem, and derive both sufficient condition and necessary condition for a penalized minimization procedure to achieve the oracle property, where both the loss function (the leading term of the objective function) and the penalty function can take a very general form. We propose the penalized GMM (PGMM) and penalized empirical likelihood (PEL), and study their asymptotic behaviors. We first consider the case where only the important covariates are required to be exogenous, allowing for arbitrary unimportant covariates. This setting has immediate application interests in finance, biology and machine learnings, and our results significantly contribute to the recent statistical literature on variable selections. We then allow the important covariates to be also endogenous, and construct the GMM objective function using the valid instrumental variables. It will be shown that both PGMM and PEL achieve the oracle properties in the presence of endogeneity. In particular, the estimator converges in probability to β_{0S} at the near *oracle rate* $O_p(\sqrt{(s\log s)/n})$ (Belloni and Chernozhukov (2009), Fan and Lv (2011)). This is achieved because both of the two procedures fully take advantages of the correctly specified moment conditions defined by the exogenous variables, while the penalized OLS does cannot.

In addition, we develop new results for estimating the optimal instruments in nonlinear conditional moment restricted models. It is well known that the optimal instruments involve a conditional expectation, whose functional form is unknown unless strong assumptions on the conditional distribution of the endogenous variables are imposed. The idea of using nonparametric estimates of the optimal instruments was proposed by Newey (1990), where he used a slowly growing sieve approximation. The most important feature of our method is that we consider many more instruments (ultra high dimensional) for estimation, but only a few of the instruments are important while most of others' contributions are negligible. In addition, we let the identities of the important instruments be unknown as in Belloni, Chen, Chernozhukov and Hansen (2010), which substantially generalizes the classical parametric models as well as the slowly-growing sieve models for the optimal instruments.

Recently, there are some related works in the shrinkage GMM literature that allows for endogeneity such as Caner and Zhang (2009) and Liao (2010). They assume the number of covariates and/or the moment conditions are either finite or growing slowly with the sample size, and include all the candidate moment conditions to construct the GMM objective function. However, it is a completely different story in the ultra high dimensional models since the candidate moment conditions (possibly misspecified due to the possible endogeneity) is much more than the observed data. When p increases exponentially fast in n, including all the p moment conditions can lead to inconsistency as p can be much larger than n. It requires us develop a new technique to show the consistency of PGMM and PEL. Therefore our results also constitute essential contributions to the GMM with many moment condition literature. The objective function is designed in a way that uses only a portion of the candidate covariates or instruments, depending on the subspace of the nonzero components of the function argument. This leads our objective function to being discontinuous. We apply a kernel smoothing technique to smooth the objective function for the numerical implementation.

In the literature of model selection, a penalty term attached to an objective function is commonly used. Such a technique can date back at least to the seminal paper of Akaike (1974). In applied statistics, the famous Lasso (Tibshirani (1996)) and its various modifications, e.g., adaptive Lasso (Zou (2006)), bridge estimators (Huang, Horowitz and Ma (2008)), post-Lasso (Belloni and Chernozhukov (2009)), have been widely used for penalization. Fan and Li (2001) pioneered in proposing properties that a penalized estimator should possess, and introduced the *smoothly clipped absolute deviation* (SCAD) penalty that satisfies these properties (See Fan and Li (2001) for details). In addition, a number of researchers proposed various penalty functions for different application problems, for example, hard-thresholding in Antoniadis (1996), minimax concave penalty (MCP) in Zhang (2009), Dantzig selector in

(Candes and Tao (2007)), among others (See Fan and Lv (2010) as an excellent review). In the econometrics literature, Andrews and Lu (2001) used an L_0 penalty (penalizes the *degree of overidentification*) to select the true combination of moment conditions and parameter components. Caner and Zhang (2009) developed penalized GMM method when the parameter is identified by a set of unconditional moment conditions. More recent work is found in Huang, Horowitz and Wei (2010) for selecting the nonparametric additive components, and in Belloni et al (2010) for selecting the instrumental variables using Lasso and Post-Lasso to estimate the optimal instruments, and Belloni and Chernozhukov (2011) for the Lasso quantile regression. Other related works in ultra high dimensional models are done by Fan and Song (2010) and Fan and Lv (2011). It is also worth to mention that, the penalization technique has been aware of for a long history in the Bayesian literature, as the prior distribution plays a natural role of the penalty attached to the log-likelihood.

The remainder of this paper is as follows: Section 2 defines a general class of penalty functions that to be used in this paper, and give model-robust sufficient and necessary conditions for a general penalized optimization procedure to achieve the oracle property. Section 3 and 4 show respectively how penalized GMM and penalized EL are constructed to select the important covariates when there are ultra-high many unimportant ones, potentially endogenous. We also show that in this case, the penalized OLS is inconsistent in variable selection. Section 5 extends the previous results to the generalized sparsity condition, in which the coefficients of the unimportant covariates are not zero but small. It also allows the conditional moment restrictions are subject to local perturbations. Section 6 studies the case when the important covariates are also endogenous, with the help of instrumental variables. We also demonstrate how sparse models can be used to estimate the optimal instruments in nonlinear models. Numerical implementations and simulation results are demonstrated in Sections 7 and 8. Finally, Section 9 concludes.

Throughout the paper, we denote by $\|A\| = \sqrt{tr(AA^T)}$ as the Frobenius norm of a matrix A, $\|\alpha\| = \sqrt{\alpha^T\alpha}$ as the Euclidean norm of a vector α. For two sequences $a_n, b_n \neq 0$, write $a_n \prec b_n$ (equivalently, $b_n \succ a_n$) if $a_n = o(b_n)$. $|\beta|_0$ denotes the number of nonzero components of vector β. For $\beta_S \in \mathbb{R}^s$, let $B(\beta_S, r_n) = \{\beta \in \mathbb{R}^s : \|\beta - \beta_S\| < r_n\}$. 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.1. as brevity for "with probability one".

2 Penalized Optimization

2.1 Penalty function

The penalized optimization takes the general form:

$$\hat{\beta} = \arg\min_{\beta} L_n(\beta) + Pen(\beta),$$

in which L_n is the objective function (or loss function) such as negative log-likelihood, least square, GMM, and EL, and $Pen(\beta)$ serves as the penalty term. It is usually assumed that the true structural parameter β_0 is identified through: $L_n(\beta_0) < L_n(\beta) + o_p(1)$ for any $\beta \in \mathbb{R}^{\dim(\beta_0)}/\{\beta_0\}$ with probability approaching 1. Minimizing L_n directly may not yield a consistent estimator in L_2 norm in high dimensional models. For example, if each component of the L_n minimizer $\hat{\beta}$ is root-n consistent, the overall L_2 distance is $\|\hat{\beta} - \beta_0\| = O_p(\sqrt{p/n})$, which does not converge when $p \succ n$. Therefore, a penalization is necessary.

Over the past decades, many penalty functions have been introduced in high dimensional variable selection problems to serve as the regularization. Some of the most popular penalty functions are: the adaptive Lasso (Zou 2006), elastic net (Zou and Hastie 2005), SCAD (Fan and Li 2001, Fan and Lv 2011), Dantzig selector (Candes and Tao 2007, Fan and Lv 2008), and the weighted l_1- penalty (Bradic, Fan and Wang 2010). Recently, Fan and Lv (2011) 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 $\beta = (\beta_1, ..., \beta_s)^T \in \mathbb{R}^s$, and $|\beta_j| \neq 0, j = 1, ..., s$, define

$$\eta(\beta) = \limsup_{\epsilon \to 0^+} \max_{j \leq s} \sup_{\substack{t_1 < t_2 \\ (t_1, t_2) \in (|\beta_j| - \epsilon, |\beta_j| + \epsilon)}} -\frac{P_n'(t_2) - P_n'(t_1)}{t_2 - t_1}, \tag{2.1}$$

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

Let

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

Assumption 2.1. *$d_n \succ s/\sqrt{n}$, and d_n is bounded away from infinity.*

Let $Pen(\beta) = \sum_{j=1}^{p} P_n(|\beta_j|)$, where $P_n(t)$ is a prespecified penalty function. We now define a class of penalty functions to be used throughout the paper:

Assumption 2.2. *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) $\liminf_{t\to 0^+} P_n'(t) \succ s\sqrt{(\log s)/n}$.
(iv) $P_n'(d_n) = o(s^{-1/2})$
(v) There exists $c > 0$ *such that* $\sup_{\beta\in B(\beta_{0S}, cd_n)} \eta(\beta) = o(1)$.

Note that the concavity of P_n implies that $\eta(\beta) \geq 0$ for all $\beta \in \mathbb{R}^s$. These conditions are needed for establishing the oracle properties of the penalized optimization. They are standard and are satisfied by many commonly used penalty functions. We check these conditions for some popular penalties in the following examples.

Example 2.1 (Lasso (Tibshirani (1986)))**.** Consider

$$P_n(t) = \lambda_n t.$$

We have $P_n'(t) = \lambda_n$ and $P_n''(t) = 0$. Condition (i)(ii) are satisfied naturally. Condition (iii)(iv) hold as long as $s/\sqrt{n} \prec \lambda_n \prec s^{-1/2}$. Finally, for all $\beta \in \mathbb{R}^s$, $\eta(\beta) = 0$. Therefore Lasso satisfies Assumption 2.2 if $s/\sqrt{n} \prec \lambda_n \prec s^{-1/2}$, which also requires $s^3 = o(n)$.

Example 2.2 (l_q Penalty for $q \leq 1$)**.** For $q \in (0, 1]$, and some $\lambda_n > 0$, consider

$$P_n(t) = \lambda_n t^q, t \geq 0.$$

Note that when $q = 1$, P_n corresponds to Lasso. When $q < 1$, $P_n'(t) = \lambda_n q t^{q-1}$ and $P_n''(t) = \lambda_n q(q-1)t^{q-2}$. Condition (i)(ii) are satisfied naturally. For each fixed n, $\liminf_{t\to 0^+} P_n'(t) = \infty$, which implies (iii) as long as $\lambda_n > 0$ for all n. In addition, $P_n'(d_n) = \lambda_n q d_n^{q-1} = o(s^{-1/2})$ if $\lambda_n = o(d_n^{1-q} s^{-1/2})$. Finally, $\eta(\beta_{0S}) = \lambda_n q(1-q)\frac{1}{\min_{j\in A_S} |\beta_{0Sj}|^{2-q}}$.

There exists a neighborhood of β_{0S}, such that $\sup_{\beta\in\mathcal{N}_1} \eta(\beta) \leq \lambda_n q(1-q)d_n/2 = o(1)$ as $\lambda_n \to 0$, which implies (v). Hence when $q < 1$, l_q penalty satisfies Assumption 2.2 if $0 < \lambda_n \prec \min\{d_n^{1-q} s^{-1/2}, d_n^{2-q}\}$.

Example 2.3 (SCAD (Fan and Li 2001))**.** For some $a > 2$, and $\lambda_n > 0$, consider

$$P_n'(t) = \lambda_n \left[I(t \leq \lambda_n) + \frac{(a\lambda_n - t)_+}{(a-1)\lambda_n} I(t > \lambda_n) \right],$$

with $P_n(0) = 0$, $t \geq 0$. We have $P_n''(t) = 0$ when $t > a\lambda_n$. All the conditions can be easily verified as long as $s/\sqrt{n} = o(\lambda_n)$ and $\lambda_n \to 0$.

Example 2.4 (MCP (Zhang 2009))**.** For $a \geq 1$ and $\lambda_n > 0$, consider

$$P_n'(t) = \frac{(a\lambda_n - t)_+}{a}, t \geq 0.$$

with $P_n(0)=0$. We have $P''(t)=0$ when $t>a\lambda_n$. All the conditions are satisfied as long as $s/\sqrt{n}=o(\lambda_n)$ and $\lambda_n\to 0$. In particular, when $a=1$, this is the hard-thresholding penalty introduced by Antoniadis (1996).

2.2 Ultra high dimensional variable selection consistency

In this subsection, we give general consistency results for the ultra high dimensional variable selection and estimation. The following theorems summarize the variable selection consistency theorems in the literature, which provide sufficient conditions for the penalized optimization (GMM, MLE, LS, etc) to have oracle properties in ultra high dimension.

Define $A_S=\{j\in\{1,...,p\}:\beta_{0j}\neq 0\}$, and $\mathcal{B}=\{\beta\in\mathbb{R}^p:\beta_j=0 \text{ if } j\notin A_S\}$. The variable selection aims to recover A_S with high probability. Our first theorem restricts the penalized optimization onto the s-dimensional subspace $\mathcal{B}$. In what follows, denoted by $\lambda_{\min}(A),\lambda_{\max}(A)$ as the smallest and largest eigenvalues of a square matrix A. For any $\beta=(\beta_S^T,0)^T\in\mathcal{B}$, write $L_n(\beta_S,0)=L_n(\beta)$.

Theorem 2.1 (Oracle Consistency)**.** *Suppose $L_n(\beta_S,0)$ is twice differentiable with respect to β_S in a neighborhood of β_{0S} restricted on the subspace $\mathcal{B}$, and there exist a positive sequence $\{a_n\}_{n=1}^{\infty}$ such that $a_n/d_n\to 0$, and a constant $c>0$ such that:*
(i) $\|\partial_{\beta_S}L_n(\beta_{0S},0)\|=O_p(a_n)$,
(ii) $\partial^2_{\beta_S}L_n(\beta_S,0)=\Sigma(\beta_S)+M(\beta_S)$, where $\lambda_{min}(\Sigma(\beta_{0S}))>c$, and $\|M(\beta_{0S})\|<\frac{c}{2}$ with probability approaching 1, and $\Sigma(.),M(.)$ are element-wise continuous on a neighborhood of β_{0S}.
In addition, suppose Assumption 2.1, 2.2 are satisfied. Then there exists a strictly local minimizer $(\hat{\beta}_S^T,0)^T$ of $Q_n(\beta_S,0)=L_n(\beta_S,0)+\sum_{j\in A_S}P_n(|\beta_j|)$ subject to $(\beta_S^T,0)^T\in\mathcal{B}$ such that

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

In penalized optimization estimator, normally the rate of convergence depends on both $\|\partial_{\beta_S}L_n(\beta_{0S},0)\|$ 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 $L_n(\beta_S,0)$ is positive definite at β_{0S}, and also in a neighborhood of β_{0S} because of the continuity of $\Sigma(.)$ and $M(.)$. Both conditions are satisfied by the likelihood-type loss function considered in Fan and Lv (2011) and Bradic, Fan and Wang (2009). It is shown in their papers that in both GLM (generalized linear model) with penalized likelihood and simple linear model with composite quasi-likelihood, $a_n=\sqrt{s/n}$. It will be shown in the subsequent sections that both PGMM and PEL can achieve the near-oracle rate $O_p(\sqrt{(s\log s)/n})$.

The previous theorem assumes that the true support A_S were known, which is not actually. We therefore need to derive the conditions under which A_S can be recovered from the data with probability approaching one. This can be done by demonstrate 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 penalized optimization estimator, defined as the local solution to a penalized optimization problem on $\mathbb{R}^p$. For any $\beta \in \mathbb{R}^p$, define the projection function

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

Theorem 2.2 (Sparsity recovery)**.** *Suppose $L_n : \mathbb{R}^p \to \mathbb{R}$ satisfies the conditions in Theorem 2.1, and Assumptions 2.1 and 2.2 hold. In addition, for $\hat{\beta}_S$ as in Theorem 2.1, $a_n + \sqrt{s}P_n'(d_n) = o(1)$, and suppose there exists a neighborhood $\mathcal{N} \subset \mathbb{R}^p$ of $(\hat{\beta}_S^T, 0)^T$, such that for all $\gamma \in \mathcal{N}$,*

$$L_n(\mathbb{T}\gamma) - L_n(\gamma) \leq \sum_{j=1}^{p} P_n(|\gamma_j|) - \sum_{j=1}^{p} P_n(|(\mathbb{T}\gamma)_j|). \tag{2.2}$$

Then with probability approaching 1, $(\hat{\beta}_S^T, 0)^T$ is a strict local minimizer of $Q_n(\beta) = L_n(\beta) + \sum_{j=1}^{p} P_n(|\beta_j|)$ on $\mathbb{R}^p$.

In particular, if L_n is continuously differentiable in a neighborhood of β_0, then (2.2) holds with probability approaching one, if for all $l \notin A_S$,

$$\left| \frac{\partial L_n(\beta_0)}{\partial \beta_l} \right| = o_p(P_n'(0)).$$

Condition (2.2) ensures that the constrained minimizer of Q_n on $\mathcal{B}$ is also a local minimizer on $\mathbb{R}^p$. This condition is satisfied by the log-likelihood in Fan and Lv (2011) and Bradic, Fan and Wang (2009), and also by GMM and EL criterion functions.

These sufficient conditions for the variable selection and parameter estimation are general enough and do not restrict on any specific model. We will see in the subsequent sections that, with mild regularity conditions on the moments, all the conditions in both Theorem 2.1 and 2.2 are satisfied by PGMM and PEL in conditional moment restricted model. Therefore, while imposing weaker distributional assumptions on the data generating process than other competing methods (such as penalized OLS and penalized likelihood), PGMM and PEL estimators can achieve the oracle property at the oracle convergence rate.

2.3 Necessary condition

While the current literature has been focusing on the sufficient conditions for the penalized estimator to achieve the oracle properties, there is relatively much less attention on the necessary condition for the sparsity recovery in high dimensional problems. Zhao and Yu (2006) derived an *almost necessary* condition for the sign consistency, e.g., the signs of the penalized estimator and the true parameter are equal with probability approaching one. Zou (2006) provided a necessary condition for the variable selection consistency of the OLS estimator with Lasso penalty for slowly-growing p. To the authors' best knowledge, so far there is no necessary condition for the selection consistency of general penalized optimization 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 to a consistent variable selection. We try to answer this question by the following theorem.

Theorem 2.3 (Necessary Condition)**.** *Suppose:*
(i) With probability one, for all n, $L_n(\beta)$ is differentiable at β_0; for all $l \notin A_S$, both $\liminf_{n\to\infty} \frac{\partial L_n(\beta_0)}{\partial \beta_l}$ *and* $\limsup_{n\to\infty} \frac{\partial L_n(\beta_0)}{\partial \beta_l}$ *are continuous at β_0.*
(ii) There is a local minimizer $\hat{\beta} = (\hat{\beta}_S, \hat{\beta}_N)^T$ of $L_n(\beta) + \sum_{j=1}^n P_n(|\beta_j|)$ that recovers the sparsity of β_0, i.e., $\hat{\beta}_N = 0$ with probability approaching one, and $\|\hat{\beta} - \beta_0\| = o_p(1)$.
(iii) The penalty satisfies: $P_n(.) \geq 0$, $P_n(0) = 0$, $P_n'(t)$ is non-increasing when $t \in (0, u)$ for some $u > 0$, and $\lim_n \limsup_{t\to 0^+} P_n'(t) = 0$.
Then for any $l \notin A_S$, with probability one,

$$\liminf_{n\to\infty} \frac{\partial L_n(\beta_0)}{\partial \beta_l} \leq 0 \leq \limsup_{n\to\infty} \frac{\partial L_n(\beta_0)}{\partial \beta_l} \tag{2.3}$$

The last inequality (2.3) requires all the score functions corresponding to the unimportant covariates should be unbiased. Condition (iii) is satisfied by most of the commonly used penalty functions such as the l_1-penalty (Lasso), SCAD and MCP.

Theorem 2.3 will be applied in Section 3.2 to show that in the simple linear model, in the presence of endogenous covariates, the penalized GMM can result to the variable selection consistency, but the penalized OLS cannot.

3 Penalized GMM

3.1 Oracle property

Our model is put under the *conditional moment restriction* framework:

$$E[g(y, \mathbf{x}^T\beta_0)|\mathbf{x}_S] = 0, \tag{3.1}$$

where y stands for the dependent variable, for some k-dimensional known function $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}^k$ with fixed k. Here g can be thought of as a residual function. The conditional moment restricted model was extensively studied by a number of authors: Newey (1993), Donald, Imbens and Newey (2003), Kitamura, Tripathi and Ahn (2004), etc. Since this model covers most of the popular statistical models as special cases, such as generalized linear model and many nonlinear models, it has many important applications in economics, finance, and many other fields. In the simplest case, $g(y, \mathbf{x}^T\beta_0) = y - \mathbf{x}^T\beta_0$ as in the examples above, equation (3.1) therefore represents an exogeneity condition on the residual term, i.e., the regression residual is uncorrelated with the important regressors.

Recently, Caner (2009) considered Lasso-type GMM with the dimension of the structural parameter p fixed, which was later extended to the elastic-net-penalized (Zou and Hastie (2005)) GMM by Caner and Zhang (2009), allowing p to grow with n but $p/n \to 0$. Technically, the oracle property with ultra-high dimension is completely different, because the dimension of $\mathbf{x}_N$ and the number of candidate moment conditions is much larger than the sample size, whose norm, as a result, cannot be bounded even if the support of each component is compact. We will follow a similar technique as in Fan and Lv (2011) and Bradic, Fan and Wang (2009), by first restricting the penalized GMM problem onto the s-dimensional subspace $\mathcal{B}$, and then extend to the entire parameter space.

The conditional moment restriction (3.1) implies that

$$E[g(y, \mathbf{x}^T\beta_0) \otimes \mathbf{x}_S] = 0, \tag{3.2}$$

where $A \otimes B$ denotes the Kronecker product of two matrices. However, this moment condition cannot be used directly to construct the GMM criterion function since the true identities of $\mathbf{x}_S$ is unknown to us, and p, the number of candidate moment conditions formed by $E[g(y, \mathbf{x}^T\beta_0) \otimes \mathbf{x}]$, is much larger than n in the ultra high dimensional setting.

