

# Curve Forecasting by Functional Autoregression

V. Kargin\*(Courant Institute of Mathematical Sciences) and A. Onatski†(Columbia University)

February 1, 2007

## Abstract

This paper deals with the prediction of curve-valued autoregression processes. It develops a novel technique, the predictive factor decomposition, for the estimation of the autoregression operator. The technique is based on finding a reduced-rank approximation to the autoregression operator that minimizes the expected squared norm of the prediction error.

Implementing this idea, we relate the operator approximation problem to the singular value decomposition of a combination of the cross-covariance and covariance operators. We develop an estimation method based on regularization of the empirical counterpart of this singular value decomposition, prove its consistency and evaluate convergence rates.

The method is illustrated by an example of the term structure of the Eurodollar futures rates. In the sample corresponding to the period of normal growth, the predictive factor technique outperforms the principal components method and performs on par with the custom-designed prediction methods.

KEY WORDS: Functional data analysis; Dimension reduction; Reduced-rank regression; Principal component; Singular Value Decomposition; Predictive factor; Term structure; Interest rates

---

\*kargin@cims.nyu.edu; 109-20 71st Road Apt. 4A, Forest Hills, NY 11375

†ao2027@columbia.edu; Economics Department, Columbia University, 1011 IAB, 420 West 118th St., New York, NY 10027

# 1 Introduction

As evidenced in books by Ramsay and Silverman (1997 and 2002), statistical analysis increasingly relies on functional data. For example, Cavallini et al. (1994), Besse and Cardot (1996), Besse et al. (2000), Bernard (1997), and Damon and Guillas (2002) study the forecasting of electricity consumption, traffic, climatic variations, electrocardiograms, and ozone concentration, respectively, using a generalization of the autoregression model to functional data. Specifically, the data generating process in these studies is assumed to be the autoregressive Hilbertian process of order 1 extensively studied by Bosq (2000):

$$f_{t+1} = \rho[f_t] + \varepsilon_{t+1}. \quad (1)$$

Here for each integer  $t$ ,  $f_t$  is a random element of a Hilbert space  $H$ ,  $\rho$  is a linear bounded operator on  $H$  and  $\varepsilon_t$  is a strong H-white noise.

Similarly to the papers cited above, the focus of this paper is forecasting. We seek to forecast the future value  $f_{n+1}$  of the functional autoregression process (1) as a function of data  $f_1, \dots, f_n$ , so as to minimize the mean squared forecast error:

$$\min_{\hat{f}_{n+1}} E \left\| f_{n+1} - \hat{f}_{n+1}(f_1, f_2, \dots, f_n) \right\|^2. \quad (2)$$

As follows from Theorem 3.1 of Bosq (2000) and our assumption that  $\varepsilon_t$  is a *strong* H-white noise, the conditional expectation of  $f_{n+1}$  given  $f_1, f_2, \dots, f_n$  equals  $\rho[f_n]$ . Hence, for known  $\rho$ , the best predictor is given by  $\rho[f_n]$ . However, typically,  $\rho$  is unknown and this solution is infeasible. A feasible approximation to the solution is  $\hat{\rho}[f_n]$ , where  $\hat{\rho}$  is a consistent estimator of  $\rho$ . The main difficulty in estimating  $\rho$  is its infinite dimensionality. A standard way to overcome this difficulty is to estimate

the action of  $\rho$  only along a few chosen directions in  $H$ .

In practice, the usual directions chosen are those of a few principal components, although some other choices, including those based on wavelet bases, were suggested in the literature (Antoniadis and Sapatinas (2003)). However, the direction choices based on principal components or wavelet bases are not justified directly by efficiency in the problem of prediction. In this paper we suggest a dimension-reduction technique that is derived directly from the minimization of the prediction error. We call this technique predictive factors.

The main idea of the predictive factors is to focus on estimation of those linear functionals of the data that can most reduce the expected squared error of prediction. In particular we are looking for a  $k$ -rank operator  $\rho_k$  that minimizes  $E \|(\rho - \rho_k) f_n\|^2$ . Such an operator can be decomposed as  $\rho_k = \sum_{i=1}^k |a_i\rangle \langle b_i|$  and under certain normalization the decomposition is essentially unique<sup>1</sup> and has an interpretation in terms of the prediction problem. Roughly speaking, the number  $\langle b_1, f_n \rangle$ , the first *predictive factor*, contains the most information about the future value of the curve. The *predictive factor loading*  $a_1$  determines the direction along which this factor predicts. Similar interpretation holds for  $b_i$  and  $a_i$  with  $i > 1$ . The decomposition  $\rho_k = \sum_{i=1}^k |a_i\rangle \langle b_i|$  is related to the singular value decomposition of a combination of the covariance operators of process (1). We suggest estimating  $\rho_k$  by estimating the first  $k$  terms of this singular value decomposition.

Our results are collected in Theorems 1, 2, 3, and 4. Theorem 1 relates the population predictive factors to the components of the singular value decomposition of a certain combination of the covariance operators of process (1). Theorem 2 shows that, as the rank of the predictive factor approximation grows, the approximation

---

<sup>1</sup>Non-uniqueness comes from the fact that  $a_i$  and  $b_i$  can be both multiplied by  $-1$  and the resulting  $\rho_k$  will not change. The decomposition is also non-unique if an eigenvalue of a certain operator has a multiplicity greater than 1.

converges to the true operator,  $\rho_k \rightarrow \rho$ . The convergence here can be understood in many different ways, which are made precise in Theorem 2.

Next, we suggest estimating  $\rho_k$  by finding a truncated singular value decomposition  $\sum_{i=1}^k \hat{\sigma}_{\alpha,i} |\hat{y}_{\alpha,i}\rangle \langle \hat{x}_{\alpha,i}|$  of the operator  $\hat{F} \left( \hat{\Gamma} + \alpha I \right)^{-1/2}$ . Here  $\hat{F}$  and  $\hat{\Gamma}$  denote certain empirical covariance operators computed from data, and  $\alpha$  and  $k$  are regularization and truncation parameters, which can be adapted to data by using cross-validation methods. For prediction, we compute the scalar products of  $f_n$  with  $\hat{b}_{\alpha,1}, \dots, \hat{b}_{\alpha,k}$ , defined as  $\left( \hat{\Gamma} + \alpha I \right)^{-1/2} \hat{x}_{\alpha,1}, \dots, \left( \hat{\Gamma} + \alpha I \right)^{-1/2} \hat{x}_{\alpha,k}$ , respectively. Finally, we map  $f_n$  to  $\sum_{i=1}^k \langle f_n, \hat{b}_{\alpha,i} \rangle \hat{F} \hat{b}_{\alpha,i}$ . The result is  $\hat{f}_{n+1} = \hat{\rho}_{\alpha,k} f_n$ , that is, a feasible predictor of  $f_{n+1}$  based on  $f_1, f_2, \dots, f_n$ .

Theorem 3 proves that with a certain choice of the regularization parameters  $\alpha$  and  $k$ , the operator  $\hat{\rho}_{\alpha,k}$  is a consistent estimator of  $\rho$ . Theorem 4 shows that predictor  $\hat{\rho}_{\alpha,k} [f_n]$  of  $f_{n+1}$  converges to the optimal infeasible predictor  $\rho [f_n]$ . We also provide convergence rates.

Results by Bosq (2000) and our Theorem 3 show that both the principal components and the predictive factors methods provide consistent estimators of  $\rho$ . The rates of convergence in these methods are difficult to compare because they depend on intricate interrelations of the operator  $\rho$  and the covariance of noise. The intuitive reason why the predictive factors may be better suited for forecasting than the principal components is that the predictive factors are designed to extract the information useful for prediction while the principal components may focus on features of  $f_t$  that are poorly predictable.

For example, in economics, while it is true that more than 95% percent of the variation in the yield curve of bonds can be explained by the first three principal components, recent research (Cochrane and Piazzesi (2002)) suggests that some components that do not contribute much to the overall interest rate variation are

better predictors of interest rate movements. When compared to the principal components, the predictive factors are less likely to miss these better predictors.

Let us describe some related work. The predictive factor technique is similar to the simultaneous linear predictions introduced in the static finite-dimensional context by Fortier (1966). The idea is to find uncorrelated linear combinations of predictors that have the largest, the second largest, etc., cumulative explanatory power for a set of predicted variables. The same idea has been reintroduced under the name of redundancy analysis by van den Wollenberg (1977). For the time series data, the predictive factor method extends the reduced-rank autoregression studied by Reinsel (1983). This extension is similar to the extension of the classical canonical correlation analysis to the functional data performed by Leurgans et al. (1993). The estimation of operator  $\rho$  in functional autoregression was considered by Bosq (2000) and Mas (1999). They focus mainly on the convergence of the principal components method.

As an illustration, we apply the predictive factor method to data on Eurodollar futures contracts. At each trading date the available contracts can be arranged by their delivery date from one month to 10 years into the future. Then, the term structure curve is obtained by plotting the rate of return on the contracts against the delivery time and interpolating by cubic splines. We analyze the time evolution of these curves.

Restricting the sample to the period of normal economic growth and forecasting three months into the future, we find that the predictive factor technique not only outperforms the principal components method but also performs on par with the best available prediction methods.

The rest of the paper is organized as follows. The outline of the predictive factor analysis and our main results are in Section 2. The empirical example is in

Section 3. Section 4 concludes. Proofs of the four main theorems are relegated to the Appendix.

## 2 The predictive factor analysis

Let  $H = L^2[0, \bar{X}]$  be the real Hilbert space of the square-summable functions of  $x \in [0, \bar{X}]$ , where  $0 < \bar{X} < \infty$ . For any  $g, h \in H$ , we will denote their scalar product  $\int_0^{\bar{X}} g(x) h(x) dx$  as  $\langle g, h \rangle$ . We will denote the norm of  $g$ ,  $\langle g, g \rangle^{1/2}$ , as  $\|g\|$ . For any bounded linear operator  $A$  acting on  $H$ , we will denote its uniform norm  $\sup_{\|g\|=1} \|Ag\|$  as  $\|A\|$ , and its adjoint operator as  $A'$ .

Let  $\mathcal{B}_H$  be the Borel  $\sigma$ -algebra of subsets of  $H$  and let  $(\Omega, \mathcal{F}, P)$  be a probability space. We will call any  $\mathcal{F} - \mathcal{B}_H$  measurable mapping from  $\Omega$  to  $H$  an  $H$ -valued random variable. It can be shown (see Bosq, 2000, Lemma 1.2) that  $\xi$  is an  $H$ -valued random variable if and only if  $\langle g, \xi \rangle$  is a real random variable for any  $g \in H$ . Moreover, it is possible to introduce the following definitions:

**Definition 1** *If an  $H$ -valued random variable  $\xi$  is such that  $E \|\xi\| < \infty$ , then there exists an element of  $H$ , denoted as  $\mathcal{E}(\xi)$  and called the **expectation** of  $\xi$ , such that  $E \langle g, \xi \rangle = \langle g, \mathcal{E}(\xi) \rangle$ , for any  $g \in H$ .*

**Definition 2** *If an  $H$ -valued random variable  $\xi$  is such that  $E \|\xi\|^2 < \infty$  and  $\mathcal{E}(\xi) = 0$ , then the **covariance operator** of  $\xi$  is defined by  $C_\xi : g \rightarrow \mathcal{E}(\langle g, \xi \rangle \xi)$  for any  $g \in H$ .*

**Definition 3** *If  $H$ -valued random variables  $\xi$  and  $\eta$  are such that  $E \|\xi\|^2 < \infty$ ,  $E \|\eta\|^2 < \infty$ , and  $\mathcal{E}(\xi) = \mathcal{E}(\eta) = 0$ , then the **cross-covariance operator** of  $\xi$  and  $\eta$  is defined by  $C_{\xi, \eta} : g \rightarrow \mathcal{E}(\langle g, \xi \rangle \eta)$  for any  $g \in H$ .*

It is known that the covariance and cross-covariance operators are trace-class operators (that is, that they are compact and their singular values are absolutely summable). See for details Section 1.5 in Bosq (2000).

