Curve Forecasting by Functional Autoregression

V. Kargin[1] (Cornerstone Research)

A. Onatski[2] (Columbia University)

January, 2005

Abstract

Data in which each observation is a curve occur in many applied problems. This paper explores prediction in time series in which the data is generated by a curve-valued autoregression process. It develops a novel technique, the predictive factor decomposition, for estimation of the autoregression operator, which is designed to be better suited for prediction purposes than the principal components method. The technique is based on finding a reduced-rank approximation to the autoregression operator that minimizes the norm of the expected prediction error.

Implementing this idea, we relate the operator approximation problem to an eigenvalue problem for an operator pencil that is formed by the cross-covariance and covariance operators of the autoregressive process. We develop an estimation method based on regularization of the empirical counterpart of this eigenvalue problem, and prove that with a certain choice of parameters, the method consistently estimates the predictive factors. In addition, we show that forecasts based on the estimated predictive factors converge in probability to the optimal forecasts.

The new method is illustrated by an analysis of the dynamics of the term structure of Eurodollar futures rates. We restrict the sample to the period of normal growth and find that in this subsample the predictive factor technique not only outperforms the principal components method but also performs on par with the best available prediction methods.

Key Words: Functional data analysis; Dimension reduction, Reduced-rank regression; Principal component; Predictive factor, Generalized eigenvalue problem; Term structure; Interest rates

[1] skarguine@cornerstone.com ;
Cornerstone Research, 599 Lexington Avenue floor 43, New York, NY 10022

[2] ao2027@columbia.edu;
Economics Department, Columbia University, 1007 International Affairs Building MC 3308, 420 West 118th Street , New York, NY 10027

1. INTRODUCTION

The statistical analysis of problems from different disciplines increasingly relies on functional data, where each observation is a curve as opposed to a finite-dimensional vector. Numerous examples of functional data analysis can be found in the books by Ramsay and Silverman (1997 and 2002). In this paper we study the problem of curve forecasting when the data generating process is the autoregressive Hilbertian process of order 1 introduced by Bosq (1991):

$$f_{t+h} = \rho[f_t] + \varepsilon_{t+h}. \tag{1}$$

Here for each integer t, f_t is an element of a Hilbert space H, ρ is a linear bounded operator on H, ε_t is a strong H-white noise, and h is the lag length. (Appendix A briefly describes the formalism of Hilbert space valued random variables.) Model (1) has been successfully used by Cavallini et al (1994), Besse and Cardot (1996), Besse et al (2000), Bernard (1997), and Damon and Guillas (2002) for forecasting of electricity consumption, traffic, climatic variations, electrocardiograms, and ozone concentration respectively.

Forecasting in the functional autoregression framework calls for estimation of the infinite-dimensional operator ρ. Since only a finite number of data points is observed, what is needed is a dimension reduction technique. All above-mentioned studies use the first few eigenvectors of the sample covariance operator as the basis for the dimension reduction. We argue that this method is not well suited for forecasting. The reason is that the largest eigenvectors of the covariance operator for f_t may have nothing to do with the best predictors of f_{t+h}. For example, in economics, while it is true that more than 95% percent of the variation in the nominal bonds' yield curve can be explained by the first three principal components, recent research (Cochrane and Piazzesi (2002)) suggests that the best predictors of interest rate movements are among those components that do not contribute much to the overall interest rate variation.

This paper develops a novel technique, the predictive factor decomposition, for the estimation of the autoregression operator, which is designed to be better suited for prediction purposes than the principal components method. The main idea of the predictive factor method is to focus on estimation of those linear functionals of the data that can contribute most to the reduction of the expected error of prediction. To describe such functionals, we approximate ρ by a reduced-rank operator so that the norm of the expected error from prediction using the approximating operator is minimized. We call the functions forming a particular orthogonal basis

in the image of the approximating operator *predictive factor loadings* and the random coordinates (in this basis) of the reduced-rank prediction *predictive factors*. Relative to the forecasting based on the principal components dimension reduction, the predictive factors are less likely to miss those linear functionals of the data having much predictive power. This creates a potential for the predictive factors to work better than the principal components in finite samples.

The new technique is an equivalent of the simultaneous linear predictions introduced in the static finite-dimensional context by Fortier (1966). For the time series data, the method extends the reduced-rank autoregression studied by Reinsel (1983) to the infinite-dimensional case. This extension parallels in many respects the extension of the classical canonical correlation analysis to the functional data performed by Leurgans, Moyeed and Silverman (1993).

Our main theoretical results are in Theorems 2, 3, and 4. Theorem 2 relates the predictive factors to eigenvectors of a certain generalized eigenvalue problem. Since the Courant-Fischer theorem characterizes the eigenvectors as solutions of a minimax problem, the results of Theorem 2 suggest estimating the predictive factors as solutions of a regularized minimax problem. Theorem 3 proves that with a certain choice of regularization the minimax estimates of the predictive factors are consistent. To the extent that generalized eigenvalue problems often arise in different research areas, this consistency result has an independent interest. Finally, Theorem 4 shows that the forecasts obtained using the estimated predictive factors are also consistent in the sense that they converge to the optimal forecasts.

As an application, we illustrate the method using ten years of data on Eurodollar futures contracts. At each particular point in time, the available contracts have different delivery dates ranging from one month to 10 years into the future. Plotting the rate of return on the contracts against the corresponding delivery days and interpolating by cubic splines, we obtain the term structure of Eurodollar futures rates. Making such plots for every day in our sample we get our functional data set.

The futures contracts are interesting because their prices approximate forward interest rates, and therefore provide information about the interest rate term structure. Both economists and investors believe that the shape of the term structure reflects the market's future expectation for interest rates and the conditions for monetary policy. Accurate forecasting of the term structure is, therefore, a subject of tremendous practical and theoretical interest.

We find that model (1) does not provide us with a structurally stable representation of the Eurodollar futures price dynamics for the whole sample. Our preliminary analysis indicates that there might be a structural break that occurred around the onset of the recent US recession. However, restricting the sample to the period of normal growth and forecasting three months into

the future, we find that the predictive factor technique not only outperform the principal components method but also perform on par with the best available prediction methods.

Our empirical analysis contributes to the long-standing problem of whether interest rates are predictable. Some research – Duffee (2002) and Ang and Piazzesi (2003) – indicates that it is hard to predict better than simply by random walk evolution. This means that today's interest rate is the best predictor for tomorrow's interest rate, or, for that matter, for the interest rate three months from now. The subject, however, is rife with controversy. Cochrane and Piazzesi (2002) and Diebold and Li (2003) report, for example, that their methods outperform the random walk prediction. We confirm that, for our sample, the Diebold and Li outperforms the random walk and find that our predictive factors outperform the random walk for maturities larger than 4 years.

Meant to be an illustration of the predictive factors technique, our empirical analysis has several limitations. We do not attempt to use non-interest rate macroeconomic variables for interest rate forecasting. We do not aim to derive implications of interest rate predictability for the control of the economy by interest rate targeting. We also do not address the question whether financial portfolios that correspond to the predictable combinations of interest rates generate excess returns that cannot be explained by traditional risk factors. Overcoming these limitations would be a separate research effort.

The rest of the paper is organized as follows. The principal component method of estimation of the functional autoregression operator ρ is described in section 2. The predictive factor analysis is in Section 3. The data are described in Section 4. The results of estimation of the predictive factors for the interest rate curve are in Section 5. And Section 6 concludes. Proofs of three main theorems are relegated to Appendices B, C, and D, respectively.

2. THE ESTIMATION PROBLEM

In this paper, we focus on the prediction of curves $f_t(x)$ that belong to the Hilbert space of the square-summable functions of $x \in [0, \overline{X}]$. We assume that the curve dynamics is governed by a stationary functional autoregression (1). According to Theorem 3.1 of Bosq (2000), the stationarity is guaranteed by the following:

Assumption 1 *There exists an integer* $j \geq 1$ *such that* $\left\|\rho^j\right\|_{L^2} < 1$.

Here $\|\cdot\|_{L^2}$ denotes the operator norm induced by the L^2 norm.

To forecast f_{t+h} we need to estimate ρ. Let Γ_{11} be the covariance operator of random curve f_t and Γ_{21} be the cross-covariance operator for curves f_t and f_{t+h}. It is easy to see that the following useful operator relationship holds:

(2) $$\Gamma_{12} = \rho\Gamma_{11}.$$

To estimate ρ, it is tempting to substitute the covariance and cross-covariance operators with their estimates in (2) and solve the resulting equation for ρ. Unfortunately, this will not work. Indeed, the empirical covariance and cross-covariance operators are

$$\hat{\Gamma}_{11} : g \to \frac{1}{n}\sum_{i=1}^{n}\langle f_i, g\rangle f_i, \qquad \hat{\Gamma}_{12} : g \to \frac{1}{n-h}\sum_{i=1}^{n-h}\langle f_i, g\rangle f_{i+h},$$

where $<\cdot,\cdot>$ denotes the scalar product in L^2, and n is the number of available curves. Consequently, the empirical covariance operator $\hat{\Gamma}_{11}$ has a finite rank, and therefore is singular and cannot be inverted. Intuitively, the estimation problem that we are trying to handle is ill-posed: we estimate a functional dependence using a discrete set of data. As a consequence, obtaining a consistent estimate of ρ requires a regularization of the problem.

One possible regularization method has been suggested by several researchers including Ramsay and Silverman (1997) and Bosq (2000), and consists of projecting on principal components of $\hat{\Gamma}_{11}$. The idea is to determine how the operator ρ acts on those linear combinations of f_t that have the largest variation. In more detail, denote the span of k_n eigenvectors of $\hat{\Gamma}_{11}$ associated with the largest eigenvalues as H_{k_n}, and let π_{k_n} be the orthogonal projector on this subspace. Define the regularized covariance and cross-covariance estimates as follows: $\widetilde{\Gamma}_{11} = \pi_{k_n}\hat{\Gamma}_{11}\pi'_{k_n}$ and $\widetilde{\Gamma}_{12} = \pi_{k_n}\hat{\Gamma}_{12}\pi'_{k_n}$. These are simply the empirical covariance and cross-covariance operators restricted to H_{k_n}. Then define

$$\widetilde{\rho} = \pi'_{k_n}\widetilde{\Gamma}_{12}\widetilde{\Gamma}_{11}^{-1}\pi_{k_n}.$$

Note that $\widetilde{\rho}$ is $\widetilde{\Gamma}_{12}\widetilde{\Gamma}_{11}^{-1}$ on H_{k_n}, and zero on the orthogonal complement to H_{k_n}. The claim is that under certain assumptions on the covariance operator, this estimator is consistent. Here is the precise result:

Assumption 2 *All eigenvalues of* Γ_{11} *are positive and distinct.*

Assumption 3 *The first* k_n *eigenvalues of* $\hat{\Gamma}_{11}$ *are almost surely positive for any* n.

Let $a_1 = (\lambda_1 - \lambda_2)^{-1}$, and $a_i = \max\{(\lambda_{i-1} - \lambda_i)^{-1}, (\lambda_i - \lambda_{i+1})^{-1}\}$ for $i > 1$, where λ_i are eigenvalues of the covariance operator Γ_{11} ordered in the decreasing order.

Theorem 1 *Suppose that assumptions 1, 2, and 3 hold, that process f_t has a finite fourth unconditional moment, and that ρ is Hilbert-Schmidt. If for some $\beta > 1$*

$$\lambda_{k_n}^{-1} \sum_{1}^{k_n} a_j = O\left(n^{1/4} (\log n)^{-\beta}\right),$$

then we have:

$$\left\| \widetilde{\rho}_n - \rho \right\|_{L^2} \to 0$$

almost surely.

Remark: The conditions of the theorem require that the eigenvalues of the covariance matrix do not approach zero too fast, and that the eigenvalues be not too close to each other.