Before formally defining the penalized GMM, let us introduce some additional notation. For any $\beta \in \mathbb{R}^p/\{0\}$, and $i = 1, ..., n$, define a $|\beta|_0$ dimensional vector $\mathbf{X}_i(\beta) = (X_{i,l_1}, ..., X_{i,l_r})^T \in \mathbb{R}^{|\beta|_0}$, where $(\beta_{l1}, ..., \beta_{lr})$ are the nonzero components of β with $r = |\beta|_0$.

The GMM weight matrix is specified as following: Let $\{\sigma_i\}_{i=1}^p$ be a bounded positive sequence that $\min_{i\leq p}\sigma_i > c > 0$ for some $c > 0$. The weight matrix is given by a diagonal matrix $W(\beta) = I_k \otimes$ diag$\{\sigma_{l_1}, ..., \sigma_{l_r}\}$, where I_k denotes the $k \times k$ identity matrix.

Our GMM criterion function is defined as

$$L_{GMM}(\beta) = \left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_i^T\beta) \otimes \mathbf{X}_i(\beta)\right]^T W(\beta) \left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_i^T\beta) \otimes \mathbf{X}_i(\beta)\right]. \tag{3.3}$$

We consider the penalized GMM criterion function:

$$Q_{PGMM}(\beta) = L_{GMM}(\beta) + \sum_{j=1}^p P_n(|\beta_j|). \tag{3.4}$$

Observe that L_{GMM} is not continuous, due to the definition of $\mathbf{X}(\beta)$, and hence to study the large sample property of the PGMM estimator, Taylor's expansion cannot be applied directly. However, the penalized minimization of Q_{PGMM} can be first constrained on $\mathcal{B} = \{\beta \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \notin A_S\}$, and consider $\tilde{L}_{GMM}(\beta_S) = L_{GMM}(\beta_S, 0)$ instead, which is assumed to be twice differentiable. We can then show that if $\hat{\beta}_S$ is a local solution to $\min_{\beta_S} \tilde{L}_{GMM}(\beta_S) + \sum_{j=1}^s P_n(|\beta_j|)$, and that $\|\hat{\beta}_S - \beta_{0S}\| = o_p(1)$, then $(\hat{\beta}_S^T, 0)^T$ is also a local solution to $\min_{\beta\in\mathbb{R}^p} Q_{PGMM}(\beta)$.

In the following assumptions, let $\mathbf{X}_i = (\mathbf{X}_{iS}^T, \mathbf{X}_{iN}^T)^T$ be the partition of $\mathbf{X}_i$, where $\mathbf{X}_{iN}$ and $\mathbf{X}_{iS}$ respectively denote the subvector of $\mathbf{X}_i$ formed by the indices in A_S and A_S^c.

Assumption 3.1. *(i) The true* β_0 *is identified by* $E(g(y, \mathbf{x}^T\beta_0)|\mathbf{x}_S) = 0$.
(ii) $(y_1, X_1), ..., (y_n, X_n)$ *are independent and identically distributed.*

The identification condition (i) ensures that the important covariates are linearly independent with the unimportant covariates. It has the same spirit of the *Strong Irrepresentable Condition* in Zhao and Yu (2006), which rules out not only the existence of the same components in $\mathbf{x}_S$ and $\mathbf{x}_N$, but also the possibility that the components of $\mathbf{x}_S$ can be linearly represented by components of $\mathbf{x}_N$. The i.i.d. assumption can be easily relaxed to stationary ergodic data series with strong mixing structures.

Assumption 3.2. *There exist* $b_1, b_2 > 0$ *and* $r_1, r_2 > 0$ *such that for any* $t > 0$,
(i) $\forall j \leq k$, $P(|g_j(y, \mathbf{x}^T\beta_0)| > t) \leq \exp(-(t/b_1)^{r_1})$,
(ii) $\forall l \in A_s$, $P(|x_l| > t) \leq \exp(-(t/b_2)^{r_2})$.

This assumption requires that both the regression residuals and the important covariates should have exponential tails, which enables us to apply the large deviation theory to show

$\|n^{-1}\sum_{i=1}^n g(y_i, \mathbf{x}_i^T\beta_0) \otimes \mathbf{x}_{iS}\| = O_p(\sqrt{s\log s/n})$. The simplest example that satisfies this assumption is that $g_j(y, \mathbf{x}^T\beta_0)$ is Gaussian and the support of each component in $\mathbf{x}_S$ is bounded.

Assumption 3.3. *(i) $Ex_i^4 < \infty$ for each component of $\mathbf{x}$.*
(ii) $E\|g(y, \mathbf{x}^T\beta_0)\|^4 < \infty$. For each $i = 1, ..., k$, $g_i(t_1, t_2)$ is second order differentiable on $\mathbb{R} \times \mathbb{R}$.
(iii) For each (x, y), $\frac{\partial^2}{\partial\beta\partial\beta^T} g_i(y, \mathbf{x}^T\beta)$ is continuous in β in a neighborhood of β_0.

The continuity of $\frac{\partial^2}{\partial\beta\partial\beta^T} g_i(y, \mathbf{x}^T\beta)$ at β_0 combined with Assumption 3.2(iii) below will be used to show that the Hessian matrix $\partial^2_{\beta_S} L_{GMM}(\beta_S, 0)$ is positive definite in a neighborhood of β_S.

In the following assumptions, let $\mathbf{X}_S = (X_{1S}, ..., X_{nS})^T$, and

$$m(t_1, t_2) = \frac{\partial g(t_1, t_2)}{\partial t_2} = (m_1(t_1, t_2), ..., m_k(t_1, t_2))^T$$

$$q_j(t_1, t_2) = \frac{\partial^2 g_j(t_1, t_2)}{\partial t_2^2}.$$

Assumption 3.4. *(i) $\max_{j\le k}\sup_{t_1,t_2} |m_j(t_1, t_2)| < \infty$,*
(ii) $\max_{j\le k}\sup_{t_1,t_2} |q_j(t_1, t_2)| < \infty$.

This assumption is satisfied by most of the interesting examples in the generalized linear model. For instance,

- 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 $\Phi(.)$ denotes the standard normal cdf.

Assumption 3.5. *There exist $C_1 > 0$ and $C_2 > 0$ such that*
(i) $C_1 < \lambda_{\min}(E\mathbf{x}_S\mathbf{x}_S^T) \le \lambda_{\max}(E\mathbf{x}_S\mathbf{x}_S^T) < C_2$.
(ii) $\min_{j\le k}\lambda_{\min}[(Em_j(y, \mathbf{x}_S^T\beta_{0S})\mathbf{x}_S\mathbf{x}_S^T)^2] > C_1$.
(iii) $\max_{j\le k, l\in A_S}\lambda_{\max}[(Ex_l q_j(y, \mathbf{x}_S^T\beta_{0S})\mathbf{x}_S\mathbf{x}_S^T)^2] < C_2$.

Condition (i) is needed for $\hat{\beta}_S$ to converge at a near oracle rate, i.e., $a_n = O_p(\sqrt{(s\log s)/n})$ for a_n in Theorem 2.1. Condition (ii)(iii) ensure that the Hessian matrix of $L_{GMM}(\beta_S, 0)$ is positive definite at β_{0S}, which implies the familiar information matrix equality to approximately hold $\partial^2 L(\beta_{0S}) = \partial L(\beta_{0S})\partial L(\beta_{0S})^T + o_p(1)$. In particular, Condition (iii) makes the oracle property be achieved when $s^{3/2}\sqrt{(\log s)/n} \to 0$. If Condition (iii) is relaxed, we can achieve the same oracle property when $s^2\sqrt{(\log s)/n}\to 0$. For example, if $g(t_1,t_2)=t_1-t_2$ as in the simple linear model, (ii) is implied by (i); (iii) automatically holds because $q_j(t_1,t_2)=0$. Similar conditions are also assumed in Bradic, Fan and Wang (2010 Condition 4), and Fan and Lv (2011, Condition 4).

Under these conditions, we can show the oracle property of the local minimizer of the penalized GMM (3.3).

Theorem 3.1. *Suppose Assumptions 2.1, 2.2, 3.1-3.4, and 3.5(i) (ii) hold. If either (a) Assumption 3.5(iii) and $s^3\log s=o(n)$, or (b) $s^4\log s=o(n)$ holds, then there exists a strictly local minimizer $\hat\beta=(\hat\beta_S^T,\hat\beta_N^T)^T$ of the penalized GMM $Q_{PGMM}(\beta)$ such that:*
(i)

$$\|\hat\beta_S-\beta_{0S}\|=O_p(\sqrt{(s\log s)/n}+\sqrt{s}P_n'(d_n)),$$

where $\hat\beta_S$ is a subvector of $\hat\beta$ formed by the components whose indices are in A_S, and
(ii) $\hat\beta_N=0$ with probability approaching one as $n\to\infty$.

If we assume $P_n'(d_n)=O(\sqrt{\log s/n})$, then $\|\hat\beta_S-\beta_{0S}\|=O_p(\sqrt{s\log s/n})$, which is very close to the oracle rate $O_p(\sqrt{s/n})$.

The asymptotic normality requires an additional assumption as follows: Define

$$V=\frac{1}{n}\sum_{i=1}^{n}(g(y_i,\mathbf{X}_{iS}^T\beta_{0S})\otimes\mathbf{X}_{iS})(g(y_i,\mathbf{X}_{iS}^T\beta_{0S})\otimes\mathbf{X}_{iS})^T. \tag{3.5}$$

Assumption 3.6. *(i) For some $c>0$, $\lambda_{\min}(V)>c$, with probability one.*
(ii) $P_n'(d_n)=o(1/\sqrt{ns})$.
(iii) There exists $C>0$, $\sup_{\|\beta-\beta_{0S}\|\le C\sqrt{(s\log s)/n}}\eta(\beta)=o((s\log s)^{-1/2})$.

Conditions (ii) and (iii) are satisfied by the penalty functions SCAD, MCP and hard-thresholding. However, they are not satisfied by l_q-penalty ($q\in(0,2)$), or the elastic net (Zou and Hastie (2005)).

Theorem 3.2 (Asymptotic Normality). *Suppose the conditions in Theorem 3.1 and Assumption 3.6 hold, then the penalized GMM estimator in Theorem 3.1 satisfies: for any unit vector $\alpha\in\mathbb{R}^s$, $\|\alpha\|=1$,*

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

where $\Gamma_n=4A_nW(\beta_0)VW(\beta_0)A_n^T$, $\Sigma_n=2A_nW(\beta_0)A_n^T$, and
$A_n=\frac{1}{n}\sum_{i=1}^n(m_1(y_i,\mathbf{X}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{X}_{iS}^T,...,m_k(y_i,\mathbf{X}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{X}_{iS}^T)$.

Remark 3.1. 1. The instrument $\mathbf{X}_i(\beta)$ in the definition of PGMM can be replaced with a cp-dimensional function vector of $f(\mathbf{X}_i, \beta)$ for any fixed integer $c \geq 1$, e.g., $f(\mathbf{X}_i, \beta) = (\mathbf{X}_i(\beta)^T, \mathbf{X}_i^2(\beta)^T)^T$ for $c = 2$. For each fixed β with $|\beta|_0 = r$ nonzero components, there are cr instruments in $f(\mathbf{X}, \beta)$ associated, and hence $cr(> r)$ moment conditions are used to construct the GMM criterion function. In most of the cases, this guarantees that the parameter is over-identified. Roughly speaking, minimizing the GMM criterion function on $\mathbb{R}^r \times \{0\}^{p-r}$ always identifies a unique solution for any $r \leq p$, and due to the over-identification, the minimum would not be close to zero unless is minimized on the exact subspace $\mathbb{R}^s \times \{0\}^{p-s}$ where β_0 lies. Similar results as in Theorem 3.1 can be still obtained. In this case,the true β_0 is the global minimizer of $E[g(y, \mathbf{x}^T\beta) \otimes f(\mathbf{x}, \beta)]^T W(\beta) E[g(y, \mathbf{x}^T\beta) \otimes f(\mathbf{x}, \beta)]$ on $\mathbb{R}^p$ due to the over-identification outside of any small neighborhood of zero.

2. As the GMM criterion function is constructed based on the moment condition $E(g(y, \mathbf{x}^T\beta_0)|\mathbf{x}_S) = 0$, it requires all the important covariates are exogenous. In many econometric applications, the endogeneity also arises from the important covariates. We will show in Section 6 that with valid instrumental variables, the PGMM still achieves the oracle property when both the important and unimportant covariates are possibly endogenous.

3.2 Simple linear model: an example

As an interesting example of application, consider the simple linear model:

$$y = \mathbf{x}^T\beta_0 + \epsilon,$$

where $E(\epsilon|\mathbf{x}_S) = 0$, which implies the moment condition

$$E(y - \mathbf{x}^T\beta_0|\mathbf{x}_S) = 0.$$

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 can contain hundreds or even thousands of variables that may be correlated to wages but are unimportant. These variables are very likelihood to be endogenous.

Consider, for example, a true linear regression model $y = \mathbf{x}_S^T\beta_{0S} + \epsilon$, where $\mathbf{x}_S$ is exogenous. Suppose some components of ϵ are observable, denoted by $\mathbf{x}_N$. The regression error term can then be represented as $\epsilon = f(\mathbf{x}_N, u)$ for some unknown function f, where u represents the unobservable noise component. Then the regression model can be equivalently written as

$$y = \mathbf{x}^T\beta_0 + \epsilon,$$

where $\mathbf{x} = (\mathbf{x}_S, \mathbf{x}_N)$, and $\beta_0 = (\beta_{0S}, 0)$. Apparently, $\mathbf{x}_N$ is correlated with ϵ.

If there exists some component $l \notin A_S$ such that $|E(\epsilon \mathbf{x}_l)|$ is bounded away from zero, the penalized OLS does not achieve the variable selection consistency. The inconsistency is not due to the choice of the penalty (none of the penalty functions in Examples 2.1-2.4 leads to the consistency), but the limitation of the least square loss function. The penalized OLS objective function is defined as:

$$\tilde{Q}_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{X}_i^T\beta)^2 + \sum_{j=1}^{p} P_n(|\beta_j|).$$

It can be shown that the necessary condition in Theorem 2.3 does not hold for the OLS loss function.

Theorem 3.3 (Inconsistency of penalized OLS)**.** *Suppose $\mathbf{x}_N$ has an endogenous component x_l, i.e., $|E(x_{Nl}\epsilon)| > c$ for some $c > 0$. Assume that $Ex_l^4 < \infty$, $E\epsilon^4 < \infty$, and $P_n(t)$ satisfies the conditions in Theorem 2.3. If $\tilde{\beta} = (\tilde{\beta}_S^T, \tilde{\beta}_N^T)^T$ is a local minimizer of $\tilde{Q}_n(\beta)$ corresponding to the coefficients of $(\mathbf{x}_S, \mathbf{x}_N)$, then either $\|\tilde{\beta}_S - \beta_{0S}\| \not\to^p 0$, or*

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

The inconsistency of penalized OLS is due to the fact that the moment condition $E[(y - \mathbf{x}^T\beta_0)x_l] = 0$ is misspecified when x_l is endogenous. We can use the penalized GMM instead. The penalized GMM crierion function is defined as:

$$Q_{PGMM}(\beta) = \left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(\beta)(y_i - \mathbf{X}_i^T\beta)\right)^T\left(\frac{1}{n}\sum_{i=1}^{n}\mathbf{X}_i(\beta)(y_i - \mathbf{X}_i^T\beta)\right) + \sum_{j=1}^{p} P_n'(|\beta_j|), \quad (3.6)$$

where we use an identity weight matrix for simplicity. We can apply the theorem in the previous section to immediately obtain the oracle property of the local minimizer of $Q_{PGMM}(\beta)$, which is stated in the following corolary:

Corollary 3.1. *Consider the simple linear model:*

$$y = \mathbf{x}^T\beta_0 + \epsilon,$$

where $\beta_0 = (\beta_{0S}^T, 0)^T$, $x = (\mathbf{x}_S^T, \mathbf{x}_N^T)^T$, *and* $E(\epsilon|\mathbf{x}_S) = 0$. *Suppose the penalty satisfies Assumption 2.2, and* $s^3 \log s = o(n)$. *In addition, suppose* $Ex_i^4 < \infty$, *for all* $i = 1, ..., p$, $E\epsilon^4 < \infty$, *and there exist* $c_1, c_2 > 0$ *such that* $c_1 < \lambda_{\min}(E\mathbf{x}_S^T\mathbf{x}_S) \leq \lambda_{\max}(E\mathbf{x}_S^T\mathbf{x}_S) \leq c_2$. *Then there exists a strict local minimizer* $\hat{\beta} = (\hat{\beta}_S^T, \hat{\beta}_N^T)^T$ *of (3.6), satisfying:*

1. $\hat{\beta}_N = 0$ *with probability approaching one, and*

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

Note that we only require $\mathbf{x}_S$ to be uncorrelated with the error term, in other words, if some of the components in $\mathbf{x}_N$ are endogenous, i.e., correlated with the error term, the GMM loss function combined with the penalty function that satisfies Assumption 2.2, (SCAD, Lasso, MCP, Hard-thresholding, etc) can achieve the variable selection consistency. This is because it is well known that we can select the correct moment conditions using methods based on GMM (Andrews (1999) and Liao (2010)).

4 Penalized empirical likelihood

4.1 Definition

This section studies the oracle property of the penalized empirical likelihood (PEL), as an alternative to the penalized GMM. The Empirical likelihood in the conditional moment restricted model was studied by Kitamura, et al. (2004), Donald, et al. (2003), and Otsu (2007). In particular, Kitamura, et al. (2004) and Otsu (2007) used a localized empirical likelihood by imposing a Nadaraya-Watson kernel weight to incorporate the conditional moment restrictions. However, when the dimension of $\mathbf{x}_S$ is large, the localization method, because of the curse of dimensionality, is practically difficulty to handel. To just focus on the oracle property of PEL (sparsity recovery and the oracle rate of the estimator), we employ the regular unconditional empirical likelihood, as used in Qin and Lawless (1994). Otsu (2007) added a penalty term to the empirical likelihood to penalize the roughness the estimator when the parameter contains a nonparametric component.

Similar to PGMM, we impose a penalty function that belongs to the same penalty class as before to obtain the PEL. To achieve the oracle property of PEL in the conditional moment restricted model, there is no need to carry out a kernel weighting for localization as in Kitamura et al. (2004) and Otsu (2007), nor do we need to introduce the basis functions of $\mathbf{x}_S$ as in Donald, et al. (2003). Our PEL objective function is defined as:

$$Q_{PEL}(\beta) = L_{EL}(\beta) + \sum_{j=1}^{p} P_n(|\beta_j|), \tag{4.1}$$

where

$$L_{EL}(\beta) = \max_{\lambda \in \mathbb{R}^{k|\beta|_0}} \frac{1}{n} \sum_{i=1}^{n} \log\{1 + \lambda^T [g(y_i, \mathbf{X}_i^T \beta) \otimes \mathbf{X}_i(\beta)]\}, \tag{4.2}$$

and $P_n(t)$ satisfies Assumption 2.2.

Similar to PGMM, L_{EL} is not continuous on $\mathbb{R}^p$. However, we can first constrain the minimization problem of $\min_\beta Q_{PEL}(\beta)$ on $\mathcal{B}$ and apply Talyor's expansion to $\tilde{L}_{EL}(\beta_S) = L_{EL}(\beta_S, 0)$. It will be then shown that a local minimization solution constrained on $\mathcal{B}$ is also a local solution on $\mathbb{R}^p$.

4.2 Oracle property of PEL

We impose the following assumptions in this section. Assumption 4.1 imposes regularity conditions on the moments. We need the eighth moment of $g_j(y, \mathbf{x}^T \beta_0)\mathbf{x}_S$ to be finite to ensure that the score function of $L_{EL}(\beta_{0S})$ is unbiased. In addition, $E\|\partial_{t_2} m(y, \mathbf{x}^T \beta_0)\|^8 < \infty$ is needed for the remaining term of $\tilde{L}_{EL}(\beta_{0S})$ to converge to zero in Frobenius norm.

Assumption 4.1. *(i) $E\|\mathbf{x}_S\|^8 = O(s^4)$, and $Ex_l^4 < \infty$ for all $l \notin A_S$.*
(ii) There exists $B > 0$, such that $E[g_j(y, \mathbf{x}^T \beta_0) x_l]^8 < B$, for all $j = 1, ..., k$, $l \in A_S$.

The next assumption requires some additional notation: For any subset $R = \{r_1, ..., r_q\} \subseteq S$, let $X_{iR} = (X_{i,r_1}, ..., X_{i,r_q})$, $i = 1, ..., n$, which is a subvector of $\mathbf{X}_{iS}$. Define

$$\begin{aligned} \hat{V}_R &= \frac{1}{n} \sum_{i=1}^{n} [g(y_i, \mathbf{X}_i^T \beta_0) g(y_i, \mathbf{X}_i^T \beta_0)^T] \otimes [X_{iR} X_{iR}^T] \\ &= \frac{1}{n} \sum_{i=1}^{n} [g(y_i, \mathbf{X}_i^T \beta_0) \otimes X_{iR}][g(y_i, \mathbf{X}_i^T \beta_0) \otimes X_{iR}]^T. \end{aligned}$$

In particular, write $\hat{V} = \hat{V}_S$.

Assumption 4.2. *(i) For any subset $R \subseteq S$, there exists $c > 0$, such that $\lambda_{\min}(\hat{V}_R) \geq c$, w.p.1.*
(ii) There exists $c_2 > 0$, $\lambda_{\max}(\mathbf{X}_S^T \mathbf{X}_S) \leq c_2 n$, w.p.1.

Theorem 4.1. *Suppose $s^4 = O(n)$. Under the assumptions of Theorem 3.1 and Assumptions 4.1, and 4.2, there exists a strictly local minimizer of $\hat{\beta} = (\hat{\beta}_S^T, \hat{\beta}_N^T)^T$ of the penalized empirical likelihood $Q_{PEL}(\beta)$ such that:*
(i)

$$\|\hat{\beta}_S - \beta_S\| = O_p(\sqrt{\frac{s \log s}{n}} + \sqrt{s} P_n'(d_n)),$$

where $\hat{\beta}_S$ is a subvector of $\hat{\beta}$ formed by the components whose indices are in S, and (ii) $\hat{\beta}_N = 0$ with probability approaching one as $n \to \infty$.

