High Dimensional Stochastic Regression with Latent Factors, Endogeneity and Nonlinearity

Jinyuan Chang* Bin Guo† Qiwei Yao‡

The University of Melbourne Peking University London School of Economics

August 12, 2014

Abstract

We consider a multivariate time series model which represents a high dimensional vector process as a sum of three terms: a linear regression of some observed regressors, a linear combination of some latent and serially correlated factors, and a vector white noise. We investigate the inference without imposing stationary conditions on the target multivariate time series, the regressors and the underlying factors. Furthermore we deal with the the endogeneity that there exist correlations between the observed regressors and the unobserved factors. We also consider the model with nonlinear regression term which can be approximated by a linear regression function with a large number of regressors. The convergence rates for the estimators of regression coefficients, the number of factors, factor loading space and factors are established under the settings when the dimension of time series and the number of regressors may both tend to infinity together with the sample size. The proposed method is illustrated with both simulated and real data examples.

Keywords: α-mixing, dimension reduction, instrument variables, nonstationarity, time series

JEL classification: C13; C32; C38.

*Department of Mathematics and Statistics, The University of Melbourne, Parkville, VIC, Australia 3010. Email: jinyuan.chang@unimelb.edu.au. This work was completed while the first author was a PhD student at Guanghua School of Management at Peking University.

†Guanghua School of Management, Peking University, Beijing, China 100871. Email: guobin1987@pku.edu.cn.

‡Corresponding Author: Department of Statistics, London School of Economics, London, WC2A 2AE, U.K. and Guanghua School of Management, Peking University, Beijing, China. Phone: +44 (0)20 7955 6767. Email: q.yao@lse.ac.uk. This work was partially supported by an EPSRC research grant.

1 Introduction

In this modern information age, the availability of large or vast time series data bring the opportunities with challenges to time series analysts. The demand of modelling and forecasting high-dimensional time series arises from various practical problems such as panel study of economic, social and natural (such as weather) phenomena, financial market analysis, communications engineering. On the other hand, modelling multiple time series even with moderately large dimensions is always a challenge. Although a substantial proportion of the methods and the theory for univariate autoregressive and moving average (ARMA) models has found the multivariate counterparts, the usefulness of unregularized multiple ARMA models suffers from the overparametrization and the lack of the identification (Lütkepohl, 2006). Various methods have been developed to reduce the number of parameters and to eliminate the non-identification issues. For example, Tiao and Tsay (1989) proposed to represent a multiple series in terms of several scalar component models based on canonical correlation analysis, Jakeman et al. (1980) adopted a two stage regression strategy based on instrumental variables to avoid using moving average explicitly. Another popular approach is to represent multiple time series in terms of a few factors defined in various ways; see, among others, Stock and Watson (2005), Bai and Ng (2002), Forni et al. (2005), Lam et al. (2011), and Lam and Yao (2012). Davis et al. (2012) proposed a vector autoregressive (VAR) model with sparse coefficient matrices based on partial spectral coherence. LASSO regularization has also been applied in VAR modelling; see, for example, Shojaie and Michailidis (2010) and Song and Bickel (2011).

This paper can be viewed as a further development of Lam et al. (2011) and Lam and Yao (2012) which express a high-dimensional vector time series as a linear transformation of a low-dimensional latent factor process plus a vector white noise. We extend their methodology and explore three new features. We only deal with the cases when the dimension is large in relation to the sample size. Hence all asymptotic theory is developed when both the sample size and the dimension of time series tend to infinity together.

Firstly, we add a regression term to the factor model. This is a useful addition as in many applications there exist some known factors which are among the driving forces for the dynamics of most the component series. For example, temperature is an important factor in forecasting household electricity consumptions. The price of a product plays a key role in its sales over different regions. The capital asset pricing model (CAPM) theory implies that the market index is a common factor for pricing different assets. When the regressor and the latent factor are uncorrelated, we estimate the regression coefficients first by the least squares method. We then estimate the number of factors and the factor loading space based on the residuals resulted from the regression estimation. We show that the latter is asymptotically adaptive to the unknown regression coefficients in the sense that the convergence rates for estimating the factor loading space and the factor process are the same as if the regression coefficients were known. We also consider the models with endogeneity in the sense that there exist correlations between the regressors and the latent

factors. We show that the factor loading space can still be identified and estimated consistently in the presence of the endogeneity. However relevant instrumental variables need to be employed if the 'original' regression coefficients have to be estimated consistently. The exploration in this direction has some overlap with Pesaran and Tosetti (2011), although the models, the inference methods and the asymptotic results in the two papers are different.

Our second contribution lies in the fact that we do not impose stationarity conditions on the regressors and the latent factor process throughout the paper. This enlarges the potential application substantially, as many important factors in practical problems (such as temperature, calendar effects) are not stationary. Different from the method of Pan and Yao (2008) which can also handle nonstationary factors but is computationally expensive, our approach is a direct extension of Lam et al. (2011) and Lam and Yao (2012) and, hence, is applicable to the cases when the dimensions of time series is in the order of thousands with an ordinary personal computer.

Finally, we focus on the factor models with a nonlinear regression term. By expressing the nonlinear regression function as a linear combination of some base functions, we turn the problem into the model with a large number of linear regressors. Now the asymptotic theory is established when the sample size, the dimension of time series and the number of regressors go to infinity together.

The rest of the paper is organized as follows. Section 2 deals with linear regression models with latent factors but without endogeneity. The models with the endogeneity are handled in Section 3. Section 4 investigates the models with nonlinear regression term. Simulation results are reported in Section 5. Illustration with some stock prices included in S&P500 is presented in Section 6. All the technical proofs are relegated to the Appendix.

2 Regression with latent factors

2.1 Models

Consider the regression model

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{1}$$

where $\mathbf{y}_t$ and $\mathbf{z}_t$ are, respectively, observable $p \times 1$ and $m \times 1$ time series, $\mathbf{x}_t$ is an $r \times 1$ latent factor process, $\boldsymbol{\varepsilon}_t \sim \mathrm{WN}(0, \boldsymbol{\Sigma}_\varepsilon)$ is a white noise with zero mean and covariance matrix $\boldsymbol{\Sigma}_\varepsilon$ and $\boldsymbol{\varepsilon}_t$ is uncorrelated with $(\mathbf{z}_t, \mathbf{x}_t)$, $\mathbf{D}$ is an unknown regression coefficient matrix, and $\mathbf{A}$ is an unknown factor loading matrix. The number of the latent factors r is an unknown (fixed) constant. With the observations $\{(\mathbf{y}_t, \mathbf{z}_t) : \ t = 1, \ldots, T\}$, the goal is to estimate $\mathbf{D}$, $\mathbf{A}$ and r, and to recover the factor process $\mathbf{x}_t$, when p is large in relation to the sample size T. As our inference will be based on the serial dependence of each and across $\mathbf{y}_t, \mathbf{z}_t$ and $\mathbf{x}_t$, we assume $E(\mathbf{z}_t) = \mathbf{0}$ and $E(\mathbf{x}_t) = \mathbf{0}$ for simplicity.

In this section, we consider the simple case when $\mathbf{z}_t$ and $\mathbf{x}_t$ are uncorrelated. This condition

ensures that the coefficient matrix $\mathbf{D}$ in (1) is identifiable. However the factor loading matrix $\mathbf{A}$ and the factor $\mathbf{x}_t$ are not uniquely determined by (1), as we may replace $(\mathbf{A}, \mathbf{x}_t)$ by $(\mathbf{AH}, \mathbf{H}^{-1}\mathbf{x}_t)$ for any invertible matrix $\mathbf{H}$. Nevertheless the linear space spanned by the columns of $\mathbf{A}$, denoted by $\mathcal{M}(\mathbf{A})$, is uniquely defined. $\mathcal{M}(\mathbf{A})$ is called the factor loading space. Hence there is no loss of the generality in assuming that $\mathbf{A}$ is a half orthogonal matrix in the sense that $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r$. In this paper, we always adhere with this assumption. Once we have specified a particular $\mathbf{A}$, $\mathbf{x}_t$ is uniquely defined accordingly. On the other hand, when $\mathrm{cov}(\mathbf{z}_t, \mathbf{x}_t) \neq \mathbf{0}$, the endogeneity makes $\mathbf{D}$ unidentifiable, which will be dealt with in Section 3 below.

2.2 Estimation

Formally the estimation for $\mathbf{D}$ may be treated as a standard least squares problem, since

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \boldsymbol{\eta}_t, \qquad \boldsymbol{\eta}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{2}$$

and $\mathrm{cov}(\mathbf{z}_t, \boldsymbol{\eta}_t) = \mathbf{0}$; see (1). Write $\mathbf{D} = (\mathbf{d}_1, \ldots, \mathbf{d}_p)^{\mathrm{T}}$. The least squares estimator for $\mathbf{D}$ can be expressed as

$$\widehat{\mathbf{D}} = (\widehat{\mathbf{d}}_1, \ldots, \widehat{\mathbf{d}}_p)^{\mathrm{T}}, \qquad \widehat{\mathbf{d}}_i = \left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} y_{i,t}\,\mathbf{z}_t\right), \tag{3}$$

where $y_{i,t}$ is the ith component of $\mathbf{y}_t$.

The estimation for $\mathcal{M}(\mathbf{A})$ is based on the residuals $\widehat{\boldsymbol{\eta}}_t = \mathbf{y}_t - \widehat{\mathbf{D}}\mathbf{z}_t$, using the same idea as Lam et al. (2011) and Lam and Yao (2012), though we do not assume that the processes concerned are stationary. To this end, we introduce some notation first. Let

$$\boldsymbol{\Sigma}_x(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} \mathrm{cov}(\mathbf{x}_{t+k}, \mathbf{x}_t), \qquad \boldsymbol{\Sigma}_{x\varepsilon}(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} \mathrm{cov}(\mathbf{x}_{t+k}, \boldsymbol{\varepsilon}_t),$$

$$\boldsymbol{\Sigma}_\eta(k) = \frac{1}{T-k}\sum_{t=1}^{T-k} \mathrm{cov}(\boldsymbol{\eta}_{t+k}, \boldsymbol{\eta}_t).$$

When, for example, $\mathbf{x}_t$ is stationary, $\boldsymbol{\Sigma}_x(k)$ is the autocovariance matrix of $\mathbf{x}_t$ at lag k. It follows from the second equation in (2) that for any $k \neq 0$,

$$\boldsymbol{\Sigma}_\eta(k) = \mathbf{A}\boldsymbol{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}} + \mathbf{A}\boldsymbol{\Sigma}_{x\varepsilon}(k). \tag{4}$$

For a prescribed fixed positive integer $\bar{k}$, define

$$\mathbf{M} = \sum_{k=1}^{\bar{k}} \boldsymbol{\Sigma}_\eta(k)\boldsymbol{\Sigma}_\eta(k)^{\mathrm{T}}. \tag{5}$$

We assume $\mathrm{rank}(\mathbf{M}) = r$. This is reasonable as it effectively assumes that the latent factor process $\mathbf{x}_t$ is genuinely r-dimensional. Since $\mathbf{M}$ is implicitly sandwiched by $\mathbf{A}$ and $\mathbf{A}^{\mathrm{T}}$, $\mathbf{Mb} = \mathbf{0}$ for any $\mathbf{b} \perp \mathcal{M}(\mathbf{A})$. Thus we may take the eigenvectors of $\mathbf{M}$ corresponding to non-zero eigenvalues as

the columns of $\mathbf{A}$, as the choice of $\mathbf{A}$ is almost arbitrary as long as $\mathcal{M}(\mathbf{A})$ does not change. Let $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$, where $\mathbf{a}_1, \ldots, \mathbf{a}_r$ be the r orthonormal eigenvectors of $\mathbf{M}$ corresponding to the r largest eigenvalues $\lambda_1 \geq \cdots \geq \lambda_r > 0$. Then $\mathbf{A}$ is a half orthogonal matrix in the sense that $\mathbf{A}^{\mathrm{T}}\mathbf{A} = \mathbf{I}_r$. In the sequel, we always use $\mathbf{A}$ defined this way. When the r non-zero eigenvalues of $\mathbf{M}$ are distinct, $\mathbf{A}$ is unique if we ignore the trivial replacements of $\mathbf{a}_j$ by $-\mathbf{a}_j$.

Let $\widehat{\boldsymbol{\eta}}_t = \mathbf{y}_t - \widehat{\mathbf{D}}\mathbf{z}_t$ and

$$\widehat{\boldsymbol{\Sigma}}_\eta(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(\widehat{\boldsymbol{\eta}}_{t+k} - \bar{\boldsymbol{\eta}})(\widehat{\boldsymbol{\eta}}_t - \bar{\boldsymbol{\eta}})^{\mathrm{T}}, \qquad \bar{\boldsymbol{\eta}} = \frac{1}{T}\sum_{t=1}^{T}\widehat{\boldsymbol{\eta}}_t.$$

The above discussion leads to a natural estimator of $\mathbf{A}$ denoted by $\widehat{\mathbf{A}} \equiv (\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_r)$. Here $\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_r$ are the orthonormal eigenvectors of $\widehat{\mathbf{M}}$ corresponding to the r largest eigenvalues $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_r$, where

$$\widehat{\mathbf{M}} = \sum_{k=1}^{\bar{k}} \widehat{\boldsymbol{\Sigma}}_\eta(k)\widehat{\boldsymbol{\Sigma}}_\eta(k)^{\mathrm{T}}. \tag{6}$$

Since $\widehat{\mathbf{A}}$ is a half orthogonal matrix, we may extract the factor process by $\widehat{\mathbf{x}}_t = \widehat{\mathbf{A}}^{\mathrm{T}}(\mathbf{y}_t - \widehat{\mathbf{D}}\mathbf{z}_t)$; see (2).

All the arguments above are based on a known r which is actually unknown in practice. The determination of r is a key step in our inference. In practice we may estimate it by the ratio estimator

$$\widehat{r} = \arg\min\left\{\frac{\widehat{\lambda}_{j+1}}{\widehat{\lambda}_j} : 1 \leq j \leq R\right\}, \tag{7}$$

where $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ are the eigenvalues of $\widehat{\mathbf{M}}$, and R is a constant which may be taken as $R = p/2$; see Lam and Yao (2012) for further discussion on this estimation method.

2.3 Asymptotic properties

We present the asymptotic theory for the estimation methods described in Section 2.2 above when $T,\ p \to \infty$ while r is fixed. We also assume m fixed now; see Section 4 below for the results when $m \to \infty$ as well. We do not impose stationarity conditions on $\mathbf{y}_t, \mathbf{z}_t$ and $\mathbf{x}_t$. Instead we assume that they are mixing processes; see Condition 2.1 below. Hence our results in the special case when $\mathbf{z}_t \equiv \mathbf{0}$ extend those in Lam et al. (2011) and Lam and Yao (2012) to nonstationary cases. Pan and Yao (2008) dealt with a different method for nonstationary factor models.

We introduce some notation first. For any matrix $\mathbf{H}$, we denote by $\|\mathbf{H}\|_F = \{\mathrm{tr}(\mathbf{H}^{\mathrm{T}}\mathbf{H})\}^{1/2}$ the Frobenius norm of $\mathbf{H}$, and by $\|\mathbf{H}\|_2 = \{\lambda_{\max}(\mathbf{H}^{\mathrm{T}}\mathbf{H})\}^{1/2}$ the L_2-norm, where $\mathrm{tr}(\cdot)$ and $\lambda_{\max}(\cdot)$ denote, respectively, the trace and the maximum eigenvalue of a square matrix. We also denote by $\|\mathbf{H}\|_{\min}$ the square-root of the minimum nonzero eigenvalue of $\mathbf{H}^{\mathrm{T}}\mathbf{H}$. Note that when $\mathbf{H} = \mathbf{h}$ is a vector, $\|\mathbf{h}\|_F = \|\mathbf{h}\|_2 = \|\mathbf{h}\|_{\min} = (\mathbf{h}^{\mathrm{T}}\mathbf{h})^{1/2}$, i.e. the conventional Euclidean norm for vector $\mathbf{h}$.

Condition 2.1. The process $\{(\mathbf{y}_t, \mathbf{z}_t, \mathbf{x}_t)\}$ is α-mixing with the mixing coefficients satisfying the condition $\sum_{k=1}^{\infty} \alpha(k)^{1-2/\gamma} < \infty$ for some $\gamma > 2$, where

$$\alpha(k) = \sup_i \sup_{A \in \mathcal{F}_{-\infty}^i,\, B \in \mathcal{F}_{i+k}^{\infty}} \big|P(A \cap B) - P(A)P(B)\big|,$$

and $\mathcal{F}_i^j$ is the σ-field generated by $\{(\mathbf{y}_t, \mathbf{z}_t, \mathbf{x}_t) : \ i \le t \le j\}$.

