Estimation of Latent Factors for High-Dimensional Time Series*

Clifford Lam and Qiwei Yao
Department of Statistics, The London School of Economics and Political Science, Houghton Street, London, WC2A 2AE, U.K.
and Neil Bathia
UBS AG, 100 Liverpool Street, London EC2M 2RH UK

April 7, 2011

Abstract

This paper deals with the dimension reduction of high-dimensional time series based on a lower-dimensional factor process. In particular we allow the dimension of time series N to be as large as, or even larger than, the length of observed time series (also refereed as the sample size) T. The estimation of the factor loading matrix and the factor process itself is carried out via an eigenanalysis of a $N \times N$ 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 $N^{1/2}$, the estimator of the factor loading matrix is weakly consistent in L_2-norm with the convergence rate independent of N. 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.

Keywords: Convergence in L_2-norm; Curse and blessing of dimensionality; Dimension reduction; Eigen-analysis; Factor model.

*Financial support from the STICERD and LSE Annual Fund, and the Engineering and Physical Sciences Research Council (UK) is gratefully acknowledged.

1 Introduction

Analysis of large data sets is an integral part of modern scientific research. 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 for 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 Autoregressive or vector Autoregressive-Moving-Average models are not practically viable when the dimension of time series N is high, as the number of parameters involved is in the order of N^2. Furthermore, one may face a serious model-identification problem in a vector Autoregressive-Moving-Average model. In fact such model has hardly been used in practice without further regularization on its matrix coefficients. Therefore dimension-reduction is a pertinent 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 using 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 N is as large as, or even larger than, the length of the time series T, more recent efforts focus on the inference when N goes to infinity together with T. 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* that affect the dynamics of most of the N component 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. 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 N (i.e. the number of the component series) goes to infinity; see Chamberlain and Rothschild (1983) and Chamberlain (1983). Hence those econometric factor models are only asymptotically identifiable when $N \to \infty$. See also Forni et al. (2000).

We adopt a different approach in this paper from a dimension-reduction point of view. Our model is similar to those in Peña and Box (1987), Peña and Poncela (2006), and Pan and Yao (2008), and we consider the inference when N is as large as, or even larger than, T. Different from the aforementioned econometric factor models, we decompose the N-dimensional time series into two parts: the dynamic part driven by r factors ($r \leq N$) and the static part which is a vector white noise. Hence the r factors in our model consist of the common factors as well as each serial-correlated idiosyncratic component in econometric factor models. Since the white noise exhibits no serial correlations, the decomposition is unique in the sense that both the number of factors r and the factor loading space in our model are

identifiable for any finite N. Furthermore, we allow the future factors to depend on past (white) noise. This substantially enlarges the capacity of the model. Such a conceptually simple decomposition also makes the statistical inference easier; the estimation for the factor loading space and the factor process itself is equivalent to an eigenanalysis of a $N \times N$ non-negative definite matrix. Therefore it is applicable when N 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 N is large, and is invalid when $N > T$. 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 N. Our approach identifies factors based on the autocorrelation structure of the data, which conceptually is more relevant in retaining the dynamics of time series than the least squares approach advocated by Bai and Ng (2002) and Bai (2003).

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 N-dimensional time series converges at a rate independent of N, provided that all the factors are strong in the sense that the norm of each column in the factor loading matrix is of order $N^{1/2}$. Our simulation indicates that the estimation errors are indeed independent of N. 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 N. 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(ii) in section 3.

2 Models and estimation methodology

2.1 Factor models and identifiability

Let $y_1, \cdots, y_n$ be T $N \times 1$ successive observations from a vector time series process. The factor model decomposes y_t into two parts:

$$y_t = Ax_t + \epsilon_t, \tag{1}$$

where $\{x_t\}$ is a $r \times 1$ unobserved factor time series which is assumed to be weakly stationary with finite first two moments, A is a $N \times r$ unknown constant factor loading matrix, $r(\leq N)$ is the number of factors,

and $\{\epsilon_t\}$ is a white noise with mean 0 and covariance matrix Σ_ϵ. Note that the decomposition (2.1) always holds. However it is only useful when $r << N$, as then the dimension-reduction is achieved in the sense that the serial correlation of y_t is driven by a much lower dimensional process x_t.

Model (1) is unchanged if we replace the pair (A, x_t) on the right hand side by $(AH, H^{-1}x_t)$ for any invertible H. However the linear space spanned by the columns of A, denoted by $\mathcal{M}(A)$ and called the factor loading space, is uniquely defined by (1). Note $\mathcal{M}(A) = \mathcal{M}(AH)$ for any invertible H. Once such an A is specified, the factor process x_t is uniquely defined accordingly. We see the lack of uniqueness of A as an advantage, as we may choose a particular A which facilitates our estimation in a simple and convenient manner. On the other hand, we can always rotate an estimated factor loading matrix whenever appropriate.

To highlight the key idea of our approach, we give a heuristic account on the estimation method now before introducing it more formally in section 2.4 below. Based on the above discussion, we may choose A to be a half orthogonal matrix in the sense that $A'A = I_r$, where I_r is the $r \times r$ identity matrix. Let B be a $N \times (N - r)$ matrix for which (A, B) forms a $N \times N$ orthogonal matrix. Hence $A'B = 0$, i.e. the columns of A are perpendicular to the columns of B. For simplicity we assume, for the time being, factor x_t and white noise ϵ_t are uncorrelated of each other across all lags. (This condition will be relaxed; see conditions (C) and (D) below.) It follows from (1) that

$$\Sigma_y(k) = A\Sigma_x(k)A', \qquad k = 1, 2, \cdots,$$

where $\Sigma_y(k) = \text{cov}(y_{t+k}, y_t)$ and $\Sigma_x(k) = \text{cov}(x_{t+k}, x_t)$. Hence $\Sigma_y(k)B = 0$, i.e. the columns of B are the orthonormal eigenvectors of $\Sigma_y(k)$ corresponding to the zero eigenvalues. Hence as long as $\Sigma_y(k)$ is full-ranked (see condition (B) below), $\mathcal{M}(A)$ is the orthogonal compliment of the linear space spanned by the eigenvectors of $\Sigma_y(k)$ corresponding to the zero eigenvalues. Based on this observation, we introduce a non-negative definite matrix

$$L^\star = \sum_{k=1}^{k_0} \Sigma_y(k)\Sigma_y(k)',$$

where $k_0 \geq 1$ is a prescribed integer. Since $L^\star B = 0$ and the eigenvectors of $L^\star$ corresponding to different eigenvalues are orthogonal of each other, we conclude that the number of factors r is the number of non-zero eigenvalues of $L^\star$, and the columns of A may be taken as the r orthonormal eigenvectors of $L^\star$ corresponding to its non-zero eigenvalues. Hence the estimators for both r and A may be obtained by performing an eigenanalysis for the sample version of $L^\star$, which can be easily obtained by replacing $\Sigma_y(k)$ by their sample counterparts.

Taking the sum in the definition of $L^\star$ enables us to accumulate the information over different lags together, which is particularly helpful when the sample size is small. However the choice of k_0 is not sensitive for the estimation, as the equation $L^\star B = 0$ holds for any $k_0 \geq 1$. When k_0 increases, the added terms are all non-negative definite matrices. Hence the information from different lags will not cancel off from each other, which is further confirmed in our simulation study in Example 2 in section 4 below. Note

that the term with $k = 0$ should be excluded from the sum in $L^\star$, as $\Sigma_y(0) = A\Sigma_x(0)A' + \text{var}(\epsilon_t)$ and, therefore, $\Sigma_y(0)B \neq 0$.

