Clifford Lam, Qiwei Yao and Neil Bathia

Estimation of latent factors for high-dimensional time series

Article (Accepted version) (Refereed)

Estimation of Latent Factors for High-Dimensional Time Series

By Clifford Lam, Qiwei Yao and Neil Bathia

Department of Statistics, London School of Economics and Political Science
Houghton Street, London WC2A 2AE, U.K.

c.lam2@lse.ac.uk q.yao@lse.ac.uk n.bathia@lse.ac.uk

Summary

This paper deals with the dimension reduction of high-dimensional time series based on common factors. In particular we allow the dimension of time series p to be as large as, or even larger than, the sample size n. The estimation of the factor loading matrix and the factor process itself is carried out via an eigenanalysis of a $p \times p$ non-negative definite matrix. We show that when all the factors are strong in the sense that the norm of each column in the factor loading matrix is of the order $p^{1/2}$, the estimator of the factor loading matrix is weakly consistent in L_2-norm with the convergence rate independent of p. This result exhibits clearly that the 'curse' is canceled out by the 'blessing' of dimensionality. We also establish the asymptotic properties of the estimation when factors are not strong. The proposed method together with their asymptotic properties are further illustrated in a simulation study. An application to an implied volatility data set, together with a trading strategy derived from the fitted factor model, is also reported.

Short Title: Estimation of Large Latent Time Series Factors.

Some key words: Convergence in L_2-norm; Curse and blessing of dimensionality; Dimension reduction; Eigenanalysis; Factor model.

1 Introduction

In the modern information age, analysis of large data sets is an integral part of both scientific research and practical problem-solving. In particular, high-dimensional time series analysis is commonplace in many fields including, among others, finance, economics, environmental and medical studies. For example, understanding the dynamics of the returns of large number of assets is the key to asset pricing, portfolio allocation, and risk management. Panel time series are frequently encountered in studying economic and business phenomena. Environmental time series are often of a high dimension because of the large number of indices monitored across many different locations. However the standard multiple time series models such as vector AR or vector ARMA are not practically viable when the dimension of time series p is high, as the number of parameters involved is in the order of p^2. Furthermore, one may face a serious model-identification problem in a vector ARMA model. In fact the vector ARMA model has hardly been used in practice without further regularization in its matrix coefficients. Therefore dimension-reduction is an important step in order to achieve an efficient and effective analysis of high-dimensional time series data. In relation to the dimension-reduction for independent observations, the added challenge here is to retain the dynamical structure of time series.

Modeling by common factors is one of the most frequently used methods to achieve dimension-reduction in analyzing multiple time series. Early attempts in this direction include Anderson (1963), Priestley et al. (1974), Brillinger (1981) and Peña and Box (1987). To deal with the situations when the number of time series p is as large as, or even larger than, the length of the time series n, more recent efforts focus on the inference when p goes to infinity together with n. See, e.g. Chamberlain and Rothschild (1983), Chamberlain (1983), Bai (2003) Forni et al. (2000, 2004, 2005). Furthermore, in analyzing economic and financial phenomena, most econometric factor models seek to identify the common factors such that each of them affects the dynamics of *most* the original p time series. These common factors are separated from the so-called idiosyncratic noise components; each idiosyncratic noise component may at most affect the dynamics of *a few* original time series. Note that an idiosyncratic noise series is not necessarily white noise. The rigorous definition of the common factors and the idiosyncratic noise can only be established asymptotically when the number of time series p goes to infinity; see Chamberlain and Rothschild (1983) and Chamberlain (1983). Hence those econometric factor models are only asymptotically identifiable when $p \to \infty$. See also Forni et al.

(2000).

We adopt a different and more statistical approach in this paper from a dimension-reduction point of view. Our model is similar to those in Peña and Box (1987), Bai and Ng (2002), Peña and Poncela (2006), and Pan and Yao (2008), and we consider the inference when p is as large as, or even larger than, n. Different from the aforementioned econometric factor models, we decompose the p-dimensional time series into two parts: the dynamic part driven by low-dimensional factors and the static part which is a vector white noise. Furthermore, we allow the future factors to depend on past (white) noise. Such a conceptually simple decomposition is convenient for both model identification and statistical inference. In fact, the model is identifiable for any finite p. Furthermore the estimation for the factor loading matrix and the factor process itself is equivalent to an eigenanalysis of a $p \times p$ non-negative definite matrix. Therefore it is applicable when p is in the order of a few thousands. Our approach is rooted in the same idea on which the methods of Peña and Poncela (2006) and Pan and Yao (2008) were based. However, our method is radically different and is substantially simpler. For example, Peña and Poncela (2006) requires the computation of the inverse of the sample covariance matrix for the data, which is computationally costly when p is large, and is invalid when $p > n$. (See also Peña and Box (1987).) Moreover, in contrast to performing eigenanalysis for one autocovariance matrix each time, our method only requires to perform one single eigenanalysis on a matrix function of several autocovariance matrices, and it augments the information on the dynamics along different lags. The method of Pan and Yao (2008) involves solving several nonlinear optimization problems, which is designed to handle non-stationary factors and is only feasible for moderately large p. Our approach identifies factors based on the autocorrelation structure of the data, which, we argue, is more relevant than the least squares approach advocated by Bai and Ng (2002) and Bai (2003) in the context of identifying time series factors.

The major theoretical contribution of this paper is to reveal an interesting and somehow intriguing feature in factor modeling: the estimator for the factor loading matrix of the original p-dimensional time series converges at a rate independent of p, provided that all the factors are strong in the sense that the norm of each column in the factor loading matrix is of order $p^{1/2}$. Our simulation indicates that the estimation errors are indeed independent of p. This result exhibits clearly that the 'curse' is canceled out by the 'blessing' in dimensionality. In the presence of weak factors, the convergence rate of

the estimated factor loading matrix depends on p. In spite of this, we have shown that the optimal convergence rate is obtained under some additional conditions on the white noise, which include Gaussian white noise as a special case.

Although we focus on stationary processes only in this paper, our approach is still relevant for the nonstationary processes for which a generalized autocovariance matrix is well-defined; see remark 1(v) in section 3.

The rest of the paper is organized as follows. The model, its presentational issues and the estimation method are presented in section 2. Section 3 introduces the asymptotic properties of the proposed estimation method. Our simulation results are presented in section 4. A detailed analysis of a set of implied volatility data is reported in section 5. All technical proofs are relegated to section 6.

2 Models and estimation methodology

2.1 Factor models

Let $\mathbf{y}_1, \cdots, \mathbf{y}_n$ be n $p \times 1$ successive observations from a vector time series process. The factor model assumes

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_t, \tag{2.1}$$

where $\{\mathbf{x}_t\}$ is a $r \times 1$ unobserved factor time series which is assumed to be strictly stationary with finite first two moments, $\mathbf{A}$ is a $p \times r$ unknown constant factor loading matrix, $r(\leq p)$ is the number of factors, and $\{\boldsymbol{\epsilon}_t\}$ is a white noise with mean $\mathbf{0}$ and covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$.

We introduce some notation first. For $k \geq 0$, let $\boldsymbol{\Sigma}_{\mathbf{x}}(k) = \mathrm{Cov}(\mathbf{x}_{t+k}, \mathbf{x}_t)$, $\boldsymbol{\Sigma}_{\mathbf{x},\boldsymbol{\epsilon}}(k) = \mathrm{Cov}(\mathbf{x}_{t+k},\ \boldsymbol{\epsilon}_t)$, and

$$\widetilde{\boldsymbol{\Sigma}}_{\mathbf{x}}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(\mathbf{x}_{t+k} - \bar{\mathbf{x}})(\mathbf{x}_t - \bar{\mathbf{x}})^T, \quad \widetilde{\boldsymbol{\Sigma}}_{\mathbf{x},\boldsymbol{\epsilon}}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(\mathbf{x}_{t+k} - \bar{\mathbf{x}})(\boldsymbol{\epsilon}_t - \bar{\boldsymbol{\epsilon}})^T,$$

where $\bar{\mathbf{x}} = n^{-1}\sum_{t=1}^{n}\mathbf{x}_t$, $\bar{\boldsymbol{\epsilon}} = n^{-1}\sum_{t=1}^{n}\boldsymbol{\epsilon}_t$. The autocovariance matrices $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}(k)$, $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon},\mathbf{x}}(k)$, and their sample versions are defined in a similar manner. Some assumptions on model (2.1) are now in order.

(A) No linear combination of the components of $\mathbf{x}_t$ is white noise.

(B) For $k = 0, 1, \cdots, k_0$, where $k_0 \geq 1$ is a small positive integer, $\boldsymbol{\Sigma}_{\mathbf{x}}(k)$ is full-ranked.

(C) For $k \geq 0$, the cross autocovariance matrix $\boldsymbol{\Sigma}_{\mathbf{x},\boldsymbol{\epsilon}}(k)$ and the covariance matrix $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ have elements of order $O(1)$.

(D) $\mathrm{Cov}(\boldsymbol{\epsilon}_t, \mathbf{x}_s) = \mathbf{0}$ for all $s \leq t$.

(E) $\mathbf{y}_t$ is strictly stationary and ψ-mixing with the mixing coefficients $\psi(\cdot)$ satisfying the condition that $\sum_{t\geq 1} t\psi(t)^{1/2} < \infty$. Furthermore $E\{\|\mathbf{y}_t\|^4\} < \infty$.

Assumption (A) is natural, as all the white noise linear combinations of $\mathbf{x}_t$ should be absorbed into $\boldsymbol{\epsilon}_t$. It implies that there exists at least one $k \geq 1$ for which $\boldsymbol{\Sigma}_{\mathbf{x}}(k)$ is full ranked. Assumption (B) strengthens this statement for all $1 \leq k \leq k_0$, which entails that the non-negative definite matrix $\mathbf{L}$, defined in (2.4) below, has only r positive eigenvalues. Assumption (C) is also a natural condition which ensures each element in $\boldsymbol{\Sigma}_{\mathbf{x},\boldsymbol{\epsilon}}(k)$ and $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ behaves normally when p increases. Assumption (D) relaxes independence assumption between $\{\mathbf{x}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$, which is present in most factor model literature. It allows future factors to be correlated with past white noise. Finally, assumption (E) is not the weakest possible. The ψ-mixing condition may be replaced by the α-mixing condition at the expenses of more lengthy technical argument.

In this paper, we always assume that the number of factors r is known and fixed. It is reasonable to assume r fixed while $p \to \infty$, as model (2.1) is practically useful only when $r \ll p$. There is a large body of literature on the determination of r. See, for example, Bai and Ng (2002, 2007), Hallin and Liška (2007), Pan and Yao (2008), Bathia et al. (2010) and Lam and Yao (2010). We use the information criterion proposed by Bai and Ng (2002) to determine r in our numerical examples in section 4.

