Adaptive GMM Shrinkage Estimation with Consistent Moment Selection*

Zhipeng Liao†
Department of Economics, Yale University

November 18, 2010

Abstract

This paper proposes a GMM shrinkage method to efficiently estimate the unknown parameters θ_o identified by some moment restrictions, when there is another set of possibly misspecified moment conditions. I show that my method enjoys oracle-like properties, i.e. it consistently selects the correct moment conditions in the second set and at the same time, its estimator achieves the semi-parametric efficiency bound implied by all correct moment conditions. For empirical implementation, I provide both consistent and conservative data-driven procedures for selecting the tuning parameter of the penalty function. Several extensions are also studied. First, I show that the shrinkage technique can be used in GMM to perform grouped variable selection and moment selection simultaneously. Second, I establish oracle properties of the GMM shrinkage method in the practically important scenario where the moment conditions in the first set fail to strongly identify θ_o. The simulation results show that the method works remarkably well in terms of correct moment selection and the finite sample properties of its estimators. As an empirical illustration, I apply my method to estimate the life-cycle labor supply equation studied in MaCurdy (1981) and Altonji (1986). My empirical findings support the validity of the IVs used in both papers and confirm that wage is an endogenous variable in the labor supply equation.

Keywords: GMM, Model/Moment Selection, Oracle Properties, Semi-parametric Efficiency, Shrinkage Estimation, Sparsity.

*I am deeply indebted to my advisors Peter Phillips and Xiaohong Chen for guidance, inspiration and encouragement. I am also grateful to Donald Andrews for valuable suggestions. I have benefited from insightful comments made by Joseph Altonji, Yuichi Kitamura, Oliver Linton and Edward Vytlacil. Financial support from a Carl Arvid Anderson Prize of the Cowles Foundation is acknowledged.

†Email: zhipeng.liao@yale.edu. Tel: 1 203 432 3717. Comments are welcome.

1 Introduction

The generalized method of moments (GMM) is a popular methodology for estimating structural equations in economics and finance. It is particularly attractive when moment conditions appear naturally in model formulation. The statistical properties of the GMM estimators rely heavily on the quality of these moment conditions. For example, the GMM estimator based on misspecified moment conditions is inconsistent. On the other hand, including strong and valid moment conditions in GMM can help to reduce finite-sample bias and improve efficiency of the GMM estimator. Hence, whenever an empirical researcher has a set of moment conditions and there is no prior information about their validity, it is important to have some procedure to select the correctly specified moment conditions in that set and include them in estimation. This paper proposes a new method to achieve this goal.

Specifically, we are interested in estimating some unknown parameter θ_o identified by the following moment restrictions

$$E\left[g_q(Z,\theta_o)\right]=0, \tag{1.1}$$

where $\{Z_i\}_{i\leq n}$ is stationary and ergodic, Z is used generically for Z_i, the subscript q of $g_q(\cdot,\cdot)$ denotes the number of moment conditions in (1.1) and $g_q(\cdot,\cdot): R^{d_z}\times R^{d_\theta}\to R^q$. Suppose there is another set of possibly misspecified moment conditions

$$E\left[g_k(Z,\theta_o)\right]\overset{?}{=}0, \tag{1.2}$$

where "$\overset{?}{=}$" signifies that equality may hold for some elements but not others, the subscript k of $g_k(\cdot,\cdot)$ denotes the number of moment conditions in (1.2) and $g_k(\cdot,\cdot): R^{d_z}\times R^{d_\theta}\to R^{k}$[1]. When the moment conditions in set-2 (or some of them) are correctly specified, including them into estimation can improve the asymptotic efficiency of the estimator for θ_o. However, if they are misspecified, then using these moment conditions will lead to inconsistent estimation. The goal of this paper is to consistently select the correct moment conditions in set-2 and automatically include them into GMM estimation to improve the efficiency of estimating θ_o.

To reduce the risk of misspecification, one can employ the Sargan/Hansen test (Sargan

[1] Hereafter, the moment conditions in (1.1) are cited as set-1 moment conditions and those in (1.2) are cited as set-2 moment conditions.

(1958) and Eichenbaum, Hansen, and Singleton (1988)) to check the validity of the set-2 moment conditions and then estimate θ_o using the set-1 moment conditions and empirically supported moment conditions in set-2. In addition to the Sargan/Hansen's test, there are other moment selection procedures in the literature for empirical researchers to use. For example, Andrews (1999) defines moment selection criterion (MSC) using the J-test statistic and shows that consistent moment selection can be achieved by choosing the moment selection vector minimizing the MSC. Based on the J-test statistic, Andrews (1999) also proposes downward testing (DT) and upward testing (UT) moment selection procedures and shows their consistency. Hong, Preston, and Shum (2003) construct the MSC, DT and UT procedures using generalized empirical likelihood (GEL) statistic and show that these procedures are consistent in moment selection.

The above methods perform moment selection in a stepwise manner and break the moment selection and efficient estimation into two separate procedures. Moreover, when the number of moment conditions in set-2 is large, there may be too many candidate subsets of moment conditions for these methods to investigate, which makes them computationally intensive in practice. This paper embeds the moment selection in GMM estimation and once certain moment condition is selected, our method will automatically include it into estimating θ_o. Hence, our method not only selects the correct moment conditions in set-2 in one step, but also deals with the moment selection issue and efficient estimation simultaneously.

To incorporate moment selection into estimation, we first introduce a set of auxiliary unknown parameters β_o and reparametrize the moment conditions in set-2 to be

$$E\left[g_k(Z,\theta_o)-\beta_o\right]=0. \tag{1.3}$$

From (1.3), we see that if the j-th ($j=1,...,k$) moment condition in (1.2) is correctly specified (misspecified), then $\beta_{o,j}=0$ ($\beta_{o,j}\neq 0$). Hence, the zero/nonzero components in β_o can be used to identify the correctly specified/misspecified moment conditions in set-2 and consistent moment selection is equivalent to consistent selection of the zero components in ${\beta_o}$[2].

[2] Andrews (1999) notes that one can specify different β which takes some of its components as zero and the rest as unknown. θ_o and the unknown components in β can be estimated using GMM. Different specifications of β will give different sets of GMM estimators $(\widehat{\theta}_n,\widehat{\beta}_n)$ and different values of the MSC. Consistent moment selection is indicated by the zero components in $\widehat{\beta}_n$, if $(\widehat{\theta}_n,\widehat{\beta}_n)$ asymptotically minimizes the MSC. Instead of using different specifications of β, this paper treats β_o as unknown nuisance parameters

We can stack the moment conditions in (1.1) and (1.3) to get

$$E\left[\rho(Z,\theta_o,\beta_o)\right] := E\left[\begin{pmatrix} g_q(Z,\theta_o) \\ g_k(Z,\theta_o)-\beta_o \end{pmatrix}\right] = 0. \tag{1.4}$$

The GMM shrinkage estimator $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ of (θ_o, β_o) is defined as

$$(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) = \underset{(\theta,\beta)\in\Theta\times\mathcal{B}}{\arg\min} \left[\frac{\sum_{i=1}^n \rho(Z_i,\theta,\beta)}{\sqrt{n}}\right]' W_n \left[\frac{\sum_{i=1}^n \rho(Z_i,\theta,\beta)}{\sqrt{n}}\right] + n\sum_{j=1}^k \widehat{P}_{\lambda_n}(\beta_j), \tag{1.5}$$

where $\Theta\times\mathcal{B}$ is the parameter space where (θ_o, β_o) lies; W_n is a $(q+k)\times(q+k)$ weight matrix, λ_n is the tuning parameter in some general penalty function $\widehat{P}_{\lambda_n}(\cdot)$. The success of our method in simultaneous moment selection and efficient estimation relies on the "oracle properties" of the shrinkage techniques. That is to say, if $\beta_{o,j} = 0$ for some $j \in \{1, ..., k\}$, our method will estimate $\beta_{o,j}$ as zero with probability approaching 1 (w.p.a.1.). When $\beta_{o,j}$ is estimated as zero w.p.a.1., the information contained in the j-th moment condition of (1.2) is used in estimating θ_o w.p.a.1. On the other hand, the nonzero components in β_o are consistently estimated and their estimators are nonzero w.p.a.1. Hence our method can consistently distinguish the zero and nonzero components in β_o and is consistent in moment selection. Moreover, it estimates θ_o as if we knew all potentially correct moment conditions in set-2.

There are many popular choices for the penalty function $\widehat{P}_{\lambda_n}(\cdot)$. For example, the bridge penalty is defined as

$$\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma, \tag{1.6}$$

where $\gamma \in (0, 1)$; the adaptive Lasso penalty is defined as

$$\widehat{P}_{\lambda_n}(\beta) = \lambda_n \widehat{w}_\beta |\beta|, \tag{1.7}$$

where $\widehat{w}_\beta = |\widehat{\beta}_n|^{-\omega}$ $(\omega > 0)$ and $\widehat{\beta}_n$ is some first-step consistent estimator of β_o; and the

and we use the shrinkage method to consistently identify the zero components in β_o. Hence, in place of multiple sets of GMM estimations, the shrinkage approach uses only a single step revised GMM estimation.

smoothly clipped absolute deviation (SCAD) penalty is defined as

$$\widehat{P}_{\lambda_n}(\beta) = \begin{cases} \lambda_n|\beta| & |\beta| \leq \lambda_n \\ \frac{\lambda_n a|\beta|}{a-1} - \frac{\beta^2+\lambda_n^2}{2(a-1)} & \lambda_n < |\beta| \leq a\lambda_n \\ \frac{(a+1)\lambda_n^2}{2} & a\lambda_n < |\beta| \end{cases}, \tag{1.8}$$

where a is some positive real number strictly larger than 2. In the variable selection literature (i.e., when an investigator seeks to select the relevant variables to appear in the statistic model), Knight and Fu (2000) show that least squares (LS) shrinkage estimation based on the bridge penalty has positive possibility of shrinking the estimators of zero regression coefficients towards zero. Zou (2006) shows that the LS shrinkage estimator based on the adaptive Lasso penalty has the oracle properties. In a more general framework, Fan and Li (2001) study penalized maximum likelihood estimation (PMLE) using the SCAD penalty and they establish the oracle properties of their procedure in variable selection.

In the GMM framework, Caner (2009) studies variable selection using the bridge penalty function. However, there is no conservative or consistent variable selection result derived in that paper[3]. Caner and Zhang (2009) study variable selection in a scenario where the number of moment conditions and the number of structural coefficients grow with the sample size, where an adaptive elastic net penalty function is used to achieve consistent variable selection[4]. Moment selection is not addressed in Caner (2009) and Caner and Zhang (2009) either in the theory development or in the simulation studies of these papers. When the penalty function is used to perform both variable and moment selection (as in Section 4 of this paper), our results overlap those in Caner and Zhang (2009). However, as we show in Section 4, when the number of moment conditions and the number of structural coefficients are fixed, our consistent variable selection result is derived under weaker conditions on the adaptive Lasso penalty function and on the tuning parameter than those in Caner and Zhang (2009)[5].

[3] Theorem 2 of Caner (2009) shows that the centered GMM bridge estimator converges in distribution to some nonstandard random variable at the $\sqrt{n}$ rate. The nonstandard random variable has positive probability measure on the point zero. Hence, Theorem 2 of Caner (2009) only implies that the GMM bridge estimator of the zero coefficients converge to zero at a rate faster than $\sqrt{n}$ and therefore does not explain why zero coefficients are estimated as zero with positive probability in finite samples.

[4] The adaptive elastic net penalty function is defined as $\widehat{P}_{\lambda_n}(\beta) = \lambda_{1,n}|\beta|/|\widehat{\beta}_n|^{\omega} + \lambda_{2,n}|\beta|^2$, where $\omega > 0$ and $\widehat{\beta}_n$ is some first-step estimator of β_o.

[5] When the number of moment conditions and the number of structural coefficients are fixed, Caner and Zhang (2009) require the condition $\omega \in (2,4)$ on the adaptive Lasso penalty function and $n^{\frac{\omega}{2}}\lambda_n = o(1)$ and $n^{\omega-1}\lambda_n \to \infty$ on the tuning parameter to derive consistent variable selection. In contrast, the conditions

Let $\mathcal{S}_\beta = \{j : \beta_{o,j} \neq 0,\ j = 1, ..., k\}$ and $\mathcal{S}_{\beta,n} = \{j : \widehat{\beta}^{\mathcal{S}}_{n,j} \neq 0,\ j = 1, ..., k\}$ to be the index sets of the non-zero components in β_o and $\widehat{\beta}^{\mathcal{S}}_n$ respectively. Under some regularity conditions, we show that the GMM shrinkage estimator $(\widehat{\theta}^{\mathcal{S}}_n, \widehat{\beta}^{\mathcal{S}}_n)$ enjoys oracle-like properties in the sense that

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1 \tag{1.9}$$

and

$$\sqrt{n}\left(\widehat{\theta}^{\mathcal{S}}_n - \theta_o\right) \to_d N(0, \Sigma^*), \tag{1.10}$$

where Σ^* is the semi-parametric efficiency bound, implied by all the correct moment conditions. The results in (1.9) and (1.10) imply both consistent moment selection and efficient estimation.

The tuning parameter λ_n plays a key role in deriving the oracle properties of our method. Although the asymptotic properties in (1.9) and (1.10) hold if the rate that λ_n converges to zero satisfies some restrictions specified in Section 3, in finite samples it is not clear which λ should be used in the GMM shrinkage estimation. This paper proposes data-driven procedures of selecting λ_n by minimizing the MSC defined as

$$MSC_n(\lambda) = \Phi_n(\lambda) - \kappa_n h\left(\left|\mathcal{S}^c_\beta(\lambda)\right|\right), \tag{1.11}$$

where $\Phi_n(\lambda)$ is some function depending on the moment conditions selected by the GMM shrinkage method given λ, κ_n is a sequence of positive real numbers, $h(\cdot)$ is some strictly increasing function, $\mathcal{S}^c_\beta(\lambda) = \{j : \widehat{\beta}^{\mathcal{S}}_{n,j,\lambda} = 0, j = 1, ..., k\}$ is the index set of the zero components in the GMM shrinkage estimator $\widehat{\beta}^{\mathcal{S}}_{n,\lambda}$ given the tuning parameter λ and $\left|\mathcal{S}^c_\beta(\lambda)\right|$ denotes the cardinality of the index set $\mathcal{S}^c_\beta(\lambda)$. If $\Phi_n(\lambda)$ takes the form of the J-test statistic

$$J_n(\lambda) = \left[n^{-\frac{1}{2}}\sum_{i=1}^n g_{q+s_\lambda}(Z_i, \widehat{\theta}^*_{\lambda,n})\right]' W_{q+s_\lambda,n}\left[n^{-\frac{1}{2}}\sum_{i=1}^n g_{q+s_\lambda}(Z_i, \widehat{\theta}^*_{\lambda,n})\right], \tag{1.12}$$

where $g_{q+s_\lambda}(\cdot,\cdot) = [g'_q(\cdot,\cdot), g'_{s_\lambda}(\cdot,\cdot)]'$, $g_{s_\lambda}(\cdot,\cdot)$ denotes the moment conditions in set-2 selected by the GMM shrinkage estimation given λ, $W_{q+s_\lambda,n}$ is a $(q+s_\lambda)\times(q+s_\lambda)$ weight matrix and $\widehat{\theta}^*_{\lambda,n}$ is the GMM estimator based on the moment conditions implied by $g_{q+s_\lambda}(\cdot,\cdot)$, then we get the MSC defined in Andrews (1999). Under some regularity conditions, we

needed in this paper are $\omega \in (0,\infty)$, $n^{\frac{1}{2}}\lambda_n = o(1)$ and $n^{\frac{\omega+1}{2}}\lambda_n \to \infty$.

show the minimizer $\widehat{\lambda}_n$ of $MSC_n(\lambda)$ can consistently select the correctly specified moment conditions in set-2.

In some econometric models, variables or moment conditions should be selected in groups. Examples include lagged differences selection in the vector auto regression (VAR) model and moment selection in dynamic panel model where one specification assumption usually implies several moment conditions. Under some regularity conditions, we show that our method can perform grouped variable selection and grouped moment selection consistently. As an another extension, we study the properties of the GMM shrinkage estimation in the scenario where the moment conditions in set-1 are nearly weak. The GMM estimators based on nearly weak moment conditions usually have a convergence rate slower than $\sqrt{n}$. However, we show that if there are potentially valid and strong moment conditions in set-2, the GMM shrinkage estimator $\widehat{\theta}_n^{\mathcal{S}}$ can retain the $\sqrt{n}$ convergence rate.

The rest of this paper is organized as follows. Section 2 gives some examples in macroeconomics and labor economics, which arise naturally from the framework of this paper. Section 3 establishes the asymptotic properties of the GMM shrinkage estimators. Section 4 investigates the grouped variable selection and moment selection method using an adaptive group Lasso penalty in the GMM shrinkage estimation. Section 5 studies the GMM shrinkage estimation with nearly weak moment conditions in set-1. Section 6 provides data-driven procedures of selecting the tuning parameters λ_n in finite samples. Section 7 conducts several Monte Carlo experiments to check the finite sample properties of our method. Section 8 applies our method to a life-cycle labor supply model to illustrate how the GMM shrinkage method works with real data. Section 9 concludes and provides some future research directions. Proofs, technical derivations, figures and tables are included in the Appendix.

2 Examples

The first example is a dynamic panel model with fixed effects. In this model, different specification assumptions give different sets of moment conditions. As noted in Arellano and Bover (1995) and Blundell and Bond (1998), the first-differenced moment conditions may contain weak information about the structural coefficient. On the other hand, the moment conditions based on the initial value are strong but their validity depends on a stationarity assumption. We use the empirical growth model to illustrate these points.

Example 2.1 (Empirical Growth Model) *Consider the following dynamic panel data model with fixed effects*

$$\Delta Y_{i,t} = Y_{i,t-1}\theta_{1,o} + X'_{i,t}\theta_{2,o} + \varepsilon_{i,t} \text{ and } \varepsilon_{i,t} = \nu_i + u_{i,t}, \tag{2.1}$$

where $|\theta_{1,o}| < 1$, $Y_{i,t}$ *is the log GDP of country* i *at the beginning of period* t, $\Delta Y_{i,t} = Y_{i,t} - Y_{i,t-1}$, $X_{i,t}$ *is a set of predetermined control variables including education of the population, index of life expectancy, investment ratio, government consumption ratio, black market premium and the number of revolutions,* ν_i *is an unobservable individual effect and* $u_{i,t}$ *is a time varying error term. Barro and Lee (1993) estimate the equation in (2.1) using three stage least squares (3SLS) with a pooled two-period panel data set. Their estimator of* $\theta_{1,o}$ *is -0.0255, which implies a relatively slow rate of convergence in the cross-country economic growth. Caselli, Esquivel, and Lefort (1996) argues that Barro and Lee (1993) fail to take the individual effect* ν_i *and the measurement errors in* $X_{i,t}$ *into account, hence their results may be inconsistent due to the endogeneity of* $Y_{i,t-1}$ *and* $X_{i,t}$. *Under the assumption*

$$E[u_{i,t}Y_{i,0}] = 0,\ E[u_{i,t}u_{i,s}] = 0 \text{ and } E\left[(X_{i,1}, ..., X_{i,t-1})u_{i,t}\right] = 0 \tag{2.2}$$

for all i, t *and* $t \neq s$, *Caselli, Esquivel, and Lefort (1996) estimate the equation in (2.1) using GMM with the following moment conditions*

$$E[(Y_{i,0}, ..., Y_{i,t-2})\Delta\varepsilon_{i,t}] = 0 \text{ and } E[(X_{i,1}, ..., X_{i,t-2})\Delta\varepsilon_{i,t}] = 0. \tag{2.3}$$

Their estimator of $\theta_{1,o}$ *is around -0.065, which implies a faster rate of economic convergence. However, Bond, Hoeffler, and Temple (2001) argues that the results in Caselli, Esquivel, and Lefort (1996) may suffer from large finite-sample bias, because the moment conditions in (2.3) are weak if* $\theta_{1,o}$ *is close to zero. They suggest to estimate the equation in (2.1) by GMM with moment conditions in (2.3) and the following moment conditions*

$$E[\varepsilon_{i,t}\Delta Y_{i,t-1}] = 0, \tag{2.4}$$

which is implied by specification assumptions

$$E[\nu_i u_{i,t}] = 0 \text{ and } E[\eta_i Y_{i,1}] = E[\eta_i Y_{i,0}]$$

for all t and i[6]*. In this example, the moment conditions in (2.3) and (2.4) are our set-1 and set-2 moment conditions respectively. We can use our method to pick up the valid moment conditions in (2.4) and automatically include them into the estimation of* $(\theta_{1,o}, \theta_{2,o})$.

Our second example is the linear instrumental variable (IV) model. Empirical researchers sometimes can find credibly valid IVs from a so-called natural experiment. For example, one can use quarter of birth as the IV for education (Angrist and Krueger (1991)) and rainfall as the IV for the economic growth (Miguel, Satyanath, and Sergenti (2004)). On the other hand, they may have other candidate IVs, which are strongly correlated with the endogenous variables, but may not be exogenous. We use the Mincer equation to illustrate these points.

Example 2.2 (Linear IV Model) *Consider the Mincer equation*

$$Y_i = S_i\theta_{1,o} + X_i\theta_{2,o} + X_i^2\theta_{3,o} + u_i \text{ with } E[u_i|X_i] = 0, \tag{2.5}$$

where Y_i*,* S_i *and* X_i *denote individual i's log-wage, education and experience in the labor market respectively. Education* S_i *is endogenous, because unobservable ability* a_i *may be included in* u_i *and* $E[a_i S_i] \neq 0$*. IVs based on features of the school system are usually regarded to be correctly specified, e.g. the quarter of birth (Angrist and Krueger (1991)) and the tuition of college and distance to the school (Kane and Rouse (1995)). These IVs, although they may be valid, only contain limited information about* $\theta_{1,o}$*, as their correlations with* S_i *are usually small. On the other hand, one could use family background variables as IVs for* S_i*, e.g. the parent's education, sibling's education and parent's economic status when the individual was young. These variables typically have strong correlation with* S_i*, as illustrated in Card (1999), but they may be correlated with* u_i *too. In this example, we use the IVs based on the features of the school system to construct the set-1 moment conditions and the family background variables to construct the set-2 moment conditions. Our method can be used to check whether the family background variables are valid IVs or not. If they are, then their information will be automatically included into estimating* θ_o*.*

Hausman specification test can be used to check the exogeneity of some control variables in linear models. If they are exogenous, OLS estimators will be consistent and efficient.

[6] Bond, Hoeffler and Temple (2001) only estimate a simple augmented Solow model, where only the log of education is included in the basic Solow model. Hence their results are not directly comparable to these we described above.

On the other hand, if they are endogenous, then IV estimators are consistent but may have large standard errors. In the third example, we show that the shrinkage method can be used to perform exogeneity test and efficient estimation simultaneously.

Example 2.3 (Hausman Pre-test) *Consider the following structural equations*

$$\begin{aligned} y_1 &= Y_2\theta_{1,o} + X\theta_{2,o} + u, \\ Y_2 &= Z\pi + X\phi + v, \end{aligned} \tag{2.6}$$

where y_1 is the dependent variable, Y_2 denotes a vector of possibly endogenous variables, X and Z (Z =(Z_1, Z_2)) are exogenous variables, u and v are error terms. Suppose that one has the following moment conditions

$$E[Z_1 u] = 0 \text{ and } E[Xu] = 0 \tag{2.7}$$

which can be used to identify and consistently estimate the coefficients $(\theta_{1,o}, \theta_{2,o})$. However, if Y_2 is exogenous, then the OLS estimator is consistent and more efficient. The OLS estimator of $(\theta_{1,o}, \theta_{2,o})$ can be viewed as a GMM estimator based on the moment conditions in (2.7) and the following possibly misspecified moment conditions

$$E[Y_2 u] \stackrel{?}{=} 0. \tag{2.8}$$

In this example, we have the moment conditions in (2.7) and (2.8) as the set-1 and set-2 moment conditions, respectively. Our method can be used to check whether the moment conditions in (2.8) are correctly specified or not. The GMM shrinkage estimator will asymptotically become the OLS estimator if the moment conditions in (2.8) are valid and the IV estimator otherwise.

3 Asymptotic Properties of the GMM Shrinkage Estimator

This section establishes the oracle properties of the GMM shrinkage estimator. For ease of notation, we sort the elements in β_o in the following way $\beta_o = (\beta_{o,+}, \beta_{o,-})$, where $\beta_{o,+} \neq 0$ and $\beta_{o,-} = 0$. We denote the unknown parameter (θ_o, β_o) as α_o, i.e. $\alpha_o = (\theta_o, \beta_o)$. Accordingly, the GMM shrinkage estimator of α_o is denoted as $\widehat{\alpha}_n^{\mathcal{S}} = (\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$. We use $\|\cdot\|_E$ to denote the Euclidean norm in the Euclidean space.

3.1 Consistency and the Rate of Convergence

We first present and discuss the sufficient conditions for consistency of $\widehat{\alpha}_n^{\mathcal{S}}$. The assumptions imposed on the moment functions are similar to these ensuring the consistency of the GMM estimator, while some extra conditions are needed to make sure that attaching a penalty function to the GMM criterion function will not lead to inconsistent estimation.

