Statistica Sinica **30** (2020), 1463-1484
doi:https://doi.org/10.5705/ss.202017.0250

ERROR-CORRECTION FACTOR MODELS FOR HIGH-DIMENSIONAL COINTEGRATED TIME SERIES

Yundong Tu, Qiwei Yao and Rongmao Zhang

Peking University, London School of Economics and Zhejiang University

Abstract: Cointegration inferences often rely on a correct specification for the short-run dynamic vector autoregression. However, this specification is unknown, a priori. A lag length that is too small leads to an erroneous inference as a result of the misspecification. In contrast, using too many lags leads to a dramatic increase in the number of parameters, especially when the dimension of the time series is high. In this paper, we develop a new methodology which adds an error-correction term for the long-run equilibrium to a latent factor model in order to model the short-run dynamic relationship. The inferences use the eigenanalysis-based methods to estimate the cointegration and latent factor process. The proposed error-correction factor model does not require an explicit specification of the short-run dynamics, and is particularly effective for high-dimensional cases, in which the standard error-correction suffers from overparametrization. In addition, the model improves the predictive performance of the pure factor model. The asymptotic properties of the proposed methods are established when the dimension of the time series is either fixed or diverging slowly as the length of the time series goes to infinity. Lastly, the performance of the model is evaluated using both simulated and real data sets.

Key words and phrases: Cointegration, eigenanalysis, factor models, nonstationary processes, vector time series.

1. Introduction

Cointegration refers to the existence of 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); Engle and Granger (1987), cointegration has attracted increasing attention in the fields of econometrics and statistics. An excellent survey on early works on cointegration can be found in Johansen (1995).

To date, considerable effort has been devoted to inferences on the long-run trend (cointegration) restrictions in vector autoregressions (VARs); see, among others, Engle and Granger (1987); Johansen (1991); Phillips (1991), for estima-

tion and testing, and Engle and Yoo (1987); Lin and Tsay (1996) for forecasting. As shown in Engle and Granger (1987), a VAR with cointegration restrictions can be represented by a vector error-correction model (VECM) that reflects the correction on the long-run relationship using short-run dynamics. One of the remarkable features of a VECM is that it identifies clearly the gain in prediction from using the cointegrated variables over that of the standard ARIMA approach, as noted by Engle and Yoo (1987); Lin and Tsay (1996); Peña and Poncela (2004). However, it does require that we specify a finite autoregressive order for the short-run dynamic before an 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. In particular, 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 the AIC, BIC, and HQIC have been applied to determine both the autoregressive order and the cointegration rank; see, for example, Chao and Phillips (1999); Athanasopoulos et al. (2011). While appealing to practitioners, these methods are nevertheless subject to pre-test biases and post model selection inferential errors (Liao and Phillips (2015)). Furthermore, a VECM is ineffective when the dimension of the time series is high, largely as a result of the overparametrization of the VAR specification.

Relative to the considerable number of studies on long-run restrictions, one may argue that the importance of short-run restrictions has not received due attention in the cointegration literature. On the other hand, common cyclical movements exist extensively in the field of macroeconomics. For example, Engle and Kozicki (1993) found common international cycles in GNP data for OECD countries. Issler and Vahid (2001) reported common cycles for macroeconomic aggregates and sectoral and regional outputs in the United States. It has been shown that using (short-run) rank restrictions in a stationary VAR can improve its short-term forecasting ability, as documented by Ahn and Reinsel (1988); Vahid and Issler (2002); 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 utilize it in their subsequent inference procedure. Issler and Vahid (2001) used a similar argument to cointegration for the short-run effect. Based on a VECM, they proposed modeling the common cycles using

sample squared canonical correlations. In addition, they use Johansen's likelihood method to identify the cointegration relationship.

When the dimension of a time series is high, VAR models suffer from having too many parameters, even after imposing rank restrictions. Furthermore, most classical inference methods for cointegration, including Johansen's likelihood method, either do not work, or do not work effectively; see the numerical studies reported in Gonzalo and Pitarakis (1995); Ho and Sørensen (1996). Although high-dimensional problems exist extensively in macroeconomic and financial data, the development of related theory and methodologies in the context of cointegration is still in its infancy.

We propose an error-correction factor model (ECFM) that identifies the linear dynamic structures, in a parsimonious and robust fashion, in a high-dimensional cointegrated series. Specifically, the long-run equilibrium relationship among all nonstationary components is represented by a cointegration vector, that is, the correction term to the equilibrium. This term is then utilized to improve a factor representation for the short-run dynamics of the differenced processes. In contrast to the classical VECM, our setting does not require explicitly specifying the short-run dynamics, thus avoiding erroneous inferences on cointegration due to, for example, a misspecification of the autoregressive order.

Factor models have beome a popular way of modeling high-dimensional time series in order to achieve dimension reduction; see, for example, Bai (2004); Bai and Ng (2004); Banerjee, Marcellino and Masten (2014a,b); Barigozzi, Lippi and Luciani (2016a,b). In this paper, we adopt a latent and low-dimensional factor process to represent the high-dimensional short-run dynamics. Compared with the pure factor model, the cointegration term improves the modelling and the prediction for short run dynamics. For inferences, we first adopt the eigenanalysis-based method of Zhang, Robinosn and Yao (2019) (ZRY, hereafter) to identify both the cointegration rank and 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 an appropriate time series model. Owning to the errors that accumulate during the estimation, fitting a dynamic model for the factor process turns out to be an error-in-observation problem in autoregression. This problem has not been thoroughly investigated in the literature; thus we propose a version of the corrected Yule-Walker method (see Section 2.2.3).

The proposed methodology is further supported by the newly established asymptotic theory and numerical evidence. In particular, our numerical results corroborate the findings from the asymptotic theory. The results of Monte Carlo simulation show that the cointegration rank, cointegration space, number of factors, and factor co-feature space can all be estimated reasonably well with typical sizes of observed samples. Our empirical example on forecasting 12 U.S. industrial production indices shows that the proposed ECFM outperforms both the VECM and the univariate AR models for each component in the post-sample forecasting, for most forecast horizons considered.

The rest of the paper is organized as follows. We describe the proposed ECFM and the associated estimation methods in Section 2. In Section 3, the asymptotic properties of the estimation methods are established with the dimension of the time series fixed or diverging slowly, as the length of the time series goes to infinity. The proposed methodology is illustrated numerically in Section 4 using both simulated and real data sets. Furthermore, we compare the forecasting performance of the proposed ECFM to that of the reduced-rank VECM and the univariate AR models for each component. The forecasting performance for the real data is evaluated for different forecast horizons based on the criterion of Clements and Hendry (1993). Section 5 concludes the paper. All technical lemmas and proofs are provided in the online Supplementary Material.

