AUTOMATED ESTIMATION OF VECTOR ERROR CORRECTION MODELS

by

Zhipeng Liao and Peter C. B. Phillips

COWLES FOUNDATION PAPER NO. 1476

COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

2015

http://cowles.econ.yale.edu/

Econometric Theory, **31**, 2015, 581–646.
doi:10.1017/S026646661500002X

AUTOMATED ESTIMATION OF VECTOR ERROR CORRECTION MODELS

ZHIPENG LIAO
UC Los Angeles

PETER C. B. PHILLIPS
Yale University, University of Auckland, University of Southampton, and Singapore Management University

Model selection and associated issues of post-model selection inference present well known challenges in empirical econometric research. These modeling issues are manifest in all applied work but they are particularly acute in multivariate time series settings such as cointegrated systems where multiple interconnected decisions can materially affect the form of the model and its interpretation. 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 paper 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. 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. The paper develops the new methods, derives their limit theory, discusses implementation, reports simulations, and presents an empirical illustration with macroeconomic aggregates.

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 (VEC) 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

Comments from two referees and the Co-Editor helped guide the revision of this paper and are gratefully acknowledged. Support from the NSF under Grant Nos. SES 09-56687 and SES 12-58258 is gratefully acknowledged. Address correspondence to Zhipeng Liao, Department of Economics, UC Los Angeles, 8379 Bunche Hall, Mail Stop: 147703, Los Angeles, CA 90095; e-mail: zhipeng.liao@econ.ucla.edu.

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). Both approaches are popular in empirical research.

Information criteria offer certain advantages such as joint determination of the cointegrating rank and autoregressive order, consistent estimation of both order parameters (Chao and Phillips, 1999; Athanasopoulos, Guillen, Issler, and Vahid, 2011), 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, 2009, 2012). Sequential testing procedures have recent enhancements including bootstrap modifications to improve test performance and under certain conditions provide consistent order estimation by adaptation if test size is driven to zero as the sample size expands to infinity. However, these adaptive methods have not been systematically investigated in the VEC framework and there is little research on rate control and testing order, and no asymptotics for such adaptive procedures to offer guidance for empirical implementation. More importantly in the VEC setting, sequential tests involve different test statistics for lags and cointegrating rank, and model selection is inevitably unstable in the sense that different models may be selected when different sequential orders are used. Moreover, general to specific and specific to general testing algorithms encounter obstacles to consistent model selection even when test size is driven to zero (see Section 9 for an example). Finally, while they are appealing to practitioners, all of these methods are nonetheless subject to pre-test bias and post model selection inferential problems (Leeb and Pötscher, 2005).

The present paper explores a different approach. The goal is to liberate the empirical researcher from some of the difficulties of sequential testing and order estimation procedures in inference about cointegrated systems and in policy work that relies on associated 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 nonzero regression coefficients. Caner and Knight (2013) first showed how this type of estimator may be used in a univariate autoregressive model with a potential unit root. 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. When data-based tuning parameter selection is employed, the methods also enable automated implementation making them convenient for empirical practice.

One of the contributions of this paper is to develop new adaptive versions of 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 in this econometric setting is by no means immediate. In particular, multivariate models with some unit roots and cointegration involve dimension reductions and nonlinear restrictions which present new difficulties of both formulation and asymptotics in the Lasso framework that go beyond existing work in the statistics literature such as Yuan, Ekici, Lu, and Monteiro (2007). The present paper contributes to the Lasso and econometric literatures by providing a new penalty function that handles these complications, developing a rigorous limit theory of order selection and estimation for this multivariate nonlinear nonstationary setting, and devising a straightforward method of implementation that is well suited to empirical econometric research. When reduced to the univariate case, our results cover the methodology and implicit unit root test procedure suggested in Caner and Knight (2013) and extend their univariate results to cases where there is misspecification in the transient dynamics.

The paper 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. This formulation is necessary because of the nonlinearities involved in potential reduced rank structures and the interdependence of decision making concerning the form of the transient dynamics and the cointegrating rank structure. The resulting 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 all of the order parameters as a consequence of the fitting procedure.

A novel contribution of the paper in this nonlinear setting where eigenvalues play a key role is the use of a penalty which is a simple convex function of the coefficient matrix. The new penalty makes penalized estimation stable and accurate, facilitates the limit theory, and simplifies implementation because existing code for grouped L-1 penalized estimation can be used for computation. All the theoretical results are rigorously derived in a general nonstationary set-up that allows for unit roots, cointegration, and transient dynamics, which combines with the new penalty formulation to complement recent asymptotic theory for Lasso estimation in stationary vector autoregressive (VAR) models (Song and Bickel, 2009; Kock and Callot, 2012) and multivariate regression (Yuan et al., 2007; Peng et al., 2010).

The paper is organized as follows. Section 2 lays out the model and assumptions and shows how to implement adaptive shrinkage methods in VEC systems. Section 3 considers a simplified first order version of the vector error correction

model (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 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 5 deals with the practically important case of a general VEC system driven by independent identically distributed (iid) shocks, where shrinkage estimation simultaneously performs consistent lag selection, cointegrating rank selection, and optimal estimation of the system coefficients. Section 6 considers adaptive selection of the tuning parameter and Section 7 reports some simulation findings. Section 8 applies our method to an empirical example. Section 9 concludes and outlines some useful extensions of the methods and limit theory to other models. Proofs are given in the Appendix. A Supplement to the paper (Liao and Phillips, 2013) provides supporting lemmas and technical results.

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 corresponding sample average. The notation $\|\cdot\|$ denotes the Euclidean norm and $|A|$ is the determinant of a square matrix A. A' refers to the transpose of any matrix A and $\|A\|_B \equiv ||A'BA||$ for any conformable matrices A and B. I_k and $\mathbf{0}_l$ are used to denote $k \times k$ identity matrix and $l \times l$ zero matrices respectively. The symbolism $A \equiv B$ 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. Following standard convention we frequently write integrals of stochastic processes (V, W) over $[0, 1]$ such as $\left(\int_0^1 V(r)\,dW(r)', \int_0^1 V(r)\,V(r)'\,dr\right)$ in the simple form $\left(\int V dW', \int V V'\right)$.

2. VECTOR ERROR CORRECTION AND ADAPTIVE SHRINKAGE

Throughout this paper we consider the following parametric VEC representation of a cointegrated system

$$\Delta Y_t = \Pi_o Y_{t-1} + \sum_{j=1}^{p} B_{o,j}\,\Delta Y_{t-j} + u_t, \tag{2.1}$$

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 $0 \leq r_o \leq m$, $B_{o,j}$ $(j = 1, \ldots, p)$ are $m \times m$ (transient) coefficient matrices, u_t is an m-vector error term with mean zero and nonsingular covariance

matrix Σ_{uu}, m and p are fixed positive integers. 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 index set of nonzero matrices $B_{o,j}$ $(j = 1, \ldots, p)$ 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, i.e. $\alpha_{o,\perp}$ and $\beta_{o,\perp}$ are full rank $m \times (m - r_o)$ matrices satisfying $\alpha_{o,\perp}' \alpha_o = 0_{(m-r_o)\times m}$ and $\beta_{o,\perp}' \beta_o = 0_{(m-r_o)\times m}$ respectively. Without loss of generality, assume that $\alpha_{o,\perp}' \alpha_{o,\perp} = I_{m-r_o}$ and $\beta_{o,\perp}' \beta_{o,\perp} = I_{m-r_o}$.[1]

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

$$\alpha_o \left(\beta_o' \alpha_o\right)^{-1} \beta_o' + \beta_{o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp}\right)^{-1} \alpha_{o,\perp}' = I_m, \tag{2.2}$$

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

$$Q\Pi_o = \begin{bmatrix} \beta_o' \alpha_o \beta_o' \\ 0 \end{bmatrix} \text{ and } Q\Pi_o Q^{-1} = \begin{bmatrix} \beta_o' \alpha_o & 0 \\ 0 & 0 \end{bmatrix}. \tag{2.3}$$

Under Assumption RR in Section 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 nonzero 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 (2.1) 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 transient dynamics structure, 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 (2.1), the cointegrating vectors, and the transient dynamics. This method does not require knowledge of r_o but does require knowledge of the transient dynamics structure. In addition, a semiparametric approach can be adopted in which r_o is estimated semiparametrically by order selection as in Cheng and Phillips (2010, 2012) followed by fully modified least squares regression to estimate the cointegrating matrix. That approach achieves asymptotically efficient estimation of the long run relations (under Gaussianity) but does not estimate the transient relations.

The present paper explores direct estimation of the parameters of (2.1) by Lasso-type regression. The resulting estimator is a shrinkage estimator that takes account of potential degeneracies in the system involving both long run reduced

rank structures and transient dynamics. Specifically, the least squares (LS) shrinkage estimator of (Π_o, B_o), where $B_o = (B_{o,1}, \ldots, B_{o,p})$ is defined as

$$\left(\widehat{\Pi}_n, \widehat{B}_n\right) = \underset{\Pi, B_1, \ldots, B_p \in R^{m\times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum\nolimits_{j\le p} B_j \Delta Y_{t-j} \right\|^2 + n\sum_{j=1}^{p} \lambda_{b,j,n} \left\| B_j \right\| + n\sum_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\| \right\}, \tag{2.4}$$

where $\lambda_{b,j,n}$ and $\lambda_{r,k,n}$ ($j = 1, \ldots, p$ and $k = 1, \ldots, m$) are tuning parameters that directly control the penalization, $\Phi_{n,k}(\Pi)$ is the k-th row vector of $Q_n\Pi$, and Q_n denotes the normalized left eigenvector matrix of eigenvalues of $\widehat{\Pi}_{1st}$. The matrix $\widehat{\Pi}_{1st}$ is some first step (e.g., OLS) estimate of Π_o. The penalty function on the coefficients B_j ($j = 1, \ldots, p$) of the lagged differences is called a group Lasso penalty (see, Yuan and Lin, 2006). On the other hand, the penalty function on Π is different from the group Lasso, because it works on the rows of the adaptively transformed matrix $Q_n\Pi$, not the rows (or any deterministic functions such as eigenvalues) of Π directly.[2]

Given the tuning parameters, this procedure delivers a one step estimator of the model (2.1) with an implied estimate of the cointegrating rank (based on the number of nonzero rows of $Q_n\widehat{\Pi}_n$) and an implied estimate of the transient dynamic structure (that is, $B_{o,j}$ in B_o with $\left\| B_{o,j} \right\| = 0$ for $j = 1, \ldots, p$) based on the fitted value $\widehat{B}_n$. It is therefore well suited to empirical implementation where information is limited concerning model specification. By definition, the penalized LS estimate is invariant to permutation of the lag differences, which implies that the rank and lag differences selected in the penalized LS estimation are stable regardless the potential structure of the true model. This feature is a particular advantage of Lasso-type model selection methods over traditional sequential testing procedures which typically work from general to specific formulations.

A novel contribution of this paper is that it provides an adaptive penalty function $f(\Pi) = \sum_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\|$, which enables penalized LS estimation in (2.4) to perform rank selection.[3] Importantly, this penalty function differs from those proposed in the statistics literature for dimension reduction in multivariate regression with iid data. For example, Peng et al. (2010) assume that the coefficient matrix has many zero components and suggest dimension reduction by penalizing the estimates of the components in the coefficient matrix with L-1 and L-2 penalty functions. Yuan et al. (2007) propose to penalize the singular values of the estimate of the coefficient matrix with an L-1 penalty to achieve dimension reduction. While this approach is intuitive and the idea of working through the eigenvalues of Π was used independently in our own earlier work, Yuan et al. (2007) provide theory only under an orthonormal regressor design, which is unrealistic in VEC structures with nonstationary data[4].

Let $\Phi'(\Pi_o) = \left[\Phi_1'(\Pi_o), \ldots, \Phi_m'(\Pi_o)\right]$ denote the row vectors of $Q\Pi_o$. When $\{u_t\}_{t\ge 1}$ 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 $\Phi(\Pi_o)$.[5] We show that the shrinkage LS estimator defined in (2.4) enjoys these oracle-like properties, in the sense that the zero components in B_o and $\Phi(\Pi_o)$ are estimated as zeros with probability approaching 1 (w.p.a.1). Thus, Π_o and the nonzero 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 (2.1) 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 nonzero 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 nonzero eigenvalues of Π_o. Specifically, we show that if the estimator of some nonzero 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 tuning parameters $\{\lambda_{r,k,n}\}_{k=1}^m$ can be constructed in the way such that 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 parameters $\{\lambda_{r,k,n}\}_{k=1}^m$ to ensure that only r_o rows of $Q_n\widehat{\Pi}_n$ are adaptively shrunk to zero w.p.a.1.

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 paper progresses.

(1) After preliminary LS estimation of the system, perform a first step GLS shrinkage estimation with adaptive Lasso (c.f. Zou, 2006) type of tuning parameters

$$\lambda_{r,k,n} = \frac{2\log(n)}{n}\left|\left|\phi_k\left(\widehat{\Pi}_{1st}\right)\right|\right|^{-2} \text{ and } \lambda_{b,j,n} = \frac{2m^2\log(n)}{n}\left|\left|\widehat{B}_{j,1st}\right|\right|^{-2}$$

for $k = 1, \ldots, m$ and $j = 1, \ldots, p$, where $||\phi_k(\Pi)||$ denotes the k-th largest modulus of the eigenvalues $\{\phi_k(\Pi)\}_{k=1}^m$ of the matrix Π [6] and $\widehat{B}_{j,1st}$ is some first step (OLS) estimates of $B_{o,j}$ $(j = 1, \ldots, p)$.

(2) Construct adaptive tuning parameters using the first step GLS shrinkage estimates and the formulas in (6.10) and (6.11). Using the adaptive tuning parameters, obtain the GLS shrinkage estimator $(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n})$ of (Π_o, B_o) - see (5.12). 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 $\widehat{B}_{g,n}$.

(3) 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 (2).

3. FIRST ORDER VECM ESTIMATION

This section considers the following simplified first order version of (2.1),

$$\Delta Y_t = \Pi_o Y_{t-1} + u_t = \alpha_o \beta_o' Y_{t-1} + u_t. \tag{3.1}$$

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 (2.1).

We start with the following condition on the innovation u_t.

Assumption 3.1 (WN)**.** $\{u_t\}_{t\geq 1}$ is an m-dimensional iid process with zero mean and nonsingular covariance matrix $\mathbf{\Omega}_u$.

Assumption 3.1 ensures that the full parameter matrix Π_o is consistently estimable in this simplified system. Under Assumption 3.1, partial sums of u_t satisfy the functional law

$$n^{-\frac{1}{2}} \sum_{t=1}^{[n\cdot]} u_t \to_d B_u(\cdot), \tag{3.2}$$

where $B_u(\cdot)$ is a vector of Brownian motion with variance matrix $\mathbf{\Omega}_u$. With no material changes in what follows, the iid condition in **WN** could be weakened to a martingale difference sequence condition provided the functional law (3.2) still holds together with some related weak convergence results needed for the limit theory. Cheng and Phillips (2012) developed such a limit theory while exploring the properties of model selection methods based on information criteria but did not consider penalized regression approaches.

Assumption 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.

Assumption 3.2 leads to the following partial sum Granger representation,

$$Y_t = C\sum_{s=1}^{t} u_s + \alpha_o\left(\beta_o'\alpha_o\right)^{-1} R(L)\beta_o' u_t + CY_0, \tag{3.3}$$

where $C = \beta_{o,\perp}(\alpha_{o,\perp}'\beta_{o,\perp})^{-1}\alpha_{o,\perp}'$. Using the matrix Q, (3.1) transforms as

$$\Delta Z_t = \Xi_o Z_{t-1} + w_t, \tag{3.4}$$

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 $\Xi_o = Q\Pi_o Q^{-1}$. Under Assumption 3.2 and (3.2), we have the functional law

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

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 virtue of Assumption RR and the properties of Q, we know that $\mathcal{S}_\phi = \{1, \ldots, r_o\}$ and $\mathcal{S}_\phi^c = \{r_o + 1, \ldots, m\}$. It follows that consistent selection of the rank of Π_o is equivalent to the consistent recovery of the zero rows in $\Phi(\Pi_o) = Q\Pi_o$.

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

$$\widehat{\Pi}_n = \underset{\Pi \in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \|\Delta Y_t - \Pi Y_{t-1}\|^2 + n\sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \left\|\Phi_{n,k}(\Pi)\right\|. \tag{3.5}$$

We first show the consistency of the LS shrinkage estimate $\widehat{\Pi}_n$.

THEOREM 3.1 (Consistency). *Let* $\delta_{r,n} = \max_{k\in\mathcal{S}_\phi} \lambda_{r,k,n}$, *then under Assumptions WN, RR, and* $\delta_{r,n} = o_p(1)$, *the LS shrinkage estimator* $\widehat{\Pi}_n$ *is consistent, i.e.* $\widehat{\Pi}_n - \Pi_o = o_p(1)$.

When consistent shrinkage estimators are considered, Theorem 3.1 extends Theorem 1 of Caner and Knight (2013) 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 $k = 1, \ldots, m$. Theorem 3.1 implies that the nonzero eigenvalues of Π_o are estimated as nonzeros, which means that the rank of Π_o will not be under-selected. However, consistency of the estimates of the nonzero 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 (nonzero) 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.

THEOREM 3.2 (Rate of Convergence). *Define* $D_n = diag(n^{-\frac{1}{2}} I_{r_o}, n^{-1} I_{m-r_o})$, *then under the conditions of Theorem 3.1, the LS shrinkage 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* $\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} = O_p\left(1 + n^{\frac{1}{2}}\delta_{r,n}\right)$.

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 3.2 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 convergence rate of the LS estimator $\widehat{\Pi}_{1st}$ (see, Lemma A.2 in the appendix). 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.

Let $\mathcal{S}_{n,\phi}$ denote the index set of the nonzero rows of $Q_n\widehat{\Pi}_n$ and its complement $\mathcal{S}_{n,\phi}^c$ be the index set of the zero rows of $Q_n\widehat{\Pi}_n$. We subdivide the matrix Q_n as $Q_n' = \left[Q_{\alpha,n}', Q_{\alpha_\perp,n}'\right]$, where $Q_{\alpha,n}$ and $Q_{\alpha_\perp,n}$ are the first r_o rows and the last $m - r_o$ rows of Q_n respectively. Under Lemma A.2 and Theorem 3.1,

$$Q_{\alpha,n}\widehat{\Pi}_n = Q_{\alpha,n}\widehat{\Pi}_{1st} + o_p(1) = \Lambda_{\alpha,n} Q_{\alpha,n} + o_p(1) \tag{3.6}$$

and similarly

$$Q_{\alpha_\perp,n}\widehat{\Pi}_n = Q_{\alpha_\perp,n}\widehat{\Pi}_{1st} + o_p(1) = \Lambda_{\alpha_\perp,n} Q_{\alpha_\perp,n} + o_p(1) = o_p(1), \tag{3.7}$$

where $\Lambda_{\alpha,n} = diag[\phi_1(\widehat{\Pi}_{1st}), \ldots, \phi_{r_o}(\widehat{\Pi}_{1st})]$ and $\Lambda_{\alpha_\perp,n} = diag[\phi_{r_o+1}(\widehat{\Pi}_{1st}), \ldots, \phi_m(\widehat{\Pi}_{1st})]$. Result in (3.6) implies that the first r_o rows of $Q_n\widehat{\Pi}_n$ are nonzero w.p.a.1., while the results in (3.7) 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 (3.6) we deduce that $\mathcal{S}_\phi \subseteq \mathcal{S}_{n,\phi}$. However, (3.7) is insufficient for showing that $\mathcal{S}_\phi^c \subseteq \mathcal{S}_{n,\phi}^c$, 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 (Super Efficiency). *Suppose that Assumptions WN and RR are satisfied. If* $n^{\frac{1}{2}}\delta_{r,n} = O_p(1)$ *and* $\lambda_{r,k,n} \to_p \infty$ *for* $k \in \mathcal{S}_\phi^c$, *then*

$$\Pr\left(Q_{\alpha_\perp,n}\widehat{\Pi}_n = 0\right) \to 1 \text{ as } n \to \infty. \tag{3.8}$$

Theorem 3.3 requires the tuning parameters related to the zero and nonzero components have different asymptotic behaviors. As we do not have any prior information about the zero and nonzero components, it is clear that some sort of adaptive penalization should appear in the tuning parameters $\{\lambda_{r,k,n}\}_{k=1}^{m}$. Such an adaptive penalty is constructed in (6.1) of Section 6 and sufficient conditions for $n^{\frac{1}{2}}\delta_{r,n} = O_p(1)$ and $\lambda_{r,k,n} \to_p \infty$ for $k \in \mathcal{S}_\phi^c$ are provided in Lemma 6.1.

Combining Theorem 3.1 and Theorem 3.3, we deduce that

$$\Pr\left(\mathcal{S}_{n,\phi} = \mathcal{S}_\phi\right) \to 1, \tag{3.9}$$

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

COROLLARY 3.4. *Under the conditions of Theorem 3.3, we have*

$$\Pr\left(r(\widehat{\Pi}_n) = r_o\right) \to 1 \tag{3.10}$$

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

From Corollary 3.4, 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 3.5 (Limiting Distribution). *Suppose that conditions of Theorem 3.3 and* $n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$ *are satisfied. We have*

$$\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \left(B_{m,1}\, \alpha_o (\alpha_o' \alpha_o)^{-1} \alpha_o' B_{m,2}\right), \tag{3.11}$$

where

$$B_{m,1} \equiv N\left(0, \Omega_u \otimes \Sigma_{z_1 z_1}^{-1}\right) \text{ and } B_{m,2} \equiv \int d B_u B_{w_2}' \left(\int B_{w_2} B_{w_2}'\right)^{-1}.$$

From (3.11) and the continuous mapping theorem (CMT),

$$Q\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} \beta_o' B_{m,1} & \beta_o' \alpha_o (\alpha_o' \alpha_o)^{-1} \alpha_o' B_{m,2} \\ \alpha_{o,\perp}' B_{m,1} & 0 \end{pmatrix}. \tag{3.12}$$

Similarly, from Lemma A.2.(a) in Appendix and CMT

$$Q\left(\widehat{\Pi}_{1st} - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} \beta_o' B_{m,1} & \beta_o' B_{m,2} \\ \alpha_{o,\perp}' B_{m,1} & \alpha_{o,\perp}' B_{m,2} \end{pmatrix}. \tag{3.13}$$

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 (3.11) 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. Johansen, 1995; Phillips, 1998; Anderson, 2002), the LS shrinkage estimator suffers from second order bias in the limit distribution (3.11), 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

$$\widehat{\Pi}_{g,n} = \underset{\Pi \in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \| \Delta Y_t - \Pi Y_{t-1} \|^2_{\widehat{\Omega}^{-1}_{u,n}} + n \sum_{k=1}^{m} \lambda_{r,k,n} ||\Phi_{n,k}(\Pi)||, \tag{3.14}$$

where $\widehat{\Omega}_{u,n}$ is some consistent estimator of Ω_u. GLS methods enable efficient estimation in cointegrating systems with known rank (Phillips, 1991a, 1991b). Here they are used to achieve efficient estimation with unknown rank. In fact, the asymptotic distribution of $\widehat{\Pi}_{g,n}$ is the same as that of the oracle RRR estimator.

COROLLARY 3.6 (Oracle Properties). *Suppose Assumptions 3.1 and 3.2 hold. If* $\widehat{\Omega}_{u,n} \to_p \Omega_u$ *and the tuning parameter satisfies* $n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$ *and* $\lambda_{r,k,n} \to_p \infty$ *for* $k \in \mathcal{S}^c_\phi$, *then*

$$\Pr\left(r(\widehat{\Pi}_{g,n}) = r_o\right) \to 1 \text{ as } n \to \infty \tag{3.15}$$

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

$$\begin{aligned}&\left(\widehat{\Pi}_{g,n} - \Pi_o\right) Q^{-1} D_n^{-1} \\ &\quad \to_d \left(B_{m,1}\, \alpha_o (\beta_o' \alpha_o)^{-1} \int dB_{u\cdot w_2} B'_{w_2} \left(\int B_{w_2} B'_{w_2}\right)^{-1} \right),\end{aligned} \tag{3.16}$$

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

From (3.16), we can invoke the CMT to obtain

$$Q\left(\widehat{\Pi}_{g,n} - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \begin{pmatrix} \beta_o' B_{m,1} & \int dB_{u\cdot w_2} B'_{w_2} \left(\int B_{w_2} B'_{w_2}\right)^{-1} \\ \alpha'_{o,\perp} B_{m,1} & 0 \end{pmatrix}, \tag{3.17}$$

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

Remark 3.7. In the triangular representation of a cointegration system studied in Phillips (1991a), 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

$$\Pi_o = \begin{pmatrix} -I_{r_o} & O_o \\ 0 & \mathbf{0}_{m-r_o} \end{pmatrix}, \; Q = \begin{pmatrix} -I_{r_o} & O_o \\ 0 & I_{m-r_o} \end{pmatrix} \text{ and } Q^{-1} = \begin{pmatrix} -I_{r_o} & O_o \\ 0 & I_{m-r_o} \end{pmatrix}.$$

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 3.6, we deduce that

$$\sqrt{n}\left(\widehat{A}_{g,n} - I_{r_o}\right) \to_d N(0, \Omega_{u_1} \otimes \Sigma^{-1}_{z_1 z_1}) \tag{3.18}$$

and

$$n\widehat{A}_{g,n}\left(\widehat{O}_{g,n} - O_o\right) \to_d \int dB_{u_{1\cdot 2}} B'_{u_2} \left(\int B_{u_2} B'_{u_2}\right)^{-1}, \tag{3.19}$$

where B_{u_1} and B_{u_2} denotes the first r_o and last $m - r_o$ vectors of B_u, and $B_{u_{1\cdot 2}} = B_{u_1} - \Omega_{u,12}\Omega^{-1}_{u,22}B_{u_2}$. Under (3.18), (3.19), and CMT, we deduce that

$$n\left(\widehat{O}_{g,n} - O_o\right) \to_d \int dB_{u_{1\cdot 2}} B'_{u_2} \left(\int B_{u_2} B'_{u_2}\right)^{-1}. \tag{3.20}$$

Evidently from (3.20) the GLS estimator $\widehat{O}_{g,n}$ of the cointegration matrix O_o is asymptotically equivalent to the maximum likelihood estimator studied in Phillips (1991a) and has the usual mixed normal limit distribution, facilitating inference.