**Definition 4** *A sequence  $\{\varepsilon_t, t \in Z\}$  of  $H$ -valued random variables is said to be a strong **H-white noise** if  $0 < E \|\varepsilon_t\|^2 = \sigma^2 < \infty$ ,  $\mathcal{E}(\varepsilon_t) = 0$ , and  $\varepsilon_t$  are i.i.d.*

In this paper we assume that data have a form of  $n$  observations  $\{f_1, \dots, f_n\}$  of  $H$ -valued random variables that belong to a strictly stationary sequence  $\{f_t, t \in Z\}$ , which satisfies Assumption 1 formulated below. We study the problem of prediction of  $f_{n+1}$  based on the data.

**Notation 1** *Let us denote the covariance operator  $C_{f_t}$  of  $f_t$  as  $\Gamma$  and the cross-covariance operator  $C_{f_t, f_{t+1}}$  of  $f_t$  and  $f_{t+1}$  as  $F$ .*

**Assumption 1** *The sequence  $\{f_t, t \in Z\}$  satisfies a functional auto-regression equation (1), and there exists an integer  $j \geq 1$  such that  $\|\rho^j\| < 1$ . Furthermore,  $E \|f_t\|^4 < \infty$ .*

Sections 4.1 and 4.3 in Bosq (2000) show that if Assumption 1 holds, then the covariance and cross-covariance operators  $\Gamma$  and  $F$  can be consistently estimated by their respective sample versions  $\hat{\Gamma}$  and  $\hat{F}$ , which are defined as follows:

**Definition 5** *The empirical covariance and cross-covariance operators of the sequence  $\{f_t, t \in Z\}$  are  $\hat{\Gamma} : g \rightarrow \frac{1}{n} \sum_{t=1}^n \langle g, f_t \rangle f_t$ , and  $\hat{F} : g \rightarrow \frac{1}{n-1} \sum_{t=1}^{n-1} \langle g, f_t \rangle f_{t+1}$ , respectively, where  $n$  is the number of observations.*

Let us turn to the subject of most interest to us. Denote by  $\mathfrak{R}_k$  the set of all finite-rank operators acting on  $H$ . We would like to find an operator  $A \in \mathfrak{R}_k$ , approximating  $\rho$ , that minimizes  $E \|f_{t+1} - Af_t\|^2$ . It is easy to see that this problem

is equivalent to the problem of finding an  $A \in \mathfrak{R}_k$  that minimizes  $E \|(\rho - A) f_t\|^2$ . Further, Formula 1.59 in Bosq (2000) implies that this latter problem can be reduced to the following:

$$\min_{A \in \mathfrak{R}_k} \|\rho - A\|_{\Gamma, 2}. \quad (3)$$

Here  $\|X\|_{\Gamma, 2}$  denotes a modified Hilbert-Schmidt norm, which is defined as  $\|X\|_{\Gamma, 2}^2 := \text{tr}(X\Gamma X')$ . Note that  $\|X\|_{\Gamma, 2}$  is in general a seminorm. It is a norm if  $\text{Ker } \Gamma = 0$ .

In this form, it is obvious that our problem is a problem of approximation of a given operator by a finite-rank operator relative to a specific operator seminorm. For infinite-dimensional integral operators and the usual Hilbert-Schmidt norm (i.e., the norm  $\|X\|_2^2 := \text{tr}(X'X)$ ), this problem was solved by Schmidt (1907). For the finite-dimensional case this solution was independently rediscovered by Eckart and Young (1936). (See also Problem 10.6, p. 458, in Eaton (1983) with the solution outlined on p. 497.)

The Schmidt-Eckart-Young solution is given by the  $k$ -element partial sum of the singular value decomposition (SVD) of the operator, which is being approximated. The main difference between our problem and this classical case is that the operator  $\Gamma$  which modifies the classical norm does not have a bounded inverse. This difference can be illustrated by the fact that even some non-compact operators (e.g., the identity operator) can be approximated by finite-rank operators in the modified seminorm, while such an approximation is impossible in the usual uniform or Hilbert-Schmidt norm.

We solve problem (3) in two steps.

### **Step 1. Solving a Classical Approximation Problem**

We first solve the following problem

$$\min_{X \in \mathfrak{R}_k} \left\| \rho\Gamma^{1/2} - X \right\|_2. \quad (4)$$

The classical SVD-based solution is valid for such an approximation problem. Indeed, since  $\Gamma$  is a trace-class operator,  $\Gamma^{1/2}$  is a Hilbert-Schmidt operator. Then,  $\rho\Gamma^{1/2}$  is also a Hilbert-Schmidt operator because Hilbert-Schmidt operators form a two-sided ideal in the algebra of all linear bounded operators (see Chapter III in Gohberg and Krein, 1969). Since any Hilbert-Schmidt operator in  $L^2[0, \bar{X}]$  can be represented in the form of an integral operator (see Gohberg and Krein, 1969, p. 111), Schmidt's (1907) solution to the approximation problem is valid. Hence,

$$\min_{X \in \mathfrak{R}_k} \left\| \rho\Gamma^{1/2} - X \right\|_2 = \left\| \rho\Gamma^{1/2} - r_k \right\|_2,$$

where  $r_k$  is the  $k$ -element partial sum of a singular value decomposition of  $\rho\Gamma^{1/2}$ .

Recall that a singular value decomposition of  $\rho\Gamma^{1/2}$  is constructed as follows (see Gohberg and Krein, 1969, p. 28). Let  $\rho\Gamma^{1/2} = U\Phi^{1/2}$  be the polar representation of  $\rho\Gamma^{1/2}$ , where  $\Phi := \Gamma^{1/2}\rho'\rho\Gamma^{1/2}$  and  $U$  is a partially isometric operator, which maps  $(\text{Ker } \Phi^{1/2})^\perp$  isometrically onto the range of  $\rho\Gamma^{1/2}$  and sends any function from  $\text{Ker } \Phi^{1/2}$  to zero. Then the eigenvalues of  $\Phi^{1/2}$   $\sigma_1 \geq \sigma_2 \geq \dots$  (represented in this inequality chain as many times as their multiplicities) are called the singular values of  $\rho\Gamma^{1/2}$ . Let  $\{x_i, i = 1, \dots, \text{rank}(\Phi^{1/2})\}$  be an orthonormal system of eigenfunctions of  $\Phi^{1/2}$  corresponding to all non-zero eigenvalues  $\{\sigma_i, i = 1, \dots, \text{rank}(\Phi^{1/2})\}$ . Note that  $\text{rank}(\Phi^{1/2}) = \text{rank}(\Phi)$  and that this rank is allowed to be infinite. Note also that  $x_i$  are eigenfunctions of  $\Phi$ , corresponding to its eigenvalues  $\sigma_i^2$ . Let  $\{y_i, i = 1, \dots, \text{rank}(\Phi)\}$  be an orthonormal system of functions

defined as  $y_i := Ux_i$ . Then a singular value decomposition of  $\rho\Gamma^{1/2}$  has the form  $\rho\Gamma^{1/2} = \sum_{i=1}^{\text{rank}(\Phi)} \sigma_i |y_i\rangle \langle x_i|$ , which means that  $\rho\Gamma^{1/2} : g \rightarrow \sum_{i=1}^{\text{rank}(\Phi)} \sigma_i \langle g, x_i\rangle y_i$  for any  $g \in H$ .

Therefore, Schmidt's (1907) solution to (4) is given by the operator

$$r_k = \sum_{i=1}^{\min\{k, \text{rank}(\Phi)\}} \sigma_i |y_i\rangle \langle x_i|. \quad (5)$$

Moreover,

$$\left\| \rho\Gamma^{1/2} - r_k \right\|_2^2 = \varphi(k), \quad (6)$$

where  $\varphi(k)$  is defined as  $\varphi(k) = \sum_{i=k+1}^{\infty} \sigma_i^2$ . (Note that since  $\rho\Gamma^{1/2}$  is a Hilbert-Schmidt operator, the sum  $\sum_{i=1}^{\infty} \sigma_i^2$  converges, and consequently function  $\varphi(k)$  converges to zero as  $k \rightarrow \infty$ .) For a modern proof of (5) and (6) see Lemma 6.1 of Gohberg and Krein (1969), p. 87.

### Step 2: Transforming the SVD Solution

Note that the operator  $r_k$  defined by (5) can be represented as  $\rho_k\Gamma^{1/2}$ , where  $\rho_k \in \mathfrak{R}_k$ . Indeed, define  $\rho_k$  as

$$\rho_k = \sum_{i=1}^{\min\{k, \text{rank}(\Phi)\}} \sigma_i |y_i\rangle \langle \Gamma^{-1/2}x_i|. \quad (7)$$

This definition makes sense because, for any  $i = 1, 2, \dots, \text{rank}(\Phi)$ ,  $x_i$  is an eigenfunction of  $\Phi$  corresponding to a non-zero eigenvalue  $\sigma_i^2$ , and therefore it belongs to the range of  $\Gamma^{1/2}$ . Hence, for any  $i = 1, 2, \dots, \text{rank}(\Phi)$ ,  $\Gamma^{-1/2}x_i$  is well defined and

equal to  $\sigma_i^{-2}\rho'\rho\Gamma^{1/2}x_i$ . Now, for any  $g \in H$ , we have:

$$\begin{aligned}\rho_k\Gamma^{1/2}g &= \sum_{i=1}^{\min\{k,\text{rank}(\Phi)\}} \sigma_i \left\langle \Gamma^{1/2}g, \Gamma^{-1/2}x_i \right\rangle y_i \\ &= \sum_{i=1}^{\min\{k,\text{rank}(\Phi)\}} \sigma_i \left\langle g, \Gamma^{1/2}\Gamma^{-1/2}x_i \right\rangle y_i = r_k g.\end{aligned}$$

Hence,  $r_k = \rho_k\Gamma^{1/2}$ , which, together with the fact that  $r_k$  is a solution to (4), implies that  $\rho_k$  is a solution to (3).

Our first Theorem shows that the solution of (3),  $\rho_k$ , is a function of the covariance and cross-covariance operators of process (1). Such a representation suggests an estimation strategy that we follow in Theorems 3 and 4 below.

