CONSISTENT FUNCTIONAL PCA FOR FINANCIAL TIME-SERIES

Sebastian Jaimungal
Department of Statistics and Mathematical Finance Program,
University of Toronto, Toronto, ON, Canada
sebastian.jaimungal@utoronto.ca

Eddie K. H. Ng
The Edward S. Rogers Sr. Department of
Electrical and Computer Engineering
University of Toronto, Toronto, ON, Canada
eddie@psi.toronto.edu

Date: April 30, 2007

ABSTRACT
Functional Principal Component Analysis (FPCA) provides a powerful and natural way to model functional financial data sets (such as collections of time-indexed futures and interest rate yield curves). However, FPCA assumes each sample curve is drawn from an independent and identical distribution. This assumption is axiomatically inconsistent with financial data; rather, samples are often interlinked by an underlying temporal dynamical process. We present a new modeling approach using Vector auto-regression (VAR) to drive the weights of the principal components. In this novel process, the temporal dynamics are first learned and then the principal components extracted. We dub this method the VAR-FPCA. We apply our method to the NYMEX light sweet crude oil futures curves and demonstrate that it contains significant advantages over the conventional FPCA in applications such as statistical arbitrage and risk management.

KEY WORDS
Vector auto-regression, functional principal component analysis, risk management, statistical arbitrage, forward curve modeling

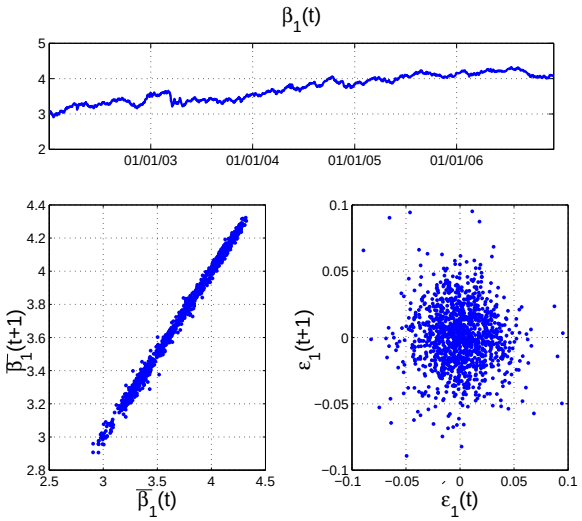

Figure 1. Removal of temporal structure. (Top) Shows the data curve fitting coefficient $\boldsymbol{\beta}_1(t)$; (Bottom left) Scatter plot of $\bar{\boldsymbol{\beta}}_1(t) = \boldsymbol{\beta}_1(t) - \widehat{\mathbf{m}} - \widehat{\mathbf{d}}\, t$ and its lag revealing its AR nature; (Bottom right) The residuals $\widehat{\boldsymbol{\varepsilon}}_1(t)$ scatter plot – after the AR structure is removed from $\boldsymbol{\beta}_1(t)$.

1 Introduction

Principal component analysis (PCA) has been widely used to analyze coupled time-series [2], [4], [3] and recently, the functional nature of certain data sets has received mounting attention [9], [8]. Functional principal component analysis (FPCA) provides a natural and powerful way to model coupled time-series when the data are entire curves but are observable only at discrete, not necessarily uniform, intervals.

The current applications of PCA and FPCA on financial time-series assume that each sample in the data set are drawn independently from a stationary distribution. As such, the principal components (PCs) are invariant under a permutation of the data set – defying the time-indexed nature of financial data. For such data, the validity of the PCs extracted in this manner is highly questionable. In the top panel of Figure 1 we present a time-series resulting from the NYMEX light sweet crude oil futures price data set (to be explained in more detail in Section 2.2). This series is typical of many types of financial data: there is an unmistakable trend together with mean-reversion to this trend.