For the asymptotic normality, define

$$\Sigma(\beta_{0S}) = \left(\frac{1}{n}\sum_{i=1}^{n} m(y_i, \mathbf{X}_{iS}^T\beta_{0S})^T \otimes (\mathbf{X}_{iS}\mathbf{X}_{iS})\right)\hat{V}^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}[m(y_i, \mathbf{X}_{iS}^T\beta_{0S})^T \otimes (\mathbf{X}_{iS}\mathbf{X}_{iS})]^T\right).$$

Assumption 4.3. *There exists $c > 0$, such that $\lambda_{\min}(\Sigma(\beta_{0S})) > c$ w.p.1.*

A sufficient condition for this assumption is that all the eigenvalues of $\hat{V}$ are bounded by a fixed constant which does not depend on s, p, n, and the eigenvalues of $\left(\frac{1}{n}\sum_{i=1}^{n} m(y_i, \mathbf{X}_{iS}^T\beta_{0S})^T \otimes (\mathbf{X}_{iS}\mathbf{X}_{iS})\right)\left(\frac{1}{n}\sum_{i=1}^{n}[m(y_i, \mathbf{X}_{iS}^T\beta_{0S})^T \otimes (\mathbf{X}_{iS}\mathbf{X}_{iS})]^T\right)$ are bounded away from zero. For the simple linear model as an example, this is satisfied for all large enough n if the eigenvalues of $E\mathbf{x}_S\mathbf{x}_S^T$ are bounded away from both zero and infinity, $E\epsilon^2 < \infty$, and ϵ and $\mathbf{x}_S$ are independent.

Under the above assumptions, we can show the asymptotic normality of the PEL estimator:

Theorem 4.2. *Under the assumptions of Theorem 4.1, and Assumptions 4.3, the penalized empirical likelihood estimator in Theorem 4.1 satisfies: for any unit vector $\alpha \in \mathbb{R}^s$, $\|\alpha\| = 1$,*

$$\sqrt{n}\alpha^T(A_n\hat{V}^{-1}A_n)^{1/2}(\hat{\beta}_S - \beta_S) \to^d N(0,1).$$

5 Generalized Sparsity and Local Perturbation

So far we have made the assumptions that the important covariates are exogenous, i.e., uncorrelated with the error term, and that the coefficients of the unimportant covariates are exactly zero. These conditions can be relaxed to allow for local perturbations. We allow the conditional moment restrictions to have local perturbations:

$$E[g(y, \mathbf{x}^T\beta_0)|x_s] = O_p(n^{-\alpha}), \text{for some } \alpha > 0. \tag{5.1}$$

In the linear regression model, this means the important covariates can be weakly dependent on the error term. In addition, Suppose β_0 can be partitioned as: $\beta_0 = (\beta_{0S}^T, \beta_{0N}^T)^T$, where β_{0S} and β_{0N} correspond to the "large" and "small" coefficient components in some sense to be defined later. Let $A_S \subset \{1, ..., p\}$ be a subset containing the indices of the large coefficients, and let $A_N \subset \{1, ..., p\}$ contain the indices of the small coefficients, $A_N \cap A_N = \emptyset$. Hence β_{0S} and β_{0N} correspond to the components of β_0 in A_S and A_N. Instead of $\beta_{0N} = 0$ as in the previous sections, the generalized sparsity condition assumes that (Zhang and Huang (2008) and Horowitz and Huang (2010)):

$$\|\beta_{0N}\|_1 < \eta_n,$$

for some $\eta_n \to 0$ under l_1 norm $\|.\|_1$.

Under the generalized sparse condition, the endogenous unimportant covariates arise naturally in the simple linear regression.

Example 5.1. Consider linear regression

$$y = \mathbf{x}_S^T \beta_{0S} + \mathbf{x}_N^T \beta_{0N} + \epsilon.$$

where $\mathbf{x}_S$ and $\mathbf{x}_N$ are uncorrelated with ϵ. Suppose a component of $\mathbf{x}_N$, denoted by x_l, has a small but nonzero coefficient $|\beta_l| = a_n$, a_n decays to zero fast, and $Ex_l = 0$. In addition, there is another component of $\mathbf{x}_N$, denoted by x_k, correlated with x_l. We can put x_1 into the error term, by defining $\tilde{\epsilon} = x_l \beta_l + \epsilon$, and writing

$$y = \mathbf{x}_S^T \beta_{0S} + (\mathbf{x}_N^{-l,-k})^T \beta_{0N}^{-l,-k} + x_k \beta_k + \tilde{\epsilon}, \tag{5.2}$$

where $\mathbf{x}_N^{-l,-k}$ is the vector of unimportant covariates excluding x_l and x_k. Since x_k is correlated with x_l, x_k is endogenous in model (5.2). But the correlation is weak: $E\tilde{\epsilon}x_k = \beta_l E x_l x_k = O_p(a_n)$, assuming $|Ex_l x_k|$ is bounded away from infinity. □

In the linear regression model $g(y, \mathbf{x}^T\beta_0) = y - \mathbf{x}^T\beta_0$, under the assumption that all the components in x are uncorrelated of $y - \mathbf{x}^T\beta_0$, Zhang and Huang (2008) gave conditions under which the Lasso selects exactly the set of nonzero regression coefficients, provided that these coefficients are bounded away from zero at a certain rate. More recently, Horowitz and Huang (2010) showed that the adaptive Lasso distinguishes correctly between large and small coefficients with probability approaching one.

All these results concerning about the generalized sparsity in the literature allow for only the weak endogeneity, i.e., the correlations between the components of x_N and ϵ decay to zero. Essentially, the unimportant covariates considered are still exogenous for large enough n. This section extends the results in the literature to the case when these correlations are bounded away from zero. It also extends the results in the previous sections to models with local perturbation (5.1), which then also allows the important covariates to be weakly correlated with the regression error, without introducing instrumental variables.

Let $|\beta_{(1)}| \leq ... \leq |\beta_{(p)}|$ be the ordered components of β_0 in absolute value. Fix a decaying sequence η_n, and let

$$t = \max\{k : \sum_{j=1}^{k} |\beta_{(j)}| \leq \eta_n\}.$$

Define

$$A_N = \{j \in \{1, ..., p\} : |\beta_{0,j}| \leq |\beta_{(t)}|\}, \quad A_S = \{1, ..., p\} \cap A_N^c.$$

As before, define

$$d_n = \frac{1}{2} \min\{|\beta_{0,j}| : j \in A_S\}, \quad s = \#A_S.$$

We aim to estimate the small coefficients of β_0 to be exactly zero with high probability, under the assumption that $g(y, \mathbf{x}^T\beta_0)$ is weakly correlated with $\mathbf{x}_S$, but allow for endogenous unimportant covariates whose correlation with $g(y, \mathbf{x}^T\beta_0)$ can be bounded away from zero. Theorem 2.3 and Theorem 3.3 can be extended to the generalized sparsity case, which show that in the presence of endogenous unimportant covariates, the penalized OLS is generally not consistent for variable selection. However, we can still apply either PGMM or PEL. For simplicity and brevity, we consider the case when $\dim(g) = 1$, and show the oracle property of PGMM only. Consistency results of PEL like those in Theorem 4.1 can be naturally obtained as well.

As in Section 3, the objective function of PGMM is given by:

$$Q(\beta) = [\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{X}_i^T\beta)\mathbf{X}_i(\beta)]^T W(\beta)[\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{X}_i^T\beta)\mathbf{X}_i(\beta)] + \sum_{j=1}^{p} P_n(|\beta_j|).$$

In addition to the assumptions in Sections 2 and 3, we impose the following conditions on the generalized sparsity, local perturbation, as well as the penalty function:

Assumption 5.1. *(i)* $s^3(\eta_n^2 + n^{-2\alpha}) = o(1)$.
(ii) $\sqrt{s}(\eta_n + n^{-\alpha}) = o(\liminf_{t\to 0^+} P_n'(t))$.

For Lasso, SCAD and MCP, $\liminf_{t\to 0^+} P_n'(t) = O(\lambda_n)$. Hence Condition (ii) puts a restriction of the tuning parameter of the penalty function to depend on the unknown degree of local perturbations $\eta_n + n^{-\alpha}$. Such a condition (that the tuning parameter depends on the unknown model parameters) is not uncommon in the literature, which is often used in sensitivity studies and the regularization literature, as in Hall and Horowitz (2005), Chen and Pouzo (2011).

The effect of the local perturbation and the generalized sparsity condition on the rate of convergence is demonstrated in the following theorem.

Theorem 5.1. *Under Assumption 5.1 and the assumptions of Theorem 3.1, there exists a strictly local minimizer* $\hat{\beta} = (\hat{\beta}_S^T, \hat{\beta}_N^T)^T$ *of* $Q(\beta)$ *such that*

$$\|\hat{\beta}_S - \beta_{0S}\| = O_p(\sqrt{\frac{s \log s}{n}} + \frac{\sqrt{s}}{n^{\alpha}} + \sqrt{s}\eta_n + \sqrt{s}P_n'(d_n)),$$

and

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

6 Endogenous Important Covariates and Selection of Optimal Instruments

6.1 Selection of covariates

In many empirical applications, the important covariates are also endogenous. In this case the moment condition (3.2) is misspecified. However, suppose econometricians observe a set of instrumental variables $\mathbf{w}$ such that

$$E(g(y, \mathbf{x}^T\beta_0)|\mathbf{w}) = 0, \tag{6.1}$$

With the help of the moment condition (6.1), we can also achieve the oracle property of the estimator, allowing the important covariates to be endogenous.

In the presence of possibly endogenous important covariates, recently Caner and Zhang (2009) proposed a penalized GMM procedure for variable selection when p diverges but $p = o(n)$, based on the elastic net of Zou and Zhang (2009). This section extends their results to the ultra high dimensional case with general penalty functions. The extension is not trivial because we allow $p = O(\exp(\alpha n))$ for some $\alpha \in (0, 1)$.

Let $\mathbf{v} = (v_1, ..., v_p)^T = (f_1(\mathbf{w}), ..., f_p(\mathbf{w}))^T$, which can be either a subset of $\mathbf{w}$ if a large set of instrumental variables in $\mathbf{w}$ is available, or a p-dimensional vector of instruments transformed from $\mathbf{w}$ by the basis functions $(f_1, ..., f_p)$, the moment condition (6.1) then implies

$$E(g(y, \mathbf{x}^T\beta_0) \otimes \mathbf{v}) = 0. \tag{6.2}$$

For any $\beta \in \mathbb{R}^p/\{0\}$, let $\mathbf{v}(\beta) = (v_{l1}, ..., v_{lr}) \in \mathbb{R}^r$ be a subset of $\mathbf{v}$ such that $(\beta_{l1}, ..., \beta_{lr})$ are the nonzero components of β with $r = |\beta|_0$. In particular, we denote by $\mathbf{v}_S = \mathbf{v}(\beta_{0S})$. The GMM weight matrix $W(\beta)$ is a diagonal matrix defined similarly as in Section 3.1.

Suppose we have $(y_i, \mathbf{X}_i, \mathbf{V}_i)_{i=1}^n$ as n i.i.d. observations of $(y, \mathbf{x}, \mathbf{v})$. The penalized GMM objective function is constructed based on (6.2) as follows: for a penalty function $P_n(.)$ that belongs to the family described in Section 2,

$$Q_{IV}(\beta)=(\frac{1}{n}\sum_{i=1}^{n}g(y_i,\mathbf{X}_i^T\beta)\otimes\mathbf{V}_i(\beta))^TW(\beta)(\frac{1}{n}\sum_{i=1}^{n}g(y_i,\mathbf{X}_i^T\beta)\otimes\mathbf{V}_i(\beta))+\sum_{j=1}^{p}P_n(|\beta_j|).$$

Technically, the oracle property of the PGMM procedure is attained by an application of Theorem 2.1.

For simplicity, in this section we still assume $\beta_0=(\beta_{0S}^T,0)^T$, where $\dim(\beta_{0S})=s=o(n)$. The results can be also extended to the generalized sparsity condition using the same techniques in Section 5. We impose the following assumptions.

Assumption 6.1. *(i)β_0 is identified by $E(g(y,\mathbf{x}^T\beta_0)|\mathbf{v})=0$.*
(ii) $\max_{i\leq p}Ev_i^2<\infty$, and there exist $b>0$ and $r>0$ such that for any $t>0$,

$$\max_{i\leq p}P(|v_i|>t)\leq\exp(-(t/b)^r).$$

Let

$$\tilde{V}=E(g(y,\mathbf{x}\beta_0)\otimes\mathbf{v})(g(y,\mathbf{x}\beta_0)\otimes\mathbf{v})^T.$$

Assumption 6.2. *There exists $C>0$ such that,*
(i) $\lambda_{\min}(\tilde{V})>C$,
(ii) $\lambda_{\max}((E\mathbf{x}_S\mathbf{v}^T)(E\mathbf{x}_S\mathbf{v}^T)^T)=O(1)$.
(iii) $\min_{j\leq k}\lambda_{\min}(E(m_j(y,\mathbf{x}^T\beta_0)\mathbf{x}_S\mathbf{v}_S^T)E(m_j(y,\mathbf{x}^T\beta_0)\mathbf{x}_S\mathbf{v}_S^T)^T)>C$,
(iv) $\max_{j\leq k,l\in A_S}\lambda_{\max}((Ex_lq_j(y,\mathbf{x}^T\beta_0)\mathbf{x}_S\mathbf{v}_S^T)(Ex_lq_j(y,\mathbf{x}^T\beta_0)\mathbf{x}_S\mathbf{v}_S^T)^T)=o(n/(s^2\log s))$.

These conditions are parallel to those in Assumption 3.5 when $\mathbf{v}$ is used as the instrumental variables. Note that in the linear regression model, (iv) is naturally satisfied as $q_j\equiv 0$.

Assumption 6.3. *(i) For some $C>0$, $\sup_{\|\beta-\beta_0\|\leq C\sqrt{(s\log s)/n}}\eta(\beta)=o((s\log s)^{-1/2})$.*
(ii) $\liminf_{t\to0^+}P_n'(t)\succ s\sqrt{(\log s)/n}$.

In the assumption $\eta(\beta)$ is defined in (2.2). This assumption is satisfied for Lasso as long as $\lambda_n\succ s\sqrt{\log s/n}$. For SCAD and MCP, $P_n''(t)=0$ when $t>a\lambda_n$, and hence the assumption is satisfied if $d_n\succ\lambda_n\succ s\sqrt{\log s/n}$, where d_n denotes the minimal signal $\min\{\beta_{0j}:j\in A_S\}$.

Theorem 6.1. *Under Assumptions 2.1, 3.2(i), 3.3, 3.4, 6.1-6.2, and 6.3(ii), there exists a strictly local minimizer $\hat{\beta}=(\hat{\beta}_S^T,\hat{\beta}_N^T)^T$ of $Q_{IV}(\beta)$ such that*

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

and

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

In addition, if Assumptions 6.3(i) and 3.6(ii) are satisfied, for any $\alpha \in \mathbb{R}^s$, $\|\alpha\| = 1$,

$$\sqrt{n}\alpha^T \tilde{\Gamma}_n^{-1/2} \tilde{\Sigma}_n (\hat{\beta}_S - \beta_{0S}) \to^d N(0,1),$$

where $\tilde{\Gamma}_n = 4\tilde{A}_n W_S \tilde{V} W_S \tilde{A}_n^T$, $\tilde{\Sigma}_n = 2\tilde{A}_n W_S \tilde{A}_n^T$, *and*
$\tilde{A}_n = \frac{1}{n}\sum_{i=1}^n (m_1(y_i, \mathbf{X}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{V}_{iS}^T, ..., m_k(y_i, \mathbf{X}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{V}_{iS}^T)$.

Remark 6.1.

1. The instrumental variable $\mathbf{v}$ can be made of *cp*-dimensional for any fixed integer $c \geq 1$, when a larger set of transformed instruments are included. Then for each fixed β with $|\beta|_0 = r$ nonzero components, there are cr instruments in $\mathbf{v}(\beta)$ associated. In most of the cases, this guarantees that the parameter is over-identified. Roughly speaking, minimizing the GMM criterion function on $\mathbb{R}^r \times \{0\}^{p-r}$ always identifies a unique solution for any $r \leq p$, and due to the over-identification, the minimum would not be close to zero unless is minimized on the exact subspace $\mathbb{R}^s \times \{0\}^{p-s}$ where β_0 lies. Similar results as in Theorem 6.1 can be still obtained.

2. An alternative GMM criterion function is constructed using all the candidate instruments (see, e.g., Liao (2011), and Caner and Zhang (2009)). In order for the GMM criterion function to identify a unique minimizer, we need the number of used instruments to be at least as many as the dimension of the parameter. When p grows much faster than n as in the ultra high dimensional variable selection problem, however, this will lead to the inconsistency of GMM. In contrast, our approach uses only a subset $\mathbf{v}(\beta)$ that depends on the support of each fixed argument, which avoids the inconsistency introduced by using too many moment conditions, and at the same time, always guarantees the identification of the solution on each subspace $\mathbb{R}^r \times \{0\}^{p-r}$.

3. We can also conduct the variable selection using penalized empirical likelihood as in Section 4, where the EL objective function is given by

$$L_{EL}(\beta) = \max_{\lambda \in \mathbb{R}^{k|\beta|_0}} \frac{1}{n}\sum_{i=1}^n \log\{1 + \lambda^T[g(y_i, \mathbf{X}_i^T\beta) \otimes \mathbf{V}_i(\beta)]\}$$

   Similar conditions as those in Section 4 can be derived to achieve the oracle properties of the PEL procedure.

6.2 Selection of the optimal instruments

When β_{0S} is of fixed dimension, we can obtain the semiparametric efficient estimator of β_{0S} in two steps. In the first step, apply the PGMM procedure as described above to select the important covariates $\mathbf{x}_S$, and obtain a consistent initial estimator $\hat{\beta}_S$. In the second step, apply GMM with the estimated optimal weight matrix and instrument using $\hat{\beta}_S$. In the linear regression model when Lasso is chosen as the penalty function, the two-step procedure described above is called *post-Lasso* in Belloni et al (2010).

After the important covariates are selected in step one, by Theorem 6.1, with probability approaching one, we identify the following model

$$E(\rho(Z, \beta_{0S})|\mathbf{w}) = 0, \tag{6.3}$$

where $\rho(Z, \beta_{0S}) = g(y, \mathbf{x}_S^T \beta_{0S})$, and obtain a consistent estimator $\hat{\beta}_S$. Since we have identified the support of important covariates A_S with high probability, we treat it as a known set. Hence in the following, we do not distinguish the notation $l \leq s$ from $l \in A_S$.

Suppose $\rho(Z, .)$ is continuously differentiable in β. It is well known that the optimal instrument that leads to the semiparametric efficient estimation of β_{0S} is given by $A(\mathbf{w}) = D(\mathbf{w})^T \Omega(\mathbf{w})^{-1}$ (see, e.g., Chamberlain (1987), Newey (1993)), where

$$D(\mathbf{w}) = E(\frac{\partial \rho(\beta_{0S})}{\partial \beta_S}|\mathbf{w}), \Omega(\mathbf{w}) = E(\rho(Z, \beta_{0S})\rho(Z, \beta_{0S})'|\mathbf{w}).$$

Our goal is to estimate $D(\mathbf{w})$ and $\Omega(\mathbf{w})$ nonparametrically in the presence of many instrumental variables.

For simplicity, we consider only the single moment condition case $\dim(\rho) = 1$, and restrict ourselves to the homoskedastic case where $\Omega(\mathbf{w}) = \Omega$ is a constant matrix independent of $\mathbf{w}$, which can be consistently estimated by

$$\hat{\Omega} = \frac{1}{n} \sum_{i=1}^{n} g(y_i, \hat{\mathbf{X}}_{iS}^T \hat{\beta}_S) g(y_i, \hat{\mathbf{X}}_{iS}^T \hat{\beta}_S)^T.$$

Here $\hat{\mathbf{X}}_{iS}$ is the vector of covariates selected by the penalized GMM described previously.

Suppose there is a very large list of technical instruments $\mathbf{v} = (v_1, ..., v_{p_1})^T$. To avoid introducing redundant notation, we set $p_1 = p$, and $\mathbf{v}$ can be thought of as the instruments used to select the important covariates in Section 6.1. In this subsection, we aim to provide a consistent instrumental selection procedure to estimate $D(\mathbf{w})$ in the presence of many instruments, which can be even ultra high dimensional. Recently Belloni et al (2010) and

Belloni et al (2011) proposed a Lasso procedure to estimate $D(\mathbf{w})$ in linear models, and derived an asymptotic theory for the resulting IV estimators. However, to our best knowledge, there has not been a formal study of the ultra high dimensional instrument selection problem for the possibly nonlinear models as well as the effect of using more general (and possibly more design-adaptive) penalty functions. Some of the alternative penalties allowed are more design-adaptive than Lasso, as illustrated in Fan and Li (2001), and Antoniadis and Fan (2001). In addition, the effect of the generalized sparsity condition under which the optimal instruments are allowed to weakly depend on many unimportant technical instrumental variables is still unclear yet.