2. Methodology

2.1. ECFMs

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. We denote the difference operator as ∇, that is, $\nabla\mathbf{u}_t = \mathbf{u}_t - \mathbf{u}_{t-1}$. We use the convention $\nabla^0\mathbf{u}_t = \mathbf{u}_t$. A process $\mathbf{u}_t$ is said to be a weakly integrated process with order 1, abbreviated as weak $I(1)$, if $\nabla\mathbf{u}_t$ is weakly stationary with a finite spectral density and is positive definite at frequency 0, but $\mathbf{u}_t$ itself is not. Because we deal only with weak $I(1)$ processes in this study, we refer to them as weakly integrated processes.

Let $\mathbf{y}_t$ be an observable $p \times 1$ weak $I(1)$ process with initial values $\mathbf{y}_t = 0$, for $t \leq 0$. Suppose that cointegration exists; that is, there are r (≥ 1) stationary linear combinations of $\mathbf{y}_t$, where r is the cointegration rank, and is often unknown.

The ECFM 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, $\mathbf{C}\mathbf{y}_t$ is weakly stationary, $\mathbf{f}_t$ is an $m \times 1$ weakly stationary process, $\mathbf{B}$ is a $p \times m$ matrix, and $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ white noise with mean zero and covariance matrix Σ_ε, and uncorrelated with $\mathbf{y}_{t-1}$ and $\{\mathbf{f}_t\}$. In contrast to VECM, (2.1) represents the short-run dynamics using the latent process $\mathbf{f}_t$. Its linear dynamic structure is completely unspecified. Note that $\mathbf{f}_t$ does not enter the inference for the error-correction term $\mathbf{C}\mathbf{y}_{t-1}$. Model (2.1) is particularly useful when p is large and m is small, which is often the case with many real data sets, because it leads to effective dimension reduction when modeling high-dimensional time series.

Without loss of generality, in (2.1), we assume $\mathbf{B}$ is an orthogonal matrix, that is, $\mathbf{B}'\mathbf{B} = \mathbf{I}_m$, where $\mathbf{I}_m$ denotes the $m \times m$ identity matrix. This is because 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. Thus we can replace $(\mathbf{B}, \mathbf{f}_t)$ in (2.1) with $(\mathbf{Q}, \mathbf{U}\mathbf{f}_t)$.

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 $p \times r$ matrices. Furthermore, the columns of $\mathbf{A}_2$ are the cointegration vectors, and 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, $\mathbf{D}$ can then be uniquely determined. Thus, to fit model (2.1), we need 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 forecasting can be based on a fitted time series model for $\mathbf{f}_t$.

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

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

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 $\mathbf{M}(\mathbf{A}_2)$, that is, 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 $\mathbf{A} = (\mathbf{A}_1, \mathbf{A}_2)$ is 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 components of $\mathbf{x}_{t,1}$ are weak $I(1)$.

We adopt the eigenanalysis-based method proposed by ZRY to estimate r and $\mathbf{A}_2$. 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 ensure that each term in the sum is nonnegative definite, and that there is no information cancellation over different lags. Let $\widetilde{\lambda}_1 \geq \cdots \geq \widetilde{\lambda}_p$ be the eigenvalues of $\widehat{\mathbf{W}}$, and let $\widetilde{\boldsymbol{\gamma}}_1, \ldots, \widetilde{\boldsymbol{\gamma}}_p$ be the corresponding eigenvectors. Then, $\mathbf{A}_2$ is estimated by $\widehat{\mathbf{A}}_2 = (\widetilde{\boldsymbol{\gamma}}_{p-r+1}, \ldots, \widetilde{\boldsymbol{\gamma}}_p)$, and the cointegration rank is estimated by

$$\widehat{r} = \arg\min_{1 \leq l \leq p} IC(l), \tag{2.2}$$

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 (because we allow ω_n to be data dependent). ZRY has shown that $\mathcal{M}(\widehat{\mathbf{A}}_2)$ and $\widehat{r}$ are consistent estimators for $\mathcal{M}(\mathbf{A}_2)$ and r, respectively.

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 $\mathbf{d}_i$, for $i = 1, \ldots, p$ be a row vector of $\mathbf{D}$, and let $\nabla\mathbf{y}_t = (\nabla y_t^1, \ldots, \nabla y_t^p)'$. The least squares estimator for $\mathbf{d}_i$ is defined as

$$\widehat{\mathbf{d}}_i = \arg\min_{\mathbf{d}_i} \sum_{t=1}^{n} (\nabla y_t^i - \mathbf{d}_i \widehat{\mathbf{A}}_2'\mathbf{y}_{t-1})^2, \tag{2.3}$$

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 $\mathbf{f}_t$ using the residuals $\widehat{\mathbf{v}}_t \equiv \nabla \mathbf{y}_t - \widehat{\mathbf{D}}\widehat{\mathbf{A}}_2'\mathbf{y}_{t-1}$, for $t = 1, \ldots, n$. To this end, let

$$\widehat{\mathbf{W}}_v = \sum_{j=1}^{j_0} \widehat{\mathbf{\Sigma}}_v(j)\widehat{\mathbf{\Sigma}}_v'(j), \tag{2.4}$$

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

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

One advantage of using the quadratic form $\widehat{\mathbf{\Sigma}}_v(j)\widehat{\mathbf{\Sigma}}_v(j)'$ instead of $\widehat{\mathbf{\Sigma}}_v(j)$ in (2.4) is that there is no information cancellation over different lags. Therefore, this approach is insensitive to the choice of j_0 in (2.4). Often, small values, such as $j_0 = 5$, are sufficient to identify the relevant characteristics, because serial dependence is usually predominant at small lags; see Lam and Yao (2012); Chang, Guo and Yao (2015). Let $(\widehat{\gamma}_1, \ldots, \widehat{\gamma}_m)$ 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

$$\widehat{\mathbf{B}} = (\widehat{\gamma}_1, \ldots, \widehat{\gamma}_m), \quad \text{and} \quad \widehat{\mathbf{f}}_t = \widehat{\mathbf{B}}'\widehat{\mathbf{v}}_t. \tag{2.5}$$

Because m is usually unknown and the last $p - m$ eigenvalues of $\widehat{\mathbf{W}}_v$ may not be exactly zero owning to the random fluctuation, we need to determine m. We propose doing so 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:

$$\widetilde{m} = \arg\min_{1 \leq i \leq R} \frac{\widehat{\lambda}_{i+1}}{\widehat{\lambda}_i}, \tag{2.6}$$

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

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

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

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. Then it can be shown that as $\omega_n \to 0$ and $\omega_n n^{1/2}/p \to \infty$, $\widehat{m}$ is consistent for m.