4. EXTENSION I: 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 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

$$u_t = D(L)\varepsilon_t = \sum_{j=0}^{\infty} D_j \varepsilon_{t-j}, \text{ with } \sum_{j=0}^{\infty} j^{\frac{1}{2}} ||D_j|| < \infty, \tag{4.1}$$

where ε_t is iid $(0, \Sigma_{\varepsilon\varepsilon})$ with $\Sigma_{\varepsilon\varepsilon}$ positive definite and finite fourth moments.

The i.i.d. assumption on ε_t can be relaxed to allow for martingale difference innovations and to allow for some mild heterogeneity in the innovations without disturbing the limit theory in a material way (see Phillips and Solo, 1992).

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 (4.1), 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.

As expected, under general weak dependence assumptions on u_t, the simple reduced rank regression models (2.1) and (3.1) 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 (A.62), 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 (see, Lemma A.4 in the appendix),

$$\widehat{\Pi}_{1st} \to_p \Pi_1 \equiv Q^{-1} H_o Q + \Pi_o, \tag{4.2}$$

where $H_o = Q\left[\Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}, 0_{m\times(m-r_o)}\right]$. Note that

$$\Pi_1 = Q^{-1}H_oQ + \Pi_o = \left[\alpha_o + \Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}\right]\beta_o' = \widetilde{\alpha}_o\beta_o', \tag{4.3}$$

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 the rank of Π_1 by r_1. Then, by virtue of the expression $\Pi_1 = \widetilde{\alpha}_o\beta_o'$, we have $r_1 \le r_o$ 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 $m \times r_1$ matrices 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 Q_1 denote the ordered[7] left eigenvector matrix of Π_1 and define $\Phi_{1,k}(\Pi) = Q_1(k)\Pi$, where $Q_1(k)$ denotes the k-th row of Q_1. It is clear that the index set $\widetilde{\mathcal{S}}_\phi \equiv \{k : \Phi_{1,k}(\Pi_1) \neq 0\} = \{1, ...r_1\}$ is a subset of $\mathcal{S}_\phi = \{k : \Phi_k(\Pi_o) \neq 0\} = \{1, ...r_o\}$. We next derive the "consistency" of $\widehat{\Pi}_n$.

COROLLARY 4.1. *Let* $\widetilde{\delta}_{r,n} = \max_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}$, *then under Assumptions RR, LP and* $\widetilde{\delta}_{r,n} = o_p(1)$, *the LS shrinkage estimator* $\widehat{\Pi}_n$ *is consistent, i.e.* $\widehat{\Pi}_n \to_p \Pi_1$.

Corollary 4.1 implies that the shrinkage estimator $\widehat{\Pi}_n$ has the same probability limit as that of the OLS estimator $\widehat{\Pi}_{1st}$. As the pseudo limit Π_1 may have more zero eigenvalues, compared with Theorem 3.1, Corollary 4.1 imposes weaker condition on the tuning parameters $\{\lambda_{r,k,n}\}_{k=1}^m$. The next corollary provides the convergence rate of the LS shrinkage estimate to the pseudo true parameter matrix Π_1.

COROLLARY 4.2. *Under Assumptions RR, LP, and* $\widetilde{\delta}_{r,n} = o_p(1)$, *the LS shrinkage 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* $0 < r_o \le m$, *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 respectively. Under Corollary 4.1 and Lemma A.4.(a),

$$Q_{\widetilde{\alpha},n}\widehat{\Pi}_n = Q_{\widetilde{\alpha},n}\widehat{\Pi}_{1st} + o_p(1) = \Lambda_{\widetilde{\alpha},n}Q_{\widetilde{\alpha},n} + o_p(1), \tag{4.4}$$

where $\Lambda_{\widetilde{\alpha},n}$ is a diagonal matrix with the ordered first (largest) r_1 eigenvalues of $\widehat{\Pi}_{1st}$. (4.4) and Lemma A.4.(b) implies that the first r_1 rows of $Q_n\widehat{\Pi}_n$ are estimated as nonzero w.p.a.1. On the other hand, by Corollary 4.1 and Lemma A.4.(a),

$$Q_{\widetilde{\alpha}_\perp,n}\widehat{\Pi}_n = Q_{\widetilde{\alpha}_\perp,n}\widehat{\Pi}_{1st} + o_p(1) = \Lambda_{\widetilde{\alpha}_\perp,n}Q_{\widetilde{\alpha}_\perp,n} + o_p(1), \tag{4.5}$$

where $\Lambda_{\widetilde{\alpha}_\perp,n}$ is a diagonal matrix with the ordered last (smallest) $m-r_1$ eigenvalues of $\widehat{\Pi}_{1st}$. Under Lemmas A.4.(b) and (c), we know that $Q_{\widetilde{\alpha}_\perp,n}\widehat{\Pi}_n$ converges to zero in probability, while its first $r_o - r_1$ rows and the last $m-r_o$ rows have the convergence rates $n^{\frac{1}{2}}$ and n respectively. 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 4.3 (Super Efficiency). *Under Assumptions LP and RR, if* $\lambda_{r,k,n} \to_p \infty$ *for* $k \in \mathcal{S}_\phi^c$ *and* $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = O_p(1)$*, then we have*

$$\Pr\left(Q_n(k)\widehat{\Pi}_n = 0\right) \to 1 \text{ as } n \to \infty, \tag{4.6}$$

for any $k \in \mathcal{S}_\phi^c$.

Corollary 4.3 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 nonzeros w.p.a.1. From Lemma A.4, 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 4.4. *Under Assumptions LP and RR, if* $n^{\frac{1}{2}}\lambda_{r,k,n} = o_p(1)$ *for* $k \in \{r_1+1,\ldots,r_o\}$ *and* $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = O_p(1)$*, then we have*

$$\Pr\left(Q_n(k)\widehat{\Pi}_n \neq 0\right) \to 1 \text{ as } n \to \infty, \tag{4.7}$$

for any $k \in \{r_1+1,\ldots,r_o\}$.

In the proof of Corollary 4.4, we show that $n^{\frac{1}{2}}Q_n(k)\widehat{\Pi}_n$ converges in distribution to some nondegenerated continuous random vectors, which is a stronger result than (4.7). Corollary 4.2 and Corollary 4.4 implies that $\widehat{\Pi}_n$ has at least $m-r_o$ eigenvalues not estimated as zeros w.p.a.1. Hence Corollary 4.2, Corollary 4.3, and Corollary 4.4 give us the following result immediately.

THEOREM 4.5. *Suppose that Assumptions LP and RR are satisfied. If* $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = O_p(1)$, $n^{\frac{1}{2}}\lambda_{r,k,n} = o_p(1)$ *for* $k \in \{r_1+1, \ldots, r_o\}$ *and* $\lambda_{r,k',n} \to_p \infty$ *for* $k' \in \mathcal{S}_\phi^c$, *then we have*

$$\Pr\left(r(\widehat{\Pi}_n) = r_o\right) \to 1 \text{ as } n \to \infty, \tag{4.8}$$

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

Compared with Theorem 3.3, Theorem 4.5 imposes similar conditions on the tuning parameters $\{\lambda_{r,k,n}\}_{k=1}^m$. It is clear that when the pseudo limit Π_1 preserves the rank of Π_o, i.e. $r_o = r_1$, we do not need to show Corollary 4.4 because Theorem 4.5 follows by Corollary 4.2 and Corollary 4.3. In that case, Theorem 4.5 imposes the same conditions on the tuning parameters, i.e. $n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = O_p(1)$ and $\lambda_{r,k,n} \to_p \infty$ for $k \in \mathcal{S}_\phi^c$, where $\widetilde{\delta}_{r,n} = \delta_{r,n}$. On the other hand, when $r_1 < r_o$, the conditions in Theorem 4.5 is stronger, because it requires $n^{\frac{1}{2}}\lambda_{r,k,n} = o_p(1)$ for $k \in \{r_1+1, \ldots, r_o\}$. In Section 6, we construct empirically available tuning parameters which are shown to satisfy the conditions of Theorem 4.5 without knowing whether $r_1 = r_o$ or $r_1 < r_o$.

Theorem 4.5 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 4.5 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 4.4, 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.

5. EXTENSION II: ESTIMATION WITH EXPLICIT TRANSIENT DYNAMICS

This section considers estimation of the general model

$$\Delta Y_t = \Pi_o Y_{t-1} + \sum_{j=1}^{p} B_{o,j}\Delta Y_{t-j} + u_t \tag{5.1}$$

with simultaneous cointegrating rank selection and lag order selection. Recall that the unknown parameters (Π_o, B_o) are estimated by penalized LS estimation

$$(\widehat{\Pi}_n, \widehat{B}_n) = \underset{\Pi, B_1, \ldots, B_p \in R^{m\times m}}{\arg\min} \left\{ \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum\nolimits_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2 + n\sum_{j=1}^{p} \lambda_{b,j,n} \left\| B_j \right\| + n\sum_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\| \right\}. \tag{5.2}$$

For consistent lag order selection the model should be consistently estimable and it is assumed that the given p in (5.1) is such that the error term u_t satisfies Assumption 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 3.2 to accommodate the general structure in (5.1).

Assumption 5.1 (GRR)**.**

(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

$$\alpha'_{o,\perp}\left(I_m - \sum_{j=1}^{p} B_{o,j}\right)\beta_{o,\perp} \tag{5.3}$$

is nonsingular.

Under Assumption 5.1, the time series Y_t has the following partial sum representation,

$$Y_t = C_B \sum_{s=1}^{t} u_s + \Xi(L)u_t + C_B Y_0, \tag{5.4}$$

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 (5.4), we deduce that $\beta'_o Y_t = \beta'_o \Xi(L)u_t$ and ΔY_{t-j} $(j = 0,\dots,p)$ 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 $\Delta X_{t-1} = \left[\Delta Y'_{t-1},\dots,\Delta Y'_{t-p}\right]'$ and then the model in (5.1) can be written as

$$\Delta Y_t = \left[\Pi_o\ B_o\right]\begin{bmatrix} Y_{t-1} \\ \Delta X_{t-1} \end{bmatrix} + u_t. \tag{5.5}$$

Let

$$Z_{t-1} = Q_B \begin{bmatrix} Y_{t-1} \\ \Delta X_{t-1} \end{bmatrix} = \begin{bmatrix} Z_{3,t-1} \\ Z_{2,t-1} \end{bmatrix}, \tag{5.6}$$

where $Z'_{3,t-1} = \left[Y'_{t-1}\beta_o \ \Delta X'_{t-1} \right]$ is a stationary process and $Z_{2,t-1} = \alpha'_{o,\perp} Y_{t-1}$ comprises the $I(1)$ components. 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 (5.2).

LEMMA 5.1. *Suppose that Assumptions WN and GRR are satisfied. If* $\delta_{r,n} = o_p(1)$ *and* $\delta_{b,n} = o_p(1)$ *where* $\delta_{b,n} \equiv \max_{j \in \mathcal{S}_B} \lambda_{b,j,n}$, *then the LS shrinkage estimator* $(\widehat{\Pi}_n, \widehat{B}_n)$ *satisfies*

$$\left[(\widehat{\Pi}_n, \widehat{B}_n) - (\Pi_o, B_o) \right] Q_B^{-1} D_{n,B}^{-1} = O_p\left(1 + n^{\frac{1}{2}} \delta_{r,n} + n^{\frac{1}{2}} \delta_{b,n} \right), \tag{5.7}$$

where $D_{n,B} = diag\left(n^{-\frac{1}{2}} I_{r_o + mp}, n^{-1} I_{m - r_o} \right)$.

Lemma 5.1 implies that the LS shrinkage estimators $(\widehat{\Pi}_n, \widehat{B}_n)$ have the same convergence rates as the OLS estimators $(\widehat{\Pi}_{1st}, \widehat{B}_{1st})$ (see, Lemma A.6.a). 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 $Q_n \widehat{\Pi}_n$ 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 5.1. *Suppose that Assumptions WN and GRR are satisfied. If the tuning parameters satisfy* $n^{\frac{1}{2}}(\delta_{r,n} + \delta_{b,n}) = O_p(1)$, $\lambda_{r,k,n} \to_p \infty$ *and* $n^{\frac{1}{2}} \lambda_{b,j,n} \to_p \infty$ *for* $k \in \mathcal{S}_\phi^c$ *and* $j \in \mathcal{S}_B^c$, *then we have*

$$\Pr\left(Q_{\alpha,n} \widehat{\Pi}_n = 0 \right) \to 1 \text{ as } n \to \infty; \tag{5.8}$$

and for all $j \in \mathcal{S}_B^c$

$$\Pr\left(\widehat{B}_{n,j} = \mathbf{0}_{m \times m} \right) \to 1 \text{ as } n \to \infty. \tag{5.9}$$

Theorem 5.1 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 5.1 and Theorem 5.1 imply consistent cointegration rank selection and consistent lag order selection.

We next derive the asymptotic distribution of $\widehat{\Theta}_{\mathcal{S}} = \left(\widehat{\Pi}_n, \widehat{B}_{\mathcal{S}_B} \right)$, where $\widehat{B}_{\mathcal{S}_B}$ denotes the LS shrinkage estimator of the nonzero matrices in B_o. Let $I_{\mathcal{S}_B} = diag(I_{1,m}, \ldots, I_{d_{\mathcal{S}_B},m})$ where the $I_{j,m}$ ($j = 1, \ldots, d_{\mathcal{S}_B}$) are $m \times m$ identity matrices and $d_{\mathcal{S}_B}$ is the dimensionality of the index set $\mathcal{S}_B$. Define

$$Q_{\mathcal{S}} \equiv \begin{pmatrix} \beta'_o & 0 \\ 0 & I_{\mathcal{S}_B} \\ \alpha'_{o,\perp} & 0 \end{pmatrix} \text{ and } D_{n,\mathcal{S}} \equiv diag(n^{-\frac{1}{2}} I_{r_o}, n^{-\frac{1}{2}} I_{\mathcal{S}_B}, n^{-1} I_{m - r_o}),$$

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 (5.1), then the true model can be written as

$$\Delta Y_t = \Pi_o Y_{t-1} + B_{o,\mathcal{S}_B} \Delta X_{\mathcal{S},t-1} + u_t = \Theta_{o,\mathcal{S}} Q_{\mathcal{S}}^{-1} Z_{\mathcal{S},t-1} + u_t, \tag{5.10}$$

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 $Z'_{3\mathcal{S},t-1} = \left[Y'_{t-1}\beta_o \; \Delta X'_{\mathcal{S},t-1} \right]$ and $Z_{2,t-1} = \alpha'_{o,\perp} Y_{t-1}$. From Lemma A.5, 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 5.2. *Under conditions of Theorem 5.1, if* $n^{\frac{1}{2}}(\delta_{r,n} + \delta_{b,n}) = o_p(1)$, *then*

$$\left(\widehat{\Theta}_{\mathcal{S}} - \Theta_{o,\mathcal{S}}\right) Q_{\mathcal{S}}^{-1} D_{n,\mathcal{S}}^{-1} \to_d \left(B_{m,\mathcal{S}} \, \alpha_o (\alpha'_o \alpha_o)^{-1} \alpha'_o B_{m,2} \right), \tag{5.11}$$

where

$$B_{m,\mathcal{S}} \equiv N\left(0, \Omega_u \otimes \Sigma^{-1}_{z_{3\mathcal{S}} z_{3\mathcal{S}}}\right) \text{ and } B_{m,2} \equiv \int dB_u B'_{w_2} \left(\int B_{w_2} B'_{w_2} \right)^{-1}.$$

Theorem 5.2 extends the result of Theorem 3.5 to the general VECM with lagged differences. From Theorem 5.2, 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 asymptotic bias, evident in the component $B_{m,2}$ of the limit (5.11). As in the simpler model this asymptotic bias is eliminated by GLS estimation. Accordingly we define the GLS shrinkage estimator of the general model as

$$\begin{aligned} (\widehat{\Pi}_{g,n}, \widehat{B}_{g,n}) = \underset{\Pi, B_1, \ldots, B_p \in R^{m \times m}}{\arg\min} \Bigg\{ & \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum\nolimits_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2_{\widehat{\Omega}^{-1}_{u,n}} \\ & + n \sum_{j=1}^{p} \lambda_{b,j,n} \left\| B_j \right\| + n \sum_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\| \Bigg\}. \end{aligned} \tag{5.12}$$

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 5.3 (Oracle Properties of GLS). *Suppose the conditions of Theorem 5.2 are satisfied. If* $\widehat{\Omega}_{u,n} \to_p \Omega_u$, *then*

$$\Pr\left(r(\widehat{\Pi}_{g,n}) = r_o\right) \to 1 \text{ and } \Pr\left(\widehat{B}_{g,j,n} = 0\right) \to 1 \tag{5.13}$$

for $j \in \mathcal{S}_B^c$ *as* $n \to \infty$; *moreover,* $\widehat{\Theta}_{\mathcal{S}}$ *has the following limit distribution*

$$\begin{aligned}&\left(\widehat{\Theta}_{\mathcal{S}} - \Theta_{o,\mathcal{S}}\right) Q_{\mathcal{S}}^{-1} D_{n,\mathcal{S}}^{-1} \\ &\quad \to_d \left(B_{m,\mathcal{S}}\, \alpha_o \left(\beta_o' \alpha_o\right)^{-1} \int dB_{u\cdot w_2} B_{w_2}' \left(\int B_{w_2} B_{w_2}'\right)^{-1}\right)\end{aligned} \tag{5.14}$$

where $B_{u\cdot w_2}$ *is defined in Theorem 3.6.*

Corollary 5.3 is proved using the same arguments as Corollary 3.6 and Theorem 5.2. Its proof is omitted. The asymptotic distributions of the penalized LS/GLS estimates can be used to conduct inference on Π_o and B_o. However, use of these asymptotic distributions implies that the true cointegrating rank and lag order are selected with probability one. In consequence, these distributions may provide poor approximations to the finite sample distributions of the penalized LS/GLS estimates when model selection errors occur in finite samples, leading to potential size distortions in inference based on (5.11) or (5.14). The development of robust approaches to confidence interval construction therefore seems an important task for future research.

Remark 5.4. Although the grouped Lasso penalty function $P(B) = \|B\|$ is used in LS shrinkage estimation (5.2) and GLS shrinkage estimation (5.12), we remark that a full Lasso penalty function can also be used and the resulting GLS shrinkage estimate enjoys the same properties stated in Corollary 5.3. The GLS shrinkage estimation using the (full) Lasso penalty takes the following form

$$\begin{aligned}(\widehat{\Pi}_{g,n}, \widehat{B}_{g,n}) = \underset{\Pi, B_1, \ldots, B_p \in R^{m\times m}}{\arg\min} \Bigg\{ &\sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum\nolimits_{j=1}^{p} B_j \Delta Y_{t-j} \right\|_{\widehat{\Omega}_{u,n}^{-1}}^2 \\ &+ n \sum_{j=1}^{p} \sum_{l=1}^{m} \sum_{s=1}^{m} \lambda_{b,j,l,s,n} |B_{j,ls}| + n \sum_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\| \Bigg\},\end{aligned} \tag{5.15}$$

where $B_{j,ls}$ denotes the (l,s)-th element of B_j. The advantage of the grouped Lasso penalty $P(B)$ is that it shrinks elements in B to zero groupwisely, which makes it a natural choice for the lag order selection (as well as lag elimination) in VECMs. The Lasso penalty is more flexible and when used in shrinkage estimation, it can do more than select the zero matrices. It can also select the nonzero elements in the nonzero matrices $B_{o,j}$ $(j \in \mathcal{S}_B)$ w.p.a.1.

Remark 5.5. The flexibility of the Lasso penalty enables GLS shrinkage estimation to achieve more goals in one-step, in addition to 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

$$\begin{bmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{bmatrix} = \begin{bmatrix} \Pi_o^{11} & \Pi_o^{12} \\ \Pi_o^{21} & \Pi_o^{22} \end{bmatrix} \begin{bmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{bmatrix} + \sum_{j=1}^{p} \begin{bmatrix} B_{o,j}^{11} & B_{o,j}^{12} \\ B_{o,j}^{21} & B_{o,j}^{22} \end{bmatrix} \begin{bmatrix} \Delta Y_{1,t-j} \\ \Delta Y_{2,t-j} \end{bmatrix} + u_t, \quad \textbf{(5.16)}$$

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) Lasso penalty of Π^{12} in (5.16) such that the causality test is automatically executed in GLS shrinkage estimation.

Remark 5.6. In this paper, we only consider the Lasso penalty function in the LS or GLS shrinkage estimation. The main advantage of the Lasso penalty is that it is a convex function, which combines the convexity of the LS or GLS criterion, making the computation of the shrinkage estimate faster and more accurate. It is clear that as long as the tuning parameter satisfies certain rate requirements, our main results continue to hold if other penalty functions (e.g., the bridge penalty) are used in the LS or GLS shrinkage estimation.

6. ADAPTIVE SELECTION OF THE TUNING PARAMETERS

This section develops a data-driven procedure of selecting the tuning parameters $\{\lambda_{r,k,n}\}_{k=1}^{m}$ and $\{\lambda_{b,j,n}\}_{j=1}^{p}$. As presented in previous sections, the conditions ensuring oracle properties in GLS shrinkage estimation require that the tuning parameters of the estimates of zero and nonzero components have different asymptotic behavior. For example, in Theorem 3.3, we need $\lambda_{r,k,n} = O_p(n^{-\frac{1}{2}})$ for any $k \in \mathcal{S}_\phi$ and $\lambda_{r,k,n} \to_p \infty$ for $k \in \mathcal{S}_\phi^c$, which implies that some sort of known adaptive penalty should appear in $\lambda_{r,k,n}$. One popular choice of such a penalty is the adaptive Lasso penalty (c.f., Zou, 2006), which in our model can be defined as

$$\lambda_{r,k,n} = \frac{\lambda_{r,k,n}^*}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} \text{ and } \lambda_{b,j,n} = \frac{m^\omega \lambda_{b,j,n}^*}{||\widehat{B}_{1st,j}||^\omega}, \quad \textbf{(6.1)}$$

where $\lambda_{r,k,n}^*$ and $\lambda_{b,j,n}^*$ are nonincreasing positive sequences and ω is some positive finite constant.

The adaptive penalty in $\lambda_{r,k,n}$ is $||\phi_k(\widehat{\Pi}_{1st})||^{-\omega}$ $(k = 1, \ldots, m)$, because for any $k \in \mathcal{S}_\phi^c$, there is $||\phi_k(\widehat{\Pi}_{1st})||^{-\omega} \to_p \infty$ and for any $k \in \mathcal{S}_\phi$, there is $||\phi_k(\widehat{\Pi}_{1st})||^{-\omega} \to_p ||\phi_k(\Pi_o)||^{-\omega} = O(1)$ under Assumption WN[8]. Similarly, the adaptive penalty in $\lambda_{b,j,n}$ is $m^\omega ||\widehat{B}_{1st,j}||^{-\omega}$, where the extra term m^ω is used to adjust the effect of dimensionality of B_j on the adaptive penalty. Such adjustment

does not effect the asymptotic properties of the LS/GLS shrinkage estimation, but it is used to improve their finite sample performances. To see the effect of the dimensionality on the adaptive penalty, we write

$$||\widehat{B}_{1st,j}||^{\omega} = \left[\sum_{l=1}^{m}\sum_{h=1}^{m}\left|\widehat{B}_{1st,j,lh}\right|^2\right]^{\frac{\omega}{2}}.$$

Although each individual $|\widehat{B}_{1st,j,lh}|^2$ may be close to zero, $||\widehat{B}_{1st,j}||^2$ could be large in magnitude in finite samples because it is the sum of m^2 such terms (i.e. $|\widehat{B}_{1st,j,lh}|^2$). As a result, the adaptive penalty $||\widehat{B}_{1st,j}||^{-\omega}$ without any adjustment tends to be smaller than the value it should be. One straightforward adjustment for the dimensionality effect is to use the average, instead of the sum, of the square terms $|\widehat{B}_{1st,j,lh}|^2$, i.e.

$$\left[m^{-2}\sum_{l=1}^{m}\sum_{h=1}^{m}|\widehat{B}_{1st,j,lh}|^2\right]^{\frac{\omega}{2}} = m^{-\omega}||\widehat{B}_{1st,j}||^{\omega}$$

in the adaptive penalty. Under some general rate conditions on $\lambda^*_{r,k,n}$ and $\lambda^*_{b,j,n}$, the following lemma shows that the tuning parameters specified in (6.1) satisfy the conditions in our theorems of super efficiency and oracle properties.