**Theorem 1** *Under Assumption 1, the operator  $\rho_k$  can be represented in the form*

$$\rho_k = \sum_{i=1}^{\min\{k,\text{rank}(\Phi)\}} |a_i\rangle \langle b_i|. \quad (8)$$

Here  $a_i = Fb_i$  and  $b_i = \Gamma^{-1/2}x_i$ , where  $i = 1, \dots, \text{rank}(\Phi)$ , and  $\{x_i, i = 1, \dots, \text{rank}(\Phi)\}$  is an orthonormal system of eigenfunctions of  $\Phi$  corresponding to the eigenvalues  $\{\sigma_i^2, i = 1, \dots, \text{rank}(\Phi)\}$ , counted with multiplicity. Operator  $\Phi$  coincides with operator  $\Gamma^{-1/2}F'F\Gamma^{-1/2}$  on the range of  $\Gamma^{1/2}$ .

Recall that  $\rho_k[f_n]$  is a predictor of  $f_{n+1}$  which minimizes the expected squared error of prediction in the class of predictors of the form  $Af_n$ ,  $A \in \mathfrak{R}_k$ . Theorem 1 tells us that such an optimal (infeasible) predictor has the form  $\sum_{i=1}^{\min\{k,\text{rank}(\Phi)\}} \langle f_n, b_i \rangle a_i$ . In this sense, if we want to summarize the infinite-dimensional dynamics of the functional autoregression by the dynamics of no more than  $k$  univariate processes, the best summary would be provided by processes  $\{\langle f_t, b_i \rangle, i = 1, \dots, \min\{k, \text{rank}(\Phi)\}\}$ . These processes would approximate the infinite-dimensional dynamics along the di-

reactions given by  $\{a_i, i = 1, \dots, \min\{k, \text{rank}(\Phi)\}\}$ .

**Definition 6** *We will call the processes  $\{\langle f_t, b_i \rangle, i = 1, \dots, \text{rank}(\Phi)\}$  the **predictive factors** and the functions  $\{a_i, i = 1, \dots, \text{rank}(\Phi)\}$  the corresponding **predictive loadings** for the functional autoregressive process (1).*

Unfortunately, the predictive factors and the predictive loadings are not uniquely determined. This non-uniqueness is a consequence of the fact that the eigenfunctions  $x_i$ , on which the definition of  $b_i$  is based, are not unique. First, multiplying  $x_i$  by  $-1$ , we still get an eigenfunction of norm one. Second, if the multiplicity of some of the non-zero eigenvalues of the symmetric operator  $\Phi$  is larger than one, the choice of the corresponding eigenfunctions is essentially non-unique. We could eliminate the first cause of the ambiguity by requiring that the first non-zero coordinate of  $b_i$  in a particular basis in  $H$ , say in the basis that consists of functions  $1, \cos(2\pi x/\bar{X}), \sin(2\pi x/\bar{X}), \cos(4\pi x/\bar{X}), \sin(4\pi x/\bar{X})$ , and so on, is positive. However, then the second cause of the non-uniqueness still remains.

Instead of artificially fixing the ambiguity, we leave the predictive factors non-uniquely defined. We note, however, that in the case when all non-zero eigenvalues of  $\Phi$  have multiplicity one, which is a generic case, the predictive factors and the corresponding predictive loadings can be uniquely defined relative to a particular basis in  $H$ , as explained above. Furthermore, if  $\sigma_k \neq \sigma_{k+1}$ , the operator  $\rho_k$  is uniquely defined without any reference to a particular basis. Since predictive factors and the corresponding loadings enter the definition of  $\rho_k$  multiplicatively, their non-uniqueness due to the freedom of choice of the sign does not create non-uniqueness for  $\rho_k$ .

Our next theorem describes the quality of the approximation of  $\rho$  provided by  $\rho_k$ .

**Theorem 2** *Suppose Assumption 1 holds. Then, as  $k \rightarrow \infty$ , we have:*

(a)  $\|\rho - \rho_k\|_{\Gamma, 2}^2 = \varphi(k) \rightarrow 0$ ;

(b) *If  $f_t$  is from process (1), then  $E \|(\rho - \rho_k) f_t\|^2 = \varphi(k) \rightarrow 0$ . If, in addition,  $\varphi(k) = O(k^{-1-\theta})$  for a  $\theta > 0$ , then  $\|(\rho - \rho_k) f_t\| = O(k^{-\theta/2+\varepsilon})$  almost surely for every  $\varepsilon > 0$ ;*

(c) *If  $\text{Ker } \Gamma = 0$ , then  $\|(\rho - \rho_k) f\| \rightarrow 0$  for every  $f \in H$ , and*

(d) *If  $\text{Ker } \Gamma = 0$  and  $\rho$  is compact, then  $\|\rho - \rho_k\| \rightarrow 0$ .*

We are mostly interested in how the operator approximation behaves on functions drawn from process (1) as opposed to arbitrary functions from  $L^2(0, \bar{X})$ . This is important because the approximation  $\rho_k$  has no chance to converge to  $\rho$  in the uniform norm in those cases when  $\rho$  is non-compact. This is simply because non-compact operators cannot be approximated by operators of finite rank in the uniform norm. That is, for any sequence of approximations  $\rho_k$  we can find a sequence of functions  $f_k$  such that  $\|(\rho - \rho_k) f_k\| > \varepsilon$ . However, property (b) says that with respect to process (1) the convergence is uniform. This means that for any positive  $\varepsilon$  and  $\delta$  we can find such  $K$  that for any  $k > K$ ,  $\|(\rho - \rho_k) f_t\| < \varepsilon$  with a probability greater than  $1 - \delta$ .

Property (b) also provides an estimate for the rate of convergence and indicates that the speed of the convergence is determined mainly by the speed of decline of the eigenvalues of operator  $\Phi = \Gamma^{1/2} \rho' \rho \Gamma^{1/2}$ . This can be interpreted as saying that the speed of convergence is determined by interaction of the noise and the coefficient operator.

From Theorems 2 and 1, it seems natural to estimate  $\rho$  by computing  $k$  eigenfunctions of the operator  $\hat{\Gamma}^{-1/2} \hat{F}' \hat{F} \hat{\Gamma}^{-1/2}$ , which is well-defined on the subspace of  $H$  spanned by  $\{f_1, \dots, f_n\}$ , and then use a sample analog of  $\rho_k$  as an estimate. We

might hope that when  $k$  approaches infinity as the sample size grows, such an estimator converges in some sense to  $\rho$ . Unfortunately, this is not so. The reason is that even though  $\hat{F}$  and  $\hat{\Gamma}$  converge to  $F$  and  $\Gamma$ , this does not imply that eigenfunctions of  $\hat{\Phi} := \hat{\Gamma}^{-1/2} \hat{F}' \hat{F} \hat{\Gamma}^{-1/2}$  converge to eigenfunctions of  $\Gamma^{-1/2} F' F \Gamma^{-1/2}$ .

We will prove that a way to get a consistent estimator of  $\rho$  is to use the following regularized version of the above estimation strategy. First, define  $\hat{\Gamma}_\alpha$  as  $\hat{\Gamma} + \alpha I$ , where  $\alpha$  is a positive real number. Next, compute  $k$  eigenfunctions  $\{\hat{x}_{\alpha,i}, i = 1, \dots, k\}$  of the operator  $\hat{\Phi}_\alpha := \hat{\Gamma}_\alpha^{-1/2} \hat{F}' \hat{F} \hat{\Gamma}_\alpha^{-1/2}$ , which correspond to its  $k$  largest eigenvalues  $\hat{\sigma}_{\alpha,1}^2 \geq \dots \geq \hat{\sigma}_{\alpha,k}^2$  counted with multiplicity. Finally, based on these eigenfunctions, construct a sample analog of  $\rho_k$  :

$$\hat{\rho}_{\alpha,k} = \sum_{i=1}^k |\hat{a}_{\alpha,i}\rangle \langle \hat{b}_{\alpha,i}|,$$

where  $\hat{b}_{\alpha,i} = \hat{\Gamma}_\alpha^{-1/2} \hat{x}_{\alpha,i}$  and  $\hat{a}_{\alpha,i} = \hat{F} \hat{b}_{\alpha,i}$ . The following Theorem shows that as  $\alpha$  decreases to zero and  $k$  increases to infinity at an appropriate rate, as the sample size grows, the estimator  $\hat{\rho}_{\alpha,k}$  approaches  $\rho$  in the seminorm  $\|\cdot\|_{\Gamma,2}$ .

**Theorem 3** *Let Assumption 1 hold and suppose that  $\rho$  is a Hilbert-Schmidt operator. Let  $\{\alpha_n\}$  be a sequence of positive real numbers such that  $\alpha_n \sim n^{-1/6}$  as  $n \rightarrow \infty$ , and let  $\{k_n\}$  be any sequence of positive integers such that  $n \geq k_n \geq Kn^{1/4}$  for some  $K > 0$ . Then, for all  $\beta > 1/2$  we have:*

$$\|\rho - \hat{\rho}_{\alpha_n k_n}\|_{\Gamma,2} = o\left(n^{-1/12} \log^\beta n\right) \text{ a.s.}$$

Theorem 3 shows that, under the additional assumption that  $\rho$  is a Hilbert-Schmidt operator, estimator  $\hat{\rho}_{\alpha_n, k_n}$  approaches  $\rho$  in the  $(\Gamma, 2)$ -seminorm at a particular rate. Note that without additional assumptions, even the exact solution to

problem (3),  $\rho_{k_n}$ , may converge to  $\rho$  arbitrarily slowly. Indeed, according to Theorem 2a), the  $(\Gamma, 2)$ -seminorm of  $\rho - \rho_{k_n}$  equals  $\varphi(k_n)$ , which may converge to zero arbitrarily slowly. The additional assumption that  $\rho$  is Hilbert-Schmidt makes  $\varphi(k_n)$  decay as fast as  $k_n^{-1}$ , which makes possible the non-trivial convergence result of Theorem 3.

Theorem 3 implies that  $E \left\| (\hat{\rho}_{\alpha_n, k_n} - \rho) f_t \right\|^2 \rightarrow 0$ , if  $f_t$  is a curve from an independent realization of process (1). Hence, when we have two independent and identically distributed functional autoregression processes (1), we can obtain an estimate of  $\rho$  based on the observations of the first process, and use this estimate to predict the future value of the second process. The mean squared error of such a prediction will converge to zero as the number of observations grows. In practice, however, we are interested in the situation when estimation and prediction are done using the same sample. The next theorem shows that predictor  $\hat{\rho}_{\alpha_n, k_n} [f_n]$  converges to the optimal infeasible predictor  $\rho [f_n]$  even if we use the same sample for both estimation and prediction.

**Theorem 4** *Under the assumptions of Theorem 3, for all  $\beta > 1/2$ , we have:*

$$E \left\| (\hat{\rho}_{\alpha_n, k_n} - \rho) f_n \right\|^2 = O \left( n^{-1/6} \log^{2\beta} n \right) \text{ as } n \rightarrow \infty \text{ a.s.}$$

### 3 Empirical Example

#### 3.1 Description of data.

We use daily settlement data on the Eurodollar futures contracts that we obtained from the Commodity Research Bureau. Each Eurodollar futures contract is an obligation to deliver a 3-month deposit of \$1,000,000 to a bank account outside of the United States at a specified time. The available contracts have monthly delivery

dates for the first six months after the current date, and then the delivery dates become quarterly up to 10 years into the future.