2.2 Identifiability and factor strength

Model (2.1) is unchanged if we replace the pair $(\mathbf{A}, \mathbf{x}_t)$ on the RHS by $(\mathbf{AH}, \mathbf{H}^{-1}\mathbf{x}_t)$ for any invertible $\mathbf{H}$. However the linear space spanned by the columns of $\mathbf{A}$, denoted by $\mathcal{M}(\mathbf{A})$ and called the factor loading space, is uniquely defined by (2.1). Note $\mathcal{M}(\mathbf{A}) = \mathcal{M}(\mathbf{AH})$ for any invertible $\mathbf{H}$. Once such an $\mathbf{A}$ is specified, the factor process $\mathbf{x}_t$ is uniquely defined accordingly. We see the lack of uniqueness of $\mathbf{A}$ as an advantage, as we may choose a particular $\mathbf{A}$ which facilitates our estimation in a simple and convenient manner. Before we specify explicitly such an $\mathbf{A}$ in section 2.3 below, we introduce an index δ to measure the strength of the factors. We always use the notation $a \asymp b$ to denote $a = O_P(b)$ and $b = O_P(a)$.

(F) $\mathbf{A} = (\mathbf{a}_1 \cdots \mathbf{a}_r)$ such that $\|\mathbf{a}_i\|^2 \asymp p^{1-\delta}$, $i = 1, \cdots, r$, $0 \le \delta \le 1$.

(G) For each $i = 1, \cdots, r$ and δ given in (F), $\min_{\theta_j, j \ne i} \|\mathbf{a}_i - \sum_{j \ne i} \theta_j \mathbf{a}_j\|^2 \asymp p^{1-\delta}$.

When $\delta = 0$ in assumption (F), the corresponding factors are called strong factors since it includes the case where each element of $\mathbf{a}_i$ is $O(1)$, implying that the factors are shared (strongly) by the majority of the p time series. When $\delta > 0$, the factors are called weak factors. In fact the smaller the δ is, the stronger the factors are. This definition is different from Chudik et al. (2009) which defined the strength of factors by the finiteness of the mean absolute values of the component of $\mathbf{a}_i$. One advantage of using index δ is to link the convergence rates of the estimated factors explicitly to the strength of factors. In fact the convergence is slower in the presence of weak factors. Assumptions (F) and (G) together ensure that all r factors in the model are of the equal strength δ.

To facilitate our estimation, we use the QR decomposition $\mathbf{A} = \mathbf{QR}$ to normalize the factor loading matrix, so that (2.1) becomes

$$\mathbf{y}_t = \mathbf{QRx}_t + \boldsymbol{\epsilon}_t = \mathbf{Qf}_t + \boldsymbol{\epsilon}_t, \tag{2.2}$$

where $\mathbf{f}_t = \mathbf{Rx}_t$, and $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}_r$. Note that the pair $(\mathbf{Q}, \mathbf{f}_t)$ in the above model can be replaced by $(\mathbf{QU}, \mathbf{U}^T\mathbf{f}_t)$ for any $r \times r$ orthogonal matrix $\mathbf{U}$. In the following section we will specify explicitly such a $\mathbf{Q}$ to be used in our estimation.

2.3 Estimation

For $k \ge 1$, model (2.2) implies that

$$\boldsymbol{\Sigma}_{\mathbf{y}}(k) = \mathrm{Cov}(\mathbf{y}_{t+k}, \mathbf{y}_t) = \mathbf{Q}\boldsymbol{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \mathbf{Q}\boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k), \tag{2.3}$$

where $\boldsymbol{\Sigma}_{\mathbf{f}}(k) = \mathrm{Cov}(\mathbf{f}_{t+k}, \mathbf{f}_t)$ and $\boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k) = \mathrm{Cov}(\mathbf{f}_{t+k}, \boldsymbol{\epsilon}_t)$. For $k_0 \ge 1$ given in condition (B), define

$$\begin{aligned} \mathbf{L} &= \sum_{k=1}^{k_0} \boldsymbol{\Sigma}_{\mathbf{y}}(k)\boldsymbol{\Sigma}_{\mathbf{y}}(k)^T \\ &= \mathbf{Q}\Big(\sum_{k=1}^{k_0} \{\boldsymbol{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\}\{\boldsymbol{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\}^T \Big)\mathbf{Q}^T. \end{aligned} \tag{2.4}$$

Obviously $\mathbf{L}$ is a $p \times p$ non-negative definite matrix. Now we are ready to specify the factor loading matrix $\mathbf{Q}$ to be used in our estimation. Apply the spectral decomposition

to the positive-definite matrix sandwiched by $\mathbf{Q}$ and $\mathbf{Q}^T$ on the RHS of (2.4), i.e.

$$\sum_{k=1}^{k_0}\{\mathbf{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\}\{\mathbf{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\}^T = \mathbf{U}\mathbf{D}\mathbf{U}^T,$$

where $\mathbf{U}$ is an $r\times r$ orthogonal matrix, and $\mathbf{D}$ is a diagonal matrix with the elements on the main diagonal in descending order. This leads to $\mathbf{L} = \mathbf{Q}\mathbf{U}\mathbf{D}\mathbf{U}^T\mathbf{Q}^T$. As $\mathbf{U}^T\mathbf{Q}^T\mathbf{Q}\mathbf{U} = \mathbf{I}_r$, the columns of $\mathbf{Q}\mathbf{U}$ are the eigenvectors of $\mathbf{L}$ corresponding to its r non-zero eigenvalues. We take $\mathbf{Q}\mathbf{U}$ as the $\mathbf{Q}$ to be used in our inference, i.e.

> *the columns of the factor loading matrix* $\mathbf{Q}$ *are the* r *orthonormal eigenvectors of the matrix* $\mathbf{L}$ *corresponding to its* r *non-zero eigenvalues, and the columns are arranged such that the corresponding eigenvalues are in the descending order.*

A natural estimator for the $\mathbf{Q}$ specified above is defined as $\widehat{\mathbf{Q}} = (\widehat{\mathbf{q}}_1, \cdots, \widehat{\mathbf{q}}_r)$, where $\widehat{\mathbf{q}}_i$ is the eigenvector of $\widetilde{\mathbf{L}}$ corresponding to its i-th largest eigenvalue, $\widehat{\mathbf{q}}_1, \cdots, \widehat{\mathbf{q}}_r$ are orthonormal, and

$$\widetilde{\mathbf{L}} = \sum_{k=1}^{k_0}\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k)^T, \quad \widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) = \frac{1}{n-k}\sum_{t=1}^{n-k}(\mathbf{y}_{t+k} - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^T, \tag{2.5}$$

where $\bar{\mathbf{y}} = n^{-1}\sum_{t=1}^{n}\mathbf{y}_t$.

Consequently, we estimate the factors and the residuals respectively by

$$\widehat{\mathbf{f}}_t = \widehat{\mathbf{Q}}^T\mathbf{y}_t, \qquad \mathbf{e}_t = \mathbf{y}_t - \widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t = (\mathbf{I}_p - \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T)\mathbf{y}_t. \tag{2.6}$$

3 Asymptotic theory

In this section we present the rates of convergence for the estimators $\widehat{\mathbf{Q}}$ for model (2.2), and also for the estimated factor $\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t$. It goes without saying explicitly that we may replace some $\widehat{\mathbf{q}}_j$ by $-\widehat{\mathbf{q}}_j$ in order to match the direction of $\mathbf{q}_j$. Denote by $\|M\|$ the spectral norm of a matrix M (i.e. the positive square root of the maximum eigenvalue of MM^T), and denote by $\|M\|_{\min}$ the positive square root of the minimum eigenvalue of MM^T or M^TM, whichever is a smaller matrix. For model (2.2), define

$$\kappa_{\min} = \min_{1\le k\le k_0}\|\mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\|_{\min}, \quad \kappa_{\max} = \max_{1\le k\le k_0}\|\mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\|.$$

Both $\kappa_{\max}$ and $\kappa_{\min}$ may be viewed as the measures of the strength of the cross-correlation between the factor process and the white noise.

Theorem 1 *Let assumptions (A) - (G) hold, and the r positive eigenvalues of matrix* $\mathbf{L}$*, defined in (2.4), be distinct. Then,*

(i) $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(p^{\delta} n^{-1/2})$ *provided* $\kappa_{\max} = o(p^{1-\delta})$ *and* $p^{\delta} n^{-1/2} = o(1)$*, and*

(ii) $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(\kappa_{\min}^{-2} \kappa_{\max} \cdot p n^{-1/2})$ *provided* $p^{1-\delta} = o(\kappa_{\min})$ *and* $\kappa_{\min}^{-2} \kappa_{\max} \cdot p n^{-1/2} = o(1)$.

Remark 1. (i) When all the factors are strong (i.e. $\delta = 0$), Theorem 1(i) reduces to $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(n^{-1/2})$ provided $\kappa_{\max}/p \to 0$. The standard root-n rate might look too good to be true, as the dimension p goes to infinity together with the sample size n. But this is the case when 'blessing of dimensionality' is at its clearest. Note that the strong factors pool together the information from most, if not all, of the original p component series. When p increases, the curse of dimensionality is offset by the increase of the information from more component series. The condition $\kappa_{\max}/p \to 0$ is very mild. It implies that the linear dependence between the factors and the white noise is not too strong.

(ii) When $\delta > 0$, Theorem 1(i) shows that the stronger the factors are, the faster the convergence rate is. The condition $\kappa_{\max} = o(p^{1-\delta})$ ensures that the matrix $\mathbf{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)$, in (2.4), is dominated by the first term.

(iii) Theorem 1(ii) represents the cases that there are strong cross-correlations between the factors and the white noise, as $\kappa_{\min}/p^{1-\delta} \to \infty$. However this does not necessarily imply a slow convergence rate in estimating $\mathbf{Q}$. For instance, when $\kappa_{\max} \asymp p^{1-\delta/2} \asymp \kappa_{\min}$ (see Lemma 1 in section 6 below), $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(p^{\delta/2} n^{-1/2})$. This convergence rate is even faster than the rate $p^{\delta} n^{-1/2}$. This is not surprising, as we assume that r is known and we estimate $\mathbf{Q}$ by extracting the information on the autocorrelation of the data, including the cross-autocorrelation between $\{\mathbf{f}_t\}$ and $\{\boldsymbol{\epsilon}_t\}$. See the definition of $\mathbf{L}$ in (2.4). However, this may create difficulties for estimating r; see the relevant asymptotic results in Lam and Yao (2010).

(iv) The assumption that all the non-zero eigenvalues of $\mathbf{L}$ are different is not essential, and is merely introduced to simplify the presentation in the sense that Theorem 1 now can deal with the convergence of the estimator for $\mathbf{Q}$ directly. Otherwise a discrepancy measure for two linear spaces has to be introduced in order to make statements on the convergence rate of the estimator for the factor loading space $\mathcal{M}(\mathbf{A})$; see Pan and Yao (2008).

(v) Theorem 1 can be extended to the cases when the factor $\mathbf{x}_t$ in model (2.1) is non-stationary, provided that a generalized sample (auto)covariance matrix

$$n^{-\alpha}\sum_{t=1}^{n-k}(\mathbf{x}_{t+k}-\bar{\mathbf{x}})(\mathbf{x}_t-\bar{\mathbf{x}})^T$$

converges weakly, where $\alpha > 1$ is a constant. This weak convergence has been established when, for example, $\{\mathbf{x}_t\}$ is a two times integrated process (i.e. $\{\mathbf{x}_t\}$ is $I(2)$) by Peña and Poncela (2006). It can also be proved for other processes with linear trends, random walk or long memories. In this paper we do not pursue further in this direction.