Our method is based on a key assumption on the generalized sparse model of instrumental variables

Assumption 6.4. *(i)There exists an $s \times p$ matrix $\Theta_0 = (\theta_{01}, ..., \theta_{0s})^T$, a vector function $a(\mathbf{w})$, and a nonnegative sequence $c_n \to 0$ such that*

$$D(\mathbf{w}) = \Theta_0 \mathbf{v} + a(\mathbf{w}), \qquad \max_{l \leq s} (\frac{1}{n} \sum_{i=1}^{n} a_l(\boldsymbol{W}_i)^2) = O_p(c_n^2). \tag{6.4}$$

(ii) There exist $\alpha_1 \in (\frac{1}{2}, \infty]$ and $\alpha_2 \in (0, \frac{1}{2})$ such that, for each $l \leq s$, we have partition $\{1, ..., p\} = T_l \cap T_l^c$ with

$$\max_{l \leq s} \sum_{i \notin T_l} |\theta_{0l,i}| < n^{-\alpha_1}, \quad \min_{l \leq s, i \in T_l} |\theta_{0l,i}| = h_n > n^{-\alpha_2},$$
$$\max_{l \leq s} \#\{i : i \in T_l\} = s_1 = o(n). \tag{6.5}$$

When $\mathbf{v}$ is a vector of functions of $\mathbf{w}$, it can be taken as a large number of series terms with respect to $\mathbf{w}$ such as B-splines, dummies, polynomials, and various interactions. Then Condition (i) is simply the nonparametric sieve approximation assumption as in Newey (1990) and Belloni et al (2011). Condition (ii) states that for each p-dimensional sieve coefficient θ_{0l}, only a few number of its components are "big", whose indices are collected in the important set T_l. The rest components are comparatively much smaller, and satisfy the generalized l_1 sparsity condition $\sum_{i \notin T_l} |\theta_{0l,i}| < n^{-\alpha_1}$. Roughly speaking, it allows the sieve approximation of each component of $D(\mathbf{w})$ to be a function of only a few important instruments and many unimportant instruments, and lets the identities of the important instruments (whose indices support is T_l) be unknown. Hence we substantially generalize the classical parametric model of optimal instruments.

For each l, let $\theta_{0l,S} = (\theta_{0l,j} : j \in T_l)$, and $\theta_{0l,N} = (\theta_{0l,j} : j \notin T_l)$, corresponding to the subvectors of large and small coefficients in θ_{0l}. The sieve coefficient is then partitioned into

$\theta_{0l} = (\theta_{0l,S}^T, \theta_{0l,N}^T)^T$. Accordingly, define $\mathbf{v}_{lS} = (v_i : i \in T_l)$ as the subvector of $\mathbf{v}$ consisting of the instruments that are important to $D_l(\mathbf{w})$. Our goal is to identify $\theta_{0l,S}$ from $\theta_{0l,N}$ with high probability, and consistently estimate $D_l(\mathbf{w})$ using only the important instruments for each l.

Consider the penalized least square criterion function:

$$Q_l(\theta) = \frac{1}{n}\sum_{i=1}^{n}(\frac{\partial\rho(Z_i, \hat{\beta}_S)}{\partial\beta_{Sl}} - \mathbf{V}_i^T\theta)^2 + \sum_{j=1}^{p} P_n(|\theta_j|). \tag{6.6}$$

The penalized least square estimator of Θ_0 as well as the optimal instrument $D(\mathbf{w})$ is defined as:

$$\hat{\theta}_l = \arg\min_{\theta} Q_l(\theta),$$

$$\hat{D}(\mathbf{w}) = \hat{\Theta}\mathbf{v}, \quad \hat{\Theta} = (\hat{\theta}_1, ..., \hat{\theta}_s)^T.$$

We impose the following conditions:

Assumption 6.5. *(i)* $\|\hat{\beta}_S - \beta_{0S}\| = O_p(\sqrt{(s\log s)/n})$.
(ii) There exists $M > 0$ *such that for any* $\beta_1, \beta_2 \in \mathbb{R}^s$ *and each* z,

$$\max_{l\leq s}|\frac{\partial\rho(z, \beta_1)}{\partial\beta_{Sl}} - \frac{\partial\rho(z, \beta_2)}{\partial\beta_{Sl}}| \leq M\|\beta_1 - \beta_2\|.$$

Condition (i) requires the initial estimator of β_{0S} be consistent, and this is guaranteed by Theorem 6.1 in the last subsection. In addition, if $O(P_n'(d_n))$ is dominated by $O(\sqrt{\log s/n})$, the rate of convergence then simplifies to $O_p(\sqrt{(s\log s)/n})$. We only consider the case when the model structural function is differentiable, whose derivative is Lipschitz continuous.

Assumption 6.6. *(i)* $C_1 < \min_{l\leq s}\lambda_{\min}(E\mathbf{v}_{lS}\mathbf{v}_{lS}^T) < \max_{l\leq s}\lambda_{\max}(E\mathbf{v}_{lS}\mathbf{v}_{lS}^T) < C_2$, *and* $\max_{l\leq p} Ev_l^2 < C_2$ *for some* $C_1 > 0$, $C_2 > 0$.
(ii) $\max\{c_n, \sqrt{\log p/n}\} = o(P_n'(0^+))$.
(iii) For each l, *let* $e_l = \partial_{\beta_l}\rho(\boldsymbol{Z}, \beta_{0S}) - D_l(\boldsymbol{W})$. *Then* $\max_{l\leq s} P(|e_l| > t) \leq \exp(-(t/b)^r)$ *for some* $b > 0$ *and* $r > 0$.

Condition (i) states that $E\mathbf{v}_{lS}\mathbf{v}_{lS}^T$ should be well-behaved. While it is a standard assumption in the sieve approximation literature that the population Gram matrix of the fourier basis functions has eigenvalues bounded from above and below (e.g., Newey (1997)), our condition here requires only a small proportion of the transformed IV's satisfy this assumption. Sufficient conditions for (i) can be found, for example, in Belloni et al (2010). Condition (ii) places the regularity condition on the penalty function as before, and Condition (iii) requires an exponential tail for $\partial_{\beta_l}\rho(\mathbf{Z}, \beta_{0S}) - D_l(\mathbf{W})$.

Let $\hat{\theta}_l = (\hat{\theta}_{lS}, \hat{\theta}_{lS})$ be the partition corresponding to the positions of $\theta_{0l,S}$ and $\theta_{0l,N}$. Therefore, $\hat{\theta}_{lS} = \{\hat{\theta}_{lj} : j \in T_l\}$, and $\hat{\theta}_{lN} = \{\hat{\theta}_{lj} : j \in T_l^c\}$. The following theorem derives the asymptotic properties of the penalized least square estimation procedure.

Theorem 6.2. *Under the assumptions of Theorem 6.1 and Assumptions 6.4-6.6, for each l, $Q_l(\theta_l)$ has a strictly local minimizer $\hat{\theta}_l = (\hat{\theta}_{lS}, \hat{\theta}_{lS})$, such that*

$$\|\hat{\theta}_{lS} - \theta_{0l,S}\| = O_p(\sqrt{\frac{s \log s}{n}} + \sqrt{\frac{s_1 \log s_1}{n}} + \sqrt{s_1} n^{-\alpha_1} + \sqrt{s_1} c_n + \sqrt{s_1} P_n'(h_n)).$$

$$\lim_{n\to\infty} P(\hat{\theta}_{lN} = 0) = 1.$$

In addition,

$$\frac{1}{n}\sum_{i=1}^{n} |\hat{D}_l(\boldsymbol{W}_i) - D_l(\boldsymbol{W}_i)|^2 = O_p(\frac{s_1 s \log s_1}{n} + \frac{s_1^2 \log s_1}{n} + s_1^2 n^{-2\alpha_1} + s_1^2 c_n^2 + s_1^2 P_n'(h_n)^2).$$

With the estimate $\hat{D}(\mathbf{w})$ and $\hat{\Omega}$, it is straightforward to construct the optimal GMM and obtain the semiparametric efficient estimator of β_{0S}.

7 Implementation

In this section we discuss the implementation for numerically minimizing the penalized objective function in PGMM and PEL.

7.1 Smoothed PGMM and smoothed PEL

The GMM objective function (3.3) is given by

$$\begin{aligned} L_{GMM}(\beta) &= \left[\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{X}_i^T\beta) \otimes \mathbf{X}_i^{\beta}\right]^T W(\beta) \left[\frac{1}{n}\sum_{i=1}^{n} g(y_i, \mathbf{X}_i^T\beta) \otimes \mathbf{X}_i^{\beta}\right] \\ &= \sum_{j=1}^{p} w_j \left[\frac{1}{n}\sum_{i=1}^{n} g(y_i, x_i^T\beta) x_{ij} I(\beta_j \neq 0)\right]^T \left[\frac{1}{n}\sum_{i=1}^{n} g(y_i, x_i^T\beta) x_{ij} I(\beta_j \neq 0)\right] \end{aligned}$$

where $I(\beta_j \neq 0)$ is the indicator function. Note that for each fixed subset $\mathcal{S} \subset \{1, ..., p\}$, this objective function is continuous in β on $\{\beta \in \mathbb{R}^p : \beta_j = 0 \text{ if } j \in \mathcal{S}\}$, but is not continuous in β globally on $\mathbb{R}^p$. As there are 2^p subsets like $\mathcal{S}$, minimizing $Q_{GMM}(\beta) = L_{GMM}(\beta)$+Penalty is generally NP-hard, i.e., there are no polynomial time algorithms to solve the problem[1]. The same discontinuity problem also applies to the PEL objective function.

We overcome this discontinuity problem by applying the "smoothing" technique as in Horowitz (1992), which approximates the indicator function by a continuous smooth function $K : [0, \infty) \to \mathbb{R}$ such that:
(i) $0 \leq K(t) < M$ for some finite M and all $t \geq 0$.
(ii) $K(0) = 0$ and $\lim_{t\to\infty} K(t) = 1$.
(iii) $\lim_{t\to\infty} K'(t)t = 0$, 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 continuous cumulative distribution function. For a pre-determined small number h_n, L_{GMM} is approximated by a continuous function in β:

$$L_K(\beta) = \sum_{j=1}^{p} w_j \left[\frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i^T \beta) x_{ij} K(\frac{\beta_j^2}{h_n}) \right]^T \left[\frac{1}{n} \sum_{i=1}^{n} g(y_i, x_i^T \beta) x_{ij} K(\frac{\beta_j^2}{h_n}) \right].$$

The objective function of PEL can be approximated in a similar manner. Note that

$$\begin{aligned} L_{EL}(\beta) &= \max_{\lambda} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \lambda^T (g(y_i, \mathbf{X}_i^T \beta) \otimes \mathbf{X}_i(\beta))) \\ &= \max_{\lambda_j \in \mathbb{R}^k, j=1,\ldots,p} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \sum_{j=1}^{p} \lambda_j^T g(y_i, \mathbf{X}_i^T \beta) x_{ij} I(\beta_j \neq 0)), \end{aligned}$$

which can be replaced with

$$L_K(\beta) = \max_{\lambda_j \in \mathbb{R}^k, j=1,\ldots,p} \frac{1}{n} \sum_{i=1}^{n} \log(1 + \sum_{j=1}^{p} \lambda_j^T g(y_i, \mathbf{X}_i^T \beta) x_{ij} K(\frac{\beta_j^2}{h_n})).$$

The smoothed version of the objective function based on the instrumental variables in Section 6.1 is straightforward.

Remark 7.1. If $h_n = o(\min\{\beta_{0j} : j \in A_S\})$, it can be shown that minimizing either the smoothed PGMM or the smoothed PEL also leads to the oracle property, and the results in Theorems 3.1, 3.1, 4.1, 4.2 and 6.1 still hold. This can be done by directly checking the sufficient conditions derived in Section 2. The detailed proof is omitted here, and is available from the authors.

[1] In a special case when $g(y, \mathbf{x}^T\beta) = y - \mathbf{x}^T\beta$, and $\mathbf{x}_S$ is independent of $\mathbf{x}_N$, the problem can be solved in polynomial time. It can be shown that, with the identification condition: $\forall \epsilon > 0, \exists \delta > 0$ such that $\inf_{\|\beta_{0S}-\beta_S\|>\epsilon, \beta=(\beta_S^T,\beta_N^T)^T \neq 0} L_{GMM}(\beta_S, \beta_N) > \delta$, minimizing $L_{GMM} + SCAD$ can be carried out by a backward elimination procedure.

7.2 Second order approximation

As remarked in Section 3 and Section 6, the instrument in the definition of PGMM can be replaced with a cp-dimensional function vector of $f(\mathbf{X}_i, \beta)$ (or $f(\mathbf{V}_i, \beta)$ if the IV is available). In most of the cases, this guarantees that the parameter is over-identified. Roughly speaking, minimizing the GMM criterion function on $\mathbb{R}^r \times \{0\}^{p-r}$ always identifies a unique solution for any $r \leq p$, and due to the over-identification, the minimum would not be close to zero unless is minimized on the exact subspace $\mathbb{R}^s \times \{0\}^{p-s}$ where β_0 lies. In this case,the true β_0 is the global minimizer of $E[g(y, \mathbf{x}^T\beta) \otimes f(\mathbf{x}, \beta)]^T W(\beta) E[g(y, \mathbf{x}^T\beta) \otimes f(\mathbf{x}, \beta)]$ on $\mathbb{R}^p$ due to the over-identification outside of any small neighborhood of zero.

We employ the iterative coordinate algorithm (Fan and Lv (2011)): minimize one coordinate of β at a time with fixed other coordinates obtained from previous steps and successive replacements. The penalty function is approximated by local linear approximation as in Zou and Li (2008). Specifically, suppose we have obtained $\beta^{(l)}$ at step l. For $k \in \{1, ..., p\}$, denote by $\beta^{(l)}_{(-k)}$ as a $(p-1)$-dimensional vector consisting of all the components of $\beta^{(l)}$ but $\beta^{(l)}_k$. Write $(\beta^{(l)}_{(-k)}, t)$ as the p-dimensional vector that replaces the kth component of $\beta^{(l)}$ with t. Optimization (7.1) is a univariate minimization problem, which can be carried out by golden section search. To speed up the convergence, we can also use the second order approximation of $L_K(\beta^{(l)}_{(-k)}, t)$ along the kth component:

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

We solve for

$$\beta^{(l+1)}_k = \arg\min_t \hat{L}_K(\beta^{(l)}_{(-k)}, t) + P'_n(|\beta^{(l)}_k|)|t|. \tag{7.1}$$

For the remaining component at this step, let $\beta^{(l+1)}_{(-k)} = \beta^{(l)}_{(-k)}$. We accept $\beta^{(l+1)}_k$ as the updated kth component of $\beta^{(l+1)}$ only if $L_K(\beta^{(l+1)}_{(-k)}, t) + P'_n(|\beta^{(l+1)}_k|)|t|$ strictly decreases. Update $k \to k+1$, $l \to l+1$.

When the second order approximation is combined with SCAD, the local linear approximation of SCAD is not needed. As demonstrated in the appendix of Fan and Lv (2011), 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.

8 Monte Carlo Experiments

8.1 Design 1

To test our proposed method for variable selection, we simulate a simple linear model:

$$y = \mathbf{x}^T \beta_0 + \epsilon, \quad \epsilon \sim N(0, 1).$$

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

For simplicity, the distribution of the error term is set to be homoskedastic. The p-dimensional vector of covariates x is generated from the following process:

$$z = (z_1, ..., z_p)^T \sim N_p(0, \Sigma), \quad (\Sigma)_{ij} = 0.5^{|i-j|},$$

$$z \text{ is independent of } \epsilon,$$

$$(x_1, ..., x_5) = (z_1, ..., z_5), \quad x_j = (z_j + 5)(\epsilon + 1), \text{ for } 6 \leq j \leq p.$$

The unimportant covariates are correlated with both important covariates and the error term.

The data contains $n = 200$ i.i.d. copies of (y, x). Penalized OLS and PGMM are carried out separately for comparison. The simulation results in Fan and Lv (2011) favored SCAD over Lasso in high dimensional variable selection when all the candidate covariates are exogenous. Hence in our simulation we use SCAD with pre-determined tuning parameters of λ as the penalty function, and look at its behavior when endogenous covariates are present.

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

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

After the minimization procedure, a coefficient β_j is selected if $|\beta_j| > 10^{-4}$. There are 100 replications per experiment. Three performance measures are used to compare the methods. The first measure is the mean squared error (MSE) of the important covariates, determined by the average of $\|\hat{\beta}_S - \beta_{0S}\|$, where $A_S = \{1, ..., 5\}$. The second measure is the number of correctly selected non-zero coefficients, i.e., the true positive (TP), and the third measure is the number of incorrectly selected coefficients, i,e., the false positive (FP). In addition, the standard error over the 100 replications of each measure is also reported. In each simulation, we initiate $\beta^{(0)} = (0, ..., 0)^T$, and run a penalized OLS for $\lambda = 0.01$ to obtain the initial value for the penalized GMM procedure. The results of the simulation are summarized in Table

1-3, which compare the performance measures of penalized OLS and PGMM for three values of p.

Table 1: Performance Measures of POLS and PGMM when $p = 15$

	POLS				PGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE-Mean	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)
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)

POLS 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 covariates. For $\lambda_n = 1$, the median of the number of selected nonzero parameters is only 4. In contrast, PGMM performs quite well in selecting the important covariates, and in correctly eliminating the unimportant covariates. Note that the average MSE of PGMM is only slightly larger than that of POLS when $\lambda = 0.05$ and 0.1. However, it has error-free selection of the important covariates, and almost no false positives. Note that $\lambda = 0.4$ is a large tuning parameter that results to some incorrectly eliminated important covariates, and a larger MSE.

Table 2: Performance Measures of POLS and PGMM when $p = 50$

	POLS				PGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE-Mean	0.145	0.133	0.629	1.417	0.261	0.176	0.204	0.979
	(0.053)	(0.043)	(0.301)	(0.329)	(0.094)	(0.069)	(0.069)	(0.245)
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	0.03	0.02	0.14
Median	38	35	8	2	0	0	0	0
	(2.902)	(3.045)	(3.334)	(1.557)	(0.337)	(0.171)	(0.141)	(0.569)

Table 3: Performance Measures of POLS and PGMM when $p = 300$

	POLS				PGMM			
	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.5$	$\lambda = 1$	$\lambda = 0.05$	$\lambda = 0.1$	$\lambda = 0.2$	$\lambda = 0.4$
MSE-Mean	0.186	0.159	0.650	1.430	0.274	0.187	0.187	1.009
	(0.073)	(0.054)	(0.304)	(0.310)	(0.086)	(0.102)	(0.068)	(0.276)
TP-Mean	5	5	4.82	3.62	5	5	5	4.45
Median	5	5	5	4	5	5	5	4
	(0)	(0)	(0.384)	(0.487)	(0)	(0)	(0)	(0.557)
FP-Mean	227.96	210.47	42.78	7.94	0.11	0	0	0.05
Median	227	211	42	7	0	0	0	0
	(10.767)	(11.38)	(11.773)	(5.635)	(0.37)	(0)	(0)	(0.330)

8.2 Design 2

Consider the model

$$\begin{aligned} y &= \mathbf{x}^T\beta_0 + \epsilon, \\ \mathbf{x} &= \mathbf{v} + \mathbf{u}, \end{aligned}$$

where $(\beta_{01}, \beta_{02}, \beta_{03}, \beta_{04}, \beta_{05}) = (5, -4, 7, -1, 1.5), \beta_{0j} = 0$, for $6 \leq j \leq p$. Here $\mathbf{x}$ is a $p \times 1$ vector of covariates, and $\mathbf{v}$ is a $p \times 1$ vector of instrumental variables, generated independently from $N(0, 1)$, also independently of ϵ. The error terms $(\epsilon, \mathbf{u})$ are generated from $N_{p+1}(0, \Sigma)$, where $\Sigma = (0.95^{|i-j|})_{(p+1)\times(p+1)}$. All the components in $\mathbf{x}$ are endogenous.

The simulations are carried out for $p = 10, 50$ and 300 three levels. One hundred replications are conducted for each p, with $n = 200$ observations generated each time. We still use SCAD as the penalty function. In each simulation, we initiate $\beta^{(0)} = (0, ..., 0)^T$, and run a penalized OLS for $\lambda = 0.01$ to obtain the initial value for the penalized GMM procedure. The results are summarized in the following table for different choices of the tuning parameters of SCAD in the PGMM step.

The performance of the estimators is quite consistent for the three levels of p. The magnitude of the penalty $\lambda = 0.01$ is relatively small so that there are a small number of false positives. However, the penalty $\lambda = 1$ is a bit too much which results to a nonzero coefficient to be falsely eliminated. Also, the minimal nonzero signal $|\beta_4| = 1$ turns out to be large enough so that the penalized GMM can do a perfect job in identifying all the nonzero and zero coefficients under some appropriate penalty level, i.e., $\lambda = 0.1$ for $p = 10$ and $\lambda = 0.5$ for $p = 50$ and 300.

Table 4: Performance Measures of Penalized GMM

	$p = 10$			$p = 50$			$p = 300$		
	$\lambda = 0.01$	$\lambda = 0.1$	$\lambda = 1$	$\lambda = 0.01$	$\lambda = 0.3$	$\lambda = 1$	$\lambda = 0.01$	$\lambda = 0.3$	$\lambda = 1$
MSE-Mean	0.111	0.290	0.896	0.104	0.113	0.825	0.111	0.156	0.873
	(0.039)	(0.117)	(0.244)	(0.037)	(0.039)	(0.205)	(0.040)	(0.116)	(0.219)
TP-Mean	5	5	4.82	5	5	4.88	5	5	4.77
Median	5	5	5	5	5	5	5	5	5
	(0)	(0)	(0.412)	(0)	(0)	(0.356)	(0)	(0)	(0.423)
FP-Mean	0.43	0	0	1.150	0	0	1.663	0	0
Median	0	0	0	1	0	0	1	0	0
	(0.624)	(0)	(0)	(1.067)	(0)	(0)	(2.081)	(0)	(0)

To study the sensitivity of our procedure to the minimal nonzero signals, we run another set of simulations in which we change $\beta_4 = -0.5$ and keep all the remaining parameters the same as before. The minimal nonzero signal becomes $|\beta_4| = 0.5$, and we run for $p = 20$. Table 5 indicates that the minimal signal is too small so that it is not as easily distinguishable from the zero coefficients as before.

