Factor Modeling for High Dimensional Time Series

Neil Bathia

Department of Statistics

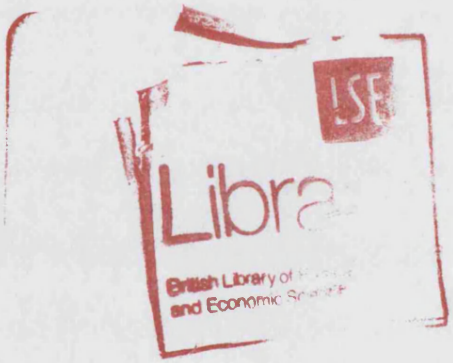
London School of Economics

A thesis submitted for the degree of

Doctor of Philosophy

January 2010

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346



To my loving parents...

Acknowledgements

Firstly, I would like to say that this thesis is not a result of my efforts alone. My advisor, Professor Qiwei Yao, provided me with immense guidance and support with seemingly infinite patience. A student could not wish for a more caring advisor. I will forever be in his intellectual debt.

I would also like to thank all of the staff and students in the LSE Statistics department for making the last three years one of the most memorable experiences of my life. In particular, I am indebted to all of the members of the of the Center for the Analysis of Time Series for their hospitality and countless stimulating discussions. Dr Flavio Ziegelmann is also due a special mention for his assistance with some of the computations in this thesis.

I am incredibly fortunate to have such amazing friends and family, without whose encouragement none of this would have been possible. I owe a great deal to my grandparents for the tremendous role they have played in my upbringing. In particular, my late grandfather Babulal strived to ensure that every dream I had was made a reality. I wish that he was here to celebrate this day with me. My uncle and aunt, Ajit and Nayna, have been like second parents to me and I hope that one day I will be able to repay them for their endless love and support. Also, my life would not be the same without my brothers and sisters, Raj, Krishma, Priyankaa and Nishal, who never fail to bring a smile to my face.

Last, but by no means least, I wish to thank my parents Ramesh and Vandana. They bore me, raised me, supported me, taught me and loved me. To them I dedicate this thesis.

Abstract

Chapter 1: Identifying the finite dimensionality of curve time series

The curve time series framework provides a convenient vehicle to model some types of nonstationary time series in a stationary framework. We propose a new method to identify the finite dimensionality of curve time series based on the autocorrelation between different curves. Based upon the duality relation between row and column subspaces of a data matrix, we show that the practical implementation of our methodology reduces to the eigenanalysis of a real matrix. Furthermore, the determination of the dimensionality is equivalent to indentifying the number of non-zero eigenvalues of this same matrix. For this purpose we propose a simple bootstrap test. Asymptotic properties of our methodology are investigated. The proposed methodology is illustrated with some simulation studies as well as an application to IBM intraday return densities.

Chapter 2: Methodology and convergence rates for factor modeling of multiple time series

An important task in modeling multiple time series is to obtain some form of dimension reduction. We tackle this problem using a factor model where the estimation of the factor loading space is constructed via eigenanalysis of a matrix which is a simple function of the sample autocovariance matrices. The number of factors is then equal to the number of "non-zero" eigenvalues of this matrix. We use the term "non-zero" loosely because in practice it is unlikely that there will be any eigenvalues which are exactly zero. However, our theoretical

results suggest that the sample eigenvalues whose population counterparts are zero are "super-consistent" (i.e. they converge to zero at a n rate) whereas the sample eigenvalues whose population counterparts are non-zero converge at an ordinary parametric rate of root-n. Here n denotes the sample size. This striking result is supported by simulation evidence and consequences for inference are discussed. In addition, we study the properties of the factor loading space under very general conditions (including possible non-stationarity) and a simple white noise test for empirically determining the number of non-zero eigenvalues is proposed and theoretically justified. We also provide an example of a heuristic threshold based estimator for the number of factors and prove that it yields a consistent estimator provided that the threshold is chosen to be of an appropriate order. Finally we conclude with an analysis of some implied volatility datasets.

Contents

Chapter 1

Identifying the finite dimensionality of curve time series

1.1 Introduction

A curve time series may consist of, for example, annual weather record charts, annual production charts or daily volatility curves (from morning to evening). In these examples the curves are segments of a single long time series. One advantage to view them as a curve series is to accommodate some nonstationary features (such as seasonal cycles or diurnal volatility patterns) into a stationary framework in a Hilbert space. There are other types of curve series that cannot be pieced together into a single long time series; for example, daily mean-variance efficient frontiers of portfolios, yield curves and intraday asset return distributions. The goal of this chapter is to identify the finite dimensionality of curve time series in the sense that the serial dependence across different curves is driven by a finite number of scalar components. Therefore the problem of modeling curve dynamics is reduced to that of modeling a finite-dimensional vector time series.

Throughout this chapter we assume that the observed curve time series, which we denote by $Y_1(\cdot), \cdots, Y_n(\cdot)$, are defined on a compact interval $\mathcal{I}$ and are subject

to errors in the sense that

$$Y_t(u) = X_t(u) + \varepsilon_t(u), \qquad u \in \mathcal{I}, \tag{1.1}$$

where $X_t(\cdot)$ is the curve process of interest. The existence of the noise term $\varepsilon(\cdot)$ reflects the fact that curves $X_t(\cdot)$ are seldom perfectly observed. They are often only recorded on discrete grids and are subject to both experimental error and numerical rounding. These noisy discrete data are smoothed to yield the 'observed' curves $Y_t(\cdot)$. Note that both $X_t(\cdot)$ and $\varepsilon_t(\cdot)$ in (1.1) are unobservable.

We assume that $\varepsilon_t(\cdot)$ is a white noise sequence in the sense that $E\{\varepsilon_t(u)\} = 0$ for all t and $\text{Cov}\{\varepsilon_t(u), \varepsilon_s(v)\} = 0$ for any $u, v \in \mathcal{I}$ as long as $t \neq s$. This is guaranteed since we may include all of the dynamic parts of $Y_t(\cdot)$ into $X_t(\cdot)$. Likewise, we may also assume that no parts of $X_t(\cdot)$ are white noise since these parts should be absorbed into $\varepsilon_t(\cdot)$. We also assume that

$$\int_{\mathcal{I}} E\{X_t(u)^2 + \varepsilon_t(u)^2\} du < \infty, \tag{1.2}$$

and both

$$\mu(u) \equiv E\{X_t(u)\}, \qquad M_k(u, v) \equiv \text{Cov}\{X_t(u), X_{t+k}(v)\} \tag{1.3}$$

do not depend on t. Furthermore, we assume that $X_t(\cdot)$ and $\varepsilon_{t+k}(\cdot)$ are uncorrelated for all integer k. Under condition (1.2), $X_t(\cdot)$ admits the Karhunen-Loéve expansion

$$X_t(u) - \mu(u) = \sum_{j=1}^{\infty} \xi_{tj} \varphi_j(u), \tag{1.4}$$

where $\xi_{tj} = \int_{\mathcal{I}} \{X_t(u) - \mu(u)\} \varphi_j(u) du$ are a sequence of scalar random variables with $E(\xi_{tj}) = 0$, $\text{Var}(\xi_{tj}) = \lambda_j$ and $\text{Cov}(\xi_{ti}, \xi_{tj}) = 0$ if $i \neq j$. We rank $\{\xi_{tj},\ j \geq 1\}$ such that λ_j is monotonically decreasing as j increases.

We say that $X_t(\cdot)$ is d-dimensional if $\lambda_d \neq 0$ and $\lambda_{d+1} = 0$, where $d \geq 1$ is a finite integer; see Hall & Vial (2006). The primary goal of this chapter is to identify d and to estimate the dynamic space $\mathcal{M}$ spanned by the (deterministic) eigenfunctions $\varphi_1(\cdot), \cdots, \varphi_d(\cdot)$.

Hall & Vial (2006) tackle this problem under the assumption that the curves $Y_1(\cdot), \cdots, Y_n(\cdot)$ are independent. Then the problem is insoluble in conventional

terms as one cannot separate $X_t(\cdot)$ from $\varepsilon_t(\cdot)$ in (1.1). This difficulty was resolved in Hall & Vial (2006) under a 'low noise' setting which assumes that the noise $\varepsilon_t(\cdot)$ goes to zero as the sample size goes to infinity. This condition is reasonable under an infill asymptotic scheme, i.e. when the observations on each curve are relatively dense. However, when dealing with a sparse design the condition is far from adequate and penalties are incurred in terms of the convergence rates of the resulting estimators; see Hall *et al.* (2006). Our approach is different and it does not require the low noise condition, since we identify d and $\mathcal{M}$ in terms of the serial dependence of the curves. Our method relies on a simple fact that $M_k(u,v) = \text{Cov}\{Y_t(u), Y_{t+k}(v)\}$ for any $k \neq 0$, which automatically filters out the noise $\varepsilon_t(\cdot)$; see (1.3). In this sense, the existence of dynamic dependence across different curves makes the problem tractable without the low noise argument.

Dimension reduction plays an important role in functional data analysis. The most frequently used method is functional principal component analysis in the form of computing the spectral decomposition of the empirical covariance operator. The literature in this field is vast and dates back to the early work of Besse & Ramsay (1986), Dauxois *et al.* (1982), Ramsay & Dalzell (1991) and Rice & Silverman (1991). Much of the work is described in Ramsay & Silverman (2005). However, despite the methodological advancements in functional data analysis with independent observations, the work on functional time series has been of a more theoretical nature; see e.g. Bosq (2000). The available inference methods focus mostly on nonparametric estimation of some characteristics of functional series (Part IV of Ferraty & Vieu (2006)). As far as we are aware, this work represents the first attempt on the dimension reduction based on dynamic dependence. Although we confine ourselves to square integrable curve series in this chapter, the methodology may be extended to a more general functional framework including, for example, a surface series which is particularly important for environmental studies; see Guillas & Lai (2008).

The rest of the chapter is organized as follows. Section 1.2 introduces the proposed method for identifying the dimensionality d and for estimating the dynamic space $\mathcal{M}$. Although an empirical analogue of the Karhunen-Loéve decomposition in (1.4) serves as a starting point of our approach, computationally our method boils down to an eigenanalysis of a finite matrix thus requiring no computing of

eigenfunctions in a functional space directly. The theoretical results of the estimation are presented in Section 1.3. Numerical illustration using both simulated and real datasets is provided in Section 1.4 and Section 1.5 closes with a brief discussion. We relegate all of the technical proofs to Section 1.6.

1.2 Methodology

1.2.1 Characterization of d and $\mathcal{M}$ via serial dependence

Let $\mathcal{L}_2(\mathcal{I})$ denote the Hilbert space consisting of all square integrable curves defined on $\mathcal{I}$ equipped with the inner product

$$\langle f, g \rangle = \int_{\mathcal{I}} f(u)g(u)du, \qquad f, g \in \mathcal{L}_2(\mathcal{I}). \tag{1.5}$$

Now M_k defined in (1.3) may be viewed as the kernel of a linear operator acting on $\mathcal{L}_2(\mathcal{I})$, i.e. for any $g \in \mathcal{L}_2(\mathcal{I})$, M_k maps $g(u)$ to $\breve{g}(u) \equiv \int_{\mathcal{I}} M_k(u,v)g(v)dv$. For notational economy, we will use M_k to denote both the kernel and the operator. Some relevant facts about operators acting on Hilbert spaces are listed in Appendix A.

For M_0 defined in (1.3), we have a spectral decomposition of the form

$$M_0(u,v) = \sum_{j=1}^{\infty} \lambda_j \varphi_j(u)\varphi_j(v), \qquad u, v \in \mathcal{I}, \tag{1.6}$$

where $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ are its eigenvalues and $\varphi_1, \varphi_2, \cdots$ are the corresponding orthonormal eigenfunctions ($\langle \varphi_i, \varphi_j \rangle = 1$ for $i = j$, and 0 otherwise), i.e.

$$\int_{\mathcal{I}} M_0(u,v)\varphi_j(v)dv = \lambda_j \varphi_j(u), \qquad j \geq 1.$$

Furthermore the random curves $X_t(\cdot)$ admit the representation (1.4). We assume in this chapter that $X_t(\cdot)$ is d-dimensional (i.e. $\lambda_{d+1} = 0$). Therefore

$$M_0(u,v) = \sum_{j=1}^{d} \lambda_j \varphi_j(u)\varphi_j(v), \qquad X_t(u) = \mu(u) + \sum_{j=1}^{d} \xi_{tj}\varphi_j(u). \tag{1.7}$$

It follows from (1.1) that

$$Y_t(u) = \mu(u) + \sum_{j=1}^{d} \xi_{tj}\varphi_j(u) + \varepsilon_t(u). \tag{1.8}$$

Hence the serial dependence of $Y_t(\cdot)$ is entirely determined by that of the d-vector process $\boldsymbol{\xi}_t \equiv (\xi_{t1}, \cdots, \xi_{td})'$ since $\varepsilon_t(\cdot)$ is white noise. By the virtue of the Karhunen-Loéve decomposition $E(\boldsymbol{\xi}_t) = 0$ and $\mathrm{Var}(\boldsymbol{\xi}_t) = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$.

Denote our estimator of M_k in (1.3) by

$$\widehat{M}_k(u,v) = \frac{1}{n-p}\sum_{j=1}^{n-p}\left\{Y_j(u) - \bar{Y}(u)\right\}\left\{Y_{j+k}(v) - \bar{Y}(v)\right\}, \quad 0 \le k \le p, \tag{1.9}$$

where $\bar{Y}(\cdot) = n^{-1}\sum_{1\le j\le n} Y_j(\cdot)$ and $p \ge 1$ is a prescribed integer. The reason for truncating the sums in (1.9) at $n-p$ as opposed to $n-k$ is to ensure a duality operation which simplifies the computation of the eigenfunctions; see Remark 2.2 at the end of Section 1.2.2.2. The conventional approach to estimate d and $\mathcal{M} = \mathrm{span}\{\varphi_1(\cdot), \cdots \varphi_d(\cdot)\}$ is to perform an eigenanalysis on $\widehat{M}_0$ and let $\widehat{d}$ be the number of non-zero eigenvalues and $\widehat{\mathcal{M}}$ be spanned by the $\widehat{d}$ corresponding eigenfunctions; see for example, Ramsay & Silverman (2005) and references therein. However this approach suffers from complications due to fact that $\widehat{M}_0$ is not a consistent estimator for M_0 since $\mathrm{Cov}\{Y_t(u), Y_t(v)\} = M_0(u,v) + \mathrm{Cov}\{\varepsilon_t(u), \varepsilon_t(v)\}$. Therefore $\widehat{M}_0$ needs to be adjusted to remove the part due to $\varepsilon_t(\cdot)$ before the eigenanalysis may be performed which is a non-trivial matter. An alternative is to let the variance of $\varepsilon_t(\cdot)$ decay to zero as the sample size n goes to infinity; see Hall & Vial (2006).

We adopt a different approach based on the fact that $\mathrm{Cov}\{Y_t(u), Y_{t+k}(v)\} = M_k(u,v)$ for any $k \neq 0$, which ensures that $\widehat{M}_k$ is a legitimate estimator for M_k; see (1.3) and (1.9).

Let $\boldsymbol{\Sigma}_k = E(\boldsymbol{\xi}_t\boldsymbol{\xi}'_{t+k}) \equiv (\sigma_{ij}^{(k)})$ be the autocovariance matrix of $\boldsymbol{\xi}_t$ at lag k. It is easy to see from (1.3) and (1.7) that $M_k(u,v) = \sum_{i,j=1}^{d}\sigma_{ij}^{(k)}\varphi_i(u)\varphi_j(v)$. Define a non-negative operator

$$N_k(u,v) = \int_{\mathcal{I}} M_k(u,z)M_k(v,z)dz = \sum_{i,j=1}^{d} w_{ij}^{(k)}\varphi_i(u)\varphi_j(u), \tag{1.10}$$

where $\mathbf{W}_k = (w_{ij}^{(k)}) = \mathbf{\Sigma}_k \mathbf{\Sigma}_k'$ is a non-negative definite matrix. Then it holds for any integer k that

$$\int_{\mathcal{I}} N_k(u,v)\zeta(v)dv = 0, \quad \text{for any } \zeta(\cdot) \in \mathcal{M}^{\perp}, \tag{1.11}$$

where $\mathcal{M}^{\perp}$ denotes the orthogonal complement of $\mathcal{M}$ in $\mathcal{L}_2(\mathcal{I})$. Note (1.11) also holds if we replace N_k by the operator

$$K(u,v) = \sum_{k=1}^{p} N_k(u,v) \tag{1.12}$$

which is also a non-negative operator on $\mathcal{L}_2(\mathcal{I})$.

Proposition 1.1 *Let the matrix $\mathbf{\Sigma}_k$ be full-ranked for some $k \geq 1$. Then the assertions below hold.*

(i) The operator N_k has exactly d non-zero eigenvalues, and $\mathcal{M}$ is the linear space spanned by the corresponding d eigenfunctions.

(ii) For $p \geq k$, (i) also holds for the operator K.

Remark 1.1 (i) The condition that $\text{rank}(\mathbf{\Sigma}_k) = d$ for some $k \geq 1$ is implied by the assumption that $X_t(\cdot)$ is d-dimensional. In the case where $\text{rank}(\mathbf{\Sigma}_k) < d$ for all k, the component with no serial correlations in $X_t(\cdot)$ should be absorbed into the white noise term $\varepsilon_t(\cdot)$; see similar arguments on modeling vector time series in Pena & Box (1987) and Pan & Yao (2008).

(ii) The introduction of the operator K in (1.12) is to pull together the information at different lags. Using a single N_k may lead to spurious choices of $\widehat{d}$; see Section 1.2.2.3

(iii) Note that $\int_{\mathcal{I}} K(u,v)\zeta(v)dv = 0$ if and only if $\int_{\mathcal{I}} N_k(u,v)\zeta(v)dv = 0$ for all $1 \leq k \leq p$. However, we cannot use M_k directly in defining K since it does not necessarily hold that $\int_{\mathcal{I}} \sum_{1\leq k\leq p} M_k(u,v)g(v) \neq 0$ for all $g \in \mathcal{M}$. This is due to the fact that M_k are not non-negative definite operators.

1.2.2 Estimation of d and $\mathcal{M}$

1.2.2.1 Estimators and fitted dynamic models

As we have stated above, M_k for $k \neq 0$ may be directly estimated from the observed curves Y_t; see (1.9). Hence a natural estimator of K may be defined as

$$\begin{aligned}\widehat{K}(u,v) &= \sum_{k=1}^{p}\int_{\mathcal{I}}\widehat{M}_k(u,z)\widehat{M}_k(v,z)dz \qquad (1.13)\\ &= \frac{1}{(n-p)^2}\sum_{t,s=1}^{n-p}\sum_{k=1}^{p}\{Y_t(u)-\bar{Y}(u)\}\{Y_s(v)-\bar{Y}(v)\}\langle Y_{t+k}-\bar{Y}, Y_{s+k}-\bar{Y}\rangle,\end{aligned}$$

see (1.12), (1.10), (1.9) and (1.5).

By Proposition 1.1 we define $\widehat{d}$ to be the number of non-zero eigenvalues of $\widehat{K}$ (see Section 1.2.2.3 below) and $\widehat{\mathcal{M}}$ to be the linear space spanned by the $\widehat{d}$ corresponding orthonormal eigenfunctions $\widehat{\psi}_1(\cdot), \cdots \widehat{\psi}_{\widehat{d}}(\cdot)$. This leads to the fitting

$$\widehat{Y}_t(u) = \bar{Y}(u) + \sum_{j=1}^{\widehat{d}}\widehat{\eta}_{tj}\widehat{\psi}_j(u), \quad u \in \mathcal{I}, \qquad (1.14)$$

where

$$\widehat{\eta}_{tj} = \int_{\mathcal{I}}\{Y_t(u)-\bar{Y}(u)\}\widehat{\psi}_j(u)du, \qquad j = 1, \cdots, \widehat{d}. \qquad (1.15)$$

See (1.8). Although $\widehat{\mathcal{M}} = \text{span}\{\widehat{\psi}_1(\cdot), \cdots \widehat{\psi}_{\widehat{d}}(\cdot)\}$ is a consistent estimator for $\mathcal{M} = \text{span}\{\varphi_1(\cdot), \cdots, \varphi_d(\cdot)\}$ (Theorem 1.2 in Section 1.3 below), $\widehat{\psi}_j$ are the estimators for the eigenfunctions of K defined in (1.12), which are different from the eigenfunctions φ_j of M_0 defined in (1.6). Therefore $\widehat{\boldsymbol{\eta}}_t \equiv (\widehat{\eta}_{t1}, \cdots, \widehat{\eta}_{t\widehat{d}})'$ is not an estimator for $\boldsymbol{\xi}_t$ used in (1.8).

Now in order to model the dynamic behavior of $Y_t(\cdot)$, we only need to model the $\widehat{d}$-dimensional vector process $\widehat{\boldsymbol{\eta}}_t$ which may be done using VARMA or any other multivariate time series models. See also Tiao & Tsay (1989) for applying linear transformations in order to obtain a more parsimonious model for $\widehat{\boldsymbol{\eta}}_t$.

1.2.2.2 Eigenanalysis

To perform an eigenanalysis in a Hilbert space is not a trivial matter. A popular pragmatic approach is to use an approximation via discretization, i.e. to evaluate

the observed curves at a fine grid and consequently replace them with the resulting vectors. This effectively transforms the problem to the eigenanalysis of a finite matrix. See, e.g. Section 8.4 of Ramsay & Silverman (2005). Below we also transform the problem into the eigenanalysis of a finite matrix but not via any approximations. Instead we make use of the well-known duality property that $\mathbf{AB}'$ and $\mathbf{B}'\mathbf{A}$ share the same non-zero eigenvalues for any matrices $\mathbf{A}$ and $\mathbf{B}$ of the same sizes. Furthermore if $\boldsymbol{\gamma}$ is an eigenvector of $\mathbf{B}'\mathbf{A}$, $\mathbf{A}\boldsymbol{\gamma}$ is an eigenvector of $\mathbf{AB}'$ with the same eigenvalue. In fact this duality also holds for operators acting on function spaces. This scheme was adopted in Kneip & Utikal (2001) and Benko *et al.* (2009) but the adaptation to our setting is more involved so we provide a detailed exposition.

We present a heuristic argument first. To view the operator $\widehat{K}(\cdot,\ \cdot)$ defined in (1.13) in the form of $\mathbf{AB}'$, let us denote the curve $Y_t(\cdot) - \bar{Y}(\cdot)$ as an $\infty \times 1$ vector $\mathbf{Y}_t$ with $\mathbf{Y}_t'\mathbf{Y}_s = \langle Y_t - \bar{Y}, Y_s - \bar{Y}\rangle$; see (1.5). Put $\mathcal{Y}_k = (\mathbf{Y}_{1+k}, \cdots, \mathbf{Y}_{n-p+k})$. Then $\widehat{K}(\cdot,\ \cdot)$ may be represented as an $\infty \times \infty$ matrix

$$\widehat{\mathbf{K}} = \frac{1}{(n-p)^2}\mathcal{Y}_0 \sum_{k=1}^{p} \mathcal{Y}_k'\mathcal{Y}_k\mathcal{Y}_0'.$$

Applying the duality stated the above with $\mathbf{A} = \mathcal{Y}_0$ and $\mathbf{B}' = \sum_{1\le k\le p} \mathcal{Y}_k'\mathcal{Y}_k\mathcal{Y}_0'$, $\widehat{\mathbf{K}}$ shares the same non-zero eigenvalues with the $(n-p)\times(n-p)$ matrix

$$\mathbf{K}^* = \frac{1}{(n-p)^2} \sum_{k=1}^{p} \mathcal{Y}_k'\mathcal{Y}_k\mathcal{Y}_0'\mathcal{Y}_0, \tag{1.16}$$

where the (t,s)-th element of $\mathcal{Y}_k'\mathcal{Y}_k$ is $\mathbf{Y}_{t+k}'\mathbf{Y}_{s+k} = \langle Y_{t+k} - \bar{Y}, Y_{s+k} - \bar{Y}\rangle$ and $k = 1, \cdots, p$. Furthermore, let $\boldsymbol{\gamma}_j = (\gamma_{1j}, \cdots, \gamma_{n-p,j})'$, $j = 1, \cdots, \widehat{d}$, be the eigenvectors of $\mathbf{K}^*$ corresponding to the $\widehat{d}$ largest eigenvalues. Then

$$\sum_{t=1}^{n-p} \gamma_{tj}\{Y_t(\cdot) - \bar{Y}(\cdot)\}, \quad j = 1, \cdots, \widehat{d} \tag{1.17}$$

are the $\widehat{d}$ eigenfunctions of $\widehat{K}(\cdot,\ \cdot)$. In practice, it is likely that the functions in (1.17) may not be exactly orthogonal. Thus the orthonormal eigenfunctions $\widehat{\psi}_1(\cdot), \cdots, \widehat{\psi}_{\widehat{d}}(\cdot)$ used in (1.14) may be obtained by applying a Gram-Schmidt algorithm to (1.17).

This heuristic argument is justified by the result below. The formal proof is relegated to Section 1.6.

Proposition 1.2 *The operator $\widehat{K}(\cdot,\cdot)$ shares the same non-zero eigenvalues with matrix $\mathbf{K}^*$ defined in (1.16) with the corresponding eigenfunctions given in (1.17).*

Remark 1.2 The truncation of the sums in (1.9) at $(n-p)$ for different k is necessary to ensure the applicability of the above duality operation. If we truncated the sum for $\widehat{M}_k$ at $(n-k)$ instead, $\mathcal{Y}_k'\mathcal{Y}_k$ would be of different sizes for different k, and $\mathbf{K}^*$ in (1.16) would not be well-defined.

1.2.2.3 Determination of d via statistical tests

Although the number of nonzero eigenvalues of the operator $K(\cdot,\ \cdot)$ defined in (1.12) is d (see Proposition 1.1), the number of nonzero eigenvalues of its estimator $\widehat{K}(\cdot,\ \cdot)$ defined in (1.13) may be much greater than d due to random fluctuation in the sample. One empirical approach is to take $\widehat{d}$ to be the number of 'large' eigenvalues of $\widehat{K}$ in the sense that the $(\widehat{d}+1)$-th largest eigenvalue drops significantly; see Figure 1.1. A more formal method of determining the value of d is given in the bootstrap test below.

Let $\theta_1 \geq \theta_2 \geq \cdots \geq 0$ be the eigenvalues of K. Suppose we are interested in testing the null hypothesis

$$H_0:\ \theta_{d_0+1} = 0,$$

where d_0 is a known integer, obtained, for example, by visual observation of the estimated eigenvalues $\widehat{\theta}_1 \geq \widehat{\theta}_2 \geq \cdots \geq 0$ of $\widehat{K}$. Hence we reject H_0 if $\widehat{\theta}_{d_0+1} > l_\alpha$, where l_α is the critical value at the $\alpha \in (0,1)$ significance level. To evaluate the critical value l_α, we propose the following bootstrap procedure.

1. Let $\widehat{Y}_t(\cdot)$ be defined as in (1.14) with $\widehat{d} = d_0$. Let $\widehat{\varepsilon}_t(\cdot) = Y_t(\cdot) - \widehat{Y}_t(\cdot)$.

2. Generate a bootstrap sample from the model

$$Y_t^*(\cdot) = \widehat{Y}_t(\cdot) + \varepsilon_t^*(\cdot),$$

where ε_t^* are drawn independently (with replacement) from $\{\widehat{\varepsilon}_1, \cdots, \widehat{\varepsilon}_n\}$.

3. Form the operator K^* in the same manner as $\widehat{K}$ with $\{Y_t\}$ replaced by $\{Y_t^*\}$ and compute the (d_0+1)-th largest eigenvalue $\theta^*_{d_0+1}$ of K^*.

Then the conditional distribution of $\theta^*_{d_0+1}$, given the observations $\{Y_1, \cdots, Y_n\}$, is taken as the distribution of $\widehat{\theta}_{d_0+1}$ under H_0. In practical implementation, we repeat steps 1 and 2 above B times for some large integer B, and we reject H_0 if the event that $\theta^*_{d_0+1} > \widehat{\theta}_{d_0+1}$ occurs no more than $[\alpha B]$ times. The simulation results reported in Section 1.4.1 below indicate that the above bootstrap method works well. A full theoretical justification of the bootstrap is beyond the scope of this current work.

1.3 Theoretical properties

We introduce some regularity conditions for model (1.8) first.