In this work, we propose a natural marriage between vector auto-regressive (VAR) models and FPCA. First, the underlying temporal dynamics is estimated and removed. Then FPCA is performed on the detrended data, which now satisfies the *iid* requirements of PCA. The rationale for employing a modified FPCA as opposed to a modification of the usual PCA is quite simple: futures price data on a given day are available for contracts maturing at the start of each calendar month for the next two and a half years, and then quarterly for the next year, and finally semi-annually for the last two years. Consequently, the data points on a given curve move to the left as time progresses and are also unequally spaced. Through the use of basis functions, these discrete data points are transformed into a continuous function; FPCA then extracts the PCs from the now functional data. Data sets of the sort above are quite common in functional financial time-series. Furthermore, standard PCA does not guarantee smooth PCs, cannot be used to extrapolate beyond or interpolate between data points, and does

not incorporate domain specific knowledge which in FPCA enters through the choice of the basis functions.

This paper is organized as follows: Section 2 describes our modeling assumptions, our VAR-FPCA methodology and the specific commodities data set used in the analysis; Section 3 discusses the results of the data analysis and the resulting financial intuition; and Section 4 concludes and discusses ongoing and future work.

2 The VAR-FPCA Methodology

2.1 Model

Please consult Ramsay & Silverman [10] and Hamilton [6] for an excellent introduction to FPCA and VAR.

As usual in FPCA analysis, the data is first transformed into functional form. This involves expressing the original data as a linear combination of a set of basis functions $\{\phi_k(\tau) : k = 1, \ldots, K\}$. The curve-fitting error is minimized in the least-square sense, on a curve-by-curve basis, providing a time-series of coefficients β_{tk} for each basis function. The data is now represented in functional form:

$$F_t(\tau) = \sum_{k=1}^{K} \beta_{tk}\phi_k(\tau) \ ,$$

or equivalently

$$\mathbf{F}(\tau) = \mathbf{B}\boldsymbol{\phi}(\tau) \ .$$

Here, $F_t(\tau)$ denotes the $t + \tau$-maturity futures price at time t, τ is the time to maturity, and β_{tk} is the coefficient of the k-th basis function ϕ_k at time t. $\mathbf{F}$ and $\mathbf{B} = (\boldsymbol{\beta}_{t=1} \ldots \boldsymbol{\beta}_{t=N})^T$ are NxK matrices, where N is number of observed curves and $\boldsymbol{\beta}_t = (\beta_{t1} \ldots \beta_{tK})^T$. Finally, $\boldsymbol{\phi}$ is a column vector of the K basis functions. As mentioned earlier, the $\boldsymbol{\beta}_t$ coefficients are obtained by least squares minimization on a curve-by-curve basis

$$\boldsymbol{\beta}_t = \arg\min_{\boldsymbol{\beta}_t} \sum_{i=1}^{n_t} \|\boldsymbol{\beta}_t\boldsymbol{\phi}(\tau_{ti}) - F_t(\tau^*_{ti})\|^2$$

for every t, where $\{F_t(\tau^*_{ti}) : i = 1, \ldots, n_t\}$ is the observed futures price curve at time t.

As evident in Figure 1, there is a strong underlying temporal structure embedded in the $\boldsymbol{\beta}$ coefficients. We use the following first order VAR process with a linear trend to model this dynamics:

$$\boldsymbol{\beta}_t = \mathbf{m} + \mathbf{d}\, t + \mathbf{A}\boldsymbol{\beta}_{t-1} + \boldsymbol{\varepsilon}_t \ , \tag{1}$$

where, $\mathbf{m}$ is a constant mean vector process, $\mathbf{d}$ is the linear trend vector process, $\mathbf{A}$ is the KxK cross-factor interaction matrix, and $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Omega})$ is the *iid* K-dimensional zero mean normal innovation. Let the estimators of the VAR parameters be denoted by $\{\widehat{\mathbf{m}},\ \widehat{\mathbf{d}},\ \widehat{\mathbf{A}},\ \widehat{\boldsymbol{\Omega}}\}$ (see Hamilton [6] for standard VAR estimation procedures). Based on these estimates, we then extract the residuals

$$\widehat{\boldsymbol{\varepsilon}}_t = \boldsymbol{\beta}_t - \left(\widehat{\mathbf{m}} + \widehat{\mathbf{d}}\, t + \widehat{\mathbf{A}}\boldsymbol{\beta}_{t-1}\right) \ . \tag{2}$$

The scatter plot of the residuals $\widehat{\boldsymbol{\varepsilon}}_1$ versus their one-period lag are shown in the bottom right panel in Figure 1. Comparing with the left panel, it is evident that the autoregression has been removed. The residuals therefore contain the fluctuation of the basis loadings free of the temporal structure – precisely what is demanded in PC analysis whether it be standard PCA or FPCA.