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 $\mathbf{f}_t$ with a VAR model.

Let

$$\mathbf{f}_t = \sum_{i=1}^{s} \mathbf{E}_i \mathbf{f}_{t-i} + \mathbf{e}_t, \tag{2.7}$$

where $\mathbf{E}_i$, for $1 \le i \le s$ is an $m \times m$ matrix and $\{\mathbf{e}_t\}$ is a sequence of independent vectors with mean zero and that are independent of $\{\mathbf{x}'_{t2}, \mathbf{f}'_t, \boldsymbol{\varepsilon}'_t\}$. In our setting, $\mathbf{f}_t$ contains unobservable latent factors, and is estimated by $\widehat{\mathbf{f}}_t = \widehat{\mathbf{B}}'\widehat{\mathbf{v}}_t$ as given in (2.5). It can be shown that

$$\widehat{\mathbf{f}}_t = \mathbf{f}_t + \mathbf{B}'\boldsymbol{\varepsilon}_t + \sum_{i=2}^{4} \boldsymbol{\zeta}_{t,i}.$$

If we ignore the term $\sum_{i=2}^{4} \boldsymbol{\zeta}_{t,i}$, $\widehat{\mathbf{f}}_t$ can be viewed as the observation of $\mathbf{f}_t$ with a measurement error. Thus, $\mathbf{E}_i$ can be estimated using a VAR model with observations in errors. This is an interesting and important topic, and has been actively pursued in various contexts; seex for example, Carroll, Ruppert and Stefanski (1995). However, time series models with measurement errors have not received sufficient attention. Note that when $\widehat{\mathbf{f}}_t = \mathbf{f}_t + \mathbf{B}'\boldsymbol{\varepsilon}_t$, (2.7) can be written as a vector ARMA model (VARMA) with the same order of AR and MA parts. We can estimate $\mathbf{E}_i$ using a VARMA model. An alternative method is to use the

classic least squares procedure, which estimates $\mathbf{E}_i$ based on $\{\widehat{\mathbf{f}}\}$; that is,

$$(\widetilde{\mathbf{E}}_1, \ldots, \widetilde{\mathbf{E}}_s) = \operatorname{argmin}_{\mathbf{E}_1,\ldots,\mathbf{E}_s} \sum_{t=s+1}^{n} ||\widehat{\mathbf{f}}_t - \sum_{i=1}^{s} \mathbf{E}_i \widehat{\mathbf{f}}_{t-i}||^2. \tag{2.8}$$

However, just as in a simple linear regression for independent data, $\widetilde{\mathbf{E}}_i$ can not be used to estimate $\mathbf{E}_i$ consistently when the spectral norm of the covariance of $\mathbf{B}'\boldsymbol{\varepsilon}_t + \sum_{i=2}^{4} \boldsymbol{\zeta}_{t,i}$ has the same order as that of $\widehat{\mathbf{f}}_t$ and a correcting factor is required. To see this, we simply assume $\widehat{\mathbf{f}}_t = \mathbf{f}_t + \mathbf{B}'\boldsymbol{\varepsilon}_t$ and $s = 1$; then,

$$\begin{aligned}
\widetilde{\mathbf{E}}_1' - \mathbf{E}_1' &= (\sum_{t=2}^{n} \widehat{\mathbf{f}}_{t-1}\widehat{\mathbf{f}}'_{t-1})^{-1} \sum_{t=1}^{n} (\mathbf{f}_{t-1} + \mathbf{B}'\boldsymbol{\varepsilon}_{t-1})(\mathbf{e}_t + \mathbf{B}'\boldsymbol{\varepsilon}_t - \mathbf{E}_1\mathbf{B}'\boldsymbol{\varepsilon}_{t-1})' \\
&= (\sum_{t=2}^{n} \widehat{\mathbf{f}}_{t-1}\widehat{\mathbf{f}}'_{t-1})^{-1} \sum_{t=1}^{n} [(\mathbf{f}_{t-1} + \mathbf{B}'\boldsymbol{\varepsilon}_{t-1})\mathbf{e}_t' + \mathbf{f}_{t-1}(\boldsymbol{\varepsilon}_t'\mathbf{B} - \boldsymbol{\varepsilon}_{t-1}\mathbf{B}\mathbf{E}_1')] \\
&\quad -(\sum_{t=2}^{n} \widehat{\mathbf{f}}_{t-1}\widehat{\mathbf{f}}'_{t-1})^{-1} \sum_{t=1}^{n} \mathbf{B}'\boldsymbol{\varepsilon}_{t-1}\boldsymbol{\varepsilon}_{t-1}'\mathbf{B}\mathbf{E}_1'.
\end{aligned}$$

Under some regular condition, $(\sum_{t=1}^{n-1} \widehat{\mathbf{f}}_t\widehat{\mathbf{f}}'_t)^{-1} \sum_{t=1}^{n-1} \mathbf{B}'\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t'\mathbf{B} \xrightarrow{p} [\operatorname{Var}(\mathbf{f}_1 + \mathbf{B}'\boldsymbol{\varepsilon}_1)]^{-1} \operatorname{Var}(\mathbf{B}'\boldsymbol{\varepsilon}_1)$. Thus, a corrected factor is required, and we can use the modified LSE: $\widehat{\mathbf{E}}_1' = [\operatorname{Var}(\mathbf{f}_1)]^{-1}[\operatorname{Var}(\mathbf{f}_1 + \mathbf{B}'\boldsymbol{\varepsilon}_1)]\widetilde{\mathbf{E}}_1'$ to estimate $\mathbf{E}_1'$. One simple method is to correct the LSE given in (2.8) by

$$\begin{aligned}
(\widehat{\mathbf{E}}_1, \ldots, \widehat{\mathbf{E}}_s)' = &\left[\sum_{t=s+1}^{n} (\widehat{\mathbf{f}}'_{t-1}, \ldots, \widehat{\mathbf{f}}'_{t-s})'(\widehat{\mathbf{f}}'_{t-1}, \ldots, \widehat{\mathbf{f}}'_{t-s}) - M\right]^{-1} \\
&\left[\sum_{t=s+1}^{n} \widehat{\mathbf{f}}_t(\widehat{\mathbf{f}}'_{t-1}, \ldots, \widehat{\mathbf{f}}'_{t-s})\right]',
\end{aligned} \tag{2.9}$$