The available data start in 1982; however, we use only the data starting in 1994, when the trading in the 10-year contract appeared. We interpolated available data points by cubic splines to obtain smooth contract rate curves. To speed up the estimation, we restricted each curve to points that are 30 days apart. We also removed datapoints with fewer than 90 or more than 3,480 days to expiration. That left us with 114 points per curve and 2,507 valid dates. Figure 1 illustrates the evolution of the Eurodollar futures rate curves.

[Put Figure 1 here.]

**Figure 1** Eurodollar Futures Rates Evolution

Note: The time to maturity (in months) is on the left axis.

### **3.2** *Comparison of predictive factors with other methods*

We restrict our attention to the subsample corresponding to the normal growth period from 3-Jan-94 to 28-Feb-01. We hope that for this period, the functional autoregression describes the term structure dynamics reasonably well. Using this subsample, we compare the predictive performance of our method with those of four other methods. The first one is the same functional autoregression but estimated using the principal components technique. The second method is the random walk. The third method is the mean forecast, where the term structure three months ahead is predicted to be equal to the average past term structure. Finally, we consider the Diebold-Li forecasting procedure (2006), which was specifically designed for forecasting of the term structure of interest rates.

Before making predictions we have to choose the value of the regularization parameter  $\alpha$  and the number of predictive factors  $N_{PF}$  for the predictive factor

method, the number of principal components  $N_{PC}$  for the principal components method, and a parameter,  $\lambda$ , for the Diebold-Li method. We used the following cross-validation procedure to optimize our choice of these parameters. The first half of the subsample, that is, the period from 3-Jan-94 to 25-Jul-97, was considered as a learning subset. The optimal parameter values,  $\alpha=0.73$ ,  $N_{PF}=3$ ,  $N_{PC}=2$ , and  $\lambda=0.0147$ , minimized the mean squared error of the three-months-ahead pseudo-out-of-sample prediction for the next year, from 28-Jul-97 to 28-Jul-98.

The first five eigenvalues of the operator  $\hat{\Phi}_{0.73}$  are 37.12, 0.93, 0.04, 0.00, and 0.00. Recall that these eigenvalues can be interpreted as estimates of the reduction in the mean squared error of forecasting. We see that the error reduction due to the first predictive factor is much larger than the reduction corresponding to the other factors. The contributions of the fourth and fifth factors are essentially zero, which agrees well with our cross-validation choice  $N_{PF}=3$ .

[Put Figure 2 here.]

**Figure 2** “Weights” and Loadings of the First Three Predictive Factors

Figure 2 shows functions  $\hat{b}_{0.73,i}$ ,  $i = 1, 2, 3$  and  $\hat{a}_{0.73,i}$ ,  $i = 1, 2, 3$ , respectively. The shapes of the predictive factor loadings  $\hat{a}_{0.73,i}$  roughly correspond to the “level”, “slope”, and “curvature” factors typically used to describe the term structure. The “weights” of the predictive factors,  $\hat{b}_{0.73,i}$ , describe which linear combinations have the most predictive power. We see that the first predictive factor places relatively large “weights” on the contracts of short maturities. This fact is not surprising, as short-term interest rates are typically associated with the monetary policy stance, which strongly affects rates on the contracts of all maturities.

To assess the predictive performance of the alternative methods considered above, we run the following experiment. We first estimate the functional autoregressions and the Diebold-Li model using the pooled learning and cross-validation sample,

from 3-Jan-94 to 28-Jul-98, and make forecasts of the term structure three months ahead. The next step is to extend the first subsample to include one more day, re-estimate the models, and forecast the term structure three months ahead. We continue adding data to the first sample until we add the day which is three months before the end of the normal growth subsample.

Our measure of the predictive performance is the root mean squared error based on the difference between the actual term structure and the forecasted one. In Figure 3 we report curves of the root mean squared errors of the alternative methods.

[Put Figure 3 here.]

**Figure 3** Predictive Performances of Different Forecasting Methods

The thick-dashed line in Figure 3 corresponds to the Diebold-Li method. It outperforms all the other methods, which could be expected, as this model was custom-designed for this particular problem. The thick-solid line is for our predictive factors method. It is the second best method for contracts of maturities longer than 4 years and the third best, losing to the random walk (thick-dotted line), for shorter maturities. The thin-solid and -dashed lines correspond to the principal components method with  $N_{PC} = 2$  and  $N_{PC} = 3$ , respectively. The root mean squared forecast error for the principal components is uniformly worse than that for the predictive factors. We do not report the results for the mean prediction method because it worked much worse than the rest of the methods.

## 4 Conclusion

We have shown that prediction of function-valued autoregressive processes can benefit from a novel dimension-reduction technique, predictive factor decomposition. This technique differs from the usual principal components method by focusing on

the estimation of those linear combinations of variables that matter most for the prediction, as opposed to those that matter most for describing the variance. It turns out that the predictive factor approximation of the true autoregression operator can be consistently estimated using a regularization of a singular value decomposition problem.

In an empirical illustration we applied the new method to the interest rate curve dynamics. The results demonstrate that the new method is easy to estimate numerically and performs reasonably well. The predictive factors method not only outperforms the principal components method but also performs on par with the best of the other prediction methods.

## 5 Appendix

Throughout the Appendix, we use  $C$  to denote possibly different constants that may depend only on  $\rho$  and  $\Gamma$ .

### 5.1 Proof of Theorem 1

The predictive factor representation (8) of the operator  $\rho_k$  is a consequence of the representation (7) and the fact that  $x_i$  and  $y_i$  in representation (7) satisfy equality  $F\Gamma^{-1/2}x_i = \sigma_i y_i$ . This equality can be established as follows. Since  $F = \rho\Gamma$ , we have  $F\Gamma^{-1/2}x_i = \rho\Gamma^{1/2}x_i = U\Phi^{1/2}x_i = \sigma_i Ux_i = \sigma_i y_i$ .

Next, by definition,  $\Phi = \Gamma^{1/2}\rho'\rho\Gamma^{1/2}$ . Therefore, on the range of  $\Gamma^{1/2}$ ,  $\Phi = \Gamma^{-1/2}\Gamma\rho'\rho\Gamma^{-1/2} = \Gamma^{-1/2}F'F\Gamma^{-1/2}$ . QED.

## 5.2 Proof of Theorem 2

**Proof of Theorem 2a):** Statement a) of the Theorem follows from (6) and from the fact, established in Section 2, that  $r_k = \rho_k \Gamma^{1/2}$ . QED.

**Proof of Theorem 2b):** Since  $E \|\rho - \rho_k\|_{\Gamma, 2}^2 = \|\rho - \rho_k\|_{\Gamma, 2}^2$ , the first part of b) follows from a). For the second part, note that by Chebyshev's inequality and the assumption that  $\varphi(k) = O(k^{-1-\theta})$ :  $P\{\|\rho - \rho_k\|_{\Gamma, 2} \geq k^{-\theta/2+\varepsilon}\} \leq E \|\rho - \rho_k\|_{\Gamma, 2}^2 / k^{-\theta+2\varepsilon} \leq Ck^{-1-2\varepsilon}$  for all sufficiently large  $k$ . Using the Borel-Cantelli lemma, we get:  $\|\rho - \rho_k\|_{\Gamma, 2} = O(k^{-\theta/2+\varepsilon})$  a.s. as  $k \rightarrow \infty$ . QED.

**Proof of Theorem 2c):** We start with a useful Lemma. Let  $\{x_i, i = 1, 2, \dots\}$  extend a system of eigenfunctions  $\{x_i, i = 1, \dots, \text{rank}(\Phi)\}$  of operator  $\Phi^{1/2}$  corresponding to the non-zero eigenvalues to an orthonormal basis in  $H$ . Note that, for  $i > \text{rank}(\Phi)$ ,  $x_i \in \text{Ker}(\Phi^{1/2})$  and, therefore,  $x_i \in \text{Ker}(\rho \Gamma^{1/2})$ . Define an operator  $\Pi_k$  as  $\Pi_k = \sum_{i=1}^k |x_i\rangle \langle x_i|$ . Clearly,  $\Pi_k$  is the orthogonal projector on the subspace spanned by  $\{x_i, i = 1, \dots, k\}$  and it commutes with  $\Phi^{1/2} = \sum_{i=1}^{\text{rank}(\Phi)} \sigma_i |x_i\rangle \langle x_i|$ .

**Lemma 1** *Suppose that Assumption 1 holds and  $\text{Ker} \Gamma = 0$ . Then*

- i)  $\|\rho - \rho_k\| = \sup_{\|x\| \leq 1} \|(I - \Pi_k) \Phi^{1/2} x\|$ .
- ii)  $\|\rho - \rho_k\| \leq \|\rho\|$

**Proof:** Note that  $\overline{\text{Im}} \Gamma^{1/2} = H$  because  $\text{Ker} \Gamma^{1/2} = 0$ , and therefore:

$$\|\rho - \rho_k\| = \sup_{\|z\| \leq 1} \|(\rho - \rho_k) z\| = \sup_{\|x\| \leq 1} \left\| \left( \rho - \sum_{i=1}^{\min\{k, \text{rank}(\Phi)\}} |a_i\rangle \langle b_i| \right) \Gamma^{1/2} x \right\|. \quad (9)$$

We can write:  $\sum_{i=1}^{\min\{k, \text{rank}(\Phi)\}} |a_i\rangle \langle b_i| \Gamma^{1/2} = \sum_{i=1}^{\min\{k, \text{rank}(\Phi)\}} |F b_i\rangle \langle \Gamma^{1/2} b_i|$   
 $= \sum_{i=1}^{\min\{k, \text{rank}(\Phi)\}} |\rho \Gamma^{1/2} x_i\rangle \langle x_i| = \rho \Gamma^{1/2} \Pi_k$ . Substituting this into (9), we have:  
 $\|\rho - \rho_k\| = \sup_{\|x\| \leq 1} \|\rho \Gamma^{1/2} (I - \Pi_k) x\|$ . Recall that the operator  $\rho \Gamma^{1/2}$  can be

written as  $U\Phi^{1/2}$ , where  $U$  is a partial isometry. We continue:

$$\begin{aligned} \sup_{\|\Gamma^{1/2}x\|\leq 1} \left\| \rho\Gamma^{1/2}(I - \Pi_k)x \right\| &= \sup_{\|\Gamma^{1/2}x\|\leq 1} \left\| \Phi^{1/2}(I - \Pi_k)x \right\| \\ &= \sup_{\|\Gamma^{1/2}x\|\leq 1} \left\| (I - \Pi_k)\Phi^{1/2}x \right\|, \end{aligned} \quad (10)$$

where the first equality follows from the fact that  $U$  maps the range of  $\Phi^{1/2}$  isometrically onto the range of  $\rho\Gamma^{1/2}$  (see Gohberg and Krein, 1969, p. 7). This proves i) of the Lemma. Since  $(I - \Pi_k)$  is an orthogonal projector, we can further estimate expression (10) as follows:  $\sup_{\|\Gamma^{1/2}x\|\leq 1} \left\| (I - \Pi_k)\Phi^{1/2}x \right\| \leq \sup_{\|\Gamma^{1/2}x\|\leq 1} \left\| \Phi^{1/2}x \right\| = \sup_{\|\Gamma^{1/2}x\|\leq 1} \left\| U'\rho\Gamma^{1/2}x \right\| \leq \|\rho\|$ . QED.