Assumption 3.1 *(i)* $E[g_k(Z,\theta)]$ *is continuous in* θ *and for any* $\varepsilon > 0$*, there exists some* $\delta_\varepsilon > 0$ *such that*

$$\inf_{\{\theta\in\Theta:\|\theta-\theta_o\|_E\geq\varepsilon\}} \|E[g_q(Z,\theta)]\|_E > \delta_\varepsilon; \tag{3.1}$$

(ii) the following uniform law of large numbers (ULLN) holds

$$\sup_{\theta\in\Theta}\left[n^{-1}\sum_{i=1}^{n}\{g_l(Z_i,\theta) - E\left[g_l(Z_i,\theta)\right]\}\right] = o_p(1) \tag{3.2}$$

for $l = q, k$*; (iii) there exists some symmetric, nonrandom and positive definite matrix* W_o *such that*

$$W_n \to_p W_o; \tag{3.3}$$

(iv) the penalty function $\widehat{P}_{\lambda_n}(\cdot)$ *is non-negative and* $\widehat{P}_{\lambda_n}(\beta_{o,j}) = o_p(1)$ *for all* $j = 1, ..., k$*.*

Condition (3.1) in assumption 3.1.(i) is the identifiable uniqueness condition for θ_o. By definition $\beta_o = E[g_k(Z,\theta_o)]$, thus β_o is locally uniquely identified under (3.1) and the continuity of $E[g_k(Z,\theta)]$. Assumption 3.1.(ii) is a high-level condition, because it does not specify the data structure and the properties of the moment functions. The advantage of this high-level condition is it makes our results applicable to models with a general data structure (e.g., i.i.d., i.ni.d. or weakly dependent data) and general moment functions (e.g., non-smooth moment functions). Assumption 3.1.(iii) is also a high-level condition, because it does not specify the form of the weight matrix W_n and its probability limit W_o. It is clear that when W_n is an identity matrix, this assumption holds automatically. Assumption 3.1.(iv) implies that the shrinkage effect of the penalty function on the moment selection coefficients converges in probability to zero as $n \to \infty$. We show that the bridge, adaptive Lasso and SCAD penalty functions satisfy assumption 3.1.(iv) in Appendix F.

Lemma 3.1 *Under assumption 3.1, the GMM shrinkage estimator is consistent, i.e.,* $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) \to_p (\theta_o, \beta_o)$*.*

From the consistency of $\widehat{\beta}_n^{\mathcal{S}}$, we can deduce that if $j \in \mathcal{S}_\beta$, then $\widehat{\beta}_{n,j}^{\mathcal{S}}$ will be estimated as nonzero w.p.a.1 and we have $j \in \mathcal{S}_{\beta,n}$ w.p.a.1. Hence under assumption 3.1, the misspecified moment conditions in set-2 will not be selected asymptotically. However consistent moment selection also requires that if $j \in \mathcal{S}_\beta^c$, then $j \in \mathcal{S}_{\beta,n}^c$ w.p.a.1. The latter result can not be deduced from the consistency of $\widehat{\beta}_n^{\mathcal{S}}$, because what we need to show is $\widehat{\beta}_{n,j}^{\mathcal{S}}$ $(j \in \mathcal{S}_\beta^c)$ concentrates on zero w.p.a.1, while the consistency only indicates that $\widehat{\beta}_{n,j}^{\mathcal{S}}$ $(j \in \mathcal{S}_\beta^c)$ concentrates on local neighborhoods of zero w.p.a.1.

Remark 3.1 *In Corollary F.1 in Appendix F, we show that if $\lambda_n = o(1)$, then the bridge, adaptive Lasso and SCAD penalty functions satisfy assumption 3.1.(iv). We next show that $\lambda_n = o(1)$ is also a necessary condition for the consistency of the GMM shrinkage estimator based on these penalty functions. First, note that if $\lambda_n \to \infty$, then $\widehat{P}_{\lambda_n}(\beta) \to \infty$ unless $\beta = 0$. Hence, if $\lambda_n \to \infty$, we can invoke the epi-convergence theorem in Geyer (1994) and Knight (1999) to deduce that the GMM shrinkage estimator $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ satisfies $\widehat{\beta}_n^{\mathcal{S}} \to_p 0$ and*

$$\widehat{\theta}_n^{\mathcal{S}} \to_p \arg\min_{\theta \in \Theta} E\left[\rho(Z_i, \theta, 0)\right]' W_o E\left[\rho(Z_i, \theta, 0)\right]. \tag{3.4}$$

From (3.4), we see that $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ is inconsistent if β_o is a non-zero vector. On the other hand, if $\lambda_n \to \lambda_0 \in (0, \infty)$, then using the argmax continuous mapping theorem (ACMT) we can show that the GMM shrinkage estimator based on the bridge penalty satisfies

$$\widehat{\alpha}_n^{\mathcal{S}} \to_p \arg\min_{\alpha \in \mathcal{A}} \left\{ E\left[\rho(Z_i, \alpha)\right]' W_o E\left[\rho(Z_i, \alpha)\right] + \lambda_0 \sum_{j=1}^{k} |\beta_j|^\gamma \right\}, \tag{3.5}$$

where $\alpha = (\theta, \beta)$ and $\mathcal{A} = \Theta \times \mathcal{B}$, and the GMM shrinkage estimator based on the SCAD penalty satisfies

$$\widehat{\alpha}_n^{\mathcal{S}} \to_p \arg\min_{\alpha \in \mathcal{A}} \left\{ E\left[\rho(Z_i, \alpha)\right]' W_o E\left[\rho(Z_i, \alpha)\right] + \sum_{j=1}^{k} P_{\lambda_0}(\beta_j) \right\}, \tag{3.6}$$

where

$$P_{\lambda_0}(\beta_j) = \begin{cases} \lambda_0 |\beta_j| & |\beta_j| \le \lambda_0 \\ \frac{\lambda_0 a |\beta_j|}{a-1} - \frac{\beta_j^2 + \lambda_0^2}{2(a-1)} & \lambda_0 < |\beta_j| \le a\lambda_0 \\ \frac{(a+1)\lambda_0^2}{2} & a\lambda_0 < |\beta_j| \end{cases}. \tag{3.7}$$

Using the epi-convergence theorem, we can show that the GMM shrinkage estimator based on the adaptive Lasso penalty satisfies

$$\widehat{\alpha}_n^{\mathcal{S}} \rightarrow_p \arg\min_{\alpha\in\mathcal{A}} \begin{cases} E\left[\rho(Z_i,\alpha)\right]' W_o E\left[\rho(Z_i,\alpha)\right] + \lambda_0 \sum_{j\in\mathcal{S}_\beta} \frac{|\beta_j|}{|\beta_{o,j}|^\omega} & if\ \beta_j = 0\ \forall j \in \mathcal{S}_\beta^c \\ \infty & otherwise \end{cases}. \tag{3.8}$$

From the results in (3.5), (3.6) and (3.8), we see that if $\lambda_n \to \lambda_0 \in (0,\infty)$*, then* $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ *is inconsistent. Thus for the bridge, SCAD and adaptive Lasso penalty functions,* $\lambda_n = o(1)$ *is also a necessary condition for the consistency of* $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$*.*

We next present conditions needed to derive the convergence rate of $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$.

Assumption 3.2 *(i) The following functional central limit theorem (FCLT) holds*

$$\sup_{\theta\in\Theta} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \left\{ g_l(Z_i,\theta) - E\left[g_l(Z_i,\theta)\right] \right\} \right] = O_p(1), \tag{3.9}$$

for $l = q, k$*; (ii)* $E[g_l(Z,\theta)]$ *is continuously differentiable in some neighborhood of* θ_o *for* $l = q, k$*; (iii)* $\frac{\partial E[g_q(Z,\theta_o)]}{\partial \theta'}$ *has full column rank; (iv) the penalty function* $\widehat{P}_{\lambda_n}(\cdot)$ *satisfies* $\widehat{P}_{\lambda_n}(0) = 0$ *and is continuously twice differentiable at* $\beta_{o,j}$ *with*

$$\left| \widehat{P}''_{\lambda_n}(\beta_{o,j}) \right| = o_p(1) \tag{3.10}$$

for all $j \in \mathcal{S}_\beta$*.*

Assumption 3.2.(i) is a high-level condition, which can be verified by applying the Donsker's theorem in specific models. Assumption 3.2.(ii) imposes a local differentiability condition on the expectation of the moment function $g_l(Z,\theta)$ $(l = q, k)$. Assumption 3.2.(iii) is a local identification condition for θ_o. If this assumption fails, the resulting estimator $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ may not be $\sqrt{n}$-consistent. Assumption 3.2.(iv) imposes some local smoothness conditions on the penalty function $\widehat{P}_{\lambda_n}(\cdot)$. Intuitively, this condition implies that attaching a penalty function to the GMM criterion function does not cause any local identification problem for the unknown parameter (θ_o, β_o). We show that the bridge, adaptive Lasso and SCAD penalty functions satisfy Assumption 3.2.(iv) in Appendix F.

Lemma 3.2 *Under the assumptions 3.1-3.2, we have*

$$(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) = (\theta_o, \beta_o) + O_p(\delta_n), \tag{3.11}$$

where $\delta_n = \max\left\{b_n, n^{-\frac{1}{2}}\right\}$ *and* $b_n = \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|$.

It is interesting to see that the convergence rate of $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ may also depend on the rate of the tuning parameter λ_n converging to zero. Intuitively, the finite-sample bias of the shrinkage estimator comes from two sources. The first is the stochastic error, which converges (in probability) to zero with the $\sqrt{n}$ rate. The second bias is due to the shrinkage effect of the penalty function on the estimators of the non-zero components in β_o. The shrinkage bias converges (in probability) to zero with the rate b_n. Hence, the convergence rate of $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ is of the order $b_n \vee n^{-\frac{1}{2}}$.

Remark 3.2 *If* $\widehat{P}_{\lambda_n}(\cdot)$ *is the bridge or adaptive Lasso penalty function, then* $b_n = O_p(\lambda_n)$. *The condition imposed on* λ_n *to show the consistency of* $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$, *i.e.* $\lambda_n = o(1)$, *is insufficient to deduce that* $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ *is* $\sqrt{n}$*-consistent. For example, if* $\lambda_n = o(1)$ *and* $\sqrt{n}\lambda_n \to \infty$, *then from Lemma 3.2, we have*

$$(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) = (\theta_o, \beta_o) + O_p(\lambda_n)$$

which implies the convergence rate of $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$ *may be slower than* $\sqrt{n}$. *Hence, we need to impose stronger condition on* λ_n *to ensure that the GMM shrinkage estimator is* $\sqrt{n}$*-consistent. However, if* $\widehat{P}_{\lambda_n}(\cdot)$ *is the SCAD penalty function, then under the condition* $\lambda_n = o(1)$, *we can deduce that* $b_n = \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = 0$ *when* n *is sufficiently large. So in this case,* $\lambda_n = o(1)$ *is a sufficient condition for the* $\sqrt{n}$*-consistency of* $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$.

3.2 Sparsity and Asymptotic Normality

In this sub-section, we derive the sparsity (i.e., the zero components in β_o are estimated as zero w.p.a.1.) of $\widehat{\beta}_n^{\mathcal{S}}$ and the centered joint limiting distribution of $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_{n,+}^{\mathcal{S}})$, where $\widehat{\beta}_{n,+}^{\mathcal{S}}$ denotes the GMM shrinkage estimator of the nonzero components $\beta_{o,+}$ in β_o. We first present and discuss the assumptions needed to show sparsity.

Assumption 3.3 *(i) The tuning parameter λ_n satisfies*

$$\sqrt{n}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = o_p(1); \tag{3.12}$$

(ii) the penalty function satisfies

$$\liminf_{n\to\infty}\left[\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{r_n\lambda_n}\right] > 0 \ a.e. \tag{3.13}$$

for all $j \in \mathcal{S}^c_\beta$, where r_n is some non-negative sequence such that $n^{\frac{1}{2}}\lambda_n r_n \to \infty$.

Assumption 3.3.(i) indicates that the convergence rate of $\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|$ for all $j \in \mathcal{S}_\beta$ is faster than $\sqrt{n}$. Under this assumption, Lemma 3.2 implies that

$$\sqrt{n}\left(\widehat{\alpha}^{\mathcal{S}}_n - \alpha_o\right) = O_p(1) \tag{3.14}$$

i.e., the convergence rate of $\widehat{\alpha}^{\mathcal{S}}_n$ is $\sqrt{n}$. Assumption 3.3.(ii) is a generalized version of the condition (3.5) in Fan and Li (2001). Intuitively, assumption 3.3.(ii) implies that the shrinkage estimator $\widehat{\beta}^{\mathcal{S}}_{n,j}$ of $\beta_{o,j}$ $(j \in \mathcal{S}^c_\beta)$ is the minimizer of $\widehat{P}_{\lambda_n}(\cdot)$ w.p.a.1. From assumptions 3.1.(iv) and 3.2.(iv), we know that $\widehat{P}_{\lambda_n}(\cdot)$ is locally minimized at 0. Hence assumption 3.3.(ii) is the key condition needed for showing sparsity. We show that the bridge, adaptive Lasso and SCAD penalty functions satisfy Assumption 3.3 in Appendix F.

Theorem 3.3 (sparsity) *Under the assumptions 3.1-3.3, we have*

$$\Pr\left(\widehat{\beta}^{\mathcal{S}}_{n,j} = \beta_{o,j}\right) \to 1 \tag{3.15}$$

for all $j \in \mathcal{S}^c_\beta$.

From the consistency and sparsity of $\widehat{\beta}^{\mathcal{S}}_n$, we can immediately deduce that

$$\lim_{n\to\infty}\Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1, \tag{3.16}$$

i.e. the consistent moment selection. The sparsity of $\widehat{\beta}^{\mathcal{S}}_n$ also facilitates the way of deriving the centered joint limiting distribution of $(\widehat{\theta}^{\mathcal{S}}_n, \widehat{\beta}^{\mathcal{S}}_{n,+})$, because we can concentrate on the

event $\{\widehat{\beta}^{\mathcal{S}}_{n,j} = 0,\ j \in \mathcal{S}^c_\beta\}$, which has probability measure 1 with the sample size n goes to infinity.

Assumption 3.4 *The following central limit theorem (CLT) holds*

$$n^{-\frac{1}{2}} \sum_{i \leq n} \{\rho(Z_i, \theta_o, \beta_o) - E\left[\rho(Z_i, \theta_o, \beta_o)\right]\} \to_d \Psi(\theta_o, \beta_o), \tag{3.17}$$

where $\Psi(\theta_o, \beta_o)$ is some Gaussian random variable.

Assumption 3.4 is a high-level condition, which can be verified by applying CLTs in models with specific moment functions and data structure.

Let $g_{d_{\beta\_}}(Z, \theta_o)$ and $g_{d_{\beta+}}(Z, \theta_o)$ denote the potentially valid and misspecified moment functions in set-2 respectively. Denote

$$\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} = \begin{bmatrix} \frac{\partial E[g_{q+d_{\beta\_}}(Z,\theta_o)]}{\partial \theta'} & 0 \\ \frac{\partial E[g_{d_{\beta+}}(Z,\theta_o)]}{\partial \theta'} & -I_{d_{\mathcal{S}_\beta}} \end{bmatrix},$$

where $g_{q+d_{\beta\_}}(Z, \theta_o) = \left[g'_q(Z, \theta_o), g'_{d_{\beta\_}}(Z, \theta_o)\right]'$, $d_{\mathcal{S}_\beta}$ is the cardinality of the index set $\mathcal{S}_\beta$ and $I_{d_{\mathcal{S}_\beta}}$ denotes a $d_{\mathcal{S}_\beta} \times d_{\mathcal{S}_\beta}$ identity matrix. If we define $M_{11} = \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}}\right]$, then under assumptions 3.1.(iii) and 3.2.(iii), M_{11} is non-singular.

Theorem 3.4 ($\sqrt{n}$-Normality) *Under the assumptions 3.1-3.4, we have*

$$\sqrt{n}\left[(\widehat{\theta}^{\mathcal{S}}_n, \widehat{\beta}^{\mathcal{S}}_{n,+}) - (\theta_o, \beta_{o,+})\right] \to_d N(0, M_{11}^{-1}\Sigma_{11}M_{11}^{-1}), \tag{3.18}$$

where

$$\Sigma_{11} = \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o E\left[\Psi(\theta_o, \beta_o)\Psi'(\theta_o, \beta_o)\right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}}\right]$$

and $\Psi(\theta_o, \beta_o)$ is defined in (3.17).

3.3 Oracle Properties

The oracle properties state that the GMM shrinkage estimation can consistently identify all potentially valid moment conditions in set-2 and its estimator of θ_o can attain the semi-parametric efficiency bound implied by all correct moment conditions. As the consistent

moment selection is directly implied by the consistency of $\widehat{\beta}_n^{\mathcal{S}}$ established in Lemma 3.1 and the sparsity of $\widehat{\beta}_n^{\mathcal{S}}$ established in Lemma 3.3, the oracle properties follow if we can show that $\widehat{\theta}_n^{\mathcal{S}}$ is semi-parametric efficient.

If we had prior information about the validities of the moment conditions in set-2, then there would be $q + d_{\mathcal{S}_\beta}$ moment conditions to estimate θ_o. We can stack these moment conditions as

$$m_e(\theta_o) = E \left[\begin{array}{c} g_q(Z, \theta_o) \\ g_{d_{\beta^-}}(Z, \theta_o) \end{array} \right]_{(q+d_{\mathcal{S}_\beta})\times 1} = 0. \tag{3.19}$$

From the moment conditions in (3.19), we can compute the semiparametric efficiency bound of θ_o as

$$(\Sigma^*)^{-1} = \left[\frac{\partial m_e(\theta)}{\partial \theta_o'}\right]' \{V_{e,o}\}^{-1} \left[\frac{\partial m_e(\theta)}{\partial \theta_o}\right]', \tag{3.20}$$

where $V_{e,o}$ is the leading $(q + d_{\mathcal{S}_\beta}) \times (q + d_{\mathcal{S}_\beta})$ sub-matrix of $E\left[\Psi(\theta_o, \beta_o)\Psi'(\theta_o, \beta_o)\right]$.

If we choose the asymptotically efficient weight matrix W_n^* in the GMM shrinkage estimation such that

$$W_n^* \to_p W_o = \left\{E\left[\Psi(\theta_o, \beta_o)\Psi'(\theta_o, \beta_o)\right]\right\}^{-1}, \tag{3.21}$$

then an interesting question is whether the resulting GMM shrinkage estimator $\widehat{\theta}_n^{\mathcal{S}}$ of θ_o asymptotically attains the semi-parametric efficiency bound in (3.20). The answer to the above question is affirmative, as illustrated in the following theorem.

Theorem 3.5 (Oracle Properties) *Under the assumptions 3.1-3.3, we have*

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1. \tag{3.22}$$

Furthermore, if the weight matrix W_n satisfies (3.21) and the assumption 3.4 holds, then we have

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right) \to_d N(0, \Sigma^*), \tag{3.23}$$

where Σ^* is defined in (3.20).

Remark 3.6 *In finite samples, the naive confidence intervals (CIs) for θ_o are constructed using the asymptotic distribution in (3.23) and "plug-in" estimator of the variance covariance matrix Σ^* based on the shrinkage estimator $\widehat{\theta}_n^{\mathcal{S}}$ and the selected moment conditions.*

It should be noted that the results in (3.22) and (3.23) are pointwise asymptotic properties. Hence in finite samples, the naive CIs fail to take the moment selection errors into account, though the moment selection errors are integrated into the plug-in estimator of Σ^**. That is to say, the naive CIs may be mistakenly centered if some misspecified moment conditions are selected in finite samples. One should note that the post moment selection estimators based on other moment selection procedures also suffer from this problem. Ignoring the errors in moment selection may lead to poor coverage probabilities of naive CIs and size distortion of hypothesis tests, which represents a well-known challenge in the model/moment selection literature as recently overviewed in Leeb and Potscher (2005). The treatment of the PMS inference is beyond the scope of this paper and we will address this issue in a different paper.*

4 Extension I: Grouped Variable and Grouped Moment Selection via Adaptive Group Lasso

In this section, we study the grouped variable selection and grouped moment selection in GMM using the shrinkage method. In some econometric models, variables and moment conditions are selected in groups, instead of being selected individually. One example is the selection of lagged differences in the VAR model. Another example is the selection of moment conditions in dynamic panel models, where one specification assumption usually implies several moment conditions. In the latter example, the moment conditions implied by the same specification assumption should be accepted or rejected altogether. However, if these moment conditions are treated individually, contradictory results may appear in empirical studies, when some of the moment conditions are accepted and the others are rejected.

To perform the grouped variable and moment selection, we need to impose some extra restrictions on the general penalty function $\widehat{P}_{\lambda_n}(\cdot)$. For the brevity of this paper, we only consider the adaptive group Lasso penalty function in this section. The adaptive group Lasso penalty is defined as

$$\widehat{P}_{\lambda_n}(\alpha) = \lambda_n \widehat{w}_\alpha \left\|\alpha\right\|_2, \tag{4.1}$$

where $\widehat{w}_\alpha = \left\|\widehat{\alpha}_n\right\|_2^{-\omega}$ $(\omega > 0)$, $\widehat{\alpha}_n$ is a first-step consistent estimator of α and $\left\|\cdot\right\|_2$ denotes the l_2-norm. The adaptive group Lasso is originally proposed in Wang and Leng (2008) to perform consistent grouped variable selection and efficient estimation in LS regression

models. It is clear that when α is a scale, the adaptive group Lasso penalty is reduced to be the adaptive Lasso penalty defined in (1.7). Intuitively the adaptive group Lasso can perform grouped variable or moment selection, because by definition it delivers the shrinkage effect groupwisely and the estimators of the grouped parameters will be shrunk to zero only when all of them are zero.

Suppose that the unknown parameters θ_o can be decomposed into J_θ groups i.e. $\theta_o = (\theta_{1,o}, ..., \theta_{J_\theta,o})$. There are J_{θ^+} sub-groups indexed by $\mathcal{S}_\theta$ such that $\|\theta_{o,j}\|_2 \neq 0$ for all $j \in \mathcal{S}_\theta$ and J_{θ^-} sub-groups indexed by $\mathcal{S}_\theta^c$ such that $\|\theta_{o,j}\|_2 = 0$ for all $j \in \mathcal{S}_\theta^c$. Similarly, suppose that the moment selection coefficients β_o can be decomposed into J_β groups i.e. $\beta_o = (\beta_{1,o}, ..., \beta_{J_\beta,o})$ with J_{β^+} sub-groups (indexed by $\mathcal{S}_\beta$) such that $\left\|\beta_{o,j}\right\|_2 \neq 0$ for all $j \in \mathcal{S}_\beta$ and J_{β^-} sub-groups (indexed by $\mathcal{S}_\beta^c$) such that $\left\|\beta_{o,j}\right\|_2 = 0$ for all $j \in \mathcal{S}_\beta^c$. Denote $\mathcal{S}_\alpha = \{j : \|\alpha_{o,j}\|_2 \neq 0,\ j = 1, ..., J_\theta + J_\beta\}$ be the index set of the grouped non-zero components in α_o, then by definition there is $\mathcal{S}_\alpha = \mathcal{S}_\theta \cup \mathcal{S}_\beta$ and $\mathcal{S}_\alpha^c = \mathcal{S}_\theta^c \cup \mathcal{S}_\beta^c$.

The GMM shrinkage estimator $\widehat{\alpha}_n^{\mathcal{S}}$ with grouped variable and moment condition selection is defined as

$$\widehat{\alpha}_n^{\mathcal{S}} = \arg\min_{\alpha \in \mathcal{A}} \left[\frac{\sum_{i=1}^n \rho(Z_i, \alpha)}{\sqrt{n}}\right]' W_n \left[\frac{\sum_{i=1}^n \rho(Z_i, \alpha)}{\sqrt{n}}\right] + \sum_{j=1}^{J_\theta + J_\beta} \widehat{P}_{\lambda_n}(\alpha_j), \tag{4.2}$$

where $\mathcal{A}$ is parameter space where α_o lies and α_j denotes the j-th group of parameters in α. Let $\mathcal{S}_{n,\alpha} = \{j : \left\|\widehat{\alpha}_{n,j}^{\mathcal{S}}\right\|_2 \neq 0,\ j = 1, ..., J_\theta + J_\beta\}$ denote the index set of groups of nonzero components in $\widehat{\alpha}_n^{\mathcal{S}}$. For the ease of notation, we sort the groups in θ_o in the following way $\theta_o = (\theta_{o,+}, \theta_{o,-})$, where $\theta_{o,+} = \{\theta_{o,j} : j \in \mathcal{S}_\theta\}$ and $\theta_{o,-} = \{\theta_{o,j} : j \in \mathcal{S}_\theta^c\}$. Under some regularity conditions, we show the GMM shrinkage estimation can perform consistent grouped variable selection and moment selection, i.e.

$$\lim_{n \to \infty} \Pr\left(\mathcal{S}_\alpha = \mathcal{S}_{n,\alpha}\right) = 1, \tag{4.3}$$

and the GMM shrinkage estimator $\widehat{\theta}_{n,+}^{\mathcal{S}}$ of $\theta_{o,+}$ is semi-parametric efficient, i.e.

$$\sqrt{n}\left(\widehat{\theta}_{n,+}^{\mathcal{S}} - \theta_{o,+}\right) \to_d N(0, \Sigma^+), \tag{4.4}$$

where Σ^+ is the semi-parametric efficiency bound, implied by the true model with all correct moment conditions. We first derive the convergence rate of $\widehat{\alpha}_n^{\mathcal{S}}$.

Lemma 4.1 *If $\lambda_n = o(1)$ and assumptions 3.1.(i)-(iii) are satisfied, then the GMM shrinkage estimator defined in (4.2) is consistent. If we further assume that the assumptions 3.2.(i)-(iii) are satisfied, then*

$$\widehat{\alpha}_n^{\mathcal{S}} = \alpha_o + O_p(\delta_n), \tag{4.5}$$

where $\delta_n = \max\left\{\lambda_n, n^{-\frac{1}{2}}\right\}$.

From the convergence rate in (4.5), we can deduce that if $\sqrt{n}\lambda_n = O_p(1)$, then

$$\sqrt{n}\left(\widehat{\alpha}_n^{\mathcal{S}} - \alpha_o\right) = O_p(1)$$

and hence $\widehat{\alpha}_n^{\mathcal{S}}$ is $\sqrt{n}$-consistent. We next establish the sparsity of $\widehat{\alpha}_n^{\mathcal{S}}$.