C1. $\{Y_t(\cdot)\}$ is strictly stationary and satisfies the condition $E[\{\int_{\mathcal{I}} Y_t(u)^2 du\}^{2+\delta}] < \infty$ for some $\delta > 0$.

C2. For p given in (1.12), the sequence $\{Y_t(\cdot), \ldots, Y_{t+p}(\cdot)\}$ is strongly mixing in the sense that $\alpha(m) \longrightarrow 0$ as $m \longrightarrow \infty$, where

$$\alpha(m) = \sup_{l \geq 1} \sup_{U \in \mathcal{F}^l_{-\infty},\, V \in \mathcal{F}^\infty_{m+l}} |P(U \cap V) - P(U)P(V)|,$$

and $\mathcal{F}_i^j = \sigma\{Y_i(\cdot), \ldots, Y_{j+p}(\cdot)\}$. In addition, it holds that $\sum_{j=1}^{\infty} \alpha(j)^{\delta/(2+\delta)} < \infty$ with δ given in C1 above.

C3. $\mathrm{Cov}\{X_s(u), \varepsilon_t(v)\} = 0$ for all s, t and $u, v \in \mathcal{I}$.

C4. The d non-zero eigenvalues of K defined in (1.12) satisfy $\theta_1 > \cdots > \theta_d > 0$, i.e. they are all unique.

Remark 1.3 (i) Strong mixing is one of the weakest mixing conditions for weakly dependent random variables but since probability measure on function spaces are complex objects (see Delaigle & Hall (2009) and Hall & Heckman (2002)) one may wonder about it's validity in this context. However, in light of the

Karhunen-Loeve expansion of $X_t(\cdot)$ in (1.4) (with an analogous representation readily available for $\varepsilon_t(\cdot)$), we may think of the mixing condition on the random function $Y_t(\cdot)$ as being equivalent to placing mixing conditions on it's (scalar) Karhunen-Loeve coefficients.

(ii) We note that conditions C1 and C2 are in fact stronger than required (see Merlevede *et al.* (1997)). However, the necessary and sufficient conditions in this context would be messy and perhaps unintuitive so we have stated only a set of sufficient conditions here.

(iii) The condition that all eigenvalues of K are different ensures that its eigenfunctions are identifiable. It is still possible to obtain a form of consistency of the empirical eigenfunctions without this assumption but the proofs become a lot more technical and little further insight is to be gained.

We now solidify some notation before presenting the asymptotic results. Denote by (θ_j, ψ_j) and $(\widehat{\theta}_j, \widehat{\psi}_j)$ the (eigenvalue, eigenfunction) pairs of K and $\widehat{K}$ respectively (see (1.12) and (1.13)). We will always arrange the eigenvalues in descending order, i.e. $\theta_j \geq \theta_{j+1}$ and $\widehat{\theta}_j \geq \widehat{\theta}_{j+1}$. As the eigenfunction of K and $\widehat{K}$ are only unique up to sign changes, in the sequel it will go without saying that the right versions are used. Finally, for any operator L acting on $\mathcal{L}_2(\mathcal{I})$, denote by $\|L\|_{\mathcal{N}}$ the sum of the absolute eigenvalues of L; see also Appendix A. All of the results in this section require that p is a fixed and finite integer.

Theorem 1.1 *Let conditions C1 - C3 hold. Then as* $n \to \infty$, *it holds that* $\|\widehat{K} - K\|_{\mathcal{N}} = O_P(n^{-1/2})$ *and* $\sup_{j \geq 1} |\widehat{\theta}_j - \theta_j| = O_P(n^{-1/2})$. *In addition, if C4 also holds, then*

$$\left(\int_{\mathcal{I}} \left\{ \widehat{\psi}_j(u) - \psi_j(u) \right\}^2 du \right)^{1/2} = O_P(n^{-1/2}), \qquad 1 \leq j \leq d.$$

Note that the results of Theorem 1 hold even if $d = \infty$, i.e. when the dynamic space $\mathcal{M}$ is infinite dimensional.

With d known, put $\widetilde{\mathcal{M}} = \text{span}\{\widehat{\psi}_1(\cdot), \cdots, \widehat{\psi}_d(\cdot)\}$ where $\widehat{\psi}_1(\cdot), \cdots, \widehat{\psi}_d(\cdot)$ are the eigenfunctions of $\widehat{K}$ corresponding to the d largest eigenvalues. (See also Section 1.2.2.1.) To measure the error in estimating $\mathcal{M}$ by $\widetilde{\mathcal{M}}$, we introduce a measure for the discrepancy of any two d-dimensional subspaces $\mathcal{N}_1$ and $\mathcal{N}_2$ in $\mathcal{L}_2(\mathcal{I})$. Let

$\{\zeta_{i1}(\cdot), \cdots, \zeta_{id}(\cdot)\}$ be an orthonormal basis of $\mathcal{N}_i$, $i = 1, 2$. Then the projection of ζ_{1k} onto $\mathcal{N}_2$ may be expressed as

$$\sum_{j=1}^{d} \langle \zeta_{2j}, \zeta_{1k} \rangle \zeta_{2j}(u).$$

Its squared norm is $\sum_{j=1}^{d} (\langle \zeta_{2j}, \zeta_{1k} \rangle)^2 \leq 1$. The discrepancy measure is defined as

$$D(\mathcal{N}_1, \mathcal{N}_2) = \sqrt{1 - \frac{1}{d} \sum_{j,k=1}^{d} (\langle \zeta_{2j}, \zeta_{1k} \rangle)^2}. \tag{1.18}$$

It is clear that this is a symmetric measure between 0 and 1. It is independent of the choice of the orthonormal bases used in the definition and it equals 0 if and only if $\mathcal{N}_1 = \mathcal{N}_2$. Let $\mathcal{Z}$ be the set consisting of all d-dimensional subspaces of $\mathcal{L}_2(\mathcal{I})$. Then $(\mathcal{Z}, D)$ forms a metric space in the sense that D is a well defined distance measure on $\mathcal{Z}$ (see Lemma 1.4 in Section 1.6 below).

We are now in a position to state consistency results about $\widetilde{\mathcal{M}}$. We also consider the asymptotic properties of $\widehat{\mathcal{M}} = \text{span}\{\widehat{\psi}_1, \ldots, \widehat{\psi}_{\widehat{d}}\}$ where $\widehat{d}$ is some estimator of d. Since $\widehat{d}$ may differ from d, we use a modified version of D (D' given in (1.19)) to measure the distance between $\widehat{\mathcal{M}}$ and $\mathcal{M}$.

Theorem 1.2 *Let conditions C1 - C4 hold. In addition, suppose that the conditions of Proposition 1.1 are satisfied and d is fixed, finite and known. (i) Then as* $n \longrightarrow \infty$, *it holds that* $D(\widetilde{\mathcal{M}}, \mathcal{M}) = O_p(n^{-1/2})$. *(ii) In addition if* $\widehat{d} \xrightarrow{p} d$, *then it holds that* $|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| = o_p(n^{-1/2})$.

A remarkable feature of Theorem 1.2 is the adaptivity property, i.e. we do not suffer any penalty in our estimation of $\mathcal{M}$ when d is unknown provided that it can be estimated consistently. Thus it is reasonable to assume that d is known when considering the asymptotic properties of our estimation of $\mathcal{M}$. As remarked earlier, it is beyond the scope of this current work to provide a full theoretical justification of the bootstrap method for determining d given in Section 1.2.2.3.

1.4 Numerical properties

1.4.1 Simulations

We illustrate the proposed method simulating data generated from model (1.1) with

$$X_t(u) = \sum_{i=1}^{d} \xi_{ti}\varphi_i(u), \qquad \varepsilon_t(u) = \sum_{j=1}^{100} \frac{Z_{tj}}{2^{j-1}}\zeta_j(u), \qquad u \in [0,1],$$

where $\{\xi_{ti},\ t \geq 1\}$ is a linear AR(1) process with the coefficient $(-1)^i(0.9-0.5i/d)$, the innovations Z_{tj} are independent $N(0,1)$ variables and

$$\varphi_i(u) = \sqrt{2}\cos(\pi i u), \qquad \zeta_j(u) = \sqrt{2}\sin(\pi j u).$$

We set the sample size to be $n = 100$, 300 or 600, and the dimension parameter $d = 10$, 20 or 40. For each setting we repeated the simulation 200 times. We used $p = 5$ in defining the operator $\widehat{K}$ in (1.13) and for each of the 200 simulations, we replicated the bootstrap method 200 times. Note that for this simulation experiment, the conditions of Proposition 1.1 are satisfied even if $p = 1$ so by taking $p > 1$ in constructing $\widehat{K}$ we are accumulating estimation error. However, when analyzing a real dataset we would not know whether or not such a small value of p would be sufficient so taking $p = 1$ may lead to a spurious choice of $\widehat{d}$. Thus our aim in taking a value of p that is larger than necessary is to demonstrate that the methodology still performs well even when this is the case.

The average of the ordered eigenvalues of $\widehat{K}$ obtained from the 200 replications are plotted in Figure 1.1. For better illustration, we only plotted eleven eigenvalues (i.e. five on the each side of the d-th largest eigenvalue). It is clear that drop from the d-th largest eigenvalue to the $(d+1)$-st is very pronounced. We applied the bootstrap method to test the hypothesis that the d-th or the $(d+1)$-st largest eigenvalue of K (θ_d and θ_{d+1} respectively) are 0. The results are summarized in Table 1.1 and Figure 1.2. The bootstrap test could not reject the true null hypothesis $\theta_{d+1} = 0$. In fact among the 200 replications for all the settings, the P-value was invariably greater than 10%; see panel (b) of Figure 1.2. The false null hypothesis $\theta_d = 0$ was routinely rejected by the bootstrap when $n = 600$ or 300; see panel (a) of Figure 1.2 and Table 1.1. However the test does not work when the sample size is as small as 100.

Null hypothesis	$\theta_d = 0$		
d	10	20	40
$n = 100$	0.35 (0.30)	0.33 (0.24)	0.36 (0.28)
$n = 300$	0.02 (0.09)	0.14 (0.20)	0.02 (0.04)
$n = 600$	0.00 (0.00)	0.00 (0.03)	0.05 (0.12)
Null hypothesis	$\theta_{d+1} = 0$		
d	10	20	40
$n = 100$	0.95 (0.12)	0.99 (0.05)	0.82 (0.16)
$n = 300$	0.96 (0.10)	1.00 (0.01)	1.00 (0.00)
$n = 600$	0.97 (0.09)	1.00 (0.00)	1.00 (0.00)

Table 1.1: *The means and standard deviations (in parentheses) of the P-values of the bootstrap test.*

To measure the accuracy of our estimation of the factor loading space $\mathcal{M}$, we need to modify the metric D defined in (1.18) as $\widehat{d}$ may be different from d. Let $\mathcal{N}_1$, $\mathcal{N}_2$ be two subspaces of $\mathcal{L}_2(\mathcal{I})$ with dimension d_1 and d_2 respectively. Let $\{\zeta_{i1}, \cdots, \zeta_{id_i}\}$ be an orthonormal basis of $\mathcal{N}_i$, $i = 1, 2$. The discrepancy measure between the two subspaces is defined as

$$D'(\mathcal{N}_1, \mathcal{N}_2) = \sqrt{1 - \frac{1}{\max(d_1,\, d_2)} \sum_{k=1}^{d_1} \sum_{j=1}^{d_2} (\langle \zeta_{2j}, \zeta_{1k} \rangle)^2}. \tag{1.19}$$

It can be shown that $D'(\mathcal{N}_1, \mathcal{N}_2) \in [0,\, 1]$. It equals 0 if and only if $\mathcal{N}_1 = \mathcal{N}_2$, and 1 if and only if $\mathcal{N}_1 \cap \mathcal{N}_2 = \emptyset$. Obviously $D'(\mathcal{N}_1, \mathcal{N}_2) = D(\mathcal{N}_1, \mathcal{N}_2)$ when $d_1 = d_2 = d$. We set $\widehat{d} = \min\{k : \theta_k = 0\} - 1$ where we use the bootstrap test at the 5% level to determine whether or not $\theta_k = 0$. We computed the $D'(\widehat{\mathcal{M}}, \mathcal{M})$ in the 200 replications for each setting. Figure 1.3 presents the boxplots of those D' values. It is noticeable that the D' measure decreases as the sample size n increases. It is also interesting to note that the accuracy of the estimation is independent of the dimension d.

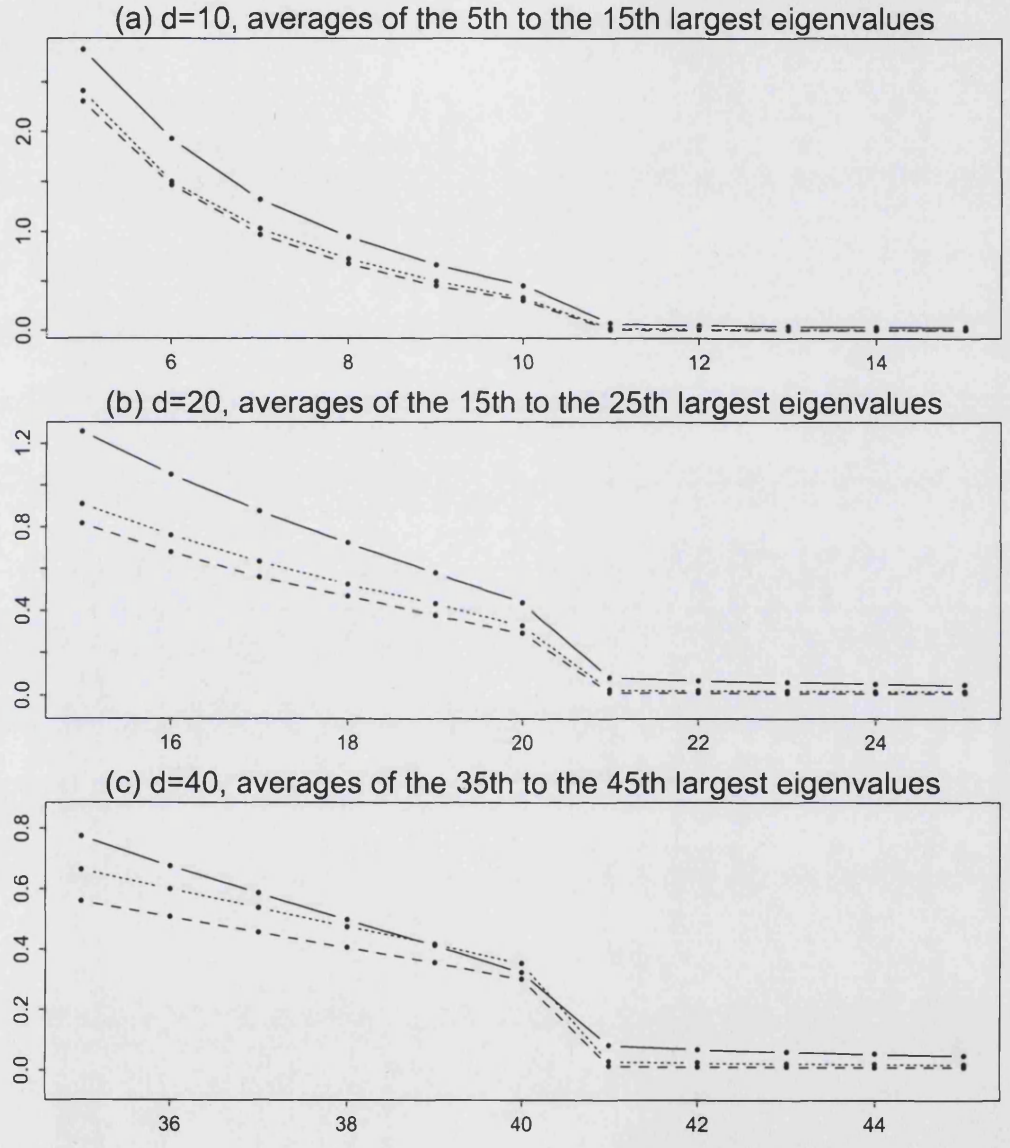


Figure 1.1: *The average estimated eigenvalues of the operator $\widehat{\theta}_j$ over 200 replications with sample sizes $n = 100$ (solid lines), 300 (dotted lines) and 600 (dashed lines). Recall that by Proposition 1.1 $\theta_j = 0$ for all $j \geq d + 1$.*

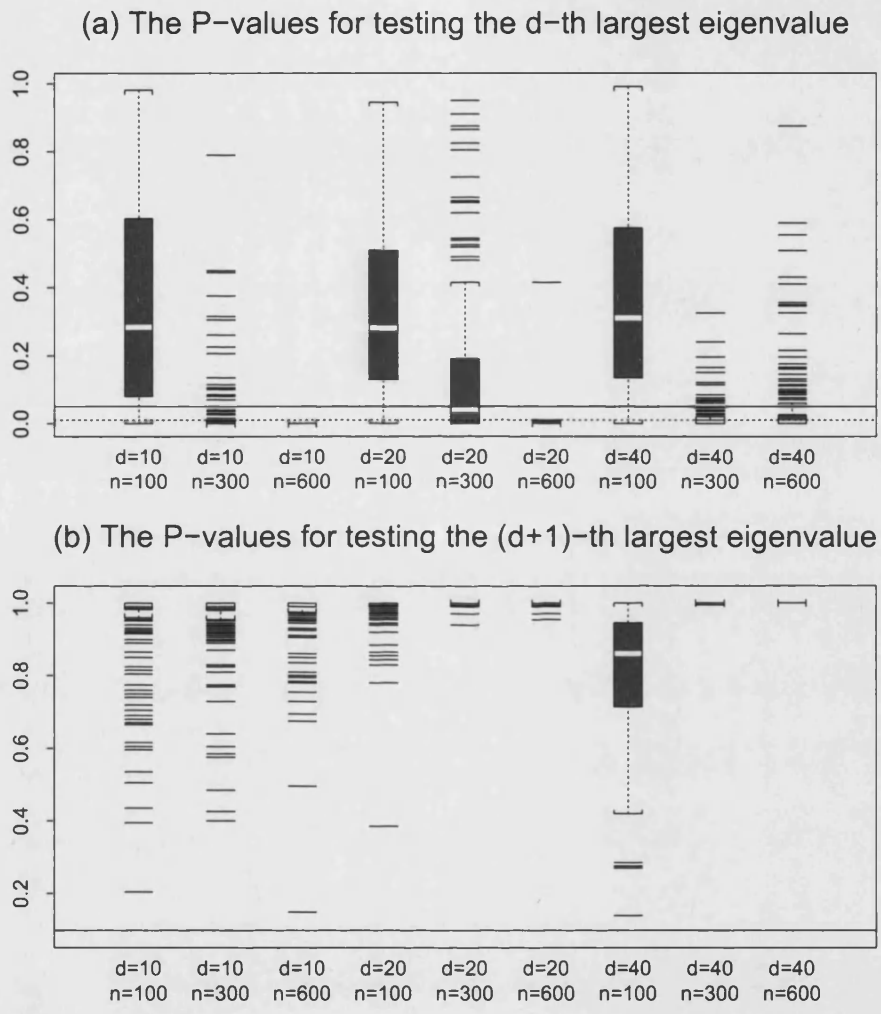


Figure 1.2: *The boxplots of the P-values for the bootstrap tests of the hypothesis that (a) the d-th largest eigenvalue of K is 0, and (b) the $(d+1)$-th largest eigenvalue of K is 0. In (a), the solid horizontal line marks the position of the 5% significance level, and the dotted line marks the 1% significance level. In (b), the solid horizontal line marks the position of the 10% significance level.*

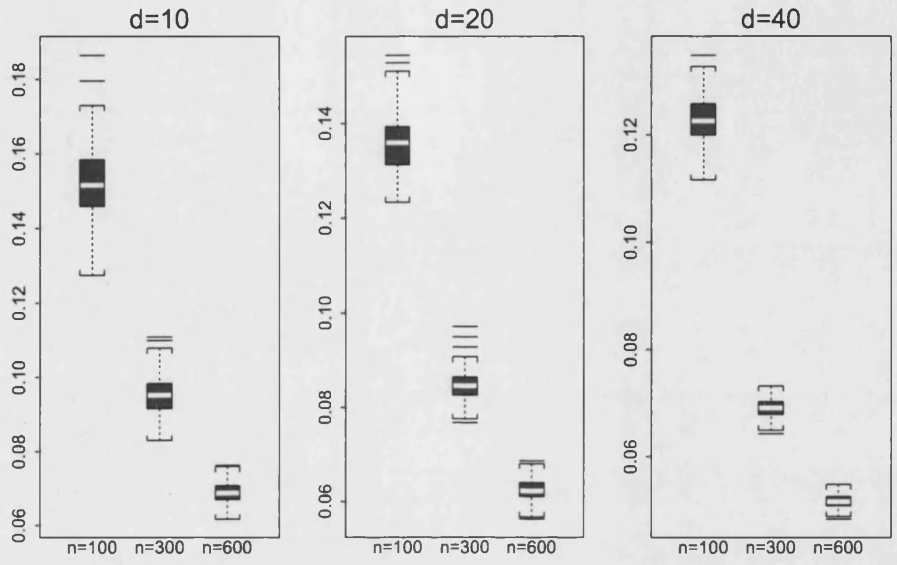


Figure 1.3: *Boxplots of the estimated error* $D'(\widehat{\mathcal{M}}, \mathcal{M})$ *defined in (1.19).*

1.4.2 Intraday return densities

To illustrate the methodology developed in this chapter, we set upon the task of modeling the intraday return densities for the IBM stock in 2006. To this end, we have obtained the intraday prices from NYSE TAQ via the WRDS database. We have only used prices between 09:30 - 16:00 since the market is not particularly active outside of these times. There are $n = 251$ trading days in the sample and a total of 2786650 observations. The dataset is 73.7 MB in total.

Since high frequency prices are not equally spaced in time, it is worth mentioning how we compute the returns. We adopt a standard 'clock time' scheme for which we have decided to work with 5 minute returns. The clock time strategy samples the prices as follows. Let $X_i(t_{ij})$ denote the observed stock price on the i-th day, $i = 1, \ldots, n$, at time t_{ij}, $j = 1, \ldots, n_i$. Put $X_{i0} = X_i(t_{i1})$. Now let τ_l, $l = 1, \ldots, m$ be the times at which we would like to collect prices. In our case with the requirement of 5 minute returns we would have $\tau_1 =$ 09:35, $\tau_2 =$ 09:40, $\ldots$, $\tau_m =$ 16:00 and $m = 78$. Now for day i, define the previous tick time

$$\tau_{il} = \max\{t_{ij} : t_{ij} < \tau_l, l = 1, \ldots, n_i\}, \qquad l = 1, \ldots, m.$$

Let $X_{il} = X_i(\tau_{il})$. Then the l-th return on day i is given by $Z_{il} = \log(X_{il}/X_{i,l-1})$, $l = 1, \ldots, m$. Note that sampling at 5 minutes in clock time for this dataset yields a total of $n \times m = 19578$ effective observations.

One may also sample the prices based on some activity based scheme. For example, rather than sampling the prices every few minutes, an alternative may be to perform the sampling every few ticks. This type of strategy is known as sampling in 'business time'. We performed the analysis using the business time sampling strategy as well but we have not reported the results here since they were very similar to the clock time based method.

Now given a set of high frequency returns, we estimate the intraday return densities by using a simple kernel density estimator

$$Y_t(u) = (nh_t)^{-1} \sum_{j=1}^{m} K\left(\frac{Z_{tj} - u}{h_t}\right), \qquad t = 1, \ldots, n, \tag{1.20}$$

where $K(u) = (\sqrt{2\pi})^{-1} \exp(-u^2/2)$ is a Gaussian kernel and h_t is a bandwidth. For all values of t, we take the support of $Y_t(\cdot)$ to be $\mathcal{I} = [-0.002, 0.002]$.

Let $\widehat{\sigma}_t$ be the sample standard deviation of Z_{tj} and $\widehat{h}_t = 1.06\widehat{\sigma}_t m^{-1/5}$ be Silverman's rule of thumb bandwidth choice for day t. Then for each t, we employ three different levels of smoothing (low, medium and high) by setting h_t in (1.20) equal to $0.5\widehat{h}_t$, $\widehat{h}_t$ and $2\widehat{h}_t$. More elaborate smoothing techniques are the subject of further study; see also the discussion in Section 1.5. Figure 1.4 displays the observed densities for the first 8 days of the sample.

Using the observed densities, $Y_t(\cdot)$, we apply the methodology developed in this chapter. In defining the operator K in (1.12), we take $p = 5$ for all levels of smoothing. Estimation results for different values of p are very similar and thus not reported here.

Figure 1.5 displays the estimated eigenvalues of the operator K. A feature that stands out from this graphic is that the eigenvalues are uniformly smaller when larger bandwidths are used. Since the size of the eigenvalues determines the strength of the dynamics in their corresponding eigensubspace, this finding is intuitive as one would expect that the more the data is smoothed, the less dynamics there will be.

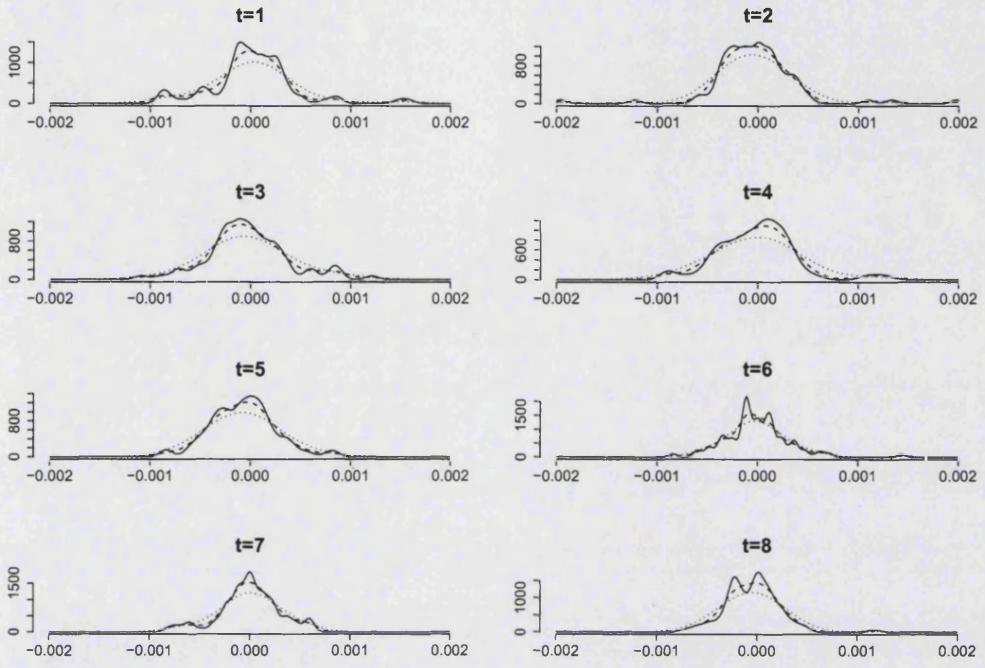


Figure 1.4: *Observed densities, $Y_t(\cdot)$, using bandwidths $h_t = 0.5\widehat{h}_t$ (solid lines), $\widehat{h}_t$ (dashed lines) and $2\widehat{h}_t$ (dotted lines).*

It is clear from Figure 1.5 that for all levels of smoothing, the first two eigenvalues are much larger than the remaining ones. Indeed, from the third eigenvalue onwards it appears that there is no clear cut-off. This would lead one to conclude that for all levels of smoothing, we have two factors driving the dynamic behavior in the density functions. These findings are supported by Table 1.2 which contains the P-values from 100 replications of the bootstrap test in Section 1.2.2.3. For all levels of smoothing, we reject the null $H_0 : \theta_2 = 0$ but cannot reject the hypothesis that $\theta_j = 0$ for $j = 3, 4, 5$. We can of course continue to test $\theta_j = 0$ for $j \geq 6$ but there is little point in doing so since if $\theta_j = 0$ this automatically implies that $\theta_{j+1} = 0$; recall these eigenvalues are always arranged in descending order. Indeed, we only proceeded beyond testing $\theta_3 = 0$ for illustrative purposes. Collating all these findings, we set $\widehat{d} = 2$ for all levels of smoothing in proceeding analysis.