where $M = \operatorname{diag}(\widehat{\Sigma}_{\mathbf{B}\boldsymbol{\varepsilon}}(1), \ldots, \widehat{\Sigma}_{\mathbf{B}\boldsymbol{\varepsilon}}(s))$ and $\widehat{\Sigma}_{\mathbf{B}\boldsymbol{\varepsilon}}(i) = \sum_{t=s+1}^{n} \mathbf{B}'\boldsymbol{\varepsilon}_{t-i}\boldsymbol{\varepsilon}_{t-i}'\mathbf{B}$. This is in the same spirit as the corrected Yule–Walker estimator proposed by Staudenmayer and Buonaccorsi (2005) for the AR model with a measurement error. The autoregressive order s can be determined using standard criteria such as the AIC or BIC; see, for example, Section 4.2.3 of Fan and Yao (2015).

Combining (2.1), (2.7) and (2.9), we have the following h-step-ahead forecast,

for $h = 1, 2$:

$$\begin{aligned}
\mathbf{y}_{t+1|t} &= (\mathbf{I} + \widehat{\mathbf{C}})\mathbf{y}_t + \widehat{\mathbf{B}}\widehat{\mathbf{f}}_{t+1} = (\mathbf{I} + \widehat{\mathbf{C}})\mathbf{y}_t + \widehat{\mathbf{B}}\left(\sum_{i=1}^{s} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i}\right), \\
\mathbf{y}_{t+2|t} &= (I + \widehat{\mathbf{C}})\mathbf{y}_{t+1|t} + \widehat{\mathbf{B}}\widehat{\mathbf{f}}_{t+2|t} \\
&= (\mathbf{I} + \widehat{\mathbf{C}})^2\mathbf{y}_t + (\mathbf{I} + \widehat{\mathbf{C}})\widehat{\mathbf{B}}\left(\sum_{i=1}^{s} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i}\right) \\
&\quad + \widehat{\mathbf{B}}\left[\sum_{i=1}^{s-1} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i} + \widehat{\mathbf{E}}_1\left(\sum_{i=1}^{s} \widehat{\mathbf{E}}_i \widehat{\mathbf{f}}_{t+1-i}\right)\right].
\end{aligned} \tag{2.10}$$

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. For a given m, we calculate the distance between the co-feature space $\mathcal{M}(\mathbf{B})$ and its estimate as

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

Then, $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B})) \in [0, 1]$, taking the value zero if and only if $\mathcal{M}(\widehat{\mathbf{B}}) = \mathcal{M}(\mathbf{B})$, and one if and only if $\mathcal{M}(\widehat{\mathbf{B}})$ and $\mathcal{M}(\mathbf{B})$ are orthogonal. We consider two asymptotic modes: (i) p is fixed and $n \to \infty$; and (ii) both p and n diverge, but r is fixed.

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

We introduce the regularity conditions first.

Condition 1. The process $\{\mathbf{x}'_{t2}, \nabla\mathbf{y}'_t, \boldsymbol{\varepsilon}'_t\}$ is a stationary α-mixing process with mean zero, $\mathrm{E}\|(\mathbf{x}'_{t2}, \nabla\mathbf{y}'_t, \boldsymbol{\varepsilon}'_t)\|_\infty^{4\gamma} < \infty$ for some constant $\gamma > 1$, and the mixing coefficients α_t satisfy the condition $\sum_{t=1}^{\infty} \alpha_t^{1-1/\gamma} < \infty$, where $\|\mathbf{x}\|_\infty$ denotes the maximum norm of a vector $\mathbf{x} = (x_1, \ldots, x_n)$, that is, $\|\mathbf{x}\|_\infty = \max(|x_1|, \ldots, |x_n|)$.

Condition 2. The characteristic polynomial of VAR model (2.7) has no roots on or outside of the unit circle so that it is a causal VAR model.

Theorem 1. *Let Condition* 1 *hold.*

(a) Let $vech(\mathbf{D}) = (\mathbf{d}_1, \ldots, \mathbf{d}_p)'$. *As* $n \to \infty$ *and* p *is fixed, it holds that*

$$\sqrt{n}(vech(\widehat{\mathbf{D}}) - vech(\mathbf{D})) \xrightarrow{d} N(0, \mathbf{\Omega}_1),$$

where $\mathbf{\Omega}_1$ *is an* $rp \times rp$ *positive-definite matrix,* $||\widehat{\mathbf{C}} - \mathbf{C}||_2 = O_p(n^{-1/2})$, *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) In addition, if Condition 2 *and* $\mathrm{E}\|\mathbf{e}_t\|^{2\gamma} < \infty$ *hold, then*

$$||(\widehat{\mathbf{E}}_1 - \mathbf{E}_1, \ldots, \widehat{\mathbf{E}}_s - \mathbf{E}_s)||_2 = O_p(n^{-1/2}).$$

Theorem 2. *Let* $1 \le m < p$ *and Condition* 1 *hold. For* $\widetilde{m}$ *defined in* (2.6),

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

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

Let $z_t^j \equiv \nabla x_t^j$, for $j = 1, \ldots, p - r$, and let $\mathbf{z}_t = (z_t^1, \ldots, z_t^{p-r})'$ and $\boldsymbol{\nu}_t = (\mathbf{z}_t', \mathbf{x}_{t2}')'$. In this subsection, we extend the asymptotic results of 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 $k \ge p$ and $c_1 \le \lambda_{\min}(\mathbf{M}) \le \lambda_{\max}(\mathbf{M}) \le c_2$, where c_1, c_2 are two positive constants. Suppose that $\boldsymbol{\nu}_t = \mathbf{M}\mathbf{v}_t$, and all components of $\mathbf{v}_t = (v_t^1, \ldots, v_t^k)'$ are independent and have mean zero.