2.2 Regularity conditions

We introduce some notation first. For $k \geq 0$, let $\Sigma_{x,\epsilon}(k) = \text{cov}(x_{t+k},\ \epsilon_t)$, and

$$\widetilde{\Sigma}_x(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(x_{t+k} - \bar{x})(x_t - \bar{x})', \quad \widetilde{\Sigma}_{x,\epsilon}(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(x_{t+k} - \bar{x})(\epsilon_t - \bar{\epsilon})',$$

where $\bar{x} = T^{-1}\sum_{t=1}^{T} x_t$, $\bar{\epsilon} = T^{-1}\sum_{t=1}^{T}\epsilon_t$. The autocovariance matrices Σ_ϵ, $\Sigma_{\epsilon,x}(k)$, and their sample versions are defined in a similar manner. Some regularity conditions are now in order. In the following and thereafter we denote $\|M\|_2$ the L_2 norm of the matrix or vector M. (If M is a matrix, it means the positive square root of the maximum eigenvalue of MM'.)

(A) No linear combination of the components of x_t is white noise.

(B) For $k = 0, 1, \cdots, k_0$, where $k_0 \geq 1$ is a positive integer, $\Sigma_x(k)$ is full-ranked.

(C) For $k \geq 0$, each element of $\Sigma_{x,\epsilon}(k)$ or Σ_ϵ remains bounded as T, N increase to infinity.

(D) The covariance matrix $\text{cov}(\epsilon_t, x_s) = 0$ for all $s \leq t$.

(E) The data series $\{y_t\}$ 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 $E(|y_t|^4) < \infty$ elementwisely.

Assumption (A) is natural, as all the white noise linear combinations of x_t should be absorbed into ϵ_t. This ensures that there exists at least one $k \geq 1$ for which $\Sigma_x(k)$ is full ranked. Assumption (B) strengthens this statement for all $1 \leq k \leq k_0$, which entails that the non-negative definite matrix L, defined in (4) below, has r positive eigenvalues. Assumption (D) relaxes independence assumption between $\{x_t\}$ and $\{\epsilon_t\}$ imposed 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.

2.3 Factor strength

Since only y_t is observable in model (1), how well we can recover the factor x_t from y_t depends on the factor 'strength' reflected by the coefficients in the factor loading matrix A. For example in case $A = 0$, y_t carries no information on x_t. We now 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) $A = (a_1 \cdots a_r)$ such that $\|a_i\|_2^2 \asymp N^{1-\delta}$, $\ i = 1, \cdots, r,\ \ 0 \leq \delta \leq 1$.

(G) For each $i = 1, \cdots, r$ and δ given in (F), $\ \min_{\theta_j, j\neq i}\|a_i - \sum_{j\neq i}\theta_j a_j\|_2^2 \asymp N^{1-\delta}$.

When $\delta = 0$ in assumption (F), the corresponding factors are called strong factors since it includes the case where each element of a_i is $O(1)$, implying that the factors are shared (strongly) by the majority of the N 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 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.

2.4 Estimation

To facilitate our estimation, we use the QR decomposition $A = QR$ to normalize the factor loading matrix, so that (1) becomes

$$y_t = QRx_t + \epsilon_t = Qf_t + \epsilon_t, \tag{2}$$

where $f_t = Rx_t$, and $Q'Q = I_r$. The pair (Q, f_t) in the above model can be replaced by $(QU, U'f_t)$ for any $r \times r$ orthogonal matrix U.

For $k \geq 1$, it follows from (2) that

$$\Sigma_y(k) = \text{cov}(y_{t+k}, y_t) = Q\Sigma_f(k)Q' + Q\Sigma_{f,\epsilon}(k), \tag{3}$$

where $\Sigma_f(k) = \text{cov}(f_{t+k}, f_t)$ and $\Sigma_{f,\epsilon}(k) = \text{cov}(f_{t+k}, \epsilon_t)$. For $k_0 \geq 1$ given in condition (B), define

$$L = \sum_{k=1}^{k_0} \Sigma_y(k)\Sigma_y(k)' = Q\Big\{ \sum_{k=1}^{k_0} (\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k))(\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k))' \Big\} Q'. \tag{4}$$

Obviously L is a $N \times N$ non-negative definite matrix. Now we are ready to specify the factor loading matrix Q to be used in our estimation. Apply the spectral decomposition to the positive-definite matrix sandwiched by Q and Q' on the right hand side of (4), i.e.

$$\sum_{k=1}^{k_0} (\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k))(\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k))' = UDU',$$

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

> the columns of the factor loading matrix Q are the r orthonormal eigenvectors of the matrix 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 Q specified above is defined as $\widehat{Q} = (\widehat{q}_1, \cdots, \widehat{q}_r)$, where $\widehat{q}_i$ is the eigenvector of $\widetilde{L}$ corresponding to its i-th largest eigenvalue, $\widehat{q}_1, \cdots, \widehat{q}_r$ are orthonormal, and

$$\widetilde{L} = \sum_{k=1}^{k_0} \widetilde{\Sigma}_y(k)\widetilde{\Sigma}_y(k)', \quad \widetilde{\Sigma}_y(k) = \frac{1}{T-k}\sum_{t=1}^{T-k}(y_{t+k} - \bar{y})(y_t - \bar{y})', \tag{5}$$

where $\bar{y} = T^{-1}\sum_{t=1}^{T} y_t$. Consequently, we estimate the factors and the residuals respectively by

$$\widehat{f}_t = \widehat{Q}' y_t, \qquad e_t = y_t - \widehat{Q}\widehat{f}_t = (I_N - \widehat{Q}\widehat{Q}')y_t. \tag{6}$$

With $\widehat{Q}$ and the estimated factor series $\{\widehat{f}_t\}$, we can make an h-step ahead forecast for the y_t-series using the formula $\widehat{y}_{T+h}^{(h)} = \widehat{Q}\widehat{f}_{T+h}^{(h)}$, where $\widehat{f}_{T+h}^{(h)}$ is an h-step ahead forecast for $\{f_t\}$ based on the estimated past values $\widehat{f}_1, \cdots, \widehat{f}_T$. It can be obtained, for example, by fitting a vector-autoregressive model to $\{\widehat{f}_1, \cdots, \widehat{f}_T\}$.

Due to the random fluctuation in a finite sample, all the eigenvalues of $\widetilde{L}$ may not be exactly 0. We use the ratio-based estimator proposed in Lam and Yao (2011) to determine r. Let $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_N$ be the eigenvalues of $\widetilde{L}$. The ratio-based estimator for r is defined as

$$\widehat{r} = \arg\min_{1 \leq j \leq R} \widehat{\lambda}_{j+1}/\widehat{\lambda}_j, \tag{7}$$

where $r \leq R < N$ is an integer. In practice we may take $R = N/3$ or $N/2$ for example. Both the theoretical and empirical properties of this method can be found in Lam and Yao (2011). There is a large body of literature on the determination of r under various settings. See, for example, Bai and Ng (2002, 2007), Hallin and Liska (2007), Pan and Yao (2008) and Bathia et al. (2010).