Proof: This is a restatement of Theorem 8.7 in Bosq (2000).

While consistent, the principal component estimation method may perform very badly in small samples if the best predictors of future evolution have little to do with the largest principal components. To see why, consider a k-factor version of the Vasicek (1977) model of the term structure of interest rates. The term structure of interest rates refers to the relationship between bonds of different maturities. It can be used to compute forward interest rates, that is, interest rates which are specified now for loans that will occur at a specified future date. A plot of the forward rates against the maturities of the corresponding loans is called the forward rate curve. Economists agree that the shape of the forward rate curve reflects the market's future expectation for interest rates and the conditions for monetary policy, which makes it an interesting object of study.

We chose the Vasicek model as an example with two goals in mind. First, we demonstrate that functional autoregression (1) is consistent with a classical and widely used financial model. Second, the example prepares a background for the application of the predictive factors technique in Section 4.

Our k-factor version of the Vasicek (1977) model begins by postulating that the short-term interest rate (spot rate) process r_t can be represented as a sum of k independent factors z_{it} that follow an Ornstein-Uhlenbeck process:

$$r_t = \sum_{i=1}^{k} z_{it},$$
$$dz_i = \alpha_i(\gamma_i - z_i)dt + \sigma_i dw_i, \quad i = 1,...,k.$$

The original model considers the case $k = 1$. Using an arbitrage argument, Vasicek (1977) shows that the entire term structure dynamics is determined by the dynamics of r_t, and gives a formula for the forward rate curve.

As explained by Dybvig (1997), the forward rate curve in a multifactor Vasicek model will be simply a sum of the forward rate curves implied by the single-factor models based on z_{it}. Therefore, for the forward rate curves (net of their means) we have (see formula (29) of Vasicek (1977)):

$$f_t(x) = \sum_{i=1}^{k} (z_{it} - \gamma_i)(1 - e^{-\alpha_i x}), \tag{3}$$

where x denotes time to maturity of the forward contract.

Since the discrete time sampling of z_{it} follows an autoregression:

$$z_{i,t+h} = \gamma_i + e^{-\alpha_i h}(z_{i,t} - \gamma_i) + \eta_{i,t+h},$$
$$\eta_i \sim i.i.d.N(0, s_i^2),$$
$$s_i^2 = \sigma_i^2 \frac{1 - e^{-2\alpha_i h}}{2\alpha_i},$$

the model falls in the functional autoregression framework. We can, for example, define the Hilbert space H as the space of functions on the positive semi-axis that are square integrable with respect to the exponential density e^{-x} so that the norm of an element of H has the following form:

$$\|f\|^2 = \int_0^\infty e^{-x} f(x)^2 \, dx.$$

The functional autoregression operator ρ is then equal to the composition of a projection on and scaling along the subspace S spanned by $1 - e^{-\alpha_i x}$, $i = 1,...,k$, and the strong H-white noise ε_t has a singular covariance operator with eigenvectors that span S.

In this example we will ignore estimation issues and simply assume that we observe all the factors and are able to estimate well the parameters of the corresponding Ornstein-Uhlenbeck processes. However, to illustrate problems with the principal components method we assume that we can use only $r < k$ factors for prediction and set the rest of the factors equal to their mean. Which factors should we use?

Let the loss from predicting f_{t+1} by $\hat{f}_{t+1}$ be $L_t = E\left\| f_{t+1} - \hat{f}_{t+1} \right\|^2$. Formula (3) implies that forecasting of the factor z_i leads to the reduction in L_t equal to the explained portion of variance of z_i, $Var(z_i) - Var(\eta_i)$, times the squared norm of $1 - e^{-\alpha_i x}$. A simple calculation reveals:

$$\Delta L_t = \frac{\sigma_i^2 \alpha_i e^{-2\alpha_i h}}{(\alpha_i + 1)(2\alpha_i + 1)}. \tag{4}$$

Consequently, the optimal choice of the factors to be used for forecasting should be based on the ranking of the loss reductions computed in (4). The first factor to be included should correspond to the largest value of ΔL_t, the second one should correspond to the second largest value of ΔL_t, and so on.

For comparison, let us check how the principal components method would rank the factors. In this example, Γ_{11} acts as follows:

$$\Gamma_{11} : g(x) \to \sum_{i=1}^{k} Var(z_i) \left\langle 1 - e^{-\alpha_i u}, g(u) \right\rangle \left(1 - e^{-\alpha_i x}\right),$$

Therefore, eigenvectors corresponding to non-zero eigenvalues of Γ_{11} are equal to $1 - e^{-\alpha_i x}$, where $i = 1,...,k$ and the eigenvalues are equal to $Var(z_i)\left\| 1 - e^{-\alpha_i x} \right\|^2$ respectively. The explicit formula for the eigenvalues is:

$$\lambda_i = \frac{\sigma_i^2 \alpha_i}{(\alpha_i + 1)(2\alpha_i + 1)}. \tag{5}$$

Hence, the principal components method chooses the factors according to the ranking induced by (5).

Clearly, the choice of the factors made by the principal components method may be very different from the optimal choice based on the ranking of (4). For example, if factor z_i has a huge instantaneous variance σ_i^2 and a large mean reversion parameter α_i, it may well happen that the principal components method would rank z_i first to include, and the optimal method would rank it last to include. In such a case, although z_i would explain almost all variation in the forward curve, its predictive power would be miniscule because z_i lacks persistence. Factors that better predict the curve would be hidden among more distant principal components.

Note that the optimal choice of factors depends on the horizon h of our forecasting problem. When the horizon goes to infinity, the first factor becomes equal to the most persistent

factor. If the most persistent factor has a small instantaneous variance then it is unlikely to be captured by a few largest principal components of the curve variation.

The above example suggests that we might be better off by searching for good predictors directly without first projecting a curve on the largest principal components. The next section develops a method for this search.

3. PREDICTIVE FACTORS

To start with, note that the principal components method is a particular way to approximate a full-ranked ρ by a reduced-rank operator. In general, a rank k approximation to ρ has form

$$\rho \approx A_k B_k^{'},$$

where $A_k : R^k \to L^2$ and $B_k^{'} : L^2 \to R^k$ are linear operators. We can think about B_k as a vector of k functionals on L^2, which we can represent by the Riesz theorem as k square summable functions $b_1(x),...,b_k(x)$. Similarly we can think about A_k as a vector of k square summable functions $a_1(x),...,a_k(x)$. The operator $A_k B_k^{'}$ acts in the following way:

$$A_k B_k^{'} : f(x) \to \sum_{i=1}^{k} \left[\int b_i(t) f(t) dt \right] a_i(x).$$

In section 2 we argued that the principal components method would not choose the approximation optimally from the forecasting point of view. We would like, therefore, to find an A_k and a $B_k^{'}$ that minimize the mean squared error of the prediction

$$E \left\| f_{t+1} - A_k B_k^{'} f_t \right\|^2 \to \min, \tag{6}$$

subject to the following normalizing constraints: i) elements of the vector B_k are orthogonal in the metric Γ_{11}, that is to say, $b_i^{'} \Gamma_{11} b_j = \delta_{ij}$, where δ is the Kronecker delta, and ii) $A_k^{'} A_k$ is diagonal with non-increasing elements on the diagonal. This particular form of normalization is chosen for its analytical convenience.

Fortier (1966) considers such problem in the static context, when predictors are not the lagged values of the forecasted series, and calls the corresponding variables $B_k^{'} f_t$ simultaneous linear predictions. In what follows, we will call $B_k^{'} f_t$ the first k **predictive factors** and A_k the corresponding **predictive factor loadings**.

Similar to principal components, the predictive factors can be defined recursively. The first predictive factor, $b_1^{'} f_t$, and the first predictive factor loading, a_1, correspond to solution of (6)

for $k=1$. (In what follows, we write $f'g$ to denote scalar products like $\int_0^{\bar{X}} f(x)g(x)dx$.) The second predictive factor and factor loading are defined as solving the same problem subject to an additional constraint, that b_2 must be orthogonal to b_1 in the metric Γ_{11}, that is to say, $b_2'\Gamma_{11}b_1=0$. And so on for the third, fourth, etc., factors and factor loadings.

Let us define an operator $\Phi=\Gamma_{11}^{1/2}\rho'\rho\Gamma_{11}^{1/2}$. We will make the following assumption:

Assumption 2a *All eigenvalues of* Φ *are positive and distinct.*

Note that since $\Gamma_{12}=\rho\Gamma_{11}$, operator Φ has an alternative representation, $\Phi=\Gamma_{11}^{-1/2}\Gamma_{21}\Gamma_{12}\Gamma_{11}^{-1/2}$, reminiscent of the cross-correlation operator $\Gamma_{XX}^{-1/2}\Gamma_{XY}\Gamma_{YY}^{-1/2}$ playing the key role in He et al. (2003) study of the existence of functional canonical correlations for functional processes X and Y. He et al. (2003) argue that a natural condition for the existence of the canonical correlations is compactness of the cross-correlation operator and derive conditions on X and Y under which the operator is well-defined and compact. In our study, the functional autoregression relationship between f_t and f_{t+h} insures compactness of Φ and problems analogous to those addressed by He et al. (2003) do not arise. The existence and the structure of solution to (5) are described by the following theorem. Its proof is relegated to Appendix B.

Theorem 2 *Under assumptions 1 and 2a we have:*

i) For any integer $k\geq 1$*, there exist* A_k *and* B_k*, solving (6). This solution is unique up to a simultaneous change in sign of* A_k *and* B_k*. Vector* B_k *consists of the first k eigenfunctions of* $\Gamma_{21}\Gamma_{12}-\lambda\Gamma_{11}$*, where eigenfunctions arranged in the order of declining eigenvalues. Vector* A_k *is equal to* $\Gamma_{12}B_k$, where Γ_{12} *acts component-wise.*

ii) The i^{th} *eigenvalue of* $\Gamma_{21}\Gamma_{12}-\lambda\Gamma_{11}$ *is equal to the reduction in the mean squared error of forecasting due to the i-th predictive factor.*

iii) If ρ *is compact,* $\left\|\rho-A_kB_k^{'}\right\|_{L^2}\to 0$ *as* $k\to\infty$.

Remark: For A_k and B_k to be well defined for a given k, it is enough to require that the first k eigenvalues of Φ are positive and distinct.

For illustration, let us return to the example of the multifactor Vasicek model. In this example, the cross-covariance operator Γ_{12} acts as follows:

$$\Gamma_{12} : g(x) \to \sum_{i=1}^{k} Cov\left(z_{it}, z_{i,t+h}\right)\left\langle 1 - e^{-\alpha_i u}, g(u)\right\rangle\left(1 - e^{-\alpha_i x}\right).$$

The non-zero eigenvalues of $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$ are equal to

$$\frac{Cov^2\left(z_{it}, z_{i,t+h}\right)}{Var\left(z_{it}\right)}\left\|1 - e^{-\alpha_i u}\right\|^2,$$

which is exactly equal to the ratio in (4) that optimally ranks the factors.

The significance of Theorem 2 is twofold. First, it relates the problem of optimal prediction to a well studied area of generalized eigenvalue problems. Second, it suggests a method for estimation of the optimal predictive factors that proceeds by solving a regularized version of the eigenvalue problem.

It seems natural to estimate A_k and B_k by computing the eigenvectors of $\hat{\Gamma}_{21}\hat{\Gamma}_{12} - \lambda\hat{\Gamma}_{11}$ and using Theorem 2. Unfortunately, similar to the situation with the canonical covariates studied by Leurgans, Moyeed and Silverman (1993), such a method of estimation would be inconsistent and the corresponding estimators meaningless. That is because the predictive factors are designed to extract those linear combinations of the data that have small variance relative to their covariance with the next period's data. Linear combinations with small variance are poorly estimated and a seemingly strong covariance (in relative terms) with the next period's data may easily be an artifact of the sample.