LEMMA 6.1. *(i) If* $n^{\frac{1}{2}}\lambda^*_{r,k,n} = o(1)$ *and* $n^{\omega}\lambda^*_{r,k,n} \to \infty$*, then under Assumptions WN and RR we have*

$n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$ *and* $\lambda_{r,k,n} \to_p \infty$

for any $k \in \mathcal{S}^c_{\phi}$*; (ii) if* $n^{\frac{1+\omega}{2}}\lambda^*_{r,k,n} = o(1)$ *and* $n^{\omega}\lambda^*_{r,k,n} \to \infty$*, then under Assumptions LP and RR*

$n^{\frac{1}{2}}\widetilde{\delta}_{r,n} = o_p(1)$, $n^{\frac{1}{2}}\lambda_{r,k,n} = o_p(1)$ *and* $\lambda_{r,k',n} \to_p \infty$

for any $k \in \{r_1+1,\ldots,r_o\}$ *and* $k' \in \mathcal{S}^c_{\phi}$*; (iii) if* $n^{\frac{1}{2}}\lambda^*_{r,k,n} = o(1)$ *and* $n^{\omega}\lambda^*_{r,k,n} \to \infty$ *for any* $k = 1,\ldots,m$, *and* $n^{\frac{1}{2}}\lambda^*_{b,j,n} = o(1)$ *and* $n^{\frac{1+\omega}{2}}\lambda^*_{b,j,n} \to \infty$ *for any* $j = 1,\ldots,p$*, then under Assumptions WN and GRR*

$n^{\frac{1}{2}}(\delta_{r,n}+\delta_{b,n}) = o_p(1)$, $\lambda_{r,k,n} \to_p \infty$ *and* $\lambda_{b,j,n} \to_p \infty$

for any $k \in \mathcal{S}^c_{\phi}$ *and* $j \in \mathcal{S}^c_B$.

It is notable that, when u_t is iid, $\lambda^*_{r,k,n}$ is required to converge to zero with the rate faster than $n^{-\frac{1}{2}}$, while when u_t is weakly dependent, $\lambda^*_{r,k,n}$ has to converge to zero with the rate faster than $n^{-\frac{1+\omega}{2}}$. The convergence rate of $\lambda^*_{r,k,n}$ in Lemma 6.1.(ii) is faster to ensure that the pseudo $r_o - r_1$ zero eigenvalues in Π_1

are estimated as nonzeros w.p.a.1. When $r_1 = r_o$, Π_1 contains no pseudo zero eigenvalues and it has the true rank r_o. It is clear that in this case, we only need $n^{\frac{1}{2}}\lambda^*_{r,k,n} = o(1)$ and $n^{\omega}\lambda^*_{r,k,n} \to \infty$ to show that the tuning parameters in (6.1) satisfy $n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$ and $\lambda_{r,k',n} \to_p \infty$ for any $k' \in \mathcal{S}^c_{\phi}$.

From Lemma 6.1, we see that the conditions imposed on $\{\lambda^*_{r,k,n}\}_{k=1}^m$ and $\{\lambda^*_{b,j,n}\}_{j=1}^p$ to ensure oracle properties in GLS shrinkage estimation only restrict the rates at which the sequences $\lambda^*_{r,k,n}$ and $\lambda^*_{b,j,n}$ go 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 these sequences 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 $\lambda^*_{r,k,n} = 0$ and $\lambda^*_{b,j,n} = 0$, LS shrinkage estimation 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, these sequences should converge to zero as slow as possible so that in finite samples zero components in the model are estimated as zeros with higher probability. In the opposite extremity $\lambda^*_{r,k,n} = \infty$ and $\lambda^*_{b,j,n} = \infty$, 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 sequences in $\{\lambda^*_{r,k,n}\}_{k=1}^m$ and $\{\lambda^*_{b,j,n}\}_{j=1}^p$.

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

$$\left\| \frac{1}{n}\sum_{t=1}^{n} Q_n(k)\widehat{\Omega}^{-1}_{u,n}\left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p}\widehat{B}_{n,j}\Delta Y_{t-j}\right)Y'_{t-1}\right\| < \frac{\lambda^*_{r,k,n}}{2||\phi_k\left(\widehat{\Pi}_{1st}\right)||^{\omega}}. \tag{6.2}$$

Let $T \equiv Q\Pi_o$ and $T(k)$ be the k-th row of the matrix $Q\Pi_o$. If a nonzero $T(k)$ $(k \le r_o)$ 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 (6.2) is satisfied. As $n\phi_k(\widehat{\Pi}_{1st}) = O_p(1)$, we can rewrite the inequality in (6.2) as

$$\left\| \frac{1}{n}\sum_{t=1}^{n} Q_n(k)\widehat{\Omega}^{-1}_{u,n}\left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p}\widehat{B}_{n,j}\Delta Y_{t-j}\right)Y'_{t-1}\right\| < \frac{n^{\omega}\lambda^*_{r,k,n}}{2||n\phi_k\left(\widehat{\Pi}_{1st}\right)||^{\omega}}. \tag{6.3}$$

The sample average in the left side of this inequality is asymptotically a vector of linear combinations of nondegenerate 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 $\lambda^*_{r,k,n} = c_{r,k}n^{-\frac{\omega}{2}}$ (here $c_{r,k}$ is some positive constant whose selection is discussed later) 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

$$\left\| n^{-\frac{1}{2}} \sum_{t=1}^{n} \widehat{\Omega}_{u,n}^{-1} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j} \right) \Delta Y'_{t-j} \right\| < \frac{n^{\frac{1}{2}} \lambda^*_{b,j,n}}{2||\widehat{B}_{1st,j}||^{\omega}}. \tag{6.4}$$

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 (6.4) is satisfied. As $\widehat{B}_{1st,j} = O_p(n^{-\frac{1}{2}})$ the inequality in (6.4) can be written as

$$\left\| n^{-\frac{1}{2}} \sum_{t=1}^{n} \widehat{\Omega}_{u,n}^{-1} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j} \right) \Delta Y'_{t-j} \right\| < \frac{n^{\frac{1+\omega}{2}} \lambda^*_{b,j,n}}{2||n^{\frac{1}{2}} \widehat{B}_{1st,j}||^{\omega}}. \tag{6.5}$$

The sample average on the left side of this inequality is asymptotically a vector of linear combinations of nondegenerated 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 $\lambda^*_{b,j,n} = c_{b,j}n^{-\frac{1}{2}-\frac{\omega}{4}}$ (again $c_{b,j}$ is some positive constant whose selection is discussed later) 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 $\lambda^*_{b,j,n}$. Note that the sample average on the left hand side of (6.3) can be written as

$$F_{\pi,n}(k) \equiv \frac{Q_n(k)\widehat{\Omega}_{u,n}^{-1}}{n} \sum_{t=1}^{n} \left[u_t - \left(\widehat{\Theta}_n - \Theta_o \right) Q_B^{-1} Z_{t-1} \right] Y'_{t-1}.$$

Similarly, the sample average on the left hand side of (6.5) can be written as

$$F_{b,n}(j) \equiv \frac{\widehat{\Omega}_{u,n}^{-1}}{\sqrt{n}} \sum_{t=1}^{n} \left[u_t - \left(\widehat{\Theta}_n - \Theta_o \right) Q_B^{-1} Z_{t-1} \right] \Delta Y_{t-j}'.$$

The next lemma provides the asymptotic distributions of $F_{\pi,n}(k)$ and $F_{b,n}(j)$ for $k = 1, \ldots, m$ and $j = 1, \ldots, p$.[9]

LEMMA 6.2. *Suppose that the conditions of Corollary 5.3 are satisfied, then*

$$F_{\pi,n}(k) = Q_n(k) T_{1,\pi_o} \int dB_u B_u' T_{2,\pi_o} + o_p(1) \tag{6.6}$$

for $k = 1, \ldots, m$, *where*

$$T_{1,\pi_o} = \Omega_u^{-1} - \Omega_u^{-1} \alpha_o (\alpha_o' \Omega_u^{-1} \alpha_o)^{-1} \alpha_o' \Omega_u^{-1} \text{ and } T_{2,\pi_o} = \alpha_{o,\perp} (\beta_{o,\perp}' \alpha_{o,\perp})^{-1} \beta_{o,\perp}';$$

further, for $j = 1, \ldots, p$,

$$F_{b,n}(j) \to_d \Omega_u^{-\frac{1}{2}} B_{m \times m}(1) \Sigma_{\Delta y_j | z_{3\mathcal{S}}}^{\frac{1}{2}}, \tag{6.7}$$

where $B_{m,m} = N(0, I_m \otimes I_m)$,

$$\begin{aligned} \Sigma_{\Delta y_j | z_{3\mathcal{S}}} &= E\left[\left(\Delta Y_{t-j} \middle| Z_{3\mathcal{S}} \right) \left(\Delta Y_{t-j}' \middle| Z_{3\mathcal{S}} \right) \right] \text{ and } \Delta Y_{t-j} \middle| Z_{3\mathcal{S}} \\ &= \Delta Y_{t-j} - \Sigma_{\Delta y_j z_{3\mathcal{S}}} \Sigma_{z_{3\mathcal{S}} z_{3\mathcal{S}}}^{-1} Z_{3\mathcal{S},t-1}. \end{aligned}$$

We propose to select $c_{r,k}$ to normalize the random sum in (6.6), i.e.

$$\widehat{c}_{r,k} = 2 \left\| Q_n(k) \widehat{T}_{1,\pi} \widehat{\Omega}_{u,n}^{1/2} \right\| \times \left\| \widehat{\Omega}_{u,n}^{1/2} \widehat{T}_{2,\pi} \right\|, \tag{6.8}$$

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 initial estimates of the rank r_o and the 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 propose to select c_b to normalize the random sum in (6.6), i.e.

$$\widehat{c}_{b,j} = 2 \left\| \widehat{\Omega}_{u,n}^{-1/2} \right\| \times \left\| \widehat{\Sigma}_{\Delta y_j \Delta y_j}^{\frac{1}{2}} \right\|, \tag{6.9}$$

where $\widehat{\Sigma}_{\Delta y_j \Delta y_j} = \frac{1}{n} \sum_{t=1}^{n} \Delta Y_{t-j} \Delta Y_{t-j}'$. From the expression in (6.7), it seems that the empirical analog of $\Sigma_{\Delta y_j | z_{3\mathcal{S}}}$ is a more propriate term to normalize $F_{b,n}(j)$. However, if ΔY_{t-j} is a redundant lag and the residual of its projection

on $\beta_o' Y_{t-1}$ and nonredundant lagged differences is close to zero, then $\Sigma_{\Delta y_j|z_{3\mathcal{S}}}$ and its estimate will be close to zero. As a result, $\widehat{c}_{b,j}$ tends to be small, which will increase the probability of including ΔY_{t-j} in the selected model with higher probability in finite samples. To avoid such unappealing scenario, we use $\widehat{\Sigma}_{\Delta y_j \Delta y_j}$ instead of the empirical analog of $\Sigma_{\Delta y_j|z_{3\mathcal{S}}}$ in (6.9). It is clear that $\widehat{c}_{b,j}$ can be directly constructed from the preliminary LS estimation.

The choice of ω is a more complicated issue which is not pursued in this paper. For the empirical applications, we propose to choose $\omega = 2$ because such a choice is popular in the Lasso-based variable selection literature, it satisfies all our rate criteria, and simulations show that the choice works remarkably well. Based on all the above results, we propose the following data dependent tuning parameters for LS shrinkage estimation:

$$\lambda_{r,k,n} = \frac{2}{n} \left\| Q_n(k) \widehat{T}_{1,\pi} \widehat{\Omega}_{u,n}^{1/2} \right\| \times \left\| \widehat{\Omega}_{u,n}^{1/2} \widehat{T}_{2,\pi} \right\| \times ||\phi_k(\widehat{\Pi}_{1st})||^{-2} \tag{6.10}$$

and

$$\lambda_{b,j,n} = \frac{2m^2}{n} \left\| \widehat{\Omega}_{u,n}^{-1/2} \right\| \times \left\| \widehat{\Sigma}_{\Delta y_j \Delta y_j}^{\frac{1}{2}} \right\| \times ||\widehat{B}_{1st,j}||^{-2} \tag{6.11}$$

for $k = 1, \ldots, m$ and $j = 1, \ldots, p$. The above discussion is based on the general VECM with *iid* u_t. In the simple error correction model where the cointegration rank selection is the only concern, the adaptive tuning parameters proposed in (6.10) are still valid. The expression in (6.10) will be invalid when u_t is weakly dependent and $r_1 < r_o$. In that case, we propose to replace the leading term $2n^{-1}$ in (6.10) by $2n^{-3/2}$.

7. SIMULATION STUDY

We conducted simulations to assess the finite sample performance of the shrinkage estimates in terms of cointegrating rank selection and efficient estimation. Three models were investigated. In the first model, the simulated data are generated from

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} u_{1,t} \\ u_{2,t} \end{pmatrix}, \tag{7.1}$$

where $u_t \equiv iid\ N(0, \Omega_u)$ 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 is specified as follows

$$\begin{pmatrix} \pi_{11,o} & \pi_{12,o} \\ \pi_{21,o} & \pi_{22,o} \end{pmatrix} = \begin{pmatrix} 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} -1 & -0.5 \\ 1 & 0.5 \end{pmatrix} \text{ and } \begin{pmatrix} -0.5 & 0.1 \\ 0.2 & -0.4 \end{pmatrix} \tag{7.2}$$

to allow for the cointegration rank to be 0, 1 and 2 respectively.

In the second model, the simulated data $\{Y_t\}_{t=1}^n$ are generated from equations (7.1)–(7.2), 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 $\varepsilon_t \equiv iid\ N(0, \Omega_\varepsilon)$ 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

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + B_{1,o} \begin{pmatrix} \Delta Y_{1,t-1} \\ \Delta Y_{2,t-1} \end{pmatrix} + B_{3,o} \begin{pmatrix} \Delta Y_{1,t-3} \\ \Delta Y_{2,t-3} \end{pmatrix} + u_t, \tag{7.3}$$

where u_t is generated under the same condition in (7.1), Π_o is specified similarly in (7.2), $B_{1,o}$ and $B_{3,o}$ are taken to be $diag(0.4, 0.4)$ such that Assumption 5.1 is satisfied. The initial values (Y_t, ε_t) $(t = -3, \ldots, 0)$ are set to be zero. In the above three cases, 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

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + u_t, \tag{7.4}$$

where u_t is $iid(0, \Omega_u)$ with some unknown positive definite matrix Ω_u. The above empirical model is correctly specified under the data generating assumption (7.1), but is misspecified under (7.2). 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 (7.1).

In the third model, we assume that the econometrician specifies the following model

$$\begin{pmatrix} \Delta Y_{1,t} \\ \Delta Y_{2,t} \end{pmatrix} = \Pi_o \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \sum_{j=1}^{3} B_{j,o} \begin{pmatrix} \Delta Y_{1,t-j} \\ \Delta Y_{2,t-j} \end{pmatrix} + u_t, \tag{7.5}$$

where u_t is $iid(0, \Omega_u)$ with some unknown positive definite matrix Ω_u. The above empirical model is over-parameterized according to (7.3). 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_o.

Table 1 presents finite sample probabilities of rank selection under different model specifications. Overall, the GLS 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 close to 1 (around 0.96) and the probabilities of selecting the true ranks $r_o = 1$ and $r_o = 2$ are almost

TABLE 1. Cointegration rank selection with adaptive Lasso penalty

	Model 1					
	$r_o = 0, \lambda_o = (0\ 0)$		$r_o = 1, \lambda_o = (0\ -0.5)$		$r_o = 2, \lambda_o = (-0.6\ -0.5)$	
	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
$\widehat{r}_n = 0$	0.9588	0.9984	0.0000	0.0002	0.0000	0.0000
$\widehat{r}_n = 1$	0.0412	0.0016	0.9954	0.9996	0.0000	0.0000
$\widehat{r}_n = 2$	0.0000	0.0000	0.0046	0.0002	1.0000	1.0000
	Model 2					
	$r_o = 0, \lambda_1 = (0\ 0)$		$r_o = 1, \lambda_1 = (0\ -0.25)$		$r_o = 2, \lambda_1 = (-0.30\ -0.15)$	
	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
$\widehat{r}_n = 0$	0.9882	0.9992	0.0010	0.0000	0.0006	0.0000
$\widehat{r}_n = 1$	0.0118	0.0008	0.9530	0.9962	0.1210	0.0008
$\widehat{r}_n = 2$	0.0010	0.0000	0.0460	0.0038	0.8784	0.9992

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in eqation (6.15). λ_o represents the eigenvalues of the true matrix Π_o, while λ_1 represents the eigenvalues of the pseudo true matrix Π_1.

equal to 1. When the sample size is increased to 400, the probabilities of selecting the true ranks $r_o = 0$ and $r_o = 1$ are almost equal to 1 and the probability of selecting the true rank $r_o = 2$ equals 1. Similar results show up when the data are weakly dependent (model 2). The only difference is that when the pseudo true eigenvalues are close to zero, the probability of falsely selecting these small eigenvalues is increased, as illustrated in the weakly dependent case with $r_o = 2$. However, as the sample size grows, the probability of selecting the true rank moves closer to 1.

Tables 3, 4 and 5 provide finite sample properties of the GLS 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 GLS shrinkage estimate clearly dominates the LS estimate due to the high probability of the shrinkage method selecting the true rank. When the true rank $r_o = 1$, we do not observe an efficiency advantage of the GLS shrinkage estimator over the LS estimate, but the finite sample bias of the shrinkage estimate is remarkably smaller (Table 4). From Corollary 3.6, we see that the GLS shrinkage estimator is free of high order bias, which explains its smaller bias in finite samples. Moreover, Lemma A.2 and Corollary 3.6 indicate that the OLS estimator and the GLS shrinkage estimator (and hence the oracle estimator) have almost the same variance. This explains the phenomenon that the GLS shrinkage estimate does not look more efficient than the OLS estimate. To better compare the OLS estimate, the GLS shrinkage estimate, and the oracle estimate, we transform the three estimates using the matrix Q and its inverse (i.e. the estimate $\widehat{\Pi}$ is transformed to $Q\widehat{\Pi}Q^{-1}$). Note that in this case, $Q\Pi_o Q^{-1} = diag(\text{-}0.5, 0)$.

The finite sample properties of the transformed estimates are presented in the last two panels of Table 4. We see that the elements in the last column of the transformed GLS 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 as the sample size grows, the GLS shrinkage estimator approaches the oracle estimator in terms of overall performance. When the true rank $r_o = 2$, the LS estimator is better than the shrinkage estimator as the latter suffers from shrinkage bias in finite samples. If shrinkage bias is a concern, one can run a reduced rank regression based on the rank selected by the GLS shrinkage estimation to get the so called post-Lasso estimator (c.f. Belloni and Chernozhukov, 2013). The post-Lasso estimator also enjoys oracle properties and it is free of shrinkage bias in finite samples.

Table 2 shows finite sample probabilities of the new shrinkage method in joint rank and lag order selection for the third model. Evidently, the method performs very well in selecting the true rank and true lagged differences (and thus the true

TABLE 2. Rank selection and lagged order selection with adaptive Lasso penalty

	Cointegration rank selection					
	$r_o = 0, \lambda_o = (0\ 0)$		$r_o = 1, \lambda_o = (0\ -0.5)$		$r_o = 2, \lambda_o = (-0.6\ -0.5)$	
	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
$\widehat{r}_n = 0$	0.9818	1.0000	0.0000	0.0000	0.0000	0.0000
$\widehat{r}_n = 1$	0.0182	0.0000	0.9980	1.0000	0.0000	0.0008
$\widehat{r}_n = 2$	0.0000	0.0000	0.0020	0.0000	1.0000	0.9992
	Lagged difference selection					
	$r_o = 0, \lambda_o = (0\ 0)$		$r_o = 1, \lambda_o = (0\ -0.5)$		$r_o = 2, \lambda_o = (-0.6\ -0.5)$	
	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
$\widehat{p}_n \in T$	0.9856	0.9976	0.9960	0.9998	0.9634	1.0000
$\widehat{p}_n \in C$	0.0058	0.0004	0.0040	0.0002	0.0042	0.0000
$\widehat{p}_n \in I$	0.0086	0.0020	0.0000	0.0000	0.0324	0.0000
	Model Selection					
	$r_o = 0, \lambda_o = (0\ 0)$		$r_o = 1, \lambda_o = (0\ -0.5)$		$r_o = 2, \lambda_o = (-0.6\ -0.5)$	
	$n = 100$	$n = 400$	$n = 100$	$n = 400$	$n = 100$	$n = 400$
$\widehat{m}_n \in T$	0.9692	0.9976	0.9942	0.9998	0.9634	0.9992
$\widehat{m}_n \in C$	0.0222	0.0004	0.0058	0.0002	0.0042	0.0000
$\widehat{m}_n \in I$	0.0086	0.0020	0.0000	0.0000	0.0324	0.0008

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in (6.15) and (6.16). λ_o in each column represents the eigenvalues of Π_o. "T" denotes selection of the true lags model, "C" denotes the selection of a consistent lags model (i.e., a model with no incorrect shrinkage), and "I" denotes the selection of an inconsistent lags model (i.e. a model with incorrect shrinkage).

TABLE 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}	−0.0005	0.0073	0.0073	−0.0251	0.0361	0.0440	0.0000	0.0000	0.0000
Π_{12}	0.0000	0.0052	0.0052	0.0005	0.0406	0.0406	0.0000	0.0000	0.0000
Π_{21}	0.0000	0.0035	0.0035	0.0002	0.0301	0.0301	0.0000	0.0000	0.0000
Π_{22}	0.0004	0.0069	0.0069	−0.0244	0.0349	0.0426	0.0000	0.0000	0.0000
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}	0.0000	0.0000	0.0000	−0.0084	0.0118	0.0145	0.0000	0.0000	0.0000
Π_{12}	0.0000	0.0000	0.0000	−0.0001	0.0101	0.0101	0.0000	0.0000	0.0000
Π_{21}	0.0000	0.0000	0.0000	−0.0001	0.0134	0.0134	0.0000	0.0000	0.0000
Π_{22}	0.0000	0.0000	0.0000	−0.0082	0.0116	0.0142	0.0000	0.0000	0.0000

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (6.15). 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.

model) in all scenarios.[10] It is interesting to see that the probabilities of selecting the true ranks are not negatively affected either by adding lags to the model or by the lagged order selection being simultaneously performed with rank selection. Tables 6, 7, and 8 present the finite sample properties of GLS shrinkage, OLS, and oracle estimation. When compared with the oracle estimates, some components in the GLS shrinkage estimate even have smaller variances, though their finite sample biases are slightly larger. As a result, their root mean square errors are smaller than these of their counterparts in oracle estimation. Moreover, the GLS 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, as expected. The intuition that explains how the GLS shrinkage estimate can outperform the oracle estimate lies in the fact that there are some zero components in B_o and shrinking their estimates towards zero (but not exactly to zero) helps to reduce their bias and variance. From this perspective, the shrinkage estimates of the zero components in B_o share features similar to traditional shrinkage estimates, revealing that finite sample shrinkage bias is not always harmful.