3 Asymptotic theory

In this section we present the rates of convergence for the estimators $\widehat{Q}$ for model (2), and also for the estimated factor $\widehat{Q}\widehat{f}_t$, when both T and N tend to infinity while r is fixed and known. Note that the factor decomposition (1) is practically useful only when $r \ll N$. It goes without saying explicitly that we may replace some $\widehat{q}_j$ by $-\widehat{q}_j$ in order to match the direction of q_j. Denote by $\|M\|_{\min}$ the positive square root of the minimum eigenvalue of MM' or $M'M$, whichever is a smaller matrix. For model (2), define

$$\kappa_{\min} = \min_{1 \leq k \leq k_0} \|\Sigma_{f,\epsilon}(k)\|_{\min}, \quad \kappa_{\max} = \max_{1 \leq k \leq k_0} \|\Sigma_{f,\epsilon}(k)\|_2.$$

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 L, defined in (4), be distinct. Then,*

(i) $\|\widehat{Q} - Q\|_2 = O_P(N^{\delta}T^{-1/2})$ *provided* $\kappa_{\max} = o(N^{1-\delta})$, $N^{\delta}T^{-1/2} = o(1)$, *and*

(ii) $\|\widehat{Q} - Q\|_2 = O_P(\kappa_{\min}^{-2}\kappa_{\max}NT^{-1/2})$ *provided* $N^{1-\delta} = o(\kappa_{\min})$, $\kappa_{\min}^{-2}\kappa_{\max}NT^{-1/2} = o(1)$.

When all the factors are strong (i.e. $\delta = 0$), Theorem 1(i) reduces to $\|\widehat{Q} - Q\|_2 = O_P(T^{-1/2})$ provided $\kappa_{\max}/N \to 0$. The standard root-T rate might look too good to be true, as the dimension N goes to infinity together with the sample size T. But this is the case when 'blessing of dimensionality' is at its clearest. The strong factors pool together the information from most, if not all, of the original N component series. When N increases, the curse of dimensionality is offset by the increase of the information from more component series. The condition $\kappa_{\max}/N \to 0$ is mild. It implies that the linear dependence between the factors and the white noise is not too strong to distort the information on the serial dependence of the factors.

When $\delta > 0$, the rate of convergence in Theorem 1(i) depends on N. It is also clear that the stronger the factors are, the faster the convergence rate is. The condition $\kappa_{\max} = o(N^{1-\delta})$ ensures that the first term in matrix $\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k)$ is the dominating part; see (4).

When $N^{1-\delta} = o(\kappa_{\min})$, representing the cases that there are strong cross-correlations between the factors and the white noise, the second term in $\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k)$ dominates, and the conclusion of Theorem 1(ii) applies. The convergence rate, however, is not necessarily slower in estimating Q. For instance, when $\kappa_{\max} \asymp N^{1-\delta/2} \asymp \kappa_{\min}$ (see Lemma 1 in section 6 below), $\|\widehat{Q} - Q\|_2 = O_P(N^{\delta/2}T^{-1/2})$. This convergence rate is even faster than the rate $N^{\delta}T^{-1/2}$. This is not surprising, as we assume that r is known and we estimate Q by extracting the information on the autocorrelation of the data, including the cross-autocorrelation between $\{f_t\}$ and $\{\epsilon_t\}$. See the definition of L in (4). However, this may create difficulties for estimating r; see the relevant asymptotic results in Lam and Yao (2011).

Remark 1. (i) The assumption that all the non-zero eigenvalues of 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 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 $\mathcal{M}(A)$; see Pan and Yao (2008).

(ii) Theorem 1 can be extended to the cases when the factor x_t in model (1) is non-stationary, provided that a generalized sample (auto)covariance matrix

$$T^{-\alpha}\sum_{t=1}^{T-k}(x_{t+k} - \bar{x})(x_t - \bar{x})'$$

converges weakly, where $\alpha > 1$ is a constant. This weak convergence has been established when, for example, $\{x_t\}$ is an integrated process of order 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 $N \asymp T$, 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 ϵ_t as follows. In particular, both assumptions (H) and (I) are fulfilled when ϵ_t are independent and $N(0, \sigma^2 I_N)$. See also Péché (2009).

(H) Let ϵ_{jt} denote the j-th component of ϵ_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 $1 \leq j \leq N$ 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* $T = O(N)$*, then*
(i) $\|\widehat{Q} - Q\|_2 = O_P(N^{\delta/2}T^{-1/2})$ *provided* $N^{\delta/2}T^{-1/2} = o(1)$, $\kappa_{\max} = o(N^{1-\delta})$, *and*
(ii) $\|\widehat{Q} - Q\|_2 = O_P(\kappa_{\min}^{-2}\kappa_{\max}N^{1-\delta/2}T^{-1/2})$ *provided* $\kappa_{\min}^{-2}\kappa_{\max}N^{1-\delta/2}T^{-1/2} = o(1)$, $N^{1-\delta} = o(\kappa_{\min})$.

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

Theorem 3 *If all the eigenvalues of* Σ_ϵ *are uniformly bounded from infinity (as* $N \to \infty$*), it holds that*

$$N^{-1/2}\|\widehat{Q}\widehat{f}_t - Ax_t\|_2 = N^{-1/2}\|\widehat{Q}\widehat{f}_t - Qf_t\|_2 = O_P(N^{-\delta/2}\|\widehat{Q} - Q\|_2 + N^{-1/2}). \tag{8}$$

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 $\|\widehat{Q} - Q\|_2 = O_P(T^{-1/2})$. Now it follows Theorem 3 that

$$N^{-1/2}\|\widehat{Q}\widehat{f}_t - Ax_t\|_2 = O_P(T^{-1/2} + N^{-1/2}). \tag{9}$$

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 $\|\widehat{Q} - Q\|_2 = O_P(N^{\delta/2}T^{-1/2})$. Plugging this into the right hand side of (8), we obtain (9).

4 Simulation

Example 1. We start with a simple one factor model $y_t = Ax_t + \epsilon_t$, where ϵ_{tj} are independent $N(0, 4)$ random variables, A is a $N \times 1$ vector with $2\cos(2\pi i/N)$ as its i-th element, the factor is defined as $x_t = 0.9x_{t-1} + \eta_t$, and η_t are independent $N(0, 4)$ random variables. Hence we have a strong factor for this model with $\delta = 0$. We set $T = 200, 500$ and $N = 20, 180, 400, 1000$. We set $k_0 = 5$ in (5). (The results with $k_0 = 1, 2, 3, 4$ are similar, and are not presented to save the space.) For each (T, N) combination, we generate from the model 50 samples and calculate the estimation errors. The results are listed in table 1 below. It indicates clearly that the estimation error in L_2 norm for $\widehat{Q}$ is independent of N, as shown in Theorem 1(i) with $\delta = 0$.

