Shrinkage Methods for Automated Econometric Model Determination

A Dissertation
Presented to the Faculty of the Graduate School
of
Yale University
in Candidacy for the Degree of
Doctor of Philosophy

by
Zhipeng Liao

Dissertation Director: Professor Peter C.B. Phillips and
Professor Xiaohong Chen

March 2012

Abstract

Shrinkage Methods for Automated Econometric Model Determination

Zhipeng Liao

2012

The first chapter of my dissertation 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 a simple data-driven procedure for selecting the tuning parameters of the penalty function. Several extensions are also studied. First, I establish the 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. Second, I show that this shrinkage technique can be used in GMM to perform grouped variable selection and moment selection simultaneously. 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. Moreover, my estimate of the labor supply elasticity enjoys the smallest standard error with its value being very close to those in the literature.

The moment selection methods proposed in the first chapter focus only on the orthogonality restriction of a moment condition. As a result, the moment conditions which satisfy the orthogonality restriction but fail to improve the efficiency of the GMM estimate may be selected by the GMM shrinkage estimation with high probability in the finite samples. Such moment conditions are called irrelevant moment conditions. Irrelevant moment conditions may enlarge the finite sample bias of the GMM estimate, although they do not affect its asymptotic properties. Chapter 2 provides a new adaptive penalty that serves as the basis for which the GMM shrinkage estimation can consistently select the correctly specified and relevant moment conditions. As a result, misspecified and irrelevant moment conditions are not selected with probability approaching 1 and the GMM shrinkage estimate is not only asymptotically efficient but also robust against the irrelevant moment conditions in finite samples.

The third chapter studies the joint determination of cointegrating rank and autoregressive lag order in vector error correction (VEC) models. In cointegrated system modeling, empirical estimation typically proceeds in a stepwise manner that involves the determination of cointegrating rank and autoregressive lag order in a reduced rank vector autoregression followed by estimation and inference. This chapter proposes an automated approach to cointegrated system modeling that uses adaptive shrinkage techniques to estimate VEC models with unknown cointegrating rank structure and unknown transient lag dynamic order. These methods enable simultaneous order estimation of the cointegrating rank and autoregressive order in conjunction with oracle-like efficient estimation of the cointegrating matrix and transient dynamics. As such, they offer considerable advantages to the practitioner as an automated approach to the estimation of cointegrated systems. This chapter develops the new methods, derives their limit theory and reports simulation results.

Contents

Acknowledgements

I would like to express my sincere gratitude to Donald Andrews, Xiaohong Chen, Peter Phillips and Edward Vytlacil. I was amazingly fortunate to have them as my committee members. Their guidance, support and friendship have given me the most valuable experience during my years at Yale. I am deeply indebted to Peter and Xiaohong who encouraged me to explore on my own, and were always there whenever I needed advice. I hope that one day I will become as good an advisor to my students as they have been to me.

I have benefited from discussions with Joseph Altonji, Xu Cheng, Jinyong Hahn, Yuichi Kitamura, Oliver Linton, Taisuke Otsu and Xiaoxia Shi. I also appreciate the helpful comments from the seminar participants in all the schools where the first chapter of the dissertation was presented as a job market paper. I want to thank Tom Lewis for proofreading some parts of my dissertation. I am also grateful to Pam O'Donnell for all the care and help she has given me over the last five years. Financial support from a Carl Arvid Anderson Prize of the Cowles Foundation is acknowledged.

I would also like to thank my parents whose love and support have brought me to where I am today. Finally I want to thank my wife Ruige and my son Henry for their love, patience and understanding and for the endless joy they bring to my life. This work is dedicated to them.

Introduction

Model selection presents a primary challenge in all applied econometric research. Sometimes the choice is between competing models, sometimes the choice involves the restrictions to be imposed, sometimes it can involve the selection of moment conditions – as in generalized method of moment (GMM) estimation – and sometimes it may be the number of equilibrium relations (or cointegrating rank) as in an error correction model (ECM). The danger of misspecification from the imposition of an incorrect choice is biased and inconsistent empirical results, while the risk from over-specification is loss of information and inefficiency in estimation.

In my dissertation, I propose a new automated approach to address selection issues of this type. My method employs adaptive shrinkage techniques that have the interesting and novel property that they can perform consistent model selection and efficient estimation simultaneously. In effect, potential problems of misspecification and inefficiency are resolved by successfully shrinking an unrestricted specification towards the correct restricted specification in a finite sample of data.

The first chapter of my dissertation addresses the issue of selecting valid moment conditions from a set that includes both correctly and incorrectly specified moment conditions. Suppose that there is a set of correctly specified moment conditions (set-1) to identify unknown parameters θ_o and at the same time, one has another set of possibly misspecified moment conditions (set-2). Using only the moment conditions

in set-1 ignores the information contained in the potentially valid moment conditions of set-2, while including the misspecified moment conditions of set-2 into GMM will lead to inconsistent estimation. This chapter proposes a method for simultaneously selecting the correct moment conditions in set-2 and efficiently estimating θ_o.

To proceed, I reparametrize the moment functions in set-2 by taking deviations from their expectations. The parameterized expectations β_o are treated as potential nuisance parameters, whose zero/nonzero components signal correct/incorrect moment conditions corresponding to those in set-2. I estimate (θ_o, β_o) by minimizing a GMM shrinkage criterion. This construction adds a penalty function of β_o to the usual GMM criterion of the moment functions. In the estimation, any zero components in β_o are shrunk to zero by the penalty function and this information about exclusion of the corresponding moment functions is used in estimating θ_o.

I show that my method can consistently select the correct moment conditions and, at the same time, the estimator can achieve the semi-parametric efficiency bound implied by all correct moment conditions. I establish similar properties of my method in scenarios where there are grouped variable and moment selection formulations and when there are potentially weak moment conditions in set-1. Moreover, I provide simple data-driven procedures for selecting the tuning parameters in the penalty function, thereby making the procedures fully adaptive for empirical implementation. I apply my method to estimate the life-cycle labor supply equation, where wage is assumed to be endogenous. Using parents' economic status as a valid IV, my method selects the IVs used in MaCurdy (1981) and Altonji (1986), while discarding wage as an IV for itself. My estimators of the labor supply elasticity have the smallest variance and their values are close to those in the literature.

The second chapter of my dissertation studies the GMM shrinkage estimation with ℓ-1 type of penalty functions, which includes the Lasso and adaptive Lasso

penalty functions as special examples. I show that the GMM Lasso estimation is conservative in moment selection. A similar result is established for the GMM adaptive Lasso estimation, when the tuning parameter converges to zero fast enough. Both the consistent moment selection procedures proposed in chapter 1 and the conservative moment selection methods presented in this chapter focus only on the orthogonality restriction of a moment condition, ignoring the possibility that the selected moment conditions may contain no information about the structural coefficient θ_o. The uninformative moment conditions are called irrelevant moment conditions, because they fail to improve the efficiency of the GMM estimate.

The key difference between the GMM Lasso and GMM adaptive Lasso estimations gives us the inspiration for devising a new adaptive penalty to ensure that the valid and relevant moment conditions are consistently selected in the GMM shrinkage estimation. The new adaptive penalty depends on a measure of information contained in the moment condition. I show that such an information measure can be consistently estimated and its estimate is termed the empirical information measure. The new adaptive penalty is constructed as a product of a power function of the empirical information measure and the adaptive Lasso penalty. Under certain regularity conditions, I show the GMM shrinkage estimation based on the new adaptive penalty is consistent in selecting the valid and relevant moment conditions. As a result, the misspecified and irrelevant moment conditions are not selected with probability approaching 1, the GMM shrinkage estimate is robust against the irrelevant moment conditions in the finite samples and it is asymptotically efficient in the large samples.

The third chapter (joint with Peter C. B. Phillips) of my dissertation studies the joint determination of cointegrating rank and autoregressive lag order in vector error correction (VEC) models. In cointegrated system modeling, empirical estimation

typically proceeds in a stepwise manner that involves the determination of cointegrating rank and autoregressive lag order in a reduced rank vector autoregression followed by estimation and inference. This chapter proposes an automated approach to cointegrated system modeling that uses adaptive shrinkage techniques to estimate vector error correction models with unknown cointegrating rank structure and unknown transient lag dynamic order.

We first study the cointegration rank selection and efficient estimation in a simple first-order VEC model with *iid* innovation term. The cointegrating matrix is estimated as the minimizer of the penalized generalized least square (GLS) criterion, which is constructed by attaching adaptive Lasso penalty functions of the eigenvalues of the cointegrating matrix to the GLS criterion. Under some regularity conditions on the cointegrating matrix, we show that the GLS shrinkage estimation can consistently select the cointegration rank. More importantly, the GLS shrinkage estimate has the same asymptotic distribution as the oracle reduced rank regression (RRR) estimate informed by knowledge of the true rank (see, e.g., Phillips, 1998 and Anderson, 2002). We extend the result of consistent cointegration selection to a scenario in which the innovation term is weakly dependent. In this case, the cointegrating matrix can not be consistently estimated, but we show that the GLS shrinkage estimation is consistent in cointegration rank selection, even if the rank of the probability limit of the LS shrinkage estimate of the cointegrating matrix is strictly less than that of the true cointegrating matrix. As a further extension, we show that in the general VEC models our methods enable simultaneous order estimation of the cointegrating rank and autoregressive order in conjunction with oracle-like efficient estimation of the cointegrating matrix and transient dynamics. As such they offer considerable advantages to the practitioner as an automated approach to the estimation of cointegrated systems.

Chapter 1

Adaptive GMM Shrinkage Estimation with Consistent Moment Selection

1.1 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 chapter 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 $g_q(\cdot,\cdot): R^{d_z}\times R^{d_\theta}\rightarrow R^q$. Suppose there is another set of possibly misspecified moment conditions

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

where "$\stackrel{?}{=}$" 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 $g_k(\cdot,\cdot): R^{d_z}\times R^{d_\theta}\rightarrow R^{k1}$. 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 chapter 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 (1958) and Eichenbaum, Hansen, and Singleton (1988)) to check the validity of the set-2 moment conditions. In addition to the Sargan/Hansen 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

[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.

J-test statistic and shows that consistent moment selection [2] 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 chapter embeds the moment selection in GMM estimation and once a 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.

The automatic moment selection method proposed in this chapter is closely related to the Lasso-type of variable selection methods in the statistics literature. This chapter shows how the Lasso-type of variable selection techniques can be generalized in a GMM framework to perform moment selection. Under some regularity conditions, we show that the penalized GMM estimator (which is called as GMM shrinkage estimator thereafter in this chapter) is root-n consistent and asymptotically normal. Moreover, we show that consistent moment selection is automatically achieved in the penalized GMM estimation and the GMM shrinkage estimator is asymptotically

[2] In this chapter, we call a moment selection procedure is consistent if it can select the set of valid moment conditions in set-2 with probability approaching 1.

oracle-efficient (i.e. as efficient as the GMM estimator based on all potentially valid moment conditions). As an extension of the main results of this chapter, 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 $\sqrt{n}$ (see e.g. Hahn and Kuersteiner (2002)). However, we show that if there are potentially valid and strong moment conditions in set-2, the GMM shrinkage estimator can retain the $\sqrt{n}$ convergence rate. As an another extension, we show that the shrinkage technique can be used in GMM to perform grouped variable selection and moment selection simultaneously. For the empirical implementation, we provide simple and data-dependent tuning parameters which are easy to compute in practice. The results from our simulation studies and empirical application show that the GMM shrinkage estimation based on the data-driven tuning parameters works well in terms of the moment selection and finite sample properties of the GMM shrinkage estimator.

There are some recent works in the econometrics literature which are related to this chapter. In the linear instrumental variable (IV) models, Belloni, Chernozhukov and Hansen (2011) and Belloni, Chen, Chernozhukov and Hansen (2011) apply the Lasso-type of estimation to the first-stage high dimensional reduced form equations. They show the optimal IV can be well approximated by the selected IVs from the first-stage Lasso-type of estimation and the resulting IV estimators based on these selected IVs are asymptotically oracle-efficient. This chapter is different from the above two papers, because when specified in linear IV models, our goal is to distinguish the potentially valid IVs from the invalid ones. To the best of our knowledge, this chapter is the first work to show how the Lasso-type technique can be used in GMM to select potentially valid moment conditions. In a more recent work, Gautier and Tsybakov (2011) propose a Danzig selector based IV estimator in high dimensional models.

They derive an upper bound of estimation error for the structural coefficients under the assumption that all IVs are valid. When there exist consistent estimates of the structural coefficients, Gautier and Tsybakov (2011) also derive an upper bound of estimation error for the moment selection coefficients (see the definition in Section 1.2) for the invalid IVs.

The rest of this chapter is organized as follows. Section 1.2 describes our method and gives some examples from applied econometrics, which arise naturally from the framework of this chapter. Section 1.3 establishes the main asymptotic properties of the GMM shrinkage estimators. We show that our method can consistently select the valid moment conditions in set-2 and the GMM shrinkage estimator is asymptotically oracle-efficient. We also give the data-dependent tuning parameter in this section. Section 1.4 studies the GMM shrinkage estimation with nearly weak moment conditions in set-1. Section 1.5 investigates the grouped variable selection and moment selection method using an adaptive group Lasso penalty in the GMM shrinkage estimation. Section 1.6 conducts several Monte Carlo experiments to check the finite sample properties of our method. Section 1.7 applies our method to a life-cycle labor supply model to illustrate how the GMM shrinkage method works with real data. Section 1.8 concludes this chapter. Proofs and technical derivations are included in Appendix 1.9.

1.2 GMM Shrinkage Estimation and Some Examples

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 (or misspecified), then $\beta_{o,j}=0$ (or $\beta_{o,j}\neq 0$). 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 β_o[3].

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

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

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

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

where $\Theta\times\mathcal{B}$ is the parameter space where (θ_o,β_o) lies; W_n is a $(q+k)\times(q+k)$

[3]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 chapter treats β_o as unknown nuisance parameters 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.

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 $\widehat{P}_{\lambda_n}(\cdot)$ in the statistics literature. For example, the bridge penalty is defined as $\widehat{P}_{\lambda_n}(\beta) = \lambda_n |\beta|^\gamma$, where $\gamma \in (0, 1)$; the adaptive Lasso penalty is defined as $\widehat{P}_{\lambda_n}(\beta) = \lambda_n \widehat{w}_\beta |\beta|$, where $\widehat{w}_\beta = |\widehat{\beta}_{1st}|^{-\omega}$ $(\omega > 0)$ and $\widehat{\beta}_{1st}$ is some first-step consistent estimator of β_o; and the smoothly clipped absolute deviation (SCAD) penalty is defined as

$$\widehat{P}_{\lambda_n}(\beta) = \begin{cases} \lambda_n |\beta| & |\beta| \leq \lambda_n \\ \frac{\lambda_n a |\beta|}{a-1} - \frac{\beta^2 + \lambda_n^2}{2(a-1)} & \lambda_n < |\beta| \leq a\lambda_n \\ \frac{(a+1)\lambda_n^2}{2} & a\lambda_n < |\beta| \end{cases}, \tag{1.6}$$

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 [4]. 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 [5]. 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.

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

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_{\beta,n} = \mathcal{S}_\beta\right) = 1 \tag{1.7}$$

and

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_o\right) \to_d N(0, \Sigma^*), \tag{1.8}$$

where Σ^* is the semi-parametric efficiency bound, implied by all the correct moment conditions. The results in (1.7) and (1.8) imply both consistent moment selection

[4] 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.

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

and efficient estimation.

We use two well-known examples to illustrate the moment selection issue and show how our method can be applied in practice. 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.

Example 1.2.1 (Dynamic Panel Model) *Consider the following dynamic panel model with fixed effects*

$$\Delta Y_{i,t} = Y_{i,t-1}\theta_{1,o} + X'_{i,t}\theta_{2,o} + \varepsilon_{i,t} \text{ and } \varepsilon_{i,t} = \nu_i + u_{i,t}, \tag{1.9}$$

where $|\theta_{1,o}| < 1$, $Y_{i,t}$ *is the dependent variable of individual* 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,* ν_i *is an unobservable individual effect and* $u_{i,t}$ *is a time varying error term. Under the assumption*

$$E[u_{i,t}Y_{i,0}] = 0,\ E[u_{i,t}u_{i,s}] = 0 \text{ and } E\left[(X_{i,1}, ..., X_{i,t-1})u_{i,t}\right] = 0 \tag{1.10}$$

for all i, t *and* $t \neq s$, *one can estimate the equation in (1.9) using GMM with the following moment conditions*

$$E[(Y_{i,0}, ..., Y_{i,t-2})\Delta\varepsilon_{i,t}] = 0 \text{ and } E[(X_{i,1}, ..., X_{i,t-2})\Delta\varepsilon_{i,t}] = 0. \tag{1.11}$$

However, Arellano and Bover (1995) argues that the GMM estimate based on (1.11)

may suffer from large finite-sample bias, because the moment conditions in (1.11) are weak if $\theta_{1,o}$ is close to zero. They suggest to estimate the equation in (1.9) by GMM with moment conditions in (1.11) and the following moment conditions

$$E[\varepsilon_{i,t}\Delta Y_{i,t-1}] \stackrel{?}{=} 0, \tag{1.12}$$

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. In this example, the moment conditions in (1.11) and (1.12) are our set-1 and set-2 moment conditions respectively. We can use our method to pick up the valid moment conditions in (1.12) and automatically include them into the estimation of $(\theta_{1,o}, \theta_{2,o})$.

Our second example is the linear 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.

Example 1.2.2 (Linear IV Model) *Consider the following model*

$$\begin{aligned} Y_i &= X_i\theta_{1,o} + W_i'\theta_{2,o} + u_i, & (1.13)\\ X_i &= Z_{1,i}\pi_{1,o} + Z_{2,i}'\pi_{2,o} + W_i'\delta_o + v_i, & (1.14) \end{aligned}$$

where Y_i, X_i are scaler endogenous variables, W_i contains a set of exogenous variable

and $Z_{1,i}$ denotes an instrumental variable for X_i. Suppose that the following moment conditions hold

$$E[u_i Z_{1,i}] = 0 \text{ and } E[u_i W_i] = 0 \tag{1.15}$$

which can be used to identify and consistently estimate $\theta_o = (\theta_{1,o}, \theta_{2,o})$. $Z_{2,i}$ can also be a valid IV for estimating θ_o under the following condition $E[u_i Z_{2,i}] \stackrel{?}{=} 0$. Moreover, if X_i 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 (1.15) and the following possibly misspecified moment conditions $E[u_i X_i] \stackrel{?}{=} 0$. In this example, set-1 moment conditions are in (1.15) and the moment conditions constructed using $Z_{2,i}$ and X_i are in set-2. Our method can be used to check whether $Z_{2,i}$ are valid IVs. Moreover, our method can automatically detect the endogeneity of X_i. The GMM shrinkage estimator will asymptotically become the GLS estimator if the moment condition $E[u_i X_i] \stackrel{?}{=} 0$ is valid.

1.3 Asymptotic Properties of the GMM Shrinkage Estimator

This section establishes the asymptotic 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 $\widehat{\alpha}_n = (\widehat{\theta}_n, \widehat{\beta}_n)$. We use $\|\cdot\|_E$ to denote the Euclidean norm in the Euclidean space.

1.3.1 Consistency and the Rate of Convergence

We first present and discuss the sufficient conditions for consistency of $\widehat{\alpha}_n$. 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 1.3.1 *(i)* $E[g_k(Z,\theta)]$ *is continuous in* θ *and for any* $\varepsilon > 0$*, there exists some* $\delta_\varepsilon > 0$ *such that*

$$\inf_{\{\theta\in\Theta:\|\theta-\theta_o\|_E\geq\varepsilon\}} \|E[g_q(Z,\theta)]\|_E > \delta_\varepsilon; \tag{1.16}$$

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

$$\sup_{\theta\in\Theta}\left[n^{-1}\sum_{i=1}^{n}\{g_l(Z_i,\theta) - E[g_l(Z_i,\theta)]\}\right] = o_p(1) \tag{1.17}$$

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

$$W_n \to_p W_o; \tag{1.18}$$

(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 (1.16) in Assumption 1.3.1.(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 (1.16) and the continuity of $E[g_k(Z,\theta)]$. Assumption 1.3.1.(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 applica-

ble to models with a general data structure and general moment functions (e.g., non-smooth moment functions). Assumption 1.3.1.(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 1.3.1.(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 1.3.1.(iv) in Appendix 1.9.

Lemma 1.3.1 *Under Assumption 1.3.1, the GMM shrinkage estimator is consistent, i.e.,* $(\widehat{\theta}_n, \widehat{\beta}_n) \to_p (\theta_o, \beta_o)$.

From the consistency of $\widehat{\beta}_n$, we can deduce that if $j \in \mathcal{S}_\beta$, then $\widehat{\beta}_{n,j}$ 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 1.3.1, 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 $\widehat{\beta}_n$, because what we need to show is $\widehat{\beta}_{n,j}$ $(j \in \mathcal{S}_\beta^c)$ concentrates on zero w.p.a.1, while the consistency only indicates that $\widehat{\beta}_{n,j}$ $(j \in \mathcal{S}_\beta^c)$ concentrates on local neighborhoods of zero w.p.a.1.

Remark 1.3.2 *In Corollary 1.10.1 in Appendix 1.10, we show that if* $\lambda_n = o(1)$, *then the bridge, adaptive Lasso and SCAD penalty functions satisfy Assumption 1.3.1.(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 $(\widehat{\theta}_n, \widehat{\beta}_n)$ *satisfies* $\widehat{\beta}_n \to_p 0$ *and*

$$\widehat{\theta}_n \to_p \arg\min_{\theta \in \Theta} E\left[\rho(Z,\theta,0)\right]' W_o E\left[\rho(Z,\theta,0)\right]. \tag{1.19}$$

From (1.19), we see that $(\widehat{\theta}_n, \widehat{\beta}_n)$ *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*

$$\widehat{\alpha}_n \to_p \arg\min_{\alpha \in \mathcal{A}} \left\{ E\left[\rho(Z,\alpha)\right]' W_o E\left[\rho(Z,\alpha)\right] + \lambda_0 \sum_{j=1}^{k} |\beta_j|^{\gamma} \right\}, \tag{1.20}$$

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

$$\widehat{\alpha}_n \to_p \arg\min_{\alpha \in \mathcal{A}} \left\{ E\left[\rho(Z_i,\alpha)\right]' W_o E\left[\rho(Z_i,\alpha)\right] + \sum_{j=1}^{k} P_{\lambda_0}(\beta_j) \right\}, \tag{1.21}$$

where

$$P_{\lambda_0}(\beta_j) = \begin{cases} \lambda_0 |\beta_j| & |\beta_j| \le \lambda_0 \\ \frac{\lambda_0 a |\beta_j|}{a-1} - \frac{\beta_j^2 + \lambda_0^2}{2(a-1)} & \lambda_0 < |\beta_j| \le a\lambda_0 \\ \frac{(a+1)\lambda_0^2}{2} & a\lambda_0 < |\beta_j| \end{cases}. \tag{1.22}$$

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

$$\widehat{\alpha}_n \to_p \arg\min_{\alpha \in \mathcal{A}} \begin{cases} E\left[\rho(Z,\alpha)\right]' W_o E\left[\rho(Z,\alpha)\right] + \lambda_0 \sum_{j \in \mathcal{S}_\beta} \frac{|\beta_j|}{|\beta_{o,j}|^{\omega}} & if\ \beta_j = 0\ \forall j \in \mathcal{S}_\beta^c \\ \infty & otherwise \end{cases}. \tag{1.23}$$

From the results in (1.20), (1.21) and (1.23), we see that if $\lambda_n \to \lambda_0 \in (0,\infty)$*,*

then $(\widehat{\theta}_n, \widehat{\beta}_n)$ *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* $(\widehat{\theta}_n, \widehat{\beta}_n)$.

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

Assumption 1.3.2 *(i) The following functional central limit theorem (FCLT) holds*

$$\sup_{\theta \in \Theta} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \{ g_l(Z_i, \theta) - E\left[g_l(Z_i, \theta) \right] \} \right] = O_p(1) \tag{1.24}$$

for $l = q, k$*; (ii)* $E[g_l(Z, \theta)]$ *is continuously differentiable in some neighborhood of* θ_o *for* $l = q, k$*; (iii)* $\partial E[g_q(Z, \theta_o)] / \partial \theta'$ *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*

$$\left| \widehat{P}''_{\lambda_n}(\beta_{o,j}) \right| = o_p(1) \tag{1.25}$$

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

Assumption 1.3.2.(i) is a high-level condition, which can be verified by applying the Donsker's theorem in specific models. Assumption 1.3.2.(ii) imposes a local differentiability condition on the expectation of the moment function $g_l(Z, \theta)$ $(l = q, k)$. Assumption 1.3.2.(iii) is a local identification condition for θ_o. If this assumption fails, the resulting estimator $(\widehat{\theta}_n, \widehat{\beta}_n)$ may not be $\sqrt{n}$-consistent. Assumption 1.3.2.(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 1.3.2.(iv) in Appendix 1.10.

Lemma 1.3.3 *Under the Assumption 1.3.1 and Assumption 1.3.2, we have*

$$(\widehat{\theta}_n, \widehat{\beta}_n) = (\theta_o, \beta_o) + O_p(\delta_n) \tag{1.26}$$

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

It is interesting to see that the convergence rate of $(\widehat{\theta}_n, \widehat{\beta}_n)$ 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 $(\widehat{\theta}_n, \widehat{\beta}_n)$ is of the order δ_n.

Remark 1.3.4 *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* $(\widehat{\theta}_n, \widehat{\beta}_n)$, *i.e.* $\lambda_n = o(1)$, *is insufficient to deduce that* $(\widehat{\theta}_n, \widehat{\beta}_n)$ *is* $\sqrt{n}$*-consistent. For example, if* $\lambda_n = o(1)$ *and* $\sqrt{n}\lambda_n \to \infty$, *then from Lemma 1.3.3, we have*

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

which implies the convergence rate of $(\widehat{\theta}_n, \widehat{\beta}_n)$ *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* $b_n = \max_{j \in \mathcal{S}_\beta} |\widehat{P}'_{\lambda_n}(\beta_{o,j})| = 0$ *when* n *is sufficiently large. So in this case,* $\lambda_n = o(1)$ *is a sufficient condition for the* $\sqrt{n}$*-consistency of* $(\widehat{\theta}_n, \widehat{\beta}_n)$.

1.3.2 Consistent Moment Selection and Asymptotic Normality

In this subsection, we derive the consistent moment selection and the centered joint limiting distribution of $(\widehat{\theta}_n, \widehat{\beta}_{n,+})$, where $\widehat{\beta}_{n,+}$ denotes the GMM shrinkage estimator of the nonzero components $\beta_{o,+}$ in β_o. We first present and discuss the assumptions needed to show consistent moment selection.

Assumption 1.3.3 *(i) The tuning parameter λ_n satisfies*

$$\sqrt{n} \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| = o_p(1); \tag{1.27}$$

(ii) the penalty function satisfies

$$\liminf_{n \to \infty} \left[\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{r_n \lambda_n} \right] > 0 \ a.e. \tag{1.28}$$

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

Assumption 1.3.3.(i) indicates that the convergence rate of $|\widehat{P}'_{\lambda_n}(\beta_{o,j})|$ for all $j \in \mathcal{S}_\beta$ is faster than $\sqrt{n}$. Under this assumption, Lemma 1.3.3 implies that

$$\sqrt{n} \left(\widehat{\alpha}_n - \alpha_o \right) = O_p(1) \tag{1.29}$$

i.e., the convergence rate of $\widehat{\alpha}_n$ is $\sqrt{n}$. Assumption 1.3.3.(ii) is a generalized version of the condition (3.5) in Fan and Li (2001). Intuitively, Assumption 1.3.3.(ii) implies that the shrinkage estimator $\widehat{\beta}_{n,j}$ of $\beta_{o,j}$ $(j \in \mathcal{S}_\beta^c)$ is the minimizer of $\widehat{P}_{\lambda_n}(\cdot)$ w.p.a.1. From Assumptions 1.3.1.(iv) and 1.3.2.(iv), we know that $\widehat{P}_{\lambda_n}(\cdot)$ is locally minimized at 0. Hence Assumption 1.3.3.(ii) is the key condition needed for showing consistent

moment selection. We show that the bridge, adaptive Lasso and SCAD penalty functions satisfy Assumption 1.3.3 in Appendix 1.10.

Theorem 1.3.5 *Under the Assumption 1.3.1, Assumption 1.3.2 and Assumption 1.3.3, we have*

$$\Pr\left(\widehat{\beta}_{n,j} = \beta_{o,j}\right) \to 1 \tag{1.30}$$

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

From the consistency of $\widehat{\beta}_n$ and Theorem 1.3.5, we can immediately deduce that

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1, \tag{1.31}$$

i.e. the consistent moment selection.

Assumption 1.3.4 *Let* $g_{q+k}(Z,\theta)$ *denote the stacked moment functions from set-1 and set-2, i.e.* $g'_{q+k}(Z,\theta) \equiv \left[g'_q(Z,\theta), g'_k(Z,\theta)\right]$. *The following central limit theorem (CLT) holds*

$$n^{-\frac{1}{2}} \sum_{i\leq n} \left\{g_{q+k}(Z_i,\theta_o) - E\left[g_{q+k}(Z_i,\theta_o)\right]\right\} \to_d \Psi(\theta_o) \tag{1.32}$$

where $\Psi(\theta_o)$ *is a* $q+k$ *dimensional Gaussian random vector.*

Assumption 1.3.4 is a high-level condition, which can be verified by applying CLTs in specific models with specific moment functions and data structure. Let $g_{d_{\beta_-}}(Z,\theta)$ and $g_{d_{\beta+}}(Z,\theta)$ 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 $g_{q+d_{\beta_-}}(Z,\theta) = \left[g_q'(Z,\theta), g_{d_{\beta_-}}'(Z,\theta)\right]'$, $d_{\mathcal{S}_\beta}$ 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 Assumption 1.3.1.(iii) and Assumption 1.3.2.(iii), M_{11} is non-singular.

Theorem 1.3.6 ($\sqrt{n}$-Normality) *Under the Assumption 1.3.1, Assumption 1.3.2, Assumption 1.3.3 and Assumption 1.3.4, we have*

$$\sqrt{n}\left[(\widehat{\theta}_n, \widehat{\beta}_{n,+}) - (\theta_o, \beta_{o,+})\right] \to_d N(0, M_{11}^{-1}\Sigma_{11}M_{11}^{-1}) \tag{1.33}$$

where

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

and $\Psi(\theta_o)$ is defined in (1.32).

1.3.3 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 $\widehat{\beta}_n$ established in Lemma 1.3.1 and the sparsity of $\widehat{\beta}_n$ established in Lemma 1.3.5, the oracle properties follow if we can show that $\widehat{\theta}_n$ 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

$$m_e(\theta_o) = E\begin{bmatrix} g_q(Z,\theta_o) \\ g_{d_{\beta^-}}(Z,\theta_o) \end{bmatrix}_{(q+d_{\mathcal{S}_\beta})\times 1} = 0. \tag{1.34}$$

From the moment conditions in (1.34), we can compute the semiparametric efficiency bound of θ_o as

$$(\Sigma^*)^{-1} = \left[\frac{\partial m_e(\theta)}{\partial \theta_o'}\right]' V_{e,o}^{-1} \left[\frac{\partial m_e(\theta)}{\partial \theta_o}\right]', \tag{1.35}$$

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

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

$$W_n^* \to_p W_o = \left\{E\left[\Psi(\theta_o)\Psi'(\theta_o)\right]\right\}^{-1}, \tag{1.36}$$

then an interesting question is whether the resulting GMM shrinkage estimator $\widehat{\theta}_n$ of θ_o asymptotically attains the semi-parametric efficiency bound in (1.35). The answer to the above question is affirmative, as illustrated in the following theorem.

Theorem 1.3.7 (Oracle Properties) *Under the Assumption 1.3.1, Assumption 1.3.2 and Assumption 1.3.3, we have*

$$\lim_{n\to\infty} \Pr\left(\mathcal{S}_\beta = \mathcal{S}_{\beta,n}\right) = 1. \tag{1.37}$$

Furthermore, if the weight matrix W_n satisfies (1.36) and the Assumption 1.3.4 holds, then we have

$$\sqrt{n}\left(\widehat{\theta}_n - \theta_o\right) \to_d N(0, \Sigma^*), \tag{1.38}$$

where Σ^* is defined in (1.35).

Remark 1.3.8 *In finite samples, the naive confidence intervals (CIs) for θ_o are constructed using the asymptotic distribution in (1.38) and "plug-in" estimator of the variance covariance matrix Σ^* based on the shrinkage estimator $\widehat{\theta}_n$ and the selected moment conditions. It should be noted that the results in (1.37) and (1.38)*

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 Pötscher (2005). The treatment of the PMS inference is beyond the scope of this chapter and is recently studied in Liao (2012).

1.3.4 Selection of the Tuning Parameter

From the results of previous subsections, we see that the tuning parameter λ_n plays an important role in deriving the asymptotic properties of GMM shrinkage estimator. Assumptions 1.3.1.(iv), 1.3.2.(iv) and 1.3.3.(i)-(ii) are sufficient conditions imposed on λ_n for the oracle properties. However, these conditions do not give a straightforward way of selecting the tuning parameter λ_n in finite samples. In this subsection, we provide explicit data-dependent tuning parameters for the GMM shrinkage estimation. The proposed tuning parameters not only satisfy Assumption 1.3.1.(iv), Assumption 1.3.2.(iv) and Assumption 1.3.3.(i)-(ii) but also take the finite sample performance of the GMM shrinkage estimates into account. We use the adaptive Lasso penalty as the illustrating example, as this penalty is used in the simulation study and empirical application of this chapter. The same idea applies to the other penalty functions.

From the Karush-Kuhn-Tucker (KKT) optimality condition, we know that the

nonzero $\beta_{o,j}$ will be estimated as zero only if the following inequality hold

$$\left| W_n(j) \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n) \right] \right| < \frac{\sqrt{n}\lambda_n}{2|\widehat{\beta}_{1st,j}|^{\omega}} \tag{1.39}$$

where $W_n(j)$ denotes the j-th row of the weight matrix W_n. However, if nonzero $\beta_{o,j}$ is estimated as zero, then the left handside of the above inequality will be asymptotically close to a nonzero real number because the invalid moment condition is used in estimation. To ensure the shrinkage bias and error of selecting the invalid moment conditions are small in the finite samples, one would like to have $\sqrt{n}\lambda_n/2$ converge to zero as fast as possible.

On the other hand, the zero $\beta_{o,j}$ will be estimated as zero only if the same inequality in (1.39) is satisfied. As $\widehat{\beta}_{ols,j} = O_p(n^{-\frac{1}{2}})$, we can rewrite the inequality in (1.39) as

$$\left| W_n(j) \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n) \right] \right| < \frac{n^{\frac{1+\omega}{2}}\lambda_n}{2|\sqrt{n}\widehat{\beta}_{1st,j}|^{\omega}}. \tag{1.40}$$

As the left handside of the above inequality is asymptotically a linear combination of Gaussian random variables, one would like to have $n^{\frac{1+\omega}{2}}\lambda_n/2$ diverge to infinite as fast as possible to ensure that valid moment conditions are selected with high probability in finite samples. We choose $\lambda_n = 2cn^{-\frac{1}{2}-\frac{\omega}{4}}$ by balancing the requirement that $\sqrt{n}\lambda_n/2$ converges to zero and $n^{\frac{1+\omega}{2}}\lambda_n/2$ diverges to infinite as fast as possible. We next discuss the selection of the loading term c in the above λ_n.

As $\lambda_n = 2cn^{-\frac{1}{2}-\frac{\omega}{4}}$ satisfies Assumption 1.3.1.(iv), Assumption 1.3.2.(iv) and Assumption 1.3.3.(i)-(ii), we can use the arguments in the proof of Theorem 1.3.6 and

Theorem 1.3.7 to deduce that

$$\begin{aligned} n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) &= v_n \left[\rho(Z, \widehat{\alpha}_n) - \rho(Z, \alpha_o)\right] + v_n \left[\rho(Z, \alpha_o)\right] + n^{\frac{1}{2}} E\left[\rho(Z, \widehat{\alpha}_n)\right] \\ &= v_n \left[\rho(Z, \alpha_o)\right] + \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'} \left[n^{\frac{1}{2}} (\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}})\right] + o_p(1) \\ &= \left\{ I_{q+k} - \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right] M_{11}^{-1} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o \right\} W_o^{-\frac{1}{2}} B_{q+k}(1) + o_p(1), \end{aligned} \tag{1.41}$$

where $B_{q+k}(1)$ denotes a $q+k$ dimensional standard Brownian motion. The consistency of W_n and (1.41) imply that

$$W_n(j) \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) \right] = W_o^{\frac{1}{2}}(j) \Pi_{q+k} B_{q+k}(1) + o_p(1), \tag{1.42}$$

where

$$\Pi_{q+k} = \left\{ I_{q+k} - W_o^{\frac{1}{2}} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right] M_{11}^{-1} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}}\right] W_o^{\frac{1}{2}} \right\} \tag{1.43}$$

is an idempotent matrix with rank $q + d_{\beta\_} - d_\theta$. We propose to choose the loading term $\widehat{c}_{j,n} = ||W_n^{\frac{1}{2}}(j) \widehat{\Pi}_{q+k}||_E$ to normalize the linear combination of the projected Brownian motion in (1.42), where $\widehat{\Pi}_{q+k}$ can be estimated by a first step GMM shrinkage estimation with $\lambda_n = 2\sqrt{\log(q)} n^{-\frac{1}{2} - \frac{\omega}{4}}$ for example. To sum up the above discussion, we propose to select the tuning parameter

$$\widehat{\lambda}_{n,j} = 2||W_n^{\frac{1}{2}}(j) \widehat{\Pi}_{q+k}||_E n^{-\frac{1}{2} - \frac{\omega}{4}} \tag{1.44}$$

for the j-th moment selection coefficient.

1.4 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

$$E\left[g_{n,q}(Z,\theta)\right] = n^{-\tau} G_{n,q}(\theta), \tag{1.45}$$

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

$$E\left[g_{n,k}(Z,\theta_o)\right] = G_k(\theta_o) \stackrel{?}{=} 0, \tag{1.46}$$

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 the IVs are similar to these in (1.45). 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 (1.45) has the same rate.

In this section, we show that if there are strong and correctly specified moment conditions in (1.46), then these moment conditions can be consistently selected by the shrinkage method. More importantly, we show that the GMM shrinkage estimator $\widehat{\theta}_n$ has faster rate of convergence and hence better stochastic properties than the estimators only using the moment conditions in (1.45). The intuition is that when the valid and strong moment conditions in (1.46) are consistently selected, the information contained in these moment conditions is included into estimating θ_o w.p.a.1.

Assumption 1.4.1 *(i) $G_k(\theta)$ is continuous in θ and there exists some continuous function $G_q(\theta)$ such that*

$$\sup_{\theta\in\Theta}\left\|G_{n,q}(\theta)-G_q(\theta)\right\|_E=o(1) \tag{1.47}$$

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

$$\inf_{\{\theta\in\Theta:\|\theta-\theta_o\|_E\geq\varepsilon\}}\left\|G_q(\theta)\right\|_E>\delta_\varepsilon; \tag{1.48}$$

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

$$\sup_{\theta\in\Theta}\left|n^{-1}\sum_{i=1}^{n}\left\{g_{n,l}(Z_i,\theta)-E\left[g_{n,l}(Z_i,\theta)\right]\right\}\right|=O_p(n^{-\frac{1}{2}}); \tag{1.49}$$

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

$$\sup_{\theta\in\Theta}\left\|\frac{\partial G_{n,q}(\theta)}{\partial\theta}-\frac{\partial G_q(\theta)}{\partial\theta}\right\|_E=o(1), \tag{1.50}$$

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 1.4.1.(i) imposes continuity condition on $G_k(\theta)$ and assumes the existence of uniform limit function $G_q(\theta)$ for $G_{n,q}(\theta)$. The uniform approximation in (1.47) is a regularity condition in the weak moment condition literature (e.g., Stock and Wright (2000)). Assumption 1.4.1.(ii) is a identifiable uniqueness condition of θ_o. Assumption 1.4.1.(iii) and (iv) are the generalized versions of Assumption 1.3.1.(ii) and Assumption 1.3.2.(ii)-(iii) respectively. Compared with Assumption 1.3.1.(iv), Assumption 1.4.1.(v) imposes a stronger restriction on the tuning parameter λ_n. When the moment conditions in (1.45) 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 $n^{-2\tau}$, hence $\widehat{P}_{\lambda_n}(\beta_{o,j})$ must converge to zero faster than $n^{-2\tau}$ to ensure the consistency of the GMM shrinkage estimator $\widehat{\alpha}_n$.

Lemma 1.4.1 (Rate of Convergence) *Under Assumption 1.3.1.(iii), Assumption 1.3.2.(iv) and Assumption 1.4.1, we have*

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

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

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

which gives the optimal convergence rate of the estimators based on the moment conditions in (1.45). 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 valid moment conditions in (1.46), but also the GMM shrinkage estimator $\widehat{\theta}_n$ has the rate of convergence faster than $n^{\tau - \frac{1}{2}}$.

Assumption 1.4.2 *(i) The penalty function $\widehat{P}_{\lambda_n}(\cdot)$ satisfies*

$$n^{\frac{1}{2}+\tau} \max_{j \in \mathcal{S}_\beta} \left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = o_p(1); \tag{1.51}$$

(ii) there exists some sequence r_n such that

$$\liminf_{n \to \infty} \left[\frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{r_n \lambda_n}\right] > 0 \ a.e. \tag{1.52}$$

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

Intuitively, Assumption 1.4.2.(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. However, Assumption 1.4.2.(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. We verify Assumption 1.4.1.(v) and Assumption 1.4.2 using the bridge, adaptive Lasso and SCAD penalty functions in Appendix 1.10.

Lemma 1.4.2 *Under the conditions of Lemma 1.4.1 and Assumption 1.4.2, there is*

$$\Pr\left(\widehat{\beta}_{n,j}=0\right)\to 1, \tag{1.53}$$

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

Lemma 1.4.2 implies that $\beta_{o,j}$ $(j\in\mathcal{S}_\beta^c)$ 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 $\widehat{\theta}_n$. Denote the potentially valid moment functions and misspecified moment functions in (1.46) to be $g_{d_{\beta_-}}(Z,\theta)$ and $g_{d_{\beta_+}}(Z,\theta)$ respectively.

Assumption 1.4.3 *Denote* $G_k(\theta)=\left[G'_{d_{\beta_-}}(\theta)\ G'_{d_{\beta_+}}(\theta)\right]'$, *then* $\partial G_{d_{\beta_+}}(\theta_o)/\partial\theta'$ *has full column rank.*

Assumption 1.4.3 is important for deriving the $\sqrt{n}$ convergence rate of $\widehat{\theta}_n$. 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 $\widehat{\theta}_n$ is not $\sqrt{n}$, but is still faster than the rate $n^{\frac{1}{2}-\tau}$.

Lemma 1.4.3 *Under the conditions of Lemma 1.4.2 and Assumption 1.4.3, we have*

$$(\widehat{\theta}_n,\widehat{\beta}_{n,+})-(\theta_o,\beta_{o,+})=O_p(n^{-\frac{1}{2}}). \tag{1.54}$$

Let $W_{o,kk}$ denote 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 $\partial G_{d_{\beta_-}}(\theta_o)/\partial \theta'$ has full column rank, so combined with the Assumption 1.3.1.(iii), we can deduce that the matrix M_+ is invertible.

Corollary 1.4.4 ($\sqrt{n}$-Normality) *Under the conditions of Lemma 1.4.3 and Assumption 1.3.4, we have*

$$\sqrt{n}\left[(\widehat{\theta}_n, \widehat{\beta}_{n,+}) - (\theta_o, \beta_{o,+})\right] \to_d N(0, M_+^{-1}\Sigma_+ M_+^{-1}), \tag{1.55}$$

where

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

and $\Psi(\theta_o)$ is defined in (1.32).

Remark 1.4.5 *If the weight matrix W_n satisfies (1.36), 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 1.4.4, we can deduce that

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

We can similarly prove the efficiency result for the GMM shrinkage estimator $\widehat{\theta}_n$, where the semiparametric efficiency bound of θ_o is implied by the potentially correct and strong moment conditions in (1.46).

1.5 Grouped Variable and Moment Selection via Adaptive Group Lasso

In this section, we study the grouped variable selection and 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 chapter, we only consider the adaptive group Lasso penalty function in this section. The adaptive group Lasso penalty is defined as

$$\widehat{P}_{\lambda_n}(\alpha) = \lambda_n \widehat{w}_\alpha \left\|\alpha\right\|_2, \tag{1.56}$$

where $\widehat{w}_\alpha = \|\widehat{\alpha}_n\|_2^{-\omega}$ $(\omega > 0)$, $\widehat{\alpha}_n$ is a first-step consistent estimator of α and $\|\cdot\|_2$ denotes the l_2-norm. The adaptive group Lasso is originally proposed in Wang (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. 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 parameter 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 there are 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 coefficient 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 $\widehat{\alpha}_n$ with grouped variable and moment condition selection is defined as

$$\widehat{\alpha}_n = \arg\min_{\alpha \in \mathcal{A}} \left[\frac{\sum_{i=1}^n \rho(Z_i, \alpha)}{\sqrt{n}}\right]' W_n \left[\frac{\sum_{i=1}^n \rho(Z_i, \alpha)}{\sqrt{n}}\right] + \sum_{j=1}^{J_\theta + J_\beta} \widehat{P}_{\lambda_n}(\alpha_j), \tag{1.57}$$

where $\mathcal{A}$ is parameter space where α_o lies and α_j denotes the j-th group of parameters in α. Let $\mathcal{S}_{n,\alpha} = \{j : \|\widehat{\alpha}_{n,j}\|_2 \neq 0,\ j = 1, ..., J_\theta + J_\beta\}$ denote the index set of groups of nonzero components in $\widehat{\alpha}_n$. 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.

$$\lim_{n \to \infty} \Pr\left(\mathcal{S}_\alpha = \mathcal{S}_{n,\alpha}\right) = 1, \tag{1.58}$$

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

$$\sqrt{n}\left(\widehat{\theta}_{n,+} - \theta_{o,+}\right) \to_d N(0, \Sigma^+), \tag{1.59}$$

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 $\widehat{\alpha}_n$.

Lemma 1.5.1 *If $\lambda_n = o(1)$ and Assumption 1.3.1.(i)-(iii) are satisfied, then the GMM shrinkage estimator defined in (1.57) is consistent. If we further assume that Assumption 1.3.2.(i)-(iii) are satisfied, then*

$$\widehat{\alpha}_n = \alpha_o + O_p(\delta_n), \tag{1.60}$$

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

From the convergence rate in (1.60), we can deduce that if $\sqrt{n}\lambda_n = O(1)$, then

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

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

Assumption 1.5.1 *(i) For $l = c, k$, the moment function $g_l(z, \theta)$ is continuously differentiable in θ for almost all z; (ii) the following ULLNs hold*

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial g_l(Z, \theta)}{\partial \theta} - E\left[\frac{\partial g_l(Z, \theta)}{\partial \theta}\right] \right\} \right| = o_p(1). \tag{1.61}$$

Assumption 1.5.1.(i) imposes differentiability condition on the moment function $g_l(z, \theta)$ $(l = c, k)$, which is needed to invoke the Karush-Kuhn-Tucker (KKT) optimality condition to derive the sparsity of $\widehat{\alpha}_n$. Note that by definition, $\rho(Z, \alpha) = \rho(Z, \theta, \beta)$

and $\rho(Z, \theta, \beta)$ is trivially differentiable in β. Hence from Assumption 1.5.1.(i), we can deduce that the stacked moment function $\rho(z, \alpha)$ is differentiable in α for almost all z. The ULLNs in (1.61) are useful to derive the probability limit of the score process of the GMM criterion function evaluated at the GMM shrinkage estimator.

Theorem 1.5.2 *Suppose that Assumption 1.3.1.(i)-(iii), Assumption 1.3.2.(i)-(iii) and Assumption 1.5.1 are satisfied and the first-step estimator $\widehat{\alpha}_n$ is root-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*

$$\Pr\left(\|\widehat{\alpha}_{n,j}\|_2 = 0\right) \to 1 \tag{1.62}$$

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

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

$$\Pr\left(\|\widehat{\alpha}_{n,j}\|_2 \neq 0\right) \to 1 \tag{1.63}$$

for all $j \in \mathcal{S}_\alpha$. Hence, the results in (1.62) and (1.63) imply that

$$\Pr\left(\mathcal{S}_{\alpha,n} = \mathcal{S}_\alpha\right) \to 1 \tag{1.64}$$

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

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

where $g_{c+d_{\beta_-}}(Z,\theta_o) = \left[g_c'(Z,\theta_o), g_{d_{\beta_-}}'(Z,\theta_o)\right]'$, $g_{d_{\beta_-}}(Z,\theta_o)$ 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 $(\widehat{\theta}_{n,+}, \widehat{\beta}_{n,+})$.

Corollary 1.5.3 *Under the conditions of Theorem 1.5.2 and Assumption 1.3.4, we have*

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

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

$$m_e(\theta_{o,+}) = E\left[\rho_e(Z,\theta_{o,+})\right] = E\begin{bmatrix} g_c(Z,\theta_{o,+}) \\ g_{d_{\beta^-}}(Z,\theta_{o,+}) \end{bmatrix}_{(q+d_{\mathcal{S}_\beta^c})\times 1} = 0. \tag{1.65}$$

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

$$\left(\Sigma_+^*\right)^{-1} = \left[\frac{\partial m_e(\theta_{o,+})}{\partial \theta_+}\right]\left\{E\left[\rho_e(Z,\theta_{o,+})\rho_e(Z,\theta_{o,+})'\right]\right\}^{-1}\left[\frac{\partial m_e(\theta_{o,+})}{\partial \theta_+'}\right]. \tag{1.66}$$

Next, we show that if the weight matrix W_n in the GMM shrinkage estimation satisfies (1.36), then the GMM shrinkage estimator $\widehat{\theta}_{n,+}$ can asymptotically attain the semiparametric efficiency bound defined in (1.66).

Corollary 1.5.4 *Suppose that the assumptions in Theorem 1.5.3 are satisfied. If the weight matrix W_n satisfies (1.36), then we have*

$$\sqrt{n}\left(\widehat{\theta}_{n,+} - \theta_{o,+}\right) \to_d N\left(0, \Sigma_+^*\right), \tag{1.67}$$

where Σ_+^* is defined in (1.66).

The proof of this corollary is similar to the proof of Theorem 1.38 and thus is omitted. The limiting distribution established in (1.67) is also a pointwise asymptotic property. 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 $\widehat{\theta}_{n,+}$ goes to zero when sample size n goes to infinity. Hence in finite samples, the naive CIs constructed using (1.67) 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 $\widehat{\theta}_{n,j}$ 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 chapter and is recently studied in Liao (2012).

1.6 Simulation Study

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

$$Y_i = X_i\theta_{1,o} + W_{1,i}\theta_{2,o} + W_{2,i}\theta_{3,o} + u_i, \tag{1.68}$$

$$X_i = Z_{1,i}\pi_{1,o} + W_i'\pi_{2,o} + Z_{2,i}'\pi_{3,o} + v_i, \tag{1.69}$$

where Y_i is a scaler dependent variable, X_i is a scaler endogenous variable, $Z_{1,i}$ is the IV whose validity is assumed to be known, $W_i = (W_{1,i}, W_{2,i})$ contains two exogenous variables, $Z'_{2,i} = (Z_{21,i}, ..., Z_{25,i})$ is a set of potentially valid IVs, u_i and v_i are error terms which are correlated with each other.

Suppose a econometrician specifies the model

$$Y_i = X_i\theta_o + W_{1,i}\theta_{2,o} + W_{2,i}\theta_{3,o} + u_i$$

with the following set-1 moment conditions

$$E\left[u_i(Z_{1,i}, W_{1,i}, W_{2,i})\right] = 0 \tag{1.70}$$

to identify and consistently estimate $\theta_o = (\theta_{1,o}, \theta_{2,o}, \theta_{3,o})$. The potentially valid IVs in $Z_{2,i}$ are mixed with 20 invalid IVs $F'_i = (F_{1,i}, ..., F_{20,i})$ to construct the set-2 moment conditions

$$E\left[u_i(Z'_{2,i}, F'_i)\right] \stackrel{?}{=} 0. \tag{1.71}$$

To generate the simulated data, we first generate $(W'_i, Z_{1,i}, Z'_{2,i}, u_i, v_i, F^*_i)$ from a multivariate normal distribution with mean 0 and variance-covariance matrix Σ, where $F^*_i = (F^*_{1,i}, ..., F^*_{20,i})$, $\Sigma = diag(\Sigma_Z, \Sigma_{u,v}, \mathbf{I}_{20})$, Σ_Z is a 8×8 matrix with the i,j-th element being $0.2^{|i-j|}$, $\Sigma_{u,v}$ is a 2×2 matrix with diagonal elements 1 and off-diagonal elements 0.6, $\mathbf{I}_{20}$ is a 20×20 identity matrix. Let c_l to be some value between 0 and 0.8 and l be a 1×20 vector with the j-th ($j = 1, ..., 20$) element being $c_l + (0.8 - c_l) * (j-1)/19$. The invalid IV is generated in the following way

$$F'_i = F^*_i + u_i \times l.$$

It is clear that when c_l is close to zero, the IVs in F_i with larger index numbers (e.g. $F_{20,i}$ and $F_{19,i}$) behave more like valid IVs and it becomes more difficult to distinguish these IVs from the potentially valid ones. We choose different values for c_l ($c_l = 0.2$ or 0.5) to see how our method works in different scenarios. The parameters in the model (1.68)-(1.69) take the following values

$$\theta_o = (0.5, 0.4, 0.3),\ \pi_{2,o} = (0.1, 0.1) \text{ and } \pi_{3,o} = (0.15, 0.15, 0.2, 0.2, 0.25).$$

When $\pi_{1,o}$ is close to zero, $Z_{1,i}$ may contain weak information about the unknown parameter $\theta_{1,o}$, which may also effect the performance of our method in moment selections. In the simulation studies, we choose different values for $\pi_{1,o}$ ($\pi_{1,o} = 0.2$ or 0.8) to see how our method is effected by the signal strength of $Z_{1,i}$.

For each specification of $(c_l, \pi_{1,o})$, we use the simulated samples with sample sizes $n = 250$, 500 and 1000 respectively and for each sample size, 10,000 simulated samples are drawn from the data generating mechanism. The adaptive Lasso penalty with $\omega = 2$ is used to construct the criterion of GMM shrinkage estimation. We use the projected scaled sub-gradient method (active-set variant) method proposed in Schmidt (2010) to solve the minimization problem in the GMM shrinkage estimation. It is remarkable that in this simulation study, there are 25 moment conditions in set-2 and hence 2^{25} subsets of moment conditions to be investigated by the traditional moment selection procedures.

With each simulated sample, we calculate four different GMM estimators, which include the oracle estimator, GMM estimator with set-1 moment conditions, GMM shrinkage estimator and GMM estimate based on set-1 and the moment conditions selected by GMM shrinkage estimation. The oracle estimator is a GMM estimator based on the moment conditions in set-1 and all valid moment conditions in set-2. We

call the second and the last estimates as feasible GMM estimate and post-shrinkage GMM (PSGMM) estimate respectively. Given the specification of $(c_l, \pi_{1,o})$ and the sample size n, we get 10,000 estimators of θ_o for each of the four estimators using the 10,000 simulated samples. We use these 10,000 values of each estimator to calculate its finite sample bias, standard error and root of mean square error.