Additional simulations were conducted to compare the performance of our least squares (LS) shrinkage techniques with the direct use of information criteria for model determination. The results are summarized here and presented in full in the Supplemental Appendix (Liao and Phillips, 2013). Amongst the usual information criteria, we find that BIC outperforms AIC and HQ and does well in selecting cointegrating rank even when the sample size is as small as $n = 100$, corroborating earlier findings in Cheng and Phillips (2009, 2012). In the determination of

TABLE 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}	0.0032	0.0609	0.0610	−0.0067	0.0551	0.0555	−0.0046	0.0548	0.0550
Π_{12}	−0.0023	0.0308	0.0308	−0.0066	0.0285	0.0293	−0.0023	0.0275	0.0276
Π_{21}	0.0015	0.0617	0.0617	−0.0035	0.0478	0.0480	−0.0018	0.0476	0.0477
Π_{22}	−0.0012	0.0308	0.0308	−0.0045	0.0246	0.0250	−0.0009	0.0238	0.0238
	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}	0.0008	0.0343	0.0343	−0.0027	0.0307	0.0308	−0.0020	0.0306	0.0307
Π_{12}	0.0004	0.0171	0.0171	−0.0013	0.0155	0.0157	−0.0007	0.0153	0.0154
Π_{21}	−0.0007	0.0312	0.0312	−0.0025	0.0276	0.0277	−0.0010	0.0275	0.0275
Π_{22}	−0.0004	0.0156	0.0156	−0.0016	0.0140	0.0140	−0.0003	0.0138	0.0138
	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}	0.0022	0.0833	0.0833	0.0008	0.0728	0.0728	−0.0055	0.0712	0.0714
Q_{12}	−0.0003	0.0069	0.0069	−0.0130	0.0243	0.0276	0.0000	0.0033	0.0033
Q_{21}	0.0008	0.0778	0.0779	0.0012	0.0658	0.0658	−0.0046	0.0643	0.0644
Q_{22}	−0.0003	0.0052	0.0052	−0.0119	0.0220	0.0251	0.0000	0.0004	0.0004
	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}	0.0004	0.0415	0.0415	−0.0003	0.0405	0.0405	−0.0023	0.0401	0.0401
Q_{12}	0.0000	0.0010	0.0010	0.0000	0.0081	0.0092	−0.0019	0.0010	0.0010
Q_{21}	0.0000	0.0371	0.0371	−0.0044	0.0368	0.0368	0.0000	0.0364	0.0364
Q_{22}	0.0000	0.0001	0.0001	−0.0040	0.0073	0.0083	0.0000	0.0001	0.0001

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (6.15). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is the RRR estimate with rank restriction r = 1.

transient dynamic structure, information criteria typically proceed by way of sequential selection working from the most general model to the most restrictive, largely for convenience and computational simplicity. Accordingly, these methods commonly miss true transient dynamic structures in which some subsets of lag coefficients are zero. In such cases, BIC and the other criteria may select the maximum lag correctly but miss the more complex dynamic structure. In comparison, LS shrinkage estimation performs well in selecting the true transient dynamic structure, the maximum lag in the transient dynamics, and the cointegrating rank.

TABLE 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}	−0.0228	0.0897	0.0926	−0.0104	0.0934	0.0940	−0.0104	0.0934	0.0940
Π_{12}	0.0384	0.0914	0.0992	−0.0008	0.0904	0.0904	−0.0008	0.0904	0.0904
Π_{21}	−0.0247	0.0995	0.1025	0.0016	0.0813	0.0813	0.0016	0.0813	0.0813
Π_{22}	0.0505	0.1459	0.1544	−0.0099	0.0780	0.0786	−0.0099	0.0780	0.0786
	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}	−0.0058	0.0524	0.0527	−0.0025	0.0523	0.0523	−0.0025	0.0523	0.0523
Π_{12}	0.0051	0.0545	0.0547	0.0009	0.0508	0.0509	0.0009	0.0508	0.0509
Π_{21}	−0.0049	0.0546	0.0548	−0.0019	0.0459	0.0459	−0.0019	0.0459	0.0459
Π_{22}	0.0075	0.0750	0.0754	−0.0037	0.0438	0.0440	−0.0037	0.0438	0.0440

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (6.15). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is simply the OLS estimate.

8. AN EMPIRICAL EXAMPLE

This section reports an empirical example to illustrate the application of these techniques to time series modeling of long-run and short-run behavior of aggregate income, consumption, and investment in the US economy. The sample[11] used in the empirical study is quarterly data over the period 1947–2009 from the *Federal Reserve Economic Data (FRED)*.

The sample data are shown in Figure 1 Evidently, the time series display long-term trend growth, which is especially clear in GNP and consumption, and some commonality in the growth mechanism over time. In particular, the series show evidence of some co-movement over the entire period. We therefore anticipate that modeling the series in terms of a VECM might reveal some nontrivial cointegrating relations. That is to say, we would expect cointegration rank r_o to satisfy $0 < r_o < 3$. These data were studied in Athanasopoulos et. al. (2011) who found on the same sample period and data that information criteria model selection produced a zero rank estimate for r_o and a single lag (ΔY_{t-1}) in the VECM.

Let $Y_t = (C_t, G_t, I_t)$, where C_t, G_t and I_t denote the logarithms of real consumption per capita, real GNP per capita and real investment per capita at period t respectively. For the same data as Athanasopoulos et. al. (2011) we applied our shrinkage methods to estimate the following system[12]

$$\Delta Y_t = \Pi Y_{t-1} + \sum_{k=1}^{3} B_k \Delta Y_{t-k} + u_t. \tag{8.1}$$

TABLE 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}	0.0000	0.0000	0.0000	−0.0019	0.0029	0.0035	0.0000	0.0000	0.0000
Π_{21}	0.0000	0.0000	0.0000	0.0000	0.0025	0.0025	0.0000	0.0000	0.0000
Π_{12}	0.0000	0.0000	0.0000	0.0000	0.0033	0.0033	0.0000	0.0000	0.0000
Π_{22}	0.0000	0.0000	0.0000	−0.0018	0.0029	0.0035	0.0000	0.0000	0.0000
$B_{1,11}$	−0.0301	0.0493	0.0577	−0.0069	0.0535	0.0540	−0.0044	0.0477	0.0479
$B_{1,21}$	−0.0006	0.0334	0.0334	−0.0007	0.0462	0.0462	−0.0008	0.0409	0.0409
$B_{1,12}$	−0.0006	0.0428	0.0428	−0.0017	0.0630	0.0631	−0.0011	0.0569	0.0569
$B_{1,22}$	−0.0304	0.0502	0.0587	−0.0079	0.0543	0.0549	−0.0048	0.0486	0.0489
$B_{2,11}$	0.0000	0.0013	0.0013	−0.0048	0.0575	0.0577	0.0000	0.0000	0.0000
$B_{2,21}$	0.0000	0.0001	0.0001	−0.0001	0.0502	0.0502	0.0000	0.0000	0.0000
$B_{2,12}$	−0.0000	0.0004	0.0004	0.0009	0.0664	0.0664	0.0000	0.0000	0.0000
$B_{2,22}$	0.0000	0.0009	0.0009	−0.0043	0.0577	0.0579	0.0000	0.0000	0.0000
$B_{3,11}$	−0.0315	0.0482	0.0576	−0.0068	0.0535	0.0539	−0.0061	0.0474	0.0478
$B_{3,21}$	0.0005	0.0337	0.0337	0.0004	0.0457	0.0458	0.0002	0.0411	0.0411
$B_{3,12}$	0.0009	0.0413	0.0413	0.0004	0.0612	0.0612	0.0011	0.0551	0.0552
$B_{3,22}$	−0.0318	0.0486	0.0581	−0.0073	0.0532	0.0537	−0.0058	0.0478	0.0482

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in equations (6.15) and (6.16). λ_o in each column represents the eigenvalues of Π_o. The oracle estimate in this case is simply the OLS estimate assuming that Π_o and B_{2o} are zero matrics.

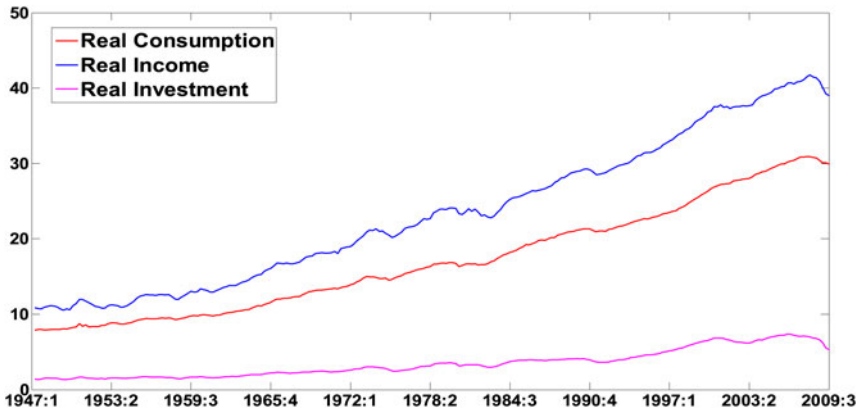

FIGURE 1. US GNP, consumption and investment in logarithms and in 2005 dollars. Data source: Federal Reserve Economic Data (FRED) St. Louis Fed.

TABLE 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}	−0.0012	0.0653	0.0653	−0.0015	0.0653	0.0653	−0.0006	0.0647	0.0647
Π_{21}	−0.0005	0.0564	0.0564	−0.0011	0.0563	0.0563	−0.0003	0.0558	0.0558
Π_{12}	−0.0006	0.0326	0.0326	−0.0009	0.0327	0.0327	−0.0003	0.0324	0.0324
Π_{22}	−0.0002	0.0282	0.0282	−0.0007	0.0282	0.0282	−0.0002	0.0279	0.0279
$B_{1,11}$	−0.1086	0.0536	0.1211	−0.0028	0.0572	0.0572	−0.0022	0.0532	0.0533
$B_{1,21}$	−0.0766	0.0432	0.0880	−0.0024	0.0490	0.0491	−0.0021	0.0461	0.0462
$B_{1,12}$	−0.0351	0.0660	0.0747	−0.0019	0.0769	0.0769	−0.0022	0.0727	0.0728
$B_{1,22}$	−0.0281	0.0643	0.0702	−0.0018	0.0672	0.0672	−0.0019	0.0633	0.0633
$B_{2,11}$	0.0000	0.0000	0.0000	−0.0010	0.0438	0.0438	0.0000	0.0000	0.0000
$B_{2,21}$	0.0000	0.0000	0.0000	−0.0012	0.0378	0.0378	0.0000	0.0000	0.0000
$B_{2,12}$	0.0000	0.0000	0.0000	−0.0015	0.0789	0.0789	0.0000	0.0000	0.0000
$B_{2,22}$	0.0000	0.0000	0.0000	−0.0005	0.0674	0.0674	0.0000	0.0000	0.0000
$B_{3,11}$	−0.1206	0.0336	0.1252	−0.0032	0.0424	0.0425	−0.0023	0.0375	0.0375
$B_{3,21}$	−0.0825	0.0295	0.0876	−0.0029	0.0373	0.0374	−0.0021	0.0327	0.0328
$B_{3,12}$	−0.1010	0.0388	0.1082	−0.0020	0.0701	0.0701	−0.0017	0.0523	0.0523
$B_{3,22}$	−0.0730	0.0460	0.0862	−0.0029	0.0611	0.0611	−0.0020	0.0461	0.0462

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in equations (6.15) and (6.16). 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 $B_{2o} = 0$.

Unrestricted LS estimation of this model produced eigenvalues 0.0025 and $-0.0493 \pm 0.0119i$, which indicates that Π might contain at least one zero eigenvalue as the positive eigenvalue estimates 0.0025 is close to zero. The LS estimates of the lag coefficients B_k are

$$\widehat{B}_{1,1st} = \begin{pmatrix} .14 & -.03 & .16 \\ .72 & -.18 & .97 \\ .19 & .02 & .35 \end{pmatrix}, \ \widehat{B}_{2,1st} = \begin{pmatrix} .33 & -.09 & .10 \\ .43 & -.06 & .23 \\ .16 & -.06 & .07 \end{pmatrix}, \ \widehat{B}_{3,1st} = \begin{pmatrix} .31 & -.20 & .24 \\ .19 & -.11 & -.15 \\ .09 & -.03 & .06 \end{pmatrix}.$$

From these estimates it is by no means clear which lagged differences should be ruled out from (8.1). From their magnitudes, it seems that ΔY_{t-1}, ΔY_{t-2} and ΔY_{t-3} might all be included in the empirical model.

We applied LS shrinkage estimation to the model (8.1). Using the LS estimate, we constructed an adaptive penalty for GLS shrinkage estimation. We first tried GLS shrinkage estimation with tuning parameters

$$\lambda_{r,k,n} = \frac{2\log(n)}{n} ||\phi_k(\widehat{\Pi}_{1st})||^{-2} \text{ and } \lambda_{b,j,n} = \frac{18\log(n)}{n} ||\widehat{B}_{j,1st}||^{-2}$$

TABLE 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}	0.0489	0.0521	0.0715	−0.0024	0.0637	0.0637	−0.0034	0.0514	0.0515
Π_{21}	0.0140	0.0488	0.0508	0.0009	0.0552	0.0552	0.0001	0.0441	0.0441
Π_{12}	−0.0214	0.0432	0.0482	0.0010	0.0486	0.0486	0.0013	0.0407	0.0407
Π_{22}	0.0124	0.0531	0.0545	−0.0009	0.0416	0.0416	−0.0008	0.0349	0.0350
$B_{1,11}$	−0.0852	0.0528	0.1003	−0.0019	0.0644	0.0644	−0.0004	0.0579	0.0579
$B_{1,21}$	−0.0089	0.0436	0.0445	−0.0020	0.0559	0.0560	−0.0013	0.0504	0.0505
$B_{1,12}$	0.0093	0.0426	0.0437	−0.0020	0.0580	0.0580	−0.0023	0.0540	0.0540
$B_{1,22}$	−0.0480	0.0490	0.0686	−0.0025	0.0500	0.0501	−0.0021	0.0469	0.0469
$B_{2,11}$	−0.0000	0.0000	0.0000	−0.0008	0.0577	0.0577	0.0000	0.0000	0.0000
$B_{2,21}$	0.0000	0.0000	0.0000	−0.0011	0.0501	0.0501	0.0000	0.0000	0.0000
$B_{2,12}$	0.0000	0.0000	0.0000	0.0002	0.0573	0.0573	0.0000	0.0000	0.0000
$B_{2,22}$	−0.0000	0.0000	0.0000	−0.0001	0.0498	0.0498	0.0000	0.0000	0.0000
$B_{3,11}$	−0.0728	0.0484	0.0875	−0.0051	0.0545	0.0547	−0.0038	0.0518	0.0519
$B_{3,21}$	−0.0011	0.0367	0.0367	−0.0008	0.0478	0.0478	−0.0004	0.0450	0.0450
$B_{3,12}$	−0.0014	0.0439	0.0439	0.0009	0.0559	0.0559	0.0008	0.0555	0.0555
$B_{3,22}$	−0.0565	0.0524	0.0770	−0.0033	0.0479	0.0480	−0.0029	0.0475	0.0476

Note: Replications = 5000, $\omega = 2$, adaptive tuning parameter λ_n given in equation (6.15) and (6.16). 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 $B_{2o} = 0$.

for $k, j = 1, 2, 3$. The eigenvalues of the GLS shrinkage estimate of Π are 0.0000394, −0.0001912, and 0, which implies that Π contains one zero eigenvalue. There are two nonzero eigenvalue estimates which are both close to zero. The effect of the adaptive penalty on these two estimates is substantial because of the small magnitudes of the eigenvalues of the original LS estimate of Π. As a result, the shrinkage bias in the two nonzero eigenvalue estimates is likely to be large. The GLS shrinkage estimates of B_2 and B_3 are zero, while the GLS shrinkage estimate of B_1 is

$$\widehat{B}_1 = \begin{pmatrix} .0687 & .1076 & .0513 \\ .4598 & .1212 & .4053 \\ .0986 & .1123 & .2322 \end{pmatrix}.$$

Using the results from the above GLS shrinkage estimation, we construct the adaptive loading parameters in (6.8) and (6.9). Using the adaptive tuning parameters in (6.10) and (6.11), we perform a further GLS shrinkage estimation of the empirical model (8.1). The eigenvalues of the new GLS shrinkage estimate of Π are $-0.0226 \pm 0.0158i$ and 0, which again imply that Π contains one zero eigenvalue. Of course, the new nonzero eigenvalue estimates also contains nontrivial

shrinkage bias. The new GLS shrinkage estimates of B_2 and B_3 are zero, but the estimate of B_1 becomes

$$\widehat{B}_1 = \begin{pmatrix} .0681 & .1100 & .0115 \\ .4288 & .1472 & .4164 \\ .1054 & .1136 & .1919 \end{pmatrix}.$$

Finally, we run a post-Lasso RRR estimation based on the cointegration rank and lagged difference selected in the above GLS shrinkage estimation. The RRR estimates are the following

$$\Delta Y_t = \begin{pmatrix} .026 & -.022 \\ .082 & -.026 \\ -.012 & .013 \end{pmatrix} \begin{pmatrix} .822 & -.555 & -.128 \\ -.265 & .378 & -.887 \end{pmatrix} Y_{t-1} + \begin{pmatrix} .127 & .028 & .312 \\ .598 & -.088 & 1.098 \\ .161 & .055 & .364 \end{pmatrix} \Delta Y_{t-1} + \widehat{u}_t,$$

where the eigenvalues of the RRR estimate of Π are -0.0262, -0.0039, and 0. To sum up, this empirical implementation of our approach estimates cointegrating rank r_o to be 2 and selects one lagged difference in the VECM (8.1). These results corroborate the manifestation of co-movement in the three time series G_t, C_t, and I_t through the presence of two cointegrating vectors in the fitted model, whereas traditional information criteria fail to find any co-movement in the data and set cointegrating rank to be zero.

9. 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 paper 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.

As indicated in the Introduction, sequential methods can encounter obstacles to consistent order estimation even when test size is driven to zero as the sample size $n \to \infty$. For instance, in the model (7.3) considered earlier

$$\Delta Y_t = \Pi_o Y_{t-1} + B_{o,1}\Delta Y_{t-1} + B_{o,2}\Delta Y_{t-2} + B_{o,3}\Delta Y_{t-3} + u_t,$$

where $\|B_{o,2}\| = 0$, $\|B_{o,1}\| \neq 0$, and $\|B_{o,3}\| \neq 0$. It is clear that in this model both upward and downward sequential testing procedures either include the second lag difference or exclude it together with the third lag difference. As a result, the true model is never correctly selected by such standard algorithms - much more intensive searches are required. In the more general model

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

where p is large but fixed, the model selection limitations of standard sequential testing are inevitably worse, although these may be mitigated by orthonormalization, parsimonious encompassing, and other automated devices (Hendry and Krolzig, 2005; Hendry and Johansen, 2015). The methods of the present paper do not require any specific order or format of the lag differences to ensure consistent model selection. As a result, the approach is invariant to permutations of the order of the lag differences. Moreover, the method is easier to implement in empirical work, requires no intensive cross lag search procedures, is automated with data-based tuning parameter selection, and is computationally straightforward.

Various extensions of the methods developed here seem desirable. One rather obvious (and simple) 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. Third, it will be interesting and useful, given the growing availability of large dimensional data sets in macroeconomics and finance, to extend the results of the paper to high dimensional VEC systems where the dimension m of the matrix Π_o and the length p of the lag order are large. 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.

NOTES

1. Note that when $m - r_o > 1$, the normalizations $\alpha'_{o,\perp}\alpha_{o,\perp} = I_{m-r_o}$ and $\beta'_{o,\perp}\beta_{o,\perp} = I_{m-r_o}$ are not sufficient to ensure the uniqueness of $\alpha_{o,\perp}$ and $\beta_{o,\perp}$. In the paper, we only need the existence of normalized $\alpha_{o,\perp}$ and $\beta_{o,\perp}$ such that $\alpha'_{o,\perp}\alpha_o = 0$ and $\beta'_{o,\perp}\beta_o = 0$.

2. The tranform of the matrix Π is important for rank selection because, by virtue of the consistency of the first step estimator $\widehat{\Pi}_{1st}$, $Q_n\Pi_o$ has (and only has) $m - r_o$ rows which are asymptotically

nonzero. Note that the fact that a matrix Π does not have full rank does not necessarily mean that any element in Π should be zero. Hence penalized LS regression in (2.4) with a group Lasso penalty on Π does not deliver any implication for rank selection in general case.

3. The new penalty is defined as a function on $R^{m\times m}$, i.e. on the square matrix Π. While this formulation is relevant in the present setting, it is clear that the approach can be trivially extended to the general case with any matrix.

4. As indicated, the idea in Yuan et al. (2007) is related to the original approach pursued in an earlier version (2010) of the present paper. In that version, we showed that when adding the L-1 penalty on the eigenvalues to the LS criterion, the $m - r_o$ smallest eigenvalues of the penalized LS estimate of the cointegration matrix Π_o have convergence rate faster than n^{-1}. This result has implications for efficient estimation of the VECM when the true model is nested. But it does not necessarily imply model selection because selection requires that zero eigenvalues be estimated as zeros with positive probability. That is a challenging problem due to the highly nonlinear relation between Π_o and its eigenvalues. The approach pursued in the present paper is far simpler, enhancing implementation and leading directly to the required asymptotic result.

5. The adaptive penalization means that the penalization on the estimators of zero components (e.g., zero matrices $B_{o,j}$) is large, while the penalization on the estimators of nonzero components (e.g., nonzero matrices $B_{o,j}$) is small.

6. 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.

7. The eigenvectors in Q_1 are ordered according to the magnitudes of the eigenvalues, i.e. the ordering of the eigenvalues of Π_1.

8. The same intuition applies to the scenario where Assumption LP holds.

9. The proof of Lemma 6.2 is in the supplemental appendix of this paper.

10. Joint determination of the lagged differences and cointegration rank can also be performed using information criteria like AIC and BIC, as suggested in Phillips and McFarland (1997) and Chao and Phillips (1999). As discussed below, the supplemental appendix provides simulation comparisons between information criteria and LS shrinkage estimation.

11. We thank George Athanasopoulos for providing the data.

12. The system (8.1) was fitted with and without an intercept. The findings were very similar and in both cases cointegrating rank was found to be 2. Results are reported here for the fitted intercept case. Of course, Lasso methods can also be applied to determine whether an intercept should appear in each equation or in any long-run relation that might be found. That extension of Lasso is not considered in the present paper. It is likely to be important in forecasting.

REFERENCES

Anderson, T.W. (2002) Reduced rank regression in cointegrated models. *Journal of Econometrics* 106, 203–216.

Athanasopoulos, G., O.T.C. Guillen, J.V. Issler, & F. Vahid (2011) Model selection, estimation and forecasting in VAR models with short-run and long-run restrictions. *Journal of Econometrics* 164(1), 116–129.

Belloni, A. & V. Chernozhukov (2013) Least squares after model selection in high-dimensional sparse models. *Bernoulli* 19, 521–547.

Caner, M. & K. Knight (2013) An alternative to unit root tests: Bridge estimators differentiate between nonstationary versus stationary models and select optimal lag. *Journal of Statistical Planning and Inference* 143, 691–715.

Chao, J. & 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. & P.C.B. Phillips (2009) Semiparametric cointegrating rank selection. *Econometrics Journal* 12, S83–S104.

Cheng, X. & P.C.B. Phillips (2012) Cointegrating rank selection in models with time-varying variance. *Journal of Econometrics* 142(1), 201–211.

Hendry, D.F. & S. Johansen (2015) Model discovery and Trygve Haavelmo.s legacy. *Econometric Theory* 31, 93.114.

Hendry, D.F. & H.-M. Krolzig (2005) The properties of automatic gets modelling. *Economic Journal* 115, C32–C61.

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.

Kock, A. & L. Callot (2012) Oracle Inequalities for High Dimensional Vector Autoregressions. CREATES Research Paper 2012–16,

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

Leeb, H. & B.M. Pötscher (2008). Sparse estimators and the oracle property, or the return of the Hodges estimator. *Journal of Econometrics* 142(1), 201–211.

Liao, Z. & P.C.B. Phillips (2013) Supplemental Material for 'Automated Estimation of Vector Error Correction Models'. Mimeo, Yale University and UCLA. Unpublished paper.

Peng, J., J. Zhu, A. Bergamaschi, W. Han, D.-Y. Noh, J.R. Pollack, & P. Wang (2010) Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. *Annals of Applied Statistics* 4, 53–77.

Phillips, P.C.B. (1991a) Optimal inference in cointegrated systems. *Econometrica* 59(2), 283–306.

Phillips, P.C.B. (1991b) Spectral regression for cointegrated time series. In W. Barnett, J. Powell, & G. Tauchen (eds.), *Nonparametric and Semiparametric Methods in Economics and Statistics*, pp. 413–435. Cambridge University Press.

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 asymtotics in nonstationary VARs. *Journal of Econometrics* 83, 21–56.

Phillips, P.C.B. & J.W. McFarland (1997) Forward exchange market unbiasedness: The case of the Australian dollar since 1984. *Journal of International Money and Finance* 16, 885–907.

Phillips, P.C.B. & V. Solo (1992) Asymptotics for linear processes. *Annals of Statistics* 20(2), 971–1001.

Song, S. & P. Bickel (2009) Large Vector Auto Regressions. SFB 649 Discussion Paper 2011–048.

Yuan, M. & Y. Lin (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B* 68, 49–67.

Model selection and estimation in the Gaussian graphical model. *Biometrika* 94, 19–35.

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

APPENDIX

We start with some standard preliminary results and then prove the main results in each of the sections of the paper in turn, together with various lemmas that are useful in those derivations. Additional technical results are provided in the Supplemental Appendix.