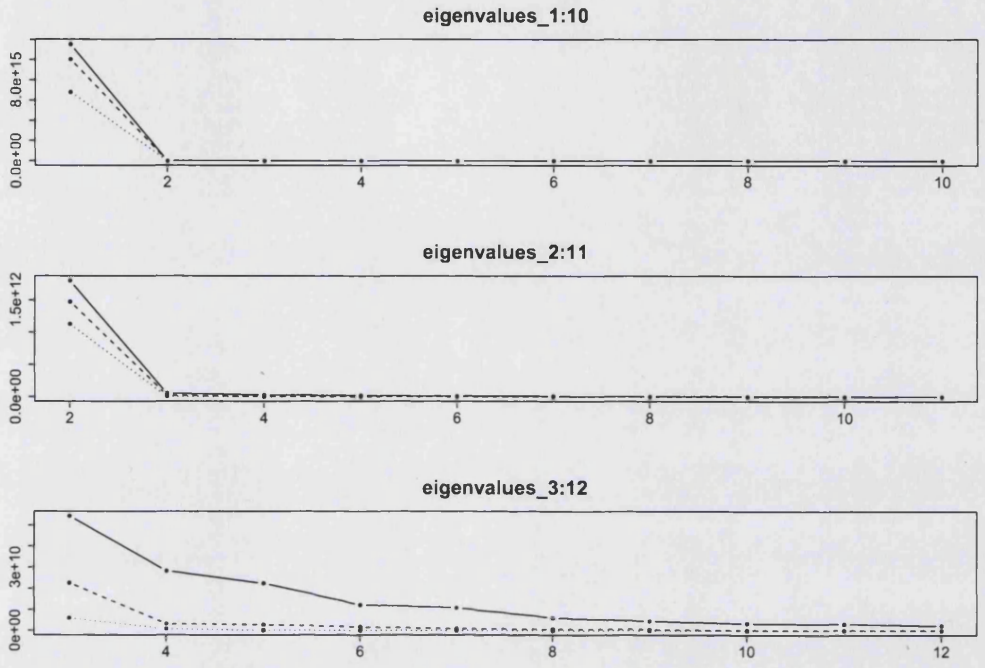


Figure 1.5: *Estimated eigenvalues* $\widehat{\theta}_j$ *using bandwidths* $h_t = 0.5\widehat{h}_t$ *(solid lines),* $\widehat{h}_t$ *(dashed lines) and* $2\widehat{h}_t$ *(dotted lines).*

Figure 1.6 displays the first $\widehat{d} = 2$ estimated eigenfunctions $\widehat{\psi}_j$ in (1.14) for all three levels of smoothing. A striking feature of this graphic is that although the observed densities, $Y_t(\cdot)$ in Figure 1.4 look very different for each level of smoothing, the estimated eigenfunctions appear to be very insensitive to the level of smoothing in terms of their general shape. Of course, for smaller values of $\widehat{h}_t$ the estimated eigenfunctions will appear more wiggly. The intuition behind this finding is fairly simple; estimation of the eigenfunctions can be thought of as a standard semi-parametric estimation problem where the density functions themselves are nuisance parameters. In this sense, it is fairly well known that the feasible set for h_t in (1.20) is larger when our parameters of interest are the eigenfunctions rather than the density functions themselves. Another possible reason behind this finding which is unique to our methodology is discussed in Section 1.5.

	$h_t = 0.5\widehat{h}_t$	$h_t = \widehat{h}_t$	$h = 2\widehat{h}_t$
$H_0 : \theta_1 = 0$	0.00	0.00	0.00
$H_0 : \theta_2 = 0$	0.00	0.00	0.00
$H_0 : \theta_3 = 0$	0.35	0.15	0.18
$H_0 : \theta_4 = 0$	0.62	0.73	0.74
$H_0 : \theta_5 = 0$	0.68	0.91	0.93

Table 1.2: *P-values from applying the bootstrap test in Section 1.2.2.3 to the intraday return density example.*

Figure 1.7 displays time series plots of the estimated loadings $\widehat{\eta}_{t1}$ and $\widehat{\eta}_{tj}$ in (1.15). Again, a remarkable feature of this graphic is that although the observed densities in Figure 1.4 are clearly dissimilar for different choices of h_t, the estimated loadings are almost indistinguishable from one another. Furthermore, the ACF and PACF of the series $\widehat{\boldsymbol{\eta}}_t = (\widehat{\eta}_{t1}, \widehat{\eta}_{t2})'$ are also virtually identical for all three levels of smoothing. These graphics are displayed in Figure's 1.8 and 1.9.

Next we fit some VAR models to the estimated loadings, $\widehat{\boldsymbol{\eta}}_t$;

$$\widehat{\boldsymbol{\eta}}_t = \sum_{k=1}^{\tau} \mathbf{A}_k \widehat{\boldsymbol{\eta}}_{t-k} + \mathbf{e}_t. \tag{1.21}$$

Since the estimated loadings have mean zero by construct, we do not fit an intercept term. We choose the order of the VAR models, τ in (1.21), by minimizing the AIC. The AIC values for the order $\tau = 0, 1, \ldots, 5$ are given in Table 1.3. For all three levels of smoothing, the AIC criterion chose $\tau = 3$ and the multivariate portmanteau test (with lag values 1, 3, and 5) of Li & McLeod (1981) for the residuals of the fitted VAR models are insignificant at the 10% level. The Yule-Walker estimates of the parameter matrices, $\mathbf{A}_k = (a_{k,ij})$ in (1.21), with the order $\tau = 3$ are given in Table 1.4.

To summarize, we found that the dynamic behavior of the IBM intraday return densities in 2006 was driven by $\widehat{d} = 2$ factors. These factors are modeled well using VAR models of order $\tau = 3$. A striking feature of this analysis was that for all three levels of smoothing that we applied, these conclusions were identical and much of the resulting inference was also very similar. Some further insights on this last point are given in Section 1.5.

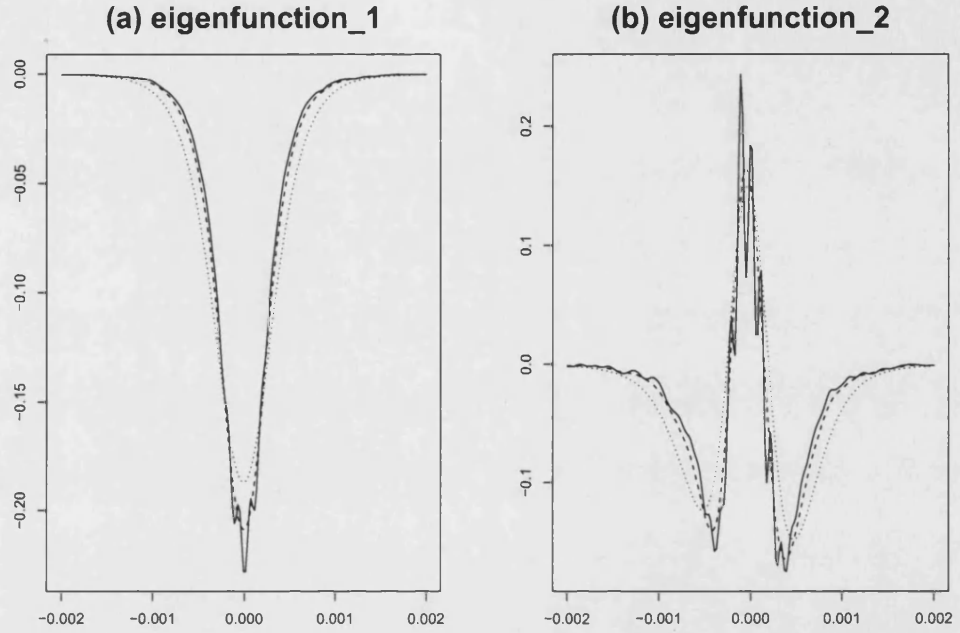


Figure 1.6: *Estimated eigenfunctions (a)* $\widehat{\psi}_1$ *and (b)* $\widehat{\psi}_2$ *using bandwidths* $h_t = 0.5\widehat{h}_t$ *(solid lines),* $\widehat{h}_t$ *(dashed lines) and* $2\widehat{h}_t$ *(dotted lines).*

	$\tau = 0$	$\tau = 1$	$\tau = 2$	$\tau = 3$	$\tau = 4$	$\tau = 5$
$h_t = 0.5\widehat{h}_t$	131.33	40.39	9.98	0.00	7.86	10.38
$h_t = \widehat{h}_t$	133.04	41.32	9.53	0.00	7.47	10.08
$h_t = 2\widehat{h}_t$	135.47	40.83	9.58	0.00	7.00	8.94

Table 1.3: *AIC values from fitting the VAR model in (1.21). The figures in this table have been centered at the minimum AIC value.*

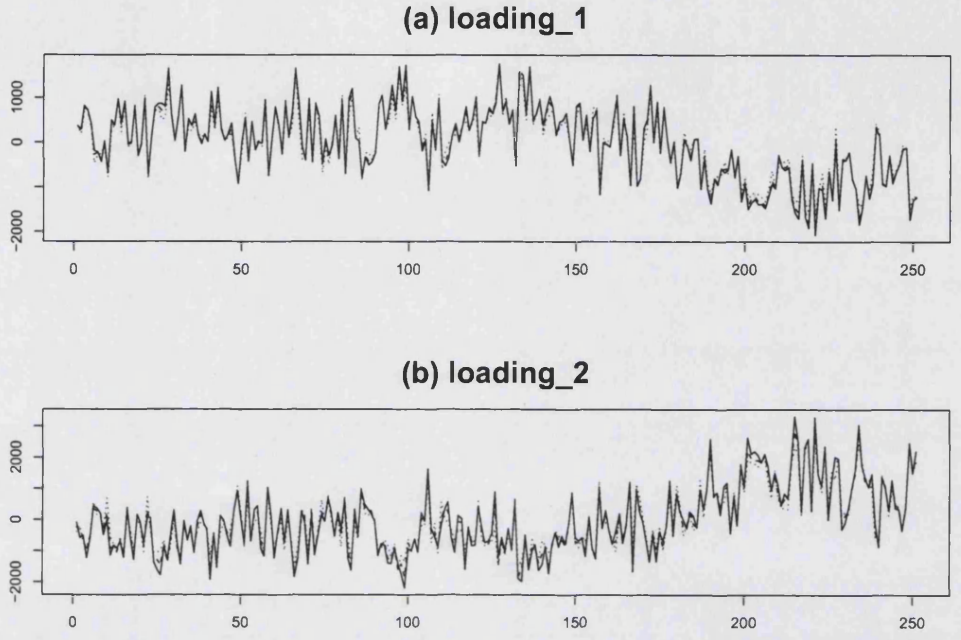


Figure 1.7: *Estimated loadings (a) $\widehat{\eta}_{t1}$ and (b) $\widehat{\eta}_{t2}$ using bandwidths $h_t = 0.5\widehat{h}_t$ (solid lines), $\widehat{h}_t$ (dashed lines) and $2\widehat{h}_t$ (dotted lines).*

j	1			2		
h_t	$0.5\widehat{h}_t$	$\widehat{h}_t$	$2\widehat{h}_t$	$0.5\widehat{h}_t$	$\widehat{h}_t$	$2\widehat{h}_t$
$a_{1,1j}$	0.08	0.07	0.01	-0.14	-0.16	-0.22
$a_{1,2j}$	-0.08	-0.05	0.03	0.24	0.26	0.33
$a_{2,1j}$	0.35	0.39	0.38	0.06	0.09	0.08
$a_{2,2j}$	-0.36	-0.43	-0.43	-0.05	-0.10	-0.11
$a_{3,1j}$	0.08	0.05	0.02	-0.13	-0.15	-0.18
$a_{3,2j}$	-0.16	-0.13	-0.11	0.14	0.15	0.17

Table 1.4: *Estimated parameter matrices $\mathbf{A}_k = (a_{k,ij})$ from fitting the VAR model in (1.21).*

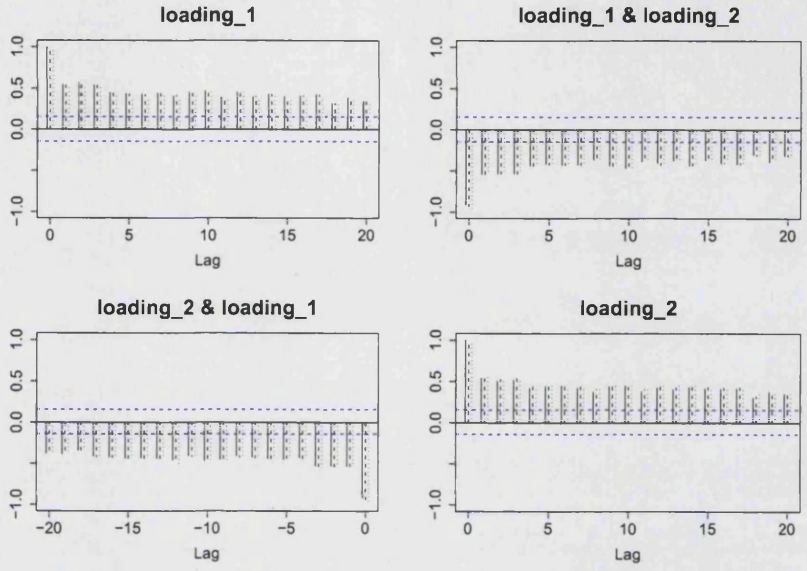


Figure 1.8: *ACF of* $\widehat{\boldsymbol{\eta}}_{tj}$ *using bandwidths* $h = 0.5\widehat{h}_t$ *(solid lines),* $\widehat{h}_t$ *(dashed lines) and* $2\widehat{h}_t$ *(dotted lines).*

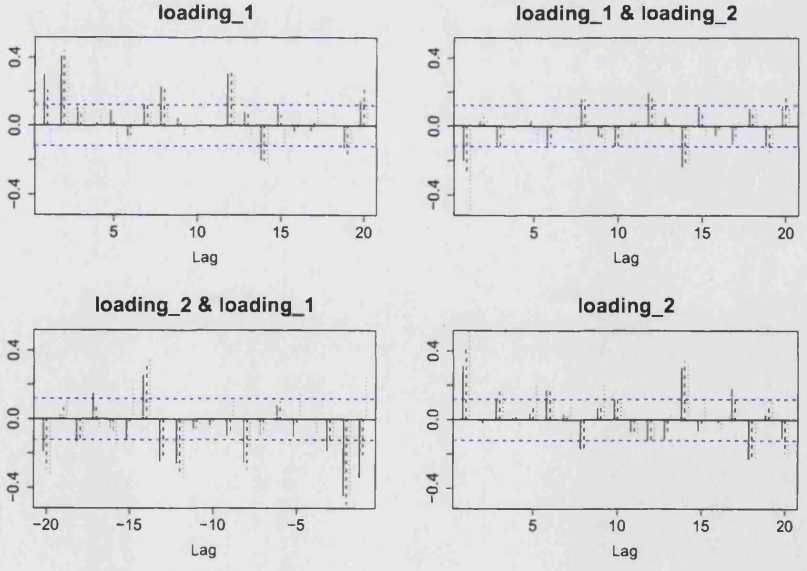


Figure 1.9: *PACF of* $\widehat{\boldsymbol{\eta}}_{tj}$ *using bandwidths* $h = 0.5\widehat{h}_t$ *(solid lines),* $\widehat{h}_t$ *(dashed lines) and* $2\widehat{h}_t$ *(dotted lines).*

1.5 Discussion

In this chapter, we have developed a method for identifying the finite dimensionality of curve time series for the realistic setting where the curves of interest are observed with error. Based upon a computational shortcut we have developed, the practical implementation of our methodology is trivial even for higher dimensional applications.

We conclude with some remarks which may spur future research. As we saw in the IBM density function example in Section 1.4.2, much of our inference was unaffected by the level of smoothing we applied to the observed densities. In particular, the estimated eigenfunctions had an identical shape for the different bandwidth values we applied.

Consider the standard observational setting

$$Y_{ij} = X_i(u_{ij}) + \varepsilon_{ij}, \qquad i = 1, \ldots, n, j = 1, \ldots, m_i, \tag{1.22}$$

where ε_{ij} is some random noise. Now let $Y_i(\cdot)$ be a local polynomial estimate of $X_i(\cdot)$. Then in traditional functional data analysis, one is typically interested in estimating the eigenfunctions of $M_0(u,v) = \text{Cov}\{X_i(u), X_i(v)\}$ for which we often use the eigenfunctions of $\widehat{M_0} = n^{-1}\sum_{i=1}^{n}\{Y_i(u) - \bar{Y}(u)\}\{Y_i(v) - \bar{Y}(v)\}$ as our estimates. In this case, it was shown in Hall *et al.* (2006) that the resulting estimates are root-n consistent, in the $\mathcal{L}_2(\mathcal{I})$ sense, provides $m = \min m_i$ diverges with n. However, if the observations on each function are sparse, i.e. m is fixed, and provided that t_{ij} are random enough, then the minimax optimal rate of convergence empirical eigenfunctions is $n^{-r/(2r+1)}$. Here r is the order of the local polynomial estimator used and thus the number of derivatives assumed on the population functions. The intuition behind this result is simple; under an appropriate set of regularity conditions, if $m \longrightarrow \infty$, $Y_i(\cdot)$ is consistent for $X_i(\cdot)$ and thus $\widehat{M_0}$ is root-n consistent for M_0. From here, one can use standard techniques (see Hall & Hosseini-Nasab (2006) and Hall & Hosseini-Nasab (2008)) to show that the corresponding eigenfunctions are also root-n consistent. However, when the observations on each function are sparse, the problem becomes a nonparametric one and the familiar rates of convergence from this paradigm are obtained.

Our setting is different in that we do not require our estimate $Y_i(\cdot)$ to be consistent for $X_i(\cdot)$. In fact to obtain root-n convergence of the eigenfunctions and thus the dynamic space $\mathcal{M}$, all we require is that $Y_i(\cdot)$ has the same auto-correlation structure as $X_i(\cdot)$ or more generally that $\widehat{M_k}$ in (1.9) is root-n consistent for M_k in (1.3). This leads to an open question;

> *Given the discrete data design in (1.22), is it possible to construct a root-n consistent estimate of $\mathcal{M}$ for the setting where the observations on each curve $X_i(\cdot)$ are sparse, i.e. m is fixed?*

We believe that if this can be achieved, then it would be a significant achievement in the modern era of high dimensional data analysis.

1.6 Proofs

In this section we provide the proofs of the propositions in Section 1.2 and the theorems in Section 1.3. Throughout the proofs we may use C to denote a positive and finite constant which may vary from line to line. We refer the reader to Appendix A for some background on operator theory which is required for the proofs. We introduce some technical lemmas first.

Lemma 1.1 *Let L be a finite dimensional operator such that for some sequences of orthonormal vectors $\{e_j\}$, $\{f_j\}$, $\{g_j\}$ and $\{h_j\}$ and some sequences of decreasing scalars $\{\theta_j\}$ and $\{\lambda_j\}$, L admits the spectral decompositions $L = \sum_{j=1}^{d} \theta_j e_j \otimes f_j = \sum_{j=1}^{d'} \lambda_j g_j \otimes h_j$. Then it holds that $d' = d$.*

Proof of Lemma 1.1 Note that if $d \neq d'$ then both $\mathrm{Im}(L)$ and $\mathrm{Im}(L_k^*)$ will be of different dimensions under the alternative characterizations due to linear independence of $\{e_j\}$, $\{f_j\}$, $\{g_j\}$ and $\{h_j\}$. Thus it must hold that $d = d'$. □

Lemma 1.2 *Let L be a linear operator from $\mathcal{H}$ to $\mathcal{H}$, where $\mathcal{H}$ is a separable Hilbert space. Then it holds that $\overline{Im(LL^*)} = \overline{Im(L)}$.*

Proof of Lemma 1.2 First note that $\mathrm{Ker}(L^*) = (\mathrm{Im}(L))^\perp$, $\mathrm{Ker}(L) = (\mathrm{Im}(L^*))^\perp$ and $\mathrm{Ker}(L^*) = \mathrm{Ker}(LL^*)$. Thus

$$\begin{aligned}
\overline{\mathrm{Im}(LL^*)} &= (\mathrm{Im}(LL^*))^{\perp\perp} \\
&= (\mathrm{Im}((LL^*)^*))^{\perp\perp} \\
&= (\mathrm{Ker}(LL^*))^\perp \\
&= (\mathrm{Ker}(L^*))^\perp \\
&= (\mathrm{Im}(L))^{\perp\perp} \\
&= \overline{\mathrm{Im}(L)},
\end{aligned}$$

which concludes the proof. □

For the sake of the simplicity in presentation of the remaining proofs, we adopt the standard notation for operators acting on Hilbert spaces. For any $f \in \mathcal{L}_2(\mathcal{I})$, we write $\|f\| = \sqrt{\langle f, f\rangle}$ (see (1.5)), and denote $M_k f \in \mathcal{L}_2(\mathcal{I})$ the image of f under the operator M_k in the sense that

$$(M_k f)(u) = \int_{\mathcal{I}} M_k(u, v) f(v) dv.$$

The operators $N_k, K, \widehat{M}_k$ and $\widehat{K}$ may be expressed in the same manner. Note now that the adjoint operator of M_k is

$$(M_k^* f)(u) = \int_{\mathcal{I}} M_k(v, u) f(v) dv.$$

Furthermore $N_k = M_k M_k^*$ in the sense that $N_k f = M_k M_k^* f$; see (1.10). In the same way $K = \sum_{k=1}^p M_k M_k^*$ and $\widehat{K} = \sum_{k=1}^p \widehat{M}_k \widehat{M}_k^*$; see (1.12) and (1.13).

Proof of Proposition 1.1 (i) We only need to show $\mathrm{Im}(N_k) = \mathcal{M}$. Since $N_k = M_k M_k^*$, it follows by Lemma 1.2 that $\mathrm{Im}(N_k) = \mathrm{Im}(M_k M_k^*) = \mathrm{Im}(M_k)$ since both N_k and M_k are finite dimensional and thus their images are closed.

Now, recall from Section 1.2.1 that M_k may be decomposed as

$$M_k = \sum_{i,j=1}^{d} \sigma_{ij}^{(k)} \varphi_i \otimes \varphi_j. \tag{1.23}$$

Thus from (1.23), we may write

$$M_k = \sum_{i=1}^{d} \lambda_i^{(k)} \varphi_i \otimes \rho_i^{(k)}, \tag{1.24}$$

where

$$\rho_{ik} = \frac{\sum_{j=1}^{d} \sigma_{ij}^{(k)} \varphi_j}{\| \sum_{j=1}^{d} \sigma_{ij}^{(k)} \varphi_j \|}, \qquad \lambda_k^{(}k) = \| \sum_{j=1}^{d} \sigma_{ij}^{(k)} \varphi_i \|.$$

From (1.23), it is clear that $\mathrm{Im}(M_k) \subseteq \mathcal{M}$, which is finite dimensional. Thus M_k is compact and therefore admits a spectral decomposition of the form

$$M_k = \sum_{j=1}^{d_k} \theta_j^{(k)} \psi_j^{(k)} \otimes \phi_j^{(k)}, \tag{1.25}$$

with $(\phi_j^{(k)}, \psi_j^{(k)})$ forming an adjoint pair of singular functions of M_k corresponding to the singular value $\theta_j^{(k)}$. Clearly $d_k \leq d$. Thus if $d_k < d$, $\mathrm{Im}(M_k) \subset \mathcal{M}$ since from (1.25), $\mathrm{Im}(M_k) = \mathrm{span}\{\psi_j^{(k)} : j = 1, \ldots, d_k\}$ and any subset of $d_k (< d)$ linearly independent elements in a d-dimensional space can only span a proper subset of the original space.

Now to complete the proof, we only need to show that the set of $\{\rho_j^{(k)}\}$ in (1.24) is linearly independent for some k. If this can be done then we are in a position to apply Lemma 1.1. Let $\boldsymbol{\beta}$ be an arbitrary vector in $\mathbb{R}^d$ and put $\boldsymbol{\varphi} = (\varphi_1, \ldots, \varphi_d)'$ and $\boldsymbol{\rho}_k = (\rho_1^{(k)}, \ldots, \rho_d^{(k)})'$, then the linear independence of the set $\{\rho_i^{(k)}\}$ can easily be seen as the equation

$$\boldsymbol{\beta}\boldsymbol{\rho}_k = \boldsymbol{\beta}\boldsymbol{\Sigma}_k\boldsymbol{\varphi} = 0,$$

has a nontrivial solution if and only if $\boldsymbol{\beta}\boldsymbol{\Sigma}_k = \mathbf{0}$. However since $\boldsymbol{\Sigma}_k$ is of full rank by assumption, it follows that it is invertible and the only solution is the trivial one $\boldsymbol{\beta} = \mathbf{0}$. Thus Lemma 1.1 implies $d_k = d$ and the result follows from noting that any linearly independent set of d elements in a d-dimensional vector space forms a basis for that space.

(ii) Similarly to the proof of part (i) above, we only need to show $\mathrm{Im}(K) = \mathcal{M}$. Note that for any $f \in L_2(\mathcal{I})$, $< M_k M_k^* f,\ f > = < M_k^* f,\ M_k^* f > = \|M_k^* f\|^2 \geq 0$, thus the composition $N_k = M_k M_k^*$ is non-negative definite which implies that

$K = \sum_{k=1}^{p} N_k$ is also non-negative definite. Therefore, $\mathrm{Im}(K) = \bigcup_{k=1}^{p} \mathrm{Im}(N_k)$. From here, the result given in part (i) of the proposition concludes the proof. □

Proof of Proposition 1.2 Let $\widehat{\theta}_j$ be a non-zero eigenvalue of $\mathbf{K}^*$, and $\boldsymbol{\gamma}_j = (\gamma_{1j}, \cdots, \gamma_{n-p,j})'$ be the corresponding eigenvector, i.e. $\mathbf{K}^* \boldsymbol{\gamma}_j = \boldsymbol{\gamma}_j \widehat{\theta}_j$. Writing this equation component by component, we obtain that

$$\frac{1}{(n-p)^2} \sum_{i,s=1}^{n-p} \sum_{k=1}^{p} \langle Y_{t+k} - \bar{Y}, Y_{s+k} - \bar{Y} \rangle \langle Y_s - \bar{Y}, Y_i - \bar{Y} \rangle \gamma_{ij} = \gamma_{tj} \widehat{\theta}_j, \tag{1.26}$$

for $t = 1, \ldots, n-p$; see (1.16). For $\widehat{\psi}_j$ defined in (1.17),

$$\begin{aligned}
(\widehat{K}\widehat{\psi}_j)(u) &= \int_{\mathcal{I}} \widehat{K}(u,v)\widehat{\psi}_j(v)dv \\
&= \frac{1}{(n-p)^2} \sum_{t,s=1}^{n-p} \sum_{k=1}^{p} \{Y_t(u) - \bar{Y}(u)\} \langle Y_s - \bar{Y}, \widehat{\psi}_j \rangle \langle Y_{t+k} - \bar{Y}, Y_{s+k} - \bar{Y} \rangle \\
&= \frac{1}{(n-p)^2} \sum_{t,s,i=1}^{n-p} \sum_{k=1}^{p} \{Y_t(u) - \bar{Y}(u)\} \gamma_{ij} \langle Y_s - \bar{Y}, Y_i - \bar{Y} \rangle \langle Y_{t+k} - \bar{Y}, Y_{s+k} - \bar{Y} \rangle,
\end{aligned}$$

see (1.13). Plugging (1.26) into the right hand side of the above expression, we obtain that

$$(\widehat{K}\widehat{\psi}_j)(u) = \sum_{t=1}^{n-p} \{Y_t(u) - \bar{Y}(u)\} \gamma_{tj} \widehat{\theta}_j = \widehat{\psi}_j(u) \widehat{\theta}_j,$$

i.e. $\widehat{\psi}_j$ is an eigenfunction of $\widehat{K}$ corresponding to the eigenvalue $\widehat{\theta}_j$. □

Before presenting the proof of Theorem 1.1, we present a further technical results which proves useful in the derivations.

