Error-Correction Factor Models*

Yundong Tu† Qiwei Yao‡,† Rongmao Zhang⋆

†Guanghua School of Management and Center for Statistical Science,
Peking University, Beijing, 100871, China

‡Department of Statistics, London School of Economics, London, WC2A 2AE, U.K.

⋆School of Mathematics, Zhejiang University, Hangzhou, 310058, China

yundong.tu@gsm.pku.edu.cn q.yao@lse.ac.uk rmzhang@zju.edu.cn

8 August 2015

Abstract

Cointegration inference is often built on the correct specification for the short-run dynamic vector autoregression. However, this specification is unknown in prior. A too small lag length leads to erroneous inference due to size distortions, while using too many lags leads to dramatic increase of the number of parameters, especially when the dimension of time series is high. In this paper, we develop a new methodology which adds an error correction term for long-run equilibrium to a latent factor model for modeling short-run dynamic relationship. Two eigenanalysis based methods for estimating, respectively, cointegration and latent factor process consists of the cornerstones of the inference. The proposed error correction factor model does not require to specify the short-run dynamics explicitly, and increases the predictability over a pure factor model. Asymptotic properties of the proposed methods are established when the dimension of the time series is either fixed or diverges slowly as the length of time series goes to infinity. Illustration with both simulated and real data sets is also reported.

KEYWORDS: α-mixing, cointegration, eigenanalysis, nonstationary processes, weak stationarity, vector time series.

*Partially supported by NSFC grants 71301004 and 71472007 (YT), EPSRC grant EP/L01226X/1 (QY), and NSFC grants 11371318 (RZ).

1 Introduction

Cointegration refers to the phenomenon that there exists a long-run equilibrium among several distinct nonstationary series, as illustrated in, for example, Box and Tiao (1977). Since the seminal work of Granger (1981), Granger and Weiss (1983) and Engle and Granger (1987), it has attracted increasing attention in econometrics and statistics. An excellent survey on the early developments of cointegration can be found in Johansen (1995).

Up to present, considerable effort has been devoted to the inference on the long-run trend (cointegration) restrictions in vector autoregression (VAR); see, among others, Engle and Granger (1987), Johansen (1991), Phillips (1991) for estimation and testing, and Engle and Yoo (1987), Lin and Tsay (1996) for forecasting. As shown in Engle and Granger (1987), VAR with cointegration restrictions can be represented as a vector error correction model (VECM) which reflects the correction on the long-run relationship by short-run dynamics. One of the remarkable features of VECM is that it identifies clearly the gain in prediction from using the cointegrated variables over the standard ARIMA approach, as noted by Engle and Yoo (1987), Lin and Tsay (1996) and Peña and Poncela (2004). However, it is a prerequisite to specify a finite autoregressive order for the short-run dynamic before the inference can be carried out on the cointegration part of the model. In many applications, using different orders for the VAR results in different conclusions on the cointegration. Especially when the VAR order is under-specified or the process lies outside the VAR class, the optimal inference on the unknown cointegration will lose validity (Hualde and Robinson, 2010). To overcome this shortcoming, information criteria such as AIC, BIC and HQIC have been applied to determine both the autoregressive order and the cointegration rank. See, for example, Chao and Phillips (1999) and Athanssopoulos, et al. (2011). While appealing for practitioners, all these methods are nevertheless subject to pre-test biases and post model selection inferential errors (Liao and Phillips, 2015).

Relative to considerable effort on long-run restriction, one may argue that the importance of short-run restrictions has not received due attention in cointegrated literature. On the other hand, common cyclical movements exist extensively in macroeconomics. For example, Engle and Kozicki (1993) found common international cycles in GNP data for OECD countries. Issler and Vahid (2001) reported the common cycles for macroeconomic aggregates and sectoral and regional outputs in US. It has been shown that using (short-run) rank restrictions in stationary VAR can improve short-term forecasting ability, as documented by Ahn and Reinsel (1988), Vahid and Isser (2002) and Athanasopoulos and Vahid (2008), Athanasopoulos et al (2011). Hence it is reasonable to expect that imposing appropriate short-run structures will improve the model performance in cointegrated systems. Note that Athanasopoulos et al (2011) recognized the factor structure in

the short-run dynamics, but did not utilized it in their sequential inference procedure.

When the dimension of time series is high, VAR models suffer from having too many parameters even with some imposed rank restrictions. Furthermore most the classical inference methods for cointegration, including Johansen's likelihood method, will not work or not work effectively. See the numerical studies reported in Gonzalo and Pitarakis (1994) and Ho and Sørensen (1996). Although high dimensional problems exist extensively in macroeconomic and financial data, the development in both theory and methodology in the context of cointegration is still in its infancy.

We propose in this paper an error correction factor model which is designed for catching high-dimensional linear dynamical structures exhibiting cointegration in a parsimonious and robust fashion. More specifically the long-run equilibrium relationship among all nonstationary components is represented by a cointegration vector, i.e. the correction term to equilibrium. This term is then utilized to improve a factor representation for the short run dynamics for the differenced processes. Comparing to the classical VECM, our setting does not require to specify the short run dynamics explicitly, avoiding the erroneous inference on cointegration due to, for example, a misspecification of the autoregressive order. Furthermore the high-dimensional short run dynamics is represented by a latent and low-dimensional factor process, avoiding the difficulties due to too many parameters in a high-dimensional VAR setting. Comparing to a pure factor model, the cointegration term improves the modelling and the prediction for short run dynamics. In terms of inference, we first adopt the eigenanalysis based method of Zhang, Robinson and Yao (2015) (ZRY, hereafter) to identify both the cointegration rank and cointegration space; no prespecification on the short run dynamics is required. We then calculate the regression estimation for the error correction term, and recover the latent factor process from the resulting residuals using the eigenanalysis based method of Lam and Yao (2012). Once the latent factor process has been recovered, we can model separately its linear dynamics using whatever an appropriate time series model.

The proposed methodology is further supported by the newly established asymptotic theory and numerical evidences. Especially our numerical results corroborate the findings from the asymptotic theory. In particular, Monte Carlo simulation reveals that the cointegration rank, the cointegration space, the number of factors and the factor co-feature space can all be estimated reasonably well with typical sizes of observed samples. Our empirical example on forecasting the ten U.S. industrial production indices shows that the proposed error correction factor model outperforms both VECM and vector AR models in post-sample forecasting. This finding is also robust with respect to the different forecast horizons.

The rest of the paper is organized as follows. We spell out the proposed error correction model and the associated estimation methods in Section 2. In Section 3 the asymptotic properties for

the estimation methods are established with the dimension of time series both fixed or diverging slowly, when the length of time series goes to infinity. The proposed methodology is further illustrated numerically in Section 4 with both simulated and real data sets. Furthermore we compare the forecasting performance of the proposed error correction factor model to those of the VECM with cointegration rank determined by Johansen's procedure and lag length selected by AIC, and unrestricted VAR in level model. The forecasting performances for real data were evaluated for different forecast horizons based on the criterion of Clements and Hendry (1993).

2 Methodology

2.1 Error Correction Factor Models

We call a vector process $\mathbf{u}_t$ weakly stationary if (i) $E\mathbf{u}_t$ is a constant vector independent of t, and (ii) $E\|\mathbf{u}_t\|^2 < \infty$, and $\mathrm{Cov}(\mathbf{u}_t, \mathbf{u}_{t+s})$ depends on s only for any integers t, s, where $\|\cdot\|$ denotes the Euclidean norm. Denoted by ∇ the difference operator, i.e. $\nabla\mathbf{u}_t = \mathbf{u}_t - \mathbf{u}_{t-1}$. We use the convention $\nabla^0\mathbf{u}_t = \mathbf{u}_t$. A process $\mathbf{u}_t$ is said to be weakly integrated process with order 1, abbreviated as weak $I(1)$, if $\nabla\mathbf{u}_t$ is weakly stationary with spectral density finite and positive definite at frequency 0 but $\mathbf{u}_t$ itself is not. Since we only deal with weak $I(1)$ processes in this paper, we simply call them weakly integrated processes.

Let $\mathbf{y}_t$ be observable $p \times 1$ weakly $I(1)$ process with the initial values $\mathbf{y}_t = 0$ for $t \leq 0$ and $\mathrm{Var}(\mathbf{y}_t)$ be full-ranked. Suppose that cointegration exists, i.e., there are r (≥ 1) stationary linear combinations of $\mathbf{y}_t$, where r is called the cointegration rank and is often unknown. The error correction factor model is defined as

$$\nabla\mathbf{y}_t = \mathbf{C}\mathbf{y}_{t-1} + \mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t, \tag{2.1}$$

where $\mathbf{C}$ is a $p \times p$ matrix with rank r and $\mathbf{C}\mathbf{y}_t$ is weakly stationary, $\mathbf{f}_t$ is an $m \times 1$ weakly stationary process and $\mathbf{B}$ is a $p \times m$ matrix, $\{\boldsymbol{\varepsilon}_t\}$ is a $p \times 1$ white noise with mean zero and finite fourth moments. Comparing with VECM, (2.1) represents the short-run dynamics by the latent process $\mathbf{f}_t$. Its linear dynamic structure is completely unspecificed. Note that $\mathbf{f}_t$ does not enter the inference for the error correction term $\mathbf{C}\mathbf{y}_{t-1}$.

Without loss of generality, we assume in (2.1) $\mathbf{B}$ to be an orthogonal matrix, i.e., $\mathbf{B}'\mathbf{B} = \mathbf{I}_m$, where $\mathbf{I}_m$ denotes the $m \times m$ identify matrix. This is due to the fact that any non-orthogonal $\mathbf{B}$ admits the decomposition $\mathbf{B} = \mathbf{Q}\mathbf{U}$, where $\mathbf{Q}$ is an orthogonal matrix and $\mathbf{U}$ is an upper-triangular matrix, and we may then replace $(\mathbf{B}, \mathbf{f}_t)$ in (2.1) by $(\mathbf{Q}, \mathbf{U}\mathbf{f}_t)$.

Before the end of this section, we illustrate with a simple toy model (i.e. a drunk and a dog model) the gain in prediction of the proposed error correction factor model over a pure factor

model. Let $y_{t,1}$ be the position of the drunk at time t, i.e.

$$y_{0,1} \equiv 0, \quad y_{t,1} = y_{t-1,1} + \varepsilon_t \quad \text{for} \quad t = 1, 2, \cdots,$$

where $\varepsilon_t \sim \text{IID}(0, 1)$. Let $y_{t,2} = y_{t,1} + z_t$ denote the position of the dog, as the dog always wanders around his master (i.e. the drunk), where $z_t \sim \text{IID}(0, \sigma^2)$, and $\{\varepsilon_t\}$ and $\{z_t\}$ are independent with each other. Then both $y_{t,1}, y_{t,2}$ are $I(1)$, and $y_{t,1}$ is a random walk but $y_{t,2}$ is not. Let $\mathbf{y}_t = (y_{t,1}, y_{t,2})'$. It holds that

$$\nabla \mathbf{y}_t = (0, 1)' f_t + (\varepsilon_t, \varepsilon_t)', \tag{2.2}$$

where $f_t = z_t - z_{t-1}$ can be viewed as a latent factor process. Thus (2.2) is a pure factor model. Now there is a clear cointegration:

$$z_t = (-1, 1)\mathbf{y}_t = y_{t,2} - y_{t,1}.$$

Consequently, the error correction factor model is

$$\nabla \mathbf{y}_t = (0, -1)' z_{t-1} + (\varepsilon_t, \varepsilon_t + z_t). \tag{2.3}$$

Note that for this simple example, the factor is identical to 0, as both ε_t and z_t are independent sequences. Using the info available upto time t, the best predictor for $\nabla \mathbf{y}_{t+1}$ based on error correction factor model is $\widehat{\nabla \mathbf{y}}_{t+1} = (0, -1)' z_t$ with the mean squared predictive error:

$$E(\|\nabla \mathbf{y}_{t+1} - \widehat{\nabla \mathbf{y}}_{t+1}\|^2) = \text{Var}(\varepsilon_{t+1}) + \text{Var}(\varepsilon_{t+1} + z_{t+1}) = 2 + \sigma^2. \tag{2.4}$$

On the other hand, the best predictor for $\nabla \mathbf{y}_{t+1}$ based on factor model (2.2) is $\widetilde{\nabla \mathbf{y}}_{t+1} = (0, 1)' \widehat{f}_{t+1}$ with the mean squared predictive error:

$$E(\|\nabla \mathbf{y}_{t+1} - \widetilde{\nabla \mathbf{y}}_{t+1}\|^2) = E(\|f_{t+1} - \widehat{f}_{t+1}\|^2) + 2\text{Var}(\varepsilon_{t+1}) = E(\|f_{t+1} - \widehat{f}_{t+1}\|^2) + 2,$$

where $\widehat{f}_{t+1}$ is the best predictor for f_{t+1} based on $\mathbf{y}_t, \mathbf{y}_{t-1}, \cdots$. Since $f_t = z_t - z_{t-1}$ is a latent MA(1) process. The best predictor for f_{t+1} is the one based on f_t and is equal to

$$\widehat{f}_{t+1} = -\frac{E(f_t f_{t+1})}{E(f_t^2)} f_t = f_t/2$$

with $E(\|f_{t+1} - \widehat{f}_{t+1}\|^2) = 3\sigma^2/2$. Hence

$$E(\|\nabla \mathbf{y}_{t+1} - \widetilde{\nabla \mathbf{y}}_{t+1}\|^2) = 3\sigma^2/2 + 2,$$

which is greater than $E(\|\nabla \mathbf{y}_{t+1} - \widehat{\nabla \mathbf{y}}_{t+1}\|^2)$; see (2.4). This shows that adding the error correction term increases the predictibility of factor model (2.2).