Condition 2.2. For any $i = 1, \ldots, m$, $j = 1, \ldots, p$ and t, $E(|z_{i,t}|^{2\gamma}) \le C_1$, $E(|\zeta_{j,t}|^{2\gamma}) \le C_1$ and $E(|\varepsilon_{j,t}|^{2\gamma}) \le C_1$, where $C_1 > 0$ is a constant, γ is given in Condition 2.1, and $z_{i,t}$ is the ith element of $\mathbf{z}_t$, $\zeta_{j,t}$ and $\varepsilon_{j,t}$ are the jth element of, respectively, $\mathbf{A}\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$.

Condition 2.3. There exists a constant $C_2 > 0$ such that $\lambda_{\min}\{E(\mathbf{z}_t \mathbf{z}_t^{\mathrm{T}})\} > C_2$ for all t.

Condition $E(|\zeta_{j,t}|^{2\gamma}) \le C_1$ in Condition 2.2 can be guaranteed by some suitable conditions on each $x_{i,t}$, as $\mathbf{A}$ is a half orthogonal matrix. For example, it holds if $\max_{i,t} E(|x_{i,t}|^{2\gamma}) < \infty$. Proposition 2.1 below establishes the convergence rate of the estimator for the $p \times m$ coefficient matrix $\mathbf{D}$. Since $p \to \infty$ together with the sample size T, the convergence rate depends on p. Especially when $p/T \to 0$, the least squares estimator $\widehat{\mathbf{D}}$ is a consistent estimator for $\mathbf{D}$. This condition can be relaxed if we impose some sparse condition on $\mathbf{D}$, and then apply appropriate thresholding on $\widehat{\mathbf{D}}$. We do not pursue this further here. When p is fixed, the convergence rate is $T^{1/2}$ which is the optimal rate for the regression with the dimension fixed.

Proposition 2.1. *Let Conditions 2.1-2.3 hold. As $T \to \infty$ and $p \to \infty$, it holds that*

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_F = O_p\big(p^{1/2} T^{-1/2}\big).$$

To state the results for estimating factor loadings, we introduce more conditions.

Condition 2.4. There exist positive constants C_i $(i = 3, 4)$ and $\delta \in [0, 1]$ such that $C_3 p^{1-\delta} \le \|\boldsymbol{\Sigma}_x(k)\|_{\min} \le \|\boldsymbol{\Sigma}_x(k)\|_2 \le C_4 p^{1-\delta}$ for all $k = 1, \ldots, \bar{k}$.

Condition 2.5. Matrix $\mathbf{M}$ admits r distinct positive eigenvalues $\lambda_1 > \cdots > \lambda_r > 0$.

The constant δ in Condition 2.4 controls the strength of the factors. When $\delta = 0$, the factors are strong. When $\delta > 0$, the factors are weak. In fact the value of δ reflects the sparse level of the factor loading matrix $\mathbf{A}$, and a certain degree of sparsity is present when $\delta > 0$. Therefore not all components of $\mathbf{y}_t - \mathbf{D}\mathbf{z}_t$ carry the information for all factor components. This causes difficulties in recovering the factor process. This argument will be verified in Theorem 2.2. See also Remark 1 in Lam and Yao (2012). Condition 2.5 implies that $\mathbf{A}$ defined as in Section 2.2 above is unique. This simplifies the presentation significantly, as Theorem 2.1 below can present the convergence rates of the estimator for $\mathbf{A}$ directly. Without Condition 2.5, the same convergence rates can be obtained for the estimation of the linear space $\mathcal{M}(\widehat{\mathbf{A}})$; see (9) below. Let

$$\kappa_1 = \min_{1 \le k \le \bar{k}} \|\boldsymbol{\Sigma}_{x\varepsilon}(k)\|_{\min} \ \text{ and } \ \kappa_2 = \max_{1 \le k \le \bar{k}} \|\boldsymbol{\Sigma}_{x\varepsilon}(k)\|_2.$$

Note that both κ_1 and κ_2 may diverge as $p \to \infty$.

Theorem 2.1. *Let Conditions 2.1-2.5 hold. Suppose that r is known and fixed, then*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 = \begin{cases} O_p(p^{\delta}T^{-1/2}), & \text{if } \kappa_2 = o(p^{1-\delta}) \text{ and } p^{2\delta}T^{-1} = o(1); \\ O_p(\kappa_1^{-2}\kappa_2 pT^{-1/2}), & \text{if } p^{1-\delta} = o(\kappa_1) \text{ and } \kappa_1^{-2}\kappa_2 pT^{-1/2} = o(1). \end{cases}$$

The convergence rates in Theorem 2.1 above are exactly the same as Theorem 1 of Lam et al. (2011) which deals with a pure factor model, i.e. model (2) with $\mathbf{z}_t \equiv \mathbf{0}$. In this sense, the estimator $\widehat{\mathbf{A}}$ is asymptotically adaptive to unknown $\mathbf{D}$.

Theorem 2.2. *Let Conditions 2.1-2.5 hold, and r be known and fixed. If $\|\mathbf{\Sigma}_\varepsilon\|_2$ is bounded as $p \to \infty$, then*

$$p^{-1/2}\|\widehat{\mathbf{A}}\widehat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_t\|_2 = O_p(\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 + p^{-1/2} + T^{-1/2}).$$

Theorem 2.2 deals with the convergence of the extracted factor term. Combining it with Theorem 2.1, we obtain

$$\begin{aligned} & p^{-1/2}\|\widehat{\mathbf{A}}\widehat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_t\|_2 \\ = & \begin{cases} O_p(p^{\delta}T^{-1/2} + p^{-1/2}), & \text{if } \kappa_2 = o(p^{1-\delta}) \text{ and } p^{2\delta}T^{-1} = o(1); \\ O_p(\kappa_1^{-2}\kappa_2 pT^{-1/2} + p^{-1/2} + T^{-1/2}), & \text{if } p^{1-\delta} = o(\kappa_1) \text{ and } \kappa_1^{-2}\kappa_2 pT^{-1/2} = o(1). \end{cases} \end{aligned}$$

Thus when all the factors are strong (i.e. $\delta = 0$) and $\kappa_2 = o(p)$, it holds that $p^{-1/2}\|\widehat{\mathbf{A}}\widehat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_t\|_2 = O_p(p^{-1/2} + T^{-1/2})$, which is the optimal convergence rate specified in Theorem 3 of Bai (2003).

In general the choice of $\mathbf{A}$ in model (1) is not unique, we consider the error in estimating $\mathcal{M}(\mathbf{A})$ instead of a particular $\mathbf{A}$, as $\mathcal{M}(\mathbf{A})$ is uniquely defined by (1) and does not vary with different choices of $\mathbf{A}$. To this end, we adopt the discrepancy measure used by Pan and Yao (2008): for two $p \times r$ half orthogonal matrices $\mathbf{H_1}$ and $\mathbf{H}_2$ satisfying the condition $\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_1 = \mathbf{H}_2^{\mathrm{T}}\mathbf{H}_2 = \mathbf{I}_r$, the difference between the two linear spaces $\mathcal{M}(\mathbf{H}_1)$ and $\mathcal{M}(\mathbf{H}_2)$ is measured by

$$D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2)) = \sqrt{1 - \frac{1}{r}\mathrm{tr}(\mathbf{H}_1\mathbf{H}_1^{\mathrm{T}}\mathbf{H}_2\mathbf{H}_2^{\mathrm{T}})}. \tag{8}$$

In fact $D(\mathcal{M}(\mathbf{H}_1), \mathcal{M}(\mathbf{H}_2))$ always takes values between 0 and 1. It is equal to 0 if and only if $\mathcal{M}(\mathbf{H}_1) = \mathcal{M}(\mathbf{H}_2)$, and to 1 if and only if $\mathcal{M}(\mathbf{H}_1) \perp \mathcal{M}(\mathbf{H}_2)$.

Theorem 2.3. *Let Conditions 2.4-2.5 hold. Suppose that r is known and fixed, then*

$$\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2 \asymp \|(\widehat{\mathbf{A}} - \mathbf{A})^{\mathrm{T}}(\widehat{\mathbf{A}} - \mathbf{A}) - \mathbf{A}^{\mathrm{T}}(\widehat{\mathbf{A}} - \mathbf{A})(\widehat{\mathbf{A}} - \mathbf{A})^{\mathrm{T}}\mathbf{A}\|_2.$$

This theorem establishes the link between $D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))$ and $\widehat{\mathbf{A}} - \mathbf{A}$ when r is known. Obviously, the RHS of the above expression can be bounded by $2\|\widehat{\mathbf{A}} - \mathbf{A}\|_2^2$. This implies that $D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A})) = O_p(\|\widehat{\mathbf{A}} - \mathbf{A}\|_2)$. In fact, the convergence of $D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))$ does not depend on Condition 2.5. Even when $\mathbf{M}$ admits multiple non-zero eigenvalues, and, therefore, $\mathbf{A}$ is not

uniquely defined, it can be shown based on the similar arguments as for Theorem 1 in Chang et al. (2014) that

$$D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A})) = \begin{cases} O_p(p^{\delta}T^{-1/2}), & \text{if } \kappa_2 = o(p^{1-\delta}) \text{ and } p^{2\delta}T^{-1} = o(1); \\ O_p(\kappa_1^{-2}\kappa_2 pT^{-1/2}), & \text{if } p^{1-\delta} = o(\kappa_1) \text{ and } \kappa_1^{-2}\kappa_2 pT^{-1/2} = o(1), \end{cases} \tag{9}$$

which is the same as that followed by Theorem 2.3 when Condition 2.5 holds.

Theorems 2.1-2.3 above present the asymptotic properties when the number of factors r is assumed to be known. However, in practice we need to estimate r as well. Lam and Yao (2012) showed that for the ratio estimator $\widehat{r}$ defined in (7), $P(\widehat{r} \geq r) \to 1$. In spite of favorable finite sample evidences reported in Lam and Yao (2012), it remains as a unsolved challenge to establish the consistency $\widehat{r}$. Following the idea of Xia et al. (2013), we adjust the ratio estimator as follows

$$\widetilde{r} = \arg\min\left\{\frac{\widehat{\lambda}_{j+1} + C_T}{\widehat{\lambda}_j + C_T} : 1 \leq j \leq R\right\}, \tag{10}$$

where $C_T = (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T$. Theorem 2.4 shows that $\widetilde{r}$ is a consistent estimator for r.

Theorem 2.4. *Let Conditions 2.1-2.5 hold, and* $(p^{1-\delta} + \kappa_2)pT^{-1/2}\log T = o(1)$. *Then* $P(\widetilde{r} \neq r) \to 0$.

With the estimator $\widetilde{r}$, we may define an estimator for $\mathbf{A}$ as $\widetilde{\mathbf{A}} = (\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_{\widetilde{r}})$, where $\widehat{\mathbf{a}}_1, \ldots, \widehat{\mathbf{a}}_{\widetilde{r}}$ are the orthonormal eigenvectors of $\widehat{\mathbf{M}}$, defined in (6), corresponding to the $\widetilde{r}$ largest eigenvalues. Then $\widetilde{\mathbf{A}} = \widehat{\mathbf{A}}$ when $\widetilde{r} = r$. To measure the error in estimating the factor loading space, we use

$$\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) = \sqrt{1 - \frac{1}{\max(\widetilde{r}, r)}\mathrm{tr}(\widetilde{\mathbf{A}}\widetilde{\mathbf{A}}^{\mathrm{T}}\mathbf{A}\mathbf{A}^{\mathrm{T}})}.$$

This is a modified version of (8). It takes into account the fact that the dimensions of $\mathcal{M}(\widetilde{\mathbf{A}})$ and $\mathcal{M}(\mathbf{A})$ may be different. Obviously $\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) = D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A}))$ if $\widetilde{r} = r$. We show below that $\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) \to 0$ in probability at the same rate as $D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A}))$. Hence even without knowing r, $\mathcal{M}(\widetilde{\mathbf{A}})$ is a consistent estimator for $\mathcal{M}(\mathbf{A})$. Let $\rho = \rho(T,p)$ denote the convergence rate of $D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A}))$, i.e. $\rho D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A})) = O_p(1)$, see Theorems 2.1 and 2.3. For any $\epsilon > 0$, there exists a positive constant M_ϵ such that $P\{\rho D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A})) > M_\epsilon\} < \epsilon$. Then,

$$\begin{aligned} & P\{\rho\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) > M_\epsilon\} \\ \leq\ & P\{\rho D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A})) > M_\epsilon, \widetilde{r} = r\} + P\{\rho\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) > M_\epsilon, \widetilde{r} \neq r\} \\ \leq\ & P\{\rho D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A})) > M_\epsilon\} + o(1) \\ \leq\ & \epsilon + o(1) \to \epsilon \end{aligned}$$

which implies $\rho\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) = O_p(1)$. Hence, $\widetilde{D}(\mathcal{M}(\widetilde{\mathbf{A}}),\mathcal{M}(\mathbf{A})) \to 0$ shares the same convergence rate of $D(\mathcal{M}(\widehat{\mathbf{A}}),\mathcal{M}(\mathbf{A}))$ which means that $\mathcal{M}(\widetilde{\mathbf{A}})$ has the oracle property in estimating the factor loading space $\mathcal{M}(\mathbf{A})$.

3 Models with endogeneity

In last section, the consistent estimation for the coefficient matrix $\mathbf{D}$ is used in identifying the latent factor process. The consistency is guaranteed by the assumption that $\mathrm{cov}(\mathbf{z}_t, \mathbf{x}_t) = E(\mathbf{z}_t\mathbf{x}_t^{\mathrm{T}}) = \mathbf{0}$. However when the endogeneity exists in model (1) in the sense that the regressor $\mathbf{z}_t$ and the latent factor $\mathbf{x}_t$ are contemporaneously correlated with each other, $\mathbf{D}$ is no longer identifiable. Nevertheless (1) can be written as

$$\begin{aligned} \mathbf{y}_t &= [\mathbf{D} + \mathbf{A}E(\mathbf{x}_t\mathbf{z}_t^{\mathrm{T}})\{E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\}^{-1}]\mathbf{z}_t + \mathbf{A}[\mathbf{x}_t - E(\mathbf{x}_t\mathbf{z}_t^{\mathrm{T}})\{E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\}^{-1}\mathbf{z}_t] + \boldsymbol{\varepsilon}_t \\ &\equiv \mathbf{D}^\star\mathbf{z}_t + \mathbf{A}\mathbf{x}_t^\star + \boldsymbol{\varepsilon}_t, \end{aligned} \tag{11}$$

where the latent factor $\mathbf{x}_t^\star = \mathbf{x}_t - E(\mathbf{x}_t\mathbf{z}_t^{\mathrm{T}})\{E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\}^{-1}\mathbf{z}_t$ is uncorrelated with the regressor $\mathbf{z}_t$. Hence if we apply the methods presented in Section 2 to model (1) in the presence of the endogeneity, $\widehat{\mathbf{D}}$ defined in (3) is a consistent estimator for $\mathbf{D}^\star = \mathbf{D} + \mathbf{A}E(\mathbf{x}_t\mathbf{z}_t^{\mathrm{T}})\{E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\}^{-1}$ instead of the original regression coefficient $\mathbf{D}$, provided that $\mathbf{D}^\star$ so defined is a constant matrix independent of t. The latter is guaranteed when both $\mathbf{x}_t$ and $\mathbf{z}_t$ are stationary. Furthermore, the recovered factor process $\widehat{\mathbf{x}}_t$ is an estimator for $\mathbf{x}_t^\star$. Hence in the presence of the endogeneity and if $\mathbf{D}^\star$ defined in (11) is a constant matrix, the factor loading space $\mathcal{M}(\mathbf{A})$ can still be estimated consistently although the ordinary least squares estimator for the regression coefficient matrix $\mathbf{D}$ is no longer consistent.