Some conditions in Theorem 1 may be too restrictive. For instance when $p \asymp n$, Theorem 1(i) requires $\delta < 1/2$. This rules out the cases in the presence of weaker factors with $\delta \geq 1/2$. The convergence rates in Theorem 1 are also not optimal. They can be further improved under additional assumptions on $\boldsymbol{\epsilon}_t$ as follows. Note that in particular, both assumptions (H) and (I) are fulfilled when $\boldsymbol{\epsilon}_t$ are independent and $N(0, \sigma^2\mathbf{I}_p)$. See also Péché (2009).

(H) Let ϵ_{jt} denote the j-th component of $\boldsymbol{\epsilon}_t$. Then ϵ_{jt} are independent for different t and j, and have mean 0 and common variance $\sigma^2 < \infty$.

(I) The distribution of each ϵ_{jt} is symmetric. Furthermore $E(\epsilon_{jt}^{2k+1}) = 0$, and $E(\epsilon_{jt}^{2k}) \leq (\tau k)^k$ for all $1 \leq j \leq p$ and $t, k \geq 1$, where $\tau > 0$ is a constant independent of j, t, k.

Theorem 2 *In addition to the assumptions of Theorem 1, we assume (H) and (I). If* $n = O(p)$, *then*

(i) $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(p^{\delta/2}n^{-1/2})$ *provided* $\kappa_{\max} = o(p^{1-\delta})$ *and* $p^{\delta/2}n^{-1/2} = o(1)$, *and*

(ii) $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(\kappa_{\min}^{-2}\kappa_{\max} \cdot p^{1-\delta/2}n^{-1/2})$ *provided* $p^{1-\delta} = o(\kappa_{\min})$ *and* $\kappa_{\min}^{-2}\kappa_{\max} \cdot p^{1-\delta/2}n^{-1/2} = o(1)$.

By comparing with Theorem 1, the rates provided in Theorem 2 are improved by a factor $p^{-\delta/2}$. This also relaxes the condition on the strength of the factors. For instance, when $p \asymp n$, Theorem 2(i) only requires $\delta < 1$ while Theorem 2(i) requires $\delta < 1/2$.

Theorem 3 *If all the eigenvalues of* $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ *are uniformly bounded from infinity (as* $p \to \infty$*), it holds that*

$$p^{-1/2}\|\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t - \mathbf{A}\mathbf{x}_t\| = p^{-1/2}\|\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t - \mathbf{Q}\mathbf{f}_t\| = O_P(p^{-\delta/2}\|\widehat{\mathbf{Q}} - \mathbf{Q}\| + p^{-1/2}). \tag{3.7}$$

Theorem 3 specifies the convergence rate for the estimated factors. When all factors are strong (i.e. $\delta = 0$), both Theorems 1 and 2 imply $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(n^{-1/2})$. Now it follows Theorem 3 that

$$p^{-1/2}\|\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t - \mathbf{A}\mathbf{x}_t\| = O_P(n^{-1/2} + p^{-1/2}). \tag{3.8}$$

This is the optimal convergence rate specified in Theorem 3 of Bai (2003). This optimal rate is still attained when the factors are weaker (i.e. $\delta > 0$) but the white noise fulfils assumptions (H) and (I), as then Theorem 2(i) implies $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = O_P(p^{\delta/2}n^{-1/2})$. Plugging this into the RHS of (3.7), we obtain (3.8).

4 Simulation

In this section, we illustrate our estimation method and their properties via two simulated examples.

Example 1. We start with a simple one factor model

$$\mathbf{y}_t = \mathbf{A}x_t + \boldsymbol{\epsilon}_t, \;\; \epsilon_{tj} \sim \text{ i.i.d. } N(0, 2^2),$$

where the factor loading matrix $\mathbf{A}$ is a $p \times 1$ vector with $2\cos(2\pi i/p)$ as its i-th element, and the factor time series is defined as $x_t = 0.9x_{t-1} + \eta_t$, where η_t are independent $N(0, 2^2)$ random variables. Hence we have a strong factor for this model with $\delta = 0$. We set $n = 200, 500$ and $p = 20, 180, 400, 1000$. For each (n, p) combination, we generate from the model 50 samples and calculate the estimation errors. The results are listed in Table 1 below. Table 1 indicates clearly that the estimation error in L_2 norm for $\widehat{\mathbf{Q}}$ is independent of p, as shown in Theorem 1(i) with $\delta = 0$.

$\|\widehat{\mathbf{Q}} - \mathbf{Q}\|$	$n = 200$	$n = 500$
$p = 20$	.022(.005)	.014(.003)
$p = 180$	.023(.004)	.014(.002)
$p = 400$	.022(.004)	.014(.002)
$p = 1000$	.023(.004)	.014(.002)

Table 1: *Means and standard errors (in brackets) of* $\|\widehat{\mathbf{Q}} - \mathbf{Q}\|$ *for Example 1.*

Example 2. We consider model (2.1) with three factors now ($r = 3$). We compare the performance of our estimators with the principle component (PC) method of Bai and Ng

(2002) under two different scenarios: (I) $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{I}_p)$, and (II) $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$, where $\boldsymbol{\Sigma}_{\boldsymbol{\epsilon}}$ has elements σ_{ij} following the fractional Gaussian noise, defined by

$$\sigma_{ij} = \frac{1}{2}((|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}),$$

and $H \in [0.5, 1]$ is the Hurst parameter. We set $H = 0.9$ in (II) to simulate strong cross-sectional dependence for the elements of $\boldsymbol{\epsilon}_t$, which violates the weak cross-sectional dependence assumption in Bai and Ng (2002), but is allowed in our setting.

The factors are defined by

$$\begin{aligned} x_{1,t} &= -0.8x_{1,t-1} + 0.9e_{1,t-1} + e_{1,t}, \\ x_{2,t} &= -0.7x_{2,t-1} + 0.85e_{2,t-1} + e_{2,t}, \\ x_{3,t} &= \quad 0.8x_{2,t} - 0.5x_{3,t-1} + e_{3,t}, \end{aligned}$$

where $e_{i,t}$ are independent $N(0,1)$ random variables. For each column of $\mathbf{A}$, we generate the first $p/2$ elements randomly from the $U(-2,2)$ distribution; the rest are set to zero. This increases the difficulty in detecting the signals from the factors. We then adjust the strength of the factors by normalizing the columns, setting $\mathbf{a}_i/p^{\delta_i/2}$ as the i-th column of $\mathbf{A}$ with $\delta_2 = \delta_3$.

We estimate $\widehat{\mathbf{Q}}$ either using the true number of factors $r = 3$, or $\widehat{r}$ obtained by minimizing the BIC type of information criterion proposed in Bai and Ng (2002):

$$\widehat{r} = \arg\min_k \mathrm{IC}(k) = \arg\min_k \log\left(p^{-1}n^{-1}\sum_{j=1}^{p}\|\widehat{\boldsymbol{\epsilon}}_j\|^2\right) + k\left(\frac{p+n}{pn}\right)\log\left(\frac{pn}{p+n}\right).$$

We set $n = 100, 200, 400$ and $p = 200, 400, 800$. We use $k_0 = 4$ in the definition of $\widetilde{\mathbf{L}}$ in (2.5). The first factor has strength index δ_1 and the last two factors have strength index δ_2. For each combination of $(n, p, \delta_1, \delta_2)$, we replicate the simulation 100 times, and calculate the mean and the standard deviation of the root-mean-square error (RMSE):

$$\mathrm{RMSE} = \left(\frac{\sum_{t=1}^{n}\|\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t - \mathbf{Q}\mathbf{f}_t\|^2}{pn}\right)^{1/2}.$$

We also use $\widehat{\mathbf{y}}_n^{(1)} = \widehat{\mathbf{Q}}\widehat{\mathbf{f}}_n^{(1)}$ to forecast the factor $\mathbf{Q}\mathbf{f}_t$, where $\widehat{\mathbf{f}}_n^{(1)}$ is the one-step predictor for $\mathbf{f}_n$ derived from a fitted AR(4) model based on $\widehat{\mathbf{f}}_1, \cdots, \widehat{\mathbf{f}}_{n-1}$. We then calculate the mean and standard deviation of the factor forecast error (FFE) and the forecast error (FE):

$$\mathrm{FFE} = p^{-1/2}\|\widehat{\mathbf{y}}_n^{(1)} - \mathbf{Q}\mathbf{f}_n\|, \quad \mathrm{FE} = p^{-1/2}\|\widehat{\mathbf{y}}_n^{(1)} - \mathbf{y}_n\|.$$

(I): $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{I}_p)$		PC method				Our method			
$\delta_1 = \delta_2 = 0$		$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
n=100	p=200	$3_{(0)}$	$.21_{(.005)}$	$1.55_{(.76)}$	$1.87_{(.62)}$	$3.0_{(.2)}$	$.28_{(.02)}$	$1.54_{(.75)}$	$1.87_{(.62)}$
	p=400	$3_{(0)}$	$.19_{(.003)}$	$1.61_{(.77)}$	$1.93_{(.65)}$	$3.1_{(.3)}$	$.27_{(.02)}$	$1.61_{(.77)}$	$1.94_{(.66)}$
	p=800	$3_{(0)}$	$.18_{(.003)}$	$1.61_{(.82)}$	$1.95_{(.71)}$	$3.1_{(.3)}$	$.26_{(.02)}$	$1.64_{(.87)}$	$1.97_{(.76)}$
n=200	p=200	$3_{(0)}$	$.17_{(.004)}$	$1.58_{(.74)}$	$1.90_{(.61)}$	$3_{(0)}$	$.21_{(.008)}$	$1.58_{(.74)}$	$1.90_{(.61)}$
	p=400	$3_{(0)}$	$.15_{(.003)}$	$1.44_{(.71)}$	$1.80_{(.59)}$	$3_{(0)}$	$.19_{(.01)}$	$1.44_{(.70)}$	$1.80_{(.59)}$
	p=800	$3_{(0)}$	$.14_{(.001)}$	$1.28_{(.64)}$	$1.67_{(.51)}$	$3_{(0)}$	$.18_{(.01)}$	$1.28_{(.64)}$	$1.67_{(.51)}$
n=400	p=200	$3_{(0)}$	$.15_{(.003)}$	$1.47_{(.74)}$	$1.82_{(.62)}$	$3_{(0)}$	$.17_{(.004)}$	$1.47_{(.74)}$	$1.82_{(.62)}$
	p=400	$3_{(0)}$	$.12_{(.002)}$	$1.59_{(.73)}$	$1.92_{(.62)}$	$3_{(0)}$	$.15_{(.004)}$	$1.59_{(.73)}$	$1.92_{(.62)}$
	p=800	$3_{(0)}$	$.11_{(.001)}$	$1.37_{(.61)}$	$1.73_{(.50)}$	$3_{(0)}$	$.13_{(.004)}$	$1.37_{(.61)}$	$1.73_{(.50)}$

Table 2: *Means and standard deviations (in brackets) of estimation errors and forecast errors for Example 2:* $\boldsymbol{\epsilon}_t \sim N(0, \mathbf{I}_p)$*, and all three factors are strong* ($\delta_1 = \delta_2 = 0$).