Table 1.1 Performance of GMM Shrinkage Method in Moment Selection[1]

	$\pi_{1,o} = 0.2$		
	$n = 250$	$n = 500$	$n = 1000$
$c_l = 0.2$	(.5429 .3699 .0872)	(.0857 .8182 .0961)	(.0015 .9503 .0482)
$c_l = 0.5$	(.0022 .8145 .1833)	(.0000 .9012 .0988)	(.0000 .9541 .0459)
	$\pi_{1,o} = 0.8$		
	$n = 250$	$n = 500$	$n = 1000$
$c_l = 0.2$	(.4973 .4709 .0318)	(.0743 .9012 .0245)	(.0010 .9862 .0128)
$c_l = 0.5$	(.0000 .9312 .0688)	(.0000 .9726 .0274)	(.0000 .9872 .0128)

Table 1.1: 1. The three numbers in each bracket are the estimated finite sample probabilities of selecting subsets of moment conditions in set-2 from three different categories respectively. The first category includes the subsets of moment conditions which contain at least one invalid moment condition. The second category contains and only contains the subset of all potentially valid moment conditions in set-2. The third category includes the subsets which do not have the invalid moment conditions, but fail to contain all valid moment conditions in set-2. The finite sample probabilities are computed based on 10,000 replications.

For the ease of describing the simulation results, we divide all possible subsets of moment conditions in set-2 into three categories. The first category includes subsets of moment conditions which contain at least one invalid moment condition. The second category has only one subset which contains and only contains all potentiality valid moment conditions in set-2. The third category includes the subsets of moment conditions which do not have any invalid moment conditions but fail to include all valid ones in set-2. The GMM estimates based on the subsets of moment conditions from the first category are inconsistent. On the other hand, the GMM estimates based on the subsets of moment conditions from the second or third categories (and

Table 1.2 Finite Sample Bias (BS), Standard Deviations (SD) and RMSEs (RE)[1]
$(c_l, \pi_{1,o}) = (0.2, 0.2)$

	Oracle GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0316	.1125	.1169	.0158	.0804	.0819	.0066	.0569	.0573
θ_{2o}	-.0035	.0671	.0672	-.0016	.0460	.0460	-.0008	.0329	.0329
θ_{3o}	-.0029	.0673	.0674	-.0024	.0478	.0478	-.0016	.0333	.0334
	GMM Shrinkage Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.1184	.2467	.2737	.0293	.0931	.0976	.0067	.0588	0592
θ_{2o}	-.0124	.0714	.0725	-.0029	.0462	.0463	-.0008	.0329	.0329
θ_{3o}	-.0160	.0754	.0771	-.0043	.0482	.0484	-.0016	.0334	.0334
	Post-Shrinkage GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0539	.2640	.2694	.0155	.0918	.0931	.0057	.0594	.0597
θ_{2o}	-.0061	.0727	.0730	-.0015	.0463	.0463	.0007	.0329	.0329
θ_{3o}	-.0061	.0781	.0783	.0023	.0483	.0483	-.0015	.0335	.0335
	Consistent GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	-.0633	.4173	.4221	-.0228	.2178	.2190	-.0133	.1462	.1468
θ_{2o}	.0053	.0841	.0843	.0022	.0521	.0521	.0012	.0358	.0358
θ_{3o}	.0110	.0948	.0954	.0034	.0582	.0583	.0012	.0390	.0390

Table 1.2: 1. The finite sample bias, standard error and mean square error are computed using the corresponding estimates from 10,000 replications.

the moment conditions in set-1) are consistent, while the subset of moment conditions in the second category (and the moment conditions in set-1) gives the most efficient estimate.

Table 1.1 presents the finite sample (estimated) probabilities of our method selecting subsets from the first category (the first number in each bracket), the second category (the second number in each bracket) and the third category (the third number in each bracket) respectively. Within each row, we see that the probability of selecting the subset of moment conditions from the second category increases with

Table 1.3 Finite Sample Bias (BS), Standard Deviations (SD) and RMSEs (RE)[1]
$(c_l, \pi_{1,o}) = (0.5, 0.2)$

	Oracle GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0316	.1125	.1169	.0158	.0804	.0819	.0066	.0569	.0573
θ_{2o}	-.0035	.0671	.0672	-.0016	.0460	.0460	-.0008	.0329	.0329
θ_{3o}	-.0029	.0673	.0674	-.0024	.0478	.0478	-.0016	.0333	.0334
	GMM Shrinkage Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0247	.2290	.2304	.0134	.0859	.0869	.0058	.0588	.0591
θ_{2o}	-.0030	.0708	.0708	-.0014	.0462	.0462	-.0007	.0329	.0329
θ_{3o}	-.0020	.0731	.0731	-.0020	.0480	.0481	-.0015	.0334	.0335
	Post-Shrinkage GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0151	.2514	.2518	.0117	.0888	.0896	.0055	.0594	.0596
θ_{2o}	-.0020	.0719	.0719	-.0012	.0463	.0463	-.0007	.0329	.0329
θ_{3o}	-.0005	.0750	.0750	-.0017	.0482	.0482	-.0014	.0335	.0335
	Consistent GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	-.0633	.4173	.4221	-.0228	.2178	.2190	-0133	.1462	.1468
θ_{2o}	.0053	.0841	.0843	.0022	.0521	.0521	.0012	.0358	.0358
θ_{3o}	.0110	.0948	.0954	.0034	.0582	.0583	.0012	.0390	.0390

Table 1.3: 1. The finite sample bias, standard error and mean square error are computed using the corresponding estimates from 10,000 replications.

Table 1.4 Finite Sample Bias (BS), Standard Deviations (SD) and RMSEs (RE)[1]
$(c_l, \pi_{1,o}) = (0.2, 0.8)$

	Oracle GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0110	.0661	.0670	.0058	.0466	.0470	.0020	.0332	.0333
θ_{2o}	-.0016	.0672	.0672	-.0007	.0457	.0457	-.0004	.0325	.0326
θ_{3o}	-.0012	.0680	.0680	-.0015	.0482	.0482	-.0012	.0334	.0334
	GMM Shrinkage Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0378	.0721	.0814	.0094	.0474	.0483	.0020	.0333	.0333
θ_{2o}	-.0043	.0684	.0685	-.0010	.0458	.0458	-.0004	.0326	.0326
θ_{3o}	-.0087	.0694	.0699	-.0024	.0483	.0484	-.0012	.0334	.0334
	Post-Shrinkage GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0173	.0692	.0713	.0060	.0470	.0474	.0019	.0333	.0334
θ_{2o}	-.0023	.0677	.0678	-.0006	.0457	.0457	-.0003	.0326	.0326
θ_{3o}	-.0030	.0687	.0688	-.0016	.0483	.0483	-.0012	.0334	.0335
	Consistent GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	-.0031	.0780	.0781	.0000	.0552	.0552	-.0012	.0391	.0392
θ_{2o}	-.0001	.0667	.0667	.0000	.0455	.0455	.0000	.0325	.0325
θ_{3o}	.0025	.0683	.0683	.0003	.0488	.0488	-.0003	.0337	.0337

Table 1.4: 1. The finite sample bias, standard error and mean square error are computed using the corresponding estimates from 10,000 replications.

Table 1.5 Finite Sample Bias (BS), Standard Deviations (SD) and RMSEs (RE)[1]
$(c_l, \pi_{1,o}) = (0.5, 0.8)$

	Oracle GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0110	.0661	.0670	.0058	.0466	.0470	.0020	.0332	.0333
θ_{2o}	-.0016	.0672	.0672	-.0007	.0457	.0457	-.0004	.0325	.0326
θ_{3o}	-.0012	.0680	.0680	-.0015	.0482	.0482	-.0012	.0334	.0334
	GMM Shrinkage Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0101	.0663	.0671	.0056	.0466	.0469	.0020	.0332	.0333
θ_{2o}	-.0015	.0671	.0671	-.0006	.0457	.0457	-.0004	.0326	.0326
θ_{3o}	-.0010	.0679	.0679	-.0015	.0482	.0482	-.0012	.0334	.0335
	Post-Shrinkage GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	.0096	.0670	.0677	.0054	.0468	.0472	.0019	.0333	.0334
θ_{2o}	-.0015	.0671	.0672	-.0006	.0457	.0457	-.0003	.0326	.0326
θ_{3o}	-.0008	.0681	.0681	-.0014	.0483	.0483	-.0012	.0334	.0335
	Consistent GMM Estimate								
	n=250			n=500			n=1000		
	BS	SD	RE	BS	SD	RE	BS	SD	RE
θ_{1o}	-.0031	.0780	.0781	.0000	.0552	.0552	-.0012	.0391	.0392
θ_{2o}	-.0001	.0667	.0667	.0000	.0455	.0455	.0000	.0325	.0325
θ_{3o}	.0025	.0683	.0683	.0003	.0488	.0488	-.0003	.0337	.0337

Table 1.5: 1. The finite sample bias, standard error and mean square error are computed using the corresponding estimates from 10,000 replications.

the sample size growing. When the sample size is 1000, this probability is close to 1, which is predicted by the consistent moment selection result established in Section 1.3. Both the signal strength of the IVs in set-1 and the severity of misspecification of IVs in set-2 effect the finite sample performance of our method in moment selection. When the misspecified moment conditions in set-2 are close to be valid (i.e. $c_l = 0.2$) and the sample size is small (i.e. $n = 250$), the probabilities of selecting the subset of moment conditions from the second category become small. This problem is exacerbated by the weak information contained in the IV in set-1. On the other hand, when the invalid moment conditions in set-2 are severely misspecified, i.e. $c_l = 0.5$, the probabilities of selecting invalid moment conditions are small. Finally, when the IVs in set-1 are strong (i.e. $\pi_{1,o} = 0.8$) and the invalid IVs in set-2 are severely misspecified, the GMM shrinkage method performs very well even when the sample size is small (i.e. $n = 250$).

Table 1.2 to Table 1.5 describe the finite sample properties of the above four GMM estimates in terms of finite sample bias, standard deviation and root of mean square error. We summarize the main results in these tables by comparing the GMM shrinkage estimate with the oracle estimate, the feasible GMM estimate and the post-shrinkage GMM estimate respectively. In all scenarios, the finite sample bias, standard error and mean square error of the GMM shrinkage estimate approximate these of the oracle estimate with the sample size growing. The finite sample properties of the two estimates are almost identical when the sample size $n = 1000$. Compared with the GMM estimator, the GMM shrinkage estimator has much smaller standard error, particularly when the IV in set-1 is weak, i.e. $\pi_{1,o} = 0.2$. The finite sample bias of the GMM shrinkage estimate is larger than the GMM estimate in some scenarios (e.g. when $c_l = 0.2$). For the GMM shrinkage estimate, one source of the finite sample bias is the shrinkage bias. To get rid of this shrinkage bias, one

can apply GMM estimation based on the set-1 and moment conditions in set-2 selected by the GMM shrinkage estimation to get the post-shrinkage GMM estimator. Although generally speaking, the post-shrinkage GMM estimate enjoys smaller finite sample bias, its finite sample standard error may be larger than the GMM shrinkage estimate. This phenomenon is recovered in Table 1.2 and 1.3, when the potentially valid moment conditions contain strong information about the endogenous variable. The underlying reason for this phenomenon is that when the moment selection coefficients of invalid moment conditions are shrunk towards zero, part of the information contained in these moment conditions are used in the GMM shrinkage estimation, which helps to reduce the variance but at the same time introduces new bias to the GMM shrinkage estimate.

1.7 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

$$\Delta \log(h_{i,t}) = \alpha_{o,t} + \Delta \log(w_{i,t})\delta_o + \varepsilon_{i,t}, \tag{1.72}$$

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, $\alpha_{o,t}$ is a time varying constant (invariant across the individuals) 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 (1.72) may be inconsistent. MaCurdy (1981) proposes to use a set of 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 of wage $w^*_{i,t}$ to construct a 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[6].

Instead of using all the 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 of wage $w^*_{i,t}$ and the wage $w_{i,t}$ itself into set-2. This specification at least enables us to answer the following four questions. First, are the other IVs in MaCurdy (1981), especially the age and education 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 $w^*_{i,t}$, 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. Fourth, are the lagged wages rates, i.e. $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$ $(s < t)$, potentially valid IVs for $\Delta \log(w_{i,t})$?

[6] In MaCurdy (1981) and Altonji (1986), $w_{i,t}$ is constructed by dividing the annual labor income of individual i 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.

Table 1.6. 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	-.0164	-.0123	-.0233	-.0125	-.0217
	(.0040)	(.0189)	(.0069)	(.0204)	(.0061)	(.0212)
θ_o	-.3937	-.3992	.0703	.2743	.1638	.2032
	(.0276)	(.0278)	(.2730)	(.4396)	(.1967)	(.2234)
d_t?[4]	No	Yes	No	Yes	No	Yes

Table 1.6: 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, age square and the interaction between age and education; 3. the moment condition is constructed using an alternative measure of hour wage; 4. d_t refers to the set of time dummy variables for the years from 1971 to year 1981.

These lagged wage variables would provide some extra information about $\Delta \log(w_{i,t})$ if the shocks to the wage process were dependent. To investigate the validities of $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$ and at the same time avoid the potential pitfall of using weak IVs, we include the previous 3 period differenced wages (i.e., $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$, $s = t-1$, $t-2$ and $t-3$) variables in set-2.

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 [7].

Table 1.6 presents the GMM estimators of δ_o based on the moment conditions

[7]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.

constructed by $w_{i,t}$, IVs used in MaCurdy (1981) and IV used in Altonji (1986) respectively. The results in Table 1.6 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 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 apply the GMM shrinkage estimation to the labor supply equation (1.72). The estimation results are presented in Table 1.7. As a comparison, we also include the GMM estimators based on the moment condition in set-1 and the post-shrinkage GMM estimators in Table 1.7. Columns (1)-(2) of Table 1.7 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 1.7, the GMM estimators in columns (1)-(2) not only are larger in magnitude, but also have larger standard errors. Columns (3) and (5) of Table 1.7 contain the results of GMM shrinkage estimation with the rest IVs (i.e. education, square of education, age and age square, father's education and the interaction between education and age of individual i) from MaCurdy (1981), the IV (i.e. $\Delta \log(w^*_{i,t})$) from Altonji (1986), the first difference of wage $\Delta \log(w_{i,t})$ and the lagged differences of wage rates $\Delta \log(w_{i,s})$ and $\Delta \log(w^*_{i,s})$, $s = t-1$, $t-2$ and $t-3$ in set-2. All moment conditions in set-2, except the $\Delta \log(w_{i,t})$, are selected in the GMM shrinkage estimation. The GMM shrinkage estimates of the moment selection coefficients of $\Delta \log(w_{i,t})$ are -0.0301 (with standard error 0.0079) and -0.0340 (with standard error 0.0098) in columns (3) and (5) respectively. The GMM shrinkage

estimators have much smaller standard errors, because the moment conditions in set-2 are selected and automatically included into estimation by the GMM shrinkage method. The post-shrinkage GMM estimates in columns (4) and (6) are free of shrinkage bias, but are only slightly different from the GMM shrinkage estimates in columns (3) and (5) respectively.

From columns (3)-(6), we see that when $\alpha_{o,t}$ is treated as a time invariant constant term, the standard errors of the GMM estimates are relatively small, though the differences between the estimates in columns (3) and (5) (or columns (4) and (6)) are nontrivial. The smaller standard errors in columns (3) and (4) benefit from the assumption that $\alpha_{o,t}$ is constant across the time, which reduces the number of parameters to be estimated. If such assumption is miss-specified, then extra bias will be introduced to the GMM estimates. We next use the shrinkage method to test if $\alpha_{o,t}$ is constant or variant across the time, i.e. we penalize $|\alpha_t - \alpha_{t-1}|$ $(t = 2, ..., 11)$ together with the moment selection coefficients in the GMM shrinkage estimation. The results of this GMM shrinkage estimation are included in column (7). The constraints found in the GMM shrinkage estimation are $\alpha_1 = ... = \alpha_5$ and $\alpha_6 = ... = \alpha_{11}$ and all IVs, except $\Delta \log(w_{i,t})$, are selected. The GMM shrinkage estimate of δ_o in column (7) is between the GMM shrinkage estimates of δ_o in column (3) and (5) (or column (4) and (6)) in the magnitude and standard errors. The post-shrinkage GMM estimates based on the selected moment conditions and the selected constraints on $\alpha_{o,t}$ are included in column (8).

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

Table 1.7. GMM and GMM Shrinkage Estimation for the Labor Supply Equation[1,3]

	GMM	GMM	SGMM	PSGMM	SGMM	PSGMM	SGMM	PSGMM
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
a_1	-.0159	-.0224	-.0135	-.0139	-.0323	-.0313	-.0200	-.0103
	(.0257)	(.0236)	(.0051)	(.0051)	(.0192)	(.0190)	(.0185)	(.0047)
δ_o	.3241	.2682	.1648	.1659	.2395	.2509	.1615	.2096
	(1.225)	(1.143)	(.1454)	(.1450)	(.1826)	(.1814)	(.1356)	(.1663)
d_t?[2]	No	Yes	No	No	Yes	Yes	Yes	Yes

Table 1.7: 1. Standard errors are in parentheses and sample size n=3487. 2. $d_t$refers to the set of time dummy variables for the years from 1971 to year 1981. 3. In columns (3), (5) and (7), the IVs in set-2 are: father's education, education, education square, age, age square, the interaction between age and education, the first difference of alternative measure of hour wage, the first difference of constructed wage and the lagged (3 period) differences of alternative measure wage and constructed wage. The constraints on constant term are also selected together with the IVs in set-2 in the GMM shrinkage estimation in column (7).

by $w_{i,t}$, which indicates that $\Delta \log(w_{i,t})$ is an endogenous variable in the labor supply equation. Fourth, our method selects the lagged differences of wage rates $\Delta \log(w^*_{i,s})$ and $\Delta \log(w_{i,s})$ as valid IVs and hence provides extra moment conditions to more efficiently estimate δ_o. Fifth, our method does not assume that $\alpha_{o,t}$ is constant across time and detects a structural break in $\alpha_{o,t}$. Hence our estimate of δ_o is more robust compared with the estimates based on the time invariant assumption of $\alpha_{o,t}$, and at the same time more efficient than the estimates based on the assumption that $\alpha_{o,t}$ changes over time. 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-shrinkage GMM estimator, which is as efficient as the GMM shrinkage estimator but has smaller finite sample bias.

1.8 Conclusion

This chapter 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. 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 also show that the shrinkage technique can be used in GMM to perform grouped variable selection and moment selection simultaneously. We provide a simple

and data-driven procedure of selecting the tuning parameters 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. Simulations show that our method performs well in terms of the correct moment selection and the finite sample properties of its estimators. 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), thereby supporting 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. In addition to the existing IVs in MaCurdy (1981) and Altonji (1986), our method also finds that lagged wage rates are valid IVs for $\Delta \log(w_{i,t})$. Moreover, our estimators of the intertemporal substitution elasticity have smaller standard deviations, though their values are close to those in the literature.

1.9 Appendix A

Throughout the appendix, the symbols "$\rightarrow_p$" and "$\rightarrow_d$" stand for "convergence in probability" and "convergence in distribution" respectively. Let $m(\alpha) = E\left[\rho(Z, \alpha)\right]$ and define an empirical process indexed by the function g as $\upsilon_n(g) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} [g(Z_i) - E(g(Z_i)]$. 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.

1.9.1 Two Useful Lemmas

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

$$\begin{aligned} V_1^{(n)}(\theta,\beta) &= \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_n \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] + \sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_j) \\ &= V_0^{(n)}(\theta,\beta) + \sum_{j=1}^{k}\widehat{P}_{\lambda_n}(\beta_j) \end{aligned} \tag{1.73}$$

and

$$V_0(\theta,\beta) = \{E[\rho(Z,\theta,\beta)]\}' W_o \{E[\rho(Z,\theta,\beta)]\}. \tag{1.74}$$

Lemma 1.9.1 *Under Assumption 1.3.1.(iii), we have*

$$V_0^{(n)}(\theta,\beta) \geq c_1 V_0(\theta,\beta) - c_2 R_n \tag{1.75}$$

and

$$V_0^{(n)}(\theta,\beta) \leq c_3 V_0(\theta,\beta) + c_4 R_n, \tag{1.76}$$

for all $(\theta,\beta) \in \Theta \times \mathcal{B}$ *w.p.a.1, where*

$$R_n \equiv \sup_{(\theta,\beta)\in\Theta\times\mathcal{B}} \frac{1}{n}\{v_n[\rho(Z,\theta,\beta)]\}' W_o \{v_n[\rho(Z,\theta,\beta)]\} \tag{1.77}$$

and c_i $(i = 1, ..., 4)$ *denotes some generic positive constant.*

Proof. By Assumption 1.3.1.(iii), we can deduce that

$$V_0^{(n)}(\theta,\beta) \geq c\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \tag{1.78}$$

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

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

which can be rewritten as

$$\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \geq \frac{1}{2}V_0(\theta,\beta) - R_n. \tag{1.79}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$. Now the result in (1.75) can be deduced from the inequalities in (1.78) and (1.79).

For the second result, note that Assumption 1.3.1.(iii) also implies

$$V_0^{(n)}(\theta,\beta) \leq c\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] \tag{1.80}$$

w.p.a.1 and

$$\left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta) - 2E\left[\rho(Z,\theta,\beta)\right]\right]' W_o \left[n^{-1}\sum_{i=1}^{n}\rho(Z_i,\theta,\beta) - 2E\left[\rho(Z,\theta,\beta)\right]\right] \geq 0 \tag{1.81}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$. The inequality in (1.81) can be rewritten as

$$\left[\frac{\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)}{n}\right]' W_o \left[\frac{\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)}{n}\right] \leq 2V_0(\theta,\beta) + 2R_n. \tag{1.82}$$

for all $(\theta, \beta) \in \Theta \times \mathcal{B}$. Now the result in (1.76) can be deduced from the inequalities in (1.80) and (1.81). ■

The next lemma establishes the local quadratic approximation of $V_0(\theta, \beta)$ in terms of $(||\theta - \theta_o||_E^2 + ||\beta - \beta_o||_E^2)^{\frac{1}{2}}$ 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 1.9.2 *Under Assumption 1.3.1.(iii) and Assumption 1.3.2.(ii)-(iii), we have*

$$||\theta-\theta_o||_E^2+||\beta-\beta_o||_E^2 \asymp E\left[\rho(Z,\theta,\beta)\right]' W_o E\left[\rho(Z,\theta,\beta)\right] \tag{1.83}$$

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 1.3.2.(ii)

$$E\left[\rho(Z,\theta,\beta)\right]=\left(\begin{array}{c} E\left[g_q(Z,\theta)\right] \\ E\left[g_k(Z,\theta)\right]-\beta \end{array}\right)=\left(\begin{array}{cc} \frac{\partial E\left[g_q(Z,\widetilde{\theta})\right]}{\partial\theta'} & 0 \\ \frac{\partial E\left[g_k(Z,\widetilde{\theta})\right]}{\partial\theta'} & -I_k \end{array}\right)\left(\begin{array}{c} \theta-\theta_o \\ \beta-\beta_o \end{array}\right), \tag{1.84}$$

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

$$\frac{\partial E\left[g_l(Z,\widetilde{\theta})\right]}{\partial\theta'}=\frac{\partial E\left[g_l(Z,\theta_o)\right]}{\partial\theta'}+o(1), \tag{1.85}$$

for $l = q, k$. Denote $\frac{\partial m(\theta)}{\partial \alpha'} = \begin{pmatrix} \frac{\partial E[g_q(Z,\theta)]}{\partial \theta'} & 0 \\ \frac{\partial E[g_k(Z,\theta]}{\partial \theta'} & -I_k \end{pmatrix}$, then by (1.84), (1.85) and the Cauchy-Schwarz inequality, we have

$$E\left[\rho(Z,\theta,\beta)\right] = \frac{\partial m(\alpha_o)}{\partial \alpha'}(\alpha - \alpha_o) + o(||(\alpha - \alpha_o||_E). \tag{1.86}$$

Using Assumption 1.3.1.(iii), the result in (1.86) and the Cauchy-Schwarz inequality, we get

$$E\left[\rho(Z,\alpha)\right]' W_o E\left[\rho(Z,\alpha)\right] = (\alpha - \alpha_o)' \frac{\partial m(\alpha_o)}{\partial \alpha} W_o \frac{\partial m(\alpha_o)}{\partial \alpha'}(\alpha - \alpha_o) + o(||(\alpha - \alpha_o||_E^2). \tag{1.87}$$

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 $\frac{\partial m(\theta_o)}{\partial \alpha} W_o \frac{\partial m(\theta_o)}{\partial \alpha'}$ is strictly positive definite. Let c_1 and c_2 ($c_1, c_2 > 0$) denote the smallest and largest eigenvalues of $\frac{\partial m(\theta_o)}{\partial \alpha} W_o \frac{\partial m(\theta_o)}{\partial \alpha'}$. From (1.87), we can deduce that

$$\begin{aligned} c_2||(\alpha - \alpha_o||_E^2 + o(||(\alpha - \alpha_o||_E^2) &\geq E\left[\rho(Z,\alpha)\right]' W_o E\left[\rho(Z,\alpha)\right] \\ &\geq c_1||(\alpha - \alpha_o||_E^2 + o(||(\alpha - \alpha_o||_E^2). \end{aligned} \tag{1.88}$$

Now, result in (1.83) follows directly from (1.88). ■

1.9.2 Proof of the Main Results in Section 1.3

Proof of Lemma 1.3.1. By the definition of $(\widehat{\theta}_n, \widehat{\beta}_n)$, we have

$$V_0^{(n)}(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_n) \leq V_0^{(n)}(\theta_o, \beta_o) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}). \tag{1.89}$$

Applying Lemma 1.9.1 and Assumption 1.3.1.(iv), we deduce from (1.89) that

$$V_0(\widehat{\theta}_n, \widehat{\beta}_n) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + 2cR_n, \tag{1.90}$$

w.p.a.1, where R_n is defined in Lemma 1.9.1 and c is some generic constant.

From Assumption 1.3.1.(ii) and the definition of $\rho(Z, \theta, \beta)$, we get

$$\sup_{(\theta,\beta)\in\Theta\times\mathcal{B}} \frac{v_n\left[\rho(Z,\theta,\beta)\right]}{\sqrt{n}} = \sup_{\theta\in\Theta} n^{-1} \sum_{i=1}^{n} \left\{g_{q+k}(Z_i,\theta) - E\left[g_{q+k}(Z_i,\theta)\right]\right\} = o_p(1) \tag{1.91}$$

where $g'_{q+k}(Z,\theta) = \left[g'_q(Z,\theta), g'_k(Z,\theta)\right]$. By the triangle inequality, ULLN in (1.91), Assumption 1.3.1.(iii)-(iv), we have

$$R_n = o_p(1) \text{ and } \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) = o_p(1). \tag{1.92}$$

From the Assumption 1.3.1.(iii) and results in (1.90) and (1.92), we can deduce that

$$\left\|E[g_q(Z,\widehat{\theta}_n)]\right\|_E = o(1) \text{ and } \left\|E[g_k(Z,\widehat{\theta}_n)] - \widehat{\beta}_n\right\|_E = o_p(1) \tag{1.93}$$

Now, the first result in (1.93) and Assumption 1.3.1.(i) imply that $\widehat{\theta}_n \to_p \theta_o$. From the second result in (1.93), triangle inequality, consistency of $\widehat{\theta}_n$ and the continuity of $E[g_k(Z,\theta)]$, we can deduce that

$$\begin{aligned} o_p(1) &= \left\|E[g_k(Z,\widehat{\theta}_n)] - \widehat{\beta}_n\right\|_E \\ &\geq \left|\left\|E[g_k(Z,\widehat{\theta}_n)] - E[g_k(Z,\theta_o)]\right\|_E - \left\|\widehat{\beta}_n - \beta_o\right\|_E\right| \\ &= \left|\left\|\widehat{\beta}_n - \beta_o\right\|_E + o_p(1)\right| \end{aligned} \tag{1.94}$$

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

Proof of Lemma 1.3.3. By the definition of $(\widehat{\theta}_n, \widehat{\beta}_n)$, we have

$$V_0^{(n)}(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) \leq V_0^{(n)}(\theta_o, \beta_o) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}). \tag{1.95}$$

Using the inequalities in (1.75), (1.76) and (1.95), we get

$$c_1 V_0(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + c_2 R_n, \tag{1.96}$$

w.p.a.1, where c_1 and c_2 are some generic positive constants and R_n is defined in Lemma 1.9.1.

Next, by Assumption 1.3.2.(iv), Taylor expansion, the triangle inequality and Cauchy Schwarz inequality, we get

$$\begin{aligned}
& \left| \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}_{\lambda_n}(\beta_{o,j}) - \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) \right] \right| \\
= & \left| \sum_{j \in \mathcal{S}_\beta} \left[\widehat{P}'_{\lambda_n}(\beta_{o,j})(\widehat{\beta}_{n,j} - \beta_{o,j}) + \frac{1}{2} \widehat{P}''_{\lambda_n}(\widetilde{\beta}_j)(\widehat{\beta}_{n,j} - \beta_{o,j})^2 \right] \right| \\
\leq & \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| \left\| \widehat{\alpha}_n - \alpha_o \right\|_E + \max_{j \in \mathcal{S}_\beta} \left| \frac{\widehat{P}''_{\lambda_n}(\beta_{o,j})}{2} + o_p(1) \right| \left\| \widehat{\alpha}_n - \alpha_o \right\|_E^2 . \qquad (1.97)
\end{aligned}$$

w.p.a.1, where $\widetilde{\beta}_j$ lies between $\beta_{o,j}$ and $\widehat{\beta}_{n,j}$ for $j \in \mathcal{S}_\beta$. From Assumption 1.3.1.(iv) and Assumption 1.3.2.(iv), inequalities in (1.96) and (1.97), we can apply Lemma 1.9.2 to deduce that

$$\left\| \widehat{\alpha}_n - \alpha_o \right\|_E^2 \leq c_3 b_n \left\| \widehat{\alpha}_n - \alpha_o \right\|_E + c_4 R_n \tag{1.98}$$

w.p.a.1, where c_3, c_4 are some positive constants. The inequality in (1.98) implies

$$\|\widehat{\alpha}_n - \alpha_o\|_E \leq \frac{c_3 b_n + (c_3^2 b_n^2 + 4c_4 R_n)^{\frac{1}{2}}}{2} = O_p\left(b_n + n^{-\frac{1}{2}}\right), \tag{1.99}$$

where c_5 is some positive constant. Now, for any positive M, inequality in (1.99) enables us to deduce that

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

which establishes the desired rate. ■

Proof of Theorem 1.3.5. On the event $\{\widehat{\beta}_{n,j} \neq 0\}$ for some $j \in \mathcal{S}_\beta^c$, we have the following Karush-Kuhn-Tucker (KKT) optimality condition:

$$2\left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \frac{\partial \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n)}{\partial \beta_j}\right]' W_n \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n)\right] + n\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}) = 0, \tag{1.100}$$

which implies

$$\left|W_n^{(j)} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\theta}_n, \widehat{\beta}_n)\right]\right| = \frac{\sqrt{n}\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{2}. \tag{1.101}$$

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

Note that

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

where

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

and $\widetilde{\theta}_{j,n}$ $(j = 1, ..., q+k)$ lies between θ_o and $\widehat{\theta}_n$. From Assumption 1.3.2.(i), we have $v_n \left[\rho(Z, \widehat{\alpha}_n)\right] = O_p(1)$. By Lemma 1.3.3 and Assumption 1.3.2.(ii), we have $n^{\frac{1}{2}}(\widehat{\alpha}_n - \alpha_o) = O_p(1)$. From Assumption 1.3.3.(iii) and the consistency of $\widehat{\alpha}_n$, we can deduce that

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

Hence we have $n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) = O_p(1)$, which combined with Assumption 1.3.1.(iii), implies that

$$\left| W_n^{(j)} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) \right] \right| = O_p(1). \tag{1.102}$$

While from Assumption 1.3.3.(iii), we get

$$\frac{\sqrt{n}\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{2} = \frac{r_n \lambda_n n^{\frac{1}{2}}}{2} \frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{r_n \lambda_n} \to_p \infty. \tag{1.103}$$

Now, using the results in (1.102) and (1.103), we can deduce that $\Pr\left(\widehat{\beta}_{n,j} = 0\right) \to 1$ for $j \in \mathcal{S}_\beta^c$. ■

Proof of Theorem 1.3.6. Define $\alpha_{o,\mathcal{S}} = (\theta_o, \beta_{o,+})$ and accordingly $\widehat{\alpha}_{n,\mathcal{S}} = \left(\widehat{\theta}_n, \widehat{\beta}_{n,+}\right)$. 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

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

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 1.3.5, we know that $\widehat{\beta}_{n,-} = 0$ w.p.a.1. Thus,$\sqrt{n}\,(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}})$ is the minimizer of $V_{2,n}(u_{\mathcal{S}})$ w.p.a.1.

If we denote

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

then the compactness of K, Assumption 1.3.2.(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

$$\upsilon_n\left\{\rho^s(Z,\alpha_{o,\mathcal{S}}+\frac{u_{\mathcal{S}}}{\sqrt{n}}) - \rho(Z,\alpha_o)\right\} = o_p(1), \tag{1.104}$$

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

Next note that by Assumption 1.3.2.(iii) and the compactness of K, we have

$$\sqrt{n}\left\{E\left[\rho^s(Z,\alpha_{o,\mathcal{S}}+\frac{u_{\mathcal{S}}}{\sqrt{n}})\right] - E\left[\rho(Z,\alpha_o)\right]\right\} = \frac{\partial m(\theta_o)}{\partial \alpha'_{\mathcal{S}}}u_{\mathcal{S}} + o(1), \tag{1.105}$$

uniformly over $u_{\mathcal{S}} \in K$. Thus, (1.104) and (1.105) imply that uniformly over $u_{\mathcal{S}} \in K$, there is

$$\begin{aligned} & n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) \\ = \; & \upsilon_n \left\{ \rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) - \rho(Z_i, \alpha_o) \right\} + \upsilon_n \left[\rho(Z_i, \alpha_o) \right] \\ & + \sqrt{n} \left\{ E \left[\rho^s(Z_i, \alpha_{o,\mathcal{S}} + \frac{u_{\mathcal{S}}}{\sqrt{n}}) \right] - E \left[\rho(Z_i, \alpha_o) \right] \right\} \\ = \; & \upsilon_n \left[\rho(Z, \alpha_o) \right] + \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'} u_{\mathcal{S}} + o_p(1). \end{aligned} \tag{1.106}$$

Now, we can use the result in (1.106) to deduce that

$$V_{2,n}^*(u_{\mathcal{S}}) = u_{\mathcal{S}}' \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}} W_o \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}'} u_{\mathcal{S}} + 2u_{\mathcal{S}}' \frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}} W_o \left\{ \upsilon_n \left[\rho(Z, \alpha_o) \right] \right\} + o_p(1), \tag{1.107}$$

uniformly over $u_{\mathcal{S}} \in K$. If $j \in \mathcal{S}_\beta$, then by Assumptions 1.3.2.(iv) and 1.3.3.(i)

$$\begin{aligned} & n \left[\widehat{P}_{\lambda_n}(\beta_{o,j} + \frac{u_{\beta_+,j}}{\sqrt{n}}) - \widehat{P}_{\lambda_n}(\beta_{o,j}) \right] \\ = \; & \sqrt{n} \widehat{P}_{\lambda_n}'(\beta_{o,j}) u_{\beta_+,j} + \left[\widehat{P}_{\lambda_n}''(\beta_{o,j}) + o_p(1) \right] u_{\beta_+,j}^2 \to_p 0 \end{aligned} \tag{1.108}$$

uniformly in $u_{\beta_+,j}$. Using the results in (1.107)-(1.108) and triangle inequality, we get

$$V_{2,n}(u_{\mathcal{S}}) \to_d V_2(u_{\mathcal{S}}) = u_{\mathcal{S}}' M_{11} u_{\mathcal{S}} + 2u_{\mathcal{S}}' \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}} \right] W_o \Psi(\alpha_o) \tag{1.109}$$

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

$$u_{\mathcal{S}}^* = -M_{11}^{-1} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\mathcal{S}}} \right] W_o \Psi(\alpha_o) \tag{1.110}$$

By Lemma 1.3.3 and Assumption 1.3.3.(i), there is

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) = O_p(1) \tag{1.111}$$

Now, the asymptotic tightness of $\widehat{\alpha}_{n,\mathcal{S}}$ in (1.111), the uniform convergence in distribution in (1.109) and unique minimization in (1.110) enable us to invoke the ACMT to deduce that

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

■

Proof of Theorem 1.3.7. The first result is implied by Lemma 1.3.1 and Lemma 1.3.5, so we only need to show the second claim. First note that if $W_n \to_p W_o = \{E\left[\Psi(\theta_o)\Psi(\theta_o)'\right]\}^{-1}$, then the centered limiting distribution in (1.33) will be simplified to

$$\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right) \to_d N(0, M_{11}^{-1}) \tag{1.112}$$

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

$$\begin{aligned}
M_{11} &= \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial \theta} & \frac{\partial E[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial \theta} \\ 0 & -I_{d_{\beta_o^+} \times d_{\beta_o^+}} \end{pmatrix} W_o \begin{pmatrix} \frac{\partial E[g_{q+d_{\beta_-}}(Z,\theta_o)]}{\partial \theta'} & 0 \\ \frac{\partial E[g_{d_{\beta_+}}(Z,\theta_o)]}{\partial \theta'} & -I_{d_{\beta_o^+} \times d_{\beta_o^+}} \end{pmatrix} \\
&= \begin{pmatrix} \frac{\partial m_e(\theta_o)}{\partial \theta} & \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta} \\ 0 & -I_{d_{\beta_o^+} \times d_{\beta_o^+}} \end{pmatrix} \begin{pmatrix} W_{11} & W_{12} \\ W_{21} & W_{22} \end{pmatrix} \begin{pmatrix} \frac{\partial m_e(\theta_o)}{\partial \theta'} & 0 \\ \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} & -I_{d_{\beta_o^+} \times d_{\beta_o^+}} \end{pmatrix} \\
&= \begin{pmatrix} \Sigma_{11}^{11} & -\frac{\partial m_e(\theta_o)}{\partial \theta} W_{12} - \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta} W_{22} \\ -W_{21}\frac{\partial m_e(\theta_o)}{\partial \theta'} - W_{22}\frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} & W_{22} \end{pmatrix},
\end{aligned} \tag{1.113}$$

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 (1.113), it is easy to get

$$\begin{aligned}
\Omega_{\theta_o}^{-1} &= \frac{\partial m_e(\theta_o)}{\partial \theta} W_{11} \frac{\partial m_e(\theta_o)}{\partial \theta'} + 2 \frac{\partial m_e(\theta_o)}{\partial \theta} W_{12} \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} \\
&\quad + \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta} W_{22} \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} - \frac{\partial m_e(\theta_o)}{\partial \theta} W_{12} W_{22}^{-1} W_{21} \frac{\partial m_e(\theta_o)}{\partial \theta'} \\
&\quad - 2 \frac{\partial m_e(\theta_o)}{\partial \theta} W_{12} \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} - \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta} W_{22} \frac{\partial m_{\beta_{o,+}}(\theta_o)}{\partial \theta'} \\
&= \frac{\partial m_e(\theta_o)}{\partial \theta} \left(W_{11} - W_{12} W_{22}^{-1} W_{21} \right) \frac{\partial m_e(\theta_o)}{\partial \theta'} \\
&= \left[\frac{\partial m_e(\theta_o)}{\partial \theta} \right] V_{e,o}^{-1} \left[\frac{\partial m_e(\theta_o)}{\partial \theta} \right]' = (\Sigma^*)^{-1}, \qquad (1.114)
\end{aligned}$$

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 (1.112), (1.114) and the Continuous Mapping Theorem (CMT), we can deduce that

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

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

■

1.9.3 Proof of the Main Results in Section 1.4

Proof of Lemma 1.4.1. Using similar arguments in the proof of Lemma 1.3.1, we get

$$V_0(\widehat{\theta}_n, \widehat{\beta}_n) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + 2cR_n, \qquad (1.115)$$

w.p.a.1, where R_n is defined in Lemma 1.9.1. and c is some generic constant. Under Assumption 1.3.1.(iii), Assumption 1.4.1.(iii) and (v), we have

$$\left\|n^{-\tau}G_{q,n}(\widehat{\theta}_n)\right\|_E^2 + \left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\|_E^2 = O_p(n^{-1} + \max_{j\in\mathcal{S}_\beta}\widehat{P}_{\lambda_n}(\beta_{o,j})) \tag{1.116}$$

which implies that $\left\|G_{q,n}(\widehat{\theta}_n)\right\|_E^2 = O_p(n^{2\tau-1} + n^{2\tau}\max_{j\in\mathcal{S}_\beta}\widehat{P}_{\lambda_n}(\beta_{o,j})) = o_p(1)$. Hence, using the uniform approximation in Assumption 1.4.1.(i), we get $\left\|G_q(\widehat{\theta}_n)\right\|_E = o_p(1)$, which combined with the identification condition in 1.4.1.(i), implies the consistency of $\widehat{\theta}_n$. The consistency of $\widehat{\beta}_n$ can be proved using similar arguments in the proof of Lemma 1.3.1.

Next, we derive the convergence rate of $\widehat{\alpha}_n$. Using similar arguments in the proof of Lemma 1.9.2, we can apply Assumptions 1.3.1.(iii) and 1.4.1.(iv) to deduce

$$V_0(\widehat{\theta}_n, \widehat{\beta}_n) \geq cn^{-2\tau}\left\|\widehat{\theta}_n - \theta_o\right\|_E^2 + c\left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\|_E^2 \tag{1.117}$$

w.p.a.1, where c is some generic positive constant. From (1.115) and (1.117), we can deduce that w.p.a.1,

$$\left\|\widehat{\theta}_n - \theta_o\right\|_E^2 \leq n^{2\tau}\sum_{j\in\mathcal{S}_\beta}\widehat{P}_{\lambda_n}(\beta_{o,j}) + cn^{2\tau}R_n, \tag{1.118}$$

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

Using Assumption 1.3.2.(iii), we obtain

$$\begin{aligned}\left|\widehat{P}_{\lambda_n}(\beta_{o,j}) - \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j})\right| \;\leq\; & \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right|\left\|\widehat{\beta}_n - \beta_o\right\|_E \\ & + \max_{j\in\mathcal{S}_\beta}\left|\frac{\widehat{P}''_{\lambda_n}(\beta_{o,j})}{2} + o_p(1)\right|\left\|\widehat{\beta}_n - \beta_o\right\|_E^2. \end{aligned} \tag{1.119}$$

By the triangular inequality,

$$\left\|G_k(\widehat{\theta}_n) - \widehat{\beta}_n\right\|_E^2 \geq \left\|\widehat{\beta}_n - \beta_o\right\|_E^2 - 2\left\|\widehat{\beta}_n - \beta_o\right\|_E J_{1,n} + J_{1,n}^2, \tag{1.120}$$

where $J_{1,n} = \left\|G_k(\widehat{\theta}_n) - G_k(\theta_o)\right\|_E = O_p\left(\left\|\widehat{\theta}_n - \theta_o\right\|_E\right)$. Denote $J_{2,n} = cn^{-2\tau}\left\|\widehat{\theta}_n - \theta_o\right\|_E^2$, then from (1.116), (1.117) and (1.119), we get

$$\left\|\widehat{\beta}_n - \beta_o\right\|_E^2 - c\left(\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + J_{1,n}\right)\left\|\widehat{\beta}_n - \beta_o\right\|_E \leq R_n - J_{1,n}^2 - J_{2,n}, \tag{1.121}$$

w.p.a.1. As $J_{1,n}^2 = O_p\left(\left\|\widehat{\theta}_n - \theta_o\right\|_E^2\right)$ and $J_{2,n} = O_p\left(\left\|\widehat{\theta}_n - \theta_o\right\|_E^2\right)$, hence from the inequality in (1.121), we can deduce that

$$\left\|\widehat{\beta}_n - \beta_o\right\|_E = O_p\left(\left\|\widehat{\theta}_n - \theta_o\right\|_E,\ \max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| + n^{-\frac{1}{2}}\right). \tag{1.122}$$

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

$$\left\|\widehat{\beta}_n - \beta_o\right\|_E = O_p\left(\left\|\widehat{\theta}_n - \theta_o\right\|_E\right). \tag{1.123}$$

Now, from (1.116) and (1.117), we have

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

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

Proof of Lemma 1.4.2. On the event $\left\{\widehat{\beta}_{n,j} \neq 0\right\}$ for some $j \in \mathcal{S}_\beta^c$, we have the

following KKT optimality condition:

$$\left| W_n^{(j)} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) \right] \right| = \frac{n^{\frac{1}{2}} \widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{2}. \tag{1.124}$$

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

$$n^{-\frac{1}{2}-\tau} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) = n^{-\tau} v_n \left[\rho(Z_i, \widehat{\alpha}_n) \right] + n^{\frac{1}{2}-\tau} E \left[\rho(Z_i, \widehat{\alpha}_n) \right] \tag{1.125}$$

By Lemma 1.4.1, Assumption 1.4.1.(iv) and Assumption 1.4.2.(i), we get

$$n^{\frac{1}{2}-\tau} E \left[g_{n,q}(Z_i, \widehat{\theta}_n) \right] = n^{\frac{1}{2}-2\tau} \frac{\partial G_{n,q}(\widetilde{\theta}_n)}{\partial \theta'} \left(\widehat{\theta}_n - \theta_o \right) = O_p(1) \tag{1.126}$$

where

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

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

$$\begin{aligned} & n^{\frac{1}{2}-\tau} E \left[g_k(Z_i, \widehat{\theta}_n) - \widehat{\beta}_n \right] \\ = \; & n^{\frac{1}{2}-\tau} \left\{ E \left[g_k(Z_i, \widehat{\theta}_n) - g_k(Z_i, \theta_o) \right] - \left(\widehat{\beta}_n - \beta_o \right) \right\} \\ = \; & \frac{\partial G_k(\widetilde{\theta}_n)}{\partial \theta} \left[n^{\frac{1}{2}-\tau} \left(\widehat{\theta}_n - \theta_o \right) \right] + O_p(1) = O_p(1) \end{aligned} \tag{1.127}$$

where

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

and $\widetilde{\theta}_{j,n}$ $(j = q+1, ..., q+k)$ lies between $\widehat{\theta}_n$ and θ_o. Hence, from the results in (1.125)-(1.127), Assumption 1.3.1.(iii) and Assumption 1.4.1.(iii), we can deduce

that

$$\left| W_n^{(j)} \left[n^{-\frac{1}{2}-\tau} \sum_{i=1}^{n} \rho(Z_i, \widehat{\alpha}_n) \right] \right| = O_p(1). \tag{1.128}$$

While from Assumption 1.4.2.(ii), we can deduce that

$$\frac{n^{\frac{1}{2}-\tau} \widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{2} = \frac{n^{\frac{1}{2}-\tau} r_n \lambda_n}{2} \frac{\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})}{r_n \lambda_n} \to_p \infty \tag{1.129}$$

Now, the KKT condition in (1.124), and (1.128) and (1.129) imply that $\Pr\left(\widehat{\beta}_{n,j} = 0\right) \to 1$ for any $j \in \mathcal{S}_\beta^c$. ■

Proof of Lemma 1.4.3. Applying Lemma 1.9.1, we get w.p.a.1

$$V_0(\widehat{\theta}_n, \widehat{\beta}_n) + \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\widehat{\beta}_{n,j}) \leq \sum_{j=1}^{k} \widehat{P}_{\lambda_n}(\beta_{o,j}) + R_n. \tag{1.130}$$

Now, conditional on the event $\{\widehat{\beta}_{j,n} = 0, j \in \mathcal{S}_\beta^c\}$, by Assumption 1.4.1.(iv) and Assumption 1.4.3, we can use similar arguments in the proof of Lemma 1.9.2 to deduce that

$$V_0(\widehat{\theta}_n, \widehat{\beta}_n) \geq c\left(1 + n^{-2\tau}\right) ||\widehat{\theta}_n - \theta_o||_E^2 + c||\widehat{\beta}_{n,+} - \beta_{o,+}||_E^2 \geq c||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E^2 \tag{1.131}$$

w.p.a.1, where c is some generic positive constant. Following the similar arguments in the proof of Lemma 1.3.3, we get

$$||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E^2 - \max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| \|\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\|_E \leq R_n \tag{1.132}$$

which implies that

$$||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E = O_p\left(\max_{j \in \mathcal{S}_\beta} \left| \widehat{P}'_{\lambda_n}(\beta_{o,j}) \right| + n^{-\frac{1}{2}} \right) \tag{1.133}$$

By Assumption 1.4.2.(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 (1.133) gives us $||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E = O_p(n^{-\frac{1}{2}})$. From the sparsity of $\widehat{\beta}_n$, we know that the event $\{\widehat{\beta}_{j,n} = 0, j \in \mathcal{S}_\beta^c\}$ has probability measure approaching 1. Hence we can deduce that $||\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}||_E = O_p(n^{-\frac{1}{2}})$, which finishes the proof. ■

Proof of Corollary 1.4.4. 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 1.3.5, we know that $\widehat{\beta}_{n,-} = 0$ w.p.a.1. Hence,$\sqrt{n}\,(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}})$ is the minimizer of $V_{3,n}(u_{\mathcal{S}})$ w.p.a.1. Using similar arguments in the proof of Lemma 1.9.1, one can deduce that

$$\upsilon_n \left\{ \left[\rho^s \left(Z, \alpha_{o,\mathcal{S}} + n^{-\frac{1}{2}} u_{\mathcal{S}} \right) - \rho \left(Z, \alpha_{o,\mathcal{S}} \right) \right] \right\} = o_p(1) \tag{1.134}$$

uniformly over K. From (1.134), we get

$$\begin{aligned} V^*_{3,n}(u_{\mathcal{S}}) &= n \left\{ E[\rho^s(Z, \alpha_{o,\mathcal{S}} + n^{-\frac{1}{2}} u_{\mathcal{S}})] \right\}' W_o \left\{ E[\rho^s(Z, \alpha_{o,\mathcal{S}} + n^{-\frac{1}{2}} u_{\mathcal{S}})] \right\} \\ &\quad + 2n^{\frac{1}{2}} \left\{ \upsilon_n \left[\rho(Z, \alpha_o) \right] \right\}' W_o \left\{ E[\rho^s(Z, \alpha_{o,\mathcal{S}} + n^{-\frac{1}{2}} u_{\mathcal{S}})] \right\} + o_p(1) \end{aligned} \tag{1.135}$$

uniformly over K. By Assumption 1.4.1.(iv), we have

$$E\left[g_{n,q}\left(Z,\theta_o+\frac{u_\theta}{\sqrt{n}}\right)\right]=n^{-\tau-\frac{1}{2}}\left(\begin{array}{cc}\frac{\partial G_{n,q}(\theta_o)}{\partial\theta'}, & 0\end{array}\right)u_{\mathcal{S}}+o(1) \tag{1.136}$$

and

$$E\left[g_k\left(Z_i,\theta_o+\frac{u_\theta}{\sqrt{n}}\right)-\left(\beta_{o,+}+\frac{u_{\beta_+}}{\sqrt{n}}\right)\right]=n^{-\frac{1}{2}}\frac{\partial m_k(\theta_o)}{\partial\alpha'_{\mathcal{S}}}u_{\mathcal{S}}+o(1). \tag{1.137}$$

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 (1.135)-(1.137) and Assumption 1.4.1.(iv), we can deduce that

$$V^*_{3,n}(u_{\mathcal{S}})=\left[\frac{\partial m(\theta_o)}{\partial\alpha'_{\mathcal{S}}}u_{\mathcal{S}}+2\upsilon_n\left[\rho(Z_i,\alpha_o)\right]\right]'W_o\left[\frac{\partial m(\theta_o)}{\partial\alpha'_{\mathcal{S}}}u_{\mathcal{S}}\right]+o_p(1), \tag{1.138}$$

uniformly over K. Note that there is

$$\left[\frac{\partial m(\theta_o)}{\partial\alpha_{\mathcal{S}}}\right]W_o\left[\frac{\partial m(\theta_o)}{\partial\alpha_{\mathcal{S}}}\right]'=\left[\frac{\partial m_k(\theta_o)}{\partial\alpha_{\mathcal{S}}}\right]W_{o,kk}\left[\frac{\partial m_k(\theta_o)}{\partial\alpha_{\mathcal{S}}}\right]'=M_+. \tag{1.139}$$

Using the same arguments used in the proof of Theorem 1.3.6, we can show that under Assumption 1.3.2.(iv) and Assumption 1.4.2.(i), there is

$$P_n=\sqrt{n}\widehat{P}'_{\lambda_n}(\beta_{o,j})u_{\beta_+,j}+\widehat{P}''_{\lambda_n}(\widetilde{\beta}_{o,j})u^2_{\beta_+,j}\to 0 \tag{1.140}$$

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

From the results in (1.138), (1.140) and triangle inequality, we can deduce that

$$V_{3,n}(u_{\mathcal{S}})\to_d V_3(u_{\mathcal{S}})\equiv u'_{\mathcal{S}}M_+u_{\mathcal{S}}+2u'_{\mathcal{S}}\left(\frac{\partial m(\theta_o)}{\partial\alpha'_{\mathcal{S}}}W_o\Psi(\theta_o,\beta_o)\right) \tag{1.141}$$

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

$$u_{\mathcal{S}} = -M_+^{-1}\left(\left[0, \frac{\partial m_k(\theta_o)}{\partial \alpha_{\mathcal{S}}'}\right] W_o \Psi(\theta_o, \beta_o)\right) \tag{1.142}$$

and $\sqrt{n}\left(\widehat{\alpha}_{n,\mathcal{S}} - \alpha_{o,\mathcal{S}}\right)$ is asymptotically tight. Now result in (1.55) follows by ACMT.

■

1.9.4 Proof of the Main results in Section 1.5

Proof of Lemma 1.5.1. 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

$$\widehat{P}_{\lambda_n}(\alpha_{o,j}) = \lambda_n \widehat{w}_{\alpha_j} \|\alpha_{o,j}\|_2 \to_p 0, \tag{1.143}$$

for any $j \in \mathcal{S}_\alpha$. Hence, Assumption 1.3.1.(iv) holds for the adaptive group Lasso penalty function. Now, the consistency of $\widehat{\alpha}_n$ follows by the similar arguments used in the proof of Lemma 1.3.1.