For some applications, the interest lies in estimating the 'original' $\mathbf{D}$ and $\mathbf{x}_t$; see, e.g., Angrist and Krueger (1991). Then we may employ a set of instrument variables $\mathbf{w}_t$ in the sense that $\mathbf{w}_t$ is correlated with $\mathbf{z}_t$ but uncorrelated with both $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$. Usually, we require that $\mathbf{w}_t$ is $q \times 1$ with $q \geq m$. It follows from (1) that

$$\mathbf{y}_t\mathbf{w}_t^{\mathrm{T}} = \mathbf{D}\mathbf{z}_t\mathbf{w}_t^{\mathrm{T}} + \boldsymbol{\varepsilon}_t^\star, \quad \boldsymbol{\varepsilon}_t^\star = \mathbf{A}\mathbf{x}_t\mathbf{w}_t^{\mathrm{T}} + \boldsymbol{\varepsilon}_t\mathbf{w}_t^{\mathrm{T}}. \tag{12}$$

Since $E(\mathbf{x}_t\mathbf{w}_t^{\mathrm{T}}) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}_t\mathbf{w}_t^{\mathrm{T}}) = \mathbf{0}$, we may view the first equation in the above expression as similar to a 'normal equation' in a least squares problem by ignoring $\boldsymbol{\varepsilon}_t^\star$. This leads to the following estimator for $\mathbf{D}$:

$$\widehat{\mathbf{D}} = \left(\frac{1}{T}\sum_{t=1}^{\mathrm{T}} \mathbf{y}_t\mathbf{w}_t^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t\mathbf{w}_t^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\right)^{-1}. \tag{13}$$

where $\mathbf{R}$ is any $m \times q$ constant matrix with $\mathrm{rank}(\mathbf{R}) = m$, to match the lengths of $\mathbf{w}_t$ and $\mathbf{z}_t$. When $q = m$, we can choose $\mathbf{R} = \mathbf{I}_m$. This is the 'instrument variables method' widely used in econometrics. We refer to Morimune (1983), Bound et al. (1996), Donald and Newey (2001), Hahn and Hausman (2002) and Caner and Fan (2012) for further discussion on the choice of instrument variables and the related issues. It follows from (12) and (13) that

$$\widehat{\mathbf{D}} - \mathbf{D} = \left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\varepsilon}_t^\star\mathbf{R}^{\mathrm{T}}\right)\left(\frac{1}{T}\sum_{t=1}^{T} \mathbf{z}_t\mathbf{w}_t^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\right)^{-1}.$$

The proposition below shows that $\widehat{\mathbf{D}}$ is a consistent estimator with the optimal convergence rate. See also Proposition 2.1.

Condition 3.1. For any $i = 1, \ldots, q$ and t, $E(|w_{i,t}|^{2\gamma}) \leq C_1$ for $\gamma > 2$ and $C_1 > 0$ specified in, respectively, Conditions 2.1 and 2.2.

Condition 3.2. The smallest eigenvalue of $\{E(\mathbf{w}_t\mathbf{z}_t^{\mathrm{T}})\}^{\mathrm{T}}\mathbf{R}^{\mathrm{T}}\mathbf{R}\{E(\mathbf{w}_t\mathbf{z}_t^{\mathrm{T}})\}$ is uniformly bounded away from zero for all t.

Condition 3.2 implies that all the components of the instrument variables $\mathbf{w}_t$ are correlated with the regressor $\mathbf{z}_t$. When $q = m$ and $\mathbf{R} = \mathbf{I}_m$, it reduces to the condition that all the singular values of $E(\mathbf{w}_t\mathbf{z}_t^{\mathrm{T}})$ are uniformly bounded away from zero for all t.

Proposition 3.1. *Let Conditions 2.1-2.2 and 3.1-3.2 hold. As $T \to \infty$ and $p \to \infty$, it holds that*

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_F = O_p(p^{1/2}T^{-1/2}).$$

With the consistent estimator $\widehat{\mathbf{D}}$ in (13), the factor loading space and the latent factor process may be estimated in the same manner as in Section 2.2. The asymptotic properties presented in Theorems 2.1-2.3 can be reproduced in the similar manner.

4 Models with nonlinear regression functions

Now we consider the model with nonlinear regression term:

$$\mathbf{y}_t = \mathbf{g}(\mathbf{u}_t) + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t, \tag{14}$$

where $\mathbf{g}(\cdot)$ is an unknown nonlinear function, $\mathbf{u}_t$ is an observed process with fixed dimension, and other terms are the same as in model (1). One way to handle a nonlinear regression is to transform it into a high-dimensional linear regression problem. To this end, let $\mathbf{g} = (g_1, \ldots, g_p)^{\mathrm{T}}$, and

$$g_i(\mathbf{u}) = \sum_{j=1}^{\infty} d_{i,j} l_j(\mathbf{u}), \qquad i = 1, 2, \ldots,$$

where $\{l_j(\cdot)\}$ is a set of base functions. Suppose we use the approximation with the first m terms only. Let $\mathbf{z}_t = (l_1(\mathbf{u}_t), \ldots, l_m(\mathbf{u}_t))^{\mathrm{T}}$, and $\mathbf{D}$ be the $p \times m$ matrix with $d_{i,j}$ as its (i,j)-th element, then (14) can be expressed as

$$\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t + \mathbf{e}_t, \tag{15}$$

where the additional error term $\mathbf{e}_t$ collects the residuals in approximating $\mathbf{g}(\cdot)$ by the first m terms only, i.e. the ith component of $\mathbf{e}_t$ is $\sum_{j>m} d_{i,j} l_j(\mathbf{u}_t)$. This makes (15) formally different from model (1). Furthermore a fundamentally new feature in (15) is that m may be large in relation to p or/and T. Hence the new asymptotic theory with all $T, p, m \to \infty$ together will be established in order to take into account those non-trivial changes. Due to (11), we may always assume that $\mathrm{cov}(\mathbf{z}_t, \mathbf{x}_t) = \mathbf{0}$. Condition 4.2 below ensures that $\mathbf{e}_t$ in (15) is asymptotically negligible. Hence

model (15) is as identifiable as (1) at least asymptotically when $m \to \infty$. Consequently we may estimate $\mathbf{D}$ using the ordinary least squares estimator:

$$\widehat{\mathbf{D}} = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{y}_t\mathbf{z}_t^{\mathrm{T}}\right)\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right)^{-1}.$$

We introduce some regularity conditions first.

Condition 4.1. Supports of the process $\mathbf{u}_t$ are subsets of $\mathcal{U}$, where $\mathcal{U}$ is compact with nonempty interior. Furthermore the density function of $\mathbf{u}_t$ is uniformly bounded and bounded away from zero for all t.

Condition 4.2. It holds for all large m that

$$\sup_i \sup_{\mathbf{u}\in\mathcal{U}} \left| g_i(\mathbf{u}) - \sum_{j=1}^{m} d_{i,j} l_j(\mathbf{u}) \right| = O(m^{-\lambda})$$

where $\lambda > 1/2$ is a constant.

Condition 4.3. The eigenvalues of $E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})$, are uniformly bounded away from zero and infinity for all t, where $\mathbf{z}_t = (l_1(\mathbf{u}_t), \ldots, l_m(\mathbf{u}_t))^{\mathrm{T}}$.

Condition 4.4. $E(\mathbf{A}\mathbf{x}_t|\mathbf{u}_t) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}_t|\mathbf{u}_t) = \mathbf{0}$ for all t.

Condition 4.5. For each $j = 1, \ldots, m$, $E(|l_j(\mathbf{u}_t)|^{2\gamma}) \le C_1$, where $\gamma > 2$ and $C_1 > 0$ are specified in, respectively, Conditions 2.1 and 2.2.

Condition 4.1 is often assumed in nonparametric estimation, it can be weakened at the cost of lengthier proofs. Condition 4.2 quantifies the approximation error for regression function $\mathbf{g}(\cdot)$. It is fulfilled by commonly used sieve basis functions such as spline, wavelets, or the Fourier series, provided that all components of $\mathbf{g}(\cdot)$ are in the Hölder space. See Ai and Chen (2003) for further detail on the sieve method.

Proposition 4.1. *Let Conditions 2.1-2.2 and 4.2-4.5 hold, and* $mT^{-1/2} = o(1)$. *Then*

$$\|\widehat{\mathbf{D}} - \mathbf{D}\|_F = O_p(p^{1/2}m^{1/2}T^{-1/2} + p^{1/2}m^{1/2-\lambda}).$$

Comparing this proposition with Propositions 2.1 and 3.1, m enters the convergence rates, and the term $O_p(p^{1/2}m^{1/2-\lambda})$ is due to approximating $\mathbf{g}(\mathbf{u}_t)$ by $\mathbf{D}\mathbf{z}_t$. Based on the estimator $\widehat{\mathbf{D}}$, we can define an estimator for the nonlinear regression function

$$\widehat{\mathbf{g}}(\mathbf{u}) = \widehat{\mathbf{D}}(l_1(\mathbf{u}), \ldots, l_m(\mathbf{u}))^{\mathrm{T}}.$$

The theorem below follows from Proposition 4.1. It gives the convergence rate for $\widehat{\mathbf{g}}$.

Theorem 4.1. *Let Conditions 2.1-2.2 and 4.1-4.5 hold, and* $mT^{-1/2} = o(1)$. *Then*

$$\int_{\mathbf{u}\in\mathcal{U}} \|\widehat{\mathbf{g}}(\mathbf{u}) - \mathbf{g}(\mathbf{u})\|_2^2 \, d\mathbf{u} = O_p(pmT^{-1} + pm^{-2\lambda}).$$

It is easy to see from Theorem 4.1 that the best rate for $\widehat{\mathbf{g}}(\cdot)$ is attained if we choose $m \asymp T^{1/(2\lambda+1)}$, which fulfills the condition $mT^{-1/2} = o(1)$ as $\lambda > 1/2$. When $\mathbf{g}(\cdot)$ is twice differentiable, $\lambda = 2$ for some basis functions, the convergence rate is $pT^{-4/5}$. This is the optimal rate for the nonparametric regression of p functions (Stone, 1985). Hereafter, we always set $m \asymp T^{1/(2\lambda+1)}$.

With the estimator $\widehat{\mathbf{D}}$, we may proceed as in Section 2.2 to estimate the factor loading space and to recover the latent factor process. However there is a distinctive new feature now: the number of lags $\bar{k}$ used in defining both $\mathbf{M}$ in (5) and $\widehat{\mathbf{M}}$ in (7) may tend to infinity together with m in order to achieve good convergence rates.

Theorem 4.2. *Let conditions 2.1-2.2, 2.4 and 4.2-4.5 hold, $\lambda \geq 1$, $\bar{k}T^{-1/2} = o(1)$, and $m \asymp T^{1/(2\lambda+1)}$. Suppose that r is known, and the r positive eigenvalues of $\mathbf{M}$ are distinct. Then*

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 = \begin{cases} O_p\{p^{\delta}[\bar{k}^{1/2}T^{-1/2} + \bar{k}^{-1}T^{(1-\lambda)/(2\lambda+1)}]\}, \\ \quad \textit{if } \kappa_2 = o(p^{1-\delta}) \textit{ and } p^{2\delta}[\bar{k}T^{-1} + T^{(2-2\lambda)/(2\lambda+1)}] = o(1); \\ O_p\{p\kappa_2\kappa_1^{-2}[\bar{k}^{1/2}T^{-1/2} + \bar{k}^{-1}T^{(1-\lambda)/(2\lambda+1)}]\}, \\ \quad \textit{if } p^{1-\delta} = o(\kappa_1) \textit{ and } p^2\kappa_2^2\kappa_1^{-4}[\bar{k}T^{-1} + T^{(2-2\lambda)/(2\lambda+1)}] = o(1). \end{cases}$$

From Theorem 4.2, the best convergence rate for $\widehat{\mathbf{A}}$ is attained when we choose $\bar{k} \asymp T^{1/(2\lambda+1)}$. The model with linear regression considered in Section 2.3 corresponds to the cases with $\lambda = \infty$. Note Theorem 4.2 implies that $\bar{k} \asymp 1$ should be used when $\lambda = \infty$ and m is fixed in order to attain the best possible rates. This is consistent with the procedures used in Section 2.2.

Now we comment on the impact of p on the convergence rate, which depends critically on the factor strength $\delta \in [0, 1]$ specified in Condition 2.4. To simplify the notation, let $\kappa_1 \asymp \kappa_2 \asymp \kappa$ which is a mild assumption in practice. Suppose $p^{\delta}T^{(1-\lambda)/(2\lambda+1)} = o(1)$ and $\bar{k} \asymp T^{1/(2\lambda+1)}$, Theorem 4.2 then reduces to

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 = \begin{cases} O_p(p^{\delta}T^{-\lambda/(2\lambda+1)}), \text{ if } \kappa = o(p^{1-\delta}); \\ O_p(p\kappa^{-1}T^{-\lambda/(2\lambda+1)}), \text{ if } p^{1-\delta} = o(\kappa). \end{cases}$$

If $\kappa p^{-(1-\delta)} \to \infty$, there is an additional factor $\kappa p^{-(1-\delta)}$ in the convergence rate of $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2$ than that under the setting $\kappa p^{-(1-\delta)} \to 0$, which implies that $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2$ converges to zero faster in the case $\kappa = o(p^{1-\delta})$. The dimension p must satisfy the condition $p^{\delta}T^{(1-\lambda)/(2\lambda+1)} = o(1)$, which is automatically fulfilled when $\delta = 0$, i.e. the factors are strong. However when the factors are weak in the sense $\delta \neq 0$, p can only be in the order $p = o(T^{(\lambda-1)/\{(2\lambda+1)\delta\}})$ to ensure the consistency in estimating the factor loading matrix.

Theorem 4.3. *Let the condition of Theorem 4.2 hold. In addition, if $\|\mathbf{\Sigma}_{\varepsilon}\|_2$ is bounded as $p \to \infty$, then*

$$p^{-1/2}\|\widehat{\mathbf{A}}\widehat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_t\|_2 = O_p(\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 + p^{-1/2} + T^{-(2\lambda-1)/(4\lambda+2)}).$$

Comparing the above theorem with Theorem 2.2, it has one more term $T^{(2\lambda-1)/(4\lambda+2)}$ in the convergence rate. When the dimension m is fixed and $\lambda = \infty$, it reduces to Theorem 2.2. On the

other hand, we can also consider the model (1) with diverging number of regressors (i.e., $m \to \infty$). Noting Proposition 4.1 with $\lambda = \infty$ and using the same argument of Theorem 2.1, it holds that

$$\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 = \begin{cases} O_p\{\bar{k}^{-1}p^{\delta}(\bar{k}^{3/2} + m^{1/2})T^{-1/2}\}, \\ \quad \text{if } \kappa_2 = o(p^{1-\delta}) \text{ and } p^{\delta}(\bar{k}^{1/2} + m^{1/2})T^{-1/2} = o(1); \\ O_p\{\bar{k}^{-1}p\kappa_2\kappa_1^{-2}(\bar{k}^{3/2} + m^{1/2})T^{-1/2}\}, \\ \quad \text{if } p^{1-\delta} = o(\kappa_1) \text{ and } p\kappa_2\kappa_1^{-2}(\bar{k}^{1/2} + m^{1/2})T^{-1/2} = o(1); \end{cases}$$

provided that $m = o(T^{1/2})$ and $\bar{k} = o(T^{1/3})$. Theorem 2.1 can be regarded as the special case of this result with fixed $\bar{k}$ and m. Note that the best convergence rate for $\|\widehat{\mathbf{A}} - \mathbf{A}\|_2$ is attained under such setting if we choose $\bar{k} \asymp m^{1/3}$.

5 Numerical properties

In this section, we illustrate the finite sample properties of the proposed methods in two simulated models, one with a linear regression term and one with a nonlinear regression term. For the linear model, both stationary and nonstationary factors were employed. In each model, we set the dimension of $\mathbf{y}_t$ at $p = 100, 200, 400, 600, 800$ and the sample size $T = 0.5p$, p, $1.5p$ respectively. For each setting, 200 samples were generated.