Lemma 1.3 *Let $A, B \in \mathcal{S}$. Then it holds that $\|AA^* - BB^*\|_{\mathcal{N}} \leq \{\|A\|_{\mathcal{S}} + \|B\|_{\mathcal{S}}\}\|A - B\|_{\mathcal{S}}$.*

Proof of Lemma 1.3 First note that for any $A, B \in \mathcal{S}$, $\|A^*\|_{\mathcal{S}} = \|A\|_{\mathcal{S}}$ and by the Cauchy-Schwarz inequality $\|AB\|_{\mathcal{N}} \leq \|A\|_{\mathcal{S}}\|B\|_{\mathcal{S}}$. Thus

$$\begin{aligned}
\|AA^* - BB^*\|_{\mathcal{N}} &= \|(A-B)A^* + B(A^* - B^*)\|_{\mathcal{N}} \\
&\leq \|(A-B)A^*\|_{\mathcal{N}} + \|B(A^* - B^*)\|_{\mathcal{N}} \\
&\leq \|A^*\|_{\mathcal{S}}\|A - B\|_{\mathcal{S}} + \|B\|_{\mathcal{S}}\|A^* - B^*\|_{\mathcal{S}} \\
&= \{\|A\|_{\mathcal{S}} + \|B\|_{\mathcal{S}}\}\|A - B\|_{\mathcal{S}},
\end{aligned}$$

as required. □

Proof of Theorem 1.1 First notice that

$$\begin{aligned}\|\widehat{K} - K\|_{\mathcal{N}} &= \|\sum_{k=1}^{p} \widehat{N}_k - N_k\|_{\mathcal{N}} \\ &\leq \sum_{k=1}^{p} \|\widehat{M}_k \widehat{M}_k^* - M_k M_k^*\|_{\mathcal{N}} \\ &\leq \sum_{k=1}^{p} \{\|\widehat{M}_k\|_{\mathcal{S}} + \|M_k\|_{\mathcal{S}}\} \|\widehat{M}_k - M_k\|_{\mathcal{S}},\end{aligned}$$

where the final inequality follows from Lemma 1.3. Now if $\widehat{M}_k \xrightarrow{p} M_k$ in the topology of $\mathcal{S}$, then $\|\widehat{M}_k\|_{\mathcal{S}} \xrightarrow{p} \|M_k\|_{\mathcal{S}} < \infty$ since the existence of $E(Y_t^2)$ guarantees that M_k is Hilbert-Schmidt. Thus we may write

$$\|\widehat{K} - K\|_{\mathcal{N}} \leq \Delta \sum_{k=1}^{p} \|\widehat{M}_k - M_k\|_{\mathcal{S}}, \tag{1.27}$$

where $\Delta = \max_{k \geq 1}\{\|\widehat{M}_k\|_{\mathcal{S}} + \|M_k\|_{\mathcal{S}}\}$. From (1.27) it is clear that we are only required to control $\|\widehat{M}_k - M_k\|_{\mathcal{S}}$ since if this quantity converges to zero, then Δ will be bounded in probability and $\widehat{K}$ will be consistent in the $\|\cdot\|_{\mathcal{N}}$ sense.

Now, some straightforward calculations show that

$$\begin{aligned}M_k &= \frac{1}{n-p} \sum_{t=1}^{n-p} (Y_t - \bar{Y}) \otimes (Y_{t+k} - \bar{Y}) \\ &= \frac{1}{n-p} \sum_{t=1}^{n-p} (Y_t - \mu) \otimes (Y_{t+k} - \mu) + O_p(n^{-1}),\end{aligned} \tag{1.28}$$

since $n \sim n - p$ (as p is fixed and finite). Put $Z_{tk} = (Y_t - \mu) \otimes (Y_{t+k} - \mu) - M_k$. Then in light of (1.28), it holds that $(n-p)^{-1} \sum_{t=1}^{n-p} Z_{tk} = \widehat{M}_k - M_k + O_p(n^{-1})$. Furthermore, since $\mathrm{Cov}\{X_s, \varepsilon_t\} = 0$ for all s, t, it holds that $E(Z_{tk}) = 0$. Now by using the fact that $\{Y_t\}$ (and thus also $\{Z_{kt}\}$) is strictly stationary, Lemma B.3 yields

$$E\|\sum_{t=1}^{n-p} Z_{tk}\|_{\mathcal{S}}^2 \leq 36n \int_0^1 \alpha^{-1}(u) Q_{\|Z_{tk}\|_{\mathcal{S}}}^2 du = O(n), \tag{1.29}$$

as $Z_{tk} \in \mathcal{F}_{-\infty}^{t}$ (which, by conditions C1 and C2, guarantees integrability of $\alpha^{-1}(u)Q^2_{\|Z_{tk}\|_S}(u)$). Thus via an application of the Chebyshev inequality to (1.29) it follows that

$$\widehat{M}_k - M_k = (n-p)^{-1}\sum_{t=1}^{n-p} Z_{tk} + O_p(n^{-1}) = O_p(n^{-1/2}). \tag{1.30}$$

Now (1.27) and (1.30) together yield $\|\widehat{K} - K\|_{\mathcal{N}} = O_p(n^{-1/2})$. This concludes the first part of the assertion.

Given $\|\widehat{K} - K\|_{\mathcal{N}} = O_p(n^{-1/2})$, Lemma B.1 implies that $\sup_{j\geq 1}|\widehat{\theta}_j - \theta_j| = O_p(n^{-1/2})$. Finally, condition C4 implies that ψ_j is an identifiable statistical parameter from which Lemma B.2 in yields $\|\widehat{\psi}_j - \psi_j\| = O_p(n^{-1/2})$ for $j = 1, \ldots, d$ since we always assume that the right versions (in terms of sign) of $\widehat{\psi}_j$ and ψ_j are used. □

Lemma 1.4 *The function D defined in (1.18) is a well defined distance measure on $\mathcal{Z}$.*

Proof of Lemma 1.4 Non-negativity, symmetry and the identity of indiscernibles are obvious. It only remains to prove the subadditivity property. For any $L \in \mathcal{S}$, note that $\|L\|_S = \sqrt{\mathrm{tr}(L^*L)}$, where tr denotes the trace operator. Now, for any $\mathcal{X}_i \in \mathcal{Z}$, $i = 1, 2, 3$, let $\Pi_{\mathcal{X}_i}$ denote its corresponding d dimensional projection operators defined as follows

$$\Pi_{\mathcal{X}_i} = \sum_{j=1}^{d} \zeta_{ij} \otimes \zeta_{ij},$$

where $\{\zeta_{ij} : j = 1, \ldots, d\}$ is some orthonormal basis of $\mathcal{X}_i$.

Now the triangle inequality for the Hilbert-Schmidt norm yields

$$\|\Pi_{\mathcal{X}_1} - \Pi_{\mathcal{X}_3}\|_S \leq \|\Pi_{\mathcal{X}_1} - \Pi_{\mathcal{X}_2}\|_S + \|\Pi_{\mathcal{X}_2} - \Pi_{\mathcal{X}_3}\|_S.$$

Since the projection operators are self adjoint, we have

$$\begin{aligned} & \sqrt{\mathrm{tr}(\Pi^2_{\mathcal{X}_1}) + \mathrm{tr}(\Pi^2_{\mathcal{X}_3}) - 2\mathrm{tr}(\Pi_{\mathcal{X}_1}\Pi_{\mathcal{X}_3})} \\ \leq\ & \sqrt{\mathrm{tr}(\Pi^2_{\mathcal{X}_1}) + \mathrm{tr}(\Pi^2_{\mathcal{X}_2}) - 2\mathrm{tr}(\Pi_{\mathcal{X}_1}\Pi_{\mathcal{X}_2})} + \sqrt{\mathrm{tr}(\Pi^2_{\mathcal{X}_2}) + \mathrm{tr}(\Pi^2_{\mathcal{X}_3}) - 2\mathrm{tr}(\Pi_{\mathcal{X}_2}\Pi_{\mathcal{X}_3})}. \end{aligned}$$

Now $\mathrm{tr}(\Pi_{\mathcal{X}_i}^2) = \mathrm{tr}(\Pi_{\mathcal{X}_i}) = d$ and $\mathrm{tr}(\Pi_{\mathcal{X}_i}\Pi_{\mathcal{X}_j}) = \sum_{l,k=1}^{d} \langle \zeta_{il}, \zeta_{jk} \rangle^2$ for $i, j = 1, 2, 3$. These last facts along with the definition of D in (1.18) give

$$D(\mathcal{X}_1, \mathcal{X}_3) \leq D(\mathcal{X}_1, \mathcal{X}_2) + D(\mathcal{X}_2, \mathcal{X}_3),$$

which concludes the proof. □

Proof of Theorem 1.2 (i) First note that from (1.18)

$$\sqrt{2d} D(\widehat{\mathcal{M}}, \mathcal{M}) = ||\Pi_{\widehat{\mathcal{M}}} - \Pi_{\mathcal{M}}||_S, \tag{1.31}$$

where $\Pi_{\widehat{\mathcal{M}}} = \sum_{j=1}^{d} \widehat{\psi}_j \otimes \widehat{\psi}_j$ and $\Pi_{\mathcal{M}} = \phi_j \otimes \phi_j$ with $\phi_1, \ldots, \phi_d$ forming any orthonormal basis of $\mathcal{M}$. Now if $\Pi_{\mathcal{M}}^1$ and $\Pi_{\mathcal{M}}^2$ are any projection operators onto $\mathcal{M}$, then by virtue of Lemma 1.4 it holds that $||\Pi_{\mathcal{M}}^1 - \Pi_{\mathcal{M}}^2||_S = \sqrt{2} D(\mathcal{M}, \mathcal{M}) = 0$. Thus we may proceed as if $\Pi_{\mathcal{M}}$ in (1.31) was formed with eigenfunctions of K, i.e. $\phi_j = \psi_j$ for $j = 1, \ldots, d$.

Now we have

$$||\sum_{j=1}^{d} \widehat{\psi}_j \otimes \widehat{\psi}_j - \sum_{j=1}^{d} \psi_j \otimes \psi_j||_S \leq \sum_{j=1}^{d} ||\widehat{\psi}_j \otimes \widehat{\psi}_j - \psi_j \otimes \psi_j||_S, \tag{1.32}$$

i.e. $\widehat{\psi}_j \otimes \widehat{\psi}_j$ (resp. $\psi_j \otimes \psi_j$) is the projection operator onto the eigensubspace generated by $\widehat{\theta}_j$ (resp. θ_j). Now by Theorem 1.1, $\|\widehat{K} - K\|_N \geq \|\widehat{K} - K\|_S = O_p(n^{-1/2})$. Thus an application of Lemma B.4 yields $||\widehat{\psi}_j \otimes \widehat{\psi}_j - \psi_j \otimes \psi_j||_S = O_p(n^{-1/2})$ for $j = 1, \ldots, d$. This last fact along with (1.31) and (1.32) yield $D(\widehat{\mathcal{M}}, \mathcal{M}) = O_p(n^{-1/2})$.

(ii) It remains to prove the adaptivity property. For any constant $C > 0$

$$\begin{aligned}
& P\{n^{1/2}|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| > C\} \\
= \; & P\{n^{1/2}|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| > C, \, \widehat{d} = d\} \\
& + P\{n^{1/2}|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M}), \, \widehat{d} \neq d\} \\
= \; & P\{n^{1/2}|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| > C \,|\, \widehat{d} = d\} P(\widehat{d} = d) \\
& + P\{n^{1/2}|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| > C \,|\, \widehat{d} \neq d\} P(\widehat{d} \neq d) \\
\leq \; & P\{n^{1/2}|D'(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| > C \,|\, \widehat{d} = d\} P(\widehat{d} = d) + P(\widehat{d} \neq d)
\end{aligned} \tag{1.33}$$

Now by assumption $P(\widehat{d} = d) \longrightarrow 1$ and thus $P(\widehat{d} \neq d) = o(1)$. Furthermore, if $\widehat{d} = d$ it holds that $D(\widehat{\mathcal{M}}, \mathcal{M}) = D(\widehat{\mathcal{M}}, \mathcal{M})$. These last relations along with (1.33) yield $P\{n^{1/2}|D(\widehat{\mathcal{M}}, \mathcal{M}) - D(\widetilde{\mathcal{M}}, \mathcal{M})| > C\} = o(1)$ as required. □

Chapter 2

Methodology and convergence rates for factor modeling of multiple time series

2.1 Introduction

When modelling a time series $\mathbf{Y}_t \in \mathbb{R}^p$, a crucial task in any pre-analysis is to try and reduce the dimensionality if p is large. To see why, if one tries to fit a VAR model to $\mathbf{Y}_t$ then we are required to estimate $O(p^2)$ parameters and in many modern statistical data sets we may have $p \approx n$ or even $p >> n$, where n is the sample size. In these settings, our estimation is likely to be inaccurate and we may end up making misleading inferences as a result. This is the so called "curse of dimensionality" and is at the forefront of modern statistical research; see Donoho (2000) and Fan & Li (2006).

There have been many attempts in reducing the dimensionality for multiple time series which include principal component based approaches (Preistley *et al.* (1974) and Stock & Watson (2002)), canonical correlation based methods (Box & Tiao (1977) and Tiao & Tsay (1989)) and factor models (Bai (2003), Engle & Watson (1981), Forni *et al.* (2000), Lam & Yao (2009), Pena & Box (1987) and Pan & Yao (2008)).

As suggested by the title, the focus of this chapter will be on factor models for multiple time series. The form of the model that we consider is identical to that in

Pan & Yao (2008) who identified the factor loading space by expanding the white noise space step by step and used portmanteau tests as a stopping rule. Although our estimation procedure is based upon the same principle as Pan & Yao (2008), our implementation is much more efficient and can handle cases with very large p. See also Lam & Yao (2009). An interesting feature of our methodology is that nonstationarity is permitted and does not necessarily have to be driven by unit roots. The latter was consider in Ahn (1997) and Pena & Poncela (2006). In fact, all that is required to obtain consistency of the factor loading space is that the sample autocovariance matrices are consistent for their population counterparts. This is often implied by ergodicity of a stationary process and may also be fulfilled by some non-stationary mixing processes; see the remarks made in Section 2.3.

The main focus of this chapter is on identifying the number of latent factors. To this end, we present a simple white noise test and provide some theory to support it. In particular, we argue that the number of factors is equal to the number of non-zero eigenvalues of a $p \times p$ matrix, which is simply a function of the population autocovariance matrices. Although the sample analogue of this matrix is likely to be full ranked, we prove a striking result which is unique to our methodology; the eigenvalues whose population counterparts are truly zero are "super-consistent" under an ideal set of condition, i.e. they converge to zero at the rate n^{-1}. An example of a consistent threshold based estimator of the number of latent factors is also given.

The rest of this chapter is organized as follows. Section 2.2 introduces the methodology for estimating the factor loading space and the white noise test for determining the number of latent factors. The theoretical results are presented in Section 2.3 and some simulations investigating the white noise test and also demonstrating the fast convergence rate of the eigenvalues are given in Section 2.4. We conclude with the analysis of some implied volatility surface data. All of the technical proofs are relegated to Section 2.6.

2.2 Methodology

2.2.1 Factor models

Let $\mathbf{Y}_t$ be a $p \times 1$ time series generated by the factor model

$$\mathbf{Y}_t = \mathbf{A}\mathbf{X}_t + \boldsymbol{\varepsilon}_t, \qquad t = 1, \ldots, n, \tag{2.1}$$

where $\mathbf{X}_t$ is a $d \times 1$ time series with $d \leq p$ unknown, $\mathbf{A}$ is a $p \times d$ unknown constant matrix and $\boldsymbol{\varepsilon}_t$ is a $p \times 1$ white noise process in the sense that $E(\boldsymbol{\varepsilon}_s \boldsymbol{\varepsilon}_t') = 0$ for any $s \neq t$. Note that we do not lose any generality by assuming that $\boldsymbol{\varepsilon}_t$ is a white noise sequence since if this was not the case then any parts of $\boldsymbol{\varepsilon}_t$ which possess serial correlation should be absorbed into $\mathbf{X}_t$. Conversely, we may also assume that there exists no linear combination of $\mathbf{X}_t$ which is a white noise process otherwise such a linear combination may be absorbed into $\boldsymbol{\varepsilon}_t$. We only observe $\mathbf{Y}_1, \ldots, \mathbf{Y}_n$ from the factor model (2.1). To simplify the presentation, we assume that $E(\mathbf{Y}_t) = 0$. In practice this amounts to replacing $\mathbf{Y}_t$ by $\mathbf{Y}_t - \bar{\mathbf{Y}}$ before the analysis, where $\bar{\mathbf{Y}} = n^{-1} \sum_{t=1}^{n} \mathbf{Y}_t$.

The components of $\mathbf{X}_t$ are called the common factors and $\mathbf{A}$ is called the factor loading matrix. We may assume that the rank of $\mathbf{A}$ is d since if this is not the case then (2.1) may be expressed equivalently using a smaller number of factors. Note that model (2.1) is unchanged if we replace $\mathbf{A}$ and $\mathbf{X}_t$ by $\mathbf{AH}$ and $\mathbf{H}^{-1}\mathbf{X}_t$ for any invertible $d \times d$ matrix $\mathbf{H}$. Therefore, we may assume that the columns vectors of $\mathbf{A} = (\mathbf{a}_1, \ldots, \mathbf{a}_d)$ are orthonormal, i.e.

$$\mathbf{A}'\mathbf{A} = \mathbf{I}_d, \tag{2.2}$$

where $\mathbf{I}_d$ denotes that $d \times d$ identity matrix. Note that even with the constraint in (2.2), $\mathbf{A}$ and $\mathbf{X}_t$ are still not uniquely determined in (2.1) as the aforementioned replacement is still applicable for any invertible $\mathbf{H}$. However, the linear space spanned by the columns of $\mathbf{A}$, denoted $\mathcal{M}$ and called the factor loading space, is a uniquely defined d-dimensional subspace of $\mathbb{R}^p$; see the arguments in Section 2.2.2.

2.2.2 Estimation of $\mathcal{M}$

Let $\mathbf{\Sigma}_y(k) = E(\mathbf{Y}_t\mathbf{Y}'_{t+k})$ and $\mathbf{\Sigma}_x(k) = E(\mathbf{X}_t\mathbf{X}'_{t+k})$ denote the lag k autocovariance matrices of $\mathbf{Y}_t$ and $\mathbf{X}_t$ respectively. We impose the following identifiability conditions:

A1. $\boldsymbol{\varepsilon}_s$ is a white noise sequence uncorrelated with $\mathbf{X}_t$ for all s, t.

A2. $\text{rank}\{\mathbf{\Sigma}_x(k)\} = d$ for some $1 \leq k \leq q$.

Then under conditions A1 and A2, Proposition 1.1 in Chapter 1 implies that the matrix $\mathbf{L}_q$ defined by

$$\mathbf{L}_q = \sum_{k=1}^{q} \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)', \qquad q \geq 1, \tag{2.3}$$

has exactly d non-zero eigenvalues and $\mathcal{M}$ is the linear space spanned by the corresponding eigenvectors.

A formal proof of this result is given in Chapter 1 in the more general setting where the observations are functional. However, in the vector case it is much easier to see the reasoning behind this result. Under the assumption that $\mathbf{X}_s$ and $\boldsymbol{\varepsilon}_t$ are uncorrelated for all s, t it follows that

$$\mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)' = \mathbf{A}\mathbf{\Sigma}_x(k)\mathbf{\Sigma}_x(k)'\mathbf{A}', \tag{2.4}$$

since $\mathbf{A}'\mathbf{A} = \mathbf{I}_d$; see (2.2). Thus from (2.4) it is clear that the eigenvectors of $\mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)'$ are elements of $\mathcal{M}$. In addition, it is clear that $\text{rank}\{\mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)'\}$ $\leq d$ with equality if and only if $\text{rank}\{\mathbf{\Sigma}_x(k)\} = d$. Thus if $\text{rank}\{\mathbf{\Sigma}_x(k)\} = d$, it holds that $\mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)'$ has exactly d uniquely determined eigenvectors (up to sign provided that the eigenvalues are all different) corresponding to d non-zero eigenvalues. By noting these last few facts, the result above follows by recalling that d linearly independent vectors in a d-dimensional space form a basis for that space.

The reason for considering the matrix $\mathbf{L}_q$ in (2.3) is because in practice we do not know which values of k for which $\mathbf{\Sigma}_x(k)$ is full ranked. Thus by using $\mathbf{L}_q$ we effectively aggregate the information over q different lags. Also note that

$\mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)'$ may be replaced by $\mathbf{\Sigma}_y(k)$ in all of the discussion in the preceding paragraph. However it does not necessarily hold that $\sum_{k=1}^{q} \mathbf{\Sigma}_y(k)$ has exactly d non-zero eigenvalues due to the fact that the matrices $\mathbf{\Sigma}_y(k)$ are not non-negative definite.

Remark 2.1 (i) A1 is relaxed in Pan & Yao (2008) to allow for correlations between $\mathbf{X}_s$ and $\boldsymbol{\varepsilon}_t$. Their methodology relies on estimating $\mathcal{M}^{\perp} = \text{span}\{\mathbf{e}_j \in \mathbb{R}^p : \mathbf{e}_j'\mathbf{f}_i = 0 \, \forall \, \mathbf{f}_i \in \mathcal{M}\}$ via a stepwise expansion algorithm. However, the minimization problem implicit to their algorithm is p dimensional at each step. Thus if p is large the methodology developed there may not be computationally feasible.

(ii) The condition that $\text{rank}\{\mathbf{\Sigma}_x(k)\} = d$ for some $k \geq 1$ holds without loss of generality since if this was not the the case then the parts of $\mathbf{X}_t$ without any serial correlation should be confounded into $\boldsymbol{\varepsilon}_t$.

In light of the arguments above, we may estimate $\mathcal{M}$ as follows. Let $\widehat{\mathbf{\Sigma}}_y(k) = (n-k)^{-1} \sum_{t=1}^{n-k} \mathbf{Y}_t \mathbf{Y}_{t+k}'$, then we define our estimator of $\mathbf{L}_q$ by

$$\widehat{\mathbf{L}}_q = \sum_{k=1}^{q} \widehat{\mathbf{\Sigma}}_y(k) \widehat{\mathbf{\Sigma}}_y(k)'. \tag{2.5}$$

Now our estimator of d is the number of "non-zero" eigenvalues of $\widehat{\mathbf{L}}_q$ (see Section 2.2.3 below) and our estimator of $\mathcal{M}$ is the linear space spanned by the corresponding eigenvectors. Note that both $\mathbf{L}_q$ in (2.3) and $\widehat{\mathbf{L}}_q$ in (2.5) are non-negative definite, thus all of their eigenvalues will be greater than or equal to zero.

2.2.3 White noise test for d

Let $\lambda_1 \geq \cdots \geq \lambda_p \geq 0$ denote the ordered eigenvalues of $\mathbf{L}_q$ in (2.3) and recall that under identifiability conditions A1 and A2, it holds that $\text{rank}\{\mathbf{L}_q\} = d$, that is $\lambda_j = 0$ for all $j \geq d+1$; see the discussion in Section 2.2.2. Thus if $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p$ denote the eigenvalues of our estimate $\widehat{\mathbf{L}}_q$ in (2.5), then a simple method of determining d would be to plot these ordered eigenvalues, $\widehat{\lambda}_j$, and decide how many of them are "large". Formally, for some threshold ε decided from graphically observing the estimated eigenvalues, we may estimate d by $\#\{j :$

$\widehat{\lambda}_j \geq \varepsilon\}$. Provided ε tends to zero at an appropriate rate, Theorem 2.3 shows that such a method for estimating d is consistent. Further theoretical evidence in favor of this method is supported by the striking result in Theorem 2.2 which suggests that the convergence rates of the eigenvalues $\widehat{\lambda}_j$ for $j \geq d+1$ is n^{-1} when an ideal set of conditions are satisfied; i.e. the estimated eigenvalues whose population counterparts are truly zero are "super-consistent". Some intuition behind this result is given in Section 2.3 and simulation evidence is provided in Section 2.4.

A more useful data driven method of estimating d is given by the following "white noise test". Suppose that we are interested in testing the hypothesis

$$H_0 : d = d_0, \qquad 1 \leq d_0 \leq p. \tag{2.6}$$

Then since $\text{rank}\{\mathbf{L}_q\} = d$, testing the hypothesis in (2.6) is equivalent to testing how many non-zero eigenvalues the matrix $\mathbf{L}_q$ in (2.5) has. Therefore, we may transfer our attention to testing

$$H_0 : \lambda_{d_0+1} = 0, \tag{2.7}$$

then it turns out that testing the hypothesis in (2.7) is in fact quite simple based upon the proceeding argument. Let $\mathbf{f} \in \mathcal{M}^{\perp}$. Then it holds that $\mathbf{Y}_t'\mathbf{f} = \mathbf{X}_t'\mathbf{A}'\mathbf{f} + \boldsymbol{\varepsilon}_t'\mathbf{f} = \boldsymbol{\varepsilon}_t'\mathbf{f}$ which is simply a scalar white noise; see (2.1), condition A1 and recall that $\mathcal{M}$ is the linear space spanned by the columns of $\mathbf{A}$ so that $\mathbf{A}'\mathbf{f} = 0$ for any $\mathbf{f} \in \mathcal{M}^{\perp}$. Now let $\widehat{\mathbf{e}}_{d_0+1}$ be the eigenvector of $\widehat{\mathbf{L}}_q$ corresponding to the eigenvalue $\widehat{\theta}_{d_0+1}$. Then under H_0 in (2.7), Theorem 2.1 implies that $\widehat{\mathbf{e}}_{d_0+1} \xrightarrow{p} \mathbf{f}_{d_0+1} \in \mathcal{M}^{\perp}$ and thus $Z_t = \mathbf{Y}_t'\widehat{\mathbf{e}}_{d_0+1} \xrightarrow{p} \mathbf{Y}_t'\mathbf{f}_{d_0+1}$, which is a white noise sequence. Hence we may reject the hypothesis that $\lambda_{d_0+1} = 0$ if the Ljung-Box-Pierce statistic

$$n(n+2)\sum_{k=1}^{k_0} \rho_k^2/(n-k), \tag{2.8}$$

is greater than the upper α percentage point of the $\chi^2_{k_0}$ distribution. In (2.8), ρ_k is the sample correlation coefficient of Z_t at lag k and k_0 is a prescribed integer. Effectively, we reject the hypothesis that $\lambda_{d_0+1} = 0$, if the series Z_t possesses a significant amount of autocorrelation.

In practice, we may decide upon the value of d_0 by graphical observation of the estimated eigenvalues, $\widehat{\lambda}_j$. However, determination of d may also be done in a data driven way by performing the white noise test sequentially until we find an integer i such that $\lambda_i \neq 0$ and $\lambda_{i+1} = 0$.

2.2.4 Modeling via common factors

Let $\widetilde{\mathbf{A}} \in \mathbb{R}^{p \times \widehat{d}}$ be the matrix whose columns are formed by the eigenvectors of $\widehat{\mathbf{L}}_q$ corresponding to the $\widehat{d}$ largest eigenvalues. Then from (2.1) we have

$$\mathbf{Y}_t = \widetilde{\mathbf{A}}\widetilde{\mathbf{X}}_t + \widetilde{\boldsymbol{\varepsilon}}_t, \tag{2.9}$$

where

$$\widetilde{\mathbf{X}}_t = \widetilde{\mathbf{A}}'\mathbf{Y}_t = \widetilde{\mathbf{A}}'\mathbf{A}\mathbf{X}_t + \widetilde{\mathbf{A}}'\boldsymbol{\varepsilon}_t, \qquad \widetilde{\boldsymbol{\varepsilon}}_t = \left(\mathbf{I}_p - \widetilde{\mathbf{A}}\widetilde{\mathbf{A}}'\right)\mathbf{Y}_t,$$

with $\mathbf{I}_p$ denoting the $p \times p$ identity matrix. Now in order to model the dynamic behavior of the $p \times 1$ time series $\mathbf{Y}_t$, we only need to model the $\widehat{d} \times 1$ process $\widetilde{\mathbf{X}}_t$. Thus if $\widehat{d} << p$ our task is substantially simplified.

Let $\mathbf{A}^* \in \mathbb{R}^{p \times d}$ be the matrix whose columns are the eigenvectors of $\mathbf{L}_q$ corresponding to the d non-zero eigenvalues. Then although the linear space spanned by the columns of $\widetilde{\mathbf{A}}$ is consistent for the linear space spanned by the columns of $\mathbf{A}$ (see (2.12)), $\widetilde{\mathbf{A}}$ is only an estimator of the particular rotation $\mathbf{A}^*$. In the same spirit, $\widetilde{\mathbf{X}}_t$ is then an estimate of $\mathbf{X}_t^* = \mathbf{A}^{*\prime}\mathbf{Y}_t$. Thus, analogously to our remarks earlier, for any invertible matrix $\mathbf{H} \in \mathbb{R}^{\widehat{d} \times \widehat{d}}$, $\widetilde{\mathbf{A}}$ and $\widetilde{\mathbf{X}}_t$ may be replaced by $\widetilde{\mathbf{A}}\mathbf{H}$ and $\mathbf{H}^{-1}\widetilde{\mathbf{X}}_t$ without altering (2.9). This indeterminism may be exploited to further simplify our analysis by choosing the rotation matrix $\mathbf{H}$ so that a more parsimonious model is obtained for the estimated common factors; see Tiao & Tsay (1989).