Let $\mathbf{E} = (\widehat{\boldsymbol{\varepsilon}}_{t=1} \ldots \widehat{\boldsymbol{\varepsilon}}_{t=N})^T$ denote the matrix form of all K residuals. FPCA, much like multi-variate PCA, seeks to diagonalize the now continuous variance-covariance function (instead of a finite dimensional matrix)

$$v(\tau_1, \tau_2) := \frac{1}{N}\, \boldsymbol{\phi}(\tau_1)^T\mathbf{E}^T\mathbf{E}\boldsymbol{\phi}(\tau_2) \ ,$$

via the eigen-problem

$$\langle v, \xi\rangle(\tau) = \rho\, \xi(\tau) \ , \tag{3}$$

subject to the orthonormality constraints $\langle \xi_k, \xi_l\rangle = \delta_{kl}$, where the inner product is defined as follows

$$\langle f, g\rangle(\tau) = \int_{\tau_{min}}^{\tau_{max}} f(\tau, s)\, g(s)\, ds \ .$$

The vector of eigen-functions $\boldsymbol{\xi}(\tau)$ are the functional PCs representing the modes of largest variation in the functional data. The eigen-problem is solved by expanding a given ξ_k in terms of the basis functions $\boldsymbol{\phi}$:

$$\xi_k(\tau) = \mathbf{z}_k^T\boldsymbol{\phi}(\tau) \ , \tag{4}$$

transforming the eigen-equation (3) into one for $\mathbf{z}_k$:

$$\frac{1}{N}\mathbf{W}^{1/2}\mathbf{E}^T\mathbf{E}\mathbf{W}^{1/2}\mathbf{u}_k = \rho_k\mathbf{u}_k \ , \tag{5}$$

where $\mathbf{z}_k = \mathbf{W}^{-1/2}\mathbf{u}_k$ and $(\mathbf{W})_{ij} = \langle\phi_i, \phi_j\rangle$ (see the appendix for details). The domain specific knowledge of the basis functions are now embedded in the $\mathbf{W}$ matrix.

Once the functional PCs are extracted, the original data $\mathbf{F}$ can be projected into the new basis functions via

$$\mathbf{F}(\tau) = \mathbf{B}\phi(\tau) = \mathbf{G}\xi(\tau) \quad \text{or} \quad F_t(\tau) = \sum_{k=1}^{K} \gamma_{tk}\xi_k(\tau) \ ,$$

where the matrix $\mathbf{G} = \mathbf{B}(\mathbf{Z}^{-\mathbf{1}})^T$ and $\mathbf{Z} = (\mathbf{z}_1 \ldots \mathbf{z}_K)$ is the matrix of eigenvectors stacked column wise. In the above, the time-series of the PC's coefficients $\gamma_t = \mathbf{Z}^{-1}\beta_t$ naturally appears. Stacking these loadings columnwise produces $\mathbf{G}$.

Given that the basis function loadings β_t follow the VAR(1) model in equation (1), the PC loadings γ_t therefore follow the VAR(1) process

$$\boldsymbol{\gamma}_t = \mathbf{Z}^{-1}\mathbf{m} + \mathbf{Z}^{-1}\mathbf{d}\ t + \mathbf{Z}^{-1}\mathbf{A}\mathbf{Z}\ \boldsymbol{\gamma}_t + \boldsymbol{\varepsilon}'_t \ , \tag{6}$$

where $\boldsymbol{\varepsilon}'_t \sim N(\mathbf{0}; \boldsymbol{\Omega}' = \mathbf{Z}^{-1}\boldsymbol{\Omega}(\mathbf{Z}^{-1})^T)$ are the *iid* K-dimensional normal innovations. In light of this fact, the innovations will not be independent – $\mathbf{Z}$ is not an orthogonal matrix and therefore cannot diagonalize the positive