Table 1: Relative frequency estimates of $P(\widehat{r} = r)$ for Example 1 with stationary factors.

		p	100	200	400	600	800
$\delta = 0$	**D** known	$T = 0.5p$	0.700	0.960	0.990	0.995	1
		$T = p$	0.900	0.985	1	1	1
		$T = 1.5p$	0.980	1	1	1	1
	D unknown	$T = 0.5p$	0.615	0.940	0.990	0.995	1
		$T = p$	0.865	0.985	1	1	1
		$T = 1.5p$	0.960	1	1	1	1
$\delta = 0.5$	**D** known	$T = 0.5p$	0.105	0.805	0.950	0.805	0.930
		$T = p$	0.285	0.880	0.940	1	0.975
		$T = 1.5p$	0.895	0.975	1	1	1
	D unknown	$T = 0.5p$	0.065	0.765	0.930	0.780	0.910
		$T = p$	0.280	0.880	0.940	1	0.975
		$T = 1.5p$	0.870	0.975	0.995	1	1

Example 1. Consider the linear model $\mathbf{y}_t = \mathbf{D}\mathbf{z}_t + \mathbf{A}\mathbf{x}_t + \boldsymbol{\varepsilon}_t$, in which $\mathbf{z}_t$ follows the VAR(1) model:

$$\mathbf{z}_t = \begin{pmatrix} 5/8 & 1/8 \\ 1/8 & 5/8 \end{pmatrix} \mathbf{z}_{t-1} + \mathbf{e}_t, \tag{16}$$

where $\mathbf{e}_t \sim N(\mathbf{0}, \mathbf{I}_2)$. Let $\mathbf{D}$ be a $p \times 2$ matrix of which the elements were generated independently from the uniform distribution $U(-2, 2)$, $\mathbf{x}_t$ be 3×1 VAR(1) process with independent $N(\mathbf{0}, \mathbf{I}_3)$ innovations and the diagonal autoregressive coefficient matrix with 0.6, -0.5 and 0.3 as the main

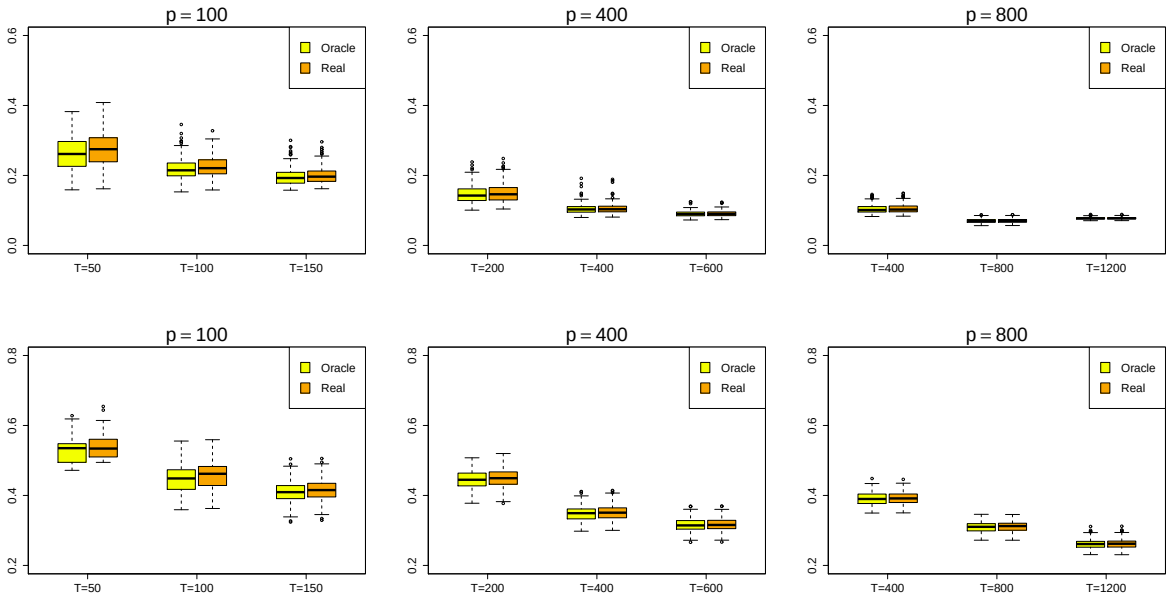


Figure 1: Boxplots of $\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}$ for Example 1 with stationary factor, and $\delta = 0$ (3 top panels) and $\delta = 0.5$ (3 bottom panels). Errors obtained using true $\mathbf{D}$ are marked with 'oracle', and using $\widehat{\mathbf{D}}$ are marked with 'real'.

diagonal elements. This is a stationary factor process with $r = 3$ factors. The elements of $\mathbf{A}$ were drawn independently from $U(-2, 2)$ resulting a strong factor case with $\delta = 0$. Also we considered a weak factor case with $\delta = 0.5$ for which randomly selected $p - \lfloor p^{1/2} \rfloor$ elements in each column of $\mathbf{A}$ were set to 0. Let $\boldsymbol{\varepsilon}_t$ be independent and $N(\mathbf{0}, \mathbf{I}_p)$. To show the impact of the estimated coefficient matrix $\widehat{\mathbf{D}}$ on the estimation for the factors, we also report the results from using the true $\mathbf{D}$. We report the results with $\bar{k} = 1$ only, since the results with $1 \leq \bar{k} \leq 10$ are similar. The relative frequency estimates of $P(\widehat{r} = r)$ are reported in Table 1. It shows that the defect in estimating r due to the errors in estimating $\mathbf{D}$ is almost negligible. Fig.1 displays the boxplots of the estimation errors $\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}$. Again the performance with the estimated coefficient matrix $\widehat{\mathbf{D}}$ is only slightly worse than that with the true $\mathbf{D}$. When the factors are weaker (i.e. when $\delta = 0.5$), it is harder to estimate both the number of factors and the factor loading space. All those findings are in line with the asymptotic results presented in Section 2.3.

Now we consider the case with the endogeneity. To this end, we changed the definition for the regressor process $\mathbf{z}_t$ in the above setting. Instead of (16), we let

$$z_{1,t} = 0.1x_{1,t} + 0.1u_t + 0.1u_t^2, \quad z_{2,t} = 0.1x_{2,t} - 0.1u_t + 0.1u_t^2,$$

where u_t is an AR(1) process defined by $u_t = 0.5u_{t-1} + \epsilon_t$ and $\epsilon_t \sim N(0, 1)$. The ordinary least squares estimator of $\mathbf{D}$ is no longer consistent now. We employ two different instrument variables $\mathbf{w}_t = (u_t, u_t^2)^{\mathrm{T}}$ and $\widetilde{\mathbf{w}}_t = (u_t, u_t^2, u_t^3, u_t^4)^{\mathrm{T}}$, as they are correlated with $\mathbf{z}_t$ but uncorrelated with $\mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$. The estimation error for $\widehat{\mathbf{D}}$ is measured by the normalized Frobenius norm $p^{-1/2}\|\widehat{\mathbf{D}} - \mathbf{D}\|_F$.

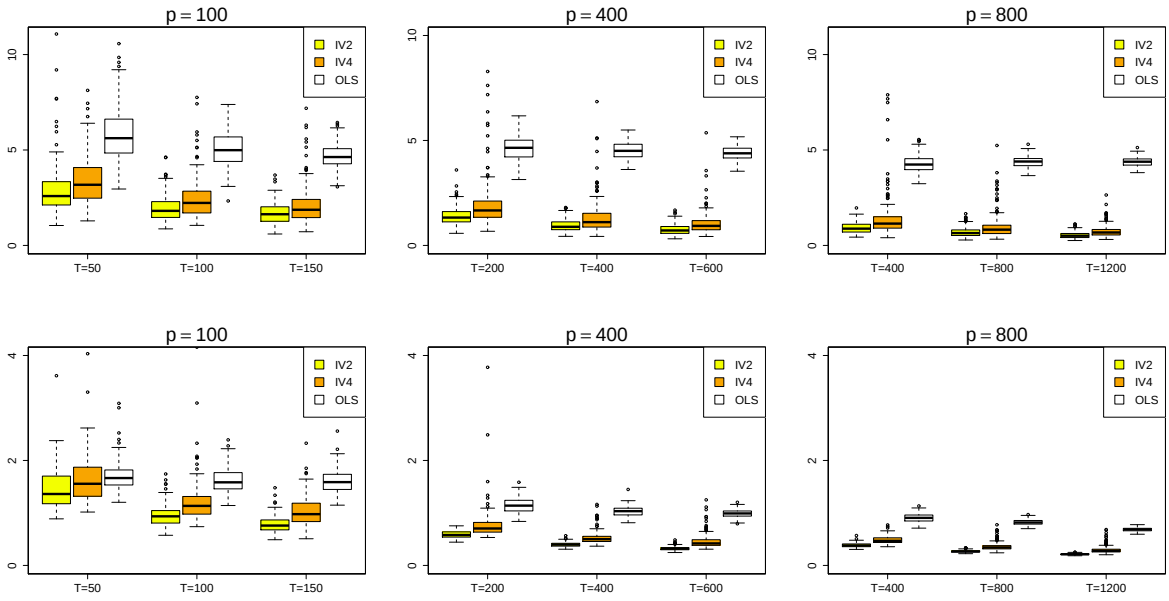


Figure 2: Boxplots of $p^{-1/2}\|\widehat{\mathbf{D}} - \mathbf{D}\|_F$ for Example 1 with endogeneity, and $\delta = 0$ (3 top panels) and $\delta = 0.5$ (3 bottom panels).

Setting $\mathbf{R} = \mathbf{I}_2$ for $\mathbf{w}_t$ and the elements of $\widetilde{\mathbf{R}}$ are generated from $U(-2, 2)$ for $\widetilde{\mathbf{w}}_t$ in (13), we computed first both the ordinary least squares (OLS) estimates and the instrument variable method (IV) estimates for $\mathbf{D}$, and then the estimates for the number of factors r and the factor loading matrix $\mathbf{A}$ based on, respectively, the two sets of residuals resulted from the two regression estimation methods. The results are reported in Figs.2 and 3 and Table 2 where IV2 and IV4 represent the estimation using $\mathbf{w}_t$ and $\widetilde{\mathbf{w}}_t$ respectively. Those simulation results reinforce the findings in Section 3, which indicate that the existence of the endogeneity has no impact in identifying and in estimating the factor loading space. More precisely, Fig.2 shows that the errors $p^{-1/2}\|\widehat{\mathbf{D}} - \mathbf{D}\|_F$ for the OLS method are unusually large, as it effectively estimates $\mathbf{D}^\star$ in (11) instead of $\mathbf{D}$. On the other hand, the IV method provides accurate estimates for $\mathbf{D}$. However the differences of the two methods on the subsequent estimation for the number of factors r and the factor loading space $\mathcal{M}(\mathbf{A})$ are small; see Table 2 and Fig.3. Since the IV method uses extra information, it tends to offer slightly better performance. Nevertheless Table 2 indicates that this improvement in estimating r is almost negligible. Also, the results are not sensitive to the choice of $\mathbf{R}$ as long as the instrument variables are properly selected.

Now we consider the model with nonstationary factors $\mathbf{x}_t = 3(x_{1,t}, x_{2,t}, x_{3,t})^{\mathrm{T}}$:

$$x_{1,t} - 2t/T = 0.8(x_{1,T-1} - 2t/T) + e_{1,t}, \quad x_{2,t} = 3t/T, \quad x_{3,t} = x_{3,t-1} + \sqrt{\frac{10}{T}}e_{3,t}, \tag{17}$$

where $e_{j,t}$ are independent and $N(0, 1)$. The other settings are the same as the first part of this example. The results are reported in Table 3 and Fig.4. The patterns are similar to those in Table 1 and Fig.1, except that for a fixed p, the performance does not necessarily improve when the

Table 2: Relative frequency estimates of $P(\widehat{r} = r)$ for Example 1 with endogeneity.

		p	100	200	400	600	800
$\delta = 0$	IV2	$T = 0.5p$	0.660	0.885	0.995	0.995	1
		$T = p$	0.855	0.990	1	1	1
		$T = 1.5p$	0.960	1	1	1	1
	IV4	$T = 0.5p$	0.590	0.865	0.975	0.970	0.965
		$T = p$	0.845	0.970	0.970	0.990	0.985
		$T = 1.5p$	0.930	0.990	0.980	0.970	0.975
	OLS	$T = 0.5p$	0.580	0.865	0.990	1	1
		$T = p$	0.855	0.980	1	1	1
		$T = 1.5p$	0.945	1	1	0.995	1
$\delta = 0.5$	IV2	$T = 0.5p$	0.280	0.665	0.625	0.630	0.620
		$T = p$	0.600	0.715	0.980	1	1
		$T = 1.5p$	0.550	0.980	0.990	1	1
	IV4	$T = 0.5p$	0.205	0.570	0.550	0.605	0.550
		$T = p$	0.510	0.650	0.890	0.960	0.925
		$T = 1.5p$	0.580	0.915	0.950	0.935	0.940
	OLS	$T = 0.5p$	0.225	0.405	0.635	0.625	0.640
		$T = p$	0.535	0.705	1	1	0.995
		$T = 1.5p$	0.630	0.955	1	1	0.995

sample size T increases; see Fig.4. This is due to the nonstationary nature of the factors defined in (17): new observations bring in the information on the new and time-varying underlying structure as far as the factor processes are concerned.

Example 2. We now consider a model with nonlinear regression function. Let $\mathbf{u}_t = u_t$ be a univariate AR(1) process defined by $u_t = 0.5u_{t-1} + e_t$ with independent $N(0,1)$ innovations e_t. The nonlinear regression function $\mathbf{g}(u_t) = (g_1(u_t), \ldots, g_p(u_t))^{\mathrm{T}}$ was defined as

$$g_i(u_t) = \frac{\exp(\alpha_i^{(1)} u_t)}{1 + \exp(\alpha_i^{(1)} u_t)}, \quad i = 1, \ldots, \frac{p}{2} \quad \text{and} \quad g_i(u_t) = \sin(\alpha_i^{(2)} u_t), \quad i = \frac{p}{2} + 1, \ldots, p,$$

where the parameters $\alpha_i^{(1)}$ were drawn independently from $N(0,4)$, and $\alpha_i^{(2)}$ were drawn independently from $U(-2,2)$ respectively. We used the same $\mathbf{A}, \mathbf{x}_t$ and $\boldsymbol{\varepsilon}_t$ as in the first part of Example 1.

We used the polynomial expansion to approximate $\mathbf{g}(u_t)$, i.e. $g_i(u_t) \approx \sum_{j=1}^{m} d_{i,j} l_j(u_t)$ with $l_j(u_t) = u_t^{j-1}$, where the order m was set as $\lfloor 2T^{1/5} \rfloor$. We obtained $\hat{d}_{i,j}$ by the least square estimation. Put $\widehat{\mathbf{g}}(u_t) = (\hat{g}_1(u_t), \ldots, \hat{g}_p(u_t))^{\mathrm{T}}$ for $\hat{g}_i(u_t) = \sum_{j=1}^{m} \hat{d}_{i,j} l_j(u_t)$. The residuals $\widehat{\boldsymbol{\eta}}_t = \mathbf{y}_t - \widehat{\mathbf{g}}(\mathbf{u}_t)$ were then used to estimate the latent factors. We set $\bar{k} = \lfloor 2T^{1/5} \rfloor$; see Theorem 4.2. The simulation results are reported in Table 4 and Fig.5, which present similar patterns as in the first part of Example 1.

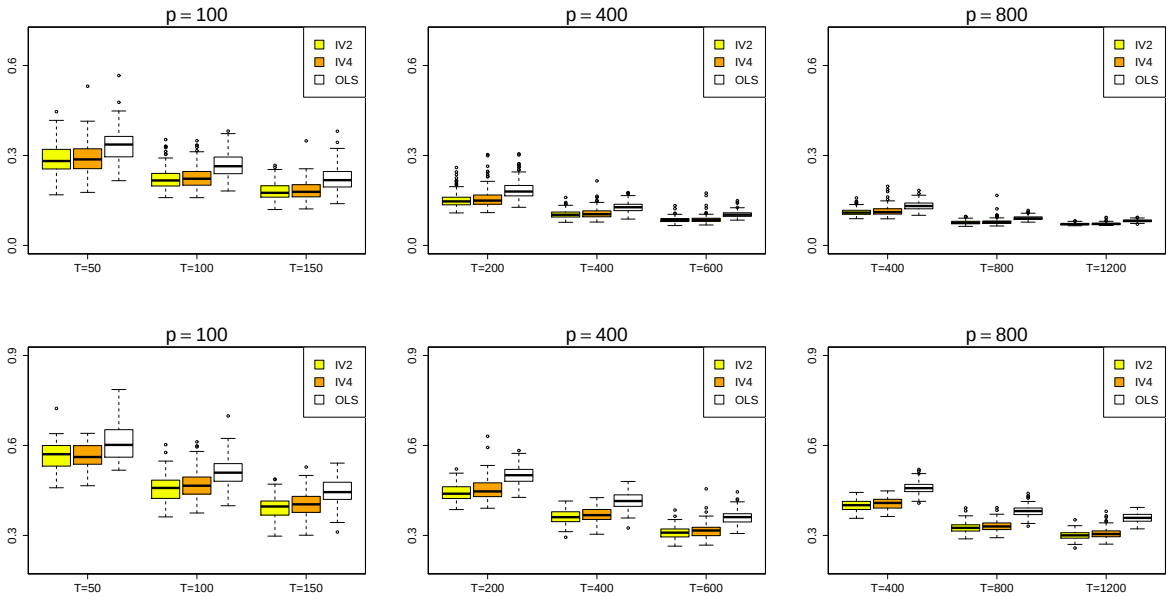


Figure 3: Boxplots of $\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}$ for Example 1 with endogeneity, and $\delta = 0$ (3 top panels) and $\delta = 0.5$ (3 bottom panels).