Table 5: Performance Measures of Penalized GMM when $p = 20$, $\beta_4 = -0.5$

λ	0.001	0.005	0.01	0.05	0.1	0.5
MSE-Mean	0.112	0.136	0.137	0.156	0.142	0.433
	(0.090)	(0.117)	(0.102)	(0.117)	(0.083)	(0.158)
TP-Mean	4.96	4.92	4.94	4.910	4.960	4.250
Median	5	5	5	5	5	4
	(0.197)	(0.273)	(0.239)	(0.288)	(0.197)	(0.435)
FP-Mean	11.28	3.88	1.135	0.020	0	0
Median	11	3	1	1	0	0
	(1.545)	(2.447)	(2.139)	(0.141)	(0)	(0)

9 Conclusion

We consider the ultra high dimensional variable selection problem in which the number of regressors grows exponentially fast with the sample size. The true parameter is assumed to be sparse in the sense that many components are exactly zero. 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 conditional moment restricted model, which covers a board range of statistical models in application.

An interesting finding is that, when there exists an endogenous variable whose true regression coefficient is zero, the penalized OLS does not satisfy the necessary condition of variable selection regardless of the penalty selected from a large family of penalty functions. We then propose two alternative solutions to the above inconsistency problem, by either penalized GMM or penalized EL. It is shown that both of the procedures possess the oracle property asymptotically.

The oracle property can be also achieved when the important covariates are also potentially endogenous, with the help of instrumental variables. In addition, in the presence of many instruments (possibly ultra-high dimensional), the optimal instrument is estimated by a sparse model, where the only a few instruments are important. We allow for the generalized sparsity condition on the nonparametric sieve approximation to the optimal instrument, and derive the oracle properties.

A Proofs for Section 2

Throughout the Appendix, C will denote a generic positive constant that may be different in different uses, and $|\alpha|$ will denote the "absolute value" of a vector α taken coordinately.

A.1 Proof of Theorem 2.1

Lemma A.1. *Under Assumptions 2.1 and 2.2, if $\beta \in \mathbb{R}^s$ and $\|\beta - \beta_{0S}\| = o(1)$, then for all large n,*

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

Proof. By Taylor's expansion, there exists β^* lies on the line segment joining β and β_{0S}, $\sum_{j=1}^{s}(P_n(|\beta_j|) - P_n(|\beta_{0S,j}|) = (P_n'(|\beta_1^*|), ..., P_n'(|\beta_s^*|))(\beta - \beta_{0S}) \leq \|\beta - \beta_{0S}\|\sqrt{s}\max_{j\leq s} P_n'(|\beta_j^*|)$. If $\|\beta - \beta_{0S}\| = o(1)$, then $\max_{j\leq s}|\beta_j^* - \beta_{0S,j}| = o(1)$. Hence for all large n, $\min\{|\beta_j^*| : j \leq s\} > d_n$. 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 $\sum_{j=1}^{s}(P_n(|\beta_j|) - P_n(|\beta_{0S,j}|) \leq \|\beta - \beta_{0S}\|\sqrt{s}P_n'(d_n)$. Q.E.D.

Proof of Theorem 2.1

The proof is a generalization of the proof of Theorem 3 in Fan and Lv (2011). Let $k_n = a_n + \sqrt{s}P_n'(d_n)$, and write $Q_1(\beta_S) = Q_n(\beta_S, 0)$, and $L_1(\beta_S) = L_n(\beta_S, 0)$. Then $\partial^j L_1(\beta_S) = \partial_{\beta_S}^j L_n(\beta_S, 0)$, for $j = 1, 2$. Define $\mathcal{N}_a = \{\beta \in \mathbb{R}^s : \|\beta - \beta_{0S}\| \leq k_n a\}$ for some $a > \frac{8}{c} > 0$ where c is such that $\lambda_{\min}(\Sigma(\beta_{0S})) > c$. Let $\partial\mathcal{N}_a$ denotes the boundary of $\mathcal{N}_a$. If $Q_1(\beta_{0S}) < \min_{\beta_S\in\partial\mathcal{N}_a} Q_1(\beta_S)$, then by the continuity of Q_1, there exists a local minimizer of Q_1 inside $\mathcal{N}_a$. Equivalently, there exists a local minimizer of Q_n restricted on $\mathcal{B}$ inside $\{\beta = (\beta_S^T, 0)^T : \beta_S \in \mathcal{N}_a\}$. Hence it suffices to show that $P(Q_1(\beta_{0S}) < \min_{\beta_S\in\partial\mathcal{N}_a} Q_1(\beta_S)) \to^p 1$, and that the local minimizer is strict.

For any $\beta_S \in \partial\mathcal{N}_a$, there exists β^* lying on the segment joining β_S and β_{0S} such that by the Taylor's expansion on L_1:

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

By Condition (i), $(\beta_S - \beta_{0S})^T\partial L_1(\beta_{0S}) \geq -\|\beta_S - \beta_{0S}\|a_n$ w.p.a.1. In addition, Condition (ii) yields $(\beta_S - \beta_{0S})^T\Sigma(\beta_{0S})(\beta_S - \beta_{0S}) > c\|\beta_S - \beta_{0S}\|^2$, and $|(\beta_S - \beta_{0S})^T M(\beta_{0S})(\beta_S - \beta_{0S})| \leq \|\beta_S - \beta_{0S}\|^2\frac{c}{2}$. Hence by the continuity of Σ and M, and that $\|\beta_S - \beta_{0S}\| \to 0$, $(\beta_S - \beta_{0S})^T\partial^2 L_1(\beta^*)(\beta_S - \beta_{0S}) > \frac{c}{2}\|\beta_S - \beta_{0S}\|^2$. By Lemma A.1, $\sum_{j=1}^{s}[P_n(|\beta_{Sj}|) - P_n(|\beta_{0S,j}|)] \geq -\sqrt{s}P_n'(d_n)\|\beta_S - \beta_{0S}\|$. Hence w.p.a.1,

$$\min_{\beta\in\partial\mathcal{N}_a} Q_1(\beta) - Q_1(\beta_{0S}) \geq k_n a(\frac{c}{4}k_n a - a_n - \sqrt{s}P_n'(d_n)) > k_n a(2k_n - a_n - \sqrt{s}P_n'(d_n)) \geq 0$$

It remains to show that the local minimizer in $\mathcal{N}_a$ (denoted by $\hat{\beta}_S$) is strict. For each $h \in \mathbb{R}/\{0\}$, define $\tau(h) = \limsup_{\epsilon\to 0^+}\sup_{\substack{t_1<t_2\\(t_1,t_2)\in(|h|-\epsilon,|h|+\epsilon)}} -\frac{P_n'(t_2)-P_n'(t_1)}{t_2-t_1}$. By the concavity of P_n, $\tau \geq 0$. For $\beta_S \in \mathcal{N}_a$, we know that L_1 is twice differentiable. Let $A(\beta_S) = \partial^2 L_1(\beta_S) - diag\{\tau(\beta_{S1}), ..., \tau(\beta_{Ss})\}$. Since $\|\hat{\beta}_S - \beta_{0S}\| = o_p(1)$, by Condition (ii), for any nonzero $\alpha \in \mathbb{R}^s$,

$$\alpha^T A(\hat{\beta}_S)\alpha \geq \frac{c}{2}\alpha^T\alpha - \alpha^T\alpha\max_{j\leq s}\tau(\hat{\beta}_{Sj})$$

By Assumption 2.2, $\max_{j\leq s}\tau(\hat{\beta}_{Sj})\leq\sup_{\beta\in\mathcal{N}_1}\eta(\beta)=o_p(1)$. Therefore $A(\hat{\beta}_S)$ is positive definite w.p.a.1. Q.E.D.

A.2 Proof of Theorem 2.2

By Theorem 2.1, there exists a neighborhood $\mathcal{N}_2\subset\mathcal{N}$, such that, for any $\gamma\in\mathcal{N}_2$, we can write $\mathbb{T}\gamma=(\gamma_S^T,0)$. Note that $Q_n(\mathbb{T}\gamma)\geq Q_n(\hat{\beta})$, where $\hat{\beta}=(\hat{\beta}_S,0)^T$. Thus it suffices to show that $Q_n(\mathbb{T}\gamma)\leq Q_n(\gamma)$. In fact, $Q_n(\mathbb{T}\gamma)-Q_n(\gamma)=L_n(\mathbb{T}\gamma)-L_n(\gamma)-(\sum_{j=1}^p P_n(\gamma_j)-\sum_{j=1}^s P_n(|(\mathbb{T}\gamma)_j|))\leq 0$, by Condition (2.2).

If L_n is continuously differentiable in a neighborhood of β_0, by the mean value theorem, there exists $\lambda>0$ such that for $h=\lambda\gamma+(1-\lambda)\mathbb{T}\gamma$,

$$Q_n(\mathbb{T}\gamma)-Q(\gamma)=\sum_{l\notin A_S}\frac{\partial L_n(h)}{\partial\beta_l}(-\gamma_l)-\sum_{l\notin A_S}P_n'(|h_l|)|\gamma_l|\leq\sum_{l\notin A_S}\left(\left|\frac{\partial L_n(h)}{\partial\beta_l}\right|-P_n'(|h_l|)\right)\gamma_l|.$$

It thus suffices to show $|\frac{\partial L_n(h)}{\partial\beta_l}|\leq P_n'(|h_l|)$ for each $l\notin A_S$. By assumption, $|\frac{\partial L_n(\beta_0)}{\partial\beta_l}|=o_p(P_n'(0))$, and $\beta_{0N}=0$. Therefore, there exists $\delta>0$ such that if $\|\beta-\beta_0\|<\delta$, $|\frac{\partial L_n(\beta)}{\partial\beta_l}|<P_n'(|\beta_l|)$. We know that w.p.a.1, $\|\hat{\beta}_S-\beta_{0S}\|<\delta/2$, thus as long as $\|\gamma-(\hat{\beta}_{0S}^T,0)^T\|<\delta/2$ almost surely, the triangular inequality then implies $\|h-\beta_0\|<\delta$ w.p.a.1, which gives the desired result.

Q.E.D.

A.3 Proof of Theorem 2.3

Proof. Write $P_n'(0^+)=\limsup_{t\to0^+}P_n'(t)$. Suppose the necessary condition does not hold, then there exists $l\notin A_S$, such that either one of the two cases holds with probability bounded away from zero:
(i) $\liminf_{n\to\infty}\frac{\partial L_n(\beta_0)}{\partial\beta_l}>P'(0)$, or
(ii) $\limsup_{n\to\infty}\frac{\partial L_n(\beta_0)}{\partial\beta_l}<-P'(0)$.

We show that both cases lead to contradiction:

Case (i): By the continuity of $\liminf\partial_{\beta_l}L_n(.)$, there exists a convex neighborhood U of β_0 such that for all $t\in U$, $\liminf_{n\to\infty}\frac{\partial L_n(t)}{\partial\beta_l}>P'(0)$. Let $r=-c$ for some $c>0$. Define $\beta=(\hat{\beta}_S^T,\beta_2^T)^T$, where $\beta_2=(0,...,0,r,0,...0)^T$, with r on the lth position of β. Since $\hat{\beta}=(\hat{\beta}_S^T,0)^T$ is a local minimizer, there exists a neighborhood $\mathcal{N}$ of $\hat{\beta}$, such that when $c>0$ is small enough, $\beta\in\mathcal{N}$, and $Q_n(\hat{\beta})\leq Q_n(\beta)$, which is $L_n(\hat{\beta})-L_n(\beta)\leq P_n(|r|)$. As $\|\hat{\beta}-\beta_0\|=o_p(1)$, $\mathcal{N}$ can be made small enough such that both β and $\hat{\beta}$ are inside U ($\hat{\beta}$ converges to be inside the interior of U). Applying the mean value theorem to both sides of $L_n(\hat{\beta})-L_n(\beta)\leq P_n(|r|)$ yields

$$c\partial_{\beta_l}L_n(h)=-r\partial_{\beta_l}L_n(h)\leq P_n'(|u|)|r|\leq P_n'(0^+)c$$

for some $|u|<|r|$, and h lying on the segment joining $\hat{\beta}$ and β. The last inequality follows from the fact that P_n' is nonincreasing.

By the convexity of U, $h\in U$. These arguments imply that the following event occurs with probability bounded away from zero:

$$P'(0)<\liminf_{n\to\infty}\frac{\partial L_n(h)}{\partial\beta_l}\leq\lim_n P_n'(0^+)$$

which is a contradiction.

Case (ii): Let $r = c > 0$. Define the same β as in case (i). For small enough c, $L_n(\hat{\beta}) - L_n(\beta) \leq P_n(|r|)$. In addition, $-\limsup_{n\to\infty} \frac{\partial L_n(t)}{\partial \beta_l} > P'(0)$ for all t in some convex neighborhood U of β_0. On the other hand, by mean value theorem and the fact that P_n' is decreasing, $-r\partial_{\beta_l} L_n(h) \leq P_n'(0^+)c$, which is $-c\partial_{\beta_l} L_n(h) \leq P_n'(0^+)c$, for some h described as before. Hence with probability bounded away from zero,

$$-\limsup_{n\to\infty} \frac{\partial L_n(h)}{\partial \beta_l} \leq \lim_n P_n'(0^+)$$

By the same argument, h can be made inside U, which implies a contradiction. Q.E.D.

B Proofs for Section 3

B.1 Proof of Theorem 3.1

Lemma B.1. *Suppose $(A_1, ..., A_k), (W_1, ..., W_k)$ are $2k$ $a \times a$ matrices. Let $\lambda_1 = \max\{|\lambda_{ij}| : i = 1, ..., k, j = 1, ..., a\}$, and $\lambda_2 = \max\{|\lambda_{ij}^*| : i = 1, ..., k, j = 1, ..., a\}$, where $|\lambda_{ij}|$, $|\lambda_{ij}^*|$ denote the jth eigenvalues of A_i and W_i. Let B_1, B_2 be matrices so that the products in the follows are defined, $A = (A_1, ..., A_k)$and $W = diag\{W_1, ..., W_k\}$. Then*
(i) $\|AB_1\|^2 \leq k\lambda_1^2 \|B_1\|^2$,
(ii) $\|WB_2\|^2 \leq \lambda_2^2 \|B_2\|^2$.

Proof. (i) Write $B_1 = (M_1^T, ..., M_k^T)^T$, then $AB_1 = \sum_{i=1}^k A_i M_i$. By Cauchy-Schwarz inequality, $\|AB_1\| \leq \sum_{i=1}^k \|A_i M_i\| \leq |\lambda_1| \sum_{i=1}^k \|M_i\| \leq |\lambda_1|\sqrt{k}(\sum_{i=1}^k \|M_i\|^2)^{1/2} = |\lambda_1|\sqrt{k}\|B_1\|$.
(ii) Write $B_2 = (H_1^T, ..., H_k^T)^T$, then $\|WB_2\|^2 = \sum_{i=1}^k \|W_i H_i\|^2 \leq \lambda_2^2 \sum_{i=1}^k \|H_i\|^2 = \lambda_2^2 \|B_2\|^2$. Q.E.D.

Theorem 3.1: Consistency

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

$$\tilde{L}_{GMM}(\beta_S) = \left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T \beta_S) \otimes \mathbf{X}_{iS}\right]^T W(\beta_0) \left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T \beta_S) \otimes \mathbf{X}_{iS}\right]$$

Then $\tilde{L}_{GMM}(\beta_S) = L_{GMM}(\beta_S, 0)$. We proceed by verifying the conditions in Theorem 2.1.
Condition (i): $\partial \tilde{L}_{GMM}(\beta_{0S}) = 2A_n(\beta_{0S})W(\beta_0)\left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T \beta_{0S}) \otimes \mathbf{X}_{iS}\right]$, where

$$\begin{aligned} A_n(\beta) &\equiv \frac{1}{n}\sum_{i=1}^n m(y_i, \mathbf{X}_i^T \beta)^T \otimes (\mathbf{X}_{iS}\mathbf{X}_{iS}^T) \\ &= \frac{1}{n}\sum_{i=1}^n (m_1(y_i, \mathbf{X}_i^T \beta)\mathbf{X}_{iS}\mathbf{X}_{iS}^T, ..., m_k(y_i, \mathbf{X}_i^T \beta)\mathbf{X}_{iS}\mathbf{X}_{iS}^T). \end{aligned} \quad \text{(B.1)}$$

By Assumptions 3.4 and 3.5(i), the absolute values of the eigenvalues of $\{\frac{1}{n}\sum_{i=1}^n m_j(y_i, \mathbf{X}_i^T \beta)\mathbf{X}_{iS}\mathbf{X}_{iS}^T\}_{j=1}^k$ are uniformly bounded across $j = 1, ..., k$ by a constant $C > 0$ with probability approaching one. In addition, the elements in $W(\beta_0)$ are bounded. Hence by Lemma B.1,

$$\|\partial \tilde{L}_{GMM}(\beta_{0S})\| \leq O_p(1)\|\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T \beta_{0S}) \otimes \mathbf{X}_{iS}\|.$$

Using the Bernstein inequality with Assumption 3.2, it can be shown that

$$\max_{l\in A_S, j\leq k}|\frac{1}{n}\sum_{i=1}^n g_j(y_i, \mathbf{X}_{iS}^T\beta_{0S})x_{li}| = O_p(\sqrt{\frac{\log s}{n}}).$$

Hence $\|\partial\tilde{L}_{GMM}(\beta_{0S})\| = O_p(\sqrt{(s\log s)/n})$.

Condition (ii) Straightforward but tedious calculation yields $\partial^2\tilde{L}_{GMM}(\beta_{0S}) = \Sigma(\beta_{0S}) + M(\beta_{0S})$, where $\Sigma(\beta_{0S}) = 2A_n(\beta_{0S})W(\beta_{0S})A_n(\beta_{0S})^T$, and $M(\beta_{0S}) = 2\sum_{j=1}^k B_j(\beta_{0S})H_j(\beta_{0S})$, with (suppose $\mathbf{X}_{iS} = (x_{il_1}, ..., x_{il_s})^T$)

$$\begin{aligned}
H_j(\beta_{0S})_{s^2\times s} &= I_s\otimes\left[W(\beta_{0S})\frac{1}{n}\sum_{i=1}^n g_j(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS}\right],\\
B_j(\beta_{0S}) &= \frac{1}{n}\sum_{i=1}^n(x_{il_1}q_j(y_i, \mathbf{x}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{X}_{iS}^T, ..., x_{il_s}q_j(y_i, \mathbf{x}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{X}_{iS}^T).
\end{aligned}$$

It is not hard to obtain $\|M(\beta_{0S})\| = O_p(s^2\sqrt{(\log s)/n})$. Given Assumption 3.5(iii), we can achieve a sharper bound of $\|M(\beta_{0S})\|$ as following: By Lemma B.1 and Assumption 3.5 (iii),

$$\|B_j(\beta_{0S})H_j(\beta_{0S})\|^2 \leq O_p(s)\|H_j(\beta_{0S})\|^2 \leq C_2^2 s\|H_j(\beta_{0S})\|^2 = O_p(\frac{s^3\log s}{n}).$$

Thus $\|M(\beta_{0S})\| \leq \sum_{j=1}^k\|B_j(\beta_{0S})H_j(\beta_{0S})\| = O_p(\sqrt{s^3\log s}/n^{1/2})$. By Theorem 2.1, we have

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

Theorem 3.1: Sparsity: To show the sparsity, we check (2.2) in Theorem 2.2.
For some neighborhood $\mathcal{N}$ of $(\hat{\beta}_S^T, 0)^T$, and $\forall\gamma\in\mathcal{N}$, write $\gamma = (\gamma_S^T, \gamma_N^T)^T$, $\mathbb{T}\gamma = (\gamma_S^T, 0)^T$. For all $\theta\in\mathbb{R}^p$, define

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

Hence $L_{GMM}(\mathbb{T}(\gamma)) = F(\mathbb{T}\gamma)$. One can then check that $L_{GMM}(\gamma) = F(\gamma) + \xi_2(\gamma)$, where

$$\xi_2(\gamma) = \sum_{j=1}^k(\frac{1}{n}\sum_{i=1}^n g_j(y_i, \mathbf{X}_i^T\gamma)\mathbf{X}_i(\gamma_N))^T W^{\gamma_N}(\frac{1}{n}\sum_{i=1}^n g_j(y_i, \mathbf{X}_i^T\gamma)\mathbf{X}_i(\gamma_N)) \geq 0,$$

and $W^{\gamma_N} = diag\{\sigma_l : l\in A_N\}$. Hence $L_{GMM}(\mathbb{T}\gamma) - L_{GMM}(\gamma) \leq F(\mathbb{T}\gamma) - F(\gamma)$.

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

$$\begin{aligned}
F(\mathbb{T}\gamma) - F(\gamma) &= -\sum_{l\notin A_S, \gamma_l\neq 0}\gamma_{Nl}\left[\frac{1}{n}\sum_{i=1}^n\frac{\partial}{\partial\beta_l}g(y_i, \mathbf{X}_i^T h)\otimes\mathbf{X}_i(\gamma_S)\right]^T W(\gamma_S)\left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_i^T h)\otimes\mathbf{X}_i(\gamma_S)\right]\\
&\equiv \sum_{l\notin A_S, \gamma_l\neq 0}\gamma_{Nl}a_l(h).
\end{aligned}$$

By the mean value theorem, there exists $\lambda_2 \in (0,1)$,

$$\sum_{j=1}^{p}(P_n(|\gamma_j|) - P_n(|(\mathbb{T}\gamma)_j|)) = \sum_{l \notin A_S, \gamma_l \neq 0} |\gamma_l| P_n'(\lambda_2|\gamma_l|).$$

Hence it suffices to show that for each $l \notin A_S$, and $\gamma_l \neq 0$,

$$|\gamma_l a_l(h)| \leq |\gamma_l| P_n'(\lambda_2|\gamma_l|). \tag{B.2}$$

Since $E(g(y, \mathbf{x}^T\beta_0)|\mathbf{x}_S) = 0$, by Assumptions 2.1, 3.3, $|a_l(\beta_0)| \leq C\sqrt{s}\sqrt{(s\log s)/n} = O_p(s\log s/\sqrt{n}) \prec \liminf_{t\to 0^+} P_n'(t)$. By the continuity of a_l, $|a_l(\hat{\beta}^T, 0)| < \liminf_{t\to 0^+} P_n'(t)$ with probability approaching 1. Note that $h \in \mathcal{N}$. For small enough $\mathcal{N}$, again by continuity, $|a_l(h)| < \frac{1}{2}\liminf_{t\to 0^+} P_n'(t)$ w.p.a.1. Hence $|a_l(h)| < \frac{1}{2}P_n'(\lambda_2|\gamma_l|)$, which yields (D.1). Q.E.D.