Example 2. We consider a model of the form (1), with moving average factors $x_t = (x_{1,t}, x_{2,t}, x_{3,t})'$ defined by

$$x_{1,t} = w_t, \quad x_{2,t} = w_{t-1}, \quad x_{3,t} = w_{t-2},$$

$\|\widehat{Q}-Q\|_2$	$T=200$	$T=500$
$N=20$	$22_{(5)}$	$14_{(3)}$
$N=180$	$23_{(4)}$	$14_{(2)}$
$N=400$	$22_{(4)}$	$14_{(2)}$
$N=1000$	$23_{(4)}$	$14_{(2)}$

Table 1: *Means and standard errors (in brackets) of $\|\widehat{Q}-Q\|_2$ for Example 1. The values presented are the true values multiplied by 1000.*

where $w_t = 0.2z_{t-1} + z_t$, and z_t are independent $N(0,1)$ random variables. Hence the true number of dynamic factors is $q = 1$ in the GDFM context, and the number of factors is $r = 3$ for our model. Each column of the factor loading matrix A has the first $N/2$ elements generated 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 consider two scenarios for noise ϵ_t. In Scenario I, $\epsilon_t \sim N(0, I)$. In Scenario II, ϵ_t are independent $N(0, \Sigma_\epsilon)$ random vectors, where the (i,j)-th element of Σ_ϵ is defined as

$$\sigma_{ij} = \frac{1}{2}\{(|i-j|+1)^{2H} - 2|i-j|^{2H} + (|i-j|-1)^{2H}\}, \tag{10}$$

and $H = 0.9$ is the Hurst parameter.

Setting $T = 100, 200$ and $N = 100, 200, 400$, we compare the performance of our estimators with the principal components method of Bai and Ng (2002), and both the one and two-sided GDFM (see Forni et al. (2000) and Forni et al. (2005)). We report the results with $k_0 = 1$ and $k_0 = 5$ in the definition of $\widetilde{L}$ in (5). The number of dynamic factors for both of the GDFM is determined by the method of Hallin and Liska (2007). For the principal components method, the number of factors is determined by the BIC-type criterion of Bai and Ng (2002), defined by

$$\widehat{r} = \arg\min_k \Big\{ \log\left(N^{-1}T^{-1}\sum_{j=1}^{N}\|\widehat{\epsilon}_j\|_2^2\right) + k\left(\frac{N+T}{NT}\right)\log\left(\frac{NT}{N+T}\right)\Big\}. \tag{11}$$

For our model, the number of factors is estimated by the ratio-based method of (7).

For each combination of (T, N), we replicate the simulation 100 times, and calculate the mean and the standard deviation of the root-mean-square error (RMSE):

$$\text{RMSE} = \left(\frac{\sum_{t=1}^{T}\|\widehat{Q}\widehat{f}_t - Qf_t\|_2^2}{NT}\right)^{1/2}.$$

For our method and the principal components method of Bai and Ng (2002), we also use $\widehat{y}_T^{(1)} = \widehat{Q}\widehat{f}_T^{(1)}$ to forecast the factor Qf_t, where $\widehat{f}_T^{(1)}$ is the one-step predictor for f_T derived from a fitted AR(4) model based on $\widehat{f}_1, \cdots, \widehat{f}_{T-1}$. They are then compared with the one-step ahead forecast for the one-sided GDFM. For all three methods, we calculate the mean and standard deviation of the forecast error (FE):

$$\text{FE} = N^{-1/2}\|\widehat{y}_T^{(1)} - y_T\|_2.$$

(I): $\epsilon_t \sim N(0, I)$	GDFM			Principal components		Our method		
	2 sided	1 sided				$k_0 = 1$		$k_0 = 5$
	RMSE	RMSE	FE	RMSE	FE	RMSE	FE	RMSE
$(T, N) = (100, 100)$	$371_{(12)}$	$269_{(8)}$	$123_{(26)}$	$267_{(7)}$	$124_{(27)}$	$268_{(7)}$	$124_{(27)}$	$268_{(7)}$
$(T, N) = (200, 100)$	$309_{(7)}$	$226_{(6)}$	$122_{(25)}$	$224_{(6)}$	$122_{(24)}$	$225_{(6)}$	$122_{(24)}$	$225_{(6)}$
$(T, N) = (200, 400)$	$284_{(5)}$	$167_{(3)}$	$121_{(22)}$	$166_{(3)}$	$121_{(22)}$	$167_{(3)}$	$121_{(22)}$	$167_{(3)}$
(II): $\epsilon_t \sim N(0, \Sigma_\epsilon)$								
$(T, N) = (100, 100)$	$762_{(98)}$	$735_{(172)}$	$137_{(36)}$	$508_{(182)}$	$134_{(40)}$	$509_{(178)}$	$134_{(40)}$	$506_{(183)}$
$(T, N) = (200, 100)$	$740_{(80)}$	$685_{(214)}$	$140_{(39)}$	$441_{(165)}$	$132_{(35)}$	$444_{(170)}$	$132_{(35)}$	$444_{(169)}$
$(T, N) = (200, 400)$	$531_{(60)}$	$297_{(109)}$	$128_{(34)}$	$222_{(51)}$	$126_{(35)}$	$222_{(51)}$	$126_{(34)}$	$223_{(52)}$

Table 2: *Means and standard deviations (in brackets) of estimation and forecast errors for the moving average model of example 2. True number of factors $q = 1, r = 3$ are used throughout. Upper table: $\epsilon_t \sim N(0, I)$. Lower table: $\epsilon_t \sim N(0, \Sigma_\epsilon)$. The* RMSE *and their respective standard deviations reported are actual values multiplied by 1000, while the* FE *and their respective standard deviations reported are actual values multiplied by 100.*

From table 2, it is clear that the two-sided GDFM performs worse than the one-sided one in general, and has the worst performance in RMSE among all methods. Under scenario (I) (the upper table), our method, the principal components one and the one-sided GDFM all perform very similarly in terms of RMSE and FE. Moreover, we have estimated the number of factors under scenario (I) using the three different methods described earlier. We obtain consistently that $\widehat{q} = 1$ for the number of dynamic factors for the GDFM, and $\widehat{r} = 3$ for the other two methods. Hence the results are identical to that in table 2 under scenario (I) when the number of factors is estimated.

Under scenario (II), however, the one-sided GDFM performs worse than our method in terms of RMSE. It is expected since the GDFM requires weak cross-correlations in the noise in their theories, while our method do not. The principal components method has similar performance to our method in both scenarios, showing that when the number of factors is given, it is less sensitive to strong cross-correlations in the noise than both the one and two-sided gdfm do.

It is also clear from the table that the results are almost identical for $k_0 = 1$ and $k_0 = 5$ for our method in terms of RMSE, and not shown here they are also almost identical to the results when $k_0 = 2, 3, 4$. Same goes for the FE, therefore it is not shown in the table. Hence our method is not sensitive to the choice of k_0 when RMSE and FE are concerned.