2.2 Estimation

In model (2.1), $\mathbf{C}$ is a $p \times p$ matrix with the reduced rank $r(< p)$. Hence it can be expressed as $\mathbf{C} = \mathbf{D}\mathbf{A}_2'$, where $\mathbf{D}$, $\mathbf{A}_2$ are two $p \times r$ matrices. Furthermore $\mathbf{A}_2'\mathbf{y}_{t-1}$ is the cointegration vector, r is the cointegration rank. Although $\mathbf{A}_2$ is not unique, the coefficient matrix $\mathbf{C}$ is uniquely determined by (2.1). Once we specify an $\mathbf{A}_2$ such that $\mathbf{A}_2'\mathbf{y}_{t-1}$ is weakly stationary, consequently $\mathbf{D}$ can be uniquely determined. Thus, to fit model (2.1), the key is to estimate r, $\mathbf{A}_2$, the factor dimension m and the factor loading matrix $\mathbf{B}$. Then the coefficient matrix $\mathbf{D}$ can be estimated by a multiple regression, the latent factors $\mathbf{f}_t$ can be recovered easily, and the forecasting can be based on a fitted time series model for $\mathbf{f}_t$.

To simplify the inference, in the sequel we always assume that $\mathbf{C}\mathbf{y}_{t-1}$ and $\mathbf{f}_t$ are uncorrelated. This avoids the identification issues due to possible endogeneity. Note that this condition is always fulfilled if we replace $(\mathbf{C}, \mathbf{f}_t)$ in (2.1) by $(\mathbf{C}^*, \mathbf{f}_t^*)$, where

$$\mathbf{C}^* = \{\mathbf{D} + \mathbf{B}\mathrm{E}[\mathbf{f}_t(\mathbf{A}_2'\mathbf{y}_{t-1})'][\mathrm{E}((\mathbf{A}_2'\mathbf{y}_{t-1})(\mathbf{A}_2'\mathbf{y}_{t-1})')]^{-1}\}\mathbf{A}_2',$$
$$\mathbf{f}_t^* = \mathbf{f}_t - \mathrm{E}(\mathbf{f}_t(\mathbf{A}_2'\mathbf{y}_{t-1})')[\mathrm{E}((\mathbf{A}_2'\mathbf{y}_{t-1})(\mathbf{A}_2'\mathbf{y}_{t-1})')]^{-1}(\mathbf{A}_2'\mathbf{y}_{t-1}).$$

2.2.1 Estimation for cointegration

While the representation of the cointegration vector $\mathbf{A}_2'\mathbf{y}_t$ is not unique, the cointegration space $\mathbf{M}(\mathbf{A}_2)$, i.e. the linear space spanned by the columns of $\mathbf{A}_2$, is uniquely determined by the process $\mathbf{y}_t$; see ZRY. In fact we can always assume that $\mathbf{A}_2$ is a half-orthogonal matrix in the sense that $\mathbf{A}_2'\mathbf{A}_2 = \mathbf{I}_r$. Let $\mathbf{A}_1$ be a $p \times (p-r)$ half orthogonal matrix such that $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$ be a $p \times p$ orthogonal matrix. Let $\mathbf{x}_{t,i} = \mathbf{A}_i'\mathbf{y}_t$ for $i = 1, 2$. Then $\mathbf{x}_{t,2}$ is a weakly stationary process, and all the components of $\mathbf{x}_{t,1}$ are weak $I(1)$.

We adopt the eigenanalysis based method proposed by ZRY to estimate r as well as $\mathbf{A}_2$. To this end, let

$$\widehat{\mathbf{W}} = \sum_{j=0}^{j_0} \widehat{\mathbf{\Sigma}}_j \widehat{\mathbf{\Sigma}}_j',$$

where $j_0 \geq 1$ is a prescribed and fixed integer, and

$$\widehat{\mathbf{\Sigma}}_j = \frac{1}{n}\sum_{t=1}^{n-j}(\mathbf{y}_{t+j} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})', \qquad \bar{\mathbf{y}} = \frac{1}{n}\sum_{t=1}^{n}\mathbf{y}_t.$$

We use the product $\widehat{\mathbf{\Sigma}}_j\widehat{\mathbf{\Sigma}}_j'$ instead of $\widehat{\mathbf{\Sigma}}_j$ to make sure that each term in the sum is non-negative definite, and that there is no information cancelation over different lags. Let $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{W}}$, and $\widetilde{\gamma}_1, \cdots, \widetilde{\gamma}_p$ be the corresponding eigenvectors. Then $\mathbf{A}_2$ is estimated by $\widehat{\mathbf{A}}_2 = (\widetilde{\gamma}_{p-r+1}, \cdots, \widetilde{\gamma}_p)$, and the cointegration rank is estimated by

$$\widehat{r} = \arg\min_{1 \leq l \leq p} IC(l), \tag{2.5}$$

where $IC(l) = \sum_{j=1}^{l} \widetilde{\lambda}_{p+1-j} + (p-l)\omega_n$, and $\omega_n \to \infty$ and $\omega_n/n^2 \to 0$ in probability (as we allow ω_n to be data-dependent). ZRY has shown that both $\mathcal{M}(\widehat{\mathbf{A}}_2)$ and $\widehat{r}$ are consistent estimators for, respectively, $\mathcal{M}(\mathbf{A}_2)$ and r.

Having obtained the estimated cointegration vector $\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1}$, the coefficient matrix $\mathbf{D}$ can be estimated using the standard least squares estimation. Let $\mathbf{d}_i$, $i = 1, 2, \cdots, p$ be the row vectors of $\mathbf{D}$ and $\nabla\mathbf{y}_t = (\nabla y_t^1, \cdots, \nabla y_t^p)'$. The least square estimator for $\mathbf{d}_i$ is defined as

$$\widehat{\mathbf{d}}_i = \arg\min_{\mathbf{d}_i} \sum_{t=1}^{n} (\nabla y_t^i - \mathbf{d}_i \widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})^2, \tag{2.6}$$

which leads to $\widehat{\mathbf{d}}_i = \sum_{t=1}^{n} \nabla y_{ti}(\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})' \left(\sum_{i=1}^{n}(\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})(\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})'\right)^{-1}$. Consequently, the estimator for the coefficient matrix $\mathbf{D}$ can be written as

$$\widehat{\mathbf{D}} = \sum_{t=1}^{n} \nabla\mathbf{y}_t(\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})' \left(\sum_{i=1}^{n}(\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})(\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})'\right)^{-1}.$$

2.2.2 Estimation for latent factors

We adopt the eigenanalysis based method of Lam and Yao (2012) to estimate the factor loading space $\mathcal{M}((\mathbf{B})$ and the latent factor process $\mathbf{f}_t$ based on the residuals $\widehat{\mathbf{v}}_t \equiv \nabla\mathbf{y}_t - \widehat{\mathbf{D}}\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1}$, $t = 1, \cdots, n$. To this end, let

$$\widehat{\mathbf{W}}_v = \sum_{j=1}^{j_0} \widehat{\mathbf{\Sigma}}_v(j)\widehat{\mathbf{\Sigma}}_v'(j), \tag{2.7}$$

where $j_0 \geq 1$ is a prespecified and fixed integer, and

$$\widehat{\mathbf{\Sigma}}_v(j) = \frac{1}{n}\sum_{t=1}^{n-j}(\widehat{\mathbf{v}}_{t+j} - \bar{\mathbf{v}})(\widehat{\mathbf{v}}_t - \bar{\mathbf{v}})', \qquad \bar{\mathbf{v}} = \frac{1}{n}\sum_{t=1}^{n}\widehat{\mathbf{v}}_t.$$

where $j_0 \geq 1$ is a prespecified and fixed integer. One distinctive advantage of using the quadratic form $\widehat{\mathbf{\Sigma}}_v(j)\widehat{\mathbf{\Sigma}}_v(j)'$ instead of $\widehat{\mathbf{\Sigma}}_v(j)$ in (2.7) is that there is no information cancellation over different lags. Therefore this approach is insensitive to the choice of j_0 in (2.7). Often small values such as $j_0 = 5$ are sufficient to catch the relevant characteristics, as serial dependence is usually most predominant at small lags. See Lam and Yao (2012) and Chang et al. (2014). Let $(\widehat{\gamma}_1, \cdots, \widehat{\gamma}_m)$ be the orthonormal eigenvectors of $\widehat{\mathbf{W}}_v$ corresponding to the m largest eigenvalues. Consequently, we estimate $\mathbf{B}$ and $\mathbf{f}_t$ by

$$\widehat{\mathbf{B}} = (\widehat{\gamma}_1, \cdots, \widehat{\gamma}_m), \quad \text{and} \quad \widehat{\mathbf{f}}_t = \widehat{\mathbf{B}}'\widehat{\mathbf{v}}_t. \tag{2.8}$$

Since m is usually unknown and the last $p - m$ eigenvalues of $\widehat{\mathbf{W}}_v$ may not be exactly 0 due to the random fluctuation, the determination of m is required. We propose to select m by using

the ratio-based method of Lam and Yao (2012). In particular, let $\widehat{\lambda}_1 \geq \widehat{\lambda}_2 \geq \cdots \geq \widehat{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{W}}_v$. We define an estimator for the number of factors m as follows:

$$\widetilde{m} = \arg\min_{1 \leq i \leq R} \widehat{\lambda}_{i+1}/\widehat{\lambda}_i, \tag{2.9}$$

with $m < R < p$. In practice we may pick, for example, $R = p/2$, following the recommendation of Lam and Yao (2012).

Remark 1. *The above ratio estimator of m is not necessarily consistent, though it works fine in practice. See Lam and Yao (2012), and also Tables 1, 2 and 3 in Section 4.1 below. To establish the consistency, one can estimate m using the information criterion defined as*

$$\widehat{m} = \arg\min_{1 \leq i \leq p} IC(l),$$

where $IC(l) = \sum_{j=l+1}^{p} \widehat{\lambda}_j + l\omega_n$, is the information criterion and ω_n is the turning parameter. It can be shown as $\omega_n \to 0$ and $\omega_n n^{1/2}/p \to \infty$, $\widehat{m}$ is consistent.

2.2.3 Fitting linear dynamics for factors

Once we have recovered the factor process $\widehat{\mathbf{f}}_t$, we can fit an appropriate model to represent its linear dynamic structure. As an illustration, below we fit $\widehat{\mathbf{f}}_t$ with a VAR model.