Leurgans, Moyeed and Silverman (1993) deal with this problem of the canonical correlation analysis by introducing a penalty for roughness of the estimated canonical covariates. We use the same idea to obtain a consistent estimate of the predictive factors.

Let us denote the j-th eigenvalue and eigenvector of the operator pencils

$\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$, $\Gamma_{21}\Gamma_{12} - \lambda\left(\Gamma_{11} + \alpha I\right)$ and $\hat{\Gamma}_{21}\hat{\Gamma}_{12} - \lambda\left(\hat{\Gamma}_{11} + \alpha I\right)$ as $\lambda_j, \lambda_{\alpha j}, \hat{\lambda}_{\alpha j}$ and $b_j, b_{\alpha j}, \hat{b}_{\alpha j}$ respectively. Here $\alpha > 0$ is a regularization parameter. We assume that the eigenvectors are normalized so that

$$b_j'\Gamma_{11}b_i = \delta_{ji},$$
$$b_{\alpha j}'\left(\Gamma_{11} + \alpha I\right)b_{\alpha i} = \delta_{ji},$$
$$\text{and} \quad \hat{b}_{\alpha j}'\left(\hat{\Gamma}_{11} + \alpha I\right)\hat{b}_{\alpha i} = \delta_{ji}$$

where δ_{ji} is the Kronecker delta.

Further, for any integer $j \geq 1$, define

$$\mu_j = \min_{b \in sp(b_1,\ldots,b_j)} \frac{b'\Gamma_{11}b}{b'b},$$

and

$$g_i = 8\lambda_1(\lambda_i - \lambda_{i+1})^{-1}.$$

In Appendix C we prove the following theorem:

Theorem 3 *Suppose that assumptions 1 and 2a hold and that process* f_t *has bounded support. If* α_n *approaches zero sufficiently slowly, so that* $\alpha_n \to 0$ *and* $(n/\log n)^{1/2}\alpha_n \to \infty$ *as* $n \to \infty$*, and if* k_n *increases sufficiently slowly, so that* $k_n \to \infty$ *and* $\alpha_n \mu_{k_n}^{-1} \to 0$ *as* $n \to \infty$*, then*

i) $\sup_{j \le k_n} \left| \hat{\lambda}_{\alpha_n j} - \lambda_j \right| \to 0$ *almost surely as* $n \to \infty$.

If in addition k_n *is chosen so that*

$$\left(\alpha_n^{-1} \left(\frac{\log n}{n} \right)^{1/2} + \alpha_n \mu_{k_n}^{-1} \right) g_{k_n} \prod_{i=1}^{k_n - 1} (1 + g_i) \to 0,$$

then

ii) $\sup_{j \le k_n} \left(\hat{b}_{\alpha_n j} - b_j \right)' \Gamma_{11} \left(\hat{b}_{\alpha_n j} - b_j \right) \to 0$

almost surely as $n \to \infty$.

Remarks:

1) When f_t does not have a bounded support but its fourth moment is finite, the theorem remains true if $(n/\log n)^{1/2}$ is replaced by $(\log n)^{\beta}(n/\log n)^{1/4}$ for some $\beta < -1/4$.

2) Of course, what can be consistently estimated is not the eigenvector itself, but the subspace generated by this eigenvector. For this reason, statement ii) holds for a particular choice of the sign of the eigenvectors $\hat{b}_{\alpha j}$ and b_j.

3) Accurate estimation of a fixed finite number of the predictive factors seems to have more practical relevance than the ability to estimate well ever-increasing number of factors. Clearly, Theorem 3 can be relaxed to have the following form:

Corollary 1 *Suppose that assumptions 1 and 2a hold and that process* f_t *has bounded support. If* $(n/\log n)^{1/2}\alpha_n \to \infty$ *and* $\alpha_n \to 0$ *as* $n \to \infty$*, then for any integer* $k \geq 1$

i) $\left|\hat{\lambda}_{\alpha_n k} - \lambda_k\right| \to 0$ *almost surely as* $n \to \infty$.

ii) $\left(\hat{b}_{\alpha_n k} - b_k\right)\Gamma_{11}\left(\hat{b}_{\alpha_n k} - b_k\right) \to 0$ *almost surely as* $n \to \infty$.

Corollary 2 *Under assumptions of Corollary 1, estimates* $\hat{b}_{\alpha j}' f_t$ *and* $\hat{\Gamma}_{12}\hat{b}_{\alpha j}$ *are consistent estimates of the predictive factor* $b_j' f_t$ *and the predictive factor loadings* $\Gamma_{12} b_j$.

Proof is in Appendix C.

In sum, Theorem 2 and its two corollaries say that by maximizing a regularized Rayleigh criterion we can consistently estimate the factors, the corresponding factor loadings, and the reduction in the mean squared error achievable by using the factors. Hence, the concept of predictive factors can be effectively used for data exploration purposes and may be a better tool for the finite-dimensional approximation than the principal components.

Moreover, when the number of the observed curves and the number of the estimated predictive factors simultaneously go to infinity, the predictive power of the autoregressive operator estimate converges to the theoretical maximum achieved by the true autoregressive operator. We formulate this precisely in Theorem 4. Suppose that f_t is chosen at random from its unconditional distribution and the task is to forecast f_{t+h}, given f_t. The best, but infeasible, forecast is ρf_t. We approximate this forecast by $\hat{A}\hat{B}' f_t$, where $\hat{B} = [\hat{b}_{\alpha 1}, ..., \hat{b}_{\alpha k_n}]$ and $\hat{A} = \hat{\Gamma}_{12}\hat{B}$.

Theorem 4 *Suppose that assumptions 1 and 2a hold, the process* f_t *has bounded support, and* ρ *is compact . If* $(n/\log n)^{1/2}\alpha_n \to \infty$, $\alpha_n \to 0$, *and* k_n *increases to infinity slowly, so that*

$$k_n g_{k_n} \prod_{i=1}^{k_n - 1}(1 + g_i)\left(\alpha_n^{-1}(n/\log n)^{-1/2} + \alpha_n \mu_{k_n}^{-1}\right) \to 0 \text{ as } n \to \infty,$$

then for any $\varepsilon > 0$,

$$\Pr\left(\left\|\rho f_t - \hat{A}\hat{B} f_t\right\| > \varepsilon \mid \hat{A}, \hat{B}\right) \to 0$$

almost surely as $n \to \infty$.

Proof is in Appendix D.

The need for regularization of the Rayleigh criterion makes estimation of the predictive factors a harder problem than estimation of the principal components. Consequently, despite the

theoretical appeal of the predictive factors technique, its performance should be judged on the basis of empirical investigations. It could conceivably happen that with a realistic amount of data theoretical advantages are washed out by estimation problems. In the rest of the paper, we use the data on the term structure of Eurodollar futures prices to illustrate the predictive factors method and to compare its predictive performance with several alternatives.

4. EMPIRICAL APPLICATION

4.1 Description of Data

We use daily settlement data on 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.

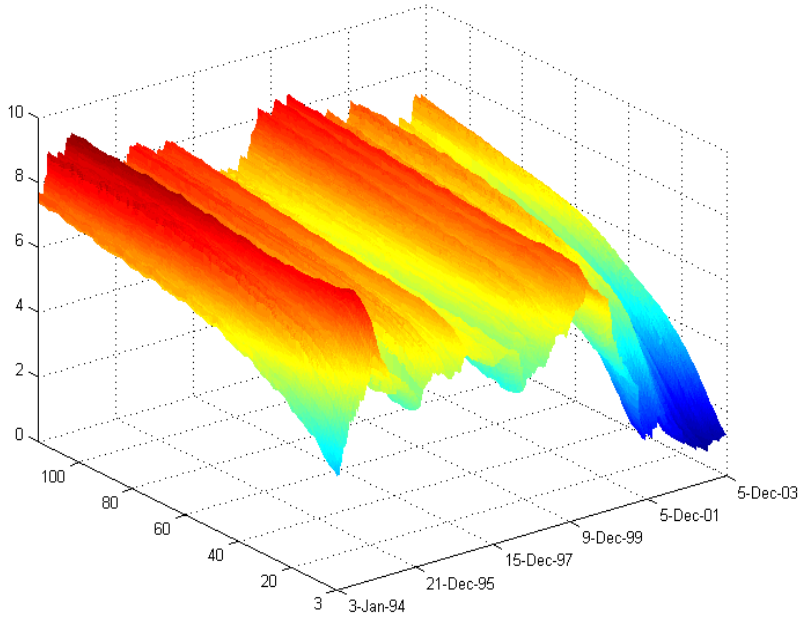
The available data start in 1982; however, we use only the data starting in 1994 when the trading in the 10-year contract appeared. We interpolated available data points by cubic splines to obtain smooth contract rate curves. To speed up the estimation, we restricted each curve to points that are 30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

The futures contracts are interesting because they provide information about interest forward rates. The main difference of the futures contract from the forward contract is that it settles during the entire life of the contract, while the forward contract settles only on the settlement date. This difference and variability of short-term interest rates make the values of the forward and futures contracts different. While the difference is small for short maturities, it can be significant for long maturities.

4.2 Three-Months-Ahead Prediction of Futures Rates

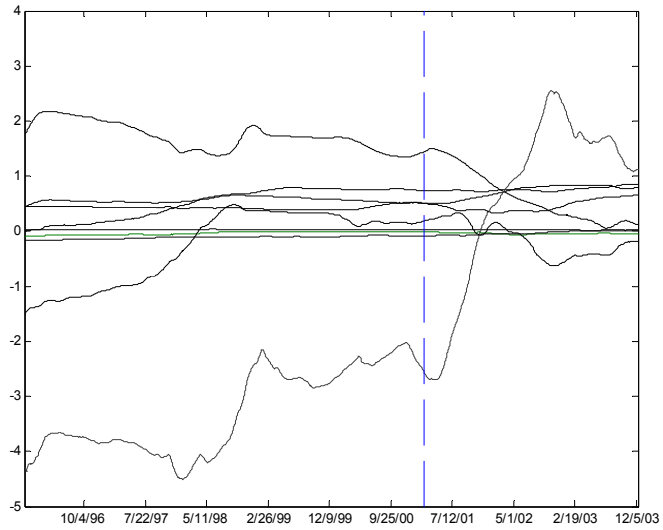
We first investigate whether the data can be sensibly represented by the functional autoregression model (1) with lag length h equal to three months. To this goal, we estimate the autoregressive operator ρ on a rolling basis using daily data. We start from the subsample that extends from 3-Jan-94 to 2-Jan-96 and increase this subsample to the full sample. We restrict the estimates to the subspace spanned by the basis of the three principal components of the sample covariance operator. In this basis, our estimate of the autoregressive operator ρ can be represented by a 3 by 3 matrix. Figure 2 presents the results of the estimation as the amount of data increases.

Figure 1 Eurodollar Futures Rates Evolution



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

Figure 2 Evolution of Matrix Entries of the Estimate of Operator ρ



Note: The operator ρ is estimated using the daily data on Eurodollar futures rates. The estimation is on a rolling basis so it uses all the information available at the time of estimation.

The dashed vertical line on the chart corresponds to the NBER's beginning date of the last US recession. The coefficients' estimates are visibly unstable between the normal growth and the

recession period. In the rest of the paper, therefore, 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.

4.3 Comparison of Predictive Factors with Other Methods

Using this subsample, we compare the predictive performance of our method with four different methods. The first one is the same functional autoregression but estimated using the principal components dimension reduction technique as discussed in Section 2. The second method is the random walk. The third method is the mean forecast, when the term structure three months ahead is predicted to be equal to the average term structure so far. Finally, we consider the Diebold-Li forecasting procedure.