Now, let us return to the proof of Theorem 2c). Since by ii) of Lemma 1,  $\rho - \rho_k$  is uniformly bounded, it is enough to consider only functions  $f$  from everywhere dense subspace  $\text{Im}\Gamma$ . For these  $f$ , we can write:  $\|(\rho - \rho_k)f\| = \|(F\Gamma^{-1} - F\Gamma^{-1/2}\Pi_k\Gamma^{-1/2})f\| = \|F\Gamma^{-1}(I - \Gamma^{1/2}\Pi_k\Gamma^{-1/2})f\| \leq \|\rho\| \|(I - \Gamma^{1/2}\Pi_k\Gamma^{-1/2})f\|$ . But  $\Pi_k\Gamma^{-1/2}f \rightarrow \Gamma^{-1/2}f$  as  $k \rightarrow \infty$ , and, consequently,  $\Gamma^{1/2}\Pi_k\Gamma^{-1/2}f \rightarrow f$  as  $k \rightarrow \infty$ . QED.

**Proof of Theorem 2d):** By i) of Lemma 1,  $\|\rho - \rho_k\| = \sup_{\|\Gamma^{1/2}z\|\leq 1} \|(I - \Pi_k)\Phi^{1/2}z\|$ .

Suppose that the supremum is greater than  $\varepsilon > 0$  for an infinite number of positive integers  $k$ . Then there exists such an infinite sequence of  $z_k$  that  $\|\Gamma^{1/2}z_k\| \leq 1$  and  $\|(I - \Pi_k)\Phi^{1/2}z_k\| > \varepsilon$ .

By the compactness of  $\rho$ ,  $\Phi^{1/2}z_k = U'\rho\Gamma^{1/2}z_k$  has a limiting point, say,  $z$ . By the triangle inequality,  $\|(I - \Pi_k)z\| \geq \|(I - \Pi_k)\Phi^{1/2}z_k\| - \|(I - \Pi_k)(\Phi^{1/2}z_k - z)\| \geq \varepsilon - \|(\Phi^{1/2}z_k - z)\|$ . Consequently,  $\|(I - \Pi_k)z\| \geq \varepsilon/2$  for infinitely many  $k$ . However,  $\|(I - \Pi_k)z\|^2 = \sum_{i=k+1}^{\infty} \langle z, x_i \rangle^2 \rightarrow 0$ , because  $\{x_i, i = 1, 2, \dots\}$  form an orthonormal basis. We have a contradiction. QED.

### 5.3 Proof of Theorem 3

We will often use the following norm inequalities in our proof. For any Hilbert-Schmidt operator  $A$  and any linear bounded operator  $B$ , we have:

$$\|BA\|_2 \leq \|B\| \|A\|_2, \text{ and } \|AB\|_2 \leq \|A\|_2 \|B\|. \quad (11)$$

These inequalities easily follow from inequalities about singular values of  $AB$  in Theorem 1.6 (p. 3) of Simon (2005). Below we will first formulate and prove several auxiliary Lemmas and then will use these Lemmas to establish Theorem 3.

Denote  $\Gamma + \alpha I$  as  $\Gamma_\alpha$ . We have the following:

**Lemma 2** *If Assumption 1 holds then, for any  $\alpha > 0$ ,  $\|\Gamma^{1/2} - \Gamma\Gamma_\alpha^{-1}\Gamma^{1/2}\| < \alpha^{1/2}$*

**Proof:** Since  $\Gamma\Gamma_\alpha^{-1}\Gamma^{1/2} = (\Gamma_\alpha - \alpha I)\Gamma_\alpha^{-1}\Gamma^{1/2} = \Gamma^{1/2} - \alpha\Gamma_\alpha^{-1}\Gamma^{1/2}$ , we have:  $\|\Gamma^{1/2} - \Gamma\Gamma_\alpha^{-1}\Gamma^{1/2}\| = \|\alpha\Gamma_\alpha^{-1}\Gamma^{1/2}\|$ . Further,  $\|\alpha\Gamma_\alpha^{-1}\Gamma^{1/2}\| \leq \|\alpha\Gamma_\alpha^{-1/2}\| \|\Gamma_\alpha^{-1/2}\Gamma^{1/2}\| \leq \alpha^{1/2} \|\Gamma_\alpha^{-1/2}\Gamma^{1/2}\|$ . The norm of  $\Gamma_\alpha^{-1/2}\Gamma^{1/2}$  is equal to the square root of the largest eigenvalue of  $(\Gamma_\alpha^{-1/2}\Gamma^{1/2})' \Gamma_\alpha^{-1/2}\Gamma^{1/2} = \Gamma^{1/2}\Gamma_\alpha^{-1}\Gamma^{1/2}$ . All eigenvalues of  $\Gamma^{1/2}\Gamma_\alpha^{-1}\Gamma^{1/2}$  have form  $\lambda^{1/2}(\lambda + \alpha)^{-1}\lambda^{1/2}$ , where  $\lambda > 0$  is an eigenvalue of  $\Gamma$ . But  $\lambda^{1/2}(\lambda + \alpha)^{-1}\lambda^{1/2}$  is smaller than one for all  $\alpha > 0$ . Hence,  $\|\Gamma_\alpha^{-1/2}\Gamma^{1/2}\| < 1$  and therefore,  $\|\Gamma^{1/2} - \Gamma\Gamma_\alpha^{-1}\Gamma^{1/2}\| < \alpha^{1/2}$ . QED.

**Notation 2** *Let us define  $\Delta_1^{(n)} = \hat{\Gamma} - \Gamma$ ,  $\Delta_2^{(n)} = \hat{F} - F$ , and  $\delta_n = \max_{i=1,2} \left( \|\Delta_i^{(n)}\|_2 \right)$ .*

**Lemma 3** *If Assumption 1 holds, then, for any  $\alpha > 0$ ,  $\|I - \Gamma_\alpha^{1/2}\hat{\Gamma}_\alpha^{-1/2}\| \leq \alpha^{-1}\delta_n$ .*

**Proof:** Since the uniform norm of a Hilbert-Schmidt operator is no larger than its 2-norm, it is enough to show that  $I - \Gamma_\alpha^{1/2}\hat{\Gamma}_\alpha^{-1/2}$  is a Hilbert-Schmidt operator and to prove the inequality of the Lemma for  $\|I - \Gamma_\alpha^{1/2}\hat{\Gamma}_\alpha^{-1/2}\|_2$ .

Proposition 3.2 of van Hemmen and Ando (1980) shows that for any positive definite self-adjoint operators  $A$  and  $B$  such that their difference is a Hilbert-Schmidt operator, the operator  $A^{1/2} - B^{1/2}$  is also Hilbert-Schmidt. Furthermore, for any non-negative constant  $\mu$ , such that  $A^{1/2} + B^{1/2} \geq \mu I$ , the following inequality holds:  $\|A - B\|_2 \geq \mu \|A^{1/2} - B^{1/2}\|_2$ .

Taking  $A = \hat{\Gamma}_\alpha$  and  $B = \Gamma_\alpha$  and noting that  $\hat{\Gamma}_\alpha - \Gamma_\alpha = \Delta_1^{(n)}$  is a Hilbert-Schmidt operator, we conclude that  $\hat{\Gamma}_\alpha^{1/2} - \Gamma_\alpha^{1/2}$  is also a Hilbert-Schmidt operator. Since Hilbert-Schmidt operators form a two-sided ideal in the algebra of bounded operators, the equality  $I - \Gamma_\alpha^{1/2} \hat{\Gamma}_\alpha^{-1/2} = (\hat{\Gamma}_\alpha^{1/2} - \Gamma_\alpha^{1/2}) \hat{\Gamma}_\alpha^{-1/2}$  implies that  $I - \Gamma_\alpha^{1/2} \hat{\Gamma}_\alpha^{-1/2}$  is a Hilbert-Schmidt operator. Further, noting that  $\hat{\Gamma}_\alpha^{1/2} + \Gamma_\alpha^{1/2} \geq 2\alpha^{1/2}I$  because the function  $x^{1/2}$  is operator monotonic, we get:  $\|\hat{\Gamma}_\alpha^{1/2} - \Gamma_\alpha^{1/2}\|_2 \leq \frac{1}{2}\alpha^{-1/2} \|\hat{\Gamma}_\alpha - \Gamma_\alpha\|_2 = \frac{1}{2}\alpha^{-1/2} \|\hat{\Gamma} - \Gamma\|_2 \leq \alpha^{-1/2}\delta_n$ . Using this fact and inequality (11), we obtain:  $\|I - \Gamma_\alpha^{1/2} \hat{\Gamma}_\alpha^{-1/2}\|_2 \leq \|\hat{\Gamma}_\alpha^{1/2} - \Gamma_\alpha^{1/2}\|_2 \|\hat{\Gamma}_\alpha^{-1/2}\|_2 \leq \alpha^{-1/2}\delta_n \alpha^{-1/2} = \alpha^{-1}\delta_n$ . QED.

**Definition 7** Let us define  $\hat{\rho}_\alpha = \hat{F} \hat{\Gamma}_\alpha^{-1}$ .

**Lemma 4** If Assumption 1 holds and  $\rho$  is Hilbert-Schmidt, then for a positive constant  $C$  which depends only on  $\rho$  and  $\Gamma$  and for every  $\alpha > 0$ ,  $\|\rho - \hat{\rho}_\alpha\|_{\Gamma,2} \leq C(\alpha^{1/2} + \alpha^{-1}\delta_n + \alpha^{-2}\delta_n^2)$ .

**Proof:** By the triangle inequality and norm inequalities (11),

$$\begin{aligned}
& \|\rho - \hat{\rho}_\alpha\|_{\Gamma,2} \\
&= \left\| \left( \rho - F\Gamma_\alpha^{-1} + F\Gamma_\alpha^{-1} - F\hat{\Gamma}_\alpha^{-1/2}\Gamma_\alpha^{-1/2} + F\hat{\Gamma}_\alpha^{-1/2}\Gamma_\alpha^{-1/2} - F\hat{\Gamma}_\alpha^{-1} + F\hat{\Gamma}_\alpha^{-1} - \hat{\rho}_\alpha \right) \Gamma^{1/2} \right\|_2 \\
&= \left\| \rho \left( \Gamma^{1/2} - \Gamma\Gamma_\alpha^{-1}\Gamma^{1/2} \right) + F\Gamma_\alpha^{-1/2} \left( I - \Gamma_\alpha^{1/2}\hat{\Gamma}_\alpha^{-1/2} \right) \Gamma_\alpha^{-1/2}\Gamma^{1/2} \right. \\
&\quad \left. + F\hat{\Gamma}_\alpha^{-1/2} \left( I - \hat{\Gamma}_\alpha^{-1/2}\Gamma_\alpha^{1/2} \right) \Gamma_\alpha^{-1/2}\Gamma^{1/2} + (F - \hat{F}) \hat{\Gamma}_\alpha^{-1}\Gamma^{1/2} \right\|_2 \\
&\leq \|\rho\|_2 \left\| \Gamma^{1/2} - \Gamma\Gamma_\alpha^{-1}\Gamma^{1/2} \right\| + \left\| F\Gamma_\alpha^{-1/2} \right\|_2 \left\| I - \Gamma_\alpha^{1/2}\hat{\Gamma}_\alpha^{-1/2} \right\| \left\| \Gamma_\alpha^{-1/2}\Gamma^{1/2} \right\| \\
&\quad + \left\| F\hat{\Gamma}_\alpha^{-1/2} \right\|_2 \left\| I - \hat{\Gamma}_\alpha^{-1/2}\Gamma_\alpha^{1/2} \right\| \left\| \Gamma_\alpha^{-1/2}\Gamma^{1/2} \right\| + \|F - \hat{F}\|_2 \left\| \hat{\Gamma}_\alpha^{-1}\Gamma^{1/2} \right\|.
\end{aligned}$$

Lemma 2 implies that the first term in the above sum is no larger than  $\|\rho\|_2 \alpha^{1/2}$ .