Let

$$\widehat{\mathbf{f}}_t = \sum_{i=1}^{s} \mathbf{E}_i \widehat{\mathbf{f}}_{t-i} + \mathbf{e}_t, \tag{2.10}$$

where $\mathbf{E}_i$, $1 \leq i \leq s$ are $m \times m$ matrices and $\{\mathbf{e}_t\}$ is a sequence of independent vectors with mean zero and independent of $\{\mathbf{x}'_{t2}, \mathbf{f}'_t, \boldsymbol{\varepsilon}'_t\}$. We estimate the parameters $\mathbf{E}_i$ by least squares method, i.e.

$$(\widehat{\mathbf{E}}_1, \cdots, \widehat{\mathbf{E}}_s) = \text{argmin}_{\mathbf{E}_1, \cdots, \mathbf{E}_s} \sum_{t=s+1}^{n} ||\widehat{\mathbf{f}}_t - \sum_{i=1}^{s} \mathbf{E}_i \widehat{\mathbf{f}}_{t-i}||^2, \tag{2.11}$$

where $\widehat{\mathbf{f}}_t = \widehat{\mathbf{B}}'\widehat{\mathbf{v}}_t$ is given in (2.8). The autoregressive order s may be determined by, for example, the standard criteria such as AIC or BIC. See, for example, Section 4.2.3 of Fan and Yao (2015).

Combining (2.1), (2.10) and (2.11), we have h-step ahead forecast, for $h = 1, 2$, as:

$$\begin{aligned}
\mathbf{y}_{t+1|t} &= (\mathbf{I} + \widehat{\mathbf{C}})\mathbf{y}_t + \widehat{\mathbf{B}}\widehat{\mathbf{f}}_{t+1} = (\mathbf{I} + \widehat{\mathbf{C}})\mathbf{y}_t + \widehat{\mathbf{B}} \left(\sum_{i=1}^{s} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i} \right), \\
\mathbf{y}_{t+2|t} &= (I + \widehat{\mathbf{C}})\mathbf{y}_{t+1|t} + \widehat{\mathbf{B}}\widehat{\mathbf{f}}_{t+2|t} \\
&= (\mathbf{I} + \widehat{\mathbf{C}})^2 \mathbf{y}_t + (\mathbf{I} + \widehat{\mathbf{C}})\widehat{\mathbf{B}} \left(\sum_{i=1}^{s} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i} \right) + \widehat{\mathbf{B}} \left[\sum_{i=1}^{s-1} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i} + \widehat{\mathbf{E}}_1 \left(\sum_{i=1}^{s} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i} \right) \right].
\end{aligned}$$

We can similarly deduce any h-step ahead forecast $\mathbf{y}_{t+h|t}$, for $h \geq 3$, by recursive iteration.

3 Asymptotic Theory

In this section, we investigate the asymptotic properties of the proposed estimators. We first show that the estimator $\widehat{\mathbf{C}}$ is consistent. And given m, we measure the distance between the cofeature space $\mathcal{M}(\mathbf{B})$ and its estimate by

$$D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) = \sqrt{1 - \frac{1}{m}\text{tr}(\widehat{\mathbf{B}}\widehat{\mathbf{B}}'\mathbf{B}\mathbf{B}')}. \tag{3.1}$$

Then $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) \in [0, 1]$, being 0 if and only if $\mathcal{M}(\widehat{\mathbf{B}}) = \mathcal{M}(\mathbf{B})$, and 1 if and only if $\mathcal{M}(\widehat{\mathbf{B}})$ and $\mathcal{M}(\mathbf{B})$ are orthogonal. Furthermore, we show that the estimator $\widetilde{m}$, defined in (2.9), is a consistent estimator for the true number of factors m. We consider two asymptotic modes: (i) p is fixed while $n \to \infty$, and (ii) both p and n diverge, but r is fixed.

3.1 When $n \to \infty$ and p is fixed

We introduce a regularity condition first.

> **Condition 1**. The process $\{\mathbf{x}'_{t2}, \mathbf{f}'_t, \boldsymbol{\varepsilon}'_t\}$ is a stationary α-mixing process with mean zero, $\text{E}||(\mathbf{x}'_{t2}, \mathbf{f}'_t, \boldsymbol{\varepsilon}'_t)||^{4\gamma} < \infty$ for some constant $\gamma > 1$ and the mixing coefficients α_t satisfying the condition $\sum_{t=1}^{\infty} \alpha_t^{1-1/\gamma} < \infty$.
>
> **Condition 2**. The characteristic polynomial of VAR model (2.10) has no roots on or outside of the unit circle so that it is a casual VAR model.

Theorem 1. *Let Condition 1 hold.*

> *(a) Let vech*$(\mathbf{D}) = (\mathbf{d}_1, \cdots, \mathbf{d}_p)'$. *As* $n \to \infty$ *and* p *fixed, it holds that*
>
> $$\sqrt{n}(vech(\widehat{\mathbf{D}}) - vech(\mathbf{D})) \xrightarrow{d} N(0, \boldsymbol{\Omega}_1),$$
>
> *where* $\boldsymbol{\Omega}_1$ *is a* $rp \times rp$ *positive definite matrix and* $||\widehat{\mathbf{C}} - \mathbf{C}||_2 = O_p(n^{-1/2})$, *and* $||\cdot||_2$ *denotes the spectral norm of a matrix.*
>
> *(b) Let* m *be known, then* $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) = O_p(n^{-1/2})$.
>
> *(c) If Condition 2 and* $\text{E}\|\mathbf{e}_t\|^{2\gamma} < \infty$ *hold in addition, there exists a positive definite matrix* $\boldsymbol{\Omega}_2$ *such that*
>
> $$\sqrt{n}(vech(\widehat{\mathbf{E}}_1, \cdots, \widehat{\mathbf{E}}_s) - vech(\mathbf{E}_1, \cdots, \mathbf{E}_s)) \xrightarrow{d} N(0, \boldsymbol{\Omega}_2).$$

Theorem 2. *Let* $1 \leq m < p$ *and Condition 1 hold. For* $\widetilde{m}$ *defined in (2.9),*

$$\lim_{n\to\infty} P(\,\widetilde{m} > m\,) = 1.$$

3.2 When $n \to \infty$ and $p = o(n^c)$

Let $z_t^j \equiv \nabla x_t^j$, $j = 1, \cdots, p-r$, $\mathbf{z}_t = (z_t^1, \cdots, z_t^{p-r})'$ and $\boldsymbol{\nu}_t = (\mathbf{z}_t', \mathbf{x}_{t2}')'$. In this subsection, we extend the asymptotic results in the previous section to the cases when $p \to \infty$ and $p = o(n^c)$ for some $c \in (0, 1/2)$. Technically we employ a normal approximation method to establish the results.

Condition 3.

(i) Let $\mathbf{M}$ be a $p \times k$ constant matrix with $k \geq p$ and $c_1 \leq \lambda_{\min}(\mathbf{M}) \leq \lambda_{\max}(\mathbf{M}) \leq c_2$, where c_1, c_2 are two positive constants. Suppose that $\boldsymbol{\nu}_t = \mathbf{M}\mathbf{v}_t$, all the components of $\mathbf{v}_t = (v_t^1, \cdots, v_t^k)'$ are independent and with mean zero.

(ii) The process $\{\mathbf{v}_t, \mathbf{f}_t, \boldsymbol{\varepsilon}_t\}$ is a stationary α-mixing process with $\mathrm{E}||(\mathbf{v}_t', \mathbf{f}_t', \boldsymbol{\varepsilon}_t')||^{2\theta} < \infty$ for some $\theta > \eta \in (2, 4]$ and the mixing coefficients α_m satisfying

$$\sum_{m=1}^{\infty} \alpha_m^{(\theta-\eta)/(\theta\eta)} < \infty. \tag{3.2}$$

(iii) There exist two positive constants c_3, c_4 such that $c_3 \leq \lambda_{\min}(\mathbf{D}) \leq \lambda_{\max}(\mathbf{D}) \leq c_4$.

Theorem 3. *Let m be known. Suppose Condition 3 holds with $k = o(n^{1/2-1/\eta})$ and $p = O\left(n^{1/2-1/\eta}/(\log n)^2\right)$, then the following assertions hold.*

(a) $\max\{||\widehat{\mathbf{D}} - \mathbf{D}||_2, ||\widehat{\mathbf{C}} - \mathbf{C}||_2\} = O_p((pr)^{1/2}n^{-1/2} + p^{1/2}k^2n^{-1})$.

(b) $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) = O_p(pn^{-1/2})$.

(c) $||(\widehat{\mathbf{E}}_1, \cdots, \widehat{\mathbf{E}}_s)||_2 = O_p((pm)^{1/2}n^{-1/2} + p^{1/2}k^2n^{-1})$, *provided that Condition 2 and* $\mathrm{E}||\mathbf{e}_t||^\theta < \infty$ *hold in addition.*

Theorem 4. *Let $1 \leq m < p$, Condition 3 holds with $k = o(n^{1/2-1/\eta})$ and $p = O\left(n^{1/2-1/\eta}/(\log n)^2\right)$. For $\widetilde{m}$ defined in (2.9),*

$$\lim_{n\to\infty} P(\,\widetilde{m} > m\,) = 1.$$

Remark 2. *All the above asymptotic theorems can be generalized to other stationary noise $\boldsymbol{\nu}_t$ considered by ZRY.*

4 Numerical Studies

In this section, we first evaluate the finite sample performance of our proposed inference procedure via Monte Carlo simulation. We then illustrate the advantage in forecasting of the proposed error correction factor model via a real data example.

4.1 Monte Carlo Simulations

In our simulation, we let $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t$, where $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$ is an orthogonal matrix which was drawn elementwisely from $U[0,1]$ independently first and was then orthogonalized, and $\mathbf{x}_t = (\mathbf{x}_{t1}', \mathbf{x}_{t2}')'$ in which the r components of $\mathbf{x}_{t2}$ are independent Gaussian AR(1) processes with identical autoregressive coefficient 0.5, and the $(p-r)$ vector $\mathbf{x}_{t1}$ is $I(1)$ according to a factor augmented AR(1) defined as

$$\mathbf{x}_{t1} = \mathbf{x}_{t-1,1} + \boldsymbol{\Upsilon}\mathbf{f}_t + \boldsymbol{e}_t. \tag{4.3}$$

In the above expression, $\boldsymbol{\Upsilon}$ is a $(p-r)\times m$ half orthogonal matrix (i.e. $\boldsymbol{\Upsilon}'\boldsymbol{\Upsilon} = \mathbf{I}_m$) generated in the same manner as $\mathbf{A}$, the components of factor $\mathbf{f}_t$ are independent stationary Gaussian AR(1) with identical autoregressive coefficient 0.5, and $\boldsymbol{e}_t$ are independent and $N(0, \mathbf{I}_p)$. Then it is to see that $\mathbf{y}_t$ satisfies equation (2.1) with $\mathbf{C} = 0.5\mathbf{A}_2\mathbf{A}_2'$ and $\mathbf{B} = \mathbf{A}_1\boldsymbol{\Upsilon}$.

With $p = 5, 10, 20, 40, 60$, $r = 1, 2, 4, 6, 8, 10$, and $m = 1, 2, 4, 6, 8, 10$ $(m \le p - r)$, we generate a time series $\mathbf{y}_t$ with length $n = 100, 200, 400, 800, 1200, 1600, 2000, 2400$ and estimate $r, \mathbf{C}, m$ and $\mathbf{B}$. For estimating r, we use the IC criterion (2.5) with the penalty $w_n = \log n\widetilde{\lambda}_p$. The number of factor m is estimated using the ratio method (2.9). For each setting we replicated the experiment 1000 times.

Tables 1-3 list the relative frequencies of the occurrence of the events $(\widehat{r} = r)$ and $(\widetilde{m} = m)$ in simulation with 1000 replications. We make the following observations from Table 1 which contains the results with $p = 5, 10$ and 20. First, with $p = 5$ or 10, the relative frequencies for the correct specification for the cointegration rank r and the number of factors m are as high as 85% even for the sample size n as small as 200. When n increases to 400, those relative frequencies increase to 100%. Secondly, with fixed n and r the correct estimation rates for m increases when dimension p increases, a phenomenon coined as the "blessing-of-dimensionality". This is consistent with the findings in Lam and Yao (2012) which only dealt with purely stationary processes. Thirdly, the inference on r tends to be more challenging when p increases. For example, the relative frequency for correct estimation of $r(=2)$, when $m = 1$ and $n = 200$, decreases from 68.5% to 65.4% with p increasing from 5 to 10. This is in line with the findings in ZRY. Lastly, we note that the increase in p, r and m would generally demand a larger n to maintain the same level of estimation accuracy. This is consistent with our theory that requires $p = o(n^c)$ for $c \in (0, 1/2)$.