Table 3 reports the results under Scenario II when the number of factors is estimated. It is clear that it is overestimated consistently by all methods, although our ratio-based method (7) overestimates the true value by merely 1. This is not necessarily disadvantageous in terms of forecasting, as the results for all the methods actually slightly outperform the corresponding simulation results in table 2 in terms of FE (with the exception of one-sided GDFM when $(T, N) = (200, 400)$). Our method outperforms all others in terms of RMSE and FE. The two-sided GDFM now exhibits a better performance than the one-sided counterpart and the principal components method in terms of RMSE. Note that under the setting (10) the strong cross-

(II): $\epsilon_t \sim N(0, \Sigma_\epsilon)$	GDFM ($\widehat{q}=2, r=3$)			Principal components			Our method $k_0 = 1$		
	2 sided	1 sided							
	RMSE	RMSE	FE	$\widehat{r}$	RMSE	FE	$\widehat{r}$	RMSE	FE
$T=100, N=100$	$690_{(36)}$	$857_{(53)}$	$129_{(30)}$	$7_{(7)}$	$770_{(37)}$	$123_{(29)}$	$4_{(0)}$	$670_{(39)}$	$116_{(29)}$
$N=200$	$629_{(33)}$	$824_{(60)}$	$137_{(34)}$	$7_{(6)}$	$716_{(34)}$	$132_{(39)}$	$4_{(0)}$	$624_{(36)}$	$127_{(37)}$
$N=400$	$607_{(36)}$	$740_{(128)}$	$142_{(34)}$	$6_{(6)}$	$667_{(39)}$	$138_{(41)}$	$4_{(0)}$	$584_{(43)}$	$133_{(37)}$
$T=200, N=100$	$675_{(31)}$	$836_{(46)}$	$130_{(32)}$	$8_{(7)}$	$770_{(32)}$	$121_{(34)}$	$4_{(0)}$	$654_{(35)}$	$117_{(33)}$
$N=200$	$632_{(30)}$	$820_{(43)}$	$135_{(38)}$	$8_{(6)}$	$730_{(26)}$	$127_{(39)}$	$4_{(0)}$	$613_{(28)}$	$123_{(35)}$
$N=400$	$582_{(30)}$	$708_{(137)}$	$132_{(36)}$	$8_{(7)}$	$684_{(24)}$	$125_{(33)}$	$4_{(0)}$	$571_{(24)}$	$122_{(32)}$

Table 3: *Means and standard deviations (in brackets) of estimation and forecast errors for the moving average model of example 2 under scenario (II). Number of dynamic factors for* GDFM *is estimated to be* $\widehat{q}=2$ *throughout. For one-sided* GDFM *we used* $r=3$ *throughout. The standard deviations for the number of factors are actual values multiplied by 10. The* RMSE *and their respective standard deviations reported are actual values multiplied by 1000, while the* FE *and their respective standard deviations reported are actual values multiplied by 100.*

sectional dependence among the components of ϵ_t violates the assumptions of Forni et al. (2000), Forni et al. (2005), and Bai and Ng (2002). However it is permitted in our model.

5 A Real Data Example: Implied Volatility Surfaces

5.1 The data, and method of estimation

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 $i = 1, \cdots, 10$ 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 $j = 1, \cdots, 13$ respectively. We collect these implied volatilities as $W_t = \{W_t(u_i, v_i)\} \in \mathbb{R}^{10\times 13} = \mathbb{R}^{130}$ for $t = 0, 1, \cdots, 249$. 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-sections in the maturity or delta axis display the well documented volatility smile. It is a well documented stylized fact that implied volatilities are unit-root non-stationary (see, e.g. Fengler et al. (2007)). Therefore, we choose to work with the differences $y_t = W_t - W_{t-1} \in \mathbb{R}^{130}$, $t = 1, \cdots, 249$.

We perform the factor modeling on each of the 150 rolling windows each of length 100 days, defined from the i-th day to the $(i + 99)$-th day for $i = 1, \cdots, 150$. For each window, we applied the three methods: the method proposed in section 2.4, the principal components method of Bai and Ng (2002), and the one-sided GDFM of Forni et al. (2005). The BIC (11) is applied to estimate r for the principal

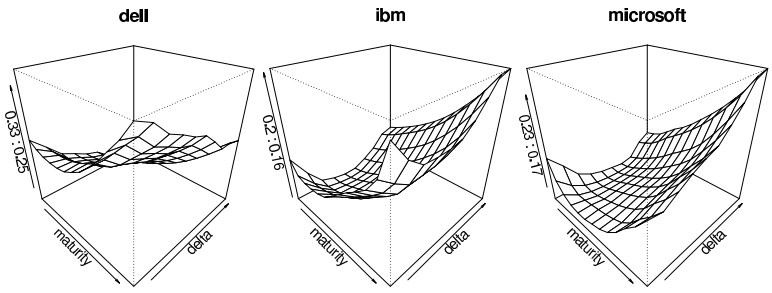

Figure 1: *Mean implied volatility surfaces over all trading days of year 2006.*

components method, and the ratio-based estimator $\widehat{r}$ in 7 is used for estimating r for our method.

For the i-th window, we use an autoregressive model to forecast the $(i+100)$-th value of the estimated factor series $x_{i+100}^{(1)}$, so as to obtain a one-step ahead forecast $y_{i+100}^{(1)} = \widehat{A}x_{i+100}^{(1)}$ for y_{i+100}. We also compare with the one-step ahead forecast using the one-sided GDFM. For comparison, we calculate the forecast error for the $(i+100)$-th day for each method, defined by

$$\text{FE} = N^{-1/2}\|y_{i+100}^{(1)} - y_{i+100}\|_2.$$

5.2 Estimation results

In forming the matrix $\widetilde{L}$ for each window, we take $k_0 = 5$ in (5) , though similar results (not reported here) are obtained for smaller k_0. This is consistent with our simulation results that estimation is not sensitive to the choice of k_0.

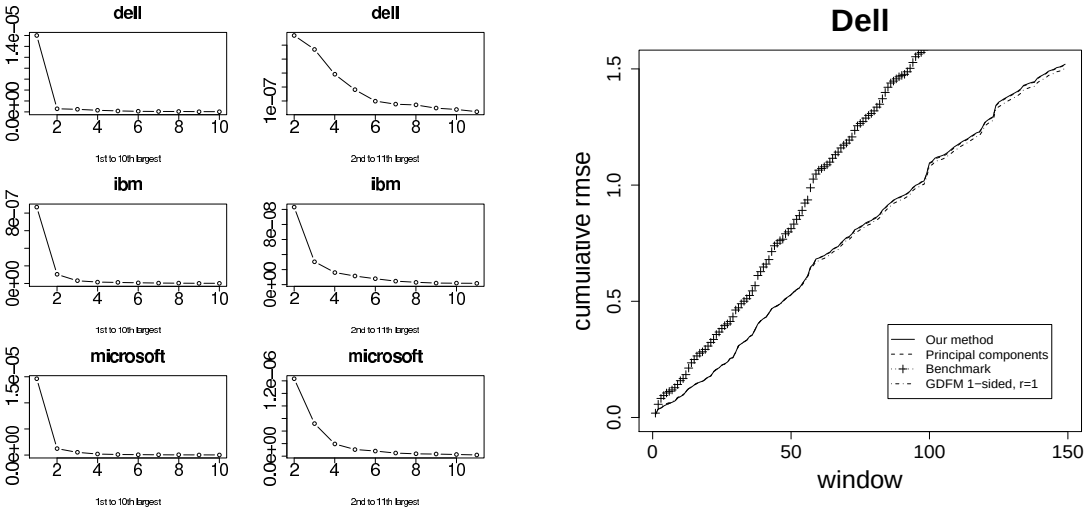

Figure 2: *Left: Averages of ordered eigenvalues of $\widetilde{L}$ over the 150 windows, with the left panel showing the ten largest, and the right panel showing the second to eleventh largest. Right: Plot of the cumulative* RMSE *for Dell.*

After performing the factor model estimation for a 100-days window (150 windows in total), we order the $N = 130$ eigenvalues obtained. For these 150 sets of ordered eigenvalues, we calculate the average of the largest eigenvalue across the windows, the average of the second largest and so on. The left panel of figure 2 displays these averages in descending order. The left hand side shows the the largest to the tenth largest for Dell, IBM and Microsoft for our method, whereas the right hand side shows the second to eleventh largest. We obtain similar results for the principal components method of Bai and Ng (2002) and thus the corresponding figure is not shown.