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

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

(iii) $c_3 \le \lambda_{\min}(\mathbf{D}) \le \lambda_{\max}(\mathbf{D}) \le c_4$, for some positive constants c_3, 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) $||(\widehat{\mathbf{E}}_1 - \mathbf{E}_1, \ldots, \widehat{\mathbf{E}}_s - \mathbf{E}_s)||_2 = O_p((pm)^{1/2}n^{-1/2} + p^{1/2}k^2n^{-1})$, *provided that Condition 2 and* $\mathrm{E}||\mathbf{e}_t||^\theta < \infty$ *also hold.*

Theorem 4. *Let* $1 \le m < p$ *and Condition* 3 *hold with* $k = o(n^{1/2-1/\eta})$ *and* $p = O\left(n^{1/2-1/\eta}/(\log n)^2\right)$. *Then, for* $\widetilde{m}$ *defined in* (2.6), *we have*

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

Remark 2. 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 using a Monte Carlo simulation. We then illustrate the forecasting ability of the proposed ECFM using a real data example.

4.1. Monte Carlo simulation

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 that is first drawn elementwise from $U[0,1]$ independently, and is then orthogonalized. In addition, $\mathbf{x}_t = (\mathbf{x}_{t1}', \mathbf{x}_{t2}')'$, where the r components of $\mathbf{x}_{t2}$ are independent Gaussian AR(1) processes with identical autoregressive coefficients equal to 0.5, and the $(p-r)$ vector $\mathbf{x}_{t1}$ is $I(1)$, according to a factor-augmented AR(1) defined as

$$\mathbf{x}_{t1} = \mathbf{x}_{t-1,1} + \boldsymbol{\Upsilon}\mathbf{f}_t + \boldsymbol{e}_t. \tag{4.1}$$

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 $\mathbf{A}$. Furthermore, the components of $\mathbf{f}_t$ are independent stationary Gaussian AR(1) with identical autoregressive coefficients equal to 0.5, and $\boldsymbol{e}_t$ is independent and $N(0, \mathbf{I}_{p-r})$. Then, it is easy 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 $n = 100, 200, 400, 800, 1{,}200, 1{,}600, 2{,}000, 2{,}400$, and then estimate $r, \mathbf{C}, m$ and $\mathbf{B}$. To estimate r, we use the IC criterion (2.2) with the penalty $w_n = \log n\widetilde{\lambda}_p$. The number of factors m is estimated using the ratio method (2.6), with $j_0 = 5$. For each setting, we replicated the experiment 1,000 times.

Tables 1–3 list the relative frequencies of the occurrence of the events $(\widehat{r} = r)$

Table 1. Relative frequencies ($\times 100$) of the occurrences of events $\widehat{r} = r$ (first entries in parentheses) and $\widetilde{m} = m$ (second entries in parentheses).

$p = 5$		$n = 100$	$n = 200$	$n = 400$	$n = 800$
$m = 1$	$r = 1$	(92.0, 93.5)	(100 , 99.3)	(100 , 99.9)	(100 ,100)
	$r = 2$	(44.6, 89.3)	(68.5, 96.6)	(83.7, 99.8)	(98.6,100)
$p = 10$		$n = 200$	$n = 400$	$n = 800$	$n = 1,200$
$m = 1$	$r = 1$	(85.3,100)	(100 ,100)	(100 ,100)	(100 ,100)
	$r = 2$	(65.4,100)	(82.0,100)	(95.4,100)	(99.6,100)
$m = 2$	$r = 1$	(86.5, 82.2)	(100 , 97.7)	(100 , 99.9)	(100 ,100)
	$r = 2$	(62.4, 83.4)	(75.1, 97.8)	(94.3,100)	(98.8,100)
$p = 20$		$n = 400$	$n = 800$	$n = 1,200$	$n = 1,600$
$m = 2$	$r = 2$	(85.5, 99.7)	(92.8,100)	(96.7,100)	(98.9,100)
	$r = 4$	(20.5, 95.0)	(43.3, 99.8)	(68.8,100)	(86.3,100)
$m = 4$	$r = 2$	(82.0, 93.2)	(89.5, 99.9)	(93.8, 99.9)	(96.3,100)

Table 2. Relative frequencies ($\times 100$) of the occurrences of events $\widehat{r} = r$ (first entries in parentheses) and $\widetilde{m} = m$ (second entries in parentheses).

$p = 40$		$n = 800$	$n = 1,200$	$n = 1,600$	$n = 2,000$
$m = 2$	$r = 2$	(72.8,100)	(94.7,100)	(100 ,100)	(100 ,100)
	$r = 4$	(64.0, 99.9)	(99.5,100)	(99.3,100)	(99.7,100)
	$r = 6$	(86.4, 93.8)	(95.2, 98.9)	(96.2, 99.7)	(97.5,100)
	$r = 8$	(53.8,100)	(77.4,100)	(82.2,100)	(89.6,100)
$m = 4$	$r = 2$	(73.3,100)	(89.5,100)	(99.9,100)	(100 ,100)
	$r = 4$	(66.8, 99.9)	(99.3,100)	(99.5,100)	(99.2,100)
	$r = 6$	(75.1, 99.5)	(88.3,100)	(89.5,100)	(91.0,100)
	$r = 8$	(27.1, 99.7)	(59.0,100)	(64.4,100)	(75.9,100)
$m = 6$	$r = 2$	(72.7, 99.6)	(86.2,100)	(99.6,100)	(100 ,100)
	$r = 4$	(69.2, 96.5)	(98.6, 99.4)	(98.3,100)	(98.4,100)
	$r = 6$	(65.6, 99.7)	(83.1,100)	(86.1,100)	(88.8,100)
	$r = 8$	(16.9, 98.7)	(41.3,100)	(50.8,100)	(62.4,100)
$m = 8$	$r = 2$	(73.7, 99.9)	(81.1,100)	(99.8,100)	(100 ,100)
	$r = 4$	(71.0, 89.1)	(98.3, 99.2)	(98.2, 99.9)	(98.0,100)
	$r = 6$	(60.8, 98.7)	(82.1, 99.9)	(82.1,100)	(87.0,100)
	$r = 8$	(12.7, 83.7)	(37.0, 96.5)	(45.3, 98.5)	(52.6, 99.6)