Some similar conclusions can be drawn from results reported in Table 2-3. In particular, the inference on the number of factor (when m is relatively small compared to p) is relatively easy when $p = 40$ and 60, with a sample size equal to 800. Unreported results for $n = 200, 400$ also corroborate this conclusion. However, the inference on the cointegration rank is more difficult when n is small or/and r is large.

To evaluate the performance of the estimation for both cointegration space and factor cofeature space, we present the boxplots of $D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\mathbf{A}_2))$ and that of $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B}))$ in Figures 1, for a few (selected) combinations of p, r and m, with $n = 400, 800, 1600, 3200$. The overall profile of the estimation accuracy is similar to those in Tables 1-3. For example, when p increase, the estimation accuracy of cointegration space becomes worse, while that of factor cofeature space tends to improve. That is, the "curse-of-dimensionality" in inferring cointegration space is coupled with the "blessing-of-dimensionality" in estimating the factor cofeature space. It is further observed that the estimation in general improves as n increases, which confirms our consistency theory.

Table 1: Relative frequencies ($\times 100$) of the occurrences of events $\widehat{r} = r$ (1st entries in parebtheses) and $\widetilde{m} = m$ (2nd entries in parentheses) in a simulation with 1000 replications.

$p = 5$		$n = 100$	$n = 200$	$n = 400$	$n = 800$
$m = 1$	$r = 1$	(92.0, 93.5)	(100.0, 99.3)	(100.0, 99.9)	(100.0, 100.0)
	$r = 2$	(44.6, 89.3)	(68.5, 96.6)	(83.7, 99.8)	(98.6, 100.0)
$p = 10$		$n = 200$	$n = 400$	$n = 800$	$n = 1200$
$m = 1$	$r = 1$	(85.3, 100.0)	(100.0, 100.0)	(100.0, 100.0)	(100.0, 100.0)
	$r = 2$	(65.4, 100.0)	(82.0, 100.0)	(95.4, 100.0)	(99.6, 100.0)
$m = 2$	$r = 1$	(86.5, 82.2)	(100.0, 97.7)	(100.0, 99.9)	(100.0, 100.0)
	$r = 2$	(62.4, 83.4)	(75.1, 97.8)	(94.3, 100.0)	(98.8, 100.0)
$p = 20$		$n = 400$	$n = 800$	$n = 1200$	$n = 1600$
$m = 2$	$r = 2$	(85.5, 99.7)	(92.8, 100.0)	(96.7, 100.0)	(98.9, 100.0)
	$r = 4$	(20.5, 95.0)	(43.3, 99.8)	(68.8, 100.0)	(86.3, 100.0)
$m = 4$	$r = 2$	(82.0, 93.2)	(89.5, 99.9)	(93.8, 99.9)	(96.3, 100.0)

4.2 A Real Data Example

To further illustrate the proposed approach, we apply the proposed error correction factor model (ECFM) to the ten U.S. Industrial Production monthly indices in January 1959 — December 2006, extracted from Stock and Watson (2008), namely, products total, final products, consumer goods, durable consumer goods, nondurable consumer goods, materials, durable goods materials, nondurable goods materials, manufacturing, and residential utilities. The estimated cointegration rank is $\widehat{r} = 5$, and the number of factor is $\widetilde{m} = 1$. We also fit the data with a vector error correction model (VECM) using Johansen's trace test to determine the cointegration rank r for each given

Table 2: Relative frequencies ($\times 100$) of the occurrences of events $\widehat{r} = r$ (1st entries in parebtheses) and $\widetilde{m} = m$ (2nd entries in parentheses) in a simulation with 1000 replications.

$p = 40$		$n = 800$	$n = 1200$	$n = 1600$	$n = 2000$
$m = 2$	$r = 2$	(72.8, 100.0)	(94.7, 100.0)	(100.0, 100.0)	(100.0, 100.0)
	$r = 4$	(64.0, 99.9)	(99.5, 100.0)	(99.3, 100.0)	(99.7, 100.0)
	$r = 6$	(86.4, 93.8)	(95.2, 98.9)	(96.2, 99.7)	(97.5, 100.0)
	$r = 8$	(53.8, 100.0)	(77.4, 100.0)	(82.2, 100.0)	(89.6, 100.0)
$m = 4$	$r = 2$	(73.3, 100.0)	(89.5, 100.0)	(99.9, 100.0)	(100.0, 100.0)
	$r = 4$	(66.8, 99.9)	(99.3, 100.0)	(99.5, 100.0)	(99.2, 100.0)
	$r = 6$	(75.1, 99.5)	(88.3, 100.0)	(89.5, 100.0)	(91.0, 100.0)
	$r = 8$	(27.1, 99.7)	(59.0, 100.0)	(64.4, 100.0)	(75.9, 100.0)
$m = 6$	$r = 2$	(72.7, 99.6)	(86.2, 100.0)	(99.6, 100.0)	(100.0, 100.0)
	$r = 4$	(69.2, 96.5)	(98.6, 99.4)	(98.3, 100.0)	(98.4, 100.0)
	$r = 6$	(65.6, 99.7)	(83.1, 100.0)	(86.1, 100.0)	(88.8, 100.0)
	$r = 8$	(16.9, 98.7)	(41.3, 100.0)	(50.8, 100.0)	(62.4, 100.0)
$m = 8$	$r = 2$	(73.7, 99.9)	(81.1, 100.0)	(99.8, 100.0)	(100.0, 100.0)
	$r = 4$	(71.0, 89.1)	(98.3, 99.2)	(98.2, 99.9)	(98.0, 100.0)
	$r = 6$	(60.8, 98.7)	(82.1, 99.9)	(82.1, 100.0)	(87.0, 100.0)
	$r = 8$	(12.7, 83.7)	(37.0, 96.5)	(45.3, 98.5)	(52.6, 99.6)

autoregressive order between 1 and 8, and then using the Akaike Information Criterion (AIC) to select the optimal autoregressive order 6. The corresponding estimated cointegration rank is also 5. Hence both the fitted models suggest the same cointegration rank 5, while VECM represents the short-run dynamics in terms of a ten-dimensional vector AR(6) process (with 6 10×10 autocoefficient matrices), and, in contrast, the newly proposed ECFM captures this dynamics in a univariate latent factor process, achieving a massive reduction in the number of parameters required. The difference between the cointegration space estimated by our ECFM and that produced by Johansen's method is computed as

$$D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\widetilde{\mathbf{A}}_2))^2 = 1 - \frac{1}{5}\mathrm{tr}\{\widehat{\mathbf{A}}_2\widehat{\mathbf{A}}_2'(\widetilde{\mathbf{A}}_2(\widetilde{\mathbf{A}}_2'\widetilde{\mathbf{A}}_2)^{-1}\widetilde{\mathbf{A}}_2)'\} = 0.0157,$$

where columns of $\widehat{\mathbf{A}}_2$ denote the loadings of the five cointegrated variables identified by our method and those of $\widetilde{\mathbf{A}}_2$ by Johansen's. This suggests that the estimated cointegration spaces by both approaches be effectively equivalent.

Table 3: Relative frequencies ($\times 100$) of the occurrences of events $\widehat{r} = r$ (1st entries in parebtheses) and $\widetilde{m} = m$ (2nd entries in parentheses) in a simulation with 1000 replications.

$p = 60$		$n = 1200$	$n = 1600$	$n = 2000$	$n = 2400$
$m = 2$	$r = 2$	(20.8, 100.0)	(34.3, 100.0)	(97.3, 100.0)	(100.0, 100.0)
	$r = 4$	(16.2, 100.0)	(87.3, 100.0)	(100.0, 100.0)	(99.9, 100.0)
	$r = 6$	(63.4, 100.0)	(99.1, 100.0)	(99.5, 100.0)	(99.5, 100.0)
	$r = 8$	(88.4, 100.0)	(98.9, 100.0)	(97.5, 100.0)	(97.1, 100.0)
	$r = 10$	(72.0, 100.0)	(92.4, 100.0)	(89.7, 100.0)	(89.6, 100.0)
$m = 4$	$r = 2$	(19.8, 100.0)	(23.3, 100.0)	(94.3, 100.0)	(99.9, 100.0)
	$r = 4$	(16.7, 100.0)	(78.4, 100.0)	(100.0, 100.0)	(100.0, 100.0)
	$r = 6$	(59.3, 100.0)	(97.7, 100.0)	(99.1, 100.0)	(98.7, 100.0)
	$r = 8$	(80.1, 100.0)	(95.3, 100.0)	(92.7, 100.0)	(92.5, 100.0)
	$r = 10$	(51.0, 100.0)	(77.8, 100.0)	(73.4, 100.0)	(71.5, 100.0)
$m = 6$	$r = 2$	(20.4, 100.0)	(29.6, 100.0)	(86.6, 100.0)	(99.5, 100.0)
	$r = 4$	(13.4, 100.0)	(72.5, 100.0)	(99.8, 100.0)	(100.0, 100.0)
	$r = 6$	(58.9, 100.0)	(97.2, 100.0)	(98.6, 100.0)	(98.1, 100.0)
	$r = 8$	(73.3, 100.0)	(91.7, 100.0)	(87.0, 100.0)	(87.0, 100.0)
	$r = 10$	(29.9, 100.0)	(62.5, 100.0)	(59.2, 100.0)	(57.2, 100.0)
$m = 8$	$r = 2$	(20.7, 100.0)	(24.9, 100.0)	(79.3, 100.0)	(99.3, 100.0)
	$r = 4$	(33.2, 100.0)	(70.1, 100.0)	(99.5, 100.0)	(99.7, 100.0)
	$r = 6$	(59.3, 100.0)	(95.6, 100.0)	(98.8, 100.0)	(98.2, 100.0)
	$r = 8$	(67.9, 100.0)	(89.9, 100.0)	(84.3, 100.0)	(85.4, 100.0)
	$r = 10$	(23.7, 99.7)	(54.0, 100.0)	(50.9, 100.0)	(51.6, 100.0)
$m = 10$	$r = 2$	(20.3, 100.0)	(21.2, 100.0)	(76.6, 100.0)	(98.5, 100.0)
	$r = 4$	(33.8, 100.0)	(65.8, 100.0)	(99.4, 100.0)	(100.0, 100.0)
	$r = 6$	(60.0, 100.0)	(94.7, 100.0)	(98.7, 100.0)	(98.3, 100.0)
	$r = 8$	(61.6, 100.0)	(87.6, 100.0)	(84.7, 100.0)	(85.4, 100.0)
	$r = 10$	(18.6, 99.9)	(49.5, 100.0)	(48.0, 100.0)	(48.2, 100.0)

We further examine the forecasting performance of the proposed ECFM. To this end, we compare the out-of-sample forecasting performance of our ECFM with those of (i) unrestricted VAR in log-levels with lag length selected by the standard Schwarz criterion, and (ii) VECM with cointegration rank chosen by Johansen's procedure (trace test with 5% critical values) and

Table 4: Percentage improvement in forecast accuracy measures: US IP indices

Horizon (h)	ECFM versus VAR in level			VECM(AIC+J) versus VAR in level			ECFM versus VECM(AIC+J)		
	TMSFE	| MSFE|	GFESM	TMSFE	| MSFE|	GFESM	TMSFE	| MSFE|	GFESM
1	55.5	99.5	27.4	28.1	91.6	14.3	38.2	94.8	15.2
4	66.6	97.8	77.6	46.5	90.3	58.7	37.5	78.0	45.6
8	78.3	89.8	88.0	48.2	74.2	69.0	58.1	60.7	61.4
12	81.9	94.5	89.0	51.3	72.5	70.4	62.7	80.2	63.0
16	82.6	95.4	90.8	58.1	69.2	71.1	58.5	85.4	68.4

lag length selected by AIC. For each of the last 10% of data points, we fit the models using the data upto its previous month and forecast the values using the three fitted models. Following Athanasopoulos *et al.* (2011), we measure the forecast accuracy using traditional trace of the mean-squared forecast error matrix (TMSFE) and the determinant of the mean-squared forecast error matrix |MSFE| at each forecast horizon $h = 1, \cdots, 16$. We also calculate the generalized forecast error second moment (GFESM), i.e., the determinant of the expected value of the outer product of the vector of stacked forecast errors of all future times up to the horizon of interest, of Clements and Hendry (1993). GFESM is invariant to elementary operations that involve different variables, and also to elementary operations that involve the same variable at different horizons. The forecasting comparison results are presented in Table 4.