B.2 Proof of Theorem 3.2

Let $P_n'(|\hat{\beta}_S|) = (P_n'(|\hat{\beta}_{S1}|), ..., P_n'(|\hat{\beta}_{Ss}|))^T$. The asymptotic normality builds on the following lemma.

Lemma B.2. *Let $Q_1(\beta_1)$, $L_1(\beta_1)$ and $\hat{\beta}_S$ satisfy the conditions in Theorem 2.1. Suppose there exists an $s \times s$ matrix Ω_n, such that:*
(i) For any unit vector $\alpha \in \mathbb{R}^s$, $\|\alpha\| = 1$,

$$\alpha^T \Omega_n \partial L_1(\beta_{0S}) \to^d N(0,1)$$

(ii) $\|\Omega_n[P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S)]\| = o_p(1)$.
Then for any unit vector $\alpha \in \mathbb{R}^s$,

$$\alpha^T \Omega_n \Sigma_n(\beta_{0S})(\hat{\beta}_S - \beta_{0S}) \to^d N(0,1).$$

Proof. The KKT condition of $\hat{\beta}_S$ is given by

$$-P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S) = \partial L_1(\hat{\beta}_S)$$

where $\circ$ denotes the Hadamard product of two vectors. By the mean value theorem, there exists β^* lying on the segment joining β_{0S} and $\hat{\beta}_S$ such that $\partial L_1(\hat{\beta}_S) = \partial L_1(\beta_{0S}) + (\Sigma(\beta^*) + M(\beta^*))(\hat{\beta}_S - \beta_{0S})$. Since $\|\hat{\beta}_S - \beta_{0S}\| = o_p(1)$, $\|\beta_* - \beta_{0S}\| = o_p(1)$. By the continuity of $\Sigma(.)$ and $M(.)$, we have $\Sigma(\beta^*) + M(\beta^*) = \Sigma(\beta_{0S}) + M(\beta_{0S}) = \Sigma(\beta_{0S}) + o_p(1)$. Therefore,

$$(\Sigma(\beta_{0S}) + o_p(1))(\hat{\beta}_S - \beta_{0S}) = -P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S) - \partial L_1(\beta_{0S}). \tag{B.3}$$

For any unit vector $\alpha \in \mathbb{R}^s$, by Condition (ii), $\|\alpha^T \Omega_n[P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S)]\| = o_p(1)$. Hence the result follows immediately from B.3 and Condition (i). Q.E.D.

Lemma B.3. *Under Assumption 2.1, 2.2, for $a_n, \hat{\beta}_S$ defined in Theorem 2.1,*

$$\|P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S)\| = O_p(\max_{\beta \in \mathcal{N}_1} \eta(\beta) a_n + \sqrt{s} P_n'(d_n)),$$

where $\mathcal{N}_1 = \{\beta \in \mathbb{R}^s : \|\beta - \beta_{0S}\| \leq C\sqrt{(s \log s)/n}\}$, *for* $C > 0$ *in Assumption 3.6.*

Proof. Write $P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S) = (v_1, ..., v_s)^T$, with $v_i = P_n'(|\hat{\beta}_{Si}|)sgn(\hat{\beta}_{Si})$. $|v_i| \leq |P_n'(|\hat{\beta}_{Si}|) - P_n'(|\beta_{Si}|)| + P_n'(|\beta_{Si}|) \leq \max_{\beta \in \mathcal{N}_1} \eta(\beta)|\hat{\beta}_{Si} - \beta_{Si}| + P_n'(d_n)$. By Minkowski's inequality, $\|P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S)\| \leq \sqrt{\sum_{i=1}^s \max_{\mathcal{N}_1} \eta(\beta)^2 |\hat{\beta}_{Si} - \beta_{Si}|^2} + \sqrt{s}P_n'(d_n) = \max_{\beta \in \mathcal{N}_1} \eta(\beta)\|\hat{\beta}_S - \beta_{0S}\| + \sqrt{s}P_n'(d_n) = O_p(\max_{\beta \in \mathcal{N}_1} \eta(\beta)(a_n + \sqrt{s}P_n'(d_n)) + \sqrt{s}P_n'(d_n))$. Q.E.D.

Lemma B.4. *Let* $\Omega_n = \sqrt{n}\Gamma_n^{-1/2}$. *Then for any unit vector* $\alpha \in \mathbb{R}^s$,

$$\alpha^T \Omega_n \partial \tilde{L}_{GMM}(\beta_{0S}) \to^d N(0,1)$$

Proof. $\partial \tilde{L}_{GMM}(\beta_{0S}) = 2A_n W^{\beta_{0S}} B_n$, where $B_n = \frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T \beta_{0S}) \otimes \mathbf{X}_{iS}$. $Var(\sqrt{n}B_n) = Var(g(y_i, \mathbf{X}_{iS}^T \beta_{0S}) \otimes \mathbf{X}_{iS}) \equiv V_0$. Let $H = -2\sqrt{n}EA_n W^{\beta_{0S}} B_n$, then

$$Var(H) = 4EA_n W^{\beta_{0S}} V_0 W^{\beta_{0S}} EA_n^T \equiv G$$

By CLT, $\alpha^T G^{-1/2} H \to^d N(0, \sigma^2)$, where $\sigma^2 = \alpha^T G^{-1/2} G G^{-1/2} \alpha = 1$. Note that $A_n \to^p EA_n$ and $V \to^p V_0$, hence $\Gamma_n \to^p G$. By Slutsky's theorem, $\alpha^T \sqrt{n}\Gamma_n^{-1/2} \partial \tilde{L}_{GMM}(\beta_{0S}) \to^d N(0,1)$.

Proof of Theorem 3.2: It remains to check that for $\Omega_n = \sqrt{n}\Gamma_n^{-1/2}$, Condition (ii) in Lemma B.2 holds.

For $M = P_n'(|\hat{\beta}_S|) \circ sgn(\hat{\beta}_S)$, by Assumption 3.5 and Lemma B.3, $\sqrt{n}\lambda_{\min}(\Gamma_n)^{-1/2}\|M\| \leq \sqrt{cn}(\max \eta(\beta)\sqrt{s \log s/n} + \sqrt{s}P_n'(d_n)) = O_p(\sqrt{s} \max \eta(\beta) + \sqrt{ns}P_n'(d_n)) = o_p(1)$. Hence $\|\Omega_n M\| \leq \sqrt{n}\lambda_{\min}(\Gamma_n)^{-1/2}\|M\| = o_p(1)$. Q.E.D.

B.3 Proof of Corollary 3.1

The theorem is proved by straightforward checking Assumptions 3.1-3.4, and applying Theorem 3.1.

B.4 Proof of Theorem 3.3

Proof. Let $\{x_{il}\}_{i=1}^n$ be the i.i.d. data of x_{Nl}. Under the theorem assumptions, by the strong law of large number $\frac{1}{n}\sum_{i=1}^n \epsilon_i x_{il} \to E(x_{Nl}\epsilon) \neq 0$ with probability one. Note that in penalized OLS, $L_n(\beta) = \frac{1}{n}\sum_{i=1}^n (y_i - \mathbf{X}_i^T \beta)^2$. Hence $\partial_{\beta_l} L_n(\beta_0) = -\frac{2}{n}\sum_{i=1}^n x_{il}(y_i - \mathbf{X}_i^T \beta_0) = -\frac{2}{n}\sum_{i=1}^n x_{il}\epsilon_i$. Thus $\partial_{\beta_l} L_n(\beta_0) \to -2E(x_{Nl}\epsilon)$ almost surely. Therefore, with probability one, either $\limsup_n \partial_{\beta_l} L_n(\beta_0) = \liminf_n \partial_{\beta_l} L_n(\beta_0) > 0$, or $\limsup_n \partial_{\beta_l} L_n(\beta_0) = \liminf_n \partial_{\beta_l} L_n(\beta_0) < 0$. This contradicts with the necessary condition of Theorem 2.3. Q.E.D.

C Proofs for Section 4

C.1 Preliminary results

Write $\psi(z_i, \beta) = g(y_i, \mathbf{X}_i^T \beta) \otimes \mathbf{X}_i^\beta = (\psi_1, .., \psi_{sk})^T$. Then $\psi(z_i, \beta_0) = g(y_i, \mathbf{X}_i^T \beta_0) \otimes \mathbf{X}_{iS}$, and $\partial_{\beta_S} \psi(z_i, \beta_0) = (m_1(y_i, \mathbf{X}_{iS}^T \beta_S)\mathbf{X}_{iS}\mathbf{X}_{iS}^T, ..., m_k(y_i, \mathbf{X}_{iS}^T \beta_S)\mathbf{X}_{iS}\mathbf{X}_{iS}^T)$. Let
$|\partial_{\beta_S} \psi(z_i, \beta_0)| = (|m_1(y_i, \mathbf{X}_{iS}^T \beta_S)|\mathbf{X}_{iS}\mathbf{X}_{iS}^T, ..., |m_k(y_i, \mathbf{X}_{iS}^T \beta_S)|\mathbf{X}_{iS}\mathbf{X}_{iS}^T)$. For any matrix $A = (q_j(y_i, \mathbf{x}_i^T \beta_0))$ other than $\partial_{\beta_S} \psi(z_i, \beta_0)$, write $|A| = (|q_j(y_i, \mathbf{x}_i^T \beta_0)|)$. Finally, let $\hat{V} = \frac{1}{n}\sum_{i=1}^n \psi(z_i, \beta_0)\psi(z_i, \beta_0)^T$.

Lemma C.1. *For some $a > 0$, and a random matrix X, if $E\|X\|^a < \infty$, then* $\max_{1\leq i\leq n}\|\mathbf{X}_i\| = o(n^{1/a})$ *almost surely.*

Proof. Note that $\|X\| = \|vec(X)\|$. The result then follows immeidately from Lemma D.2 in Kitamura, et al. (2004).

Lemma C.2. *If there exists $B > 0$ such that for some $p > 1$, $E\psi_j(z,\beta_0)^{2p} < B$ for all $j = 1, ..., sk$, then* $\max_{i\leq n}\|\psi(z_i,\beta_0)\| = o(n^{1/2p}s^{1/2})$ *almost surely.*

Proof. Applying Holder inequality, we have $E\|\psi(z,\beta_0)\|^{2p} = E(\sum_{j=1}^{ks}\psi_j(z,\beta_0)^2)^p$ $\leq (ks)^{p/q}\sum_{j=1}^{ks}E\psi_j(z,\beta_0)^{2p} \leq (sk)^{p/q+1}B$, where $1/p+1/q=1$. Therefore, $E\|Z\|^{2p} < \infty$, where $Z = \psi(z,\beta_0)/(sk)^{1/2}$. By Lemma C.1, $\max_{i\leq n}\|Z\| = o(n^{1/2p})$ almost surely, which is $\max_{i\leq n}\|\psi(z_i,\beta_0)\| = o(s^{1/2}n^{1/2p})$ w.p.1.

In this section, we assume $s^4 = O(n)$, and $p = 4$. Hence $\max_{i\leq n}\|\psi(z_i,\beta_0)\| = o(n^{1/8}s^{1/2})$

Lemma C.3. *Under Assumptions 4.1(ii), 4.2(i),* $\|\lambda(\beta_0)\| = O_p(\sqrt{s\log s/n})$.

Proof. Write $\lambda(\beta_0) = \rho\theta$, where $\rho = \|\lambda(\beta_0)\|$, and $\|\theta\| = 1$. Since $\lambda(\beta_0) = \arg\max\sum_{i=1}^n\log(1+\lambda^T\psi(z_i,\beta_0))$, the first order condition implies $\frac{1}{n}\sum_{i=1}^n\frac{\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} = 0$, hence

$$\begin{aligned}
0 &= \|\theta\|\times\|\frac{1}{n}\sum_{i=1}^n\frac{\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\| \geq |\theta^T\frac{1}{n}\sum_{i=1}^n\frac{\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}| \\
&= \left|\theta^T\frac{1}{n}\sum_{i=1}^n\psi(z_i,\beta_0) - \theta^T\frac{1}{n}\sum_{i=1}^n\frac{\psi(z_i,\beta_0)\lambda(\beta_0)^T\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\right| \\
&\geq \theta^T\frac{1}{n}\sum_{i=1}^n\frac{\psi(z_i,\beta_0)\psi(z_i,\beta_0)^T\lambda(\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} - \left|\theta^T\frac{1}{n}\sum_{i=1}^n\psi(z_i,\beta_0)\right| \\
&\geq \rho\theta^T\frac{1}{n}\sum_{i=1}^n\frac{\psi(z_i,\beta_0)\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\theta - \|\frac{1}{n}\sum_{i=1}^n\psi(z_i,\beta_0)\|.
\end{aligned}$$

Note that $0 \leq 1+\lambda(\beta_0)^T\psi(z_i,\beta_0) \leq 1+\rho\max_{i\leq n}\|\psi(z_i,\beta_0)\|$. By Assumption 4.2(i), $\lambda_{\min}(\frac{1}{n}\sum_i\psi(z_i,\beta_0)\psi(z_i,\beta_0)^T) = \lambda_{\min}(\hat{V}) \geq c$, and $\|\frac{1}{n}\sum_{i=1}^n\psi(z_i,\beta_0)\| = O_p(\sqrt{s\log s/n})$, hence

$$0 \geq \frac{c\rho}{1+\rho\max_{i\leq n}\|\psi(z_i,\beta_0)\|} + O_p(\sqrt{\frac{s\log s}{n}})$$

which implies $\rho = O_p(\sqrt{s\log s/n}/(1-\sqrt{s\log s/n}\max_i\|\psi(z_i,\beta_0)\|)) = O_p(\sqrt{s\log s/n})$, by Lemma C.2. In addition,

$$\|\lambda(\beta_0)\|\max_{i\leq n}\|g(z_i,\beta_0)\otimes\mathbf{X}_{iS}\| = o_p(sn^{-3/8}) = O(s^{-1/2}) \tag{C.1}$$

$$1+\lambda(\beta_0)^T[g(z_i,\beta_0)\otimes\mathbf{X}_{iS}] \geq 1-\|\lambda(\beta_0)\|\max_{i\leq n}\|g(z_i,\beta_0)\otimes\mathbf{X}_{iS}\| \geq \frac{1}{2}. \tag{C.2}$$

Lemma C.4. *Under Assumptions 3.5(i), 4.1(ii), 4.2(i),*

$$\|\frac{1}{n}\sum_{i=1}^n\frac{\partial_{\beta_S}\psi(z_i,\beta_0)\lambda(\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\| = o_p(1/\sqrt{s})$$

Proof. Write $a = \frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)\lambda(\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}$, then by (C.2), $\|a\| \leq \|\frac{1}{n}\sum_{i=1}^n \frac{|\partial_{\beta_S}\psi(z_i,\beta_0)||\lambda(\beta_0)\psi(z_i,\beta_0)^T|}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\|$ $\leq C\|\frac{1}{n}\sum_{i=1}^n |\partial_{\beta_S}\psi(z_i,\beta_0)\|\lambda(\beta_0)\psi(z_i,\beta_0)^T\|\|$. Since $\frac{1}{n}\sum_{i=1}^n |\partial\psi(z_i,\beta_0)| = \frac{1}{n}(\mathbf{X}_S^T|\Lambda_1|\mathbf{X}_S, ..., \mathbf{X}_S^T|\Lambda_k|\mathbf{X}_S)$, by Assumption3.2(i) and Lemma B.1, $\|a\| \leq C\|\lambda(\beta_0)\| \max_{i\leq n}\|\psi(z_i,\beta_0)\| = o_p(1/\sqrt{s})$, where the last equality is due to (C.1).

Lemma C.5. *Under Assumptions 3.5(i), 4.1(ii), 4.2(i),*

$$\|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\| = O_p(\sqrt{s})$$

Proof. Since $\hat{V}^{-1}$ is positive definite,

$$\begin{aligned}&\|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\| \leq \|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{|\partial\psi(z_i,\beta_0)|^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\| \\ &\leq C\lambda_{\min}(\hat{V})^{-1}\|\frac{1}{n}\sum_{i=1}^n |\partial_{\beta_S}\psi(z_i,\beta_0)|^T\| = C\|\frac{1}{n}(X_S^T|\Lambda_1|\mathbf{X}_S, ..., \mathbf{X}_S^T|\Lambda_k|\mathbf{X}_S)\| = O_p(\sqrt{s}).\end{aligned}$$

Lemma C.6. *Under Assumptions 4.1(ii), 4.2(i),*

$$\|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_0)[\lambda(\beta_0)^T\psi(z_i,\beta_0)]^2}{1+\lambda(\beta_0)^T\psi(z,\beta_0)}\| = o_p(1/\sqrt{n})$$

Proof. Let $A = \frac{1}{n}\sum_{i=1}^n \frac{[\lambda(\beta_0)^T\psi(z_i,\beta_0)]^2}{1+\lambda(\beta_0)^T\psi(z,\beta_0)}$. One can check that $A = \frac{1}{n}\sum_{i=1}^n \lambda(\beta_0)^T\psi(z_i,\beta_0)$. Hence the left-hand-side $\leq C\max_{i\leq m}\|\psi(z_i,\beta_0)\|\|A\| \leq o_p(1/\sqrt{s})\|\frac{1}{n}\sum_{i=1}^n \psi(z_i,\beta_0)\| = o_p(1/\sqrt{n})$

Lemma C.7. *Under Assumptions 4.1(ii), 4.2(i),*

$$\partial\lambda(\beta_0) = \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} - \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)\lambda(\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\right)\hat{V}^{-1}(1+o_p(1))$$

Proof. Since $\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta)}{1+\lambda(\beta)^T\psi(z_i,\beta)} = 0$, for all $\beta \in \mathbb{R}^p$, taking derivative with respect to β, and plugging-in β_0, we have $\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\partial\lambda(\beta_0)^T = \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} - \frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_0)\lambda(\beta_0)^T\partial_{\beta_S}\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}$. Since $c_1 \leq \lambda_{\min}(\hat{V}) \leq \lambda_{\max}(\hat{V}) \leq c_2$, and $\lambda(\beta_0)^T\psi(z_i,\beta_0) \to^p 0$ uniformly in $i \leq n$, we have

$$\left(\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\right)^{-1} = \hat{V}^{-1}(1+o_p(1))$$

Lemma C.8. *Under Assumptions 3.5(i), 4.1(ii), 4.2(i),* $\|\partial\lambda(\beta_0)\| = O_p(\sqrt{s})$.

Proof. By Lemma C.7, $\|\partial\lambda(\beta_0)\| = A+B+o_p(A+B)$, where $A = \|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\| = O_p(\sqrt{s})$, by Lemma C.5, and $B = \|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_0)\lambda(\beta_0)^T\partial_{\beta_S}\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\| = o_p(s^{-1/2})$, by Lemma C.4. Q.E.D.

Lemma C.9. *For any vector function* $f(z,\beta)$*, differentiable w.r.t.* β*, let* $F(\beta) = \frac{1}{n}\sum_{i=1}^n \log\{1+\lambda(\beta)^T f(z,\beta)\}$*, where* $\lambda(\beta)$ *is such that* $\frac{1}{n}\sum_{i=1}^n \frac{f(z_i,\beta)}{1+\lambda(\beta)^T f(z_i,\beta)} = 0$*, then* $\partial_\beta F(\beta) = (\Delta_1+\Delta_2+\Delta_3)(\beta)$, *where:*

$$\begin{aligned}
\Delta_1(\beta) &= \left(\frac{1}{n}\sum_{i=1}^n \partial_\beta f(z_i,\beta)\right)\left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)f(z_i,\beta)^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)\right)\\
\Delta_2(\beta) &= \frac{1}{n}\sum_{i=1}^n \frac{\partial_\beta f(z_i,\beta)}{1+\lambda(\beta)^T f(z_i,\beta)}\left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)f(z_i,\beta)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^n \frac{f(z_i,\beta)[\lambda(\beta)^T f(z_i,\beta)]^2}{1+\lambda(\beta)^T f(z_i,\beta)}\\
\Delta_3(\beta) &= \frac{1}{n}\sum_{i=1}^n \frac{-\partial_\beta f(z_i,\beta)\lambda(\beta)^T f(z_i,\beta)}{1+\lambda(\beta)^T f(z_i,\beta)}\left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)f(z_i,\beta)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^n f(z_i,\beta).
\end{aligned}$$

Proof. It is not hard to verify that $\Delta_1+\Delta_3 = \partial F(\beta) - A$, where $A = \frac{1}{n}\sum_{i=1}^n \frac{\partial f(z_i,\beta)}{1+\lambda(\beta)^T f(z_i,\beta)}[\lambda(\beta) - \left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)f(z_i,\beta)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)]$. Hence it remains to show $\Delta_2 = A$. Note that $0 = \frac{1}{n}\sum_{i=1}^n \frac{f(z_i,\beta)}{1+\lambda(\beta)^T f(z_i,\beta)} = \frac{1}{n}\sum_{i=1}^n f(z_i,\beta)\{1 - f(z_i,\beta)^T\lambda(\beta) + \frac{[\lambda(\beta)^T f(z_i,\beta)]^2}{1+\lambda(\beta)^T f(z_i,\beta)}\}$, which implies $\lambda(\beta) - \left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)f(z_i,\beta)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^n f(z_i,\beta) = \left(\frac{1}{n}\sum_{i=1}^n f(z_i,\beta)f(z_i,\beta)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^n \frac{f(z_i,\beta)[\lambda(\beta)^T f(z_i,\beta)]^2}{1+\lambda(\beta)^T f(z_i,\beta)}$. Q.E.D.