A.1. Some Auxiliary Results

Denote

$$\widehat{S}_{12} = \sum_{t=1}^{n} \frac{Z_{1,t-1} Z'_{2,t-1}}{n}, \; S_{21} = \sum_{t=1}^{n} \frac{Z_{2,t-1} Z'_{1,t-1}}{n},$$

$$\widehat{S}_{11} = \sum_{t=1}^{n} \frac{Z_{1,t-1} Z'_{1,t-1}}{n} \text{ and } \widehat{S}_{22} = \sum_{t=1}^{n} \frac{Z_{2,t-1} Z'_{2,t-1}}{n}.$$

The following lemma is standard and useful.

LEMMA A.1. *Under Assumptions 3.1 and 3.2, we have*

(a) $\widehat{S}_{11} \to_p \Sigma_{z_1 z_1}$;
(b) $\widehat{S}_{21} \to_d -\int B_{w_2} dB'_{w_1} (\alpha'_o \beta_o)^{-1} + \Gamma_{w_2 z_1}$;
(c) $n^{-1} \widehat{S}_{22} \to_d \int B_{w_2} B'_{w_2}$;
(d) $n^{-\frac{1}{2}} \sum_{t=1}^{n} u_t Z'_{1,t-1} \to_d N(0, \Omega_u \otimes \Sigma_{z_1 z_1})$;
(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 (b), (c), (d), and (e) converge jointly.

Proof of Lemma A.1. See Johansen (1995) and Cheng and Phillips (2009). ■

A.2. Proofs of Main Results in Section 3

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

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

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

LEMMA A.2. *Under Assumptions 3.1 and 3.2, we have:*

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

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

where $B_{m,1}$ *and* $B_{m,2}$ *are defined in Theorem 3.5;*

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

$$n \left(\phi_1 \left(\widehat{\Pi}_{1st} \right), \ldots, \phi_{m-r_o} \left(\widehat{\Pi}_{1st} \right) \right) \to_d \left(\widetilde{\phi}_{o,1}, \ldots, \widetilde{\phi}_{o,m-r_o} \right), \tag{A.3}$$

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

$$\left| \mu I_{m-r_0} - \left(\int dB_{w_2} B'_{w_2} \right) \left(\int B_{w_2} B'_{w_2} \right)^{-1} \right| = 0. \tag{A.4}$$

The proof of Lemma A.2 is in the supplemental appendix of this paper. Lemma A.2 is useful because the OLS estimate $\widehat{\Pi}_{1st}$ and the related eigenvalue estimates can be used to construct adaptive penalty in the tuning parameters. 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 subdivide the matrices P_n and Q_n as $P_n = \left[P_{\alpha,n}, P_{\alpha_\perp,n}\right]$ 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,

$$Q_{\alpha_\perp,n} P_{\alpha_\perp,n} = I_{m-r_o},\ Q_{\alpha,n} P_{\alpha_\perp,n} = \mathbf{0}_{r_o\times(m-r_o)} \text{ and}$$
$$Q_{\alpha_\perp,n} \widehat{\Pi}_{1st} = \Lambda_{\alpha_\perp,n} Q_{\alpha_\perp,n}, \tag{A.5}$$

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 (A.5), we can define a useful estimator of Π_o as

$$\Pi_{n,f} = \widehat{\Pi}_{1st} - P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n} Q_{\alpha_\perp,n}. \tag{A.6}$$

The estimator $\Pi_{n,f}$ is infeasible because r_o is unknown. $\Pi_{n,f}$ may be interpreted as a modification to the unrestricted estimate $\widehat{\Pi}_{1st}$ which removes components in the eigen-representation of the unrestricted estimate that correspond to the smallest $m - r_o$ eigenvalues.

By definition

$$Q_{\alpha,n}\Pi_{n,f} = Q_{\alpha,n}\widehat{\Pi}_{1st} - Q_{\alpha,n}P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n} = \Lambda_{\alpha,n}Q_{\alpha,n}, \tag{A.7}$$

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

$$Q_{\alpha_\perp,n}\Pi_{n,f} = Q_{\alpha_\perp,n}\widehat{\Pi}_{1st} - Q_{\alpha_\perp,n}P_{\alpha_\perp,n}\Lambda_{\alpha_\perp,n}Q_{\alpha_\perp,n} = \mathbf{0}_{(m-r_o)\times m}. \tag{A.8}$$

From Lemma A.2.(b), (A.7), and (A.8), 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 Lemmas A.2.(a) and (c),

$$\left(\Pi_{n,f} - \Pi_o\right) Q^{-1} D_n^{-1} = O_p(1). \tag{A.9}$$

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 (A.9) we can compare the LS shrinkage estimator $\widehat{\Pi}_n$ with $\Pi_{n,f}$ to establish the consistency and convergence rate of $\widehat{\Pi}_n$.

Proof of Theorem 3.1. Define

$$V_n(\Pi) = \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} \right\|^2 + n \sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\|.$$

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 $\Delta y = vec(\Delta Y)$, $\Delta Y = (\Delta Y_1, \ldots, \Delta Y_n)_{m \times n}$ and $Y_{-1} = \left(Y_0, \ldots, Y_{T-1}\right)_{m \times n}$.

By definition, $V_n\left(\widehat{\Pi}_n\right) \leq V_n\left(\Pi_{n,f}\right)$ and thus

$$\begin{aligned} & vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)' \left(\textstyle\sum_{t=1}^{n} Y_{t-1} Y'_{t-1} \otimes I_m \right) vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right) \\ & \quad + 2 vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)' vec\left(\textstyle\sum_{t=1}^{n} Y_{t-1} u'_t \right) \\ & \quad + 2 vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)' \left(\textstyle\sum_{t=1}^{n} Y_{t-1} Y'_{t-1} \otimes I_m \right) vec\left(\Pi_o - \Pi_{n,f}\right) \\ & \leq n \sum_{k=1}^{m} \lambda_{r,k,n} \left[||\Phi_{n,k}\left(\Pi_{n,f}\right)|| - ||\Phi_{n,k}\left(\widehat{\Pi}_n\right)|| \right]. \end{aligned} \tag{A.10}$$

When $r_o = 0$, ΔY_t is stationary and Y_t is full rank $I(1)$, so that

$$n^{-2} \sum_{t=1}^{n} Y_{t-1} Y'_{t-1} \to_d \int_0^1 B_u(a) B'_u(a) da \text{ and}$$

$$n^{-2} \sum_{t=1}^{n} Y_{t-1} u'_t = O_p\left(n^{-1}\right). \tag{A.11}$$

From the results in (A.10) and (A.11), we get

$$\mu_{n,\min} ||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2(c_{1,n} + c_{2,n}) ||\widehat{\Pi}_n - \Pi_{n,f}|| - d_n \leq 0, \tag{A.12}$$

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,

$$\begin{aligned} c_{1,n} &= ||n^{-2} \textstyle\sum_{t=1}^{n} Y_{t-1} u'_t||, \\ c_{2,n} &= m \left\| n^{-2} \textstyle\sum_{t=1}^{n} Y_{t-1} Y'_{t-1} \right\| ||\Pi_{n,f} - \Pi_o||, \text{ and} \\ d_n &= n^{-1} \sum_{k=1}^{m} \lambda_{r,k,n} ||\Phi_{n,k}\left(\Pi_{n,f}\right)||. \end{aligned} \tag{A.13}$$

Under (A.9) and (A.11), $c_{1,n} = o_p(1)$ and $c_{2,n} = o_p(1)$. Under (A.7), (A.8), and $\lambda_{r,k,n} = o_p(1)$ for all $k \in \mathcal{S}_\phi$,

$$d_n = n^{-1} \sum_{k=1}^{r_o} \lambda_{r,k,n} ||\Phi_{n,k}(\Pi_{n,f})|| = o_p\left(n^{-1}\right). \tag{A.14}$$

From (A.12), (A.13), and (A.14), 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

$$n^{-1}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \to_p \Sigma_{yy} = R(1)\Omega_u R(1)' \text{ and}$$

$$n^{-1}\sum_{t=1}^{n} Y_{t-1}u_t' = O_p\left(n^{-\frac{1}{2}}\right). \tag{A.15}$$

From the results in (A.10) and (A.15), we get

$$\mu_{n,\min}||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2n(c_{1,n}+c_{2,n})||\widehat{\Pi}_n - \Pi_{n,f}|| - nd_n \le 0, \tag{A.16}$$

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 (A.14). It is clear that $nc_{1,n} = o_p(1)$ and $nc_{2,n} = o_p(1)$ under (A.15) and (A.9), and $nd_n = o_p(1)$ under (A.14). So, consistency of $\widehat{\Pi}_n$ follows directly from the inequality in (A.16), 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 A.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

$$\begin{aligned}&vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)'\left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right)vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right) \\ &\quad \ge \mu_{n,\min}||\left(\widehat{\Pi}_n - \Pi_{n,f}\right)B_n||^2,\end{aligned} \tag{A.17}$$

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

$$\left|\left[vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)\right]'vec\left(B_nD_n\sum_{t=1}^{n} Z_{t-1}u_t'\right)\right| \le ||\left(\widehat{\Pi}_n - \Pi_{n,f}\right)B_n||e_{1,n} \tag{A.18}$$

and

$$\begin{aligned}&\left|vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)'\left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right)vec\left(\Pi_o - \Pi_{n,f}\right)\right| \\ &\quad \le ||\left(\widehat{\Pi}_n - \Pi_{n,f}\right)B_n||e_{2,n},\end{aligned} \tag{A.19}$$

where

$$\begin{aligned}&e_{1,n} = ||D_n\textstyle\sum_{t=1}^{n} Z_{t-1}u_t'|| \text{ and} \\ &e_{2,n} = m||D_n\textstyle\sum_{t=1}^{n} Z_{t-1}Z_{t-1}'D_n|| \times ||(\Pi_{n,f} - \Pi_o)B_n||.\end{aligned} \tag{A.20}$$

Under Lemma A.1 and (A.9), $e_{1,n} = O_p(1)$ and $e_{2,n} = O_p(1)$. From (A.10), (A.17), (A.18), (A.19), we have the inequality

$$\mu_{n,\min}||\left(\widehat{\Pi}_n - \Pi_{n,f}\right)B_n||^2 - 2\left(e_{1,n}+e_{2,n}\right)||\left(\widehat{\Pi}_n - \Pi_{n,f}\right)B_n|| - nd_n \leq 0, \tag{A.21}$$

which implies

$$\left(\widehat{\Pi}_n - \Pi_{n,f}\right)B_n = O_p\left(1+\sqrt{n}d_n^{\frac{1}{2}}\right). \tag{A.22}$$

By the definition of B_n, (A.9) and (A.22), we deduce that

$$\widehat{\Pi}_n - \Pi_o = O_p\left(n^{-\frac{1}{2}} + d_n^{\frac{1}{2}}\right) = o_p(1),$$

which implies the consistency of $\widehat{\Pi}_n$. ■

Proof of Theorem 3.2. By the triangle inequality and (A.8), we have

$$\begin{aligned}
&\sum_{k=1}^{m}\lambda_{r,k,n}\left[||\Phi_{n,k}(\Pi_{n,f})|| - ||\Phi_{n,k}\left(\widehat{\Pi}_n\right)||\right] \\
&\quad\leq \sum_{k=1}^{r_o}\lambda_{r,k,n}\left[||\Phi_{n,k}(\Pi_{n,f})|| - ||\Phi_{n,k}\left(\widehat{\Pi}_n\right)||\right] \\
&\quad\leq r_o \max_{k\in\mathcal{S}_\phi}\lambda_{r,k,n}||\widehat{\Pi}_n - \Pi_{n,f}||.
\end{aligned} \tag{A.23}$$

Using (A.23) and invoking the inequality in (A.10) we get

$$\begin{aligned}
&vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)'\left(\textstyle\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'\otimes I_m\right)vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right) \\
&\quad + 2vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)'vec\left(\textstyle\sum_{t=1}^{n} Y_{t-1}u_t'\right) \\
&\quad + 2vec\left(\Pi_{n,f} - \widehat{\Pi}_n\right)'\left(\textstyle\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'\otimes I_m\right)vec\left(\Pi_o - \Pi_{n,f}\right) \\
&\quad\leq nr_o\delta_{r,n}||\widehat{\Pi}_n - \Pi_{n,f}||.
\end{aligned} \tag{A.24}$$

When $r_o = 0$, we use (A.13) and (A.24) to obtain

$$\mu_{n,\min}||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2\left(c_{1,n}+c_{2,n}+n^{-1}r_o\delta_{r,n}\right)||\widehat{\Pi}_n - \Pi_{n,f}|| \leq 0, \tag{A.25}$$

where under (A.11) $c_{1,n} = O_p\left(n^{-1}\right)$ and $c_{2,n} = O_p\left(n^{-1}\right)$. We deduce from the inequality (A.25) and (A.9) that

$$\widehat{\Pi}_n - \Pi_o = O_p\left(n^{-1} + n^{-1}\delta_{r,n}\right). \tag{A.26}$$

When $r_o = m$, we use (A.24) to obtain

$$\mu_{n,\min}||\widehat{\Pi}_n - \Pi_{n,f}||^2 - 2n\left(c_{1,n}+c_{2,n}+n^{-1}r_o\delta_{r,n}\right)||\widehat{\Pi}_n - \Pi_{n,f}|| \leq 0, \tag{A.27}$$

where $nc_{1,n} = ||\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}u_t'|| = O_p\left(n^{-\frac{1}{2}}\right)$ and $nc_{2,n} = O_p\left(n^{-\frac{1}{2}}\right)$ by Lemma A.1 and (A.9). The inequality (A.27) and (A.9) lead to

$$\widehat{\Pi}_n - \Pi_o = O_p\left(n^{-\frac{1}{2}} + \delta_{r,n}\right). \tag{A.28}$$

When $0 < r_o < m$, we can use the results in (A.17), (A.18), (A.19), (A.20), and (A.24) to deduce that

$$\begin{aligned}&\mu_{n,\min}||\left(\Pi_{n,f} - \widehat{\Pi}_n\right)B_n||^2 - 2\left(e_{1,n} + e_{2,n}\right)||\left(\Pi_{n,f} - \widehat{\Pi}_n\right)B_n|| \\ &\quad \le r_o n\delta_{r,n}||\Pi_{n,f} - \widehat{\Pi}_n||,\end{aligned} \tag{A.29}$$

where $e_{1,n} = \left\| D_n Q \sum_{t=1}^{n} Y_{t-1}u_t' \right\| = O_p(1)$ and $e_{2,n} = O_p(1)$ by Lemma A.1 and (A.9). By the definition of B_n,

$$||\left(\Pi_{n,f} - \widehat{\Pi}_n\right)B_n B_n^{-1}|| \le cn^{-\frac{1}{2}}||\left(\Pi_{n,f} - \widehat{\Pi}_n\right)B_n||, \tag{A.30}$$

where c is some finite positive constant. Using (A.29), (A.30) and (A.9), we get

$$\left(\widehat{\Pi}_n - \Pi_o\right)B_n = O_p\left(1 + n^{\frac{1}{2}}\delta_{r,n}\right), \tag{A.31}$$

which finishes the proof. ■

Proof of Theorem 3.3. To facilitate the proof, we rewrite the LS shrinkage estimation problem as

$$\widehat{T}_n = \underset{T\in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \left\| \Delta Y_t - P_n T Y_{t-1} \right\|^2 + n\sum_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(P_n T) \right\|. \tag{A.32}$$

By definition, $\widehat{\Pi}_n = P_n\widehat{T}_n$ and $\widehat{T}_n = Q_n\widehat{\Pi}_n$ for all n. Under (3.6) and (3.7),

$$\widehat{T}_n = \begin{pmatrix} Q_{\alpha,n}\widehat{\Pi}_n \\ Q_{\alpha_\perp,n}\widehat{\Pi}_n \end{pmatrix} = \begin{pmatrix} Q_{\alpha,n}\widehat{\Pi}_{1st} \\ Q_{\alpha_\perp,n}\widehat{\Pi}_{1st} \end{pmatrix} + o_p(1). \tag{A.33}$$

Results in (3.8) 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, $\Phi_{n,k}(P_n T) = Q_n(k)P_n T = T(k)$ and the problem in (A.32) can be rewritten as

$$\widehat{T}_n = \underset{T\in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \left\| \Delta Y_t - P_n T Y_{t-1} \right\|^2 + n\sum_{k=1}^{m} \lambda_{r,k,n} \left\| T(k) \right\|, \tag{A.34}$$

which has the following Karush-Kuhn-Tucker (KKT) optimality conditions

$$\begin{cases} \frac{1}{n}\sum_{t=1}^{n}\left(\Delta Y_t - P_n\widehat{T}_n Y_{t-1}\right)' P_n(k)Y_{t-1}' = \frac{\lambda_{r,k,n}}{2}\frac{\widehat{T}_n(k)}{||\widehat{T}_n(k)||} & \text{if } \widehat{T}_n(k) \ne 0 \\ \left\| \frac{1}{n}\sum_{t=1}^{n}\left(\Delta Y_t - P_n\widehat{T}_n Y_{t-1}\right)' P_n(k)Y_{t-1}' \right\| \le \frac{\lambda_{r,k,n}}{2} & \text{if } \widehat{T}_n(k) = 0 \end{cases}, \tag{A.35}$$

for $k = 1, \ldots, m$. Conditional on the event $\{Q_n(k_o)\widehat{\Pi}_n \ne 0\}$ for some k_o satisfying $r_o < k_o \le m$, we obtain the following equation from the KKT optimality conditions

$$\left\| \frac{1}{n}\sum_{t=1}^{n}\left(\Delta Y_t - P_n\widehat{T}_n Y_{t-1}\right)' P_n(k_o)Y_{t-1}' \right\| = \frac{\lambda_{r,k_o,n}}{2}. \tag{A.36}$$

The sample average in the left hand side of (A.36) can be rewritten as

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n}\left(\Delta Y_t - P_n\widehat{T}_n Y_{t-1}\right)' P_n(k_o)Y'_{t-1} \\
&\quad = \frac{1}{n}\sum_{t=1}^{n}\left[u_t - \left(\widehat{\Pi}_n - \Pi_o\right)Y_{t-1}\right]' P_n(k_o)Y'_{t-1} \\
&\quad = \frac{P'_n(k_o)\sum_{t=1}^{n} u_t Y'_{t-1}}{n} - \frac{P'_n(k_o)\left(\widehat{\Pi}_n - \Pi_o\right)\sum_{t=1}^{n} Y_{t-1}Y'_{t-1}}{n}. \qquad \textbf{(A.37)}
\end{aligned}$$

Under Lemmas A.2 and A.1 and Theorem 3.2

$$\frac{P'_n(k_o)\sum_{t=1}^{n} u_t Y'_{t-1}}{n} = O_p(1) \qquad \textbf{(A.38)}$$

and

$$\begin{aligned}
&\frac{P'_n(k_o)\left(\widehat{\Pi}_n - \Pi_o\right)\sum_{t=1}^{n} Y_{t-1}Y'_{t-1}}{n} \\
&\quad = P'_n(k_o)\left(\widehat{\Pi}_n - \Pi_o\right)Q^{-1}D_n^{-1}\frac{D_n\sum_{t=1}^{n} Z_{t-1}Z'_{t-1}}{n}Q'^{-1} = O_p(1). \qquad \textbf{(A.39)}
\end{aligned}$$

Using the results in (A.37), (A.38), and (A.39), we deduce that

$$\left\|\frac{1}{n}\sum_{t=1}^{n}\left(\Delta Y_t - P_n\widehat{T}_n Y_{t-1}\right)' P_n(k_o)Y'_{t-1}\right\| = O_p(1). \qquad \textbf{(A.40)}$$

By the assumption on the tuning parameters, we have $\frac{\lambda_{r,k_o,n}}{2} \to_p \infty$, which together with the results in (A.36) and (A.40) implies that

$$\Pr\left(Q_n(k_o)\widehat{\Pi}_n = 0\right) \to 1 \text{ as } n \to \infty.$$

As the above result holds for any k_o such that $r_o < k_o \leq m$, this finishes the proof. ■

Proof of Theorem 3.5. From Corollary 3.4, 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

$$\Pi_o = \alpha_o\beta'_o = [\alpha_o, \alpha_o O_{r_o}]. \qquad \textbf{(A.41)}$$

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. 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

$$\beta'_{1,o,\perp} + \beta'_{2,o,\perp}O'_{r_o} = 0 \text{ and } \beta'_{1,o,\perp}\beta_{1,o,\perp} + \beta'_{2,o,\perp}\beta_{2,o,\perp} = I_{m-r_o} \qquad \textbf{(A.42)}$$

which implies that

$$\beta'_{1,o,\perp} = -\beta'_{2,o,\perp} O'_{r_o} \text{ and } \beta_{2,o,\perp} = \left(I_{m-r_o} + O'_{r_o} O_{r_o}\right)^{-\frac{1}{2}}. \tag{A.43}$$

From Theorem 3.2 and $n^{\frac{1}{2}}\delta_{r,n} = o_p(1)$, we have

$$\begin{aligned} O_p(1) &= \left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \\ &= \left(\widehat{\Pi}_n - \Pi_o\right) \left[\sqrt{n}\alpha_o \left(\beta'_o \alpha_o\right)^{-1}, n\beta_{o,\perp} \left(\alpha'_{o,\perp}\beta_{o,\perp}\right)^{-1}\right], \end{aligned} \tag{A.44}$$

which implies that

$$\begin{aligned} O_p(1) &= \sqrt{n}\left(\widehat{\Pi}_n - \Pi_o\right)\alpha_o(\beta'_o\alpha_o)^{-1} \\ &= \sqrt{n}\left[\left(\widehat{\alpha}_n - \alpha_o\right)\widehat{\beta}'_n + \alpha_o\left(\widehat{\beta}_n - \beta_o\right)'\right]\alpha_o\left(\beta'_o\alpha_o\right)^{-1} \end{aligned} \tag{A.45}$$

and

$$n\widehat{\alpha}_n\left(\widehat{\beta}_n - \beta_o\right)'\beta_{o,\perp}\left(\alpha'_{o,\perp}\beta_{o,\perp}\right)^{-1} = O_p(1). \tag{A.46}$$

By the definitions of $\widehat{\beta}_n$ and $\beta_{o,\perp}$ and the result in (A.46), we get

$$O_p(1) = \beta'_o\widehat{\alpha}_n\left[n\left(\widehat{O}_n - O_{r_o}\right)\right]\beta_{2,o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1}$$

which implies that

$$\begin{aligned} n(\widehat{O}_n - O_{r_o}) &= \left[\beta'_o\alpha_o + o_p(1)\right]^{-1} O_p(1)\left(\alpha'_{o,\perp}\beta_{o,\perp}\right)\left(I_{m-r_o} + O'_{r_o}O_{r_o}\right)^{\frac{1}{2}} \\ &= O_p(1), \end{aligned} \tag{A.47}$$

where $\beta'_o\widehat{\alpha}_n = \beta'_o\alpha_o + o_p(1)$ is by the consistency of $\widehat{\alpha}_n$. By the definition of $\widehat{\beta}_n$, (A.47) means that $n(\widehat{\beta}_n - \beta_o) = O_p(1)$, which together with (A.45) implies that

$$\sqrt{n}\left(\widehat{\alpha}_n - \alpha_o\right) = \left[O_p(1) - \alpha_o\sqrt{n}\left(\widehat{\beta}_n - \beta_o\right)'\alpha_o\right]\left[\beta'_o\alpha_o + o_p(1)\right]^{-1} = O_p(1). \tag{A.48}$$

From Corollary 3.4, we can deduce that $\widehat{\alpha}_n$ and $\widehat{\beta}_n$ minimize the following criterion function w.p.a.1

$$V_n(\alpha,\beta) = \sum_{t=1}^{n}\left\|\Delta Y_t - \alpha\beta' Y_{t-1}\right\|^2 + n\sum_{k=1}^{r_o}\lambda_{r,k,n}||\Phi_{n,k}\left(\alpha\beta'\right)||. \tag{A.49}$$

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

$$\begin{aligned} \left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} &= \left[\widehat{\alpha}_n\left(\widehat{\beta}_n - \beta_o\right)' + \left(\widehat{\alpha}_n - \alpha_o\right)\beta'_o\right] Q^{-1} D_n^{-1} \\ &= \left[n^{-\frac{1}{2}}\widehat{\alpha}_n U^*_{3,n}\alpha_o\left(\beta'_o\alpha_o\right)^{-1} + U^*_{1,n},\ \widehat{\alpha}_n U^*_{3,n}\beta_{o,\perp}\left(\alpha'_{o,\perp}\beta_{o,\perp}\right)^{-1}\right]. \end{aligned}$$

Define

$$\Pi_n(U) = \left[n^{-\frac{1}{2}}\widehat{\alpha}_n U_3\alpha_o(\beta'_o\alpha_o)^{-1} + U_1, \widehat{\alpha}_n U_3\beta_{o,\perp}(\alpha'_{o,\perp}\beta_{o,\perp})^{-1}\right],$$

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

$$V_n(U) = \sum_{t=1}^{n} \left(\left\| \Delta Y_t - \Pi_o Y_{t-1} - \Pi_n(U) D_n Z_{t-1} \right\|^2 - \left\| \Delta Y_t - \Pi_o Y_{t-1} \right\|^2 \right)$$
$$+ n \sum_{k=1}^{r_o} \lambda_{r,k,n} \left[||\Phi_{n,k}(\Pi_n(U) D_n Q + \Pi_o)|| - ||\Phi_{n,k}(\Pi_o)|| \right].$$

For any compact set $K \subset R^{m \times r_o} \times R^{r_o \times (m - r_o)}$ and any $U \in K$, we have

$$\Pi_n(U) D_n Q = O_p(n^{-\frac{1}{2}}).$$

Hence, from the triangle inequality, we can deduce that for all $k \in \mathcal{S}_\phi$

$$n \left| \lambda_{r,k,n} \left[||\Phi_{n,k}(\Pi_n(U) D_n Q + \Pi_o)|| - ||\Phi_{n,k}(\Pi_o)|| \right] \right|$$
$$\leq n \lambda_{r,k,n} ||\Phi_{n,k}(\Pi_n(U) D_n Q)|| = O_p(n^{\frac{1}{2}} \lambda_{r,k,n}) = o_p(1), \tag{A.50}$$

uniformly over $U \in K$.