It is observed from Table 4 that both ECFM and VECM provide more accurate forecasts than the VAR in level model. For example, for 12 month ahead forecast, ECFM achieves improvement in TMSFE, |MSFE| and GFESM by, respectively, 81.9%, 94.5%, 89.0%, compared to VAR in level model. The improvement from using VECM over VAR is obvious though less substantial than that of ECFM. The direct comparison between EDFM and VECM in the right panel of Table 4 shows superiority of EDFM in forecasting across all the forecasting horizons.

5 Conclusions

We conclude the paper with two open questions.

First, in order to apply the result of Zhang, Robinson and Yao (2015), the dimension p cannot be too large (i.e. not greater than than $O(n^{1/4})$). It would be interesting and more challenging to consider the cases with larger p. Note that the rank of the matrix $\mathbf{C}$ is r. One possible solution is to replace the first step in the procedure via sparse shrinkage technique by solving the following optimal problem:

$$\widehat{\mathbf{C}} = \text{argmin}_{\mathbf{C}\in R^{p\times p}} \left\{ \sum_{t=1}^{n} ||\nabla \mathbf{y}_t - \mathbf{C}\mathbf{y}_{t-1}||^2 + \lambda_n |||\mathbf{C}|||_{s_1} \right\}, \tag{5.4}$$

where $||\mathbf{C}||_{s_1} = \sum_{j=1}^{p} \lambda_j(\mathbf{C}),$ and $\lambda_1(\mathbf{C}), \lambda_2(\mathbf{C}), \cdots, \lambda_p(\mathbf{C})$ denote the singular values of $\mathbf{C}$.

Secondly, since in this paper our purpose is in prediction and inference for the cofeatures, we can impose the condition that $\mathbf{C}\mathbf{y}_{t-1}$ and $\mathbf{f}_t$ are uncorrelated; see the beginning of Section 2.2. However, for some applications the main concern may be on the original $\mathbf{C}$ and $\mathbf{f}_t$. Since $\mathbf{C}\mathbf{y}_{t-1}$ and $\mathbf{f}_t$ may be correlated with each other, the inference method proposed in this paper will lead to insistent estimators. It would be interesting to consider the inference based on some iterative equations as in Bai (2009), i.e., estimate $\{\mathbf{C}, \mathbf{F}, \mathbf{B}\}$ via the least squares objective function defined as

$$\text{SSR}(\mathbf{C}, \mathbf{F}, \mathbf{B}) = \sum_{t=1}^{n} (\nabla \mathbf{y}_t - \mathbf{C}\mathbf{y}_{t-1} - \mathbf{B}\mathbf{f}_t)'(\nabla \mathbf{y}_t - \mathbf{C}\mathbf{y}_{t-1} - \mathbf{B}\mathbf{f}_t) \tag{5.5}$$

subject to the constraint $\mathbf{B}'\mathbf{B} = \mathbf{I}_m$.

6 Appendix: Technical proofs

Lemma 5. *Under Condition 1 or conditions of Theorem 3, we have*

$$\frac{1}{n}\sum_{t=1}^{n-1} (\widehat{\mathbf{A}}_2' \mathbf{y}_t \mathbf{y}_t' \widehat{\mathbf{A}}_2 - \mathbf{A}_2' \mathbf{y}_t \mathbf{y}_t' \mathbf{A}_2) = o_p(1). \tag{6.1}$$

Proof. We first show the case with fixed p. Since $\{\mathbf{x}_{t2}, \mathbf{f}_t, \boldsymbol{\varepsilon}_t\}$ is α mixing with mixing coefficients α_m satisfying

$$\sum_{m=1}^{\infty} \alpha_m^{1-1/\gamma} < \infty, \tag{6.2}$$

it follows that $\{\nabla \mathbf{y}_t\}$ is a α mixing process with mixing coefficients satisfying (6.2). Thus, by Theorem 3.2.3 of Lin and Lu (1997), there exists a p-dimensional Gaussian process $\mathbf{g}(t)$ such that

$$\mathbf{y}_{[nt]}/\sqrt{n} \xrightarrow{d} \mathbf{g}(t), \text{ on } D[0,1]. \tag{6.3}$$

From (6.3) and the continuous mapping theorem, it follows that

$$\frac{1}{n^2}\sum_{t=1}^{n} \mathbf{y}_t \mathbf{y}_t' \xrightarrow{d} \int_0^1 \mathbf{g}(t)\mathbf{g}'(t)\, dt. \tag{6.4}$$

Further, by $\text{E}||\mathbf{x}_{t2}||^{2\gamma} < \infty$ for some $\gamma > 1$, we have

$$\max_{1\le t\le n} ||\mathbf{x}_{t2} - \text{E}\mathbf{x}_{t2}||/\sqrt{n} = o_p(1), \text{ and } \frac{1}{n}\sum_{t=1}^{n} ||\mathbf{x}_{t2} - \text{E}\mathbf{x}_{t2}|| = O_p(1). \tag{6.5}$$

Combining (6.3) and (6.5) (see Lemma 7 of ZRY) yields

$$\frac{1}{n^{3/2}}||\sum_{t=1}^{n} \mathbf{y}_t \mathbf{x}_{t2}'||_2 = o_p(1). \tag{6.6}$$

On the other hand, by $\nabla \mathbf{x}_{t1} = \mathbf{A}_1' \nabla \mathbf{y}_t$, we know $(\nabla \mathbf{x}_{t1}, \mathbf{x}_{t2})$ is also α mixing with mixing coefficients satisfying (6.2). As a result, by the proof of Theorem 1 in ZRY,

$$||\widehat{\mathbf{A}}_2 - \mathbf{A}_2||_2 = O_p(1/n). \tag{6.7}$$

By (6.4), (6.6) and (6.7), we have

$$\begin{aligned}
& ||\frac{1}{n}\sum_{t=1}^{n-1}(\widehat{\mathbf{A}}_2'\mathbf{y}_t\mathbf{y}_t'\widehat{\mathbf{A}}_2 - \mathbf{A}_2'\mathbf{y}_t\mathbf{y}_t'\mathbf{A}_2)||_2 \\
= \; & ||(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\frac{\sum_{t=1}^{n-1}\mathbf{y}_t(\mathbf{A}_2'\mathbf{y}_t)'}{n} + \frac{\sum_{t=1}^{n-1}(\mathbf{A}_2'\mathbf{y}_t)\mathbf{y}_t'}{n}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2) + (\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\frac{\sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{y}_t'}{n}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)||_2 \\
= \; & ||(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\frac{\sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{x}_{t2}'}{n} + \frac{\sum_{t=1}^{n-1}\mathbf{x}_{t2}\mathbf{y}_t'}{n}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2) + (\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\frac{\sum_{t=1}^{n-1}\mathbf{y}_t\mathbf{y}_t'}{n}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'||_2 \\
= \; & o_p(1).
\end{aligned} \tag{6.8}$$

Next, consider the case $p = o(n^c)$. Let $\boldsymbol{\varsigma}_t$ be a k-dimensional $I(1)$ process such that $\nabla \boldsymbol{\varsigma}_t = \mathbf{v}_t$. By Remark 2 of ZRY, we know that Condition 3 (i) and Remark 3 of ZRY hold for $\boldsymbol{\varsigma}_t$. Let $\mathbf{M}_1, \mathbf{M}_2$ be $k \times (p-r)$ and $k \times r$ matrices such that $\mathbf{M}$ given in (i) of Condition 3 satisfying $\mathbf{M}' = (\mathbf{M}_1, \mathbf{M}_2)$. Let $\mathbf{F}(t) = (F^1(t), \cdots, F^k(t))'$ be defined as in ZRY and $\bar{\boldsymbol{\varsigma}} = \frac{1}{n}\sum_{t=1}^n \boldsymbol{\varsigma}_t$, then

$$\begin{aligned}
& ||\frac{1}{n^2}\sum_{t=1}^{n}(\mathbf{x}_{t1} - \bar{\mathbf{x}}_1)(\mathbf{x}_{t1} - \bar{\mathbf{x}}_1)' - \mathbf{M}_1'\int_0^1 \mathbf{F}(t)\mathbf{F}'(t)\,dt\mathbf{M}_1||_2 \\
= \; & ||\mathbf{M}_1'\left(\frac{1}{n^2}\sum_{t=1}^{n}(\boldsymbol{\varsigma}_t - \bar{\boldsymbol{\varsigma}})(\boldsymbol{\varsigma}_t - \bar{\boldsymbol{\varsigma}})' - \int_0^1 \mathbf{F}(t)\mathbf{F}'(t)\,dt\right)\mathbf{M}_1||_2 = o_p(1).
\end{aligned} \tag{6.9}$$

By Remark 3 of ZRY, we have $\lambda_{\min}\left(\int_0^1 \mathbf{F}(t)\mathbf{F}'(t)\,dt\right) \geq 1/k$ in probability. Since $c_1 \leq \lambda_{\min}(\mathbf{M}) \leq \lambda_{\max}(\mathbf{M}) \leq c_2$, it follows $\lambda_{\min}\left(\mathbf{M}_1'\int_0^1 \mathbf{F}(t)\mathbf{F}'(t)\,dt\mathbf{M}_1'\right) \geq 1/k$ in probability. Further, for any given $j \geq 0$,

$$\begin{aligned}
& ||\frac{1}{n}\sum_{t=1}^{n-j}(\mathbf{x}_{t+j,2} - \bar{\mathbf{x}}_2)(\mathbf{x}_{t2} - \bar{\mathbf{x}}_2)' - \mathrm{Cov}(\mathbf{x}_{t+j,2}, \mathbf{x}_{t2})||_2 \\
= \; & ||\mathbf{M}_2'\Big(\frac{1}{n}\sum_{t=1}^{n}[(\mathbf{v}_{t+j} - \bar{\mathbf{v}})(\mathbf{v}_t - \bar{\mathbf{v}})' - \mathrm{Cov}(\mathbf{v}_{t+j}, \mathbf{v}_t)]\Big)\mathbf{M}_2||_2 = o_p(1), \text{ and}
\end{aligned} \tag{6.10}$$

$$\begin{aligned}
||\frac{1}{n^{3/2}}\sum_{t=1}^{n-j}(\mathbf{x}_{t+j,1} - \bar{\mathbf{x}}_2)(\mathbf{x}_{t2} - \bar{\mathbf{x}}_2)'||_2 &= ||\mathbf{M}_1'\left(\frac{1}{n^{3/2}}\sum_{t=1}^{n}(\boldsymbol{\varsigma}_{t+j} - \bar{\boldsymbol{\varsigma}})(\mathbf{v}_t - \bar{\mathbf{v}})'\right)\mathbf{M}_2||_2 \\
&= O_p(k/n^{1/2}),
\end{aligned} \tag{6.11}$$

where $\mathbf{v}_t$ is given in (i) of Condition 3.