C.2 Proof of Theorem 4.1

As in the proof of Theorem 3.1, we check the conditions in Theorem 2.1. For any $\beta\in\mathbb{R}^p$, let $\mathbb{T}\beta = (\beta_S^T, 0)^T$. Define $\tilde{L}_{EL}(\beta_S) = L_{EL}(\beta_S, 0) = L_{EL}(\mathbb{T}\beta)$.

Lemma C.10. *Under Assumptions 3.5(i), 4.1(ii), 4.2(i),*

$$\partial\tilde{L}_{EL}(\beta_{0S}) = \Xi(\beta_{0S})(1+o_p(1))$$

where $\Xi(\beta_{0S}) = \frac{1}{n}\sum_{i=1}^n \partial\psi(z_i,\beta_{0S})\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \psi(z_i,\beta_{0S})$.

Proof. $\tilde{L}_{EL}(\beta_S) = \max_\lambda \frac{1}{n}\sum_{i=1}^n \log(1+\lambda^T\psi(z_i,\beta_S))$, where $\psi(z_i,\beta_S) = g(y_i, \mathbf{X}_{iS}^T\beta_1)\otimes\mathbf{X}_{iS}$, which is differentiable w.r.t. β_S. Hence by Lemma C.9, $\partial\tilde{L}_{EL}(\beta_{0S}) = \sum_{i=1}^3 \Delta_i(\beta_{0S})$, with $f(z_i,\beta_{0S}) = \psi(z_i,\beta_{0S})$, and $\Delta_1(\beta_{0S}) = \Xi(\beta_{0S})$. The result follows from equations (C.3)-(C.5) below.

Lemma C.11. *Under Assumptions 3.5(i), 4.1(ii), 4.2(i),* $\|\partial\tilde{L}_{EL}(\beta_{0S})\| = O_p(\sqrt{s\log s/n})$

Proof. From the proof of Lemma C.9, $\partial\tilde{L}_{EL}(\beta_{0S}) = \Xi(\beta_{0S}) + \Delta_2(\beta_{0S}) + \Delta_3(\beta_{0S})$, with $f(z,\beta)$ replaced with $\psi(z,\beta_{0S})$. By Assumption 3.2(i) and Lemma B.1(i),

$$\|\Xi(\beta_{0S})\| \le \|\frac{1}{n}(\mathbf{X}_{iS}|\Lambda_1|\mathbf{X}_{iS}^T, ..., \mathbf{X}_{iS}|\Lambda_k|\mathbf{X}_{iS}^T)\hat{V}^{-1}|\frac{1}{n}\sum_{i=1}^n \psi(z_i,\beta_{0S})|\| = O_p(\sqrt{s\log s/n}). \tag{C.3}$$

By Lemma C.6, and the fact that $\frac{1}{n}\sum_i |\partial_{\beta_S}\psi(z_i,\beta_{0S})| = \frac{1}{n}(\mathbf{X}_{iS}|\Lambda_1|\mathbf{X}_{iS}^T, ..., \mathbf{X}_{iS}|\Lambda_k|\mathbf{X}_{iS}^T)$,

$$\begin{aligned}
\|\Delta_2(\beta_{0S})\| &= \|\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_{0S})}{1+\lambda(\beta_{0S})^T\psi(z_i,\beta_{0S})}\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_{0S})[\lambda(\beta_{0S})^T\psi(z_i,\beta_{0S})]^2}{1+\lambda(\beta_{0S})^T\psi(z_i,\beta_{0S})}\|\\
&\le C\|\frac{1}{n}(\mathbf{X}_{iS}|\Lambda_1|\mathbf{X}_{iS}^T, ..., \mathbf{X}_{iS}|\Lambda_k|\mathbf{X}_{iS}^T)|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\psi(z_i,\beta_{0S})[\lambda(\beta_{0S})^T\psi(z_i,\beta_{0S})]^2}{1+\lambda(\beta_{0S})^T\psi(z_i,\beta_{0S})}|\|\\
&= o_p(1/\sqrt{n}). \qquad (C.4)
\end{aligned}$$

Finally,

$$\begin{aligned}
\|\Delta_3(\beta_0)\| &\leq C\|\lambda(\beta_{0S})\| \max_{i\leq n}\|\psi(z_i,\beta_{0S})\|\|\|\frac{1}{n}\sum_{i=1}^n |\partial_{\beta_S}\psi(z_i,\beta_{0S})\|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \psi(z_i,\beta_{0S})\|\| \\
&= o_p(1/\sqrt{n}). \qquad \text{(C.5)}
\end{aligned}$$

Lemma C.12. *$\partial^2\tilde{L}_{EL}(\beta_{0S}) = \Sigma(\beta_{0S}) + M(\beta)$, where $\|M(\beta_{0S})\| = o_p(1)$.*

Proof. Straightforward but tedious calculation yields $\partial^2\tilde{L}_{EL}(\beta_{0S}) = \sum_{i=1}^4 T_i$, where

$$\begin{aligned}
T_1 &= \frac{1}{n}\sum_{i=1}^n \frac{\partial\lambda(\beta_0)\partial\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} \\
T_2 &= -\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)\lambda(\beta_0)\lambda(\beta_0)^T\partial_{\beta_S}\psi(z_i,\beta_0)^T}{[1+\lambda(\beta_0)^T\psi(z_i,\beta_0)]^2} \\
T_3 &= -\frac{1}{n}\sum_{i=1}^n \frac{\partial\lambda(\beta_0)\psi(z_i,\beta_0)\lambda(\beta_0)^T\partial_{\beta_S}\psi(z_i,\beta_0)^T}{[1+\lambda(\beta_0)^T\psi(z_i,\beta_0)]^2} \\
T_4 &= \frac{1}{n}\sum_{i=1}^n \frac{1}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\sum_{j=1}^{ks}\nabla^2_{\beta_S}\psi_j(z_i,\beta_0)\lambda_j(\beta_0).
\end{aligned}$$

The result follows from Lemma C.13-C.16 below. Q.E.D.

Lemma C.13. *Under Assumptions 3.5(i), 4.1(ii), 4.2(i), $T_1 = \Sigma(\beta_{0S}) + T_{11}$, where $\|T_{11}\| = o_p(1)$.*

Proof. By Lemma C.7,

$$T_1 = \left(\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} - \frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)\lambda(\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\right)\hat{V}^{-1}(1+o_p(1))$$

$$\times\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} = A + B + R_n$$

where $\|R_n\| = o_p(1)$, and

$$A = \frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}$$

$$B = -\frac{1}{n}\sum_{i=1}^n \frac{\partial\psi(z_i,\beta_0)\lambda(\beta_0)\psi(z_i,\beta_0)^T}{(1+\lambda(\beta_0)^T\psi(z_i,\beta_0))^2}\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}$$

By Lemma C.4, C.5, $\|B\| = o_p(s^{-1/2}\sqrt{s}) = o_p(1)$. In addition, $A = \Sigma(\beta_{0S}) + A_1 + A_2$, where $\Sigma(\beta_{0S}) = \frac{1}{n}\sum_{i=1}^n \partial_{\beta_S}\psi(z_i,\beta_0)\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \partial_{\beta_S}\psi(z_i,\beta_0)^T$,

$$\begin{aligned}
\|A_1\| &= \|\frac{1}{n}\sum_{i=1}^n\left(\frac{\partial_{\beta_S}\psi(z_i,\beta_0)}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} - \partial_{\beta_S}\psi(z_i,\beta_0)\right)\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\|\| \\
&\leq C\|\lambda(\beta_0)\|\max_{i\leq n}\|\psi(z_i,\beta_0)\|\|\|\frac{1}{n}\sum_{i=1}^n \partial_{\beta_S}\psi(z_i,\beta_0)|\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \frac{\partial_{\beta_S}\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)}\| \\
&= o_p(s^{-1/2}\sqrt{s}) = o_p(1) \text{ (by Lemma C.5, B.1 and Assumption 3.2(1)).}
\end{aligned}$$

Likewise,

$$\|A_2\| = \|\frac{1}{n}\sum_{i=1}^n \partial_{\beta_S}\psi(z_i,\beta_0)\hat{V}^{-1}\frac{1}{n}\sum_{i=1}^n \left(\frac{\partial_{\beta_S}\psi(z_i,\beta_0)^T}{1+\lambda(\beta_0)^T\psi(z_i,\beta_0)} - \partial_{\beta_S}\psi(z_i,\beta_0)^T\right)\| = o_p(1)$$

Q.E.D.

Lemma C.14. *Under Assumptions 4.1(i)(iii),* $\|T_2\| = o_p(1)$

Proof. By Lemma C.3, and Cauchy-Schwarz's inequality,

$$\begin{aligned}
\|T_2\| &\leq C\|\lambda(\beta_0)\|^2\|\frac{1}{n}\sum_{i=1}^n \partial_{\beta_S}\psi(z_i,\beta_0)\partial_{\beta_S}\psi(z_i,\beta_0)^T\| \\
&= O_p(s\log s/n)\|\frac{1}{n}\sum_{i=1}^n (\mathbf{X}_{iS}\mathbf{X}_{iS}^T)^2\|m(y_i,\mathbf{X}_{iS}^T\beta_{0S})\|^2\| \text{ where} m(t_1,t_2) = \partial_{t_2}g(t_1,t_2) \\
&\leq O_p(s\log s/n)\frac{1}{n}\sum_{i=1}^n \|(\mathbf{X}_{iS}\mathbf{X}_{iS}^T)^2\|\cdot\|m(y_i,\mathbf{X}_{iS}^T\beta_{0S})\|^2 \\
&\leq O_p(s\log s/n)\sqrt{\frac{1}{n}\sum_{i=1}^n \|\mathbf{X}_{iS}\mathbf{X}_{iS}^T\|^4\frac{1}{n}\sum_{i=1}^n \|m(y_i,\mathbf{X}_{iS}^T\beta_{0S})\|^4} \\
&= O_p(s\log s/n\sqrt{E\|\mathbf{x}_S\mathbf{x}_S^T\|^4}) = O_p(s^3/n) = o_p(1). \qquad \text{(C.6)}
\end{aligned}$$

Lemma C.15. $\|T_3\| = o_p(1)$

Proof. By Lemma C.4, C.8, $\|T_3\| = O_p(\sqrt{s})o_p(s^{-1/2}) = o_p(1)$.

Lemma C.16. *Under Assumptions 4.1(ii)(iv), 4.2(i)(ii),* $\|T_4\| = o_p(1)$

Proof. $\sum_{j=1}^{ks}\nabla_\beta^2\psi_j(z_i,\beta_0)\lambda_j(\beta_0) = \mathbf{X}_{iS}\mathbf{X}_{iS}^T\lambda(\beta_0)^T[\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\otimes\mathbf{X}_{iS}]$. Hence by (C.1), and Lemma C.3,

$$\begin{aligned}
\|T_4\| &\leq C\|\frac{1}{n}\sum_{i=1}^n \mathbf{X}_{iS}\mathbf{X}_{iS}^T\lambda(\beta_0)^T[\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\otimes\mathbf{X}_{iS}]\| \\
&\leq C\lambda_{\max}(\mathbf{X}_S^T\mathbf{X}_S)/n\max_{i\leq n}\|\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\otimes\mathbf{X}_{iS}\|O_p(\sqrt{s\log s/n}) \\
&= O_p(\sqrt{s\log s/n})\max_{i\leq n}\|\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\otimes\mathbf{X}_{iS}\|.
\end{aligned}$$

Note that $\|\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\otimes\mathbf{X}_{iS}\|^2 = \|\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\|^2\|\mathbf{X}_{iS}\|^2$, and by Cauchy-Schwarz inequality, $E\|\partial_{t_2}m(y,\mathbf{x}^T\beta_0)\|^4\|x_S\|^4 \leq \sqrt{E\|\partial_{t_2}m(y,\mathbf{x}^T\beta_0)\|^8E\|x_S\|^8}$. By Assumption 4.1, $E\|\partial_{t_2}m(y,\mathbf{x}^T\beta_0)\|^8 < \infty$, and $\sqrt{E\|x_S\|^8} = O(s^2)$. Therefore, $E\|\partial_{t_2}m(y,\mathbf{x}^T\beta_0)\otimes x_S/\sqrt{s}\|^4 < \infty$, which implies $\max_{i\leq n}\|\partial_{t_2}m(y_i,\mathbf{X}_i^T\beta_0)\otimes\mathbf{X}_{iS}\| = o(s^{1/2}n^{1/4})$ with probability one. Hence $\|T_4\| = o_p(sn^{-1/4}) = o_p(1)$, as $s^4 = O(n)$.

Proof of Theorem 4.1

It remains to verify (2.2) in Theorem 2.2. For any $\gamma = (\gamma_S^T,\gamma_N^T)^T$ in a neighborhood $\mathcal{N}$ of $\hat{\beta} = (\hat{\beta}_S^T, 0)^T$, $\mathbb{T}(\gamma) = (\gamma_S^T, 0)^T$. For $\theta\in\mathbb{R}^p$, define

$$F(\theta) \quad = \quad \max_\lambda \frac{1}{n}\sum_{i=1}^n \log\{1+\lambda^T[g(y_i,\mathbf{X}_i^T\theta)\otimes\mathbf{X}_i^{\gamma_S}]\}$$

$$= \quad \frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda(\theta)^T[g(y_i,\mathbf{X}_i^T\theta)\otimes\mathbf{X}_i^{\gamma_S}]\}. \tag{C.7}$$

Then $L_{EL}(\mathbb{T}\gamma)=F(\mathbb{T}\gamma)$. By Lemma C.17 below, $F(\mathbb{T}\gamma)-F(\gamma)\leq\sum_{j=1}^{p}P_n(|\gamma_j|)-\sum_{j=1}^{p}P_n(|(\mathbb{T}\gamma)_j|)$. In addition, as $g(y_i,\mathbf{X}_i^T\theta)\otimes\mathbf{X}_i^{\gamma_S}$ is a subvector of $g(y_i,\mathbf{X}_i^T\theta)\otimes\mathbf{X}_i^{\gamma}$, it follows that

$$\begin{aligned}
F(\gamma) &= \max_{\lambda}\frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda^T[g(y_i,\mathbf{X}_i^T\gamma)\otimes\mathbf{X}_i^{\gamma_S}]\}\\
&= \max_{\tilde\lambda=(\lambda^T,0)^T}\frac{1}{n}\sum_{i=1}^{n}\log\{1+\tilde\lambda^T[g(y_i,\mathbf{X}_i^T\gamma)\otimes\mathbf{X}_i^{\gamma}]\}\\
&\leq \max_{\tilde\lambda\in\mathbb{R}^{|\gamma|_0 k}}\frac{1}{n}\sum_{i=1}^{n}\log\{1+\tilde\lambda^T[g(y_i,\mathbf{X}_i^T\gamma)\otimes\mathbf{X}_i^{\gamma}]\}=L_{EL}(\gamma).
\end{aligned}$$

Therefore, $L_{EL}(\mathbb{T}\gamma)-L_{EL}(\gamma)\leq F(\mathbb{T}\gamma)-F(\gamma)\leq\sum_{j=1}^{p}P_n(|\gamma_j|)-\sum_{j=1}^{p}P_n(|(\mathbb{T}\gamma)_j|)$. Q.E.D.

Lemma C.17. *There exists a neighborhood $\mathcal{N}$ of $\hat\beta=(\hat\beta_S^T,0)^T$, such that for all $\gamma\in\mathcal{N}$, $F(\mathbb{T}\gamma)-F(\gamma)\leq\sum_{j=1}^{p}P_n(|\gamma_j|)-\sum_{j=1}^{p}P_n(|(\mathbb{T}\gamma)_j|)$*

Proof. Let $\lambda(\theta)=\arg\max_{\lambda}\frac{1}{n}\sum_{i=1}^{n}\log\{1+\lambda^T[g(y_i,\mathbf{X}_i^T\theta)\otimes\mathbf{X}_i^{\gamma_S}]\}$. The implicit function theorem applying on the first order condition of λ implies that $\lambda(\theta)$ is continuous on $\mathbb{R}^p$. We then have, by Taylor's expansion, $F(\mathbb{T}\gamma)-F(\gamma)=\sum_{l\notin A_S,\gamma_{Nj}\neq0}\gamma_{Nl}a_l(h)$, where, by Lemma C.9, $a_l(h)=\sum_{i=1}^{3}\Delta_{il}(h)$, and h lies on the segment joining $\mathbb{T}\gamma$ and γ. Here Δ_{il} are given by:

$$\begin{aligned}
\Delta_{1l}(h) &= \left(\frac{1}{n}\sum_{i=1}^{n}\partial_{\beta_l}f(z_i,h)\right)\left(\frac{1}{n}\sum_{i=1}^{n}f(z_i,h)f(z_i,h)^T\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}f(z_i,h)\right)\\
\Delta_{2l}(h) &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial_{\beta_l}f(z_i,h)}{1+\lambda(h)^Tf(z_i,h)}\left(\frac{1}{n}\sum_{i=1}^{n}f(z_i,h)f(z_i,h)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{f(z_i,h)[\lambda(\beta)^Tf(z_i,h)]^2}{1+\lambda(h)^Tf(z_i,h)}\\
\Delta_{3l}(h) &= \frac{1}{n}\sum_{i=1}^{n}\frac{-\partial_{\beta_l}f(z_i,h)\lambda(h)^Tf(z_i,h)}{1+\lambda(h)^Tf(z_i,h)}\left(\frac{1}{n}\sum_{i=1}^{n}f(z_i,h)f(z_i,h)^T\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}f(z_i,h),
\end{aligned}$$

where $f(z_i,h)=g(y_i,\mathbf{X}_i^Th)\otimes\mathbf{X}_i^{\gamma_S}$, and $\partial_{\beta_l}f(z_i,h)=(m_1(y_i,\mathbf{X}_i^Th),...,m_k(y_i,X_i^Th))X_{il}\mathbf{X}_i^{\gamma_S}$. By Assumption 4.2(i), 4.1(i)(iii), and Lemma B.1, and the fact that $\frac{1}{n}\sum_i f(z_i,\beta_0)=O_p(\sqrt{s\log s/n})$, $\|\Delta_{1l}(\beta_0)\|=O_p(s/\sqrt{n})$. Note that the first order condition of λ implies $\sum_{i=1}^{n}\frac{f(z_i,h)}{1+\lambda(h)^Tf(z_i,h)}=0$, hence $\frac{1}{n}\sum_{i=1}^{n}\frac{[\lambda(\beta)^Tf(z_i,h)]^2}{1+\lambda(h)^Tf(z_i,h)}=\frac{1}{n}\sum_{i=1}^{n}\lambda(h)^Tf(z_i,h)$. In addition, using a similar proof of Lemma C.3, $\|\lambda(\beta_0)\|\max_{i\leq n}\|f(z_i,\beta_0)\|=o_p(1/\sqrt{s})$. Hence $\|\Delta_{2l}(\beta_0)\|=o_p(\sqrt{s\log s/n})$. Finally, $\|\Delta_{3l}(\beta_0)\|=o_p(\sqrt{s\log s/n})$. Therefore, $\|a_l(\beta_0)\|=O_p(s/\sqrt{n})$.

By the continuity of P_n', $a_l(.)$, the facts that $O_p(s/\sqrt{n})\prec\liminf_{t\to0^+}P_n'(t)$, and that $\|\hat\beta-\beta_0\|=o_p(1)$, similar to the proof of sparsity in Theorem 3.1 Condition (ii), for small enough $\mathcal{N}$, we have $|a_l(h)|<P_n'(b|\gamma_{Nl}|)$ for any $b\in(0,1)$. Q.E.D.

C.3 Proof of Theorem 4.2

Lemma C.18. *Let $\Gamma_n=A_n\hat V^{-1}A_n$, and $\Omega_n=\sqrt{n}\Gamma_n^{-1/2}$, then for any unit vector $\alpha\in\mathbb{R}^s$, $\alpha^T\Omega_n\partial\tilde L_{EL}(\beta_{0S})\to^d N(0,1)$.*

Proof. By Lemma C.10, $\partial\tilde{L}_{EL}(\beta_{0S}) = A_n\hat{V}^{-1}B_n + o_p(1)$, where $B_n = \frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_{0S}) \otimes \mathbf{X}_{iS}$. Then $Var(\sqrt{n}B_n) = Var(g(y_i, \mathbf{X}_{iS}^T\beta_{0S}) \otimes \mathbf{X}_{iS}) \equiv V_0$. Let $H = \sqrt{n}EA_nV_0^{-1}B_n$, then $Var(H) = EA_nV_0^{-1}EA_n^T \equiv G$. By CLT, $\alpha^TG^{-1/2}H \to N(0,1)$. Note that $\hat{V} \to^p V_0$ and $A_n \to^p EA_n$, and thus $\Gamma_n = A_n\hat{V}^{-1}A_n^T \to^p G$ pointwisely. Hence by Slutsky's theorem, $\alpha^T\Omega_n\partial\tilde{L}_{EL}(\beta_{0S}) \to^d N(0,1)$.

Proof of Theorem 4.2 By Lemma B.3 and Assumption 4.3, Condition (ii) in Lemma B.2 holds for $\Omega_n = \sqrt{n}(A_n\hat{V}^{-1}A_n)^{-1/2}$. Then the asymptotic normality follows immediately from Lemma C.18 and Lemma B.2. Q.E.D.

D Proofs for Section 5

Note that the results of Theorems 2.1 and 2.2 still hold under the generalized sparsity condition and the presence of local perturbation, since the objective function $L_n(.)$ defined in these two theorems are not model-specific. As before, we proceed by verifying the conditions therein.