Diebold and Li's (2003) procedure consists of the following steps. First, we regress the term structure on three deterministic curves, the components of the Nelson and Siegel (1987) forward rate curve:

$$f_t(T) = \beta_{1t} + \beta_{2t} e^{-\lambda T} + \beta_{3t} \lambda T e^{-\lambda T}.$$

(We fix parameter λ so it does not depend on time, as Diebold and Li do.) This regression is run for each day in a subsample. Then, the time series for the coefficients of the regression are modeled as three separate autoregressive processes of order 1 (each of the current coefficients is regressed on the corresponding coefficient from three months before). A three-months-ahead forecast of the coefficients is made, and the corresponding Nelson-Siegel forward curve is taken as the three-months-ahead forecast of the term structure.

Before making predictions we have to choose the value of the regularization parameter α and the number of the predictive factors N_{PF} for the predictive factor method, the number of the principal components N_{PC} for the principal components method, and the parameter λ 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, α=0.73, N_{PF}=3, N_{PC}=2, λ=0.0147, minimized the mean squared error of three months ahead pseudo-out-of-sample prediction for the next year, from 28-Jul-97 to 28-Jul-98.

Table 1 shows the first 5 eigenvalues of the operator pencil $\hat{\Gamma}_{21}\hat{\Gamma}_{12} - \lambda\left(\hat{\Gamma}_{11} + 0.73I\right)$, where the sample covariance and cross-covariance operators correspond to the entire normal growth subsample. Recall that eigenvalues of the pencil can be interpreted as estimates of the reductions in the mean squared error of forecasting due to the corresponding predictive factors. We see that the error reduction due to the first predictive factor is much larger than the reductions

corresponding to the other factors. The contribution of the fourth factor is essentially zero which agrees well with our cross-validation choice N_{PF}=3.

Table 1 Eigenvalues of $\hat{\Gamma}_{21}\hat{\Gamma}_{12} - \lambda\left(\hat{\Gamma}_{11} + 0.73I\right)$.

Eigenvalue	$\hat{\lambda}_{0.73,1}$	$\hat{\lambda}_{0.73,2}$	$\hat{\lambda}_{0.73,3}$	$\hat{\lambda}_{0.73,4}$	$\hat{\lambda}_{0.73,5}$
	37.12	0.93	0.04	0.00	0.00

Figure 3 shows the estimate of the first predictive factor weight b_1 when no regularization is performed, $\alpha = 0$. As expected, the non-regularized estimate makes no sense.

Figure 3 Weights of the First Predictive Factor, $\alpha = 0$.

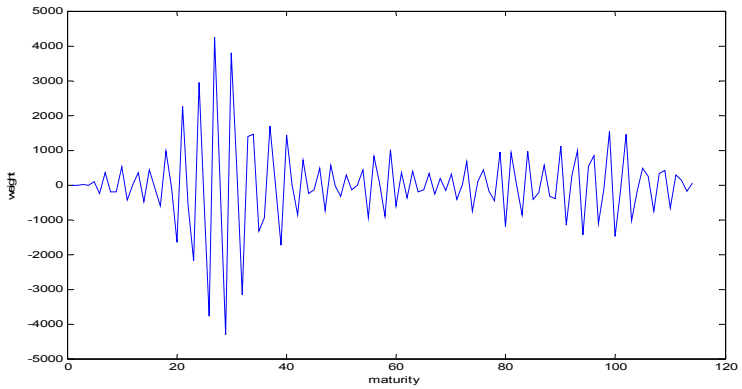
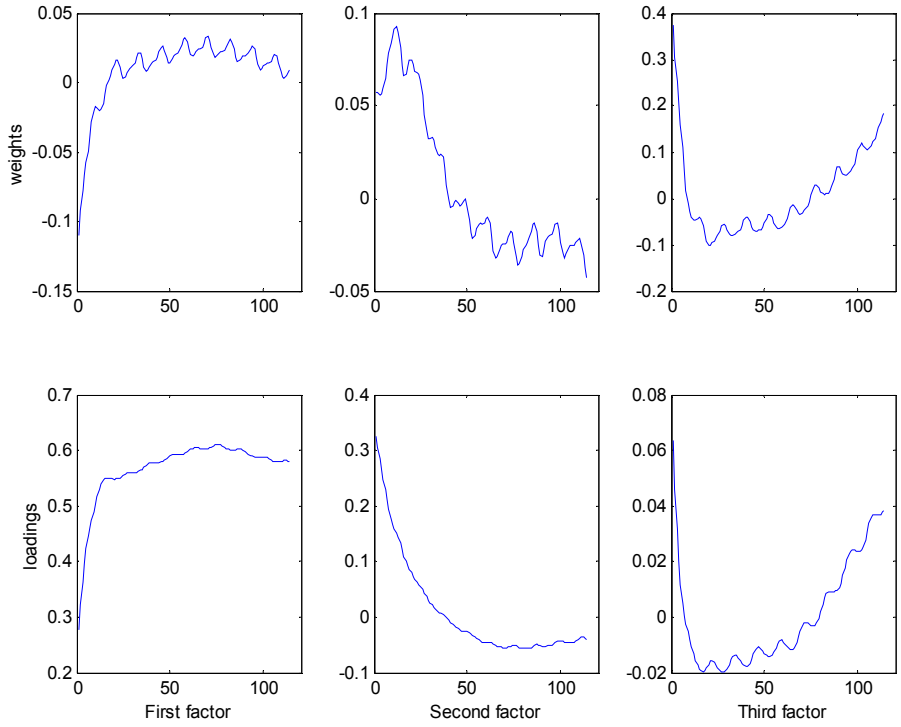


Figure 4 shows the regularized estimates of the weights of the first three predictive factors and the corresponding factor loadings, that is to say, we show functions $\hat{b}_{0.73,i}$, $i = 1,2,3$ and $\hat{a}_{0.73,i}$, $i = 1,2,3$ respectively, in the terminology of Section 3. (The entire normal growth subsample is used for these estimates.) The shapes of the predictive factor loadings roughly correspond to the "level", "slope", and "curvature" shapes of the factor loadings typically found in the literature using the classical factor analysis to study the term structure (see for example Bliss (1997)). The weights of the predictive factors correspond to the functions representing the linear functionals having the best predictive power for the entire curve. We see that the first predictive factor is essentially a linear combination of the futures contracts rates with most of the weights close to zero but relatively large weights on the rates for the contracts of short maturities. This fact is not surprising as the short-term interest rates are typically associated with the monetary policy stance, which strongly affects rates on the contracts of all maturities.

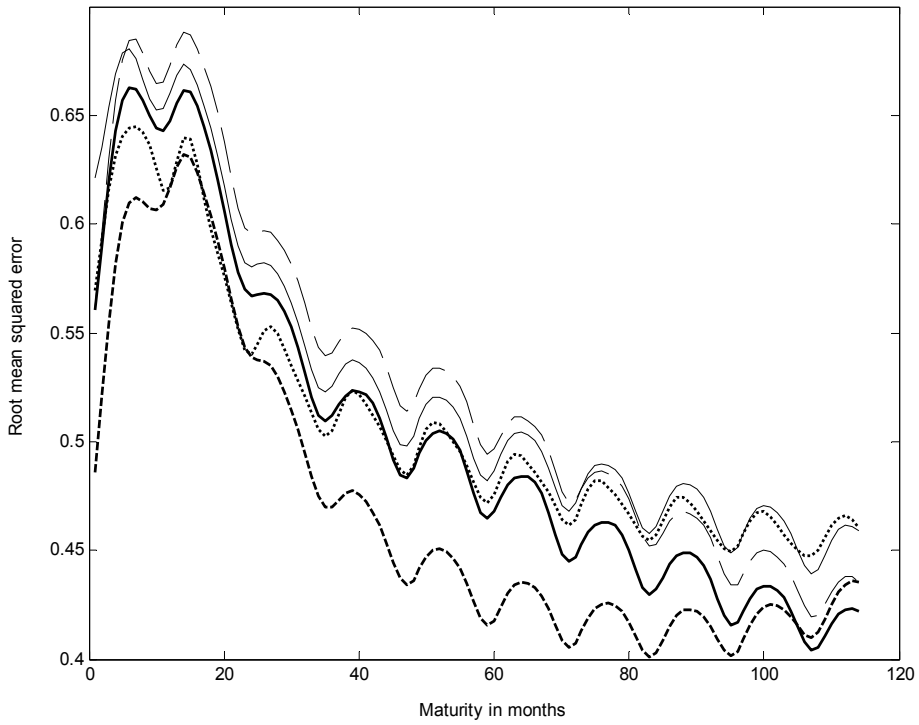
Figure 4 Weights and Loadings of the First Three Predictive Factors



To assess the predictive performance of the alternative methods considered above, we run the following experiment. We first estimate the functional autoregression and the Diebold-Li model (using the optimized parameter values) on 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 three months before the end of the normal growth subsample. After that, our forecasting would correspond to the term structures beyond the normal growth period, and therefore we stop the exercise.

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. This measure will be different for different maturities. Therefore, in figure 5 we report whole curves of the root mean squared errors of the alternative methods considered.

Figure 5 Predictive Performances of Different Forecasting Methods



The thick dashed line on the above graph corresponds to the Diebold and Li method. It outperforms all the other methods. The thick solid line is for our predictive factors method. It is the second best for the contracts of maturities longer than 4 years and the third best, losing to the random walk (thick dotted line), for the shorter maturities. The thin solid and dashed lines correspond to the principal components method with $N_{PC} = 2$ and $N_{PC} = 3$ respectively. We include the case $N_{PC} = 3$ even though our optimized parameter is $N_{PC} = 2$ to be sure that the poor performance of the principal components method relative to the predictive factors method is not caused by the fact that $N_{PC} < N_{PF}$. For our sample, three principal components work worse than 2 principal components in accordance to the cross-validation result. Note that the root mean squared error forecast error for the principal components method is uniformly worse than that for the predictive factor method. We do not report the results for the mean prediction method because it worked much worse than the rest of the methods.

5. CONCLUSION

We have shown that prediction of function-valued autoregressive processes can benefit from a novel dimension-reduction technique, the predictive factor decomposition. The 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 factors can be consistently estimated using a regularization of a generalized eigenvalue problem. To the extent that such problems often arise in different research areas, our theoretical results on consistency of the estimation procedure have an independent interest.

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.

The possible direction for further developing the new method is to investigate whether it can help in making inferences about the autoregressive operator.

APPENDIX A

Consider an abstract real Hilbert space H. Let function f_n map a probability space (Ω,A,P) to H. We call this function an H-valued random variable if the scalar product $\langle g, f_n \rangle$ is a standard random variable for every g from H. The definitions that follow are slight modifications of those in Chapters 2 and 3 of Bosq (2000).

Definition 1. If $E\|f\| < \infty$, then there exists an element of H, denoted as Ef and called the ***expectation*** of f, such that

$$E\langle g, f \rangle = \langle g, Ef \rangle, \text{ for any } g \in H.$$

Definition 2. Let f be an H-valued random variable, such that $E\|f\|^2 < \infty$ and $Ef = 0$. The ***covariance operator*** of f is the bounded linear operator on H, defined by

$$C_f(g) = E[\langle g, f \rangle f], \quad g \in H.$$

If $Ef \neq 0$, one sets $C_f = C_{f-Ef}$.