By (6.9)–(6.11), similar to the proof of Theorem 3 in ZRY, it can be shown that when $k = o(n^{1/2-1/\eta})$,

$$||\widehat{\mathbf{A}}_2 - \mathbf{A}_2||_2 = O_p(p^{1/2}k/n). \tag{6.12}$$

Similar to (6.9), there exists a k-dimensional Gaussian process $\mathbf{w}(t)$ such that

$$||\frac{1}{n^2}\sum_{t=1}^{n}\mathbf{y}_t\mathbf{y}_t' - \mathbf{A}_1\mathbf{M}_1'\int_0^1 \mathbf{w}(t)\mathbf{w}'(t)\,dt\mathbf{M}_1\mathbf{A}_1'||_2 = o_p(1) \tag{6.13}$$

and similar to (6.11), we can show (6.6) holds provided $k/n^{1/2} \to 0$ as $n \to \infty$. Thus, by (6.12) and (6.13), we also have (6.8) and complete the proof of Lemma 5. □

Lemma 6. *Under Condition 1,*

$$||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)||_2 = o_p(1),$$

and under the conditions of Theorem 3,

$$||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)||_2 = O_p(p^{1/2}k^2/n^{1/2}). \tag{6.14}$$

Proof. When p is fixed, similar to (6.6), we have

$$\frac{1}{n^{3/2}}||\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'||_2 = o_p(1).$$

As a result, it follows from (6.7) that

$$||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)||_2 = o_p(1). \tag{6.15}$$

When p tends to infinity as $n \to \infty$, using the same idea as in (6.11), we can show

$$\frac{1}{n^{3/2}}||\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'||_2 = O_p(k/n^{1/2}). \tag{6.16}$$

Thus, by (6.12) and $p \le k = o(n^{1/2})$, it follows that

$$||\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)||_2 = O_p(p^{1/2}k^2/n^{1/2}).$$

Thus, we have Lemma 6. □

Lemma 7. *Let* $\mathbf{\Sigma} = \mathrm{E}\{[(\mathbf{f}_{t-1})', \cdots, (\mathbf{f}_{t-s})']'[(\mathbf{f}_{t-1})', \cdots, (\mathbf{f}_{t-s})']\} + diag(\mathbf{B}\Sigma_\varepsilon\mathbf{B}', \cdots, \mathbf{B}\mathbf{\Sigma}_\varepsilon\mathbf{B})$, *where* $\mathbf{\Sigma}_\varepsilon$ *is the variance of* $\boldsymbol{\varepsilon}_t$. *Under Condition 1 , for any given positive integer* s,

$$\frac{1}{n}\sum_{t=s}^{n}[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})']'[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})'] \xrightarrow{p} \mathbf{\Sigma} \tag{6.17}$$

and under the condition of Theorem 3, in probability

$$\frac{1}{n}\sum_{t=s}^{n}[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})']'[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})'] \ge \mathbf{\Sigma}, \tag{6.18}$$

where $\mathbf{A} \ge \mathbf{B}$ *means that* $\mathbf{A} - \mathbf{B}$ *is a nonnegative definition matrix.*

Proof. By some elementary computation, we have

$$\widehat{\mathbf{f}}_t = [\mathbf{f}_t + \mathbf{B}'\boldsymbol{\varepsilon}_t] + [(\widehat{\mathbf{B}} - \mathbf{B})'(\mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t)] + [\widehat{\mathbf{B}}'(\mathbf{D} - \widehat{\mathbf{D}})\mathbf{x}_{t2}] + [\widehat{\mathbf{D}}(\mathbf{A}_2 - \widehat{\mathbf{A}}_2)'\mathbf{y}_{t-1}] = \sum_{i=1}^{4} \boldsymbol{\zeta}_{t,i}. \quad (6.19)$$

Next, we first show (6.17) holds for fixed p. By (6.33) (see below), we have

$$||\widehat{\mathbf{B}} - \mathbf{B}||_2 = O_p(n^{-1/2}), \quad (6.20)$$

which gives

$$||\frac{1}{n}\sum_{t=s}^{n}(\boldsymbol{\zeta}'_{t-1,2}, \cdots, \boldsymbol{\zeta}'_{t-s,2})'(\boldsymbol{\zeta}'_{t-1,2}, \cdots, \boldsymbol{\zeta}'_{t-s,2})||_2 = o_p(1). \quad (6.21)$$

Similarly, by (6.29) (see below) and (6.7), we have

$$\sum_{i=3}^{4} ||\frac{1}{n}\sum_{t=s}^{n}(\boldsymbol{\zeta}'_{t-1,i}, \cdots, \boldsymbol{\zeta}'_{t-s,i})'(\boldsymbol{\zeta}'_{t-1,i}, \cdots, \boldsymbol{\zeta}'_{t-s,i})||_2 = o_p(1). \quad (6.22)$$

On the other hand, by large number law for α-mixing process, we get

$$\frac{1}{n}\sum_{t=s}^{n}(\boldsymbol{\zeta}'_{t-1,1}, \cdots, \boldsymbol{\zeta}'_{t-s,1})'(\boldsymbol{\zeta}'_{t-1,1}, \cdots, \boldsymbol{\zeta}'_{t-s,1}) \xrightarrow{p} \boldsymbol{\Sigma}. \quad (6.23)$$

Combining (6.21)–(6.23) yields that

$$\begin{aligned} & \frac{1}{n}\sum_{t=s}^{n}[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})']'[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})'] \\ = & \frac{1}{n}\sum_{t=s}^{n}(\sum_{i=1}^{4}\boldsymbol{\zeta}'_{t-1,i}, \cdots, \sum_{i=1}^{4}\boldsymbol{\zeta}'_{t-s,i})'(\sum_{i=1}^{4}\boldsymbol{\zeta}'_{t-1,i}, \cdots, \sum_{i=1}^{4}\boldsymbol{\zeta}'_{t-s,i}) \\ = & \frac{1}{n}\sum_{t=s}^{n}(\boldsymbol{\zeta}'_{t-1,1}, \cdots, \boldsymbol{\zeta}'_{t-s,1})'(\boldsymbol{\zeta}'_{t-1,1}, \cdots, \boldsymbol{\zeta}'_{t-s,1}) + o_p(1) \xrightarrow{p} \boldsymbol{\Sigma} \end{aligned}$$

and (6.17) follows.

Now, we turn to show the case with p varying with n. Since $p = o(n^{1/2})$, (6.23) still holds. Note that $\frac{1}{n}\sum_{t=s}^{n}(\boldsymbol{\zeta}'_{t-1,i}, \cdots, \boldsymbol{\zeta}'_{t-s,i})'(\boldsymbol{\zeta}'_{t-1,i}, \cdots, \boldsymbol{\zeta}'_{t-s,i}) \geq \mathbf{0}$ for $i = 1, \cdots, 4$. For the proof of (6.18), it is enough to show for all $1 \leq i \neq j \leq 4$,

$$||\frac{1}{n}\sum_{t=s}^{n}(\boldsymbol{\zeta}'_{t-1,i}, \cdots, \boldsymbol{\zeta}'_{t-s,i})'(\boldsymbol{\zeta}'_{t-1,j}, \cdots, \boldsymbol{\zeta}'_{t-s,j})||_2 = o_p(1). \quad (6.24)$$

We only give $i = 1, j = 4$ in details, other cases can be shown similarly. Since $\mathbf{y}_t = \mathbf{A}\mathbf{x}_t$, it follows from (2.1) that

$$\boldsymbol{\zeta}_{t,1} = \mathbf{B}'(\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1,2}) = \mathbf{B}'\mathbf{A}\mathbf{e}_t - \mathbf{B}'(\mathbf{D} + \mathbf{A}_2)\mathbf{x}_{t-1,2} = \mathbf{B}'\mathbf{A}\mathbf{M}\mathbf{v}_t - \mathbf{B}'(\mathbf{D} + \mathbf{A}_2)\mathbf{M}'_2\mathbf{v}_{t-1}.$$

Thus, by the fact that for any $-s-1 \le j \le s+1$,

$$||\sum_{t=1}^{n}\sum_{s=1}^{t}\mathbf{v}_s\mathbf{v}_{t+j}||_2 = O_p(kn) \tag{6.25}$$

and (6.12), we have the left-hand side of (6.24) is of order $O_p(p^{1/2}k^2/n) = o_p(1)$, where (6.25) holds because the components of $\mathbf{v}_t$ are independent. Thus, we have (6.18) and complete the proof of Lemma 7. □

Proof of Theorem 1. Let $\mathbf{b}_i, i = 1, \cdots, p$ be the rows of $\mathbf{B}$. Lemmas 5 and 6 implies that for any $1 \le i \le p$,

$$\begin{aligned}\sqrt{n}(\widehat{\mathbf{d}}_i - \mathbf{d}_i) &= \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\mathbf{b}_i\mathbf{f}_t + \varepsilon_t^i)\mathbf{y}_{t-1}'\mathbf{A}_2\right)\left(\frac{1}{n}\sum_{i=1}^{n}(\mathbf{A}_2'\mathbf{y}_{t-1})(\mathbf{A}_2'\mathbf{y}_{t-1})'\right)^{-1} + o_p(1)\\ &= \left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n}(\mathbf{b}_i\mathbf{f}_t + \varepsilon_t^i)\mathbf{x}_{t-1,2}'\right)\left(\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{x}_{t2}\mathbf{x}_{t2}'\right)^{-1} + o_p(1). \end{aligned}\tag{6.26}$$

Since $\{\mathbf{x}_{t2}\}$ is α mixing with mixing coefficients satisfying (6.2), it follows that

$$\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{x}_{t2}\mathbf{x}_{t2}' \xrightarrow{p} \mathrm{E}(\mathbf{x}_{t2}\mathbf{x}_{t2}') =: \mathbf{\Pi}. \tag{6.27}$$

On the other hand, by central limit theory (CLT) for α-mixing process $\{(\mathbf{b}_i\mathbf{f}_t + \varepsilon_t^i)\mathbf{x}_{t-1,2}', 1 \le i \le p\}$, there exists a $pr \times pr$ matrix $\mathbf{\Lambda}$ such that

$$\frac{1}{\sqrt{n}}\left(\sum_{t=1}^{n}(\mathbf{b}_1\mathbf{f}_t + \varepsilon_t^1)\mathbf{x}_{t-1,2}', \cdots, \sum_{t=1}^{n}(\mathbf{b}_p\mathbf{f}_t + \varepsilon_t^p)\mathbf{x}_{t-1,2}'\right) \xrightarrow{d} N(0, \mathbf{\Lambda}). \tag{6.28}$$

Thus, by (6.27) and (6.28), we have

$$\sqrt{n}(\mathrm{vech}(\widehat{\mathbf{D}}) - \mathrm{vech}(\mathbf{D})) \xrightarrow{d} N(0, \mathbf{\Pi}^{-1}\mathbf{\Lambda}\mathbf{\Pi}^{-1}). \tag{6.29}$$

Further, by (6.29) and (6.7), it is easy to show that

$$||\widehat{\mathbf{C}} - \mathbf{C}||_2 = ||(\widehat{\mathbf{D}} - \mathbf{D})\mathbf{A}_2' + \widehat{\mathbf{D}}(\widehat{\mathbf{A}}_2' - \mathbf{A}_2')||_2 = O_p(n^{-1/2}).$$

Next, we show (b) of Theorem 1. Observe that

$$\widehat{\mathbf{v}}_t = \nabla\mathbf{y}_t - \widehat{\mathbf{D}}\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1} = (\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1,2}) - (\widehat{\mathbf{D}} - \mathbf{D})[(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t-1} + \mathbf{x}_{t-1,2}] - \mathbf{D}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t-1},$$

which means that

$$
\begin{aligned}
& \frac{1}{n}\sum_{t=1}^{n-j}\left[\widehat{\mathbf{v}}_{t+j}\widehat{\mathbf{v}}_t' - \mathrm{E}(\nabla\mathbf{y}_{t+j} - \mathbf{D}\mathbf{x}_{t+j-1})(\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1})'\right] \\
= \; & \frac{1}{n}\sum_{t=1}^{n-j}\left[(\nabla\mathbf{y}_{t+j} - \mathbf{D}\mathbf{x}_{t+j-1})(\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1})' - \mathrm{E}(\nabla\mathbf{y}_{t+j} - \mathbf{D}\mathbf{x}_{t+j-1})(\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1})'\right] \\
& +(\widehat{\mathbf{D}} - \mathbf{D})\left(\frac{1}{n}\sum_{t=1}^{n-j}[(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t+j-1} + \mathbf{x}_{t+j-1,2}][(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t-1} + \mathbf{x}_{t-1,2}]'\right)(\widehat{\mathbf{D}} - \mathbf{D})' \\
& +\mathbf{D}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\left(\frac{1}{n}\sum_{t=1}^{n-j}\mathbf{y}_{t+j-1}\mathbf{y}_{t-1}'\right)(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\mathbf{D}' \\
& -\frac{1}{n}\sum_{t=1}^{n-j}(\nabla\mathbf{y}_{t+j} - \mathbf{D}\mathbf{x}_{t+j-1,2})\{[\mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2) + \mathbf{x}_{t-1,2}'](\widehat{\mathbf{D}} - \mathbf{D})' + \mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\mathbf{D}'\} \\
& -\frac{1}{n}\sum_{t=1}^{n-j}\{(\widehat{\mathbf{D}} - \mathbf{D})[(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t+j-1} + \mathbf{x}_{t+j-1,2}] + \mathbf{D}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t+j-1}\}(\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1,2})' \\
& +\frac{1}{n}\sum_{t=1}^{n-j}[(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'\mathbf{y}_{t+j-1}\mathbf{y}_{t-1}' + \mathbf{x}_{t+j-1,2}\mathbf{y}_{t-1}'](\widehat{\mathbf{A}}_2 - \mathbf{A}_2)\mathbf{D}' \\
& +\frac{1}{n}\sum_{t=1}^{n-j}\mathbf{D}(\widehat{\mathbf{A}}_2 - \mathbf{A}_2)'[\mathbf{y}_{t+j-1}\mathbf{y}_{t-1}'(\widehat{\mathbf{A}}_2 - \mathbf{A}_2) + \mathbf{y}_{t+j-1}\mathbf{x}_{t-1,2}'](\widehat{\mathbf{D}} - \mathbf{D})'. \qquad (6.30)
\end{aligned}
$$