From (A.48),

$$\Pi_n(U) \to_p \left[U_1, \alpha_o U_3 \beta_{o,\perp} (\alpha_{o,\perp}' \beta_{o,\perp})^{-1} \right] \equiv \Pi_\infty(U) \tag{A.51}$$

uniformly over $U \in K$. By Lemma A.1 and (A.51), we deduce that

$$\sum_{t=1}^{n} \left(\left\| \Delta Y_t - \Pi_o Y_{t-1} - \Pi_n(U) D_n Z_{t-1} \right\|_E^2 - \left\| \Delta Y_t - \Pi_o Y_{t-1} \right\|_E^2 \right)$$
$$= vec[\Pi_n(U)]' \left(D_n \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_n \otimes I_m \right) vec[\Pi_n(U)]$$
$$- 2 vec[\Pi_n(U)]' vec \left(\sum_{t=1}^{n} u_t Z_{t-1}' D_n \right)$$
$$\to_d vec[\Pi_\infty(U)]' \left[\begin{pmatrix} \Sigma_{z_1 z_1} & 0 \\ 0 & \int B_{w_2} B_{w_2}' \end{pmatrix} \otimes I_m \right] vec[\Pi_\infty(U)]$$
$$- 2 vec[\Pi_\infty(U)]' vec\left[(V_{1,m}, V_{2,m}) \right] \equiv V(U) \tag{A.52}$$

uniformly over $U \in K$, where $V_{1,m} \equiv N(0, \Omega_u \otimes \Sigma_{z_1 z_1})$ and $V_{2,m} \equiv \left(\int B_{w_2} dB_u' \right)'$.

By definition $\Pi_\infty(U) = \left[U_1, \alpha_o U_2 \beta_{2,o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp} \right)^{-1} \right]$, thus

$$vec[\Pi_\infty(U)] = \left[vec(U_1)', vec\left(\alpha_o U_2 \beta_{2,o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp} \right)^{-1} \right)' \right]'$$

and

$$vec\left(\alpha_o U_2 \beta_{2,o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp} \right)^{-1} \right) = \left[\left(\beta_{o,\perp}' \alpha_{o,\perp} \right)^{-1} \beta_{2,o,\perp}' \otimes \alpha_o \right] vec(U_2).$$

Using above expression, we can rewrite $V(U)$ as

$$\begin{aligned} V(U) = {} & vec(U_1)' \left[\Sigma_{z_1 z_1} \otimes I_m\right] vec(U_1) + vec(U_2)' \\ & \times \left[\beta_{2,o,\perp} \left(\alpha'_{o,\perp}\beta_{o,\perp}\right)^{-1} \int B_{w_2} B'_{w_2} \left(\beta'_{o,\perp}\alpha_{o,\perp}\right)^{-1} \beta'_{2,o,\perp} \otimes \alpha'_o \alpha_o \right] vec(U_2) \\ & -2vec(U_1)' vec\left(V_{1,m}\right) \\ & -2vec(U_2)' vec\left[\alpha'_o V_{2,m} \left(\beta'_{o,\perp}\alpha_{o,\perp}\right)^{-1} \beta'_{2,o,\perp}\right]. \end{aligned} \tag{A.53}$$

The expression in (A.53) 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

$$U_1^* = B_{m,1} \text{ and } U_2^* = \left(\alpha'_o\alpha_o\right)^{-1}\alpha'_o B_{m,2}. \tag{A.54}$$

From (A.47) and (A.48), we can see that U_n^* is asymptotically tight. Invoking the Argmax Continuous Mapping Theorem (ACMT), we can deduce that

$$U_n^* = \left(U_{1,n}^*, U_{2,n}^*\right) \to_d \left[U_1^*,\ U_2^*\left(\alpha'_{o,\perp}\beta_{o,\perp}\right)\beta_{2,o,\perp}^{-1}\right]$$

which together with (A.51) and CMT implies that

$$\left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} \to_d \left(B_{m,1}\ \alpha_o \left(\alpha'_o\alpha_o\right)^{-1}\alpha'_o B_{m,2}\right).$$

This finishes the proof. ■

Proof of Corollary 3.6. The consistency, convergence rate, and super efficiency of $\widehat{\Pi}_{g,n}$ can be established using similar arguments in the proof of Theorems 3.1, 3.2, and 3.3.

Under the super efficiency of $\widehat{\Pi}_{g,n}$, the true rank r_o is imposed on $\widehat{\Pi}_{g,n}$ 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

$$\sum_{t=1}^{n} \left(\Delta Y_t - \alpha\beta' Y_{t-1}\right)' \widehat{\Omega}_{u,n}^{-1} \left(\Delta Y_t - \alpha\beta' Y_{t-1}\right) + n \sum_{k=1}^{r_o} \lambda_{r,k,n} ||\Phi_{n,k}(\alpha\beta')||. \tag{A.55}$$

Using the similar arguments in the proof of Theorem 3.5, 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 $U_{3,n}^* = n\left(\widehat{\beta}_{g,n} - \beta_o\right)' = \left[\mathbf{0}_{r_o}, n\left(\widehat{O}_{g,n} - O_o\right)\right] \equiv \left[\mathbf{0}_{r_o}, U_{2,n}^*\right]$, then

$$\begin{aligned} \left(\widehat{\Pi}_n - \Pi_o\right) Q^{-1} D_n^{-1} &= \left[\widehat{\alpha}_{g,n}\left(\widehat{\beta}_{g,n} - \beta_o\right)' + \left(\widehat{\alpha}_{g,n} - \alpha_o\right)\beta'_o\right] Q^{-1} D_n^{-1} \\ &= \left[n^{-\frac{1}{2}}\widehat{\alpha}_{g,n} U_{3,n}^* \alpha_o \left(\beta'_o\alpha_o\right)^{-1}\right. \\ &\quad \left. + U_{1,n}^*,\ \widehat{\alpha}_{g,n} U_{3,n}^* \beta_{o,\perp}\left(\alpha'_{o,\perp}\beta_{o,\perp}\right)^{-1}\right]. \end{aligned}$$

Define

$$\Pi_n(U) = \left[n^{-\frac{1}{2}} \widehat{\alpha}_{g,n} U_3 \alpha_o \left(\beta_o' \alpha_o\right)^{-1} + U_1, \widehat{\alpha}_{g,n} U_3 \beta_{o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp}\right)^{-1} \right],$$

then by definition, $U_n^* = \left(U_{1,n}^*, U_{2,n}^*\right)$ minimizes the following criterion function w.p.a.1

$$\begin{aligned} V_n(U) = &\sum_{t=1}^{n} \left[\left(u_t - \Pi_n(U) D_n Z_{t-1}\right)' \widehat{\Omega}_{u,n}^{-1} \left(u_t - \Pi_n(U) D_n Z_{t-1}\right) - u_t' \widehat{\Omega}_{u,n}^{-1} u_t \right] \\ &+ n \sum_{k=1}^{r_o} \lambda_{r,k,n} \left[||\Phi_{n,k}(\Pi_n(U) D_n Q + \Pi_o)|| - ||\Phi_{n,k}(\Pi_o)|| \right]. \end{aligned} \tag{A.56}$$

Following similar arguments in the proof of Theorem 3.5, we can deduce that for any $k \in \mathcal{S}_\phi$

$$n \left| \lambda_{r,k,n} \left[||\Phi_{n,k}(\Pi_n(U) D_n Q + \Pi_o)|| - ||\Phi_{n,k}(\Pi_o)|| \right] \right| = o_p(1), \tag{A.57}$$

and

$$\begin{aligned} &\sum_{t=1}^{n} \left(u_t - \Pi_n(U) D_n Z_{t-1}\right)' \widehat{\Omega}_{u,n}^{-1} \left(u_t - \Pi_n(U) D_n Z_{t-1}\right) - \sum_{t=1}^{n} u_t' \widehat{\Omega}_{u,n}^{-1} u_t \\ &\quad \to_d \upsilon ec(U_1)' \left(\Sigma_{z_1 z_1} \otimes \Omega_u^{-1}\right) \upsilon ec(U_1) + \upsilon ec(U_2)' \\ &\qquad \times \left[\beta_{2,o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp}\right)^{-1} \int B_{w_2} B_{w_2}' \left(\beta_{o,\perp}' \alpha_{o,\perp}\right)^{-1} \beta_{2,o,\perp}' \otimes \alpha_o' \Omega_u^{-1} \alpha_o \right] \upsilon ec(U_2) \\ &\qquad - 2 \upsilon ec(U_1)' \upsilon ec\left(\Omega_u^{-1} V_{1,m}\right) - 2 \upsilon ec(U_2)' \upsilon ec\left[\alpha_o' \Omega_u^{-1} V_{2,m} \left(\beta_{o,\perp}' \alpha_{o,\perp}\right)^{-1} \beta_{2,o,\perp}' \right] \\ &\quad \equiv V(U) \end{aligned} \tag{A.58}$$

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 $U_{g,1}^* = B_{1,m} \Sigma_{z_1 z_1}^{-1}$ and

$$U_{g,2}^* = \left(\alpha_o' \Omega_u^{-1} \alpha_o\right)^{-1} \left(\alpha_o' \Omega_u^{-1} V_{2,m}\right) \left(\int B_{w_2} B_{w_2}' \right)^{-1} \left(\alpha_{o,\perp}' \beta_{o,\perp}\right)^{-1} \beta_{2,o,\perp}^{-1}.$$

Invoking the ACMT, we obtain

$$\begin{aligned} \left(\widehat{\Pi}_{g,n} - \Pi_o\right) Q^{-1} D_n^{-1} &= \left[\widehat{\alpha}_{g,n} \left(\widehat{\beta}_{g,n} - \beta_o\right)' + \left(\widehat{\alpha}_{g,n} - \alpha_o\right) \beta_o' \right] Q^{-1} D_n^{-1} \\ &\to_d \left[V_{1,m} \Sigma_{z_1 z_1}^{-1},\ \alpha_o (\alpha_o' \Omega_u^{-1} \alpha_o)^{-1} \right. \\ &\qquad \left. \times \left(\alpha_o' \Omega_u^{-1} V_{2,m}\right) \left(\int B_{w_2} B_{w_2}' \right)^{-1} \right]. \end{aligned} \tag{A.59}$$

By the definition of w_1 and w_2, we can define $\Omega_{\widetilde{u}} = Q \Omega_u Q'$ such that

$$\Omega_{\widetilde{u}} = \begin{pmatrix} \Sigma_{w_1 w_1} & \Sigma_{w_1 w_2} \\ \Sigma_{w_2 w_1} & \Sigma_{w_2 w_2} \end{pmatrix} \text{ and } \Omega_{\widetilde{u}}^{-1} = \begin{pmatrix} \Omega_{\widetilde{u}}(11) & \Omega_{\widetilde{u}}(12) \\ \Omega_{\widetilde{u}}(21) & \Omega_{\widetilde{u}}(22) \end{pmatrix}.$$

Note that

$$\begin{aligned}\left(\alpha_o'\Omega_u^{-1}\alpha_o\right)^{-1}\alpha_o'\Omega_u^{-1} &= \left(\alpha_o'Q'\Omega_{\widetilde{u}}^{-1}Q\alpha_o\right)^{-1}\alpha_o'Q'\Omega_{\widetilde{u}}^{-1}Q \\ &= \left[\left(\alpha_o'\beta_o\right)\Omega_{\widetilde{u}}(11)\left(\beta_o'\alpha_o\right)\right]^{-1}\left[\left(\alpha_o'\beta_o\right),0\right]\Omega_{\widetilde{u}}^{-1}Q \\ &= \left(\beta_o'\alpha_o\right)^{-1}\Omega_{\widetilde{u}}^{-1}(11)\left[\Omega_{\widetilde{u}}(11)\beta_o' + \Omega_{\widetilde{u}}(12)\alpha_{o,\perp}'\right]. \end{aligned} \tag{A.60}$$

Under $\Omega_{\widetilde{u}}(12) = -\Omega_{\widetilde{u}}(11)\,\Sigma_{w_1w_2}\Sigma_{w_2w_2}^{-1}$,

$$\left(\alpha_o'\Omega_u^{-1}\alpha_o\right)^{-1}\alpha_o'\Omega_u^{-1} = \left(\beta_o'\alpha_o\right)^{-1}\left(\beta_o' - \Sigma_{w_1w_2}\Sigma_{w_2w_2}^{-1}\alpha_{o,\perp}'\right). \tag{A.61}$$

Now, using (A.59) and (A.61), we can deduce that

$$\left(\widehat{\Pi}_{g,n} - \Pi_o\right)Q^{-1}D_n^{-1} \to_d \left(B_{m,1}\,\alpha_o(\beta_o'\alpha_o)^{-1}\left(\int B_{w_2}dB_{u\cdot w_2}'\right)'\left(\int B_{w_2}B_{w_2}'\right)^{-1}\right).$$

This finishes the proof. ■

A.3. Proofs of Main Results in Section 4

The following lemma is useful in establishing the asymptotic properties of the shrinkage estimator with weakly dependent innovations.

LEMMA A.3. *Under Assumption 3.2 and 4.1, (a), (b), and (c) of Lemma A.1 are unchanged, while Lemma A.1.(d) becomes*

$$n^{-\frac{1}{2}}\sum_{t=1}^{n}\left[u_tZ_{1,t-1}' - \Sigma_{uz_1}(1)\right] \to_d N(0, V_{uz_1}), \tag{A.62}$$

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* $u_t \otimes Z_{1,t-1}$*; and Lemma A.1.(e) becomes*

$$n^{-1}\sum_{t=1}^{n}u_tZ_{2,t-1}' \to_d \left(\int B_{w_2}dB_u'\right)' + (\Gamma_{uu} - \Sigma_{uu})\,\alpha_{o\perp}. \tag{A.63}$$

The proof of Lemma A.3 is in the supplemental appendix of this paper. Let $P_1 = (P_{11}, P_{12})$ be the orthonormalized right eigenvector matrix of Π_1 and Λ_1 be a $r_1 \times r_1$ diagonal matrix of nonzero eigenvalues of Π_1, where P_{11} is an $m \times r_1$ matrix (of eigenvectors of nonzero eigenvalues) and P_{12} is an $m \times (m - r_1)$ matrix (of eigenvectors of zero eigenvalues). By the eigenvalue decomposition,

$$\Pi_1 = (P_{11}, P_{12})\begin{pmatrix}\Lambda_1 & 0\\ 0 & \mathbf{0}_{m-r_1}\end{pmatrix}\begin{pmatrix}Q_{11}\\ Q_{12}\end{pmatrix} = P_{11}\Lambda_1Q_{11}, \tag{A.64}$$

where $Q' = \left(Q_{11}', Q_{12}'\right)$ and $Q = P^{-1}$. By definition

$$\begin{pmatrix}Q_{11}\\ Q_{12}\end{pmatrix}(P_{11}, P_{12}) = \begin{pmatrix}Q_{11}P_{11} & Q_{11}P_{12}\\ Q_{12}P_{11} & Q_{12}P_{12}\end{pmatrix} = I_m, \tag{A.65}$$

which implies that $Q_{11}P_{11} = I_{r_1}$. From (A.64), without loss of generality, we can define $\widetilde{\alpha}_1 = P_{11}$ and $\widetilde{\beta}_1 = Q_{11}'\Lambda_1$. By (A.65), we deduce that

$$\widetilde{\beta}_1'\widetilde{\alpha}_1 = \Lambda_1 Q_{11}P_{11} = \Lambda_1 \text{ and } \widetilde{\alpha}_1'\widetilde{\beta}_1 = P_{11}'Q_{11}'\Lambda_1 = \Lambda_1$$

which imply that $\widetilde{\beta}_1'\widetilde{\alpha}_1$ and $\widetilde{\alpha}_1'\widetilde{\beta}_1$ are nonsingular $r_1 \times r_1$ matrix. Without loss of generality, we let $\widetilde{\alpha}_{1\perp} = P_{12}$ and $\widetilde{\beta}_{1\perp} = Q_{12}'$, then $\widetilde{\beta}_{1\perp}'\widetilde{\beta}_{1\perp} = I_{m-r_1}$ and under (A.65),

$$\widetilde{\beta}_{1\perp}'\widetilde{\alpha}_1 = Q_{12}P_{11} = 0,$$

which implies that $\widetilde{\beta}_\perp'\widetilde{\alpha}_1 = 0$ as $\widetilde{\beta}_{1\perp} = (\widetilde{\beta}_\perp, \beta_{o\perp})$.

Let $[\phi_1(\widehat{\Pi}_{1st}), \ldots, \phi_m(\widehat{\Pi}_{1st})]$ and $[\phi_1(\Pi_1), \ldots, \phi_m(\Pi_1)]$ be the ordered eigenvalues of $\widehat{\Pi}_{1st}$ and Π_1 respectively. For the ease of notation, we define

$$\mathcal{N}_1 \equiv \left[N(0, V_{uz_1}) + \Sigma_{uz_1}(1)\Sigma_{z_1z_1}^{-1}N(0, V_{z_1z_1})\right]\Sigma_{z_1z_1}^{-1}\beta_o',$$

where $N(0, V_{uz_1})$ is a random matrix defined in (A.62) and $N(0, V_{z_1z_1})$ denotes the matrix limit distribution of $\sqrt{n}(\widehat{S}_{11} - \Sigma_{z_1z_1})$. We also define

$$\mathcal{N}_2 \equiv \left[\int dB_u B_u' + (\Gamma_{uu} - \Sigma_{uu})\right]\alpha_{o\perp}\left(\int B_{w_2}B_{w_2}'\right)^{-1}\alpha_{o\perp}'.$$

The next lemma provides asymptotic properties of the OLS estimate and its eigenvalues when the data is weakly dependent.

LEMMA A.4. *Under Assumption 3.2 and 4.1, we have the following results:*

(a) *the OLS estimator $\widehat{\Pi}_{1st}$ satisfies*

$$(\widehat{\Pi}_{1st} - \Pi_1)Q^{-1}D_n^{-1} = O_p(1), \tag{A.66}$$

where Π_1 is defined in (4.2);

(b) *the eigenvalues of $\widehat{\Pi}_{1st}$ satisfy $\phi_k(\widehat{\Pi}_{1st}) \to_p \phi_k(\Pi_1)$ for $k = 1, \ldots, m$;*

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

$$n[\phi_{r_o+1}(\widehat{\Pi}_{1st}), \ldots, \phi_m(\widehat{\Pi}_{1st})] \to_d [\widetilde{\phi}_{r_o+1}', \ldots, \widetilde{\phi}_m'], \tag{A.67}$$

where $\widetilde{\phi}_j'$ $(j = r_o + 1, \ldots, m)$ are the ordered solutions of

$$\left|uI_{m-r_o} - \beta_{o\perp}'\left[\mathcal{N}_2 + \mathcal{N}_1\widetilde{\beta}_\perp(\widetilde{\beta}_\perp'\mathcal{N}_1\widetilde{\beta}_\perp)^{-1}\widetilde{\beta}_\perp'\mathcal{N}_2\right]\beta_{o\perp}\right| = 0; \tag{A.68}$$

(d) *$\widehat{\Pi}_{1st}$ has $r_o - r_1$ eigenvalues satisfying*

$$\sqrt{n}[\phi_{r_1+1}(\widehat{\Pi}_{1st}), \ldots, \phi_{r_o}(\widehat{\Pi}_{1st})] \to_d [\widetilde{\phi}_{r_1+1}', \ldots, \widetilde{\phi}_{r_o}'], \tag{A.69}$$

where $\widetilde{\phi}_j'$ $(j = r_1 + 1, \ldots, r_o)$ are the ordered solutions of

$$\left|uI_{r_o-r_1} - \widetilde{\beta}_\perp'\mathcal{N}_1\widetilde{\beta}_\perp\right| = 0. \tag{A.70}$$

The proof of Lemma A.4 is in the supplemental appendix of this paper. Recall that P_n is defined as the inverse of Q_n. We divide P_n and Q_n as $P_n = \left[P_{\widetilde{\alpha},n}, P_{\widetilde{\alpha}_\perp,n}\right]$ 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,

$$Q_{\widetilde{\alpha}_\perp,n} P_{\widetilde{\alpha}_\perp,n} = I_{m-r_1},\ Q_{\widetilde{\alpha},n} P_{\widetilde{\alpha}_\perp,n} = \mathbf{0}_{r_1\times(m-r_1)} \text{ and}$$
$$Q_{\widetilde{\alpha}_\perp,n} \widehat{\Pi}_{1st} = \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n}, \tag{A.71}$$

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 (A.71), we can define a useful estimator of Π_1 as

$$\widetilde{\Pi}_{n,f} = \widehat{\Pi}_{1st} - P_{\widetilde{\alpha}_\perp,n} \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n}. \tag{A.72}$$

By definition

$$Q_{\widetilde{\alpha},n} \widetilde{\Pi}_{n,f} = Q_{\widetilde{\alpha},n} \widehat{\Pi}_{1st} - Q_{\widetilde{\alpha},n} P_{\widetilde{\alpha}_\perp,n} \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} = \Lambda_{\widetilde{\alpha},n} Q_{\widetilde{\alpha},n}, \tag{A.73}$$

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

$$Q_{\widetilde{\alpha}_\perp,n} \widetilde{\Pi}_{n,f} = Q_{\widetilde{\alpha}_\perp,n} \widehat{\Pi}_{1st} - Q_{\widetilde{\alpha}_\perp,n} P_{\widetilde{\alpha}_\perp,n} \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} = \mathbf{0}_{(m-r_1)\times m}. \tag{A.74}$$

From Lemma A.4.(b), (A.73), and (A.74), 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 (A.71), we can write

$$\begin{aligned}\widetilde{\Pi}_{n,f} - \Pi_1 &= \left(\widehat{\Pi}_{1st} - \Pi_1\right) - P_{\widetilde{\alpha}_\perp,n} \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \\ &= \left(\widehat{\Pi}_{1st} - \Pi_1\right) - P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \left(\widehat{\Pi}_{1st} - \Pi_1\right) - P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \Pi_1,\end{aligned} \tag{A.75}$$

where Lemma A.4.(a),

$$\left(\widehat{\Pi}_{1st} - \Pi_1\right) Q^{-1} D_n^{-1} = O_p(1) \tag{A.76}$$

and by Lemmas A.4.(a), (c) and (d)

$$\begin{aligned}P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \Pi_1 Q^{-1} D_n^{-1} &= \sqrt{n} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \Pi_1 Q^{-1} \\ &= -\sqrt{n} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \left(\widehat{\Pi}_{1st} - \Pi_1\right) Q^{-1} \\ &\quad + \sqrt{n} P_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} \widehat{\Pi}_{1st} Q^{-1} \\ &= \sqrt{n} P_{\widetilde{\alpha}_\perp,n} \Lambda_{\widetilde{\alpha}_\perp,n} Q_{\widetilde{\alpha}_\perp,n} Q^{-1} + O_p(1) = O_p(1).\end{aligned} \tag{A.77}$$

Thus under (A.75), (A.76), and (A.77), we get

$$\left(\widetilde{\Pi}_{n,f} - \Pi_1\right) Q^{-1} D_n^{-1} = O_p(1). \tag{A.78}$$

Comparing (A.76) with (A.78), 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 4.1. First, when $r_o = 0$, then $\Pi_1 = \widetilde{\alpha}_o \beta_o' = 0 = \Pi_o$. Hence, the consistency of $\widehat{\Pi}_n$ follows by the similar arguments to those in the proof of Theorem 3.1. 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 3.1, by definition we have $V_n(\widehat{\Pi}_n) \leq V_n(\widetilde{\Pi}_{n,f})$, which implies

$$\begin{aligned}
&\left[vec\left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right)\right]' \left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right) \left[vec\left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right)\right] \\
&\quad + 2\left[vec\left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right)\right]' vec\left[\sum_{t=1}^{n} u_t Y_{t-1}' - (\Pi_1 - \Pi_o)\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'\right] \\
&\quad - 2\left[vec\left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right)\right]' \left(\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \otimes I_m\right) vec\left(\widetilde{\Pi}_{n,f} - \Pi_1\right) \\
&\leq n\left\{\sum_{k=1}^{m} \lambda_{r,k,n}\left[||\Phi_{n,k}\left(\widetilde{\Pi}_{n,f}\right)|| - ||\Phi_{n,k}\left(\widehat{\Pi}_n\right)||\right]\right\}. \qquad \textbf{(A.79)}
\end{aligned}$$

When $r_o = m$, Y_t is stationary and we have

$$\frac{1}{n}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}' \to_p \Sigma_{yy} = R(1)\Omega_u R(1)'. \qquad \textbf{(A.80)}$$

From the results in (A.79) and (A.80), we get w.p.a.1,

$$\mu_{n,\min}||\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}|| - ||\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}||(c_{1n} + c_{2n}) - d_n \leq 0, \qquad \textbf{(A.81)}$$

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,

$$\begin{aligned}
c_{1n} &= \left\| \frac{\sum_{t=1}^{n} u_t Y_{t-1}'}{n} - (\Pi_1 - \Pi_o)\frac{\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'}{n} \right\| \\
&\to_p \left\| \Sigma_{uy}(1) - \Sigma_{uy}(1)\Sigma_{yy}^{-1}\Sigma_{yy} \right\| = 0 \qquad \textbf{(A.82)}
\end{aligned}$$

by Lemma A.3 and the definition of Π_1, and

$$c_{2n} = m\left\|n^{-1}\sum_{t=1}^{n} Y_{t-1}Y_{t-1}'\right\| ||\widetilde{\Pi}_{n,f} - \Pi_1|| = o_p(1) \qquad \textbf{(A.83)}$$

by Lemma A.3 and (A.78), and

$$\begin{aligned}
d_n &= \sum_{k=1}^{m} \lambda_{r,k,n}\left[||\Phi_{n,k}(\widetilde{\Pi}_{n,f})|| - ||\Phi_{n,k}(\widehat{\Pi}_n)||\right] \\
&\leq \sum_{k=1}^{r_1} \lambda_{r,k,n}||\Phi_{n,k}(\widetilde{\Pi}_{n,f})|| = o_p(1) \qquad \textbf{(A.84)}
\end{aligned}$$

by Lemma A.4, (A.74) and $\lambda_{r,k,n} = o_p(1)$ for $k = 1, \ldots, r_1$. So the consistency of $\widehat{\Pi}_n$ follows directly from (A.78), the inequality in (A.81) and the triangle inequality.