It is clear from Table 2 that the information criterion for estimating r performed very well on both methods under scenario (I). The PC method performs better in estimating the factors, reflected by the smaller RMSE in most cases. As n increases the RMSE for the two methods become closer. Moreover, the two methods perform equally well in terms of the forecast errors. Table 3 shows the results under the same scenario when the factors have different strength indices δ_1 and δ_2, and $p = 2n$. It is clear that r is estimated very well even in the presence of weak factors, and the relative performance of the two methods is about the same as in Table 2.

Table 4 shows the results under scenario (II). The information criterion leads to over-estimation of r for both methods, with a more adverse effect for the PC. Our method outperforms the PC method under this scenario in general for all the measures RMSE, FFE and FE. This is well-expected since $\{\boldsymbol{\epsilon}_t\}$ exhibits strong cross-sectional dependence, which violates the condition imposed in Bai and Ng (2002).

Since r is overestimated in all cases in Table 4, we repeat the simulation with the number of factors set at the true value $r = 3$. The results are reported in Table 5. Our method outperforms the PC method in all cases except when $\delta_1 = \delta_2 = 0$. When all factors are strong, the PC method can pick up the signals from all the three factors and still gives better performance. However, in the presence of weaker factors coupled with strong cross-sectional dependence of $\{\boldsymbol{\epsilon}_t\}$, the PC method cannot identify correctly the signals from all the factors, evidenced by the sharp increase in RMSE and their large standard deviations. On the other hand, our method can detect the presence of weaker factors even under strong cross-sectional dependence.

We repeat the above experiments with all the components of $\boldsymbol{\epsilon}_t$ being i.i.d and each follows an AR(1) with parameter $\phi = 0.2$. This creates weak serial correlation in $\{\boldsymbol{\epsilon}_t\}$. The results are very similar to that in Table 2, and are therefore omitted. This demonstrates that weak serial correlation in $\{\boldsymbol{\epsilon}_t\}$ does not affect the performance of both methods.

(I): $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \mathbf{I}_p)$	PC method				Our method			
$\delta_1 = 0, \delta_2 = 1/4$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$3_{(0)}$	$.21_{(.006)}$	$1.20_{(.55)}$	$1.61_{(.43)}$	$3.0_{(.1)}$	$.28_{(.02)}$	$1.22_{(.54)}$	$1.62_{(.43)}$
$(n,p) = (200, 400)$	$3_{(0)}$	$.15_{(.003)}$	$1.01_{(.44)}$	$1.45_{(.32)}$	$3_{(0)}$	$.19_{(.01)}$	$1.01_{(.44)}$	$1.46_{(.32)}$
$(n,p) = (400, 800)$	$3_{(0)}$	$.11_{(.002)}$	$.89_{(.42)}$	$1.37_{(.29)}$	$3_{(0)}$	$.13_{(.005)}$	$.89_{(.42)}$	$1.38_{(.29)}$
$\delta_1 = 1/4, \delta_2 = 0$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$3_{(0)}$	$.21_{(.005)}$	$1.31_{(.62)}$	$1.68_{(.49)}$	$3.0_{(.1)}$	$.28_{(.02)}$	$1.31_{(.61)}$	$1.68_{(.49)}$
$(n,p) = (200, 400)$	$3_{(0)}$	$.15_{(.003)}$	$1.38_{(.80)}$	$1.76_{(.64)}$	$3_{(0)}$	$.19_{(.01)}$	$1.38_{(.80)}$	$1.77_{(.65)}$
$(n,p) = (400, 800)$	$3_{(0)}$	$.11_{(.001)}$	$1.30_{(.72)}$	$1.69_{(.58)}$	$3_{(0)}$	$.13_{(.004)}$	$1.30_{(.72)}$	$1.69_{(.58)}$
$\delta_1 = 0, \delta_2 = 1/2$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$3.0_{(.2)}$	$.22_{(.01)}$	$.79_{(.42)}$	$1.30_{(.30)}$	$3.0_{(.2)}$	$.27_{(.02)}$	$.80_{(.41)}$	$1.30_{(.29)}$
$(n,p) = (200, 400)$	$3_{(0)}$	$.15_{(.003)}$	$.77_{(.49)}$	$1.32_{(.34)}$	$3_{(0)}$	$.19_{(.008)}$	$.78_{(.49)}$	$1.32_{(.33)}$
$(n,p) = (400, 800)$	$3_{(0)}$	$.11_{(.002)}$	$.64_{(.41)}$	$1.23_{(.27)}$	$3_{(0)}$	$.13_{(.004)}$	$.64_{(.41)}$	$1.23_{(.27)}$

Table 3: *Means and standard deviations (in brackets) of estimation errors and forecast errors for Example 2:* $\boldsymbol{\epsilon}_t \sim N(0, \mathbf{I}_p)$.

(II): $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$	PC method				Our method			
$\delta_1 = 0, \delta_2 = 0$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$6.7_{(.6)}$	$.71_{(.04)}$	$1.84_{(.79)}$	$2.14_{(.69)}$	$4.8_{(1.0)}$	$.63_{(.05)}$	$1.73_{(.80)}$	$2.03_{(.71)}$
$(n,p) = (200, 400)$	$8.4_{(.6)}$	$.68_{(.02)}$	$1.55_{(.69)}$	$1.88_{(.61)}$	$5.2_{(1.3)}$	$.57_{(.04)}$	$1.48_{(.63)}$	$1.82_{(.55)}$
$(n,p) = (400, 800)$	$11.2_{(.7)}$	$.66_{(.02)}$	$1.42_{(.76)}$	$1.82_{(.63)}$	$7.0_{(1.8)}$	$.54_{(.03)}$	$1.35_{(.74)}$	$1.77_{(.59)}$
$\delta_1 = 0, \delta_2 = 1/4$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$6.8_{(.7)}$	$.72_{(.04)}$	$1.18_{(.53)}$	$1.57_{(.46)}$	$4.7_{(.7)}$	$.64_{(.05)}$	$1.07_{(.45)}$	$1.48_{(.38)}$
$(n,p) = (200, 400)$	$8.3_{(.6)}$	$.68_{(.02)}$	$1.02_{(.46)}$	$1.44_{(.37)}$	$5.4_{(1.0)}$	$.58_{(.03)}$	$.96_{(.40)}$	$1.40_{(.30)}$
$(n,p) = (400, 800)$	$11.3_{(.6)}$	$.66_{(.02)}$	$.98_{(.43)}$	$1.43_{(.35)}$	$6.1_{(1.0)}$	$.54_{(.02)}$	$.93_{(.42)}$	$1.40_{(.33)}$
$\delta_1 = 1/4, \delta_2 = 0$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$6.6_{(.6)}$	$.71_{(.04)}$	$1.47_{(.70)}$	$1.83_{(.62)}$	$4.7_{(.9)}$	$.63_{(.05)}$	$1.30_{(.65)}$	$1.71_{(.57)}$
$(n,p) = (200, 400)$	$8.4_{(.6)}$	$.69_{(.02)}$	$1.42_{(.80)}$	$1.78_{(.68)}$	$5.9_{(1.4)}$	$.59_{(.03)}$	$1.38_{(.78)}$	$1.74_{(.66)}$
$(n,p) = (400, 800)$	$11.2_{(.6)}$	$.66_{(.02)}$	$1.25_{(.65)}$	$1.64_{(.54)}$	$7.3_{(1.3)}$	$.55_{(.03)}$	$1.25_{(.65)}$	$1.64_{(.53)}$
$\delta_1 = 0, \delta_2 = 1/2$	$\widehat{r}$	RMSE	FFE	FE	$\widehat{r}$	RMSE	FFE	FE
$(n,p) = (100, 200)$	$6.7_{(.7)}$	$.71_{(.03)}$	$.96_{(.46)}$	$1.42_{(.43)}$	$4.7_{(1.0)}$	$.64_{(.04)}$	$.96_{(.47)}$	$1.42_{(.44)}$
$(n,p) = (200, 400)$	$8.4_{(.7)}$	$.68_{(.03)}$	$.84_{(.46)}$	$1.30_{(.37)}$	$5.6_{(1.1)}$	$.59_{(.03)}$	$.82_{(.46)}$	$1.29_{(.35)}$
$(n,p) = (400, 800)$	$11.2_{(.7)}$	$.66_{(.02)}$	$.88_{(.55)}$	$1.36_{(.44)}$	$10.7_{(4.1)}$	$.60_{(.05)}$	$.90_{(.56)}$	$1.37_{(.45)}$

Table 4: *Means and standard deviations (in brackets) of estimation errors and forecast errors for Example 2:* $\boldsymbol{\epsilon}_t \sim N(0, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$.

(II): $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$	PC method			Our method		
$\delta_1 = 0, \delta_2 = 0$	RMSE	FFE	FE	RMSE	FFE	FE
$(n,p) = (100,200)$	$.28_{(.08)}$	$1.58_{(.74)}$	$1.93_{(.62)}$	$.34_{(.07)}$	$1.58_{(.74)}$	$1.93_{(.63)}$
$(n,p) = (200,400)$	$.16_{(.02)}$	$1.44_{(.66)}$	$1.80_{(.57)}$	$.20_{(.02)}$	$1.44_{(.65)}$	$1.80_{(.56)}$
$(n,p) = (400,800)$	$.11_{(.02)}$	$1.35_{(.71)}$	$1.76_{(.56)}$	$.14_{(.02)}$	$1.35_{(.71)}$	$1.76_{(.56)}$
$\delta_1 = 0, \delta_2 = 1/4$	RMSE	FFE	FE	RMSE	FFE	FE
$(n,p) = (100,200)$	$.71_{(.13)}$	$1.10_{(.43)}$	$1.49_{(.39)}$	$.40_{(.15)}$	$1.02_{(.45)}$	$1.45_{(.37)}$
$(n,p) = (200,400)$	$.68_{(.11)}$	$0.99_{(.43)}$	$1.43_{(.31)}$	$.23_{(.05)}$	$0.91_{(.42)}$	$1.37_{(.30)}$
$(n,p) = (400,800)$	$.63_{(.12)}$	$1.00_{(.41)}$	$1.45_{(.34)}$	$.15_{(.02)}$	$0.92_{(.40)}$	$1.39_{(.32)}$
$\delta_1 = 1/4, \delta_2 = 0$	RMSE	FFE	FE	RMSE	FFE	FE
$(n,p) = (100,200)$	$.55_{(.21)}$	$1.21_{(.57)}$	$1.63_{(.48)}$	$.36_{(.10)}$	$1.18_{(.58)}$	$1.60_{(.48)}$
$(n,p) = (200,400)$	$.54_{(.21)}$	$1.31_{(.74)}$	$1.68_{(.62)}$	$.22_{(.05)}$	$1.26_{(.75)}$	$1.65_{(.62)}$
$(n,p) = (400,800)$	$.52_{(.22)}$	$1.23_{(.67)}$	$1.62_{(.54)}$	$.15_{(.03)}$	$1.18_{(.66)}$	$1.59_{(.53)}$
$\delta_1 = 0, \delta_2 = 1/2$	RMSE	FFE	FE	RMSE	FFE	FE
$(n,p) = (100,200)$	$.65_{(.03)}$	$.93_{(.43)}$	$1.40_{(.40)}$	$.59_{(.10)}$	$.93_{(.42)}$	$1.40_{(.39)}$
$(n,p) = (200,400)$	$.60_{(.03)}$	$.82_{(.44)}$	$1.29_{(.34)}$	$.51_{(.10)}$	$.81_{(.44)}$	$1.29_{(.34)}$
$(n,p) = (400,800)$	$.56_{(.02)}$	$.86_{(.56)}$	$1.36_{(.45)}$	$.42_{(.12)}$	$.85_{(.56)}$	$1.35_{(.44)}$