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 1.3.2.(iv). Now the convergence rate in (1.60) follows by similar arguments in the proof of Lemma 1.3.3. ■

Proof of Theorem 1.5.2. On the event $\left\{\|\widehat{\alpha}_{n,j}\|_2 \neq 0\right\}$, for some $j \in \mathcal{S}_\alpha^c$, we have

the following KKT optimality condition

$$2\left\|\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n)}{\partial\alpha_j}\right]W_n\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n)\right]\right\|_E=\left\|\frac{n\lambda_n\widehat{w}_{\alpha_j}\widehat{\alpha}_{n,j}}{\|\widehat{\alpha}_{n,j}\|_2}\right\|_E. \quad (1.144)$$

Following similar arguments used in the proof of Theorem 1.3.5, we can show that

$$n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n)=O_p(1). \quad (1.145)$$

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

$$\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n)}{\partial\beta_j}\right]W_n=W_{n,j} \quad (1.146)$$

where $W_{n,j}$ denotes the j-th component-wise rows of W_n. Hence by Assumption 1.3.1.(iii) and the result in (1.145), we get

$$\left\|\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n)}{\partial\alpha_j}\right]W_n\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\widehat{\alpha}_n)\right]\right\|_E=O_p(1). \quad (1.147)$$

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

$$\begin{aligned}
&\left\|n^{-1}\sum_{i=1}^{n}\frac{\partial\rho(Z_i,\widehat{\alpha}_n)}{\partial\theta_j}\right\|_E\\
\leq\ &\left\|n^{-1}\sum_{i=1}^{n}\left\{\frac{\partial g(Z,\widehat{\theta}_n)}{\partial\theta_j}-E\left[\frac{\partial g(Z,\widehat{\theta}_n)}{\partial\theta_j}\right]\right\}\right\|_E+\left\|E\left[\frac{\partial g(Z,\widehat{\theta}_n)}{\partial\theta_j}\right]\right\|_E\\
\leq\ &\left\|E\left[\frac{\partial g(Z,\widehat{\theta}_n)}{\partial\theta_j}\right]-E\left[\frac{\partial g(Z,\theta_o)}{\partial\theta_j}\right]\right\|_E+\left\|E\left[\frac{\partial g(Z,\theta_o)}{\partial\theta_j}\right]\right\|_E+o_p(1)\\
=\ &O_p(1), \quad (1.148)
\end{aligned}$$

where the first inequality is due to the triangle inequality, the second inequality is by the triangle inequality and Assumption 1.5.1, the last equality is by Assumption 1.3.2.(ii). By (1.148) and Cauchy-Schwarz inequality, result in (1.147) also holds for

$j \in \mathcal{S}_\theta^c$. Hence, by definition, (1.147) 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

$$\frac{1}{2}\left\|\frac{\sqrt{n}\lambda_n \widehat{w}_{\alpha_j}\widehat{\alpha}_{n,j}}{\|\widehat{\alpha}_{n,j}\|_2}\right\|_E = \frac{n^{\frac{1+\omega}{2}}\lambda_n}{2}\frac{1}{\|\sqrt{n}\widehat{\alpha}_{n,j}\|_2^{\omega}} \to_p \infty. \tag{1.149}$$

Now, using the results in (1.144), (1.147) and (1.149), we can deduce $\Pr\left(\|\widehat{\alpha}_{n,j}\|_2 = 0\right) \to 1$ for $j \in \mathcal{S}_\alpha^c$. ■

1.10 Appendix B

In this appendix, 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 1.10.1 *Suppose that $\lambda_n = o(1)$, then the bridge and SCAD penalty functions satisfy Assumption 1.3.1.(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 1.3.1.(iv).*

Proof of Corollary 1.10.1. 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 $\widehat{\beta}_{1st}$ and the Slutsky Theorem, we can deduce that

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

for all j. ■

Remark 1.10.2 *Compared with the results in Knight and Fu (2000) and Caner (2009) where bridge penalty is used, Corollary 1.10.1 imposes the same condition on λ_n to derive the consistency of the shrinkage estimator $\widehat{\alpha}_n$. 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 $\widehat{\alpha}_n$ based on adaptive Lasso penalty, our condition imposed on λ_n to derive the consistency is not stronger than that of Zou (2006).*

Corollary 1.10.3 *Suppose that $\lambda_n = o(1)$, then the bridge and SCAD penalty functions satisfy Assumption 1.3.2.(iv). If we further assume that $\widehat{\beta}_{1st}$ is a consistent estimator of β_o, then the adaptive Lasso penalty function also satisfies Assumption 1.3.2.(iv).*

Proof of Corollary 1.10.3. 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 1.3.2.(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 1.10.4 *Suppose that Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(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 Assumption 1.3.3 (i)-(ii); (ii) suppose that $\lambda_n n^{\frac{1}{2}} = o(1)$, $\lambda_n n^{\frac{1}{2}(1+\omega)} \to \infty$ and $\widehat{\beta}_{1st}$ is $\sqrt{n}$ consistent, then the adaptive Lasso penalty function satisfies Assumption 1.3.3 (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 1.3.3 (i)-(ii).*

Proof of Corollary 1.10.4. First if $\widehat{P}_{\lambda_n}(\beta_j) = \lambda_n \left|\beta_j\right|^\gamma$, then

$$\sqrt{n}\max_{j\in\mathcal{S}_\beta}\left|\widehat{P}'_{\lambda_n}(\beta_{o,j})\right| = \max_{j\in\mathcal{S}_\beta}\gamma n^{\frac{1}{2}}\lambda_n\left|\beta_{o,j}\right|^{\gamma-1} = o(1). \tag{1.150}$$

As Assumption 1.3.1.(iv), Assumption 1.3.2.(iv) and Assumption 1.3.3.(i) are satisfied, from Lemma 1.3.3, we can deduce that $\widehat{\beta}_{n,j} = O_p(n^{-\frac{1}{2}})$. 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

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

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

Next, if $\widehat{P}_{\lambda_n}(\beta_j) = \lambda_n \widehat{w}_{\beta_j}|\beta_j|$, then by the consistency of $\widehat{\beta}_{1st}$ 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

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

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 Assumption 1.3.1.(iv), Assumption 1.3.2.(iv) and Assumption 1.3.3.(i) are satisfied, from Lemma 1.3.3, we can deduce that $\widehat{\beta}_{n,j} = O_p(n^{-\frac{1}{2}})$. Let $r_n = 1 > 0$, then $n^{\frac{1}{2}}\lambda_n r_n = \lambda_n n^{\frac{1}{2}} \to \infty$. As $\widehat{\beta}_{n,j} = O_p(n^{-\frac{1}{2}})$, we have $|\sqrt{n}\widehat{\beta}_{n,j}| \leq \sqrt{n}\lambda_n$ w.p.a.1. From the definition of $\widehat{P}'_{\lambda_n}(\beta)$, i.e.

$$\widehat{P}'_{\lambda_n}(\beta) = \lambda_n \left\{ I(\beta \leq \lambda_n) + \frac{(\lambda_n a - \beta)_+}{(a-1)\lambda_n} I(\beta > \lambda_n) \right\}, \tag{1.151}$$

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

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

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

We next check Assumption 1.4.1.(v) and Assumption 1.4.2 using the bridge, adaptive Lasso and SCAD penalty functions. For the bridge penalty, Assumption 1.4.1.(v) and Assumption 1.4.2.(i) require $n^{\frac{1}{2}+\tau}\lambda_n = o(1)$. Under this condition and the assumptions of Lemma 1.4.1, we can deduce that $\widehat{\beta}_{n,j} = O_p(n^{\tau-\frac{1}{2}})$ for all $j \in \mathcal{S}_\beta^c$. Note that

$$\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j}) \Big/ \lambda = \gamma n^{(\frac{1}{2}-\tau)(1-\gamma)} |n^{\frac{1}{2}-\tau}\widehat{\beta}_{n,j}|^{\gamma-1}.$$

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 Assumption 1.4.1.(v) and Assumption 1.4.2.(i)-(ii) hold simultaneously. Secondly for the adaptive Lasso penalty, Assumption 1.4.1.(v) and Assumption 1.4.2.(i) require that $n^{\frac{1}{2}+\tau}\lambda_n = o(1)$. Note that

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

The first step estimator of $\beta_{o,j}$ $(j \in \mathcal{S}_\beta^c)$ 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, Assumption 1.4.1.(v) and Assumption 1.4.2.(i) require $n^{\frac{1}{2}+\tau}\lambda_n = o(1)$. Under the assumptions of Lemma 1.4.1, we can deduce that $\widehat{\beta}_{n,j} = O_p(n^{\tau-\frac{1}{2}})$ for all $j \in \mathcal{S}_\beta^c$. Note that

$$\widehat{P}'_{\lambda_n}(\widehat{\beta}_{n,j})\Big/ \lambda = I(\widehat{\beta}_{n,j} \leq \lambda_n) + \frac{(a\lambda_n - \widehat{\beta}_{n,j})_+}{(a-1)\lambda_n} I(\widehat{\beta}_{n,j} > \lambda_n).$$

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 Assumption 1.4.1.(v) and Assumption 1.4.2.(i)-(ii) hold simultaneously. ■

Chapter 2

Robust GMM Estimation with Irrelevant and Misspecified Moment Conditions

2.1 Introduction

It is well-know that the statistical properties of the generalized method of moments (GMM) estimate heavily rely on the quality of moment conditions. For example, misspecified moment conditions lead to inconsistent estimation. On the other hand, when the moment conditions only contain weak information about the structural coefficients, the GMM estimate will have slow rate of convergence and may not even be consistent. In both scenarios, the GMM estimate is highly biased in the finite samples. The moment selection methods proposed in chapter 1 are useful in reducing the risk of using the misspecified moment conditions in the GMM estimation in the finite samples. However, if some moment conditions in the second set are irrelevant [1],

[1] In this chapter, a moment condition is called irrelevant if including it into the GMM estimation neither produces consistent estimation nor improves the efficiency of the resulting GMM estimate.

these methods will select the irrelevant moment conditions and automatically include them into GMM estimation with probability approaching 1 (w.p.a.1). The irrelevant moment conditions do not effect the asymptotic properties of the GMM shrinkage estimate, but they may enlarge its bias in the finite samples.

In this chapter, we propose a new adaptive penalty based on which the GMM shrinkage estimation can consistently select the valid and relevant moment conditions. As a result, both the misspecified moment conditions and irrelevant moment conditions are not included in estimating the structural coefficients w.p.a.1. This chapter shares similar set-up of chapter 1. Specifically, we are interested in estimating some unknown parameter θ_o identified by the following moment restrictions

$$E\left[g_q(Z,\theta_o)\right]=0, \tag{2.1}$$

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

$$E\left[g_k(Z,\theta_o)\right]\stackrel{?}{=}0, \tag{2.2}$$

where "$\stackrel{?}{=}$" signifies that equality may hold for some but not others and $g_k(\cdot,\cdot)$: $R^{d_z}\times R^{d_\theta}\to R^k$. The goal of this chapter is to consistently select the valid and relevant moment conditions from (2.2) and include them into estimation of θ_o to improve efficiency.

Following the practice in chapter 1, we introduce a set of nuisance parameters β_o (which are also called moment selection coefficients in this chapter) and reparametrize

A more formal definition can be found in section 2.3.

the moment conditions in (2.2) to be

$$E\left[g_k(Z,\theta_o)-\beta_o\right]=0. \tag{2.3}$$

We stack the moment conditions in (2.1) and (2.3) to get

$$E\left[\rho(Z,\theta_o,\beta_o)\right]\equiv E\left[\begin{pmatrix} g_c(Z,\theta_o) \\ g_k(Z,\theta_o)-\beta_o \end{pmatrix}\right]=0. \tag{2.4}$$

The Lasso-type of GMM estimate $(\widehat{\theta}_n,\widehat{\beta}_n)$ of (θ_o,β_o) is defined as

$$(\widehat{\theta}_n,\widehat{\beta}_n)=\underset{(\theta,\beta)\in\Theta\times\mathcal{B}}{\arg\min}\frac{1}{n}\left[\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]'W_n\left[\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]+n\lambda_n\sum_{j=1}^{k}\omega_{n,j}\left|\beta_j\right|, \tag{2.5}$$

where $\Theta\times\mathcal{B}$ is the parameter space where (θ_o,β_o) lies; λ_n is the tuning parameter which directly controls the general penalty to all β_j, W_n is a $(q+k)\times(q+k)$ weight matrix and $\omega_{n,j}$ is some adaptive penalty related to individual β_j $(j=1,...,k)$. The adaptive penalty $\omega_{n,j}$ plays the key role in consistent selection of valid and relevant moment conditions. Ideally, we hope that $\omega_{n,j}$ would be large when the j-th moment condition in (2.2) is valid and relevant such that β_j is estimated as zero, otherwise $\omega_{n,j}$ should be small such that adding the penalty function to the GMM criterion does not effect the asymptotic properties of the estimates of moment selection coefficients of misspecified or irrelevant moment conditions.

When the adaptive penalty $\omega_{n,j}=1$ for all j, the penalty function in (2.5) reduces to the well-known Lasso penalty. As we have already discussed in the previous chapter, when $\omega_{n,j}=|\widehat{\beta}_{n,j,1st}|^{-\omega}$ $(\omega>0)$ with some first-step consistent estimator $\widehat{\beta}_{n,j,1st}$ of $\beta_{j,o}$, the penalty function in (2.5) reduces to the adaptive Lasso penalty. In this chapter, we show that if the tuning parameter satisfies the condition such

that the GMM Lasso estimate is root-n consistent, the GMM shrinkage estimation is conservative in moment selection, i.e. the probability of selecting the invalid moment conditions approaches 0 and the probability of selecting the valid moment conditions is strictly less than 1 when the sample size goes to infinity. Similar result is established for the GMM adaptive Lasso estimation, when the tuning parameter satisfies $\lambda_n = o(1)$ and $n^{\frac{1+\omega}{2}}\lambda_n \to \lambda_* \in R_+$. The key difference between the Lasso and adaptive Lasso penalty gives us the inspiration of designing the new adaptive penalty to ensure consistent selection of relevant moment conditions.

In section 2.3, we introduce an empirical measure of the information contained in moment condition. We show that the empirical measure converges to zero in probability, if the corresponding moment condition fails to improve the efficiency of the GMM estimate. On the other hand, if a moment condition is helpful to reduce the asymptotic variance of the GMM estimate, then its empirical information measure will converge to some nonzero constant in probability. The different probability limits of the empirical information measures of relevant and irrelevant moment conditions are useful for us to design a new adaptive penalty $\omega_{n,j}$ in (2.5) such that $\omega_{n,j}$ is small for the misspecified or irrelevant moment condition, and $\omega_{n,j}$ is large for the valid and relevant moment condition. We show that the GMM shrinkage estimation based on the new adaptive penalty is consistent in selecting the valid and strong moment conditions. As a result, the GMM shrinkage estimate is free of the misspecification risk in large samples and is robust to the irrelevant moment conditions in the finite samples.

The remainder of this chapter is organized as follows. Section 2.2 studies conservative moment selection in the GMM Lasso/adaptive Lasso estimation. In section 2.3, we first define a measure of information contained in moment condition and then, we show that such information measure can be consistently estimated. Based on

the empirical information measure, we propose a new adaptive penalty $\omega_{n,j}$, which enables the GMM shrinkage estimation in (2.5) to select the valid and relevant moment conditions in (2.2) w.a.p.1. Section 2.4 conducts some simulation experiments to investigate the finite sample properties of the GMM shrinkage estimation based on the new adaptive penalty. there. Section 2.5 concludes this chapter. Proofs and technical derivations are included in the Appendix.

2.2 GMM Shrinkage Estimation with Conservative Moment Selection

We call the GMM shrinkage estimate based on the Lasso (adaptive Lasso) penalty as Lasso (adaptive Lasso) GMM estimate. In this section, we derive the asymptotic properties of the GMM Lasso and GMM adaptive Lasso estimates. Following the notation in chapter 1, we use $\mathcal{S}_\beta \equiv \{j : \beta_{o,j} \neq 0,\ j = 1, ..., k\}$ to denote the index set of the non-zero components in β_o. Note that when the tuning parameter satisfies $\lambda_n = o(1)$ and the adaptive penalty satisfies $\omega_{n,j} = O_p(1)$ for any $j \in \mathcal{S}_\beta$, the penalty function $\widehat{P}_{\lambda_n}(\beta_j) = \lambda_n \omega_{n,j}|\beta_j|$ trivially satisfies Assumption 1.3.1.(iv) and Assumption 1.3.2.(iv). Applying Lemma 1.3.3 in chapter 1, we immediately get the following results.

Corollary 2.2.1 *Suppose that Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii) in chapter 1 are satisfied. If the tuning parameter satisfies* $\sqrt{n}\lambda_n = O(1)$ *and the adaptive penalty satisfies* $\omega_{n,j} = O_p(1)$ *for any* $j \in \mathcal{S}_\beta$*, then*

$$(\widehat{\theta}_n, \widehat{\beta}_n) = (\theta_o, \beta_o) + O_p(n^{-\frac{1}{2}}). \tag{2.6}$$

It is clear that both the Lasso penalty and the adaptive Lasso penalty satisfy the requirement that $\omega_{n,j} = O_p(1)$ for any $j \in \mathcal{S}_\beta$. Thus if the tuning parameter λ_n converges to zeros at the rate not slower than $n^{-1/2}$, the GMM Lasso and GMM adaptive Lasso estimates are root-n consistent. Recall that

$$\frac{\partial m(\alpha_o)}{\partial \alpha'} = \begin{pmatrix} \frac{\partial E[g_q(Z,\theta_o)]}{\partial \theta'} & 0 \\ \frac{\partial E[g_k(Z,\theta_o)]}{\partial \theta'} & -I_k \end{pmatrix} \text{ and } M = \frac{\partial m(\alpha_o)}{\partial \alpha} W_o \frac{\partial m(\alpha_o)}{\partial \alpha'}.$$

We next derive the limiting distribution of the GMM Lasso estimate.

Proposition 2.2.2 *Suppose Assumption 1.3.1.(i)-(iii), Assumption 1.3.2.(i)-(iii) and Assumption 1.3.4 in chapter 1 are satisfied. If the tuning parameter satisfies* $n^{\frac{1}{2}}\lambda_n \to \lambda_o \in [0, \infty)$ *as* $n \to \infty$, *then*

$$\sqrt{n}\left[(\widehat{\theta}_n, \widehat{\beta}_n) - (\theta_o, \beta_o)\right] \to_d \arg\min \left\{ V^*(u) + \lambda_o \left[\sum_{j \in \mathcal{S}_\beta} sgn(\beta_{j,o}) u_{\beta_j} + \sum_{j \in \mathcal{S}_\beta^c} |u_{\beta_j}| \right] \right\}$$

where $V^*(u) = u'Mu + 2u'\left[\frac{\partial m(\alpha_o)}{\partial \alpha}\right] W_o \Psi(\theta_o)$, $\Psi(\theta_o)$ *is defined in Assumption 1.3.4 and* $u' = (u'_\theta, u'_\beta)$, u_θ *and* u_β *are* d_θ *and* k *dimensional real vectors respectively.*

From Proposition 2.2.2, we see that when $\lambda_o = 0$, the GMM Lasso estimate has the same limiting distribution as that of the regular GMM estimate. However when $\lambda_o > 0$, the GMM Lasso estimate has non-standard limiting distribution which puts non-zero probability measure on the point zero. Moreover, Proposition 2.2.2 indicates that when $\lambda_o > 0$, the GMM Lasso estimates of zero components in β_o have nonzero high-order bias, which may contaminate other estimates by the interaction of the moment conditions.

Although Proposition 2.2.2 provides the joint asymptotic distribution of $(\widehat{\theta}_n, \widehat{\beta}_n)$, it does not give a complete story of the asymptotic properties of the GMM Lasso

estimate. For example, no moment selection results can be deduce from Proposition 2.2.2. Because what we can learn from Proposition 2.2.2 is that with nonzero probability, the estimates of the zero moment selection coefficients converge to zero with rate faster than $n^{-1/2}$. However Proposition 2.2.2 does not tell us if the zero moment selection coefficients are estimated exactly as zero with any nontrivial probability or not. The following proposition shows that GMM Lasso estimation is conservative in the moment selection.

Proposition 2.2.3 *Under the conditions of Proposition 2.2.2, we have*

$$\limsup_{n\to\infty} \Pr\left(\widehat{\beta}_{j,n} = 0, \text{ for all } j \in \mathcal{S}_\beta^c\right) \leq c(\lambda_o)$$

where $c(\lambda_o)$ *is some constant in* $[0,1)$ *for all* $\lambda_o \in [0,\infty)$.

Proposition 2.2.3 indicates that under the condition $n^{\frac{1}{2}}\lambda_n \to \lambda_o \in [0,\infty)$, the GMM Lasso estimation can only achieve conservative moment selection. Moreover, from the proof of Proposition 2.2.3, we see that even if one is willing to sacrifice the root-n consistency of the GMM Lasso estimate and let $n^{\frac{1}{2}}\lambda_n \to \infty$ (i.e. $\lambda_o = \infty$), the consistent moment selection may not be necessarily achieved. The conservative feature of the Lasso penalty is well-known in the variable selection literature. Proposition 2.2.3 shows that similar property holds when the Lasso penalty is used to perform moment selection.

We next derive the asymptotic distribution of the adaptive Lasso GMM estimate under the condition that $n^{\frac{1+\omega}{2}}\lambda_n \to \lambda_* \in [0,\infty)$ as $n \to \infty$. Define the first step estimates of the moment selection coefficients as

$$\widehat{\beta}_{1st,n} = n^{-1}\sum_{i=1}^{n} g_k(Z_i, \widehat{\theta}_{1st,n})$$

where $\widehat{\theta}_{1st,n}$ denotes the GMM estimate based on the moment conditions in (2.1), i.e.

$$\widehat{\theta}_{1st,n} = \arg\min_{\theta\in\Theta} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} g_q'(Z_i,\theta) \right] W_{q,n} \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} g_q(Z_i,\theta) \right] \tag{2.7}$$

and $W_{q,n}$ is some q by q weighting matrix which converges in probability to $W_{q,o}$, the asymptotic variance of the empirical process $\upsilon_n\left[g_q(Z,\theta_o)\right]$.

Under Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii) in chapter 1, it is a tedious excise to show that

$$\sqrt{n}(\widehat{\beta}_{1st,n} - \beta_o) = \overline{M}_k \upsilon_n\left[g(Z,\theta_o)\right] + o_p(1) \tag{2.8}$$

where

$$\overline{M}_k = \left(I_k, -\frac{\partial E[g_k(Z,\theta_o)]}{\partial\theta} M_q^{-1} \frac{\partial E[g_q(Z,\theta_o)]}{\partial\theta'} W_{q,o} \right) \tag{2.9}$$

and $M_q = \frac{\partial E[g_q(Z,\theta_o)]}{\partial\theta} W_{q,o} \frac{\partial E[g_q(Z,\theta_o)]}{\partial\theta'}$.

Proposition 2.2.4 *Suppose that Assumption 1.3.1.(i)-(iii), Assumption 1.3.2.(i)-(iii) and Assumption 1.3.4 in chapter 1 are satisfied. If the tuning parameter satisfies* $n^{\frac{1+\omega}{2}}\lambda_n \to \lambda_* \in [0,\infty)$ *as* $n\to\infty$, *then*

$$\sqrt{n}\left[(\widehat{\theta}_n,\widehat{\beta}_n) - (\theta_o,\beta_o)\right] \to_d \arg\min \left\{ V^*(u) + \lambda_* \sum\nolimits_{j\in\mathcal{S}_\beta^c} \frac{|u_{\beta_j}|}{|\overline{M}_k(j)\Psi(\theta_o)|^\omega} \right\}$$

where $V^*(u)$ *is defined in Proposition 2.2.3,* $\Psi(\theta_o)$ *is defined in Assumption 1.3.4 and* $\overline{M}_k(j)$ *denotes the* j*-th row of* $\overline{M}_k$.

Unlike the GMM Lasso estimate, the GMM adaptive Lasso estimate does not have the high order bias. Moreover, as implied by Proposition 2.2.4, the asymptotic distribution of the GMM adaptive Lasso estimate puts non-zero probability measure on the zero point. We can use similar arguments in the proof of Proposition 2.2.3 to

show that the GMM adaptive Lasso estimation is also conservative in the moment selection, if the tuning parameter λ_n satisfies $n^{\frac{1+\omega}{2}}\lambda_n \to \lambda_* \in [0, \infty)$.

Comparing Proposition 2.2.2 with Proposition 2.2.4, we see the adaptive penalty $\omega_{n,j}$ $(j = 1, ..., k)$ plays an important role in determining the asymptotic properties of the GMM shrinkage estimate. The Lasso penalty has the same effect on all moment selection coefficients, because $\omega_{n,j} = 1$ for all j. As a result, the GMM Lasso estimation fails to achieve consistent moment selection and its estimate suffers from second order bias. On the other hand, the adaptive Lasso penalty assigns large penalty to the estimates of zero components in β_o and small penalty to the estimates of nonzero components in β_o. As a result, the GMM adaptive Lasso estimation can achieve consistent moment selection. Moreover, even if the tuning parameter converges to zero fast enough such that only conservative moment selection can be achieved, the GMM adaptive Lasso estimate of θ_o is free of the second order asymptotic bias. However, the adaptive Lasso penalty does not take the information of the moment conditions into account, which is the main reason that it can not consistently select the relevant moment conditions. To achieve consistent selection of the relevant moment conditions, we have to revise the adaptive Lasso penalty such that the penalty would be large only if the related moment condition is valid and relevant, otherwise, it should be small. This intuition motivates us to find a measure of the relevance or information of the moment condition in (2.2). The measure of information should be zero if the related moment condition is irrelevant and nonzero otherwise. In the next section, we provide one of such measures and construct the adaptive penalty based on its estimate.

2.3 Moment Selection with Possibly Irrelevant and Misspecified Moment Conditions

In this section, we study the moment selection problem in the scenario that there may be irrelevant moment conditions in (2.2). We first introduce a measure of the information contained in the moment conditions. The relevant/irrelevant moment condition is defined using this information measure. We show that such information measure can be consistently estimated and we derive its asymptotic properties in the subsection 2.3.1. Using the estimate of the information measure, we construct a new adaptive penalty for the GMM shrinkage estimation. The new adaptive penalty is large when the corresponding moment condition is valid and relevant. On the other hand, it is small when the related moment condition is misspecified or irrelevant. In subsection 2.3.3, we show that the GMM shrinkage estimation based on the new adaptive penalty can consistently select the valid and relevant moment conditions.

2.3.1 Measure the Information of Moment Conditions

If we only use the moment conditions in (2.1), the asymptotic variance of the GMM estimate will be

$$Var_q \equiv \left[\frac{\partial m_q(\theta_o)}{\partial \theta} W_{q,o} \frac{\partial m_q(\theta_o)}{\partial \theta'}\right]^{-1}$$

where $m_q(\theta) \equiv E\left[g_q(Z,\theta)\right]$. The moment conditions in (2.2) are expected to be able to improve the efficiency of the GMM estimate. Thus any set of valid moment conditions $E\left[g_r(Z,\theta)\right] = 0$ from (2.2) is relevant and should be included into the GMM estimation only if

$$Var_{q+r} \equiv \left[\frac{\partial m_{q+r}(\theta_o)}{\partial \theta} W_{q+r,o} \frac{\partial m_{q+r}(\theta_o)}{\partial \theta'}\right]^{-1} \leq Var_q$$

and there exists some non-zero vector $\lambda \in R^{d_\theta}$ such that $\lambda'(Var_q - Var_{q+r})\lambda > 0$, where $W_{q+r,o}$ is the probability limit of the efficient weight matrix. As the matrix $Var_q - Var_{q+r}$ is always positive definite, its eigenvalues are always non-negative and the relevance requires that at least one of its eigenvalues is strictly larger than zero. Thus we can use the largest eigenvalue of $Var_q - Var_{q+r}$ as the measure of the information of the moment conditions $E[g_r(Z,\theta)] = 0$.

Definition 2.3.1 *Any subset of moment conditions* $E[g_r(Z,\theta)] \stackrel{?}{=} 0$ *in (2.2) is called redundant or irrelevant with respect to the moment conditions in (2.1) if the largest eigenvalue* τ_r *of the matrix* $Var_q - Var_{q+r}$ *satisfies* $\tau_r = O(\eta_n)$ *where* $\eta_n = o(1)$. *Otherwise, such subset of moment conditions is called non-redundant or relevant.*

Definition 2.3.1 is inspired by Hall, Inoue, Jana and Shin (2007) which uses the term $\log|Var_{q+r}|$ to define the entropy of the model with moment restrictions and uses the entropy based information criterion to select relevant moment conditions. However, moment selection issue addressed in this chapter is in a different scenario, because the moment conditions in (2.2) could be misspecified, while the moment conditions to be selected in Hall, Inoue, Jana and Shin (2007) are all assumed to be valid.

We call Var_{q+r} as pseudo asymptotic variance, because it is the asymptotic variance of GMM estimate based on the moment conditions in (2.1) and $E[g_{q+r}(Z,\theta)] \stackrel{?}{=} 0$ with assuming that the moment conditions $E[g_r(Z,\theta)] \stackrel{?}{=} 0$ are valid. The information measure τ_r has the nice property that if a moment condition is irrelevant, then $\tau_r = 0$, otherwise, $\tau_r \neq 0$.

By definition, a moment condition is irrelevant if it fails to improve the efficiency of the GMM estimate based on the moment conditions in (2.1). The most obvious example of the redundant moment conditions are those constructed by linear com-

binations of the moment conditions in (2.1). The following remark presents more examples.

Remark 2.3.2 *Consider the following linear IV model*

$$\begin{aligned} Y_i &= X_i\theta_{1,o} + W_i'\theta_{2,o} + u_i \\ X_i &= Z_{1,i}\pi_{1,o} + Z_{2,i}'\pi_{2,o} + W_i'\delta_o + v_i, \end{aligned}$$

where X_i is an endogenous variable, W_i is a set of exogenous variables, $Z_{1,i}$ and $Z_{2,i}$ are valid IVs for X_i and $\pi_{1,o}$ is a fixed nonzero real number. It is noticed in Breusch, Qian, Schmidt and Wyhowski (1999) that if $\pi_{2,o} = 0$, then adding the moment conditions constructed using $Z_{2,i}$ into GMM estimation of $\theta_o = (\theta_{1,o}, \theta_{2,o})$ does not effect the asymptotic variance-covariance matrix of the GMM estimate. Thus in this case, if we use τ_{Z_2} to measure the information of the aforementioned moment conditions, then $\tau_{Z_2} = 0$. On the other hand, if $Z_{2,i}$ are weak IVs in the sense of Staiger and Stock (1997), then $\pi_{2,o} = \pi_2 n^{-\frac{1}{2}}$ where π_2 is some nonzero finite constant vector. Under some regularity conditions, one can use similar arguments in Hall, Inoue, Jana and Shin (2007) to show that $\tau_{Z_2} = O(n^{-\frac{1}{2}})$.

In the finite samples, τ_r can be estimated by the largest eigenvalue $\tau_{n,r}$ of $Var_{n,q} - Var_{n,q+r}$, where

$$Var_{n,q} = \left[\left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_q(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right) W_{q,n} \left(\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_q(Z_i, \widehat{\theta}_{1st})}{\partial \theta'} \right) \right],$$

$\widehat{\theta}_{1st}$ and $W_{q,n}$ are defined in (2.7) and $Var_{n,q+r}$ is defined similarly. The following assumption is useful for deriving the asymptotic properties of $\tau_{n,r}$.

Assumption 2.3.1 *(i) The following functional central limit theorem (FCLT) holds*

$$\sup_{\theta\in\Theta}\left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\left\{\frac{\partial g_l(Z_i,\theta)}{\partial\theta}-E\left[\frac{\partial g_l(Z,\theta)}{\partial\theta}\right]\right\}\right]=O_p(1) \tag{2.10}$$

for $l=q,k$*; (ii)* $E\left[\frac{\partial g^2(Z,\theta)}{\partial\theta\partial\theta'}\right]$ *is continuous in the local neighborhood of* θ_o *and is finite at* θ_o*; (iii)* $W_n=W_o+O_p(n^{-1/2})$ *where* $W_o=\{E[\Psi(\theta_o)\Psi'(\theta_o)]\}^{-1}$.

Assumption 2.3.1.(i) is a high-level condition, which can be verified by applying the Donsker's theorem in specific models. Assumption 2.3.1.(ii) imposes some smoothness condition on the expectation of the second derivative of the moment function $g(Z,\theta)$. Assumption 2.3.1.(iii) imposes some restriction on the convergence rate of the estimate of efficient weighting matrix W_o. Let $\mathcal{S}$ to be the index set of the moment conditions in (2.2). Let τ_j and $\tau_{n,j}$ be the theoretical and empirical information measures of the j-th moment condition in (2.2) respectively.

Lemma 2.3.3 *Under Assumption 1.3.1.(i)-(iii), 1.3.2.(i)-(iii) in chapter 1 and Assumption 2.3.1, we have* $\tau_{n,j}=\tau_j+O_p(n^{-1/2})$ *for any* $j\in\mathcal{S}$.

From Lemma 2.3.3, we know that when the j-th moment condition in (2.2) is irrelevant, the empirical information measure satisfies $\tau_{n,j}=O_p(\eta_n\vee n^{-1/2})$. That is an important feature for constructing the new adaptive penalty, as illustrated in the next subsection.

2.3.2 A New Adaptive Penalty

In the GMM shrinkage estimation defined in (2.5), the adaptive penalty controls the level of penalization on each individual moment selection coefficient. As illustrated in Section 2.2, a well designed adaptive penalty is the key element for the

GMM shrinkage estimation achieving consistent moment selection. The usual adaptive Lasso penalty only takes the sparsity of the moment selection coefficients into account. As a result, the GMM adaptive Lasso estimation selects not only the valid and relevant moment conditions, but also valid and irrelevant moment conditions with probability approaching 1.

To ensure irrelevant moment conditions are not selected in the GMM shrinkage estimation, we need to revise the adaptive penalty such that the penalty is large only if the moment condition is valid and relevant, otherwise it should go to zero as fast as possible. Based on the empirical information measure devised in the previous subsection, we propose the following adaptive penalty

$$\omega_{\tau,j,n} = \tau_{n,j}^{\omega_1} |\widehat{\beta}_{n,j,1st}|^{-\omega_2} \text{ for any } j \in \mathcal{S} \tag{2.11}$$

where the constants ω_1 and ω_2 satisfy $(\eta_n \vee n)^{\frac{-\omega_1}{2}} n^{\frac{\omega_2}{2}} = o(1)$.

If the j-th moment condition in (2.2) is valid and relevant, then $\tau_{n,j} \to_p \tau_j \neq 0$ and $\widehat{\beta}_{n,j,1st} = O_p(n^{-1/2})$. Thus in this case, we have $\omega_{\tau,j,n} \to_p \infty$. Secondly, if the j-th moment condition in (2.2) is invalid, then

$$\omega_{\tau,j,n} = \tau_{n,j}^{\omega_1} |\widehat{\beta}_{n,j,1st}|^{-\omega_2} \to_p \tau_j^{\omega_1} |\beta_{j,o}|^{-\omega_2} < \infty. \tag{2.12}$$

Finally, if the j-th moment condition in (2.2) is valid and irrelevant, then

$$\omega_{\tau,j,n} = (\eta_n \vee n)^{\frac{-\omega_1}{2}} n^{\frac{\omega_2}{2}} \left[(\eta_n \vee n)^{1/2} \tau_{n,j}\right]^{\omega_1} |n^{1/2} \widehat{\beta}_{n,j,1st}|^{-\omega_2} \to_p 0. \tag{2.13}$$

To sum up, the new adaptive penalty diverges to infinity only if the related moment condition is valid and relevant. Otherwise, it will be bounded or even converge to zero. As illustrated in the next subsection, these properties are the key

elements for the GMM shrinkage estimation being able to distinguish the relevant and irrelevant moment conditions.

2.3.3 Consistent Moment Selection and Robust Estimation

Using the new adaptive penalty, we define the new GMM shrinkage estimate as

$$(\widehat{\theta}_{\tau,n}, \widehat{\beta}_{\tau,n}) = \underset{(\theta,\beta)\in\Theta\times\mathcal{B}}{\arg\min} \frac{1}{n}\left[\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right]' W_n \left[\sum_{i=1}^{n}\rho(Z_i,\theta,\beta)\right] + n\lambda_n \sum_{j=1}^{k}\omega_{\tau,j,n}\left|\beta_j\right|. \tag{2.14}$$

From the discussion of the previous subsection, it is clear that when the tuning parameter λ_n satisfies $\lambda_n = o(1)$, the penalty function $\widehat{P}_{\lambda_n}(\beta_j) = \lambda_n\omega_{\tau,j,n}|\beta_j|$ trivially satisfies Assumption 1.3.1.(iv) and Assumption 1.3.2.(iv). The following corollary can be directly deduced from Lemma 1.3.3 in chapter 1.

Corollary 2.3.4 *Suppose that Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii) in chapter 1, and Assumption 2.3.1 are satisfied. If the tuning parameter satisfies* $\sqrt{n}\lambda_n = O(1)$*, then*

$$(\widehat{\theta}_{\tau,n}, \widehat{\beta}_{\tau,n}) = (\theta_o, \beta_o) + O_p(n^{-\frac{1}{2}}). \tag{2.15}$$

Under Assumption 1.3.1.(i)-(iii), Assumption 1.3.2.(i)-(iii) in chapter 1 and Assumption 1.3.4, we can use (2.8) to deduce that

$$\sqrt{n}(\widehat{\beta}_{1st,n} - \beta_o) \to_d \overline{M}_k\Psi(\theta_o) \tag{2.16}$$

where $\overline{M}_k$ is defined in (2.9) and $\Psi(\theta_o)$ is defined in Assumption 1.3.4. Let $\mathcal{S}_{*,\beta}$ be the index set of the misspecified or irrelevant moment conditions in (2.2), then by definition $\mathcal{S}^c_{*,\beta}$ is the index set of valid and relevant moment conditions. By definition, $\mathcal{S}_\beta \subset \mathcal{S}_{*,\beta}$ and $\mathcal{S}^c_{*,\beta} \subset \mathcal{S}^c_\beta$. Let $r_n = n^{\frac{\omega_2}{2}}$, from Lemma 2.3.3 and (2.16), we can show

that

$$\frac{\omega_{\tau,j,n}}{r_n} = \tau_{n,j}^{\omega_1} |n^{1/2} \widehat{\beta}_{n,j,1st}|^{-\omega_2} > 0 \ a.e. \tag{2.17}$$

for any $j \in \mathcal{S}_{*,\beta}^c$. Thus if the tuning parameter satisfies $n^{\frac{1+\omega_2}{2}} \lambda_n \to \infty$ as $n \to \infty$, then Assumption 1.3.3.(ii) is satisfied for any $j \in \mathcal{S}_{*,\beta}^c$. Under Lemma 2.3.3 and (2.16), if $\sqrt{n}\lambda_n = o(1)$, we can show that

$$\sqrt{n}\lambda_n \omega_{\tau,j,n} = \sqrt{n}\lambda_n \tau_{n,j}^{\omega_1} |\widehat{\beta}_{n,j,1st}| = o_p(1) \tag{2.18}$$

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

$$\sqrt{n}\lambda_n \omega_{\tau,j,n} = \frac{\sqrt{n}\lambda_n \left[(\eta_n \vee n)^{1/2} \tau_{n,j} \right]^{\omega_1} n^{\frac{\omega_2}{2}}}{(\eta_n \vee n)^{\frac{\omega_1}{2}} |n^{1/2} \widehat{\beta}_{n,j,1st}|^{\omega_2}} = o_p(1) \tag{2.19}$$

for any $j \in \mathcal{S}_\beta / \mathcal{S}_{*,\beta}$. Results in (2.18) and (2.19) imply that Assumption 1.3.3.(i) is satisfied for any $j \in \mathcal{S}_{*,\beta}$. Note the role that the empirical information measure plays here. Without the empirical information measure $\tau_{n,j}$, one can show that (2.17) also holds with $r_n = n^{\frac{\omega_2}{2}}$ for all $j \in \mathcal{S}_\beta / \mathcal{S}_{*,\beta}$. Thus, from Theorem 1.3.5, we know that $\Pr\left(\widehat{\beta}_{\tau,n,j} = \beta_{o,j}\right) \to 1$ for all $j \in \mathcal{S}_\beta / \mathcal{S}_{*,\beta}$. The empirical information measure changes the asymptotic properties of $\widehat{\beta}_{\tau,n,j}$ ($j \in \mathcal{S}_\beta / \mathcal{S}_{*,\beta}$), as we will see later in this subsection. The following corollary is an immediate result of Theorem 1.3.5.

Corollary 2.3.5 *Suppose that Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii) in chapter 1 and Assumption 2.3.1 hold. If the tuning parameter λ_n satisfies $n^{\frac{1+\omega_2}{2}} \lambda_n \to \infty$ and $\sqrt{n}\lambda_n = o(1)$, then we have*

$$\Pr\left(\widehat{\beta}_{\tau,n,j} = \beta_{o,j}\right) \to 1 \ \text{as} \ n \to \infty \tag{2.20}$$

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

Corollary 2.3.5 implies that the potentially valid and relevant moment conditions in (2.2) are selected w.p.a.1. On the other hand, Corollary 2.3.4 implies that the invalid moment conditions are not selected in the GMM shrinkage estimation w.p.a.1. Thus, to show that the irrelevant moment conditions are not selected w.p.a.1, it is sufficient to show that $\widehat{\beta}_{\tau,n,j} \neq 0$ w.p.a.1 for any $j \in \mathcal{S}_\beta/\mathcal{S}_{*,\beta}$. Such result can be trivially proved if we can show that $\sqrt{n}\widehat{\beta}_{\tau,n,j}$ $(j \in \mathcal{S}_\beta/\mathcal{S}_{*,\beta})$ has non-degenerated limiting distribution.

Without loss of generality, we sort the moment conditions in (2.2) such that $\beta_o = (\beta_{o,+}, \beta_{o,\tau,-}, \beta_{o,-})$, where $\beta_{o,+} \neq 0$ are the moment selection coefficients of the misspecified moment conditions, $\beta_{o,\tau,-} = 0$ are the moment selection coefficients of the valid and irrelevant moment conditions and $\beta_{o,-} = 0$ are the moment selection coefficients of the valid and relevant moment conditions. Denote $\alpha_{o,\tau,\mathcal{S}} = (\theta_o, \beta_{o,+}, \beta_{o,\tau,-})$ and accordingly $\widehat{\alpha}_{n,\tau,\mathcal{S}} = (\widehat{\theta}_{\tau,n}, \widehat{\beta}_{n,\tau,+}, \widehat{\beta}_{n,\tau,-})$ to be its GMM shrinkage estimate. For any vector γ, we use d_γ to denote its dimensionality.

Theorem 2.3.6 *Suppose that Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii) in chapter 1 and Assumption 2.3.1 hold. If the tuning parameter λ_n satisfies $n^{\frac{1+\omega_2}{2}}\lambda_n \to \infty$ and $\sqrt{n}\lambda_n = o(1)$, then we have*

$$\sqrt{n}\left(\widehat{\alpha}_{n,\tau,\mathcal{S}} - \alpha_{o,\tau,\mathcal{S}}\right) \to_d u^*_{\mathcal{S}_{*,\beta}} \tag{2.21}$$

where $u^*_{\mathcal{S}_{*,\beta}}$ is a $d_{\theta_o} + d_{\beta_{o,+}} + d_{\beta_{o,\tau,-}}$ dimensional random vector with joint $N(0, M^{-1}_{\tau,11})$ distribution, $M_{\tau,11} \equiv \frac{\partial m(\theta_o)}{\partial \alpha_{\tau,\mathcal{S}}} W_o \frac{\partial m(\theta_o)}{\partial \alpha'_{\tau,\mathcal{S}}}$ and $\alpha_{\tau,\mathcal{S}} \equiv (\theta, \beta_+, \beta_{\tau,-})$.

From Theorem 2.3.6, we see that $\sqrt{n}\widehat{\beta}_{\tau,n,j}$ $(j \in \mathcal{S}_\beta/\mathcal{S}_{*,\beta})$ has asymptotic normal distribution, which implies that

$$\begin{aligned} & \limsup_{n\to\infty} \Pr\left(\widehat{\beta}_{\tau,n,j} = 0, j \in \mathcal{S}_\beta/\mathcal{S}_{*,\beta}\right) \\ = & \limsup_{n\to\infty} \Pr\left(\sqrt{n}\widehat{\beta}_{\tau,n,j} = 0, j \in \mathcal{S}_\beta/\mathcal{S}_{*,\beta}\right) \\ \leq & \Pr\left(u^*_{\beta_{\tau,-}} = 0\right) = 0 \end{aligned}$$

where $u^*_{\beta_{\tau,-}}$ denotes the last $d_{\mathcal{S}_{*,\beta}} - d_{\mathcal{S}_\beta}$ elements of $u^*_{\mathcal{S}_{*,\beta}}$. Thus, we immediately have the following result.

Corollary 2.3.7 *Suppose that Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii) in Chapter 1 and Assumption 2.3.1 hold. If the tuning parameter λ_n satisfies $n^{\frac{1+\omega_2}{2}}\lambda_n \to \infty$ and $\sqrt{n}\lambda_n = o(1)$, then we have*

$$\Pr\left(\widehat{\beta}_{\tau,n,j} \neq 0\right) \to 1 \text{ as } n \to \infty \tag{2.22}$$

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

Let $\mathcal{S}^c_{n,*,\beta}$ to be the index set of the zero components in $\widehat{\beta}_{\tau,n}$. Combining the results in Corollary 2.3.4, Corollary 2.3.5 and Corollary 2.3.7, we can deduce that

$$\Pr\left(\mathcal{S}^c_{n,*,\beta} = \mathcal{S}^c_{*,\beta}\right) \to 1 \text{ as } n \to \infty \tag{2.23}$$

which implies that the valid and relevant moment conditions in (2.2) are consistently selected w.p.a.1. From Theorem 2.3.6, we can use similar arguments in the proof of Theorem 1.3.7 to show that the asymptotic variance-covariance matrix of the GMM shrinkage estimate $\widehat{\theta}_{\tau,n}$ is determined by the moment conditions in (2.1) and the valid and relevant moment conditions in (2.2). The asymptotic properties of GMM

shrinkage estimate $\widehat{\theta}_{\tau,n}$ are not effected by the misspecified or irrelevant moment conditions.

2.4 Simulation Studies

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

$$Y_i = X_i\theta_o + u_i, \tag{2.24}$$

$$X_i = Z'_{1,i}\pi_{1,o} + Z'_{2,i}\pi_{3,o} + v_i, \tag{2.25}$$

where Y_i is a scaler dependent variable, X_i is a scaler endogenous variable, $Z'_{1,i} = (Z_{11,i}, Z_{12,i})$ contains the IVs whose validity are assumed to be known, $Z'_{2,i} = (Z_{21,i}, Z_{22,i})$ is a set of potentially valid IVs, u_i and v_i are error terms which are correlated with each other.

Suppose a econometrician specifies the model

$$Y_i = X_i\theta_o + u_i$$

with the moment conditions $E[u_i Z_{1,i}] = 0$ to identify and consistently estimate θ_o. The potentially valid IVs in $Z_{2,i}$ are mixed with 4 invalid IVs $F'_{1i} = (F_{11,i}, ..., F_{14,i})$ and 4 irrelevant IVs $F'_{2i} = (F_{21,i}, ..., F_{24,i})$ to construct the following moment conditions

$$E\left[u_i(Z'_{2,i}, F'_i)\right] \stackrel{?}{=} 0 \tag{2.26}$$

where $F'_i = (F'_{1i}, F'_{2i})$.

To generate the simulated data, we first generate $(Z_{1,i}, Z'_{2,i}, u_i, v_i, F^*_i)$ from a multivariate normal distribution with mean 0 and variance-covariance matrix Σ,

where $F_i^* = (F_{1,i}^*, ..., F_{8,i}^*)$, $\Sigma = diag(\Sigma_Z, \Sigma_{u,v}, \mathbf{I}_8)$, Σ_Z is a 3×3 matrix with the i,j-th element being $0.2^{|i-j|}$, $\Sigma_{u,v}$ is a 2×2 matrix with diagonal elements $(0.5, 1)$ and off-diagonal elements $(0.6, 0.6)$, $\mathbf{I}_8$ is a 8×8 identity matrix. Let c_l to be some value between 0 and 0.8 and l be a 1×4 vector with the j-th ($j = 1, ..., 4$) element being $c_l + (0.8 - c_l) * (j-1)/3$. The invalid IVs are generated in the following way

$$F'_{1j,i} = F^*_{j,i} + u_i \times l \text{ for } j = 1, ...4.$$

The irrelevant IVs F_{2i} are simply $F^*_{j,i}$ ($j = 5, ..., 8$).

It is clear that when c_l is close to zero, the IV in $F_{1,i}$ with smallest index number (i.e. $F_{11,i}$) behaves more like valid IVs and it becomes more difficult to distinguish it from the potentially valid ones. We choose different values for c_l ($c_l = 0.2$ or 0.5) to see how our method works in different scenarios. The parameters in the model (2.24)-(2.25) take the following values

$$\theta_o = 0.5,\ \pi'_{1,o} = (\pi_{11,o}, 0.1) \text{ and } \pi'_{3,o} = (0.5, 0.5).$$

When $\pi_{1,o}$ is close to zero, $Z_{1,i}$ may contain weak information about the unknown parameter θ_o, which may also effect the performance of our method in moment selections. In the simulation studies, we choose different values for $\pi_{11,o}$ ($\pi_{11,o} = 0.1$ or 0.3) to see how our method is effected by the signal strength of $Z_{1,i}$.

For each specification of $(c_l, \pi_{11,o})$, we use the simulated samples with sample sizes $n = 250$ and 2500 respectively and for each sample size, 5,000 simulated samples are drawn from the data generating mechanism. The penalty function with $\omega_1 = 2$ and $\omega_2 = 3$ is used to construct the criterion of GMM shrinkage estimation. We use the projected scaled sub-gradient method (active-set variant) method proposed in Schmidt (2010) to solve the minimization problem in the GMM shrinkage estimation.

Table 2.1 Performance of GMM Shrinkage Method in Moment Selection

	$\pi_{11,o} = 0.1$							
	$n = 250$				$n = 2500$			
$c_l = 0.2$	(.0112	.4908	.4866	.0114)	(.0000	.9026	.0950	.0024)
$c_l = 0.5$	(.0016	.4944	.4932	.0108)	(.0000	.9028	.0946	.0026)
	$\pi_{11,o} = 0.3$							
	$n = 250$				$n = 2500$			
$c_l = 0.2$	(.0006	.6874	.1884	.1236)	(.0000	.9602	.0278	.0120)
$c_l = 0.5$	(.0000	.6888	.1878	.1234)	(.0000	.9606	.0284	.0011)

Table 2.1: The four numbers in each bracket (from left to right) are the estimated finite sample probabilities of selecting subsets of moment conditions in the second set from four different categories respectively. The first category includes the subsets of moment conditions which contain at least one invalid moment condition. The second category contains and only contains the subset of all potentially valid and relevent moment conditions in the second set. The third category includes the subsets which have all the valid and relevent moment conditions and do not have the invalid moment conditions, but fail to rule out all irrelevent moment conditions in the second set. The fourth category includes the rest of possible subsets of moment conditions from the second set. The finite sample probabilities are computed based on 5,000 replications.

Table 2.1 presents the finite sample probabilities of the GMM shrinkage estimation selecting different subsets of moment conditions from (2.2). The first number in each bracket is the probability of any invalid IVs to be selected in the finite samples. The GMM shrinkage estimation does very well in ruling out the invalid IVs. Even in the worst scenario that the signal strength of the IVs in $Z_{1,i}$ is weak (i.e. $\pi_{11,o} = 0.1$) and the sample size is small (i.e. $n = 250$), the probability that invalid moment conditions are selected is only 0.011. We see that when the sample size is small and the IVs in $Z_{1,i}$ are weak, the probability of selecting the valid and relevant moment conditions is around 0.50. Given the small sample size, this probability increases when the IVs in $Z_{1,i}$ becomes strong. When the sample size is small, the severity of misspecification plays some role in ruling out the invalid IVs. However, its effect is diminishing with the sample size growing. When the sample size is large (i.e. $n = 2500$), the probabilities of selecting the valid and relevant moment conditions

Table 2.2 Finite Sample Bias (BS), Standard Deviations (SD) and RMSEs (RE)

	GMM Shrinkage Estimate						Conservative GMM Estimate					
	n=250			n=2500			n=250			n=2500		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.1 .2)	.0030	.0847	.0848	.0001	.0248	.0248	-.0035	.2613	.2613	.0006	.0786	.0786
(.1 .5)	.0026	.0856	.0857	.0001	.0248	.0248	-.0035	.2613	.2613	.0006	.0786	.0786
(.3 .2)	.0042	.0816	.0817	.0001	.0232	.0232	-.0013	.1614	.1614	.0003	.0501	.0501
(.3 .5)	.0042	.0815	.0816	.0001	.0232	.0232	-.0013	.1614	.1614	.0003	.0501	.0501
	Pooled GMM Estimate						Aggressive GMM Estimate					
	n=250			n=2500			n=250			n=2500		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.1 .2)	.0057	.0810	.0812	.0004	.0248	.0248	.1068	.1265	.1655	.1034	.0404	.1110
(.1 .5)	.0057	.0810	.0812	.0004	.0248	.0248	.1377	.1422	.1979	.1364	.0463	.1441
(.3 .2)	.0049	.0754	.0755	.0004	.0232	.0232	.0931	.1187	.1508	.0902	.0378	.0977
(.3 .5)	.0049	.0754	.0755	.0004	.0232	.0232	.1203	.1337	.1799	.1191	.0433	.1267
	Post-Shrinkage GMM Estimate						Oracle GMM Estimate					
	n=250			n=2500			n=250			n=2500		
	BS	SD	RE	BS	SD	RE	BS	SD	RE	BS	SD	RE
(.1 .2)	.0033	.0821	.0821	.0000	.0247	.0247	.0021	.0800	.0800	.0000	.0247	.0247
(.1 .5)	.0028	.0842	.0842	.0000	.0247	.0247	.0021	.0800	.0800	.0000	.0247	.0247
(.3 .2)	.0054	.0904	.0906	.0002	.0237	.0237	.0017	.0744	.0744	.0000	.0231	.0231
(.3 .5)	.0054	.0904	.0906	.0002	.0237	.0237	.0017	.0744	.0744	.0000	.0231	.0231

Table 2.2: The finite sample bias , standard error and mean square error are computed using the corresponding estimates from 5,000 replications. The Oracle GMM estimate is based on all valid and relevent moment conditions. The conservative GMM estimate only uses the known valid IVs $Z_{1,i}$. The pooled GMM estimate uses all valid moment conditions including the irrelevant ones. The aggressive GMM estimate is based on all available IVs including the invalid ones. The post-shrinkage GMM estimate refers to the GMM estimate based on the known valid IVs $Z_{1,i}$ and selected (by GMM shrinkage estimation) IVs from the second set.

approach 1 in all scenarios. Of course, such phenomenon is predicted by the result in (2.23).