The data $\mathbf{X}_t^*$ may be thought of as the "oracle dataset", i.e. the data about which we would like to make inferences. However, we note that there are some complications that are worth mentioning when using $\widetilde{\mathbf{X}}_t$ to make inferences about $\mathbf{X}_t^*$. To see this, we consider the following simple example. Suppose that $\mathbf{X}_t^*$ follows the VAR(τ) process

$$\mathbf{X}_t^* = \sum_{k=1}^{\tau} \mathbf{C}_k \mathbf{X}_{t-k}^* + \mathbf{U}_t, \tag{2.10}$$

where $\mathbf{C}_k$ are $d \times d$ coefficient matrices and $\mathbf{U}_t$ is a $d \times 1$ white noise. Put $\mathcal{C}_\tau = (\mathbf{C}_1, \ldots, \mathbf{C}_\tau) \in \mathbb{R}^{d^2 \times d}$. Then for a fixed τ, the least squares estimator of $\mathcal{C}_\tau$ is given by

$$\widehat{\mathcal{C}}_\tau = \arg\min_{\mathcal{C}_\tau} \sum_{t=\tau+1}^{n} \|\mathbf{X}_t^* - \sum_{j=1}^{\tau} \mathbf{C}_k \mathbf{X}_{t-k}^*\|^2. \tag{2.11}$$

where $\|\mathbf{H}\| = \mathrm{tr}(\{\mathbf{H}'\mathbf{H}\})^{1/2}$ for any $\mathbf{H} \in \mathbb{R}^{a \times b}$, $a, b \geq 1$. Now if the characteristic polynomial of the VAR model in (2.10) has no roots on or inside the complex unit circle and $E\|\mathbf{U}_t\|^4 < \infty$, it holds that $n^{1/2}(\widehat{\mathcal{C}}_\tau - \mathcal{C}_\tau)$ converges in distribution to a zero mean Gaussian random variable (see Proposition 3.1 in Lutkepohl (2005)). However, in practice we do not have access to the oracle data, thus we may replace $\mathbf{X}_t^*$ in (2.11) with $\widetilde{\mathbf{X}}_t$ and denote the resulting estimate by $\widetilde{\mathcal{C}}_\tau$. Now provided that $\sup_t \|\mathbf{Y}_t\| < \infty$ and d is known (thus we set $\widehat{d} = d$), $\sup_t \|\mathbf{X}_t^* - \widetilde{\mathbf{X}}_t\| = O_p(n^{-1/2})$ from which some straightforward calculations lead to $\|\widehat{\mathcal{C}}_\tau - \widetilde{\mathcal{C}}_\tau\| = O_p(n^{-1/2})$. Here is where the problem lies, in that although the preceding arguments imply $\|\widetilde{\mathcal{C}}_\tau - \mathcal{C}_\tau\| = O_p(n^{-1/2})$, $\widetilde{\mathcal{C}}_\tau$ is not first order asymptotically equivalent to $\widehat{\mathcal{C}}_\tau$. In fact, the limiting distribution of $n^{1/2}(\widetilde{\mathcal{C}}_\tau - \mathcal{C}_\tau)$ will have an inflated variance due to the additional estimation error in approximating $\mathbf{X}_t^*$ with $\widetilde{\mathbf{X}}_t$. Thus standard asymptotic distribution approximations for least squares estimators of VAR models would not be valid in this context. An example of where this may cause a problem would be when constructing forecast confidence intervals of $\mathbf{X}_t^*$ using $\widetilde{\mathbf{X}}_t$ as a proxy since the traditional intervals would be too narrow. We remark that in this setting, tailor made bootstrap methods should be employed to suit the task at hand. However, if we are only interested in point estimates of $\mathcal{C}_\tau$, then the approximation $\widetilde{\mathcal{C}}_\tau$ is fine. Of course, the problem is still there when d is unknown and is actually even more complicated since it may hold that $\widehat{d} \neq d$ in which case the difference $\widehat{\mathcal{C}}_\tau - \widetilde{\mathcal{C}}_\tau$ will not be defined.

2.3 Theoretical results

We first solidify some notation. Let $(\lambda_j, \mathbf{e}_j)$ (resp. $(\widehat{\lambda}_j, \widehat{\mathbf{e}}_j)$ form (eigenvalue, eigenvector) pairs of $\mathbf{L}_q$ (resp. $\widehat{\mathbf{L}}_q$). Note that $\lambda_j = 0$ for all $j \geq d+1$. With d known, we may set $\widehat{\mathcal{M}} = \mathrm{span}\{\widehat{\mathbf{e}}_1, \ldots, \widehat{\mathbf{e}}_d\}$. Let $\|\mathbf{H}\| = (\mathrm{tr}\{\mathbf{H}'\mathbf{H}\})^{1/2}$ be the norm of any vector or matrix $\mathbf{H}$. Now if $\mathcal{X}$ is a subspace of $\mathbb{R}^p$, we will use $\mathbf{P}_\mathcal{X}$

to denote its orthonormal projection matrix, i.e. if $\mathbf{X} \in \mathbb{R}^{p \times d}$ is a matrix whose columns form an orthonormal basis of $\mathcal{X}$ then we may set $\mathbf{P}_{\mathcal{X}} = \mathbf{X}\mathbf{X}'$.

Throughout the presentation in this section, we will always assume that d (the number of common factors), q (the number of lags taken in defining $\mathbf{L}_q$) and p (the dimension of $\mathbf{Y}_t$) are fixed integers. The first two are reasonable conditions and the third was relaxed in Bai (2003), Lam & Yao (2009) and Fan *et al.* (2008), to name a few, in order to allow for the modern "big p" asymptotic setting in deriving the properties of the estimated factor loading space, $\widehat{\mathcal{M}}$.

Some regularity conditions are now in order.

C1. $\lambda_1 > \cdots > \lambda_d > 0$.

C2. For $k = 1, \ldots, q$ it holds that and $\|\boldsymbol{\Sigma}_y(k)\| = O(1)$ and $\|\widehat{\boldsymbol{\Sigma}}_y(k) - \boldsymbol{\Sigma}_y(k)\| = O_p(n^{-\gamma})$ for some $\gamma > 0$.

Remark 2.2 (i) The condition that the eigenvalues of $\mathbf{L}_q$ are all different is made in order to simplify the proofs of the theoretical results given in this section. Straightforward (but tedious) adjustments can be made if C1 is not met.

(ii) Condition C2 is satisfied by causal linear processes as well as many well studied weakly dependent processes. In these cases, the rate is likely to be $\gamma = 1/2$. If the factors $\mathbf{X}_t$ exhibit long range dependence, then it is likely that $\gamma < 1/2$. In particular note that $\mathbf{Y}_t$ does not need to be stationary in order for C2 to hold. For example, it may hold when $E\widehat{\boldsymbol{\Sigma}}_y(k) \longrightarrow \boldsymbol{\Sigma}_y(k)$ and $\mathbf{Y}_t$ is a non-stationary ψ-mixing process; see Zhengyan & Lu (1997). Examples of non-stationary ψ-mixing processes include, amongst others, stationary ψ-mixing processes plus non-constant trends and standardized random walks. However, we note that common types of structured non-stationarity (such as unit roots) are not permitted in this setting. See Lam & Yao (2009) and references therein for dealing with these types of processes.

Now under conditions A1, A2, C1 and C2, an identical set of manipulations to those used in the proof of Theorem 1.2 in Chapter 1 may be used to show that

$$\|\mathbf{P}_{\widehat{\mathcal{M}}} - \mathbf{P}_{\mathcal{M}}\| = O_p(n^{-\gamma}), \tag{2.12}$$

i.e $\widehat{\mathcal{M}}$ is converges in probability to $\mathcal{M}$ with rate $n^{-\gamma}$. Furthermore if $\widetilde{\mathcal{M}} = \text{span}\{\widehat{\mathbf{e}}_1, \ldots, \widehat{\mathbf{e}}_{\widehat{d}}\}$ where $\widehat{d}$ is some estimator of d satisfying $P(\widehat{d} \neq d) = o(1)$ then it holds that

$$\|\mathbf{P}_{\widehat{\mathcal{M}}} - \mathbf{P}_{\widetilde{\mathcal{M}}}\| = o_p(n^{-\gamma}), \tag{2.13}$$

i.e. our estimation of $\mathcal{M}$ is adaptive to d. The next few results provide a theoretical basis on how we may estimate d and justify much of the discussion in Section 2.2.3.

Theorem 2.1 *Let $\mathbf{P}_{\mathcal{M}}$ (resp. $\mathbf{P}_{\mathcal{M}^\perp}$) be the projection matrix onto $\mathcal{M}$ (resp. $\mathcal{M}^\perp$). Suppose that conditions A1, A2, C1 and C2 hold. Then under H_0 in (2.7), it holds that*

$$\|\mathbf{P}_{\mathcal{M}}\widehat{\mathbf{e}}_{d_0+1}\| = \|\widehat{\mathbf{e}}_{d_0+1} - \mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1}\| = O_p(n^{-\gamma}).$$

Furthermore, if $\|\mathbf{Y}_t\| = O_p(1)$ for all $t \geq 1$, then it holds that $|\mathbf{Y}_t'\widehat{\mathbf{e}}_{d_0+1} - \mathbf{Y}_t'\mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1}| = O_p(n^{-\gamma})$.

Since $\mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1} \in \mathcal{M}^\perp$, it follows from Theorem 2.1 that under H_0 in (2.7), the series $\mathbf{Y}_t'\widehat{\mathbf{e}}_{d_0+1}$ converges in probability to a white noise sequence; see the discussion of the white noise test in Section 2.2.3. Analogously, (2.12) implies that under H_0 in (2.7), the series $\mathbf{Y}_t'\widehat{\mathbf{e}}_{d_0}$ converges in probability to a sequence with significant autocorrelation. However, we note that some caution should be exercised when using the white noise test in practice since the $\mathbf{Y}_t'\widehat{\mathbf{e}}_{d_0+1}$ is only asymptotically a white noise series under H_0. Here is where the problem lies in that the χ^2 approximation to the Ljung-Box-Pierce portmanteau statistic in (2.8) requires that the series under question is exactly white noise under the null. Indeed, some simulation results presented in Section 2.4 suggest that the white noise test for the null that an eigenvalue is equal to zero has a larger than desired false positive rate. However, when the null is false and provided the autocorrelation in the latent factors $\mathbf{X}_t$ is not too weak, the test performs extremely well.

Our next result concerns the convergence rates of the empirical eigenvalues of $\mathbf{L}_q$. To this end, we require some more specific regularity conditions than those given in C1 and C2 above. Let $\mathbf{Z}_t(k) = \mathbf{Y}_t\mathbf{Y}_{t+k}'$. For $k = 1, \ldots, q$ the required regularity conditions on $\mathbf{Z}_t(k)$ are stated below.

C3. $\{\mathbf{Z}_t(k)\}$ is strictly stationary and β-mixing with β-mixing coefficients given by

$$\beta(l) = \sup_{A \in \mathcal{F}_l^\infty} E\left\{|P(A|\mathcal{F}_{-\infty}^0) - P(A)\right\},$$

where $\mathcal{F}_l^m = \sigma\{\mathbf{Z}_l(k), \ldots, \mathbf{Z}_m(k)\}$ and $\beta(l) = O(l^{-(2+\delta')/\delta'})$ for some $\delta' > 0$.

C4. It holds that $\sup_{ij} E\left\{\|\mathbf{Z}_i(k)\mathbf{Z}_j(k)'\|^{2+\delta}\right\} < \infty$ and

$$\int \|\mathbf{Z}_i(k)\mathbf{Z}_j(k)\|^{2+\delta} P_k(d\mathbf{Z}_i(k)) P_k(d\mathbf{Z}_j(k)) < \infty,$$

where $\delta > \delta'$ given in A1 above and P_k is the distribution function of $\mathbf{Z}_t(k)$.

We note that A1, A2, C1, C3 and C4 together imply that $\gamma = 1/2$ in both (2.12) and (2.13).

Theorem 2.2 *Let conditions A1, A2, C3 and C4 hold. (i) Then for $i = 1, \ldots, d$ it holds that $|\widehat{\lambda}_j - \lambda_j| = O_p(n^{-1/2})$. (ii) If in addition C1 is satisfied, then for $j \geq d+1$ it holds that $\widehat{\lambda}_j = O_p(n^{(-1-\gamma)/2})$ where $\gamma = \min\left\{1, \frac{2(\delta-\delta')}{\delta'(2+\delta)}\right\}$.*

From Theorem 2.2 it is clear that the convergence rate for $\widehat{\lambda}_j$ is strictly faster than $n^{-1/2}$ when $j \geq d+1$. This is perhaps surprising and deserves some explanation. Consider the following simple example. Let $A_1, \ldots, A_n$ be a sample scalar random variables. Suppose we are interested in estimating $\mu^2 = E(A_i)^2$ for which we use the estimator $\bar{A}^2 = (n^{-1}\sum_{t=1}^n A_t)^2 = n^{-2}\sum_{s,t=1}^n A_s A_t$. Then under an appropriate set of regularity conditions, it follows that for some constant $c \in (0, 1]$

$$|\bar{A}^2 - \mu^2| \leq |\mu||\bar{A} - \mu| + |\bar{A}^2 - \bar{A}\mu| = |\mu| \cdot O_p(n^{-1/2}) + O_p(n^{(-1-c)/2}), \quad (2.14)$$

since $|\bar{A} - \mu| = O_p(n^{-1/2})$ by an according central limit theorem and $|\bar{A}^2 - \bar{A}\mu| = O_p(n^{(-1-c)/2})$ by a simple U-statistic argument; see Lee (1990). Thus from (2.14) it is clear that $|\bar{A}^2 - \mu^2| = O_p(n^{-1/2})$ if $\mu \neq 0$ and $|\bar{A}^2 - \mu^2| = O_p(n^{(-1-c)/2})$ if $\mu = 0$. This is precisely why the rate for $\widehat{\lambda}_j$ is faster than $n^{-1/2}$ when $j \geq d+1$; the matrix $\widehat{\mathbf{L}}_q = \sum_{k=1}^q \widehat{\boldsymbol{\Sigma}}_y(k)\widehat{\boldsymbol{\Sigma}}_y(k)' = \sum_{k=1}^q (n-k)^{-2}\sum_{s,t} \mathbf{Z}_s(k)\mathbf{Z}_t(k)'$ can be thought of as being similar to $\bar{A}^2$ in the example just given. The eigenvalues $\widehat{\lambda}_j$ are then obtained via pointwise evaluation of $\widehat{\mathbf{L}}_q$. Thus when the true eigenvalues

$\lambda_j \neq 0$, as in part (i) of Theorem 2.2, the convergence rate is the ordinary $n^{-1/2}$ where as when $\lambda_j = 0$, as in part (ii) of Theorem 2.2, the rate is faster that $n^{-1/2}$.

Note that when $j \geq d+1$, the optimal rate of convergence for $\widehat{\lambda}_j$ is n^{-1}. This is known as "super-consistency" and is obtained when, for example, $\delta \geq 4\delta'/(2-\delta')$ or the condition on the mixing coefficients is strengthened to satisfy $\sum_{k=1}^{\infty} k\beta(k)^{\delta/(2+\delta)}$. In terms of inference, Theorem 2.2 has a dramatic impact since it means that the eigenvalues whose population counterparts are truly zero converge to zero very quickly; see also the simulation results in Section 2.4. This property makes it easier to identify the number of latent factors d.

Remark 2.3 Some key technical tools used in the proof of Theorem 2.2 are the U-Statistic results of Yoshihara (1976) which require the condition that $\{\mathbf{Z}_t(k)\}$ is strictly stationary. However, the results of Yoshihara (1976) can be substituted for those of Harel & Puri (1989) which don't actually require stationarity. In that case, the proof is essentially the same except the regularity conditions become much more cumbersome. For that reason, we have decided not to state the results for the non-stationary case here.

We conclude this section with an example of an estimator of d which satisfies the conditions required for the adaptivity property in (2.13).

Theorem 2.3 *Let* $\widehat{d} = \#\{j : \widehat{\lambda}_j \geq \varepsilon\}$ *for some* $\varepsilon \geq 0$. *Suppose that conditions A1, A2, C1, C3 and C4 hold. Let* $\varepsilon = \varepsilon(n) \longrightarrow 0$ *as* $n \longrightarrow \infty$. *Then* $P(\widehat{d} \neq d) = O((\varepsilon^2 n)^{-1})$ *and thus* $\widehat{d} \xrightarrow{p} d$ *provided* $\varepsilon^2 n \longrightarrow 0$.

In practice, one may select the threshold ε via cross-validation by setting our objective function to be the τ step ahead forecast error, $\tau \geq 1$. However, we note that the primary reasons for presenting the result of Theorem 2.3 is to provide a justification of the heuristic method of estimating d by setting it equal to the number of "large" eigenvalues of $\widehat{\mathbf{L}}_q$ as well as to show that it is possible to estimate d consistently. Indeed for practical purposes, the white noise test in Section 2.2.3 is perhaps more useful.

2.4 Simulation studies

We simulate data from the following one factor model

$$\mathbf{Y}_t = \mathbf{a}X_t + \boldsymbol{\varepsilon}_t \qquad t = 1, \ldots, n, \tag{2.15}$$

with $\mathbf{a} = (1, 0, \ldots, 0)' \in \mathbb{R}^p$, $X_t = \gamma X_{t-1} + \eta_t$ where $\eta_t \sim iN(0,1)$ and $\boldsymbol{\varepsilon}_t \sim iN(0, \mathbf{I}_p)$. Note that for the model in (2.15) $d = \dim(\mathcal{M}) = 1$. For all of the simulations in this section, we set $p = 100$. However, similar results to those that we will present here can be obtained for any fixed p. We do not pursue simulations results about the estimation error of $\mathcal{M}$, $D(\widehat{\mathcal{M}}, \mathcal{M})$ since those are similar to the results presented in Chapter 1.

Our first set of simulation results concerns the convergence rates of the empirical eigenvalues. To this end we simulate 1,000 replications from model (2.15) for various values of n and set $\gamma = 1/\sqrt{2}$ in simulating the AR(1) model for X_t in (2.15) and $q = 1$ in defining the matrix $\widehat{\mathbf{L}}_q$ in (2.5). Note that for this setting, the population analogue of $\widehat{\mathbf{L}}_q$, $\mathbf{L}_q$ in (2.3), has a single non-zero eigenvalue equal to 2, i.e. $\lambda_1 = 2$ and $\lambda_j = 0$ for $j \geq 2$; this can be computed analytically without much difficulty. Of course our estimate $\widehat{\mathbf{L}}_q$ will have many more than one non-zero eigenvalue $\widehat{\lambda}_j$, $j \geq 1$. Furthermore, there is little point in taking $q > 1$ for this example since the identifiability condition in A2 is met for $q = 1$. In terms of convergence rates, no harm would be done if we were to take $q > 1$ (for any fixed q) since for each k, the lag k sample autocovariance matrices of $\mathbf{Y}_t$ will be root-n consistent for it's population analogue. However, for larger choices of q we would effectively be accumulating estimation error and thus increasing the constant in the limiting variance.

Figure 2.1 displays boxplots of the absolute estimation errors of the eigenvalues. The errors in estimating the non-zero eigenvalue are considerably greater that those in estimating the zero eigenvalue's; note that $\widehat{\lambda}_2 \geq \widehat{\lambda}_j$ for all $j \geq 3$. In particular, the rate of decay in the estimation error of the non-zero eigenvalue is far slower than that of the zero eigenvalues.

Figure 2.2 shows the distribution of $\sqrt{n}(\widehat{\lambda}_1 - \lambda_1)$ over the 1,000 iterations. As suggested by Theorem 2.2, standardizing by a factor of $\sqrt{n}$ leads to a non-degenerate limiting distribution for the non-zero eigenvalue estimators. In particular, the distributions begin to look Gaussian at samples sizes of around $n = 200$.

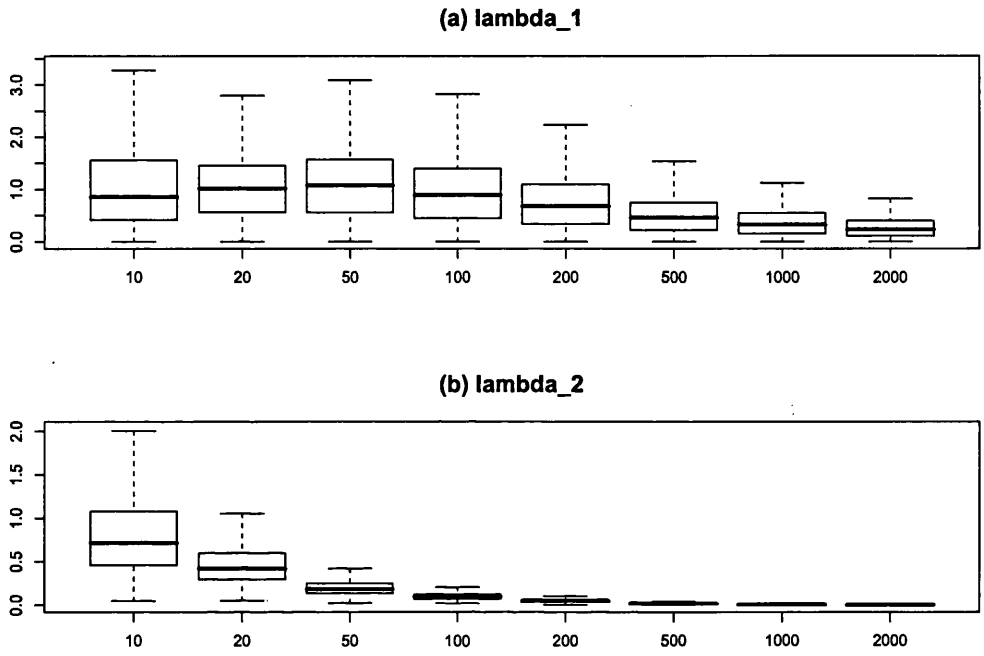


Figure 2.1: *Boxplots of estimation errors of eigenvalues; (a) non-zero eigenvalue estimation error* $|\widehat{\lambda}_1 - \lambda_1|$ *and (b) zero eigenvalue estimation error* $\widehat{\lambda}_2$. *To add clarity to the display, the outliers are not plotted.*

Figure 2.3 on the other hand, displays the distribution of $\sqrt{n}(\widehat{\lambda}_2 - \lambda_2) = \sqrt{n}\widehat{\lambda}_2$, which is an estimate of a zero population eigenvalue. As suggested by Theorem 2.2, the distribution of $\sqrt{n}\widehat{\lambda}_2$ converges to zero as n increases. The rate $\sqrt{n}$ is clearly not optimal here. In light of this, Figure 2.4 shows the distribution of $n\widehat{\lambda}_2$. In accordance to our theory, the standardizing factor n appears to be optimal. A striking feature of this graphic is how stable the distribution appears across different sample sizes; for example, the distribution for $n = 10$ looks almost identical to the distribution for $n = 2000$.

Our next study involves the performance of the white noise test in Section 2.2.3. The simulation design is identical to the one already described. Table 2.1 presents the means and standard deviations of the P-values from testing

the hypothesis $H_0 : \lambda_i = 0$ $(i = 1, 2)$. It appears that the white noise test is performing well. The false null $H_0 : \lambda_1 = 0$ is routinely rejected even for moderate sample sizes of around $n = 50$ and the true null $H_0 : \lambda_2 = 0$ cannot be rejected for samples as small as $n = 10$. Note that together, these these hypotheses are actually equivalent to the null $H_0 : d = 1$. It is not surprising that the conclusions drawn from this test are independent of the number of lags taken in computing the Ljung-Box-Pierce statistic.

Figure 2.5 displays the the simulated power of the white noise test as a function of the autoregressive parameter $\gamma \in (0, 1)$ for testing the false null $H_0 : \lambda_1 = 0$. That is, for a given γ each point on the plotted curve is an estimate of $1 - P(H_0 : \lambda_1 = 0$ is not rejected). In this graphic (and those that follow) we have taken the level of the test to be $\alpha = 0.05$ and the number of lags used in computing the Ljung-Box-Pierce statistic in (2.8) is $k_0 = 5$. The plotted power function is an average over 1000 replications. It is clear from this graphic that for larger values of α, the power converges to 1, i.e. the test appears to be consistent provided there is enough autocorrelation in the latent factor X_t. This finding is quite intuitive since if the correlation in X_t is weak, then we expect that there would be some difficulty in identifying any factors. Another way of thinking about this is that as γ tends to zero, the identifiability condition in A2 is violated, i.e. there are actually no dynamic factors in model (2.15).

Figure 2.6 displays estimates of $P(H_0 : \lambda_2 = 0$ is rejected) as a function of γ. Since the hypothesis $H_0 : \lambda_2 = 0$ is true, we would expect that the plotted curve should fall below the $\alpha = 0.05$ level if the test is unbiased. This is not the case for any $\gamma \in (0, 1)$ which is expected from the results given in Theorem 2.1 since $\mathbf{Y}_t'\widehat{\mathbf{e}}_2$ is only asymptotically a white noise sequence (recall, $\widehat{\mathbf{e}}_j$ is the eigenvector of $\widehat{\mathbf{L}}_q$ in (2.5) corresponding to the j-th largest eigenvalue λ_j). In particular, it appears that the number of rejections increases with larger values of γ. Again, this is intuitive since if the autocorrelation in the factors increases, the amount of autocorrelation in $\mathbf{Y}_t'\widehat{\mathbf{e}}_2$ would also increase due to estimation error.