The second term in the sum is no larger than  $\|\rho\Gamma^{1/2}\|_2 \alpha^{-1}\delta_n$ . This estimate follows from Lemma 3 and the facts that  $\left\| F\Gamma_\alpha^{-1/2} \right\|_2 \leq \|\rho\Gamma^{1/2}\|_2$  and  $\left\| \Gamma_\alpha^{-1/2}\Gamma^{1/2} \right\| \leq 1$ . For the first component of the third term in the sum, using the triangle inequality, inequalities (11), and Lemma 3, we have:  $\left\| F\hat{\Gamma}_\alpha^{-1/2} \right\|_2 \leq \left\| F\Gamma_\alpha^{-1/2} \right\|_2 \left\| \Gamma_\alpha^{1/2}\hat{\Gamma}_\alpha^{-1/2} - I \right\| + \left\| F\Gamma_\alpha^{-1/2} \right\|_2 \leq \|\rho\Gamma^{1/2}\|_2 (\alpha^{-1}\delta_n + 1)$ . Using Lemma 3 one more time, we find that the third term in the sum is no larger than  $\|\rho\Gamma^{1/2}\|_2 (\alpha^{-1}\delta_n + 1) \alpha^{-1}\delta_n$ . Finally, the last term of the sum is obviously no larger than  $\alpha^{-1}\delta_n$ . Combining the upper bounds for all the four terms in the sum, gives the statement of the Lemma. QED.

**Lemma 5** *If Assumption 1 holds, then, for all  $\beta > 1/2$ ,  $\delta_n = o(n^{-1/4} \log^\beta n)$  as  $n \rightarrow \infty$  a.s.*

**Proof:** Theorems 4.1 and 4.8 of Bosq (2000) imply the statement of this Lemma.

QED.

Let us now turn to the proof of Theorem 3. By the triangle inequality,

$$\|\rho - \hat{\rho}_{\alpha,k}\|_{\Gamma,2} \leq \|\rho - \hat{\rho}_\alpha\|_{\Gamma,2} + \|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\|_{\Gamma,2}. \tag{12}$$

By Lemma 4, the first term in the above sum is no larger than  $C(\alpha^{1/2} + \alpha^{-1}\delta_n + \alpha^{-2}\delta_n^2)$ .

For the second term, using norm inequalities (11), we obtain:  $\|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\|_{\Gamma,2} \leq \|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\| \|\Gamma^{1/2}\|_2$ . We now look for an estimate of  $\|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\|$ . We have:  $\|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\| = \left\| \left( I - \hat{\Pi}_{\alpha,k} \right) \hat{\Phi}_\alpha^{1/2} \hat{\Gamma}_\alpha^{-1/2} \right\| \leq \alpha^{-1/2} \left\| \left( I - \hat{\Pi}_{\alpha,k} \right) \hat{\Phi}_\alpha^{1/2} \right\|$ . Next note that  $\left\| \left( I - \hat{\Pi}_{\alpha,k} \right) \hat{\Phi}_\alpha^{1/2} \right\|$  is equal to the  $(k+1)$ -st eigenvalue of  $\hat{\Phi}_\alpha^{1/2}$ ,  $\hat{\sigma}_{\alpha,k+1}$ , or in other words, to the  $(k+1)$ -st singular value of the operator  $\hat{F} \hat{\Gamma}_\alpha^{-1/2}$ , which we denote as  $\mu_{k+1} \left( \hat{F} \hat{\Gamma}_\alpha^{-1/2} \right)$ . Using Theorem 1.6 in Simon (2005), we can estimate this as follows:  $\mu_{k+1} \left( \hat{F} \hat{\Gamma}_\alpha^{-1/2} \right) \leq \left\| \hat{\Gamma}_\alpha^{-1/2} \right\| \mu_{k+1} \left( \hat{F} \right) \leq \alpha^{-1/2} \mu_{k+1} \left( \hat{F} \right)$ . Since  $\hat{F} = F + (\hat{F} - F)$  and  $\|\hat{F} - F\| \leq \delta_n$ , we can apply Theorem 1.7 in Simon (2005) and continue this estimate:  $\mu_{k+1} \left( \hat{F} \hat{\Gamma}_\alpha^{-1/2} \right) \leq \alpha^{-1/2} \left( \mu_{k+1} \left( F \right) + \delta_n \right) = \alpha^{-1/2} \left( \mu_{k+1} \left( \rho \Gamma \right) + \delta_n \right) \leq \alpha^{-1/2} \left( \|\Gamma^{1/2}\| \mu_{k+1} \left( \rho \Gamma^{1/2} \right) + \delta_n \right) = \alpha^{-1/2} \left( \|\Gamma^{1/2}\| \sigma_{k+1} + \delta_n \right)$ . Altogether we get:  $\|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\|_{\Gamma,2} \leq \|\Gamma^{1/2}\|_2 \alpha^{-1} \left( \|\Gamma^{1/2}\| \sigma_{k+1} + \delta_n \right)$ .

Note that since, by assumption,  $\rho$  is a Hilbert-Schmidt operator,  $\rho \Gamma^{1/2}$  is a trace-class operator. This follows from the facts that  $\Gamma^{1/2}$  is Hilbert-Schmidt and that the product of two Hilbert-Schmidt operators is a trace-class operator (see Theorem 3.2 of Simon, 2005, p. 31). Let  $\rho \Gamma^{1/2} = U \Phi^{1/2}$  be the polar decomposition of  $\rho \Gamma^{1/2}$ . Then,  $\Phi^{1/2} = U' \rho \Gamma^{1/2}$ , and, since trace-class operators form a two-sided ideal in the algebra of bounded linear operators,  $\Phi^{1/2}$  is a trace-class operator. Therefore,  $\sum_{i=1}^{\infty} \sigma_i < \infty$ . Since  $\{\sigma_i\}$  is a non-increasing sequence, the latter inequality implies that  $\sigma_i = o(i^{-1})$ . Hence, we have:  $\|\hat{\rho}_\alpha - \hat{\rho}_{\alpha,k}\|_{\Gamma,2} \leq C \alpha^{-1} (k^{-1} + \delta_n)$  for some constant  $C$  that depends only on  $\rho$  and  $\Gamma$ . Combining the latter inequality with (12) and the upper bound for  $\|\rho - \hat{\rho}_\alpha\|_{\Gamma,2}$ , we get:

$$\|\rho - \hat{\rho}_{\alpha,k}\|_{\Gamma,2} \leq C \left[ \alpha^{1/2} + \alpha^{-1}\delta_n + \alpha^{-2}\delta_n^2 + \alpha^{-1}k^{-1} \right]. \quad (13)$$

Let  $\alpha_n \sim n^{-1/6}$  as  $n \rightarrow \infty$  and  $k_n \geq K n^{1/4}$  for a certain  $K > 0$ . Then, by Lemma 5,  $\alpha_n^{1/2} + \alpha_n^{-1}\delta_n + \alpha_n^{-2}\delta_n^2 = o(n^{-1/12} \log^\beta n)$  for any  $\beta > 1/2$  a.s., and

$\alpha_n^{-1}k_n^{-1} = O(n^{-1/12})$  a.s. Inequality (13) then implies the statement of Theorem 3.

QED.

We note here that the above proof also shows that

$$\|\hat{\rho}_{\alpha_n} - \hat{\rho}_{\alpha_n, k_n}\| = o\left(n^{-1/12} \log^\beta n\right) \text{ a.s.} \quad (14)$$

for any  $\beta > 1/2$ . We will refer to this fact in the proof of Theorem 4.

#### 5.4 Proof of Theorem 4

Let  $m = \lfloor n - n^\phi \rfloor$ , where  $\phi < 1/12$  and  $\lfloor \cdot \rfloor$  denotes the integer part of a number. And let  $\hat{\Gamma}^{(m)}$  and  $\hat{F}^{(m)}$  denote the empirical covariance and cross-covariance operators based on the first  $m$  observations of the process  $\{f_t\}$ ,  $\hat{\Gamma}_\alpha^{(m)} = \hat{\Gamma}^{(m)} + \alpha I$ , and  $\hat{\Phi}_\alpha^{(m)} = \left[\hat{\Gamma}_\alpha^{(m)}\right]^{-1/2} \hat{F}^{(m)} \hat{F}^{(m)} \left[\hat{\Gamma}_\alpha^{(m)}\right]^{-1/2}$ . Further, let  $\hat{\Phi}_\alpha^{(m)} = \sum_{i=1}^{\infty} \hat{\sigma}_{\alpha, i}^{2(m)} \left| \hat{x}_{\alpha, i}^{(m)} \right\rangle \left\langle \hat{x}_{\alpha, i}^{(m)} \right|$  be a spectral decomposition of  $\hat{\Phi}_\alpha^{(m)}$ , and define  $\hat{b}_{\alpha, i}^{(m)} = \left[\hat{\Gamma}_\alpha^{(m)}\right]^{-1/2} \hat{x}_{\alpha, i}^{(m)}$ ,  $\hat{a}_{\alpha, i}^{(m)} = \hat{F}^{(m)} \hat{b}_{\alpha, i}^{(m)}$ , and  $\hat{\rho}_{\alpha, k}^{(m)} = \sum_{i=1}^k \left| \hat{a}_{\alpha, i}^{(m)} \right\rangle \left\langle \hat{b}_{\alpha, i}^{(m)} \right|$ .

Using the fact that  $f_n = \rho^{n-m} f_m + \sum_{i=0}^{n-m-1} \rho^i \varepsilon_{n-i}$ , and the inequality  $(a + b + c)^2 \leq 3(a^2 + b^2 + c^2)$ , we have:  $E \left\| (\rho - \hat{\rho}_{\alpha_n, k_n}) f_n \right\|^2 \leq C_1 + C_2 + C_3$ , where  $C_1 = 3E \left\| (\rho - \hat{\rho}_{\alpha_n, k_n}^{(m)}) \rho^{n-m} f_m \right\|^2$ ,  $C_2 = 3E \left\| (\rho - \hat{\rho}_{\alpha_n, k_n}^{(m)}) \sum_{i=0}^{n-m-1} \rho^i \varepsilon_{n-i} \right\|^2$ , and  $C_3 = 3E \left\| (\hat{\rho}_{\alpha_n, k_n}^{(m)} - \hat{\rho}_{\alpha_n, k_n}) f_n \right\|^2$ .