Table 5: *Means and standard deviations (in brackets) of estimation errors and forecast errors for Example 2:* $\boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\epsilon}})$*, and the number of factors is fixed at* $r = 3$*.*

5 Data Analysis : Implied Volatility Surfaces

We illustrate our method by modeling the dynamic behavior of IBM, Microsoft and Dell implied volatility surfaces through the period $03/01/2006 - 29/12/2006$ (250 days in total). The data was obtained from OptionMetrics via the WRDS database. For each day t we observe the implied volatility $W_t(u_i, v_j)$ computed from call options. Here u_i is the time to maturity, taking values 30, 60, 91, 122, 152, 182, 273, 365, 547 and 730 for $i = 1, \cdots, p_u = 10$ respectively, and v_j is the delta, taking values 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65, 0.7, 0.75, and 0.8 for $j = 1, \cdots, p_v = 13$ respectively. We collect these implied volatilities in the matrix $\mathbf{W}_t = (W_t(u_i, v_i)) \in \mathbb{R}^{p_u \times p_v}$. Figure 1 displays the mean volatility surface of IBM, Microsoft and Dell in this period. It shows clearly that the implied volatilities surfaces are not flat. Indeed any cross-section in the maturity or delta axis display the well documented volatility smile.

It is a well documented stylized fact that implied volatilities are non-stationary (see Cont and da Fonseca (1988), Fengler et al. (2007) and Park et al. (2009) amongst others). Indeed, when applying the Dickey-Fuller test to each of the univariate time series $W_t(u_i, v_i)$, none of the $p_u \times p_v = 130$ nulls of unit roots could be rejected at the 10% level. Of course we should treat the results of these tests with some caution since we

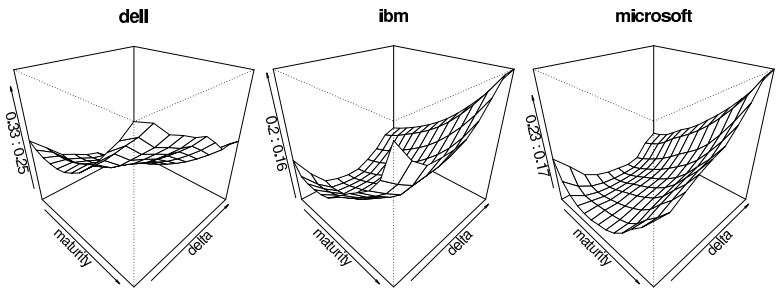


Figure 1: *Mean implied volatility surfaces.*

are performing a large number of hypothesis tests, but even still the evidence in favor of unit roots is overwhelming. Therefore, instead of working with $\mathbf{W}_t$ directly, we choose to work with $\Delta\mathbf{W}_t = \mathbf{W}_t - \mathbf{W}_{t-1}$. Our observations are then $\mathbf{y}_t = \text{vec}\{\Delta\mathbf{W}_t\}$, where for any matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_{p_v}) \in \mathbb{R}^{p_u \times p_v}$, $\text{vec}\{\mathbf{M}\} = (\mathbf{m}_1^T, \ldots, \mathbf{m}_{p_v}^T)^T \in \mathbb{R}^{p_u p_v}$. Note that $\mathbf{y}_t$ is now defined over $04/01/2006 - 29/12/2006$ since we lose an observation due to differencing. Hence altogether there are 249 time points, and the dimension of $\mathbf{y}_t$ is $p = p_v \times p_u = 130$.

We perform the factor model estimation on a rolling window of length 100 days, defined from the i-th day to the $(i + 99)$-th day for $i = 1, \cdots, 150$. The length of the window is chosen so that the stationarity assumption of the data is approximately satisfied. For each window, we compare our method with the PC method by estimating the factor loading matrix and the factor series. For the i-th window, we use an AR model to forecast the $(i + 100)$-th value of the estimated factor series $\mathbf{x}_{i+100}^{(1)}$, so as to obtain a one-step ahead forecast $\mathbf{y}_{i+100}^{(1)} = \widehat{\mathbf{A}}\mathbf{x}_{i+100}^{(1)}$ for $\mathbf{y}_{i+100}$. We then calculate the forecast error for the $(i + 100)$-th day defined by

$$FE = p^{-1/2}\|\mathbf{y}_{i+100}^{(1)} - \mathbf{y}_{i+100}\|.$$

5.1 Estimation results

In forming the matrix $\widetilde{\mathbf{L}}$ for each window, we take $k_0 = 5$ in (2.5) , taking advantage that the autocorrelations are not weak even at higher lags, though similar results (not reported here) are obtained for smaller k_0.

Figure 2 displays the average of each ordered eigenvalue over the 150 windows. The left hand side shows the average of the largest to the average of the tenth largest eigenvalue of $\widetilde{\mathbf{L}}$ for Dell, IBM and Microsoft for our method, whereas the right hand side shows the

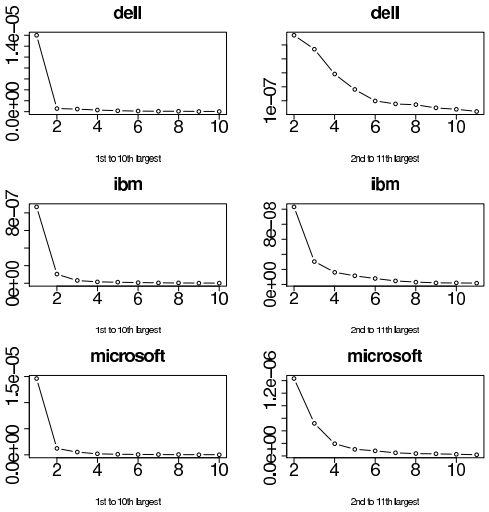


Figure 2: *Averages of ordered eigenvalues of* $\widetilde{\mathbf{L}}$ *over the 150 windows. Left: Ten largest. Right: Second to eleventh largest.*

second to eleventh largest. We obtain similar results for the Bai and Ng (2002) procedure and thus the corresponding graph is not shown.

From this diagram it is apparent that there is one eigenvalue that is much larger than the others for all three companies for each window. We have done automatic selection for the number of factors for each window using the IC criterion by Bai and Ng (2002) introduced in Example 2 in section 4, and a one factor model is consistently obtained for each window and for each company. Hence both methods choose a one factor model over the 150 windows.

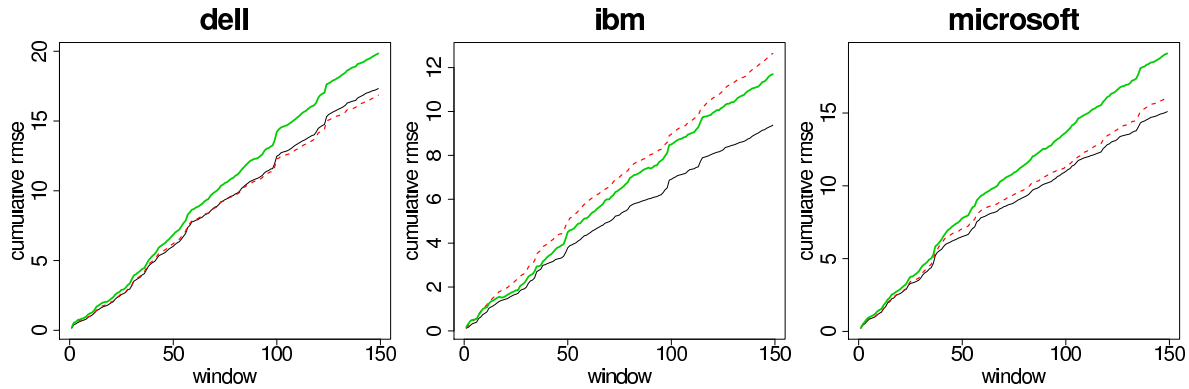
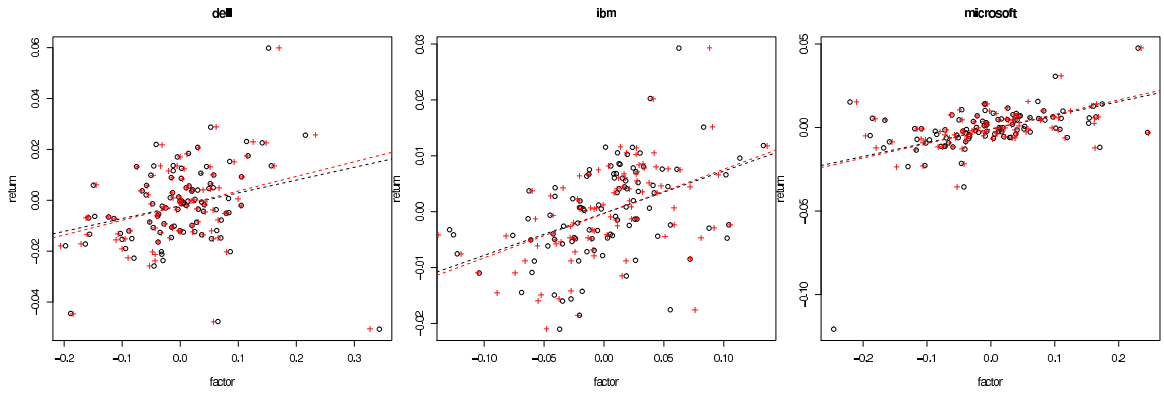


Figure 3: *The cumulative FE over the 150 windows. Red dotted: Bai and Ng (2002) procedure. Green: Taking forecast* $\mathbf{y}_{t+1}^{(1)}$ *to be* $\mathbf{y}_t$*. Black: Our method.*



(a) Dell (b) IBM (c) Microsoft

Figure 4: *Plot of return against the estimated factor for the first window. Black circle: Our method. Red "+": PC method.*

Figure 3 displays the cumulative FE over the 150 windows for each method. We choose a benchmark procedure (green line in each plot), where we just treat today's value as the one-step ahead forecast. Except for Dell where the PC method is doing marginally better, our method consistently outperforms the benchmark procedure and is better than the PC method for IBM and Microsoft.