Assumption 4.1 *(i) For $l = q, k$, the moment function $g_l(z, \theta)$ is continuously differentiable in θ for almost all z; (ii) the following SLLNs hold*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial g_l(Z, \theta)}{\partial \theta} - E\left[\frac{\partial g_l(Z, \theta)}{\partial \theta} \right] \right\} \right| = o_p(1). \tag{4.6}$$

Assumption 4.1.(i) imposes differentiability condition on the moment function $g_l(z, \theta)$ $(l = q, k)$, which is needed to invoke the Karush-Kuhn-Tucker (KKT) optimality condition to derive the sparsity of $\widehat{\alpha}_n^{\mathcal{S}}$. Note that by definition, $\rho(Z, \alpha) = \rho(Z, \theta, \beta)$ and $\rho(Z, \theta, \beta)$ is differentiable in β. Hence from assumption 4.1.(i), we can deduce that the stacked moment function $\rho(z, \alpha)$ is differentiable in α for almost all z. The SLLNs in (4.6) are useful to derive the probability limit of the score process of the GMM criterion function evaluated at the GMM shrinkage estimator.

Theorem 4.1 *Suppose that assumptions 3.1.(i)-(iii), 3.2.(i)-(iii) and 4.1 are satisfied and the first-step estimator $\widehat{\alpha}_n$ is $\sqrt{n}$-consistent. If the tuning parameter λ_n satisfies $\sqrt{n}\lambda_n = O(1)$ and $n^{\frac{1+\omega}{2}}\lambda_n \to \infty$, then*

$$\Pr\left(\left\| \widehat{\alpha}_{n,j}^{\mathcal{S}} \right\|_2 = 0 \right) \to 1 \tag{4.7}$$

for all $j \in \mathcal{S}_\alpha^c$.

By the consistency of $\widehat{\alpha}_n^{\mathcal{S}}$, we can deduce that

$$\Pr\left(\left\| \widehat{\alpha}_{n,j}^{\mathcal{S}} \right\|_2 \neq 0 \right) \to 1 \tag{4.8}$$

for all $j \in \mathcal{S}_\alpha$. Hence, the results in (4.7) and (4.8) imply that

$$\Pr\left(\mathcal{S}_{\alpha,n} = \mathcal{S}_\alpha\right) \to 1 \tag{4.9}$$

as $n \to \infty$, which gives the consistent grouped variable selection and moment selection.

Denote $\beta_{o,+} = \{\beta_{o,j} : j \in \mathcal{S}_\beta\}$ and let $(\widehat{\theta}^{\mathcal{S}}_{n,+}, \widehat{\beta}^{\mathcal{S}}_{n,+})$ be the GMM shrinkage estimator of $(\theta_{o,+}, \beta_{o,+})$. Define $M_{\mathcal{S}_\alpha} = \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}_\alpha}}\right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}_\alpha}}\right]$ and

$$\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}_\alpha}} = \begin{bmatrix} \frac{\partial E[g_{q+d_{\beta^-}}(Z,\theta_o)]}{\partial \theta'_{\mathcal{S}_\theta}} & 0 \\ \frac{\partial E[g_{d_{\beta^+}}(Z,\theta_o)]}{\partial \theta'_{\mathcal{S}_\theta}} & -I_{d_{\mathcal{S}_\beta}} \end{bmatrix},$$

where $g_{q+d_{\beta_-}}(Z,\theta_o) = \left[g'_q(Z,\theta_o), g'_{d_{\beta_-}}(Z,\theta_o)\right]'$, $g_{d_{\beta_-}}(Z,\theta_o)$ and $g_{d_{\beta_+}}(Z,\theta_o)$ are the correctly specified and misspecified moment conditions in set-2 respectively, $\theta_{\mathcal{S}_\theta} = \{\theta_j : j \in \mathcal{S}_\theta\}$, $d_{\mathcal{S}_\beta}$ is the cardinality of the index set $\mathcal{S}_\beta$ and $I_{d_{\mathcal{S}_\beta}}$ is a $d_{\mathcal{S}_\beta} \times d_{\mathcal{S}_\beta}$ identity matrix. We next derive the centered joint limiting distribution of $(\widehat{\theta}^{\mathcal{S}}_{n,+}, \widehat{\beta}^{\mathcal{S}}_{n,+})$.

Corollary 4.2 *Under the conditions of Theorem 4.1 and assumption 3.4, we have*

$$\sqrt{n}\left[(\widehat{\theta}^{\mathcal{S}}_{n,+}, \widehat{\beta}^{\mathcal{S}}_{n,+}) - (\theta_{o,+}, \beta_{o,+})\right] \to_d N\left(0, M^{-1}_{\mathcal{S}_\alpha} \Sigma_{\mathcal{S}_\alpha} M^{-1}_{\mathcal{S}_\alpha}\right),$$

where

$$\Sigma_{\mathcal{S}_\alpha} = \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}_\alpha}}\right] W_o E\left[\Psi(\theta_o, \beta_o)\Psi'(\theta_o, \beta_o)\right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}_\alpha}}\right].$$

The proof of this corollary is similar to that of Theorem 3.4 and thus is omitted. If we knew the true model and all correct moment conditions in set-2, then there would be $q + d_{\mathcal{S}^c_\beta}$ moment conditions to estimate $\theta_{o,+}$, where $d_{\mathcal{S}^c_\beta}$ is the cardinality of the index set $\mathcal{S}^c_\beta$. The moment conditions in set-1 and the valid moment conditions in set-2 can be stacked in the following way

$$m_e(\theta_{o,+}) = E\left[\rho_e(Z,\theta_{o,+})\right] = E\begin{bmatrix} g_q(Z,\theta_{o,+}) \\ g_{d_{\beta^-}}(Z,\theta_{o,+}) \end{bmatrix}_{(q+d_{\mathcal{S}^c_\beta})\times 1} = 0. \tag{4.10}$$

The semiparametric efficiency bound for $\theta_{o,+}$ is

$$\left(\Sigma_{+}^{*}\right)^{-1} = \left[\frac{\partial m_e(\theta_{o,+})}{\partial \theta_{+}}\right] \left\{E\left[\rho_e(Z,\theta_{o,+})\rho_e(Z,\theta_{o,+})'\right]\right\}^{-1} \left[\frac{\partial m_e(\theta_{o,+})}{\partial \theta_{+}'}\right]. \tag{4.11}$$

Next, we show that if the weight matrix W_n in the GMM shrinkage estimation satisfies (3.21), then the GMM shrinkage estimator $\widehat{\theta}_{n,+}^{\mathcal{S}}$ can asymptotically attain the semiparametric efficiency bound defined in (4.11).

Corollary 4.3 *Suppose that the assumptions in Theorem 4.2 are satisfied. If the weight matrix W_n satisfies (3.21), then we have*

$$\sqrt{n}\left(\widehat{\theta}_{n,+}^{\mathcal{S}} - \theta_{o,+}\right) \to_d N\left(0, \Sigma_{+}^{*}\right), \tag{4.12}$$

where Σ_{+}^{*} is defined in (4.11).

The proof of this corollary is similar to the proof of Theorem 3.23 and is omitted. The limiting distribution established in (4.12) is also a pointwise asymptotic result. The model selection and moment selection errors do not enter into this asymptotic distribution, because our consistent model/moment selection procedure implies the probability that these errors effect the limiting distribution of $\widehat{\theta}_{n,+}^{\mathcal{S}}$ goes to zero when sample size n goes to infinity. Hence in finite samples, the naive CIs constructed using (4.12) fail to take the model and moment selection errors into account and their coverage probabilities may be poor. In the extreme case, the GMM shrinkage estimator $\widehat{\theta}_{n,j}^{\mathcal{S}}$ of certain nonzero group $\theta_{o,j}$ $(j \in \mathcal{S}_\theta)$ may be shrunk to zero in finite samples. In that scenario, the naive CIs have the zero coverage probability. One should note that the PMS estimators based on other model/moment selection procedures also suffer from this problem. The treatment of the PMS inference is beyond the scope of this paper and we address this issue in a separate paper.

5 Extension II: GMM Shrinkage Estimation under Weak Identification

In this section, we study the properties of the GMM shrinkage estimator when the moment conditions in set-1 fail to strongly identify the unknown parameter θ_o. Specifically, we

assume that we have the following moment conditions in set-1

$$E\left[g_{n,q}(Z,\theta)\right] = n^{-\tau}G_{n,c}(\theta), \tag{5.1}$$

where $g_{n,q}(\cdot,\cdot) : R^{d_z} \times R^{d_\theta} \to R^q$, $G_{n,q}(\theta_o) = 0$ and $\tau \in [0, \frac{1}{2})$. At the same time, we have another set of possibly misspecified moment conditions

$$E\left[g_{n,k}(Z,\theta_o)\right] = G_k(\theta_o) \stackrel{?}{=} 0, \tag{5.2}$$

where $g_{n,k}(\cdot,\cdot) : R^{d_z} \times R^{d_\theta} \to R^k$. Hahn and Kuersteiner (2002) study a linear IV model where the moment conditions constructed from IVs are similar to these (5.1). They show that the IV estimators have the convergence rate $n^{\frac{1}{2}-\tau}$, if $\tau \in [0, \frac{1}{2})$. Caner (2009) shows that the GMM bridge estimator based on the moment condition (5.1) has the same rate.

In this section, we show that if there are strong and correctly specified moment conditions in (5.2), then these moment conditions can be consistently selected by the shrinkage method. More importantly, we show that the GMM shrinkage estimator $\widehat{\theta}_n^{\widehat{S}}$ has faster rate of convergence and hence better stochastic properties than the estimators only using the moment conditions in (5.1). The intuition is that when the correctly specified and strong moment conditions in (5.2) are consistently selected, the information contained in these moment conditions is included into estimating θ_o w.p.a.1.

Assumption 5.1 *(i) $G_k(\theta)$ is continuous in θ and there exists some continuous function $G_q(\theta)$ such that*

$$\sup_{\theta\in\Theta} \|G_{n,q}(\theta) - G_q(\theta)\|_E = o(1) \tag{5.3}$$

as $n \to \infty$; (ii) for any $\varepsilon > 0$, there exists some $\delta_\varepsilon > 0$ such that

$$\inf_{\{\theta\in\Theta:\|\theta-\theta_o\|_E\geq\varepsilon\}} \|G_q(\theta)\|_E > \delta_\varepsilon; \tag{5.4}$$

(iii) for $l = q, k$, the following FCLTs hold,

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^{n}\left\{g_{n,l}(Z_i,\theta) - E\left[g_{n,l}(Z_i,\theta)\right]\right\}\right| = O_p(n^{-\frac{1}{2}}); \tag{5.5}$$

(iv) $G_l(\theta)$ $(l = q, k)$ is continuously differentiable in the local neighborhood of θ_o and there

is

$$\sup_{\theta\in\Theta}\left\|\frac{\partial G_{n,q}(\theta)}{\partial\theta}-\frac{\partial G_q(\theta)}{\partial\theta}\right\|_E=o(1), \tag{5.6}$$

where $\frac{\partial G_q(\theta_o)}{\partial\theta'}$ *has full column rank; (v) the penalty function* $\widehat{P}_{\lambda_n}(\cdot)$ *is non-negative and satisfies* $n^{2\tau}\widehat{P}_{\lambda_n}(\beta_{o,j})=o_p(1)$ *for all* j*.*

Assumption 5.1.(i) imposes continuity condition on $G_k(\theta)$ and uniform limit function $G_q(\theta)$ of $G_{n,q}(\theta)$. The uniform approximation in (5.3) is a popular condition in the weak moment condition literature (e.g., Stock and Wright (2000)). Assumption 5.1.(ii) is a identifiable uniqueness condition of θ_o. Assumption 5.1.(iii) and (iv) are the generalized versions of assumption 3.1.(ii) and 3.2.(ii)-(iii) respectively. Compared with assumption 3.1.(iv), assumption 5.1.(v) imposes a stronger restriction on the tuning parameter λ_n. When the moment conditions in (5.1) are nearly weak, their information about θ_o is vanishing at the rate $n^{-\tau}$ and the information contained in GMM criterion function is vanishing at the rate $n^{1-2\tau}$, hence $\widehat{P}_{\lambda_n}(\beta_{o,j})$ must converge to zero faster than $n^{-2\tau}$ to ensure that $\widehat{\alpha}_n^{\mathcal{S}}$ is consistent.

Lemma 5.1 (Rate of Convergence) *Under assumptions 5.1, 3.1.(iii) and 3.2.(iv), we have*

$$\widehat{\alpha}_n^{\mathcal{S}}-\alpha_o=O_p\left(n^{2\tau}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|+n^{\tau-\frac{1}{2}}\right).$$

It's clear that if $n^{\frac{1}{2}+\tau}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|=O_p(1)$, then

$$\widehat{\theta}_n^{\mathcal{S}}-\theta_o=O_p(n^{\tau-\frac{1}{2}})$$

which gives the optimal convergence rate of the estimators based on the moment conditions in (5.1). However, we next show that if the tuning parameter λ_n converges to zero not very fast, then not only the shrinkage method can consistently select the correctly specified moment conditions in (5.2), but also the shrinkage estimator $\widehat{\theta}_n^{\mathcal{S}}$ has the rate of convergence faster than $n^{\tau-\frac{1}{2}}$.

Assumption 5.2 *(i) The penalty function* $\widehat{P}_{\lambda_n}(\cdot)$ *satisfies*

$$n^{\frac{1}{2}+\tau}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|=o_p(1); \tag{5.7}$$

(ii) there exists some sequence r_n such that

$$\liminf_{n\to\infty}\left[\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{r_n\lambda_n}\right] > 0 \ a.e. \tag{5.8}$$

for any $j \in \mathcal{S}^c_\beta$, and $n^{\frac{1}{2}-\tau}\lambda_n r_n \to \infty$.

Intuitively, assumption 5.2.(i) requires that the tuning parameter λ_n converge to zero fast enough such that the shrinkage bias converges to zero faster than the stochastic error does. However, assumption 5.2.(ii) requires that λ_n converge to zero slow enough such that the estimators of zero components in β_o are shrunk to zero w.p.a.1. Hence it is important to check if there is any penalty function which satisfies assumptions 5.1.(v) and 5.2.(i)-(ii) simultaneously.

Remark 5.1 *For the bridge penalty, assumptions 5.1.(v) and 5.2.(i) require $n^{\frac{1}{2}+\tau}\lambda_n = o(1)$. Under this condition and the assumptions of Lemma 5.1, we can deduce that $\widehat{\beta}^{\mathcal{S}}_{n,j} = O_p(n^{\tau-\frac{1}{2}})$ for all $j \in \mathcal{S}^c_\beta$. Note that*

$$\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{\lambda_n} = \gamma n^{(\frac{1}{2}-\tau)(1-\gamma)}\left|n^{\frac{1}{2}-\tau}\widehat{\beta}^{\mathcal{S}}_{n,j}\right|^{\gamma-1}.$$

Hence $r_n = n^{(\frac{1}{2}-\tau)(1-\gamma)}$ and to get

$$n^{\frac{1}{2}-\tau}\lambda_n r_n = n^{\frac{1}{2}+\tau}\lambda_n \times n^{(\frac{1}{2}-\tau)(1-\gamma)-2\tau} \to \infty$$

we need $\gamma < \frac{1-6\tau}{1-2\tau}$. It is clear that when $\tau \geq \frac{1}{3}$, then there is no such sequence λ_n which makes assumptions 5.1.(v) and 5.2.(i)-(ii) hold simultaneously. Secondly for the adaptive Lasso penalty, assumptions 5.1.(v) and 5.2.(i) require that $n^{\frac{1}{2}+\tau}\lambda_n = o(1)$. Note that

$$\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{\lambda_n} = n^{(\frac{1}{2}-\tau)\omega}\left|n^{\frac{1}{2}-\tau}\widehat{\beta}_{n,j}\right|^{-\omega}.$$

The first step estimator of $\beta_{o,j}$ ($j \in \mathcal{S}^c_\beta$) typically has the convergence rate $n^{\tau-\frac{1}{2}}$. Hence if we take $r_n = n^{(\frac{1}{2}-\tau)\omega}$ and $\omega > \frac{4\tau}{1-2\tau}$, then we can select λ_n such that

$$n^{\frac{1}{2}-\tau}\lambda_n r_n = n^{\frac{1}{2}+\tau}\lambda_n \times n^{\omega(\frac{1}{2}-\tau)-2\tau} \to \infty.$$

Finally, for the SCAD penalty, assumptions 5.1.(v) and 5.2.(i) require $n^{\frac{1}{2}+\tau}\lambda_n = o(1)$. *Under the assumptions of Lemma 5.1, we can deduce that* $\widehat{\beta}^{\mathcal{S}}_{n,j} = O_p(n^{\tau-\frac{1}{2}})$ *for all* $j \in \mathcal{S}^c_\beta$. *Note that*

$$\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{\lambda_n} = I\left(\widehat{\beta}^{\mathcal{S}}_{n,j} \leq \lambda_n\right) + \frac{(a\lambda_n - \widehat{\beta}^{\mathcal{S}}_{n,j})_+}{(a-1)\lambda_n} I\left(\widehat{\beta}^{\mathcal{S}}_{n,j} > \lambda_n\right).$$

Hence $r_n = 1$ *and to get*

$$n^{\frac{1}{2}-\tau}\lambda_n r_n = n^{\tau}\lambda_n \times n^{\frac{1}{2}-2\tau} \to \infty$$

we need $\tau < \frac{1}{4}$. *It is clear that when* $\tau > \frac{1}{4}$, *then there is no such sequence* λ_n *which makes assumptions 5.1.(v) and 5.2.(i)-(ii) hold simultaneously.*

Lemma 5.2 *Under the conditions of Lemma 5.1 and assumption 5.2, there is*

$$\Pr\left(\widehat{\beta}^{\mathcal{S}}_{n,j} = 0\right) \to 1, \tag{5.9}$$

for all $j \in \mathcal{S}^c_\beta$.

The sparsity of $\widehat{\beta}^{\mathcal{S}}_n$ implies that $\beta_{o,j}$ ($j \in \mathcal{S}^c_\beta$) is estimated as zero w.p.a.1. This result, combined with the following local identification assumption, enables us to improve the convergence rate of $\widehat{\theta}^{\mathcal{S}}_n$. Denote the potentially valid moment functions and misspecified moment functions in (5.2) to be $g_{d_{\beta_-}}(Z,\theta)$ and $g_{d_{\beta_+}}(Z,\theta)$ respectively.

Assumption 5.3 *Denote* $G_k(\theta) = \left[G'_{d_{\beta_-}}(\theta)\ G'_{d_{\beta_+}}(\theta)\right]'$, *then* $\frac{\partial G_{d_{\beta_+}}(\theta_o)}{\partial \theta'}$ *has full column rank.*

Assumption 5.3 is important for deriving the $\sqrt{n}$ convergence rate of $\widehat{\theta}^{\mathcal{S}}_n$. If this condition does not hold, then the moment conditions in set-1 are needed to achieve the local identification of θ_o. In that case, the convergence rate of $\widehat{\theta}^{\mathcal{S}}_n$ is not $\sqrt{n}$, but is still faster than the rate $n^{\frac{1}{2}-\tau}$.

Lemma 5.3 *Under the conditions of Lemma 5.2 and assumption 5.3, we have*

$$(\widehat{\theta}^{\mathcal{S}}_n, \widehat{\beta}^{\mathcal{S}}_{n,+}) - (\theta_o, \beta_{o,+}) = O_p(n^{-\frac{1}{2}}). \tag{5.10}$$

It is clear that from Lemma 5.3

$$\sqrt{n}\left[(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_{n,+}^{\mathcal{S}}) - (\theta_o, \beta_{o,+})\right] = O_p(1) \tag{5.11}$$

i.e. $\widehat{\theta}_n^{\mathcal{S}}$ has $\sqrt{n}$ convergence rate and $\sqrt{n}\left[(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_{n,+}^{\mathcal{S}}) - (\theta_o, \beta_{o,+})\right]$ is asymptotically tight. Denote $W_{o,kk}$ to be the right-lower $k \times k$ sub-matrix of W_o,

$$M_+ = \left[\frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_{o,kk} \left[\frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right] \text{ and } \frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'} = \begin{pmatrix} \frac{\partial G_{d_{\beta_-}}(\theta_o)}{\partial \theta'} & 0 \\ \frac{\partial G_{d_{\beta_+}}(\theta_o)}{\partial \theta'} & -I_{d_{\mathcal{S}_\beta}} \end{pmatrix}.$$

As $\frac{\partial G_{d_{\beta_-}}(\theta_o)}{\partial \theta'}$ has full column rank, so combined with the Assumption 3.1.(iii), we can deduce that the matrix M_+ is invertible.

Corollary 5.2 ($\sqrt{n}$-Normality) *Under the conditions of Lemma 5.3 and assumption 3.4, we have*

$$\sqrt{n}\left[(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_{n,+}^{\mathcal{S}}) - (\theta_o, \beta_{o,+})\right] \to_d N(0, M_+^{-1}\Sigma_+ M_+^{-1}), \tag{5.12}$$

where

$$\Sigma_+ = \left[0, \frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o E\left[\Psi(\theta_o, \beta_o)\Psi'(\theta_o, \beta_o)\right] W_o \left[0, \frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right]'$$

and $\Psi(\theta_o, \beta_o)$ is defined in (3.17).

Remark 5.3 *If the weight matrix W_n satisfies (3.21), then there is*

$$\Sigma_+ = \left[0, \frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o \left[0, \frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right]' = \left[\frac{\partial m_k(\alpha_o)}{\partial \alpha_{\mathcal{S}}}\right] W_{o,kk} \left[\frac{\partial m_k(\alpha_o)}{\partial \alpha_{\mathcal{S}}'}\right] = M_+.$$

So from Corollary 5.2, we can deduce that

$$\sqrt{n}\left[(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_{n,+}^{\mathcal{S}}) - (\theta_o, \beta_{o,+})\right] \to_d N(0, M_+^{-1}).$$

Note that we can similarly prove the efficiency result for the GMM shrinkage estimator $\widehat{\theta}_n^{\mathcal{S}}$, where the semiparametric efficiency bound of θ_o is implied by the potentially correct and strong moment conditions in (5.2).

6 Adaptive Selection of the Tuning Parameter

From the results of previous sections, we see that the tuning parameter λ_n plays an important role in deriving the oracle properties of the GMM shrinkage estimator. Assumptions 3.1.(iv), 3.2.(iv) and 3.3.(i)-(ii) are sufficient conditions imposed on λ_n for the oracle properties to hold. However, these conditions do not provide a straightforward mechanism for choosing the tuning parameter λ_n in finite samples. For practical implementation of the shrinkage techniques, it is important to have some procedures of selecting λ_n such that the GMM shrinkage estimator not only enjoys the oracle properties asymptotically, but also has good finite-sample properties.

In the LS shrinkage estimation Wang, Li, and Tsai (2007) propose to select the tuning parameter by minimizing the Bayesian information criterion (BIC). They show that the minimizer of BIC can consistently select the true model. They also find that the generalized cross validification (GCV) method proposed in Fan and Li (2001) to select λ_n is equivalent to the Akaike's information criterion (AIC) method. Hence the λ_n selected by GCV has non-trivial asymptotic probability of producing over-fitted models in the LS shrinkage estimation. More recently, Zhang, Li, and Tsai (2010) study the tuning parameter selection in PMLE with the SCAD penalty function. They propose to select the tuning parameter by minimizing a generalized information criterion (GIC), which includes the traditional AIC and BIC as special examples. Zhang, Li, and Tsai (2010) provide high level conditions to ensure consistent/conservative variable selection of PMLE based on the minimizer of GIC.

In this section, we extend the results in Wang, Li, and Tsai (2007) and Zhang, Li, and Tsai (2010) to the GMM framework with moment selection. For any given λ, denote the GMM shrinkage estimator of (θ_o, β_o) to be $(\widehat{\theta}_{\lambda,n}, \widehat{\beta}_{\lambda,n})$. We propose to select the tuning parameter by minimizing the GMM-type of information criterion, which is defined as

$$MSC_n(\lambda) = \Phi_n(\lambda) - \kappa_n h\left(\left|\mathcal{S}^c_{\beta,\lambda}\right|\right), \tag{6.1}$$

where $\Phi_n(\lambda)$ is some function depending on the tuning parameter λ, κ_n is a sequence of positive numbers, $\mathcal{S}_{\beta,\lambda}$ is the index set of nonzero elements in $\widehat{\beta}_{\lambda,n}$ and $\mathcal{S}^c_{\beta,\lambda}$ is the complement set of $\mathcal{S}_{\beta,\lambda}$, $\left|\mathcal{S}^c_{\beta,\lambda}\right|$ denotes the cardinality of the index set $\mathcal{S}^c_{\beta,\lambda}$ and it stands for the number of moment conditions selected by the GMM shrinkage method given λ, $h(\cdot)$ is some increasing function. The function $\Phi_n(\lambda)$ measures the fit of the model based on the moment conditions selected by the GMM shrinkage method given λ, while the term

$\kappa_n h\left(\left|\mathcal{S}^c_{\beta,\lambda}\right|\right)$ gives extra bonus to selecting more moment conditions.