D.1 Consistency

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

$$\tilde{L}_{GMM}(\beta_S) = \left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_S)\mathbf{X}_{iS}\right]^T W(\beta_0)\left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_S)\mathbf{X}_{iS}\right].$$

Condition (i): The same arguments in the proof of Theorem 3.1 implies

$$\begin{aligned}
\|\partial\tilde{L}_{GMM}(\beta_{0S})\| &\leq O_p(1)\|\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS}\| \\
&\leq O_p(1)\|\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS} - g(y_i, \mathbf{X}_i^T\beta_0)\mathbf{X}_{iS}\| \\
&\quad +O_p(1)\|\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_i^T\beta_0)\mathbf{X}_{iS} - Eg(y, \mathbf{x}^T\beta_0)\mathbf{x}_S\| + O_p(1)\|Eg(y, \mathbf{x}^T\beta_0)\mathbf{x}_S\| \\
&= O_p(1)(A + B + C).
\end{aligned}$$

Using the Bernstein inequality with Assumption 3.2, it can be shown that $B = O_p(\sqrt{s\log s/n})$. In addition, given that $Ex_l^2 < \infty$ for $l \in A_S$, (5.1) implies $\|Eg(y, \mathbf{x}^T\beta_0)\mathbf{x}_S\| = n^{-\alpha}\sqrt{s}$. In addition, by the mean value theorem, for some r,

$$\begin{aligned}
A &= O_p(\sqrt{\frac{s\log s}{n}}) + \|E[g(y, \mathbf{x}_S^T\beta_{0S}) - g(y, \mathbf{x}^T\beta_0)]\mathbf{x}_S\| = O_p(\sqrt{\frac{s\log s}{n}}) + \|Em(y,r)\mathbf{x}_S\mathbf{x}_N^T\beta_{0N}\| \\
&\leq O_p(\sqrt{\frac{s\log s}{n}}) + \sqrt{s}\|Em(y,r)\mathbf{x}_S\mathbf{x}_N^T\|_\infty\|\beta_{0N}\|_1 = O_p(\sqrt{\frac{s\log s}{n}} + \sqrt{s}\|\beta_{0N}\|_1),
\end{aligned}$$

where we used the fact that if $H = (h_{ij})$ is an $s \times p$ matrix, then $\|H\beta\| \leq \sqrt{s}\|H\|_\infty\|\beta\|_1 = \sqrt{s}\max_{ij}|h_{ij}|\|\beta\|_1$. The last equality is due to $\sup_{t_1,t_2}|m(t_1, t_2)| < K$ for some $K > 0$, and

$$\|Em(y,r)\mathbf{x}_S\mathbf{x}_N^T\|_\infty = \max_{i\in A_S, j\in A_N}|Em(y,r)x_ix_j| \leq K\max_{i\in A_S, j\in A_N}E|x_ix_j| < \infty.$$

It then follows that

$$\|\partial \tilde{L}_{GMM}(\beta_{0S})\| = O_p(\sqrt{(s\log s)/n} + n^{-\alpha}\sqrt{s} + \|\beta_{0N}\|_1\sqrt{s}).$$

Condition (ii) The decomposition of $\partial^2 \tilde{L}_{GMM}(\beta_{0S}) = \Sigma(\beta_{0S}) + M(\beta_{0S})$ is as before, where $\Sigma(\beta_{0S}) = 2A_n(\beta_{0S})W(\beta_{0S})A_n(\beta_{0S})^T$, and $M(\beta_{0S}) = 2B(\beta_{0S})H(\beta_{0S})$,

$$\begin{aligned} H(\beta_{0S})_{s^2\times s} &= I_s \otimes \left[W(\beta_{0S})\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS} \right], \\ B(\beta_{0S}) &= \frac{1}{n}\sum_{i=1}^n (x_{il_1} q(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS}\mathbf{X}_{iS}^T, ..., x_{il_s} q(y_i, , \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS}\mathbf{X}_{iS}^T). \end{aligned}$$

In the presence of local perturbation and generalized sparsity condition, we have

$$\begin{aligned} \|H(\beta_{0S})\| &\leq O_p(\sqrt{s})\|\frac{1}{n}\sum_{i=1}^n (g(y_i, \mathbf{X}_{iS}^T\beta_{0S}) - g(y_i, \mathbf{X}_S^T\beta_0))\mathbf{X}_{iS}\| \\ &\quad + O_p(\sqrt{s})\|\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_S^T\beta_0)\mathbf{X}_{iS} - Eg(y, \mathbf{x}^T\beta_0)\mathbf{x}_S\| + O_p(\sqrt{s})\|Eg(y, \mathbf{x}^T\beta_0)\mathbf{x}_S\| \\ &= O_p(s^2(\|\beta_{0N}\|_1^2 + \frac{\log s}{n} + n^{-2\alpha})). \end{aligned}$$

Hence $\|M(\beta_{0S})\| = o_p(1)$ as long as $s^3(\|\beta_{0N}\|_1^2 + \frac{\log s}{n} + n^{-2\alpha}) = o(1)$. By Theorem 2.1, we have

$$\|\hat{\beta}_S - \beta_{0S}\| = O_p(\sqrt{(s\log s)/n} + n^{-\alpha}\sqrt{s} + \|\beta_{0N}\|_1\sqrt{s} + \sqrt{s}P_n'(d_n)).$$

D.2 Sparsity Recovery

For some neighborhood $\mathcal{N}$ of $(\hat{\beta}_S^T, 0)^T$, and $\forall \gamma \in \mathcal{N}$, write $\gamma = (\gamma_S^T, \gamma_N^T)^T$, $\mathbb{T}\gamma = (\gamma_S^T, 0)^T$. For all $\theta \in \mathbb{R}^p$, define

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

Hence $L_{GMM}(\mathbb{T}(\gamma)) = F(\mathbb{T}\gamma)$. The same argument of the proof of Theorem 3.1 implies $L_{GMM}(\mathbb{T}\gamma) - L_{GMM}(\gamma) \leq F(\mathbb{T}\gamma) - F(\gamma)$.

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

$$\begin{aligned} F(\mathbb{T}\gamma) - F(\gamma) &= -\sum_{l\notin A_S, \gamma_l \neq 0} \gamma_{Nl} \left[\frac{1}{n}\sum_{i=1}^n \frac{\partial}{\partial\beta_l} g(y_i, \mathbf{X}_i^T h)\mathbf{X}_i(\gamma_S)\right]^T W(\gamma_S) \left[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_i^T h)\mathbf{X}_i(\gamma_S)\right] \\ &\equiv \sum_{l\notin A_S, \gamma_l \neq 0} \gamma_{Nl} a_l(h). \end{aligned}$$

There exists $\lambda_2 \in (0,1)$, $\sum_{j=1}^p (P_n(|\gamma_j|) - P_n(|(\mathbb{T}\gamma)_j|)) = \sum_{l\notin A_S, \gamma_l\neq 0} |\gamma_l| P_n'(\lambda_2|\gamma_l|)$. Hence it suffices to show that for each $l \notin A_S$, and $\gamma_l \neq 0$,

$$|\gamma_l a_l(h)| \leq |\gamma_l| P_n'(\lambda_2|\gamma_l|). \tag{D.1}$$

Note that $|a_l(\beta_{0S}, 0)| = O_p(\sqrt{s})\|n^{-1}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_i(\gamma_S)\| = O_p(\sqrt{s}(\sqrt{s\log s/n} + \|\beta_N\|_1 + n^{-\alpha}))$. By assumption, $|a_l(\beta_{0S}, 0)| = o_p(\liminf_{t\to 0^+} P_n'(t))$. By the continuity of a_l, $|a_l(\hat{\beta}^T, 0)| < \liminf_{t\to 0^+} P_n'(t)$

with probability approaching 1. Note that $h \in \mathcal{N}$. For small enough $\mathcal{N}$, again by continuity, $|a_l(h)| < \frac{1}{2}\liminf_{t\to 0^+} P_n'(t)$ w.p.a.1. Hence $|a_l(h)| < P_n'(\lambda_2|\gamma_l|)$, which yields (D.1). Q.E.D.

E Proofs for Section 6

E.1 Proof of Theorem 6.1

We can verify the conditions in Theorems 2.1,2.2 and Lemma B.2.

Conditions in Theorem 2.1

Define $L_{IV}(\beta_S) = (\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_S) \otimes \mathbf{V}_{iS})^T W(\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_S) \otimes \mathbf{V}_{iS})$. For Condition (i), we have $\partial L_{IV}(\beta_{0S}) = 2\tilde{A}_n(\beta_{0S})W[\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_{iS}^T\beta_{0S}) \otimes \mathbf{V}_{iS}]$, where

$$\tilde{A}_n = \frac{1}{n}\sum_{i=1}^n (m_1(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS}\mathbf{V}_{iS}^T, ..., m_k(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{X}_{iS}\mathbf{V}_{iS}^T).$$

Since $m_i(.)$ is bounded, we have $\lambda_{\max}(\tilde{A}_n\tilde{A}_n^T) = O_p(\lambda_{\max}((E\mathbf{x}_S\mathbf{v}_S^T)(E\mathbf{x}_S\mathbf{v}_S^T)^T)) = O_p(1)$. Hence

$$\|\partial L_{IV}(\beta_{0S})\| = O_p(\sqrt{\frac{s\log s}{n}}).$$

For Condition (ii), straightforward calculation yields $\partial^2 L_{IV}(\beta_{0S}) = 2\tilde{A}_n W\tilde{A}_n^T + 2\sum_{j\leq k}\tilde{B}_j\tilde{H}_j$, where ($\mathbf{X}_{iS} = (x_{il_1}, ..., x_{il_s})$)

$$\tilde{H}_j = I_s \otimes [W\frac{1}{n}\sum_{i=1}^n g_j(y_i, \mathbf{X}_{iS}^T\beta_{0S})\mathbf{V}_{iS}],$$

$$\tilde{B}_j = \frac{1}{n}\sum_{i=1}^n (x_{il_1}q_j(y_i, \mathbf{x}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{V}_{iS}^T, ..., x_{il_s}q_j(y_i, \mathbf{x}_i^T\beta_0)\mathbf{X}_{iS}\mathbf{V}_{iS}^T).$$

Assumption 6.2 and Lemma B.1 imply that $\|\tilde{B}_j\tilde{H}_j\| = O_p(\lambda\sqrt{s^2\log s/n}) = o_p(1)$, where

$$\lambda = \max_{l\in A_S, j\leq k}\lambda_{\max}((Ex_{il}q_j(y_i, \mathbf{x}^T\beta_0)\mathbf{x}_S\mathbf{v}^T)(Ex_{il}q_j(y_i, \mathbf{x}^T\beta_0)\mathbf{x}_S\mathbf{v}^T)^T).$$

Hence all the eigenvalues of the Hessian matrix are bounded away from zero.

Conditions in Theorem 2.2 This condition can be checked using a similar argument as in the proof of Theorem 3.1. Hence we omit the details but simply check: for any $l \notin A_S$,

$$\begin{aligned}|\frac{\partial L_{IV}(\beta_0)}{\partial\beta_l}| &= |\frac{1}{n}\sum_{i=1}^n x_{il}(m(y_i, \mathbf{X}_i^T\beta_0) \otimes \mathbf{V}_{iS})^T W(\frac{1}{n}\sum_{i=1}^n g(y_i, \mathbf{X}_i^T\beta_0) \otimes \mathbf{V}_{iS})| \\ &= O_p(s\sqrt{\frac{\log s}{n}}).\end{aligned}$$

Hence Condition (2.2) is satisfied as $P_n'(0^+) \succ s\sqrt{\log s/n}$ by Assumption 6.3.

Asymptotic Normality

Condition (i) of Lemma B.2 can be verified by the same arguments as those in Lemma B.4. For Condition (ii), note that $\lambda_{\min}(\tilde{\Gamma}_n)$ is bounded away from zero by Assumption 6.2, hence by Lemma B.3, it suffices to verify that, there exists $c > 0$, for $\mathcal{N} = \{\beta \in \mathbb{R}^s : \|\beta - \beta_{0S}\| \leq c\sqrt{s \log s/n}\}$,

$$\sqrt{n}(\max_{\beta \in \mathcal{N}} \eta(\beta)\sqrt{\frac{s \log s}{n}} + \sqrt{s} P'_n(d_n)) = o(1).$$

This holds given Assumption 6.3. Q.E.D.

E.2 Proof of Theorem 6.2

Again, we prove this theorem by checking the conditions in Theorem 2.1 and 2.2. For each l, let $L_l(\theta_{lS}) = \frac{1}{n}\sum_{i=1}^n (\partial_l \rho(Z_i, \hat{\beta}_S) - \mathbf{V}_{lS,i}^T \theta_{lS})^2$. In addition, let $\hat{u}_i = \partial_l \rho(Z_i, \hat{\beta}_S) - \mathbf{V}_{lS,i}^T \theta_{0l,S}$, $u_i = \partial_l \rho(Z_i, \beta_{0S}) - \mathbf{V}_{lS,i}^T \theta_{0l,S}$, and $e_i = \partial_l \rho(Z_i, \beta_{0S}) - D_l(\mathbf{w}_i)$. By definition, $E(e_i|\mathbf{W}_i) = 0$, and $u_i = \mathbf{V}_{lN,i}^T \theta_{0l,N} + a_l(\mathbf{W}_i) + e_i$. Then we have

$$\begin{aligned} \|\partial_{\theta_{lS}} L_l(\theta_{0l,S})\| &= \|\frac{2}{n}\sum_{i=1}^n \hat{u}_i \mathbf{V}_{lS,i}\| \\ &\leq \|\frac{2}{n}\sum_{i=1}^n (u_i - \hat{u}_i)\mathbf{V}_{lS,i}\| + \|\frac{2}{n}\sum_{i=1}^n e_i \mathbf{V}_{lS,i}\| + \|\frac{2}{n}\sum_{i=1}^n (\mathbf{V}_{lN,i}^T \theta_{0l,N} + a_l(\mathbf{W}_i))\mathbf{V}_{lS,i}\| \\ &= A + B + C. \end{aligned}$$

By the Lipschitz continuity of $\partial_l \rho(z, .)$, and Theorem 6.1, $A = O_p(\|\beta_{0S} - \hat{\beta}_S\|) = O_p(\sqrt{s \log s/n})$. It follows from $E(e_i|\mathbf{W}_i) = 0$ that $B = O_p(\sqrt{s_1 \log s_1/n})$. In addition,

$$\|\frac{2}{n}\sum_{i=1}^n \mathbf{V}_{lN,i}^T \theta_{0l,N} \mathbf{V}_{lS,i}\| \leq O_p(\sqrt{s_1}) \sum_{j \notin T_l} |\theta_{0l,j}| = O_p(\sqrt{s_1} n^{-\alpha_1}).$$

Finally, by Cauchy Schwarz inequality, $\|\frac{2}{n}\sum_{i=1}^n a_l(\mathbf{W}_i)\mathbf{V}_{lS,i}\| = O_p(\sqrt{s_1} c_n)$. Thus,

$$\|\partial_{\theta_{lS}} L_l(\theta_{0l,S})\| = O_p(\sqrt{\frac{s \log s}{n}} + \sqrt{\frac{s_1 \log s_1}{n}} + \sqrt{s_1} n^{-\alpha_1} + \sqrt{s_1} c_n).$$

The positive definiteness of the Hessian matrix is easy to verify since $\partial^2_{\theta_{l_s}} L_l(\theta_{0l,S}) = \frac{2}{n}\sum_{i=1}^n \mathbf{V}_{lS,i}\mathbf{V}_{lS,i}^T$. For the condition in Theorem 2.2, let $L'_l(\theta_l) = \frac{1}{n}\sum_{i=1}^n (\partial_l \rho(Z_i, \hat{\beta}_S) - \mathbf{V}_i^T \theta_l)^2$. Note that L'_l is differentiable, and $\forall j \notin T_l$, $|\frac{1}{n}\sum_i e_i v_{ij}| = O_p(\sqrt{\log p/n})$.

$$|\frac{\partial L'_l(\theta_{0l})}{\partial \theta_{lj}}| = |\frac{2}{n}\sum_{i=1}^n (\partial_l \rho(Z_i, \hat{\beta}_S) - \mathbf{V}_i^T \theta_{0l}) v_{ij}| = O_p(c_n + \sqrt{s \log s/n} + \sqrt{\log p/n}) = o_p(P'_n(0)).$$

Finally,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^n |\hat{D}_l(\mathbf{W}_i) - D_l(\mathbf{W}_i)|^2 &\leq \frac{2}{n}\sum_{i=1}^n |\hat{\theta}_l^T \mathbf{V}_i - \theta_{0l}^T \mathbf{V}_i|^2 + \frac{2}{n}\sum_{i=1}^n a_l(\mathbf{W}_i)^2 \\ &\leq \|\hat{\theta}_{lS} - \theta_{0l,S}\|^2 \frac{4}{n}\sum_{i=1}^n \|\mathbf{V}_{lS,i}\|^2 + \frac{4}{n}\sum_{i=1}^n (\theta_{l0,N}^T \mathbf{V}_{lN,i})^2 + \frac{2}{n}\sum_{i=1}^n a_l(\mathbf{W}_i)^2 \\ &= O_p(\frac{s_1 s \log s_1}{n} + s_1^2(\frac{\log s_1}{n} + n^{-2\alpha_1} + c_n^2 + P'_n(h_n)^2)). \end{aligned}$$

Q.E.D.

References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* , **19** 716-723

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. *Journal of Econometrics*, **101** 123-164

Andrews, D. and Soares(2010). Inference for parameters defined by moment inequalities using generalized moment selection *Econometrica*, **78** 119-157

Angrist, J. and Krueger, A. (1991). Does compulsory school attendance affect schooling and earnings? *The Quarterly Journal of Economics*, **106**, 979-1014.

Antoniadis, A. (1996). Smoothing noisy data with tapered coiflets series. *Scandinavian Journal of Statistics*, **23** 313-330

Antoniadis, A. and Fan, J. (2001). Regularization of wavelet approximations. *Journal of American Statistical Association*, **96**, 939-967.

Belloni, A., Chen, D., Chernozhukov, V. and Hansen, C.(2010). Sparse models and methods for optimal instruments with an application to eminent domain. *Manuscript.* MIT.

Belloni, A. and Chernozhukov, V. (2009). Post-l_1- penalized estimators in high dimensional linear regression models. *Manuscript.* MIT.

Belloni, A. and Chernozhukov, V. (2011). High-Dimensional Sparse Econometric Models, an Introduction. *Manuscript.* MIT.

Belloni, A. and Chernozhukov, V. (2011). l_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics*, **39**, 82-130.

Bradic, J., Fan, J. and Wang, W. (2010). Penalized composite quasi-likelihood for ultrahigh-dimensional variable selection. To appear in *Journal of the Royal Statistical Association*, Ser. B.

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

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

Chen, X. and Pouzo, D. (2011) Estimation of nonparametric conditional moment models with possibly nonsmooth moments. Forthcoming in *Econometrica.*

Chernozhukov, V. and Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica,* **73**, 245-261.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics.* **35** 2313-2404

Donald, S., Imbens, G. and Newey, W. (2003). Empirical likelihood estimation and consistent tests with conditional moment restrictions. *Journal of Econometrics*,**117** 55-93

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

Fan, J. and Lv, J. (2008). Sure independence screening for ultra-high dimensional feature space. *Journal of the Royal Statistical Society*, Ser. B. **70** 849-911

Fan, J. and Lv, J. (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* , **20** 101-148

Fan, J. and Lv, J. (2011). Non-concave penalized likelihood with NP-dimensionality. To appear in *IEEE Transactions on Information Theory.*

Fan, J., Lv, J. and Qi, L. (2010). Sparse High Dimensional Models in Economics. *Manuscript* , Princeton University

Fan, J. and Song, R. (2010). Sure independence screening in generalized linear models with NP-dimensionality. *Manuscript* Princeton University

Hall, P. and Horowitz, J. (2005). Nonparametric methods for inference in the presence of instrumental variables. *Annals of Statistics* **33** 2904-2929.

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

HOROWITZ, J. and HUANG, J. (2010). The adaptive lasso under a generalized sparsity condition. *Manuscript*, Northwestern University.

HONG, H., PRESTON, B. and SHUM, M. (2003). Generalized empirical likelihood-based model selection criteria for moment condition models. *Econometric Theory*, **19** 923-943.

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

HUANG, J., HOROWITZ, J. and WEI, F. (2010). Variable selection in nonparametric additive models. *The Annals of Statistics* **38** 2282-2313

KITAMURA, Y. (2006). Empirical likelihood methods in econometrics: theory and practice. *Cowles foundation discussion paper* Yale University.

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.

DOMINGUEZ, M. and LOBATO, I. (2004). Consistent estimation of models defined by conditional moment restrictions. *Econometrica*, **72** 1601-1615

NEWEY, W. (1990a). Efficient instrumental variables estimation of nonlinear models. *Econometrica.* **58** 809-837

NEWEY, W. (1990b). Semiparametric efficiency bound *Journal of applied 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. (1997). Convergence rates and asymptotic normality for series estimators, *Journal of Econometrics*, **79**, 147-168.

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.

Newey, W. and Smith, R. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica*, **72** 219-255

Ostu, T. (2006). Asymptotic optimality of empirical likelihood for selecting moment restrictions. *Manuscript.* Yale University.

Ostu, T. (2007). Penalized empirical likelihood estimation of semiparametric models. *Journal of Multivariate Analysis,* **98** 1923-1954

Smith, R. (1997). Alternative semiparametric likelihood approaches to generalized method of moments estimation. *Economic Journal*, **107** 503-519

Smith, R. (2007). Efficient information theoretic inference for conditional moment restrictions. *Journal of Econometrics*, **138** 430-460

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society,* Ser. B, **58** 267-288

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty, *The Annals of Statistics*, **38** 894-942

Zhang, C. and Huang, J. (2008). The sparsity and bias of the Lasso selection in high-dimensional linear models. *The Annals of Statistics*, **36** 1567-1594.

Zhao, P. and Yu, B. (2006). One model selection consistency of Lasso. *Journal of Machine Learning Research.* **7** 2541-2563

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of American Statistical Association*, **101** 1418-1429

Zou, H. and Hastie, t. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society*, Ser. B. **67** 301-320

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