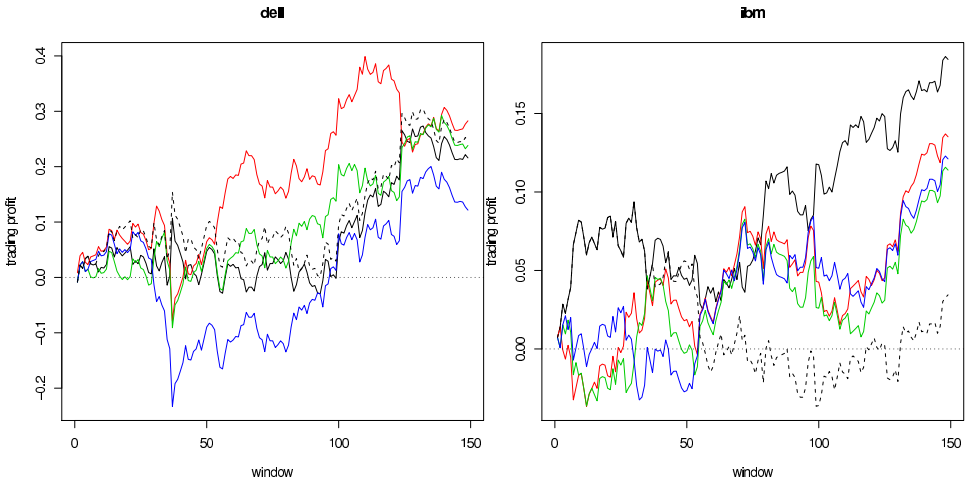
5.2 A simple trading exercise

We use the one-step ahead forecast above to forecast the next day return of the three stocks. Figure 4 shows the plots of return against the estimated factor for all three companies for the data in the first window. Simple linear regression suggests that the slope of the regression lines are significant. Hence we can plug in the one-step ahead forecast of the factor into the estimated linear function to estimate the next day return. All other windows for the three companies show linear pattern with similar plots, and hence we can do this for all the 150 windows.

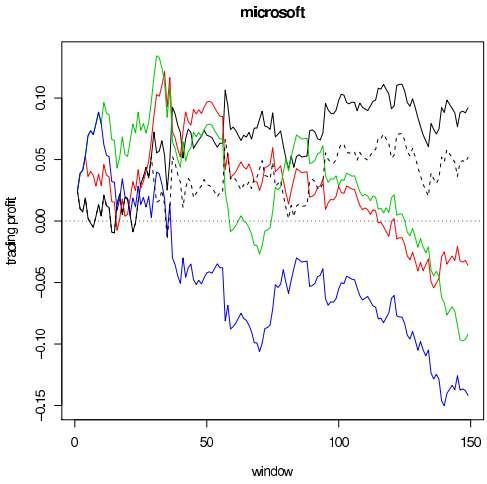
After forecasting the return of the $(t+1)$-th day, if it is higher than that of the t-th day, we buy \$1; otherwise, we sell \$1. Ignoring all trading costs, the accumulated return is calculated at the end of the whole time period. This is done for our method and the PC method. For the benchmark procedure, we calculated the average of the price of a stock for the past 5 days, and compare that to the price today. If the average is higher than the price today, we sell \$1; otherwise we buy \$1. We have two more similar benchmark

procedures, which look at the average price of the past 10 and 15 days respectively.

Figure 5 shows the results for the three companies. Our method outperforms others for IBM and Microsoft. For Dell, both the returns of our method and the PC method stay flat for around the first 100 days, and then gradually go up to perform similarly to other moving average benchmarks in the end.



(a) Dell (b) IBM



(c) Microsoft

Figure 5: *Plot of accumulated returns over time. Black: Our method. Black dotted: PC method. Red, Green and Blue lines are respectively benchmark procedure looking at 5, 10 and 15 days moving average price.*

6 Proofs

Before proving the theorems in section 3, we need to have three lemmas.

Lemma 1 *Under model (2.2) with assumptions (A) - (G) in sections 2.1 and 2.2, we have*

$$\|\mathbf{\Sigma_f}(k)\| \asymp p^{1-\delta} \asymp \|\mathbf{\Sigma_f}(k)\|_{\min}, \quad \|\mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\| = O(p^{1-\delta/2}).$$

Proof. Model (2.2) is an equivalent representation of model (2.1), where

$$\mathbf{y}_t = \mathbf{A}\mathbf{x}_t + \boldsymbol{\epsilon}_t = \mathbf{Q}\mathbf{f}_t + \boldsymbol{\epsilon}_t,$$

with $\mathbf{A} = \mathbf{QR}$ and $\mathbf{f}_t = \mathbf{R}\mathbf{x}_t$. With assumptions (F) and (G), the diagonal entries of $\mathbf{R}$ are all asymptotic to $p^{\frac{1-\delta}{2}}$ (which is the order of $\|\mathbf{a}_i\|$), and the off-diagonal entries are of smaller order. Hence, as r is a constant, using

$$\|\mathbf{R}\| = \max_{\|\mathbf{u}\|=1} \|\mathbf{R}\mathbf{u}\|, \quad \|\mathbf{R}\|_{\min} = \min_{\|\mathbf{u}\|=1} \|\mathbf{R}\mathbf{u}\|,$$

we can conclude that

$$\|\mathbf{R}\| \asymp p^{\frac{1-\delta}{2}} \asymp \|\mathbf{R}\|_{\min}.$$

This, together with $\mathbf{\Sigma_f}(k) = \mathrm{Cov}(\mathbf{f}_{t+k}, \mathbf{f}_t) = \mathrm{Cov}(\mathbf{R}\mathbf{x}_{t+k}, \mathbf{R}\mathbf{x}_t) = \mathbf{R}\mathbf{\Sigma_x}(k)\mathbf{R}^T$ for $k = 1, \cdots, k_0$, implies

$$p^{1-\delta} \asymp \|\mathbf{R}\|^2_{\min} \cdot \|\mathbf{\Sigma_x}(k)\|_{\min} \le \|\mathbf{\Sigma_f}(k)\|_{\min} \le \|\mathbf{\Sigma_f}(k)\| \le \|\mathbf{R}\|^2 \cdot \|\mathbf{\Sigma_x}(k)\| \asymp p^{1-\delta},$$

where we used assumption (B) to arrive at $\|\mathbf{\Sigma_x}(k)\| \asymp 1 \asymp \|\mathbf{\Sigma_x}(k)\|_{\min}$, so that

$$\|\mathbf{\Sigma_f}(k)\| \asymp p^{1-\delta} \asymp \|\mathbf{\Sigma_f}(k)\|_{\min}.$$

We used the inequality $\|\mathbf{AB}\|_{\min} \ge \|\mathbf{A}\|_{\min} \cdot \|\mathbf{B}\|_{\min}$ for any square matrices $\mathbf{A}$ and $\mathbf{B}$, which can be proved by noting

$$\begin{aligned}\|\mathbf{AB}\|_{\min} &= \min_{\mathbf{u}\neq\mathbf{0}} \frac{\mathbf{u}^T\mathbf{B}^T\mathbf{A}^T\mathbf{ABu}}{\|\mathbf{u}\|^2} \ge \min_{\mathbf{u}\neq\mathbf{0}} \frac{(\mathbf{Bu})^T\mathbf{A}^T\mathbf{A}(\mathbf{Bu})}{\|\mathbf{Bu}\|^2} \cdot \frac{\|\mathbf{Bu}\|^2}{\|\mathbf{u}\|^2} \\ &\ge \min_{\mathbf{w}\neq\mathbf{0}} \frac{\mathbf{w}^T\mathbf{A}^T\mathbf{Aw}}{\|\mathbf{w}\|^2} \cdot \min_{\mathbf{u}\neq\mathbf{0}} \frac{\|\mathbf{Bu}\|^2}{\|\mathbf{u}\|^2} = \|\mathbf{A}\|_{\min} \cdot \|\mathbf{B}\|_{\min}. \end{aligned} \tag{6.1}$$

Finally, using assumption (C) that $\mathbf{\Sigma}_{\mathbf{x},\boldsymbol{\epsilon}}(k) = O(1)$ elementwisely, and that it has $rp \asymp p$ elements, we have

$$\|\mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\| = \|\mathbf{R}\mathbf{\Sigma}_{\mathbf{x},\boldsymbol{\epsilon}}(k)\| \le \|\mathbf{R}\| \cdot \|\mathbf{\Sigma}_{\mathbf{x},\boldsymbol{\epsilon}}(k)\|_F = O(p^{\frac{1-\delta}{2}}) \cdot O(p^{1/2}) = O(p^{1-\delta/2}),$$

where $\|M\|_F := \mathrm{trace}(MM^T)$ denotes the Frobenius norm of the matrix M. □

Lemma 2 *Under model (2.2) and assumption (E) in section 2.1, we have for* $0 \le k \le k_0$,

$$\|\widetilde{\mathbf{\Sigma}}_{\mathbf{f}}(k) - \mathbf{\Sigma}_{\mathbf{f}}(k)\| = O_P(p^{1-\delta}n^{-1/2}), \qquad \|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k) - \mathbf{\Sigma}_{\boldsymbol{\epsilon}}(k)\| = O_P(pn^{-1/2}),$$
$$\|\widetilde{\mathbf{\Sigma}}_{\mathbf{f},\boldsymbol{\epsilon}}(k) - \mathbf{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\| = O_P(p^{1-\delta/2}n^{-1/2}) = \|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon},\mathbf{f}}(k) - \mathbf{\Sigma}_{\boldsymbol{\epsilon},\mathbf{f}}(k)\|,$$

Moreover, $\|\mathbf{f}_t\|^2 = O_P(p^{1-\delta})$ *for all integers* $t \ge 0$.

Proof. From (2.1) and (2.2), we have the relation $\mathbf{f}_t = \mathbf{R}\mathbf{x}_t$, where $\mathbf{R}$ is an upper triangular matrix with $\|\mathbf{R}\| \asymp p^{\frac{1-\delta}{2}} \asymp \|\mathbf{R}\|_{\min}$ (see the proof of Lemma 1). Then we immediately have $\|\mathbf{f}_t\|^2 \le \|\mathbf{R}\|^2 \cdot \|\mathbf{x}_t\|^2 = O_P(p^{1-\delta}r) = O_P(p^{1-\delta})$.

Also, the covariance matrix and the sample covariance matrix for $\{\mathbf{f}_t\}$ are respectively

$$\mathbf{\Sigma}_{\mathbf{f}}(k) = \mathbf{R}\mathbf{\Sigma}_{\mathbf{x}}(k)\mathbf{R}^T, \quad \widetilde{\mathbf{\Sigma}}_{\mathbf{f}}(k) = \mathbf{R}\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}(k)\mathbf{R}^T.$$

Hence

$$\begin{aligned}\|\widetilde{\mathbf{\Sigma}}_{\mathbf{f}}(k) - \mathbf{\Sigma}_{\mathbf{f}}(k)\| &\le \|\mathbf{R}\|^2 \cdot \|\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}(k) - \mathbf{\Sigma}_{\mathbf{x}}(k)\| \\ &= O(p^{1-\delta}) \cdot O_P(n^{-1/2} \cdot r) \\ &= O_P(p^{1-\delta}n^{-1/2}),\end{aligned}$$

which is the rate specified in the lemma. We used the fact that the matrix $\widetilde{\mathbf{\Sigma}}_{\mathbf{x}}(k) - \mathbf{\Sigma}_{\mathbf{x}}(k)$ has r^2 elements, with elementwise rate of convergence being $O(n^{-1/2})$ which is implied by assumption (E) and that $\{\boldsymbol{\epsilon}_t\}$ is white noise. Other rates can be derived similarly using the Frobenius norm as an upper bound. □

The following is Theorem 8.1.10 in Golub and Van Loan (1996), which is stated explicitly since our main theorems are based on this. See Johnstone and Arthur (2009) also.