6 Real data analysis

We illustrate our method by modeling the daily returns of 123 stocks from 2 January 2002 to 11 July 2008. The stocks were selected among those contained in the S&P500 which were traded everyday during this period. The returns were calculated based on the daily close prices. We have in total $T = 1642$ observations with the dimension $p = 123$. This data has been analyzed in Lam and Yao (2012). They identified two factors under a pure factor model setting, i.e. model (1) with $\mathbf{z}_t \equiv \mathbf{0}$. Furthermore the estimated factor loading space contains the return of the S&P500. Hence it can be regarded as one of the two factors. Since the S&P500 index is often viewed as a proxy of the market index, it is reasonable to take its return as a known factor z_t in our model (1). We calculated the ordinary least square estimator for the regression coefficient matrix $\mathbf{D}$ which is now a 123×1 vector with each element representing the impact of the S&P500 index to the return of the corresponding stock. As all the estimated elements are positive, indicating the positive correlations between the returns of market index and the those 123 stocks.

Fig.6 displays the first 30 eigenvalues of $\widehat{\mathbf{M}}$, defined as in (6) with $\bar{k} = 1$, sorted in the descending order. The ratio of $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ in the right panel indicates that there is only one latent factor. Varying $\bar{k}$ between 1 to 20 did not alter this result. Fig.6(c) shows that the sparks of the estimated factor process occur around 22 July, 2002, which is consistent with the oscillations of S&P500 index, although the S&P500 are less volatile. The autocorrelations of the estimated factors $\widehat{\gamma}_j^{\mathrm{T}}(\mathbf{y}_t - \widehat{\mathbf{D}}\mathbf{z}_t)$, where $\widehat{\gamma}_j$ is the unit eigenvector of $\widehat{\mathbf{M}}$ corresponding to its jth largest eigenvalue, are plotted in Fig.7 for $j = 1, 2, 3$. The autocorrelations of the first factor is significant non-zero. On the other

Table 3: Relative frequency estimates of $P(\widehat{r} = r)$ for Example 1 with nonstationary factors.

		p	100	200	400	600	800
$\delta = 0$	**D** known	$T = 0.5p$	0.155	0.525	0.855	0.925	0.970
		$T = p$	0.465	0.800	0.940	0.990	0.990
		$T = 1.5p$	0.625	0.890	0.995	0.985	1
	D unknown	$T = 0.5p$	0.110	0.525	0.835	0.920	0.970
		$T = p$	0.430	0.780	0.940	0.990	0.990
		$T = 1.5p$	0.595	0.890	0.995	0.985	1
$\delta = 0.5$	**D** known	$T = 0.5p$	0	0.070	0.175	0.385	0.525
		$T = p$	0.025	0.235	0.535	0.705	0.765
		$T = 1.5p$	0.145	0.475	0.740	0.815	0.860
	D unknown	$T = 0.5p$	0	0.055	0.160	0.380	0.520
		$T = p$	0.025	0.215	0.520	0.685	0.760
		$T = 1.5p$	0.125	0.465	0.740	0.805	0.850

Table 4: Relative frequency estimates of $P(\widehat{r} = r)$ for Example 2 (with nonlinear regression).

		p	100	200	400	600	800
$\delta = 0$	**g** known	$T = 0.5p$	0.780	0.865	0.965	0.975	0.985
		$T = p$	0.840	0.920	0.990	1	1
		$T = 1.5p$	0.820	0.990	1	1	1
	g unknown	$T = 0.5p$	0.750	0.860	0.955	0.975	0.980
		$T = p$	0.830	0.890	0.990	1	1
		$T = 1.5p$	0.780	0.990	1	1	1
$\delta = 0.5$	**g** known	$T = 0.5p$	0.270	0.665	0.725	0.430	0.650
		$T = p$	0.390	0.700	0.850	0.810	0.800
		$T = 1.5p$	0.390	0.720	0.885	0.960	1
	g unknown	$T = 0.5p$	0.260	0.625	0.665	0.390	0.600
		$T = p$	0.390	0.655	0.760	0.810	0.795
		$T = 1.5p$	0.335	0.700	0.875	0.950	1

hand, there are hardly any significant non-zero autocorrelations for both the second and the third factors.

To gain some appreciation of the latent factor, we divide the 123 stocks into eight sectors: Financial, Basic Materials, Industrial Goods, Consumer Goods, Healthcare, Services, Utilities and Technology. We estimated the latent factor for each of those eight sectors. Those estimated sector factors are plotted in Fig.8. We observe that those estimated sector factors behave differently for the different sectors. Especially the Basic Materials sector exhibits the largest fluctuation. Consequently, we may deduce that the oscillations, especially the sparks, of the estimated factor in Fig.6(c) are largely due to changes in the Basic Materials sector. This is consistent with the relevant economics and finance principles. Basic Materials sector includes mainly the stocks of energy companies such as oil, gas, coal et al. The energy, especially oil, is the foundation for economic and social development. Hence, the changes in oil price are often considered as important

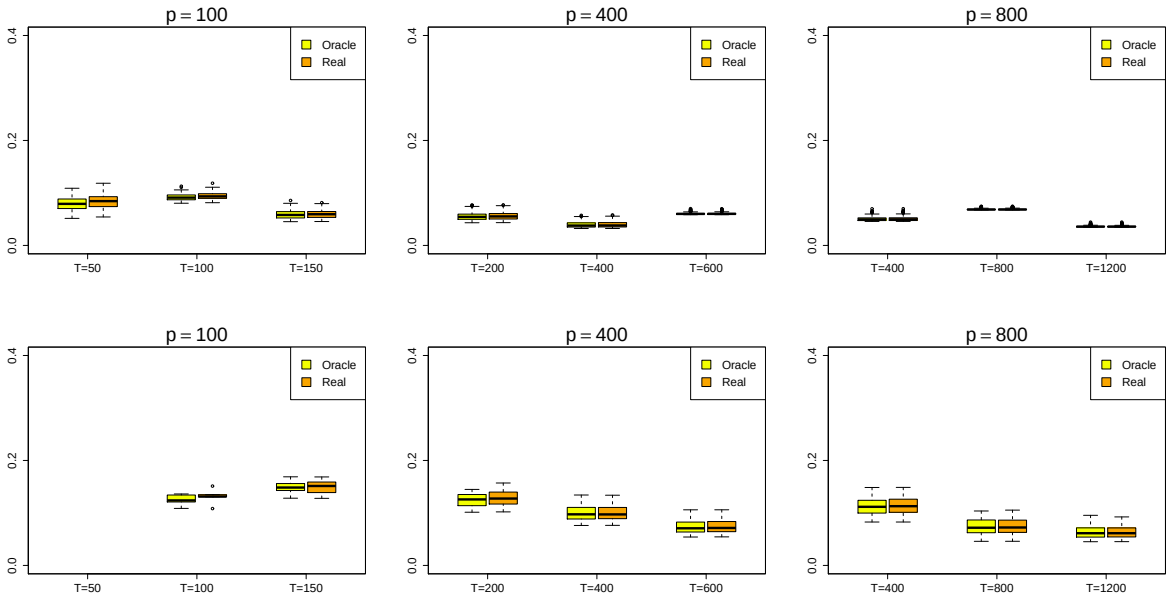


Figure 4: Boxplots of $\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}$ for Example 1 with nonstationary factors, and $\delta = 0$ (3 top panels) and $\delta = 0.5$ (3 bottom panels). Errors obtained using true $\mathbf{D}$ are marked with 'oracle', and using $\widehat{\mathbf{D}}$ are marked with 'real'.

events which underpin stock market fluctuations, see, e.g. Jones and Kaul (1996) and Kilian and Park (2009). During January 2002 to December 2003, international oil price had a huge increase. It rose 19% from the average in 2002. The 2003 invasion of Iraq marks a significant event as Iraq possesses a significant portion of the global oil reserve. Hence, the returns of the Basic Materials sector oscillate dramatically during that period. Among other sectors, Industrial and Consumer Goods have similar behaviors. However, the returns of both the sectors have little changes around zero, thus they have little contributions to the estimated factor. The same arguments hold for the Utilities sector. Also note that the returns for the Financial, Healthcare, Services and Technology sectors are much less volatile in comparison to that of the Basic Materials sector. We may conclude that, the estimated factor mainly reflects the feature of stocks in Basic Materials sector. The factor also contains some market information about the Financial, Healthcare, Services and Technology sectors, but less so on the Industrial Goods, Consumer Goods and Utilities sectors.

We repeat the above exercise for another set of return data in 14 July 2008 – 11 July 2014 from the 196 stocks contained in S&P500. Now $T = 1510$ and $p = 196$. The ratios of $\widehat{\lambda}_{i+1}/\widehat{\lambda}_i$ shown in Fig.9(b) indicate that there is still only one latent factor, in addition to S&P500. The estimated latent factor shown in Fig.9(c) fluctuated widely around 2009, which is consistent with the pronounced decline the stock market due to the global financial crisis. While the latent factor process seems to resemble the returns of S&P500 (see Figs.9(c) and (d)), the two series are orthogonal with each other (with the sample correlation coefficient equal to 0.00047). The estimated factors for each of the eight sectors are plotted in Fig.11. In contrast to the findings in 2002-2008,

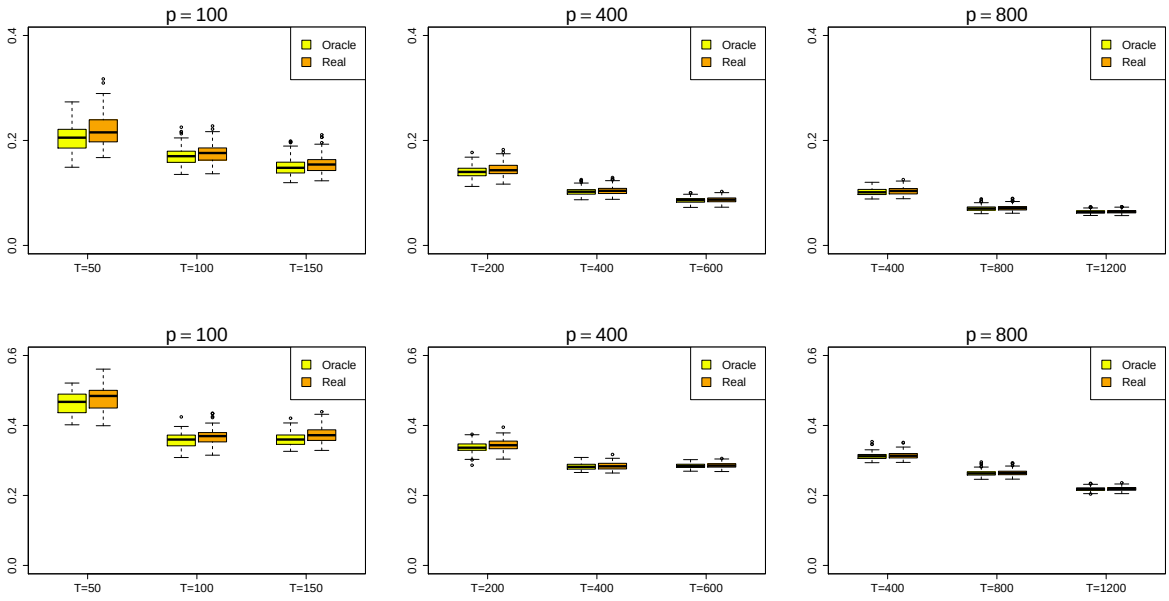


Figure 5: Boxplots of $\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}$ for Example 2 with nonlinear regression, and $\delta = 0$ (3 top panels) and $\delta = 0.5$ (3 bottom panels). Errors obtained using true $\mathbf{g}$ are marked with 'oracle', and using $\widehat{\mathbf{g}}$ are marked with 'real'.

all the eight sectors contributed to the fluctuation around 2009, though the financial sector was most predominant. The crisis caused by the sharp down-turn of financial industry in early 2009 impacted all sectors in the society.

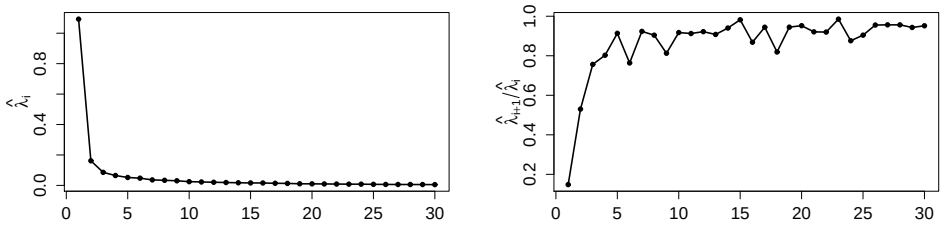
Acknowledgements

The authors would like to thank two referees for their helpful comments and suggestions.

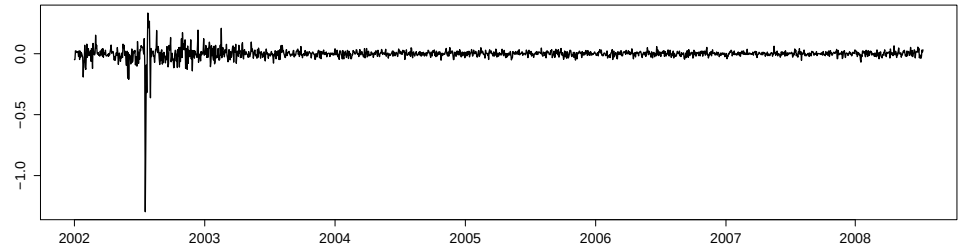
Appendix

Throughout the Appendix, we use Cs to denote generic uniformly positive constants only depends on the parameters C_is appear in the technical conditions which may be different in different uses. Meanwhile, we denote $\mathbf{A}\mathbf{x}_t$ by $\boldsymbol{\zeta}_t$. We first present the following lemmas which are used in proofs of the propositions and theorems.

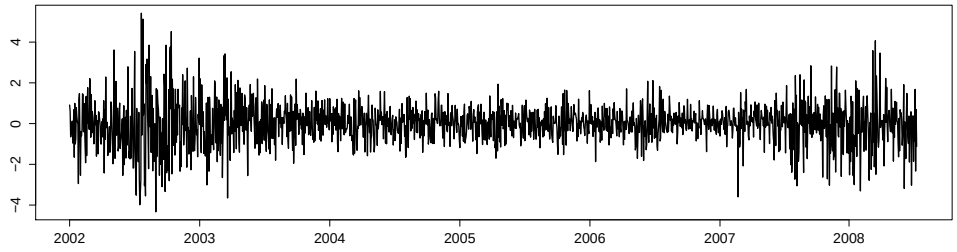
Lemma 6.1. *Under Conditions 2.1-2.2,* $\|T^{-1}\sum_{t=1}^{T}\{\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\}\|_F = O_p(mT^{-1/2})$.



(a) The eigenvalues of $\widehat{\mathbf{M}}$ (b) Ratio of the eigenvalues of $\widehat{\mathbf{M}}$



(c) The estimated latent factor



(d) S&P 500 returns

Figure 6: The estimated eigenvalues (multiplied by 10^6), the ratio of eigenvalues, the estimated latent factor and the S&P500 returns in 2 January 2002 – 11 July 2008.