H_0	$\lambda_1 = 0$		
k_0	1	3	5
$n = 10$	0.35 (0.28)	0.48 (0.27)	0.56 (0.27)
$n = 20$	0.22 (0.26)	0.30 (0.29)	0.36 (0.30)
$n = 50$	0.05 (0.15)	0.07 (0.17)	0.09 (0.19)
$n = 100$	0.00 (0.05)	0.00 (0.04)	0.01 (0.04)
$n = 200$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$n = 500$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$n = 1000$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
$n = 2000$	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
H_0	$\lambda_2 = 0$		
k_0	1	3	5
$n = 10$	0.49 (0.28)	0.55 (0.26)	0.65 (0.25)
$n = 20$	0.40 (0.30)	0.47 (0.29)	0.52 (0.29)
$n = 50$	0.32 (0.30)	0.38 (0.29)	0.42 (0.30)
$n = 100$	0.31 (0.30)	0.36 (0.29)	0.40 (0.30)
$n = 200$	0.31 (0.30)	0.36 (0.30)	0.38 (0.30)
$n = 500$	0.31 (0.30)	0.34 (0.30)	0.36 (0.30)
$n = 1000$	0.32 (0.31)	0.36 (0.31)	0.37 (0.30)
$n = 2000$	0.31 (0.31)	0.34 (0.30)	0.37 (0.30)

Table 2.1: *The means and standard deviations (in parentheses) of the P-values for the white noise test. The parameter k_0 defined is the number of lags used in computing the Ljung-Box-Pierce statistic in (2.8).*

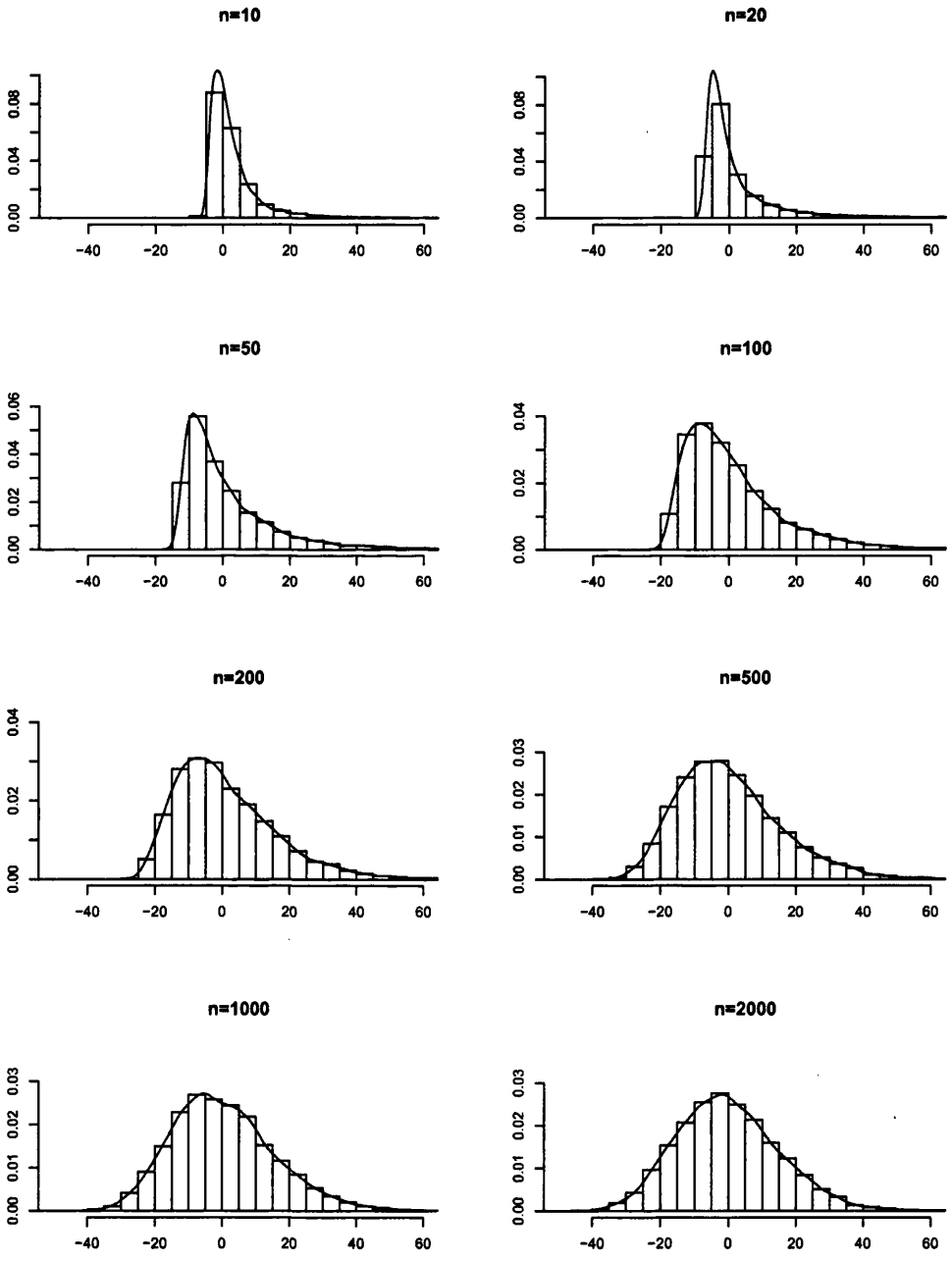


Figure 2.2: *Histograms overlaid by kernel density estimates of* $\sqrt{n}\,(\widehat{\lambda}_1 - \lambda_1)$.

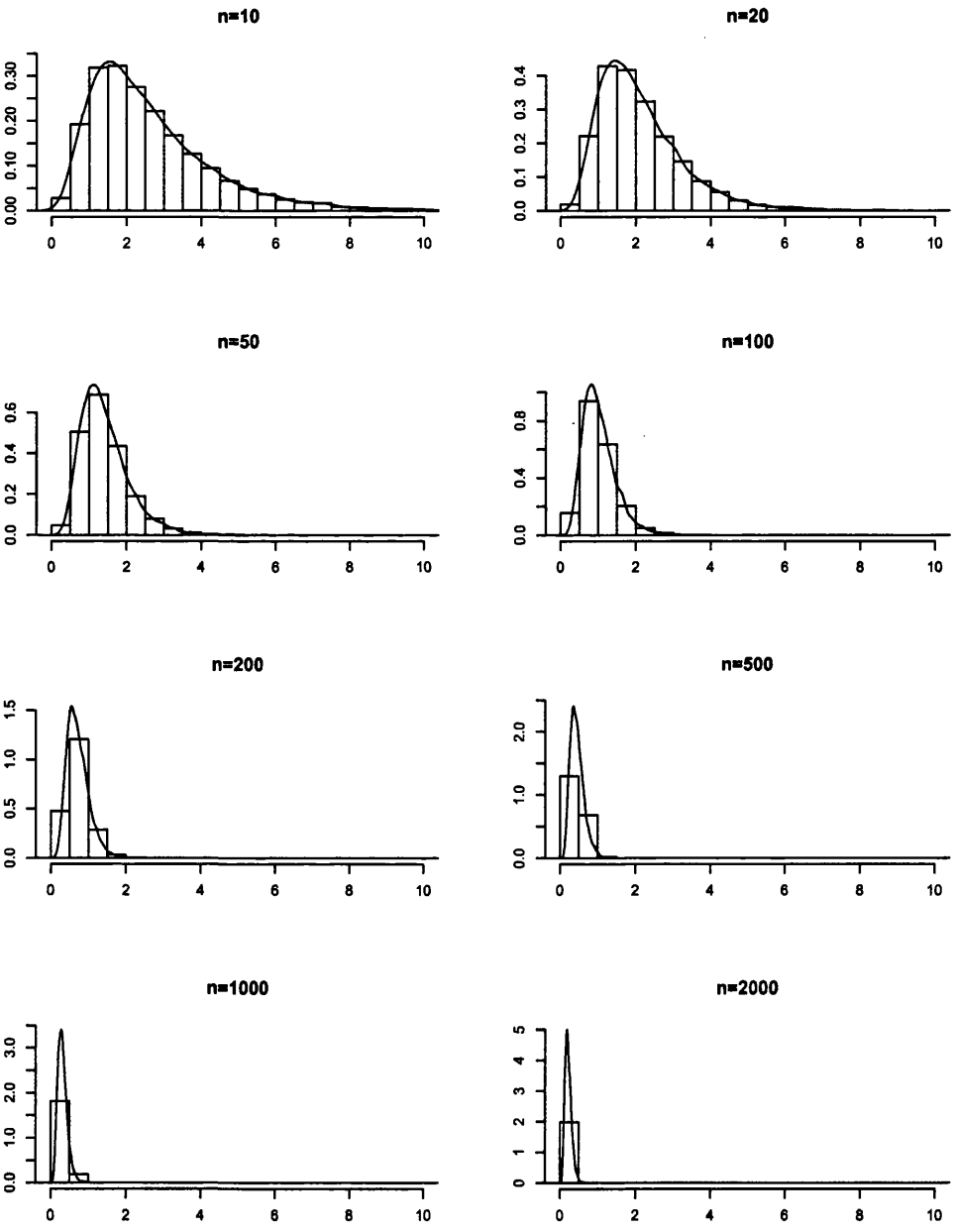


Figure 2.3: *Histograms overlaid by kernel density estimates of* $\sqrt{n}\widehat{\lambda}_2$.

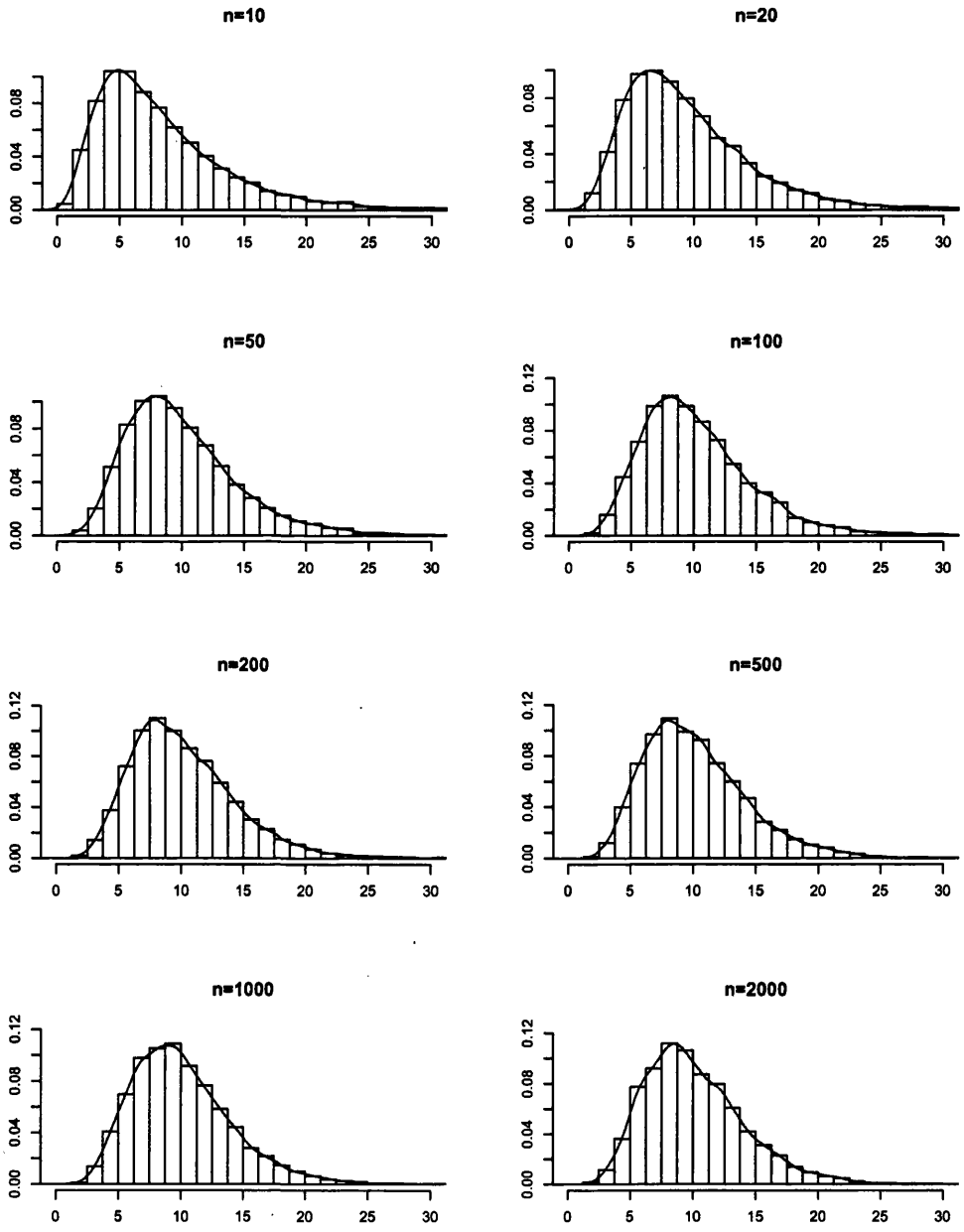


Figure 2.4: *Histograms overlaid by kernel density estimates of* $n\widehat{\lambda}_2$.

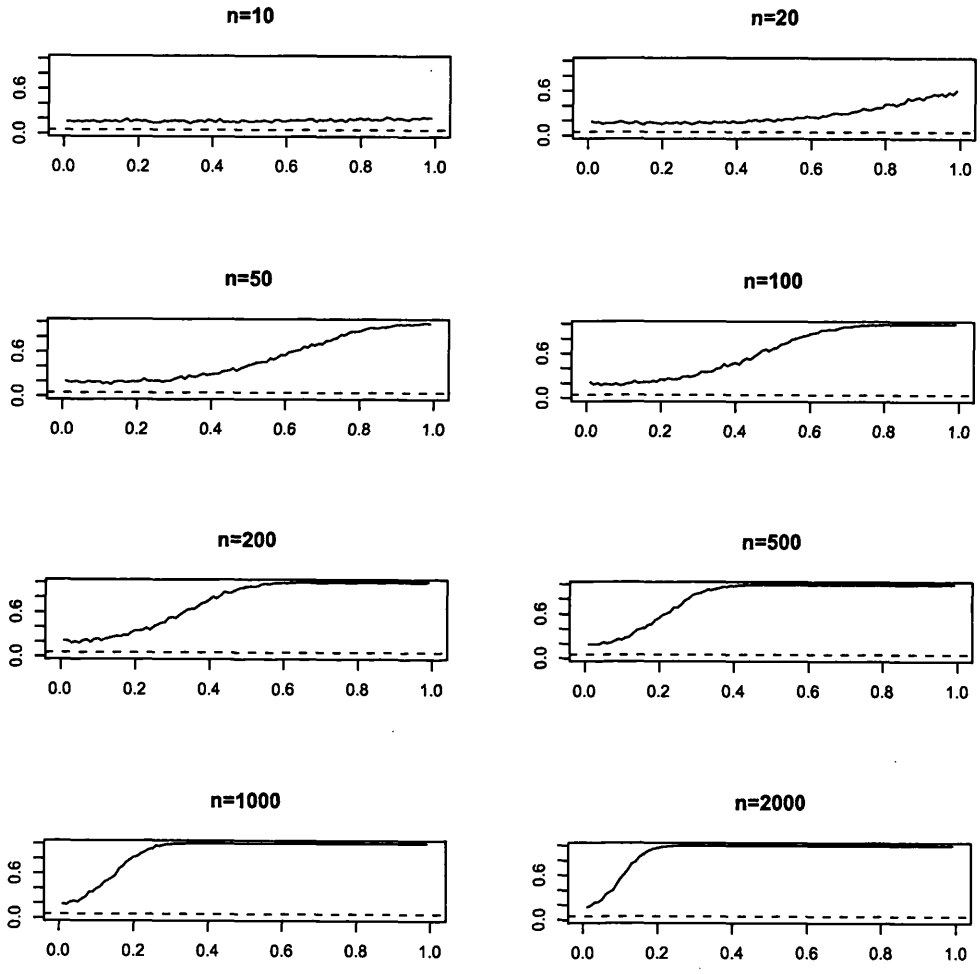


Figure 2.5: *Simulated power* $1 - P(H_0 : \lambda_1 = 0$ *is not rejected) (y-axis) as a function of the autoregressive parameter* $\gamma \in (0, 1]$ *(x-axis). The dashed horizontal line denotes the* $\alpha = 0.05$ *level and the plotted curve is an average over 1000 replications. The number of lags used in computing the Ljung-Box-Pierce statistic in (2.8) is taken to be* $k_0 = 5$.

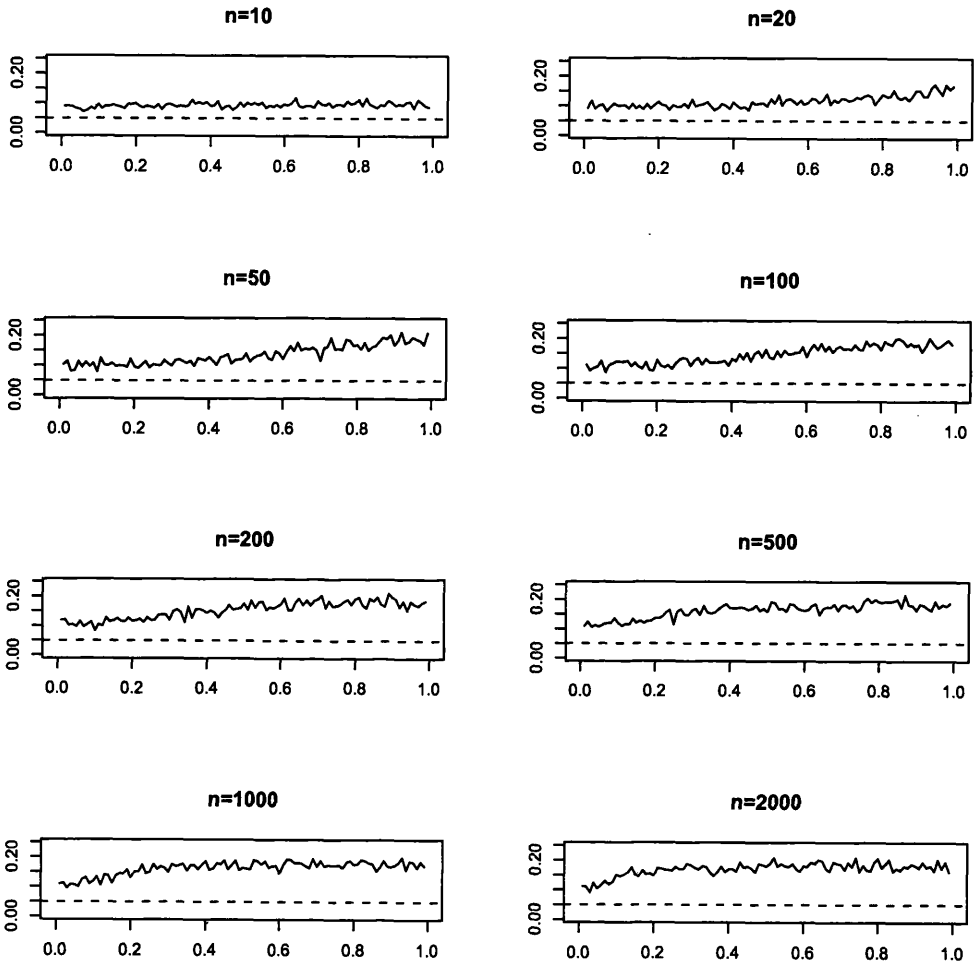


Figure 2.6: *Simulated probabilities* $P(H_0 : \lambda_2 = 0$ *is rejected)* *(y-axis) as a function of the autoregressive parameter* $\gamma \in (0, 1]$ *(x-axis). The dashed horizontal line denotes the* $\alpha = 0.05$ *level and the plotted curve is an average over 1000 replications. The number of lags used in computing the Ljung-Box-Pierce statistic in (2.8) is taken to be* $k_0 = 5$.

2.5 Implied volatility surfaces

To illustrate the methodology developed in this chapter, we set upon the task of modeling the dynamic behavior of IBM, Microsoft and Dell implied volatility surfaces. The implied volatility of an option (typically a "vanilla" call or put) is value of the volatility parameter obtained from inverting the Black-Scholes equation for the observable option price and a set of variables related to the option (typically any two from time to maturity, delta, moneyness or strike). If the Black-Scholes equation was correct, then the the implied volatility would be independent of the values of the parameters that went into its computation. However, this is far from accurate for real market data with many empirical studies demonstating that the profiles of implied volatility surfaces display a "smile" or "skew" (see Cont & da Fonseca (2002), Fengler *et al.* (2007) and Park *et al.* (2009) amongst others). Another prominent feature of implied volatility surfaces is that their level changes from day to day. The evolution in time of this surface captures the evolution of prices in the option market.

Since implied volatilities are linked directly to market prices via the Black-Scholes formula, many investment professional argue that they are better measures of volatility than historical (or "realized") volatility estimators. Implied volatilities are often said to be "forward looking" since they are based on current prices which presumably have expectations of the future contained within them. Thus there is great interest in modeling the dynamic behavior of implied volatility for risk management purposes.

2.5.1 Data description

We begin with a brief description of the dataset under consideration. The data was obtained from OptionMetrics via the WRDS database. The dates in question are 03/01/2006 – 29/12/2006 (250 days in total). For each day t we observe the implied volatility $W_t(u_i, v_j)$ computed from call options as a function of time to maturity of 30, 60, 91, 122, 152, 182, 273, 365, 547 and 730 calendar days which we denote by u_i, $i = 1, \ldots, p_u$ ($p_u = 10$) and deltas of 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 which we denote by v_j, $j = 1, \ldots, p_v$ ($p_v = 13$). 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 2.7 displays the mean volatility surface of IBM, Microsoft and Dell over the period in question. It is clear from this graphic that the implied volatilities surfaces are not flat. Indeed any cross-section in the maturity or delta axis displays the well documented volatility smile.

Now it is a well documented stylized fact that implied volatilities are non-stationary (see the aforementioned references). 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. The P-values from these test are displayed in Figure 2.8. Of course we should treat the results of these tests with some caution since we 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 in the form of (2.1) are then $\mathbf{Y}_t = \text{vec}\{\Delta\mathbf{W}_t\}$, where for any matrix $\mathbf{M} = (\mathbf{m}_1, \ldots, \mathbf{m}_{p_v}) \in \mathbb{R}^{p_u \times p_v}$, $\text{vec}\{\mathbf{M}\} = (\mathbf{m}_1', \ldots, \mathbf{m}_{p_v}')' \in \mathbb{R}^{p_u p_v}$. For convenience, we also introduce the inverse function vec^{-1} which satisfies $\text{vec}^{-1}(\text{vec}\{\mathbf{M}\}) = \mathbf{M}$. Note that $\mathbf{Y}_t$ is now defined over $04/01/2006 - 29/12/2006$ since we lose an observation due to differencing. The effective sample size is thus $n = 249$ and the dimension of $\mathbf{Y}_t$ is $p = p_v \times p_u = 130$.

2.5.2 Estimation results

In forming the matrix $\widehat{\mathbf{L}}_q$ in 2.5 we take $q = 5$. Estimation results for different q are similar and thus not reported here. Figure 2.9 displays the ten largest eigenvalues of $\widehat{\mathbf{L}}_q$ for Dell, IBM and Microsoft. From this graphic it is apparent that there is one eigenvalue that is much larger than the others. Indeed, the white noise test, whose results are summarized in Table 2.2, also suggests that we should reject the null $\lambda_1 = 0$ but we cannot reject $\lambda_2 = 0$. Thus we set $\widehat{d} = 1$ for all three companies.

Figure 2.10 displays the estimated factor loading surfaces $\text{vec}^{-1}\{\widehat{\mathbf{e}}_1\}$. A common feature is evident in the factor loading surface for all three companies in that it is relatively flat apart from at long maturities and small delta's where there is a large spike. This discovery has a simple explanation based on market activity.

H_0	$\lambda_1 = 0$			$\lambda_2 = 0$		
k_0	1	3	5	1	3	5
Dell	0.00	0.00	0.00	0.54	0.32	0.39
IBM	0.00	0.00	0.01	0.57	0.83	0.81
Microsoft	0.00	0.00	0.00	0.34	0.57	0.64

Table 2.2: *P-values from testing the null* $H_0 : Z_{t,k,1:200}$ *is a white noise sequence. The parameter* k_0 *defined is the number of lags used in computing the Ljung-Box-Pierce statistic in (2.8).*

At longer maturities and smaller delta's, there are fewer market participants so a single transaction can cause a large movement in the option price which in turn induces a large change in the implied volatility. The interaction of these two variables is what produces the very pronounced spike in Figure 2.10.

Figure 2.11 displays a scatterplots of the stock returns (which we denote by R_t, i.e. R_1 is the return on 04/01/2006) against the estimated common factor $Z_{t1} = \mathbf{Y}_t'\widehat{\mathbf{e}}_1$. It is clear that there is a strong linear relationship between R_t and Z_{t1} for all three companies. Indeed, when regressing R_t on Z_{t1} the slope parameter is highly significant. This opens up the opportunity of a statistical arbitrage since there is a great deal of linear predictability in Z_{t1} (see Table 2.2). Thus we may forecast Z_{t1} using a model of our choice (for example an ARMA model) and consequently take positions on the IBM stock based on this forecast. Further empirical studies are required to determine whether or not this is a common feature in other datasets.

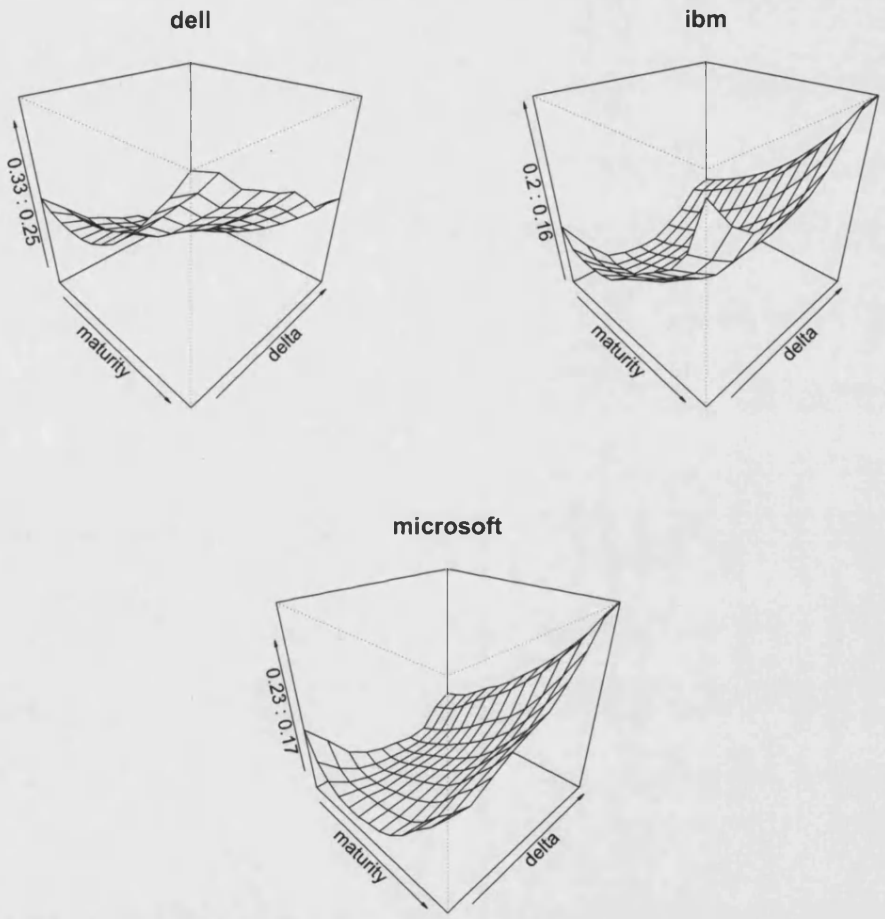


Figure 2.7: *Mean implied volatility surfaces.*

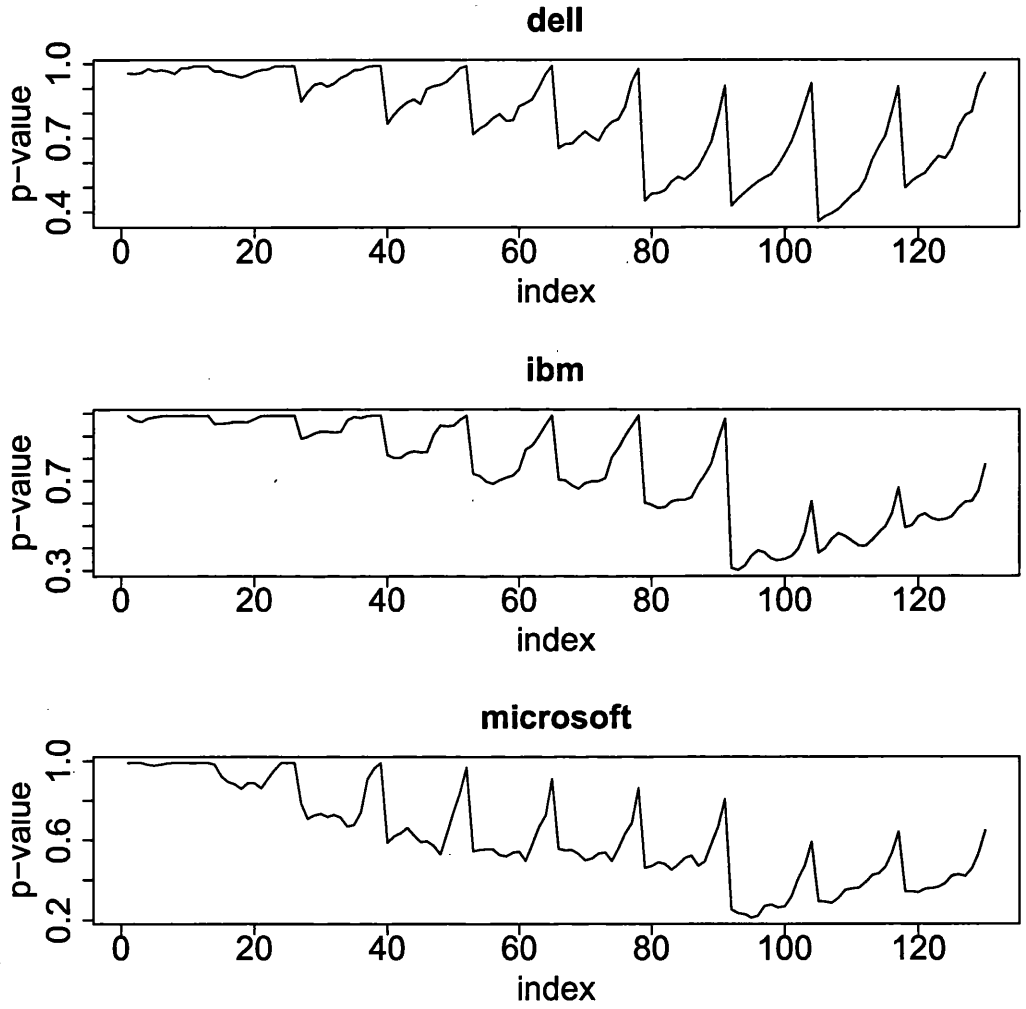


Figure 2.8: *P-values from the Dickey-Fuller test for the null* $H_0 : \delta = 0$ *in the regression* $\Delta W_t(u_i, v_j) = \delta W_t(u_i, v_j) + \varepsilon_t$. *The indices are* $i = 1, \ldots, p_u = 10$ *and* $j = 1, \ldots, p_v = 13$. *Thus the plotted P-values are from a total of* $p = p_u \times p_v = 130$ *hypothesis tests.*

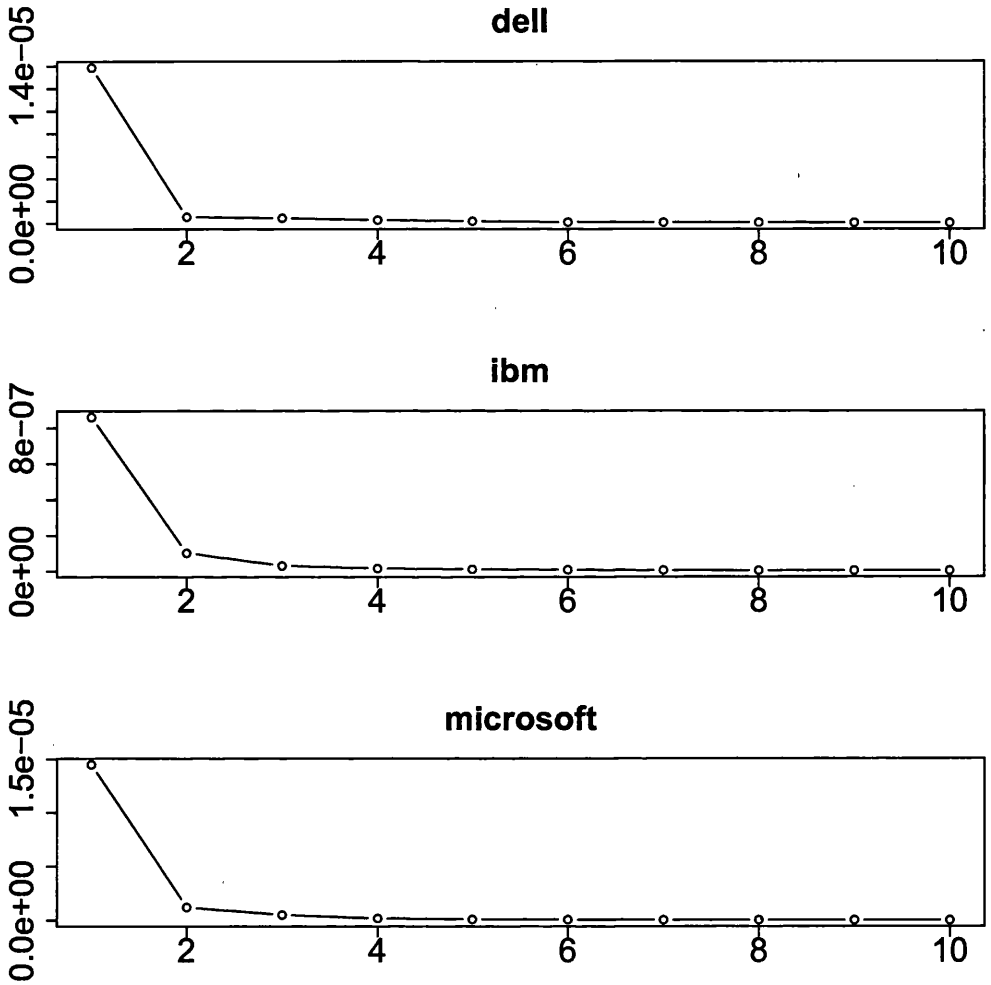


Figure 2.9: *Ten largest eigenvalues of* $\widehat{\mathbf{L}}_q$.