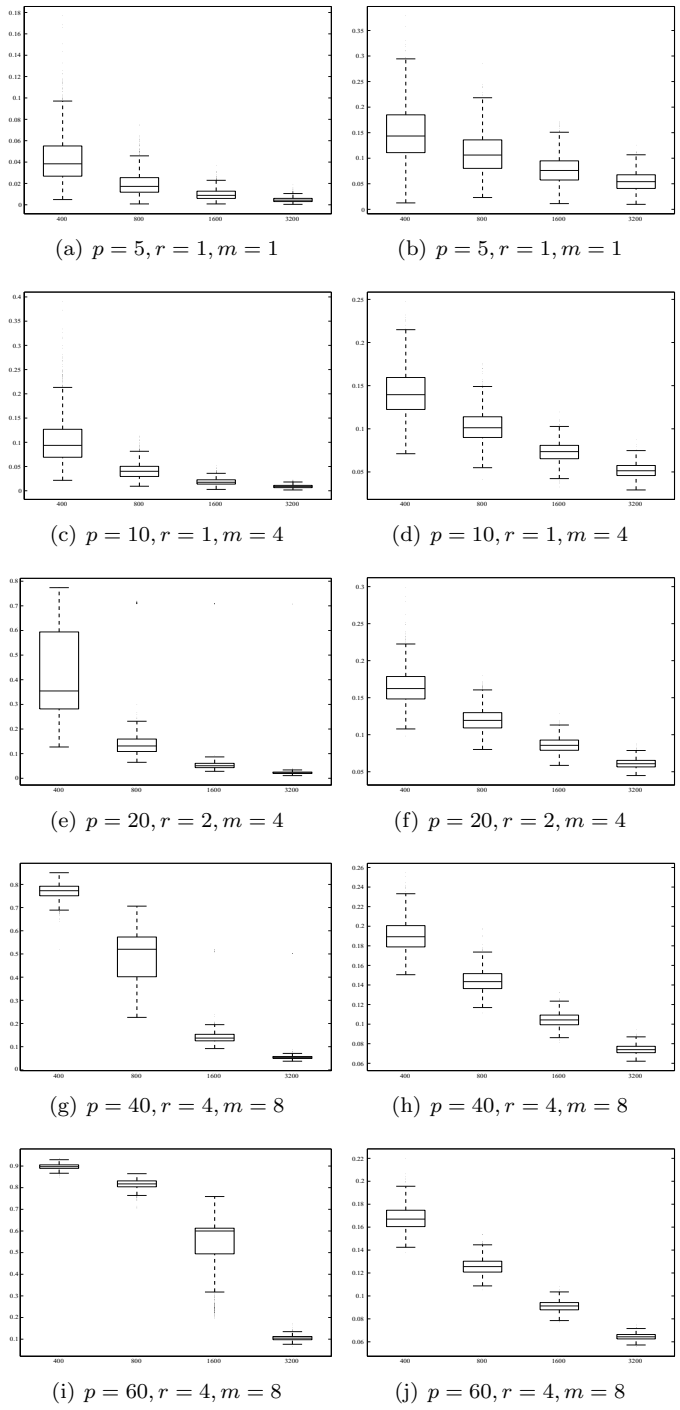
and $(\widetilde{m} = m)$ in a simulation with 1,000 replications. We make the following observations from Table 1 which contains the results for $p = 5, 10$ and 20. First, for $p = 5$ or 10, the relative frequencies for the correct specification of the cointegration rank r and the number of factors m are as high as 85%, even for a sample size n as small as 200. When n increases to 400, those relative frequencies increase to 100%. Second, with fixed n and r, the correct estimation rates for m increases as the dimension p increases, a phenomenon known as the "blessing of

Table 3. Relative frequencies ($\times 100$) of the occurrences of events $\widehat{r} = r$ (first entries in parentheses) and $\widetilde{m} = m$ (second entries in parentheses).

$p = 60$		$n = 1,200$	$n = 1,600$	$n = 2,000$	$n = 2,400$
$m = 2$	$r = 2$	(20.8,100)	(34.3,100)	(97.3,100)	(100 ,100)
	$r = 4$	(16.2,100)	(87.3,100)	(100 ,100)	(99.9,100)
	$r = 6$	(63.4,100)	(99.1,100)	(99.5,100)	(99.5,100)
	$r = 8$	(88.4,100)	(98.9,100)	(97.5,100)	(97.1,100)
	$r = 10$	(72.0,100)	(92.4,100)	(89.7,100)	(89.6,100)
$m = 4$	$r = 2$	(19.8,100)	(23.3,100)	(94.3,100)	(99.9,100)
	$r = 4$	(16.7,100)	(78.4,100)	(100 ,100)	(100 ,100)
	$r = 6$	(59.3,100)	(97.7,100)	(99.1,100)	(98.7,100)
	$r = 8$	(80.1,100)	(95.3,100)	(92.7,100)	(92.5,100)
	$r = 10$	(51.0,100)	(77.8,100)	(73.4,100)	(71.5,100)
$m = 6$	$r = 2$	(20.4,100)	(29.6,100)	(86.6,100)	(99.5,100)
	$r = 4$	(13.4,100)	(72.5,100)	(99.8,100)	(100 ,100)
	$r = 6$	(58.9,100)	(97.2,100)	(98.6,100)	(98.1,100)
	$r = 8$	(73.3,100)	(91.7,100)	(87.0,100)	(87.0,100)
	$r = 10$	(29.9,100)	(62.5,100)	(59.2,100)	(57.2,100)
$m = 8$	$r = 2$	(20.7,100)	(24.9,100)	(79.3,100)	(99.3,100)
	$r = 4$	(33.2,100)	(70.1,100)	(99.5,100)	(99.7,100)
	$r = 6$	(59.3,100)	(95.6,100)	(98.8,100)	(98.2,100)
	$r = 8$	(67.9,100)	(89.9,100)	(84.3,100)	(85.4,100)
	$r = 10$	(23.7, 99.7)	(54.0,100)	(50.9,100)	(51.6,100)
$m = 10$	$r = 2$	(20.3,100)	(21.2,100)	(76.6,100)	(98.5,100)
	$r = 4$	(33.8,100)	(65.8,100)	(99.4,100)	(100 ,100)
	$r = 6$	(60.0,100)	(94.7,100)	(98.7,100)	(98.3,100)
	$r = 8$	(61.6,100)	(87.6,100)	(84.7,100)	(85.4,100)
	$r = 10$	(18.6, 99.9)	(49.5,100)	(48.0,100)	(48.2,100)

dimensionality". This is consistent with the findings of Lam and Yao (2012), who dealt with purely stationary processes only. Third, the inference on r tends to become more challenging as p increases. For example, the relative frequency for a 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 of ZRY. Lastly, note that an increase in p, r and/or 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 $c \in (0, 1/2)$.

Similar conclusions can be drawn from the results reported in Table 2–3. In particular, the inference on the number of factors (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 $n = 200, 400$ corroborate this conclusion. However,



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

Figure 1. Box plot 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), $400 \leq n \leq 3{,}200$.

the inference on the cointegration rank is more difficult when n is small and/or r is large.