Table 2.2 summarizes the finite sample properties of the GMM shrinkage estimate and various GMM estimates based on different sets of moment conditions. Compared with the GMM estimate using only the known valid IVs (i.e. the conservative GMM estimate in the table), the GMM shrinkage estimate enjoys smaller finite sample bias and at the same time, smaller standard error. In all of the scenarios we considered in the simulation, the GMM shrinkage estimate dominates the conservative GMM estimate. Compared with the GMM estimate using all valid moment conditions including the irrelevant ones (i.e. the pooled GMM estimate in the table), the GMM shrinkage estimate has smaller finite sample bias, though when the sample size is small (i.e. $n = 250$), its standard error is slightly larger than that of the pooled GMM estimate. From Table 2.1, we see that the GMM shrinkage estimation rules out the invalid and irrelevant moment conditions with large probability, as a result, its estimate is robust to the finite sample bias incurred by these moment conditions. The pooled GMM estimate is not effected by the invalid IVs, but it suffers from the finite sample bias caused by the irrelevant moment conditions. This explains why the GMM shrinkage enjoys smaller finite sample bias. Its standard error is slightly larger when $n = 250$, because there is nontrivial probability that at least one of the valid and relevant IVs in the second set are not selected in the GMM shrinkage estimation. When the sample size increases, we see that the GMM shrinkage estimate and the pooled GMM estimate have the same standard error, but the finite sample bias of the GMM shrinkage estimate remains to be smaller. The GMM shrinkage estimate using all available IVs (i.e. the aggressive GMM estimate) is inconsistent and suffers from large finite sample bias, as illustrated in Table 2.2. Compared with the post-shrinkage GMM estimate, it is interesting to see that the GMM shrinkage estimate

has smaller bias, though its standard error is slightly large. The difference between these two estimates is very small even when the sample size is small. When the sample size is large, it is clear that the GMM shrinkage estimate, the post-shrinkage GMM estimate and the GMM estimate based all valid and relevant moment conditions (i.e. the Oracle GMM estimate) are almost the same in terms of the finite sample properties.

2.5 Conclusion

This chapter studies the GMM shrinkage estimation with ℓ-1 type of penalty functions, which includes the GMM Lasso/adaptive Lasso estimation as special examples. We show that the GMM Lasso estimation is conservative in moment selection, which means that with the sample size approaching infinity, the misspecified moment conditions are not selected with probability approaching 1 and the valid moment conditions are selected with probability strictly less than 1. The similar result is established for the GMM adaptive Lasso estimation, when the tuning parameter converges to zero fast enough (i.e. $n^{\frac{1+\omega}{2}}\lambda_n = O(1)$). Both the consistent moment selection procedures proposed in chapter 1 and the conservative moment selection methods presented in this chapter can not distinguish the relevant moment conditions from the irrelevant moment conditions. However, the key difference between the GMM Lasso and GMM adaptive Lasso estimations gives us the inspiration for designing a new adaptive penalty based on which, the GMM shrinkage estimation can consistently select the valid and relevant moment conditions from (2.2).

The new adaptive penalty depends on the measure of information contained in the moment conditions. We show that such information measure can be consistently estimated and its estimate is called as empirical information measure. The new

adaptive penalty is constructed as the product of a power function of the empirical information measure and the adaptive Lasso penalty. We show that the GMM shrinkage estimation based on the new adaptive penalty is consistent in selecting the valid and relevant moment conditions. As a result, the misspecified and irrelevant moment conditions are not selected with probability approaching 1 and the GMM shrinkage estimate is not only asymptotically efficient, but also robust against the irrelevant moment conditions in finite samples.

2.6 Appendix

2.6.1 Proof of the Main Result in Section 2.2

Proof of Proposition 2.2.2. Denote

$$\begin{aligned} V_{L,n}(u) &= \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\alpha_o+\frac{u}{\sqrt{n}})\right]' W_n \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\alpha_o+\frac{u}{\sqrt{n}})\right] \\ &\quad - \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\alpha_o)\right]' W_n \left[n^{-\frac{1}{2}}\sum_{i=1}^{n}\rho(Z_i,\alpha_o)\right] \\ &\quad + n\lambda_n \sum_{j=1}^{k}\left(|\beta_{j,o}+n^{-\frac{1}{2}}u_{\beta_j}| - |\beta_{j,o}|\right) \\ &\equiv V_n^*(u) + n\lambda_n \sum_{j=1}^{k}\left(|\beta_{j,o}+n^{-\frac{1}{2}}u_{\beta_j}| - |\beta_{j,o}|\right), \end{aligned} \tag{2.27}$$

then by definition, $u_{\theta,n}^* = \sqrt{n}(\widehat{\theta}_n - \theta_o)$ and $u_{\beta,n}^* = \sqrt{n}(\widehat{\beta}_n - \beta_o)$ constitute the minimizer $u_n^{*\prime} = (u_{\theta,n}^{*\prime}, u_{\beta,n}^{*\prime})$ of $V_{L,n}(u)$.

Using the similar arguments in the proof of Theorem 1.3.6 in Chapter 1, we can show that

$$V_n^*(u) \to_d V^*(u) \equiv u'Mu + 2u'\left[\frac{\partial m(\alpha_o)}{\partial \alpha}\right] W_o \Psi(\theta_o) \tag{2.28}$$

uniformly in $u \in K$, where K denotes any compact subset in $R^{d_\theta+k}$. If $\beta_{j,o} = 0$, then

$$n\lambda_n \left(|\beta_{j,o} + n^{-\frac{1}{2}} u_{\beta_j}| - |\beta_{j,o}| \right) \to \lambda_o \left| u_{\beta_j} \right| \tag{2.29}$$

uniformly in u_{β_j}. On the other hand, if $\beta_{j,o} \neq 0$, then

$$n\lambda_n \left(|\beta_{j,o} + n^{-\frac{1}{2}} u_{\beta_j}| - |\beta_{j,o}| \right) \to \lambda_o sgn(\beta_{j,o}) u_{\beta_j} \tag{2.30}$$

uniformly in u_{β_j}. From the results in (2.27)-(2.30), we can deduce that

$$V_{L,n}(u) \to_d V^*(u) + \lambda_o \left[\sum\nolimits_{j \in \mathcal{S}_\beta} sgn(\beta_{j,o}) u_{\beta_j} + \sum\nolimits_{j \in \mathcal{S}_\beta^c} |u_{\beta_j}| \right]. \tag{2.31}$$

From the root-n consistency of $(\widehat{\theta}_n, \widehat{\beta}_n)$, we know that $u^*_{\theta,n}$ and $u^*_{\beta,n}$ are asymptotically tight. The claimed result now follows by the argmax continuous mapping theorem (ACMT). ■

Proof of Proposition 2.2.3. First note that

$$\Pr\left(\widehat{\beta}_{j,n} = 0, \, \forall j \in \mathcal{S}_\beta^c \right) = \Pr\left(\sqrt{n}(\widehat{\beta}_{j,n} - \beta_{j,o}) = 0, \, \forall j \in \mathcal{S}_\beta^c \right). \tag{2.32}$$

Using the weak convergence result in Proposition 2.2.2,

$$\limsup_{n \to \infty} \Pr\left(\sqrt{n}(\widehat{\beta}_{j,n} - \beta_{j,o}) = 0, \, \forall j \in \mathcal{S}_\beta^c \right) \leq \Pr\left(u^*_{\beta_j} = 0, \, \forall j \in \mathcal{S}_\beta^c \right). \tag{2.33}$$

When $\lambda_o = 0$, $u^* = -M^{-1} \left[\frac{\partial m(\alpha_o)}{\partial \alpha} \right] W_o \Psi(\theta_o)$ is a continuous random vector and thus $\Pr\left(u^*_{\beta_j} = 0, \, \forall j \in \mathcal{S}_\beta^c \right) = 0$. We next consider the case that $\lambda_o > 0$. For the ease of the notation, we use $M_{\mathcal{SS}}$ to denote the leading $(d_\theta + d_{\beta_+}) \times (d_\theta + d_{\beta_+})$ submatrix of M, $M_{\mathcal{S}^c\mathcal{S}}$ to denote the left-lower $d_{\beta_-} \times (d_\theta + d_{\beta_+})$ submatrix of M, $u^*_{\mathcal{S}}$ and $u^*_{\mathcal{S}^c}$ to denote the first $d_\theta + d_{\beta_+}$ and last d_{β_-} element of u^*. Conditional on the event

$\{\widehat{\beta}_{j,n} = 0, j \in \mathcal{S}_\beta^c\}$, we have the following optimality condition for u^*

$$\begin{aligned} M_{\mathcal{S}\mathcal{S}} u^*_{\mathcal{S}} + \Psi_{m,\mathcal{S}}(\theta_o) &= -\frac{\lambda_o sgn(\beta_{\mathcal{S}})}{2} \text{ componentwise} \\ |M_{\mathcal{S}^c\mathcal{S}} u^*_{\mathcal{S}} + \Psi_{m,\mathcal{S}^c}(\theta_o)| &\leq \frac{\lambda_o}{2} \qquad \text{componentwise,} \end{aligned}$$

where $\Psi_{m,\mathcal{S}}(\theta_o) = \left[\frac{\partial m(\alpha_o)}{\partial \alpha} W_o\right](\mathcal{S})\Psi(\theta_o)$ and $\left[\frac{\partial m(\alpha_o)}{\partial \alpha} W_o\right](\mathcal{S})$ denotes the $\mathcal{S}$-th row of the matrix $\frac{\partial m(\alpha_o)}{\partial \alpha} W_o$. From the above two conditions, we can deduce that

$$\left| M_{\mathcal{S}^c\mathcal{S}} M_{\mathcal{S}\mathcal{S}}^{-1} \left[\Psi_{m,\mathcal{S}}(\theta_o) + \lambda_o \frac{sgn(\beta_{\mathcal{S}})}{2} \right] - \Psi_{m,\mathcal{S}^c}(\theta_o) \right| \leq \frac{\lambda_o}{2} \text{ componentwise.}$$

Hence there exists some constant $c(\lambda_o) \in (0,1)$ such that

$$0 < \Pr\left(\left| M_{\mathcal{S}^c\mathcal{S}} M_{\mathcal{S}\mathcal{S}}^{-1} \left[\Psi_{m,\mathcal{S}}(\theta_o) + \lambda_o \frac{sgn(\beta_{\mathcal{S}})}{2} \right] - \Psi_{m,\mathcal{S}^c}(\theta_o) \right| \leq \frac{\lambda_o}{2} \right) \leq c(\lambda_o)$$

which finishes the proof. ■

Proof of Proposition 2.2.4. Denote

$$V_{AL,n}(u) = V_n^*(u) + n\lambda_n \sum_{j=1}^{k} \frac{|\beta_{j,o} + n^{-\frac{1}{2}} u_{\beta_j}| - |\beta_{j,o}|}{|\widehat{\beta}_{j,1st}|^\omega}, \tag{2.34}$$

where $V_n^*(u)$ is defined in the proof of proposition 2.2.2, $u' = (u'_\theta, u'_\beta)$, u_θ and u_β are d_θ and d_β dimensional real vectors respectively. By definition, $u^*_{\theta,n} = \sqrt{n}(\widehat{\theta}_n - \theta_o)$ and $u^*_{\beta,n} = \sqrt{n}(\widehat{\beta}_n - \beta_o)$ constitute the minimizer of $V_{AL,n}(u)$.

If $\beta_{j,o} = 0$, then

$$n\lambda_n \frac{|\beta_{j,o} + n^{-\frac{1}{2}} u_{\beta_j}| - |\beta_{j,o}|}{|\widehat{\beta}_{j,1st}|^\omega} = \frac{n^{\frac{1+\omega}{2}} \lambda_n |u_{\beta_j}|}{|\sqrt{n}(\widehat{\beta}_{j,1st} - \beta_{j,o})|^\omega} \rightarrow_d \frac{\lambda_* \left| u_{\beta_j} \right|}{|\overline{M}_k(j)\Psi(\theta_o)|^\omega} \tag{2.35}$$

uniformly in u_{β_j}. On the other hand, if $\beta_{j,o} \neq 0$, then

$$n\lambda_n \frac{|\beta_{j,o} + n^{-\frac{1}{2}} u_{\beta_j}| - |\beta_{j,o}|}{|\widehat{\beta}_{j,1st}|^{\omega}} \to_p 0 \tag{2.36}$$

uniformly in u_{β_j}. In the proof of Proposition 2.2.2, we have shown that $V_n^*(u) \to_d V^*(u)$ uniformly in $u \in K$ for any compact subset $K \in R^{d_\theta + k}$. From the results in (2.34), (2.35) and (2.36), we can deduce that

$$V_{AL,n}(u) \to_d V^*(u) + \lambda_* \sum\nolimits_{j \in \mathcal{S}_\beta^c} \frac{|u_{\beta_j}|}{|\overline{M}_k(j)\Psi(\theta_o)|^{\omega}}.$$

Now the claimed result follows from the ACMT. ■

2.6.2 Proof of the Main Result in Section 2.3

Proof of Lemma 2.3.3. First note that under Assumption 1.3.1.(i)-(iii) and Assumption 1.3.2.(i)-(iii), it is tedious to show that $\widehat{\theta}_{1st,n} = \theta_o + O_p(n^{-1/2})$.

$$\begin{aligned}
& |Var_{n,q+j} - Var_{q+j}| \\
= & \left| \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial g_{q+j}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} - E\left[\frac{\partial g_{q+j}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] \right\} W_{q+j,n} \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_{q+j}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] \right| \\
& + \left| E\left[\frac{\partial g_{q+r}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] (W_{q+j,n} - W_{q+j,o}) \left[\frac{1}{n} \sum_{i=1}^{n} \frac{\partial g_q(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] \right| \\
& + \left| E\left[\frac{\partial g_{q+r}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] W_{q+j,o} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ \frac{\partial g_{q+j}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} - E\left[\frac{\partial g_{q+j}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] \right\} \right\} \right| \\
& + 2 \left| E\left[\frac{\partial^2 g_{q+r}(Z_i, \widetilde{\theta}_n)}{\partial \theta \partial \theta'} \right] W_{q+j,o} E\left[\frac{\partial g_{q+j}(Z_i, \widehat{\theta}_{1st})}{\partial \theta} \right] \right| (\widehat{\theta}_{1st,n} - \theta_o)
\end{aligned} \tag{2.37}$$

where $\widetilde{\theta}_n$ lies between $\widehat{\theta}_{1st}$ and θ_o. From the decomposition in (2.37), we can apply Assumption 2.3.1 and $\widehat{\theta}_{1st,n} = \theta_o + O_p(n^{-1/2})$ to deduce that $|Var_{n,q+j} - Var_{q+j}| =$

$O_p(n^{-1/2})$. As τ_j and $\tau_{n,j}$ are the largest eigenvalues of Var_{q+j} and $Var_{n,q+j}$ respectively, we can invoke the Bauer-Fiker Theorem on eigenvalue sensitivity to deduce that $\tau_{n,j} = \tau_j + O_p(n^{-1/2})$. ■

Proof of Theorem 2.3.6. For any compact subset K in $R^{d_\theta + d_{\mathcal{S}_{*,\beta}}}$, we denote any element $u_{\tau,\mathcal{S}} \in K$ as $u_{\tau,\mathcal{S}} = (u_\theta, u_{\beta_+}, u_{\beta_{\tau,-}})$, where u_θ contains the first d_θ elements in $u_{\tau,\mathcal{S}}$ and $(u_{\beta_+}, u_{\beta_{\tau,-}})$ contains the last $d_{\mathcal{S}_{*,\beta}}$ elements in $u_{\tau,\mathcal{S}}$. Denote

$$\begin{aligned} V_{2,n}(u_{\tau,\mathcal{S}}) &= \frac{1}{n}\left[\sum_{i=1}^{n} \rho^s(Z_i, \alpha_{o,\tau,\mathcal{S}} + \frac{u_{\tau,\mathcal{S}}}{\sqrt{n}})\right]' W_n \left[\sum_{i=1}^{n} \rho^s(Z_i, \alpha_{o,\tau,\mathcal{S}} + \frac{u_{\tau,\mathcal{S}}}{\sqrt{n}})\right] \\ &\quad - \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \alpha_o)\right]' W_n \left[n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(Z_i, \alpha_o)\right] \\ &\quad + n\lambda_n \sum_{j \in \mathcal{S}_{*,\beta}} \omega_{\tau,j,n} \left(\left|\beta_{o,j} + n^{-1/2} u_{\beta,j}\right| - \left|\beta_{o,j}\right|\right) \\ &\equiv V^*_{2,n}(u_{\tau,\mathcal{S}}) + n\lambda_n \sum_{j \in \mathcal{S}_{*,\beta}} \omega_{\tau,j,n} \left(\left|\beta_{o,j} + n^{-1/2} u_{\beta,j}\right| - \left|\beta_{o,j}\right|\right) \qquad (2.38) \end{aligned}$$

where $\rho^s(Z_i, \alpha_{o,\tau,\mathcal{S}} + \frac{u_{\tau,\mathcal{S}}}{\sqrt{n}}) = \rho(Z_i, \theta_o + \frac{u_\theta}{\sqrt{n}}, \beta_{o,+} + \frac{u_{\beta_+}}{\sqrt{n}}, \beta_{o,\tau,-} + \frac{u_{\beta_{\tau,-}}}{\sqrt{n}}, \beta_{o,-})$. From Theorem 2.3.6, we know that $\widehat{\beta}_{n,-} = 0$ w.p.a.1. Thus $\sqrt{n}\,(\widehat{\alpha}_{n,\tau,\mathcal{S}} - \alpha_{o,\tau,\mathcal{S}})$ is the minimizer of $V_{2,n}(u_{\tau,\mathcal{S}})$ w.p.a.1.

Using similar arguments to those in the proof of Theorem 1.3.5, we can show that

$$V^*_{2,n}(u_{\tau,\mathcal{S}}) = u'_{\tau,\mathcal{S}} M_{\tau,11} u_{\tau,\mathcal{S}} + 2u'_{\tau,\mathcal{S}} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\tau,\mathcal{S}}}\right] W_o \Psi(\alpha_o) + o_p(1). \qquad (2.39)$$

If $j \in \mathcal{S}_\beta$, then by (2.12)

$$n\lambda_n \omega_{\tau,j,n} \left(\left|\beta_{o,j} + n^{-1/2} u_{\beta,j}\right| - \left|\beta_{o,j}\right|\right) = \sqrt{n}\lambda_n \omega_{\tau,j,n} u_{\beta,j} + o(1) = o_p(1) \qquad (2.40)$$

uniformly in $u_{\beta,j}$. On the other hand, if $j \in \mathcal{S}_\beta \backslash \mathcal{S}_{*,\beta}$, then by (2.13)

$$n\lambda_n \omega_{\tau,j,n} \left(\left| \beta_{o,j} + n^{-1/2} u_{\beta,j} \right| - \left| \beta_{o,j} \right| \right) = \sqrt{n} \lambda_n \omega_{\tau,j,n} u_{\beta,j} = o_p(1) \tag{2.41}$$

uniformly in $u_{\beta,j}$.

Using the results in (2.38), (2.39), (2.40) and (2.41) and triangle inequality, we get

$$V_{2,n}(u_{\tau,\mathcal{S}}) \to_d V_2(u_{\tau,\mathcal{S}}) = u'_{\tau,\mathcal{S}} M_{\tau,11} u_{\tau,\mathcal{S}} + 2u'_{\tau,\mathcal{S}} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\tau,\mathcal{S}}} \right] W_o \Psi(\alpha_o) \tag{2.42}$$

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

$$u^*_{\mathcal{S}_{*,\beta}} = -M^{-1}_{\tau,11} \left[\frac{\partial m(\theta_o)}{\partial \alpha_{\tau,\mathcal{S}}} \right] W_o \Psi(\alpha_o) \tag{2.43}$$

By Corollary 2.3.4, there is

$$\sqrt{n} \left(\widehat{\alpha}_{n,\tau,\mathcal{S}} - \alpha_{o,\tau,\mathcal{S}} \right) = O_p(1) \tag{2.44}$$

Now, the asymptotic tightness of $\widehat{\alpha}_{n,\tau,\mathcal{S}}$ in (2.44), the uniform convergence in distribution in (2.42) and unique minimization in (2.43) enable us to invoke the ACMT to deduce that

$$\sqrt{n} \left(\widehat{\alpha}_{n,\tau,\mathcal{S}} - \alpha_{o,\tau,\mathcal{S}} \right) \to_d N(0, M^{-1}_{\tau,11}) \tag{2.45}$$

where we use the assumption that $W_o = E\left[\Psi(\alpha_o)\Psi'(\alpha_o)\right]$. This finishes the proof. ■

Chapter 3

Automated Estimation of Vector Error Correction Models (joint with Peter C. B. Phillips)

3.1 Introduction

Cointegrated system modeling is now one of the main workhorses in empirical time series research. Much of this empirical research makes use of vector error correction (VECM) formulations. While there is often some prior information concerning the number of cointegrating vectors, most practical work involves (at least confirmatory) pre-testing to determine the cointegrating rank of the system as well as the lag order in the autoregressive component that embodies the transient dynamics. These order selection decisions can be made by sequential likelihood ratio tests (e.g. Johansen, 1988, for rank determination) or the application of suitable information criteria (Phillips, 1996). The latter approach offers several advantages such as joint determination of the cointegrating rank and autoregressive order, consistent estima-

tion of both order parameters (Chao and Phillips, 1999), robustness to heterogeneity in the errors, and the convenience and generality of semi-parametric estimation in cases where the focus is simply the cointegrating rank (Cheng and Phillips, 2010). While appealing for practitioners, all of these methods are nonetheless subject to pre-test bias and post model selection inferential problems (Leeb and Pötscher, 2005).

The present chapter explores a different approach. The goal is to liberate the empirical researcher from sequential testing procedures in inference about cointegrated systems and in policy work that relies on impulse responses. The ideas originate in recent work on sparse system estimation using shrinkage techniques such as lasso and bridge regression. These procedures utilize penalized least squares criteria in regression that can succeed, at least asymptotically, in selecting the correct regressors in a linear regression framework while consistently estimating the non-zero regression coefficients. While apparently effective asymptotically these procedures do not avoid post model selection inference issues in finite samples because the estimators implicitly carry effects from the implementation of shrinkage which can result in bias, multimodal distributions and difficulty discriminating local alternatives that can lead to unbounded risk (Leeb and Pötscher, 2008). On the other hand, the methods do radically simplify empirical research with large dimensional systems where order parameters must be chosen and sparsity is expected.

One of the contributions of this chapter is to show how to develop adaptive versions of these shrinkage methods that apply in vector error correction modeling which by their nature involve reduced rank coefficient matrices and order parameters for lag polynomials and trend specifications. The implementation of these methods is not immediate. This is partly because of the nonlinearities involved in potential reduced rank structures and partly because of the interdependence of decision making concerning the form of the transient dynamics and the cointegrating rank structure.

This chapter designs a mechanism of estimation and selection that works through the eigenvalues of the levels coefficient matrix and the coefficient matrices of the transient dynamic components. The methods apply in quite general vector systems with unknown cointegrating rank structure and unknown lag dynamics. They permit simultaneous order estimation of the cointegrating rank and autoregressive order in conjunction with oracle-like efficient estimation of the cointegrating matrix and transient dynamics. As such they offer considerable advantages to the practitioner: in effect, it becomes unnecessary to implement pre-testing procedures because the empirical results reveal the order parameters as a consequence of the fitting procedure. In this sense, the methods provide an automated approach to the estimation of cointegrated systems. In the scalar case, the methods reduce to estimation in the presence or absence of a unit root and thereby implement an implicit unit root test procedure, as suggested in earlier work by Caner and Knight (2009).

This chapter is organized as follows. Section 3.2 lays out the model and assumptions and shows how to implement adaptive shrinkage methods in VECM systems. Section 3.3 considers a simplified first order version of the VECM without lagged differences which reveals the approach to cointegrating rank selection and develops key elements in the limit theory. Here we show that the cointegrating rank r_o is identified by the number of zero eigenvalues of Π_o and the latter is consistently recovered by suitably designed shrinkage estimation. Section 3.4 extends this system and its asymptotics to the general case of cointegrated systems with weakly dependent errors. Here it is demonstrated that the cointegration rank r_o can be consistently selected despite the fact that Π_o itself may not be consistently estimable. Section 3.5 deals with the practically important case of a general VECM system driven by independent identically distributed (iid) shocks, where shrinkage estimation simultaneously performs consistent lag selection, cointegrating rank selection, and optimal estima-

tion of the system coefficients. Section 3.6 considers adaptive selection of the tuning parameter and Section 3.7 reports some simulation findings. Section 3.8 concludes and outlines some useful extensions of the methods and limit theory to other models. Proofs and some supplementary technical results are given in the Appendix.

Notation is standard. For vector-valued, zero mean, covariance stationary stochastic processes $\{a_t\}_{t\geq 1}$ and $\{b_t\}_{t\geq 1}$, $\Sigma_{ab}(h) = E[a_t b'_{t+h}]$ and $\Gamma_{ab} = \sum_{h=0}^{\infty} \Sigma_{ab}(h)$ denote the lag h autocovariance matrix and one-sided long-run covariance matrix. Moreover, we use Σ_{ab} for $\Sigma_{ab}(0)$ and $\Sigma_{n,ab} = n^{-1} \sum_{t=1}^{n} a_t b'_t$ as the sample average of Σ_{ab} throughout the chapter. $\|\cdot\|$ denotes the Euclidean norm on any Euclidean space and $|A|$ is the determinant of square matrix A. A' refers to the transpose of any matrix A and $\|A\|_B \equiv ||A'BA||$ for any matrices A and B. I_k and $\mathbf{0}_{l_1 \times l_2}$ are used to denote $k \times k$ identity matrix and $l_1 \times l_2$ zero matrices respectively. $A \equiv B$ means that A is defined as B; the expression $a_n = o_p(b_n)$ signifies that $\Pr(|a_n/b_n| \geq \epsilon) \to 0$ for all $\epsilon > 0$ as n go to infinity; and $a_n = O_p(b_n)$ when $\Pr(|a_n/b_n| \geq M) \to 0$ as n and M go to infinity. As usual, "$\to_p$" and "$\to_d$" imply convergence in probability and convergence in distribution, respectively.

3.2 Vector Error Correction and Adaptive Shrinkage

Throughout this chapter we consider the following parametric VECM representation of a cointegrated system

$$\Delta Y_t = \Pi_o Y_{t-1} + \sum_{j=1}^{p} B_{o,j} \Delta Y_{t-j} + u_t, \tag{3.1}$$

where $\Delta Y_t = Y_t - Y_{t-1}$, Y_t is an m-dimensional vector-valued time series, $\Pi_o = \alpha_o \beta_o'$ has rank $0 \leq r_o \leq m$, $B_{o,j}$ $(j = 1, ..., p)$ are $m \times m$ (transient) coefficient matrices and u_t is an m-vector error term with mean zero and nonsingular covariance matrix Σ_{uu}. The rank r_o of Π_o is an order parameter measuring the cointegrating rank or the number of (long run) cointegrating relations in the system. The lag order p is a second order parameter, characterizing the transient dynamics in the system.

As $\Pi_o = \alpha_o \beta_o'$ has rank r_o, we can choose α_o and β_o to be $m \times r_o$ matrices with full rank. When $r_o = 0$, we simply take $\Pi_o = 0$. Let $\alpha_{o,\perp}$ and $\beta_{o,\perp}$ be the matrix orthogonal complements of α_o and β_o and, without loss of generality, assume that $\alpha_{o,\perp}' \alpha_{o,\perp} = I_{m-r_o}$ and $\beta_{o,\perp}' \beta_{o,\perp} = I_{m-r_o}$.

Suppose $\Pi_o \neq 0$ and define $Q = [\beta_o, \alpha_{o\perp}]'$. In view of the well known relation (e.g., Johansen, 1995)

$$\alpha_o (\beta_o' \alpha_o)^{-1} \beta_o' + \beta_{o,\perp} (\alpha_{o,\perp}' \beta_{o,\perp})^{-1} \alpha_{o,\perp}' = I_m,$$

it follows that $Q^{-1} = \left[\alpha_o (\beta_o' \alpha_o)^{-1}, \beta_{o,\perp} (\alpha_{o,\perp}' \beta_{o,\perp})^{-1}\right]$,

$$Q\Pi_o = \begin{bmatrix} \beta_o' \alpha_o \beta_o' \\ 0 \end{bmatrix} \text{ and } Q\Pi_o Q^{-1} = \begin{bmatrix} \beta_o' \alpha_o & 0 \\ 0 & 0 \end{bmatrix}. \tag{3.2}$$

Under Assumption RR in Section 3.3, $\beta_o' \alpha_o$ is an invertible matrix and hence the matrix $\beta_o' \alpha_o \beta_o'$ has full rank. Cointegrating rank is the number r_o of non-zero eigenvalues of Π_o or the nonzero row vector count of $Q\Pi_o$. When $\Pi_o = 0$, then the result holds trivially with $r_o = 0$ and $\beta_{o,\perp} = I_m$. The matrices $\alpha_{o\perp}$ and $\beta_{o,\perp}$ are composed of normalized left and right eigenvectors, respectively, corresponding to the zero eigenvalues in Π_o.

Conventional methods of estimation of (3.1) include reduced rank regression or

maximum likelihood based on the assumption of Gaussian u_t and a Gaussian likelihood. This approach relies on known r_o and known p, so implementation requires preliminary order parameter estimation. The system can also be estimated by unrestricted fully modified vector autoregression (Phillips, 1995), which leads to consistent estimation of the unit roots in (3.1), the cointegrating vectors and the transient dynamics. This method does not require knowledge of r_0 but does require knowledge of the lag order p. In addition, a semiparametric approach can be adopted in which r_o is estimated semiparametrically by order selection as in Cheng and Phillips (2010) followed by fully modified least squares regression to estimate the cointegrating matrix. This method achieves asymptotically efficient estimation of the long run relations (under Gaussianity) but does not estimate the transient relations.

The present chapter explores the estimation of the parameters of (3.1) by Lasso-type regression, i.e. least squares (LS) regression with penalization. The resulting estimator is a shrinkage estimator. Specifically, the LS shrinkage estimator of (Π_o, B_o) where $B_o = (B_{o,1}, ..., B_{o,p})$ is defined as

$$\begin{aligned}(\widehat{\Pi}_n, \widehat{B}_n) \quad = \quad & \underset{\Pi, B_1, ..., B_p \in R^{m\times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2 \right. \\ & \left. + \sum_{j=1}^{p} \frac{n\lambda_{b,j,n}}{||\widehat{B}_{j,1st}||^{\omega}} \|B_j\| + \sum_{k=1}^{m} \frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \|\Phi_{n,k}(\Pi)\| \right\}, \quad (3.3)\end{aligned}$$

where $\omega > 0$ is some constant, $\lambda_{b,j,n}$ and $\lambda_{r,k,n}$ $(j = 1, ..., p$ and $k = 1, ..., m)$ are tuning parameters that directly control the penalization, $||\phi_k(\Pi)||$ denotes the k-th largest modulus of the eigenvalues $\{\phi_k(\Pi)\}_{k=1}^{m}$ of the matrix Π [1], $\Phi_{n,k}(\Pi)$ is the k-th row vector of $Q_n\Pi$, and Q_n denotes the normalized left eigenvector matrix of

[1] Throughout this chapter, for any $m \times m$ matrix Π, we order the eigenvalues of Π in decreasing order by their modulus, i.e. $\|\phi_1(\Pi)\| \geq \|\phi_2(\Pi)\| \geq ... \geq \|\phi_m(\Pi)\|$. When there is a pair of complex conjugate eigenvalues, we order the one with a positive imaginary part before the other.

eigenvalues of $\widehat{\Pi}_{1st}$. The matrices $\widehat{B}_{j,1st}$ and $\widehat{\Pi}_{1st}$ are some first step (OLS) estimates of $B_{o,j}$ and Π_o $(j = 1, ..., p)$.

Let Λ_n denote the diagonal matrix with ordered eigenvalues of $\widehat{\Pi}_{1st}$. By definition $Q_n\widehat{\Pi}_{1st} = \Lambda_n Q_n$, which implies that the Euclidean norm of the k-th row of $Q_n\widehat{\Pi}_{1st}$ equals the norm of the k-th ordered eigenvalue of $\widehat{\Pi}_{1st}$. Hence the adaptive penalty $||\phi_k(\widehat{\Pi}_{1st})||^\omega$ in the penalized LS estimation is equivalent to $||\Phi_{n,k}(\widehat{\Pi}_{1st})||^\omega$ for $k = 1, ..., m$. Let $\mathcal{S}_{n,\phi} = \{k : \Phi_{n,k}(\widehat{\Pi}_n) \neq 0\}$ be the index set of nonzero rows in $Q_n\widehat{\Pi}_n$. Similarly, we index the zero rows in $Q_n\widehat{\Pi}_n$ using the (constrained) set $\mathcal{S}^c_{n,\phi} = \{k : \Phi_{n,k}(\widehat{\Pi}_n) = 0\}$. For any $k \in \mathcal{S}^c_{n,\phi}$, the k-th row of Q_n is the normalized left eigenvector of a zero eigenvalue of $\widehat{\Pi}_n$. Given λ_n, our procedure delivers a one step estimator of the model (3.1) with an implied estimate of the cointegrating rank (based on the number of non-zero rows of $Q_n\widehat{\Pi}_n$) and an implied estimate of the transient dynamic order p and transient dynamic structure (that is, the non zero elements of B_o) based on the fitted value $\widehat{B}_n$.

Let $\Phi'(\Pi_o) = [\Phi'_1(\Pi_o), ..., \Phi'_m(\Pi_o)]$ denote the row vectors of $Q\Pi_o$. When $\{u_t\}_{t\geq 1}$ is *iid* or a martingale difference sequence, the LS estimators $(\widehat{\Pi}_{1st}, \widehat{B}_{1st})$ of (Π_o, B_o) are well known to be consistent. The eigenvalues and corresponding eigenspace of Π_o can also be consistently estimated. Thus it seems intuitively clear that some form of adaptive penalization can be devised to consistently distinguish the zero and nonzero components in B_o and $\Phi(\Pi_o)$. We show that the shrinkage LS estimator defined in (3.3) enjoys these oracle-like properties, in the sense that the zero components in B_o and $\Phi(\Pi_o)$ are estimated as zeros w.p.a.1. Thus, Π_o and the non-zero elements in B_o are estimated as if the form of the true model were known and inferences can be conducted as if we knew the true cointegration rank r_o.

If the transient behavior of (3.1) is misspecified and (for some given lag order p) the error process $\{u_t\}_{t\geq 1}$ is weakly dependent and $r_o > 0$, then consistent estimators

of the full matrix (Π_o, B_o) are typically unavailable without further assumptions. However, the $m - r_o$ zero eigenvalues of Π_o can still be consistently estimated with an order n convergence rate, while the remaining eigenvalues of Π_o are estimated with asymptotic bias at a $\sqrt{n}$ convergence rate. The different convergence rates of the eigenvalues are important, because when the non-zero eigenvalues of Π_o are occasionally (asymptotically) estimated as zeros, the different convergence rates are useful in consistently distinguishing the zero eigenvalues from the biasedly estimated non-zero eigenvalues of Π_o. Specifically, we show that if the estimator of some non-zero eigenvalue of Π_o has probability limit zero under misspecification of the lag order, then this estimator will converge in probability to zero at the rate $\sqrt{n}$, while estimates of the zero eigenvalues of Π_o all have convergence rate n. Hence the adaptive penalties associated with estimates of zero eigenvalues of Π_o will diverge to infinity at a rate faster than those of estimates of the nonzero eigenvalues of Π_o, even though the latter also converge to zero in probability. As we have prior knowledge about these different divergence rates in a potentially cointegrated system, we can impose explicit conditions on the convergence rate of the tuning parameter to ensure that only $m - r_o$ zero eigenvalues are adaptively shrunk to zero in finite samples.

For the empirical implementation of our approach, we provide data-driven procedures for selecting the tuning parameter of the penalty function in finite samples. For practical purposes our method is executed in the following steps, which are explained and demonstrated in detail as the chapter progresses.

(1) After preliminary LS estimation of the system, perform a first step GLS shrinkage estimation with tuning parameters $\omega = 2$ and $\lambda_{r,k,n} = \lambda_{b,j,n} = \frac{2}{n}\log(n)$ for $j = 1, ..., p$ and $k = 1, ..., m$.

(2) Construct adaptive tuning parameters using the first step GLS shrinkage estimates and the formulas in (3.85) and (3.86).

(3) Using the adaptive tuning parameters, obtain the GLS shrinkage estimator $(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n})$ of (Π_o, B_o).

(4) The cointegration rank selected by the shrinkage method is implied by the rank of the shrinkage estimator $\widehat{\Pi}_{g,n}$ and the lagged differences selected by the shrinkage method are implied by the nonzero matrices in $\widehat{B}_{g,n}$.

(5) The GLS shrinkage estimator contains shrinkage bias introduced by the penalty on the nonzero eigenvalues of $\widehat{\Pi}_{g,n}$ and nonzero matrices in $\widehat{B}_{g,n}$. To remove this bias, run a reduced rank regression based on the cointegration rank and the model selected in the GLS shrinkage estimation in step (iv).

3.3 First Order VECM Estimation

This section considers the following simplified first order version of (3.1),

$$\Delta Y_t = \Pi_o Y_{t-1} + u_t = \alpha_o \beta_o' Y_{t-1} + u_t. \tag{3.4}$$

The model contains no deterministic trend and no lagged differences. Our focus in this simplified system is to outline the approach to cointegrating rank selection and develop key elements in the limit theory, showing consistency in rank selection and reduced rank coefficient matrix estimation. The theory is extended in subsequent sections to models of the form (3.1).

We start with the following condition on the innovation u_t.

Assumption 3.3.1 (WN) *$\{u_t\}_{t\geq 1}$ is an m-dimensional iid process with zero mean and nonsingular covariance matrix Ω_u.*

Assumption 3.3.1 ensures that the full parameter matrix Π_o is consistently estimable in this simplified system. The *iid* condition could, of course, be weakened to

martingale differences with no material changes in what follows. Under Assumption 3.3.1, partial sums of u_t satisfy the functional law

$$n^{-\frac{1}{2}} \sum_{t=1}^{[n\cdot]} u_t \to_d B_u(\cdot), \tag{3.5}$$

where $B_u(\cdot)$ is vector Brownian motion with variance matrix Ω_u.

Assumption 3.3.2 (RR) *(i) The determinantal equation $|I - (I + \Pi_o)\lambda| = 0$ has roots on or outside the unit circle; (ii) the matrix Π_o has rank r_o, with $0 \leq r_o \leq m$; (iii) if $r_o > 0$, then the matrix $R = I_{r_o} + \beta_o'\alpha_o$ has eigenvalues within the unit circle.*

Let $\mathcal{S}_\phi = \{k : \Phi_k(\Pi_o) \neq 0\}$ be the index set of nonzero rows of $Q\Pi_o$ and similarly $\mathcal{S}_\phi^c = \{k : \Phi_k(\Pi_o) = 0\}$ denote the index set of zero rows of $Q\Pi_o$. By the property of Q, we know that $S_\phi = \{1, ..., r_o\}$ and $S_\phi^c = \{r_o + 1, ..., m\}$. It follows that consistent selection of the rank of Π_o is equivalent to the consistent recovery of the zero rows in $\Phi(\Pi_o) = Q\Pi_o$.

Using the matrix Q, (3.4) transforms as

$$\Delta Z_t = \Xi_o Z_{t-1} + w_t, \tag{3.6}$$

where

$$Z_t = \begin{pmatrix} \beta_o' Y_t \\ \alpha_{o,\perp}' Y_t \end{pmatrix} \equiv \begin{pmatrix} Z_{1,t} \\ Z_{2,t} \end{pmatrix}, \; w_t = \begin{pmatrix} \beta_o' u_t \\ \alpha_{o,\perp}' u_t \end{pmatrix} \equiv \begin{pmatrix} w_{1,t} \\ w_{2,t} \end{pmatrix}$$

and $\Xi_o = Q\Pi_o Q^{-1}$**.** Assumption 3.3.2 leads to the following Wold representation for $Z_{1,t}$

$$Z_{1,t} = \beta_o' Y_t = \sum_{i=0}^{\infty} R^i \beta_o' u_{t-i} = R(L)\beta_o' u_t, \tag{3.7}$$

and the partial sum Granger representation,

$$Y_t = C\sum_{s=1}^{t} u_s + \alpha_o(\beta_o'\alpha_o)^{-1}R(L)\beta_o'u_t + CY_0, \tag{3.8}$$

where $C = \beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\alpha_{o,\perp}'$. Under Assumption 3.3.2 and (3.5), we have the functional law

$$n^{-\frac{1}{2}}\sum_{t=1}^{[n\cdot]} w_t \to_d B_w(\cdot) = QB_u(\cdot) = \begin{bmatrix} \beta_o'B_u(\cdot) \\ \alpha_{o,\perp}'B_u(\cdot) \end{bmatrix} \equiv \begin{bmatrix} B_{w_1}(\cdot) \\ B_{w_2}(\cdot) \end{bmatrix}$$

for $w_t = Qu_t$, so that

$$n^{-\frac{1}{2}}\sum_{t=1}^{[n\cdot]} Z_{1,t} = n^{-\frac{1}{2}}\sum_{t=1}^{[n\cdot]} \beta_o'Y_t \to_d -(\beta_o'\alpha_o)^{-1}B_{w_1}(\cdot), \tag{3.9}$$

since $R(1) = \sum_{i=0}^{\infty} R^i = (I-R)^{-1} = -(\beta_o'\alpha_o)^{-1}$. Also

$$n^{-1}\sum_{t=1}^{n} Z_{1,t-1}Z_{1,t-1}' = n^{-1}\sum_{t=1}^{n} \beta_o'Y_{t-1}Y_{t-1}'\beta_o \to_p \Sigma_{z_1z_1},$$

where $\Sigma_{z_1z_1} \equiv Var[\beta_o'Y_t] = \sum_{i=0}^{\infty} R^i\beta_o'\Omega_u\beta_oR^{i\prime}$.

The shrinkage LS estimator $\widehat{\Pi}_n$ of Π_o is defined as

$$\widehat{\Pi}_n = \underset{\Pi\in R^{m\times m}}{\arg\min}\sum_{t=1}^{n}\|\Delta Y_t - \Pi Y_{t-1}\|^2 + n\sum\nolimits_{k=1}^{m}\lambda_{r,k,n}\|\Phi_{n,k}(\Pi)\|/||\phi_k(\widehat{\Pi}_{1st})||^{\omega}. \tag{3.10}$$

The unrestricted LS estimator $\widehat{\Pi}_{1st}$ of Π_o is

$$\widehat{\Pi}_{1st} = \underset{\Pi\in R^{m\times m}}{\arg\min}\sum_{t=1}^{n}\|\Delta Y_t - \Pi Y_{t-1}\|^2 = \left(\sum_{t=1}^{n}\Delta Y_tY_{t-1}'\right)\left(\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\right)^{-1}. \tag{3.11}$$

The asymptotic properties of $\widehat{\Pi}_{1st}$ and its eigenvalues are described in the following

result.

Lemma 3.3.1 *Under Assumptions 3.3.1 and 3.3.2, we have:*

(a) define $D_n = diag(n^{-\frac{1}{2}} I_{r_o}, n^{-1} I_{m-r_o})$, *then* $\widehat{\Pi}_{1st}$ *satisfies*

$$\left(\widehat{\Pi}_{1st} - \Pi_o\right) Q^{-1} D_n^{-1} \to_d (B_{m,1}, B_{m,2}) \tag{3.12}$$

where $B_{m,1} \equiv N(0, \Omega_u \otimes \Sigma_{z_1 z_1}^{-1})$ *and* $B_{m,2} \equiv \int dB_u B'_{w_2} (\int B_{w_2} B'_{w_2})^{-1}$;

(b) the eigenvalues of $\widehat{\Pi}_{1st}$ *satisfy* $\phi_k(\widehat{\Pi}_{1st}) \to_p \phi_k(\Pi_o)$ *for* $k = 1, ..., m$;

(c) the last $m - r_o$ *eigenvalues of* $\widehat{\Pi}_{1st}$ *satisfy*

$$n\left(\phi_1(\widehat{\Pi}_{1st}), ..., \phi_{m-r_o}(\widehat{\Pi}_{1st})\right) \to_d \left(\widetilde{\phi}_{o,1}, ..., \widetilde{\phi}_{o,m-r_o}\right), \tag{3.13}$$

where the $\widetilde{\phi}_{o,j}$ $(j = 1, ..., m - r_o)$ *are solutions of the following determinantal equation*

$$\left|\mu I_{m-r_0} - \left(\int dB_{w_2} B'_{w_2}\right)\left(\int B_{w_2} B'_{w_2}\right)^{-1}\right| = 0. \tag{3.14}$$

The results of Lemma 3.3.1 are useful because the OLS estimate $\widehat{\Pi}_{1st}$ and the related eigenvalue estimates can be used as first step estimates in the penalty function. The convergence rates of $\widehat{\Pi}_{1st}$ and $\phi_k(\widehat{\Pi}_{1st})$ are important for delivering consistent model selection and cointegrated rank selection.

Let P_n be the inverse of Q_n. We divide P_n and Q_n as $P_n = [P_{\alpha,n}, P_{\alpha_\perp,n}]$ and $Q'_n = \left[Q'_{\alpha,n}, Q'_{\alpha_\perp,n}\right]$, where $Q_{\alpha,n}$ and $P_{\alpha,n}$ are the first r_o rows of Q_n and first r_o columns of P_n respectively ($Q_{\alpha_\perp,n}$ and $P_{\alpha_\perp,n}$ are defined accordingly). By definition,

$$Q_{\alpha_\perp,n} P_{\alpha_\perp,n} = I_{m-r_o},\ Q_{\alpha,n} P_{\alpha_\perp,n} = \mathbf{0}_{r_o \times (m-r_o)} \text{ and } Q_{\alpha_\perp,n} \widehat{\Pi}_{1st} = \Lambda_{\alpha_\perp,n} Q_{\alpha_\perp,n} \tag{3.15}$$

where $\Lambda_{\alpha_\perp,n}$ is an diagonal matrix with the ordered last (smallest) $m - r_o$ eigenvalues

of $\widehat{\Pi}_{1st}$. Using the results in (3.15), we can define a useful estimator of Π_o as

$$\Pi_{n,f} = \widehat{\Pi}_{1st} - P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n}. \tag{3.16}$$

By definition

$$Q_{\alpha,n}\Pi_{n,f} = Q_{\alpha,n}\widehat{\Pi}_{1st} - Q_{\alpha,n}P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n} = \Lambda_{\alpha,n}Q_{\alpha,n} \tag{3.17}$$

where $\Lambda_{\alpha,n}$ is an diagonal matrix with the ordered first (largest) r_o eigenvalues of $\widehat{\Pi}_{1st}$, and more importantly

$$Q_{\alpha_\perp,n}\Pi_{n,f} = Q_{\alpha_\perp,n}\widehat{\Pi}_{1st} - Q_{\alpha_\perp,n}P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n} = \mathbf{0}_{(m-r_o)\times m}. \tag{3.18}$$

From Lemma 3.3.1.(b), (3.17) and (3.18), we can deduce that $Q_{\alpha,n}\Pi_{n,f}$ is a $r_o \times m$ matrix which is nonzero w.p.a.1 and $Q_{\alpha_\perp,n}\Pi_{n,f}$ is always a $(m-r_o)\times m$ zero matrix for all n. Moreover

$$\Pi_{n,f} - \Pi_o = (\widehat{\Pi}_{1st} - \Pi_o) - P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n}$$

and so under Lemma 3.3.1.(a) and (c),

$$(\Pi_{n,f} - \Pi_o)\, Q^{-1}D_n^{-1} = O_p(1). \tag{3.19}$$

Thus the estimator $\Pi_{n,f}$ is at least as good as the OLS estimator $\widehat{\Pi}_{1st}$ in terms of its rate of convergence. Using (3.19) we can compare the LS shrinkage estimator $\widehat{\Pi}_n$ with $\Pi_{n,f}$ to establish the consistency and convergence rate of $\widehat{\Pi}_n$.

Theorem 3.3.2 (Consistency) *Suppose Assumptions WN and RR are satisfied.*

If $\lambda_{r,k,n} = o(1)$ *for all* $k \in \mathcal{S}_\phi$*, then the shrinkage LS estimator* $\widehat{\Pi}_n$ *is consistent, i.e.* $\widehat{\Pi}_n - \Pi_o = o_p(1)$.

When consistent shrinkage estimators are considered, Theorem 3.3.2 extends Theorem 1 of Caner and Knight (2009) who used shrinkage techniques to perform a unit root test. As the eigenvalues $\phi_k(\Pi)$ of the matrix Π are continuous functions of Π, we deduce from the consistency of $\widehat{\Pi}_n$ and continuous mapping that $\phi_k(\widehat{\Pi}_n) \to_p \phi_k(\Pi_o)$ for all $k = 1, ..., m$. Theorem 3.3.2 implies that the nonzero eigenvalues of Π_o are estimated as non-zeros, which means that the rank of Π_o will not be under-selected. However, consistency of the estimates of the non-zero eigenvalues is not necessary for consistent cointegration rank selection. In that case what is essential is that the probability limits of the estimates of those (non-zero) eigenvalues are not zeros or at least that their convergence rates are slower than those of estimates of the zero eigenvalues. This point will be pursued in the following section where it is demonstrated that consistent estimation of the cointegrating rank continues to hold for weakly dependent innovations $\{u_t\}_{t\geq 1}$ even though full consistency of $\widehat{\Pi}_n$ does not generally apply in that case.

Our next result gives the convergence rate of the shrinkage estimator $\widehat{\Pi}_n$.

Theorem 3.3.3 (Rate of Convergence) *Let* $\delta_{r,n} = \max_{k\in\mathcal{S}_\phi} \left\{ \lambda_{r,k,n} ||\phi_k(\widehat{\Pi}_{1st})||^{-\omega} \right\}$. *Under Assumption WN, RR and* $\lambda_{r,k,n} = o(1)$ *for all* $k \in \mathcal{S}_\phi$*, the shrinkage LS estimator* $\widehat{\Pi}_n$ *satisfies the following:*

(a) if $r_o = 0$*, then* $\widehat{\Pi}_n - \Pi_o = O_p(n^{-1} + n^{-1}\delta_{r,n})$*;*

(b) if $0 < r_o \leq m$*, then* $\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} = O_p(1 + n^{\frac{1}{2}}\delta_{r,n})$.

The term $\delta_{r,n}$ represents the shrinkage bias that the penalty function introduces to the LS shrinkage estimator. If the convergence rate of $\lambda_{r,k,n}$ $(k \in \mathcal{S}_\phi)$ is fast enough such that $n^{\frac{1}{2}}\delta_{r,n} = O_p(1)$, then Theorem 3.3.3 implies that $\widehat{\Pi}_n - \Pi_o = O_p(n^{-1})$ when

$r_o = 0$ and $\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} = O_p(1)$ otherwise. Hence, under Assumption WN, RR and $n^{\frac{1}{2}}\delta_{r,n} = O_p(1)$, the LS shrinkage estimator $\widehat{\Pi}_n$ has the same stochastic properties of the LS estimator $\widehat{\Pi}_{1st}$. However, we next show that if the tuning parameter $\lambda_{r,k,n}$ $(k \in \mathcal{S}_\phi^c)$ does not converge to zero too fast, then the correct rank restriction $r = r_o$ is automatically imposed on the LS shrinkage estimator $\widehat{\Pi}_n$ w.p.a.1.

Recall that $S_{n,\phi}$ is the index set of the nonzero rows of $Q_n\widehat{\Pi}_n$ and $S_{n,\phi}^c$ is the index set of the zero rows of $Q_n\widehat{\Pi}_n$. Under Lemma 3.3.1 and Theorem 3.3.2

$$Q_{\alpha,n}\widehat{\Pi}_n = Q_{\alpha,n}\widehat{\Pi}_{1st} + o_p(1) \tag{3.20}$$

which means that the first r_o rows of $Q_n\widehat{\Pi}_n$ are nonzero w.p.a.1. On the other hand, Lemma 3.3.1 and Theorem 3.3.2 imply that

$$Q_{\alpha_\perp,n}\widehat{\Pi}_n = Q_{\alpha_\perp,n}\widehat{\Pi}_{1st} + o_p(1) = \Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n} + o_p(1) = o_p(1) \tag{3.21}$$

which means that the last $m - r_o$ rows of $Q_n\widehat{\Pi}_n$ are arbitrarily close to zero with w.p.a.1. Under (3.20) we deduce that $S_\phi \subseteq S_{n,\phi}$. However, (3.21) is insufficient for showing that $S_\phi^c \subseteq S_{n,\phi}^c$, because in that case, what we need to show is $Q_{\alpha_\perp,n}\widehat{\Pi}_n = 0$ w.p.a.1.

Theorem 3.3.4 (Super-efficiency) *Suppose that the conditions of Theorem 3.3.3 are satisfied. If $n^{\frac{1}{2}}\delta_{r,n} = O_p(1)$ and $n^\omega\lambda_{r,k,n} \to \infty$ for $k \in \mathcal{S}_\phi^c$, then*

$$\Pr\left(Q_{\alpha_\perp,n}\widehat{\Pi}_n = 0\right) \to 1 \text{ as } n \to \infty. \tag{3.22}$$

Combining Theorem 3.3.2 and Theorem 3.3.4, we deduce that

$$\Pr\left(\mathcal{S}_{n,\phi} = \mathcal{S}_\phi\right) \to 1, \tag{3.23}$$

which implies consistent cointegration rank selection, giving the following result.

Corollary 3.3.5 *Under the conditions of Theorem 3.3.4, we have*

$$\Pr\left(r(\widehat{\Pi}_n) = r_o\right) \to 1 \tag{3.24}$$

as $n \to \infty$, *where* $r(\widehat{\Pi}_n)$ *denotes the rank of* $\widehat{\Pi}_n$.

From Corollary 3.3.5, we can deduce that the rank constraint $r(\Pi) = r_o$ is imposed on the LS shrinkage estimator $\widehat{\Pi}_n$ w.p.a.1. As $\widehat{\Pi}_n$ satisfies the rank constraint w.p.a.1, we expect it has better properties in comparison to the OLS estimator $\widehat{\Pi}_{1st}$ which assumes the true rank is unknown. This conjecture is confirmed in the following theorem.

Theorem 3.3.6 (Limiting Distribution) *Under the conditions of Theorem 3.3.4 and* $n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$, *we have*

$$\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} B_{m,1} & \alpha_o(\alpha_o'\alpha_o)^{-1}\alpha_o' B_{m,2} \end{pmatrix} \tag{3.25}$$

where $B_{m,1}$ *and* $B_{m,2}$ *are defined in Lemma 3.3.1.(a).*

From (3.25) and the continuous mapping theorem (CMT),

$$Q\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} \beta_o' B_{m,1} & \beta_o'\alpha_o(\alpha_o'\alpha_o)^{-1}\alpha_o' B_{m,2} \\ \alpha_{o,\perp}' B_{m,1} & 0 \end{pmatrix}. \tag{3.26}$$

Similarly, from Lemma 3.3.1.(a) and CMT

$$Q\left(\widehat{\Pi}_{1st} - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} \beta_o' B_{m,1} & \beta_o' B_{m,2} \\ \alpha_{o,\perp}' B_{m,1} & \alpha_{o,\perp}' B_{m,2} \end{pmatrix}. \tag{3.27}$$

Compared with the OLS estimator, we see that in the LS shrinkage estimation, the right lower $(m - r_o) \times (m - r_o)$ submatrix of $Q\Pi_o Q^{-1}$ is estimated at a faster rate than n. The improved property of the LS shrinkage estimator $\widehat{\Pi}_n$ arises from the fact that the correct rank restriction $r(\widehat{\Pi}_n) = r_o$ is satisfied w.p.a.1, leading to the lower right zero block in the limit distribution (3.25) after normalization.