From this figure it is apparent that there is one eigenvalue that is much larger than the others for all three companies for each window. In fact, $\widehat{r} = 1$ is consistently found by both the BIC method in 11 and the ratio-based estimator in 7 for each window and for each company. Hence both methods choose a one factor model over the 150 windows.

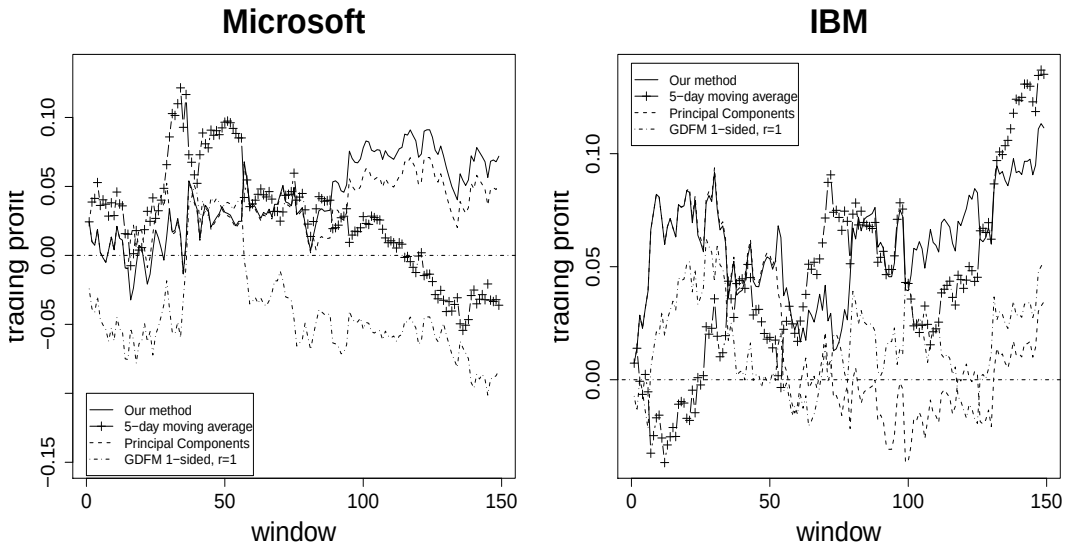

Figure 3: *Plot of accumulated return over time for the trading exercise in section 5.3. Left: Microsoft. Right: IBM. For our method* $k_0 = 5$ *is used.*

The right panel of figure 2 displays the cumulative FE over the 150 windows for each method for Dell. We choose a benchmark procedure, where we just treat today's value as the one-step ahead forecast. Our method, the principal components one and the one-sided GDFM with one factor all performed very similarly, while the benchmark procedure has a worse performance. This is similar for microsoft and IBM and so the respective figures are omitted. When the number of factors for the one sided GDFM is chosen to be 2 or 3, the performance is much worse (not shown in the figure), which lends further support that a one factor model is appropriate for all the 150 windows for each company.

5.3 A simple trading exercise

We use the one-step ahead forecast above to forecast the next day return of the three stocks. Not shown here, we have plotted the return against the estimated factor, for all three companies and for all three methods compared above. Simple linear regression suggests that the slope of the regression lines for the three methods are all 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 all three methods compared above. 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.

Our method outperforms the other two by a large margin for IBM, and slightly outperforms the principal components method for Microsoft, as shown in figure 3. For Dell (not shown) the principal components method outperforms slightly. However the one-sided GDFM always performs the worst for all three companies.

Appendix: Proofs

We introduce three technical lemmas first.

Lemma 1 *Under model (2) with assumptions (A) - (G) in sections 2.1 and 2.3, we have*

$$\|\Sigma_f(k)\|_2 \asymp N^{1-\delta} \asymp \|\Sigma_f(k)\|_{min}, \quad \|\Sigma_{f,\epsilon}(k)\|_2 = O(N^{1-\delta/2}).$$

Proof. Model (2) is an equivalent representation of model (1), where

$$y_t = Ax_t + \epsilon_t = Qf_t + \epsilon_t,$$

with $A = QR$ and $f_t = Rx_t$. With assumptions (F) and (G), the diagonal entries of R are all asymptotic to $N^{\frac{1-\delta}{2}}$ (which is the order of $\|a_i\|_2$), and the off-diagonal entries are of smaller order. Hence, as r is a constant, using

$$\|R\|_2 = \max_{\|u\|_2=1} \|Ru\|_2, \quad \|R\|_{\min} = \min_{\|u\|_2=1} \|Ru\|_2,$$

we can conclude that

$$\|R\|_2 \asymp N^{\frac{1-\delta}{2}} \asymp \|R\|_{\min}.$$

This, together with $\Sigma_f(k) = \text{cov}(f_{t+k}, f_t) = \text{cov}(Rx_{t+k}, Rx_t) = R\Sigma_x(k)R'$ for $k = 1, \cdots, k_0$, implies

$$N^{1-\delta} \asymp \|R\|^2_{\min} \times \|\Sigma_x(k)\|_{\min} \le \|\Sigma_f(k)\|_{\min} \le \|\Sigma_f(k)\|_2 \le \|R\|_2^2 \times \|\Sigma_x(k)\|_2 \asymp N^{1-\delta},$$

where we used assumption (B) to arrive at $\|\Sigma_x(k)\|_2 \asymp 1 \asymp \|\Sigma_x(k)\|_{\min}$, so that

$$\|\Sigma_f(k)\|_2 \asymp N^{1-\delta} \asymp \|\Sigma_f(k)\|_{\min}.$$

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

$$\begin{aligned}\|AB\|_{\min} &= \min_{u\neq 0} \frac{u'B'A'ABu}{\|u\|_2^2} \geq \min_{u\neq 0} \frac{(Bu)'A'A(Bu)}{\|Bu\|_2^2} \times \frac{\|Bu\|_2^2}{\|u\|_2^2} \\ &\geq \min_{w\neq 0} \frac{w'A'Aw}{\|w\|_2^2} \times \min_{u\neq 0} \frac{\|Bu\|_2^2}{\|u\|_2^2} = \|A\|_{\min} \cdot \|B\|_{\min}. \end{aligned} \tag{12}$$

Finally, using assumption (C) that $\Sigma_{x,\epsilon}(k) = O(1)$ elementwisely, and that it has $rN \asymp N$ elements, we have

$$\|\Sigma_{f,\epsilon}(k)\|_2 = \|R\Sigma_{x,\epsilon}(k)\|_2 \leq \|R\|_2 \times \|\Sigma_{x,\epsilon}(k)\|_F = O(N^{\frac{1-\delta}{2}}) \times O(N^{1/2}) = O(N^{1-\delta/2}),$$

where $\|M\|_F = \text{trace}(MM')$ denotes the Frobenius norm of the matrix M. □

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

$$\begin{aligned}\|\widetilde{\Sigma}_f(k) - \Sigma_f(k)\|_2 = O_P(N^{1-\delta}T^{-1/2}), \qquad \|\widetilde{\Sigma}_\epsilon(k) - \Sigma_\epsilon(k)\|_2 = O_P(NT^{-1/2}), \\ \|\widetilde{\Sigma}_{f,\epsilon}(k) - \Sigma_{f,\epsilon}(k)\|_2 = O_P(N^{1-\delta/2}T^{-1/2}) = \|\widetilde{\Sigma}_{\epsilon,f}(k) - \Sigma_{\epsilon,f}(k)\|_2,\end{aligned}$$

Moreover, $\|f_t\|_2^2 = O_P(N^{1-\delta})$ *for all integers* $t \geq 0$*.*

Proof. From (1) and (2), we have the relation $f_t = Rx_t$, where R is an upper triangular matrix with $\|R\|_2 \asymp N^{\frac{1-\delta}{2}} \asymp \|R\|_{\min}$ (see the proof of Lemma 1). Then we immediately have $\|f_t\|_2^2 \leq \|R\|_2^2 \times \|x_t\|_2^2 = O_P(N^{1-\delta}r) = O_P(N^{1-\delta})$.