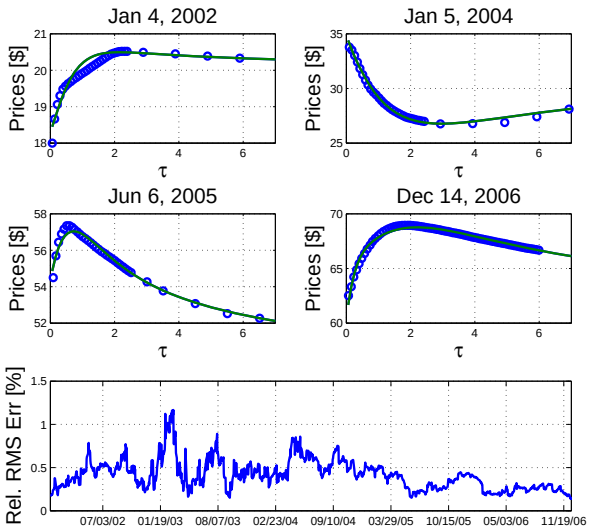

Figure 2. Transforming discrete observed data into functional form. (Top 4) A few typical examples of fitted data curves. (Bottom) The relative RMS error [%] of each data curve, averaging 0.40% with a maximum of 1.2%.

definite covariance matrix $\mathbf{\Omega}$. Curiously, we find the covariance matrix $\mathbf{\Omega}'$ is indeed essentially diagonal for the NYMEX light sweet crude oil futures data. This may not be a coincident; instead, although traditional PCA ignores the functional form of the data, it will diagonalize the covariance matrix of the residual noise terms. Since FPCA is a way of modifying PCA to correctly account for the functional nature of data, $\mathbf{Z}$ should be "close" to a diagonalizing matrix for the residual noise.

It is important to comment that the our procedure is distinct from preprocessing the data, such as the removal of seasonality trends or outliers. Instead it should be viewed as a way of extracting the true stochastic degrees of freedom from the pre-visible information embedded in the data.

2.2 Data

To illustrate the utility of our VAR-FPCA methodology, we used the daily NYMEX light sweet crude oil futures contracts data with maturities between 3 weeks to 7 years, dated from January 2002 to December 2006, totaling 1251 data curves with an average of 18 points on each curve. In general, VAR-FPCA is a powerful tool to analyze any functional time-series.

We selected a set of exponential basis functions to capture the salient features of this specific data set:

$$\phi_k(\tau) = \begin{cases} 1, & k = 1 \\ (1 - \exp\{-a_k \tau\})/a_k, & k = 2..K \end{cases} \quad (7)$$

where $a_k = \{4, 2, 1, 0.2\}$, representing the decay times of a quarter, half, one, and five years. These varying decay rates are included to provide sufficient coverage for both short and long term horizons. The exponential form of these basis functions is motivated by the implied forward prices produced by modeling the spot price as a mean-reverting multi-factor Gaussian Ornstein-Uhlenbeck process. Such models assume the spot price has a tendency to mean-revert to a given stochastic level which itself may mean-revert to another stochastic level and so on. See for example Schwartz [3] & [5] and Jaimungal & Hikspoors [7]. Earlier work on yield-curve modeling and fitting (see Nelson & Siegel [2] and Diebold & Li [4]) also used a similar (but reduced) set of basis functions. For other specific data sets, domain-specific knowledge can be incorporated in the analysis through the selection of the basis functions.

Figure 2 provides a few typical quality of fit examples using our proposed basis functions. The bottom chart of Figure 2 provides the relative RMS error for each data curve. The average relative RMS errors over all the fitted curves is about 0.4% with a maximum of 1.2%. Fitting errors of this magnitude are quite small for practical purposes. We can therefore be confident in the functional representation of our data using this basis. Interestingly, the relative RMS for the period 2005 onwards is notably smaller ($\sim 0.22\%$) than the period up to this point ($\sim 0.50\%$).

The futures curves exhibit a variety of typical shapes and are quite similar to interest rate yield curves. The curves tend to have humps, although over some periods the curves may appear flatter. These humps may be upward humps or downward humps and typically peak in the one to two year maturity region. Nonetheless, the fixed set of basis functions are able to capture all of the observed features quite well.