For the term  $C_1$ , we have:  $\left\| (\rho - \hat{\rho}_{\alpha_n, k_n}^{(m)}) \rho^{n-m} f_m \right\|^2 \leq \left\| \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\|^2 \left\| \rho^{n-m} \right\|^2 \left\| f_m \right\|^2$ . Now,  $\left\| \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\|^2 \leq 2 \left\| \rho \right\|^2 + 2 \left\| \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\|^2$ . But  $\left\| \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\| \leq \sum_{i=1}^{k_n} \left\| \hat{a}_{\alpha_n, i}^{(m)} \right\| \left\| \hat{b}_{\alpha_n, i}^{(m)} \right\| \leq k_n \alpha_n^{-1} \left\| \hat{F}^{(m)} \right\|$ , where the latter inequality follows from the definitions of  $\hat{b}_{\alpha_n, i}^{(m)}$  and  $\hat{a}_{\alpha_n, i}^{(m)}$ , and the fact that  $\left\| \hat{x}_{\alpha_n, i}^{(m)} \right\| = 1$ . Hence,  $\left\| \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\|^2 \leq C \left( 1 + k_n^2 \alpha_n^{-2} \left\| \hat{F}^{(m)} \right\|^2 \right)$ , where the constant  $C$  depends only on  $\rho$  and  $\Gamma$ . Since by assumption,  $\alpha_n \sim n^{-1/6}$  and  $k_n \leq n$ , and since, by Lemma 5,  $\left\| \hat{F}^{(m)} \right\| = O(1)$  a.s.,  $\left\| \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\|^2 = O(n^{7/3})$  a.s.

Next, by Lemma 3.1 of Bosq (2000), our Assumption 1 implies that there exists an  $a > 0$  and  $0 < b < 1$ , which do not depend on  $n$  and  $m$ , such that  $\|\rho^{n-m}\| \leq ab^{n-m}$ . Further,  $E\|f_m\|^2 < \infty$  and does not depend on  $m$ . Combining these facts with the fact that  $\left\|\rho - \hat{\rho}_{\alpha_n, k_n}^{(m)}\right\|^2 = O(n^{7/3})$  a.s., we obtain:

$$C_1 = O\left(b^{2n^\phi}\right) O(n^{7/3}) = O\left(b_0^{n^\phi}\right) \text{ a.s.}, \quad (15)$$

where  $b_0$  is any number larger than  $b^2$  but smaller than 1.

For the term  $C_2$ , we have the following Lemma.

**Lemma 6** *Let Assumption 1 hold. Denote the covariance operator of the innovation process  $\varepsilon_t$  as  $C_\varepsilon$ . Then*

$$C_2 = 3E \operatorname{tr} \left\{ \left( \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right) \left( \sum_{i=0}^{n-m-1} \rho^i C_\varepsilon (\rho^i)' \right) \left( \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right)' \right\}.$$

**Proof:** Denote the random operator  $\rho - \hat{\rho}_{\alpha_n, k_n}^{(m)}$  as  $\Psi$  and the random function  $\sum_{i=0}^{n-m-1} \rho^i \varepsilon_{n-i}$  as  $\xi$ . Let  $\{e_1, e_2, \dots\}$  be any orthonormal basis of  $H$ . Then,  $\|\Psi\xi\|^2 = \sum_{j=1}^{\infty} \langle e_j, \Psi\xi \rangle^2 = \sum_{j=1}^{\infty} \langle e_j, \langle e_j, \Psi\xi \rangle \Psi\xi \rangle = \sum_{j=1}^{\infty} \langle e_j, \langle \Psi'e_j, \xi \rangle \Psi\xi \rangle$ . Using the law of iterated expectations, we get:  $E\|\Psi\xi\|^2 = E\left(E^{(m)}\left(\sum_{j=1}^{\infty} \langle e_j, \langle \Psi'e_j, \xi \rangle \Psi\xi \rangle\right)\right)$ , where  $E^{(m)}$  is the conditional expectation relative to a sub- $\sigma$ -algebra  $\mathcal{F}^{(m)}$  of  $\mathcal{F}$  generated by the  $H$ -valued random variables  $\{f_1, \dots, f_m\}$ . (For a definition and properties of the conditional expectation operator  $\mathcal{E}^{(m)}$  relative to  $\mathcal{F}^{(m)}$  acting on the space of the square-integrable  $H$ -valued random variables, see Section 1.5 in Bosq (2000).)

By Lebesgue's monotone convergence theorem, we have:  $E^{(m)}\left(\sum_{j=1}^{\infty} \langle e_j, \langle \Psi'e_j, \xi \rangle \Psi\xi \rangle\right) = \sum_{j=1}^{\infty} E^{(m)}\langle e_j, \langle \Psi'e_j, \xi \rangle \Psi\xi \rangle$ . (Note that the theorem is applicable because  $\langle e_j, \langle \Psi'e_j, \xi \rangle \Psi\xi \rangle = \langle \Psi'e_j, \xi \rangle^2$  are non-negative). The latter expression is equal to  $\sum_{j=1}^{\infty} \langle e_j, \mathcal{E}^{(m)}(\langle \Psi'e_j, \xi \rangle \Psi\xi) \rangle$ .

Summing up,  $E \|\Psi\xi\|^2 = E \left( \sum_{j=1}^{\infty} \langle e_j, \mathcal{E}^{(m)} (\langle \Psi' e_j, \xi \rangle \Psi \xi) \rangle \right)$ .

Since  $\Psi$  is a linear operator which depends only on  $\{f_1, \dots, f_m\}$ , we have:  $\mathcal{E}^{(m)} (\langle \Psi' e_j, \xi \rangle \Psi \xi) = \Psi \mathcal{E}^{(m)} (\langle \Psi' e_j, \xi \rangle \xi)$ . Furthermore, since  $\xi$  is independent of the  $\sigma$ -algebra  $\mathcal{F}^{(m)}$ ,  $\mathcal{E}^{(m)} (\langle \Psi' e_j, \xi \rangle \xi) = \mathcal{E} (\langle \Psi' e_j, \xi \rangle \xi) = C_\xi \Psi' e_j$ . Hence,  $E \|\Psi\xi\|^2 = E \left( \sum_{j=1}^{\infty} \langle e_j, \Psi C_\xi \Psi' e_j \rangle \right) = E (\text{trace}(\Psi C_\xi \Psi'))$ , where the last equality follows from Theorem 2.14 of Simon (2005). The statement of the Lemma now follows from the fact that  $C_\xi = \sum_{i=0}^{n-m-1} \rho^i C_\varepsilon (\rho^i)'$ . QED.

Now, as shown in the proof of Theorem 3.2 of Bosq (2000),  $\sum_{i=0}^{n-m-1} \rho^i C_\varepsilon (\rho^i)' = \Gamma - \rho^{n-m} \Gamma (\rho^{n-m})'$ . Therefore,  $C_2 \leq 3E \text{tr} \left\{ \left( \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right) \Gamma \left( \rho - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right)'\right\}$ . Then Theorem 3 implies that, for any  $\beta > 1/2$ ,

$$C_2 = o \left( n^{-1/6} \log^{2\beta} n \right) \text{ a.s.} \quad (16)$$

For the term  $C_3$ , we have the following Lemma:

**Lemma 7** *Let Assumption 1 hold and suppose that  $\rho$  is a Hilbert-Schmidt operator.*

*Then  $\left\| \hat{\rho}_{\alpha_n, k_n}^{(m)} - \hat{\rho}_{\alpha_n, k_n} \right\| = o \left( n^{-1/12} \log^\beta n \right)$  as  $n \rightarrow \infty$  a.s. for any  $\beta > 1/2$ .*

**Proof:** Equation (14) established in the proof of Theorem 3 tells us that  $\left\| \hat{F} \hat{\Gamma}_{\alpha_n}^{-1} - \hat{\rho}_{\alpha_n, k_n} \right\| = o \left( n^{-1/12} \log^\beta n \right)$  as  $n \rightarrow \infty$  a.s. for any  $\beta > 1/2$ . Since  $m \sim n$  as  $n \rightarrow \infty$ , we also have:  $\left\| \hat{F}^{(m)} \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} - \hat{\rho}_{\alpha_n, k_n}^{(m)} \right\| = o \left( n^{-1/12} \log^\beta n \right)$  a.s. for any  $\beta > 1/2$ . Hence, the triangle inequality implies that to establish the Lemma, it is enough to show that  $\left\| \hat{F} \hat{\Gamma}_{\alpha_n}^{-1} - \hat{F}^{(m)} \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| = o \left( n^{-1/12} \log^\beta n \right)$  a.s. for any  $\beta > 1/2$ .

We have:  $\left\| \hat{F} \hat{\Gamma}_{\alpha_n}^{-1} - \hat{F}^{(m)} \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| \leq \|\hat{F}\| \left\| \hat{\Gamma}_{\alpha_n}^{-1} - \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| + \left\| \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| \|\hat{F} - \hat{F}^{(m)}\|$ . Since both  $\hat{F}$  and  $\hat{F}^{(m)}$  approximate  $F$ , Lemma 5 implies that  $\|\hat{F} - \hat{F}^{(m)}\|$  is  $o \left( n^{-1/4} \log^\beta n \right)$  a.s. for any  $\beta > 1/2$ . Further,  $\left\| \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| \leq$

$\alpha_n^{-1} \sim n^{1/6}$ . Combining these two facts, we find that  $\left\| \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| \left\| \hat{F} - \hat{F}^{(m)} \right\| = o\left(n^{-1/12} \log^\beta n\right)$  a.s. for any  $\beta > 1/2$ . Hence, it remains to prove the same convergence rate for  $\left\| \hat{F} \right\| \left\| \hat{\Gamma}_{\alpha_n}^{-1} - \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\|$ .

Lemma 5 implies that  $\left\| \hat{F} \right\| = O(1)$  a.s. Further, since  $\left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| \leq \alpha_n^{-1/2}$ , we have:

$$\begin{aligned} \left\| \hat{\Gamma}_{\alpha_n}^{-1} - \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| &= \left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left[ I - \left( I + \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right)^{-1} \right] \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| \\ &\leq \alpha_n^{-1} \left\| I - \left( I + \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right)^{-1} \right\| \end{aligned} \quad (17)$$

Suppose that  $\left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| \leq 1/2$ , and hence, any eigenvalue  $\lambda$  of the operator  $\hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2}$  is no larger than  $1/2$  by absolute value. Note that for any real  $\lambda$  such that  $|\lambda| \leq 1/2$ , the elementary inequality  $\left| 1 - (1+\lambda)^{-1} \right| \leq 2|\lambda|$  holds. Therefore, the absolute value of any of the eigenvalues of  $I - \left( I + \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right)^{-1}$  and hence, the norm of this operator, is no larger than  $2 \left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right\|$  whenever  $\left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| \leq 1/2$ .