Also, the covariance matrix and the sample covariance matrix for $\{f_t\}$ are respectively

$$\Sigma_f(k) = R\Sigma_x(k)R', \quad \widetilde{\Sigma}_f(k) = R\widetilde{\Sigma}_x(k)R'.$$

Hence

$$\|\widetilde{\Sigma}_f(k) - \Sigma_f(k)\|_2 \leq \|R\|_2^2\|\widetilde{\Sigma}_x(k) - \Sigma_x(k)\|_2 = O(N^{1-\delta}) \cdot O_P(T^{-1/2}) = O_P(N^{1-\delta}T^{-1/2}),$$

which is the rate specified in the lemma. We used the fact that the matrix $\widetilde{\Sigma}_x(k) - \Sigma_x(k)$ has r^2 elements, with elementwise rate of convergence being $O(T^{-1/2})$ which is implied by assumption (E) and that $\{\epsilon_t\}$ 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 A and $A+E$ are $n \times n$ symmetric matrices and that*

$$Q = [Q_1 \;\; Q_2] \quad (Q_1 \;\; is \;\; n \times r, \; Q_2 \;\; is \;\; n \times (n-r))$$

is an orthogonal matrix such that span(Q_1) *is an invariant subspace for A (that is, $A\cdot$ span$(Q_1) \subset$ span(A)). Partition the matrices $Q'AQ$ and $Q'EQ$ as follows:*

$$Q'AQ = \begin{pmatrix} D_1 & 0 \\ 0 & D_2 \end{pmatrix} \qquad Q'EQ = \begin{pmatrix} E_{11} & E_{21}' \\ E_{21} & E_{22} \end{pmatrix}.$$

If $\text{sep}(D_1, D_2) = \min_{\lambda \in \lambda(D_1),\, \mu \in \lambda(D_2)} |\lambda - \mu| > 0$*, where $\lambda(M)$ denotes the set of eigenvalues of the matrix M, and $\|E\|_2 \leq \text{sep}(D_1, D_2)/5$, then there exists a matrix $P \in \mathbb{R}^{(n-r)\times r}$ with*

$$\|P\|_2 \leq \frac{4}{\text{sep}(D_1, D_2)} \|E_{21}\|_2$$

such that the columns of $\widehat{Q}_1 = (Q_1 + Q_2 P)(I + P'P)^{-1/2}$ define an orthonormal basis for a subspace that is invariant for $A + E$.

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 $\sigma_1(M) = \|M\|_2$. We use $\lambda_j(M)$ to denote the j-th largest eigenvalue of M.

Proof of Theorem 1. Under model (2), we have shown in section 2.4 that we have $LQU = QUD$. Since U is an orthogonal matrix, we have

$$y_t = Qf_t + \epsilon_t = (QU)(U'f_t) + \epsilon_t,$$

so that we can replace QU with Q and $U'f_t$ with f_t in the model, thus making $LQ = QD$, where now D is diagonal with

$$D = \sum_{k=1}^{k_0} \{\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k)\}\{\Sigma_f(k)Q' + \Sigma_{f,\epsilon}(k)\}'.$$

If B is an orthogonal complement of Q, then $LB = 0$, and

$$\begin{pmatrix} Q' \\ B' \end{pmatrix} L (Q \;\; B) = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}, \tag{13}$$

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

To this end, define

$$W_f(k_0) = (\Sigma_f(1), \cdots, \Sigma_f(k_0)), \quad W_{f,\epsilon}(k_0) = (\Sigma_{f,\epsilon}(1), \cdots, \Sigma_{f,\epsilon}(k_0)),$$

so that we have $D = (W_f(k_0)(I_{k_0} \otimes Q') + W_{f,\epsilon}(k_0))(W_f(k_0)(I_{k_0} \otimes Q') + W_{f,\epsilon}(k_0))'$. Hence, assuming first that $\kappa_{\max} = o(N^{1-\delta})$, we have

$$\begin{aligned} \lambda_{\min}(D) &= \{\sigma_r(W_f(k_0)(I_{k_0} \otimes Q') + W_{f,\epsilon}(k_0))\}^2 \geq \{\sigma_r(W_f(k_0)(I_{k_0} \otimes Q')) - \sigma_1(W_{f,\epsilon}(k_0))\}^2 \\ &= \{\sigma_r(W_f(k_0)) - \sigma_1(W_{f,\epsilon}(k_0))\}^2 \asymp \sigma_r(W_f(k_0))^2 \asymp N^{2-2\delta}, \end{aligned}$$

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

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

Hence we have

$$\max(\kappa_{\min}^2, N^{2-2\delta}) = O(\lambda_{\min}(D)). \tag{14}$$

Next, we need to find $\|E_L\|_2$, where we define $E_L = \widetilde{L} - L$, with $\widetilde{L}$ defined in (5). Then it is easy to see that

$$\|E_L\|_2 \leq \sum_{k=1}^{k_0} \Big\{ \|\widetilde{\Sigma}_y(k) - \Sigma_y(k)\|_2^2 + 2\|\Sigma_y(k)\|_2 \times \|\widetilde{\Sigma}_y(k) - \Sigma_y(k)\|_2 \Big\}. \tag{15}$$

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

$$\|\Sigma_y(k)\|_2 = \|Q\Sigma_f(k)Q' + Q\Sigma_{f,\epsilon}(k)\|_2 \leq \|\Sigma_f(k)\|_2 + \|\Sigma_{f,\epsilon}(k)\|_2 = O(N^{1-\delta} + \kappa_{\max}). \tag{16}$$

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

$$\begin{aligned}&\|\widetilde{\Sigma}_y(k) - \Sigma_y(k)\|_2 \leq \|\widetilde{\Sigma}_f(k) - \Sigma_f(k)\|_2 + 2\|\widetilde{\Sigma}_{f,\epsilon}(k) - \Sigma_{f,\epsilon}(k)\|_2 + \|\widetilde{\Sigma}_\epsilon(k)\|_2 \\ =&O_P(N^{1-\delta}T^{-1/2} + N^{1-\delta/2}T^{-1/2} + \|\widetilde{\Sigma}_\epsilon(k)\|_2) = O_P(N^{1-\delta/2}T^{-1/2} + \|\widetilde{\Sigma}_\epsilon(k)\|_2).\end{aligned} \tag{17}$$