Compared with the oracle reduced rank regression (RRR) estimator (i.e. the RRR estimator informed by knowledge of the true rank, see e.g. Phillips, 1998 and Anderson, 2002), the LS shrinkage estimator suffers from second order bias in the limit distribution (3.25), which is evident in the endogeneity bias of the factor $\int dB_u B'_{w_2}$ in the limit matrix $B_{m,2}$. Accordingly, to remove the endogeneity bias we introduce the generalized least square (GLS) shrinkage estimator $\widehat{\Pi}_{g,n}$ which satisfies the weighted extremum problem

$$\widehat{\Pi}_{g,n} = \underset{\Pi \in R^{m \times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - \Pi Y_{t-1}\|^2_{\widehat{\Omega}^{-1}_{u,n}} + n \sum_{k=1}^{m} \frac{\lambda_{r,k,n} ||\Phi_{n,k}(\Pi)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}, \tag{3.28}$$

where $\widehat{\Omega}_{u,n}$ is some consistent estimator of Ω_u. The asymptotic distribution of $\widehat{\Pi}_{g,n}$ is the same as that of the oracle RRR estimator.

Corollary 3.3.7 (Oracle Properties) *Suppose Assumption 3.3.1 and 3.3.2 hold. If $\widehat{\Omega}_{u,n} \to_p \Omega_u$ and the tuning parameter satisfies $n^{\frac{1}{2}} \lambda_{r,k,n} = o(1)$ and $n^{\omega} \lambda_{r,k,n} \to \infty$ for $k = 1, ..., m$, then as $n \to \infty$,*

$$\Pr\left(r(\widehat{\Pi}_{g,n}) = r_o\right) \to 1 \tag{3.29}$$

and $\widehat{\Pi}_{g,n}$ *has limit distribution*

$$\left(\widehat{\Pi}_{g,n} - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \left(\begin{array}{cc} B_{m,1} & \alpha_o(\beta_o'\alpha_o)^{-1} \int dB_{u\cdot w_2} B_{w_2}' (\int B_{w_2} B_{w_2}')^{-1} \end{array} \right), \tag{3.30}$$

where $B_{u\cdot w_2} \equiv B_u - \Sigma_{uw_2}\Sigma_{w_2w_2}^{-1} B_{w_2}$.

From (3.30), we can invoke CMT to deduce that

$$Q\left(\widehat{\Pi}_{g,n} - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \left(\begin{array}{cc} \beta_o' B_{m,1} & \int dB_{u\cdot w_2} B_{w_2}' (\int B_{w_2} B_{w_2}')^{-1} \\ \alpha_{o,\perp}' B_{m,1} & 0 \end{array} \right) \tag{3.31}$$

which implies that the GLS shrinkage estimate $\widehat{\Pi}_{g,n}$ has the same limiting distribution as that of the oracle RRR estimator.

Remark 3.3.8 *In the triangular representation of a cointegration system studied in Phillips (1991), we have* $\alpha_o = [I_{r_o}, 0_{r_o\times(m-r_o)}]'$, $\beta_o = [-I_{r_o}, O_o]'$ *and* $w_2 = u_2$. *Moreover, we obtain*

$$\Pi_o = \left(\begin{array}{cc} -I_{r_o} & O_o \\ 0 & \mathbf{0}_{m-r_o} \end{array} \right), Q = \left(\begin{array}{cc} -I_{r_o} & O_o \\ 0 & I_{m-r_o} \end{array} \right) \text{ and } Q^{-1} = \left(\begin{array}{cc} -I_{r_o} & O_o \\ 0 & I_{m-r_o} \end{array} \right).$$

By the consistent rank selection, the GLS shrinkage estimator $\widehat{\Pi}_{g,n}$ *can be decomposed as* $\widehat{\alpha}_{g,n}\widehat{\beta}_{g,n}'$ *w.p.a.1, where* $\widehat{\alpha}_{g,n} \equiv [\widehat{A}_{g,n}', \widehat{B}_{g,n}']'$ *is the first* r_o *columns of* $\widehat{\Pi}_{g,n}$ *and* $\widehat{\beta}_{g,n} = [-I_{r_o}, \widehat{O}_{g,n}]'$. *From Corollary 3.3.7, we deduce that*

$$\sqrt{n}\left(\widehat{A}_{g,n} - I_{r_o}\right) \to_d N(0, \Omega_{u_1} \otimes \Sigma_{z_1z_1}^{-1}) \tag{3.32}$$

and

$$n\widehat{A}_{g,n}\left(\widehat{O}_{g,n} - O_o\right) \to_d \int dB_{u_{1\cdot 2}} B_{u_2}' \left(\int B_{u_2} B_{u_2}'\right)^{-1} \tag{3.33}$$

where B_{u_1} and B_{u_2} denotes the first r_o and last $m - r_o$ vectors of B_u, and $B_{u_{1\cdot 2}} = B_{u_1} - \Omega_{u,12}\Omega_{u,22}^{-1}B_{u_2}$. Under (3.32), (3.33) and CMT, we deduce that

$$n\left(\widehat{O}_{g,n} - O_o\right) \to_d \int dB_{u_{1\cdot 2}} B'_{u_2} \left(\int B_{u_2} B'_{u_2}\right)^{-1}. \tag{3.34}$$

From the results in (3.34), we can see that the GLS estimator $\widehat{O}_{g,n}$ of the cointegration matrix O_o is asymptotically equivalent to the maximum likelihood estimator studied in Phillips (1991) and has the usual mixed normal limit distribution, facilitating inference.

3.4 Extension I: VECM Estimation with Weakly Dependent Innovations

In this section, we study shrinkage reduced rank estimation in a scenario where the equation innovations $\{u_t\}_{t\geq 1}$ are weakly dependent. Specifically, we assume that $\{u_t\}_{t\geq 1}$ is generated by a linear process satisfying the following condition.

Assumption 3.4.1 (LP) *Let $D(L) = \sum_{j=0}^{\infty} D_j L^j$, where $D_0 = I_m$ and $D(1)$ has full rank. Let u_t have the Wold representation*

$$u_t = D(L)\varepsilon_t = \sum_{j=0}^{\infty} D_j \varepsilon_{t-j}, \text{ with } \sum_{j=0}^{\infty} j^{\frac{1}{2}} ||D_j|| < \infty, \tag{3.35}$$

where ε_t is iid $(0, \Sigma_{\varepsilon\varepsilon})$ with $\Sigma_{\varepsilon\varepsilon}$ positive definite and finite fourth moments.

Denote the long-run variance of $\{u_t\}_{t\geq 1}$ as $\Omega_u = \sum_{h=-\infty}^{\infty} \Sigma_{uu}(h)$. From the Wold representation in (3.35), we have $\Omega_u = D(1)\Sigma_{\varepsilon\varepsilon}D(1)'$, which is positive definite because $D(1)$ has full rank and $\Sigma_{\varepsilon\varepsilon}$ is positive definite. The fourth moment assumption

is needed for the limit distribution of sample autocovariances in the case of misspecified transient dynamics.

The following lemma is useful in establishing the asymptotic properties of the shrinkage estimator with weakly dependent innovations.

Lemma 3.4.1 *Under Assumption 3.3.2 and 3.4.1, (a)-(c) and (e) of Lemma 3.9.1 are unchanged, while Lemma 3.9.1.(d) becomes*

$$n^{-\frac{1}{2}} \sum_{t=1}^{n} \left[u_t Z'_{1,t-1} - \Sigma_{uz_1}(1) \right] \to_d N(0, V_{uz_1}), \tag{3.36}$$

where $\Sigma_{uz_1}(1) = \sum_{j=0}^{\infty} \Sigma_{uu}(j) \beta_o \left(R^j\right)' < \infty$ *and* V_{uz_1} *is the long run variance matrix of* $u_t \otimes Z_{1,t-1}$.

As expected, under general weak dependence assumptions on u_t, the simple reduced rank regression models (3.1) and (3.4) are susceptible to the effects of potential misspecification in the transient dynamics. These effects bear on the stationary components in the system. In particular, due to the centering term $\Sigma_{uz_1}(1)$ in (3.36), both the OLS estimator $\widehat{\Pi}_{1st}$ and the shrinkage estimator $\widehat{\Pi}_n$ are asymptotically biased. Specifically, we show that $\widehat{\Pi}_{1st}$ has the following probability limit,

$$\widehat{\Pi}_{1st} \to_p \Pi_1 \equiv Q^{-1} H_o Q + \Pi_o, \tag{3.37}$$

where $H_o = Q \left[\Sigma_{uz_1}(1) \Sigma_{z_1 z_1}^{-1}, 0_{m \times (m - r_o)} \right]$. Note that

$$Q^{-1} H_o Q + \Pi_o = \left[\alpha_o + \Sigma_{uz_1}(1) \Sigma_{z_1 z_1}^{-1} \right] \beta'_o = \widetilde{\alpha}_o \beta'_o, \tag{3.38}$$

which implies that the asymptotic bias of the OLS estimator $\widehat{\Pi}_{1st}$ is introduced via the bias in the pseudo true value limit $\widetilde{\alpha}_o$. Observe also that $\Pi_1 = \widetilde{\alpha}_o \beta'_o$ has rank at

most equal to r_o, the number of rows in β_o'. Denote

$$\begin{aligned}\widehat{S}_{12} &= \sum_{t=1}^{n} \frac{Z_{1,t-1}Z_{2,t-1}'}{n},\ S_{21} = \sum_{t=1}^{n} \frac{Z_{2,t-1}Z_{1,t-1}'}{n},\\ \widehat{S}_{11} &= \sum_{t=1}^{n} \frac{Z_{1,t-1}Z_{1,t-1}'}{n} \text{ and } \widehat{S}_{22} = \sum_{t=1}^{n} \frac{Z_{2,t-1}Z_{2,t-1}'}{n}.\end{aligned}$$

The next Lemma presents some asymptotic properties of the bias in $\widehat{\Pi}_{1st}$.

Lemma 3.4.2 *Let* $H_n = nQ\left[\Sigma_{uz_1}(1), \mathbf{0}_{m\times(m-r_o)}\right]\left(\sum_{t=1}^{n} Z_{t-1}Z_{t-1}'\right)^{-1}$ *and* $\Pi_{1,n} = Q^{-1}H_nQ + \Pi_o$. *Then*

(a) H_n *converges in probability to* H_o, *i.e.* $H_n \to_p H_o$;

(b) $nQ^{-1}H_nQ\beta_{o\perp}$ *has limit distribution* $\widetilde{\Pi}_1\beta_{o\perp}\alpha_{o\perp}'$, *where*

$$\widetilde{\Pi}_1 = \Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}(\beta_o'\alpha_o)^{-1}\left(\int dB_{w_1}B_{w_2}' + \Sigma_{w_1w_2}\right)\left(\int B_{w_2}B_{w_2}'\right)^{-1}; \tag{3.39}$$

(c) $\sqrt{n}Q^{-1}\left(H_n - H_o\right)Q\beta_o$ *has the limit distribution* $\widetilde{\Pi}_2\beta_o$, *where*

$$\widetilde{\Pi}_2 = \Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}N(0, V_{z_1z_1})\Sigma_{z_1z_1}^{-1}\beta_o' \tag{3.40}$$

and $N(0, V_{z_1z_1})$ *denotes the matrix limit distribution of* $\sqrt{n}\left(\widehat{S}_{11} - \Sigma_{z_1z_1}\right)$.

Denote the rank of Π_1 by r_1. Then, by virtue of the expression $\Pi_1 = \widetilde{\alpha}_o\beta_o'$, we have $r_1 \leq r_o$ as indicated. Without loss of generality, we decompose Π_1 as $\Pi_1 = \widetilde{\alpha}_1\widetilde{\beta}_1'$ where $\widetilde{\alpha}_1$ and $\widetilde{\beta}_1$ are $m\times r_1$ matrixes with full rank. Denote the orthogonal complements of $\widetilde{\alpha}_1$ and $\widetilde{\beta}_1$ as $\widetilde{\alpha}_{1\perp}$ and $\widetilde{\beta}_{1\perp}$ respectively. Similarly, we decompose $\widetilde{\beta}_{1\perp}$ as $\widetilde{\beta}_{1\perp} = (\widetilde{\beta}_\perp, \beta_{o\perp})$ where $\widetilde{\beta}_\perp$ is an $m \times (r_o - r_1)$ matrix. By the definition of Π_1, we know that $\beta_{o,\perp}$ is the right eigenvectors of the zero eigenvalues of Π_1. Thus, $\widetilde{\beta}_1$ lies in some subspace of the space spanned by β_o.

Let $[\phi_1(\widehat{\Pi}_{1st}), ..., \phi_m(\widehat{\Pi}_{1st})]$ and $[\phi_1(\Pi_1), ..., \phi_m(\Pi_1)]$ be the ordered eigenvalues of $\widehat{\Pi}_{1st}$ and Π_1 respectively. The next lemma provides asymptotic properties of the OLS estimate and its eigenvalues when the data is weakly dependent.

Lemma 3.4.3 *Under Assumption 3.3.2 and 3.4.1, we have the following results:*

(a) the OLS estimator $\widehat{\Pi}_{1st}$ *satisfies*

$$\left[Q\left(\widehat{\Pi}_{1st} - \Pi_o\right)Q^{-1} - H_n\right] D_n^{-1} = O_p(1); \tag{3.41}$$

(b) the eigenvalues of $\widehat{\Pi}_{1st}$ *satisfy* $\phi_k(\widehat{\Pi}_{1st}) \to_p \phi_k(\Pi_o)$ *for* $k = 1, ..., m$*;*

(c) the last $m - r_o$ *ordered eigenvalues of* $\widehat{\Pi}_{1st}$ *satisfy*

$$n[\phi_{r_o+1}(\widehat{\Pi}_{1st}), ..., \phi_m(\widehat{\Pi}_{1st})] \to_d [\widetilde{\phi}'_{r_o+1}, ..., \widetilde{\phi}'_m] \tag{3.42}$$

where $\widetilde{\phi}'_j$ $(j = r_o + 1, ..., m)$ *are the ordered solutions of*

$$\left| uI_{m-r_o} - \beta'_{o\perp}\left[\left(\int dB_u B'_{w_2}\right)\left(\int B_{w_2}B'_{w_2}\right)^{-1}\alpha'_{o\perp} + \widetilde{\Pi}_1\right]\beta_{o\perp}\right| = 0; \tag{3.43}$$

(d) $\widehat{\Pi}_{1st}$ *has* $r_o - r_1$ *eigenvalues satisfying*

$$\sqrt{n}[\phi_{r_1+1}(\widehat{\Pi}_{1st}), ..., \phi_{r_o}(\widehat{\Pi}_{1st})] \to_d [\widetilde{\phi}'_{r_1+1}, ..., \widetilde{\phi}'_{r_o}] \tag{3.44}$$

where $\widetilde{\phi}'_j$ $(j = r_1 + 1, ..., r_o)$ *are the ordered solutions of*

$$\left| uI_{r_o-r_1} - \widetilde{\beta}'_\perp\left(\widetilde{\Pi}_0 + \widetilde{\Pi}_2\right)\widetilde{\beta}_\perp\right| = 0 \tag{3.45}$$

and $\widetilde{\Pi}_0$ *is a random matrix with distribution* $N(0, V_{uz_1})\Sigma^{-1}_{z_1z_1}\beta'_o$.

We next derive the asymptotic properties of the shrinkage estimator $\widehat{\Pi}_n$ with weakly dependent innovations. By the full rank of $\widetilde{\alpha}_1$ and $\widetilde{\beta}_1$, we can define

$$Q_1 = \left[\widetilde{\beta}_1, \widetilde{\alpha}_{1\perp}\right] \text{ and } Q_1^{-1} = \left[\widetilde{\alpha}_1(\widetilde{\beta}_1'\widetilde{\alpha}_1)^{-1}, \widetilde{\beta}_{1\perp}(\widetilde{\alpha}_{1\perp}'\widetilde{\beta}_{1\perp})^{-1}\right].$$

Let $\Phi_{1,k}(\Pi) = Q_1(k)\Pi$ where $Q_1(k)$ denotes the k-th row of Q_1, then the index set $\widetilde{\mathcal{S}}_\phi \equiv \{k : \Phi_{1,k}(\Pi_1) \neq 0\} = \{1, ...r_1\}$ is a subset of $S_\phi = \{k : \Phi_k(\Pi_o) \neq 0\} = \{1, ...r_o\}$. We next derive the "consistency" of $\widehat{\Pi}_n$.

Corollary 3.4.4 *Under Assumption 3.3.2, 3.4.1 and $\lambda_{r,k,n} = o(1)$ for any $k \in \widetilde{\mathcal{S}}_\phi$, the shrinkage estimator $\widehat{\Pi}_n$ satisfies*

$$\widehat{\Pi}_n \to_p \Pi_1, \tag{3.46}$$

where Π_1 is defined in (3.37).

Corollary 3.4.4 implies that the shrinkage estimator $\widehat{\Pi}_n$ has the same probability limit as that of the OLS estimator $\widehat{\Pi}_{1st}$. The next corollary provides the convergence rate of the LS shrinkage estimate to the pseudo true parameter matrix Π_1.

Corollary 3.4.5 *Denote $\widetilde{\delta}_{r,n} = \max_{k\in\widetilde{\mathcal{S}}_\phi} ||\lambda_{r,k,n}\phi_k(\widehat{\Pi}_{1st})||^{-\omega}$. Under Assumptions RR , LP and $\lambda_{r,k,n} = o(1)$ for any $k \in \widetilde{\mathcal{S}}_\phi$, the shrinkage LS estimator $\widehat{\Pi}_n$ satisfies*

(a) if $r_o = 0$, then $\widehat{\Pi}_n - \Pi_1 = O_p(n^{-1} + n^{-1}\widetilde{\delta}_{r,n})$;

(b) if $0 < r_o \leq m$, then $\left(\widehat{\Pi}_n - \Pi_1\right) Q^{-1}D_n^{-1} = O_p(1 + n^{\frac{1}{2}}\widetilde{\delta}_{r,n})$.

Recall that Q_n is the normalized left eigenvector matrix of $\widehat{\Pi}_{1st}$. Decompose Q_n' as $\left[Q_{\widetilde{\alpha},n}', Q_{\widetilde{\alpha}_\perp,n}'\right]$, where $Q_{\widetilde{\alpha},n}$ and $Q_{\widetilde{\alpha}_\perp,n}$ are the first r_1 and last $m - r_1$ rows of Q_n. Under Corollary 3.4.4 and Lemma 3.4.1,

$$Q_{\widetilde{\alpha},n}\widehat{\Pi}_n = Q_{\widetilde{\alpha},n}\widehat{\Pi}_{1st} + o_p(1) = \Lambda_{\widetilde{\alpha},n}Q_{\widetilde{\alpha},n} + o_p(1) \tag{3.47}$$

where $\Lambda_{\widetilde{\alpha},n}$ is a diagonal matrix with the ordered first (largest) r_1 eigenvalues of $\widehat{\Pi}_{1st}$. (3.47) implies that the first r_1 rows of $Q_n\widehat{\Pi}_n$ are estimated as nonzero w.p.a.1. We next show that the last $m - r_o$ rows of $Q_n\widehat{\Pi}_n$ are estimated as zeros w.p.a.1.

Corollary 3.4.6 *Under Assumption LP and RR, if* $n^{\omega}\lambda_{r,k,n} \to \infty$ *for* $k \in S_{\phi}^{c}$ *and* $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = O_p(1)$*, then we have*

$$\Pr\left(Q_n(k)\widehat{\Pi}_n = 0\right) \to 1 \tag{3.48}$$

for any $k \in S_{\phi}^{c}$.

Corollary 3.4.6 implies that $\widehat{\Pi}_n$ has at least $m - r_o$ eigenvalues estimated as zero w.p.a.1. However, the matrix Π_1 may have more zero eigenvalues than Π_o. To ensure consistent cointegration rank selection, we need to show that the $r_o - r_1$ zero eigenvalues of Π_1 are estimated as non-zeros w.p.a.1. From Lemma 3.4.3, we see that $\widehat{\Pi}_{1st}$ has $m - r_o$ eigenvalues which converge to zero at the rate n and $r_o - r_1$ eigenvalues which converge to zero at the rate $\sqrt{n}$. The different convergence rates of the estimates of the zero eigenvalues of Π_1 enable us to empirically distinguish the estimates of the $m - r_o$ zero eigenvalues of Π_1 from the estimates of the $r_o - r_1$ zero eigenvalues of Π_1, as illustrated in the following corollary.

Corollary 3.4.7 *Under Assumption LP and RR, if* $n^{\frac{1+\omega}{2}}\lambda_{r,k,n} = o(1)$ *for* $k \in \{r_1 + 1, ..., r_o\}$ *and* $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = O_p(1)$*, then we have*

$$\Pr\left(Q_n(k)\widehat{\Pi}_n \neq 0\right) \to 1 \tag{3.49}$$

for any $k \in \{r_1 + 1, ..., r_o\}$.

In the proof of Corollary 3.4.7, we show that $n^{\frac{1}{2}}Q_n(k)\widehat{\Pi}_n$ converges in distribution to some non-degenerated continuous random vectors, which is a stronger result than

(3.49). Corollary 3.4.5 and Corollary 3.4.7 implies that $\widehat{\Pi}_n$ has at least $m - r_o$ eigenvalues not estimated as zeros w.p.a.1. Hence Corollary 3.4.5, Corollary 3.4.6 and Corollary 3.4.7 give us the following result immediately.

Theorem 3.4.8 (Super-efficiency) *Under the conditions of Corollary 3.4.6 and Corollary 3.4.7, we have*

$$\Pr\left(r(\widehat{\Pi}_n) = r_o\right) \to 1 \tag{3.50}$$

as $n \to \infty$, *where* $r(\widehat{\Pi}_n)$ *denotes the rank of* $\widehat{\Pi}_n$.

Theorem 3.4.8 states that the true cointegration rank r_o can be consistently selected, though the matrix Π_o is not consistently estimable. Moreover, when the probability limit Π_1 of the LS shrinkage estimator has rank less than r_o, Theorem 3.4.8 ensures that only r_o rank is selected in the LS shrinkage estimation. This result is new in the shrinkage based model selection literature, as the Lasso-type of techniques are usually advocated because of their ability of shrinking small estimates (in magnitude) to be zeros in estimation. However, in Corollary 3.4.7, we show the LS shrinkage estimation does not shrink the estimates of the extra $r_o - r_1$ zero eigenvalues of Π_1 to be zero.

3.5 Extension II: VECM Estimation with Explicit Transient Dynamics

This section studies estimation of the general model

$$\Delta Y_t = \Pi_o Y_{t-1} + \sum_{j=1}^{p} B_{o,j} \Delta Y_{t-j} + u_t \tag{3.51}$$

with simultaneous cointegration rank selection and lag order selection. Using $B_o = (B_{o,1}, ..., B_{o,p})$ the unknown parameters (Π_o, B_o) are estimated by penalized LS estimation

$$(\widehat{\Pi}_n, \widehat{B}_n) = \underset{\Pi, B_1, ..., B_p \in R^{m\times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2 + \sum_{j=1}^{p} \frac{n\lambda_{b,j,n}}{||\widehat{B}_{j,1st}||^{\omega}} \|B_j\| + \sum_{k=1}^{m} \frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \|\Phi_{n,k}(\Pi)\| \right\}. \quad (3.52)$$

For consistent lag order selection the model should be consistently estimable and it is assumed that the given p in (3.51) is such that the error term u_t satisfies Assumption 1.3.1. Define

$$C(\phi) = \Pi_o + \sum_{j=0}^{p} B_{o,j}(1-\phi)\phi^j, \text{ where } B_{o,0} = -I_m.$$

The following assumption extends Assumption 1.3.2 to accommodate the general structure in (3.51).

Assumption 3.5.1 (RR) *(i) The determinantal equation $|C(\phi)| = 0$ has roots on or outside the unit circle; (ii) the matrix Π_o has rank r_o, with $0 \le r_o \le m$; (iii) the $(m - r_o) \times (m - r_o)$ matrix*

$$\alpha'_{o,\perp} \left(I_m - \sum_{j=1}^{p} B_{o,j} \right) \beta_{o,\perp} \quad (3.53)$$

is nonsingular.

Under Assumption 3.5.1, the time series Y_t has following partial sum representation,

$$Y_t = C_B \sum_{s=1}^{t} u_s + \Xi(L)u_t + C_B Y_0 \quad (3.54)$$

where $C_B = \beta_{o,\perp}\left[\alpha'_{o,\perp}\left(I_m - \sum_{j=1}^{p} B_{o,j}\right)\beta_{o,\perp}\right]^{-1}\alpha'_{o,\perp}$ and $\Xi(L)u_t = \sum_{s=0}^{\infty}\Xi_s u_{t-s}$ is a stationary process. From the partial sum representation in (3.54), we deduce that $\beta'_o Y_t = \beta'_o \Xi(L)u_t$ and ΔY_{t-j} $(j = 0, ..., p)$ are stationary.

Define an $m(p+1) \times m(p+1)$ rotation matrix Q_B and its inverse Q_B^{-1} as

$$Q_B \equiv \begin{pmatrix} \beta'_o & 0 \\ 0 & I_{mp} \\ \alpha'_{o,\perp} & 0 \end{pmatrix} \text{ and } Q_B^{-1} = \begin{pmatrix} \alpha_o(\beta'_o\alpha_o)^{-1} & 0 & \beta_{o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1} \\ 0 & I_{mp} & 0 \end{pmatrix}.$$

Denote $\Delta X_{t-1} = \left[\Delta Y'_{t-1}, ..., \Delta Y'_{t-p}\right]'$ and then the model in (3.51) can be written as

$$\Delta Y_t = \begin{bmatrix} \Pi_o & B_o \end{bmatrix} \begin{bmatrix} Y_{t-1} \\ \Delta X_{t-1} \end{bmatrix} + u_t. \tag{3.55}$$

Let

$$Z_{t-1} = Q_B \begin{bmatrix} Y_{t-1} \\ \Delta X_{t-1} \end{bmatrix} = \begin{bmatrix} Z_{3,t-1} \\ Z_{2,t-1} \end{bmatrix}, \tag{3.56}$$

where $Z'_{3,t-1} = \begin{bmatrix} Y'_{t-1}\beta_o & \Delta X'_{t-1} \end{bmatrix}$ is a stationary process and $Z_{2,t-1} = \alpha'_{o,\perp}Y_{t-1}$ comprises the $I(1)$ components.

Lemma 3.5.1 *Under Assumption 3.3.1 and Assumption 3.5.1, we have*

(a) $n^{-1}\sum_{t=1}^{n} Z_{3,t-1}Z'_{3,t-1} \to_p \Sigma_{z_3 z_3}$;

(b) $n^{-\frac{3}{2}}\sum_{t=1}^{n} Z_{3,t-1}Z'_{2,t-1} \to_p 0$;

(c) $n^{-2}\sum_{t=1}^{n} Z_{2,t-1}Z'_{2,t-1} \to_d \int B_{w_2}B'_{w_2}$;

(d) $n^{-\frac{1}{2}}\sum_{t=1}^{n} u_t Z'_{3,t-1} \to_d N(0, \Omega_u \otimes \Sigma_{z_3 z_3})$;

(e) $n^{-1}\sum_{t=1}^{n} u_t Z'_{2,t-1} \to_d \left(\int B_{w_2} dB'_u\right)'$;

and the quantities in (c), (d), and (e) converge jointly.

Lemma 3.5.1 follows by standard arguments like those in Lemma 3.9.1 and its proof is omitted. We first establish the asymptotic properties of the OLS estimator $(\widehat{\Pi}_{1st}, \widehat{B}_{1st})$ of (Π_o, B_o) and the asymptotic properties of the eigenvalues of $\widehat{\Pi}_{1st}$. The estimate $(\widehat{\Pi}_{1st}, \widehat{B}_{1st})$ has the following closed-form solution

$$\left(\widehat{\Pi}_{1st}, \widehat{B}_{1st}\right) = \begin{pmatrix} \widehat{S}_{y_0y_1} & \widehat{S}_{y_0x_0} \end{pmatrix} \begin{pmatrix} \widehat{S}_{y_1y_1} & \widehat{S}_{y_1x_0} \\ \widehat{S}_{x_0y_1} & \widehat{S}_{x_0x_0} \end{pmatrix}^{-1}, \tag{3.57}$$

where

$$\begin{aligned} \widehat{S}_{y_0y_1} &= \frac{1}{n}\sum_{t=1}^{n} \Delta Y_t Y_{t-1}', \ \widehat{S}_{y_0x_0} = \frac{1}{n}\sum_{t=1}^{n} \Delta Y_t \Delta X_{t-1}', \ \widehat{S}_{y_1y_1} = \frac{1}{n}\sum_{t=1}^{n} Y_{t-1} Y_{t-1}', \\ \widehat{S}_{y_1x_0} &= \frac{1}{n}\sum_{t=1}^{n} Y_{t-1} \Delta X_{t-1}', \ \widehat{S}_{x_0y_1} = \widehat{S}_{y_1x_0}' \text{ and } \widehat{S}_{x_0x_0} = \frac{1}{n}\sum_{t=1}^{n} \Delta X_{t-1} \Delta X_{t-1}' \end{aligned} \tag{3.58}$$

Denote $Y_- = (Y_0, ..., Y_{n-1})_{m\times n}$, $\Delta Y = (\Delta Y_1, ..., \Delta Y_n)_{m\times n}$ and

$$\widehat{M_0} = I_n - n^{-1}\Delta X' \widehat{S}_{x_0x_0}^{-1} \Delta X,$$

where $\Delta X = (\Delta X_0, ..., \Delta X_{n-1})_{mp\times n}$, then $\widehat{\Pi}_{1st}$ has the explicit partitioned regression representation

$$\widehat{\Pi}_{1st} = \left(\Delta Y \widehat{M_0} Y_-'\right)\left(Y_- \widehat{M_0} Y_-'\right)^{-1} = \Pi_o + \left(U \widehat{M_0} Y_-'\right)\left(Y_- \widehat{M_0} Y_-'\right)^{-1}, \tag{3.59}$$

where $U = (u_1, ..., u_n)_{m\times n}$. Recall that $[\phi_1(\widehat{\Pi}_{1st}), ..., \phi_m(\widehat{\Pi}_{1st})]$ and $[\phi_1(\Pi_o), ..., \phi_m(\Pi_o)]$ are the ordered eigenvalues of $\widehat{\Pi}_{1st}$ and Π_o respectively, where $\phi_j(\Pi_o) = 0$ $(j = r_o + 1, ..., m)$. Let Q_n be the normalized left eigenvector matrix of $\widehat{\Pi}_{1st}$.

Lemma 3.5.2 *Suppose Assumption 3.3.1 and Assumption 3.5.1 hold.*

(a) Define $D_{n,B} = diag(n^{\frac{1}{2}} I_{r_o+mp}, nI_{m-r_o})$, *then* $\left[(\widehat{\Pi}_{1st}, \widehat{B}_{1st}) - (\Pi_o, B_o)\right] Q_B^{-1} D_{n,B}$

has the following partitioned limit distribution

$$\left[\; N(0, \Omega_u \otimes \Sigma_{z_3 z_3}^{-1}), \quad \int dB_u B_{w_2}' (\int B_{w_2} B_{w_2}')^{-1} \; \right]; \tag{3.60}$$

(b) The eigenvalues of $\widehat{\Pi}_{1st}$ *satisfy* $\phi_k(\widehat{\Pi}_{1st}) \to_p \phi_k(\Pi_o)$ *for* $\forall k = 1, ..., m$*;*

(c) For $\forall k = r_o+1, ..., m$*, the eigenvalues* $\phi_k(\widehat{\Pi}_{1st})$ *of* $\widehat{\Pi}_{1st}$ *satisfy Lemma 3.3.1.(c).*

Lemma 3.5.2 is useful, because the first step estimator $(\widehat{\Pi}_{1st}, \widehat{B}_{1st})$ and the eigenvalues of $\widehat{\Pi}_{1st}$ are used in the construction of the penalty function. Denote the index set of the zero components in B_o as $\mathcal{S}_B^c$ such that $\|B_{o,j}\| = 0$ for all $j \in \mathcal{S}_B^c$ and $\|B_{o,j}\| \neq 0$ otherwise. We next derive the asymptotic properties of the LS shrinkage estimator $(\widehat{\Pi}_n, \widehat{B}_n)$ defined in (3.52).

Lemma 3.5.3 *Suppose that Assumption 3.3.1 and Assumption 3.5.1 are satisfied. If* $\delta_{r,n} = o_p(1)$ *and* $\delta_{b,n} = o_p(1)$ *where* $\delta_{b,n} \equiv \max_{j \in \mathcal{S}_B} \left\{ \lambda_{b,j,n} ||\widehat{B}_{j,1st}||^{-\omega} \right\}$*, then the LS shrinkage estimator* $(\widehat{\Pi}_n, \widehat{B}_n)$ *satisfies*

$$\left[(\widehat{\Pi}_n, \widehat{B}_n) - (\Pi_o, B_o) \right] Q_B^{-1} D_{n,B} = O_p(1 + n^{\frac{1}{2}} \delta_{r,n} + n^{\frac{1}{2}} \delta_{b,n}). \tag{3.61}$$

Lemma 3.5.3 implies that the LS shrinkage estimators $(\widehat{\Pi}_n, \widehat{B}_n)$ have the same convergence rates as the OLS estimators $(\widehat{\Pi}_{1st}, \widehat{B}_{1st})$. We next show that if the tuning parameters $\lambda_{r,k,n}$ and $\lambda_{b,j,n}$ ($k \in \mathcal{S}_B^c$ and $j \in \mathcal{S}_\phi^c$) converge to zero but not too fast, then the zero rows of $Q\Pi_o$ and zero matrices in B_o are estimated as zero w.p.a.1. Let the zero rows of $\widehat{Q}_n \widehat{\Pi}_n$ be indexed by $\mathcal{S}_{n,\phi}^c$ and the zero matrix in $\widehat{B}_n$ be indexed by $\mathcal{S}_{n,B}^c$.

Theorem 3.5.4 *Suppose that Assumption 3.3.1 and Assumption 3.5.1 hold. If the tuning parameters satisfy* $n^{\frac{1}{2}}(\delta_{r,n} + \delta_{b,n}) = O_p(1)$**,** $n^\omega \lambda_{r,k,n} \to \infty$ *and* $n^{\frac{1+\omega}{2}} \lambda_{b,j,n} \to \infty$

for $k \in \mathcal{S}_\phi^c$ *and* $j \in \mathcal{S}_B^c$*, then we have*

$$\Pr\left(Q_{\alpha,n}\widehat{\Pi}_n = 0\right) \to 1 \text{ as } n \to \infty; \tag{3.62}$$

and for all $j \in \mathcal{S}_B^c$

$$\Pr\left(\widehat{B}_{n,j} = \mathbf{0}_{m\times m}\right) \to 1 \text{ as } n \to \infty. \tag{3.63}$$

Theorem 3.5.4 indicates that the zero rows of $Q\Pi_o$ (and hence the zero eigenvalues of Π_o) and the zero matrices in B_o are estimated as zeros w.p.a.1. Thus Lemma 3.5.3 and Theorem 3.5.4 imply consistent cointegration rank selection and consistent lag order selection.

We next derive the centered limit distribution of the shrinkage estimator $\widehat{\Theta}_{\mathcal{S}} = \left(\widehat{\Pi}_n, \widehat{B}_{\mathcal{S}_B}\right)$, where $\widehat{B}_{\mathcal{S}_B}$ denotes the LS shrinkage estimator of the nonzero matrices in B_o. Let $I_{\mathcal{S}_B} = diag(I_{1,m}, ..., I_{d_{\mathcal{S}_B},m})$ where the $I_{j,m}$ $(j = 1, ..., d_{\mathcal{S}_B})$ are $m \times m$ identity matrices and $d_{\mathcal{S}_B}$ is the dimensionality of the index set $\mathcal{S}_B$. Define

$$Q_{\mathcal{S}} \equiv \begin{pmatrix} \beta_o' & 0 \\ 0 & I_{\mathcal{S}_B} \\ \alpha_{o,\perp}' & 0 \end{pmatrix} \text{ and } D_{n,\mathcal{S}} \equiv diag(n^{\frac{1}{2}} I_{r_o}, n^{\frac{1}{2}} I_{\mathcal{S}_B}, n I_{m-r_o}),$$

where the identity matrix $I_{\mathcal{S}_B} = I_{md_{\mathcal{S}_B}}$ in Q_S serves to accommodate the nonzero matrices in B_o. Let $\Delta X_{\mathcal{S},t}$ denote the nonzero lagged differences in (3.51), then the true model can be written as

$$\Delta Y_t = \Pi_o Y_{t-1} + B_{o,\mathcal{S}} \Delta X_{\mathcal{S},t-1} + u_t = \Theta_{o,\mathcal{S}} Q_{\mathcal{S}}^{-1} Z_{\mathcal{S},t-1} + u_t \tag{3.64}$$

where the transformed and reduced regressor variables are

$$Z_{\mathcal{S},t-1} = Q_{\mathcal{S}} \begin{bmatrix} Y_{t-1} \\ \Delta X_{\mathcal{S},t-1} \end{bmatrix} = \begin{bmatrix} Z_{3\mathcal{S},t-1} \\ Z_{2,t-1} \end{bmatrix},$$

with $Z'_{3\mathcal{S},t-1} = \begin{bmatrix} Y'_{t-1}\beta_o & \Delta X'_{\mathcal{S},t-1} \end{bmatrix}$ and $Z_{2,t-1} = \alpha'_{o,\perp} Y_{t-1}$. From Lemma 3.5.1, we obtain

$$n^{-1} \sum_{t=1}^{n} Z_{3\mathcal{S},t-1} Z'_{3\mathcal{S},t-1} \to_p E\left[Z_{3\mathcal{S},t-1} Z'_{3\mathcal{S},t-1}\right] \equiv \Sigma_{z_{3\mathcal{S}} z_{3\mathcal{S}}}.$$

The centred limit theory of $\widehat{\Theta}_{\mathcal{S}}$ is given in the following result.

Theorem 3.5.5 *Under conditions of Theorem 3.5.4, if* $n^{\frac{1}{2}}(\delta_{r,n} + \delta_{b,n}) = o_p(1)$, *then*

$$\left(\widehat{\Theta}_{\mathcal{S}} - \Theta_{o,\mathcal{S}}\right) Q_{\mathcal{S}}^{-1} D_{n,\mathcal{S}}^{-1} \to_d \begin{pmatrix} B_{m,\mathcal{S}} & \alpha_o(\alpha'_o\alpha_o)^{-1}\alpha'_o B_{m,2} \end{pmatrix}, \tag{3.65}$$

where $B_{m,\mathcal{S}} \equiv N(0, \Omega_u \otimes \Sigma^{-1}_{z_{3\mathcal{S}} z_{3\mathcal{S}}})$ *and* $B_{m,2}$ *is defined in Lemma 3.3.1.(a).*

Theorem 3.5.5 extends the result of Theorem 3.3.6 to the general VEC model with lagged differences. From Theorem 3.5.5, the LS shrinkage estimator $\widehat{\Theta}_{\mathcal{S}}$ is more efficient than the OLS estimator $\widehat{\Theta}_n$ in the sense that: (i) the zero components in B_o are estimated as zeros w.p.a.1 and thus their LS shrinkage estimators are super efficient; (ii) under the consistent lagged differences selection, the true nonzero components in B_o are more efficiently estimated in the sense of smaller asymptotic variance; and (iii) the true cointegration rank is estimated and therefore when $r_o < m$ some parts of the matrix Π_o are estimated at a rate faster than root-n.

The LS shrinkage estimator $\widehat{\Pi}_n$ suffers from second order bias, evident in the component $U_{2,m}$ of the limit (3.65). Accordingly we define the GLS shrinkage estimator

of the general VEC model as

$$(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n}) = \underset{\Pi, B_1, \ldots, B_p \in R^{m\times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2_{\widehat{\Omega}^{-1}_{u,n}} + \sum_{j=1}^{p} \frac{n\lambda_{b,j,n}}{||\widehat{B}_{j,1st}||^{\omega}} \|B_j\| + \sum_{k=1}^{m} \frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \|\Phi_{n,k}(\Pi)\| \right\}. \quad (3.66)$$

To conclude this section, we show that the GLS shrinkage estimator $(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n})$ is oracle efficient in the sense that it has the same asymptotic distribution as the RRR estimate assuming the true cointegration rank and lagged differences are known.

Corollary 3.5.6 (Oracle Properties of GLS) *Suppose the conditions of Theorem 3.5.5 are satisfied. If* $\widehat{\Omega}_{u,n} \to_p \Omega_u$, *then*

$$\Pr\left(r(\widehat{\Pi}_{g,n}) = r_o\right) \to 1 \text{ and } \Pr\left(\widehat{B}_{g,j,n} = 0\right) \to 1 \quad (3.67)$$

for $j \in \mathcal{S}_B^c$ *as* $n \to \infty$; *moreover* $\widehat{\Theta}_{\mathcal{S}}$ *has the following limiting distribution*

$$\left(\widehat{\Theta}_{\mathcal{S}} - \Theta_{o,\mathcal{S}}\right) Q_{\mathcal{S}}^{-1} D_{n,\mathcal{S}}^{-1} \to_d \left(\begin{array}{cc} B_{m,\mathcal{S}} & \alpha_o(\beta_o'\alpha_o)^{-1} \int dB_{u\cdot w_2} B_{w_2}' (\int B_{w_2} B_{w_2}')^{-1} \end{array} \right) \quad (3.68)$$

where $B_{u\cdot w_2}$ *is defined in Theorem 3.3.7.*

Corollary 3.5.6 is proved using the same arguments of Corollary 3.3.7 and Theorem 3.5.5 and its proof is omitted.

Remark 3.5.7 *Although the grouped adaptive Lasso function* $P(B) = ||\widehat{B}_{1st}||^{-\omega} \|B\|$ *is used in the LS shrinkage estimations (3.52) and the GLS shrinkage estimation (3.66), we remark that the adaptive Lasso penalty function can also be used and the result GLS shrinkage estimate enjoys the same properties stated in Corollary 3.5.6.*

The GLS shrinkage estimation using the adaptive Lasso penalty takes the following form

$$(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n}) = \underset{\Pi, B_1,\ldots,B_p \in R^{m\times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2_{\widehat{\Omega}^{-1}_{u,n}} + \sum_{j=1}^{p}\sum_{l=1}^{m}\sum_{s=1}^{m} \frac{n\lambda_{b,j,n}|B_{j,ls}|}{|\widehat{B}_{j,ls,1st}|^{\omega}} + \sum_{k=1}^{m} \frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \|\Phi_{n,k}(\Pi)\| \right\} \tag{3.69}$$

where $B_{j,ls}$ and $\widehat{B}_{j,ls,1st}$ refer to the l-th row and s-th column elements in B_j and $\widehat{B}_{j,1st}$ respectively. The advantage of the grouped adaptive Lasso penalty $P(B_j)$ is that it shrinks elements in B to zero groupwisely, which makes it a natural choice for the lag order selection in VEC models. The adaptive Lasso penalty is more flexible and when used in the GMM shrinkage estimation, it can not only select the zero matrices, but also zero elements in nonzero $B_{o,j}$ ($j \in \mathcal{S}_B$) w.p.a.1.

Remark 3.5.8 *The flexibility of the adaptive Lasso penalty enables the GLS shrinkage estimation to achieve more goals in one-step, in addition to the model selection and efficient estimation. Suppose that the vector Y_t can be divided in r and $m-r$ dimensional subvectors $Y_{1,t}$ and $Y_{2,t}$, then the VECM can be rewritten as*

$$\begin{bmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{bmatrix} = \begin{bmatrix} \Pi_o^{11} & \Pi_o^{12} \\ \Pi_o^{21} & \Pi_o^{22} \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} B_{o,j}^{11} & B_{o,j}^{12} \\ B_{o,j}^{21} & B_{o,j}^{22} \end{bmatrix} \begin{bmatrix} \Delta Y_{1,t-j} \\ \Delta Y_{2,t-j} \end{bmatrix} + u_t, \tag{3.70}$$

where Π_o and $B_{o,j}$ ($j = 1,..,p$) are partitioned in line with Y_t. By definition, $Y_{2,t}$ does not Granger-cause $Y_{1,t}$ if and only if

$$\Pi_o^{12} = 0 \text{ and } B_{o,j}^{12} = 0 \text{ for any } j \in \mathcal{S}_B.$$

One can attach the (grouped) adaptive Lasso penalty of Π^{12} *in (3.70) such that the causality test is automatically executed in the GLS shrinkage estimation.*

Remark 3.5.9 *In this chapter, we only consider the adaptive Lasso type of penalty functions in the LS or GLS shrinkage estimation. The main advantage of the adaptive Lasso penalty is that it is a convex function, which combines the convexity of the LS or GLS criterion makes the computation of the shrinkage estimate faster and more accurate. It is clear that as long as the tuning parameter satisfies certain rate requirement, our main results remains to hold if one uses other penalty function (e.g., the bridge penalty) in the LS or GLS shrinkage estimation.*

3.6 Adaptive Selection of the Tuning Parameters

This section develops a data-driven procedure of selecting the tuning parameter λ_n. As presented in previous sections, the conditions imposed on λ_n to ensure oracle properties in GLS shrinkage only restrict the rate at which λ_n goes to zero. But in finite samples these conditions are not precise enough to provide a clear choice of tuning parameter for practical implementation. On one hand the tuning parameter should converge to zero as fast as possible so that shrinkage bias in the estimation of the nonzero components of the model is as small as possible. In the extreme case where $\lambda_n = 0$ LS shrinkage reduces to LS estimation and there is no shrinkage bias in the resulting estimators. (Of course there may still be finite sample estimation bias). On the other hand, the tuning parameter should converge to zero so that in finite samples zero components in the model are estimated as zeros with higher probability. In the opposite extremity the tuning parameter $\lambda_n = \infty$ and then all parameters of

the model are estimated as zeros with probability one in finite samples. Thus there is bias and variance trade-off in the selection of the tuning parameters.

By definition $\widehat{T}_n = Q_n\widehat{\Pi}_n$ and the k-th row of $\widehat{T}_n$ is estimated as zero only if the following first order condition holds

$$\left\| \frac{1}{n}\sum_{t=1}^{n} Q_n(k)\widehat{\Omega}_{u,n}^{-1}(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p}\widehat{B}_{n,j}\Delta Y_{t-j})Y_{t-1}' \right\| < \frac{\lambda_{r,k,n}}{2||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}. \quad (3.71)$$

Let $T \equiv Q\Pi_o$ and $T(k)$ be the k-th row of the matrix $Q\Pi_o$. If a nonzero $T(k)$ $(k \leq r_o)$ is estimated as zero, then the left hand side of the above inequality will be asymptotically close to a nonzero real number because the under-selected cointegration rank leads to inconsistent estimation. To ensure the shrinkage bias and errors of under-selecting the cointegration rank are small in finite samples, one would like to have $\lambda_{r,k,n}$ converge to zero as fast as possible.

On the other hand, the zero rows of T are estimated as zero only if the same inequality in (3.71) is satisfied. As $n\phi_k(\widehat{\Pi}_{1st}) = O_p(1)$, we can rewrite the inequality in (3.71) as

$$\left\| \frac{1}{n}\sum_{t=1}^{n} Q_n(k)\widehat{\Omega}_{u,n}^{-1}(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p}\widehat{B}_{n,j}\Delta Y_{t-j})Y_{t-1}' \right\| < \frac{n^{\omega}\lambda_{r,k,n}}{2||n\phi_k(\widehat{\Pi}_{1st})||^{\omega}}. \quad (3.72)$$

The sample average in the left side of this inequality is asymptotically a vector of linear combination of non-degenerated random variables, and it is desirable to have $n^{\omega}\lambda_{r,k,n}$ diverge to infinity as fast as possible to ensure that the true cointegration rank is selected with high probability in finite samples. We propose to choose $\lambda_{r,k,n} = c_{r,k}n^{-\frac{\omega}{2}}$ ($c_{r,k}$ is some positive constant and we will discuss how to select it later in this section.) to balance the requirement that $\lambda_{r,k,n}$ converges to zero and $n^{\omega}\lambda_{r,k,n}$ diverges to infinity as fast as possible.

Using similar arguments we see that the component $B_{o,j}$ in B_o will be estimated as zero if the following condition holds

$$\left\| n^{-\frac{1}{2}} \sum_{t=1}^{n} \widehat{\Omega}_{u,n}^{-1} (\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}) \Delta Y_{t-j}' \right\| < \frac{n^{\frac{1}{2}} \lambda_{b,j,n}}{2||\widehat{B}_{1st,j}||^{\omega}}. \tag{3.73}$$

As $B_{o,j} \neq 0$, the left side of the above inequality will be asymptotically close to a nonzero real number because the under-selected lagged differences also lead to inconsistent estimation. To ensure the shrinkage bias and error of under-selection of the lagged differences are small in the finite samples, it is desirable to have $n^{\frac{1}{2}} \lambda_{b,j,n}$ converge to zero as fast as possible.

On the other hand, the zero component $B_{o,j}$ in B_o is estimated as zero only if the same inequality in (3.73) is satisfied. As $\widehat{B}_{1st,j} = O_p(n^{-\frac{1}{2}})$ the inequality in (3.73) can be written as

$$\left\| n^{-\frac{1}{2}} \sum_{t=1}^{n} \widehat{\Omega}_{u,n}^{-1} (\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}) \Delta Y_{t-j}' \right\| < \frac{n^{\frac{1+\omega}{2}} \lambda_{b,j,n}}{2||n^{\frac{1}{2}} \widehat{B}_{1st,j}||^{\omega}}. \tag{3.74}$$

The sample average in the left side of this inequality is asymptotically a vector of linear combinations of non-degenerated random variables, and again it is desirable to have $n^{\frac{1+\omega}{2}} \lambda_{b,j,n}$ diverge to infinity as fast as possible to ensure that zero components in B_o are selected with high probability in finite samples. We propose to choose $\lambda_{b,j,n} = c_{b,j} n^{-\frac{1}{2} - \frac{\omega}{4}}$ ($c_{b,j}$ is some positive constant and we will discuss how to select it later in this section.) to balance the requirement that $\lambda_{b,j,n}$ converges to zero and $n^{\frac{1+\omega}{2}} \lambda_{b,j,n}$ diverges to infinity as fast as possible.

We next discuss how to choose the loading coefficients in $\lambda_{r,k,n}$ and $\lambda_{b,j,n}$. As the order of the tuning parameter ensures oracle properties of the LS shrinkage estimate, using similar arguments in the proof of Theorem 3.5.5, we rewrite the sample average

in the left hand side of (3.72) as

$$\begin{aligned} F_{\pi,n}(k) &\equiv \frac{Q_n(k)\widehat{\Omega}_{u,n}^{-1}}{n}\sum_{t=1}^{n}(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p}\widehat{B}_{n,j}\Delta Y_{t-j})Y_{t-1}' \\ &= \frac{Q_n(k)\widehat{\Omega}_{u,n}^{-1}}{n}\sum_{t=1}^{n}[u_t - \left(\widehat{\Theta}_n - \Theta_o\right)Q_B^{-1}Z_{t-1}]Y_{t-1}' \\ &= \frac{Q_n(k)\Omega_u^{-1}}{n}\left[\sum_{t=1}^{n}u_t Y_{t-1}' - (\widehat{\Theta}_{\mathcal{S},n} - \Theta_{\mathcal{S},o})Q_{\mathcal{S}}^{-1}\sum_{t=1}^{n}Z_{\mathcal{S},t-1}Y_{t-1}'\right] + o_p(1) \end{aligned} \tag{3.75}$$

where under Lemma 3.5.1, $n^{-1}\sum_{t=1}^{n}u_t Z_{t-1}' = \begin{pmatrix} \mathbf{0}_{m\times r_o} & \int dB_u B_u'\alpha_{o,\perp} \end{pmatrix} + o_p(1)$ and

$$\frac{D_{n,\mathcal{S}}\sum_{t=1}^{n}Z_{\mathcal{S},t-1}Z_{t-1}'}{n} = \begin{pmatrix} \mathbf{0}_{(mp+r_o)\times r_o} & 0 \\ 0 & \alpha_{o,\perp}'\int dB_u B_u'\alpha_{o,\perp} \end{pmatrix} + o_p(1). \tag{3.76}$$

Using (3.76) and the arguments in the proof of Theorem 3.5.5,

$$\begin{aligned} &\frac{\left(\widehat{\Theta}_{\mathcal{S},n} - \Theta_{\mathcal{S},o}\right)Q_{\mathcal{S}}^{-1}D_{n,\mathcal{S}}^{-1}D_{n,\mathcal{S}}\sum_{t=1}^{n}Z_{\mathcal{S},t-1}Z_{t-1}'}{n} \\ = &\begin{pmatrix} \mathbf{0}_{m\times r_o} & \alpha_o(\alpha_o'\Omega_u^{-1}\alpha_o)^{-1}\alpha_o'\Omega_u^{-1}\int dB_u B_u'\alpha_{o,\perp} \end{pmatrix} + o_p(1). \end{aligned} \tag{3.77}$$

From the results in (3.75), (3.76) and (3.77), we can deduce that

$$F_{\pi,n}(k) = Q_n(k)T_{1,\pi_o}\int dB_u B_u' T_{2,\pi_o} + o_p(1), \tag{3.78}$$

where $T_{1,\pi_o} = \Omega_u^{-1} - \Omega_u^{-1}\alpha_o(\alpha_o'\Omega_u^{-1}\alpha_o)^{-1}\alpha_o'\Omega_u^{-1}$ and $T_{2,\pi_o} = \alpha_{o,\perp}(\beta_{o,\perp}'\alpha_{o,\perp})^{-1}\beta_{o,\perp}'$. We propose to select $c_{r,k}$ to normalize the random sum in (3.78), i.e.

$$\widehat{c}_{r,k} = 2\left\|Q_n(k)\widehat{T}_{1,\pi}\widehat{\Omega}_{u,n}^{1/2}\right\| \times \left\|\widehat{\Omega}_{u,n}^{1/2}\widehat{T}_{2,\pi}\right\| \tag{3.79}$$

where $\widehat{T}_{1,\pi}$ and $\widehat{T}_{2,\pi}$ are some estimates of T_{1,π_o} and T_{2,π_o}. Of course, the rank of Π_o needs to be estimated before T_{1,π_o} and T_{2,π_o} can be estimated. We propose to run a first step shrinkage estimation with $\lambda_{r,k,n} = 2\log(n)n^{-\frac{\omega}{2}}$ and $\lambda_{b,j,n} = 2\log(n)n^{-\frac{1}{2}-\frac{\omega}{4}}$ to get an initial estimator of the rank r_o and order of the lagged differences. Then based on this first-step shrinkage estimation, one can construct $\widehat{T}_{1,\pi}$, $\widehat{T}_{2,\pi}$ and thus the empirical loading coefficient $\widehat{c}_{r,k}$.