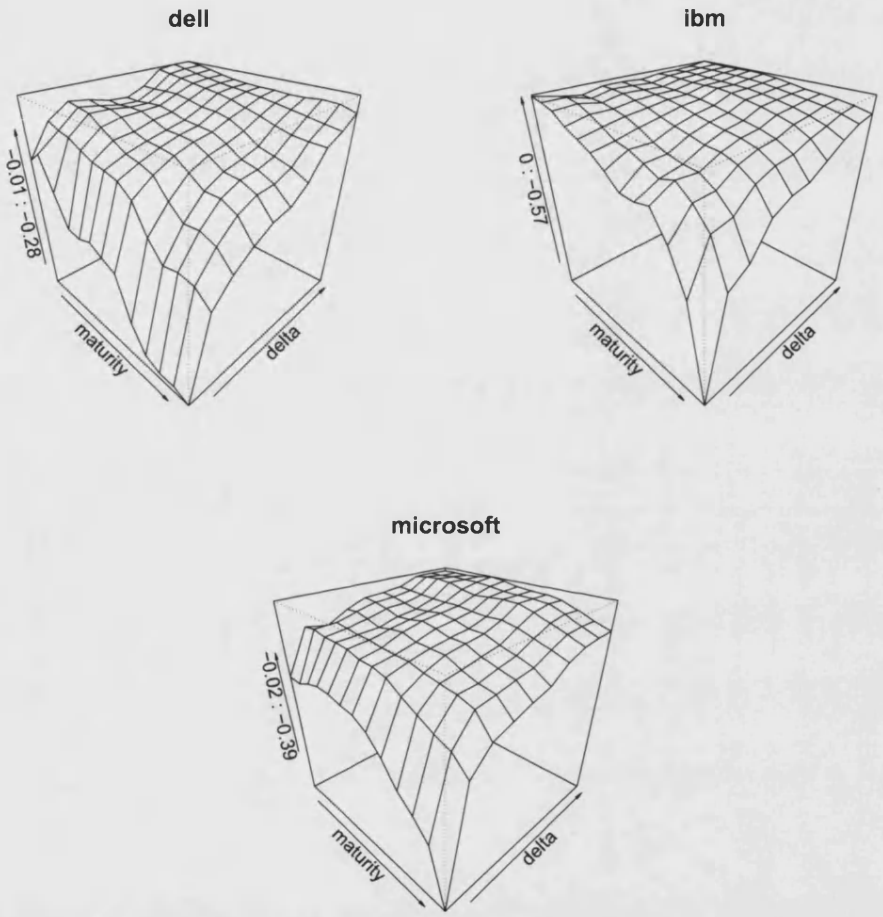


Figure 2.10: *Estimated factor loading surface* $vec^{-1}\{\widehat{\mathbf{e}}_1\}$.

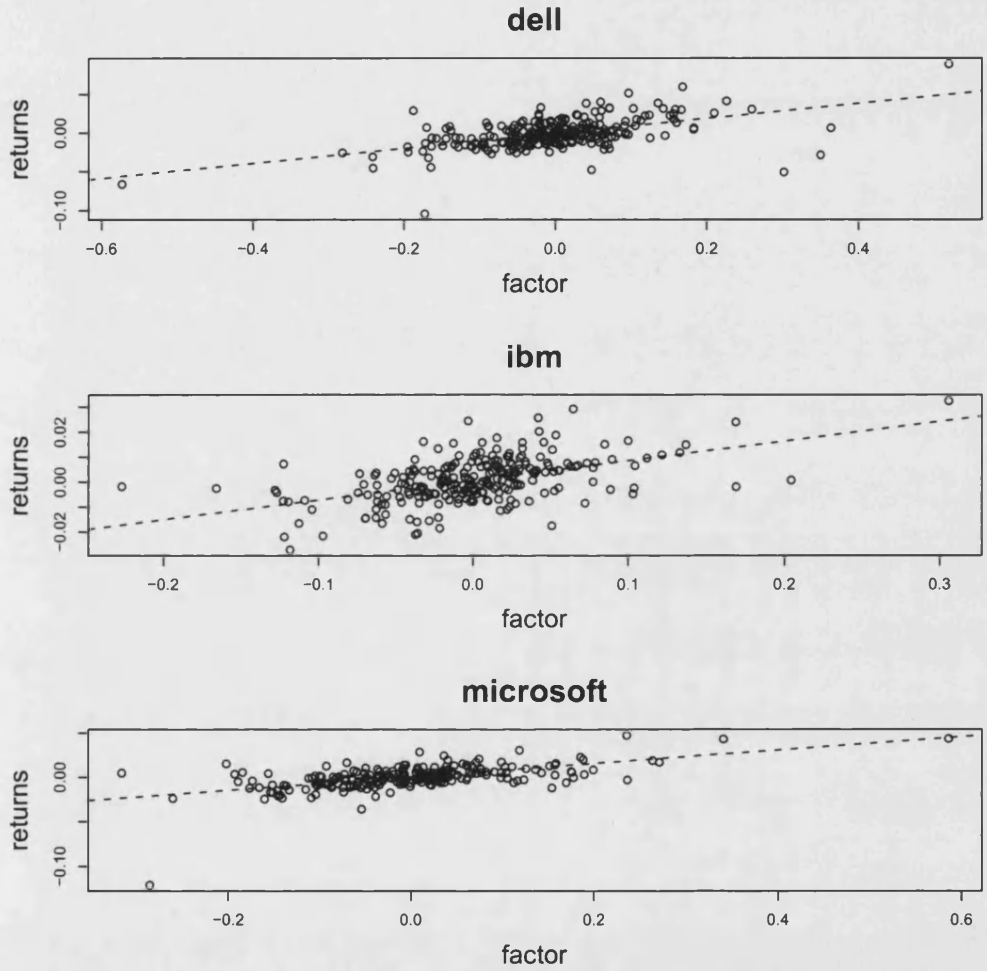


Figure 2.11: *Scatter plots of the stocks returns R_t against the estimated common factor $Z_{t1} = \mathbf{Y}_t'\widehat{\mathbf{e}}_1$. The period is 04/01/2006 - 29/12/2006. The dashed line represents an estimate of the regression $R_t = \alpha + \beta Z_{t1} + \varepsilon_t$. For all three companies, the parameter β is significant at the 1% level where as α is not significant at the 10% level.*

2.6 Proofs

In this section we provide the proofs of the results in Section 2.3. Throughout, we will use C to denote a generic constant which may vary from line to line.

Proof of Theorem 2.1 First note that $\mathbf{f} = \mathbf{P}_{\mathcal{M}}\mathbf{f} + \mathbf{P}_{\mathcal{M}^\perp}\mathbf{f}$ for any $\mathbf{f} \in \mathbb{R}^p$. Thus

$$\|\mathbf{P}_{\mathcal{M}}\widehat{\mathbf{e}}_{d_0+1}\|^2 = \|\widehat{\mathbf{e}}_{d_0+1} - \mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1}\|^2 = \sum_{j=1}^{d_0}(\mathbf{e}_j'\widehat{\mathbf{e}}_{d_0+1})^2, \tag{2.16}$$

where $\mathbf{e}_1, \ldots, \mathbf{e}_{d_0}$ are the eigenvectors of $\mathbf{L}_q$ corresponding to the non-zero eigenvalues $\lambda_1 > \cdots > \lambda_{d_0} > 0$. (Recall that $\text{span}\{\mathbf{e}_j : j = 1, \ldots, d_0\} = \mathcal{M}$ under H_0).

Now by noting that $\widehat{\mathbf{L}}_q\widehat{\mathbf{e}}_{d_0+1} = \widehat{\lambda}_{d_0+1}\widehat{\mathbf{e}}_{d_0+1}$ and $\|\widehat{\mathbf{e}}_{d_0+1}\| = 1$, we have

$$\begin{aligned}
\|\mathbf{L}_q\widehat{\mathbf{e}}_{d_0+1}\| &= \|(\mathbf{L}_q - \widehat{\mathbf{L}}_q)\widehat{\mathbf{e}}_{d_0+1} + \widehat{\mathbf{L}}_q\widehat{\mathbf{e}}_{d_0+1}\| \\
&\leq \|\mathbf{L}_q - \widehat{\mathbf{L}}_q\| \cdot \|\widehat{\mathbf{e}}_{d_0+1}\| + |\widehat{\lambda}_{d_0+1}| \cdot \|\widehat{\mathbf{e}}_{d_0+1}\| \\
&\leq 2\|\mathbf{L}_q - \widehat{\mathbf{L}}_q\|
\end{aligned} \tag{2.17}$$

where the final inequality follows from the fact $\lambda_{d_0+1} = 0$ under H_0 and thus $|\widehat{\lambda}_{d_0+1} - \lambda_{d_0+1}| = |\widehat{\lambda}_{d_0+1}| \leq \|\mathbf{L}_q - \widehat{\mathbf{L}}_q\|$ by Lemma B.1.

Let $\mathbf{e}_{d_0+1}, \ldots, \mathbf{e}_p$ be an orthonormal basis of $\mathcal{M}^\perp$. Then since $\lambda_{d_0+i} = 0$ for all $i \geq 1$ and $\mathbf{L}_q$ is symmetric, we have

$$\begin{aligned}
\|\mathbf{L}_q\widehat{\mathbf{e}}_{d_0+1}\|^2 &= \sum_{j=1}^{p}(\widehat{\mathbf{e}}_{d_0+1}'\mathbf{L}_q'\mathbf{e}_j)^2 \\
&= \sum_{j=1}^{p}(\widehat{\mathbf{e}}_{d_0+1}'\mathbf{L}_q\mathbf{e}_j)^2 \\
&= \sum_{j=1}^{d_0}\lambda_j(\widehat{\mathbf{e}}_{d_0+1}'\mathbf{e}_j)^2 \\
&\geq \min_{1\leq j\leq d_0}\{\lambda_j^2\}\sum_{j=1}^{d_0}(\widehat{\mathbf{e}}_{d_0+1}'\mathbf{e}_j)^2 \\
&= \lambda_{d_0}^2\sum_{j=1}^{d_0}(\widehat{\mathbf{e}}_{d_0+1}'\mathbf{e}_j)^2,
\end{aligned} \tag{2.18}$$

since $\lambda_1 > \cdots > \lambda_{d_0} > 0$.

Combining the bounds in (2.16), (2.17) and (2.18) we have

$$\|\mathbf{P}_{\mathcal{M}}\widehat{\mathbf{e}}_{d_0+1}\| = \|\widehat{\mathbf{e}}_{d_0+1} - \mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1}\| < C\|\mathbf{L}_q - \widehat{\mathbf{L}}_q\|. \tag{2.19}$$

Now by Lemma 1.3 in Chapter 1

$$\begin{aligned}
\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\| &\leq \sum_{k=1}^{q} \|\widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)' - \mathbf{\Sigma}_y(k)\mathbf{\Sigma}_y(k)'\| \\
&\leq \sum_{k=1}^{q} \{\|\widehat{\mathbf{\Sigma}}_y(k)\| + \|\mathbf{\Sigma}_y(k)\|\}\|\widehat{\mathbf{\Sigma}}_y(k) - \mathbf{\Sigma}_y(k)\| \\
&= O_p(n^{-\gamma}).
\end{aligned} \tag{2.20}$$

Thus from (2.19) and (2.20), the first part is proven.

Next we have

$$\begin{aligned}
|\mathbf{Y}_t'\widehat{\mathbf{e}}_{d_0+1} - \mathbf{Y}_t'\mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1}| &= |\mathbf{Y}_t'(\widehat{\mathbf{e}}_{d_0+1} - \mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1})| \\
&\leq \|\mathbf{Y}_t\|\|\widehat{\mathbf{e}}_{d_0+1} - \mathbf{P}_{\mathcal{M}^\perp}\widehat{\mathbf{e}}_{d_0+1}\| \\
&= O_p(n^{-\gamma}),
\end{aligned}$$

which concludes the proof of the assertion. □

Before proceeding to the proof of Theorem 2.2, we require some auxiliary results regarding matrix valued von Mises functionals. Let $\mathbf{A}_t \in \mathbb{R}^{p\times p}$ be a sequence of strictly stationary random matrices with distribution function denoted by $P(\mathbf{A})$, $\mathbf{A} \in \mathbb{R}^{p\times p}$. Let $\phi : (\mathbb{R}^{p\times p})^m \longrightarrow \mathbb{R}^{p\times p}$ be integrable and symmetric in each of its $m(\geq 2)$ arguments. Now consider the functional

$$\mathbf{\Theta}(P) = \int \phi(\mathbf{A}_1, \ldots, \mathbf{A}_m) \prod_{j=1}^{m} P(d\mathbf{A}_j),$$

defined over $\mathcal{P} = \{P : \|\mathbf{\Theta}(P)\| < \infty\}$. As an estimator of $\mathbf{\Theta}(P)$, consider the U-statistic based on n observations $\mathbf{A}_1, \ldots, \mathbf{A}_n$ defined as

$$\mathbf{U}_n = \binom{n}{m} \sum_{1\leq i_1<\cdots<i_m\leq n} \phi(\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_m}).$$

As another estimator, we shall also consider the von Mises functional defined by

$$\mathbf{V}_n = n^{-m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \phi(\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_m}).$$

Finally for $c = 0, 1, \ldots, m$, we define the functions

$$\phi_c(\mathbf{A}_1, \ldots, \mathbf{A}_c) = \int \phi(\mathbf{A}_1, \ldots, \mathbf{A}_c, \mathbf{A}_{c+1}, \ldots, \mathbf{A}_m) \prod_{j=c+1}^{m} P(d\mathbf{A}_j),$$

and

$$g_c(\mathbf{A}_1, \ldots, \mathbf{A}_c) = \sum_{d=0}^{c} (-1)^{c-d} \sum_{1 \le j_1 < \cdots < j_d \le c} \phi_d(\mathbf{A}_{j_1}, \ldots, \mathbf{A}_{j_d}).$$

Then via the Hoeffding decomposition we have

$$\mathbf{U}_n - \boldsymbol{\Theta}(P) = \sum_{c=1}^{m} \binom{m}{c} \mathbf{U}_{nc} \tag{2.21}$$

where $\mathbf{U}_{nc}$ is also a U-statistic with kernel g_c, i.e.

$$\mathbf{U}_{nc} = \binom{n}{c}^{-1} \sum_{1 \le i_1 < \cdots < i_c \le n} g_c(\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_c}).$$

In an analogous fashion, we may consider the canonical decomposition of $\mathbf{V}_n$ for which we use Dirac's δ-measure to define the empirical measure P_n as follows

$$P_n(\mathbf{A}) = n^{-1} \left(\delta_{\mathbf{A}_1}(\mathbf{A}) + \cdots + \delta_{\mathbf{A}_n}(\mathbf{A}) \right), \qquad \mathbf{A} \in \mathbb{R}^{p \times p}.$$

Then for $c = 1, \ldots, m$, we set

$$\begin{aligned} \mathbf{V}_{nc} &= \int \phi_c(\mathbf{A}_1, \ldots, \mathbf{A}_c) \prod_{j=1}^{c} (P_n(d\mathbf{A}_j) - P(d\mathbf{A}_j)) \\ &= n^{-c} \sum_{i_1=1}^{n} \cdots \sum_{i_c=1}^{n} g_c(\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_c}), \end{aligned}$$

then we have

$$\mathbf{V}_n - \boldsymbol{\Theta}(P) = \sum_{c=1}^{m} \binom{m}{c} \mathbf{V}_{nc}. \tag{2.22}$$

In particular note that

$$\mathbf{V}_{n1} = \mathbf{U}_{n1} = \frac{1}{n}\sum_{i=1}^{n} g_1(\mathbf{A}_i).$$

Decompositions (2.21) and (2.22) play a central role in the proof of Lemma 2.1 below. Further details about U-statistics and von Mises functionals may be found in Lee (1990).

We are now in a position to state some regularity conditions which form the basis of the results.

U1. $\{\mathbf{A}_t\}$ is strictly stationary and β-mixing with β-mixing coefficients satisfying $\beta(l) = O(l^{-(2+\delta')/\delta'})$ for some $\delta' > 0$.

U2. It holds for all $1 \le i_1 < \cdots < i_m \le i_m$ that $E\left\{\|\phi(\mathbf{A}_{i_1}, \ldots, \mathbf{A}_{i_m})\|^{2+\delta}\right\} < \infty$ and

$$\int \|\phi(\mathbf{A}_1, \ldots, \mathbf{A}_m)\|^{2+\delta} \prod_{j=1}^{m} P(d\mathbf{A}_j) < \infty,$$

for some $\delta > \delta'$ given in U1 above.

Lemma 2.1 *Let conditions U1 and U2 hold. (i) Then it holds that*

$$E\|m^{-1}(\mathbf{U}_n - \boldsymbol{\Theta}(P)) - \mathbf{U}_{n,1}\|^2 = O(n^{-1-\gamma}),$$

where $\gamma = \min\left\{1, \frac{2(\delta-\delta')}{\delta'(2+\delta)}\right\}$. *(ii) In addition*

$$E\|\mathbf{V}_n - \mathbf{U}_n\|^2 = o(n^{-1-\gamma}).$$

Thus (i) also holds for $\mathbf{U}_n$ *being replaced by* $\mathbf{V}_n$.

Proof of Lemma 2.1 (i) We make use of (2.21). Put

$$m^{-1}(\mathbf{U}_n - \boldsymbol{\Theta}(P)) = \mathbf{U}_{n1} + \mathbf{R}_n, \tag{2.23}$$

where

$$\mathbf{R}_n = m^{-1}\sum_{c=2}^{m} \binom{m}{c} \mathbf{U}_{nc}. \tag{2.24}$$

We show that $E\|\mathbf{U}_{nc}\|^2 = O(n^{-1-\gamma})$. Let $\mathbf{E}_{ij} \in \mathbb{R}^{p\times p}$ be the matrix with 1 in it's ij-th position and 0 elsewhere. Then

$$E\|\mathbf{U}_{nc}\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} E(\mathrm{tr}\{\mathbf{U}'_{nc}\mathbf{E}_{ij}\}^2), \tag{2.25}$$

where $\mathrm{tr}\{\mathbf{U}'_{nc}\mathbf{E}_{ij}\}$ is the $\mathbb{R}$ valued U-statistic

$$\mathrm{tr}\{\mathbf{U}'_{nc}\mathbf{E}_{ij}\} = \binom{n}{c}^{-1} \sum_{1\le i_1<\cdots<i_c\le n} \mathrm{tr}\{g_c(\mathbf{A}_{i_1},\ldots,\mathbf{A}_{i_c})'\mathbf{E}_{ij}\}.$$

Now under conditions U1 and U2, part (i) of Lemma B.5 yields

$$E(\mathrm{tr}\{\mathbf{U}'_{nc}\mathbf{E}_{ij}\}^2) = O(n^{-1-\gamma}), \qquad c \ge 2. \tag{2.26}$$

Thus inserting estimate (2.26) into (2.25) and by recalling that p is fixed and finite, we have $E\|\mathbf{U}_{nc}\|^2 = O(n^{-1-\gamma})$.

Now from (2.24) and Loève's c_r-inequality we have

$$E\|\mathbf{R}_n\|^2 \le C\sum_{c=2}^{m} E\|\mathbf{U}_{nc}\|^2 = O(n^{-1-\gamma}). \tag{2.27}$$

Combining equations (2.23) and (2.27) concludes the proof of part (i).

(ii) By virtue of (2.22) and part (ii) of Lemma B.5, an analogous set of manipulations to those used in the proof of part (i) above yields $E\|\mathbf{V}_{nc}^2\| = O(n^{-1-\gamma})$ for $c \ge 2$. The result now follows from the fact that $\mathbf{V}_{n1} = \mathbf{U}_{n1}$. □

We now have the following result about the rates of convergence for $E\|\mathbf{U}_n - \boldsymbol{\Theta}(P)\|^2$ and $E\|\mathbf{V}_n - \boldsymbol{\Theta}(P)\|^2$.

Lemma 2.2 *Let conditions U1 and U2 hold. (i) Then we have* $E\|\mathbf{U}_n - \boldsymbol{\Theta}(P)\|^2 = O(n^{-1})$. *(ii) In addition, the result of part (i) holds for* $\mathbf{U}_n$ *replaced by* $\mathbf{V}_n$.

Proof of Lemma 2.2 (i) First note that

$$\|\mathbf{U}_n - \boldsymbol{\Theta}(P)\| \le m\|m^{-1}(\mathbf{U}_n - \boldsymbol{\Theta}(P) - \mathbf{U}_{n1}\| + m\|\mathbf{U}_{n1}\|.$$

Thus by Lemma 2.1, we have

$$E\|\mathbf{U}_n - \boldsymbol{\Theta}(P)\|^2 = O(n^{-1-\gamma}) + O(E\|\mathbf{U}_{n1}\|^2). \tag{2.28}$$

Now using a similar argument to that in the proof of Lemma 2.1 we have

$$E\|\mathbf{U}_{n1}\|^2 = \sum_{i=1}^{p}\sum_{j=1}^{p} E(\mathrm{tr}\{\mathbf{U}'_{n1}\mathbf{E}_{ij}\}^2), \tag{2.29}$$

where $\mathrm{tr}\{\mathbf{U}'_{n1}\mathbf{E}_{ij}\}$ is the $\mathbb{R}$ valued partial sum

$$\mathrm{tr}\{\mathbf{U}'_{n1}\mathbf{E}_{ij}\} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{tr}\{g_1(\mathbf{A}_i)'\mathbf{E}_{ij}\}.$$

Then Lemma B.6 yields $E(\mathrm{tr}\{\mathbf{U}'_{n1}\mathbf{E}_{ij}\}^2) = O(n^{-1})$. Combining this last estimate with (2.28) and (2.29) produces the required result.

(ii) The proof of the second part of the assertion follows an identical set of arguments to those given above and is thus omitted. □

Proof of Theorem 2.2 (i) From Lemma B.1, we have $|\widehat{\lambda}_j - \lambda_j| \leq \|\widehat{\mathbf{L}}_q - \mathbf{L}_q\|$ for all $j \geq 1$. Thus we only need to show that $\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\| = O_p(n^{-1/2})$. To this end, consider the kernel $\phi : \mathbb{R}^{p\times p} \times \mathbb{R}^{p\times p} \longrightarrow \mathbb{R}^{p\times p}$ given by

$$\phi(\mathbf{A}, \mathbf{B}) = \mathbf{A}\mathbf{B}', \qquad \mathbf{A}, \mathbf{B} \in \mathbb{R}^{p\times p}.$$

Then from the definition of $\widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)'$ we have

$$\begin{aligned}\widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)' &= \frac{1}{(n-k)^2}\sum_{i=1}^{n-k}\sum_{j=1}^{n-k} \mathbf{Z}_i(k)\mathbf{Z}_j(k)' \\ &= \frac{1}{(n-k)^2}\sum_{i=1}^{n-k}\sum_{j=1}^{n-k} \phi(\mathbf{Z}_i(k), \mathbf{Z}_j(k)),\end{aligned} \tag{2.30}$$

which in light of the preceding discussion is simply an $\mathbb{R}^{p\times p}$ valued von Mises functional. Thus from the definition of $\widehat{\mathbf{L}}_q$ in (2.5), Lemma 2.2 and the c_r inequality yield

$$E\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\|^2 \leq C \cdot \sum_{k=1}^{q} E\|\widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)' - \mathbf{\Sigma}_y(k)\mathbf{\Sigma})y(k)'\|^2 = O(n^{-1}). \tag{2.31}$$

since $n \sim n - k$ for each fixed k. Now via an application of the Chebyshev inequality to (2.31), we have

$$\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\| = O_p(n^{-1/2}), \tag{2.32}$$

as required.