Specific examples of the function $\Phi_n(\lambda)$ include the J-test statistic proposed in Andrews (1999), which is defined as

$$J_n(\lambda) = \left[n^{-\frac{1}{2}}\sum_{i=1}^{n} g_{q+s_\lambda}(Z_i, \widehat{\theta}^*_{\lambda,n})\right]' W_{q+s_\lambda,n}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n} g_{q+s_\lambda}(Z_i, \widehat{\theta}^*_{\lambda,n})\right], \tag{6.2}$$

where $g_{q+s_\lambda}(\cdot,\cdot) = [g'_q(\cdot,\cdot), g'_{s_\lambda}(\cdot,\cdot)]'$, $g_{s_\lambda}(\cdot,\cdot)$ denotes the moment conditions in set-2 selected by the GMM shrinkage estimation given λ, $W_{q+s_\lambda,n}$ is a $(q+s_\lambda)\times(q+s_\lambda)$ weight matrix, $\widehat{\theta}^*_{\lambda,n}$ is the GMM estimator based on the moment conditions implied by $g_{q+s_\lambda}(\cdot,\cdot)$. $\Phi_n(\lambda)$ can also be the generalized empirical likelihood (GEL) statistic proposed in Hong, Preston and Shum (1999), which is defined as

$$GEL_n(\lambda) = -2\min_{\alpha_{\mathcal{S}_{\beta,\lambda}}}\max_{\pi}\sum_{i=1}^{n}\nu\left[\pi'\rho(Z_i, \alpha_{\mathcal{S}_{\beta,\lambda}})\right], \tag{6.3}$$

where $\nu(\cdot)$ is some concave function and its domain contains 0, π is some $q+k$ dimensional vector, $\alpha_{\mathcal{S}_{\beta,\lambda}} = (\theta, \beta_{\mathcal{S}_{\beta,\lambda}}, 0)$ is transferred from α by setting the elements of β whose index belong to $\mathcal{S}^c_{\beta,\lambda}$ to be zero.

From the MSC in (6.1), the GMM-type of AIC, BIC and Hannan-Quinn information criterion (HQIC) are defined as

$$GMM - AIC_n(\lambda) = \Phi_n(\lambda) - 2\left|\mathcal{S}^c_{\beta,\lambda}\right| \tag{6.4}$$

and

$$GMM - BIC_n(\lambda) = \Phi_n(\lambda) - \left|\mathcal{S}^c_{\beta,\lambda}\right|\log n \tag{6.5}$$

and

$$GMM - HQ_n(\lambda) = \Phi_n(\lambda) - Q\left|\mathcal{S}^c_{\beta,\lambda}\right|\log\log n \tag{6.6}$$

where $Q > 2$.

Note that $\mathcal{S}^c_\beta$ denotes the index set of the zero components in β_o and it stands for the index set of the correctly specified moment conditions in set-2. We say an arbitrary set of moment conditions indexed by $\mathcal{S}^c_{\beta,\lambda}$ is under-selected (over-selected), if $\mathcal{S}^c_{\beta,\lambda} \subset \mathcal{S}^c_\beta$ ($\mathcal{S}^c_{\beta,\lambda} \supsetneq \mathcal{S}^c_\beta$). Let $\lambda_{\max,n}$ be some positive sequence such that $\lambda_{\max,n} \to 0$ as $n \to \infty$. Denote $\Omega_n = [0, \lambda_{\max,n}]$ to be a bounded interval in R^+, where the potential tuning parameter λ_n

is picked up. Given each sample with sample size n, we can divide the interval Ω_n into three subsets Ω_n^+, Ω_n^- and Ω_n^o, where $\Omega_n^+ = \{\lambda \in \Omega_n : \mathcal{S}_{\beta,\lambda}^c \subset \mathcal{S}_\beta^c\}$ is the set of λs which produce under-selected sets of moment conditions, $\Omega_n^o = \{\lambda \in \Omega_n : \mathcal{S}_\beta^c = \mathcal{S}_{\beta,\lambda}^c\}$ is the set of λs which produce the set of valid moment conditions and $\Omega_n^- = \{\lambda \in \Omega_n : \mathcal{S}_{\beta,\lambda}^c \subsetneq \mathcal{S}_\beta^c\}$ is the set of λs which produce over-selected sets of moment conditions in the GMM shrinkage estimation. Denote the collection of all possible subsets of moment conditions in set-2, the collection of all possible under-selected subsets of moment conditions in set-2 and the collection of all possible over-selected subsets of moment conditions in set-2 to be $\mathcal{S}_F$, $\mathcal{S}_U$ and $\mathcal{S}_O$ respectively. As the function $\Phi_n(\lambda)$ is actually defined on different subsets of moment conditions in set-2, we can use this fact to define $\Phi_n^*(\cdot)$: $\mathcal{S}_F \to R$ such that $\Phi_n^*(\mathcal{S}_{\beta,\lambda}^c) = \Phi_n(\lambda)$ for all $\lambda \in \Omega_n$.

Assumption 6.1 *(i) $h(\cdot)$ is some strictly increasing function; (ii) for any $\mathcal{S}_j \in \mathcal{S}_F$, there exists some constant $c_j \geq 0$ such that*

$$\frac{\Phi_n^*(\mathcal{S}_j)}{n} \to_p c_j; \tag{6.7}$$

(iii) for any $\mathcal{S}_j \in \mathcal{S}_U$ and $\mathcal{S}_k \in \mathcal{S}_F \backslash \mathcal{S}_U$, there is $c_j > c_k$; (iv) for any $\mathcal{S}_m \in \mathcal{S}_O$, there is

$$2\left[\Phi_n^*(\mathcal{S}_\beta^c) - \Phi_n^*(\mathcal{S}_m)\right] \to_d \chi^2\left(\left|\mathcal{S}_\beta^c\right| - |\mathcal{S}_m|\right), \tag{6.8}$$

where $|\mathcal{S}_m|$ denotes the dimensionality of the set $\mathcal{S}_m$.

Assumption 6.1.(i) is a regularity condition which comes directly from Andrews (1999). Assumption 6.1.(ii) can be verified using ULLN. For example, when the functions $\Phi_n^*(\cdot)$ and $\Phi_n(\cdot)$ take the form of J-test statistic, (6.7) is implied by the Assumption 1.(c) in Andrews (1999). Assumption 6.1.(iii) is implied by the identification conditions of θ_o and β_o. Assumption 6.1.(iv) is an asymptotic Chi-square approximation of the likelihood ratio type of test statistic. When the functions $\Phi_n^*(\cdot)$ and $\Phi_n(\cdot)$ take the form of the J-test statistic, (6.8) can be verified using similar arguments in deriving the limiting distribution of difference-in test statistic under the null hypothesis.

Let λ_n^o denote a positive sequence which satisfies assumptions 3.1.(iv), 3.2.(iv) and 3.3.

Lemma 6.1 *Suppose assumptions 3.1.(i)–(iii), 3.2.(i)–(iii) and 6.1 are satisfied. If the*

sequence κ_n *satisfies* $\kappa_n \to \infty$ *and* $\kappa_n = o(n)$*, then we have*

$$\Pr\left(\inf_{\lambda \in \Omega_n^+ \cup \Omega_n^-} MSC_n(\lambda) > MSC_n(\lambda_n^o)\right) \to 1 \tag{6.9}$$

and

$$\Pr\left(MSC_n(\lambda_n^o) = MSC_n(\lambda^o),\ \forall \lambda^o \in \Omega_n^o\right) \to 1. \tag{6.10}$$

Lemma 6.1 implies that the λs which produce the over-selected or under-selected sets of moment conditions in GMM shrinkage estimation, fail to minimize $MSC_n(\lambda)$ w.p.a.1. Hence, the minimizer of $MSC_n(\lambda)$ could only be the one which consistently selects all potentially valid moment conditions in set-2. From the proof of Lemma 6.1, we see that the conditions $\kappa_n \to \infty$ and $\kappa_n = o(n)$ are important for showing the minimizer of $MSC_n(\lambda)$ will not produce under-selected sets of moment conditions asymptotically. In the next lemma, we show that if κ_n is bounded from above by some positive and finite constant M, then there will be some non-zero probability such that the minimizer of $MSC_n(\lambda)$ produces under-selected sets of moment conditions asymptotically.

Lemma 6.2 *Suppose Assumption 3.1.(i)–(iii), 3.2.(i)–(iii) and 6.1 are satisfied. If there exists some finite and positive constant* M *such that* $\kappa_n \leq M$ *for all* n*, then we have*

$$\Pr\left(\inf_{\lambda \in \Omega_n^-} MSC_n(\lambda) > MSC_n(\lambda_n^o)\right) \to 1 \tag{6.11}$$

and

$$\Pr\left(\inf_{\lambda \in \Omega_n^o} MSC_n(\lambda) > MSC_n(0)\right) \geq \pi, \tag{6.12}$$

where π *is some constant in* $(0, 1)$*.*

If κ_n is bounded from above, Lemma 6.2 implies that the λs which produce over-selected sets of moment conditions, fail to minimize the function $MSC_n(\lambda)$ w.p.a.1. However, the minimizer of $MSC_n(\lambda)$ has nonzero probability of producing under-selected sets of moment conditions asymptotically. In the GMM-type of BIC and HQIC, κ_n equals to $\log n$ and $\log(\log n)$ respectively. Hence by Lemma 6.1, the minimizers of these two information criterion can consistently select the potentially valid moment conditions in set-2. However, in the GMM-type of AIC, κ_n equals to 2. By Lemma 6.2, we can deduce that the minimizer of GMM-AIC has non-trivial probability of producing under-selected sets of moment

conditions, even though the probability of selecting over-selected sets of moment conditions goes to zero.

Let $\widehat{\lambda}_n$ denote the minimizer of $MSC_n(\lambda)$, i.e.

$$\widehat{\lambda}_n = \underset{\lambda \in \Omega_n}{\arg\min} \left\{ \Phi_n(\lambda) - \kappa_n h\left(\left| \mathcal{S}^c_{\beta,\lambda} \right| \right) \right\}. \tag{6.13}$$

Combining the results in Lemma 6.1 and Lemma 6.2, we get the following theorem.

Theorem 6.1 *Suppose assumptions 3.1.(i)–(iii), 3.2.(i)–(iii) and 6.1 are satisfied. (i) If there exists some finite and positive constant M such that $\kappa_n \leq M$ for all n, then $\widehat{\lambda}_n$ satisfies*

$$\Pr\left(\widehat{\lambda}_n \in \Omega_n^- \right) \to 0 \text{ and } \Pr\left(\widehat{\lambda}_n \in \Omega_n^+ \right) \geq \pi, \tag{6.14}$$

where π is some constant in $(0,1)$; (ii) if the sequence κ_n satisfies $\kappa_n \to \infty$ and $\kappa_n = o(n)$, then $\widehat{\lambda}_n$ satisfies

$$\Pr\left(\mathcal{S}^c_{\beta,\widehat{\lambda}_n} = \mathcal{S}_\beta \right) \to 1, \tag{6.15}$$

where $\mathcal{S}^c_{\beta,\widehat{\lambda}_n}$ denotes the index set of the zero elements in the GMM shrinkage estimator $\widehat{\beta}_{\widehat{\lambda}_n,n}$ given the tuning parameter $\widehat{\lambda}_n$.

From (6.15), we see that if $\kappa_n \to \infty$ and $\kappa_n = o(n)$, then the GMM shrinkage estimation procedure based on $\widehat{\lambda}_n$ can consistently select the potentially valid moment conditions in set-2. Hence, if assumption 3.4 is satisfied and the weight matrix W_n satisfies (3.21), we can deduce that the GMM shrinkage estimator of θ_o based on $\widehat{\lambda}_n$ is asymptotically efficient.

7 Monte Carlo Simulations

There are two different simulation experiments conducted in this section. In the first experiment, we apply the GMM shrinkage method to a linear IV model to select the IVs, where the set-1 has only one IV and the set-2 contains ten potential IVs. This simple specification makes it easy to study the finite sample properties of the GMM shrinkage estimation in different scenarios. For example, we can see how differently the shrinkage method behaves in terms of moment selection and efficient estimation, when the set-1 moment condition can strongly or weakly identify the unknown parameter θ_o. In the second simulation experiment, we use our method to perform variable selection and grouped moment selection in the dynamic panel model studied in Andrews and Lu (2001).

7.1 Monte Carlo Experiment I

In this simulation study, the data are generated from the following linear model,

$$Y_i = \theta_{1o} + \theta_{2o} X_i + u_i, \tag{7.1}$$

where

$$u_i \sim N(0, \sigma_u^2),\ X_i \sim N(0, \sigma_x^2)\ and\ E[X_i u_i] \neq 0 \tag{7.2}$$

for all i. The available IVs are $(Z_{1,i}, Z_{2,i})$, where $Z_{1,i}$ is a scale random variable and $Z_{2,i} = (Z_{21,i}, Z_{22,i})$ is a random vector. There are two elements in $Z_{21,i}$ which denote the potentially valid IVs and there are eight elements in $Z_{22,i}$ which are misspecified IVs.

In the equation (7.1), we take $(\theta_{1o}, \theta_{2o}) = (0.8, 0.8)$. The random variables X_i, $Z_{1,i}$, $Z_{21,i}$, $Z^*_{22,i}$ and u_i are generated from the following joint normal distribution

$$(X_i, Z_{1,i}, Z_{21,i}, u_i, Z^*_{22,i})' \sim N(0, \Sigma), \tag{7.3}$$

where the diagonal elements of Σ are 1, $E(X_i Z_{1,i}) = \sigma_{z_1 x}$, $E(X_i Z_{21,i}) = (\sigma_{z_2 x}, \sigma_{z_2 x})$, $E(X_i u_i) = 0.4$ and all other elements in Σ are zero. $Z_{22,i}$ is generated by the following equation

$$Z_{22,i} = Z^*_{22,i} + 0.5 u_i * l, \tag{7.4}$$

where l is a 1×8 vector of ones. The correlation $\sigma_{z_j x}$ of X_i and $Z_{j,i}$ $(j = 1, 2)$ measures the signal strength of the IV $Z_{j,i}$ about the endogenous variable X_i. There are two specifications of $(\sigma_{z_1 x}, \sigma_{z_2 x})$ used in the simulation, i.e. $(\sigma_{z_1 x}, \sigma_{z_2 x}) = (0.4, 0.4)$ and $(\sigma_{z_1 x}, \sigma_{z_2 x}) = (0.1, 0.3)$ respectively. In the first case, $Z_{1,i}$ has strong information about X_i, while in the second one, $Z_{1,i}$ is relatively weak.

We assume the econometrician knows that $Z_{1,i}$ is a valid IV, while being unsure about the validity of the IVs in $Z_{2,i}$. Hence the moment conditions in set-1 are

$$E[(Y_i - \theta_{1o} - \theta_{2o} X_i)] = 0 \text{ and } E[(Y_i - \theta_{1o} - \theta_{2o} X_i) Z_{1,i}] = 0, \tag{7.5}$$

while the moment conditions in set-2 are

$$E\left[(Y_i - \theta_{1o} - \theta_{2o} X_i) Z'_{2,i}\right] \stackrel{?}{=} 0. \tag{7.6}$$

The adaptive Lasso penalty is used in the GMM shrinkage estimation, where the first step

estimators of the moment selection coefficients are from the GMM estimation using the moment conditions in (7.5) and the reparametrized moment conditions in (7.6).

For each specification of $(\sigma_{z_1x}, \sigma_{z_2x})$, we use the simulated samples with sample sizes $n = 100$ and $n = 500$ respectively in our the simulation study and for each sample size, 1000 simulated samples are drawn from the data generating mechanism. With each simulated sample, we calculate four different types of estimators, which include the oracle estimator, GMM estimator, GMM shrinkage estimator using λ_n selected by GMM-AIC and GMM shrinkage estimator using λ_n selected by GMM-BIC[7]. The oracle estimator is a GMM estimator based on the moment conditions in set-1 and all valid moment conditions in set-2. The GMM estimator is a GMM estimator based only on the moment conditions in the set-1. Given the specification of $(\sigma_{z_1x}, \sigma_{z_2x})$ and the sample size n, we can get 1000 estimators of $(\theta_{1,o}, \theta_{2,o})$ for each type of estimator using the 1000 simulated sample. Hence, we can estimate the finite sample marginal densities of different estimators for $(\theta_{1,o}, \theta_{2,o})$.

There are several remarks we can make based on the simulation results presented in Figure 1 and 2. First, when the signal strength of the moment conditions in set-1 is relatively strong, i.e. $E[X_iZ_{1,i}] = 0.4$, the GMM shrinkage method selects all valid moment conditions in set-2 with high probability and selects the over-selected sets of moment conditions with low probability. However, if the moment conditions in set-1 are relatively weak, i.e. $E[X_iZ_{1,i}] = 0.1$, then the GMM shrinkage method has non-trivial probability of selecting the over-selected sets of moment conditions in set-2, especially when the sample size is small (e.g., $n = 100$). Second, when the sample size n is increased from 100 to 500, the probability of selecting the set of valid moment conditions in set-2 increases greatly and the probability of selecting the over-selected or under-selected sets of moment conditions decrease sharply. Third, if we compare the GMM shrinkage estimators based on different data-driven procedures of selecting the tuning parameter, we see that the GMM shrinkage estimation using the tuning parameters from GMM-AIC has lower probability of selecting inconsistent sets of moment conditions. But it has non-trivial probability of selecting under-selected sets of moment conditions, even when the sample size is increased from 100 to 500. On the other hand, the GMM shrinkage estimation using the tuning parameters from GMM-BIC has lower probability of selecting the over-selected sets of moment conditions and higher probability of selecting the set of correct moment conditions, but its probability of selecting the over-selected sets of moment conditions is higher. Fourth, the

[7] In the GMM-type of information criterion, the J-test statistic is used to construct the general function $\Phi_n(\lambda)$.

finite sample densities of the GMM shrinkage estimators behave much better than these of the GMM estimators in all scenarios of this simulation study. Comparing the GMM shrinkage estimator with the GMM estimator, the most obvious improvement is the reduction of the variance, as we can see from the finite sample densities depicted in Figure 1 and 2. Also note that when the sample size is increased, the finite sample densities of the GMM shrinkage estimators are approaching to these of the oracle GMM estimators. Fifth, bimodality shows up in the finite sample densities of the GMM shrinkage estimators, as we can see in Figure 1.3. Bimodality appears because in finite samples, our moment selection method has non-zero probability of selecting the misspecified moment conditions and in this scenario, the GMM shrinkage estimator has large bias due to the inconsistency. The bimodality in the finite sample densities of the GMM shrinkage estimators causes challenge for conducting stochastic inferences based on the selected moment conditions. The pointwise asymptotic distribution of the GMM shrinkage estimator does not capture the errors of moment selection, hence in finite samples, the CIs constructed using this limiting distribution may have low coverage probability and the hypothesis tests based on these CIs may have incorrect size. Finally, when the moment conditions in (7.5) are relatively strong, i.e. $E\left[X_i Z_{1,i}\right] = 0.4$, the densities of the GMM shrinkage estimators $\widehat{\theta}_{1,n}$ of $\theta_{1,o}$ are almost the same as these of the oracle estimators and the GMM estimators. This is because the moment conditions in set-2 only contain the information about $\theta_{2,o}$. Hence when $\theta_{2,o}$ could be reliably estimated using the set-1 moment conditions, the extra valid moment conditions in set-2 do not help to reduce the variances of the estimators of $\theta_{1,o}$. However, if the moment conditions in set-1 are relatively weak, i.e. $E\left[X_i Z_{1,i}\right] = 0.1$, then the valid moment conditions in set-2 can help to estimate $\theta_{2,o}$ more accurately. Hence in this case, the variances of the GMM shrinkage estimators of $\theta_{1,o}$ are also reduced, even though the valid moment conditions in set-2 do not directly help to improve the properties of the estimators for $\theta_{1,o}$.

7.2 Monte Carlo Experiment II

In this simulated study, the data are generated from the following dynamic panel model

$$y_{i,t} = \theta_{1,o} + \theta_{2,o} y_{i,t-1} + \theta_{3,o} x_{i,t} + u_{i,t}, \tag{7.7}$$

where $u_{i,t} = \eta_i + v_{i,t}$. For each i and all t, the regressor $x_{i,t}$, time varying error term $v_{i,t}$ and individual effect η_i have the following joint normal distribution

$$(x_{i,1}, ..., x_{i,T}, \eta_i, v_{i,1}, ..., v_{i,T})' \sim N(0, \Sigma), \tag{7.8}$$

where

$$\Sigma = \begin{pmatrix} \mathbf{I}_T & \sigma_{x\eta}\mathbf{1}_T & \sigma_{xv}\Gamma \\ \sigma_{x\eta}\mathbf{1}_T' & 1 & \mathbf{0}_T' \\ \sigma_{xv}\Gamma' & \mathbf{0}_T & \mathbf{I}_T \end{pmatrix}, \tag{7.9}$$

$\mathbf{I}_T$ denotes a $T \times T$ identity matrix, $\mathbf{1}_T$ denotes a $T \times 1$ vector of ones, $\mathbf{0}_T$ denotes a $T \times 1$ vector of zeros, Γ is a $T \times T$ matrix whose (j,k)-th element is one if $k = j-1$ for $j = 2, ..., T$ and zero otherwise. In the matrix defined in (7.9), we take $\sigma_{x\eta} \neq 0$ and $\sigma_{xv} \neq 0$. The L initial observations are generated by

$$\begin{aligned} y_{i,s} &= \theta_{1,o} + \theta_{2,o} y_{i,s-1} + \theta_{3,o} x_{i,s} + u_{i,s}, s = 2 - L, ..., 0, \\ y_{i,1-L} &= \frac{\theta_{1,o}}{1 - \theta_{2,o}} + \frac{\theta_{3,o}\sigma_{x\eta} + \sigma_\eta^2}{\sigma_\eta^2(1 - \theta_{2,o})}(\eta_i + v_{i,1-L}). \end{aligned} \tag{7.10}$$

From the data generating mechanism (7.7)-(7.10), it is clear that the time varying error term $v_{i,t}$ is serially uncorrelated and is uncorrelated with the individual effect η_i. The control variable $x_{i,t}$ is also serially uncorrelated and has constant variance $\sigma_x^2 = 1$. It is correlated with the individual effect as $\sigma_{x\eta} \neq 0$ and is pre-determined as $E[x_{i,t} v_{i,s}] = 0$ for $s = t+1, ..., T$, but it is not strictly exogenous with respect to $v_{i,t}$ because $E[x_{i,t} v_{i,t-1}] \neq 0$.

The true model is parametrized as

$$(\theta_{1,o}, \theta_{2,o}, \theta_{3,o}) = (0.8, 0.85, 0.5) \text{ and } (\sigma_{x\eta}, \sigma_{xv}) = (-0.2, 0.5). \tag{7.11}$$

As discussed in Andrews and Lu (2001), this parametrization has two important features. First, there is non-trivial difference in the efficiency gain between the GMM estimator which uses the correct model and all correct moment conditions and the GMM estimator which uses over-fitted models and the correct moment conditions known to the econometrician. Second, the bias in the GMM estimator is noticeable, when an under-fitted model is used or a misspecified moment condition is included into estimation.

Suppose the econometrician specifies the following model

$$\begin{aligned} y_{i,t} &= \theta_1 + \theta_2 y_{i,t-1} + \theta_4 y_{i,t-2} + \theta_3 x_{i,t} + u_{i,t}, \\ u_{i,t} &= \eta_i + v_{i,t}, \end{aligned} \tag{7.12}$$

and he/she has the data $\{x_{i,t}, y_{i,t}\}_{1 \leq i \leq N, -1 \leq t \leq T}$. Following Andrews and Lu (2001), we assume that the econometrician correctly specify the following assumption.

Assumption 7.1 *(i)* $E[\eta_i] = 0$, $E[v_{i,t}] = 0$ *and* $E[v_{i,t}\eta_i] = 0$ *for all* $i = 1, ..., N$ *and* $t = 1, ..., T$*; (ii)* $v_{i,t}$ *is serially uncorrelated for all* $i = 1, ..., N$ *and* $Var[v_{i,t}] = \sigma_i^2 > 0$ *for all* $t = 1, ..., T$*; (iii)* $E[v_{i,t}y_{i,0}] = E[v_{i,t}y_{i,-1}] = 0$ *for all* $t = 1, ..., T$*; (iv)* $E[v_{i,t}(x_{i,1}, ..., x_{i,t})] = 0$ *for all* $t = 1, ..., T$.

As discussed in Andrews and Lu (2001), assumption 7.1 imposes the standard error-component structure, constant variance for $v_{i,t}$, and predeterminedness for $x_{i,t}$. Assumption 7.1 implies the following moment conditions

$$E[(u_{i,1}, ..., u_{i,T})] = \mathbf{0}_T', \tag{7.13}$$

$$E[(y_{i,1-L}, ..., y_{i,t-2})\Delta u_{i,t}] = \mathbf{0}_t', \ \forall t = 2, ..., T \tag{7.14}$$

$$E[y_{i,t-1}\Delta u_{i,t} - y_{i,t}\Delta u_{i,t+1}] = 0, \ \forall t = 2, ..., T-1 \tag{7.15}$$

$$E[(x_{i,1}, ..., x_{i,t})\Delta u_{i,t}] = \mathbf{0}_t', \ \forall t = 2, ..., T \tag{7.16}$$

which constitute the set-1 moment conditions in our terminology. Next, suppose that the econometrician do not know whether the following assumptions are correctly specified or not.

Assumption 7.2 *(i)* $E[v_{i,t}(x_{i,t+1}, ..., x_{i,T})] = 0$ *for all* i*; (ii)* $E[\eta_i x_{i,t}] = 0$ *for all* i *and* t*; (iii)* $E[\eta_i y_{i,1}] = E[\eta_i y_{i,0}] = E[\eta_i y_{i,-1}]$ *for all* i.

As assumption 7.1.(iv) implies that $x_{i,t}$ is predetermined, under this condition, assumption 7.2.(i) requires that $x_{i,t}$ is strictly exogenous. Assumption 7.2.(i) assumes that $x_{i,t}$ is uncorrelated with η_i. Assumption 7.2.(iii) is the stationarity assumption. Assumptions

7.2.(i)-(iii), combined with assumption 7.1, imply the following moment conditions

$$E[(x_{i,1}, ..., x_{i,T})\Delta u_{i,t}] = 0, \ \forall t = 2, ..., T \tag{7.17}$$

$$E[(u_{i,1} + ... + u_{i,T})x_{i,t}] = 0, \ \forall t = 1, ..., T \tag{7.18}$$

$$E[(u_{i,2} + ... + u_{i,T})\Delta y_{i,t}] = 0, \ \forall t = 0, 1 \tag{7.19}$$

respectively, which constitute the set-2 moment conditions in our terminology. From the data generating mechanism, we know the moment conditions in (7.19) are correctly specified, while these in (7.17) and (7.18) are misspecified.