Definition 3. Let f_1 and f_2 be two H-valued random variables, such that $E\|f_1\|^2 < \infty, E\|f_2\|^2 < \infty$ and $Ef_1 = Ef_2 = 0$. Then the ***cross-covariance operators*** of f_1 and f_2 are bounded linear operators on H defined by

$$C_{f_1,f_2}(g) = E[\langle g, f_1 \rangle f_2], \quad g \in H, \quad \text{and}$$
$$C_{f_2,f_1}(g) = E[\langle g, f_2 \rangle f_1], \quad g \in H.$$

If $Ef_1 \neq 0$ or $Ef_2 \neq 0$, one sets

$$C_{f_1,f_2} = C_{f_1-Ef_1,f_2-Ef_2}, \quad \text{and}$$
$$C_{f_2,f_1} = C_{f_2-Ef_2,f_1-Ef_1}.$$

Definition 4. A sequence $\{\eta_n, \quad n \in Z\}$ of H-valued random variables is said to be ***H-white noise*** if

1) $0 < E\|\eta_n\|^2 = \sigma^2 < \infty; \quad E\eta_n = 0; \quad C_{\eta_n}$ do not depend on n, and

2) η_n is orthogonal to η_m, where $n, m \in Z, \quad n \neq m$; that is,

$$E\{\langle x, \eta_n \rangle \langle y, \eta_m \rangle\} = 0, \text{ for any } x, y \in H.$$

$\{\eta_n, \quad n \in Z\}$ is said to be a **strong H-white noise** if it satisfies 1), and

2') $\{\eta_n, \quad n \in Z\}$ is a sequence of i.i.d. H-valued random variables.

APPENDIX B

Recall that $\Phi = \Gamma_{11}^{1/2} \rho' \rho \Gamma_{11}^{1/2}$. We first prove the following Lemma:

Lemma 1 *If Assumptions 1 and 2a hold, then* λ_i *is an eigenvalue of the operator* Φ *if and only if it is an eigenvalue of the pencil* $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$. *The corresponding eigenvectors of* Φ *and* $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$, x_i *and* b_i *respectively, normalized so that* $\|x_i\| = 1$ *and* $\|\Gamma_{11}^{1/2} b_i\| = 1$, *are unique up to a change in sign and related by the formula* $x_i = \Gamma_{11}^{1/2} b_i$.

Proof: Suppose that λ_i is an eigenvalue of Φ. Assumption 2a guarantees that the corresponding normalized eigenvector x_i is unique and satisfies equation $x_i = \lambda_i^{-1} \Gamma_{11}^{1/2} \rho' \rho \Gamma_{11}^{1/2} x_i$. Using relationship $\Gamma_{12} = \rho\Gamma_{11}$, it is straightforward to check that $\lambda_i^{-1} \rho' \rho \Gamma_{11}^{1/2} x_i$ is an eigenvector of $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$ associated with eigenvalue λ_i. Now let λ_i be an eigenvalue of $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$, and b_i a corresponding normalized eigenvector. We have $\Gamma_{11}^{1/2}\left(\Phi\Gamma_{11}^{1/2} b_i - \lambda_i \Gamma_{11}^{1/2} b_i\right) = 0$. Assumption 2a implies that $\text{Ker}\,\Gamma_{11}^{1/2} = 0$, and , therefore, $\Phi\Gamma_{11}^{1/2} b_i - \lambda_i \Gamma_{11}^{1/2} b_i = 0$, which proves that λ_i is an eigenvalue of Φ, and $x_i = \Gamma_{11}^{1/2} b_i$ is the corresponding normalized eigenvector. Since x_i is unique and $\text{Ker}\,\Gamma_{11}^{1/2} = 0$, the eigenvector b_i is unique.□

Proof of Theorem 2: Transform the objective function in problem (6) as:

$$\begin{aligned} E\|f_{t+1} - AB' f_t\|^2 &= tr\left(\Gamma_{11} - AB'\Gamma_{21} - \Gamma_{12}BA' + AB'\Gamma_{11}BA'\right) \\ &= tr\left(\Gamma_{11}\right) - tr\left(AB'\Gamma_{21} + \Gamma_{12}BA'\right) + tr\left(AA'\right) \\ &= tr\left(\Gamma_{11}\right) - 2tr\left(B'\Gamma_{21}A\right) + tr\left(A'A\right), \end{aligned}$$

where the first equality follows from the fact that the expectation of the squared norm of an L^2-valued random variable is equal to the trace of its covariance operator (see Bosq (2000) p.37), and the second

equality follows from the constraint $B'\Gamma_{11}B = I_k$ imposed on B. (We omit subscript k on A_k and B_k whenever convenient to make our notations more concise.) To see that the third equality holds, write $tr(AA') = \sum_{i=1}^{\infty} e_i' AA' e_i$ and $tr(AB'\Gamma_{21}) = \sum_{i=1}^{\infty} e_i' AB'\Gamma_{21} e_i$, where $\{e_i\}$ is an arbitrary basis in L^2. Then use the fact that A and $B'\Gamma_{21}$ are finite-dimensional vectors of functions from L^2, and apply Parceval's equality.

We will first minimize the transformed objective function with respect to A, taking B as given. A necessary condition for the optimal A to exist is that the Fréchet derivative of the objective function with respect to A is equal to zero (see, for example, Proposition 2 in §7.2 and Theorem 1 in §7.4 of Luenberger (1969)). That is, $-2\Gamma_{12}B + 2A = 0$ and we have $A = \Gamma_{12}B$ in accordance with Statement i) of the theorem.

Substituting $A = \Gamma_{12}B$ into the objective function, we get

$$E\left\|f_{t+1} - AB' f_t\right\|^2 = tr(\Gamma_{11}) - tr(B'\Gamma_{21}\Gamma_{12}B) = tr(\Gamma_{11}) - tr(B'\Gamma_{11}^{1/2}\Phi\Gamma_{11}^{1/2}B).$$

We can, therefore, reformulate Problem (6) as $tr(B'\Gamma_{11}^{1/2}\Phi\Gamma_{11}^{1/2}B) \to \max$, subject to constraint $B'\Gamma_{11}^{1/2}\Gamma_{11}^{1/2}B = I_k$ and a requirement that $B'\Gamma_{11}^{1/2}\Phi\Gamma_{11}^{1/2}B$ is a diagonal matrix with non-increasing elements along the diagonal.

Assumption 2a implies that there exists a unique solution, X, to the related problem:

(B1) $$tr(X'\Phi X) \to \max$$

subject to $X'X = I_k$ and a requirement that $X'\Phi X$ is a diagonal matrix with non-increasing elements along the diagonal (see the proof of Theorem III.5.1 in Gohberg and Gohberg (1981)). The maximum is equal to the sum of the k largest eigenvalues of Φ, and the solution, X, consists of the corresponding normalized eigenvectors. By Lemma 1, $B = \Gamma_{11}^{-1/2}X$ is well defined and consists of the first k eigenvectors of $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$. It is obviously a unique solution to (5), for if $\widetilde{B}$ is another solution, then $\Gamma_{11}^{1/2}(B - \widetilde{B}) = 0$, which implies $B = \widetilde{B}$ because there are no zero eigenvalues of Φ.

Statement ii) of the theorem follows from the facts that, by Lemma 1, the eigenvalues of Φ and $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$ coincide, the maxima in (B1) and (5) are equal, and the maximum in (B1) is equal to the sum of the k largest eigenvalues of Φ.

To prove iii) note that $\overline{\mathrm{Im}}\,\Gamma_{11}^{1/2} = L^2$ because $\mathrm{Ker}\,\Gamma_{11}^{1/2} = 0$, and therefore:

(B2) $$\left\|\rho - A_k B_k'\right\| = \sup_{\|z\|\le 1}\left\|(\rho - A_k B_k')z\right\| = \sup_{\left\|\Gamma_{11}^{1/2}x\right\|\le 1}\left\|(\rho - A_k B_k')\Gamma_{11}^{1/2}x\right\|$$

Let x_i be the i-th normalized eigenvector of Φ, and let $\pi_k = \Gamma_{11}^{1/2} B_k B_k' \Gamma_{11}^{1/2}$. Note that $\{x_i\}$ forms an orthonormal basis in L^2 and $\pi_k = \sum_{i=1}^{k} \langle x_i, \cdot \rangle x_i$ by Lemma 1. We can write:

$$\begin{aligned} \rho\Gamma_{11}^{1/2}\pi_k &= \rho\Gamma_{11}^{1/2}\Gamma_{11}^{1/2} B_k B_k' \Gamma_{11}^{1/2} \\ &= \Gamma_{12} B_k B_k' \Gamma_{11}^{1/2} \\ &= A_k B_k' \Gamma_{11}^{1/2}. \end{aligned} \tag{B3}$$

Substituting (B3) into (B2), we have:

$$\left\| \rho - A_k B_k' \right\| = \sup_{\left\| \Gamma_{11}^{1/2} x \right\| \le 1} \left\| \rho\Gamma_{11}^{1/2} (I - \pi_k) x \right\|.$$

Suppose that $\left\| \rho - A_k B_k' \right\|$ does not converge to zero. Then there exists a sequence $\{z_k\}$ such that $\{\Gamma_{11}^{1/2} z_k\}$ is bounded and $\left\| \rho\Gamma_{11}^{1/2} (I - \pi_k) z_k \right\|$ does not converge to zero. Without loss of generality, we can assume that

$$\left\| \rho\Gamma_{11}^{1/2} (I - \pi_k) z_k \right\| > \varepsilon > 0 \tag{B4}$$

for any k.

Note that since, by assumption, ρ is a compact operator and $\{\Gamma_{11}^{1/2} z_k\}$ is a bounded sequence, the sequence $\{\rho\Gamma_{11}^{1/2} z_k\}$ must have a converging subsequence. Without loss of generality, let us assume that

$$\rho\Gamma_{11}^{1/2} z_k \to z \tag{B5}$$

for some $z \in L^2$.

Since x_i are the eigenvectors of $\Phi = \Gamma_{11}^{1/2} \rho' \rho \Gamma_{11}^{1/2}$, the compact operator $\rho\Gamma_{11}^{1/2}$ has a representation $\rho\Gamma_{11}^{1/2} = \sum_{i=1}^{\infty} \lambda_i^{1/2} \langle x_i, \cdot \rangle y_i$, where $\{y_i\}$ is an orthonormal basis in L^2. Let us denote $\langle x_i, z_k \rangle$ as α_{ik} and $\langle y_i, z \rangle$ as β_i. Then $(I - \pi_k) z_k = \sum_{i=k+1}^{\infty} \alpha_{ik} x_i$, and we can rewrite (B4) and (B5) as:

$$\left\| \rho\Gamma_{11}^{1/2} (I - \pi_k) z_k \right\| = \left\| \sum_{i=k+1}^{\infty} \lambda_i^{1/2} \alpha_{ik} y_i \right\| > \varepsilon \tag{B6}$$

and

$$\sum_{i=1}^{\infty} \lambda_i^{1/2} \alpha_{ik} y_i \to \sum_{i=1}^{\infty} \beta_i y_i \,. \tag{B7}$$

Now let K_1 be so large that $\left\| \sum_{i=1}^{\infty} \alpha_{ik} \lambda_i^{1/2} y_i - \sum_{i=1}^{\infty} \beta_i y_i \right\| < \varepsilon / 2$ for any $k > K_1$. Since $\{y_i\}$ is an orthonormal basis in L^2,

$$\left\|\sum_{i>k}\alpha_{ik}\lambda_i^{1/2}y_i-\sum_{i>k}\beta_i y_i\right\|\le\left\|\sum_{i=1}^{\infty}\alpha_{ik}\lambda_i^{1/2}y_i-\sum_{i=1}^{\infty}\beta_i y_i\right\|$$

and hence

(B8) $$\left\|\sum_{i>k}\alpha_{ik}\lambda_i^{1/2}y_i-\sum_{i>k}\beta_i y_i\right\|<\varepsilon/2$$

for any $k>K_1$.

Let K_2 be so large that

(B9) $$\left\|\sum_{i>k}\beta_i y_i\right\|<\varepsilon/2$$

for any $k>K_2$.