Similarly, we can rewrite the sample average in the left hand side of (3.74)

$$\begin{aligned} F_{b,n}(j) &\equiv \frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p}\widehat{B}_{n,j}\Delta Y_{t-j})\Delta Y'_{t-j} \\ &= \frac{1}{\sqrt{n}}\sum_{t=1}^{n}[u_t - \left(\widehat{\Theta}_n - \Theta_o\right)Q_B^{-1}Z_{t-1}]\Delta Y'_{t-j} \\ &= \frac{1}{\sqrt{n}}\left[\sum_{t=1}^{n}u_t\Delta Y'_{t-j} - \left(\widehat{\Theta}_{\mathcal{S},n} - \Theta_{\mathcal{S},o}\right)Q_{\mathcal{S}}^{-1}\left(\sum_{t=1}^{n}Z_{\mathcal{S},t-1}\Delta Y'_{t-j}\right)\right] \quad (3.80) \end{aligned}$$

where under Lemma 3.5.1, $n^{-\frac{1}{2}}\sum_{t=1}^{n}u_t\Delta Y'_{t-j} = N(0,\Omega_u\otimes\Sigma_{\Delta y_j\Delta y_j}) + o_p(1)$ and

$$\frac{D_{n,\mathcal{S}}\sum_{t=1}^{n}Z_{\mathcal{S},t-1}\Delta Y'_{t-j}}{\sqrt{n}} = \begin{pmatrix}\Sigma_{z_{3\mathcal{S}}\Delta y_j}\\ 0\end{pmatrix} + o_p(1). \quad (3.81)$$

Using (3.81) and the arguments in the proof of Theorem 3.5.5,

$$\frac{\left(\widehat{\Theta}_{\mathcal{S},n} - \Theta_{\mathcal{S},o}\right)Q_{\mathcal{S}}^{-1}\sum_{t=1}^{n}Z_{\mathcal{S},t-1}\Delta Y'_{t-j}}{\sqrt{n}} = N(0,\Omega_u\otimes\Sigma_{z_{3\mathcal{S}}z_{3\mathcal{S}}})\Sigma^{-1}_{z_{3\mathcal{S}}z_{3\mathcal{S}}}\Sigma_{z_{3\mathcal{S}}\Delta y_j}. \quad (3.82)$$

From the results in (3.80), (3.81) and (3.82), we can deduce that

$$F_{b,n}(j) = \Omega_u^{1/2}\left(B_{m,m}\Sigma^{1/2}_{\Delta y_j\Delta y_j} + B_{m,r_o+mp_o}\Sigma^{-1/2}_{z_{3\mathcal{S}}z_{3\mathcal{S}}}\Sigma_{z_{3\mathcal{S}}\Delta y_j}\right) \quad (3.83)$$

where $B_{m,m} \equiv N(0, I_m\otimes I_m)$ and $B_{m,r_o+mp_o} \equiv N(0, I_m\otimes I_{(p_o+1)m})$. We propose to

select c_b to normalize the random sum in (3.83), i.e.

$$\widehat{c}_{b,j} = 2\left\|\widehat{\Omega}_{u,n}^{1/2}\right\| \times \left(||\widehat{S}_{\Delta y_j \Delta y_j}^{1/2}|| + ||\widehat{S}_{z_{3\mathcal{S}} z_{3\mathcal{S}}}^{-1/2} \widehat{S}_{z_{3\mathcal{S}} \Delta y_j}^{1/2}||\right) \tag{3.84}$$

where $\widehat{S}_{\Delta y_j \Delta y_j} = n^{-1}\sum_{t=1}^{n} \Delta Y_{t-j} \Delta Y_{t-j}'$, $\widehat{S}_{z_{3\mathcal{S}} z_{3\mathcal{S}}} = n^{-1}\sum_{t=1}^{n} Z_{\mathcal{S},t-1} Z_{\mathcal{S},t-1}'$ and $\widehat{S}_{z_{3\mathcal{S}} \Delta y_j} = n^{-1}\sum_{t=1}^{n} Z_{\mathcal{S},t-1} \Delta Y_{t-j}'$. As we have discussed, $\widehat{c}_{b,j}$ can be constructed using estimates from the first-step shrinkage estimation with $\lambda_{r,k,n} = 2\log(n) n^{-\frac{\omega}{2}}$ and $\lambda_{b,j,n} = 2\log(n) n^{-\frac{1}{2}-\frac{\omega}{4}}$.

The choice of ω is a more complicated issue which is not pursued in this chapter. For the empirical application, we propose to choose $\omega = 2$ because such choice is popular in the Lasso-based variable selection literature and our simulation results based on $\omega = 2$ are remarkably well. Based on the above results, we propose the following data dependent tuning parameters for LS shrinkage estimation:

$$\lambda_{r,k,n} = \frac{2}{n}\left\|Q_n(k)\widehat{T}_{1,\pi}\widehat{\Omega}_{u,n}^{1/2}\right\| \times \left\|\widehat{\Omega}_{u,n}^{1/2}\widehat{T}_{2,\pi}\right\| \tag{3.85}$$

and

$$\lambda_{b,j,n} = \frac{2}{n}\left\|\widehat{\Omega}_{u,n}^{1/2}\right\| \times \left(||\widehat{S}_{\Delta y_j \Delta y_j}^{1/2}|| + ||\widehat{S}_{z_{3\mathcal{S}} z_{3\mathcal{S}}}^{-1/2} \widehat{S}_{z_{3\mathcal{S}} \Delta y}^{1/2}||\right) \tag{3.86}$$

for $k = 1, ..., m$ and $j = 1, ..., p$.

3.7 Simulation Study

We conduct simulation analysis to assess the finite sample performance of the shrinkage estimates in terms of cointegration rank selection and efficient estimation. Three models are investigated in this section. In the first model, the simulated data are

generated from

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix}, \tag{3.87}$$

where $u_t \equiv$i.i.d. $N(0, \Omega_u)$ with $\Omega_u = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.75 \end{pmatrix}$. The initial observation Y_0 is set to be zero for simplicity. Π_o will be specified as

$$\begin{pmatrix} \pi_{11,o} & \pi_{12,o} \\ \pi_{21,o} & \pi_{22,o} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} -1 & -0.5 \\ 1 & 0.5 \end{pmatrix} \text{ and } \begin{pmatrix} -0.5 & 0.1 \\ 0.2 & -0.4 \end{pmatrix} \tag{3.88}$$

to allow for the cointegration rank to be 2, 1 and 0 respectively. Y_0 is set to be zero for simplicity.

In the second model, the simulated data $\{Y_t\}_{t=1}^n$ are generated from equation (3.87)-(3.88), while the innovation term u_t is generated by

$$\begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 0.75 \end{pmatrix} \begin{pmatrix} u_{1,t-1} \\ u_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix},$$

where $\varepsilon_t \equiv$i.i.d. $N(0, \Omega_\varepsilon)$ with $\Omega_\varepsilon = diag(1.25, 0.75)$. The initial values Y_0 and ε_0 are set to be zero.

The third model has the following form

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + B_{1,o} \begin{pmatrix} \Delta Y_{1,t-1} \\ \Delta Y_{2,t-1} \end{pmatrix} + B_{3,o} \begin{pmatrix} \Delta Y_{1,t-3} \\ \Delta Y_{2,t-3} \end{pmatrix} + u_t, \tag{3.89}$$

where u_t is generated under the same condition in (3.87), Π_o is specified similarly in (3.88), $B_{2,o}$ is taken to be $diag(0.4, 0.4)$ such that Assumption 3.5.1 is satisfied. The initial values (Y_t, ε_t) $(t = -3, ..., 0)$ are set to be zero. In the above three cases,

Table 3.1 Probabilities of Cointegration Rank Selection

	Model 1					
	r_o=0, λ_o=(0 0)		r_o=1, λ_o=(0 -.5)		r_o=2, λ_o=(-.6 -.5)	
	n=100	n=400	n=100	n=400	n=100	n=400
$\widehat{r}_n$= 0	0.959	0.998	0.000	0.000	0.000	0.000
$\widehat{r}_n$= 1	0.041	0.002	0.995	1.000	0.000	0.000
$\widehat{r}_n$= 2	0.000	0.000	0.005	0.000	1.000	1.000
	Model 2					
	r_o=0, λ_o=(0 0)		r_o=1, λ_o^*=(0 -0.25)		r_o=2, λ_o^*=(-0.30 -0.15)	
	n=100	n=400	n=100	n=400	n=100	n=400
$\widehat{r}_n$= 0	0.000	0.000	0.001	0.000	0.001	0.000
$\widehat{r}_n$= 1	0.012	0.001	0.953	0.996	0.121	0.001
$\widehat{r}_n$= 2	0.988	0.999	0.046	0.004	0.878	0.999

Table 3.1: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in eqation (3.83). λ_o in each column represents the eigenvalues of the true matrix Π_o .

we include 50 additional observations to the simulated sample with sample size n to eliminate start-up effects from the initialization.

In the first two models, we assume that the econometrician specifies the following model

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + u_t, \tag{3.90}$$

where u_t is i.i.d.$(0, \Omega_u)$ with some unknown positive definite matrix Ω_u. The above empirical model is correctly specified under the data generating assumption (3.87), but is misspecified under (3.88). We are interested in investigating the performance of the shrinkage method in selecting the correct rank of Π_o under both data generating assumptions and efficient estimation of Π_o under Assumption (3.87).

In the third model, we assume that the econometrician specifies the following model

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \sum_{j=1}^{3} B_{j,o} \begin{pmatrix} \Delta Y_{1,t-j} \\ \Delta Y_{2,t-j} \end{pmatrix} + u_t, \tag{3.91}$$

where u_t is i.i.d.$(0, \Omega_u)$ with some unknown positive definite matrix Ω_u. The above

empirical model is over-parameterized according to (3.89). We are interested in investigating the performance of the shrinkage method in selecting the correct rank of Π_o and the order of the lagged differences, and efficient estimation of Π_o and $B_{2,o}$.

Table 3.2 Probabilities of Rank Selection and Lagged Order Selection

	Cointegration Rank Selection					
	r_o=0, λ_o=(0 0)		r_o=1, λ_o=(0 -0.5)		r_o=2, λ_o=(-.6 -.5)	
	n=100	n=400	n=100	n=400	n=100	n=400
$\widehat{r}_n = 0$	0.989	1.000	0.000	0.000	1.000	0.999
$\widehat{r}_n = 1$	0.011	0.000	0.998	1.000	0.000	0.001
$\widehat{r}_n = 2$	0.000	0.000	0.002	0.000	0.000	0.000
	Lagged Difference Selection					
	r_o=0, λ_o=(0 0)		r_o=1, λ_o=(0 -0.5)		r_o=2, λ_o=(-.6 -.5)	
	n=100	n=400	n=100	n=400	n=100	n=400
$\widehat{p}_n \in T$	0.907	0.979	0.980	1.000	0.910	0.979
$\widehat{p}_n \in C$	0.093	0.019	0.020	0.000	0.090	0.021
$\widehat{p}_n \in I$	0.000	0.002	0.000	0.000	0.000	0.000
	Model Selection					
	r_o=0, λ_o=(0 0)		r_o=1, λ_o=(0 -0.5)		r_o=2, λ_o=(-.6 -.5)	
	n=100	n=400	n=100	n=400	n=100	n=400
$\widehat{m}_n \in T$	0.899	0.979	0.978	1.000	0.910	0.978
$\widehat{m}_n \in C$	0.101	0.019	0.021	0.000	0.090	0.021
$\widehat{m}_n \in I$	0.000	0.002	0.000	0.000	0.000	0.001

Table 3.2: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in (6.15) and (3.83). λ_o in each column represents the eigenvalues of Π_o. "T", "C" and "I" denote the true lags/model, consistent lags/model and inconsistent lags/model selected by the shrinkage estimation.

Table 3.1 presents finite sample probabilities of our method in selecting the true rank under different model specifications. Overall speaking, the LS shrinkage method performs very well in selecting the true rank of Π_o. When the sample size is small (i.e. $n = 100$) and the data are *iid*, the probability of selecting the true rank $r_o = 0$ is very close to 1 (around 0.96) and the probabilities of selecting the true ranks $r_o = 1$ and $r_o = 2$ are almost equal to 1. When the sample size is increased to 400, the probabilities of our method selecting the true ranks $r_o = 0$ and $r_o = 1$ are almost equal to 1 and the probability of selecting the true ranks $r_o = 2$ equals 1. Similar

results show up in the scenario when the data are weakly dependent. The only difference is that when the pseudo true eigenvalues are close to zero, the probability of our method in false selecting these small eigenvalues is increased, as illustrated in the weakly dependent case with $r_o = 2$. However, with the sample size growing, the probabilities of our method in selecting the true ranks become close to 1.

Table 3.3 Finite Sample Properties of the Shrinkage Estimates

	Model 1 with $r_o = 0$, $\lambda_o = (0.0\ 0.0)$ and $n = 100$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	-0.001	0.007	0.007	-0.025	0.036	0.044	0.000	0.000	0.000
Π_{12}	0.000	0.005	0.005	0.001	0.041	0.041	0.000	0.000	0.000
Π_{21}	0.000	0.004	0.007	0.000	0.030	0.030	0.000	0.000	0.000
Π_{22}	0.000	0.007	0.007	-0.024	0.035	0.043	0.000	0.000	0.000
	Model 1 with $r_o = 0$, $\lambda_o = (0.0\ 0.0)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	0.000	0.000	0.000	-0.008	0.012	0.015	0.000	0.000	0.000
Π_{12}	0.000	0.000	0.000	0.000	0.010	0.010	0.000	0.000	0.000
Π_{21}	0.000	0.000	0.000	0.000	0.013	0.013	0.000	0.000	0.000
Π_{22}	0.000	0.000	0.000	-0.008	0.012	0.014	0.000	0.000	0.000

Table 3.3: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (3.83). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is simply a 4 by 4 zero matrix.

Tables 3.3, 3.4 and 3.5 provide the finite sample properties of the LS shrinkage estimate, the OLS estimate and the oracle estimate (under the first simulation design) in terms of bias, standard deviation and root of mean square error. When the true rank $r_o = 0$, the unknown parameter Π_o is a zero matrix. In this case, the LS shrinkage estimate clearly dominates the LS estimate due to the high probabilities of the shrinkage method in selecting the true rank. When the true rank $r_o = 1$, we do not observe the efficiency advantage of the LS shrinkage estimator over the LS estimate, but the finite sample bias of the shrinkage estimate is remarkably smaller. From Corollary 3.3.7, we see that the LS shrinkage estimator is free of the high

Table 3.4 Finite Sample Properties of the Shrinkage Estimates

	Model 1 with $r_o = 1$, $\lambda_o = (0.0\ -0.5)$ and $n = 100$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	0.003	0.061	0.061	-0.007	0.055	0.056	-0.005	0.055	0.055
Π_{12}	-0.002	0.031	0.031	-0.007	0.029	0.029	-0.002	0.028	0.028
Π_{21}	0.002	0.062	0.062	-0.004	0.048	0.048	-0.002	0.048	0.048
Π_{22}	-0.001	0.031	0.031	-0.005	0.025	0.025	-0.001	0.024	0.024
	Model 1 with $r_o = 1$, $\lambda_o = (0.0\ -0.5)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	0.001	0.034	0.034	-0.003	0.031	0.031	-0.002	0.031	0.031
Π_{12}	0.000	0.017	0.017	-0.001	0.016	0.016	-0.001	0.015	0.015
Π_{21}	-0.001	0.031	0.031	-0.003	0.028	0.028	-0.001	0.028	0.028
Π_{22}	-0.000	0.016	0.016	-0.002	0.014	0.014	0.000	0.014	0.014
	Model 1 with $r_o = 1$, $\lambda_o = (0.0\ -0.5)$ and $n = 100$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Q_{11}	0.002	0.083	0.083	0.001	0.073	0.073	-0.006	0.071	0.071
Q_{12}	-0.000	0.007	0.007	-0.013	0.024	0.028	0.000	0.003	0.003
Q_{21}	0.001	0.078	0.078	0.001	0.066	0.066	-0.005	0.064	0.064
Q_{22}	-0.000	0.005	0.005	-0.012	0.022	0.025	0.000	0.000	0.000
	Model 1 with $r_o = 1$, $\lambda_o = (0.0\ -0.5)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Q_{11}	0.000	0.042	0.042	0.000	0.041	0.041	-0.002	0.040	0.040
Q_{12}	0.000	0.001	0.001	0.000	0.008	0.009	-0.002	0.001	0.001
Q_{21}	0.000	0.037	0.037	-0.004	0.037	0.036	0.000	0.036	0.036
Q_{22}	0.000	0.000	0.000	-0.004	0.007	0.008	0.000	0.000	0.000

Table 3.4: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (3.83). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case are the RRR estimate with rank restriction r=1.

order bias, which explains its smaller bias in finite samples. Moreover, Lemma 3.3.1 and Corollary 3.3.7 indicate that the OLS estimator and the LS shrinkage estimator (and hence the oracle estimator) have almost the same variance. This explains the phenomena that the LS shrinkage estimate does not look more efficient than the OLS estimate. To better compare the OLS estimate, the LS shrinkage estimate and the oracle estimate, we transfer the three estimates using the matrix Q and its

Table 3.5 Finite Sample Properties of the Shrinkage Estimates

	Model 1 with $r_o = 2$, $\lambda_o = (-0.6, -0.5)$ and $n = 100$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	-0.023	0.090	0.093	-0.010	0.093	0.094	-0.010	0.093	0.094
Π_{12}	0.038	0.091	0.099	-0.001	0.090	0.090	-0.001	0.090	0.090
Π_{21}	-0.025	0.100	0.103	0.002	0.081	0.081	0.002	0.081	0.081
Π_{22}	0.051	0.146	0.154	-0.010	0.078	0.079	-0.010	0.078	0.079
	Model 1 with $r_o = 2$, $\lambda_o = (-0.6, -0.5)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	-0.006	0.052	0.053	-0.003	0.052	0.052	-0.003	0.052	0.052
Π_{12}	0.005	0.055	0.055	0.001	0.051	0.051	0.001	0.051	0.051
Π_{21}	-0.005	0.055	0.055	-0.002	0.046	0.046	-0.002	0.046	0.046
Π_{22}	0.008	0.075	0.075	-0.004	0.044	0.044	-0.004	0.044	0.044

Table 3.5: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (3.83). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is simply the OLS estimate.

inverse (i.e. the estimate $\widehat{\Pi}$ is transferred to be $Q\widehat{\Pi}Q^{-1}$). Note that in this case, $Q\Pi_o Q^{-1} = diag(\text{-}0.5, 0)$. The finite sample properties of the transferred estimates are presented in the last two panels of Table 3.4. We see that the elements in the last column of the transferred LS shrinkage estimator enjoys very small bias and small variance even when the sample size is only 100. The elements in the last column of the OLS estimator, when compared with the elements in its first column, have smaller variance but larger bias. It is clear that with the sample size growing, the LS shrinkage estimator is approaching the oracle estimator in terms of their finite sample properties. When the true rank $r_o = 2$, the LS estimator is better than the shrinkage estimator as the latter suffers from the shrinkage bias in the finite samples. If the shrinkage bias is a concern, one can run a reduced rank regression based on the rank selected by the LS shrinkage estimation to get the so called post-Lasso estimator. The post Lasso estimator also enjoys the oracle properties and it is free of the shrinkage bias in the finite samples.

Table 3.6 Finite Sample Properties of the Shrinkage Estimates

	Model 3 with $r_o = 0$, $\lambda_o = (0.0, 0.0)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	0.000	0.000	0.000	-0.002	0.003	0.004	0.000	0.000	0.000
Π_{21}	0.000	0.000	0.000	0.000	0.003	0.003	0.000	0.000	0.000
Π_{12}	0.000	0.000	0.000	0.000	0.003	0.003	0.000	0.000	0.000
Π_{22}	0.000	0.000	0.000	-0.002	0.003	0.004	0.000	0.000	0.000
$B_{1,11}$	-0.014	0.049	0.051	-0.007	0.054	0.054	-0.004	0.048	0.048
$B_{1,21}$	-0.001	0.037	0.037	-0.001	0.046	0.046	-0.001	0.041	0.041
$B_{1,12}$	-0.001	0.050	0.050	-0.002	0.063	0.063	-0.001	0.057	0.057
$B_{1,22}$	-0.014	0.050	0.052	-0.008	0.054	0.055	-0.005	0.049	0.049
$B_{2,11}$	0.000	0.003	0.003	-0.005	0.058	0.058	0.000	0.000	0.000
$B_{2,21}$	0.000	0.003	0.003	-0.000	0.050	0.050	0.000	0.000	0.000
$B_{2,12}$	0.000	0.003	0.003	0.001	0.066	0.066	0.000	0.000	0.000
$B_{2,22}$	0.000	0.004	0.004	-0.004	0.058	0.058	0.000	0.000	0.000
$B_{3,11}$	-0.016	0.048	0.051	-0.007	0.054	0.054	-0.006	0.047	0.048
$B_{3,21}$	0.000	0.038	0.038	0.000	0.046	0.046	0.000	0.041	0.041
$B_{3,12}$	0.001	0.049	0.049	0.000	0.061	0.061	0.001	0.055	0.055
$B_{3,22}$	-0.016	0.049	0.051	-0.007	0.053	0.054	-0.006	0.048	0.048

Table 3.6: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in equations (3.83) and (3.84). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is simply the OLS estimate with assuming that Π_o and B_{2o} are zero matrics.

Table 3.2 contains the finite sample probabilities of our method in rank selection and lagged order selection in model 3. We see that the shrinkage method performs very well in selecting the true rank and true lagged differences (and thus the true model) in all scenarios. It is interesting to see that the probabilities of selecting the true ranks are not negatively effected either by adding lags to the model or by the lagged order selection simultaneously performed with the rank selection. Table 3.6, 3.7 and 3.8 presents the finite sample properties of the LS shrinkage estimate, the OLS estimate and the Oracle estimate. When compared with the oracle estimates, some components in the LS shrinkage estimate even have smaller variances, though their finite sample bias are slightly larger. As a result, their root of mean square

Table 3.7 Finite Sample Properties of the Shrinkage Estimates

	Model 3 with $r_o = 1$, $\lambda_o = (0.0, -0.5)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	0.001	0.065	0.065	-0.002	0.065	0.065	-0.001	0.065	0.065
Π_{21}	-0.001	0.056	0.056	-0.001	0.056	0.056	0.000	0.056	0.056
Π_{12}	-0.001	0.033	0.033	-0.001	0.033	0.033	0.000	0.032	0.032
Π_{22}	0.000	0.028	0.028	-0.001	0.028	0.028	0.000	0.028	0.028
$B_{1,11}$	-0.048	0.056	0.073	-0.003	0.057	0.057	-0.002	0.053	0.053
$B_{1,21}$	-0.035	0.047	0.058	-0.002	0.049	0.049	-0.002	0.046	0.046
$B_{1,12}$	-0.013	0.072	0.073	-0.002	0.077	0.077	-0.002	0.073	0.073
$B_{1,22}$	-0.011	0.065	0.066	-0.002	0.067	0.067	-0.002	0.063	0.063
$B_{2,11}$	0.000	0.000	0.000	-0.001	0.044	0.044	0.000	0.000	0.000
$B_{2,21}$	0.000	0.000	0.000	-0.001	0.038	0.038	0.000	0.000	0.000
$B_{2,12}$	0.000	0.000	0.000	-0.002	0.079	0.079	0.000	0.000	0.000
$B_{2,22}$	0.000	0.000	0.000	-0.001	0.067	0.067	0.000	0.000	0.000
$B_{3,11}$	-0.057	0.038	0.069	-0.003	0.042	0.043	-0.002	0.038	0.038
$B_{3,21}$	-0.041	0.033	0.053	-0.003	0.037	0.037	-0.002	0.033	0.033
$B_{3,12}$	-0.050	0.049	0.070	-0.002	0.070	0.070	-0.002	0.052	0.052
$B_{3,22}$	-0.037	0.048	0.061	-0.003	0.061	0.061	-0.002	0.046	0.046

Table 3.7: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in equations (3.83) and (3.84). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case refers to the RRR estimate with r=1 and the restriction that $B_{2o} = 0$.

errors are smaller than these of their counterparts in the oracle estimate. Moreover, the LS shrinkage estimate generally has smaller variance when compared with the OLS estimate, though the finite sample bias of the shrinkage estimate of nonzero component is slightly larger. The main intuition behind the phenomenon that the LS shrinkage estimate beats the Oracle estimate relies on the fact that there are some zero components in B_o and shrinking their estimates towards zero (but not exactly equals zero) help to reduce their bias and the variances. From this perspective, the shrinkage estimates of the zero components in B_o share similar features of the traditional shrinkage estimates and the finite sample shrinkage bias is not always harmful.

Table 3.8 Finite Sample Properties of the Shrinkage Estimates

	Model 3 with $r_o = 2$, $\lambda_o = (-0.6, -0.5)$ and $n = 400$								
	Lasso Estimates			OLS			Oracle Estimates		
	Bias	Std	RMSE	Bias	Std	RMSE	Bias	Std	RMSE
Π_{11}	0.015	0.051	0.053	-0.002	0.064	0.064	-0.003	0.051	0.052
Π_{21}	0.004	0.049	0.049	0.001	0.055	0.055	0.000	0.044	0.044
Π_{12}	-0.005	0.044	0.044	0.001	0.049	0.049	0.001	0.041	0.041
Π_{22}	0.006	0.054	0.055	-0.001	0.042	0.042	-0.001	0.035	0.035
$B_{1,11}$	-0.031	0.056	0.064	-0.002	0.064	0.064	-0.000	0.058	0.058
$B_{1,21}$	-0.004	0.052	0.052	-0.002	0.056	0.056	-0.001	0.050	0.051
$B_{1,12}$	0.003	0.050	0.050	-0.002	0.058	0.058	-0.002	0.054	0.054
$B_{1,22}$	-0.018	0.050	0.053	-0.003	0.050	0.050	-0.002	0.047	0.047
$B_{2,11}$	0.000	0.002	0.002	-0.001	0.058	0.058	0.000	0.000	0.000
$B_{2,21}$	0.000	0.003	0.003	-0.001	0.050	0.050	0.000	0.000	0.000
$B_{2,12}$	0.000	0.002	0.002	0.000	0.057	0.057	0.000	0.000	0.000
$B_{2,22}$	0.000	0.003	0.003	0.000	0.050	0.050	0.000	0.000	0.000
$B_{3,11}$	-0.028	0.050	0.057	-0.005	0.055	0.055	-0.004	0.052	0.052
$B_{3,21}$	0.000	0.043	0.043	-0.001	0.048	0.048	-0.000	0.045	0.045
$B_{3,12}$	-0.001	0.052	0.052	0.001	0.056	0.056	0.001	0.056	0.056
$B_{3,22}$	-0.023	0.055	0.059	-0.003	0.048	0.048	-0.003	0.048	0.048

Table 3.8: Replications=5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (3.83) and (3.84). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is simply the OLS estimate with the restriction that $B_{2o} = 0$.

3.8 Conclusion

One of the main challenges in any applied econometric work is the selection of a good model for practical implementation. The conduct of inference and model use in forecasting and policy analysis are inevitably conditioned on the empirical process of model selection, which typically leads to issues of post-model selection inference. Adaptive lasso and bridge estimation methods provide a methodology where these difficulties may be partly attenuated by simultaneous model selection and estimation to facilitate empirical research in complex models like reduced rank regressions where many selection decisions need to be made to construct a satisfactory empirical model. On the other hand, as indicated in the Introduction, the methods certainly do not eliminate post-shrinkage selection inference issues in finite samples because

the estimators carry the effects of the in-built selections.

This chapter shows how to use the methodology of shrinkage in a multivariate system to develop an automated approach to cointegrated system modeling that enables simultaneous estimation of the cointegrating rank and autoregressive order in conjunction with oracle-like efficient estimation of the cointegrating matrix and the transient dynamics. As such the methods offer practical advantages to the empirical researcher by avoiding sequential techniques where cointegrating rank and transient dynamics are estimated prior to model fitting.

Various extensions of the methods developed here are possible. One rather obvious extension is to allow for parametric restrictions on the cointegrating matrix which may relate to theory-induced specifications. Lasso type procedures have so far been confined to parametric models, whereas cointegrated systems are often formulated with some nonparametric elements relating to unknown features of the model. A second extension of the present methodology, therefore, is to semiparametric formulations in which the error process in the VECM is weakly dependent, which is partly considered already in Section 4. The effects of post-shrinkage inference issues also merit detailed investigation. These matters and other generalizations of the framework will be explored in future work.

3.9 Appendix

Lemma 3.9.1 *Under Assumptions 3.3.1 and 3.3.2, we have*

(a) $n^{-1}\sum_{t=1}^{n} Z_{1,t-1}Z'_{1,t-1} \to_p \Sigma_{z_1z_1}$;

(b) $n^{-\frac{3}{2}}\sum_{t=1}^{n} Z_{1,t-1}Z'_{2,t-1} \to_p 0$;

(c) $n^{-2}\sum_{t=1}^{n} Z_{2,t-1}Z'_{2,t-1} \to_d \int B_{w_2}B'_{w_2}$;

(d) $n^{-\frac{1}{2}}\sum_{t=1}^{n} u_t Z'_{1,t-1} \to_d N(0, \Omega_u \otimes \Sigma_{z_1z_1})$;

(e) $n^{-1}\sum_{t=1}^{n} u_t Z'_{2,t-1} \to_d \left(\int B_{w_2} dB'_u\right)'$.

The quantities in (c), (d), and (e) converge jointly.

Proof of Lemma 3.9.1. See Johansen (1995) and Cheng and Phillips (2009). ■

3.9.1 Proof of Main Results in Section 3.3

Proof of Lemma 3.3.1. (a) From (3.11)

$$\begin{aligned}\left(\widehat{\Pi}_{1st} - \Pi_o\right) Q^{-1} D_n^{-1} &= \left(\sum_{t=1}^{n} u_t Y'_{t-1} Q'\right)\left(\sum_{t=1}^{n} Q Y_{t-1} Y'_{t-1} Q'\right)^{-1} D_n^{-1} \\ &= \left(\sum_{t=1}^{n} u_t Z'_{t-1} D_n\right)\left(D_n \sum_{t=1}^{n} Z_{t-1} Z'_{t-1} D_n\right)^{-1}. \quad (3.92)\end{aligned}$$

Result (a) follows directly from Lemma 3.9.1.

(b) Denote $P = [\beta_o, \beta_{o\perp}]$ and $S_n(\phi) = \phi I_m - \widehat{\Pi}_{1st}$. Then, by definition, the elements of $\phi(\widehat{\Pi}_{1st})$ are the solutions of the determinantal equation,

$$0 = \left|\phi I_m - \widehat{\Pi}_{1st}\right| = |P' S_n(\phi) P| = \begin{vmatrix} \phi\beta'_o\beta_o - \beta'_o\widehat{\Pi}_{1st}\beta_o & -\beta'_o\widehat{\Pi}_{1st}\beta_{o\perp} \\ -\beta'_{o\perp}\widehat{\Pi}_{1st}\beta_o & \phi I_{m-r_o} - \beta'_{o\perp}\widehat{\Pi}_{1st}\beta_{o\perp} \end{vmatrix}. \quad (3.93)$$

Using (a) we deduce that

$$\beta'_o\widehat{\Pi}_{1st}\beta_{o\perp} = \beta'_o\left(\widehat{\Pi}_{1st} - \Pi_o\right)\beta_{o\perp} = o_p(1), \quad (3.94)$$

$$\beta'_{o\perp}\widehat{\Pi}_{1st}\beta_{o\perp} = \beta'_{o\perp}\left(\widehat{\Pi}_{1st} - \Pi_o\right)\beta_{o\perp} = o_p(1), \quad (3.95)$$

and, similarly,

$$\beta'_{o\perp}\widehat{\Pi}_{1st}\beta_o \to_p \beta'_{o\perp}\Pi_o\beta_o \text{ and } \beta'_o\widehat{\Pi}_{1st}\beta_o \to_p \beta'_o\Pi_o\beta_o. \quad (3.96)$$

Using (3.93)-(3.96), we deduce that

$$\left|\phi I_m - \widehat{\Pi}_{1st}\right| \to_p |\phi I_{m-r_o}| \times |\phi\beta_o'\beta_o - \beta_o'\Pi_o\beta_o|\,, \tag{3.97}$$

uniformly over any compact set in R^m. By Assumption 3.3.2.(i), $\phi(\Pi_o) \in \overline{U}_1 \equiv \{\phi \in R^m,\ \|\phi\| \leq 1\}$ and $\overline{U}_1$ is a compact set in R^m. Thus, by continuous mapping, we have $\phi_k(\widehat{\Pi}_{1st}) \to_p 0$ for $k = r_o + 1, ..., m$ and

$$\left(\phi_1(\widehat{\Pi}_{1st}), ..., \phi_{r_o}(\widehat{\Pi}_{1st})\right) \to_p \phi_{\mathcal{S}_\phi}(\Pi_o), \tag{3.98}$$

where $\phi_{\mathcal{S}_\phi}(\Pi_o)$ denotes the ordered solutions of the equation $|\phi\beta_o'\beta_o - \beta_o'\Pi_o\beta_o| = 0$. The determinantal equation $|\phi\beta_o'\beta_o - \beta_o'\Pi_o\beta_o| = 0$ is equivalent to $|\phi I_{r_o} - \beta_o'\alpha_o| = 0$, so result (b) follows.

(c) Using the notation from (b), we have

$$|S_n(\phi)| = |\beta_o'S_n(\phi)\beta_o| \times \left|\beta_{o\perp}'\left\{S_n(\phi) - S_n(\phi)\beta_o\left[\beta_o'S_n(\phi)\beta_o\right]^{-1}\beta_o'S_n(\phi)\right\}\beta_{o\perp}\right|. \tag{3.99}$$

Let $\mu_k^* = n\phi_k(\widehat{\Pi}_{1st})$ $(k = r_o + 1, ..., m)$, so that μ_k^* is by definition a solution of the equation

$$0 = |\beta_o'S_n(\mu)\beta_o| \times \left|\beta_{o\perp}'\left\{S_n(\mu) - S_n(\mu)\beta_o\left[\beta_o'S_n(\mu)\beta_o\right]^{-1}\beta_o'S_n(\mu)\right\}\beta_{o\perp}\right|, \tag{3.100}$$

where $S_n(\mu) = \frac{\mu}{n}I_m - \widehat{\Pi}_{1st}$.

For any compact subset $K \subset R$, we can invoke the results in (a) to show

$$\beta_o'S_n(\mu)\beta_o = \frac{\mu}{n}\beta_o'\beta_o - \beta_o'\left(\widehat{\Pi}_{1st} - \Pi_o\right)\beta_o + \beta_o'\Pi_o\beta_o \to_p \beta_o'\Pi_o\beta_o, \tag{3.101}$$

uniformly over K. From Assumption 3.3.2.(iii), we have

$$|\beta_o'\Pi_o\beta_o| = |\beta_o'\alpha_o\beta_o'\beta_o| = |\beta_o'\alpha_o| \times |\beta_o'\beta_o| \neq 0.$$

Thus, the normalized $m - r_o$ smallest eigenvalues μ_k^* $(k = r_o + 1, ..., m)$ of $\widehat{\Pi}_{1st}$ are asymptotically the solutions of the following determinantal equation,

$$0 = \left|\beta_{o\perp}' \left\{S_n(\mu) - S_n(\mu)\beta_o \left[\beta_o' S_n(\mu)\beta_o\right]^{-1} \beta_o' S_n(\mu)\right\} \beta_{o\perp}\right|, \tag{3.102}$$

where

$$\begin{aligned}
\beta_o' S_n(\mu)\beta_{o\perp} &= \beta_o' \left(\widehat{\Pi}_{1st} - \Pi_o\right) \beta_{o\perp}, && (3.103)\\
\beta_{o\perp}' S_n(\mu)\beta_{o\perp} &= \frac{\mu}{n} I_{m-r_o} - \beta_{o\perp}' \left(\widehat{\Pi}_{1st} - \Pi_o\right) \beta_{o\perp}, && (3.104)\\
\beta_{o\perp}' S_n(\mu)\beta_o &= \beta_{o\perp}' \widehat{\Pi}_{1st}\beta_o \to_p \beta_{o\perp}'\alpha_o\beta_o'\beta_o. && (3.105)
\end{aligned}$$

Using the results in (3.101) and (3.103)-(3.105), we get

$$\begin{aligned}
& \beta_{o\perp}' \left\{S_n(\mu) - S_n(\mu)\beta_o \left[\beta_o' S_n(\mu)\beta_o\right]^{-1} \beta_o' S_n(\mu)\right\} \beta_{o\perp} \\
= \; & \frac{\mu}{n} I_{m-r_o} - \beta_{o\perp}' \left[I_m - \alpha_o \left(\beta_o'\alpha_o\right)^{-1} \beta_o' + o_p(1)\right] \left(\widehat{\Pi}_{1st} - \Pi_o\right) \beta_{o\perp}. \quad (3.106)
\end{aligned}$$

Note that

$$\beta_{o\perp}' \left[I_{m-r_o} - \alpha_o \left(\beta_o'\alpha_o\right)^{-1} \beta_o'\right] Q^{-1} = \left[\mathbf{0}_{(m-r_o)\times r_o}, (\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right] \equiv H_1, \tag{3.107}$$

and

$$Q\beta_{o\perp} = [\beta_o, \alpha_{o\perp}]' \beta_{o\perp} = \left[\mathbf{0}_{(m-r_o)\times r_o}, \beta_{o\perp}'\alpha_{o\perp}\right]' \equiv H_2'. \tag{3.108}$$

Using (3.92), we deduce that

$$\begin{aligned} & nH_1Q\left(\widehat{\Pi}_{1st}-\Pi_o\right)Q^{-1}H_2' \\ = & \left(H_1\sum_{t=1}^{n}w_tZ_{t-1}'D_n^{-1}\right)\left(D_n^{-1}\sum_{t=1}^{n}Z_tZ_t'D_n^{-1}\right)^{-1}H_2' \\ \to_d & \ (\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\left(\int B_{w_2}dB_{w_2}'\right)'\left(\int B_{w_2}B_{w_2}'\right)^{-1}(\alpha_{o,\perp}'\beta_{o,\perp}). \end{aligned} \tag{3.109}$$

Then, from (3.102)-(3.109), we obtain

$$\begin{aligned} & \left|n\beta_{o\perp}'\left\{S_n(\mu)-S_n(\mu)\beta_o\left[\beta_o'S_n(\mu)\beta_o\right]^{-1}\beta_o'S_n(\mu)\right\}\beta_{o\perp}\right| \\ \to_d & \left|\mu I_{m-r_o}-\left(\int B_{w_2}dB_{w_2}'\right)'\left(\int B_{w_2}B_{w_2}'\right)^{-1}\right|, \end{aligned} \tag{3.110}$$

uniformly over K. The result in (c) follows from (3.110) and by continuous mapping.

■

Proof of Theorem 3.3.2. Define

$$V_n(\Pi)=\sum_{t=1}^{n}\left\|\Delta Y_t-\Pi Y_{t-1}\right\|^2+n\sum\nolimits_{k=1}^{m}\lambda_{r,k,n}\left\|\Phi_{n,k}(\Pi)\right\|/||\phi_k(\widehat{\Pi}_{1st})||^{\omega}.$$

We can write

$$\sum_{t=1}^{n}\left\|\Delta Y_t-\Pi Y_{t-1}\right\|^2=\left[\Delta y-\left(Y_{-1}'\otimes I_m\right)vec(\Pi)\right]'\left[\Delta y-\left(Y_{-1}'\otimes I_m\right)vec(\Pi)\right]$$

where $\Delta y=vec\left(\Delta Y\right)$, $\Delta Y=\left(\Delta Y_1,...,\Delta Y_n\right)_{m\times n}$ and $Y_{-1}=\left(Y_0,...,Y_{T-1}\right)_{m\times n}$.

By definition, $V_n(\widehat{\Pi}_n) \leq V_n(\Pi_{n,f})$ and thus

$$\begin{aligned}
& vec(\Pi_{n,f} - \widehat{\Pi}_n)' \left(\sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m \right) vec(\Pi_{n,f} - \widehat{\Pi}_n) \\
& +2vec(\Pi_{n,f} - \widehat{\Pi}_n)' vec \left(\sum_{t=1}^{n} Y_{t-1} u_t' \right) \\
& +2vec(\Pi_{n,f} - \widehat{\Pi}_n)' \left(\sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m \right) vec(\Pi_o - \Pi_{n,f}) \\
\leq & \sum_{k=1}^{m} n\lambda_{r,k,n} \left[||\Phi_{n,k}(\Pi_{n,f})|| - ||\Phi_{n,k}(\widehat{\Pi}_n)|| \right] / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega}. \qquad (3.111)
\end{aligned}$$

When $r_o = 0$, ΔY_t is stationary and Y_t is full rank $I(1)$, so that

$$n^{-2} \sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \rightarrow_d \int_0^1 B_u(a) B_u'(a) da \text{ and } n^{-2} \sum_{t=1}^{n} Y_{t-1} u_t' = O_p(n^{-1}). \qquad (3.112)$$

From the results in (3.111) and (3.112), we get

$$\mu_{n,\min} ||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2(c_{1,n} + c_{2,n}) ||\widehat{\Pi}_n - \Pi_{n,f}|| - d_n \leq 0, \qquad (3.113)$$

where $\mu_{n,\min}$ denotes the smallest eigenvalue of $n^{-2} \sum_{t=1}^{n} Y_{t-1} Y_{t-1}'$, which is positive w.p.a.1,

$$\begin{aligned}
c_{1,n} &= ||n^{-2} \sum_{t=1}^{n} Y_{t-1} u_t'||, \; c_{2,n} = m \left\| n^{-2} \sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \right\| ||\Pi_{n,f} - \Pi_o|| \\
\text{and } d_n &= n^{-1} \sum_{k=1}^{m} \lambda_{r,k,n} ||\Phi_{n,k}(\Pi_{n,f})|| / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega}. \qquad (3.114)
\end{aligned}$$

Under (3.19) and (3.112), $c_{1,n} = o_p(1)$ and $c_{2,n} = o_p(1)$. Under (3.17), (3.18), Lemma 3.3.1.(b) and $\lambda_{r,k,n} = o(1)$,

$$d_n = n^{-1} \sum_{k=1}^{r_o} \lambda_{r,k,n} ||\Phi_{n,k}(\Pi_{n,f})|| / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega} = o_p(n^{-1}). \qquad (3.115)$$

From (3.113), (3.114) and (3.115), it is straightforward to deduce that $||\widehat{\Pi}_n - \Pi_{n,f}|| =$

$o_p(1)$. The consistency of $\widehat{\Pi}_n$ follows from the triangle inequality and the consistency of $\Pi_{n,f}$.

When $r_o = m$, Y_t is stationary and we have

$$n^{-1}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \to_p \Sigma_{yy} = R(1)\Omega_u R(1)' \text{ and } n^{-1}\sum_{t=1}^{n} Y_{t-1}u_t' = O_p(n^{-\frac{1}{2}}). \quad (3.116)$$

From the results in (3.111) and (3.116), we get

$$\mu_{n,\min}||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2n(c_{1,n} + c_{2,n})||\widehat{\Pi}_n - \Pi_{n,f}|| - nd_n \leq 0 \quad (3.117)$$

where $\mu_{n,\min}$ denotes the smallest eigenvalue of $n^{-1}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'$, which is positive w.p.a.1, $c_{1,n}$, $c_{2,n}$ and d_n are defined in (3.115). It is clear that $nc_{1,n} = o_p(1)$ and $nc_{2,n} = o_p(1)$ under (3.116) and (3.19), and $nd_n = o_p(1)$ under (3.115). So, consistency of $\widehat{\Pi}_n$ follows directly from the inequality in (3.117), triangle inequality and the consistency of $\Pi_{n,f}$.

Denote $B_n = (D_n Q)^{-1}$, then when $0 < r_o < m$, we can use the results in Lemma 3.9.1 to deduce that

$$\begin{aligned}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' &= Q^{-1}D_n^{-1}D_n\sum_{t=1}^{n} Z_{t-1}Z_{t-1}'D_nD_n^{-1}Q'^{-1} \\ &= B_n\left[\begin{pmatrix}\Sigma_{z_1z_1} & 0\\ 0 & \int B_{w_2}B_{w_2}'\end{pmatrix} + o_p(1)\right]B_n',\end{aligned}$$

and thus

$$vec(\Pi_{n,f} - \widehat{\Pi}_n)'\left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right)vec(\Pi_{n,f} - \widehat{\Pi}_n) \geq \mu_{n,\min}||(\widehat{\Pi}_n - \Pi_{n,f})B_n||^2, \quad (3.118)$$

where $\mu_{n,\min}$ is the smallest eigenvalue of $D_n\sum_{t=1}^{n} Z_{t-1}Z_{t-1}'D_n$ and is positive w.p.a.1.

Next observe that

$$\left|\left[vec(\Pi_{n,f}-\widehat{\Pi}_n)\right]' vec\left(B_nD_n\sum_{t=1}^{n}Z_{t-1}u_t'\right)\right| \leq ||(\widehat{\Pi}_n-\Pi_{n,f})B_n||e_{1,n} \tag{3.119}$$

and

$$\left|vec(\Pi_{n,f}-\widehat{\Pi}_n)'\left(\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\otimes I_m\right)vec(\Pi_o-\Pi_{n,f})\right| \leq ||(\widehat{\Pi}_n-\Pi_{n,f})B_n||e_{2,n} \tag{3.120}$$

where

$$e_{1,n}=||D_n\sum\nolimits_{t=1}^{n}Z_{t-1}u_t'|| \text{ and } e_{2,n}=m||D_n\sum\nolimits_{t=1}^{n}Z_{t-1}Z_{t-1}'D_n||\times||(\Pi_{n,f}-\Pi_o)B_n||. \tag{3.121}$$

Under Lemma 3.9.1 and (3.19), $e_{1,n}=O_p(1)$ and $e_{2,n}=O_p(1)$. From (3.111), (3.118), (3.119), (3.120), we have the inequality

$$\mu_{n,\min}||(\widehat{\Pi}_n-\Pi_{n,f})B_n||^2-2(e_{1,n}+e_{2,n})||(\widehat{\Pi}_n-\Pi_{n,f})B_n||-nd_n\leq 0, \tag{3.122}$$

which implies

$$(\widehat{\Pi}_n-\Pi_{n,f})B_n=O_p(1+\sqrt{n}d_n^{\frac{1}{2}}). \tag{3.123}$$

By the definition of B_n, (3.19) and (3.123), we deduce that

$$\widehat{\Pi}_n-\Pi_o=O_p(n^{-\frac{1}{2}}+d_n^{\frac{1}{2}})=o_p(1),$$

which implies the consistency of $\widehat{\Pi}_n$. ■

Proof of Theorem 3.3.3. By the triangle inequality and (3.18), we have

$$\begin{aligned} & \left| \sum_{k=1}^{m} \lambda_{r,k,n} \frac{||\Phi_{n,k}(\Pi_{n,f})|| - ||\Phi_{n,k}(\widehat{\Pi}_n)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \right| \\ = & \left| \sum_{k=1}^{r_o} \lambda_{r,k,n} \frac{||\Phi_{n,k}(\Pi_{n,f})|| - ||\Phi_{n,k}(\widehat{\Pi}_n)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \right| \\ \leq & r_o \max_{k \in \mathcal{S}_\phi} \left\{ \lambda_{r,k,n} ||\phi_k(\widehat{\Pi}_{1st})||^{-\omega} \right\} ||\widehat{\Pi}_n - \Pi_{n,f}||. \end{aligned} \tag{3.124}$$

Using (3.124) and invoking the inequality in (3.111) we get

$$\begin{aligned} & vec(\Pi_{n,f} - \widehat{\Pi}_n)' \left(\sum\nolimits_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m \right) vec(\Pi_{n,f} - \widehat{\Pi}_n) \\ & + 2vec(\Pi_{n,f} - \widehat{\Pi}_n)' vec \left(\sum\nolimits_{t=1}^{n} Y_{t-1} u_t' \right) \\ & + 2vec(\Pi_{n,f} - \widehat{\Pi}_n)' \left(\sum\nolimits_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m \right) vec(\Pi_o - \Pi_{n,f}) \\ \leq & nr_o \delta_{r,n} ||\widehat{\Pi}_n - \Pi_{n,f}||. \end{aligned} \tag{3.125}$$

When $r_o = 0$, we use (3.114) and (3.125) to obtain

$$\mu_{n,\min} ||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2(c_{1,n} + c_{2,n} + n^{-1} r_o \delta_{r,n}) ||\widehat{\Pi}_n - \Pi_{n,f}|| \leq 0 \tag{3.126}$$

where under (3.112) $c_{1,n} = O_p(n^{-1})$ and $c_{2,n} = O_p(n^{-1})$. We deduce from the inequality (3.126) and (3.19) that

$$\widehat{\Pi}_n - \Pi_o = O_p(n^{-1} + n^{-1} \delta_{r,n}). \tag{3.127}$$

When $r_o = m$, we use (3.125) to obtain

$$\mu_{n,\min} ||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2n(c_{1,n} + c_{2,n} + n^{-1} r_o \delta_{r,n}) ||\widehat{\Pi}_n - \Pi_{n,f}|| \leq 0 \tag{3.128}$$

where $nc_{1,n} = ||\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}u_t'|| = O_p(n^{-\frac{1}{2}})$ and $nc_{2,n} = O_p(n^{-\frac{1}{2}})$ by Lemma 3.9.1 and (3.19). The inequality (3.128) and (3.19) lead to

$$\widehat{\Pi}_n - \Pi_o = O_p(n^{-\frac{1}{2}} + \delta_{r,n}). \tag{3.129}$$

When $0 < r_o < m$, we can use the results in (3.118), (3.119), (3.120), (3.121) and (3.125) to deduce that

$$\mu_{n,\min}||(\Pi_{n,f} - \widehat{\Pi}_n)B_n||^2 - 2(e_{1,n}+e_{2,n})||(\Pi_{n,f} - \widehat{\Pi}_n)B_n|| \leq r_o n\delta_{r,n}||\Pi_{n,f} - \widehat{\Pi}_n|| \tag{3.130}$$

where $e_{1,n} = \|D_n Q\sum_{t=1}^{n} Y_{t-1}u_t'\| = O_p(1)$ and $e_{2,n} = O_p(1)$ by Lemma 3.9.1 and (3.19). By the definition of B_n,

$$||(\Pi_{n,f} - \widehat{\Pi}_n)B_n B_n^{-1}|| \leq cn^{-\frac{1}{2}}||(\Pi_{n,f} - \widehat{\Pi}_n)B_n|| \tag{3.131}$$

where c is some finite positive constant. Using (3.130), (3.131) and (3.19), we get

$$(\widehat{\Pi}_n - \Pi_o)B_n = O_p(1 + n^{\frac{1}{2}}\delta_{r,n}) \tag{3.132}$$

which finishes the proof. ■

Proof of Theorem 3.3.4. To facilitate the proof, we rewrite the LS shrinkage estimation problem as

$$\widehat{T}_n = \underset{T\in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - P_n T Y_{t-1}\|^2 + n\sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \|\Phi_{n,k}(P_n T)\| / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega}. \tag{3.133}$$

By definition, $\widehat{\Pi}_n = P_n\widehat{T}_n$ and $\widehat{T}_n = Q_n\widehat{\Pi}_n$ for all n. Under (3.20) and (3.21),

$$\widehat{T}_n = \begin{pmatrix} Q_{\alpha,n}\widehat{\Pi}_n \\ Q_{\alpha_\perp,n}\widehat{\Pi}_n \end{pmatrix} = \begin{pmatrix} Q_{\alpha,n}\widehat{\Pi}_{1st} \\ Q_{\alpha_\perp,n}\widehat{\Pi}_{1st} \end{pmatrix} + o_p(1). \tag{3.134}$$

Results in (3.22) follows if we can show that the last $m - r_o$ rows of $\widehat{T}_n$ are estimated as zeros w.p.a.1.

By definition, $\Phi_{n,k}(P_nT) = Q_n(k)P_nT = T(k)$ and the problem in (3.133) can be rewritten as

$$\widehat{T}_n = \underset{T\in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - P_nTY_{t-1}\|^2 + n\sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \|T(k)\| / ||\phi_k(\widehat{\Pi}_{1st})||^\omega, \tag{3.135}$$

which has the following Karush-Kuhn-Tucker (KKT) optimality conditions

$$\begin{aligned} \frac{1}{n}\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_nY_{t-1})'P_n(k)Y'_{t-1} &= \frac{\lambda_{r,k,n}}{2||\phi_k(\widehat{\Pi}_{1st})||^\omega}\frac{\widehat{T}_n(k)}{||\widehat{T}_n(k)||} \text{ if } \widehat{T}_n(k) \neq 0, \\ \left\|\frac{1}{n}\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_nY_{t-1})'P_n(k)Y'_{t-1}\right\| &\leq \frac{\lambda_{r,k,n}}{2||\phi_k(\widehat{\Pi}_{1st})||^\omega} \text{ if } \widehat{T}_n(k) = 0, \end{aligned} \tag{3.136}$$

for $k = 1, ..., m$. Conditional on the event $\{Q_n(k_o)\widehat{\Pi}_n \neq 0\}$ for some k_o satisfying $r_o < k_o \leq m$, we obtain the following equation from the KKT optimality conditions

$$\left\|\frac{1}{n}\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_nY_{t-1})'P_n(k_o)Y'_{t-1}\right\| = \frac{\lambda_{r,k_o,n}}{2||\phi_{k_o}(\widehat{\Pi}_{1st})||^\omega}. \tag{3.137}$$

The sample average in the left hand side of (3.137) can be rewritten as

$$\begin{aligned} & \frac{1}{n}\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_n Y_{t-1})' P_n(k_o) Y'_{t-1} \\ = \; & \frac{1}{n}\sum_{t=1}^{n}[\Delta u_t - (\widehat{\Pi}_n - \Pi_o) Y_{t-1}]' P_n(k_o) Y'_{t-1} \\ = \; & \frac{P'_n(k_o)\sum_{t=1}^{n}\Delta u_t Y'_{t-1}}{n} - \frac{P'_n(k_o)(\widehat{\Pi}_n - \Pi_o)\sum_{t=1}^{n} Y_{t-1}Y'_{t-1}}{n}. \end{aligned} \tag{3.138}$$

Under Lemma 3.3.1, Lemma 3.9.1 and Theorem 3.3.3

$$\frac{P'_n(k_o)\sum_{t=1}^{n}\Delta u_t Y'_{t-1}}{n} = O_p(1) \tag{3.139}$$

and

$$\begin{aligned} & \frac{P'_n(k_o)(\widehat{\Pi}_n - \Pi_o)\sum_{t=1}^{n} Y_{t-1}Y'_{t-1}}{n} \\ = \; & P'_n(k_o)(\widehat{\Pi}_n - \Pi_o) Q^{-1} D_n^{-1} \frac{D_n \sum_{t=1}^{n} Z_{t-1}Z'_{t-1}}{n} Q'^{-1} = O_p(1). \end{aligned} \tag{3.140}$$

Using the results in (3.138), (3.139) and (3.140), we deduce that

$$\left\| \frac{1}{n}\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_n Y_{t-1})' P_n(k_o) Y'_{t-1} \right\| = O_p(1). \tag{3.141}$$

While under Lemma 3.3.1.(c) and $n^{\omega}\lambda_{r,k_o,n} \to \infty$

$$\frac{\lambda_{r,k_o,n}}{2||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} = \frac{n^{\omega}\lambda_{r,k_o,n}}{2||n\Phi_{n,k_o}(\widehat{\Pi}_{1st})||^{\omega}} \to_p \infty. \tag{3.142}$$

Combining the results in (3.137), (3.141) and (3.142), we deduce that

$$\Pr\Big(Q_n(k_o)\widehat{\Pi}_n = 0\Big) \to 1 \text{ as } n \to \infty.$$

As the above result holds for any k_o such that $r_o < k_o \leq m$, this finishes the proof.