There are two different sample sizes used in the simulation study

$$(T, N) = (3, 250) \text{ or } (3, 500)$$

and we generate 1000 simulated samples in each case. We call the most parsimonious model with all correct moment conditions as the true model, the models with moment conditions which produce consistent estimators of the structural coefficients as consistent models and the the models with moment conditions which produce inconsistent estimators as inconsistent models. The simulation results are contained in Table 1 to Table 3. As a comparison, we also include the simulation results based on the subset selection procedures proposed in Andrews and Lu (2001). Table 1 provides the selection probabilities of the true model, consistent models and inconsistent models based on different model/moment selection procedures. Table 2 and 3 present the finite sample properties of the Oracle GMM estimators, GMM estimators, GMM shrinkage estimators and post moment selection GMM estimators with the sample size n being 250 and 500 respectively.

The results of this simulation study are summarized in the following. First, the GMM shrinkage method selects the true model with the set of valid moment conditions as well as the subset model/moment selection methods proposed in Andrews and Lu (2001). Our method has slightly higher probability of selecting the consistent models, but it has tighter control of selecting the inconsistent models. When the sample size is small, i.e. $n = 250$, the GMM shrinkage method based on the tuning parameters from GMM-BIC performs even better in terms of selecting the true model. Secondly, to check the finite sample properties of various estimators, we compute the finite sample biases, standard deviations and the root of mean square errors in Table 2 and Table3. Compared with the GMM estimators, the GMM shrinkage estimators of $(\theta_{1,o}, \theta_{2,o})$ have much smaller standard errors and the

GMM shrinkage estimators of $\theta_{2,o}$ have even smaller finite sample biases. However, the GMM estimators of $\theta_{3,o}$ beat the GMM shrinkage estimators in terms of the finite sample biases and standard errors. The main reason is the valid moment conditions in set-2 only contain information about $\theta_{2,o}$ and at the same time, the GMM shrinkage estimators of $\theta_{3,o}$ are suffered from the shrinkage bias. Thirdly, the finite sample biases of the GMM shrinkage estimators may be large because they contain not only the stochastic errors, but also the shrinkage bias. To get rid of this shrinkage bias, we estimate the structural coefficients based on the model and moment conditions selected by the shrinkage method. We call those estimators as PGMM estimators. As we can see from the right lower panels of Table 2 and Table 3, the PGMM estimators have almost the same standard deviations as these of the GMM shrinkage estimators, but they have much smaller finite sample biases. Finally, when the sample size is increased from 250 to 500, the finite sample properties of the GMM shrinkage estimators and the PGMM estimators are approaching to these of the Oracle GMM estimators. That is because the probability of the true model with the set of valid moment conditions being selected is improved greatly.

8 An Empirical Example

In this section, we apply the GMM shrinkage method to the life-cycle labor supply model studied in MaCurdy (1981) and Altonji (1986). Both papers estimate the following labor supply equation

$$\Delta \log(h_{i,t}) = \alpha_t + \Delta \log(w_{i,t})\delta_o + \varepsilon_{i,t}, \tag{8.1}$$

where $h_{i,t}$ is the annual hours working for money and $w_{i,t}$ is the hourly wage rate of individual i at period t, α_t is a time varying constant and $\varepsilon_{i,t}$ is the time varying error term. As discussed in MaCurdy (1981), the coefficient δ_o measures the intertemporal substitution elasticity of labor supply with respect to the evolutionary wage changes and the theoretical prediction for its sign is positive.

Due to the measurement errors in $w_{i,t}$, the OLS estimator of (8.1) may be inconsistent. MaCurdy (1981) proposes to use the family background variables (father's education, mother's education and parents' economic status when individual i was young), education, age and the interaction between education and age of individual i as IVs for $\Delta \log(w_{i,t})$. However, Altonji (1986) argues that the family background variables and education may only contain weak information about $\Delta \log(w_{i,t})$ and the age of individual i may not even

be a valid IV. Altonji (1986) proposes to use an alternative measure $w^*_{i,t}$ of wage to construct IV for $\Delta \log(w_{i,t})$. However, for $\Delta \log(w^*_{i,t})$ being a valid IV for $\Delta \log(w_{i,t})$, one need to impose the strong assumption that the measurement errors in $w_{i,t}$ and $w^*_{i,t}$ are independent. In MaCurdy (1981) and Altonji (1986), $w_{i,t}$ is constructed by annual labor income of individual i divided by the product of annual labor supply and GNP price deflator. In Altonji (1986), $w^*_{i,t}$ is the hourly wage rate of individual i if this person is paid based on hours.

Instead of using all IVs in MaCurdy (1981) to construct the set-1 moment conditions, we only use the parents' economic status as the credibly valid IV and include the rest of them into set-2. We also include the alternative measure $w^*_{i,t}$ of wage in Altonji (1986) and the wage $w_{i,t}$ itself into set-2. This specification enables us to answer following three questions. First, are the other IVs in MaCurdy (1981), especially the age of individual i, valid for $\Delta \log(w_{i,t})$? If they are not, then the results of MaCurdy (1981) may suffer from bias incurred not only by the weak moment conditions but also by the misspecified moment conditions. Second, is the IV $\Delta \log(w^*_{i,t})$ constructed by the alternative measure $w^*_{i,t}$ of wage valid for $\Delta \log(w_{i,t})$? If it is not, then the results of Altonji (1986) may be inconsistent. Third, is there measurement error in $\Delta \log(w_{i,t})$ which causes it to be an endogenous variable? If $\Delta \log(w_{i,t})$ is endogenous, then the OLS estimator of δ_o is inconsistent. On the other hand, if it is exogenous then OLS estimator is not only consistent but also more efficient.

Our sample is constructed from the Michigan Panel Study of Income Dynamics (PSID) data set from year 1970 to year 1981. The sample is selected according to the following criterion. First, it is limited to men with stable marriage status for the years 1970-1981. Second, individuals below age 25 in 1970 or above age 60 in 1981 are excluded to minimize the complication incurred by schooling and retirement. Third, the observations in certain year are excluded if the data are missing for the variables used in estimation[8].

Table 4 presents the GMM estimators of δ_o based on the moment conditions constructed by $w_{i,t}$, IVs used in MaCurdy (1981) and IV used in Altonji (1986) respectively. The results in Table 4 can be summarized as follows. First, the GMM estimation using moment conditions constructed by $w_{i,t}$ gives highly misleading results, because its estimators of δ_o

[8] Following the criterion used in Altonji (1986), the imputed wage $w_{i,t}$ was treated as missing if the wage measures increased by 250 percent or more than \$13 or fell by 60 percent or more than \$13 from one year to another. They were also treated as missing if the real wage was less than \$0.40 in 1972 dollars. The same criterion was applied to $w^*_{i,t}$. The 250 percent, 60 percent limits were also used for labor supply. In addition, the labor supply variable was treated as missing if annual hours exceeded 4,860.

are negative and very significant. Second, the GMM estimation using the IVs in MaCurdy (1981) provides reasonable results, as the sign of its estimators are positive. But the estimators have large standard errors, which indicates that these IVs only contain weak information about the endogenous variable $w_{i,t}$. Third, the GMM estimators using Altonji (1986)'s IV are reasonable and have smaller standard errors.

We next use the GMM shrinkage method to estimate the equation in (8.1). The estimators of the moment selection coefficients are included in Table 6. As a comparison, we also include the GMM estimators of the moment selection coefficients in different specifications of α_t in Table 6. In the first two rows of Table 6, the constant term α_t in the equation (8.1) is treated to be time variant, while in its last two rows, α_t is taken to be a time invariant constant. From Table 6, we see that the GMM estimators of the moment selection coefficients are nonzero and it is hard to determine which moment conditions are valid (misspecified) based on these estimators. On the other hand, the GMM shrinkage estimation gives the same moment selection result in the different specifications of α_t. The moment conditions constructed from the IVs in MaCurdy (1981) and Altonji (1986) are picked up by our shrinkage method. While the moment condition constructed using the imputed wage $w_{i,t}$ is not selected, which implies that $\Delta \log(w_{i,t})$ is an endogenous variable in the labor supply equation (8.1).

The results the GMM shrinkage estimation of the labor supply equation (8.1) are contained in Table 5. As a comparison, we also include the GMM estimators of δ_o based on the moment condition in set-1 and the post moment selection GMM (PGMM) estimators of δ_o in Table 5. Columns (1)-(2) of Table 5 present the GMM estimators of δ_o based on the following IV: parent's economic status when individual was young, which provides the moment condition in set-1. Compared with other estimators in Table 5, the GMM estimators in columns (1)-(2) not only are larger in magnitude, but also have larger standard errors. On the other hand, the GMM shrinkage estimators in Columns (3)-(4) have much smaller standard errors, because some moment conditions in set-2 are selected and automatically included into estimation by the GMM shrinkage method. From Table 4, we have already seen that δ_o is downward biasedly estimated in the GMM estimation using $\Delta \log(w_{i,t})$ as an IV. Hence if part of the information in the moment condition constructed by $w_{i,t}$ is used in estimation, the result estimator will be suffered from this downward bias. >From Table 6, we see that compared with the GMM estimators, the GMM shrinkage estimators of $\beta_{7,o}$ is closer to zero, which implies that part of the information in the moment condition constructed by $w_{i,t}$ is indeed used in the GMM shrinkage estimation. Based on

above reasoning, we can deduce that the shrinkage effect of the penalty function on the estimators of $\beta_{7,o}$ may introduce some bias to the estimator of δ_o. To get rid of this bias, we conduct another GMM estimation based on the moment condition in set-1 and the moment conditions in set-2 selected by our method. These PGMM estimators are included in columns (5)-(6) of Table 5. We can see that the PGMM estimators are slightly larger in magnitude than the GMM shrinkage estimators and their standard errors are almost the same.

We summarize our findings in this empirical example as follows. First, our method selects the IVs used in MaCurdy (1981), which relives the concern in Altonji (1986) that age and education may be invalid IVs. Second, our method picks up the IV used in Altonji (1986) and hence confirms the validity of $\Delta \log(w_{i,t}^*)$ as an IV for $\Delta \log(w_{i,t})$. Third, our method does not pick up the moment condition constructed by $w_{i,t}$, which indicates that $\Delta \log(w_{i,t})$ is an endogenous variable in the labor supply equation. Finally, the GMM shrinkage estimator, though consistent and asymptotically efficient, may contain some shrinkage bias in finite samples. However, we recommend to use the post moment selection GMM estimator, which is as efficient as the GMM shrinkage estimator but has smaller finite sample bias.

9 Conclusion

This paper proposes a GMM shrinkage method to efficiently estimate the unknown parameter θ_o identified by some moment restrictions, when there is another set of possibly misspecified moment conditions. We show that our GMM shrinkage method enjoys oracle properties, i.e. it consistently selects the correct moment conditions in the second set and at the same time, the estimator achieves the semi-parametric efficiency bound implied by all the correct moment conditions. We also show that the GMM shrinkage method can be used to perform grouped variable selection and grouped moment selection simultaneously. When the moment conditions in the first set fail to strongly identify θ_o, we show that the GMM shrinkage method can still consistently select the correctly specified moment conditions in the second set and more importantly, the GMM shrinkage estimator has better stochastic properties compared with estimators that only use the moment conditions in the first set. We provide both consistent and conservative data-driven procedures of selecting the tuning parameter in finite samples, which makes our method fully adaptive for empirical implementation.

We check the finite sample properties of the GMM shrinkage method in simulation experiments and in an empirical example from labor economics. The simulation studies show that our method performs remarkably well in terms of the correct model/moment selection and the finite sample properties of its estimators. Compared with existing methods, it has tighter control in selecting models and moment conditions which produce inconsistent estimation and when the sample size is small, our method even has higher probability of selecting the true model with all the correct moment conditions. As an empirical illustration, we apply the GMM shrinkage method to estimate the life-cycle labor supply equation studied in MaCurdy (1981) and Altonji (1986). Our method selects the moment conditions constructed by the IVs in MaCurdy (1981) and Altonji (1986), which supports the validity of these IVs. However, our method does not pick up the moment condition constructed by the imputed hourly wage, which indicates that $\Delta \log(w_{i,t})$ is an endogenous variable in the labor supply equation. Hence, our empirical findings support continued use of the IVs in MaCurdy (1981) and Altonji (1986) to consistently estimate the life-cycle labor supply equation. Moreover, our estimators of the intertemporal substitution elasticity have smaller standard deviations, though their values are close to those in the literature.

There are many future research directions arising from the findings of this paper. First, we can study the moment selection problem in a more complex scenario where the dimensionality of the moment conditions in (1.2) grows with the sample size. Some econometric models have IVs whose number naturally grows with the sample size. For example, the dynamic panel data model with T passing to infinity. Moment selection and efficient estimation are important issues in those models. Another research direction is to study the properties of the GMM shrinkage estimation when the moment conditions in set-1 or set-2 are locally misspecified[9]. If the moment conditions in set-1 are locally misspecified and there are correctly specified moment conditions in set-2, we would like to check whether these potentially valid moment conditions can be consistently selected and whether the GMM shrinkage estimator has better stochastic properties. It is also interesting to study the properties of the GMM shrinkage estimation when the moment conditions in set-2 are locally misspecified. In this scenario, the moment selection coefficients behave as local alternatives to zero and this specification can help us to further check the finite sample properties of the GMM shrinkage method. Finally, we have not touched the issue of constructing robust CIs in GMM shrinkage estimation. As we have discussed earlier, the

[9] A moment condition is locally misspecified if the orthogonality condition fails in finite sample, but holds when the sample size goes to infinity.

limiting distribution of the GMM shrinkage estimator is a pointwise asymptotic result and in finite samples, naive CIs constructed using this result fail to take the model and moment selection errors into account. Hence, the naive CIs may have poor coverage probabilities and the resulting hypothesis tests may have incorrect size. There are some procedures proposed in the literature to tackle this problem, for example Andrews and Guggenberger (2010) and Andrews and Guggenberger (2009). It is interesting to check if these procedures can be applied into our framework to construct robust CIs with correct coverage probability and asymptotic size. The above directions of research form part of the author's ongoing work, some preliminary results have been obtained, and these will be reported in a later paper.

References

Altonji, J. (1986): "Intertemporal substitution in labor supply: Evidence from micro data," *Journal of Political Economy*, 94(3), 176–215.

Andrews, D. (1999): "Consistent moment selection procedures for generalized method of moments estimation," *Econometrica*, 67(3), 543–563.

Andrews, D., and P. Guggenberger (2009): "Hybrid and Size-Corrected Subsampling Methods," *Econometrica*, 77(3), 721–762.

——— (2010): "The limit of finite-sample size and a problem with subsampling," *Econometric Theory*, 26(2), 426–468.

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

Angrist, J., and A. Krueger (1991): "Does compulsory school attendance affect schooling and earnings?," *The Quarterly Journal of Economics*, 106(4), 979–1014.

Arellano, M., and O. Bover (1995): "Another look at the instrumental variable estimation of error-components models* 1," *Journal of econometrics*, 68(1), 29–51.

Barro, R., and J. Lee (1993): "Losers and winners in economic growth," *NBER working paper*, NO.4341.

Blundell, R., and S. Bond (1998): "Initial conditions and moment restrictions in dynamic panel data models," *Journal of econometrics*, 87(1), 115–143.

Bond, S., A. Hoeffler, and J. Temple (2001): "GMM estimation of empirical growth models," *Unpublished manuscript*.

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

Caner, M., and H. Zhang (2009): "General Estimating Equations: Model Selection and Estimation with Diverging Number of Parameters," *Unpublished manuscript*.

Card, D. (1999): "The causal effect of education on earnings," *Handbook of Labor Economics*, 3, 1801–1863.

Caselli, F., G. Esquivel, and F. Lefort (1996): "Reopening the convergence debate: a new look at cross-country growth empirics," *Journal of Economic Growth*, 1(3), 363–389.

Chen, X., and D. Pouzo (2008): "Estimation of nonparametric conditional moment models with possibly nonsmooth moments," *Cowles Foundation Discussion Paper*, NO.1650.

Eichenbaum, M., L. Hansen, and K. Singleton (1988): "A time series analysis of representative agent models of consumption and leisure choice under uncertainty," *Quarterly Journal of Economics*, 103(1), 51–78.

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

Geyer, C. (1994): "On the asymptotics of constrained M-estimation," *Annals of Statistics*, 22(4), 1993–2010.

Hahn, J., and G. Kuersteiner (2002): "Discontinuities of weak instrument limiting distributions," *Economics Letters*, 75(3), 325–331.

Hong, H., B. Preston, and M. Shum (2003): "Generalized Empirical Likelihood–Based Model Selection Criteria for Moment Condition Models," *Econometric Theory*, 19(06), 923–943.

Kane, T., and C. Rouse (1995): "Labor-market returns to two-and four-year college," *American Economic Review*, 85(3), 600–614.

Knight, K. (1999): "Epi-convergence in distribution and stochastic equi-semicontinuity," *Unpublished manuscript*.

Knight, K., and W. Fu (2000): "Asymptotics for lasso-type estimators," *Annals of Statistics*, 28(5), 1356–1378.

Leeb, H., and B. Potscher (2005): "Model selection and inference: Facts and fiction," *Econometric Theory*, 21(01), 21–59.

MaCurdy, T. (1981): "An empirical model of labor supply in a life-cycle setting," *Journal of Political Economy*, 89(6), 1059–1085.

Miguel, E., S. Satyanath, and E. Sergenti (2004): "Economic shocks and civil conflict: An instrumental variables approach," *Journal of Political Economy*, 112(4), 725–753.

Sargan, J. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica*, 26(3), 393–415.

Stock, J., and J. Wright (2000): "GMM with weak identification," *Econometrica*, 68(5), 1055–1096.

Wang, H., and C. Leng (2008): "A note on adaptive group lasso," *Computational Statistics & Data Analysis*, 52(12), 5277–5286.

Wang, H., R. Li, and C. Tsai (2007): "Tuning parameter selectors for the smoothly clipped absolute deviation method," *Biometrika*, 94(3), 553.

Zhang, Y., R. Li, and C. Tsai (2010): "Regularization parameter selections via generalized information criterion," *Journal of the American Statistical Association*, 105(489), 312–323.

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

APPENDIX A

Throughout the appendix, the symbols "$\rightarrow_p$" and "$\rightarrow_d$" stand for "convergence in probability" and "convergence in distribution" respectively. Denote $m(\alpha) = E[\rho(Z_i, \alpha)]$ and an empirical process indexed by the function g as $\upsilon_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [g(Z_i) - E(g(Z_i)]$. We use $\|\cdot\|_E$ to denote the Euclidean norm in the Euclidean space. For any sequences $(a_n, b_n)_n$ of random variables, $a_n \asymp b_n$ means that $(1 + o_p(1))b_n = a_n$ or vice versa.

A Some Useful Lemmas

We first prove some lemmas which are useful for deriving the asymptotic properties of the GMM shrinkage estimator. Denote

$$\begin{aligned} V_1^{(n)}(\theta, \beta) &= \left[n^{-1} \sum_{i=1}^{n} \rho(Z_i, \theta, \beta) \right]' W_n \left[n^{-1} \sum_{i=1}^{n} \rho(Z_i, \theta, \beta) \right] + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_j) \\ &= V_0^{(n)}(\theta, \beta) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_j) \end{aligned} \tag{A.1}$$

and

$$V_0(\theta, \beta) = \{E[\rho(Z, \theta, \beta)]\}' W_o \{E[\rho(Z, \theta, \beta)]\}. \tag{A.2}$$

The first lemma claims that

$$V_0^{(n)}(\theta, \beta) \asymp V_0(\theta, \beta) + R_n, \tag{A.3}$$

where $R_n = \sup_{(\theta,\beta)\in\Theta\times\mathcal{B}} \left[\frac{1}{n} \{\upsilon_n[\rho(Z, \theta, \beta)]\}' W_o \{\upsilon_n[\rho(Z, \theta, \beta)]\} \right]$. Its proof follows similar arguments to the proof of Lemma B.4 in Chen and Pouzo (2008).

Lemma A.1 *Under assumption 3.1.(iii), we have*

$$V_0^{(n)}(\theta, \beta) \geq c_1 V_0(\theta, \beta) - c_2 R_n \tag{A.4}$$

and

$$V_0^{(n)}(\theta, \beta) \leq c_3 V_0(\theta, \beta) + c_4 R_n, \tag{A.5}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$, where c_i $(i = 1, ..., 4)$ denotes some generic positive constants.

Proof. By assumption 3.1.(iii), we can deduce that

$$V_0^{(n)}(\theta,\beta) \geq c\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \tag{A.6}$$

for all $(\theta,\beta) \in \Theta \times \mathcal{B}$ w.p.a.1, where c denotes some generic positive constant. As W_o is positive definite, so we have

$$\left[2\frac{\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)}{n} - E\left[\rho(Z_i,\theta,\beta)\right]\right]' W_o \left[2\frac{\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)}{n} - E\left[\rho(Z_i,\theta,\beta)\right]\right] \geq 0,$$

which can be rewritten as

$$\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \geq \frac{1}{2}V_0(\theta,\beta) - R_n. \tag{A.7}$$

for all $(\theta,\beta) \in \Theta \times \mathcal{B}$. Now the result in (A.4) can be deduced from the inequalities in (A.6) and (A.7).

For the second result, note that assumption 3.1.(iii) also implies

$$V_0^{(n)}(\theta,\beta) \leq c\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \tag{A.8}$$

and

$$\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta) - 2E\left[\rho(Z_i,\theta,\beta)\right]\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta) - 2E\left[\rho(Z_i,\theta,\beta)\right]\right] \geq 0 \tag{A.9}$$

for all $(\theta,\beta) \in \Theta \times \mathcal{B}$ w.p.a.1. The inequality in (A.9) can be rewritten as

$$\left[\frac{\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)}{n}\right]' W_o \left[\frac{\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)}{n}\right] \leq 2V_0(\theta,\beta) + 2R_n. \tag{A.10}$$

for all $(\theta,\beta) \in \Theta \times \mathcal{B}$. Now the result in (A.5) can be deduced from the inequalities in (A.8) and (A.9). ■

The next lemma establishes the local quadratic approximation of $V_0(\theta,\beta)$ in terms of $\left(||\theta-\theta_o||_E^2 + ||\beta-\beta_o||_E^2\right)^{\frac{1}{2}}$ for all (θ,β) in shrinking neighborhoods of (θ_o,β_o), which is useful to derive the convergence rate of the GMM shrinkage estimator.

Lemma A.2 *Under assumptions 3.1.(iii) and 3.2.(ii)-(iii), we have*

$$||\theta-\theta_o||_E^2+||\beta-\beta_o||_E^2 \asymp E\left[\rho(Z_i,\theta,\beta)\right]' W_o E\left[\rho(Z_i,\theta,\beta)\right] \tag{A.11}$$

for all (θ,β) *in local neighborhoods of* (θ_o,β_o).

Proof. Denote

$$g_q(Z,\theta)=\left(\begin{array}{cccc} g_{q,1}(Z,\theta), & ..., & g_{q,q}(Z,\theta) \end{array}\right)$$

and

$$g_k(Z,\theta)=\left(\begin{array}{cccc} g_{k,1}(Z,\theta), & ..., & g_{k,k}(Z,\theta) \end{array}\right).$$

First note that by assumption 3.2.(ii)

$$E\left[\rho(Z_i,\theta,\beta)\right]=\left(\begin{array}{c} E\left[g_q(Z,\theta)\right] \\ E\left[g_k(Z,\theta)\right]-\beta \end{array}\right)=\left(\begin{array}{cc} \frac{\partial E\left[g_q(Z,\widetilde{\theta})\right]}{\partial\theta'} & 0 \\ \frac{\partial E\left[g_k(Z,\widetilde{\theta})\right]}{\partial\theta'} & -I_k \end{array}\right)\left(\begin{array}{c} \theta-\theta_o \\ \beta-\beta_o \end{array}\right), \tag{A.12}$$

where

$$\begin{aligned} \frac{\partial E\left[g_q(Z,\widetilde{\theta})\right]}{\partial\theta'} &= \left(\begin{array}{cccc} \frac{\partial E\left[g_{q,1}(Z,\widetilde{\theta}_1)\right]}{\partial\theta}, & ..., & \frac{\partial E\left[g_{q,q}(Z,\widetilde{\theta}_q)\right]}{\partial\theta} \end{array}\right)', \\ \frac{\partial E\left[g_k(Z,\widetilde{\theta})\right]}{\partial\theta'} &= \left(\begin{array}{cccc} \frac{\partial E\left[g_{k,1}(Z,\widetilde{\theta}_{p+1})\right]}{\partial\theta}, & ..., & \frac{\partial E\left[g_{k,k}(Z,\widetilde{\theta}_{q+k})\right]}{\partial\theta} \end{array}\right)', \end{aligned}$$

$\widetilde{\theta}_j$ $(j=1,...,q+k)$ lies between θ and θ_o and I_k is a $k\times k$ identity matrix. As θ is in the shrinking neighborhood of θ_o and $\frac{\partial E[g_l(Z,\theta)]}{\partial\theta'}$ $(l=q,k)$ is continuous in θ, we can deduce that

$$\frac{\partial E\left[g_l(Z,\widetilde{\theta})\right]}{\partial\theta'}=\frac{\partial E\left[g_l(Z,\theta_o)\right]}{\partial\theta'}+o(1), \tag{A.13}$$

for $l=q,k$. Denote $\frac{\partial m(\theta)}{\partial\alpha'}=\left(\begin{array}{cc} \frac{\partial E[g_q(Z,\theta)]}{\partial\theta'} & 0 \\ \frac{\partial E[g_k(Z,\theta]}{\partial\theta'} & -I_k \end{array}\right)$, then by (A.12), (A.13) and the Cauchy-Schwarz inequality, we have

$$E\left[\rho(Z_i,\theta,\beta)\right]=\frac{\partial m(\alpha_o)}{\partial\alpha'}(\alpha-\alpha_o)+o(||(\alpha-\alpha_o||_E). \tag{A.14}$$

Using assumption 3.1.(iii), the result in (A.14) and the Cauchy-Schwarz inequality, we get

$$E\left[\rho(Z_i,\alpha)\right]' W_o E\left[\rho(Z_i,\alpha)\right] = (\alpha-\alpha_o)' \left[\frac{\partial m(\alpha_o)}{\partial \alpha} W_o \frac{\partial m(\alpha_o)}{\partial \alpha'}\right](\alpha-\alpha_o) + o(||(\alpha-\alpha_o||_E^2), \tag{A.15}$$

where $\alpha=(\theta,\beta)$ and $\alpha_o=(\theta_o,\beta_o)$. As $\frac{\partial E[g_q(Z,\theta_o)]}{\partial \theta'}$ has full column rank and W_o is strictly positive definite, $\frac{\partial m(\theta_o)}{\partial \alpha}$ has full rank and $\left[\frac{\partial m(\theta_o)}{\partial \alpha} W_o \frac{\partial m(\theta_o)}{\partial \alpha'}\right]$ is strictly positive definite. Let c_1 and c_2 denote the smallest and largest eigenvalues of $\left[\frac{\partial m(\theta_o)}{\partial \alpha} W_o \frac{\partial m(\theta_o)}{\partial \alpha'}\right]$. From (A.15), we can deduce that

$$\begin{aligned} c_2||(\alpha-\alpha_o||_E^2 + o(||(\alpha-\alpha_o||_E^2) &\geq E\left[\rho(Z_i,\alpha)\right]' W_o E\left[\rho(Z_i,\alpha)\right] \\ &\geq c_1||(\alpha-\alpha_o||_E^2 + o(||(\alpha-\alpha_o||_E^2). \end{aligned} \tag{A.16}$$

Now, result in (A.11) follows directly from (A.16). ■

B Proof of the Main Results in Section 3

Proof of Lemma 3.1. By the definition of $(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}})$, we have

$$V_0^{(n)}(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}}) \leq V_0^{(n)}(\theta_o,\beta_o) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}). \tag{B.1}$$

Applying Lemma A.1 and assumption 3.1.(iv), we deduce from (B.1) that

$$V_0(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + 2cR_n, \tag{B.2}$$

with probability approaching 1, where

$$R_n = \sup_{(\theta,\beta)\in\Theta\times\mathcal{B}} \left[\frac{1}{n}\left\{v_n\left[\rho(Z,\theta,\beta)\right]\right\}' W_o \left\{v_n\left[\rho(Z,\theta,\beta)\right]\right\}\right] \tag{B.3}$$

and c is some generic constant.