When $0 < r_o < m$,

$$
\begin{aligned}
&vec\left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right)' \left(\sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m\right) vec\left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right) \\
&\quad = vec\left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right)' \left(B_n D_n \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_n B_n' \otimes I_m\right) vec\left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right) \\
&\quad \geq \mu_{n,\min} || \left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right) B_n ||^2, \qquad \textbf{(A.85)}
\end{aligned}
$$

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 A.3. Next, note that

$$
\begin{aligned}
&\left\{\sum_{t=1}^{n} u_t Z_{t-1}' - \left[(\Pi_1 - \Pi_o) Q^{-1}\right] \sum_{t=1}^{n} Z_{t-1} Z_{t-1}'\right\} D_n \\
&\quad = \begin{bmatrix} n^{-\frac{1}{2}} \sum_{t=1}^{n} Z_{1,t-1} u_t' \\ n^{-1} \sum_{t=1}^{n} Z_{2,t-1} u_t' \end{bmatrix}' - \begin{bmatrix} n^{-\frac{1}{2}} \sum_{t=1}^{n} Z_{1,t-1} Z_{1,t-1}' \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) \\ n^{-1} \sum_{t=1}^{n} Z_{2,t-1} Z_{1,t-1}' \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) \end{bmatrix}'. \qquad \textbf{(A.86)}
\end{aligned}
$$

From Lemma A.3, we can deduce that

$$
n^{-1} \sum_{t=1}^{n} Z_{2,t-1} u_t' = O_p(1) \text{ and } n^{-1} \sum_{t=1}^{n} Z_{2,t-1} Z_{1,t-1}' \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) = O_p(1). \qquad \textbf{(A.87)}
$$

Similarly, we get

$$
n^{-\frac{1}{2}} \sum_{t=1}^{n} \left[Z_{1,t-1} u_t' - \Sigma_{u z_1}'(1)\right] - n^{\frac{1}{2}} \left[S_{n,11} - \Sigma_{z_1 z_1}\right] \Sigma_{z_1 z_1}^{-1} \Sigma_{u z_1}'(1) = O_p(1). \qquad \textbf{(A.88)}
$$

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 (A.86)–(A.88) we can deduce that $e_{1n} = O_p(1)$. By the Cauchy–Schwarz inequality, we have

$$
\begin{aligned}
&\left| vec\left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right)' vec\left[\sum_{t=1}^{n} u_t Y_{t-1}' - (\Pi_1 - \Pi_o) \sum_{t=1}^{n} Y_{t-1} Y_{t-1}'\right] \right| \\
&\quad = \left| vec\left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right)' vec\left[\left\{\sum_{t=1}^{n} u_t Z_{t-1}' - (\Pi_1 - \Pi_o) Q^{-1} \sum_{t=1}^{n} Z_{t-1} Z_{t-1}'\right\} D_n B_n'\right] \right| \\
&\quad \leq || \left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right) B_n || e_{1n}. \qquad \textbf{(A.89)}
\end{aligned}
$$

Under Lemma A.3 and (A.78),

$$
\begin{aligned}
e_{2n} &\equiv \left| vec\left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right)' \left(\sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \otimes I_m\right) vec\left(\widetilde{\Pi}_{n,f} - \Pi_1\right) \right| \\
&= \left| vec\left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right)' \left(B_n D_n \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_n B_n' \otimes I_m\right) vec\left(\widetilde{\Pi}_{n,f} - \Pi_1\right) \right| \\
&\leq || \left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right) B_n || \times || \left(\widetilde{\Pi}_{n,f} - \Pi_1\right) B_n || \times || D_n \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_n || \\
&= O_p(1). \qquad \textbf{(A.90)}
\end{aligned}
$$

From results in (A.79), (A.89), and (A.90), we get w.p.a.1

$$\mu_{n,\min}||\left(\widehat{\Pi}_n-\widetilde{\Pi}_{n,f}\right)B_n||^2-2||\left(\widehat{\Pi}_n-\widetilde{\Pi}_{n,f}\right)B_n||^2\left(e_{1n}+e_{2n}\right)-d_n\leq 0, \tag{A.91}$$

where $d_n=o_p(1)$ by (A.84). Now, the consistency of $\widehat{\Pi}_n$ follows by (A.91) and the same arguments in Theorem 3.1. ■

Proof of Corollary 4.2. From Lemma A.4 and Corollary 4.1, we deduce that w.p.a.1

$$\begin{aligned}
&\sum_{k=1}^{m}\lambda_{r,k,n}\left[||\Phi_{n,k}\left(\widetilde{\Pi}_{n,f}\right)||-||\Phi_{n,k}\left(\widehat{\Pi}_n\right)||\right]\\
&\quad\leq\sum_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}\left[||\Phi_{n,k}\left(\widetilde{\Pi}_{n,f}\right)||-||\Phi_{n,k}\left(\widehat{\Pi}_n\right)||\right]\\
&\quad\leq d_{\widetilde{\mathcal{S}}_\phi}\max_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}||\widehat{\Pi}_n-\widetilde{\Pi}_{n,f}||.
\end{aligned}\tag{A.92}$$

Using (A.79) and (A.92), we have

$$\begin{aligned}
&\left[vec\left(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n\right)\right]'\left(\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\otimes I_m\right)\left[vec\left(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n\right)\right]\\
&\qquad+2\left[vec\left(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n\right)\right]'vec\left[\sum_{t=1}^{n}u_tY_{t-1}'-(\Pi_1-\Pi_o)\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\right]\\
&\qquad-2\left[vec\left(\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n\right)\right]'\left(\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\otimes I_m\right)vec\left(\widetilde{\Pi}_{n,f}-\Pi_1\right)\\
&\quad\leq c\max_{k\in\widetilde{\mathcal{S}}_\phi}\lambda_{r,k,n}||\widehat{\Pi}_n-\widetilde{\Pi}_{n,f}||,
\end{aligned}\tag{A.93}$$

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 3.2. 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 3.2, we get

$$\mu_{n,\min}||\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n||^2-c||\widetilde{\Pi}_{n,f}-\widehat{\Pi}_n||\left(c_{1n}+c_{2n}+\widetilde{\delta}_{r,n}\right)\leq 0, \tag{A.94}$$

where

$$\begin{aligned}
c_{1n}&=\left\|n^{-1}\sum_{t=1}^{n}u_tY_{t-1}'-n^{-1}(\Pi_1-\Pi_o)\sum_{t=1}^{n}Y_{t-1}Y_{t-1}'\right\|\\
&=n^{-\frac{1}{2}}\left\|n^{-\frac{1}{2}}\sum_{t=1}^{n}\left[u_tY_{t-1}'-\Sigma_{uy}(1)\right]-\Sigma_{uy}(1)\Sigma_{z_1z_1}^{-1}\left[n^{\frac{1}{2}}\left(\widehat{S}_{11}-\Sigma_{z_1z_1}\right)\right]\right\|\\
&=O_p\left(n^{-\frac{1}{2}}\right)
\end{aligned}\tag{A.95}$$

by Lemma A.3, and

$$c_{2n} = \left\| n^{-1} \sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \right\| \left\| \widetilde{\Pi}_{n,f} - \Pi_1 \right\| = O_p\left(n^{-\frac{1}{2}}\right) \tag{A.96}$$

by Lemma A.3 and A.78. From the results in (A.78), (A.94), (A.95), and (A.96), we deduce that

$$\widehat{\Pi}_n - \Pi_1 = O_p\left(n^{-\frac{1}{2}} + \widetilde{\delta}_{r,n}\right). \tag{A.97}$$

When $0 < r_o < m$, we can use (A.89) and (A.90) in the proof of Corollary 4.1 and (A.93) and to get w.p.a.1

$$\begin{aligned} &\mu_{n,\min} || \left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right) B_n ||^2 - 2 || \left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right) B_n || (e_{1,n} + e_{2,n}) \\ &\quad \leq cn\delta_n || \widetilde{\Pi}_{n,f} - \widehat{\Pi}_n ||, \end{aligned} \tag{A.98}$$

where $e_{1,n} = O_p(1)$ and $e_{2,n} = O_p(1)$ as illustrated in the proof of Corollary 4.1. By the Cauchy-Schwarz inequality,

$$|| \left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right) B_n B_n^{-1} || \leq cn^{-\frac{1}{2}} || \left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right) B_n ||. \tag{A.99}$$

Using (A.98) and (A.99), we obtain

$$\mu_{n,\min} || \left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right) B_n ||^2 - c || \left(\widetilde{\Pi}_{n,f} - \widehat{\Pi}_n\right) B_n || \left(e_{1,n} + e_{2,n} + n^{\frac{1}{2}} \widetilde{\delta}_{r,n}\right) \leq 0. \tag{A.100}$$

From (A.78) and the inequality in (A.100), we obtain

$$\left(\widehat{\Pi}_n - \Pi_1\right) B_n = \left(\widehat{\Pi}_n - \widetilde{\Pi}_{n,f}\right) B_n + \left(\widetilde{\Pi}_{n,f} - \Pi_1\right) B_n = O_p\left(1 + n^{\frac{1}{2}} \widetilde{\delta}_{r,n}\right),$$

which finishes the proof. ■

Proof of Corollary 4.3. Using similar arguments in the proof of Theorem 3.3, we can rewrite the LS shrinkage estimation problem as

$$\widehat{T}_n = \underset{T \in R^{m \times m}}{\arg\min} \sum_{t=1}^{n} \left\| \Delta Y_t - P_n T Y_{t-1} \right\|^2 + n \sum_{k=1}^{m} \lambda_{r,k,n} \left\| T(k) \right\|. \tag{A.101}$$

Result in (4.6) is equivalent to $\widehat{T}_n(k) = 0$ for any $k \in \{r_o + 1, \ldots, m\}$. 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,

$$\left\| \frac{1}{n} \sum_{t=1}^{n} \left(\Delta Y_t - P_n \widehat{T}_n Y_{t-1}\right)' P_n(k_o) Y_{t-1}' \right\| = \frac{\lambda_{r,k_o,n}}{2}. \tag{A.102}$$

The sample average in the left hand side of (A.102) can be rewritten as

$$\begin{aligned} \frac{\sum_{t=1}^{n} \left(\Delta Y_t - P_n \widehat{T}_n Y_{t-1}\right)' P_n(k_o) Y_{t-1}'}{n} &= \frac{P_n'(k_o) \sum_{t=1}^{n} \left[u_t - \left(\widehat{\Pi}_n - \Pi_o\right) Y_{t-1}\right] Y_{t-1}'}{n} \\ &= \frac{P_n'(k_o)}{n} \left[\sum_{t=1}^{n} \left[u_t - (\Pi_1 - \Pi_o) Y_{t-1}\right] Y_{t-1}' - \left(\widehat{\Pi}_n - \Pi_1\right) \sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \right]. \end{aligned} \tag{A.103}$$

From the results in (A.86), (A.87), and (A.88),

$$\frac{P_n'(k_o)\sum_{t=1}^n[u_t-(\Pi_1-\Pi_o)Y_{t-1}]Y_{t-1}'}{n}=O_p(1). \tag{A.104}$$

From Corollary 4.2 and Lemma A.3,

$$\frac{\left(\widehat{\Pi}_n-\Pi_1\right)\sum_{t=1}^n Y_{t-1}Y_{t-1}'}{n}=\frac{\left(\widehat{\Pi}_n-\Pi_1\right)B_nD_n\sum_{t=1}^n Z_{t-1}Z_{t-1}'Q'^{-1}}{n}=O_p(1). \tag{A.105}$$

Using the results in (A.103), (A.104), and (A.105), we deduce that

$$\left\|\frac{1}{n}\sum\nolimits_{t=1}^n\left(\Delta Y_t-P_n\widehat{T}_nY_{t-1}\right)'P_n(k_o)Y_{t-1}'\right\|=O_p(1). \tag{A.106}$$

While by the assumption on the tuning parameters, $\lambda_{r,k_o,n}\to_p\infty$, which together with the results in (A.102) and (A.106) implies that

$$\Pr\left(Q_n(k_o)\widehat{\Pi}_n=0\right)\to 1 \text{ as } n\to\infty.$$

As the above result holds for any k_o such that $r_o<k_o\le m$, 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 A.4.(c),

$$\begin{aligned}Q_{r_o-r_1,n}\widehat{\Pi}_nB_n&=Q_{r_o-r_1,n}\left(\widehat{\Pi}_n-\widehat{\Pi}_{1st}\right)B_n+Q_{r_o-r_1,n}\left(\widehat{\Pi}_{1st}-\Pi_1\right)B_n+Q_{r_o-r_1,n}\Pi_1B_n\\&=\sqrt{n}Q_{r_o-r_1,n}\Pi_1Q^{-1}+O_p(1)\\&=\sqrt{n}Q_{r_o-r_1,n}\left(\Pi_1-\widehat{\Pi}_{1st}\right)Q^{-1}+\sqrt{n}Q_{r_o-r_1,n}\widehat{\Pi}_{1st}Q^{-1}+O_p(1)\\&=\sqrt{n}\Lambda_{r_o-r_1,n}Q_{r_o-r_1,n}Q^{-1}+O_p(1)=O_p(1),\end{aligned} \tag{A.107}$$

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

$$\left(\widehat{T}_{\alpha,n}-T_{\alpha,n}\right)B_n=\begin{bmatrix}Q_{\widetilde{\alpha},n}\left(\widehat{\Pi}_n-\Pi_1\right)B_n\\Q_{r_o-r_1,n}\widehat{\Pi}_nB_n\end{bmatrix}=O_p(1) \tag{A.108}$$

where the last equality is by Corollary 4.2 and (A.107).

Proof of Corollary 4.4. Using the results of Corollary 4.3, we can rewrite the LS shrinkage estimation problem as

$$\widehat{T}_n=\underset{T\in R^{m\times m}}{\arg\min}\sum_{t=1}^n\left\|\Delta Y_t-P_nTY_{t-1}\right\|^2+n\sum\nolimits_{k=1}^{r_o}\lambda_{r,k,n}\left\|T(k)\right\| \tag{A.109}$$

with the constraint $T(k)=0$ for $k=r_o+1,\ldots,m$. Recall that $\widehat{T}_{\alpha,n}$ is the first r_o rows of $\widehat{T}_n$, then the problem in (A.109) can be rewritten as

$$\widehat{T}_{\alpha,n}=\underset{T_\alpha\in R^{r_o\times m}}{\arg\min}\sum_{t=1}^n\left\|\Delta Y_t-P_{r_o,n}T_\alpha Y_{t-1}\right\|^2+n\sum\nolimits_{k=1}^{r_o}\lambda_{r,k,n}\left\|T_\alpha(k)\right\|, \tag{A.110}$$

where $P_{r_o,n}$ is the first r_o columns of P_n.

Let $u_n^* = \left(\widehat{T}_{\alpha,n} - T_{\alpha,n}\right) B_n$ and note that the last $r_o - r_1$ rows of $T_{\alpha,n}$ are zeros. By definition, u_n^* is the minimizer of

$$\begin{aligned} V_n(U) &= \sum_{t=1}^{n} \left[\left\| \Delta Y_t - P_{r_o,n} \left(U B_n^{-1} + T_{\alpha,n} \right) Y_{t-1} \right\|^2 - \left\| \Delta Y_t - P_{r_o,n} T_{\alpha,n} Y_{t-1} \right\|^2 \right] \\ &\quad + n \sum\nolimits_{k=1}^{r_o} \lambda_{r,k,n} \left[\left\| U B_n^{-1} + T_{\alpha,n})(k) \right\| - \left\| T_{\alpha,n}(k) \right\| \right] \\ &= V_{1,n}(U) + n \sum\nolimits_{k=1}^{r_o} \lambda_{r,k,n} \left[\left\| U B_n^{-1} + T_{\alpha,n})(k) \right\| - \left\| T_{\alpha,n}(k) \right\| \right]. \end{aligned}$$

For any U in some compact subset of $R^{r_o \times m}$, $n^{\frac{1}{2}} U D_n Q = O(1)$. Thus $n^{\frac{1}{2}} \widetilde{\delta}_{r,n} = o_p(1)$ and Lemma A.4.d imply that

$$\begin{aligned} n\lambda_{r,k,n} \left| \left\| \left(U B_n^{-1} + T_{\alpha,n} \right)(k_o) \right\| - \left\| T_{\alpha,n}(k_o) \right\| \right| &\le n^{\frac{1}{2}} \lambda_{r,k,n} \left\| n^{\frac{1}{2}} \left(U B_n^{-1} \right)(k_o) \right\| \\ &= o_p(1) \end{aligned} \tag{A.111}$$

for $k_o = 1, \ldots, r_1$. On the other hand, $n^{\frac{1}{2}} \lambda_{r,k,n} = o_p(1)$ implies that

$$\begin{aligned} n\lambda_{r,k,n} \left| \left\| \left(U B_n^{-1} + T_{\alpha,n} \right)(k_o) \right\| - \left\| T_{\alpha,n}(k_o) \right\| \right| &\le n^{\frac{1}{2}} \lambda_{r,k,n} \left\| n^{\frac{1}{2}} \left(U B_n^{-1} \right)(k_o) \right\| \\ &= o_p(1) \end{aligned} \tag{A.112}$$

for any $k_o = 1, \ldots, r_o$. Moreover, we can rewrite $V_{1,n}(U)$ as

$$V_{1,n}(U) = A_{n,t}(U) - 2B_{n,t}(U),$$

where

$$A_{n,t}(U) \equiv vec(U)' \left(B_n^{-1} \sum\nolimits_{t=1}^{n} Y_{t-1} Y_{t-1}' B_n'^{-1} \otimes P_{r_o,n}' P_{r_o,n} \right) vec(U)$$

and

$$B_{n,t}(U) \equiv vec(U)' vec \left[P_{r_o,n}' \sum\nolimits_{t=1}^{n} \left(\Delta Y_t - P_{r_o,n} T_{\alpha,n} Y_{t-1} \right) Y_{t-1}' B_n'^{-1} \right].$$

It is clear that $V_{1,n}(U)$ is minimized at

$$\begin{aligned} U_n^* &= \left(P_{r_o,n}' P_{r_o,n} \right)^{-1} P_{r_o,n}' \sum_{t=1}^{n} \left(\Delta Y_t - P_{r_o,n} T_{\alpha,n} Y_{t-1} \right) Y_{t-1}' \left(\sum_{t=1}^{n} Y_{t-1} Y_{t-1}' \right)^{-1} B_n \\ &= \left[\left(P_{r_o,n}' P_{r_o,n} \right)^{-1} P_{r_o,n}' \widehat{\Pi}_{1st} - T_{\alpha,n} \right] B_n. \end{aligned}$$

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 normalized eigenvectors of the largest r_o and smallest $m - r_o$ eigenvalues of $\widehat{\Pi}_{1st}$ respectively. From Lemmas A.4.(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

$$U_n^* = \left[\left(P_{r_o,n}' P_{r_o,n} \right)^{-1} P_{r_o,n}' P_n Q_n \widehat{\Pi}_{1st} - T_{\alpha,n} \right] B_n = \left(Q_{r_o,n} \widehat{\Pi}_{1st} - T_{\alpha,n} \right) B_n$$

w.p.a.1. Results in (A.111) and (A.112) 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

$$Q_{r_o-r_1,n}\widehat{\Pi}_n B_n = U_{r_o-r_1,n}^* + o_p(1) = \Lambda_{r_o-r_1,n} Q_{r_o-r_1,n} B_n + o_p(1), \tag{A.113}$$

where $\Lambda_{r_o-r_1,n} \equiv diag\left[\phi_{r_1+1}(\widehat{\Pi}_{1st}), \ldots, \phi_{r_o}(\widehat{\Pi}_{1st})\right]$. From (A.113) and Lemma A.4, we obtain

$$n^{\frac{1}{2}} Q_{r_o-r_1,n}\widehat{\Pi}_n = n^{\frac{1}{2}} \Lambda_{r_o-r_1,n} Q_{r_o-r_1,n} + o_p(1) = \Lambda_{r_o-r_1}\left(\widetilde{\phi}'\right) Q_{r_o-r_1,o} + o_p(1), \tag{A.114}$$

where $\Lambda_{r_o-r_1}\left(\widetilde{\phi}'\right) \equiv diag\left(\widetilde{\phi}'_{r_1+1}, \ldots, \widetilde{\phi}'_{r_o}\right)$ is a nondegenerated 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 (A.114), we deduce that

$$\lim_{n\to\infty} \sup \Pr\left(n^{\frac{1}{2}} Q_{r_o-r_1,n}\widehat{\Pi}_n = 0\right) = 0$$

which finishes the proof. ■

A.4. Proofs of Main Results in Section 5

LEMMA A.5. *Under Assumption 3.1 and Assumption 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) $n^{-\frac{1}{2}}\sum_{t=1}^n u_t Z'_{3,t-1} \to_d N\left(0, \Omega_u \otimes \Sigma_{z_3 z_3}\right)$;

(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 A.5 follows by standard arguments like those in Lemma A.1 and its proof is omitted. We next 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

$$\left(\widehat{\Pi}_{1st}, \widehat{B}_{1st}\right) = \begin{pmatrix}\widehat{S}_{y_0 y_1} & \widehat{S}_{y_0 x_0}\end{pmatrix}\begin{pmatrix}\widehat{S}_{y_1 y_1} & \widehat{S}_{y_1 x_0} \\ \widehat{S}_{x_0 y_1} & \widehat{S}_{x_0 x_0}\end{pmatrix}^{-1}, \tag{A.115}$$

where

$$\widehat{S}_{y_0 y_1} = \frac{1}{n}\sum_{t=1}^n \Delta Y_t Y'_{t-1},\ \widehat{S}_{y_0 x_0} = \frac{1}{n}\sum_{t=1}^n \Delta Y_t \Delta X'_{t-1},$$

$$\widehat{S}_{y_1 y_1} = \frac{1}{n}\sum_{t=1}^n Y_{t-1} Y'_{t-1},\ \widehat{S}_{y_1 x_0} = \frac{1}{n}\sum_{t=1}^n Y_{t-1} \Delta X'_{t-1},$$

$$\widehat{S}_{x_0 y_1} = \widehat{S}'_{y_1 x_0} \text{ and } \widehat{S}_{x_0 x_0} = \frac{1}{n}\sum_{t=1}^n \Delta X_{t-1} \Delta X'_{t-1}. \tag{A.116}$$

Denote $Y_- = \left(Y_0, \ldots, Y_{n-1}\right)_{m\times n}$, $\Delta Y = (\Delta Y_1, \ldots, \Delta Y_n)_{m\times n}$ and

$$\widehat{M}_0 = I_n - n^{-1}\Delta X' \widehat{S}_{x_0x_0}^{-1} \Delta X,$$

where $\Delta X = \left(\Delta X_0, \ldots, \Delta X_{n-1}\right)_{mp\times n}$, then $\widehat{\Pi}_{1st}$ has the explicit partitioned regression representation

$$\widehat{\Pi}_{1st} = \left(\Delta Y \widehat{M}_0 Y_-'\right)\left(Y_- \widehat{M}_0 Y_-'\right)^{-1} = \Pi_o + \left(U \widehat{M}_0 Y_-'\right)\left(Y_- \widehat{M}_0 Y_-'\right)^{-1}, \tag{A.117}$$

where $U = (u_1, \ldots, u_n)_{m\times n}$. Recall that $\left[\phi_1\left(\widehat{\Pi}_{1st}\right), \ldots, \phi_m\left(\widehat{\Pi}_{1st}\right)\right]$ and $[\phi_1(\Pi_o), \ldots, \phi_m(\Pi_o)]$ are the ordered eigenvalues of $\widehat{\Pi}_{1st}$ and Π_o respectively, where $\phi_j(\Pi_o) = 0$ $(j = r_o + 1, \ldots, m)$. Let Q_n be the normalized left eigenvector matrix of $\widehat{\Pi}_{1st}$.