Combining (B8) and (B9), we have

$$\begin{aligned}\left\|\sum_{i>k}\alpha_{ik}\lambda_i^{1/2}y_i\right\|&=\left\|\sum_{i>k}\alpha_{ik}\lambda_i^{1/2}y_i-\sum_{i>k}\beta_i y_i+\sum_{i>k}\beta_i y_i\right\|\\&\le\left\|\sum_{i>k}\alpha_{ik}\lambda_i^{1/2}y_i-\sum_{i>k}\beta_i y_i\right\|+\left\|\sum_{i>k}\beta_i y_i\right\|<\varepsilon\end{aligned}$$

for any $k>\max\{K_1,K_2\}$. But this contradicts (B6). Hence our assumption that $\left\|\rho-A_kB_k'\right\|$ does not converge to zero is wrong and Statement iii) of the theorem is established.□

APPENDIX C

Proof of Theorem 3: We first prove an extension of Lemma 1 in Leurgans et al. (1993). Let us define $\Delta_1^{(n)}=\hat{\Gamma}_{11}-\Gamma_{11}$, $\Delta_2^{(n)}=\hat{\Gamma}_{12}-\Gamma_{12}$, $\Delta_3^{(n)}=\hat{\Gamma}_{21}\hat{\Gamma}_{12}-\Gamma_{21}\Gamma_{12}$, and $\delta_n=\max_{i=1,2,3}\left(\left\|\Delta_i^{(n)}\right\|\right)$. We have:

Lemma 2 *If Assumption 1 holds and f_t has bounded support, then $\delta_n=\mathrm{O}\left((\log n/n)^{1/2}\right)$ almost surely.*

Proof: Corollary 4.1 and Theorem 4.8 of Bosq (2000) imply that for $i=1$ and $i=2$, $\left\|\Delta_i^{(n)}\right\|=\mathrm{O}\left((\log n/n)^{1/2}\right)$ almost surely. We can also write

$$\begin{aligned}\left\|\hat{\Gamma}_{21}\hat{\Gamma}_{12}-\Gamma_{21}\Gamma_{12}\right\|&\le\left\|\hat{\Gamma}_{21}\hat{\Gamma}_{12}-\Gamma_{21}\hat{\Gamma}_{12}\right\|+\left\|\Gamma_{21}\hat{\Gamma}_{12}-\Gamma_{21}\Gamma_{12}\right\|\\&\le\left\|\hat{\Gamma}_{21}-\Gamma_{21}\right\|\left\|\hat{\Gamma}_{12}\right\|+\left\|\Gamma_{21}\right\|\left\|\hat{\Gamma}_{12}-\Gamma_{12}\right\|\\&=\mathrm{O}\left((\log n/n)^{1/2}\right)\end{aligned}$$

almost surely, which completes the proof.□

Consider the Rayleigh functionals:

$$\gamma(b)=\frac{b'\Gamma_{21}\Gamma_{12}b}{b'\Gamma_{11}b},\ \gamma_\alpha(b)=\frac{b'\Gamma_{21}\Gamma_{12}b}{b'(\Gamma_{11}+\alpha I)b},\text{ and }\ \hat{\gamma}_\alpha(b)=\frac{b'\hat{\Gamma}_{21}\hat{\Gamma}_{12}b}{b'(\hat{\Gamma}_{11}+\alpha I)b}$$

for operator pencils $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$, $\Gamma_{21}\Gamma_{12} - \lambda(\Gamma_{11} + \alpha I)$, and $\hat{\Gamma}_{21}\hat{\Gamma}_{12} - \lambda(\hat{\Gamma}_{11} + \alpha I)$, respectively. According to the maxmin principle (see Eschwé and Langer (2004)), the eigenvalues of the above operator pencils solves the following problems:

$$\lambda_j = \max_{\dim M = j} \min_{b \in M} \gamma(b), \quad \lambda_{\alpha j} = \max_{\dim M = j} \min_{b \in M} \gamma_\alpha(b), \text{ and } \hat{\lambda}_{\alpha j} = \max_{\dim M = j} \min_{b \in M} \hat{\gamma}_\alpha(b).$$

The proof of the consistency of the eigenvalue estimates consists of two parts. The first is to prove that the estimates almost surely converge to eigenvalues of the regularized problem. The second part is to prove that the eigenvalues of the regularized and the initial problem converge. The proof of the first part of the plan is based on the following proposition about the Rayleigh functionals:

Proposition 1. *Suppose that* $\alpha_n \to 0$ *and* $(n/\log n)^{1/2}\alpha_n \to \infty$ *as* $n \to \infty$. *Then*

$$\sup_{b \in L^2} \left| \hat{\gamma}_{\alpha_n}(b) - \gamma_{\alpha_n}(b) \right| \le (1 + \lambda_1)\alpha_n^{-1}\delta_n + o\left(\alpha_n^{-1}\delta_n\right) \to 0$$

almost surely as $n \to \infty$.

Proposition 1 says that almost surely the estimate of the regularized Rayleigh functional uniformly converges to the true value of the regularized Rayleigh functional. The proof is based on Lemma 2. Since it is essentially the same as that of Proposition 3 in Leurgans et al. (1993), we omit it here.

Proposition 1 implies that

$$\begin{aligned} &\sup_{j \le k_n} \left| \lambda_{\alpha j} - \hat{\lambda}_{\alpha j} \right| = \sup_{j \le k_n} \left| \max_{\dim M = j} \min_{b \in M} \gamma_\alpha(b) - \max_{\dim M = j} \min_{b \in M} \hat{\gamma}_\alpha(b) \right| \\ &\le \sup_{b \in L^2} \left| \gamma_\alpha(b) - \hat{\gamma}_\alpha(b) \right| \to 0 \quad a.s. \end{aligned} \tag{C1}$$

So the convergence of the estimates to the eigenvalues of the regularized problem is established.

Next, we prove the convergence of the eigenvalues of the regularized problem to the eigenvalues of the non-regularized problem. To this end note that since $\gamma(b) \ge \gamma_\alpha(b)$ for any $b \in L^2$, we have:

$$\lambda_j = \max_{\dim M = j} \min_{b \in M} \gamma(b) \ge \lambda_{\alpha j} = \max_{\dim M = j} \min_{b \in M} \gamma_\alpha(b).$$

Indeed, if it were not the case then for a certain j-dimensional subspace M we would have

$$\begin{aligned} \lambda_j &< \lambda_{\alpha j} = \min_{b \in M} \gamma_\alpha(b) \\ &\le \min_{b \in M} \gamma(b), \end{aligned}$$

and this would contradict the maxmin property of λ_j. On the other hand, if we take the subspace spanned by eigenvalues of the pencil $\Gamma_{21}\Gamma_{12} - \lambda\Gamma_{11}$, we have the following inequality:

$$\begin{aligned}\lambda_{\alpha j} &\geq \min_{b \in sp(b_1,\ldots,b_j)} \gamma_\alpha(b) \\ &= \min_{b \in sp(b_1,\ldots,b_j)} \frac{b'\Gamma_{21}\Gamma_{12}b}{b'\Gamma_{11}b} \Big/ \frac{b'(\Gamma_{11}+\alpha I)b}{b'\Gamma_{11}b} \\ &\geq \lambda_j /(1+\alpha\mu_j^{-1}),\end{aligned}$$

where $\mu_j = \min_{b \in sp(b_1,\ldots,b_j)} \frac{b'\Gamma_{11}b}{b'b}$. Therefore,

$$\begin{aligned}\sup_{j \leq k_n} \left|\lambda_j - \lambda_{\alpha j}\right| &\leq \sup_{j \leq k_n} \left(\lambda_j - \lambda_j /(1+\alpha\mu_j^{-1})\right) \\ &< \lambda_1 \left(1 - 1/(1+\alpha\mu_{k_n}^{-1})\right) \to 0\end{aligned}, \tag{C2}$$

where the last step is by the assumption of convergence of $\alpha\mu_{k_n}^{-1}$ to zero. This establishes convergence of the eigenvalues of the regularized and non-regularized problems.

Joining the two parts of the proof together, we have

$$\sup_{j \leq k_n} \left|\lambda_j - \hat{\lambda}_{\alpha j}\right| \leq \sup_{j \leq k_n} \left|\lambda_j - \lambda_{\alpha j}\right| + \sup_{j \leq k_n} \left|\lambda_{\alpha j} - \hat{\lambda}_{\alpha j}\right| \to 0 \text{ a.s.}, \tag{C3}$$

which proves Part i) of the theorem.

Let us now turn to Part ii) of the theorem. Denote $\left(\hat{b}_{\alpha j} - b_j\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right)$ as d_j and $\alpha\hat{b}_{\alpha j}'\hat{b}_{\alpha j}$ as m_j. Below, we are going to find an upper bound on $\sup_{j \leq k_n}\left(d_j + m_j\right)$ and show that this bound tends to zero almost surely. This implies that $\sup_{j \leq k_n} d_j$ tends to 0 and therefore Part ii) of the theorem.

Consider the least squares regression of $\hat{b}_{\alpha j}$ on eigenvectors on the non-regularized problem $b_1,\ldots,b_j$ in metric Γ_{11}:

$$\hat{b}_{\alpha j} = \sum_{i=1}^{j} \beta_{ji} b_i + s_j .$$

Here for any $i \leq j$,

$$\begin{aligned}\beta_{ji} &= (B_j'\Gamma_{11}B_j)^{-1}\hat{b}_{\alpha j}'\Gamma_{11}b_i \\ &= \hat{b}_{\alpha j}'\Gamma_{11}b_i ,\end{aligned}$$

where B_j is the matrix with columns $b_1,\ldots,b_j$ and the last equality follows because of the normalization of $b_1,\ldots,b_j$. The residuals s_j are orthogonal to b_i in metric Γ_{11} and have the following properties. First,

$$\lambda_i^{-1} s_j'\Gamma_{21}\Gamma_{12}b_i = s_j'\Gamma_{11}b_i = 0 . \tag{C4}$$

Second,

$$
\begin{aligned}
s_j'\Gamma_{11}s_j &= \left(\hat{b}_{\alpha j} - \textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - \textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right) \\
&= \hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha j} - 2\left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right)\Gamma_{11}\hat{b}_{\alpha j} + \left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right)\Gamma_{11}\left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right) \\
&= \hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha j} - \left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right)\Gamma_{11}\left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right) \\
&= \hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha j} - \beta_{j1}^2 - \ldots - \beta_{jj}^2 .
\end{aligned}
$$

Subtracting the normalization equation $0 = \hat{b}_{\alpha j}'\left(\hat{\Gamma}_{11} + \alpha I\right)\hat{b}_{\alpha j} - 1$, we get:

(C5) $$s_j'\Gamma_{11}s_j = \hat{b}_{aj}'\left(\Gamma_{11} - \hat{\Gamma}_{11}\right)\hat{b}_{aj} - m_j - \beta_{j1}^2 - \ldots - \beta_{jj}^2 + 1,$$

where $m_j = \alpha\hat{b}_{\alpha j}'\hat{b}_{\alpha j}$.

Another expression for $s_j'\Gamma_{11}s_j$ follows from the following equation:

$$
\begin{aligned}
\hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha j} &= \left(\hat{b}_{\alpha j} - b_j\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right) + 2b_j'\Gamma_{11}\hat{b}_{aj} - b_j\Gamma_{11}b_j \\
&= \left(\hat{b}_{\alpha j} - b_j\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right) + 2\beta_{jj} - 1,
\end{aligned}
$$

which implies that:

(C6) $$s_j'\Gamma_{11}s_j = d_j - \beta_{j1}^2 - \ldots - \beta_{j,j-1}^2 - (1-\beta_{jj})^2,$$

where $d_j = \left(\hat{b}_{\alpha j} - b_j\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right)$.