From assumption 3.1.(ii) and the definition of $\rho(Z,\theta,\beta)$, we get

$$\begin{aligned}\sup_{(\theta,\beta)\in\Theta\times\mathcal{B}}\left\{n^{-\frac{1}{2}}v_n\left[\rho(Z,\theta,\beta)\right]\right\} &= \sup_{(\theta,\beta)\in\Theta\times\mathcal{B}}\left[n^{-1}\sum_{i=1}^{n}\left\{\rho(Z_i,\theta,\beta)-E\left[\rho(Z_i,\theta,\beta)\right]\right\}\right] \\ &= \sup_{\theta\in\Theta}\left[n^{-1}\sum_{i=1}^{n}\left\{g(Z_i,\theta)-E\left[g(Z_i,\theta)\right]\right\}\right]=o_p(1), \quad \text{(B.4)}\end{aligned}$$

where $g(Z,\theta)=\left[g_q'(Z,\theta),g_k'(Z,\theta)\right]'$. By the triangle inequality, ULLN in (B.4), assumption 3.1.(iii)-(iv), we have

$$R_n=o_p(1) \text{ and } \sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_{o,j})=o_p(1). \quad \text{(B.5)}$$

From the Assumption 3.1.(iii) and results in (B.2) and (B.5), we can deduce that

$$\left\|E[g_q(Z,\widehat{\theta}_n^{\mathcal{S}})]\right\|_E=o(1) \text{ and } \left\|E[g_k(Z,\widehat{\theta}_n^{\mathcal{S}})]-\widehat{\beta}_n^{\mathcal{S}}\right\|_E=o_p(1) \quad \text{(B.6)}$$

Now, the first result in (B.6) and Assumption 3.1.(i) imply that $\widehat{\theta}_n^{\mathcal{S}}\to_p\theta_o$. From the second result in (B.6), triangle inequality, consistency of $\widehat{\theta}_n^{\mathcal{S}}$ and the continuity of $E[g_k(Z,\theta)]$, we can deduce that

$$\begin{aligned}o_p(1) &= \left\|E[g_k(Z,\widehat{\theta}_n^{\mathcal{S}})]-\widehat{\beta}_n^{\mathcal{S}}\right\|_E \\ &\geq \left|\left\|E[g_k(Z,\widehat{\theta}_n^{\mathcal{S}})]-E[g_k(Z,\theta_o)]\right\|_E-\left\|\widehat{\beta}_n^{\mathcal{S}}-\beta_o\right\|_E\right| \\ &= \left|\left\|\widehat{\beta}_n^{\mathcal{S}}-\beta_o\right\|_E+o_p(1)\right| \quad \text{(B.7)}\end{aligned}$$

which implies $\widehat{\beta}_n^{\mathcal{S}}\to_p\beta_o$. ■

Proof of Lemma 3.2. By the definition of $(\widehat{\theta}_n^{\mathcal{S}},\widehat{\beta}_n^{\mathcal{S}})$, we have

$$V_0^{(n)}(\widehat{\theta}_n^{\mathcal{S}},\widehat{\beta}_n^{\mathcal{S}})+\sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})\leq V_0^{(n)}(\theta_o,\beta_o)+\sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_{o,j}). \quad \text{(B.8)}$$

Using the inequalities in (A.4), (A.5) and (B.8), we get

$$c_1 V_0(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}}) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + c_2 R_n, \tag{B.9}$$

w.p.a.1, where c_1 and c_2 are some generic positive constants and R_n is defined in (B.3).

Next, by assumption 3.2.(iv), Taylor expansion, the triangle inequality and Cauchy Schwarz inequality, we get

$$\begin{aligned} &\left| \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n}(\beta_{o,j}) - \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}}) \right] \right| = \left| \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}'_{\lambda_n}(\beta_{o,j}) \left(\widehat{\beta}_{n,j}^{\mathcal{S}} - \beta_{o,j} \right) + \frac{1}{2} \widehat{P}''_{\lambda_n}(\widetilde{\beta}_j) \left(\widehat{\beta}_{n,j}^{\mathcal{S}} - \beta_{o,j} \right)^2 \right] \right| \\ \leq \; & \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| \left\| \widehat{\alpha}_n^{\mathcal{S}} - \alpha_o \right\|_E + \max_{j \in \mathcal{S}_\beta} \left| \frac{\widehat{P}''_{\lambda_n}(\beta_{o,j})}{2} + o_p(1) \right| \left\| \widehat{\alpha}_n^{\mathcal{S}} - \alpha_o \right\|_E^2 . \end{aligned} \tag{B.10}$$

w.p.a.1, where $\widetilde{\beta}_j$ lies between $\beta_{o,j}$ and $\widehat{\beta}_{n,j}^{\mathcal{S}}$ for $j \in \mathcal{S}_\beta$. From assumptions 3.1.(iv) and 3.2.(iv), inequalities in (B.9) and (B.10), we can apply Lemma A.2 to deduce that

$$\left\| \widehat{\alpha}_n^{\mathcal{S}} - \alpha_o \right\|_E^2 \leq c_3 b_n \left\| \widehat{\alpha}_n^{\mathcal{S}} - \alpha_o \right\|_E + c_4 R_n \tag{B.11}$$

w.p.a.1, where c_3, c_4 are some positive constants. The inequality in (B.11) implies

$$\left\| \widehat{\alpha}_n^{\mathcal{S}} - \alpha_o \right\|_E \leq \frac{c_3 b_n + (c_3^2 b_n^2 + 4 c_4 R_n)^{\frac{1}{2}}}{2} = O_p\left(b_n + n^{-\frac{1}{2}} \right), \tag{B.12}$$

where c_5 is some positive constant. Now, for any positive M, inequality in (B.12) enables us to deduce that

$$\Pr\left(\frac{||\widehat{\alpha}_n^{\mathcal{S}} - \alpha_o||_E}{\delta_n} \geq M \right) \leq \Pr\left(\delta_n M \leq O_p\left(b_n + n^{-\frac{1}{2}} \right) \right) + o_p(1),$$

which establishes the desired rate. ■

Proof of Theorem 3.3. On the event $\{\widehat{\beta}_{n,j}^{\mathcal{S}} \neq 0\}$ for some $j \in \mathcal{S}_\beta^c$, we have the following

Karush-Kuhn-Tucker (KKT) optimality condition:

$$2\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\theta}_n^{\mathcal{S}},\widehat{\beta}_n^{\mathcal{S}})}{\partial\beta_j}\right]' W_n\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\theta}_n^{\mathcal{S}},\widehat{\beta}_n^{\mathcal{S}})\right]+n\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})=0, \tag{B.13}$$

which implies

$$\left|W_n^{(j)}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\theta}_n^{\mathcal{S}},\widehat{\beta}_n^{\mathcal{S}})\right]\right|=\frac{\sqrt{n}\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})}{2}. \tag{B.14}$$

where $W_n^{(j)}$ denotes the j-th row of the weight matrix W_n.

Note that

$$\begin{aligned}
n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}}) &= v_n\left[\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})\right]+n^{\frac{1}{2}}E\left[\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})\right] \\
&= v_n\left[\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})\right]+\frac{\partial m(\widetilde{\theta}_n^{\mathcal{S}})}{\partial\alpha'}\left[n^{\frac{1}{2}}(\widehat{\alpha}_n^{\mathcal{S}}-\alpha_o)\right],
\end{aligned}$$

where

$$\begin{aligned}
\frac{\partial m(\widetilde{\theta}_n^{\mathcal{S}})}{\partial\alpha'} &= \begin{bmatrix}\frac{\partial E[g_q(Z,\widetilde{\theta}_n^{\mathcal{S}})]}{\partial\theta'} & 0\\ \frac{\partial E[g_k(Z,\widetilde{\theta}_n^{\mathcal{S}})]}{\partial\theta'} & -I_k\end{bmatrix}, \\
\frac{\partial E[g_q(Z,\widetilde{\theta}_n^{\mathcal{S}})]}{\partial\theta'} &= \left(\frac{\partial E[g_{q,1}(Z,\widetilde{\theta}_{1,n}^{\mathcal{S}})]}{\partial\theta},...,\frac{\partial E[g_{q,q}(Z,\widetilde{\theta}_{q,n}^{\mathcal{S}})]}{\partial\theta}\right), \\
\frac{\partial E[g_k(Z,\widetilde{\theta}_n^{\mathcal{S}})]}{\partial\theta'} &= \left(\frac{\partial E[g_{k,1}(Z,\widetilde{\theta}_{q+1,n}^{\mathcal{S}})]}{\partial\theta},...,\frac{\partial E[g_{k,k}(Z,\widetilde{\theta}_{q+k,n}^{\mathcal{S}})]}{\partial\theta}\right),
\end{aligned}$$

and $\widetilde{\theta}_{j,n}^{\mathcal{S}}$ $(j=1,...,q+k)$ lies between θ_o and $\widehat{\theta}_n^{\mathcal{S}}$. From Assumption 3.2.(i), we have $v_n\left[\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})\right]=O_p(1)$. By Lemma 3.2 and Assumption 3.2.(ii), we have $n^{\frac{1}{2}}(\widehat{\alpha}_n^{\mathcal{S}}-\alpha_o)=O_p(1)$. From Assumption 3.3.(iii) and the consistency of $\widehat{\alpha}_n^{\mathcal{S}}$, we can deduce that

$$\left\|\frac{\partial m(\widetilde{\theta}_n^{\mathcal{S}})}{\partial\alpha'}\right\|_E\leq\left\|\frac{\partial m(\widetilde{\theta}_n^{\mathcal{S}})}{\partial\alpha'}-\frac{\partial m(\theta_o)}{\partial\alpha'}\right\|_E+\left\|\frac{\partial m(\theta_o)}{\partial\alpha'}\right\|_E=O_p(1).$$

Hence we have $n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})=O_p(1)$, which combined with Assumption 3.1.(iii),

implies that

$$\left| W_n^{(j)} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n^{\mathcal{S}}) \right] \right| = O_p(1). \tag{B.15}$$

While from Assumption 3.3.(iii), we get

$$\frac{\sqrt{n}\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})}{2} = \frac{r_n \lambda_n n^{\frac{1}{2}}}{2} \frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})}{r_n \lambda_n} \to_p \infty. \tag{B.16}$$

Now, using the results in (B.15) and (B.16), we can deduce that $\Pr\left(\widehat{\beta}_{n,j}^{\mathcal{S}} = 0\right) \to 1$ for $j \in \mathcal{S}_\beta^c$. ■

Proof of Theorem 3.4. Define $\alpha_{o,\mathcal{S}} = (\theta_o, \beta_{o,+})$ and accordingly $\widehat{\alpha}_{n,\mathcal{S}}^{\mathcal{S}} = \left(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_{n,+}^{\mathcal{S}}\right)$. For any compact subset K in $R^{d_\theta + d_{\mathcal{S}_\beta}}$, we denote any element $u_{\mathcal{S}} \in K$ as $u_{\mathcal{S}} = (u_\theta, u_{\beta_+})$, where u_θ is the first d_θ elements in $u_{\mathcal{S}}$ and u_{β_+} is the last $d_{\mathcal{S}_\beta}$ elements in $u_{\mathcal{S}}$. Denote

$$\begin{aligned} V_{2,n}(u_{\mathcal{S}}) &= \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) \right]' W_n \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) \right] \\ &\quad - \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \alpha_o) \right]' W_n \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \alpha_o) \right] \\ &\quad + n \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n}(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}}) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right] \\ &:= V_{2,n}^*(u_{\mathcal{S}}) + n \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n}(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}}) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right], \end{aligned}$$

where $\rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) = \rho(Z_i, \theta_o + \frac{u_\theta}{\sqrt{n}}, \beta_{o,+} + \frac{u_{\beta_+}}{\sqrt{n}}, \beta_{o,-})$. From Theorem 3.3, we know that $\widehat{\beta}_{n,-}^{\mathcal{S}} = 0$ w.p.a.1. Thus, $\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}}^{\mathcal{S}} - \alpha_{o,\mathcal{S}}\right)$ is the minimizer of $V_{2,n}(u_{\mathcal{S}})$ w.p.a.1.

If we denote

$$\mathcal{F}_n = \left\{ f_{u_{\mathcal{S}}}^n(Z) = \rho^s(Z, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) - \rho(Z, \alpha_o) : u_{\mathcal{S}} \in K \right\},$$

then the compactness of K, assumptions 3.2.(i)-(ii) imply that $\mathcal{F}_n$ is a Donsker class. As K is compact, so there exists some constant C_k, such that $\sup_{u_{\mathcal{S}} \in K} \left\| n^{-\frac{1}{2}} u_{\mathcal{S}} \right\|_E \leq n^{-\frac{1}{2}} C_k =$

$o(1)$. Now we can use Lemma 2.17 in Pakes and Pollard (1989) to deduce that

$$\upsilon_n \left\{ \rho^s(Z, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) - \rho(Z, \alpha_o) \right\} = o_p(1), \tag{B.17}$$

uniformly over $u_\mathcal{S} \in K$.

Next note that by assumption **??**.(iii) and the compactness of K, we have

$$\sqrt{n} \left\{ E\left[\rho^s(Z, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) \right] - E\left[\rho(Z, \alpha_o)\right] \right\} = \frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}'} u_\mathcal{S} + o(1), \tag{B.18}$$

uniformly over $u_\mathcal{S} \in K$. Thus, (B.17) and (B.18) imply that uniformly over $u_\mathcal{S} \in K$, there is

$$\begin{aligned} n^{-\frac{1}{2}} \sum_{i=1}^n \rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) &= \upsilon_n \left\{ \rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) - \rho(Z_i, \alpha_o) \right\} + \upsilon_n \left[\rho(Z_i, \alpha_o)\right] \\ &\quad + \sqrt{n} \left\{ E\left[\rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_\mathcal{S}}{\sqrt{n}}) \right] - E\left[\rho(Z_i, \alpha_o)\right] \right\} \\ &= \upsilon_n \left[\rho(Z, \alpha_o)\right] + \frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}'} u_\mathcal{S} + o_p(1). \end{aligned} \tag{B.19}$$

Now, we can use the result in (B.19) to deduce that

$$V_{2,n}^*(u_\mathcal{S}) = u_\mathcal{S}' \left[\frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}} \right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}'} \right] u_\mathcal{S} + 2u_\mathcal{S}' \left[\frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}} \right] W_o \left\{ \upsilon_n \left[\rho(Z, \alpha_o)\right] \right\} + o_p(1), \tag{B.20}$$

uniformly over $u_\mathcal{S} \in K$. If $j \in \mathcal{S}_\beta$, then by assumptions 3.2.(iv) and 3.3.(i)

$$n \left[\widehat{P}_{\lambda_n}(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}}) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right] = \sqrt{n} \widehat{P}_{\lambda_n}'(\beta_{o,j}) u_{\beta_+,j} + \left[\widehat{P}_{\lambda_n}''(\beta_{o,j}) + o_p(1) \right] u_{\beta_+,j}^2 \to 0 \tag{B.21}$$

uniformly in $u_{\beta_+,j}$. Using the results in (B.20)-(B.21) and triangle inequality, we get

$$V_{2,n}(u_\mathcal{S}) \to_d V_2(u_\mathcal{S}) = u_\mathcal{S}' M_{11} u_\mathcal{S} + 2u_\mathcal{S}' \left[\frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}} \right] W_o \Psi(\alpha_o) \tag{B.22}$$

in $l^\infty(K)$. It is clear that $V_2(u_\mathcal{S})$ is uniquely minimized at

$$u_\mathcal{S}^* = -M_{11}^{-1} \left[\frac{\partial m(\theta_o)}{\partial \alpha_\mathcal{S}} \right] W_o \Psi(\alpha_o) \tag{B.23}$$

By Lemma 3.2 and assumption 3.3.(i), there is

$$\sqrt{n}\left(\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) = O_p(1) \tag{B.24}$$

Now, the asymptotic tightness of $\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}}$ in (B.24), the uniform convergence in distribution in (B.22) and unique minimization in (B.23) enable us to invoke the ACMT to deduce that

$$\sqrt{n}\left(\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) \to_d N(0, M_{11}^{-1}\Sigma_{11}M_{11}^{-1})$$

■

Proof of Theorem 3.5. The first result is implied by Lemma 3.1 and Lemma 3.3, so we only need to show the second claim. First note that if $W_n \to_p W_o = \{E\left[\Psi(\alpha_o)\Psi(\alpha_o)'\right]\}^{-1}$, then the centered limiting distribution in (3.18) will be simplified to

$$\sqrt{n}\left(\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) \to_d N(0, M_{11}^{-1}) \tag{B.25}$$

Denote Ω_{θ_o} to be the first $d_{\theta_o} \times d_{\theta_o}$ sub-matrix of M_{11}^{-1} and $\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta} = \frac{\partial E[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial \theta}$. Note that

$$\begin{aligned}
M_{11} &= \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial \theta} & \frac{\partial E[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial \theta} \\ 0 & -I_{d_{\beta_o^+}\times d_{\beta_o^+}} \end{pmatrix} W_o \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial \theta'} & 0 \\ \frac{\partial E[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial \theta'} & -I_{d_{\beta_o^+}\times d_{\beta_o^+}} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial m_e(\theta_o)}{\partial \theta} & \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta} \\ 0 & -I_{d_{\beta_o^+}\times d_{\beta_o^+}} \end{pmatrix} \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \frac{\partial m_e(\theta_o)}{\partial \theta'} & 0 \\ \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} & -I_{d_{\beta_o^+}\times d_{\beta_o^+}} \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{11}^{11} & -\frac{\partial m_e(\theta_o)}{\partial \theta}W_{12} - \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta}W_{22} \\ -W_{21}\frac{\partial m_e(\theta_o)}{\partial \theta'} - W_{22}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} & W_{22} \end{pmatrix},
\end{aligned} \tag{B.26}$$

where

$$\Sigma_{11}^{11} = \frac{\partial m_e(\theta_o)}{\partial \theta}W_{11}\frac{\partial m_e(\theta_o)}{\partial \theta'} + 2\frac{\partial m_e(\theta_o)}{\partial \theta}W_{12}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} + \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta}W_{22}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'}.$$

From (B.26), it is easy to get

$$\begin{aligned}\Omega_{\theta_o}^{-1} &= \frac{\partial m_e(\theta_o)}{\partial\theta}W_{11}\frac{\partial m_e(\theta_o)}{\partial\theta'} + 2\frac{\partial m_e(\theta_o)}{\partial\theta}W_{12}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial\theta'} + \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial\theta}W_{22}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial\theta'} \\ &\quad -\frac{\partial m_e(\theta_o)}{\partial\theta}W_{12}W_{22}^{-1}W_{21}\frac{\partial m_e(\theta_o)}{\partial\theta'} - 2\frac{\partial m_e(\theta_o)}{\partial\theta}W_{12}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial\theta'} - \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial\theta}W_{22}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial\theta'} \\ &= \frac{\partial m_e(\theta_o)}{\partial\theta}\left(W_{11} - W_{12}W_{22}^{-1}W_{21}\right)\frac{\partial m_e(\theta_o)}{\partial\theta'} \\ &= \left[\frac{\partial m_e(\theta_o)}{\partial\theta}\right]\{V_{e,o}\}^{-1}\left[\frac{\partial m_e(\theta_o)}{\partial\theta}\right]' = (\Sigma^*)^{-1}, \end{aligned} \tag{B.27}$$

where the last equality is due to the fact that $\left(W_{11} - W_{12}W_{22}^{-1}W_{21}\right)^{-1} = V_{e,o}$. Now, using results in (B.25), (B.27) and the Continuous Mapping Theorem (CMT), we can deduce that

$$\sqrt{n}\left(\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right) \to_d N(0, \Sigma^*),$$

which establishes the semi-parametric efficiency of the GMM shrinkage estimator $\widehat{\theta}_n^{\mathcal{S}}$. ■

C Proof of the Main results in Section 4

Proof of Lemma 4.1. First note that by definition, $\widehat{P}_{\lambda_n}(\alpha_{o,j}) = \lambda_n \widehat{w}_{\alpha_j} \|\alpha_{o,j}\|_2 = 0$ for all $j \in \mathcal{S}_\alpha^c$. By CMT and the Slutsky Theorem, we can deduce that

$$\widehat{P}_{\lambda_n}(\alpha_{o,j}) = \lambda_n \widehat{w}_{\alpha_j} \|\alpha_{o,j}\|_2 \to_p 0, \tag{C.1}$$

for any $j \in \mathcal{S}_\alpha$. Hence, assumption 3.1.(iv) holds for the adaptive group Lasso penalty function. Now, the consistency of $\widehat{\alpha}_n^{\mathcal{S}}$ follows by the similar arguments used in the proof of Lemma 3.1.

Next note that, $\widehat{P}_{\lambda_n}(\cdot)$ is continuously twice differentiable at $\alpha_{o,j}$ for any $j \in \mathcal{S}_\alpha$ and

$$\frac{\partial^2 \widehat{P}_{\lambda_n}(\alpha_{o,j})}{\partial\alpha_j\partial\alpha_j'} = \lambda_n \widehat{w}_{\alpha_j}\left(-\frac{1}{\|\alpha_{o,j}\|_2^2}\alpha_{o,j}\alpha_{o,j}' + \frac{1}{\|\alpha_{o,j}\|_2} I_{d_{\alpha_{o,j}}}\right),$$

where $I_{d_{\alpha_{o,j}}}$ denotes a $d_{\alpha_{o,j}} \times d_{\alpha_{o,j}}$ identity matrix and $d_{\alpha_{o,j}}$ is the dimensionality of $\alpha_{o,j}$. As $\|\alpha_{o,j}\|_2 \neq 0$ for all $j \in \mathcal{S}_\alpha$ and $\lambda_n = o(1)$, by CMT and the Slutsky Theorem, we can deduce that $\left\|\frac{\partial^2 \widehat{P}_{\lambda_n}(\alpha_{o,j})}{\partial\alpha_j\partial\alpha_j'}\right\|_E = o_p(1)$ for all $j \in \mathcal{S}_\alpha$. Thus the adaptive group Lasso penalty

function satisfies Assumption 3.2.(iv). Now the convergence rate in (4.5) follows by similar arguments in the proof of Lemma 3.2. ■

Proof of Theorem 4.1. On the event $\left\{\left\|\widehat{\alpha}_{n,j}^{\mathcal{S}}\right\|_2 \neq 0\right\}$, for some $j \in \mathcal{S}_\alpha^c$, we have the following KKT optimality condition

$$2\left\|\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})}{\partial\alpha_j}\right]W_n\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})\right]\right\|_E = \left\|\frac{n\lambda_n\widehat{w}_{\alpha_j}}{\left\|\widehat{\alpha}_{n,j}^{\mathcal{S}}\right\|_2}\widehat{\alpha}_{n,j}^{\mathcal{S}}\right\|_E. \tag{C.2}$$

Following similar arguments used in the proof of Theorem 3.3, we can show that

$$n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}}) = O_p(1). \tag{C.3}$$

If $j \in \mathcal{S}_\beta^c$, then

$$\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})}{\partial\beta_j}\right]W_n = W_{n,j} \tag{C.4}$$

where $W_{n,j}$ denotes the j-th component-wise rows of W_n. Hence by Assumption 3.1.(iii) and the result in (C.3), we get

$$\left\|\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})}{\partial\alpha_j}\right]W_n\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})\right]\right\|_E = O_p(1). \tag{C.5}$$

On the other hand denote $g(Z,\theta) = [g_q'(Z,\theta), g_k'(Z,\theta)]'$, if $j \in \mathcal{S}_\theta^c$, then

$$\begin{aligned}\left\|n^{-1}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n^{\mathcal{S}})}{\partial\theta_j}\right\|_E &\leq \left\|n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial g(Z,\widehat{\theta}_n^{\mathcal{S}})}{\partial\theta_j} - E\left[\frac{\partial g(Z,\widehat{\theta}_n^{\mathcal{S}})}{\partial\theta_j}\right]\right\}\right\|_E + \left\|E\left[\frac{\partial g(Z,\widehat{\theta}_n^{\mathcal{S}})}{\partial\theta_j}\right]\right\|_E \\ &\leq \left\|E\left[\frac{\partial g(Z,\widehat{\theta}_n^{\mathcal{S}})}{\partial\theta_j}\right] - E\left[\frac{\partial g(Z,\theta_o)}{\partial\theta_j}\right]\right\|_E + \left\|E\left[\frac{\partial g(Z,\theta_o)}{\partial\theta_j}\right]\right\|_E + o_p(1) \\ &= O_p(1), \end{aligned} \tag{C.6}$$

where the first inequality is due to the triangle inequality, the second inequality is by the triangle inequality and assumption 4.1, the last equality is by assumption 3.2.(ii). By (C.6) and Cauchy-Schwarz inequality, result in (C.5) also holds for $j \in \mathcal{S}_\theta^c$. Hence, by definition, (C.5) holds for any $j \in \mathcal{S}_\alpha^c$.