Now,

$$\left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| \leq \alpha_n^{-1} \left\| \hat{\Gamma}_{\alpha_n} - \hat{\Gamma}_{\alpha_n}^{(m)} \right\|. \quad (18)$$

Further, by definition,  $\hat{\Gamma}_{\alpha_n} - \hat{\Gamma}_{\alpha_n}^{(m)} = \frac{1}{m} \sum_{t=m+1}^n |f_t\rangle \langle f_t| - \frac{n-m}{m} \hat{\Gamma}$ . For the first term in the latter difference, we have by the Chebyshev inequality:  $\Pr\left(\left\| \frac{1}{m} \sum_{t=m+1}^n |f_t\rangle \langle f_t| \right\| > n^{-5/12}\right) \leq n^{5/6} E\left(\left\| \frac{1}{m} \sum_{t=m+1}^n |f_t\rangle \langle f_t| \right\|^2\right) \leq n^{5/6} (n-m) m^{-2} \sum_{t=m+1}^n E\| |f_t\rangle \langle f_t| \|^2$ . Since  $\| |f_t\rangle \langle f_t| \|^2 = \|f_t\|^4$ , and  $\{f_t\}$  is a strictly stationary process with  $\|f_t\|^4 < \infty$ , therefore, we have  $\Pr\left(\left\| \frac{1}{m} \sum_{t=m+1}^n |f_t\rangle \langle f_t| \right\| > n^{-5/12}\right) \leq C n^{5/6+2\phi-2}$  for some positive constant  $C$ . By assumption,  $\phi < 1/12$ . Therefore, using the Borel-Cantelli

lemma, we have:

$$\left\| \frac{1}{m} \sum_{t=m+1}^n |f_t\rangle \langle f_t| \right\| = O\left(n^{-5/12}\right) \text{ a.s.} \quad (19)$$

Since, in addition,  $\frac{n-m}{m} \|\hat{\Gamma}\| = O(n^{\phi-1}) = o(n^{-11/12})$  a.s.,  $\|\hat{\Gamma}_{\alpha_n} - \hat{\Gamma}_{\alpha_n}^{(m)}\| \leq \left\| \frac{1}{m} \sum_{t=m+1}^n |f_t\rangle \langle f_t| \right\| + \frac{n-m}{m} \|\hat{\Gamma}\| = O(n^{-5/12})$  too. Inequality (18) then implies that  $\left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| = O(n^{-1/4})$  a.s., and in particular,  $\left\| \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right\| \leq 1/2$  a.s. for large enough  $n$ . As explained above, the latter two facts imply that

$$\left\| I - \left( I + \hat{\Gamma}_{\alpha_n}^{-1/2} \left( \hat{\Gamma}_{\alpha_n}^{(m)} - \hat{\Gamma}_{\alpha_n} \right) \hat{\Gamma}_{\alpha_n}^{-1/2} \right)^{-1} \right\| = O\left(n^{-1/4}\right) \text{ a.s.}$$

This equality and (17) imply in turn that  $\left\| \hat{\Gamma}_{\alpha_n}^{-1} - \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| = O\left(n^{-1/12}\right)$  a.s.

Hence,  $\|\hat{F}\| \left\| \hat{\Gamma}_{\alpha_n}^{-1} - \left( \hat{\Gamma}_{\alpha_n}^{(m)} \right)^{-1} \right\| = O\left(n^{-1/12}\right)$  a.s. as  $n \rightarrow \infty$ . QED

Lemma 7 and the fact that  $E \|f_n\|^2 < \infty$  does not depend on  $n$  imply that

$$C_3 = o\left(n^{-1/6} \log^{2\beta} n\right) \text{ a.s.} \quad (20)$$

for any  $\beta > 1/2$ . The statement of Theorem 4 follows from (15), (16), and (20). QED.

## 6 REFERENCES

ANTONIADIS, A. AND T. SAPATINAS (2003). Wavelet Methods For Continuous-Time Prediction Using Hilbert-Valued Autoregressive Processes. *J. of Multivariate Anal.* **87** 133-158.

BERNARD, P. (1997). *Analyse de Signaux Physiologiques*. Memoire Univ. Cathol. Angers.

BESSE, P. C. AND H. CARDOT (1996). Approximation Spline de la Prevision d'un Processus Fonctionnel Autoregressif d'Ordre 1. *Canad. J. Statist.* **24** 467-487.

BESSE, P. C., CARDOT, H. AND D. B. STEPHENSON (2000). Autoregressive Forecasting of Some Functional Climatic Variations. *Scand. J. Statist.* **27** 673-687.

BOSQ, D. (2000). *Linear Processes in Function Spaces: Theory And Applications*. Springer-Verlag, New York.

CAVALLINI, A., MONTANARI, G. C., LOGGINI, M., LESSI, O. AND M. CACCIARI (1994). Nonparametric Prediction of Harmonic Levels in Electrical Networks. *Proceedings of IEEE ICHPS VI*. Bologna. 165-171.

COCHRANE, J. H. AND M. PIAZZESI (2002). Bond Risk Premia. *NBER Working Paper* 9178.

DAMON, J. AND S. GUILLAS (2002). The Inclusion of Exogenous Variables in Functional Autoregressive Ozone Forecasting. *Environmetrics*. **13** 759-774.

DIEBOLD, F. X. AND C. LI (2006). Forecasting the Term Structure of Government Bond Yields. *Journal of Econometrics*. **130** 337-364.

EATON, M. L. (1983). *Multivariate Statistics: A Vector Space Approach*. John Wiley and Sons, New York.

ECKART, C. AND G. YOUNG (1936). The approximation of one matrix by another of lower rank. *Psychometrika*. **1** 211-218.

FORTIER, J. J. (1966). Simultaneous Linear Prediction. *Psychometrika*. **31** 369-381.

GOHBERG, I. C. AND M. C. KREIN (1969). *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert Space*. American Mathematical Society, Providence. (Volume 18 in Translations of Mathematical Monographs).

LEURGANS, S. E., MOYEED, R. A. AND B. W. SILVERMAN (1993). Canonical Correlation Analysis when Data are Curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*

**55** 725-740.

MAS, A. (1999). Estimation d'opérateurs de corrélation de processus fonctionnels: Lois limites, tests, déviations modérées. PhD Thesis, University Paris 6.

RAMSAY, J. O. AND B. W. SILVERMAN (1997). *Functional Data Analysis*. Springer, New York.

RAMSAY, J. O. AND B. W. SILVERMAN (2002). *Applied Functional Data Analysis*. Springer, New York.

REINSEL, O. (1983). Some Results on Multivariate Autoregressive Index Models. *Biometrika*. **70** 145-156.

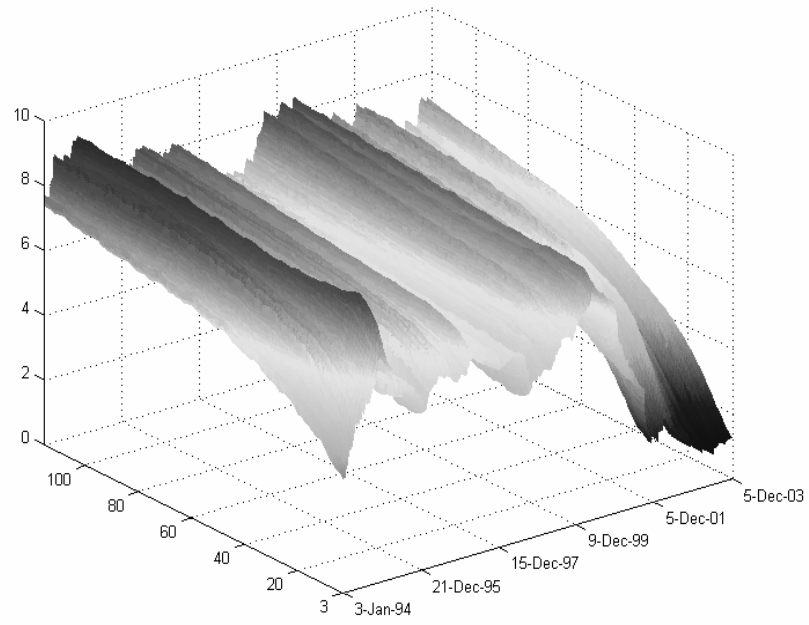
SCHMIDT, F. (1907). Zur Theorie der linearen und nichtlinearen Integralgleichungen. I Teil. Entwicklung willkürlichen Funktionen nach System vorgeschriebener. *Math. Annalen*. **63** 433-476.

SIMON, B. (2005). *Trace Ideals and Their Applications*. 2nd edition. American Mathematical Society, Providence.

VAN DEN WOLLENBERG, A.L. (1977). Redundancy Analysis: An Alternative for Canonical Correlation Analysis. *Psychometrika*. **42** 207-219.

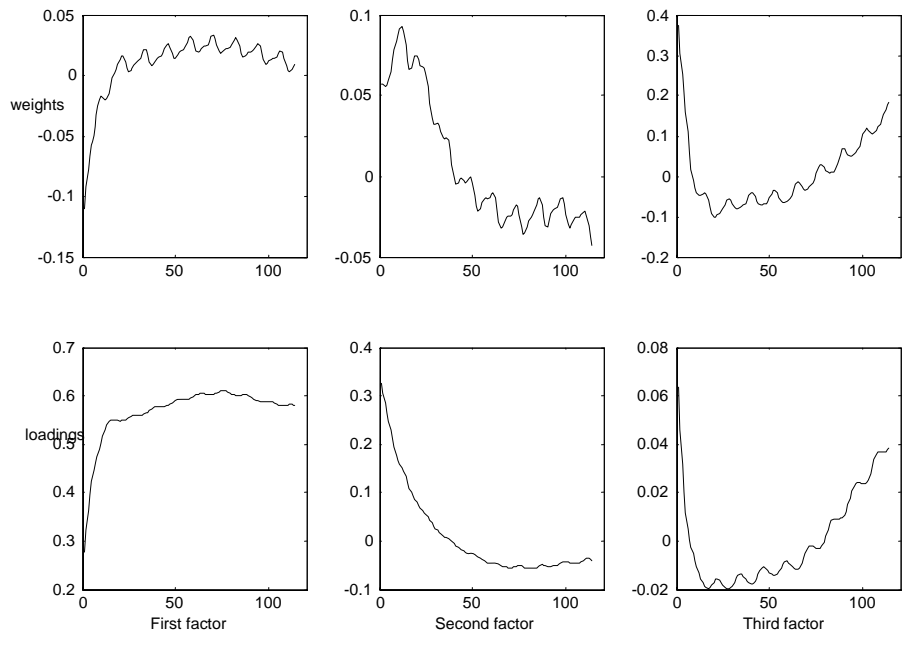
VAN HEMMEN, J.L. AND T. ANDO (1980) An Inequality for Trace Ideals. *Communications in Mathematical Physics*. **76** 143-148.

**Figure 1 Eurodollar Futures Rates Evolution**



Note: The time to maturity (in months) is on the left axis.

**Figure 2** Weights and Loadings of the First Three Predictive Factors



**Figure 3** Predictive Performances of Different Forecasting Methods