By (6.7), (6.29) and the large lumber law, we have that the spectral norm of the last six terms of the right-hand side in (6.30) is $O_p(n^{-1})$. And by CLT of α mixing process, for any given j, the first term of the right-hand side of (6.30) is $O_p(n^{-1/2})$. Similarly, we can show

$$
\Big|\Big|\frac{1}{n}\sum_{t=1}^{n-j}\bar{\mathbf{v}}\widehat{\mathbf{v}}_t'\Big|\Big|_2 = O_p(n^{-1}). \qquad (6.31)
$$

Thus,

$$
||\widehat{\mathbf{\Sigma}}_v(j) - \mathbf{\Sigma}_v(j)||_2 = O_p(n^{-1/2}), \qquad (6.32)
$$

where $\mathbf{\Sigma}_v(j) = \mathrm{E}(\nabla\mathbf{y}_{t+j} - \mathbf{D}\mathbf{x}_{t+j-1})(\nabla\mathbf{y}_t - \mathbf{D}\mathbf{x}_{t-1})'$. Since j_0 is fixed, it follows from (6.32) that

$$
||\widehat{\mathbf{W}} - \sum_{j=1}^{j_0}\mathbf{\Sigma}_v(j)\mathbf{\Sigma}_v'(j)||_2 = O_p(n^{-1/2}). \qquad (6.33)
$$

Note that $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) = O_p(||\widehat{\mathbf{W}} - \sum_{j=1}^{j_0}\mathbf{\Sigma}_v(j)\mathbf{\Sigma}_v'(j)||_2)$ (see for example, Chang, Guo and Yao (2015)), we have (b) of Theorem 1 as desired.

Now, we turn to show (c). By (6.19), we get

$$\begin{aligned}
&\sum_{t=s}^{n}[\widehat{\mathbf{f}}_t - \sum_{i=1}^{s}\mathbf{E}_i\widehat{\mathbf{f}}_{t-i}][(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})'] \\
= &\sum_{t=s}^{n}\mathbf{e}_t\left[(\mathbf{f}_{t-1})' + \boldsymbol{\varepsilon}_{t-1}'\mathbf{B}, \cdots, (\mathbf{f}_{t-s})' + \boldsymbol{\varepsilon}_{t-s}'\mathbf{B}\right] \\
&+\sum_{t=s}^{n}\mathbf{e}_t[\mathbf{x}_{t-1,2}'(\mathbf{D}-\widehat{\mathbf{D}})'\widehat{\mathbf{B}}, \cdots, \mathbf{x}_{t-s,2}'(\mathbf{D}-\widehat{\mathbf{D}})'\widehat{\mathbf{B}}] \\
&+\sum_{t=s}^{n}\mathbf{e}_t[(\mathbf{B}\mathbf{f}_{t-1}+\boldsymbol{\varepsilon}_{t-1})'(\widehat{\mathbf{B}}-\mathbf{B}), \cdots, (\mathbf{B}\mathbf{f}_{t-s}+\boldsymbol{\varepsilon}_{t-s})'(\widehat{\mathbf{B}}-\mathbf{B})] \\
&+\sum_{t=s}^{n}\mathbf{e}_t[\mathbf{y}_{t-2}'(\mathbf{A}_2-\widehat{\mathbf{A}}_2)\widehat{\mathbf{D}}', \cdots, \mathbf{y}_{t-s-1}'(\mathbf{A}_2-\widehat{\mathbf{A}}_2)\widehat{\mathbf{D}}'] =: \sum_{i=1}^{4}\Delta_{ni}. \qquad (6.34)
\end{aligned}$$

By (6.7), (6.20) and (6.29), we can show that for any given positive integer s,

$$||\Delta_{n2}||_2 + ||\Delta_{n3}||_2 + ||\Delta_{n4}||_2 = O_p(1). \qquad (6.35)$$

On the other hand, since $\text{vech}\{\mathbf{e}_t\left[(\mathbf{f}_{t-1})' + \boldsymbol{\varepsilon}_{t-1}'\mathbf{B}, \cdots, (\mathbf{f}_{t-s})' + \boldsymbol{\varepsilon}_{t-s}'\mathbf{B}\right]\}$ is a α mixing process with finite 2γ-moment and mixing coefficients satisfying (6.2), it follows that there exists a positive matrix Γ such that

$$\text{vech}\left(\frac{1}{\sqrt{n}}\sum_{t=s}^{n}\mathbf{e}_t\left[(\mathbf{f}_{t-1})' + \boldsymbol{\varepsilon}_{t-1}'\mathbf{B}, \cdots, (\mathbf{f}_{t-s})' + \boldsymbol{\varepsilon}_{t-s}'\mathbf{B}\right]\right) \xrightarrow{d} N(0, \Gamma_1). \qquad (6.36)$$

Note that

$$\begin{pmatrix} \widehat{\mathbf{E}}_1 \\ \cdots \\ \widehat{\mathbf{E}}_s \end{pmatrix} = \left(\sum_{t=s}^{n}[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})']'[(\widehat{\mathbf{f}}_{t-1})', \cdots, (\widehat{\mathbf{f}}_{t-s})']\right)^{-1} \begin{pmatrix} \sum_{t=s}^{n}\widehat{\mathbf{f}}_{t-1}\mathbf{e}_t' \\ \cdots \\ \sum_{t=s}^{n}\widehat{\mathbf{f}}_{t-s}\mathbf{e}_t' \end{pmatrix}. \qquad (6.37)$$

Thus, by (6.34)–(6.36) and Lemma 7, we have (c) and complete the proof of Theorem 1. □

Next, we first develop bounds for the estimated eigenvalues $\widehat{\lambda}_j$, $j = 1, 2, \cdots p$.

Lemma 8. *Let λ_j, $j = 1, \cdots, p$ be the eigenvalues of $\mathbf{W}_v$. Under Condition 1 or conditions of Theorem 3,*

$$|\widehat{\lambda}_m - \lambda_m| = O_p(pn^{-1/2}) \quad \textit{and} \quad |\widehat{\lambda}_{m+1}| = O_p(pn^{-1/2}). \qquad (6.38)$$

Proof. By (b) of Theorem 1 and (b) of Theorem 3, we have for any $1 \le i \le p$,

$$|\widehat{\lambda}_i - \lambda_i| \le ||\widehat{\mathbf{W}}_v - \mathbf{W}_v||_2 = O_p(pn^{-1/2}) \quad \text{and} \quad \lambda_{m+1} = \cdots = \lambda_p = 0.$$

This gives Lemma 8 as desired. □

Proof of Theorem 2. It is enough to show that

$$\lim_{n\to\infty} P\{\tilde{m} < m\} = 0. \tag{6.39}$$

Suppose $\tilde{m} < m$ is true, then by Lemma 8, there exists a positive constant c_1 such that

$$\lim_{n\to\infty} P\{\widehat{\lambda}_{\tilde{m}+1}/\widehat{\lambda}_{\tilde{m}} \geq c_1\} = 1, \quad \text{and} \quad \lim_{n\to\infty} P\{\widehat{\lambda}_{m+1}/\widehat{\lambda}_m < c_1/2\} = 1.$$

This implies that

$$\lim_{n\to\infty} P\{\widehat{\lambda}_{\tilde{m}+1}/\widehat{\lambda}_{\tilde{m}} > \widehat{\lambda}_{m+1}/\widehat{\lambda}_m\} = 1,$$

which contradicts the definition of $\widetilde{m}$. Thus, (6.39) holds. □

Proof of Theorem 3. Since $p = o(n^{1/2})$ and $\{\mathbf{x}_{t2}\}$ is a α mixing process with mixing coefficients satisfying (3.2), it follows that (6.27) also holds for this case. Further, note that for any $1 \leq i \leq p$ and $1 \leq j \leq r$, applying CLT of mixing process to $\{(\mathbf{b}_i\mathbf{f}_t + \varepsilon_t^i)x_{t-1,2}^j\}$, which is a α mixing process with coefficients satisfying (3.2), we get

$$|\sum_{t=1}^{n}(\mathbf{b}_i\mathbf{f}_t + \varepsilon_t^i)x_{t-1,2}^j| = O_p(\sqrt{n}),$$

which implies

$$||\frac{1}{n}\sum_{t=1}^{n}(\mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t)\mathbf{x}_{t-1,2}'||_2 = O_p(n^{-1/2}(pr)^{1/2}). \tag{6.40}$$

Thus, by Lemmas 5 and 6,

$$\begin{aligned}
||\widehat{\mathbf{D}} - \mathbf{D}||_2 &= \left\| \left(\frac{1}{n}\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{y}_{t-1}'\widehat{\mathbf{A}}_2\right)\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\widehat{\mathbf{A}}_2\right)^{-1} - \mathbf{D}\right\|_2 \\
&= \left\| \left(\frac{1}{n}\sum_{t=1}^{n}\nabla\mathbf{y}_t\mathbf{x}_{t-1,2}'\right)\left(\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{x}_{t-1,2}\mathbf{x}_{t-1,2}'\right)^{-1} - \mathbf{D}\right\|_2 + O_p(p^{1/2}k^2/n) \\
&= \left\| \left(\frac{1}{n}\sum_{t=1}^{n}(\mathbf{B}\mathbf{f}_t + \boldsymbol{\varepsilon}_t)\mathbf{x}_{t-1,2}'\right)\left(\frac{1}{n}\sum_{i=0}^{n-1}\mathbf{x}_{t-1,2}\mathbf{x}_{t-1,2}'\right)^{-1}\right\|_2 + O_p(p^{1/2}k^2/n) \\
&= O_p(n^{-1/2}(pr)^{1/2} + p^{1/2}k^2/n),
\end{aligned} \tag{6.41}$$

this combining with (6.12) yields

$$||\widehat{\mathbf{C}} - \mathbf{C}||_2 = ||(\widehat{\mathbf{D}} - \mathbf{D})\mathbf{A}_2' + \widehat{\mathbf{D}}'(\widehat{\mathbf{A}}_2' - \mathbf{A}_2')||_2 = O_p(n^{-1/2}(pr)^{1/2} + p^{1/2}k^2/n). \tag{6.42}$$

Thus, (a) of Theorem 3 follows from (6.41) and (6.42).

Next, we show (b). It is easy to see that

$$\|\frac{1}{n^2}\sum_{t=1}^{n-j}\mathbf{y}_{t-1}\mathbf{y}_{t-1}'\|_2 = O_p(p). \tag{6.43}$$

Thus, by (6.12), (6.41) and (iii) of Condition 3, it can be shown that $||\cdot||_2$ norm of the last six terms of the right-hand side in (6.30) are of order $o(pn^{-1/2})$, provided $k = o(n^{1/2})$ and $p = O(n^{1/4})$. On the other hand, applying CLT of α mixing process to the first term of the right-hand side of (6.30), we get for any given j, this term is of order $O_p(pn^{-1/2})$. Similarly, we can show $n^{-1}\sum_{t=1}^{n-j}\bar{\mathbf{v}}\bar{\mathbf{v}}_t' = O_p(n^{-1/2}p)$. Thus,

$$||\widehat{\mathbf{\Sigma}}_v(j) - \mathbf{\Sigma}_v(j)||_2 = O_p(n^{-1/2}p). \tag{6.44}$$

Since j_0 is fixed, it follows from (6.44) that

$$||\widehat{\mathbf{W}} - \sum_{j=1}^{j_0}\mathbf{\Sigma}_v(j)\mathbf{\Sigma}_v'(j)||_2 = O_p(n^{-1/2}p). \tag{6.45}$$

Note that $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) = O_p(||\widehat{\mathbf{W}} - \sum_{j=1}^{j_0}\mathbf{\Sigma}_v(j)\mathbf{\Sigma}_v'(j)||_2)$ (see for example, Chang, Guo and Yao (2015)), we have (b) of Theorem 3 as desired.