On the other hand, as $n^{\frac{1+\omega}{2}}\lambda_n \to \infty$ and $\widehat{\alpha}_n$ is $\sqrt{n}$-consistent, we can deduce that

$$\frac{1}{2}\left\| \frac{\sqrt{n}\lambda_n \widehat{w}_{\alpha_j}}{\left\|\widehat{\alpha}_{n,j}^{\mathcal{S}}\right\|_2}\widehat{\alpha}_{n,j}^{\mathcal{S}} \right\|_E = \frac{n^{\frac{1+\omega}{2}}\lambda_n}{2}\frac{1}{\|\sqrt{n}\widehat{\alpha}_{n,j}\|_2^{\omega}} \to_p \infty. \tag{C.7}$$

Now, using the results in (C.2), (C.5) and (C.7), we can deduce that $\Pr\left(\left\|\widehat{\alpha}_{n,j}^{\mathcal{S}}\right\|_2 = 0\right) \to 1$ for $j \in \mathcal{S}_\alpha^c$. ■

D Proof of the Main results in Section 5

Proof of Lemma 5.1. Using similar arguments in the proof of Lemma 3.1, we get

$$V_0(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + 2cR_n, \tag{D.1}$$

w.p.a.1, where $R_n = \sup_{\alpha\in\mathcal{A}}\left[\frac{1}{n}\left\{v_n\left[\rho(Z_i,\alpha)\right]\right\}' W_o \left\{v_n\left[\rho(Z_i,\alpha)\right]\right\}\right]$ and c is some generic constant. Under assumptions 3.1.(iii), 5.1.(iii) and (v), we have

$$\left\|n^{-\tau}G_{q,n}(\widehat{\theta}_n^{\mathcal{S}})\right\|_E^2 + \left\|G_k(\widehat{\theta}_n^{\mathcal{S}}) - \widehat{\beta}_n^{\mathcal{S}}\right\|_E^2 = O_p(n^{-1} + \max_{j\in\mathcal{S}_\beta}\widehat{P}_{\lambda_n}(\beta_{o,j})). \tag{D.2}$$

(D.2) implies that $\left\|G_{q,n}(\widehat{\theta}_n^{\mathcal{S}})\right\|_E^2 = O_p(n^{2\tau-1}+n^{2\tau}\max_{j\in\mathcal{S}_\beta}\widehat{P}_{\lambda_n}(\beta_{o,j})) = o_p(1)$. Hence, using the uniform approximation in assumption 5.1.(i), we can deduce that $\left\|G_q(\widehat{\theta}_n^{\mathcal{S}})\right\|_E = o_p(1)$, which combined with the identification condition in 5.1.(i), implies the consistency of $\widehat{\theta}_n^{\mathcal{S}}$. The consistency of $\widehat{\beta}_n^{\mathcal{S}}$ can be proved using similar arguments in the proof of Lemma 3.1.

Next, we derive the convergence rate of $\widehat{\alpha}_n^{\mathcal{S}}$. Using similar arguments in the proof of Lemma A.2, we can apply assumptions 3.1.(iii) and 5.1.(iv) to deduce

$$V_0(\widehat{\theta}_n^{\mathcal{S}}, \widehat{\beta}_n^{\mathcal{S}}) \geq cn^{-2\tau}\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E^2 + c\left\|G_k(\widehat{\theta}_n^{\mathcal{S}}) - \widehat{\beta}_n^{\mathcal{S}}\right\|_E^2 \tag{D.3}$$

w.p.a.1, where c is some generic positive constant. From (D.1) and (D.3), we can deduce that w.p.a.1,

$$\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E^2 \leq n^{2\tau}\sum_{j\in\mathcal{S}_\beta}\widehat{P}_{\lambda_n}(\beta_{o,j}) + cn^{2\tau}R_n, \tag{D.4}$$

which implies that $\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E = O_p(n^{\tau} \max_{j\in\mathcal{S}_\beta} \widehat{P}^{\frac{1}{2}}_{\lambda_n}(\beta_{o,j}) + n^{-\frac{1}{2}+\tau})$.

Using Assumption 3.2.(iii), we obtain

$$\begin{aligned}\left|\widehat{P}_{\lambda_n}(\beta_{o,j}) - \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})\right| \;\leq\; & \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| \left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E \\ & + \max_{j\in\mathcal{S}_\beta}\left|\frac{\widehat{P}''_{\lambda_n}(\beta_{o,j})}{2} + o_p(1)\right| \left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E^2. \qquad \text{(D.5)}\end{aligned}$$

By the triangular inequality, there is

$$\left\|G_k(\widehat{\theta}_n^{\mathcal{S}}) - \widehat{\beta}_n^{\mathcal{S}}\right\|_E^2 \geq \left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E^2 - 2\left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E J_{1,n} + J_{1,n}^2, \qquad \text{(D.6)}$$

where $J_{1,n} = \left\|G_k(\widehat{\theta}_n^{\mathcal{S}}) - G_k(\theta_o)\right\|_E = O_p\left(\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E\right)$. Denote $J_{2,n} = cn^{-2\tau}\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E^2$, then from (D.2), (D.3) and (D.5), we get

$$\left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E^2 - c\left(\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + J_{1,n}\right)\left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E \leq R_n - J_{1,n}^2 - J_{2,n}, \qquad \text{(D.7)}$$

w.p.a.1. As $J_{1,n}^2 = O_p\left(\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E^2\right)$ and $J_{2,n} = O_p\left(\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E^2\right)$, hence from the inequality in (D.7), we can deduce that

$$\left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E = O_p\left(\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E, \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + n^{-\frac{1}{2}}\right). \qquad \text{(D.8)}$$

Note that if $\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E = O_p\left(\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + n^{-\frac{1}{2}}\right)$, then the result is proved. Hence we only need to consider the case that $\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E$ has the convergence rate slower than $\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + n^{-\frac{1}{2}}$, i.e.

$$\left\|\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right\|_E = O_p\left(\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E\right). \qquad \text{(D.9)}$$

Now, from (D.2) and D.3), we have

$$n^{-2\tau}\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E^2 - O_p\left(\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|\right)\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E - R_n \leq 0,$$

which implies that $\left\|\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right\|_E = O_p\left(n^{2\tau} \max_{j\in\mathcal{S}_\beta} \widehat{P}'_{\lambda_n}(\beta_{o,j}) + n^{-\frac{1}{2}+\tau}\right)$. ■

Proof of Lemma 5.2. On the event $\left\{\widehat{\beta}_{n,j}^{\mathcal{S}} \neq 0\right\}$ for some $j \in \mathcal{S}_\beta^c$, we have the following KKT optimality condition:

$$\left| W_n^{(j)} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n^{\mathcal{S}}) \right] \right| = \frac{n^{\frac{1}{2}} \widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}^{\mathcal{S}})}{2}. \tag{D.10}$$

where $W_n^{(j)}$ denotes the j-th row of the weight matrix W_n. Note that

$$n^{-\frac{1}{2}-\tau} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n^{\mathcal{S}}) = n^{-\tau} v_n \left[\rho(Z_i, \widehat{\alpha}_n^{\mathcal{S}})\right] + n^{\frac{1}{2}-\tau} E\left[\rho(Z_i, \widehat{\alpha}_n^{\mathcal{S}})\right] \tag{D.11}$$

By Lemma 5.1, assumptions 5.1.(iv) and 5.2.(i), we get

$$n^{\frac{1}{2}-\tau} E\left[g_{n,q}(Z_i, \widehat{\theta}_n^{\mathcal{S}})\right] = n^{\frac{1}{2}-2\tau} \frac{\partial G_{n,q}(\widetilde{\theta}_n^{\mathcal{S}})}{\partial \theta'} \left(\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right) = O_p(1) \tag{D.12}$$

where

$$\frac{\partial G_{n,q}(\widetilde{\theta}_n^{\mathcal{S}})}{\partial \theta'} = \left(\begin{array}{ccc} \frac{\partial G_{n,q,1}(\widetilde{\theta}_{1,n}^{\mathcal{S}})}{\partial \theta}, & ..., & \frac{\partial G_{n,q,q}(\widetilde{\theta}_{q,n}^{\mathcal{S}})}{\partial \theta} \end{array} \right)'$$

and $\widetilde{\theta}_{j,n}^{\mathcal{S}}$ $(j = 1, ..., q)$ lies between $\widehat{\theta}_n^{\mathcal{S}}$ and θ_o, and

$$\begin{aligned} n^{\frac{1}{2}-\tau} E\left[g_k(Z_i, \widehat{\theta}_n^{\mathcal{S}}) - \widehat{\beta}_n^{\mathcal{S}}\right] &= n^{\frac{1}{2}-\tau} \left\{ E\left[g_k(Z_i, \widehat{\theta}_n^{\mathcal{S}}) - g_k(Z_i, \theta_o)\right] - \left(\widehat{\beta}_n^{\mathcal{S}} - \beta_o\right) \right\} \\ &= \frac{\partial G_k(\widetilde{\theta}_n^{\mathcal{S}})}{\partial \theta} \left[n^{\frac{1}{2}-\tau} \left(\widehat{\theta}_n^{\mathcal{S}} - \theta_o\right) \right] + O_p(1) = O_p(1) \end{aligned} \tag{D.13}$$

where

$$\frac{\partial G_k(\widetilde{\theta}_n^{\mathcal{S}})}{\partial \theta'} = \left(\begin{array}{ccc} \frac{\partial G_{k,1}(\widetilde{\theta}_{q+1,n}^{\mathcal{S}})}{\partial \theta}, & ..., & \frac{\partial G_{k,k}(\widetilde{\theta}_{q+k,n}^{\mathcal{S}})}{\partial \theta} \end{array} \right)'$$

and $\widetilde{\theta}_{j,n}^{\mathcal{S}}$ $(j = q+1, ..., q+k)$ lies between $\widehat{\theta}_n^{\mathcal{S}}$ and θ_o. Hence, from the results in (D.11)-(D.13), assumptions 3.1.(iii) and 5.1.(iii), we can deduce that

$$\left| W_n^{(j)} \left[n^{-\frac{1}{2}-\tau} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n^{\mathcal{S}}) \right] \right| = O_p(1). \tag{D.14}$$

While from assumption 5.2.(ii), we can deduce that

$$\frac{n^{\frac{1}{2}-\tau}\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{2} = \frac{n^{\frac{1}{2}-\tau}r_n\lambda_n}{2}\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{r_n\lambda_n} \to_p \infty \tag{D.15}$$

Now, the KKT condition in (D.10), and the results in (D.14) and (D.15) imply that $\Pr\left(\widehat{\beta}^{\mathcal{S}}_{n,j} = 0\right) \to 1$ for any $j \in \mathcal{S}^c_\beta$. ■

Proof of Lemma 5.3. Applying Lemma A.1, we get w.p.a.1

$$V_0(\widehat{\theta}^{\mathcal{S}}_n, \widehat{\beta}^{\mathcal{S}}_n) + \sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j}) \leq \sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_{o,j}) + R_n. \tag{D.16}$$

Now, conditional on the event $\{\widehat{\beta}^{\mathcal{S}}_{j,n} = 0, j \in \mathcal{S}^c_\beta\}$, by assumptions 5.1.(iv) and 5.3, we can use similar arguments in the proof of Lemma A.2 to deduce that

$$V_0(\widehat{\theta}^{\mathcal{S}}_n, \widehat{\beta}^{\mathcal{S}}_n) \geq c\left(1 + n^{-2\tau}\right)||\widehat{\theta}^{\mathcal{S}}_n - \theta_o||^2_E + c||\widehat{\beta}^{\mathcal{S}}_{n,+} - \beta_{o,+}||^2_E \geq c||\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||^2_E \tag{D.17}$$

w.p.a.1, where c is some generic positive constant. Following the similar arguments in the proof of Lemma 3.2, we get

$$||\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||^2 - \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|\left\|\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right\|_E \leq R_n \tag{D.18}$$

which implies that

$$||\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E = O_p\left(\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + n^{-\frac{1}{2}}\right) \tag{D.19}$$

By assumption 5.2.(i), there is $n^{\frac{1}{2}}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j,n})\right| = o_p(1)$, which combined with (D.19) gives us $||\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E = O_p(n^{-\frac{1}{2}})$. From the sparsity of $\widehat{\beta}^{\mathcal{S}}_n$, we know that the event $\{\widehat{\beta}^{\mathcal{S}}_{j,n} = 0, j \in \mathcal{S}^c_\beta\}$ has probability measure approaching 1. Hence we can deduce that $||\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E = O_p(n^{-\frac{1}{2}})$, which finishes the proof. ■

Proof of Corollary 5.2. For any compact subset K in $R^{d_\theta + d_{\mathcal{S}_\beta}}$, we denote any element $u_{\mathcal{S}} \in K$ as $u_{\mathcal{S}} = (u_\theta, u_{\beta_+})$, where u_θ is the first d_θ elements in $u_{\mathcal{S}}$ and u_{β_+} is the last $d_{\beta_{o,+}\theta}$

elements in $u_{\mathcal{S}}$. Denote

$$\begin{aligned} V_{3,n}(u_{\mathcal{S}}) &= \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho^{s}(Z_i,\alpha_{o,\mathcal{S}}+\frac{u_{\mathcal{S}}}{\sqrt{n}})\right]' W_n \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho^{s}(Z_i,\alpha_{o,\mathcal{S}}+\frac{u_{\mathcal{S}}}{\sqrt{n}})\right] \\ &\quad - \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\alpha_o)\right]' W_n \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\alpha_o)\right] \\ &\quad + n\sum_{j\in\mathcal{S}_\beta}\left[\widehat{P}_{\lambda_n}(\beta_{o,j}+\frac{u_{\beta_+,j}}{\sqrt{n}})-\widehat{P}_{\lambda_n}(\beta_{o,j})\right] := V^{*}_{3,n}(u_{\mathcal{S}})+P_n \end{aligned}$$

where $\rho^{s}(Z_i,\alpha_{o,\mathcal{S}}+\frac{u_{\mathcal{S}}}{\sqrt{n}})=\rho(Z_i,\theta_o+\frac{u_\theta}{\sqrt{n}},\beta_{o,+}+\frac{u_{\beta_+}}{\sqrt{n}},\beta_{o,-})$. From Theorem 3.3, we know that $\widehat{\beta}^{\mathcal{S}}_{n,-}=0$ w.p.a.1. Hence,$\sqrt{n}\left(\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}}-\alpha_{o,\mathcal{S}}\right)$ is the minimizer of $V_{3,n}(u_{\mathcal{S}})$ w.p.a.1. Using similar arguments in the proof of Lemma A.1, one can deduce that

$$\upsilon_n\left\{\left[\rho^{s}\left(Z,\alpha_{o,\mathcal{S}}+n^{-\frac{1}{2}}u_{\mathcal{S}}\right)-\rho\left(Z,\alpha_{o,\mathcal{S}}\right)\right]\right\}=o_p(1) \tag{D.20}$$

uniformly over K. From (D.20), we get

$$\begin{aligned} V^{*}_{3,n}(u_{\mathcal{S}}) &= n\left\{E[\rho^{s}(Z,\alpha_{o,\mathcal{S}}+n^{-\frac{1}{2}}u_{\mathcal{S}})]\right\}' W_o \left\{E[\rho^{s}(Z,\alpha_{o,\mathcal{S}}+n^{-\frac{1}{2}}u_{\mathcal{S}})]\right\} \\ &\quad +2n^{\frac{1}{2}}\left\{\upsilon_n\left[\rho(Z,\alpha_o)\right]\right\}' W_o \left\{E[\rho^{s}(Z,\alpha_{o,\mathcal{S}}+n^{-\frac{1}{2}}u_{\mathcal{S}})]\right\}+o_p(1) \end{aligned} \tag{D.21}$$

uniformly over K. By assumption 5.1.(iv), we have

$$E\left[g_{n,q}\left(Z,\theta_o+\frac{u_\theta}{\sqrt{n}}\right)\right]=n^{-\tau-\frac{1}{2}}\left(\begin{array}{cc}\frac{\partial G_{n,q}(\theta_o)}{\partial\theta'}, & 0\end{array}\right)u_{\mathcal{S}}+o(1) \tag{D.22}$$

and

$$E\left[g_k\left(Z_i,\theta_o+\frac{u_\theta}{\sqrt{n}}\right)-\left(\beta_{o,+}+\frac{u_{\beta_+}}{\sqrt{n}}\right)\right]=n^{-\frac{1}{2}}\frac{\partial m_k(\theta_o)}{\partial\alpha'_{\mathcal{S}}}u_{\mathcal{S}}+o(1). \tag{D.23}$$

Denote $\frac{\partial m(\theta_o)}{\partial\alpha'_{\mathcal{S}}}=\left(\mathbf{0},\frac{\partial m_k(\theta_o)}{\partial\alpha_{\mathcal{S}}}\right)'$, from the results in (D.21)-(D.23) and assumption 5.1.(iv), we can deduce that

$$V^{*}_{3,n}(u_{\mathcal{S}})=\left[\frac{\partial m(\theta_o)}{\partial\alpha'_{\mathcal{S}}}u_{\mathcal{S}}+2\upsilon_n\left[\rho(Z_i,\alpha_o)\right]\right]' W_o\left[\frac{\partial m(\theta_o)}{\partial\alpha'_{\mathcal{S}}}u_{\mathcal{S}}\right]+o_p(1), \tag{D.24}$$

uniformly over K. Note that there is

$$\left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right]' = \left[\frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_{o,kk} \left[\frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right]' = M_+. \tag{D.25}$$

Using the same arguments used in the proof of Theorem 3.4, we can show that under assumptions 3.2.(iv) and 5.2.(i), there is

$$P_n = \sqrt{n}\widehat{P}'_{\lambda_n}(\beta_{o,j})u_{\beta_+,j} + \widehat{P}''_{\lambda_n}(\widetilde{\beta}_{o,j})u^2_{\beta_+,j} \to 0 \tag{D.26}$$

uniformly over $u_{\beta_+,j}$ for $j \in \mathcal{S}_\beta$.

From the results in (D.24), (D.26) and triangle inequality, we can deduce that

$$V_{3,n}(u_{\mathcal{S}}) \to_d V_3(u_{\mathcal{S}}) := u'_{\mathcal{S}} M_+ u_{\mathcal{S}} + 2u'_{\mathcal{S}} \left(\frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}} W_o \Psi(\theta_o, \beta_o) \right) \tag{D.27}$$

uniformly over K. It is clear that $V_3(u_{\mathcal{S}})$ is uniquely minimized at

$$u_{\mathcal{S}} = -M_+^{-1} \left(\left[0, \frac{\partial m_k(\theta_o)}{\partial \alpha'_{\mathcal{S}}} \right] W_o \Psi(\theta_o, \beta_o) \right) \tag{D.28}$$

and $\sqrt{n}\left(\widehat{\alpha}^{\mathcal{S}}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right)$ is asymptotically tight. Now result in (5.12) follows by ACMT. ■

E Proof of the Main results in Section 6

Proof of Lemma 6.1. By Lemma 3.1 and Theorem 3.3, we know that the GMM shrinkage estimator based on the λ_n^o identifies all correct moment conditions in set-2 w.p.a.1. Hence, we can deduce that

$$\Pr\left(\inf_{\lambda \in \Omega_n^o} MSC_n(\lambda) = MSC_n(\lambda_n^o) \right) \to 1. \tag{E.1}$$

As there are only finite many candidate sets in $\mathcal{S}_F$, by definition we have

$$\inf_{\lambda \in \Omega_n^-} \frac{MSC_n(\lambda)}{n} \geq \min_{\mathcal{S}_j \in \mathcal{S}_F,\ \mathcal{S}_j \subsetneq \mathcal{S}_\beta^c} \left[\frac{\Phi_n^*(\mathcal{S}_j)}{n} - \kappa_n h(|\mathcal{S}_j|) \right],$$

which, combined with assumption 6.1.(ii), implies that

$$\begin{aligned} & \Pr\left(\inf_{\lambda\in\Omega_n^-}\frac{MSC_n(\lambda)}{n}-\frac{MSC_n(\lambda^o)}{n}>0\right) \\ \geq\ & \Pr\left(\min_{\mathcal{S}_j\in\mathcal{S}_F,\,\mathcal{S}_j\subsetneq\mathcal{S}_\beta^c}\frac{\Phi_n^*(\mathcal{S}_j)-\Phi_n^*(\mathcal{S}_\beta^c)}{n}+o(1)>0\right) \\ =\ & \Pr\left(\min_{\mathcal{S}_j\in\mathcal{S}_F,\,\mathcal{S}_j\subsetneq\mathcal{S}_\beta^c}(c_j-c_o)+o_p(1)>0\right). \end{aligned} \tag{E.2}$$

As there are only finite elements in $\mathcal{S}_F$, from assumption 6.1.(iii) we can deduce that $\min_{\mathcal{S}_j\in\mathcal{S}_F,\,\mathcal{S}_j\subsetneq\mathcal{S}_\beta^c}(c_j-c_o)>0$. Hence, from the results in (E.1)-(E.2) and assumption 6.1.(iii) we get

$$\Pr\left(\inf_{\lambda\in\Omega_n^-}MSC_n(\lambda)>MSC_n(\lambda_n^o)\right)\to 1. \tag{E.3}$$

Similarly, by definition and assumption 6.1.(iv), we can deduce that

$$\begin{aligned} & \Pr\left(\inf_{\lambda\in\Omega_n^+}MSC_n(\lambda)-MSC_n(\lambda^o)>0\right) \\ \geq\ & \Pr\left(\min_{\mathcal{S}_j\in\mathcal{S}_F,\,\mathcal{S}_j\subset\mathcal{S}_\beta^c}\left[\Phi_n^*(\mathcal{S}_j)-\kappa_n h\left(|\mathcal{S}_j|\right)\right]-\Phi_n^*(\mathcal{S}_\beta^c)+\kappa_n h\left(\left|\mathcal{S}_\beta^c\right|\right)>0\right) \\ \geq\ & \Pr\left(-\frac{1}{2}\min_{\mathcal{S}_j\in\mathcal{S}_F,\,\mathcal{S}_j\subset\mathcal{S}_\beta^c}\chi^2\left(\left|\mathcal{S}_\beta^c\right|-|\mathcal{S}_j|\right)+\kappa_n h\left(1\right)+o_p(1)>0\right)\to 1 \end{aligned} \tag{E.4}$$

From the results in (E.1) and (E.4) and assumption 6.1.(i) we get

$$\Pr\left(\inf_{\lambda\in\Omega_n^+}MSC_n(\lambda)>MSC_n(\lambda_n^o)\right)\to 1. \tag{E.5}$$

Now results in (E.3) and (E.5) imply (6.9), which finishes the proof. ■

Proof of Lemma 6.2. Note that if $\kappa_n\leq M$ for all n, then $\kappa_n=o(n)$. Hence results in (6.11) can be proved using the same arguments showing (E.3).

Next, by the definition of $MSC_n(\lambda)$, we have $\inf_{\lambda\in\Omega_n^o}MSC_n(\lambda)=MSC_n(\lambda^o)$, where λ^o is any element in Ω_n^o. Denote $d_{\lambda^o}=h\left(\left|\mathcal{S}_{\beta,\lambda^o}^c\right|\right)-h\left(0\right)$. As $\left|\mathcal{S}_{\beta,\lambda^o}^c\right|>|0|$, hence by

assumption 6.1.(i), we know that d_{λ^o} is some strictly positive number.

$$\begin{aligned}
& \Pr\left(\inf_{\lambda\in\Omega_n^o} MSC_n(\lambda) > MSC_n(0)\right) \\
= \; & \Pr\left(MSC_n(\lambda^o) - MSC_n(0) > 0\right) \\
= \; & \Pr\left(\Phi_n(\lambda^o) - \Phi_n(0) > \kappa_n d_{\lambda^o}\right) \\
= \; & \Pr\left(\chi^2\left(\left|\mathcal{S}^c_{\beta,\lambda^o}\right| - \left|\mathcal{S}^c_{\beta,\lambda^+}\right|\right) > \kappa_n d_{\lambda^o} + o_p(1)\right) \\
\geq \; & \Pr\left(\chi^2\left(\left|\mathcal{S}^c_{\beta,\lambda^o}\right| - \left|\mathcal{S}^c_{\beta,\lambda^+}\right|\right) > M d_{\lambda^o} + o_p(1)\right) \\
\to \; & \Pr\left(\chi^2\left(\left|\mathcal{S}^c_{\beta,\lambda^o}\right| - \left|\mathcal{S}^c_{\beta,\lambda^+}\right|\right) > M d_{\lambda^o}\right) := \pi.
\end{aligned}$$

As Md_{λ^o} is some finite and strictly positive constant, we can deduce that $\pi \in (0,1)$, which finishes the proof. ■

F Sufficient Conditions for the Assumptions Imposed on the Penalty Function

In this section, we check the general conditions imposed on $\widehat{P}_{\lambda_n}(\cdot)$ to derive the oracle properties of the GMM shrinkage estimators with the bridge, adaptive Lasso and SCAD penalty functions respectively.