3 Results

Figure 3 shows the fitted β_t coefficients. Most of the linear trend is loaded on the first basis function, reflecting the general upward trend of oil prices during the period of this particular data set across all maturities. Furthermore, there may have been a regime change in 2004 – in this article we will ignore the existence of this regime change and instead leave the analysis of a hidden Markov model driving regimes for future work. The mean level, trend and mixture matrix were estimated using standard VAR methods. Interestingly, the estimated mixture matrix $\widehat{\mathbf{A}}$ has heavy weights on its diagonal and lighter weights on the off-diagonal, implying there is little cross-factor interactions. This feature is illustrated in Figure 4.

We then remove the temporal structure from the $\boldsymbol{\beta}_t$ time-series and determine the principal components $\boldsymbol{\xi}$ from $\mathbf{E}$ using the FPCA procedure outlined in Section 2.1. The loadings of the three most dominant principal components are shown in Figure 5. As expected, the PC loadings $\boldsymbol{\gamma}_t$ still exhibit mean-reversion. Pleasantly, the variability of the PC loadings are significantly smaller than the basis loadings themselves. Furthermore, the covariance matrix $\mathbf{\Omega}'$ of the PC loadings coefficients $\mathbf{G}$ are essentially di-

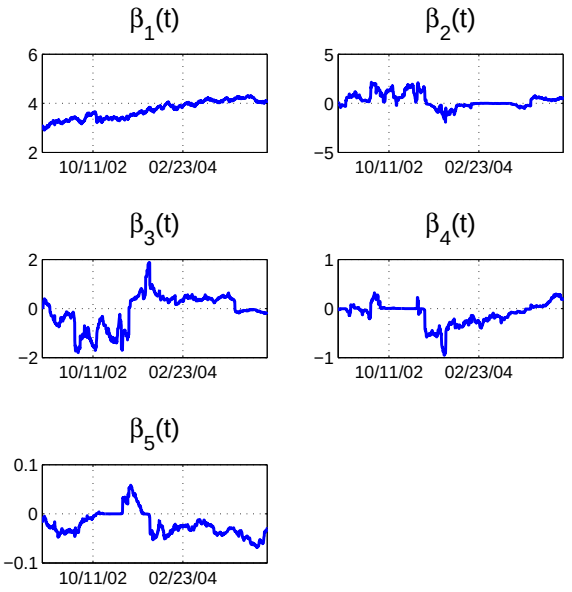

Figure 3. The basis loading coefficients $\boldsymbol{\beta}_t$ of the fitted functional data.

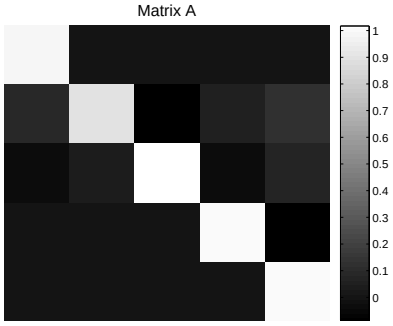

Figure 4. An image of the estimated mixture matrix $\widehat{\mathbf{A}}$. The relatively strong diagonal indicates there is little cross-factor dependency in $\boldsymbol{\beta}_t$.

agonal, with all correlations insignificantly different from zero. This is a very desirable feature. The loadings of the basis $\boldsymbol{\beta}_t$ do exhibit significant correlation, and although the VAR-FPCA algorithm does not guarantee the independence of the loadings of the PCs, this is precisely what we find. This, together with the relatively low volatility of the PC loadings, tantalizing suggests that data prediction will be more accurate when based on the PCs rather than the basis functions themselves. In addition, the low volatilities exhibited in the PC loadings can potentially lead to statistical arbitrage strategies that involve less transactions, leading to substantial savings. To complete the description, we found the volatility of the first three principal components to be 53.0% 13.8% and 4.8% respectively (annualized based on 251 trading days per year).