Lemma 3 *Suppose* $\mathbf{A}$ *and* $\mathbf{A} + \mathbf{E}$ *are* $n \times n$ *symmetric matrices and that*

$$\mathbf{Q} = [\mathbf{Q}_1 \;\; \mathbf{Q}_2] \quad (\mathbf{Q}_1 \;\; \textit{is} \;\; n \times r, \; \mathbf{Q}_2 \;\; \textit{is} \;\; n \times (n-r))$$

is an orthogonal matrix such that $\mathrm{span}(\mathbf{Q}_1)$ *is an invariant subspace for* $\mathbf{A}$ *(that is,* $\mathbf{A} \cdot \mathrm{span}(\mathbf{Q}_1) \subset \mathrm{span}(\mathbf{A})$*). Partition the matrices* $\mathbf{Q}^T\mathbf{A}\mathbf{Q}$ *and* $\mathbf{Q}^T\mathbf{E}\mathbf{Q}$ *as follows:*

$$\mathbf{Q}^T\mathbf{A}\mathbf{Q} = \begin{pmatrix} \mathbf{D}_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{D}_2 \end{pmatrix} \qquad \mathbf{Q}^T\mathbf{E}\mathbf{Q} = \begin{pmatrix} \mathbf{E}_{11} & \mathbf{E}_{21}^T \\ \mathbf{E}_{21} & \mathbf{E}_{22} \end{pmatrix}.$$

If $\text{sep}(\mathbf{D}_1, \mathbf{D}_2) := \min_{\lambda \in \lambda(\mathbf{D}_1),\, \mu \in \lambda(\mathbf{D}_2)} |\lambda - \mu| > 0$*, where* $\lambda(M)$ *denotes the set of eigenvalues of the matrix* M*, and*

$$\|\mathbf{E}\| \leq \frac{\text{sep}(\mathbf{D}_1, \mathbf{D}_2)}{5},$$

then there exists a matrix $\mathbf{P} \in \mathbb{R}^{(n-r)\times r}$ *with*

$$\|\mathbf{P}\| \leq \frac{4}{\text{sep}(\mathbf{D}_1, \mathbf{D}_2)} \|\mathbf{E}_{21}\|$$

such that the columns of $\widehat{\mathbf{Q}}_1 = (\mathbf{Q}_1 + \mathbf{Q}_2\mathbf{P})(\mathbf{I} + \mathbf{P}^T\mathbf{P})^{-1/2}$ *define an orthonormal basis for a subspace that is invariant for* $\mathbf{A} + \mathbf{E}$*.*

In the proofs thereafter, we use $\otimes$ to denote the Kronecker product of matrices, and $\sigma_j(M)$ to denote the j-th singular value of the matrix M. Hence $\sigma_1(M) = \|M\|$. We use $\lambda_j(M)$ to denote the j-th largest eigenvalue of M.

Proof of Theorem 1. Under model (2.2), we have shown in section 2.3 that we have $\mathbf{LQU} = \mathbf{QUD}$. Since $\mathbf{U}$ is an orthogonal matrix, we have

$$\mathbf{y}_t = \mathbf{Q}\mathbf{f}_t + \boldsymbol{\epsilon}_t = (\mathbf{QU})(\mathbf{U}^T\mathbf{f}_t) + \boldsymbol{\epsilon}_t,$$

so that we can replace $\mathbf{QU}$ with $\mathbf{Q}$ and $\mathbf{U}^T\mathbf{f}_t$ with $\mathbf{f}_t$ in the model, thus making $\mathbf{LQ} = \mathbf{QD}$, where now $\mathbf{D}$ is diagonal with

$$\mathbf{D} = \sum_{k=1}^{k_0} \{\boldsymbol{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\}\{\boldsymbol{\Sigma}_{\mathbf{f}}(k)\mathbf{Q}^T + \boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k)\}^T.$$

If $\mathbf{B}$ is an orthogonal complement of $\mathbf{Q}$, then $\mathbf{LB} = \mathbf{0}$, and

$$\begin{pmatrix} \mathbf{Q}^T \\ \mathbf{B}^T \end{pmatrix} \mathbf{L} (\mathbf{Q} \;\; \mathbf{B}) = \begin{pmatrix} \mathbf{D} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix}, \tag{6.2}$$

with $\text{sep}(\mathbf{D}, \mathbf{0}) = \lambda_{\min}(\mathbf{D})$ (see Lemma 3 for the definition of the function sep). We now find the order of $\lambda_{\min}(\mathbf{D})$.

To this end, define

$$\mathbf{W}_{\mathbf{f}}(k_0) = (\boldsymbol{\Sigma}_{\mathbf{f}}(1), \cdots, \boldsymbol{\Sigma}_{\mathbf{f}}(k_0)), \quad \mathbf{W}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0) = (\boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(1), \cdots, \boldsymbol{\Sigma}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0)),$$

so that we have $\mathbf{D} = (\mathbf{W}_{\mathbf{f}}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{Q}^T) + \mathbf{W}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0))(\mathbf{W}_{\mathbf{f}}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{Q}^T) + \mathbf{W}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0))^T$. Hence, assuming first that $\kappa_{\max} = o(p^{1-\delta})$, we have

$$\begin{aligned}
\lambda_{\min}(\mathbf{D}) &= \{\sigma_r(\mathbf{W}_{\mathbf{f}}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{Q}^T) + \mathbf{W}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0))\}^2 \\
&\geq \{\sigma_r(\mathbf{W}_{\mathbf{f}}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{Q}^T)) - \sigma_1(\mathbf{W}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0))\}^2 \\
&= \{\sigma_r(\mathbf{W}_{\mathbf{f}}(k_0)) - \sigma_1(\mathbf{W}_{\mathbf{f},\boldsymbol{\epsilon}}(k_0))\}^2 \\
&\asymp \sigma_r(\mathbf{W}_{\mathbf{f}}(k_0))^2 \asymp p^{2-2\delta},
\end{aligned}$$

where we use $\|\mathbf{\Sigma_f}(k)\|_{\min} \asymp p^{1-\delta}$ from Lemma 1. On the other hand, if $p^{1-\delta} = o(\kappa_{\min})$, then we have

$$\begin{aligned}\lambda_{\min}(\mathbf{D}) &\geq \{\sigma_r(\mathbf{W_{f,\epsilon}}(k_0)) - \sigma_1(\mathbf{W_f}(k_0)(\mathbf{I}_{k_0} \otimes \mathbf{Q}^T))\}^2 \\ &= \{\sigma_r(\mathbf{W_{f,\epsilon}}(k_0)) - \sigma_1(\mathbf{W_f}(k_0))\}^2 \\ &\asymp \sigma_r(\mathbf{W_{f,\epsilon}}(k_0))^2 \asymp \kappa_{\min}^2.\end{aligned}$$

Hence we have

$$\max(\kappa_{\min}^2, p^{2-2\delta}) = O(\lambda_{\min}(\mathbf{D})). \tag{6.3}$$

Next, we need to find $\|\mathbf{E_L}\|$, where we define $\mathbf{E_L} = \widetilde{\mathbf{L}} - \mathbf{L}$, with $\widetilde{\mathbf{L}}$ defined in (2.5). Then it is easy to see that

$$\|\mathbf{E_L}\| \leq \sum_{k=1}^{k_0} \Big\{ \|\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) - \mathbf{\Sigma_y}(k)\|^2 + 2\|\mathbf{\Sigma_y}(k)\| \cdot \|\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) - \mathbf{\Sigma_y}(k)\| \Big\}. \tag{6.4}$$

Consider for $k \geq 1$, using the results from Lemma 1,

$$\|\mathbf{\Sigma_y}(k)\| = \|\mathbf{Q}\mathbf{\Sigma_f}(k)\mathbf{Q}^T + \mathbf{Q}\mathbf{\Sigma_{f,\epsilon}}(k)\| \leq \|\mathbf{\Sigma_f}(k)\| + \|\mathbf{\Sigma_{f,\epsilon}}(k)\| = O(p^{1-\delta} + \kappa_{\max}). \tag{6.5}$$

Also, for $k = 1, \cdots, k_0$, using the results in Lemma 2,

$$\begin{aligned}\|\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) - \mathbf{\Sigma_y}(k)\| \leq& \|\widetilde{\mathbf{\Sigma}}_{\mathbf{f}}(k) - \mathbf{\Sigma_f}(k)\| + 2\|\widetilde{\mathbf{\Sigma}}_{\mathbf{f,\epsilon}}(k) - \mathbf{\Sigma_{f,\epsilon}}(k)\| + \|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\| \\ =& O_P(p^{1-\delta}n^{-1/2} + p^{1-\delta/2}n^{-1/2} + \|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\|) \\ =& O_P(p^{1-\delta/2}n^{-1/2} + \|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\|).\end{aligned} \tag{6.6}$$

Without further assumptions on $\{\boldsymbol{\epsilon}_t\}$, we have $\|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\| \leq \|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\|_F = O_P(pn^{-1/2})$, which implies from (6.6) that

$$\|\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) - \mathbf{\Sigma_y}(k)\| = O_P(pn^{-1/2}). \tag{6.7}$$

With (6.5) and (6.7), we can easily see from (6.4) that

$$\|\mathbf{E_L}\| = O_P(p^{2-\delta}n^{-1/2} + \kappa_{\max} \cdot pn^{-1/2}). \tag{6.8}$$