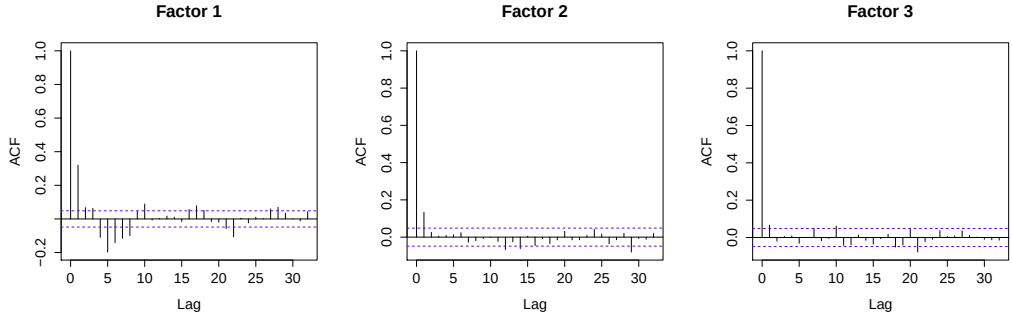


Figure 7: The ACFs of the first three estimated factors.

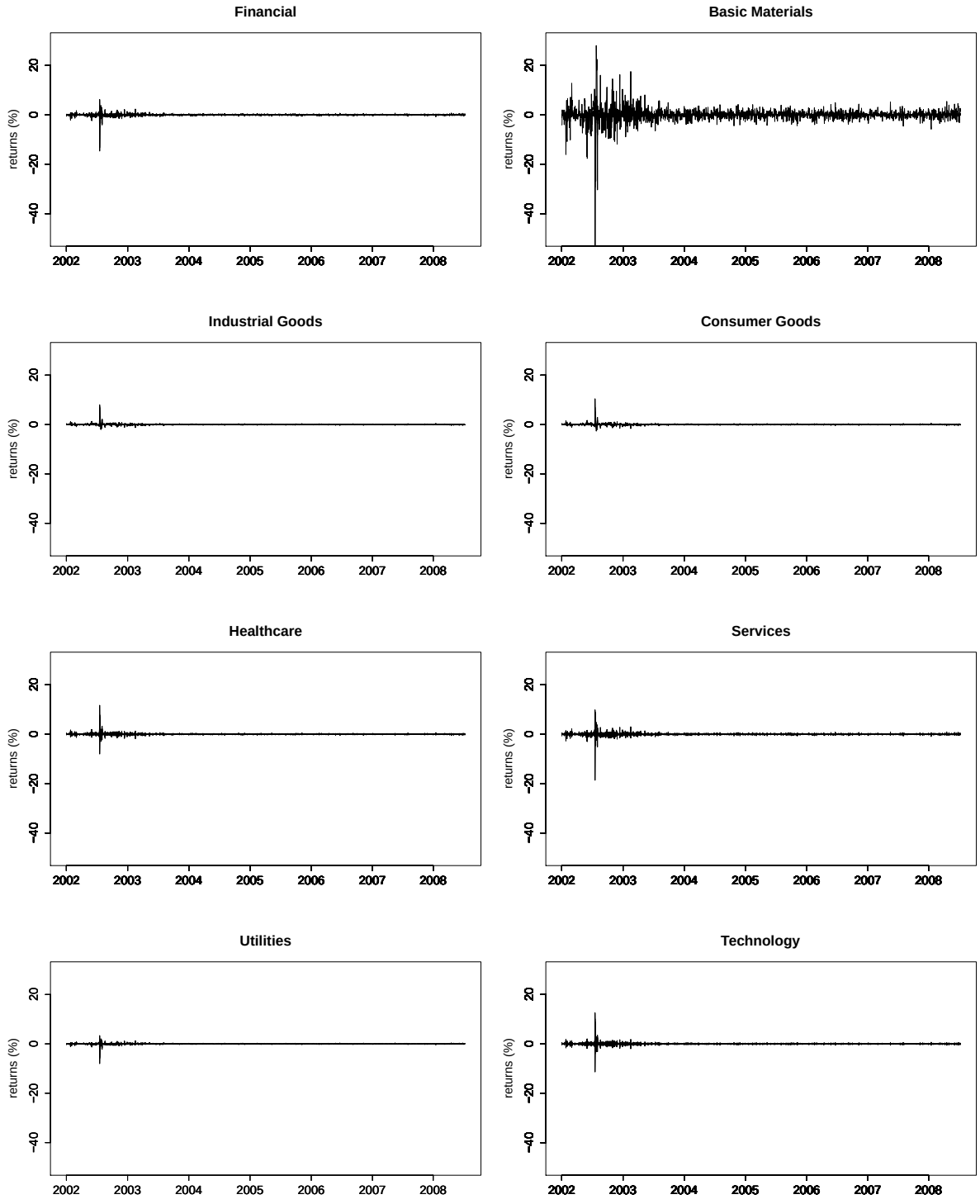
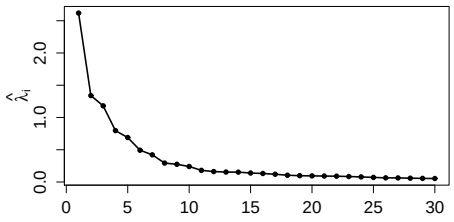
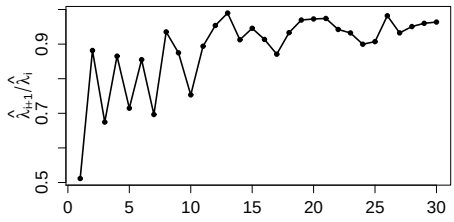


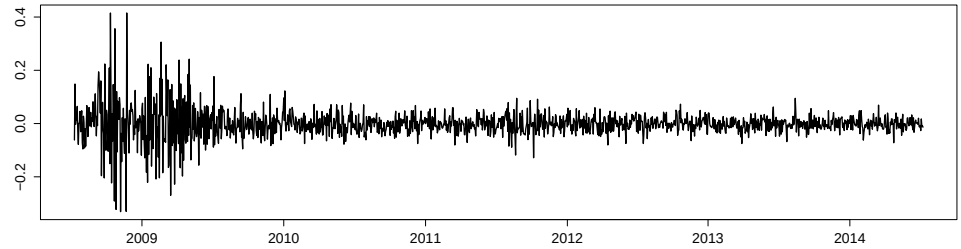
Figure 8: The estimated latent part $\mathbf{A}\mathbf{x}_t$ across different sectors.



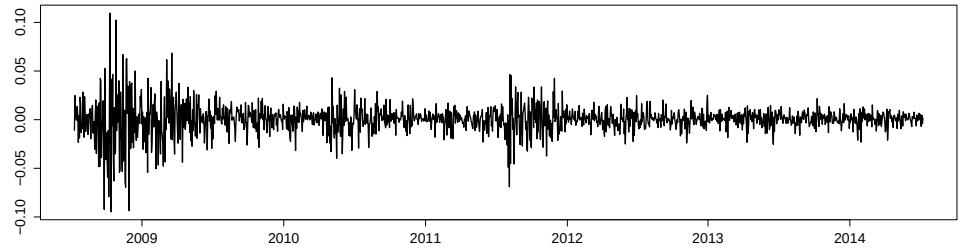
(a) The eigenvalues of $\widehat{\mathbf{M}}$



(b) Ratio of the eigenvalues of $\widehat{\mathbf{M}}$



(c) The estimated latent factor



(d) S&P 500 returns

Figure 9: The estimated eigenvalues (multiplied by 10^6), the ratio of eigenvalues, the estimated latent factor and the S&P500 returns in 14 July 2008 – 11 July 2014.

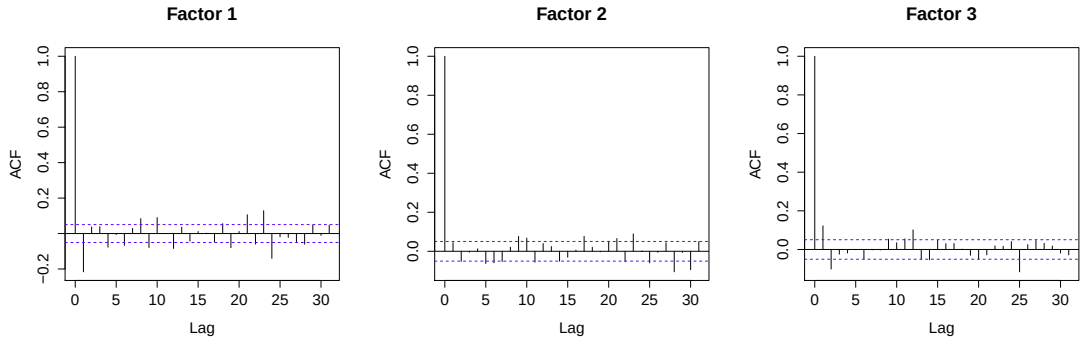


Figure 10: The ACFs of the first three estimated factors.

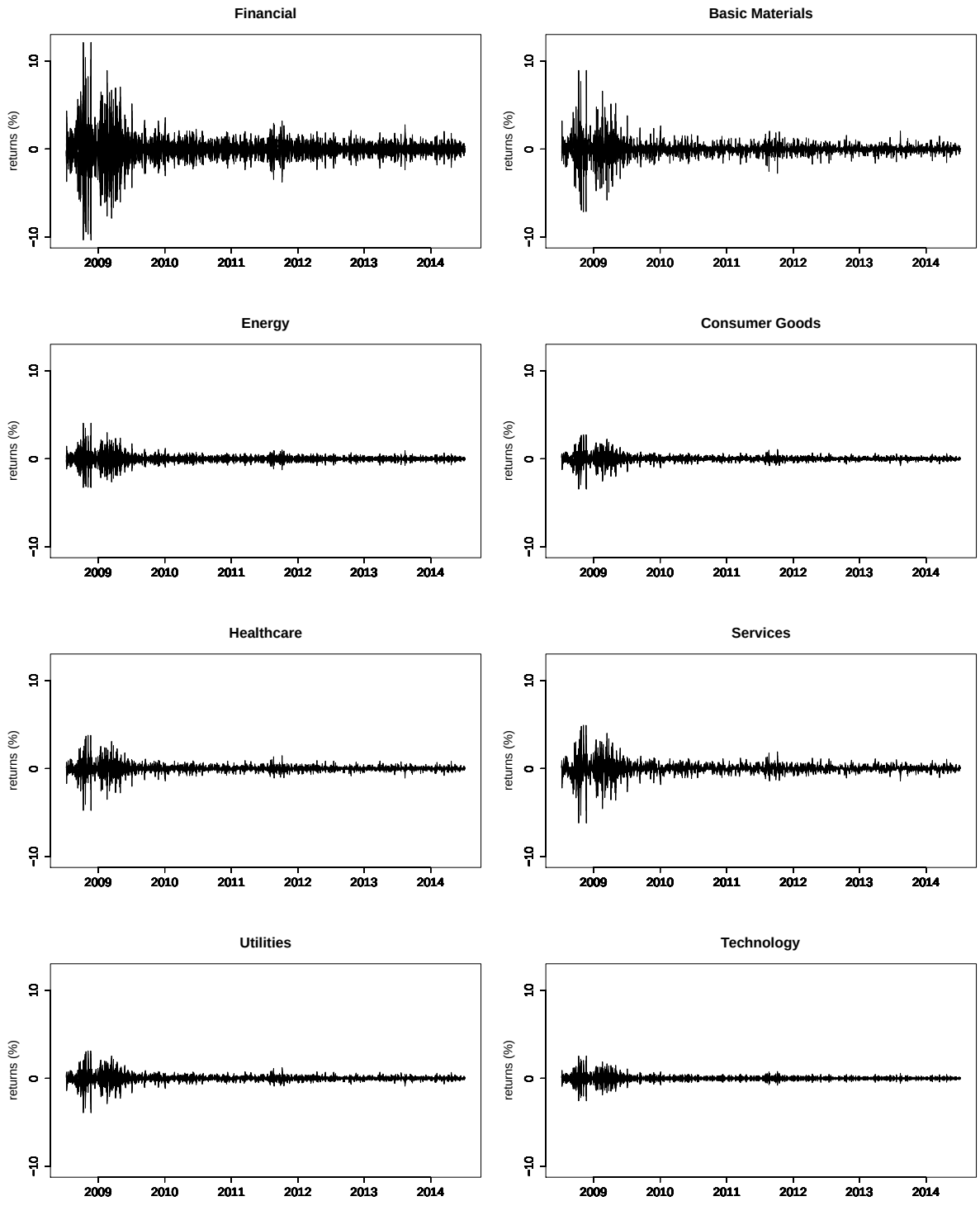


Figure 11: The estimated latent part $\mathbf{A}\mathbf{x}_t$ across different sectors.

Proof: For any $i, j = 1, \ldots, m$, by Cauchy-Schwarz inequality and Davydov inequality,

$$\begin{aligned} & E\bigg[\bigg|\frac{1}{T}\sum_{t=1}^{T}\{z_{i,t}z_{j,t} - E(z_{i,t}z_{j,t})\}\bigg|^2\bigg] \\ &= \frac{1}{T^2}\sum_{t=1}^{T} E[\{z_{i,t}z_{j,t} - E(z_{i,t}z_{j,t})\}^2] \\ &\quad + \frac{1}{T^2}\sum_{t_1\neq t_2} E[\{z_{i,t_1}z_{j,t_1} - E(z_{i,t_1}z_{j,t_1})\}\{z_{i,t_2}z_{j,t_2} - E(z_{i,t_2}z_{j,t_2})\}] \\ &\leq \frac{C}{T} + \frac{C}{T^2}\sum_{t_1\neq t_2}\alpha(|t_1 - t_2|)^{1-2/\gamma} \leq \frac{C}{T} + \frac{C}{T}\sum_{u=1}^{T}\alpha(u)^{1-2/\gamma}. \end{aligned} \tag{18}$$

Then, $E\{\|T^{-1}\sum_{t=1}^{T}\{\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}} - E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\}\|_F^2\} = O(m^2T^{-1})$ which implies the result. □

Proof of Proposition 2.1: Note that $(\widehat{\mathbf{D}} - \mathbf{D})^{\mathrm{T}} = (T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})^{-1}(T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\boldsymbol{\eta}_t^{\mathrm{T}})$ and $\lambda_{\min}(T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})$ is bounded away from zero with probability approaching one, which is implied by Condition 2.3 and Lemma 6.1, then $\|\widehat{\mathbf{D}} - \mathbf{D}\|_F = O_p(\|T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\boldsymbol{\eta}_t^{\mathrm{T}}\|_F)$. For each $i = 1, \ldots, m$ and $j = 1, \ldots, p$, from $\mathrm{cov}(\mathbf{z}_t, \boldsymbol{\eta}_t) = \mathbf{0}$ and similar to (18), we can obtain $E\{(T^{-1}\sum_{t=1}^{T}z_{i,t}\eta_{j,t})^2\} \leq CT^{-1}$. Then, $E\{\|T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\boldsymbol{\eta}_t^{\mathrm{T}}\|_F^2\} = O(pT^{-1})$. Hence, $\|\widehat{\mathbf{D}} - \mathbf{D}\|_F = O_p(p^{1/2}T^{-1/2})$. □

Lemma 6.2. *Under Conditions 2.1-2.2, if $k = o(T)$, then*

$$\|\widehat{\boldsymbol{\Sigma}}_\eta(k) - \boldsymbol{\Sigma}_\eta(k)\|_F = \|\mathbf{D} - \widehat{\mathbf{D}}\|_F^2 J_{1,k} + \|\mathbf{D} - \widehat{\mathbf{D}}\|_F J_{2,k} + J_{3,k}$$

where $E(J_{1,k}^2) \leq Ckm^2(T-k)^{-1} + Cm^2\alpha(k)^{2-4/\gamma}$, $E(J_{2,k}^2) \leq Ckpm(T-k)^{-1} + Cpm\alpha(k)^{2-4/\gamma}$ and $E(J_{3,k}^2) \leq Ckp^2(T-k)^{-1}$.