Corollary F.1 *Suppose that $\lambda_n = o(1)$, then the bridge and SCAD penalty functions defined in (1.6) and (1.8) satisfy assumption 3.1.(iv). If we further assume that $\widehat{\beta}_n$ is a consistent estimator of β_o, then the adaptive Lasso penalty function defined in (1.7) also satisfies assumption 3.1.(iv).*

Proof of Corollary F.1. First note that if $\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma$, then trivially $\widehat{P}_{\lambda_n}(0) = 0$ and

$$\lambda_n \left|\beta_{o,j}\right|^\gamma \to_p 0$$

for all j. If $\widehat{P}_{\lambda_n}(\beta)$ is the SCAD penalty function, then $\widehat{P}_{\lambda_n}(0) = 0$ and when n is sufficiently large,

$$\left|\widehat{P}_{\lambda_n}(\beta_j)\right| \leq \frac{(a+1)\lambda_n^2}{2} = o_p(1)$$

for all j. Finally for the adaptive Lasso penalty $\widehat{P}_{\lambda_n}(\beta) = \lambda_n \widehat{w}_\beta |\beta|$, $\widehat{P}_{\lambda_n}(0) = 0$ and by the consistency of $\widehat{\beta}_n$ and the Slutsky Theorem, we can deduce that

$$\lambda_n \widehat{w}_{\beta_j} |\beta_j| \to_p 0$$

for all j. ■

Remark F.2 *Compared with the results in Knight and Fu (2000) and Caner (2009) where bridge penalty is used, Corollary F.1 imposes the same condition on λ_n to derive the consistency of the shrinkage estimator $\widehat{\alpha}_n^{\mathcal{S}}$. When the penalty function is SCAD, $\lambda_n = o(1)$ is also the sufficient condition in Fan and Li (2001) to derive the consistency. In the adaptive Lasso case, Zou (2006) derives the limit distribution of the centered adaptive Lasso LS estimator under the condition that $\lambda_n = o(1)$ and $\lambda_n n^{\frac{1+\omega}{2}} \to \infty$. As we later will impose the same conditions on λ_n to derive the limit distribution of the GMM shrinkage estimator $\widehat{\alpha}_n^{\mathcal{S}}$ based on adaptive Lasso penalty, our condition imposed on λ_n to derive the consistency is not stronger than that of Zou (2006).*

Corollary F.3 *Suppose that $\lambda_n = o(1)$, then the bridge and SCAD penalty functions satisfy assumption 3.2.(iv). If we further assume that $\widehat{\beta}_n$ is a consistent estimator of β_o, then the adaptive Lasso penalty function also satisfies assumption 3.2.(iv).*

Proof of Corollary F.3. First note that if $\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma$, then for any $\beta \neq 0$, there is

$$\widehat{P}'_{\lambda_n}(\beta) = \gamma \lambda_n \beta^{\gamma-1} \text{and } \widehat{P}''_{\lambda_n}(\beta) = \gamma(\gamma-1)\lambda_n \beta^{\gamma-2}.$$

Hence

$$\widehat{P}''_{\lambda_n}(\beta_{oj}) = \gamma(\gamma-1)\lambda_n \beta_{oj}^{\gamma-2} = o(1)$$

for any $j \in \mathcal{S}_\beta$.

Next, if $\widehat{P}_{\lambda_n}(\beta) = \lambda_n \widehat{w}_\beta |\beta|$, then for any $\beta \neq 0$, one trivially has

$$\widehat{P}'_{\lambda_n}(\beta) = \lambda_n \widehat{w}_{\beta,j} \text{ and } \widehat{P}''_{\lambda_n}(\beta) = 0.$$

So assumption 3.2.(iii) is trivially satisfied.

Finally, if $\widehat{P}_{\lambda_n}(\beta)$ is the SCAD penalty function, then on the domain $(0,\infty)$, there is

$$\widehat{P}'_{\lambda_n}(\beta_j) = \begin{cases} \lambda_n & |\beta_j| \leq \lambda_n \\ \frac{a\lambda_n}{a-1} - \frac{\beta_j}{a-1} & \lambda_n < |\beta_j| \leq a\lambda_n \\ 0 & a\lambda_n < |\beta_j| \end{cases}$$

Note that for $j \in \mathcal{S}_\beta$, when n is sufficiently large such that $a\lambda_n < \left|\beta_{o,j}\right|$, $\widehat{P}_{\lambda_n}(\beta)$ is twice continuously differentiable in local neighborhood of $\beta_{o,j}$ and we trivially have $\widehat{P}''_{\lambda_n}(\beta_{o,j}) = 0$.
■

Corollary F.4 *Suppose that assumptions 3.1.(i)-(iii) and 3.2.(i)-(iii) are satisfied. (i) If $\lambda_n n^{\frac{1}{2}} = o(1)$ and $\lambda_n n^{1-\frac{\gamma}{2}} \to \infty$, then the bridge penalty function satisfies assumptions 3.3 (i)-(ii); (ii) suppose that $\lambda_n n^{\frac{1}{2}} = o(1)$, $\lambda_n n^{\frac{1}{2}(1+\omega)} \to \infty$ and $\widehat{\beta}_n$ is $\sqrt{n}$ consistent, then the adaptive Lasso penalty function satisfies assumptions 3.3 (i)-(ii); (iii) suppose that $\lambda_n = o(1)$ and $\lambda_n n^{\frac{1}{2}} \to \infty$, then the SCAD penalty function satisfies assumption 3.3 (i)-(ii).*

Proof of Corollary F.4. First if $\widehat{P}_{\lambda_n}(\beta_j) = \lambda_n \left|\beta_j\right|^\gamma$, then

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = \max_{j \in \mathcal{S}_\beta} \gamma n^{\frac{1}{2}} \lambda_n \left|\beta_{o,j}\right|^{\gamma-1} = o(1). \tag{F.1}$$

As assumptions 3.1.(iv), 3.2.(iv) and 3.3.(i) are satisfied, from Lemma 3.2, we can deduce that $\widehat{\beta}^{\mathcal{S}}_{n,j} = O_p(n^{-\frac{1}{2}})$. Let $r_n = n^{\frac{1}{2}-\frac{\gamma}{2}}$, then $n^{\frac{1}{2}}\lambda_n r_n = \lambda_n n^{1-\frac{\gamma}{2}} \to \infty$ and

$$\liminf_{n\to\infty} \frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{r_n \lambda_n} = \liminf_{n\to\infty} \gamma \left|n^{\frac{1}{2}} \widehat{\beta}^{\mathcal{S}}_{n,j}\right|^{\gamma-1} > 0, \ a.e.$$

for any $j \in \mathcal{S}^c_\beta$.

Next, if $\widehat{P}_{\lambda_n}(\beta_j) = \lambda_n \widehat{w}_{\beta_j} |\beta_j|$, then by the consistency of $\widehat{\beta}_n$ and the Slutsky Theorem

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = \max_{j \in \mathcal{S}_\beta} n^{\frac{1}{2}} \lambda_n \widehat{w}_{\beta,j} = o_p(1).$$

Let $r_n = n^{\frac{\omega}{2}}$, then $n^{\frac{1}{2}}\lambda_n r_n = \lambda_n n^{\frac{1}{2}(1+\omega)} \to \infty$ and

$$\liminf_{n\to\infty} \frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{r_n \lambda_n} = \liminf_{n\to\infty} \left|n^{\frac{1}{2}} \widehat{\beta}_{n,j}\right|^{-\omega} > 0, \ a.e.$$

for any $j \in \mathcal{S}_\beta^c$.

Finally, if $\widehat{P}_{\lambda_n}(\cdot)$ is the SCAD penalty function, by the definition of $\widehat{P}'_{\lambda_n}(\cdot)$, it is easy to see that when n is sufficiently large

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| = 0.$$

As assumptions 3.1.(iv), 3.2.(iv) and 3.3.(i) are satisfied, from Lemma 3.2, we can deduce that $\widehat{\beta}^{\mathcal{S}}_{n,j} = O_p(n^{-\frac{1}{2}})$. Let $r_n = 1 > 0$, then $n^{\frac{1}{2}} \lambda_n r_n = \lambda_n n^{\frac{1}{2}} \to \infty$. As $\widehat{\beta}^{\mathcal{S}}_{n,j} = O_p(n^{-\frac{1}{2}})$, we have $|\sqrt{n}\widehat{\beta}^{\mathcal{S}}_{n,j}| \leq \sqrt{n}\lambda_n$ w.p.a.1. From the definition of $\widehat{P}'_{\lambda_n}(\beta)$, i.e.

$$\widehat{P}'_{\lambda_n}(\beta) = \lambda_n \left\{ I(\beta \leq \lambda_n) + \frac{(\lambda_n a - \beta)_+}{(a-1)\lambda_n} I(\beta > \lambda_n) \right\}, \tag{F.2}$$

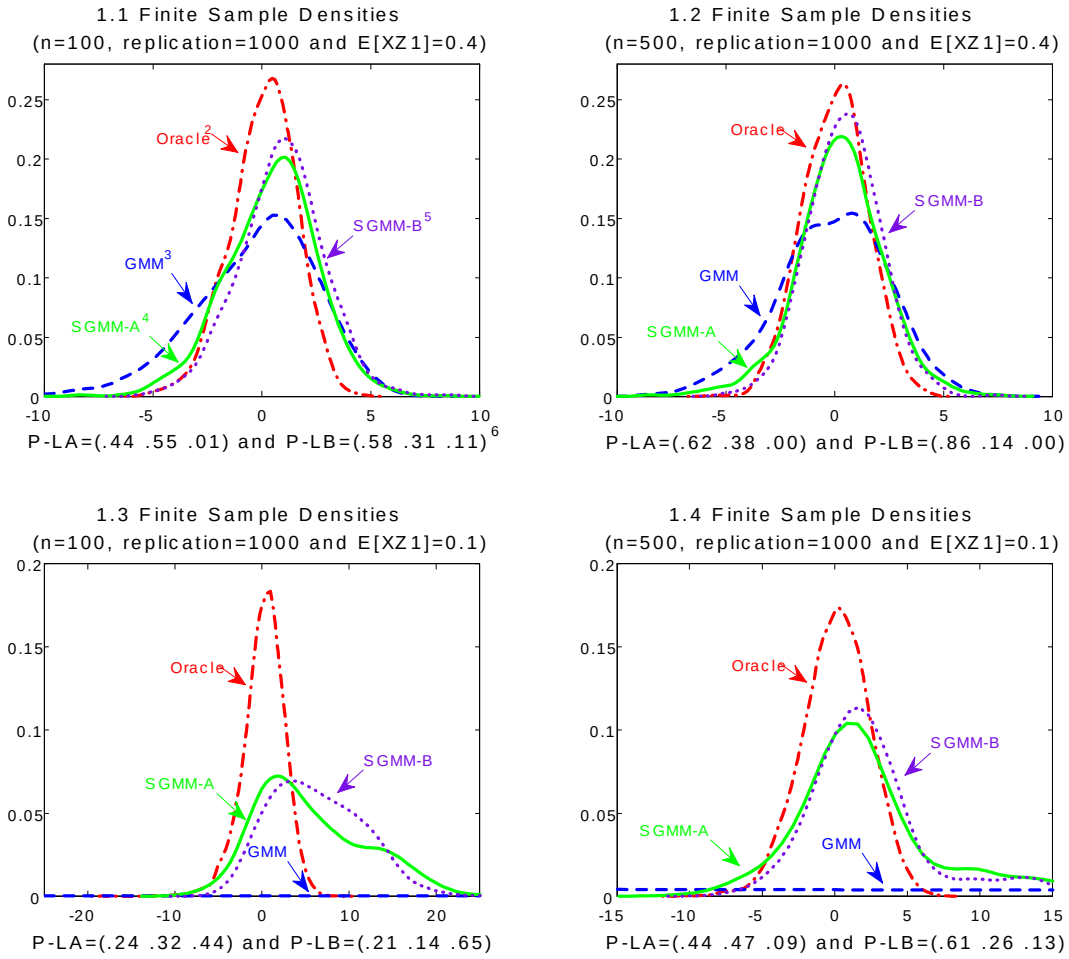
we can deduce that $\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j}) = \lambda_n$, w.p.a.1 and thus

$$\liminf_{n \to \infty} \frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}^{\mathcal{S}}_{n,j})}{r_n \lambda_n} = 1, \ a.e.$$

for all $j \in \mathcal{S}_\beta^c$. ■

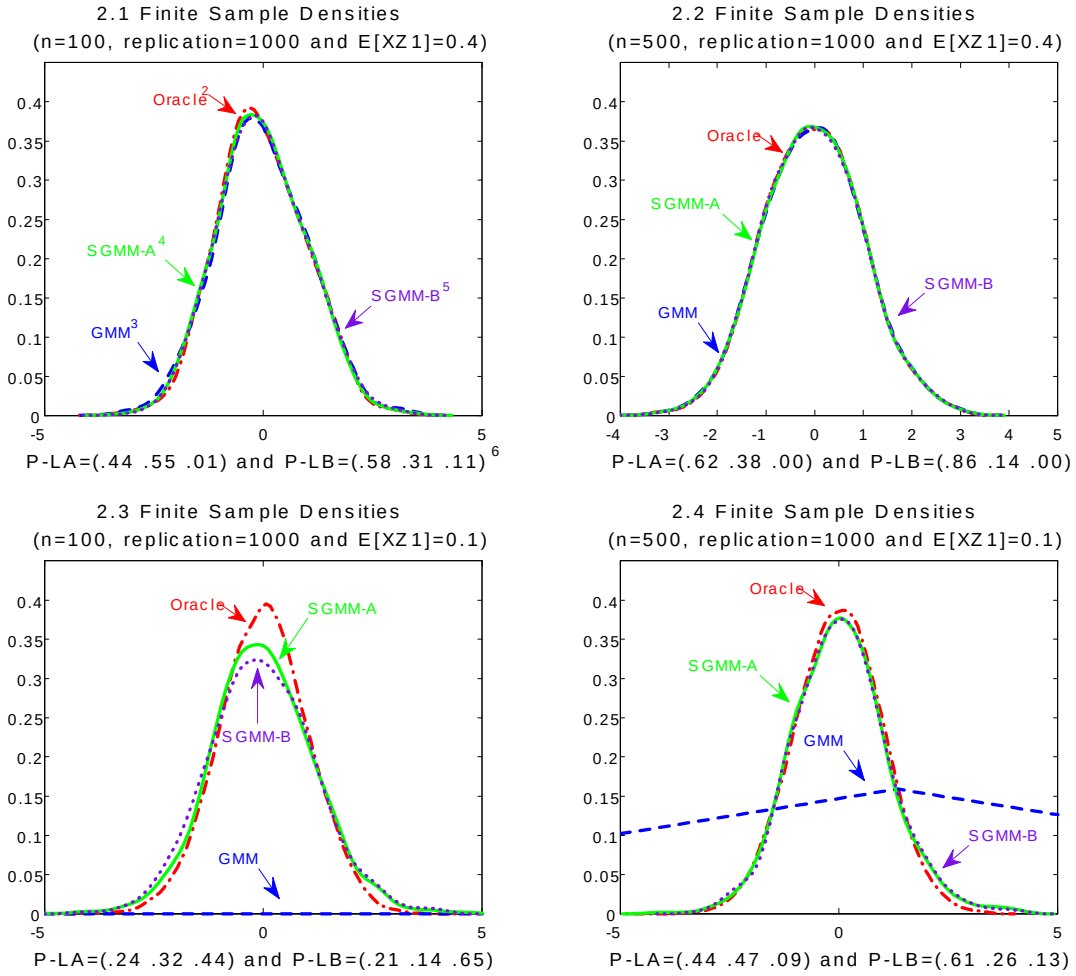
G Pictures and Figures

Figure 1. Finite Sample Densities of GMM and GMM Shrinkage Estimators for $\theta_{2,o}$[1]



1. The adaptive Lasso penalty is used in the GMM shrinkage estimation and the first-step estimators of moment selection coefficients are the GMM estimators; 2. Oracle estimators are the GMM estimators using the moment conditions in set-1 and all correct moment conditions in set-2; 3. GMM estimators only use the moment conditions in set-1; 4. SGMM-A refers the GMM shrinkage estimators using tuning parameters selected by minimizing GMM-AIC; 5. SGMM-B refers the GMM shrinkage estimators using tuning parameters selected by minimizing GMM- BIC; 6. P-LA and P-LB contain the selection probabilities of the correct, under-selected and over-selected sets of moment conditions in GMM shrinkage estimation using the tuning parameters from GMM-AIC and GMM-BIC, respectively.

Figure 2. Finite Sample Densities of GMM and GMM Shrinkage Estimators for $\theta_{1,o}$[1]



1. The adaptive Lasso penalty is used in the GMM shrinkage estimation and the first-step estimators of moment selection coefficients are the GMM estimators; 2. Oracle estimators are the GMM estimators using the moment conditions in set-1 and all correct moment conditions in set-2; 3. GMM estimators only use the moment conditions in set-1; 4. SGMM-A refers the GMM shrinkage estimators using tuning parameters selected by minimizing GMM-AIC; 5. SGMM-B refers the GMM shrinkage estimators using tuning parameters selected by minimizing GMM- BIC; 6. P-LA and P-LB contain the selection probabilities of the correct, under-selected and over-selected sets of moment conditions in GMM shrinkage estimation using the tuning parameters from GMM-AIC and GMM-BIC, respectively.

Table 1. The Probabilities of Choosing True, Consistent and Inconsistent Models[1]

n	MSC-AIC[2]	MSC-BIC[2]	MSC-HQ[2]
250	(.615 .290 .095)	(.482 .015 .503)	(.662 .076 .262)
500	(.722 .273 .005)	(.862 .007 .131)	(.898 .063 .039)
	Lasso-A[3]	Lasso-B[3]	Lasso-H[3]
250	(.584 .347 .069)	(.726 .098 .176)	(.678 .167 .155)
500	(.703 .291 .006)	(.903 .044 .053)	(.869 .104 .027)

1. Replication=1000, T=3 and the penalty function is the adaptive group Lasso. The three numbers in each bracket (from left to right) are the probabilities of true model, consistent models and inconsistent model being selected by specific model/moment selection procedure; 2. the J-test statistic is used in the construction of MSC; 3. Lasso-A, Lasso-B and Lasso-H denote GMM shrinkage estimation with the tuning parameters selected by minimizing the GMM-type of AIC, BIC and HQIC respectively.

Table 2. Bias (BS), Standard Deviations (SD) and RMSEs (RE)[1]

θ_o	BS	SD	RE	BS	SD	RE	BS	SD	RE
	GMM[2]			Lasso-A[4]			Lasso-B[4]		
θ_{1o}	.174	.434	.468	.150	.363	.393	.170	.326	.368
θ_{2o}	-.046	.109	.118	-.033	.082	.088	-.032	.063	.070
θ_{3o}	-.006	.061	.061	-.036	.072	.081	-.059	.078	.099
θ_{4o}	.012	.065	.066	.005	.040	.040	.001	.021	.021
	Oracle[3]			Lasso-H[4]			PGMM[5]		
θ_{1o}	.028	.206	.208	.164	.329	.367	.029	.317	.318
θ_{2o}	-.004	.034	.035	-.032	.068	.075	-.006	.067	.067
θ_{3o}	-.004	.059	.059	-.053	.078	.094	-.020	.073	.076
θ_{4o}	–	–	–	.002	.029	.029	.001	.029	.029

1. Sample size n=250, T=3 and replication=1000. The penalty function is the adaptive group Lasso; 2. GMM refers to the GMM estimation based on the overfitted model and set-1 moment conditions, 3. Oracle refers to the GMM estimation based on the true model with all correct moment conditions; 4. Lasso-A, Lasso-B and Lasso-H denote GMM shrinkage estimation with the tuning parameters selected by minimizing the GMM-type of AIC, BIC and HQIC respectively; 5. PGMM denotes the GMM estimation based on the model and moment conditions selected by the GMM shrinkage method in Lasso-B.

Table 3. Bias (BS), Standard Deviations (SD) and RMSEs (RE)[1]

θ_o	BS	SD	RE	BS	SD	RE	BS	SD	RE
	GMM			Lasso-A[5]			Lasso-B[3]		
θ_{1o}	.090	.294	.308	.079	.220	.234	.089	.183	.204
θ_{2o}	-.022	.073	.076	-.017	.048	.051	-.017	.035	.039
θ_{3o}	-.009	.047	.048	-.023	.048	.053	-.035	.055	.065
θ_{4o}	.006	.041	.041	.003	.022	.023	.001	.010	.010
	Oracle			Lasso-H[5]			PGMM[4]		
θ_{1o}	.020	.135	.137	.085	.192	.210	.017	.174	.175
θ_{2o}	-.003	.023	.024	-.017	.038	.042	-.004	.036	.036
θ_{3o}	-.007	.045	.046	-.030	.051	.060	-.011	.050	.051
θ_{4o}	–	–	–	.001	.014	.014	.002	.015	.015

1. Sample size n=500, T=3 and replication=1000. The penalty function is the adaptive group Lasso; 2. GMM refers to the GMM estimation based on the overfitted model and set-1 moment conditions, 3. Oracle refers to the GMM estimation based on the true model with all correct moment conditions; 4. Lasso-A, Lasso-B and Lasso-H denote GMM shrinkage estimation with the tuning parameters selected by minimizing the GMM-type of AIC, BIC and HQIC respectively; 5. PGMM denotes the GMM estimation based on the model and moment conditions selected by the GMM shrinkage method in Lasso-B.

Table 4. GMM Estimation for the Labor Supply Equation[1]

IV:	$\Delta \log w_{i,t}$		MaCurdy (1981) [2]		Altonji (1986) [3]	
	(1)	(2)	(3)	(4)	(3)	(4)
a	-.0008	-.0152	-.0123	-.0233	-.0125	-.0217
	(.0040)	(.0189)	(.0069)	(.0204)	(.0061)	(.0212)
θ_o	-.3937	-.3973	.0703	.2743	.1638	.2032
	(.0276)	(.0278)	(.2730)	(.4396)	(.1967)	(.2234)
d_t?[4]	No	Yes	No	Yes	No	Yes

1. Standard errors are in parentheses and the sample size n=3487; 2. the moment conditions are constructed using following IVs: father's education, parents' economic status when the individual was young, education, education square, age and the interaction between age and education; 3. the moment condition is constructed using an alternative measure of hour wage; 4. dt refers to the set of time dummy variables for the years from 1971 to year 1981.

Table 5. GMM and GMM Shrinkage Estimation for the Labor Supply Equation[1]

	GMM[3]	GMM[3]	SGMM[3]	SGMM[3]	P-GMM[4]	P-GMM[4]
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	-.0159	-.0224	-.0119	-.0228	-.0167	-.0190
	(.0257)	(.0236)	(.0050)	(.0198)	(.0197)	(.0195)
$\Delta \log(w_{i,t})$	.3241	.2682	.0699	.1830	.1396	.1724
	(1.225)	(1.143)	(.1472)	(.1921)	(.1454)	(.1395)
d_t?[2]	No	Yes	No	Yes	No	Yes
n	3487	3487	3487	3487	3487	3487

1. Standard errors are in parentheses and sample size n=3487. 2. dt refers to the set of time dummy variables for the years from 1971 to year 1981. 3. GMM is the GMM estimation only using the moment conditions in set-1. 4. SGMM denotes the GMM shrinkage estimation based on the adaptive Lasso penalty. In column (3) the tuning parameter equals to 0.000951 and in column (4) the tuning parameter equals to 0.000378. GMM-AIC, GMM-BIC and GMM-HQ produce the same number of the tuning parameter in each case. 4. P-GMM denotes the GMM estimation based on the moment conditions selected by the GMM shrinkage estimation. The results in columns (5) and (6) are based on the moment conditions selected in (3) and (4) respectively.

Table 6. GMM and GMM Shrinkage Estimators of the Moment Selection Coefficients [1]

IVs	edu	$edu\hat{}2$	edu_f	age	edu*age	$w^*_{i,t}$	$w_{i,t}$
β_o	$\beta_{1,o}$	$\beta_{2,o}$	$\beta_{3,o}$	$\beta_{4,o}$	$\beta_{5,o}$	$\beta_{6,o}$	$\beta_{7,o}$
GMM^2	-.0017	-.0142	-.0020	.0047	-.0120	-.0009	-.0440
	(.0072)	(.0516)	(.0059)	(.1144)	(.4693)	(.0073)	(.0998)
$SGMM^{2,4}$	.0000	.0000	.0000	.0000	.0000	.0000	-.0308
	(.0000)	(.0000)	(.0000)	(.0000)	(.0000)	(.0000)	(.0098)
GMM^3	-.0032	.0252	-.0021	.0353	.0389	-.0016	-.0494
	(.0096)	(.0557)	(.0061)	(.2164)	(.6347)	(.0092)	(.1140)
$SGMM^{3,5}$	.0000	.0000	.0000	.0000	.0000	.0000	-.0272
	(.0000)	(.0000)	(.0000)	(.0000)	(.0000)	(.0000)	(.0079)

1. Standard errors are in parentheses and n=3487. 2. GMM estimation with the time dummy variables. 3. GMM estimation without time dummy variables. 4. GMM shrinkage estimation with time dummy variables, where the penalty function is adaptive Lasso and the tuning parameter equals to 0.000378 (selected by GMM-AIC, GMM-BIC and GMM-HQ); 5. GMM shrinkage estimation without time dummy variables, where the penalty function is adaptive Lasso and the tuning parameters equals to 0.000951 (selected by GMM-AIC, GMM-BIC and GMM-HQ).