Finally, no matter $\kappa_{\max} = o(p^{1-\delta})$ or $p^{1-\delta} = o(\kappa_{\min})$, we have from (6.8) and (6.3) that

$$\begin{aligned}\|\mathbf{E_L}\| &= O_P(p^{2-\delta}n^{-1/2} + \kappa_{\max} \cdot pn^{-1/2}) = o_P(\max(p^{2-2\delta}, \kappa_{\min}^2)) \\ &= O_P(\lambda_{\min}(\mathbf{D})) = O_P(\text{sep}(\mathbf{D}, \mathbf{0})),\end{aligned}$$

since we assumed $h_n = p^\delta n^{-1/2} = o(1)$ in the former case, or $\kappa_{\min}^{-2}\kappa_{\max} \cdot pn^{-1/2} = o(1)$ in the latter. Hence for sufficient large n, we have $\|\mathbf{E_L}\| \leq \text{sep}(\mathbf{D}, \mathbf{0})/5$. This allows us to apply Lemma 3 to conclude that there exists a matrix $\mathbf{P} \in \mathbb{R}^{(p-r)\times r}$ such that

$$\|\mathbf{P}\| \leq \frac{4}{\text{sep}(\mathbf{D}, \mathbf{0})}\|(\mathbf{E_L})_{21}\| \leq \frac{4}{\text{sep}(\mathbf{D}, 0)}\|\mathbf{E_L}\|,$$

and $\widehat{\mathbf{Q}} = (\mathbf{Q} + \mathbf{BP})(\mathbf{I} + \mathbf{P}^T\mathbf{P})^{-1/2}$ is an estimator for $\mathbf{Q}$. Then we have

$$\begin{aligned}\|\widehat{\mathbf{Q}} - \mathbf{Q}\| &= \|(\mathbf{Q}(\mathbf{I} - (\mathbf{I} + \mathbf{P}^T\mathbf{P})^{1/2}) + \mathbf{BP})(\mathbf{I} + \mathbf{P}^T\mathbf{P})^{-1/2}\| \\ &\leq \|\mathbf{I} - (\mathbf{I} + \mathbf{P}^T\mathbf{P})^{1/2}\| + \|\mathbf{P}\| \\ &\leq 2\|\mathbf{P}\|,\end{aligned}$$

and using (6.3) and (6.8),

$$\|\mathbf{P}\| = O_P\left(\frac{p^{2-\delta}n^{-1/2} + \kappa_{\max} \cdot pn^{-1/2}}{\max(\kappa_{\min}^2, p^{2-2\delta})}\right) = \begin{cases} O_P(p^\delta n^{-1/2}), & \text{if } \kappa_{\max} = o(p^{1-\delta}); \\ O_P(\kappa_{\min}^{-2}\kappa_{\max} \cdot pn^{-1/2}), & \text{if } p^{1-\delta} = o(\kappa_{\min}). \end{cases}$$

This completes the proof of the theorem. □

Proof of Theorem 2. Under assumptions (H) and (I), if we can show that

$$\|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\| = O_P(pn^{-1}), \tag{6.9}$$

then (6.6) becomes

$$\|\widetilde{\mathbf{\Sigma}}_{\mathbf{y}}(k) - \mathbf{\Sigma}_{\mathbf{y}}(k)\| = O_P(p^{1-\delta/2}n^{-1/2} + pn^{-1}) = O_P(p^{1-\delta/2}n^{-1/2}),$$

where we use the assumption $p^{\delta/2}n^{-1/2} = o(1)$. This rate is smaller than that in (6.7) by a factor of $p^{\delta/2}$, which carries to other parts of the proof of Theorem 1, so that the final rates are all smaller by a factor of $p^{\delta/2}$. Hence, it remains to show (6.9).

To this end, define $\mathbf{1}_k$ the column vector of k ones, and

$$\mathbf{E}_{r,s} = (\boldsymbol{\epsilon}_r, \cdots, \boldsymbol{\epsilon}_s) \text{ for } r \leq s.$$

Since the asymptotic behavior of the three sample means

$$\bar{\boldsymbol{\epsilon}} = n^{-1}\mathbf{E}_{1,n}\mathbf{1}_n, \;\; (n-k)^{-1}\mathbf{E}_{1,n-k}\mathbf{1}_{n-k}, \;\; (n-k)^{-1}\mathbf{E}_{k+1,n}\mathbf{1}_{n-k}$$

are exactly the same as k is finite and $\{\boldsymbol{\epsilon}_t\}$ is stationary, in this proof we take the sample lag-k autocovariance matrix for $\{\boldsymbol{\epsilon}_t\}$ to be

$$\begin{aligned}\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k) &= n^{-1}(\mathbf{E}_{k+1,n} - (n-k)^{-1}\mathbf{E}_{k+1,n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}^T)(\mathbf{E}_{1,n-k} - (n - k^{-1}\mathbf{E}_{1,n-k}\mathbf{1}_{n-k}\mathbf{1}_{n-k}^T))^T \\ &= n^{-1}\mathbf{E}_{k+1,n}\mathbf{T}_{n-k}\mathbf{E}_{1,n-k}^T,\end{aligned}$$

where $\mathbf{T}_j = \mathbf{I}_j - j^{-1}\mathbf{1}_j\mathbf{1}_j'$. Then under conditions (H) and (I),

$$\begin{aligned}\|\widetilde{\mathbf{\Sigma}}_{\boldsymbol{\epsilon}}(k)\| &\leq \|n^{-1/2}\mathbf{E}_{k+1,n}\| \cdot \|\mathbf{T}_{n-k}\| \cdot \|n^{-1/2}\mathbf{E}_{1,n-k}\| \\ &= \lambda_1^{1/2}(n^{-1}\mathbf{E}_{k+1,n}^T\mathbf{E}_{k+1,n}) \cdot \lambda_1^{1/2}(n^{-1}\mathbf{E}_{1,n-k}^T\mathbf{E}_{1,n-k}) \\ &= O_P((1+(pn^{-1})^{1/2}) \cdot (1+(pn^{-1})^{1/2})) \\ &= O_P(pn^{-1}),\end{aligned}$$

where the second last line follows from Theorem 1.3 of Péché (2009) for the covariance matrices $n^{-1}\mathbf{E}_{k+1,n}^T\mathbf{E}_{k+1,n}$ and $n^{-1}\mathbf{E}_{1,n-k}^T\mathbf{E}_{1,n-k}$, and the last line follows from the assumption $n = O(p)$. This completes the proof of the theorem. □

Proof of Theorem 3. Consider

$$\begin{aligned}\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t - \mathbf{Q}\mathbf{f}_t &= \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T\mathbf{y}_t - \mathbf{Q}\mathbf{f}_t = \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T\mathbf{Q}\mathbf{f}_t - \mathbf{Q}\mathbf{f}_t + \widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T\boldsymbol{\epsilon}_t \\ &= (\widehat{\mathbf{Q}}\widehat{\mathbf{Q}}^T - \mathbf{Q}\mathbf{Q}^T)\mathbf{f}_t + \widehat{\mathbf{Q}}(\widehat{\mathbf{Q}} - \mathbf{Q})^T\boldsymbol{\epsilon}_t + \widehat{\mathbf{Q}}\mathbf{Q}^T\boldsymbol{\epsilon}_t \\ &:= K_1 + K_2 + K_3.\end{aligned}$$

Using Lemma 2, we have

$$\|K_1\| = O_P(\|\widehat{\mathbf{Q}} - \mathbf{Q}\| \cdot \|\mathbf{f}_t\|) = O_P(p^{\frac{1-\delta}{2}}\|\widehat{\mathbf{Q}} - \mathbf{Q}\|).$$

Also, since $\|\widehat{\mathbf{Q}} - \mathbf{Q}\| = o_P(1)$ and $\|\mathbf{Q}\| = 1$, we have K_2 dominated by K_3 in probability. Hence we only need to consider K_3. Now consider for $\mathbf{Q} = (\mathbf{q}_1, \cdots, \mathbf{q}_r)$, the random variable $\mathbf{q}_j^T\boldsymbol{\epsilon}_t$, with

$$E(\mathbf{q}_j^T\boldsymbol{\epsilon}_t) = 0, \quad \text{Var}(\mathbf{q}_j^T\boldsymbol{\epsilon}_t) = \mathbf{q}_j^T\mathbf{\Sigma}_{\boldsymbol{\epsilon}}\mathbf{q}_j \leq \lambda_{\max}(\mathbf{\Sigma}_{\boldsymbol{\epsilon}}) < c < \infty$$

for $j = 1, \cdots, r$ by assumption, where c is a constant independent of n and r. Hence $\mathbf{q}_j^T\boldsymbol{\epsilon}_t = O_P(1)$. We then have

$$\|K_3\| = \|\widehat{\mathbf{Q}}\mathbf{Q}^T\boldsymbol{\epsilon}_t\| \leq \|\mathbf{Q}^T\boldsymbol{\epsilon}_t\| = \sum_{j=1}^{r}(\mathbf{q}_j^T\boldsymbol{\epsilon}_t)^2 = O_P(1).$$

Hence $p^{-1/2}\|\widehat{\mathbf{Q}}\widehat{\mathbf{f}}_t - \mathbf{Q}\mathbf{f}_t\| = O_P(p^{-\delta/2}\|\widehat{\mathbf{Q}} - \mathbf{Q}\| + p^{-1/2})$, which completes the proof of the theorem. □

References

Anderson, T. (1963). The use of factor analysis in the statistical analysis of multiple time series. *Psychometrika 28*, 1–25.

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

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

Bai, J. and S. Ng (2007). Determining the number of primitive shocks in factor models. *Journal of Business & Economic Statistics 25*, 52–60.

Bathia, N., Q. Yao, and F. Zieglemann (2010). Identifying the finite dimensionality of curve time series. *Ann. Statist.*, to appear.

Brillinger, D. (1981). *Time Series Data Analysis and Theory* (Extended ed.). San Francisco: Holden-Day.

Chamberlain, G. (1983). Funds, factors, and diversification in arbitrage pricing models. *Econometrica 51*, 1305–1323.

Chamberlain, G. and M. Rothschild (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica 51*, 1281–1304.

Chudik, A., M. H. Pesaran, and E. Tosetti (2009). Weak and strong cross section dependence and estimation of large panels. Manuscript.

Cont, R. and J. da Fonseca (1988). Dynamics of implied volatility surfaces. *Quantitative Finance 2*, 45–60.

Fengler, M., W. Hardle, and E. Mammen (2007). A dynamic semiparametric factor model for implied volatility string dynamics. *Journal of Econometrics 5*, 189–218.

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

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

Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2005). The generalized dynamic factor model: One-sided estimation and forecasting. *J. Amer. Statist. Assoc. 100*, 830–840.

Golub, G. and C. Van Loan (1996). *Matrix Computations* (3rd ed.). Johns Hopkins University Press.

Hallin, M. and R. Liška (2007). Determining the number of factors in the general dynamic factor model. *J. Amer. Statist. Assoc. 102*, 603–617.

Johnstone, I. and Y. Arthur (2009). On consistency and sparsity for principal components analysis in high dimensions. *J. Amer. Statist. Assoc. 104*, 682–693.

Lam, C. and Q. Yao (2010). Factor modelling for high dimensional time series. Manuscript.

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

Park, B., E. Mammen, W. Hardle, and S. Borak (2009). Modelling dynamic semiparametric factor models. *J. Amer. Statist. Assoc.*. forthcoming.

Peña, D. and G. Box (1987). Identifying a simplifying structure in time series. *J. Amer. Statist. Assoc. 82*, 836–843.

Peña, D. and P. Poncela (2006). Nonstationay dynamic factor analysis. *Journal of Statistical Planning and Inference 136*, 1237–1257.

Péché, S. (2009). Universality results for the largest eigenvalues of some sample covariance matrix ensembles. *Probab. Theory Relat. Fields 143*, 481–516.

Priestley, M., T. Rao, and J. Tong (1974). Applications of principal component analysis and factor analysis in the identification of multivariable systems. *IEEE Trans. Automat. Control 19*, 703–704.