To evaluate the performance of the estimation for both the cointegration space and the factor cofeature space, we present the box plots for $D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\mathbf{A}_2))$ and $D(\mathcal{M}(\widehat{\mathbf{B}}), \mathcal{M}(\mathbf{B}))$ in Figure 1 for a few (selected) combinations of p, r and m, with $n = 400, 800, 1{,}600, 3{,}200$. The overall profile of the estimation accuracy is similar to those in Tables 1–3. For example, when p increases, the estimation accuracy of cointegration space becomes worse, while that of factor co-feature space tends to improve. That is, the "curse of dimensionality" in inferring the cointegration space is coupled with the "blessing of dimensionality" in estimating the factor co-feature space. Furthermore, in general the estimation improves as n increases, which confirms our consistency theory.

Next, we investigate how the autoregressive coefficient for the process of $x_{t,2}$, denoted by ρ, and the error variance of this autoregressive process, denoted by σ^2, affect the performance of the proposed method. For brevity, we report only the results for the case with $p = 60$, $m = 2, 4, 6$, $r = 2, 4$, $\rho = 0.8, 0.93$, and $\sigma^2 = 4, 8$ in Table 4. We observe the following. First, the error variance σ^2 increases, the selection for the factor seems to deteriorate. Second, as ρ increases, the performance of the cointegration rank selection procedure deteriorates, especially when ρ reaches 0.93. This drawback is inherent to the problem, and is one from which most methods suffer. Nevertheless, we observe that the performance of our procedure generally improves as the sample size increases.

4.2. A real data example

To further illustrate the proposed approach, we apply the proposed ECFM to 12 U.S. industrial production (manufacturing nondurable) monthly indices in January 1972 – August 2010, extracted from Stock and Watson (2008)[1]: namely, food, beverage, tobacco, textile mills, textile product mills, apparel, leather and allied product, paper, printing and related support activities, petroleum and coal products, chemical, plastics and rubber products. The estimated cointegration rank is $\widehat{r} = 2$, and the number of factors is $\widetilde{m} = 3$. We also fit the data using a VECM and Johansen's trace test to determine the cointegration rank r for each given autoregressive order between 1 and 8, and then using the AIC to select the optimal autoregressive order. The corresponding estimated cointegration rank is also 2. Hence, both fitted models suggest the same cointegration rank of 2, and the VECM represents the short-run dynamics in terms of a 12-dimensional vector

[1]The data are available at `http://www.princeton.edu/~mwatson/`.

Table 4. Relative frequencies (×100) of the occurrences of events $\widehat{r} = r$ (first entries in parentheses) and $\widetilde{m} = m$ (second entries in parentheses).

		$n = 1,200$	$n = 2,400$	$n = 1,200$	$n = 2,400$
$p = 60$		$\rho = 0.5, \sigma^2 = 4$		$\rho = 0.5, \sigma^2 = 8$	
$m = 2$	$r = 2$	(19.9,100)	(88.9,100)	(22.3, 99.7)	(20.7,100)
	$r = 4$	(26.2,100)	(62.3,100)	(25.5, 99.8)	(46.1,100)
$m = 4$	$r = 2$	(23.4,100)	(86.8,100)	(21.9, 93.9)	(23.3, 98.1)
	$r = 4$	(26.4, 99.9)	(52.4,100)	(23.8, 96.1)	(40.1, 99.7)
$m = 6$	$r = 2$	(23.4,100)	(79.9,100)	(25.1, 81.9)	(26.0, 98.6)
	$r = 4$	(28.2, 99.7)	(50.4,100)	(21.9, 81.5)	(39.3, 99.9)
		$\rho = 0.8, \sigma^2 = 1$		$\rho = 0.93, \sigma^2 = 1$	
$m = 2$	$r = 2$	(21.3,100)	(44.4,100)	(20.4,100)	(40.6,100)
	$r = 4$	(27.0,100)	(89.9,100)	(25.1,100)	(31.0,100)
$m = 4$	$r = 2$	(24.3,100)	(46.5,100)	(22.5,100)	(39.2,100)
	$r = 4$	(24.6,100)	(89.0,100)	(25.0,100)	(30.2,100)
$m = 6$	$r = 2$	(24.8,100)	(49.6,100)	(24.2,100)	(40.7,100)
	$r = 4$	(27.3,100)	(90.1,100)	(23.2,100)	(31.5,100)

AR(3) process (with reduced rank 2). In contrast, the newly proposed ECFM captures these dynamics in a three-dimensional latent-factor process, achieving a significant reduction in the number of parameters required, as can be seen from (2.1). The difference between the cointegration space estimated by our ECFM and that produced by Johansen's method is computed as

$$D(\mathcal{M}(\widehat{\mathbf{A}}_2), \mathcal{M}(\widetilde{\mathbf{A}}_2))^2 = 1 - \frac{1}{2}\mathrm{tr}\{\widehat{\mathbf{A}}_2\widehat{\mathbf{A}}_2'(\widetilde{\mathbf{A}}_2(\widetilde{\mathbf{A}}_2'\widetilde{\mathbf{A}}_2)^{-1}\widetilde{\mathbf{A}}_2)'\} = 0.0009,$$

where the 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 method. This suggests that the cointegration spaces estimated by the two approaches are effectively equivalent.

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 that of (i) univariate AR (UAR) models, with the lag length for each component selected using the standard Schwarz criterion, and (ii) the reduced-rank VECM with the rank and lag length selected simultaneously using the Hannan–Quinn criterion, and the cointegration rank chosen using PIC (Athanasopoulos et al. (2011)). For each of the last 10% of the data points, we fit the models using the data up to the previous month, and forecast the values using the three fitted models. Following Athanasopoulos et al. (2011), we measure the

Table 5. Percentage improvement in forecast accuracy measures: US IP indices.

Horizon (h)	ECFM versus VECM			ECFM versus UAR		
	|MSFE|	TMSFE	GFESM	|MSFE|	TMSFE	GFESM
1	-1.0	-1.9	-0.9	68.2	-3.5	68.2
4	61.2	2.5	12.4	94.8	32.6	90.8
8	40.6	-0.3	-2.1	97.2	47.5	97.1
12	83.5	2.6	55.4	98.8	54.0	98.5
16	93.9	9.0	83.3	99.1	56.7	99.3

forecast accuracy using the 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, for $h = 1, \ldots, 16$. We also calculate the generalized forecast error second moment (GFESM), that is, 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, as proposed by Clements and Hendry (1993). The GFESM is invariant to elementary operations that involve different variables, as well as to elementary operations that involve the same variable at different horizons. The forecasting comparison results are presented in Table 5, with the maximum lag lengths for order selection set as 4. The results for the maximum lag length set as 8 or 12 are very similar, and therefore are not presented here.