Subtracting (C5) from (C6), rearranging, and using the fact that

$$
\begin{aligned}
\hat{b}_{\alpha j}'\left(\Gamma_{11} - \hat{\Gamma}_{11}\right)\hat{b}_{\alpha j} &\le \hat{b}_{\alpha j}'\hat{b}_{\alpha j}\left\|\Gamma_{11} - \hat{\Gamma}_{11}\right\| \\
&\le \hat{b}_{\alpha j}'\left(\hat{\Gamma}_{11} + \alpha I\right)\hat{b}_{\alpha j}\alpha^{-1}\left\|\Gamma_{11} - \hat{\Gamma}_{11}\right\| \\
&\le \alpha^{-1}\delta_n ,
\end{aligned}
$$

we obtain:

(C7)
$$
\begin{aligned}
d_j + m_j &= \hat{b}_{\alpha j}'\left(\Gamma_{11} - \hat{\Gamma}_{11}\right)\hat{b}_{\alpha j} + (1-\beta_{jj})^2 - \beta_{jj}^2 + 1 \\
&\le 2\left(1-\beta_{jj}\right) + \alpha^{-1}\delta_n .
\end{aligned}
$$

From this expression it is clear that we can show that $d_j + m_j$ is small if we show that β_{jj} is close to 1.

The following is devoted to the proof of this property of β_{jj}.

We can write an expression for the norm of the residual in the metric given by $\Gamma_{21}\Gamma_{12}$:

(C8)
$$
\begin{aligned}
s_j'\Gamma_{21}\Gamma_{12}s_j &= \left(\hat{b}_{\alpha j} - \textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right)\Gamma_{21}\Gamma_{12}\left(\hat{b}_{\alpha j} - \textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right) \\
&= \hat{b}_{\alpha j}'\Gamma_{21}\Gamma_{12}\hat{b}_{\alpha j} - \left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right)\Gamma_{21}\Gamma_{12}\left(\textstyle\sum_{i=1}^{j}\beta_{ji}b_i\right) \\
&= \hat{b}_{\alpha j}'\Gamma_{21}\Gamma_{12}\hat{b}_{\alpha j} - \lambda_1\beta_{j1}^2 - \ldots - \lambda_j\beta_{jj}^2 \\
&= \hat{b}_{\alpha j}'\left(\Gamma_{21}\Gamma_{12} - \hat{\Gamma}_{21}\hat{\Gamma}_{12}\right)\hat{b}_{\alpha j} + \hat{\lambda}_{\alpha j} - \lambda_1\beta_{j1}^2 - \ldots - \lambda_j\beta_{jj}^2 ,
\end{aligned}
$$

where the second equality use (C4), and the fourth holds by subtraction of the normalization equality $\hat{b}_{\alpha j}\hat{\Gamma}_{21}\hat{\Gamma}_{12}\hat{b}_{\alpha j}-\hat{\lambda}_{\alpha j}=0$.

We also have:

(C9) $$s_j'\Gamma_{21}\Gamma_{12}s_j-\lambda_{j+1}s_j'\Gamma_{11}s_j\le 0.$$

This follows because s_j is orthogonal in metric Γ_{11} to the first j eigenvectors of pencil $\Gamma_{21}\Gamma_{12}-\lambda\Gamma_{11}$ and because the $(j+1)^{\text{st}}$ eigenvalue of the pencil can be characterized by the following rule: $\lambda_{j+1}=\max_{b\perp_{\Gamma_{11}} sp(b_1,\ldots,b_j)}\gamma(b)$. Consequently,

$$\frac{s_j'\Gamma_{21}\Gamma_{12}s_j}{s_j'\Gamma_{11}s_j}\le\lambda_{j+1}.$$

Expanding (C9) using (C5) and (C8) we get:

$$\hat{b}_{\alpha j}'\left(\Gamma_{21}\Gamma_{12}-\hat{\Gamma}_{21}\hat{\Gamma}_{12}\right)\hat{b}_{\alpha j}+\hat{\lambda}_{\alpha j}-\sum_{i=1}^{j}\lambda_i\beta_{ji}^2$$
$$-\lambda_{j+1}\left(\hat{b}_{aj}'\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)\hat{b}_{aj}-m_j+\sum_{i=1}^{j}\beta_{ji}^2+1\right)<0.$$

Or, after a rearrangement:

$$1-\beta_{jj}^2\le\left(\lambda_j-\lambda_{j+1}\right)^{-1}\left(\lambda_{j+1}\hat{b}_{\alpha j}'(\Gamma_{11}-\hat{\Gamma}_{11})\hat{b}_{\alpha j}-\hat{b}_{\alpha j}'(\Gamma_{21}\Gamma_{12}-\hat{\Gamma}_{21}\hat{\Gamma}_{12})\hat{b}_{\alpha j}\right.$$
$$\left.+\lambda_j-\hat{\lambda}_{\alpha j}+\textstyle\sum_{i=1}^{j-1}(\lambda_i-\lambda_{j+1})\beta_{ji}^2-\lambda_{j+1}m_j\right).$$

Recalling that

$$\hat{b}_{\alpha j}'\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)\hat{b}_{\alpha j}\le\alpha^{-1}\delta_n,$$
$$\hat{b}_{\alpha j}'\left(\Gamma_{21}\Gamma_{12}-\hat{\Gamma}_{21}\hat{\Gamma}_{12}\right)\hat{b}_{\alpha j}\le\alpha^{-1}\delta_n,$$

that $m_j>0$, and that, from (C3), (C2), (C1), and Proposition 1:

$$\left|\lambda_j-\hat{\lambda}_{\alpha j}\right|\le\lambda_j\left(\alpha\mu_j^{-1}+o(\alpha\mu_j^{-1})\right)+(1+\lambda_1)\alpha^{-1}\delta_n+o(\alpha^{-1}\delta_n),$$

we have, for all n large enough:

(C10) $$1-\beta_{jj}^2\le\left(\lambda_j-\lambda_{j+1}\right)^{-1}\left((3+\lambda_1+\lambda_{j+1})\alpha^{-1}\delta_n+2\lambda_j\alpha\mu_j^{-1}+\textstyle\sum_{i=1}^{j-1}(\lambda_i-\lambda_{j+1})\beta_{ji}^2\right).$$

We would like to write this inequality with $1-\beta_{jj}$ instead of $1-\beta_{jj}^2$. Note that the right-hand side of (C10) is positive and therefore the desired inequality holds automatically if $1-\beta_{jj}\le 0$. In the case of $1-\beta_{jj}>0$, we use the freedom in the choice of the sign of the eigenvector $\hat{b}_{\alpha j}$, and choose it so that the

coefficient β_{jj} is positive. This choice implies that $1-\beta_{jj} \leq 1-\beta_{jj}^2$ and the desired inequality holds. Therefore, we can write:

(C10a) $1-\beta_{jj} \leq \left(\lambda_j - \lambda_{j+1}\right)^{-1}\left((3+\lambda_1+\lambda_{j+1})\alpha^{-1}\delta_n + 2\lambda_j \alpha \mu_j^{-1} + \sum_{i=1}^{j-1}(\lambda_i - \lambda_{j+1})\beta_{ji}^2\right).$

Combining this inequality with (C7) and using the fact that $0 < \lambda_j \leq \lambda_1$, we get for large enough n:

(C11) $d_j + m_j \leq 2\left(\lambda_j - \lambda_{j+1}\right)^{-1}\left(3(1+\lambda_1)\alpha^{-1}\delta_n + 2\lambda_j \alpha \mu_j^{-1} + \lambda_1 \sum_{i=1}^{j-1}\beta_{ji}^2\right).$

Now, we analyze the behavior of the least squares regression coefficients $\beta_{ji}, i<j$ as $n \to \infty$. First, note that the normalization $\hat{b}_{\alpha j}'\left(\hat{\Gamma}_{11}+\alpha I\right)\hat{b}_{\alpha i} = \delta_{ji}$ implies:

(C12) $\hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha j} = \hat{b}_{\alpha j}'\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)\hat{b}_{\alpha j} + \hat{b}_{\alpha j}'\hat{\Gamma}_{11}\hat{b}_{\alpha j} \leq \alpha^{-1}\delta_n + 1.$

Second, we can write:

$$
\begin{aligned}
\left(\hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha i}\right)^2 &= \left(\hat{b}_{\alpha j}'\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)\hat{b}_{\alpha i} - \alpha \hat{b}_{\alpha j}'\hat{b}_{\alpha i}\right)^2 \\
&\leq 2\left[\hat{b}_{\alpha j}'\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)\hat{b}_{\alpha i}\right]^2 + 2\left[\alpha \hat{b}_{\alpha i}'\hat{b}_{\alpha i}\right]^2 \\
&\leq 2\left[\hat{b}_{\alpha j}'\hat{b}_{\alpha j}\right]\left[\hat{b}_{\alpha i}'\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)^2\hat{b}_{\alpha i}\right] + 2\left[\alpha \hat{b}_{\alpha i}'\hat{b}_{\alpha i}\right]^2 \\
&= 2\alpha \hat{b}_{\alpha j}'\hat{b}_{\alpha j}\left\{\alpha^{\frac{1}{2}}\hat{b}_{\alpha i}'\left[\alpha^{-1}\left(\Gamma_{11}-\hat{\Gamma}_{11}\right)\right]^2\alpha^{\frac{1}{2}}\hat{b}_{\alpha i} + \alpha \hat{b}_{\alpha i}'\hat{b}_{\alpha i}\right\} \\
&\leq 2m_i + o(\alpha^{-1}\delta_n),
\end{aligned}
\tag{C13}
$$

where the first inequality holds by the inequality $(a-b)^2 \leq 2(a^2+b^2)$, the second inequality uses the Cauchy-Schwarz inequality, and the third inequality uses the fact that $\alpha \hat{b}_{\alpha j}'\hat{b}_{\alpha j} \leq \hat{b}_{\alpha j}'\left(\Gamma_{11}+\alpha I\right)\hat{b}_{\alpha j} = 1$. Using (C12) and (C13), we have:

$$
\begin{aligned}
\beta_{ji}^2 = \left(\hat{b}_{\alpha j}'\Gamma_{11}b_i\right)^2 &= \left(\hat{b}_{\alpha j}'\Gamma_{11}\left(b_i - \hat{b}_{\alpha i}\right) + \hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha i}\right)^2 \\
&\leq 2\left(\hat{b}_{\alpha j}'\Gamma_{11}\left(b_i - \hat{b}_{\alpha i}\right)\right)^2 + 2\left(\hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha i}\right)^2 \\
&\leq 2\hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha j}\left[\left(b_i - \hat{b}_{\alpha i}\right)\Gamma_{11}\left(b_i - \hat{b}_{\alpha i}\right)\right] + 2\left(\hat{b}_{\alpha j}'\Gamma_{11}\hat{b}_{\alpha i}\right)^2 \\
&\leq 4\left(d_i + m_i\right)
\end{aligned}
\tag{C14}
$$

for large enough n. Substituting (C14) into (C11), rearranging, and using the fact that, for $j \leq k_n$, $\mu_j^{-1} \leq \mu_{k_n}^{-1}$, we obtain for large enough n:

(C15) $d_j + m_j \leq 8\lambda_1\left(\lambda_j - \lambda_{j+1}\right)^{-1}\left((1+\lambda_1^{-1})\alpha^{-1}\delta_n + \alpha \mu_{k_n}^{-1} + \sum_{i=1}^{j-1}\left(d_i + m_i\right)\right).$

It is straightforward to check that, if a sequence of real numbers $\{x_j\}$ satisfies recursive inequalities $x_1 \le g_1 f$ and $x_j \le g_j\left(f + \sum_{i=1}^{j-1} x_i\right)$ for $j \ge 2$, then $x_j \le f g_j \prod_{i=1}^{j-1}(1+g_i)$. Applying this observation to (C15), we get:

$$\text{(C16)} \qquad \sup_{j \le k_n}\left(d_j + m_j\right) \le \left((1+\lambda_1^{-1})\alpha^{-1}\delta_n + \alpha\mu_{k_n}^{-1}\right) g_{k_n} \prod_{i=1}^{k_n-1}(1+g_i),$$

where $g_i = 8\lambda_1(\lambda_i - \lambda_{i+1})^{-1}$. The right-hand side of (C16) tends to zero almost surely as $n \to \infty$ by Lemma 2, which says that $\delta_n = O((\log n / n)^{1/2})$, and by the assumptions of the theorem. This completes our proof of Statement ii). □

Proof of Corollary 2:. Suppose that we estimate a predictive factor, $b_j' f_t$, where f_t is chosen at random from its unconditional distribution, by $\hat{b}_{\alpha j}' f_t$. We can bound the probability that the difference between the factor and its estimate is greater by absolute value than ε as follows:

$$\begin{aligned}\Pr\left\{\left|\hat{b}_{\alpha j}' f_t - b_j' f_t\right| > \varepsilon \mid \hat{b}_{\alpha j}\right\} &\le \varepsilon^{-2} Var\left[\left(\hat{b}_{\alpha j} - b_j\right) f_t \mid \hat{b}_{\alpha j}\right] \\ &= \varepsilon^{-2}\left(\hat{b}_{\alpha j} - b_j\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right).\end{aligned}$$

According to Statement ii) of Corollary 1, this bound tends to zero almost surely as $n \to \infty$.