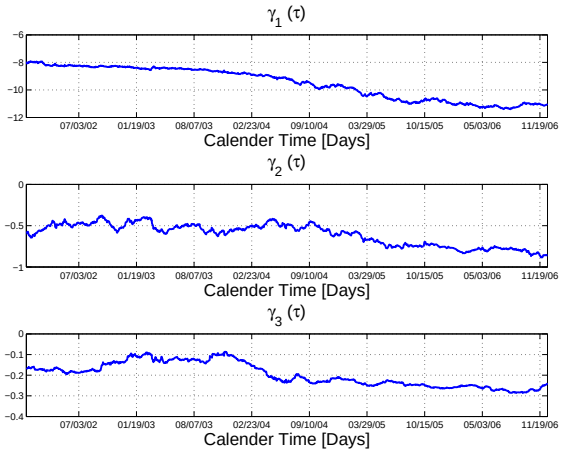

Figure 5. The loading coefficients $\boldsymbol{\gamma}_t$ of the top three principal components.

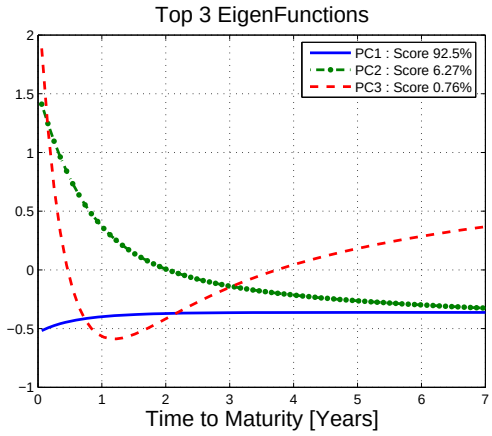

Figure 6. Superposition of the top 3 principal components.

Figure 6 depicts the three most dominant principal components. When compared with regular PCA, our VAR-FPCA algorithm produces extremely smooth and easily interpretable principal components. Similar to the PCs extracted from interest rate yield curves [2], [4], the first principal component, accounting for 92.5% of the variation, governs the overall price level of the futures price uniformly across all maturity dates with a slight biasing of short term contracts. The second PC, accounting for 6.27% of the variation, is primarily responsible for tilting the curve as it biases short term contracts relative to long term contracts. Finally, the third PC, accounting for 0.76% of the variation, accounts for introducing convexity (or bending) into the futures curves as it pushes the short and long term upwards while pulling the medium term downwards. Surprisingly, all of these principal components are more extreme or pronounced than their analogs from interest rate data. The flat PC is very flat, the titling PC is strongly tilt-

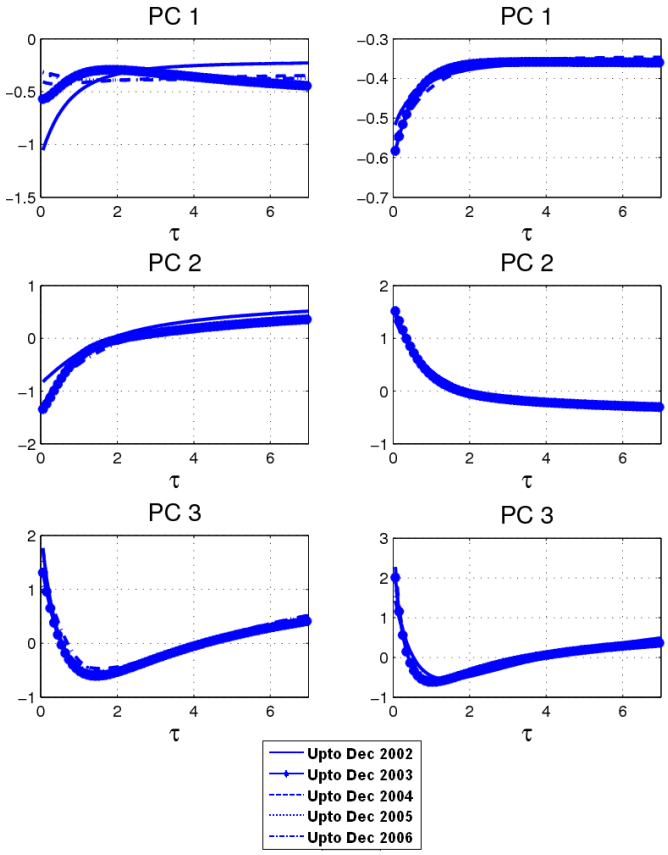

Figure 7. Time consistency of the top 3 principal components estimated using *regular* FPCA (left) versus VAR-FPCA (right) The PCs extracted by the VAR-FPCA are extremely stable over the five year life of the dataset.

ing, and the bending PC produces significant bends.