Table 5 shows that the ECFM provides forecasts that are more accurate than those of the reduced-rank VECM and the univariate AR models for most horizons. For example, for a 12-month-ahead forecast, the ECFM achieves improvement in the TMSFE, |MSFE| and GFESM of 98.8%, 54.0%, 98.5%, respectively, compared with those of the univariate AR models. In addition, the improvement from using ECFM over univariate AR models tends to increase as the forecast horizon increases. The improvement from using the ECFM over the reduced-rank VECM is obvious, especially for long horizons, but seems to be insignificant for short horizon predictions. These findings together illustrate the superiority of the ECFM in terms of forecasting.

5. Conclusion

Traditionally, cointegration inferences are built on a correct specification for the short-run dynamic vector auto-regression. It is known that choosing too short a lag length leads to size distortions, whereas choosing too many lags leads to a significant increase in the number of parameters, especially in high-dimensional

systems. To avoid this misspecification, and to address the co-feature information in the short-run dynamic, we propose modeling the dynamic relationship using a dynamic factor model and estimating the VECM using a two-step eigenanalysis: the first step estimates the long-run coefficients using the estimated cointegration space (Zhang, Robinosn and Yao (2019)); the second step estimates the loading matrix and the common factors for the short-run dynamic using a principle component analysis. The asymptotic theory and numerical studies show that the proposed procedure performs well.

The following shortcomings will be addressed in future research. First, in order to apply the result of Zhang, Robinosn and Yao (2019), the dimension p cannot be too large (i.e., not greater than $O(n^{1/4})$). It would be interesting and more challenging to consider 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 with the sparse shrinkage technique by solving the following optimal problem:

$$\widehat{\mathbf{C}} = \operatorname{argmin}_{\mathbf{C}\in R^{p\times p}} \left\{ \sum_{t=1}^{n} ||\nabla \mathbf{y}_t - \mathbf{C}\mathbf{y}_{t-1}||^2 + \lambda_n |||\mathbf{C}|||_{s_1} \right\}, \tag{5.1}$$

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

Second, because the focus of this paper is on predictions and inferences for the co-features, 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 the original $\mathbf{C}$ and $\mathbf{f}_t$. Because $\mathbf{C}\mathbf{y}_{t-1}$ and $\mathbf{f}_t$ may be correlated, the inference method proposed here will lead to inconsistent estimators. It would be interesting to consider an inference based on iterative equations as in Bai (2009). That is, estimate $\{\mathbf{C}, \mathbf{F}, \mathbf{B}\}$ using the least squares loss, defined as

$$\mathrm{SSR}(\mathbf{C}, \mathbf{F}, \mathbf{B}) = \sum_{t=1}^{n} (\nabla \mathbf{y}_t - \mathbf{C}\mathbf{y}_{t-1} - \mathbf{B}\mathbf{f}_t)'(\nabla \mathbf{y}_t - \mathbf{C}\mathbf{y}_{t-1} - \mathbf{B}\mathbf{f}_t), \tag{5.2}$$

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

Finally, our approach is relevant only if there exists a low-dimensional factor structure. Thus, it is pertinent to develop appropriate tests for the existence of such structure.

Supplementary Material

The online supplementary material contains useful lemmas and the proofs of the main theorems.

Acknowledgments

The authors would like to thank the Co-editor, Ruey Tsay, the Associate Editor, and several anonymous referees for their helpful comments and suggestions. Tu would like to acknowledge the support provided by National Natural Science Foundation of China (NSFC Grant 71472007, 71532001 and 71671002), and that from the Center for Statistical Science, Peking University, and Key Laboratory of Mathematical Economics and Quantitative Finance (Peking University), Ministry of Education. Yao received support from EPSRC grant EP/L01226X/1, and Zhang (corresponding author) received support from NSFC (Grant 11371318 and 11771390), Zhejiang Provincial Natural Science Foundation of China (Grant No. R16A010001), and the Fundamental Research Funds for the Central Universities, Ministry of Education, China.

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 versus 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. (2004). Estimating cross-section common stochastic trends in nonstationary panel data. *Journal of Econometrics* **122**, 137–183.

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

Bai, J. and Ng, S. (2004). A PANIC attack on unit roots and cointegration. *Econometrica* **72**, 1127–1177.

Banerjee, A., Marcellino, M. and Masten, I. (2014a). Forecasting with factor augmented error correction models. *International Journal of Forecasting* **30**, 589–612.

Banerjee, A., Marcellino, M. and Masten, I. (2014b). Structural FECM: Cointegration in large-scale structural FAVAR models. Working Paper 9858, CEPR.

Barigozzi, M., Lippi, M. and Luciani, M. (2016a). Dynamic factor models, cointegration, and error correction mechanisms. *FEDS*, 2016-018.

Barigozzi, M., Lippi M. and Luciani, M. (2016b). Non-stationary dynamic factor models for large- datasets. *FEDS*, 2016-024.

Box, G. and Tiao, G. (1977). A canonical analysis of multiple time series. *Biometrika* **64**,

355–365.

Carroll, R., Ruppert, D. and Stefanski, L. (1995). *Measurement Error in Nonlinear Models.* Chapman & Hall, London.

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 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.

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

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 *Studies in Econometrics, Time series and Multivariate Analysis, Academic Press*(Edited by S. Karlln, T. Amemiya and L. A. Goodman), 255-278. New York.

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, Z. and Lu, C. (1997). *Limit Theory on Mixing Dependent Random Variables.* Kluwer Academic Publishers, New York.

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.

Staudenmayer, J. and Buonaccorsi, J. (2005). Measurement error in linear autoregressive models. *Journal of the American Statistical Association* **100**, 841–852.

Stock, J. H. and Watson, M.(2008). Forecasting in dynamic factor models subject to structural instability. In *The Methodology and Practice of Econometrics, A Festschrift in Honour of Professor David F. Hendry* (Edited by J. Castle and N. Shephard), 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. (2019). Identifying cointegration by eigenanalysis. *Journal of the American Statistical Association* **114**, 916–927.

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

E-mail: yundong.tu@gsm.pku.edu.cn

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

E-mail: q.yao@lse.ac.uk

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

E-mail: rmzhang@zju.edu.cn

(Received August 2016; accepted September 2018)