Proof: For each $k = o(T)$,

$$\begin{aligned} \widehat{\boldsymbol{\Sigma}}_\eta(k) - \boldsymbol{\Sigma}_\eta(k) &= \frac{1}{T-k}\sum_{t=1}^{T-k}(\widehat{\boldsymbol{\eta}}_{t+k}\widehat{\boldsymbol{\eta}}_t^{\mathrm{T}} - \boldsymbol{\eta}_{t+k}\boldsymbol{\eta}_t^{\mathrm{T}}) + \frac{1}{T-k}\sum_{t=1}^{T-k}\{\boldsymbol{\eta}_{t+k}\boldsymbol{\eta}_t^{\mathrm{T}} - E(\boldsymbol{\eta}_{t+k}\boldsymbol{\eta}_t^{\mathrm{T}})\} \\ &\quad + \bar{\boldsymbol{\eta}}\bar{\boldsymbol{\eta}}^{\mathrm{T}} - \frac{1}{T-k}\sum_{t=1}^{T-k}\widehat{\boldsymbol{\eta}}_{t+k}\bar{\boldsymbol{\eta}}^{\mathrm{T}} - \frac{1}{T-k}\sum_{t=1}^{T-k}\bar{\boldsymbol{\eta}}\widehat{\boldsymbol{\eta}}_t^{\mathrm{T}} \\ &= I_{1,k} + I_{2,k} + I_{3,k} + I_{4,k} + I_{5,k}. \end{aligned}$$

As

$$\begin{aligned} I_{1,k} &= (\mathbf{D} - \widehat{\mathbf{D}})\bigg(\frac{1}{T-k}\sum_{t=1}^{T-k}\mathbf{z}_{t+k}\mathbf{z}_t^{\mathrm{T}}\bigg)(\mathbf{D} - \widehat{\mathbf{D}})^{\mathrm{T}} + (\mathbf{D} - \widehat{\mathbf{D}})\bigg(\frac{1}{T-k}\sum_{t=1}^{T-k}\mathbf{z}_{t+k}\boldsymbol{\eta}_t^{\mathrm{T}}\bigg) \\ &\quad + \bigg(\frac{1}{T-k}\sum_{t=1}^{T-k}\boldsymbol{\eta}_{t+k}\mathbf{z}_t^{\mathrm{T}}\bigg)(\mathbf{D} - \widehat{\mathbf{D}})^{\mathrm{T}}, \end{aligned}$$

then

$$\|I_{1,k}\|_F \le \|\mathbf{D}-\widehat{\mathbf{D}}\|_F^2\left\|\frac{1}{T-k}\sum_{t=1}^{T-k}\mathbf{z}_{t+k}\mathbf{z}_t^{\mathrm{T}}\right\|_F + \|\mathbf{D}-\widehat{\mathbf{D}}\|_F\left\|\frac{1}{T-k}\sum_{t=1}^{T-k}\mathbf{z}_{t+k}\boldsymbol{\eta}_t^{\mathrm{T}}\right\|_F$$
$$+ \|\mathbf{D}-\widehat{\mathbf{D}}\|_F\left\|\frac{1}{T-k}\sum_{t=1}^{T-k}\boldsymbol{\eta}_{t+k}\mathbf{z}_t^{\mathrm{T}}\right\|_F.$$

For any $i,j=1,\ldots,m$,

$$E\left\{\left(\frac{1}{T-k}\sum_{t=1}^{T-k}z_{i,t+k}z_{j,t}\right)^2\right\} \le 2E\left(\left[\frac{1}{T-k}\sum_{t=1}^{T-k}\{z_{i,t+k}z_{j,t}-E(z_{i,t+k}z_{j,t})\}\right]^2\right)+2\max_t\{E(z_{i,t+k}z_{j,t})\}^2.$$

By Cauchy-Schwarz inequality and Davydov inequality,

$$\begin{aligned}
&E\left(\left[\frac{1}{T-k}\sum_{t=1}^{T-k}\{z_{i,t+k}z_{j,t}-E(z_{i,t+k}z_{j,t})\}\right]^2\right)\\
&=\frac{1}{(T-k)^2}\sum_{t=1}^{T-k}E[\{z_{i,t+k}z_{j,t}-E(z_{i,t+k}z_{j,t})\}^2]\\
&\quad+\frac{1}{(T-k)^2}\sum_{t_1\neq t_2}E[\{z_{i,t_1+k}z_{j,t_1}-E(z_{i,t_1+k}z_{j,t_1})\}\{z_{i,t_2+k}z_{j,t_2}-E(z_{i,t_2+k}z_{j,t_2})\}]\\
&\le\frac{C}{T-k}+\frac{Ck}{T-k}+\frac{Ck(k-1)}{(T-k)^2}+\frac{C}{T-k}\sum_{u=1}^{T-2k-1}\alpha(u)^{1-2/\gamma}.
\end{aligned}\tag{19}$$

and $\{E(z_{i,t+k}z_{j,t})\}^2 \le C\alpha(k)^{2-4/\gamma}$. Then, $E[\{(T-k)^{-1}\sum_{t=1}^{T-k}z_{i,t+k}z_{j,t}\}^2] \le Ck(T-k)^{-1}+C\alpha(k)^{2-4/\gamma}$. Thus, $E\{\|(T-k)^{-1}\sum_{t=1}^{T-k}\mathbf{z}_{t+k}\mathbf{z}_t^{\mathrm{T}}\|_F^2\} \le Ckm^2(T-k)^{-1}+Cm^2\alpha(k)^{2-4/\gamma}$. By the same argument, we can obtain $E\{\|(T-k)^{-1}\sum_{t=1}^{T-k}\mathbf{z}_{t+k}\boldsymbol{\eta}_t^{\mathrm{T}}\|_F^2\} \le Ckpm(T-k)^{-1}+Cpm\alpha(k)^{2-4/\gamma}$ and $E\{\|(T-k)^{-1}\sum_{t=1}^{T-k}\boldsymbol{\eta}_{t+k}\mathbf{z}_t^{\mathrm{T}}\|_F^2\} \le Ckpm(T-k)^{-1}+Cpm\alpha(k)^{2-4/\gamma}$. Hence, $\|I_{1,k}\|_F = \|\mathbf{D}-\widehat{\mathbf{D}}\|_F^2 J_{1,k}+\|\mathbf{D}-\widehat{\mathbf{D}}\|_F J_{2,k}$ where $E(J_{1,k}^2) \le Ckm^2(T-k)^{-1}+Cm^2\alpha(k)^{2-4/\gamma}$ and $E(J_{2,k}^2) \le Ckpm(T-k)^{-1}+Cpm\alpha(k)^{2-4/\gamma}$. On the other hand, similar to (19), we can obtain $E(\|I_{2,k}\|_F^2) \le Ckp^2(T-k)^{-1}$. For $I_{3,k}$, we have $E(\|I_{3,k}\|_F^2) \le E(\|\bar{\boldsymbol{\eta}}\|_2^4)$. By Jensen inequality and Davydov inequality, $E(\|I_{3,k}\|_F^2) \le Cp^2T^{-1}$. Following the same way, we have both $E(\|I_{4,k}\|_F^2)$ and $E(\|I_{5,k}\|_F^2)$ can be bounded by $Ckp^2(T-k)^{-1}$. Hence, we complete the proof. □

Lemma 6.3. *Under Condition 2.4, for $k=1,\ldots,\bar{k}$,*

$$\|\boldsymbol{\Sigma}_\eta(k)\|_2 \le Cp^{1-\delta}+C\kappa_2.$$

Proof: Note that $\boldsymbol{\Sigma}_\eta(k)=\mathbf{A}\boldsymbol{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\boldsymbol{\Sigma}_{x\varepsilon}(k)$, then $\|\boldsymbol{\Sigma}_\eta(k)\|_2 \le \|\mathbf{A}\|_2^2\|\boldsymbol{\Sigma}_x(k)\|_2+\|\boldsymbol{\Sigma}_{x\varepsilon}(k)\|_2$. From Condition 2.4, we complete the proof. □

Lemma 6.4. *Under Conditions 2.1-2.4,*

$$\|\widehat{\mathbf{M}}-\mathbf{M}\|_2 = O_p\{(p^{1-\delta}+\kappa_2)pT^{-1/2}+p^2T^{-1}\}.$$

Proof: Note that

$$\|\widehat{\mathbf{M}}-\mathbf{M}\|_2 \leq \sum_{k=1}^{\bar{k}}\|\widehat{\mathbf{\Sigma}}_\eta(k)-\mathbf{\Sigma}_\eta(k)\|_2^2+2\sum_{k=1}^{\bar{k}}\|\mathbf{\Sigma}_\eta(k)\|_2\|\widehat{\mathbf{\Sigma}}_\eta(k)-\mathbf{\Sigma}_\eta(k)\|_2 = I_1+I_2.$$

By Lemmas 6.2 and 6.3, we can obtain

$$I_1 \leq 3\|\mathbf{D}-\widehat{\mathbf{D}}\|_F^4\sum_{k=1}^{\bar{k}}J_{1,k}^2+3\|\mathbf{D}-\widehat{\mathbf{D}}\|_F^2\sum_{k=1}^{\bar{k}}J_{2,k}^2+3\sum_{k=1}^{\bar{k}}J_{3,k}^2 = O_p(p^2T^{-1})$$

and

$$\begin{aligned} I_2 &\leq 2\bigg\{\|\mathbf{D}-\widehat{\mathbf{D}}\|_F^2\sum_{k=1}^{\bar{k}}J_{1,k}+\|\mathbf{D}-\widehat{\mathbf{D}}\|_F\sum_{k=1}^{\bar{k}}J_{2,k}+\sum_{k=1}^{\bar{k}}J_{3,k}\bigg\}\sup_{1\leq k\leq\bar{k}}\|\mathbf{\Sigma}_\eta(k)\|_2 \\ &= O_p\{(p^{1-\delta}+\kappa_2)pT^{-1/2}\}. \end{aligned}$$

Hence, we complete the proof. □

Lemma 6.5. *Under Condition 2.4,*

$$\lambda_r(\mathbf{M}) \geq \begin{cases} Cp^{2(1-\delta)}, & \text{if } \kappa_2 = o(p^{1-\delta}); \\ C\kappa_1^2, & \text{if } p^{1-\delta} = o(\kappa_1). \end{cases}$$

Proof: From (5), we know

$$\lambda_r(\mathbf{M}) = \lambda_{\min}\bigg[\sum_{k=1}^{\bar{k}}\{\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\mathbf{\Sigma}_{x\varepsilon}(k)\}\{\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\mathbf{\Sigma}_{x\varepsilon}(k)\}^{\mathrm{T}}\bigg].$$

For each $k=1,\ldots,\bar{k}$,

$$\begin{aligned} &\lambda_{\min}[\{\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\mathbf{\Sigma}_{x\varepsilon}(k)\}\{\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\mathbf{\Sigma}_{x\varepsilon}(k)\}^{\mathrm{T}}] \\ &\asymp \begin{cases} \lambda_{\min}\{\mathbf{\Sigma}_x(k)\mathbf{\Sigma}_x(k)^{\mathrm{T}}\}, & \text{if } \lambda_{\max}\{\mathbf{\Sigma}_{x\varepsilon}(k)\mathbf{\Sigma}_{x\varepsilon}(k)^{\mathrm{T}}\} = o(\lambda_{\min}\{\mathbf{\Sigma}_x(k)\mathbf{\Sigma}_x(k)^{\mathrm{T}}\}); \\ \lambda_{\min}\{\mathbf{\Sigma}_{x\varepsilon}(k)\mathbf{\Sigma}_{x\varepsilon}(k)^{\mathrm{T}}\}, & \text{if } \lambda_{\max}\{\mathbf{\Sigma}_x(k)\mathbf{\Sigma}_x(k)^{\mathrm{T}}\} = o(\lambda_{\min}\{\mathbf{\Sigma}_{x\varepsilon}(k)\mathbf{\Sigma}_{x\varepsilon}(k)^{\mathrm{T}}\}). \end{cases} \end{aligned}$$

Notice Condition 2.4, then

$$\lambda_{\min}[\{\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\mathbf{\Sigma}_{x\varepsilon}(k)\}\{\mathbf{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}+\mathbf{\Sigma}_{x\varepsilon}(k)\}^{\mathrm{T}}] \geq \begin{cases} Cp^{2(1-\delta)}, & \text{if } \kappa_2 = o(p^{1-\delta}); \\ C\kappa_1^2, & \text{if } p^{1-\delta} = o(\kappa_1). \end{cases}$$

Hence, we complete the proof. □

Proof of Theorem 2.1: By Lemma 6.5, $\|\widehat{\mathbf{M}}-\mathbf{M}\|_2 = o_p\{\lambda_r(\mathbf{M})\}$ provided that either case (i) $\kappa_2 = o(p^{1-\delta})$ and $p^{2\delta}T^{-1} = o(1)$ or (ii) $p^{1-\delta} = o(\kappa_1)$ and $\kappa_1^{-2}\kappa_2 pT^{-1/2} = o(1)$ hold. By Lemma 3 of Lam et al. (2011), and using the same argument of the proof of Theorem 1 in their paper,

$$\|\widehat{\mathbf{A}}-\mathbf{A}\|_2 = \begin{cases} O_p(p^{\delta}T^{-1/2}), & \text{if } \kappa_2 = o(p^{1-\delta}) \text{ and } p^{2\delta}T^{-1} = o(1); \\ O_p(\kappa_1^{-2}\kappa_2 pT^{-1/2}), & \text{if } p^{1-\delta} = o(\kappa_1) \text{ and } \kappa_1^{-2}\kappa_2 pT^{-1/2} = o(1). \end{cases}$$

Hence, we complete the proof. □

PROOF OF THEOREM 2.2: Note that

$$\begin{aligned}\widehat{\mathbf{A}}\widehat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_t &= \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}}\mathbf{A}\mathbf{x}_t + \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}}\boldsymbol{\varepsilon}_t - \mathbf{A}\mathbf{x}_t + \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}}(\widehat{\boldsymbol{\eta}}_t - \boldsymbol{\eta}_t) \\ &= (\widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}} - \mathbf{A}\mathbf{A}^{\mathrm{T}})\mathbf{A}\mathbf{x}_t + \widehat{\mathbf{A}}(\widehat{\mathbf{A}} - \mathbf{A})^{\mathrm{T}}\boldsymbol{\varepsilon}_t + \widehat{\mathbf{A}}\mathbf{A}^{\mathrm{T}}\boldsymbol{\varepsilon}_t + \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}}(\widehat{\boldsymbol{\eta}}_t - \boldsymbol{\eta}_t) \\ &= I_1 + I_2 + I_3 + I_4.\end{aligned}$$

For I_1, $\|I_1\|_2 \le 2\|\widehat{\mathbf{A}} - \mathbf{A}\|_2\|\mathbf{A}\mathbf{x}_t\|_2 \le O_p(p^{1/2}\|\widehat{\mathbf{A}} - \mathbf{A}\|_2)$. For I_2, $\|I_2\|_2 \le \|\widehat{\mathbf{A}} - \mathbf{A}\|_2\|\boldsymbol{\varepsilon}_t\|_2 = O_p(p^{1/2}\|\widehat{\mathbf{A}} - \mathbf{A}\|_2)$. For I_3, as $E(\|I_3\|_2^2) = \sum_{i=1}^r E\{(\mathbf{a}_i^{\mathrm{T}}\boldsymbol{\varepsilon}_t)^2\} \le r\lambda_{\max}(\boldsymbol{\Sigma}_\varepsilon)$, then $I_3 = O_p(1)$. For I_4, by Proposition 2.1, $\|I_4\|_2 \le \|\widehat{\mathbf{D}} - \mathbf{D}\|_2\|\mathbf{z}_t\|_2 = O_p(p^{1/2}T^{-1/2})$. Hence, $p^{-1/2}\|\widehat{\mathbf{A}}\widehat{\mathbf{x}}_t - \mathbf{A}\mathbf{x}_t\|_2 \le O_p(\|\widehat{\mathbf{A}} - \mathbf{A}\|_2 + p^{-1/2} + T^{-1/2})$. □

PROOF OF THEOREM 2.3: Let $\boldsymbol{\Sigma}_\zeta(k) = (T-k)^{-1}\sum_{t=1}^{T-k}\mathrm{cov}(\boldsymbol{\zeta}_{t+k}, \boldsymbol{\zeta}_t)$, then $\boldsymbol{\Sigma}_\zeta(k) = \mathbf{A}\boldsymbol{\Sigma}_x(k)\mathbf{A}^{\mathrm{T}}$. Note that

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} &= \mathrm{tr}\{\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}(I_r - \mathbf{A}^{\mathrm{T}}\widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}}\mathbf{A})\boldsymbol{\Sigma}_x(1)\} \\ &\ge \mathrm{tr}(I_r - \mathbf{A}^{\mathrm{T}}\widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}}\mathbf{A})\lambda_{\min}\{\boldsymbol{\Sigma}_x(1)\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}\} \\ &= r\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2\lambda_{\min}\{\boldsymbol{\Sigma}_x(1)\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}\}.\end{aligned}$$

By Condition 2.4,

$$\mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} \ge Cp^{2(1-\delta)}\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2.$$