■

Proof of Theorem 3.3.6. From Corollary 3.3.5, for large enough n the shrinkage estimator $\widehat{\Pi}_n$ can be decomposed as $\widehat{\alpha}_n\widehat{\beta}_n'$ w.p.a.1, where $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ are some $m \times r_o$ matrices. Without loss of generality, we assume the first r_o columns of Π_o are linearly independent. To ensure identification, we normalize β_o as $\beta_o = [I_{r_o}, O_{r_o}]'$ where O_{r_o} is some $r_o \times (m - r_o)$ matrix such that

$$\Pi_o = \alpha_o\beta_o' = [\alpha_o, \alpha_o O_{r_o}]. \tag{3.143}$$

Hence α_o is the first r_o columns of Π_o which is an $m \times r_o$ matrix with full rank and O_{r_o} is uniquely determined by the equation $\alpha_o O_{r_o} = \Pi_{o,2}$, where $\Pi_{o,2}$ denotes the last $m - r_o$ columns of Π_o. Correspondingly, for large enough n we can normalize $\widehat{\beta}_n$ as $\widehat{\beta}_n = [I_{r_o}, \widehat{O}_n]'$ where $\widehat{O}_n$ is some $r_o \times (m - r_o)$ matrix.

From Theorem 3.3.3 and $n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$, we have

$$O_p(1) = \left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n = \left(\widehat{\Pi}_n - \Pi_o\right) \left[\sqrt{n}\alpha_o(\beta_o'\alpha_o)^{-1}, n\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right]. \tag{3.144}$$

From (3.144), we can deduce that

$$n\widehat{\alpha}_n \left(\widehat{\beta}_n - \beta_o\right)' \beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1} = O_p(1),$$

which implies

$$n\left(\widehat{O}_n - O_o\right) = O_p(1). \tag{3.145}$$

Similarly, we have

$$\sqrt{n}\left[\left(\widehat{\alpha}_n-\alpha_o\right)\widehat{\beta}_n'\alpha_o(\beta_o'\alpha_o)^{-1}-\alpha_o\left(\widehat{\beta}_n-\beta_o\right)'\alpha_o(\beta_o'\alpha_o)^{-1}\right]=O_p(1),$$

which combined with (3.145) implies

$$\sqrt{n}\left(\widehat{\alpha}_n-\alpha_o\right)=O_p(1). \tag{3.146}$$

From Corollary 3.3.5, we can deduce that $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ minimize the following criterion function w.p.a.1

$$V_n(\alpha,\beta)=\sum_{t=1}^{n}\left\|\Delta Y_t-\alpha\beta' Y_{t-1}\right\|^2+n\sum_{k=1}^{r_o}\lambda_{r,k,n}\frac{||\Phi_{n,k}(\alpha\beta')||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}. \tag{3.147}$$

Define $U_{1,n}^*=\sqrt{n}\left(\widehat{\alpha}_n-\alpha_o\right)$ and $U_{3,n}^*=n\left(\widehat{\beta}_n-\beta_o\right)'=\left[\mathbf{0}_{r_o},n\left(\widehat{O}_n-O_o\right)\right]\equiv\left[\mathbf{0}_{r_o},U_{2,n}^*\right]$, then

$$\begin{aligned}\left(\widehat{\Pi}_n-\Pi_o\right)Q^{-1}D_n^{-1} &= \left[\widehat{\alpha}_n\left(\widehat{\beta}_n-\beta_o\right)'+\left(\widehat{\alpha}_n-\alpha_o\right)\beta_o'\right]Q^{-1}D_n^{-1}\\ &= \left[n^{-\frac{1}{2}}\widehat{\alpha}_nU_{3,n}^*\alpha_o(\beta_o'\alpha_o)^{-1}+U_{1,n}^*,\ \widehat{\alpha}_nU_{3,n}^*\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right].\end{aligned}$$

Define

$$\Pi_n(U)=\left[n^{-\frac{1}{2}}\widehat{\alpha}_nU_3\alpha_o(\beta_o'\alpha_o)^{-1}+U_1,\widehat{\alpha}_nU_3\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right],$$

where $U_3=[\mathbf{0}_{r_o},U_2]$. Then by definition, $U_n^*=\left(U_{1,n}^*,U_{2,n}^*\right)$ minimizes the following criterion function w.p.a.1

$$\begin{aligned}V_n(U) &= \sum_{t=1}^{n}\left(\left\|\Delta Y_t-\Pi_oY_{t-1}-\Pi_n(U)D_nZ_{t-1}\right\|^2-\left\|\Delta Y_t-\Pi_oY_{t-1}\right\|^2\right)\\ &\quad+n\sum_{k=1}^{r_o}\lambda_{r,k,n}\frac{||\Phi_{n,k}(\Pi_n(U)D_nQ+\Pi_o)||-||\Phi_{n,k}(\Pi_o)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}.\end{aligned}$$

For any compact set $K \in R^{m\times r_o} \times R^{r_o\times(m-r_o)}$ and any $U \in K$, we have

$$\Pi_n(U)D_nQ = O_p(n^{-\frac{1}{2}}).$$

Hence, from the triangle inequality, we can deduce that for all $k \in \mathcal{S}_\phi$

$$\begin{aligned} & n\left|\lambda_{r,k,n}\frac{||\Phi_{n,k}(\Pi_n(U)D_nQ+\Pi_o)||-||\Phi_{n,k}(\Pi_o)||}{||\phi_k(\widehat{\Pi}_{1st})||^\omega}\right| \\ \leq\ & n\lambda_{r,k,n}\frac{||\Phi_{n,k}(\Pi_n(U)D_nQ)||}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} = O_p(n^{\frac{1}{2}}\lambda_{r,k,n}) = o_p(1), \end{aligned} \tag{3.148}$$

uniformly over $U \in K$.

From (3.146),

$$\Pi_n(U) \to_p \left[U_1, \alpha_o U_3 \beta_{o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1}\right] \equiv \Pi_\infty(U) \tag{3.149}$$

uniformly over $U \in K$. By Lemma 3.9.1 and (3.149), we deduce that

$$\begin{aligned} & \sum_{t=1}^n \left(\|\Delta Y_t - \Pi_o Y_{t-1} - \Pi_n(U)D_nZ_{t-1}\|_E^2 - \|\Delta Y_t - \Pi_o Y_{t-1}\|_E^2\right) \\ =\ & vec\left[\Pi_n(U)\right]'\left(D_n\sum_{t=1}^n Z_{t-1}Z'_{t-1}D_n \otimes I_m\right)vec\left[\Pi_n(U)\right] \\ & -2vec\left[\Pi_n(U)\right]'vec\left(\sum_{t=1}^n u_t Z'_{t-1}D_n\right) \\ \to_d\ & vec\left[\Pi_\infty(U)\right]'\left[\begin{pmatrix}\Sigma_{z_1z_1} & 0\\ 0 & \int B_{w_2}B'_{w_2}\end{pmatrix}\otimes I_m\right]vec\left[\Pi_\infty(U)\right] \\ & -2vec\left[\Pi_\infty(U)\right]'vec\left[(V_{1,m}, V_{2,m})\right] \equiv V(U) \end{aligned} \tag{3.150}$$

uniformly over $U \in K$, where $V_{1,m} \equiv N(0, \Omega_u \otimes \Sigma_{z_1z_1})$ and $V_{2,m} \equiv \left(\int B_{w_2}dB'_u\right)'$.

Let $\beta_{o,\perp} = (\beta'_{1,o,\perp}, \beta'_{2,o,\perp})'$ where $\beta_{1,o,\perp}$ is a $r_o \times (m-r_o)$ matrix and $\beta_{2,o,\perp}$ is a

$(m - r_o) \times (m - r_o)$ matrix. Then by definition

$$\beta'_{1,o,\perp} + \beta'_{2,o,\perp} O_{r_o} = 0 \text{ and } \beta'_{1,o,\perp}\beta_{1,o,\perp} + \beta'_{2,o,\perp}\beta_{2,o,\perp} = 0$$

which implies that

$$\beta'_{1,o,\perp} = -\beta'_{2,o,\perp} O_{r_o} \text{ and } \beta_{2,o,\perp} = (I_{r_o} + O_{r_o} O'_{r_o})^{-\frac{1}{2}}.$$

By definition $\Pi_\infty(U) = \left[U_1, \alpha_o U_2 \beta_{2,o,\perp} (\alpha'_{o,\perp}\beta_{o,\perp})^{-1}\right]$, thus

$$vec\left[\Pi_\infty(U)\right] = \left[vec(U_1)', vec(\alpha_o U_2 \beta_{2,o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1})'\right]'$$

and

$$vec(\alpha_o U_2 \beta_{2,o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1}) = \left[(\beta'_{o,\perp}\alpha_{o,\perp})^{-1}\beta'_{2,o,\perp} \otimes \alpha_o\right] vec(U_2).$$

Using above expression, we can rewrite $V(U)$ as

$$\begin{aligned} V(U) \;=\; & vec(U_1)' \left[\Sigma_{z_1 z_1} \otimes I_m\right] vec(U_1) \\ & + vec(U_2)' \left[\beta_{2,o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1} \int B_{w_2} B'_{w_2} (\beta'_{o,\perp}\alpha_{o,\perp})^{-1}\beta'_{2,o,\perp} \otimes \alpha'_o\alpha_o\right] vec(U_2) \\ & - 2vec(U_1)' vec\left(V_{1,m}\right) - 2vec(U_2)' vec\left[\alpha'_o V_{2,m}(\beta'_{o,\perp}\alpha_{o,\perp})^{-1}\beta'_{2,o,\perp}\right]. \end{aligned} \tag{3.151}$$

The expression in (3.151) makes it clear that $V(U)$ is uniquely minimized at

$$\left[U_1^*,\; U_2^*(\alpha'_{o,\perp}\beta_{o,\perp})\beta_{2,o,\perp}^{-1}\right]$$

where

$$U_1^* = B_{m,1} \text{ and } U_2^* = (\alpha'_o\alpha_o)^{-1}\alpha'_o B_{m,2}. \tag{3.152}$$

From (3.145) and (3.146), we can see that U_n^* is asymptotically tight. Invoking the Argmax Continuous Mapping Theorem (ACMT), we can deduce that

$$U_n^* = (U_{1,n}^*, U_{2,n}^*) \to_d \left[U_1^*,\ U_2^*(\alpha_{o,\perp}'\beta_{o,\perp})\beta_{2,o,\perp}^{-1}\right]$$

which together with (3.149) and CMT implies that

$$\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} B_{m,1} & \alpha_o(\alpha_o'\alpha_o)^{-1}\alpha_o' B_{m,2} \end{pmatrix}.$$

This finishes the proof. ■

Proof of Corollary 3.3.7. The consistency, convergence rate and super-efficiency of $\widehat{\Pi}_{g,n}$ can be established using similarly arguments to those of Theorem 3.3.2, Theorem 3.3.3 and Theorem 3.3.4.

Under the super efficiency of $\widehat{\Pi}_{g,n}$, the true rank r_o is imposed in $\widehat{\Pi}_{g,n}$ w.p.a.1. Thus for large enough n, the GLS shrinkage estimator $\widehat{\Pi}_{g,n}$ can be decomposed as $\widehat{\alpha}_{g,n}\widehat{\beta}_{g,n}'$ w.p.a.1, where $\widehat{\alpha}_{g,n}$ and $\widehat{\beta}_{g,n}$ are some $m \times r_o$ matrices and they minimize the following criterion function w.p.a.1

$$\sum_{t=1}^{n} \left(\Delta Y_t - \alpha\beta' Y_{t-1}\right)' \widehat{\Omega}_{u,n}^{-1} \left(\Delta Y_t - \alpha\beta' Y_{t-1}\right) + n \sum_{k=1}^{r_o} \lambda_{r,k,n} \frac{||\Phi_{n,k}(\alpha\beta')||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}. \tag{3.153}$$

Using the similar arguments in the proof of Theorem 3.3.6, we define

$$\Pi_o = \alpha_o\beta_o' = [\alpha_o, \alpha_o O_{r_o}] \text{ and } \beta_o = [I_{r_o}, O_{r_o}]'$$

where O_{r_o} is some $r_o \times (m - r_o)$ matrix uniquely determined by the equation $\alpha_o O_{r_o} = \Pi_{o,2}$, where $\Pi_{o,2}$ denotes the last $m - r_o$ columns of Π_o.

Define $U_{1,n}^* = \sqrt{n}\left(\widehat{\alpha}_{g,n} - \alpha_o\right)$ and $U_{3,n}^* = n(\widehat{\beta}_{g,n} - \beta_o)' = \left[\mathbf{0}_{r_o}, n\left(\widehat{O}_{g,n} - O_o\right)\right] \equiv$

$\left[\mathbf{0}_{r_o}, U^*_{2,n}\right]$, then

$$\begin{aligned}\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} &= \left[\widehat{\alpha}_{g,n}\left(\widehat{\beta}_{g,n} - \beta_o\right)' + \left(\widehat{\alpha}_{g,n} - \alpha_o\right)\beta_o'\right] Q^{-1} D_n^{-1} \\ &= \left[n^{-\frac{1}{2}}\widehat{\alpha}_{g,n} U^*_{3,n}\alpha_o(\beta_o'\alpha_o)^{-1} + U^*_{1,n},\ \widehat{\alpha}_{g,n} U^*_{3,n}\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right].\end{aligned}$$

Define

$$\Pi_n(U) = \left[n^{-\frac{1}{2}}\widehat{\alpha}_{g,n} U_3\alpha_o(\beta_o'\alpha_o)^{-1} + U_1, \widehat{\alpha}_{g,n} U_3\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right],$$

then by definition, $U_n^* = \left(U^*_{1,n}, U^*_{2,n}\right)$ minimizes the following criterion function w.p.a.1

$$\begin{aligned}V_n(U) &= \sum_{t=1}^{n}\left[\left(u_t - \Pi_n(U)D_n Z_{t-1}\right)'\widehat{\Omega}^{-1}_{u,n}\left(u_t - \Pi_n(U)D_n Z_{t-1}\right) - u_t'\widehat{\Omega}^{-1}_{u,n}u_t\right] \\ &\quad + n\sum_{k=1}^{r_o}\lambda_{r,k,n}\frac{||\Phi_{n,k}(\Pi_n(U)D_nQ + \Pi_o)|| - ||\Phi_{n,k}(\Pi_o)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}. \qquad (3.154)\end{aligned}$$

Following similar arguments to those of Theorem 3.3.6, we can deduce that for any $k \in \mathcal{S}_\phi$

$$n\left|\lambda_{r,k,n}\frac{||\Phi_{n,k}(\Pi_n(U)D_nQ + \Pi_o)|| - ||\Phi_{n,k}(\Pi_o)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}\right| = o_p(1), \qquad (3.155)$$

and

$$\begin{aligned}
&\sum_{t=1}^{n}\left(u_t-\Pi_n(U)D_nZ_{t-1}\right)'\widehat{\Omega}_{u,n}^{-1}\left(u_t-\Pi_n(U)D_nZ_{t-1}\right)-\sum_{t=1}^{n}u_t'\widehat{\Omega}_{u,n}^{-1}u_t\\
\rightarrow_d\ & vec(U_1)'\left(\Sigma_{z_1z_1}\otimes\Omega_u^{-1}\right)vec(U_1)\\
&+vec(U_2)'\left[\beta_{2,o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\int B_{w_2}B_{w_2}'(\beta_{o,\perp}'\alpha_{o,\perp})^{-1}\beta_{2,o,\perp}'\otimes\alpha_o'\Omega_u^{-1}\alpha_o\right]vec(U_2)\\
&-2vec(U_1)'vec\left(\Omega_u^{-1}B_{1,m}\right)-2vec(U_2)'vec\left[\alpha_o'\Omega_u^{-1}B_{2,m}(\beta_{o,\perp}'\alpha_{o,\perp})^{-1}\beta_{2,o,\perp}'\right]\\
\equiv\ & V(U)
\end{aligned}\tag{3.156}$$

uniformly over U in any compact subspace of $R^{m\times r_o}\times R^{r_o\times(m-r_o)}$. $V(U)$ is uniquely minimized at $\left(U_{g,1}^*,U_{g,2}^*\right)$, where $U_{g,1}^*=B_{1,m}\Sigma_{z_1z_1}^{-1}$ and

$$U_{g,2}^*=(\alpha_o'\Omega_u^{-1}\alpha_o)^{-1}\left(\alpha_o'\Omega_u^{-1}B_{2,m}\right)\left(\int B_{w_2}B_{w_2}'\right)^{-1}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\beta_{2,o,\perp}^{-1}.$$

Invoking the argmax continuous mapping theorem, we obtain

$$\begin{aligned}
\left(\widehat{\Pi}_{g,n}-\Pi_o\right)Q^{-1}D_n^{-1}\ &=\ \left[\widehat{\alpha}_{g,n}\left(\widehat{\beta}_{g,n}-\beta_o\right)'+\left(\widehat{\alpha}_{g,n}-\alpha_o\right)\beta_o'\right]Q^{-1}D_n^{-1}\\
&\rightarrow\ _d\left[B_{1,m}\Sigma_{z_1z_1}^{-1},\ \alpha_o(\alpha_o'\Omega_u^{-1}\alpha_o)^{-1}\left(\alpha_o'\Omega_u^{-1}B_{2,m}\right)\left(\int B_{w_2}B_{w_2}'\right)^{-1}\right].
\end{aligned}\tag{3.157}$$

By the definition of w_1 and w_2, we can define $\Omega_{\widetilde{u}}=Q\Omega_uQ'$ such that

$$\Omega_{\widetilde{u}}=\begin{pmatrix}\Sigma_{w_1w_1} & \Sigma_{w_1w_2}\\ \Sigma_{w_2w_1} & \Sigma_{w_2w_2}\end{pmatrix}\text{ and }\Omega_{\widetilde{u}}^{-1}=\begin{pmatrix}\Omega_{\widetilde{u}}(11) & \Omega_{\widetilde{u}}(12)\\ \Omega_{\widetilde{u}}(21) & \Omega_{\widetilde{u}}(22)\end{pmatrix}.$$

Note that

$$\begin{aligned}(\alpha_o'\Omega_u^{-1}\alpha_o)^{-1}\alpha_o'\Omega_u^{-1} &= (\alpha_o'Q'\Omega_{\widetilde{u}}^{-1}Q\alpha_o)^{-1}\alpha_o'Q'\Omega_{\widetilde{u}}^{-1}Q \\ &= [(\alpha_o'\beta_o)\Omega_{\widetilde{u}}(11)(\beta_o'\alpha_o)]^{-1}[(\alpha_o'\beta_o),0]\Omega_{\widetilde{u}}^{-1}Q \\ &= (\beta_o'\alpha_o)^{-1}\Omega_{\widetilde{u}}^{-1}(11)\left[\Omega_{\widetilde{u}}(11)\beta_o' + \Omega_{\widetilde{u}}(12)\alpha_{o,\perp}'\right]. \quad (3.158)\end{aligned}$$

Under $\Omega_{\widetilde{u}}(12) = -\Omega_{\widetilde{u}}(11)\Sigma_{w_1w_2}\Sigma_{w_2w_2}^{-1}$,

$$(\alpha_o'\Omega_u^{-1}\alpha_o)^{-1}\alpha_o'\Omega_u^{-1} = (\beta_o'\alpha_o)^{-1}(\beta_o' - \Sigma_{w_1w_2}\Sigma_{w_2w_2}^{-1}\alpha_{o,\perp}'). \quad (3.159)$$

Now, using (3.157) and (3.159), we can deduce that

$$\left(\widehat{\Pi}_{g,n} - \Pi_o\right)Q^{-1}D_n^{-1} \to_d \left(\begin{array}{cc} B_{m,1} & \left(\int B_{w_2}dB_{u\cdot w_2}'\right)'\left(\int B_{w_2}B_{w_2}'\right)^{-1} \end{array}\right).$$

This finishes the proof. ■

3.9.2 Proof of Main Results in Section 3.4

Proof of Lemma 3.4.1. From the partial sum expression in (3.8), we get $Z_{1,t-1} = \beta_o'Y_{t-1} = R(L)\beta_o'u_t$, which implies that $\{\beta_o'Y_{t-1}\}_{t\geq 1}$ is a stationary process. Note that

$$E\left[u_tZ_{1,t-1}'\right] = \sum_{j=0}^{\infty}E\left[u_tu_{t-j}'\right]\beta_o\left(R^j\right)' = \sum_{j=0}^{\infty}\Sigma_{uu}(j)\beta_o\left(R^j\right)' < \infty.$$

Using a CLT for linear process time series (e.g. the multivariate version of theorem 8 and Remark 3.9 of Phillips and Solo, 1992), we deduce that

$$n^{-\frac{1}{2}}\sum_{t=1}^{n}\left[u_tZ_{1,t-1}' - \Sigma_{uz_1}(1)\right] \to_d N(0, V_{uz_1}),$$

which establishes (3.36). The results of (a)-(c) and (e) can be proved using similar arguments to those of Lemma 3.9.1. ■

Proof of Lemma 3.4.2. (a). Define $\Sigma_{wz_1} = Q\Sigma_{uz_1}(1)$. By Lemma 3.4.1, we have

$$\begin{aligned} H_n &= \left[\Sigma_{wz_1}, 0_{m\times(m-r_o)}\right] n^{\frac{1}{2}} D_n \left(D_n \sum_{t=1}^{n} Z_{t-1}Z_{t-1}' D_n \right)^{-1} D_n n^{\frac{1}{2}} \\ &= \left[\Sigma_{wz_1}, 0_{m\times(m-r_o)}\right] \left(D_n \sum_{t=1}^{n} Z_{t-1}Z_{t-1}' D_n \right)^{-1} \begin{pmatrix} I_{r_o} & 0 \\ 0 & n^{-\frac{1}{2}} I_{m-r_o} \end{pmatrix} \\ &\to_p \left(\Sigma_{wz_1}\Sigma_{z_1z_1}^{-1}, 0_{m\times(m-r_o)}\right) = H_o. \end{aligned} \tag{3.160}$$

(b). From the expression of H_n in the first line of (3.160), we get

$$\begin{aligned} nQ^{-1}H_nQ\beta_{o\perp} &= [\Sigma_{uz_1}(1), 0] \left(D_n \sum_{t=1}^{n} Z_{t-1}Z_{t-1}' D_n \right)^{-1} \begin{pmatrix} 0 \\ n^{\frac{1}{2}}\alpha_{o\perp}'\beta_{o\perp} \end{pmatrix} \\ &= -\Sigma_{uz_1}(1)\widehat{S}_{11}^{-1}\widehat{S}_{12} \left(n^{-1}\widehat{S}_{22} - n^{-1}\widehat{S}_{21}\widehat{S}_{11}^{-1}\widehat{S}_{12} \right)^{-1} \alpha_{o\perp}'\beta_{o\perp} \end{aligned} \tag{3.161}$$

By Lemma 3.4.1 and the CMT, we have

$$\left(\widehat{S}_{22} - n^{-1}\widehat{S}_{21}\widehat{S}_{11}^{-1}\widehat{S}_{12} \right)^{-1} \to_d \left(\int B_{w_2}B_{w_2}' \right)^{-1}. \tag{3.162}$$

Next note that

$$\widehat{S}_{12} \to_d -(\beta_o'\alpha_o)^{-1} \left[\left(\int B_{w_2} dB_{w_1}' \right)' + \Sigma_{w_1w_2} \right]. \tag{3.163}$$

The claimed result now follows by applying the results in (3.161)-(3.163), Lemma 3.4.1 and CMT into the expression in (3.161).

(c). From the expression of H_n in the first line of (3.160), we get

$$\begin{aligned}
& \sqrt{n}Q^{-1}\left(H_n - H_o\right)Q\beta_o \\
= \; & \sqrt{n}\left[\Sigma_{uz_1}(1), 0\right]\left\{ nD_n \begin{pmatrix} \widehat{S}_{11} & n^{-\frac{1}{2}}\widehat{S}_{12} \\ n^{-\frac{1}{2}}\widehat{S}_{21} & n^{-1}\widehat{S}_{22} \end{pmatrix}^{-1} D_n - \begin{pmatrix} \Sigma_{z_1z_1}^{-1} & 0 \\ 0 & 0 \end{pmatrix} \right\} Q\beta_o \\
= \; & \sqrt{n}\left[\Sigma_{uz_1}(1), 0\right] G_n Q\beta_o \qquad (3.164)
\end{aligned}$$

where

$$G_n = \begin{pmatrix} (\widehat{S}_{11} - \widehat{S}_{12}\widehat{S}_{22}^{-1}\widehat{S}_{21})^{-1} - \Sigma_{z_1z_1}^{-1} & -\frac{\widehat{S}_{11}^{-1}\widehat{S}_{12}}{n}\left(\frac{\widehat{S}_{22}}{n} - \frac{\widehat{S}_{21}\widehat{S}_{11}^{-1}\widehat{S}_{12}}{n}\right)^{-1} \\ -\widehat{S}_{22}^{-1}\widehat{S}_{21}(\widehat{S}_{11} - \widehat{S}_{12}\widehat{S}_{22}^{-1}\widehat{S}_{21})^{-1} & \frac{1}{n}\left(\frac{\widehat{S}_{22}}{n} - \frac{\widehat{S}_{21}\widehat{S}_{11}^{-1}\widehat{S}_{12}}{n}\right)^{-1} \end{pmatrix}. \qquad (3.165)$$

From Lemma 3.4.1,

$$\frac{\widehat{S}_{11}^{-1}\widehat{S}_{12}}{\sqrt{n}}\left(\frac{\widehat{S}_{22}}{n} - \frac{\widehat{S}_{21}\widehat{S}_{11}^{-1}\widehat{S}_{12}}{n}\right)^{-1} = o_p(1). \qquad (3.166)$$

Using (3.164), (3.165) and (3.166), we can deduce that

$$\sqrt{n}Q^{-1}\left(H_n - H_o\right)Q\beta_o = \Sigma_{uz_1}(1)\sqrt{n}\left(H_{n,11} - \Sigma_{z_1z_1}^{-1}\right)\beta_o'\beta_o + o_p(1) \qquad (3.167)$$

where $H_{n,11} = \left(\widehat{S}_{11} - \widehat{S}_{12}\widehat{S}_{22}^{-1}\widehat{S}_{21}\right)^{-1}$. Invoking the CLT and Lemma 3.4.1, we get

$$\begin{aligned}
& \Sigma_{uz_1}(1)\sqrt{n}\left(H_{n,11} - \Sigma_{z_1z_1}^{-1}\right)\beta_o'\beta_o \\
= \; & -\Sigma_{uz_1}(1)H_{n,11}\sqrt{n}\left(H_{n,11}^{-1} - \Sigma_{z_1z_1}\right)\Sigma_{z_1z_1}^{-1}\beta_o'\beta_o \\
= \; & -\Sigma_{uz_1}(1)H_{11,n}\sqrt{n}\left(\widehat{S}_{11} - \Sigma_{z_1z_1}\right)\Sigma_{z_1z_1}^{-1}\beta_o'\beta_o + o_p(1) \\
\to_d \; & \Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}N(0, V_{z_1z_1})\Sigma_{z_1z_1}^{-1}\beta_o'\beta_o, \qquad (3.168)
\end{aligned}$$

which finishes the proof. ■

Proof of Lemma 3.4.3. (a). From the expression in (3.11) and (3.160),

$$\left[Q\left(\widehat{\Pi}_{1st}-\Pi_o\right)Q^{-1}-H_n\right]D_n^{-1}=\left[\sum_{t=1}^{n}\left(w_tZ_{t-1}'-H_{1,o}\right)\right]D_n\left(D_n\sum_{t=1}^{n}Z_{t-1}Z_{t-1}'D_n\right)^{-1}, \tag{3.169}$$

where $H_{1,o}=\left[Q\Sigma_{uz_1}(1),0_{m\times(m-r_o)}\right]$. Now, the results in (a) are directly from Lemma 3.4.1 and CMT.

(b). Denote $P=\left[\widetilde{\beta}_1,\widetilde{\beta}_{1\perp}\right]$ and $S_n(\phi)=\phi I_m-\widehat{\Pi}_{1st}$, then by definition, the eigenvalues of $\widehat{\Pi}_{1st}$ are the solutions of the following determinantal equation,

$$0=|P'S_n(\phi)P|=\left|\begin{array}{cc}\phi\widetilde{\beta}_1'\widetilde{\beta}_1-\widetilde{\beta}_1'\widehat{\Pi}_{1st}\widetilde{\beta}_1 & -\widetilde{\beta}_1'\widehat{\Pi}_{1st}\widetilde{\beta}_{1\perp}\\ -\widetilde{\beta}_{1\perp}'\widehat{\Pi}_{1st}\widetilde{\beta}_1 & \phi I_{m-r_o}-\widetilde{\beta}_{1\perp}'\widehat{\Pi}_{1st}\widetilde{\beta}_{1\perp}\end{array}\right|. \tag{3.170}$$

As $\Pi_1\widetilde{\beta}_{1\perp}=0$ and $\widehat{\Pi}_{1st}=\Pi_1+o_p(1)$,

$$\widetilde{\beta}_1'\widehat{\Pi}_{1st}\widetilde{\beta}_{1\perp}=\widetilde{\beta}_1'\left(\widehat{\Pi}_{1st}-\Pi_1\right)\widetilde{\beta}_{1\perp}+\widetilde{\beta}_1'\Pi_1\widetilde{\beta}_{1\perp}=o_p(1). \tag{3.171}$$

Similarly, we have

$$\widetilde{\beta}_{1\perp}'\widehat{\Pi}_{1st}\widetilde{\beta}_{1\perp}=\widetilde{\beta}_{1\perp}'\left(\widehat{\Pi}_{1st}-\Pi_1\right)\widetilde{\beta}_{1\perp}=o_p(1) \tag{3.172}$$

and

$$\widetilde{\beta}_{1\perp}'\widehat{\Pi}_{1st}\widetilde{\beta}_1\to_p\widetilde{\beta}_{1\perp}'\Pi_1\widetilde{\beta}_1 \text{ and } \widetilde{\beta}_1'\widehat{\Pi}_{1st}\widetilde{\beta}_1\to_p\widetilde{\beta}_1'\Pi_1\widetilde{\beta}_1. \tag{3.173}$$

From the results in (3.170)-(3.173), we can invoke the Slutsky Theorem to deduce that

$$\left|\phi I_m-\widehat{\Pi}_{1st}\right|\to_p|\phi I_{m-r_1}|\times\left|\phi\widetilde{\beta}_1'\widetilde{\beta}_1-\widetilde{\beta}_1'\Pi_1\widetilde{\beta}_1\right|, \tag{3.174}$$

uniformly over any compact set in R, where $\left|\phi\widetilde{\beta}_1'\widetilde{\beta}_1-_p\widetilde{\beta}_1'\Pi_1\widetilde{\beta}_1\right|=0$ can equivalently

be written as $\left|\phi I_{r_1-r_o} - \widetilde{\beta}_1'\widetilde{\alpha}_1\right| = 0$. Hence the claimed results follow by (3.174) and the CMT.

(c). If we denote $u_{n,k}^* = n\phi_k(\widehat{\Pi}_{1st})$, then by definition, $u_{n,k}^*$ ($k \in \{r_o+1, ..., m\}$) is the solution of the following determinantal equation

$$0 = \left|\widetilde{\beta}_1' S_n(u)\widetilde{\beta}_1\right| \times \left|\widetilde{\beta}_{1\perp}' \left\{S_n(u) - S_n(u)\widetilde{\beta}_1 \left[\widetilde{\beta}_1' S_n(u)\widetilde{\beta}_1\right]^{-1} \widetilde{\beta}_1' S_n(u)\right\} \widetilde{\beta}_{1\perp}\right|, \quad (3.175)$$

where $S_n(u) = \frac{u}{n} I_m - \widehat{\Pi}_{1st}$.

From the results in (3.37) and (3.38), we have

$$\widetilde{\beta}_1' S_n(u)\widetilde{\beta}_1 = n^{-\frac{1}{2}} u\widetilde{\beta}_1'\widetilde{\beta}_1 - \widetilde{\beta}_1'\widehat{\Pi}_{1st}\widetilde{\beta}_1 \to_p -\widetilde{\beta}_1'\widetilde{\alpha}_1\widetilde{\beta}_1'\widetilde{\beta}_1, \quad (3.176)$$

where $\widetilde{\beta}_1'\widetilde{\alpha}_1\widetilde{\beta}_1'\widetilde{\beta}_1$ is a $r_1 \times r_1$ nonsingular matrix. Hence $u_{n,k}^*$ is the solution of the following determinantal equation asymptotically

$$0 = \left|\widetilde{\beta}_{1\perp}' \left\{S_n(u) - S_n(u)\widetilde{\beta}_1 \left[\widetilde{\beta}_1' S_n(u)\widetilde{\beta}_1\right]^{-1} \widetilde{\beta}_1' S_n(u)\right\} \widetilde{\beta}_{1\perp}\right|. \quad (3.177)$$

Denote $T_n(u) = S_n(u) - S_n(u)\widetilde{\beta}_1 \left[\widetilde{\beta}_1' S_n(u)\widetilde{\beta}_1\right]^{-1} \widetilde{\beta}_1' S_n(u)$, then (3.177) can be equivalently written as

$$0 = \left|\widetilde{\beta}_\perp' T_n(u)\widetilde{\beta}_\perp\right| \times \left|\beta_{o\perp}' \left\{T_n(u) - T_n(u)\widetilde{\beta}_\perp \left[\widetilde{\beta}_\perp' T_n(u)\widetilde{\beta}_\perp\right]^{-1} \widetilde{\beta}_\perp' T_n(u)\right\} \beta_{o\perp}\right|. \quad (3.178)$$

Note that

$$\begin{aligned}
n^{\frac{1}{2}}\widetilde{\beta}_\perp' S_n(u)\widetilde{\beta}_\perp &= n^{-\frac{1}{2}} u\widetilde{\beta}_\perp'\widetilde{\beta}_\perp - n^{\frac{1}{2}}\widetilde{\beta}_\perp' \left(\widehat{\Pi}_{1st} - \Pi_{1,n}\right)\widetilde{\beta}_\perp - n^{\frac{1}{2}}\widetilde{\beta}_\perp'\Pi_{1,n}\widetilde{\beta}_\perp, \\
n^{\frac{1}{2}}\widetilde{\beta}_1' S_n(u)\widetilde{\beta}_\perp &= -n^{\frac{1}{2}}\widetilde{\beta}_1' \left(\widehat{\Pi}_{1st} - \Pi_{1,n}\right)\widetilde{\beta}_\perp - n^{\frac{1}{2}}\widetilde{\beta}_1'\Pi_{1,n}\widetilde{\beta}_\perp, \\
\widetilde{\beta}_\perp' S_n(u)\widetilde{\beta}_1 &= -\widetilde{\beta}_\perp'\widehat{\Pi}_{1st}\widetilde{\beta}_1.
\end{aligned}$$

From above expressions, we can write

$$n^{\frac{1}{2}}\widetilde{\beta}_{\perp}'T_n(u)\widetilde{\beta}_{\perp} = n^{-\frac{1}{2}}u\widetilde{\beta}_{\perp}'\widetilde{\beta}_{\perp} - \widetilde{\beta}_{\perp}'\left[I_m + \widehat{\Pi}_{1st}\widetilde{\beta}_1\left[\widetilde{\beta}_1'S_n(u)\widetilde{\beta}_1\right]^{-1}\widetilde{\beta}_1'\right]M_{1,n}, \quad (3.179)$$

where $M_{1,n} = n^{\frac{1}{2}}\left(\widehat{\Pi}_{1st} - \Pi_{1,n}\right)\widetilde{\beta}_{\perp} + n^{\frac{1}{2}}\Pi_{1,n}\widetilde{\beta}_{\perp}$. Using Lemma 3.4.1 and the results in (a), we can deduce that

$$\begin{aligned} n^{\frac{1}{2}}\left(\widehat{\Pi}_{1st} - \Pi_{1,n}\right)\widetilde{\beta}_{\perp} &= Q^{-1}\left[Q\left(\widehat{\Pi}_{1st} - \Pi_{1,n}\right)Q^{-1}D_n^{-1}\right]n^{\frac{1}{2}}D_nQ\widetilde{\beta}_{\perp} \\ &\to_d N(0, V_{uz_1})\Sigma_{z_1z_1}^{-1}\beta_o'\widetilde{\beta}_{\perp} \equiv \mathcal{N}_1\widetilde{\beta}_{\perp}. \end{aligned} \quad (3.180)$$

Using the results in Lemma 3.4.2, we get

$$\begin{aligned} n^{\frac{1}{2}}\Pi_{1,n}\widetilde{\beta}_{\perp} &= n^{\frac{1}{2}}\left(\Pi_{1,n} - \Pi_1\right)\widetilde{\beta}_{\perp} = n^{\frac{1}{2}}Q^{-1}\left(H_n - H_o\right)Q\widetilde{\beta}_{\perp} \\ &\to_d \Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}N(0, V_{z_1z_1})\Sigma_{z_1z_1}^{-1}\beta_o'\widetilde{\beta}_{\perp} \equiv \mathcal{N}_2\widetilde{\beta}_{\perp}. \end{aligned} \quad (3.181)$$

From (3.179)-(3.181), we can deduce that

$$\left|\sqrt{n}\widetilde{\beta}_{\perp}'T_n(u)\widetilde{\beta}_{\perp}\right| \to_d \left|\widetilde{\beta}_{\perp}'\left(\mathcal{N}_1 + \mathcal{N}_2\right)\widetilde{\beta}_{\perp}\right| \neq 0, \ a.e. \quad (3.182)$$

Next note that

$$\beta_{o\perp}'T_n(u)\beta_{o\perp} = \frac{u\beta_{o\perp}'\beta_{o\perp}}{n} - \beta_{o\perp}'\left[I_m + \widehat{\Pi}_{1st}\widetilde{\beta}_1\left(\widetilde{\beta}_1'S_n(u)\widetilde{\beta}_1\right)^{-1}\widetilde{\beta}_1'\right]M_{2,n}, \quad (3.183)$$

$$\widetilde{\beta}_{\perp}'T_n(u)\beta_{o\perp} = -\widetilde{\beta}_{\perp}'\left[I_m + \widehat{\Pi}_{1st}\widetilde{\beta}_1\left(\widetilde{\beta}_1'S_n(u)\widetilde{\beta}_1\right)^{-1}\widetilde{\beta}_1'\right]M_{2,n}, \quad (3.184)$$

where $M_{2,n} = \left(\widehat{\Pi}_{1st} - \Pi_{1,n}\right)\beta_{o\perp} + \Pi_{1,n}\beta_{o\perp}$. By (3.173) and (3.176), we have

$$\beta_{o\perp}'T_n(u)\widetilde{\beta}_{\perp} = \beta_{o\perp}'\widehat{\Pi}_{ols}\widetilde{\beta}_{\perp} - \beta_{o\perp}'\widehat{\Pi}_{1st}\widetilde{\beta}_1\left(\widetilde{\beta}_1'S_n(u)\widetilde{\beta}_1\right)^{-1}\widetilde{\beta}_1'\widehat{\Pi}_{1st}\widetilde{\beta}_{\perp} = o_p(1). \quad (3.185)$$

Using Lemma 3.4.2, we can deduce that

$$\begin{aligned} n\left(\widehat{\Pi}_{1st}-\Pi_{1,n}\right)\beta_{o\perp} &= nQ^{-1}\left[Q\left(\widehat{\Pi}_{1st}-\Pi_{1,n}\right)Q^{-1}D_n\right]D_n^{-1}Q\beta_{o\perp} \\ &= Q^{-1}\left[Q\left(\widehat{\Pi}_{1st}-\Pi_{1,n}\right)Q^{-1}D_n\right]\begin{pmatrix}0\\ \\ \alpha'_{o\perp}\beta_{o\perp}\end{pmatrix} \\ &\to_d \int dB_u B'_{w_2}\left(\int B_{w_2}B'_{w_2}\right)^{-1}\alpha'_{o\perp}\beta_{o\perp}\equiv\widetilde{\Pi}_0\beta_{o\perp} \quad (3.186) \end{aligned}$$

Using the result in (3.39), we get

$$\begin{aligned} n\Pi_{1n}\beta_{o\perp} &= n\left[\Pi_o+Q^{-1}H_nQ\right]\beta_{o\perp} \\ &= nQ^{-1}H_nQ\beta_{o\perp}\to_d \widetilde{\Pi}_1\beta_{o\perp}. \end{aligned} \quad (3.187)$$

From (3.183)-(3.187), we deduce that

$$\begin{aligned} &\left|n\beta'_{o\perp}\left\{T_n(u)-T_n(u)\widetilde{\beta}_\perp\left[\widetilde{\beta}'_\perp T_n(u)\widetilde{\beta}_\perp\right]^{-1}\widetilde{\beta}'_\perp T_n(u)\right\}\beta_{o\perp}\right| \\ \to_d &\left|uI_{m-r_o}-\beta'_{o\perp}\left(\widetilde{\Pi}_0+\widetilde{\Pi}_1\right)\beta_{o\perp}\right|, \end{aligned} \quad (3.188)$$

uniformly over any compact set in R. Now, the results in (c) follow from (3.188) and the CMT.

(d) If we denote $u^*_{n,k}=\sqrt{n}\phi_k(\widehat{\Pi}_{1st})$, then by definition, $u^*_{n,k}$ $(k\in\{r_1+1,...,r_o\})$ is the solution of the following determinantal equation

$$0=\left|\widetilde{\beta}'_1 S_n(u)\widetilde{\beta}_1\right|\times\left|\widetilde{\beta}'_{1\perp}\left\{S_n(u)-S_n(u)\widetilde{\beta}_1\left[\widetilde{\beta}'_1 S_n(u)\widetilde{\beta}_1\right]^{-1}\widetilde{\beta}'_1 S_n(u)\right\}\widetilde{\beta}_{1\perp}\right|, \quad (3.189)$$

where $S_n(u)=\frac{u}{\sqrt{n}}I_m-\widehat{\Pi}_{1st}$.

Note that

$$\widetilde{\beta}'_{1\perp} S_n(u) \widetilde{\beta}_{1\perp} = n^{-\frac{1}{2}} u I_{m-r_1} - \widetilde{\beta}'_{1\perp} \widehat{\Pi}_{1st} \widetilde{\beta}_{1\perp}, \tag{3.190}$$

$$\widetilde{\beta}'_{1\perp} S_n(u) \widetilde{\beta}_1 = -\widetilde{\beta}'_{1\perp} \widehat{\Pi}_{1st} \widetilde{\beta}_1 \text{ and } \widetilde{\beta}'_1 S_n(u) \widetilde{\beta}_{1\perp} = -\widetilde{\beta}'_1 \widehat{\Pi}_{1st} \widetilde{\beta}_{1\perp}. \tag{3.191}$$

Using expressions (3.190) and (3.191), we have

$$\begin{aligned} & \widetilde{\beta}'_{1\perp} \left\{ S_n(u) - S_n(u) \widetilde{\beta}_1 \left[\widetilde{\beta}'_1 S_n(u) \widetilde{\beta}_1 \right]^{-1} \widetilde{\beta}'_1 S_n(u) \right\} \widetilde{\beta}_{1\perp} \\ = \; & n^{-\frac{1}{2}} u I_{m-r_1} - \widetilde{\beta}'_{1\perp} \left\{ I_m + \widehat{\Pi}_{1st} \widetilde{\beta}_1 \left[\widetilde{\beta}'_1 S_n(u) \widetilde{\beta}_1 \right]^{-1} \widetilde{\beta}'_1 \right\} \widehat{\Pi}_{1st} \widetilde{\beta}_{1\perp}. \end{aligned} \tag{3.192}$$

From (3.186), we get

$$\sqrt{n} \left(\widehat{\Pi}_{1st} - \Pi_{1n} \right) \beta_{o,\perp} = o_p(1). \tag{3.193}$$

Using (3.180), (3.39) and (3.40), we have

$$\sqrt{n} \left(\widehat{\Pi}_{1st} - \Pi_{1n} \right) \widetilde{\beta}_\perp \to_d \widetilde{\Pi}_3 \widetilde{\beta}_\perp, \tag{3.194}$$

$$\sqrt{n} \Pi_{1n} \beta_{o\perp} = o_p(1), \tag{3.195}$$

$$\sqrt{n} \Pi_{1n} \widetilde{\beta}_\perp \to_d \widetilde{\Pi}_2 \widetilde{\beta}_\perp. \tag{3.196}$$

Now, using the results in (3.193)-(3.196), we get

$$\begin{aligned} \sqrt{n} \widehat{\Pi}_{1st} \widetilde{\beta}_{1\perp} &= \sqrt{n} \left[\widehat{\Pi}_{1st} - \Pi_{1n} \right] \widetilde{\beta}_{1\perp} + \sqrt{n} \Pi_{1n} \widetilde{\beta}_{1\perp} \\ &\to_d \left[0_{m\times(m-r_o)}, \; \widetilde{\Pi}_3 \widetilde{\beta}_\perp \right] + \left[0_{m\times(m-r_o)}, \; \widetilde{\Pi}_2 \widetilde{\beta}_\perp \right] \\ &= \left[0_{m\times(m-r_o)}, \; \left(\widetilde{\Pi}_2 + \widetilde{\Pi}_3 \right) \widetilde{\beta}_\perp \right]. \end{aligned} \tag{3.197}$$

From the results in (3.192)-(3.197), it follows that

$$\begin{aligned}
&\left|\sqrt{n}\widetilde{\beta}_{1\perp}'\left\{S_n(u)-S_n(u)\widetilde{\beta}_1\left[\widetilde{\beta}_1'S_n(u)\widetilde{\beta}_1\right]^{-1}\widetilde{\beta}_1'S_n(u)\right\}\widetilde{\beta}_{1\perp}\right| \\
\rightarrow_d\ &\left|uI_{m-r_1}-\widetilde{\beta}_{1\perp}'\left[I_m+\widetilde{\alpha}_1\left(\widetilde{\beta}_1'\widetilde{\alpha}_1\right)^{-1}\widetilde{\beta}_1'\right]\left[0_{m\times(m-r_o)},\ \left(\mathcal{N}_1+\mathcal{N}_2\right)\widetilde{\beta}_\perp\right]\right| \\
=\ &|uI_{m-r_0}|\times\left|uI_{r_o-r_1}-\widetilde{\beta}_\perp'\left(\widetilde{\Pi}_2+\widetilde{\Pi}_3\right)\widetilde{\beta}_\perp\right|. \qquad (3.198)
\end{aligned}$$

Note that the determinantal equation

$$|uI_{m-r_0}|\times\left|uI_{r_o-r_1}-\widetilde{\beta}_\perp'\left(\widetilde{\Pi}_2+\widetilde{\Pi}_3\right)\widetilde{\beta}_\perp\right|=0 \qquad (3.199)$$

has $m-r_0$ zero eigenvalues, which correspond to the probability limit of $\sqrt{n}\phi_k(\widehat{\Pi}_{1st})$ ($k\in\{r_1+1,...,r_o\}$), as illustrated in (c). Equation (3.199) also has r_o-r_1 non-trivial eigenvalues as solutions of the stochastic determinantal equation

$$\left|uI_{r_o-r_1}-\widetilde{\beta}_\perp'\left(\widetilde{\Pi}_2+\widetilde{\Pi}_3\right)\widetilde{\beta}_\perp\right|=0,$$

which finishes the proof. ■

Recall that P_n is defined as the inverse of Q_n. We divide P_n and Q_n as $P_n=[P_{\widetilde{\alpha},n},P_{\widetilde{\alpha}_\perp,n}]$ and $Q_n'=\left[Q_{\widetilde{\alpha},n}',Q_{\widetilde{\alpha}_\perp,n}'\right]$, where $Q_{\widetilde{\alpha},n}$ and $P_{\widetilde{\alpha},n}$ are the first r_1 rows of Q_n and first r_1 columns of P_n respectively ($Q_{\widetilde{\alpha}_\perp,n}$ and $P_{\widetilde{\alpha}_\perp,n}$ are defined accordingly). By definition,

$$Q_{\widetilde{\alpha}_\perp,n}P_{\widetilde{\alpha}_\perp,n}=I_{m-r_1},\ Q_{\widetilde{\alpha},n}P_{\widetilde{\alpha}_\perp,n}=\mathbf{0}_{r_1\times(m-r_1)}\text{ and }Q_{\widetilde{\alpha}_\perp,n}\widehat{\Pi}_{1st}=\Lambda_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n} \qquad (3.200)$$

where $\Lambda_{\widetilde{\alpha}_\perp,n}$ is an diagonal matrix with the ordered last (smallest) $m-r_1$ eigenvalues

of $\widehat{\Pi}_{1st}$. Using the results in (3.200), we can define a useful estimator of Π_1 as

$$\widetilde{\Pi}_{n,f} = \widehat{\Pi}_{1st} - P_{\widetilde{\alpha}_\perp,n}\Lambda_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n}. \tag{3.201}$$

By definition

$$Q_{\widetilde{\alpha},n}\widetilde{\Pi}_{n,f} = Q_{\widetilde{\alpha},n}\widehat{\Pi}_{1st} - Q_{\widetilde{\alpha},n}P_{\widetilde{\alpha}_\perp,n}\Lambda_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n} = \Lambda_{\widetilde{\alpha},n}Q_{\widetilde{\alpha},n} \tag{3.202}$$

where $\Lambda_{\widetilde{\alpha},n}$ is an diagonal matrix with the ordered first (largest) r_o eigenvalues of $\widehat{\Pi}_{1st}$, and more importantly

$$Q_{\widetilde{\alpha}_\perp,n}\widetilde{\Pi}_{n,f} = Q_{\widetilde{\alpha}_\perp,n}\widehat{\Pi}_{1st} - Q_{\widetilde{\alpha}_\perp,n}P_{\widetilde{\alpha}_\perp,n}\Lambda_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n} = \mathbf{0}_{(m-r_1)\times m}. \tag{3.203}$$

From Lemma 3.4.3.(b), (3.202) and (3.203), we can deduce that $Q_{\widetilde{\alpha},n}\widetilde{\Pi}_{n,f}$ is a $r_1 \times m$ matrix which is nonzero w.p.a.1 and $Q_{\widetilde{\alpha}_\perp,n}\widetilde{\Pi}_{n,f}$ is a $(m-r_1)\times m$ zero matrix for all n. Using (3.200), we can write

$$\begin{aligned}
\widetilde{\Pi}_{n,f} - \Pi_1 &= (\widehat{\Pi}_{1st} - \Pi_1) - P_{\widetilde{\alpha}_\perp,n}\Lambda_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n} \\
&= (\widehat{\Pi}_{1st} - \Pi_1) - P_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n}(\widehat{\Pi}_{1st} - \Pi_1) - P_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n}\Pi_1. \qquad (3.204)
\end{aligned}$$

Under Lemma 3.4.3.(a),

$$\begin{aligned}
Q\left(\widehat{\Pi}_{1st} - \Pi_1\right)Q^{-1}D_n^{-1} &= Q\left(\widehat{\Pi}_{1st} - \Pi_o - Q^{-1}H_nQ\right)Q^{-1}D_n^{-1} + (H_n - H_o)D_n^{-1} \\
&= \left[Q\left(\widehat{\Pi}_{1st} - \Pi_o\right)Q^{-1} - H_n\right]D_n^{-1} + (H_n - H_o)D_n^{-1} \\
&= (H_n - H_o)D_n^{-1} + O_p(1). \qquad (3.205)
\end{aligned}$$

Under Lemma 3.4.1,

$$Q^{-1}(H_n - H_o)D_n^{-1} = \left[\Sigma_{uz_1}(1), 0_{m\times(m-r_o)}\right] G_n D_n^{-1} = O_p(1) \tag{3.206}$$

where G_n is defined in (3.165). From (3.205) and (3.206), we can deduce that

$$\left(\widehat{\Pi}_{1st} - \Pi_1\right) Q^{-1} D_n^{-1} = O_p(1). \tag{3.207}$$

Using (3.207) and Lemma 3.4.3.(c) and (d), we obtain

$$\begin{aligned} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \Pi_1 Q^{-1} D_n^{-1} &= \sqrt{n} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \Pi_1 Q^{-1} \\ &= -\sqrt{n} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \left(\widehat{\Pi}_{1st} - \Pi_1\right) Q^{-1} + \sqrt{n} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \widehat{\Pi}_{1st} \\ &= \sqrt{n} P_{\widetilde{\alpha}_\perp,n} \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} + O_p(1) = O_p(1) \end{aligned} \tag{3.208}$$

Thus under (3.204), (3.207) and (3.208), we get

$$\left(\widetilde{\Pi}_{n,f} - \Pi_1\right) Q^{-1} D_n^{-1} = O_p(1). \tag{3.209}$$

Comparing (3.207) with (3.209), we see that $\widetilde{\Pi}_{n,f}$ is as good as the OLS estimate $\widehat{\Pi}_{1st}$ in terms of its rate of convergence.

Proof of Corollary 3.4.4. First, when $r_o = 0$, then $\Pi_1 = \widetilde{\alpha}_o \beta_o' = 0 = \Pi_o$. Hence, (3.46) follows by the similar arguments to those in the proof of Theorem 3.3.2. To finish the proof, we only need to consider the scenarios where $r_o = m$ and $r_o \in (0, m)$.

Using the same notation for $V_n(\cdot)$ defined in the proof of Theorem 3.3.2, by

definition we have $V_n(\widehat{\Pi}_n) \leq V_n(\widetilde{\Pi}_{n,f})$, which implies

$$\begin{aligned}
&\left[vec(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)\right]' \left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right) \left[vec(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)\right] \\
&+2\left[vec(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)\right]' vec\left[\sum_{t=1}^{n} u_t Y_{t-1}' - (\Pi_1 - \Pi_o)\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'\right] \\
&-2\left[vec(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)\right]' \left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right) vec(\widetilde{\Pi}_{n,f} - \Pi_1) \\
\leq\ & n\left\{\sum_{k=1}^{m} \lambda_{r,k,n}\left[||\Phi_{n,k}(\widetilde{\Pi}_{n,f})|| - ||\Phi_{n,k}(\widehat{\Pi}_n)||\right] / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega}\right\}. \qquad (3.210)
\end{aligned}$$

When $r_o = m$, Y_t is stationary and we have

$$\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \to_p \Sigma_{yy} = R(1)\Omega_u R(1)'. \qquad (3.211)$$

From the results in (3.210) and (3.211), we get w.p.a.1,

$$\mu_{n,\min}||\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}|| - ||\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}||(c_{1n} + c_{2n}) - d_n \leq 0, \qquad (3.212)$$

where $\mu_{n,\min}$ denotes the smallest eigenvalue of $\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'$, which is positive w.p.a.1,

$$\begin{aligned}
c_{1n} &= \left\| \frac{\sum_{t=1}^{n} u_t Y_{t-1}'}{n} - (\Pi_1 - \Pi_o)\frac{\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'}{n} \right\| \\
&\to_p \left\|\Sigma_{uy}(1) - \Sigma_{uy}(1)\Sigma_{yy}^{-1}\Sigma_{yy}\right\| = 0 \qquad (3.213)
\end{aligned}$$

by Lemma 3.4.1 and the definition of Π_1, and

$$c_{2n} = m\left\|n^{-1}\sum\nolimits_{t=1}^{n} Y_{t-1}Y_{t-1}'\right\| ||\widetilde{\Pi}_{n,f} - \Pi_1|| = o_p(1) \qquad (3.214)$$

by Lemma 3.4.1 and (3.209), and

$$d_n = \sum_{k=1}^{m} \lambda_{r,k,n} \frac{||\Phi_{n,k}(\widetilde{\Pi}_{n,f})|| - ||\Phi_{n,k}(\widehat{\Pi}_n)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} \leq \sum_{k=1}^{r_1} \lambda_{r,k,n} \frac{||\Phi_{n,k}(\widetilde{\Pi}_{n,f})||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} = o_p(1) \tag{3.215}$$

by Lemma 3.4.3, (3.203) and $\lambda_{r,k,n} = o(1)$ for $k = 1, ..., r_1$. So result in (3.46) follows directly from (3.209), the inequality in (3.212) and the triangle inequality.