Statement ii) of Corollary 1 also implies convergence in probability of our estimates of the predictive factor loadings, $\hat{a}_{\alpha j}$. Indeed, we have:

$$\begin{aligned}\left\|\hat{a}_{\alpha j} - a_j\right\| &= \left\|\hat{\Gamma}_{12}\hat{b}_{\alpha j} - \Gamma_{12} b_j\right\| \\ &\le \left\|\left(\hat{\Gamma}_{12} - \Gamma_{12}\right)\hat{b}_{\alpha j}\right\| + \left\|\Gamma_{12}\left(\hat{b}_{\alpha j} - b_j\right)\right\|.\end{aligned}$$

Lemma 2 from Appendix C implies that the first term in the above expression tends in probability to 0. For the second term we have:

$$\begin{aligned}\left\|\Gamma_{12}\left(\hat{b}_{\alpha j} - b_j\right)\right\| &= \left\|\rho\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right)\right\| \\ &\le \sqrt{\left(\hat{b}_{\alpha j} - b_j\right)\Gamma_{11}\left(\hat{b}_{\alpha j} - b_j\right)}\left\|\rho\Gamma_{11}^{1/2}\right\|,\end{aligned}$$

which tends to zero almost surely according to Statement ii) of Corollary 1.

APPENDIX D

Proof of Theorem 4: First, note that

$$\left\|\rho f_t - \hat{A}\hat{B}' f_t\right\| \le a_1 + a_2 + a_3 + a_4,$$

where $a_1 = \left\|(\rho - AB')f_t\right\|$, $a_2 = \left\|\Gamma_{12}\left(B - \hat{B}\right)B' f_t\right\|$, $a_3 = \left\|\left(\Gamma_{12} - \hat{\Gamma}_{12}\right)\hat{B}B' f_t\right\|$, and $a_4 = \left\|\hat{\Gamma}_{12}\hat{B}\left(B - \hat{B}\right) f_t\right\|$.

Let $\xi = \frac{\varepsilon}{4}$. We have $\Pr\left(\left\|\rho f_t - \hat{A}\hat{B}' f_t\right\| > \varepsilon \mid \hat{A}, \hat{B}\right) \le \sum_{i=1}^{4} \Pr\left(a_i > \xi \mid \hat{A}, \hat{B}\right)$.

Below we will show that each of the terms in the latter expression converges to zero almost surely.

Since $\|f_t\| < \infty$, Statement iii) of theorem 2 implies that $a_1 \to 0$ and $\Pr\left(a_i > \xi \mid \hat{A}, \hat{B}\right) \to 0$ a.s.

Further, using (C13), we have:

$$\begin{aligned}
\Pr\left(a_2 > \xi \mid \hat{A}, \hat{B}\right) &\le \xi^{-2} tr\left[\Gamma_{12}\left(B - \hat{B}\right)B'\Gamma_{11}B\left(B - \hat{B}\right)'\Gamma_{21}\right] \\
&= \xi^{-2} tr\left[\left(B - \hat{B}\right)'\Gamma_{21}\Gamma_{12}\left(B - \hat{B}\right)\right] \\
&= \xi^{-2} tr\left[\left(B - \hat{B}\right)'\Gamma_{11}^{1/2}\Phi\Gamma_{11}^{1/2}\left(B - \hat{B}\right)\right] \\
&\le \lambda_1 \xi^{-2} tr\left[\left(B - \hat{B}\right)'\Gamma_{11}\left(B - \hat{B}\right)\right] \\
&\le \lambda_1 \xi^{-2} k_n g_{k_n} \prod_{i=1}^{k_n - 1}(1 + g_i)\left((1 + \lambda_1^{-1})\alpha^{-1}\delta_n + \alpha\mu_{k_n}^{-1}\right) \to 0 \quad a.s.
\end{aligned}$$

For a_3 and a_4 we have:

$$\begin{aligned}
\Pr\left(a_3 > \xi \mid \hat{A}, \hat{B}\right) &\le \xi^{-2} tr\left[\left(\Gamma_{12} - \hat{\Gamma}_{12}\right)\hat{B}B'\Gamma_{11}B\hat{B}'\left(\Gamma_{21} - \hat{\Gamma}_{21}\right)\right] \\
&= \xi^{-2} tr\left[\left(\Gamma_{12} - \hat{\Gamma}_{12}\right)\hat{B}\hat{B}'\left(\Gamma_{21} - \hat{\Gamma}_{21}\right)\right] \\
&= \xi^{-2} tr\left[\hat{B}'\left(\Gamma_{21} - \hat{\Gamma}_{21}\right)\left(\Gamma_{12} - \hat{\Gamma}_{12}\right)\hat{B}\right] \\
&\le \xi^{-2} tr\left[\alpha\hat{B}'\hat{B}\right]\frac{\left\|\left(\Gamma_{21} - \hat{\Gamma}_{21}\right)\left(\Gamma_{12} - \hat{\Gamma}_{12}\right)\right\|}{\alpha} \\
&\le \xi^{-2} k_n \frac{\delta_n^2}{\alpha} \to 0 \quad a.s.,
\end{aligned}$$

and

$$\begin{aligned}
\Pr\left(a_4 > \xi \mid \hat{A}, \hat{B}\right) &\le \xi^{-2} tr\left[\hat{\Gamma}_{12}\hat{B}\left(B - \hat{B}\right)'\Gamma_{11}\left(B - \hat{B}\right)\hat{B}'\hat{\Gamma}_{21}\right] \\
&= \xi^{-2} tr\left[\left(B - \hat{B}\right)'\Gamma_{11}\left(B - \hat{B}\right)\hat{B}'\hat{\Gamma}_{21}\hat{\Gamma}_{12}\hat{B}\right] \\
&\le \xi^{-2} tr\left[\left(B - \hat{B}\right)'\Gamma_{11}\left(B - \hat{B}\right)\right]\left\|\hat{B}'\hat{\Gamma}_{21}\hat{\Gamma}_{12}\hat{B}\right\| \\
&\le \xi^{-2} tr\left[\left(B - \hat{B}\right)'\Gamma_{11}\left(B - \hat{B}\right)\right] \\
&\le \xi^{-2} k_n g_{k_n} \prod_{i=1}^{k_n - 1}(1 + g_i)\left((1 + \lambda_1^{-1})\alpha^{-1}\delta_n + \alpha\mu_{k_n}^{-1}\right) \to 0 \quad a.s.
\end{aligned}$$

This completes the proof of theorem 4. □

REFERENCES

A. Ang and M. Piazzesi (2003) "A No-Arbitrage Vector Autoregression of Term Structure Dynamics with Macroeconomic and Latent Variables" *Journal of Monetary Economics*, **50**, 745-787

T. W. Anderson (1984) *An Introduction to Multivariate Statistical Analysis,* 2nd edition, John Wiley and Sons

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

P. C. Besse and H. Cardot (1996) "Approximation Spline de la Prevision d'un Processus Fonctionnel Autoregressif d'Ordre 1" *Canadian Journal of Statistics,* **24,** 467-487

P. C. Besse, H. Cardot and D. B. Stephenson (2000) "Autoregressive Forecasting of Some Functional Climatic Variations" *Scandinavian Journal of Statistics*, **27,** 673-687.

R. R. Bliss (1997) "Movements in the Term Structure of Interest Rates", Federal Reserve Bank of Atlanta *Economic Review*, Vol. 82, No. 4, Fourth Quarter, 16–33.

D. Bosq (1991) "Modelization, Non-parametric Estimation and Prediction for Continuous Time Processes" in *Nonparametric Functional Estimation and Related Topics* (ed. G. Roussas) 509-529. Nato, ASI series.

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

A. Cavallini, G. C. Montanari, M. Loggini, O. Lessi, and M. Cacciari (1994) "Nonparametric Prediction of Harmonic Levels in Electrical Networks" *Proceedings of IEEE ICHPS VI,* Bologna, 165-171.

J. H. Cochrane and M. Piazzesi (2002) "Bond Risk Premia" *NBER Working paper* 9178

J. Damon and S. Guillas (2002) "The Inclusion of Exogenous Variables in Functional Autoregressive Ozone Forecasting" *Environmetrcs,* **13,** 759-774

F. X. Diebold and C. Li (2003) "Forecasting the Term Structure of Government Bond Yields" Working Paper (available at http://www.ssc.upenn.edu/~diebold)

G. R. Duffee (2002) "Term Premia and Interest Rate Forecasts in Affine Models" *Journal of Finance*, **57**, 405-443

P.H. Dybvig (1997) "Bond and Bond Option Pricing Based on the Current Term Structure" in *Mathematics Of Derivative Securities* ed. by M.A.H. Dempster and S.R.Pliska, Cambridge University Press, 271-293

D. Eschwe and M. Langer (2004) "Variational Principles for Eigenvalues of Self-Adjoint Operator Functions" *Integral Equations and Operator Theory*, 49, 287-321

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

I. Gohberg and S. Gohberg (1981) "Basic Operator Theory", Birkhauser, Boston, Basel, Berlin.

G. He, H.G. Muller, and J.L. Wang (2003) "Functional Canonical Analysis for Square Integrable Stochastic Processes", *Journal of Multivariate Analysis,* **85,** 54-77.

S. E. Leurgans, R. A. Moyeed, and B. W. Silverman (1993) "Canonical Correlation Analysis when Data are Curves" *Journal of Royal Statistical Society B*, **55**, 725-740

D. G. Luenberger (1969) *Optimization by Vector Space Methods*, John Wiley & Sons, Inc. New York, Chichester, Weiheim, Brisbane, Singapore, Toronto

C. R Nelson and A. F. Siegel (1987) "Parsimonious Modeling of Yield Curves" *Journal of Business*, **60**(4), 473-489

M. Piazzesi (2003) "Bond Yields and the Federal Reserve" *Working paper*

J.O. Ramsay and B.W. Silverman (1997), *Functional Data Analysis*, Springer, New York.

J.O. Ramsay and B.W. Silverman (2002), *Applied Functional Data Analysis*, Springer, New York.

G. Reinsel (1983) "Some Results on Multivariate Autoregressive Index Models" *Biometrika*, 70, 145-156

O. A. Vasicek (1977) “An Equilibrium Characterization of the Term Structure” *Journal of Financial Economics*, **5**, 177-188