At the same time,

$$\begin{aligned}&\mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} - \mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \mathbf{A}\mathbf{A}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} \\ &= \mathrm{tr}\{\mathbf{A}\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\boldsymbol{\Sigma}_x(1)\mathbf{A}^{\mathrm{T}}\} \\ &\le \lambda_{\max}\{\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\}\mathrm{tr}\{\boldsymbol{\Sigma}_x(1)\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}\} \\ &\le Cp^{2(1-\delta)}\|\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\|_2.\end{aligned}$$

Note that $\mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \mathbf{A}\mathbf{A}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} = 0$, then

$$\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2 \le C\|\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\|_2.$$

On the other hand, we have the following two inequality,

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} &\le r\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2\lambda_{\max}\{\boldsymbol{\Sigma}_x(1)\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}\} \\ &\le Cp^{2(1-\delta)}\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2\end{aligned}$$

and

$$\begin{aligned}\mathrm{tr}\{\boldsymbol{\Sigma}_\zeta(1)^{\mathrm{T}}(I_p - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\boldsymbol{\Sigma}_\zeta(1)\} &\ge \lambda_{\min}\{\boldsymbol{\Sigma}_x(1)\boldsymbol{\Sigma}_x(1)^{\mathrm{T}}\}\mathrm{tr}\{\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\} \\ &\ge Cp^{2(1-\delta)}\|\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\|_2.\end{aligned}$$

Hence,

$$\{D(\mathcal{M}(\widehat{\mathbf{A}}), \mathcal{M}(\mathbf{A}))\}^2 \asymp \|\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A}\|_2.$$

Note that

$$\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}} - \widehat{\mathbf{A}}\widehat{\mathbf{A}}^{\mathrm{T}})\mathbf{A} = -\mathbf{A}^{\mathrm{T}}(\mathbf{A} - \widehat{\mathbf{A}})(\mathbf{A} - \widehat{\mathbf{A}})^{\mathrm{T}}\mathbf{A} + (\mathbf{A} - \widehat{\mathbf{A}})^{\mathrm{T}}(\mathbf{A} - \widehat{\mathbf{A}}),$$

then we complete the proof. □

Proof of Theorem 2.4: As $(p^{1-\delta} + \kappa_2)pT^{-1/2}\log T \to 0$, then $\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = o_p\{\lambda_r(\mathbf{M})\}$. Then $\sup_{j=1,\ldots,p} |\widehat{\lambda}_j - \lambda_j(\mathbf{M})| \le \|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = o_p\{\lambda_r(\mathbf{M})\}$. For any $j < r$,

$$\frac{\widehat{\lambda}_{j+1} + (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T}{\widehat{\lambda}_j + (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T} \xrightarrow{p} C > 0.$$

For any $j > r$, note that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = o_p\{(p^{1-\delta} + \kappa_2)pT^{-1/2}\log T\}$ which implies that $|\widehat{\lambda}_j| = o_p\{(p^{1-\delta} + \kappa_2)pT^{-1/2}\log T\}$, then

$$\frac{\widehat{\lambda}_{j+1} + (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T}{\widehat{\lambda}_j + (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T} \xrightarrow{p} 1 > 0.$$

On the other hand,

$$\frac{\widehat{\lambda}_{r+1} + (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T}{\widehat{\lambda}_r + (p^{1-\delta} + \kappa_2)pT^{-1/2}\log T} \xrightarrow{p} 0.$$

Hence, the criterion implies a consistent estimator of r. □

Proof of Proposition 3.1: Following the proof of Lemma 6.1, $\|T^{-1}\sum_{t=1}^{T}\{\mathbf{z}_t\mathbf{w}_t^{\mathrm{T}} - E(\mathbf{z}_t\mathbf{w}_t^{\mathrm{T}})\}\|_F = O_p(m^{1/2}q^{1/2}T^{-1/2})$. Note that $\mathrm{rank}(\mathbf{R}) = m$ and Condition 3.1, it yields $\lambda_{\min}(T^{-1}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{w}_t^{\mathrm{T}}\mathbf{R}^{\mathrm{T}})$ is bounded away from zero with probability approaching one. Hence, following the proof of Proposition 2.1, we can obtain the result. □

Proof of Proposition 4.1: For each $i = 1, \ldots, p$,

$$\widehat{\mathbf{d}}_i - \mathbf{d}_i = \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}\eta_{i,t}\mathbf{z}_t\right) + \left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T}e_{i,t}\mathbf{z}_t\right).$$

Then,

$$\|\widehat{\mathbf{d}}_i - \mathbf{d}_i\|_2\lambda_{\min}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right) \le \left\|\frac{1}{T}\sum_{t=1}^{T}\eta_{i,t}\mathbf{z}_t\right\|_2 + \left\|\frac{1}{T}\sum_{t=1}^{T}e_{i,t}\mathbf{z}_t\right\|_2.$$

Note that $E(\boldsymbol{\zeta}_t|\mathbf{u}_t) = \mathbf{0}$ and $E(\boldsymbol{\varepsilon}_t|\mathbf{u}_t) = \mathbf{0}$, we have $\|T^{-1}\sum_{t=1}^{T}\eta_{i,t}\mathbf{z}_t\|_2 = O_p(m^{1/2}T^{-1/2})$ and $\|T^{-1}\sum_{t=1}^{T}e_{i,t}\mathbf{z}_t\|_2 = \|T^{-1}\sum_{t=1}^{T}E(e_{i,t}\mathbf{z}_t)\|_2 + O_p(m^{1/2}T^{-1/2})$, where $O_p(m^{1/2}T^{-1/2})$s are uniformly for $i = 1, \ldots, p$. On the other hand, $\|E(e_{i,t}\mathbf{z}_t)\|_2^2 = O(m^{1-2\lambda})$. Thus, we complete the proof. □

Proof of Theorem 4.1: Let $\mathbf{z} = (l_1(\mathbf{u}), \ldots, l_m(\mathbf{u}))^{\mathrm{T}}$. For each $i = 1, \ldots, p$,

$$\widehat{g}_i(\mathbf{u}) - g_i(\mathbf{u}) = \mathbf{z}^{\mathrm{T}}\left(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\right)^{-1}\left\{\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t(e_{i,t} + \zeta_{i,t} + \varepsilon_{i,t})\right\} - e_i$$

where $g_i(\mathbf{u}) = \mathbf{d}_i^{\mathrm{T}}\mathbf{z} + e_i$. Hence,

$$\int_{\mathbf{u}\in\mathcal{U}} |\widehat{g}_i(\mathbf{u}) - g_i(\mathbf{u})|^2 \, d\mathbf{u} \leq 2\Big\{\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t^{\mathrm{T}}(e_{i,t}+\zeta_{i,t}+\varepsilon_{i,t})\Big\}\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\Big)^{-1}$$
$$\times\Big(\int_{\mathbf{u}\in\mathcal{U}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\,d\mathbf{u}\Big)\Big(\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\Big)^{-1}\Big\{\frac{1}{T}\sum_{t=1}^{T}\mathbf{z}_t(e_{i,t}+\zeta_{i,t}+\varepsilon_{i,t})\Big\}$$
$$+ Cm^{-2\lambda}.$$

Let $p(\mathbf{u}_t)$ be the density function of $\mathbf{u}_t$ and pick $\mathbf{v}^{\mathrm{T}}$ such that $\lambda_{\max}(\int_{\mathbf{u}\in\mathcal{U}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\,d\mathbf{u}) = \int_{\mathbf{u}\in\mathcal{U}}\mathbf{v}^{\mathrm{T}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\mathbf{v}\,d\mathbf{u}$, by Condition 4.1,

$$\mathbf{v}^{\mathrm{T}}E(\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}})\mathbf{v} = \int_{\mathbf{u}_t\in\mathcal{U}}\mathbf{v}^{\mathrm{T}}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\mathbf{v}p(\mathbf{u}_t)\,d\mathbf{u}_t \geq C\int_{\mathbf{u}_t\in\mathcal{U}}\mathbf{v}^{\mathrm{T}}\mathbf{z}_t\mathbf{z}_t^{\mathrm{T}}\mathbf{v}\,d\mathbf{u}_t = C\lambda_{\max}\Big(\int_{\mathbf{u}\in\mathcal{U}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\,d\mathbf{u}\Big).$$

From Condition 4.3, we know $\lambda_{\max}(\int_{\mathbf{u}\in\mathcal{U}}\mathbf{z}\mathbf{z}^{\mathrm{T}}\,d\mathbf{u}) \leq C$ which implies

$$\int_{\mathbf{u}\in\mathcal{U}} |\widehat{g}_i(\mathbf{u}) - g_i(\mathbf{u})|^2 \, d\mathbf{u} \leq O_p(mT^{-1}) + O(m^{-2\lambda}).$$

The terms $O_p(mT^{-1})$ and $O(m^{-2\lambda})$ are uniformly for $i = 1, \ldots, p$, thus we complete the proof. □

Lemma 6.6. *For nonlinear regression model (14), under Conditions 2.1-2.2, 4.2 and 4.5, if $k = o(T)$, then*

$$\|\widehat{\mathbf{\Sigma}}_\eta(k) - \mathbf{\Sigma}_\eta(k)\|_F = \|\mathbf{D} - \widehat{\mathbf{D}}\|_F^2 J_{1,k} + \|\mathbf{D} - \widehat{\mathbf{D}}\|_F J_{2,k} + J_{3,k}$$

where $E(J_{1,k}^2) \leq Ckm^2(T-k)^{-1} + Cm^2\alpha(k)^{2-4/\gamma}$, $E(J_{2,k}^2) \leq Ckpm(T-k)^{-1} + Cpm\alpha(k)^{2-4/\gamma}$ and $E(J_{3,k}^2) \leq Ckp^2(T-k)^{-1} + Cp^2m^{-2\lambda}\alpha(k)^{2-4/\gamma}$.

Proof: Noting $\sup_t \|\mathbf{e}_t\|_\infty = O(m^{-\lambda})$, similar to Lemma 6.2, we can obtain the result. □

Lemma 6.7. *Under Conditions 2.1-2.2, 2.4, 4.2-4.5, if $mT^{-1/2} = o(1)$, $\bar{k}T^{-1/2} = o(1)$ and $\lambda \geq 1$, then*

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = O_p\{(p^{1-\delta}+\kappa_2)p[(\bar{k}^{3/2}+m)T^{-1/2} + m^{1-\lambda}]\} + O_p\{p^2[(\bar{k}^2+m^2)T^{-1} + m^{2-2\lambda}]\}.$$

Proof: Note that $\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 \leq \sum_{k=1}^{\bar{k}}\{\|\widehat{\mathbf{\Sigma}}_\eta(k) - \mathbf{\Sigma}_\eta(k)\|_2^2 + 2\|\mathbf{\Sigma}_\eta(k)\|_2\|\widehat{\mathbf{\Sigma}}_\eta(k) - \mathbf{\Sigma}_\eta(k)\|_2\}$. By Lemma 6.6, we complete the proof. □

Proof of Theorem 4.2: Note that $m = O(T^{1/(2\lambda+1)})$, then

$$\|\widehat{\mathbf{M}} - \mathbf{M}\|_2 = O_p\{(p^{1-\delta}+\kappa_2)p(\bar{k}^{3/2}T^{-1/2} + T^{(1-\lambda)(2\lambda+1)}) + p^2(\bar{k}^2T^{-1} + T^{(2-2\lambda)/(2\lambda+1)})\}.$$

Similar to the proof of Lemma 6.5, we have

$$\lambda_r(\mathbf{M}) \geq \begin{cases} C\bar{k}p^{2(1-\delta)}, & \text{if } \kappa_2 = o(p^{1-\delta}); \\ C\bar{k}\kappa_1^2, & \text{if } p^{1-\delta} = o(\kappa_1). \end{cases}$$

Then, by Lemma 6.7, $\|\widehat{\mathbf{M}}-\mathbf{M}\|_2 = o_p\{\lambda_r(\mathbf{M})\}$ provided that either (i) $\kappa_2 = o(p^{1-\delta})$ and $p^{2\delta}[\bar{k}T^{-1}+T^{(2-2\lambda)/(2\lambda+1)}] = o(1)$ or (ii) $p^{1-\delta} = o(\kappa_1)$ and $p^2\kappa_2^2\kappa_1^{-4}[\bar{k}T^{-1}+T^{(2-2\lambda)/(2\lambda+1)}] = o(1)$ hold. Using the same argument of the proof of Theorem 2.1, we obtain the result. □

Proof of Theorem 4.3: Following the arguments of the proof of Theorem 2.2, we can construct the result. □

References

Ai, C. and X. Chen (2003). Efficient estimation of models with conditional moment restrictions containing unknown functions, *Econometrica*, 71, 1795–1843.

Angrist, J. and A. Krueger (1991). Does Compulsory School Attendance Affect Schooling and Earnings, *Quarterly Journal of Economics*, 106, 979–1014.

Bai, J. and S. Ng (2002). Determining the number of factors in approximate factor models, *Econometrica*, 70, 191–221.

Bai, J. (2003). Inferential theory for factor models of large dimensions, *Econometrica*, 71, 135–171.

Bound, J., D. Jaeger, and R. Baker (1996). Problems with instrumental variables estimation when the correlation between instruments and the endogenous explanatory variable is weak, *Journal of the American Statistical Association*, 90, 443–450.

Caner, M. and Q. Fan (2012). The adaptive lasso method for instrumental variable selection, *Manuscript*.

Chang, J., B. Guo, and Q. Yao (2014). Segmenting multiple time series by contemporaneous linear transformation, *Manuscript*.

Davis, R.A., P. Zhang, and T. Zheng (2012). Sparse vector autoregressive modelling. *arXiv:1207.0520v1*.

Donald, S. G. and W. Newey (2001). Choosing the number of instruments, *Econometrica*, 69, 1161–1191.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2000). The generalized dynamic-factor model: identification and estimation, *The Review of Economics and Statistics*, 82, 540–554.

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). The generalized dynamic factor model: One-sided estimation and forecasting, *Journal of the American Statistical Association*, 100, 830–840.

Hahn, J. and J. Hausman (2002). A new specification test for the validity of instrumental variables, *Econometrica*, 70, 163–189.

Hallin, M. and R. Liska (2007). Determining the number of factors in the general dynamic factor model, *Journal of the American Statistical Association*, 102, 603–617.

Jakeman, A. J., L. P. Steele, and P. C. Young (1980). Instrumental variable algorithms for multiple input systems described by multiple transfer functions, *IEEE Transactions on Systems, Man, and Cybernetics*, 10, 593-602.

Jones, C. and G. Kaul (1996). Oil and the Stock Markets, *Journal of Finance*, 51, 463–491.

Kilian, L. and Park C. (2009). The impact of oil price shocks on the U.S. stock market, *International Economic Review*, 50, 1267–1287.

Lam, C., Q. Yao, and N. Bathia (2011). Estimation of latent factors for high-dimensional time series, *Biometrika*, 98, 901–918.

Lam, C. and Q. Yao (2012). Factor modeling for high-dimensional time series: inference for the number of factors, *The Annals of Statistics*, 40, 694–726.

Lütkepohl, H. (2006). *New Introduction to Multiple Time Series Analysis*, Springer, Berlin.

Morimune, K. (1983). Approximate distributions of k-class estimators when the degree of overidentifiablity is large compared with the sample size, *Econometrica*, 51, 821–841.

Pan, J. and Q. Yao (2008). Modelling multiple time series via common factors, *Biometrika*, 95, 365–379.

Pesaran, M. H. and E. Tosetti (2011). Large panels with common factors and spatial correlation, *Journal of Econometrics*, 161, 182-202.

Shojaie, A. and G. Michailidis (2010). Discovering graphical Granger causality using the truncated lasso penalty, *Bioinformatics*, 26, 517-523.

Song, S. and P. J. Bickel (2011). Large vector auto regressions, *arXiv:1106.3519*.

Stock, J. H. and M. W. Watson (2005). Implications of dynamic factor models for VAR analysis. Available at *www.nber.org/papers/w11467*.

Stone, C. (1985). Additive Regression and Other Nonparametric Models, *The Annals of Statistics*, 13, 689–705.

Tiao, G. C. and R. S. Tsay (1989). Model specification in multivariate time series (with discussions), *Journal of the Royal Statistical Society: Series B*, 51, 157-213.

Xia, Q., W. Xu, and L. Zhu (2013). Factor modelling of multivariate volatilities for nonstationary time series, *Manuscript*.