When $0 < r_o < m$,

$$\begin{aligned} & vec(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f})' \left(\sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m \right) vec(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) \\ = \; & vec(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f})' \left(B_n D_n \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_n B_n' \otimes I_m \right) vec(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) \\ \geq \; & \mu_{n,\min} ||(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) B_n||^2 \end{aligned} \tag{3.216}$$

where $\mu_{n,\min}$ denotes the smallest eigenvalue of $D_n \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_n$ which is positive definite w.p.a.1 under Lemma 3.4.1. Next, note that

$$\begin{aligned} & \left\{ \sum_{t=1}^{n} u_t Z_{t-1}' - \left[(\Pi_1 - \Pi_o) Q^{-1} \right] \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' \right\} D_n \\ = \; & \begin{bmatrix} n^{-\frac{1}{2}} \sum_{t=1}^{n} Z_{1,t-1} u_t' \\ n^{-1} \sum_{t=1}^{n} Z_{2,t-1} u_t' \end{bmatrix}' - \begin{bmatrix} n^{-\frac{1}{2}} \sum_{t=1}^{n} Z_{1,t-1} Z_{1,t-1}' \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) \\ n^{-1} \sum_{t=1}^{n} Z_{2,t-1} Z_{1,t-1}' \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) \end{bmatrix}' \end{aligned} \tag{3.217}$$

From Lemma 3.4.1, we can deduce that

$$n^{-1} \sum_{t=1}^{n} Z_{2,t-1} u_t' = O_p(1) \text{ and } n^{-1} \sum_{t=1}^{n} Z_{2,t-1} Z_{1,t-1}' \Sigma_{\beta\beta}^{-1} \Sigma_{u z_1}'(1) = O_p(1). \tag{3.218}$$

Similarly, we get

$$n^{-\frac{1}{2}} \sum_{t=1}^{n} \left[Z_{1,t-1} u_t' - \Sigma_{u z_1}'(1) \right] - n^{\frac{1}{2}} \left[S_{n,11} - \Sigma_{z_1 z_1} \right] \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) = O_p(1). \tag{3.219}$$

Define $e_{1n} = \left\| \left\{ \sum_{t=1}^{n} u_t Z'_{t-1} - (\Pi_1 - \Pi_o) Q^{-1} \sum_{t=1}^{n} Z_{t-1} Z'_{t-1} \right\} D_n \right\|$, then from (3.217)-(3.219) we can deduce that $e_{1n} = O_p(1)$. By the Cauchy-Schwarz inequality, we have

$$\begin{aligned} & \left| vec(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f})' vec \left[\sum_{t=1}^{n} u_t Y'_{t-1} - (\Pi_1 - \Pi_o) \sum_{t=1}^{n} Y_{t-1} Y'_{t-1} \right] \right| \\ = & \left| vec(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f})' vec \left[\left\{ \sum_{t=1}^{n} u_t Z'_{t-1} - (\Pi_1 - \Pi_o) Q^{-1} \sum_{t=1}^{n} Z_{t-1} Z'_{t-1} \right\} D_n B'_n \right] \right| \\ \leq & ||(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) B_n|| e_{1n}. \end{aligned} \tag{3.220}$$

Under Lemma 3.4.1 and (3.209),

$$\begin{aligned} e_{2n} \equiv & \left| vec(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)' \left(\sum_{t=1}^{n} Y_{t-1} Y'_{t-1} \otimes I_m \right) vec(\widetilde{\Pi}_{n,f} - \Pi_1) \right| \\ = & \left| vec(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)' \left(B_n D_n \sum_{t=1}^{n} Z_{t-1} Z'_{t-1} D_n B'_n \otimes I_m \right) vec(\widetilde{\Pi}_{n,f} - \Pi_1) \right| \\ \leq & ||(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) B_n|| \times ||(\widetilde{\Pi}_{n,f} - \Pi_1) B_n|| \times ||D_n \sum_{t=1}^{n} Z_{t-1} Z'_{t-1} D_n|| = O_p(1). \end{aligned} \tag{3.221}$$

From results in (3.210), (3.220) and (3.221), we get w.p.a.1

$$\mu_{n,\min} ||(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) B_n||^2 - 2||(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}) B_n||^2 (e_{1n} + e_{2n}) - d_n \leq 0 \tag{3.222}$$

where $d_n = o_p(1)$ by (3.215). Now, result in (3.46) follows by (3.222) and the same arguments in Theorem 3.3.2. ■

Proof of Corollary 3.4.5. From Lemma 3.4.3 and Corollary 3.4.4, we deduce

that w.p.a.1

$$\begin{aligned}
&\left|\sum_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}\frac{||\Phi_{n,k}(\widetilde{\Pi}_{n,f})||-||\Phi_{n,k}(\widehat{\Pi}_n)||}{||\phi_k(\widehat{\Pi}_{1st})||^\omega}\right| \\
\leq\ &\sum_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}\left|\frac{||\Phi_{n,k}(\widetilde{\Pi}_{n,f})||-||\Phi_{n,k}(\widehat{\Pi}_n)||}{||\phi_k(\widehat{\Pi}_{1st})||^\omega}\right| \\
\leq\ &d_{\widetilde{\mathcal{S}}_\phi}\left\{\max_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}||\phi_k(\widehat{\Pi}_{1st})||^{-\omega}\right\}||\widehat{\Pi}_n-\widetilde{\Pi}_{n,f}||. \qquad (3.223)
\end{aligned}$$

Using (3.210) and (3.223), we have

$$\begin{aligned}
&\left[vec(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n)\right]'\left(\sum_{t=1}^n Y_{t-1}Y_{t-1}'\otimes I_m\right)\left[vec(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n)\right] \\
&+2\left[vec(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n)\right]'vec\left[\sum_{t=1}^n u_tY_{t-1}'-(\Pi_1-\Pi_o)\sum_{t=1}^n Y_{t-1}Y_{t-1}'\right] \\
&-2\left[vec(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n)\right]'\left(\sum_{t=1}^n Y_{t-1}Y_{t-1}'\otimes I_m\right)vec(\widetilde{\Pi}_{n,f}-\Pi_1) \\
\leq\ &c\left\{\max_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}||\phi_k(\widehat{\Pi}_{1st})||^{-\omega}\right\}||\widehat{\Pi}_n-\widetilde{\Pi}_{n,f}|| \qquad (3.224)
\end{aligned}$$

where $c>0$ is a generic positive constant. When $r_o=0$, the convergence rate of $\widehat{\Pi}_n$ could be derived using the same arguments in Theorem 3.3.3. Hence, to finish the proof, we only need to consider scenarios where $r_o=m$ or $0<r_o<m$.

When $r_o=m$, following similar arguments to those of Theorem 3.3.3, we get

$$\mu_{n,\min}||\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n||^2-c||\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n||\left(c_{1n}+c_{2n}+\widetilde{\delta}_{r,n}\right)\leq 0, \qquad (3.225)$$

where

$$\begin{aligned} c_{1n} &= \left\| n^{-1}\sum_{t=1}^{n} u_t Y'_{t-1} - n^{-1}(\Pi_1 - \Pi_o)\sum_{t=1}^{n} Y_{t-1}Y'_{t-1} \right\| \\ &= n^{-\frac{1}{2}} \left\| n^{-\frac{1}{2}}\sum_{t=1}^{n}\left[u_t Y'_{t-1} - \Sigma_{uy}(1)\right] - \Sigma_{uy}(1)\Sigma_{z_1z_1}^{-1}\left[n^{\frac{1}{2}}\left(\widehat{S}_{11} - \Sigma_{z_1z_1}\right)\right]\right\| \\ &= O_p(n^{-\frac{1}{2}}) \end{aligned} \tag{3.226}$$

by Lemma 3.4.1, and

$$c_{2n} = \left\| n^{-1}\sum\nolimits_{t=1}^{n} Y_{t-1}Y'_{t-1}\right\| \left\|\widetilde{\Pi}_{n,f} - \Pi_1\right\| = O_p(n^{-\frac{1}{2}}) \tag{3.227}$$

by Lemma 3.4.1 and 3.209. From the results in (3.209), (3.225), (3.226) and (3.227), we deduce that

$$\widehat{\Pi}_n - \Pi_1 = O_p(n^{-\frac{1}{2}} + \widetilde{\delta}_{r,n}). \tag{3.228}$$

When $0 < r_o < m$, we can use (3.220) and (3.221) in the proof of Corollary 3.4.4 and (3.224) and to get w.p.a.1

$$\mu_{n,\min}||(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)B_n||^2 - 2||(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)B_n||(e_{1,n} + e_{2,n}) \leq cn\delta_n||\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n||, \tag{3.229}$$

where $e_{1,n} = O_p(1)$ and $e_{2,n} = O_p(1)$ as illustrated in the proof of Corollary 3.4.4. By the Cauchy-Schwarz inequality,

$$||(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)B_nB_n^{-1}|| \leq cn^{-\frac{1}{2}}||(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)B_n||. \tag{3.230}$$

Using (3.229) and (3.230), we obtain

$$\mu_{n,\min}||(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)B_n||^2 - c||(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n)B_n||(e_{1,n} + e_{2,n} + n^{\frac{1}{2}}\widetilde{\delta}_{r,n}) \leq 0. \tag{3.231}$$

From (3.209) and the inequality in (3.231), we obtain

$$(\widehat{\Pi}_n - \Pi_1)B_n = (\widehat{\Pi}_n - \widetilde{\Pi}_{n,f})B_n + (\widetilde{\Pi}_{n,f} - \Pi_1)B_n = O_p(1 + n^{\frac{1}{2}}\widetilde{\delta}_{r,n}),$$

which finishes the proof. ■

Proof of Corollary 3.4.6. Using similar arguments in the proof of Theorem 3.3.4, we can rewrite the LS shrinkage estimation problem as

$$\widehat{T}_n = \underset{T \in R^{m \times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - P_n T Y_{t-1}\|^2 + n \sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \|T(k)\| / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega}. \quad (3.232)$$

Result in (3.48) is equivalent to $\widehat{T}_n(k) = 0$ for any $k \in \{r_o + 1, ..., m\}$. Conditional on the event $\{Q_n(k_o)\widehat{\Pi}_n \neq 0\}$ for some k_o satisfying $r_o < k_o \leq m$, we get the following equation from the KKT optimality conditions,

$$\left\| \frac{1}{n} \sum\nolimits_{t=1}^{n} (\Delta Y_t - P_n \widehat{T}_n Y_{t-1})' P_n(k_o) Y_{t-1}' \right\| = \frac{\lambda_{r,k_o,n}}{2||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}}. \quad (3.233)$$

The sample average in the left hand side of (3.233) can be rewritten as

$$\frac{\sum_{t=1}^{n} (\Delta Y_t - P_n \widehat{T}_n Y_{t-1})' P_n(k_o) Y_{t-1}'}{n} = \frac{P_n'(k_o) \sum_{t=1}^{n} [u_t - (\widehat{\Pi}_n - \Pi_o) Y_{t-1}] Y_{t-1}'}{n}$$

$$= \frac{P_n'(k_o)}{n} \left[\sum_{t=1}^{n} [u_t - (\Pi_1 - \Pi_o) Y_{t-1}] Y_{t-1}' - (\widehat{\Pi}_n - \Pi_1) \sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \right]. \quad (3.234)$$

From the results in (3.217), (3.218) and (3.219),

$$\frac{P_n'(k_o) \sum_{t=1}^{n} [u_t - (\Pi_1 - \Pi_o) Y_{t-1}] Y_{t-1}'}{n} = O_p(1). \quad (3.235)$$

From Corollary 3.4.5 and Lemma 3.4.1,

$$\frac{(\widehat{\Pi}_n - \Pi_1)\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'}{n} = \frac{(\widehat{\Pi}_n - \Pi_1)B_n D_n \sum_{t=1}^{n} Z_{t-1}Z_{t-1}'Q'^{-1}}{n} = O_p(1). \quad (3.236)$$

Using the results in (3.234), (3.235) and (3.236), we deduce that

$$\left\| \frac{1}{n} \sum\nolimits_{t=1}^{n} (\Delta Y_t - P_n \widehat{T}_n Y_{t-1})' P_n(k_o) Y_{t-1}' \right\| = O_p(1). \quad (3.237)$$

While under Lemma 3.4.3.(c) and $n^{\omega}\lambda_{r,k_o,n} \to \infty$,

$$\frac{\lambda_{r,k_o,n}}{2||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} = \frac{n^{\omega}\lambda_{r,k_o,n}}{2||n\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} \to_p \infty. \quad (3.238)$$

Combining the results in (3.233), (3.237) and (3.238), we deduce that

$$\Pr\Big(Q_n(k_o)\widehat{\Pi}_n = 0\Big) \to 1 \text{ as } n \to \infty.$$

As the above result holds for any k_o such that $r_o < k_o \leq m$, this finishes the proof.

■

Let $P_{r_o,n}$ and $Q_{r_o,n}$ be the first r_o columns of P_n and the first r_o rows of Q_n respectively. Let $P_{r_o-r_1,n}$ and $Q_{r_o-r_1,n}$ be the last $r_o - r_1$ columns of $P_{r_o,n}$ and the last $r_o - r_1$ rows of $Q_{r_o,n}$ respectively. Under Lemma 3.4.3.(c),

$$\begin{aligned} Q_{r_o-r_1,n}\widehat{\Pi}_n B_n &= Q_{r_o-r_1,n}(\widehat{\Pi}_n - \widehat{\Pi}_{1st})B_n + Q_{r_o-r_1,n}(\widehat{\Pi}_{1st} - \Pi_1)B_n + Q_{r_o-r_1,n}\Pi_1 B_n \\ &= \sqrt{n}Q_{r_o-r_1,n}\Pi_1 Q^{-1} + O_p(1) \\ &= \sqrt{n}Q_{r_o-r_1,n}(\Pi_1 - \widehat{\Pi}_{1st})Q^{-1} + \sqrt{n}Q_{r_o-r_1,n}\widehat{\Pi}_{1st}Q^{-1} + O_p(1) \\ &= \sqrt{n}\Lambda_{r_o-r_1,n}Q_{r_o-r_1,n}Q^{-1} + O_p(1) = O_p(1) \quad (3.239) \end{aligned}$$

where $\Lambda_{r_o-r_1,n}$ is a diagonal matrix with the (r_1+1)-th to the r_o-th eigenvalues of $\widehat{\Pi}_{1st}$. Let $\widehat{T}_{\alpha,n}$ be the first r_o rows of $\widehat{T}_n = Q_n\widehat{\Pi}_n$, then $\widehat{T}_{\alpha,n} = Q_{r_o,n}\widehat{\Pi}_n$. Define $T'_{\alpha,n} = \left[\Pi'_1 Q'_{\widetilde{\alpha},n}, \mathbf{0}_{m\times(r_o-r_1)}\right]$, then

$$\left(\widehat{T}_{\alpha,n} - T_{\alpha,n}\right) B_n = \begin{bmatrix} Q_{\widetilde{\alpha},n}\left(\widehat{\Pi}_n - \Pi_1\right) B_n \\ Q_{r_o-r_1,n}\widehat{\Pi}_n B_n \end{bmatrix} = O_p(1) \tag{3.240}$$

where the last equality is by Corollary 3.4.5 and (3.239).

Proof of Corollary 3.4.7. Using the results of Corollary 3.4.6, we can rewrite the LS shrinkage estimation problem as

$$\widehat{T}_n = \underset{T\in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - P_n T Y_{t-1}\|^2 + n\sum\nolimits_{k=1}^{r_o} \lambda_{r,k,n} \|T(k)\| / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega} \tag{3.241}$$

with the constraint $T(k) = 0$ for $k = r_o+1, .., m$. Recall that $\widehat{T}_{\alpha,n}$ is the first r_o rows of $\widehat{T}_n$, then the problem in (3.241) can be rewritten as

$$\widehat{T}_{\alpha,n} = \underset{T_\alpha\in R^{r_o\times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - P_{r_o,n} T_\alpha Y_{t-1}\|^2 + n\sum\nolimits_{k=1}^{r_o} \lambda_{r,k,n} \|T_\alpha(k)\| / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega} \tag{3.242}$$

where $P_{r_o,n}$ is the first r_o columns of P_n.

Let $u_n^* = (\widehat{T}_{\alpha,n} - T_{\alpha,n})B_n$ and note that the last $r_o - r_1$ rows of $T_{\alpha,n}$ are zeros. By definition, u_n^* is the minimizer of

$$\begin{aligned} V_n(U) &= \sum_{t=1}^{n}\left[\left\|\Delta Y_t - P_{r_o,n}(UB_n^{-1} + T_{\alpha,n})Y_{t-1}\right\|^2 - \|\Delta Y_t - P_{r_o,n}T_{\alpha,n}Y_{t-1}\|^2\right] \\ &\qquad + n\sum\nolimits_{k=1}^{r_o}\lambda_{r,k,n}\left[\left\|UB_n^{-1} + T_{\alpha,n})(k)\right\| - \|T_{\alpha,n}(k)\|\right] / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega} \\ &= V_{1,n}(U) + n\sum\nolimits_{k=1}^{r_o}\lambda_{r,k,n}\left[\left\|UB_n^{-1} + T_{\alpha,n})(k)\right\| - \|T_{\alpha,n}(k)\|\right] / ||\phi_k(\widehat{\Pi}_{1st})||^{\omega}. \end{aligned}$$

For any U in some compact subset of $R^{r_o\times m}$, $n^{\frac{1}{2}}UD_nQ = O(1)$. Thus $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} =$

$o_p(1)$ and Lemma 3.4.3.d imply that

$$n\lambda_{r,k,n}\left|\frac{\|(UB_n^{-1}+T_{\alpha,n})(k_o)\|-\|T_{\alpha,n}(k_o)\|}{||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}}\right| \leq \frac{n^{\frac{1}{2}}\lambda_{r,k,n}\left\|n^{\frac{1}{2}}(UB_n^{-1})(k_o)\right\|}{||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} = o_p(1) \tag{3.243}$$

for $k_o = 1, ..., r_1$. On the other hand, $n^{\frac{1+\omega}{2}}\lambda_{r,k,n} = o(1)$ and Lemma 3.4.3.d imply that

$$n\lambda_{r,k,n}\left|\frac{\|(UB_n^{-1}+T_{\alpha,n})(k_o)\|-\|T_{\alpha,n}(k_o)\|}{||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}}\right| \leq \frac{n^{\frac{1+\omega}{2}}\lambda_{n}\left\|n^{\frac{1}{2}}(UB_n^{-1})(k_o)\right\|}{||n^{\frac{1}{2}}\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} = o_p(1) \tag{3.244}$$

for any $k_o = 1, ..., r_o$. Moreover, we can rewrite $V_{1,n}(U)$ as

$$V_{1,n}(U) = A_{n,t}(U) - 2B_{n,t}(U)$$

where

$$A_{n,t}(U) \equiv vec\,(U)'\left(B_n^{-1}\sum\nolimits_{t=1}^{n} Y_{t-1}Y_{t-1}'B_n^{\prime-1}\otimes P_{r_o,n}'P_{r_o,n}\right)vec\,(U)$$

and

$$B_{n,t}(U) \equiv vec\,(U)'\,vec\left[P_{r_o,n}'\sum\nolimits_{t=1}^{n}\left(\Delta Y_t - P_{r_o,n}T_{\alpha,n}Y_{t-1}\right)Y_{t-1}'B_n^{\prime-1}\right].$$

It is clear that $V_{1,n}(U)$ is minimized at

$$\begin{aligned} U_n^* &= (P_{r_o,n}'P_{r_o,n})^{-1}P_{r_o,n}'\sum_{t=1}^{n}\left(\Delta Y_t - P_{r_o,n}T_{\alpha,n}Y_{t-1}\right)Y_{t-1}'\left(\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\right)^{-1}B_n \\ &= \left[(P_{r_o,n}'P_{r_o,n})^{-1}P_{r_o,n}'\widehat{\Pi}_{1st} - T_{\alpha,n}\right]B_n. \end{aligned}$$

By definition, $P_n = [P_{r_o,n}, P_{m-r_o,n}]$, where $P_{r_o,n}$ and $P_{m-r_o,n}$ are the right normal-

ized eigenvectors of the largest r_o and smallest $m-r_o$ eigenvalues of $\widehat{\Pi}_{1st}$ respectively. From Lemma 3.4.3.(c) and (d), we deduce that $P'_{r_o,n}P_{m-r_o,n} = 0$ w.p.a.1. Thus, we can rewrite U_n^* as

$$U_n^* = \left[(P'_{r_o,n}P_{r_o,n})^{-1}P'_{r_o,n}P_nQ_n\widehat{\Pi}_{1st} - T_{\alpha,n}\right]B_n = \left(Q_{r_o,n}\widehat{\Pi}_{1st} - T_{\alpha,n}\right)B_n$$

w.p.a.1. Results in (3.243) and (3.244) imply that $u_n^* = U_n^* + o_p(1)$. Thus the limiting distribution of the last $r_o - r_1$ rows of u_n^* is identical to the limiting distribution of the last $r_o - r_1$ rows of U_n^*. Let $U^*_{r_o-r_1,n}$ be the last $r_o - r_1$ rows of U_n^*, then by definition

$$Q_{r_o-r_1,n}\widehat{\Pi}_nB_n = U^*_{r_o-r_1,n} + o_p(1) = \Lambda_{r_o-r_1,n}Q_{r_o-r_1,n}B_n + o_p(1) \tag{3.245}$$

where $\Lambda_{r_o-r_1,n} \equiv diag\left[\phi_{r_1+1}(\widehat{\Pi}_{1st}), ..., \phi_{r_o}(\widehat{\Pi}_{1st})\right]$. From (3.245) and Lemma 3.4.3, we obtain

$$n^{\frac{1}{2}}Q_{r_o-r_1,n}\widehat{\Pi}_n = n^{\frac{1}{2}}\Lambda_{r_o-r_1,n}Q_{r_o-r_1,n} + o_p(1) = \Lambda_{r_o-r_1}(\widetilde{\phi}')Q_{r_o-r_1,o} + o_p(1) \tag{3.246}$$

where $\Lambda_{r_o-r_1}(\widetilde{\phi}') \equiv diag(\widetilde{\phi}'_{r_1+1}, ..., \widetilde{\phi}'_{r_o})$ is a non-degenerated full rank random matrix, and $Q_{r_o-r_1,o}$ denotes the probability limit of $Q_{r_o-r_1,n}$ and it is a full rank matrix. From (3.246), we deduce that

$$\limsup_{n\to\infty} \Pr\left(n^{\frac{1}{2}}Q_{r_o-r_1,n}\widehat{\Pi}_n = 0\right) = 0$$

which finishes the proof. ■

3.9.3 Proof of Main Results in Section 3.5

Proof of Lemma 3.5.2. (a). We start by defining $\widehat{S}_{uy_1} = \frac{1}{n}\sum_{t=1}^{n} u_t Y'_{t-1}$ and $\widehat{S}_{ux_0} = \frac{1}{n}\sum_{t=1}^{n} u_t \Delta X'_{t-1}$. From the expression in (3.58), we get

$$\begin{aligned} & \left[(\widehat{\Pi}_{1st}, \widehat{B}_{1st}) - (\Pi_o, B_o)\right] Q_B^{-1} D_{n,B} \\ = & \begin{pmatrix} \widehat{S}_{uy_1} & \widehat{S}_{ux_0} \end{pmatrix} Q'_B D_{n,B}^{-1} \left[D_{n,B}^{-1} Q_B \begin{pmatrix} \widehat{S}_{y_1y_1} & \widehat{S}_{y_1x_0} \\ \widehat{S}_{x_0y_1} & \widehat{S}_{x_0x_0} \end{pmatrix} Q'_B D_{n,B}^{-1} \right]^{-1} \end{aligned} \tag{3.247}$$

Note that

$$\begin{pmatrix} \widehat{S}_{uy_1} & \widehat{S}_{ux_0} \end{pmatrix} Q'_B D_{n,B}^{-1} = U \left[Q_B \begin{pmatrix} Y_- \\ \Delta X \end{pmatrix} \right]' D_{n,B}^{-1} = \begin{pmatrix} n^{-\frac{1}{2}} U Z'_3 & n^{-1} U Z'_2 \end{pmatrix} \tag{3.248}$$

and

$$D_{n,B}^{-1} Q_B \begin{pmatrix} \widehat{S}_{y_1y_1} & \widehat{S}_{y_1x_0} \\ \widehat{S}_{x_0y_1} & \widehat{S}_{x_0x_0} \end{pmatrix} Q'_B D_{n,B}^{-1} = \begin{pmatrix} n^{-1}\sum_{t=1}^{n} Z_{3,t} Z'_{3,t} & n^{-\frac{3}{2}}\sum_{t=1}^{n} Z_{3,t} Z'_{2,t} \\ n^{-\frac{3}{2}}\sum_{t=1}^{n} Z_{2,t} Z'_{3,t} & n^{-2}\sum_{t=1}^{n} Z_{2,t} Z'_{2,t} \end{pmatrix}, \tag{3.249}$$

where $Z_3 = (Z_{3,0}, ..., Z_{3,n-1})$ and $Z_2 = (Z_{2,0}, ..., Z_{2,n-1})$. Now the result in (3.60) follows by applying the Lemma 3.5.1.

(b). This result follows directly by the consistency of $\widehat{\Pi}_{1st}$ and CMT.

(c). Define $S_n(\phi) = \phi I_m - \widehat{\Pi}_{1st}$, then

$$|S_n(\phi)| = |\beta'_o S_n(\phi) \beta_o| \times \left| \beta'_{o\perp} \left\{ S_n(\phi) - S_n(\phi)\beta_o \left[\beta'_o S_n(\phi)\beta_o\right]^{-1} \beta'_o S_n(\phi) \right\} \beta_{o\perp} \right|. \tag{3.250}$$

Let $\mu_k^* = n\phi_k(\widehat{\Pi}_{1st})$ $(k = r_o+1, ..., m)$, using similar arguments in the proof of Lemma

3.3.1.(c), we deduce that μ_k^* is a solution of the equation

$$0 = \left| \beta_{o\perp}' \left\{ S_n(\mu) - S_n(\mu)\beta_o \left[\beta_o' S_n(\mu)\beta_o\right]^{-1} \beta_o' S_n(\mu) \right\} \beta_{o\perp} \right|, \tag{3.251}$$

where $S_n(\mu) = \frac{\mu}{n} I_m - \widehat{\Pi}_{1st}$. Using the results in (a), we can show that

$$\begin{aligned} & \beta_{o\perp}' \left\{ S_n(\mu) - S_n(\mu)\beta_o \left[\beta_o' S_n(\mu)\beta_o\right]^{-1} \beta_o' S_n(\mu) \right\} \beta_{o\perp} \\ = \; & \frac{\mu}{n} I_{m-r_o} - \beta_{o\perp}' \left[I_m - \alpha_o \left(\beta_o'\alpha_o\right)^{-1} \beta_o' + o_p(1) \right] \left(\widehat{\Pi}_{1st} - \Pi_o \right) \beta_{o\perp}. \end{aligned} \tag{3.252}$$

Using the definitions of H_1 and H_2 in the proof of Lemma 3.3.1.(c), we can deduce that

$$nH_1 Q \left(\widehat{\Pi}_{1st} - \Pi_o \right) Q^{-1} H_2' = H_1 \left(QU\widehat{M}_0 Y_-' Q' D_n^{-1} \right) \left(D_n^{-1} QY_- \widehat{M}_0 Y_-' Q' D_n^{-1} \right)^{-1} H_2' \tag{3.253}$$

where under Lemma 3.5.1,

$$\begin{aligned} D_n^{-1} QY_- \widehat{M}_0 Y_-' Q' D_n^{-1} &= D_n^{-1} Z_- Z_-' D_n^{-1} - n^{-1} D_n^{-1} Z_- \Delta X' \widehat{S}_{x_0 x_0}^{-1} \Delta X Z_-' D_n^{-1} \\ &\to_d \begin{pmatrix} \Sigma_{z_1 z_1} - \Sigma_{z_1 \Delta x} \Sigma_{\Delta x \Delta x}^{-1} \Sigma_{\Delta x z_1} & 0 \\ 0 & \int B_{w_2} B_{w_2}' \end{pmatrix} \end{aligned} \tag{3.254}$$

and

$$\begin{aligned} U\widehat{M}_0 Y_-' Q' D_n^{-1} &= U Z_-' D_n^{-1} - n^{-1} U \Delta X' \widehat{S}_{x_0 x_0}^{-1} \Delta X Z_-' D_n^{-1} \\ &\to_d \begin{pmatrix} B_{u,z_1} - B_{u,\Delta x} \Sigma_{\Delta x \Delta x}^{-1} \Sigma_{\Delta x z_1} & \left(\int B_{w_2} dB_u' \right)' \end{pmatrix}. \end{aligned} \tag{3.255}$$

Using the results in (3.254) and (3.255), we obtain

$$nH_1Q\left(\widehat{\Pi}_{1st}-\Pi_o\right)Q^{-1}H_2' \to_d (\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\left(\int B_{w_2}dB_{w_2}'\right)'\left(\int B_{w_2}B_{w_2}'\right)^{-1}(\alpha_{o,\perp}'\beta_{o,\perp}). \tag{3.256}$$

Then, from (3.252)-(3.256), we obtain

$$\begin{aligned}&\left|n\beta_{o\perp}'\left\{S_n(\mu)-S_n(\mu)\beta_o\left[\beta_o'S_n(\mu)\beta_o\right]^{-1}\beta_o'S_n(\mu)\right\}\beta_{o\perp}\right| \\ \to_d &\left|\mu I_{m-r_o}-\left(\int B_{w_2}dB_{w_2}'\right)'\left(\int B_{w_2}B_{w_2}'\right)^{-1}\right|,\end{aligned} \tag{3.257}$$

uniformly over K. The result in (c) follows from (3.257) and by continuous mapping theorem. ■

Proof of Lemma 3.5.3. Let $\Theta=(\Pi,B)$ and

$$\begin{aligned}V_n(\Theta) &= \sum_{t=1}^{n}\left\|\Delta Y_t-\Pi Y_{t-1}-\sum\nolimits_{j=1}^{p}B_j\Delta Y_{t-j}\right\|^2 \\ &\quad+\sum\nolimits_{j=1}^{p}\frac{n\lambda_{b,j,n}}{||\widehat{B}_{j,1st}||^{\omega}}\left\|B_j\right\|+\sum\nolimits_{k=1}^{m}\frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}\left\|\Phi_{n,k}(\Pi)\right\|.\end{aligned}$$

Set $\widehat{\Theta}_n=(\widehat{\Pi}_n,\widehat{B}_n)$ and define an infeasible estimator $\widetilde{\Theta}_n=(\Pi_{n,f},B_o)$, where $\Pi_{n,f}$ is defined in (3.16). Then by definition

$$(\widetilde{\Theta}_n-\Theta_o)Q_B^{-1}D_{n,B}=(\Pi_{n,f}-\Pi_o,0)Q_B^{-1}D_{n,B}=O_p(1) \tag{3.258}$$

where the last equality is by (3.19).

By definition $V_n(\widehat{\Theta}_n) \leq V_n(\widetilde{\Theta}_n)$, so that

$$\begin{aligned}
&\left\{vec\left[(\widetilde{\Theta}_n - \widehat{\Theta}_n)Q_B^{-1}D_{n,B}\right]\right\}' W_n \left\{vec\left[(\widetilde{\Theta}_n - \widehat{\Theta}_n)Q_B^{-1}D_{n,B}\right]\right\} \\
&+2\left\{vec\left[(\widetilde{\Theta}_n - \widehat{\Theta}_n)Q_B^{-1}D_{n,B}\right]\right\}' W_n \left\{vec\left(D_{n,B}^{-1}\sum\nolimits_{t=1}^{n} Z_{t-1}u_t'\right)\right\} \\
&+2\left\{vec\left[(\widetilde{\Theta}_n - \widehat{\Theta}_n)Q_B^{-1}D_{n,B}\right]\right\}' W_n \left\{vec\left[(\Theta_o - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right]\right\} \\
\leq\ & (d_{1,n} + d_{2,n}) \qquad (3.259)
\end{aligned}$$

where

$$\begin{aligned}
W_n &= D_{n,B}^{-1}\sum_{t=1}^{n} Z_{t-1}Z_{t-1}'D_{n,B}^{-1} \otimes I_{m(p+1)}, \\
d_{1,n} &= n\sum\nolimits_{j\in\mathcal{S}_B^c} \lambda_{b,j,n}\frac{\|B_{o,j}\| - ||\widehat{B}_{n,j}||}{||\widehat{B}_{j,1st}||^{\omega}}, \\
d_{2,n} &= n\sum\nolimits_{k\in\mathcal{S}_\phi^c} \lambda_{r,k,n}\frac{\|\Phi_{n,k}(\Pi_{n,f})\| - ||\Phi_{n,k}(\widehat{\Pi}_n)||}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}}.
\end{aligned}$$

Applying the Cauchy-Schwarz inequality to (3.259), we deduce that

$$\mu_n\left\|(\widehat{\Theta}_n - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\|^2 - \left\|(\widehat{\Theta}_n - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\|(c_{1,n} + c_{2,n}) \leq (d_{1,n} + d_{2,n}), \qquad (3.260)$$

where μ_n denotes the smallest eigenvalue of W_n, which is bounded away from zero w.p.a.1,

$$c_{1,n} = \left\|D_{n,B}^{-1}\sum\nolimits_{t=1}^{n} Z_{t-1}u_t'\right\| \text{ and } c_{2,n} = \|W_n\|\left\|(\Theta_o - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\|. \qquad (3.261)$$

By the definition of the penalty function, Lemma 3.5.2 and the Slutsky Theorem, we

find that

$$d_{1,n} \leq \sum_{j\in\mathcal{S}_B^c} \frac{n\lambda_{b,j,n}\left\|B_{o,j}\right\|}{||\widehat{B}_{j,1st}||^{\omega}} = O_p(n\delta_{b,n}) \text{ and} \tag{3.262}$$

$$d_{2,n} \leq \sum_{k\in\mathcal{S}_\phi^c} \frac{n\lambda_{r,k,n}\left\|\Phi_{n,k}(\Pi_{n,f})\right\|}{||\phi_k(\widehat{\Pi}_{1st})||^{\omega}} = O_p(n\delta_{r,n}). \tag{3.263}$$

Using Lemma 3.5.1 and (3.258), we obtain

$$c_{1,n} = O_p(1) \text{ and } c_{2,n} = O_p(1). \tag{3.264}$$

From the inequality in (3.260), the results in (3.262), (3.263) and (3.264), we deduce that

$$\left\|(\widehat{\Theta}_n - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\| = O_P(1 + n^{1/2}\delta_{b,n}^{1/2} + n^{1/2}\delta_{r,n}^{1/2}).$$

which implies $||\widehat{\Theta}_n - \widetilde{\Theta}_n|| = O_P(n^{-1/2} + \delta_{b,n}^{1/2} + \delta_{r,n}^{1/2}) = o_p(1)$. This shows the consistency of $\widehat{\Theta}_n$.

We next derive the convergence rate of the LS shrinkage estimator $\widehat{\Theta}_n$. Using the similar argument in the proof of Theorem 1.3.6, we get

$$|d_{1,n}| \leq cn^{\frac{1}{2}}\delta_{b,n}\left\|\left(\widehat{\Theta}_n - \Theta_o\right)Q_B^{-1}D_{n,B}\right\| \tag{3.265}$$

and

$$|d_{2,n}| \leq cn^{\frac{1}{2}}\delta_{r,n}\left\|\left(\widehat{\Theta}_n - \Theta_o\right)Q_B^{-1}D_{n,B}\right\|. \tag{3.266}$$

Combining the results in (3.265)-(3.266), we get

$$|d_{1,n} + d_{2,n}| \leq cn^{\frac{1}{2}}\delta_n\left\|\left(\widehat{\Theta}_n - \Theta_o\right)Q_B^{-1}D_{n,B}\right\| \tag{3.267}$$

where $\delta_n = \delta_{b,n} + \delta_{r,n}$. From the inequality in (3.260) and the result in (3.267),

$$\mu_n \left\|(\widehat{\Theta}_n - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\|^2 - \left\|(\widehat{\Theta}_n - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\| (c_{1,n} + c_{2,n} + n^{\frac{1}{2}}\delta_n) \leq 0, \quad (3.268)$$

which together with (3.264) implies that $\left\|(\widehat{\Theta}_n - \widetilde{\Theta}_n)Q_B^{-1}D_{n,B}\right\| = O_p(1+n^{\frac{1}{2}}\delta_n)$. This finishes the proof. ■

Proof of Theorem 3.5.4. The first result can be proved using similar arguments in the proof of Theorem 3.3.4. Specifically, we rewrite the LS shrinkage estimation problem as

$$\begin{aligned}(\widehat{T}_n, \widehat{B}_n) &= \underset{T,B_1,...,B_p \in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \left\|\Delta Y_t - P_n T Y_{t-1} - \sum\nolimits_{j=1}^{p} B_j \Delta Y_{t-j}\right\|^2 \\ &\quad + \sum\nolimits_{k=1}^{m} \frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} \|T(k)\| + \sum\nolimits_{j=1}^{p} \frac{n\lambda_{b,j,n}}{||\widehat{B}_{j,1st}||^\omega} \|B_j\|. \quad (3.269)\end{aligned}$$

By definition, $\widehat{\Pi}_n = P_n\widehat{T}_n$ and $\widehat{T}_n = Q_n\widehat{\Pi}_n$ for all n. Results in (3.62) follows if we can show that the last $m - r_o$ rows of $\widehat{T}_n$ are estimated as zeros w.p.a.1.

The KKT optimality conditions for $\widehat{T}_n$ are (i) if $\widehat{T}_n(k) \neq 0$

$$\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j}\Delta Y_{t-j})' P_n(k) Y_{t-1}' = \frac{n\lambda_{r,k,n}\widehat{T}_n(k)}{2||\phi_k(\widehat{\Pi}_{1st})||^\omega ||\widehat{T}_n(k)||}; \quad (3.270)$$

and (ii) if $\widehat{T}_n(k) = 0$

$$\left\|\frac{1}{n}\sum_{t=1}^{n}(\Delta Y_t - P_n\widehat{T}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j}\Delta Y_{t-j})' P_n(k) Y_{t-1}'\right\| < \frac{\lambda_{r,k,n}}{2||\phi_k(\widehat{\Pi}_{1st})||^\omega} \quad (3.271)$$

for $k = 1, ..., m$. Conditional on the event $\{Q_{\alpha,n}(k_o)\widehat{\Pi}_n \neq 0\}$ for some k_o satisfying

$r_o < k_o \leq m$, we obtain the following equation from the KKT optimality conditions

$$\left\| \frac{1}{n} \sum_{t=1}^{n} (\Delta Y_t - P_n \widehat{T}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j})' P_n(k_o) Y_{t-1}' \right\| = \frac{\lambda_{r,k,n}}{2||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}}. \tag{3.272}$$

The sample average in the left hand side of (3.137) can be rewritten as

$$\begin{aligned} & \frac{1}{n} \sum_{t=1}^{n} (\Delta Y_t - P_n \widehat{T}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j})' P_n(k_o) Y_{t-1}' \\ = & \frac{1}{n} \sum_{t=1}^{n} [\Delta u_t - (\widehat{\Theta}_n - \Theta_o) Q_B^{-1} Z_{t-1}]' P_n(k_o) Y_{t-1}' \\ = & \frac{P_n'(k_o) \sum_{t=1}^{n} \Delta u_t Y_{t-1}'}{n} - \frac{P_n'(k_o)(\widehat{\Theta}_n - \Theta_o) Q_B^{-1} \sum_{t=1}^{n} Z_{t-1} Y_{t-1}'}{n} = O_p(1) \end{aligned} \tag{3.273}$$

where the last equality is by Lemma 3.5.1, Lemma 3.5.2.(d) and Lemma 3.5.3. However, under Lemma 3.5.2.(c)

$$\frac{\lambda_{r,k_o,n}}{2||\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} = \frac{n^{\omega} \lambda_{r,k_o,n}}{2||n\phi_{k_o}(\widehat{\Pi}_{1st})||^{\omega}} \to_p \infty. \tag{3.274}$$

Combining the results in (3.272), (3.273) and (3.274), we deduce that

$$\Pr\Big(Q_{\alpha,n}(k_o) \widehat{\Pi}_n = 0 \Big) \to 1 \text{ as } n \to \infty.$$

As the above result holds for any k_o such that $r_o < k_o \leq m$, this finishes the proof of (3.62).

We next show the second result. The LS shrinkage estimators of the transient dynamic matrices satisfy the following KKT optimality conditions: (i) if $\widehat{B}_{n,j} \neq 0$

$$\sum_{t=1}^{n} (\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}) \Delta Y_{t-j}' = \frac{n \lambda_{b,j,n} \widehat{B}_{n,j}}{2||\widehat{B}_{1st,j}||^{\omega} ||\widehat{B}_{n,j}||}; \tag{3.275}$$

and (ii) if $\widehat{B}_{n,j} = 0$

$$\left\| \frac{1}{n} \sum_{t=1}^{n} (\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}) \Delta Y'_{t-j} \right\| < \frac{\lambda_{b,j,n} \widehat{B}_{n,j}}{2||\widehat{B}_{1st,j}||^{\omega} ||\widehat{B}_{n,j}||} \tag{3.276}$$

for any $j = 1, ..., p$. On the event $\{\widehat{B}_{n,j} \neq \mathbf{0}_{m \times m}\}$ for some $j \in \mathcal{S}_B^c$, we get the following equation from the optimality conditions,

$$\left\| n^{-\frac{1}{2}} \sum_{t=1}^{n} (\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}) \Delta Y'_{t-j} \right\| = \frac{n^{\frac{1}{2}} \lambda_{b,j,n}}{2||\widehat{B}_{1st,j}||^{\omega}}. \tag{3.277}$$

The sample average in the left hand side of (3.277) can be rewritten as

$$\begin{aligned} & n^{-\frac{1}{2}} \sum_{t=1}^{n} (\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}) \Delta Y'_{t-j} \\ = \; & n^{-\frac{1}{2}} \sum_{t=1}^{n} [\Delta u_t - (\widehat{\Theta}_n - \Theta_o) Q_B^{-1} Z_{t-1}] \Delta Y'_{t-j} \\ = \; & n^{-\frac{1}{2}} \sum_{t=1}^{n} \Delta u_t \Delta Y'_{t-j} - n^{-\frac{1}{2}} (\widehat{\Theta}_n - \Theta_o) Q_B^{-1} \sum_{t=1}^{n} Z_{t-1} \Delta Y'_{t-j} = O_p(1) \end{aligned} \tag{3.278}$$

where the last equality is by Lemma 3.5.1, Lemma 3.5.2.(d) and Lemma 3.5.3. However, as $n^{\frac{1+\omega}{2}} \lambda_{b,j,n} \to \infty$ for any $j \in \mathcal{S}_B^c$, it follows by Lemma 3.5.2 and the Slutsky Theorem that

$$\frac{n^{\frac{1+\omega}{2}} \lambda_{b,j,n}}{2||n^{\frac{1}{2}} \widehat{B}_{1st,j}||^{\omega}} \to_p \infty. \tag{3.279}$$

Now, using the results in (3.277), (3.278) (3.279), we can deduce that the event $\{\widehat{B}_{n,j} \neq \mathbf{0}_{m \times m}\}$ $\forall j \in \mathcal{S}_B^c$ has zero probability with $n \to \infty$, which finishes the proof.

■

Proof of Theorem 3.5.5. Follow the similar arguments in the proof of Theorem 3.3.6, we normalize β_o as $\beta_o = [I_{r_o}, O_{r_o}]'$ to ensure identification, where O_{r_o} is some

$r_o \times (m - r_o)$ matrix such that $\Pi_o = \alpha_o \beta_o' = [\alpha_o, \alpha_o O_{r_o}]$. From Lemma 3.5.3, we have

$$\left(\begin{array}{ccc} n^{\frac{1}{2}}(\widehat{\Pi}_n - \Pi_o)\alpha_o(\beta_o'\alpha_o)^{-1} & n^{\frac{1}{2}}(\widehat{B}_n - B_o) & n(\widehat{\Pi}_n - \Pi_o)\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1} \end{array} \right) = O_p(1),$$

which implies that

$$\begin{aligned} n\left(\widehat{O}_n - O_o\right) &= O_p(1), && (3.280) \\ n^{\frac{1}{2}}(\widehat{B}_n - B_o) &= O_p(1), && (3.281) \\ n^{\frac{1}{2}}\left(\widehat{\alpha}_n - \alpha_o\right) &= O_p(1). && (3.282) \end{aligned}$$

From the results of Theorem 3.5.4, we deduce that $\widehat{\alpha}_n$, $\widehat{\beta}_n$ and $\widehat{B}_{\mathcal{S}_B}$ minimize the following criterion function w.p.a.1,

$$\begin{aligned} V_n(\Theta_{\mathcal{S}}) &= \sum_{t=1}^{n} \left\| \Delta Y_t - \alpha\beta' Y_{t-1} - \sum_{j \in \mathcal{S}_B} B_j \Delta Y_{t-j} \right\|^2 \\ &\quad + \sum_{k \in \mathcal{S}_\phi} \frac{n\lambda_{r,k,n}}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} \left\| \Phi_{n,k}(\alpha\beta') \right\| + \sum_{j \in \mathcal{S}_B} \frac{n\lambda_{b,j,n}}{||\widehat{B}_{j,1st}||^\omega} \left\| B_j \right\|. \end{aligned}$$

Define $U_{1,n}^* = \sqrt{n}\left(\widehat{\alpha}_n - \alpha_o\right)$, $U_{2,n} = \left[\mathbf{0}_{r_o}, U_{2,n}^*\right]'$, where $U_{2,n}^* = n\left(\widehat{O}_n - O_o\right)$ and $U_{3,n}^* = \sqrt{n}\left(\widehat{B}_{\mathcal{S}_B} - B_{o,\mathcal{S}_B}\right)$, then

$$\begin{aligned} & \left[\left(\widehat{\Pi}_n - \Pi_o\right), \left(\widehat{B}_{\mathcal{S}_B} - B_{o,\mathcal{S}_B}\right)\right] Q_{\mathcal{S}}^{-1} D_{n,\mathcal{S}} \\ = & \left[n^{-\frac{1}{2}}\widehat{\alpha}_n U_{2,n}\alpha_o(\beta_o'\alpha_o)^{-1} + U_{1,n}^*, U_{3,n}^*,\ \widehat{\alpha}_n U_{2,n}\beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right]. \end{aligned}$$

Denote

$$\Pi_n(U) = \left[n^{-\frac{1}{2}}\widehat{\alpha}_n U_2 \alpha_o(\beta_o'\alpha_o)^{-1} + U_1, U_3, \widehat{\alpha}_n U_2 \beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right],$$

then by definition, $U_n^* = \left(U_{1,n}^*, U_{2,n}^*, U_{3,n}^*\right)$ minimizes the following criterion function

$$\begin{aligned} V_n(U) &= \sum_{t=1}^{n}\left(\left\|u_t - \Pi_n(U)D_{n,\mathcal{S}}^{-1}Z_{\mathcal{S},t-1}\right\|^2 - \|u_t\|^2\right) \\ &\quad + n\sum_{k\in\mathcal{S}_\phi}\lambda_{r,k,n}\frac{\left\|\Phi_{n,k}\left[\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_1 + \Pi_o\right]\right\| - \|\Phi_{n,k}(\Pi_o)\|}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} \\ &\quad + n\sum_{j\in\mathcal{S}_B}\lambda_{b,j,n}\frac{\left\|\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_{j+1} + B_{o,j}\right\| - \|B_{o,j}\|}{||\widehat{B}_{j,1st}||^\omega}. \end{aligned}$$

where $L_j = diag(A_{j,1}, ..., A_{j,d_{\mathcal{S}_B}})$ with $A_{j,j} = I_m$ and $A_{i,j} = 0$ for $i \neq j$ and $j = 1, ..., d_{\mathcal{S}_B}$.

For any compact set $K \in R^{m\times r_o} \times R^{r_o\times(m-r_o)} \times R^{m\times md_{\mathcal{S}_B}}$ and any $U \in K$, there is

$$\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}} = O_p(n^{-\frac{1}{2}}).$$

Hence using similar arguments in the proof of Theorem 3.3.6, we can deduce that

$$n\sum_{k\in\mathcal{S}_\phi}\lambda_{r,k,n}\frac{\left\|\Phi_{n,k}(\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_1 + \Pi_o)\right\| - \|\Phi_{n,k}(\Pi_o)\|}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} = o_p(1) \tag{3.283}$$

and

$$n\sum_{j\in\mathcal{S}_B}\lambda_{b,j,n}\frac{\left\|\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_{j+1} + B_{o,j}\right\| - \|B_{o,j}\|}{||\widehat{B}_{j,1st}||^\omega} = o_p(1) \tag{3.284}$$

uniformly over $U \in K$.

Next, note that

$$\Pi_n(U) \to_p \left[U_1, U_3, \alpha_o U_2(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\right] \equiv \Pi_\infty(U) \tag{3.285}$$

uniformly over $U \in K$. By Lemma 3.5.1 and (3.285), we can deduce that

$$\begin{aligned}
&\sum_{t=1}^{n}\left(\left\|u_t - \Pi_n(U)D_{n,\mathcal{S}}^{-1}Z_{\mathcal{S},t-1}\right\|^2 - \|u_t\|^2\right) \\
\rightarrow_d\ & vec\left[\Pi_\infty(U)\right]'\left[\begin{pmatrix}\Sigma_{z_{3\mathcal{S}}z_{3\mathcal{S}}} & 0\\ 0 & \int B_{w_2}B_{w_2}'\end{pmatrix}\otimes I_m\right]vec\left[\Pi_\infty(U)\right] \\
& -2vec\left[\Pi_\infty(U)\right]'vec\left[(B_{3,m},B_{2,m})\right] \equiv V(U)
\end{aligned} \tag{3.286}$$

uniformly over $U \in K$, where $B_{3,m} = N(0, \Omega_u \otimes \Sigma_{z_{3\mathcal{S}}z_{3\mathcal{S}}})$ and $B_{2,m} = \left(\int B_{w_2}dB_u'\right)'$.

Using similar arguments in the proof of Theorem 3.3.6, we can rewrite $V(U)$ as

$$\begin{aligned}
V(U) \;=\;& vec(U_1,U_3)'\left(\Sigma_{z_{3\mathcal{S}}z_{3\mathcal{S}}}\otimes I_m\right)vec(U_1,U_3) \\
& +vec(U_2)'\left[\beta_{2,o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\int B_{w_2}B_{w_2}'(\beta_{o,\perp}'\alpha_{o,\perp})^{-1}\beta_{2,o,\perp}'\otimes\alpha_o'\alpha_o\right]vec(U_2) \\
& -2vec(U_1,U_3)'vec\left(B_{3,m}\right) - 2vec(U_2)'vec\left[\alpha_o'B_{2,m}(\beta_{o,\perp}'\alpha_{o,\perp})^{-1}\beta_{2,o,\perp}'\right]
\end{aligned} \tag{3.287}$$

The expression in (3.287) makes it clear that $V(U)$ is uniquely minimized at (U_1^*, U_2^*, U_3^*), where

$$(U_1^*, U_3^*) = B_{3,m}\Sigma_{z_{3\mathcal{S}}z_{3\mathcal{S}}}^{-1} \text{ and } U_2^* = (\alpha_o'\alpha_o)^{-1}\alpha_o'B_{2,m}\left(\int B_{w_2}B_{w_2}'\right)^{-1}(\alpha_{o,\perp}'\beta_{o,\perp}). \tag{3.288}$$

From (3.280) and (3.282), we can see that U_n^* is asymptotically tight. Invoking the ACMT, we deduce that $U_n^* \rightarrow_d U^*$. The results in (3.65) follow by applying the CMT. ■

References

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

Anderson, T.W. (2002): "Reduced Rank Regression in Cointegrated Models," *Journal of Econometrics*, 106, 203-216

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

Angrist, J., and A. Krueger (1991): "Does compulsory school attendance affect schooling and earnings," *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," *Journal of Econometrics*, 68(1), 29-51.

Belloni, A., D. Chen, V. Chernozhukov, and C. Hansen (2010): "Sparse models and methods for optimal instruments with an application to eminent domain," *Preprint,* arXiv:1010.4345.

Belloni, A., V. Chernozhukov, and C. Hansen (2010): "Lasso methods for Gaussian instrumental variables models," *MIT Working Paper*.

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

Breusch, T., H. Qian, P. Schmidt and D. Wyhowski (1999): "Redundancy of moment conditions," *Journal of Econometrics*, 91 89-111.

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

Caner, M., and K. Knight (2009): "No country for old unit root tests: bridge estimators differentiate between nonstationary versus stationary models and select optimal lag," *Unpublished Manuscript*.

Caner, M., and H. Zhang (2009): "General estimating equations: model selection and estimation with diverging number of parameters," *Unpublished Manuscript.*

Chao, J., and P.C.B. Phillips (1999): "Model selection in partially nonstationary vector autoregressive processes with reduced rank structure," *Journal of Econometrics*, 91(2), 227-271.

Cheng, X., and P.C.B. Phillips (2009): "Semiparametric cointegrating rank selection," *Econometrics Journal*, 12, 83-104.

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.

Gautier, E., and A.B. Tsybakov, (2011): "High-dimensional instrumental variables regression and confidence sets," *Preprint,* arXiv: 1105.2454v2.

Geyer, C. (1994): "On the asymptotics of constrained M-estimation," *Annals of Statistics*, 28(5), 22(4), 1993–2010.

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

Hall, A. R., Inoue, A., Jana, K., and Shin, C. (2007): "Information in generalized method of moments estimation and entropy based moment selection," *Journal of Econometrics*, 138, 488–512.

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.

Johansen, S. (1988): "Statistical analysis of cointegration vectors," *Journal of*

Economic Dynamics and Control, 12(2-3), 231-254.

Johansen, S. (1995): *Likelihood-based inference in cointegrated vector autoregressive models*, Oxford University Press, USA.

Knight, K. (1999): "Epi-convergence in distribution and stochastic equisemicontinuity," *Unpublished Manuscript*.

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

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

Leeb, H., and B. Pötscher (2008): "Sparse estimators and the oracle property, or the return of the Hodges estimator," *Journal of Econometrics*, 142(1), 201-211.

Liao, Z. (2012): "On the robust inference of GMM shrinkage estimation," *Unpublished Manuscript*.

Liao Z., and P.C.B. Phillips (2010): "Reduced rank regression of partially nonstationary vector autoregressive processes under misspecification," *Unpublished Manuscript*.

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.

Pakes, A., and Pollard, D. (1989): "Simulation and the asymptotics of optimization estimators," *Econometrica*, 57, 10271057.

Phillips, P.C.B. (1991): "Optimal inference in cointegrated systems," *Econometrica*, 59(2), 283-306.

Phillips, P.C.B. (1995): "Fully modified least squares and vector autoregression,"

Econometrica, 63(5), 1023-1078.

Phillips, P.C.B. (1996): "Econometric model determination," *Econometrica*, 64(4), 763-812.

Phillips, P.C.B. (1998): "Impulse Response and Forecast Error Variance Asymptotics in Nonstationary VARs," *Journal of Econometrics*, 83, 21-56.

Phillips, P.C.B., and V. Solo (1992): "Asymptotics for linear processes," *Annals of Statistics*, 20(2), 971-1001.

Schmidt, M. (2010): "Graphical model structure learning with L1-regularization," *Thesis*, University of British Columbia.

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.

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