When the PCs were extracted using standard FPCA methods (i.e. ignoring the embedded temporal structure), we found the first PC accounted for 99% of the variability of the data. This contrasts with the 92.5% we found using the VAR-FPCA method. Such an over weighting of the importance of a single PC can have significantly negative consequences on both trading strategies and risk management decisions.

The time consistency of our VAR-FPCA methodology was also examined by computing the PCs using only the first year of data, then adding one more year of data and re-estimating the PCs, and so on until all the data was included. In this manner, five estimates for each of the principal components were produced, one for each cumulative time-frame. Figure 7 provides the results of this analysis using both our VAR-FPCA method and the regular FPCA method ignoring the time dependency in the underlying model. There is a marked difference in PC1 (which accounts for more than 90% of the variation in the data in both cases) and some significant difference in PC2 using the standard FPCA method as more data is added; contrastingly, our VAR-FPCA method produces PCs which are indistinguishable across all years.

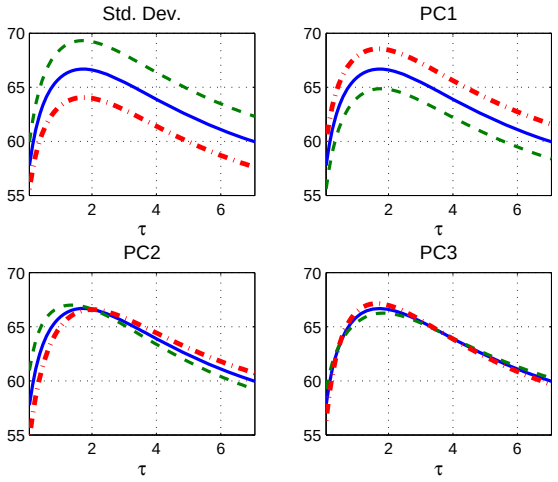

Figure 8. Perturbations of the sample price curve by principal components. (Top-left) The sample price curve is perturbed by +/- the weekly volatility of all the price curves; Similarly, the sample price curve is perturbed by +/- weekly volatility of γ_1 (Top-right), γ_2 (Bottom-left), and γ_3 (Bottom-right). The effect of the PCs on the sample curve is clearly visible – PC 1 shocks the price curve evenly up and down; PC 2 perturbs by tilting the price curve about the 2 year point; while PC 3 perturbs the bending the curve.

It is extraordinary that such time stability exists in the PCs when not only the crude oil prices had undergone highly turbulent changes between the years 2002 to 2006, the term structures had also undergone some substantial changes as evident in the shapes of the futures curves, depicted in Figure 2.

This temporal and market consistency exhibited in the PCs plays an important role in risk management. The extracted PCs allow a risk manager to quickly identify the source of exposures and map it directly into the dominating PCs [1]. Consistency in the PCs translates to more consistent risk management strategies, reducing strategy turnover and therefore the associated costs.

Figure 8 provides an alternative illustration of the effects of each principal component. Here, the solid curve represents a sample futures price curve from Nov 20, 2006. The top-left panel shows the variability intrinsic to the data set by showing the boundaries of one standard deviation of all the price curves (normalized to 1 week) away from the sample curve. In subsequent panels, the weekly volatilities (standard deviations) of the top 3 principal components are added and subtracted from the sample curve to show the perturbation from the sample curve caused by each principal component. Note that the principal components perturb the sample price curve corresponding to their shape (i.e. level shift, tilting, and bending). The first PC perturbs the sample curve more significantly than the other PCs merely because it contains 92.5% of the variance of the data.

4 Conclusion and Future Work

By first removing the temporal structure embedded in the functional time-series, we demonstrated that the principal components can be extracted in a self-consistent manner. Additionally, synergizing the temporal structure learned through VAR together with the time-consistent and smooth principal components extracted with FPCA, allows effective short-term predictions to be made. One obvious application is in statistical arbitrage or algorithmic trading strategies development. Moreover, the time and market stability of the principal components allows for efficient long-term risk management. The customary FPCA would fail to incorporate the current state of the data curves and therefore fail to correctly capture short term risks and / or statistical arbitrage strategies.