In the following, we give the proof of (c). Let Δ_{ni}, $i = 1, 2, 3, 4$ be defined as in (6.34). Since $\{\mathbf{e}_t[(\mathbf{B}\mathbf{f}_{t-1} + \boldsymbol{\varepsilon}_{t-1})', \mathbf{x}_{t-1,2}']\}$ is α mixing with mixing coefficients satisfying (3.2), it follows from conclusion (a) and (b) of Theorem 3 that

$$||\Delta_{n2} + \Delta_{n3}||_2 = O_p\left(n^{1/2}(pr)^{1/2}[n^{-1/2}(pr)^{1/2} + p^{1/2}k^2/n + pn^{-1/2}]\right). \tag{6.46}$$

By (6.12) and a similar argument as in (6.24), we have

$$||\Delta_{n4}||_2 = O_p(p^{1/2}k^2). \tag{6.47}$$

Applying CLT of α mixing to the elements of Δ_{n1}, we get

$$||\Delta_{n1}||_2 = O_p((pmn)^{1/2}). \tag{6.48}$$

Combining equations (6.46)–(6.48) with Lemma 7 and $p = o(n^{1/2})$ yield

$$||(\mathbf{E}_1, \cdots, \mathbf{E}_s)||_2 = O(p^{1/2}k^2n^{-1} + pm^{1/2}n^{-1/2}), \tag{6.49}$$

this gives (c) and completes the proof of Theorem 3. □

Proof of Theorem 4. By Lemma 8, Theorem 4 can be shown similarly as for Theorem 2. Therefore, we omit the detailed proofs. □

Proof of Remark 1. Since the proofs are similar, we only show the case with fixed p in details. It follows from the definition of $\widehat{m}$ that

$$\sum_{j=\widehat{m}+1}^{p} \widehat{\lambda}_j + \widehat{m}\omega_n \leq \sum_{j=m}^{p} \widehat{\lambda}_{p+1-j} + m\omega_n. \tag{6.50}$$

Suppose that $\widehat{m} > m$, it follows from (6.50) that

$$(\widehat{m} - m)\omega_n \leq \sum_{j=m+1}^{\widehat{m}} \widehat{\lambda}_j \leq (\widehat{m} - m)\widehat{\lambda}_{m+1}. \tag{6.51}$$

Since $\omega_n / n^{-1/2} \to \infty$, it follows from Lemma 8 that equation (6.51) holds with probability zero. This gives that

$$\lim_{n\to\infty} P\{\widehat{m} > m\} = 0. \tag{6.52}$$

On the other hand, if $\widehat{m} < m$, equation (6.50) yields

$$(m - \widehat{m})\widehat{\lambda}_m \leq \sum_{j=\widehat{m}+1}^{m} \widehat{\lambda}_j \leq (m - \widehat{m})\omega_n. \tag{6.53}$$

Lemma 8 implies $\widehat{\lambda}_m \geq \lambda_m/2 > 0$. Thus, by (6.53) and $\omega_n \to 0$ as $n \to \infty$, we have

$$\lim_{n\to\infty} P\{\widehat{m} < m\} = 0. \tag{6.54}$$

Equation (6.52) together with (6.54) give the consistency of $\widehat{m}$ as desired. □

References

Ahn, S.K. and Reinsel, G.C. (1988). Nested Reduced-rank Autoregressive Models for Multiple Time Series. *Journal of the American Statistical Association*, **83**, 849–856.

Athanasopoulos, G. and Vahid, F. (2008). VARMA vers VAR for macroeconomic forecasting. *Journal of Business & Economic Statistics*, **26**, 237–252.

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

Bai, J. (2009). Panel data models with interactive fixed effects. *Econometrica*, **77**, 1229–1279.

Box, G. and Tiao, G. (1977). A canonical analysis of multiple time series. *Biomatrika*, **64**, 355–365.

Chang, J. Y., Guo, B. and Yao, Q. (2015). Segmenting multiple time series by contemporaneous linear transformation. *A manuscript*.

Chao, J. and Phillips, P. C. B. (1999). Model selection in partially nonstationary vector autoregressive processes with reduced rank structure. *Journal of Econometrics*, **91**, 227–271.

Clements, M.P. and Hendry, D.F. (1993). On the limitations of comparing mean squared forecast errors (with discussion). *Journal of Forecasting*, **12**, 617–637.

Engle, R. and Granger, C. W. J. (1987). Cointegration and error correction: representation, estimation and testing. *Econometrica*, **55**, 251–276.

Engle, R. F. and J.V. Issler (1995). Estimating common sectoral cycles. *Journal of Monetary Economics*, **35**, 83–113.

Engle, R. F., and Kozicki, S. (1993). Testing for Common Features (with comments). *Journal of Business & Economic Statistics*, **11**, 369–395.

Engle, R. and Yoo, S. (1987). Forecasting and testing in cointegrated systems. *Journal of Econometrics*, **35**, 143–159.

Escribano, A. and Peña, D. (1994). Cointegration and Common Factors. *Journal of Time Series Analysis*, **15**, 577–586.

Fan, J. and Yao, Q. (2015). *The Elements of Financial Econometrics*. Science China Press, Beijing.

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

Forni, M., Hallin, M., Lippi, M. and Reichlin, L. (2004). The generalized dynamic factor model: consistency and rates. *Journal of Econometrics*, **119**, 231–255.

Gonzalo, J. and Pitarakis, J. (1995). Comovements in Large Systems. Working Paper, Department of Economics, Boston University.

Granger, C. W. J. (1981). Some properties of time series data and their use in econometric model specification. *Journal of Econometrics*, **16**, 121–130.

Granger, C. W. J. and Weiss, A. A. (1983). Time series analysis of error-correcting models. In: S. Karlln et al. (eds.), *Studies in Econometrics, Time series and Multivariate Analysis*, Academic Press, New York, 255–278.

Ho, M. S. and Sørensen, B. E. (1996). Finding Cointegration Rank in High Dimensional Systems Using the Johansen Test: An Illustration Using Data Based Monte Carlo Simulations. *Review of Economics and Statistics*, **78**, 726–732.

Hualde, J. and Robinson, P. (2010). Semiparametric inference in multivariate fractionally cointegrated system. *Journal of Econometrics*, **157**, 492–511.

Issler, J. V. and Vahid, F. (2001). Common cycles and the importance of transitory shocks to macroeconomic aggregates. *Journal of Monetary Economics*, **47**, 449–475.

Johansen, S. (1991). Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive model. *Econometrica*, **59**, 1551-1580.

Johansen, S. (1995). *Likelihood-Based inference in Cointegrated Vector in Gaussian Vector Autoregressive Model*. Oxford University Press, Oxford.

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

Liao, Z. and Phillips, P. C. B. (2015). Automated Estimation of Vector Error Correction Models, *Econometric Theory*, **31**, 581-646.

Lin, J. L. and Tsay, R. S. (1996). Cointegration constraints and forecasting: An empirical examination. *Journal of Applied Econometrics*, **11**, 519–538.

Peña D. and Poncela P. (2004). Forecasting with nonstationary dynamic factor models. *Journal of Econometrics*, **119**, 291–321.

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

Stock, J. H. and Watson, M. W. (2002a). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 1167–79.

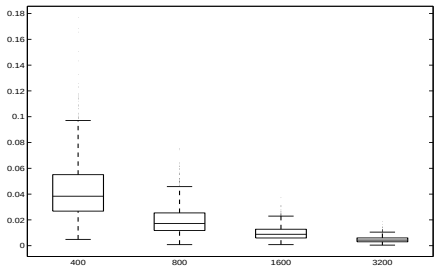
Stock, J.H. and M.W. Watson (2002b). Macroeconomic forecasting using di?usion indexes. *Journal of Business & Economic Statistics*, **20**, 147–162.

Stock, J. H. and M. W. Watson (2008). Forecasting in Dynamic Factor Models Subject to Struc- tural Instability, in *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry*, Jennifer Castle and Neil Shephard (eds), Oxford: Oxford University Press.

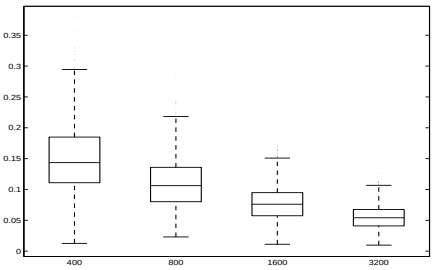
Vahid, F. and Issler, J. V. (2002). The importance of common cyclical features in VAR analysis: A Monte-Carlo study. *Journal of Econometrics*, **109**, 341–363.

Zhang, R. M., Robinosn, P. and Yao, Q. (2015). Identifying Cointegration by Eigenanalysis. *A Manuscript.*

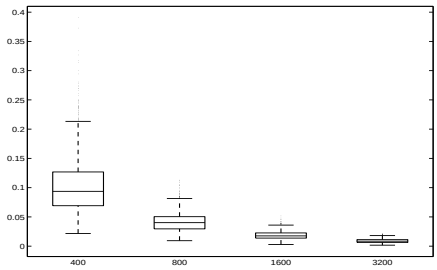
Figure 1: Boxplot of $D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\mathbf{A}_2))$ (left panel) and $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B}))$ (right panel), $400 \leq n \leq 3200$



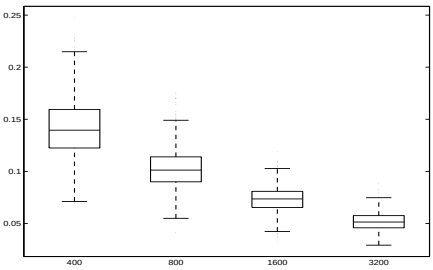
(a) $p = 5, r = 1, m = 1$



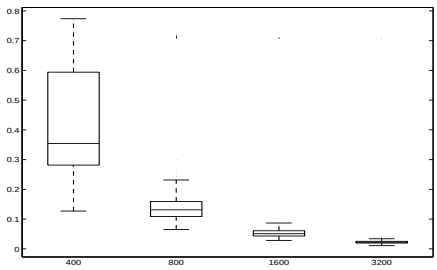
(b) $p = 5, r = 1, m = 1$



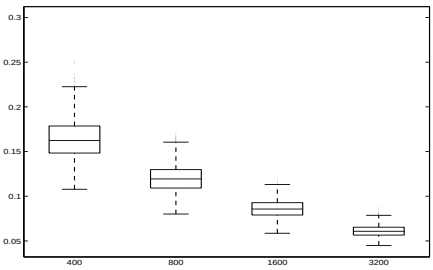
(c) $p = 10, r = 1, m = 4$



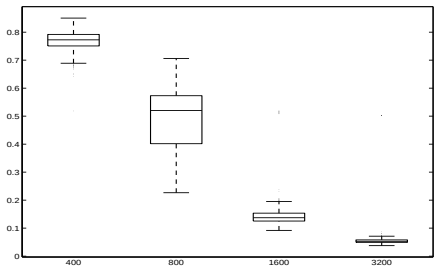
(d) $p = 10, r = 1, m = 4$



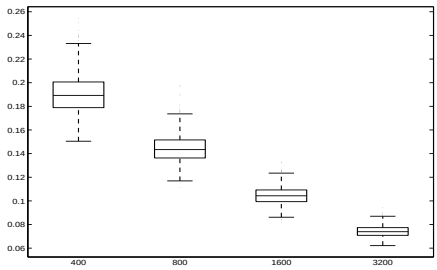
(e) $p = 20, r = 2, m = 4$



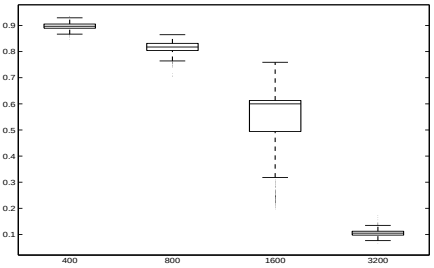
(f) $p = 20, r = 2, m = 4$



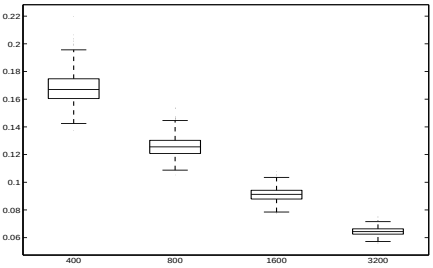
(g) $p = 40, r = 4, m = 8$



(h) $p = 40, r = 4, m = 8$



(i) $p = 60, r = 4, m = 8$



(j) $p = 60, r = 4, m = 8$