LEMMA A.6. *Suppose Assumption 3.1 and Assumption 5.1 hold.*

(a) *Recall* $D_{n,B} = diag(n^{-\frac{1}{2}} I_{r_o+mp}, n^{-1} I_{m-r_o})$, *then* $\left[\left(\widehat{\Pi}_{1st}, \widehat{B}_{1st}\right) - (\Pi_o, B_o)\right] Q_B^{-1} D_{n,B}^{-1}$ *has the following partitioned limit distribution*

$$\left[N\left(0, \Omega_u \otimes \Sigma_{z_3z_3}^{-1}\right), \int dB_u B_{w_2}' \left(\int B_{w_2} B_{w_2}'\right)^{-1}\right]; \tag{A.118}$$

(b) *The eigenvalues of* $\widehat{\Pi}_{1st}$ *satisfy* $\phi_k\left(\widehat{\Pi}_{1st}\right) \to_p \phi_k(\Pi_o)$ *for* $\forall k = 1, \ldots, m$;

(c) *For* $\forall k = r_o + 1, \ldots, m$, *the eigenvalues* $\phi_k\left(\widehat{\Pi}_{1st}\right)$ *of* $\widehat{\Pi}_{1st}$ *satisfy Lemma A.2.(c).*

The proof of Lemma A.6 is in the supplemental appendix of this paper. Lemma A.6 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.

Proof of Lemma 5.1. Let $\Theta = (\Pi, B)$ and

$$V_n(\Theta) = \sum_{t=1}^{n} \left\| \Delta Y_t - \Pi Y_{t-1} - \sum\nolimits_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2 + n \sum\nolimits_{j=1}^{p} \lambda_{b,j,n} \left\| B_j \right\| + n \sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi) \right\|.$$

Set $\widehat{\Theta}_n = \left(\widehat{\Pi}_n, \widehat{B}_n\right)$ and define an infeasible estimator $\widetilde{\Theta}_n = (\Pi_{n,f}, B_o)$, where $\Pi_{n,f}$ is defined in (A.6). Then by definition

$$\left(\widetilde{\Theta}_n - \Theta_o\right) Q_B^{-1} D_{n,B}^{-1} = \left(\Pi_{n,f} - \Pi_o, 0\right) Q_B^{-1} D_{n,B}^{-1} = O_p(1), \tag{A.119}$$

where the last equality is by (A.9).

By definition $V_n(\widehat{\Theta}_n) \le V_n(\widetilde{\Theta}_n)$, so that

$$\begin{aligned} &\left\{vec\left[\left(\widetilde{\Theta}_n - \widehat{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1}\right]\right\}' W_n \left\{vec\left[\left(\widetilde{\Theta}_n - \widehat{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1}\right]\right\} \\ &\quad + 2\left\{vec\left[\left(\widetilde{\Theta}_n - \widehat{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1}\right]\right\}' \left\{vec\left(D_{n,B} \sum\nolimits_{t=1}^{n} Z_{t-1} u_t'\right)\right\} \\ &\quad + 2\left\{vec\left[\left(\widetilde{\Theta}_n - \widehat{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1}\right]\right\}' W_n \left\{vec\left[\left(\Theta_o - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1}\right]\right\} \\ &\le \left(d_{1,n} + d_{2,n}\right) \end{aligned} \tag{A.120}$$

where

$$W_n = D_{n,B} \sum_{t=1}^{n} Z_{t-1} Z_{t-1}' D_{n,B} \otimes I_{m(p+1)},$$

$$d_{1,n} = n \sum\nolimits_{j \in \mathcal{S}_B} \lambda_{b,j,n} \left[\left\| B_{o,j} \right\| - ||\widehat{B}_{n,j}|| \right],$$

$$d_{2,n} = n \sum\nolimits_{k \in \mathcal{S}_\phi} \lambda_{r,k,n} \left[\left\| \Phi_{n,k}(\Pi_{n,f}) \right\| - ||\Phi_{n,k}(\widehat{\Pi}_n)|| \right].$$

Applying the Cauchy–Schwarz inequality to (A.120), we deduce that

$$\mu_n \left\| \left(\widehat{\Theta}_n - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\|^2 - \left\| \left(\widehat{\Theta}_n - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\| \left(c_{1,n} + c_{2,n}\right) \leq \left(d_{1,n} + d_{2,n}\right), \quad \textbf{(A.121)}$$

where μ_n denotes the smallest eigenvalue of W_n, which is bounded away from zero w.p.a.1,

$$c_{1,n} = \left\| D_{n,B} \sum\nolimits_{t=1}^{n} Z_{t-1} u_t' \right\| \text{ and } c_{2,n} = \|W_n\| \left\| \left(\Theta_o - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\|. \quad \textbf{(A.122)}$$

By the definition of the penalty function, Lemma A.6 and the Slutsky theorem, we find that

$$d_{1,n} \leq n \sum\nolimits_{j \in \mathcal{S}_B} \lambda_{b,j,n} \left\| B_{o,j} \right\| = O_p(n\delta_{b,n}) \text{ and} \quad \textbf{(A.123)}$$

$$d_{2,n} \leq n \sum\nolimits_{k \in \mathcal{S}_\phi} \lambda_{r,k,n} \left\| \Phi_{n,k}(\Pi_{n,f}) \right\| = O_p(n\delta_{r,n}). \quad \textbf{(A.124)}$$

Using Lemma A.5 and (A.119), we obtain

$$c_{1,n} = O_p(1) \text{ and } c_{2,n} = O_p(1). \quad \textbf{(A.125)}$$

From the inequality in (A.121), the results in (A.123), (A.124) and (A.125), we deduce that

$$\left\| \left(\widehat{\Theta}_n - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\| = O_P\left(1 + n^{1/2}\delta_{b,n}^{1/2} + n^{1/2}\delta_{r,n}^{1/2}\right).$$

which implies $||\widehat{\Theta}_n - \widetilde{\Theta}_n|| = O_P\left(n^{-1/2} + \delta_{b,n}^{1/2} + \delta_{r,n}^{1/2}\right) = o_p(1)$. 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 arguments in the proof of Theorem 3.2, we get

$$\left|d_{1,n}\right| \leq c n^{\frac{1}{2}} \delta_{b,n} \left\| \left(\widehat{\Theta}_n - \Theta_o\right) Q_B^{-1} D_{n,B}^{-1} \right\| \quad \textbf{(A.126)}$$

and

$$\left|d_{2,n}\right| \leq c n^{\frac{1}{2}} \delta_{r,n} \left\| \left(\widehat{\Theta}_n - \Theta_o\right) Q_B^{-1} D_{n,B}^{-1} \right\|. \quad \textbf{(A.127)}$$

Combining the results in (A.126)-(A.127), we get

$$\left|d_{1,n} + d_{2,n}\right| \leq c n^{\frac{1}{2}} \delta_n \left\| \left(\widehat{\Theta}_n - \Theta_o\right) Q_B^{-1} D_{n,B}^{-1} \right\|, \quad \textbf{(A.128)}$$

where $\delta_n = \delta_{b,n} + \delta_{r,n}$. From the inequality in (A.121) and the result in (A.128),

$$\mu_n \left\| \left(\widehat{\Theta}_n - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\|^2 - \left\| \left(\widehat{\Theta}_n - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\| \left(c_{1,n} + c_{2,n} + n^{\frac{1}{2}} \delta_n\right) \leq 0, \quad \textbf{(A.129)}$$

which together with (A.125) implies that $\left\| \left(\widehat{\Theta}_n - \widetilde{\Theta}_n\right) Q_B^{-1} D_{n,B}^{-1} \right\| = O_p\left(1 + n^{\frac{1}{2}}\delta_n\right)$. This finishes the proof. ■

Proof of Theorem 5.1. The first result can be proved using similar arguments in the proof of Theorem 3.3. Specifically, we rewrite the LS shrinkage estimation problem as

$$(\widehat{T}_n, \widehat{B}_n) = \underset{T, B_1, \ldots, B_p \in R^{m\times m}}{\arg\min} \sum_{t=1}^{n} \left\| \Delta Y_t - P_n T Y_{t-1} - \sum\nolimits_{j=1}^{p} B_j \Delta Y_{t-j} \right\|^2 + n \sum\nolimits_{k=1}^{m} \lambda_{r,k,n} \|T(k)\| + n \sum\nolimits_{j=1}^{p} \lambda_{b,j,n} \|B_j\|. \quad \textbf{(A.130)}$$

By definition, $\widehat{\Pi}_n = P_n \widehat{T}_n$ and $\widehat{T}_n = Q_n \widehat{\Pi}_n$ for all n. Results in (5.8) 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

$$\begin{cases} \sum\limits_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}\right)' P_n(k) Y_{t-1}' = \frac{n\lambda_{r,k,n} \widehat{T}_n(k)}{2\|\widehat{T}_n(k)\|} & \text{if } \widehat{T}_n(k) \neq 0 \\ \left\| n^{-1} \sum\limits_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}\right)' P_n(k) Y_{t-1}' \right\| < \frac{\lambda_{r,k,n}}{2} & \text{if } \widehat{T}_n(k) = 0 \end{cases},$$

for $k = 1, \ldots, m$. 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

$$\left\| n^{-1} \sum_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}\right)' P_n(k_o) Y_{t-1}' \right\| = \frac{\lambda_{r,k,n}}{2}. \quad \textbf{(A.131)}$$

The sample average in the left hand side of (A.36) can be rewritten as

$$\begin{aligned} &\frac{1}{n} \sum_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}\right)' P_n(k_o) Y_{t-1}' \\ &= \frac{1}{n} \sum_{t=1}^{n} \left[u_t - \left(\widehat{\Theta}_n - \Theta_o\right) Q_B^{-1} Z_{t-1}\right]' P_n(k_o) Y_{t-1}' \\ &= \frac{P_n'(k_o) \sum_{t=1}^{n} u_t Y_{t-1}'}{n} - \frac{P_n'(k_o)\left(\widehat{\Theta}_n - \Theta_o\right) Q_B^{-1} \sum_{t=1}^{n} Z_{t-1} Y_{t-1}'}{n} = O_p(1) \end{aligned} \quad \textbf{(A.132)}$$

where the last equality is by Lemmas A.5 and 5.1. However, under the assumptions on the tuning parameters $\lambda_{r,k_o,n} \to_p \infty$, which together with the results in (A.131) and (A.132) implies that

$$\Pr\left(Q_{\alpha,n}(k_o)\widehat{\Pi}_n = 0\right) \to 1 \text{ as } n \to \infty.$$

As the above result holds for any k_o such that $r_o < k_o \leq m$, this finishes the proof of (5.8).

We next show the second result. The LS shrinkage estimators of the transient dynamic matrices satisfy the following KKT optimality conditions:

$$\begin{cases} \sum\limits_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}\right) \Delta Y_{t-j}' = \frac{n\lambda_{b,j,n} \widehat{B}_{n,j}}{2\|\widehat{B}_{n,j}\|} & \text{if } \widehat{B}_{n,j} \neq 0 \\ \left\| \frac{1}{n} \sum\limits_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j}\right) \Delta Y_{t-j}' \right\| < \frac{\lambda_{b,j,n} \widehat{B}_{n,j}}{2\|\widehat{B}_{n,j}\|} & \text{if } \widehat{B}_{n,j} = 0 \end{cases},$$

for any $j = 1, \ldots, p$. 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,

$$\left\| n^{-\frac{1}{2}} \sum_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j} \right) \Delta Y_{t-j}' \right\| = \frac{n^{\frac{1}{2}} \lambda_{b,j,n}}{2}. \tag{A.133}$$

The sample average in the left hand side of (A.133) can be rewritten as

$$\begin{aligned} & n^{-\frac{1}{2}} \sum_{t=1}^{n} \left(\Delta Y_t - \widehat{\Pi}_n Y_{t-1} - \sum\nolimits_{j=1}^{p} \widehat{B}_{n,j} \Delta Y_{t-j} \right) \Delta Y_{t-j}' \\ & = n^{-\frac{1}{2}} \sum_{t=1}^{n} \left[u_t - \left(\widehat{\Theta}_n - \Theta_o \right) Q_B^{-1} Z_{t-1} \right] \Delta Y_{t-j}' \\ & = n^{-\frac{1}{2}} \sum_{t=1}^{n} u_t \Delta Y_{t-j}' - n^{-\frac{1}{2}} \left(\widehat{\Theta}_n - \Theta_o \right) Q_B^{-1} \sum_{t=1}^{n} Z_{t-1} \Delta Y_{t-j}' = O_p(1) \end{aligned} \tag{A.134}$$

where the last equality is by Lemmas A.5 and 5.1. However, by the assumptions on the tuning parameters $n^{\frac{1}{2}} \lambda_{b,j,n} \to \infty$, which together with (A.133) and (A.134) implies that

$$\Pr\left(\widehat{B}_{n,j} = \mathbf{0}_{m\times m} \right) \to 1 \text{ as } n \to \infty$$

for any $j \in \mathcal{S}_B^c$, which finishes the proof. ■

Proof of Theorem 5.2. Follow the similar arguments in the proof of Theorem 3.5, 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 5.1, we have

$$\left(n^{\frac{1}{2}} (\widehat{\Pi}_n - \Pi_o) \alpha_o (\beta_o' \alpha_o)^{-1} \; n^{\frac{1}{2}} (\widehat{B}_n - B_o) \; n (\widehat{\Pi}_n - \Pi_o) \beta_{o,\perp} (\alpha_{o,\perp}' \beta_{o,\perp})^{-1} \right) = O_p(1),$$

which implies that

$$n \left(\widehat{O}_n - O_o \right) = O_p(1), \tag{A.135}$$

$$n^{\frac{1}{2}} (\widehat{B}_n - B_o) = O_p(1), \tag{A.136}$$

$$n^{\frac{1}{2}} (\widehat{\alpha}_n - \alpha_o) = O_p(1), \tag{A.137}$$

where (A.135) and (A.137) hold with similar arguments in showing (A.47) and (A.48) in the proof of Theorem 3.5.

From the results of Theorem 5.1, we deduce that $\widehat{\alpha}_n$, $\widehat{\beta}_n$ and $\widehat{B}_{\mathcal{S}_B}$ minimize the following criterion function w.p.a.1,

$$\begin{aligned} V_n(\Theta_{\mathcal{S}}) = & \sum_{t=1}^{n} \left\| \Delta Y_t - \alpha \beta' Y_{t-1} - \sum_{j \in \mathcal{S}_B} B_j \Delta Y_{t-j} \right\|^2 \\ & + n \sum_{k \in \mathcal{S}_\phi} \lambda_{r,k,n} \left\| \Phi_{n,k}(\alpha \beta') \right\| + n \sum_{j \in \mathcal{S}_B} \lambda_{b,j,n} \left\| B_j \right\|. \end{aligned}$$

Define $U_{1,n}^* = \sqrt{n}(\widehat{\alpha}_n - \alpha_o)$ and $U_{2,n} = \left[\mathbf{0}_{r_o}, U_{2,n}^*\right]'$, 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

$$\begin{aligned}&\left[\left(\widehat{\Pi}_n - \Pi_o\right), \left(\widehat{B}_{\mathcal{S}_B} - B_{o,\mathcal{S}_B}\right)\right] Q_{\mathcal{S}}^{-1} D_{n,\mathcal{S}}^{-1} \\ &\quad = \left[n^{-\frac{1}{2}}\widehat{\alpha}_n U_{2,n}\alpha_o\left(\beta_o'\alpha_o\right)^{-1} + U_{1,n}^*, U_{3,n}^*,\ \widehat{\alpha}_n U_{2,n}\beta_{o,\perp}\left(\alpha_{o,\perp}'\beta_{o,\perp}\right)^{-1}\right].\end{aligned}$$

Denote

$$\Pi_n(U) = \left[n^{-\frac{1}{2}}\widehat{\alpha}_n U_2\alpha_o\left(\beta_o'\alpha_o\right)^{-1} + U_1, U_3, \widehat{\alpha}_n U_2\beta_{o,\perp}\left(\alpha_{o,\perp}'\beta_{o,\perp}\right)^{-1}\right],$$

then by definition, $U_n^* = \left(U_{1,n}^*, U_{2,n}^*, U_{3,n}^*\right)$ minimizes the following criterion function

$$\begin{aligned}V_n(U) = &\sum_{t=1}^{n}\left(\left\|u_t - \Pi_n(U)D_{n,\mathcal{S}}^{-1}Z_{\mathcal{S},t-1}\right\|^2 - \|u_t\|^2\right) \\ &+n\sum_{k\in\mathcal{S}_\phi}\lambda_{r,k,n}\left[\left\|\Phi_{n,k}\left[\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_1 + \Pi_o\right]\right\| - \left\|\Phi_{n,k}(\Pi_o)\right\|\right] \\ &+n\sum_{j\in\mathcal{S}_B}\lambda_{b,j,n}\left[\left\|\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_{j+1} + B_{o,j}\right\| - \left\|B_{o,j}\right\|\right].\end{aligned}$$

where $L_j = diag(A_{j,1}, \ldots, A_{j,d_{\mathcal{S}_B}+1})$ with $A_{j,j} = I_m$ and $A_{i,j} = 0$ for $i \neq j$ and $j = 1, \ldots, d_{\mathcal{S}_B+1}$.

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

$$\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}} = O_p\left(n^{-\frac{1}{2}}\right).$$

Hence using similar arguments in the proof of Theorem 3.5, we can deduce that

$$n\sum_{k\in\mathcal{S}_\phi}\lambda_{r,k,n}\left[\left\|\Phi_{n,k}\left[\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_1 + \Pi_o\right]\right\| - \left\|\Phi_{n,k}(\Pi_o)\right\|\right] = o_p(1) \tag{A.138}$$

and

$$n\sum_{j\in\mathcal{S}_B}\lambda_{b,j,n}\left[\left\|\Pi_n(U)D_{n,\mathcal{S}}^{-1}Q_{\mathcal{S}}L_{j+1} + B_{o,j}\right\| - \left\|B_{o,j}\right\|\right] = o_p(1) \tag{A.139}$$

uniformly over $U \in K$.

Next, note that

$$\Pi_n(U) \to_p \left[U_1, U_3, \alpha_o U_2\beta_{o,\perp}\left(\alpha_{o,\perp}'\beta_{o,\perp}\right)^{-1}\right] \equiv \Pi_\infty(U) \tag{A.140}$$

uniformly over $U \in K$. By Lemma A.5 and (A.140), we can deduce that

$$\begin{aligned}&\sum_{t=1}^{n}\left(\left\|u_t - \Pi_n(U)D_{n,\mathcal{S}}^{-1}Z_{\mathcal{S},t-1}\right\|^2 - \|u_t\|^2\right) \\ &\quad \to_d vec\left[\Pi_\infty(U)\right]'\left[\begin{pmatrix}\Sigma_{z_{3\mathcal{S}}z_{3\mathcal{S}}} & 0\\ 0 & \int B_{w_2}B_{w_2}'\end{pmatrix}\otimes I_m\right]vec\left[\Pi_\infty(U)\right] \\ &\qquad -2vec\left[\Pi_\infty(U)\right]'vec\left[\left(V_{3,m}, V_{2,m}\right)\right] \equiv V(U)\end{aligned} \tag{A.141}$$

uniformly over $U \in K$, where $V_{3,m} = N(0, \Omega_u \otimes \Sigma_{z_{3\mathcal{S}} z_{3\mathcal{S}}})$ and $V_{2,m} = \left(\int B_{w_2} dB_u'\right)'$.

Using similar arguments in the proof of Theorem 3.5, we can rewrite $V(U)$ as

$$\begin{aligned} V(U) = & vec(U_1, U_3)' \left(\Sigma_{z_{3\mathcal{S}} z_{3\mathcal{S}}} \otimes I_m\right) vec(U_1, U_3) \\ & + vec(U_2)' \left[\beta_{2,o,\perp} \left(\alpha_{o,\perp}' \beta_{o,\perp}\right)^{-1} \int B_{w_2} B_{w_2}' \left(\beta_{o,\perp}' \alpha_{o,\perp}\right)^{-1} \beta_{2,o,\perp}' \otimes \alpha_o' \alpha_o \right] vec(U_2) \\ & - 2vec(U_1, U_3)' vec\left(V_{3,m}\right) - 2vec(U_2)' vec\left[\alpha_o' V_{2,m} \left(\beta_{o,\perp}' \alpha_{o,\perp}\right)^{-1} \beta_{2,o,\perp}'\right]. \end{aligned} \tag{A.142}$$

The expression in (A.142) makes it clear that $V(U)$ is uniquely minimized at (U_1^*, U_2^*, U_3^*), where $\left(U_1^*, U_3^*\right) = V_{3,m} \Sigma_{z_{3\mathcal{S}} z_{3\mathcal{S}}}^{-1}$ and

$$U_2^* = \left(\alpha_o' \alpha_o\right)^{-1} \alpha_o' V_{2,m} \left(\int B_{w_2} B_{w_2}'\right)^{-1} \left(\alpha_{o,\perp}' \beta_{o,\perp}\right) \beta_{2,o,\perp}^{-1}. \tag{A.143}$$

From (A.135), (A.136), and (A.137), we see that U_n^* is asymptotically tight. Invoking the ACMT, we deduce that $U_n^* \to_d U^*$. The results in (5.11) follow by applying the CMT. ■

A.5. Proofs of Main Results in Section 6

Proof of Lemma 6.1.

(i) For any $k \in \mathcal{S}_\phi$, by Lemma A.2.(b), $||\phi_k(\widehat{\Pi}_{1st})||^\omega \to_p ||\phi_k(\Pi_o)||^\omega > 0$, which implies that

$$n^{\frac{1}{2}} \delta_{r,n} = \frac{n^{\frac{1}{2}} \lambda_{r,k,n}^*}{||\phi_k(\widehat{\Pi}_{1st})||^\omega} \to_p 0. \tag{A.144}$$

On the other hand, for any $k \in \mathcal{S}_\phi^c$, by Lemma A.2.(c), $||n\phi_k(\widehat{\Pi}_{1st})||^\omega \to_d ||\widetilde{\phi}_{o,k}||^\omega = O_p(1)$, which implies that

$$\lambda_{r,k,n} = \frac{n^\omega \lambda_{r,k,n}^*}{||n\phi_k(\widehat{\Pi}_{1st})||^\omega} \to_p \infty. \tag{A.145}$$

This finishes the proof of the first claim.

(ii) We only need to show $n^{\frac{1+\omega}{2}} \lambda_{r,k,n} = o_p(1)$ for any $k \in \{r_1 + 1, \ldots, r_o\}$, because the other two results can be proved using the same arguments showing (A.144)–(A.145). For any $k \in \{r_1 + 1, \ldots, r_o\}$, by Lemma A.4.(d), $||n^{\frac{1}{2}} \phi_k(\widehat{\Pi}_{1st})||^\omega \to_d ||\widetilde{\phi}_k'||^\omega$ which is a nondegenerated and continuous random variable. As a result, we can deduce that

$$n^{\frac{1}{2}} \lambda_{r,k,n} = \frac{n^{\frac{1+\omega}{2}} \lambda_{r,k,n}^*}{||n^{\frac{1}{2}} \phi_k(\widehat{\Pi}_{1st})||^\omega} = o_p(1), \tag{A.146}$$

which finishes the proof of the second claim.

(iii) The proof follows similar arguments to (i) and is therefore omitted. ■