(ii) Extend $\mathbf{e}_1, \ldots, \mathbf{e}_d$ to an orthonormal basis of $\mathbb{R}^p$. Then it holds that

$$\sum_{j=1}^{p} \widehat{\lambda}_j = \sum_{j=1}^{p} \mathbf{e}_j' \widehat{\mathbf{L}}_q \mathbf{e}_j, \tag{2.33}$$

and by recalling that $\lambda_j = 0$ for $j \geq d+1$

$$\sum_{j=1}^{d} \lambda_j = \sum_{j=1}^{d} \mathbf{e}_j' \mathbf{L}_q \mathbf{e}_j. \tag{2.34}$$

Note that $\text{span}\{\mathbf{e}_j : j = d+1, \ldots, p\} = \mathcal{M}^{\perp}$. Therefore $\mathbf{L}_q \mathbf{e}_j = 0$ for all $j \geq d+1$ and thus from (2.33) and (2.34) we have

$$\sum_{j=1}^{p} \widehat{\lambda}_j - \lambda_j = \sum_{j=1}^{p} \mathbf{e}_j' (\widehat{\mathbf{L}}_q - \mathbf{L}_q) \mathbf{e}_j. \tag{2.35}$$

We claim that

$$\widehat{\lambda}_j - \lambda_j = \mathbf{e}_j' (\widehat{\mathbf{L}}_q - \mathbf{L}_q) \mathbf{e}_j + O_p(n^{-1}), \qquad j = 1, \ldots, d. \tag{2.36}$$

Using the fact that $\mathbf{L}_q$ is symmetric along with the relations $\widehat{\mathbf{L}}_q \widehat{\mathbf{e}}_j = \widehat{\lambda}_j \widehat{\mathbf{e}}_j$ and $\mathbf{L}_q \mathbf{e}_j = \lambda_j \mathbf{e}_j$, we have

$$\begin{aligned} |\mathbf{e}_j' (\widehat{\mathbf{L}}_q - \mathbf{L}_q) \widehat{\mathbf{e}}_j - (\widehat{\lambda}_j - \lambda_j)| &= |\mathbf{e}_j' \widehat{\mathbf{L}}_q \widehat{\mathbf{e}}_j - \widehat{\mathbf{e}}_j' \mathbf{L}_q \mathbf{e}_j - (\widehat{\lambda}_j - \lambda_j)| \\ &= |\widehat{\lambda}_j - \lambda_j| \, |\mathbf{e}_j' \widehat{\mathbf{e}}_j - 1|. \end{aligned} \tag{2.37}$$

Now by the Cauchy-Schwarz inequality

$$|\mathbf{e}_j' \widehat{\mathbf{e}}_j - 1| = |\mathbf{e}_j' (\widehat{\mathbf{e}}_j - \mathbf{e}_j)| \leq \|\mathbf{e}_j\| \|\widehat{\mathbf{e}}_j - \mathbf{e}_j\| = \|\widehat{\mathbf{e}}_j - \mathbf{e}_j\|, \tag{2.38}$$

and from Lemma B.2 and (2.32) we have

$$\|\widehat{\mathbf{e}}_j - \mathbf{e}_j\| \leq \|\widehat{\mathbf{L}}_q - \mathbf{L}_q\| = O_p(n^{-1/2}). \tag{2.39}$$

Thus from the result of part (i), (2.37), (2.38) and (2.39) we have

$$|\mathbf{e}_j' (\widehat{\mathbf{L}}_q - \mathbf{L}_q) \widehat{\mathbf{e}}_j - (\widehat{\lambda}_j - \lambda_j)| = O_p(n^{-1}). \tag{2.40}$$

Next we have

$$\begin{aligned} |\mathbf{e}_j'(\widehat{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{e}_j - \mathbf{e}_j'(\widehat{\mathbf{L}}_q - \mathbf{L}_q)\widehat{\mathbf{e}}_j| &= |\mathbf{e}_j'(\widehat{\mathbf{L}}_q - \mathbf{L}_q)(\mathbf{e}_j - \widehat{\mathbf{e}}_j)| \\ &\leq \sup_{\|\mathbf{x}\|\leq 1}\|(\widehat{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{x}\|\|\mathbf{e}_j - \widehat{\mathbf{e}}_j\| \\ &\leq \|\widehat{\mathbf{L}}_q - \mathbf{L}_q\|\|\mathbf{e}_j - \widehat{\mathbf{e}}_j\|. \\ &= O_p(n^{-1}), \end{aligned} \tag{2.41}$$

by (2.32) and (2.39). Thus relations (2.40) and (2.41) together imply (2.36).

Now from (2.36) we have

$$\sum_{j=1}^{d}\widehat{\lambda}_j - \lambda_j = \sum_{j=1}^{d}\mathbf{e}_j'(\widehat{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{e}_j + O_p(n^{-1}). \tag{2.42}$$

Subtracting (2.42) from (2.35) yields

$$\sum_{j=d+1}^{p}\widehat{\lambda}_j = \sum_{j=d+1}^{p}\mathbf{e}_j'(\widehat{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{e}_j + O_p(n^{-1}). \tag{2.43}$$

Our final task is to derive the rate for $\sum_{j=d+1}^{p}\mathbf{e}'(\widehat{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{e}_j$. To this end, we note that an application of Lemma 2.1 and the Chebyshev inequality to the von Mises functional in (2.30) yields

$$\|\widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k) - \widehat{\mathbf{\Sigma}}_y(k)\mathbf{\Sigma}_y(k)\| = O_p(n^{(-1-\gamma)/2}) \tag{2.44}$$

since

$$\widehat{\mathbf{\Sigma}}_y(k)\mathbf{\Sigma}_y(k) = \frac{1}{n-k}\sum_{i=1}^{n-k}\int \phi(\mathbf{Z}_i(k), \mathbf{Z}_j(k))P_k(d\mathbf{Z}_j(k)).$$

Put $\widetilde{\mathbf{L}}_q = \sum_{k=1}^{q}\widehat{\mathbf{\Sigma}}_y(k)\mathbf{\Sigma}_y(k)'$. Then from (2.44) and the definition of $\widehat{\mathbf{L}}_q$ in (2.5)

$$\|\widehat{\mathbf{L}}_q - \widetilde{\mathbf{L}}_q\| \leq \sum_{k=1}^{q}\|\widehat{\mathbf{\Sigma}}_y(k)\widehat{\mathbf{\Sigma}}_y(k)' - \widehat{\mathbf{\Sigma}}_y(k)\mathbf{\Sigma}_y(k)\| = O_p(n^{(-1-\gamma)/2}). \tag{2.45}$$

Now by the relation in (2.45)

$$|\mathbf{e}_j'(\widehat{\mathbf{L}}_q - \widetilde{\mathbf{L}}_q)\mathbf{e}_j| \leq \sup_{\|\mathbf{x}\|\leq 1}\|(\widehat{\mathbf{L}}_q - \widetilde{\mathbf{L}}_q)\mathbf{x}\| \leq \|\widehat{\mathbf{L}}_q - \widetilde{\mathbf{L}}_q\| = O_p(n^{(-1-\gamma)/2}). \tag{2.46}$$

Thus from (2.43) and (2.46)

$$\sum_{j=d+1}^{p}\widehat{\lambda}_j = \sum_{j=d+1}^{p}\mathbf{e}_j'(\widetilde{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{e}_j + O_p(n^{(-1-\gamma)/2}), \tag{2.47}$$

since $\gamma \leq 1$. Now note that $\mathbf{e}_j'(\widetilde{\mathbf{L}}_q - \mathbf{L}_q)\mathbf{e}_j = 0$ for $j = d+1, \ldots, p$ since $\mathbf{e}_j \in \mathcal{M}^{\perp}$ for these values of j. This last relation together with (2.47) yield $\sum_{j=d+1}^{p} \widehat{\lambda}_j = O_p(n^{(-1-\gamma)/2})$. Finally, since $\widehat{\mathbf{L}}_q$ is positive definite it holds that $\lambda_i \leq \sum_{j=d+1}^{p} \widehat{\lambda}_j = O_p(n^{(-1-\gamma)/2})$ for $i = d+1, \ldots, p$. □

Proof of Theorem 2.3 Since $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_p \geq 0$ (with a strict inequality holding with probability one), it follows that $\{\widehat{d} > d\} = \{\widehat{\lambda}_{d+1} \geq \varepsilon\}$. Now if $\dim\{\mathcal{M}\} = d$, then it holds that $\lambda_{d+1} = 0$ and thus $\widehat{\lambda}_{d+1} = |\widehat{\lambda}_{d+1} - \lambda_{d+1}| \leq \|\widehat{\mathbf{L}}_q - \mathbf{L}_q\|$ by Lemma B.1. Collecting these last few facts and applying the Chebyshev inequality yields

$$P(\widehat{d} > d) \leq \varepsilon^{-2} \cdot E\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\|^2 = O((\varepsilon^2 n)^{-1}), \tag{2.48}$$

by (2.31).

Next we turn to $P(\widehat{d} < d)$. First note that if $d = 0$, $\{\widehat{d} < d\} = \emptyset$. Thus, we suppose that $d \geq 1$. Then due to the ordering of the eigenvalues, it holds that $\{\widehat{d} < d\} = \{\widehat{\lambda}_{d-1} < \varepsilon\}$. Therefore

$$\begin{aligned} P(\widehat{d} < d) &= P(\widehat{\lambda}_{d-1} < \varepsilon) \\ &= P(\lambda_{d-1} - \widehat{\lambda}_{d-1} > \lambda_{d-1} - \varepsilon) \\ &\leq P(|\lambda_{d-1} - \widehat{\lambda}_{d-1}| > \lambda_{d-1} - \varepsilon) \\ &\leq P(\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\| > \lambda_{d-1} - \varepsilon), \end{aligned} \tag{2.49}$$

where the final inequality follows from Lemma B.1. Now since $\lambda_{d-1} > 0$ and $\varepsilon \longrightarrow 0$ as $n \longrightarrow \infty$, it holds that $\lambda_{d-1} - \varepsilon > 0$ for large enough n. Thus by (2.31) and an application of the Chebyshev inequality to (2.49), we have

$$P(\widehat{d} < d) \leq (\lambda_{d-1} - \varepsilon)^{-2} \cdot E\|\widehat{\mathbf{L}}_q - \mathbf{L}_q\|^2 = O(n^{-1}). \tag{2.50}$$

From (2.48) and (2.50) it follows that

$$P(\widehat{d} \neq d) = P(\widehat{d} < d) + P(\widehat{d} > d) = O((\varepsilon^2 n)^{-1}),$$

as required. □

Appendix A

Background on operator theory

In this section we provide the relevant background on operator theory used in this work. More detailed accounts may be found in Dunford & Schwartz (1988).

Let $\mathcal{H}$ be a real separable Hilbert space with respect to some inner product $\langle \cdot, \cdot \rangle$. For any $\mathcal{V} \subset \mathcal{H}$, the orthogonal complement of $\mathcal{V}$ is given by

$$\mathcal{V}^{\perp} = \{x \in \mathcal{H} : \langle x, y \rangle = 0, \ \forall\, y \in \mathcal{H}\}.$$

Note that $\mathcal{V}^{\perp\perp} = \overline{\mathcal{V}}$ where $\overline{\mathcal{V}}$ denotes the closure of $\mathcal{V}$. Clearly if $\mathcal{V}$ is finite dimensional then $\mathcal{V}^{\perp\perp} = \mathcal{V}$.

Let L be a linear operator from $\mathcal{H}$ to $\mathcal{H}$. For $x \in \mathcal{H}$, denote by Lx the image of x under L. The adjoint of L is denoted by L^* and satisfies

$$\langle Lx, y \rangle = \langle x, L^* y \rangle, \qquad x, y \in \mathcal{H}.$$

L is said to be self adjoint if $L^* = L$ and non-negative definite if

$$\langle Lx, x \rangle \geq 0, \qquad \forall\, x \in \mathcal{H}.$$

The image and null space of L are defined as $\mathrm{Im}(L) = \{y \in \mathcal{H} : y = Lx, \ x \in \mathcal{H}\}$ and $\mathrm{Ker}(L) = \{x \in \mathcal{H} : Lx = 0\}$ respectively. We define the rank of L to be $\mathrm{r}(L) = \dim(\mathrm{Im}(L))$ and we say that L is finite dimensional if $r(L) < \infty$.

A linear operator L is said to be bounded if there exists some finite constant $\Delta > 0$ such that for all $x \in \mathcal{H}$

$$\|Lx\| < \Delta \|x\|,$$

where $\|\cdot\|$ is the norm induced on $\mathcal{H}$ by $\langle\cdot,\cdot\rangle$. We denote the space of bounded linear operators from $\mathcal{H}$ to $\mathcal{H}$ by $\mathcal{B} = \mathcal{B}(\mathcal{H},\mathcal{H})$ and the uniform topology on $\mathcal{B}$ is defined by

$$\|L\|_{\mathcal{B}} = \sup_{\|x\|\le 1} \|Lx\|, \qquad L \in \mathcal{B}.$$

Note that all bounded linear operators are continuous, and the converse also holds.

An operator $L \in \mathcal{B}$ is said to be compact if there exists two orthonormal sequences $\{e_j\}$ in $\{f_j\}$ of $\mathcal{H}$ and a sequence of scalars $\{\lambda_j\}$ decreasing to zero such that

$$Lx = \sum_{j=1}^{\infty} \lambda_j \langle e_j, x\rangle\, f_j, \qquad x \in \mathcal{H},$$

or more compactly

$$L = \sum_{j=1}^{\infty} \lambda_j\, e_j \otimes f_j.$$

Note that if $\mathcal{H} = \mathcal{L}_2(\mathcal{I})$ equipped with the inner product defined in (1.5), then

$$(Lx)(u) = \sum_{j=1}^{\infty} \lambda_j \langle e_j, x\rangle f_j(u).$$

Clearly $\mathrm{Im}(L) = \mathrm{sp}\{f_j : j \ge 1\}$ and $\mathrm{Ker}(L) = \mathrm{sp}\{e_j : j \ge 1\}^{\perp}$.

The Hilbert-Schmidt norm of a compact linear operator L is defined as $\|L\|_{\mathcal{S}} = (\sum_{j=1}^{\infty} \lambda_j^2)^{1/2}$ and the nuclear norm is $\|L\|_{\mathcal{N}} = \sum_{j=1}^{\infty} |\lambda_j|$. We will let $\mathcal{S}$ and $\mathcal{N}$ denote, respectively, the space consisting of all the operators with a finite Hilbert-Schmidt or nuclear norm. Clearly we have the inequalities $\|\cdot\|_{\mathcal{N}} \ge \|\cdot\|_{\mathcal{S}} \ge \|\cdot\|_{\mathcal{B}}$, and thus the inclusions $\mathcal{N} \subset \mathcal{S} \subset \mathcal{B}$. Note that $\mathcal{N}$, $\mathcal{S}$, and $\mathcal{B}$ are Banach spaces when equipped with their respective norms. Furthermore $\mathcal{S}$ is a Hilbert space with respect to the inner product

$$\langle L_1, L_2\rangle_{\mathcal{S}} = \sum_{i,j=1}^{\infty} \langle L_1 g_i, h_j\rangle\langle L_2 g_i, h_j\rangle, \qquad L_1, L_2 \in \mathcal{S},$$

where $\{g_i\}$ and $\{h_j\}$ are any orthonormal bases of $\mathcal{H}$.

Appendix B

Some useful technical Lemma's

In this section, we state some results due to other authors which are used in the proofs contained in Section's 1.6 and 2.6.

Lemma B.1 *(Bosq (2000), Lemma 4.2) Let* $L_0 = \sum_{j=1}^{\infty} \theta_{0,j} e_{0,j} \otimes f_{0,j}$ *and* $L_1 = \sum_{j=1}^{\infty} \theta_{1,j} e_{1,j} \otimes f_{1,j}$ *be compact linear operators acting on some separable Hilbert space* $\mathcal{H}$*. Suppose without loss of generality that* $\theta_{i,j} \geq \theta_{i,j+1}$ *for* $i = 0, 1$*. Then it holds that* $\sup_{j \geq 1} |\theta_{0,j} - \theta_{1,j}| \leq \|L_0 - L_1\|_{\mathcal{B}}$*.*

Lemma B.2 *(Bosq (2000), Lemma 4.3) Let* L_0 *and* L_1 *be as in Lemma B.1 except now suppose without loss of generality that they are both self adjoint, i.e.* $e_{i,j} = f_{i,j}$ *for* $i = 0, 1$ *(if this does not hold then we may instead consider the eigenvectors of* $L_i L_i^*$*). Let* $\|\cdot\|$ *be the norm on* $\mathcal{H}$ *induced by the inner product* $\langle \cdot, \cdot \rangle$*. Suppose that we have the strict inequality* $\theta_{i,j} > \theta_{i,j+1}$ *for* $i = 0, 1$*. Then it holds that* $\|e_{0,j} - \text{sign}\{\langle e_{0,j}, e_{1,j} \rangle\} e_{1,j}\| \leq C_j \|L_0 - L_1\|_{\mathcal{B}}$ *where* $C_j \leq C_{j+1} < \infty$ *is a sequence of constants depending only on* j*.*

Lemma B.3 *(Merlevede et al. (1997), Lemma 3) Let* $\{A_t\}$ *be a sequence of centered* $\mathcal{H}$ *valued random variables (not necessarily stationary) with* α*-mixing coefficients* $\alpha(u)$*. Denote by* $Q_{\|A_i\|}(u)$ *the quantile function of* $\|A_i\|$*, i.e. the inverse function of* $G(u) = P(\|A_i\| > u)$ *and assume that for each* $i = 1, \ldots, n$

$$\int_0^1 \alpha^{-1}(u) Q_i^2(u) du < \infty. \tag{B.1}$$

Then for every $n \geq 1$ it holds that

$$E\|\sum_{i=1}^{n} A_i\|^2 \leq 36 \sum_{i=1}^{n} \int_0^1 \alpha^{-1}(u) Q_i^2(u) du.$$

Note that a sufficient condition for (B.1) is that $\{A_t\}$ is strictly stationary and theres exists some $\delta > 0$ such that $E\|A_t\|^{2+\delta} < \infty$ and $\sum_{l=1}^{\infty} \alpha(l)^{\delta/(2+\delta)}$; see Theorem 2.17 in Bosq (2000).

Lemma B.4 *(Mas (2000), Proposition 6) Let L_0 and L_1 be as in Lemma B.2 and assume that we have the strict inequality $\theta_{i,j} > \theta_{i,j+1}$ for $i = 0, 1$. Let $\Pi_{i,j} = e_{i,j} \otimes e_{i,j}$ for $i = 0, 1$, i.e. $\Pi_{i,j}$ is the projection operator onto the eigensubspace of $\theta_{i,j}$. Then it holds that $\|\Pi_{0,j} - \Pi_{1,j}\|_S \leq D_j \|L_0 - L_1\|_S$ where $D_j \leq D_{j+1} < \infty$ is a sequence of constants depending only on j.*

Lemma B.5 *(Yoshihara (1976), Lemma's 2 and 4) Let $\{X_t\}$ be a sequence of strictly stationary random variables with distribution denoted by $P(X)$, $X \in \mathbb{R}$. Now consider the functional*

$$\theta(P) = \int \phi(X_1, \ldots, X_m) \prod_{j=1}^{m} P(dX_j),$$

defined over $\mathcal{P} = \{P : \|\theta(P)\| < \infty\}$. Based on a sample $X_1, \ldots, X_n$, we consider the following estimators of $\theta(P)$

$$U_n = \binom{n}{m} \sum_{1 \leq i_1 < \cdots < i_m \leq n} \phi(X_{i_1}, \ldots, X_{i_m}), \quad V_n = n^{-m} \sum_{i_1=1}^{n} \cdots \sum_{i_m=1}^{n} \phi(X_{i_1}, \ldots, X_{i_m})$$

which possess the canonical decompositions

$$U_n - \theta(P) = \sum_{c=1}^{m} \binom{m}{c} U_{nc}, \quad V_n - \theta(P) = \sum_{c=1}^{m} \binom{m}{c} V_{nc}.$$

In particular

$$U_{n1} = V_{n1} = n^{-1} \sum_{i_1=1}^{n} \int \phi(X_{i_1}, \ldots, X_{i_m}) \prod_{j=2}^{m} P(dX_{i_j}) - \theta(P).$$

Suppose that for some $\delta > 0$, $\sup_{i_1,\ldots,i_m} E|\phi(X_{i_1},\ldots,X_{i_m})|^{2+\delta} < \infty$ *and*

$$\int |\phi(X_1,\ldots,X_m)|^{2+\delta} \prod_{j=1}^{m} P(dX_j) < \infty,$$

and the beta mixing coefficients of $\{X_t\}$ *satisfy* $\beta(l) = O(l^{-(2+\delta')/\delta'})$ *with* $\delta' < \delta$. *Then for* $c = 2,\ldots,m$, *it holds that (i)* $E(U_{nc}^2) = O(n^{-1-\gamma})$ *and (ii)* $E(V_{nc}^2) = O(n^{-1-\gamma})$ *where* $\gamma = \min\left\{1, \frac{2(\delta-\delta')}{\delta'(2+\delta)}\right\}$.

Note that it is in fact an error in Yoshihara (1976) to define $\gamma = \frac{2(\delta-\delta')}{\delta'(2+\delta)}$ since even in the ideal case where $\mathbf{A}_i$ are independently and identically distributed, the rate for $E\|\mathbf{U}_{nc}\|^2$ is only n^{-2}. Thus it must hold that $\gamma < 1$.

Lemma B.6 *(Fan & Yao (2003), Proposition 2.7) Let* $\{X_t\}$ *be strictly stationary and* α*-mixing with* $E(X_t) = 0$. *Put* $S_n = \sum_{t=1}^{n} X_t$. *Then if for some* $\delta > q$ *the conditions* $E|X_t|^{\delta} < \infty$ *and* $\alpha(l) = O(l^{-\delta q/(2(\delta-q))})$ *are satisfied, it holds that* $E(S_n^q) = O(n^{q/2})$.

References

AHN, S.K. (1997). Inference of vector autoregressive models with cointegration and scalar components. *Journal of the American Statistical Association*, **93**, 350–356. 34

BAI, J. (2003). Inferential theory for factor models of large dimensions. *Econometrica*, **71**, 135–171. 33, 41

BENKO, M., HARDLE, W. & KNEIP, A. (2009). Common functional principle components. *Annals of Statistics*, **37**, 1–34. 8

BESSE, P. & RAMSAY, J.O. (1986). Principle components analysis of sampled functions. *Psychometrika*, **51**, 285–311. 3

BOSQ, D. (2000). *Linear Processes in Function Spaces*. Springer-Verlag, New York. 3, 73, 74

BOX, G. & TIAO, G. (1977). A canonical analysis of multiple time series. *Biometrika*, **64**, 355–365. 33

CONT, R. & DA FONSECA, J. (2002). Dynamics of implied volatility surfaces. *Quantitative Finance*, **2**, 45–60. 54

DAUXOIS, J., POUSSE, A. & ROMAIN, Y. (1982). Asymptotic theory for the principle component analysis of a vector random function: some applications to statistical inference. *Journal of Multivariate Analysis*, **12**, 136–154. 3

DELAIGLE, A. & HALL, P. (2009). Defining a probability density for a distribution of random functions. *Annals of Statistics*, forthcoming. 10

DONOHO, D.L. (2000). High dimensional data analysis: The curses and blessings of dimensionality. Aide-Memoire of a lecture at the American Mathematical Society conference on Math Challenges of the 21st Century. 33

DUNFORD, N. & SCHWARTZ, J.T. (1988). *Linear Operators*. Wiley, New York. 71

ENGLE, R. & WATSON, M. (1981). A one-factor multivatiate time series model of metropolitan wage rates. *Journal of the American Statistical Association*, **76**, 774–781. 33

FAN, J. & LI, R. (2006). Statistical challenges with high dimensionality: feature selecton in knowledge discovery. In *Proceedings of the International Congress of Mathematicians*, 595–622, Zurich, Switzerland. 33

FAN, J. & YAO, Q. (2003). *Nonlinear Time Series: Nonparametric and Parametric Methods*. Springer, New York. 75

FAN, J., FAN, Y. & LV, J. (2008). High dimensional covariance matrix estimation using a factor model. *Journal of Econometrics*, **147**, 186–197. 41

FENGLER, M., HARDLE, W. & MAMMEN, E. (2007). A dynamic semiparametric factor model for implied volatility string dynamics. *Journal of Financial Econometrics*, **5**, 189–218. 54

FERRATY, F. & VIEU, P. (2006). *Nonparametric Functional Data Analysis*. Springer, New York. 3

FORNI, M., HALLIN, M., LIPPI, M. & REICHLIN, L. (2000). The generalized dynamic factor model: identification and estimation. *Review of Economics and Statistics*, **82**, 540–554. 33

GUILLAS, S. & LAI, M.J. (2008). Approximation of functional spatial regression models with bivariate splines. Working Paper. 3

HALL, P. & HECKMAN, N. (2002). Estimating and depicting the structure of a distribution of random functions. *Biometrika*, **89**, 145–158. 10

HALL, P. & HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *Journal of the Royal Statistical Society, Series B*, **68**, 109–126. 25

HALL, P. & HOSSEINI-NASAB, M. (2008). Theory for high-order bounds in functional principal components analysis. *Mathematical Proceedings of the Cambridge Philosophical Societs*, **146**, 225. 25

HALL, P. & VIAL, C. (2006). Assessing the finite dimensionality of functional data. *Journal of the Royal Statistical Society, Series B*, **68**, 689–705. 2, 3, 5

HALL, P., MULLER, H.G. & WANG, J.L. (2006). Properties of principle component methods for functional and longitudinal data analysis. *Annals of Statistics*, **34**, 1493–1517. 3, 25

HAREL, M. & PURI, M.L. (1989). Limiting behaviour of u-statistics, v-statistics and one sample rank order statistics for nonstationary absolutely regular process. *Journal of multivariate analysis*, **30**, 181–204. 44

KNEIP, A. & UTIKAL, K.J. (2001). Inference for density families using functional principal component analysis (with discussion). *Journal of the American Statistical Association*, **96**, 519–542. 8

LAM, C. & YAO, Q. (2009). Estimation of large latent factor models for time series data. Working Paper. 33, 34, 41

LEE, A.J. (1990). *U-Statistics*. Marcel Dekker, New York. 43, 65

LI, W.K. & MCLEOD, A.I. (1981). Distribution of the residual autocorrelations in multivariate ARMA time series models. *Journal of the Royal Statistical Society, Series B*, **43**, 231–239. 21

LUTKEPOHL, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer Verlag, Berlin. 40

MAS, A. (2000). *Estimation of correlation operators for functional data: limiting distributions, tests and moderate deviations*. Ph.D. thesis, Universite Pierre et Marie Curis - Paris VI. 74

MERLEVEDE, F., PELIGRAD, M. & UTEV, S. (1997). Sharp conditions for the clt of linear processes in a hilbert space. *Journal of Theoretical Probability*, **10**, 681–693. 11, 73

PAN, J. & YAO, Q. (2008). Modelling multiple time series via common factors. *Biometrika*, **95**, 365–379. 6, 33, 34, 37

PARK, B., MAMMEN, E., HARDLE, W. & BORAK, S. (2009). Dynamic semi-parametrix factor models. *Journal of the American Statistical Association*, forthcoming. 54

PENA, D. & BOX, E.P. (1987). Identifying a simplifying structure in time series. *Journal of the American Statistical Associtation*, **82**, 836–843. 6, 33

PENA, D. & PONCELA, P. (2006). Nonstationary dynamic factor analysis. *Journal of Statistical Planning and Inference*, **136**, 1237–1257. 34

PREISTLEY, M.B., RAO, T. & TONG, H. (1974). Applications of principal component analysis and factor analysis in the identification of multivariate systems. *IEEE Transactions on Automatic Control*, **19**, 730–734. 33

RAMSAY, J.O. & DALZELL, C.J. (1991). Some tools for functional data analysis (with discussion). *Journal of the Royal Statistical Society, Series B*, **53**, 539–572. 3

RAMSAY, J.O. & SILVERMAN, B.W. (2005). *Functional Data Analysis*. Springer, New York. 3, 5, 8

RICE, J.A. & SILVERMAN, B.W. (1991). Estimating the mean and covariance structure nonparametrically when the data are curves. *Journal of the Royal Statistical Society, Series B*, **53**, 233–243. 3

STOCK, J.H. & WATSON, M.W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, **97**, 1167–1179. 33

TIAO, G.C. & TSAY, R.S. (1989). Model specification in multivariate time series (with discussion). *Journal of the Royal Statistical Society, Series B*, **51**, 157–213. 7, 33, 39

YOSHIHARA, K. (1976). Limiting behaviour of u-statistics for stationary absolutely regular processes. *Probability Theory and Related Fields*, **35**, 237–252. 44, 74, 75

ZHENGYAN, L. & LU, C. (1997). *Limit Theory for Mixing Dependent Random Variables*. Springer, New York. 41