Using the NYMEX light sweet crude oil futures price data, the PCs are found to have very intuitive financial interpretations. Specifically, the strongest principal component corresponds to the broadbase price movements across all contract maturities. This corresponds to the empirical fact that the price fluctuations are mostly due to parallel shifts. The second strongest PC corresponds to tilting and the third strongest corresponds to bending of the futures price curves. All PCs are smooth and have very pronounced and specific effects.

The technique proposed in this work is highly versatile. It can be readily applied to other types of data sets such as commodity futures spread, interest rate yield curves, credit-spreads, credit default swap (CDS) rates and volatility surfaces (albeit a two-dimensional version), etc.

We are currently working on using Hidden Markov Models as an alternative to the VAR process for modeling the basis loadings. This has the potential of greater flexibility and predictive power, while still maintaining the intuitive nature of the FPCA framework.

5 Acknowledgements

The authors thank the Natural Sciences and Engineering Research Council of Canada for partially funding this work.

6 Appendix: Functional Principal Component Analysis (FPCA)

This proof of the equivalence of (3) and (5) a concise version of the analysis contained in Ramsay & Silverman [10]. Insert (4) into (3) to obtain

$$\frac{1}{N}\boldsymbol{\phi}^T(\tau)\mathbf{E}^T\mathbf{E}\mathbf{W}\mathbf{z} = \rho\,\boldsymbol{\phi}^T(\tau)\mathbf{z}\ .$$

Take the inner product of the above expression with $\boldsymbol{\phi}$

$$\begin{aligned} \frac{1}{N}\langle\boldsymbol{\phi},\boldsymbol{\phi}^T\rangle\mathbf{E}^T\mathbf{E}\mathbf{W}\mathbf{z} &= \rho\,\langle\boldsymbol{\phi},\boldsymbol{\phi}^T\rangle\mathbf{z} \\ \Rightarrow \qquad \frac{1}{N}\mathbf{W}\mathbf{E}^T\mathbf{E}\mathbf{W}\mathbf{z} &= \rho\,\mathbf{W}\mathbf{z}\ . \end{aligned}$$

Finally, let $\mathbf{u} = \mathbf{W}^{1/2}\mathbf{z}$ and (5) follows. Notice that the eigen-problem is symmetric for $\mathbf{u}$ but it is not symmetric for $\mathbf{z}$. Instead the basis functions distort the metric leading to a non-orthogonal mapping of the basis functions into the PCs.

References

[1] D. Soronow C. Blanco and P. Stefiszyn. Multi-factor models of the forward price curve. In *Commidities Now*, pages 80–83.

[2] A. F. Sigel C. R. Nelson. Parimonious modeling of yield curves. *The Journal of Business*, 60(4).

[3] G. Cortazar and E. Schwartz. The valuation of commodity contingent claims. *The Journal of Derivatives*, 1:27–29, 1994.

[4] F. X. Diebold and C. Li. Forecasting the term structure of goverment bond yields. *The Journal of Econometrics*, 130:337–364, 2006.

[5] R. Gibson and E. Schwartz. Stochastic convenience yield and pricing of oil contingent claims. *Stochastic Process. Appl.*, 11:215–260, 1981.

[6] J. D. Hamilton. *Time Series Analysis*. Princeton University Press, Princeton, NJ, USA, 1994.

[7] S. Hikspoors and S. Jaimungal. Energy spot price models and spread options pricing. *International Journal of Theoretical and Applied Finance*, 2007. to appear.

[8] S. Ingrassia and G. D. Costanzo. Functional principal component analysis of financial time series. In *New Developments in Classification and Data Analysis*, pages 351–358, 2005.

[9] E. Matzner-Lober and C. Villa. Functional principal component analysis of the yield curve. *Working Paper*, 2004.

[10] J.O. Ramsay and B.W. Silverman. *Functional Data Analysis*. Springer Series in Statistics. Springer-Verlag New York, Inc., New York, NY, USA, 2005.