Without further assumptions on $\{\epsilon_t\}$, we have $\|\widetilde{\Sigma}_\epsilon(k)\|_2 \leq \|\widetilde{\Sigma}_\epsilon(k)\|_F = O_P(NT^{-1/2})$, which implies from (17) that

$$\|\widetilde{\Sigma}_y(k) - \Sigma_y(k)\|_2 = O_P(NT^{-1/2}). \tag{18}$$

With (16) and (18), we can easily see from (15) that

$$\|E_L\|_2 = O_P(N^{2-\delta}T^{-1/2} + \kappa_{\max}NT^{-1/2}). \tag{19}$$

Finally, no matter $\kappa_{\max} = o(N^{1-\delta})$ or $N^{1-\delta} = o(\kappa_{\min})$, we have from (19) and (14) that

$$\begin{aligned}\|E_L\|_2 &= O_P(N^{2-\delta}T^{-1/2} + \kappa_{\max}NT^{-1/2}) = o_P(\max(N^{2-2\delta}, \kappa_{\min}^2)) \\ &= O_P(\lambda_{\min}(D)) = O_P(\text{sep}(D, 0)),\end{aligned}$$

since we assumed $h_T = N^\delta T^{-1/2} = o(1)$ in the former case, or $\kappa_{\min}^{-2}\kappa_{\max}NT^{-1/2} = o(1)$ in the latter. Hence for sufficient large T, we have $\|E_L\|_2 \leq \text{sep}(D, 0)/5$. This allows us to apply Lemma 3 to conclude that there exists a matrix $P \in \mathbb{R}^{(p-r)\times r}$ such that

$$\|P\|_2 \leq \frac{4}{\text{sep}(D, 0)}\|(E_L)_{21}\|_2 \leq \frac{4}{\text{sep}(D, 0)}\|E_L\|_2,$$

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

$$\begin{aligned}\|\widehat{Q} - Q\|_2 &= \|(Q(I - (I + P'P)^{1/2}) + BP)(I + P'P)^{-1/2}\|_2 \\ &\leq \|I - (I + P'P)^{1/2}\|_2 + \|P\|_2 \leq 2\|P\|_2,\end{aligned}$$

and using (14) and (19),

$$\|P\|_2 = O_P\left(\frac{N^{2-\delta}T^{-1/2} + \kappa_{\max}NT^{-1/2}}{\max(\kappa_{\min}^2, N^{2-2\delta})}\right) = \begin{cases} O_P(N^{\delta}T^{-1/2}), & \text{if } \kappa_{\max} = o(N^{1-\delta}); \\ O_P(\kappa_{\min}^{-2}\kappa_{\max}NT^{-1/2}), & \text{if } N^{1-\delta} = o(\kappa_{\min}). \end{cases}$$

This completes the proof of the theorem. □

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

$$\|\widetilde{\Sigma}_{\epsilon}(k)\|_2 = O_P(NT^{-1}), \tag{20}$$

then (17) becomes

$$\|\widetilde{\Sigma}_y(k) - \Sigma_y(k)\|_2 = O_P(N^{1-\delta/2}T^{-1/2} + NT^{-1}) = O_P(N^{1-\delta/2}T^{-1/2}),$$

where we use the assumption $N^{\delta/2}T^{-1/2} = o(1)$. This rate is smaller than that in (18) by a factor of $N^{\delta/2}$, which carries to other parts of the proof of Theorem 1, so that the final rates are all smaller by a factor of $N^{\delta/2}$. Hence, it remains to show (20).

To this end, define 1_k the column vector of k ones, and

$$E_{r,s} = (\epsilon_r, \cdots, \epsilon_s) \text{ for } r \leq s.$$

Since the asymptotic behavior of the three sample means

$$\bar{\epsilon} = T^{-1}E_{1,T}1_T, \;\; (T-k)^{-1}E_{1,T-k}1_{T-k}, \;\; (T-k)^{-1}E_{k+1,T}1_{T-k}$$

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

$$\begin{aligned}\widetilde{\Sigma}_{\epsilon}(k) &= T^{-1}(E_{k+1,T} - (T-k)^{-1}E_{k+1,T}1_{T-k}1_{T-k}')(E_{1,T-k} - (T-k)^{-1}E_{1,T-k}1_{T-k}1_{T-k}'))' \\ &= T^{-1}E_{k+1,T}K_{T-k}E_{1,T-k}',\end{aligned}$$

where $K_j = I_j - j^{-1}1_j1_j'$. Then under conditions (H) and (I),

$$\begin{aligned}\|\widetilde{\Sigma}_{\epsilon}(k)\|_2 &\leq \|T^{-1/2}E_{k+1,T}\|_2 \times \|K_{T-k}\|_2 \times \|T^{-1/2}E_{1,T-k}\|_2 \\ &= \lambda_1^{1/2}(T^{-1}E_{k+1,T}'E_{k+1,T}) \times \lambda_1^{1/2}(T^{-1}E_{1,T-k}'E_{1,T-k}) \\ &= O_P((1 + (NT^{-1})^{1/2}) \times (1 + (NT^{-1})^{1/2})) = O_P(NT^{-1}),\end{aligned}$$

where the second last line follows from Theorem 1.3 of Péché (2009) for the covariance matrices $T^{-1}E_{k+1,T}'E_{k+1,T}$ and $T^{-1}E_{1,T-k}'E_{1,T-k}$, and the last line follows from the assumption $T = O(N)$. This completes the proof of the theorem. □

Proof of Theorem 3. Consider

$$\widehat{Q}\widehat{f}_t - Qf_t = \widehat{Q}\widehat{Q}'y_t - Qf_t = \widehat{Q}\widehat{Q}'Qf_t - Qf_t + \widehat{Q}\widehat{Q}'\epsilon_t = K_1 + K_2 + K_3,$$

where $K_1 = (\widehat{Q}\widehat{Q}' - QQ')Qf_t$, $K_2 = \widehat{Q}(\widehat{Q} - Q)'\epsilon_t$ and $K_3 = \widehat{Q}Q'\epsilon_t$. Using Lemma 2, we have

$$\|K_1\|_2 = O_P(\|\widehat{Q} - Q\|_2 \cdot \|f_t\|_2) = O_P(N^{\frac{1-\delta}{2}}\|\widehat{Q} - Q\|_2).$$

Also, since $\|\widehat{Q} - Q\|_2 = o_P(1)$ and $\|Q\|_2 = 1$, we have $\|K_2\|_2$ dominated by $\|K_3\|_2$ in probability. Hence we only need to consider K_3. Now consider for $Q = (q_1, \cdots, q_r)$, the random variable $q_j'\epsilon_t$, with

$$E(q_j'\epsilon_t) = 0, \quad \text{var}(q_j'\epsilon_t) = q_j'\Sigma_\epsilon q_j \leq \lambda_{\max}(\Sigma_\epsilon) < c < \infty$$

for $j = 1, \cdots, r$ by assumption, where c is a constant independent of T and r. Hence $q_j'\epsilon_t = O_P(1)$. We then have

$$\|K_3\|_2 = \|\widehat{Q}Q'\epsilon_t\|_2 \leq \|Q'\epsilon_t\|_2 = \sum_{j=1}^{r}(q_j'\epsilon_t)^2 = O_P(1).$$

Hence $N^{-1/2}\|\widehat{Q}\widehat{f}_t - Qf_t\|_2 = O_P(N^{-\delta/2}\|\widehat{Q} - Q\|_2 + N^{-1/2})$, 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. 38*, 3352–3386.

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.

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. Liska (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 (2011). Factor modelling for high dimensional time series: a dimension-reduction approach. Manuscript.

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

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.