Functional framework for high frequency financial data with focus on regression and predictability of intraday price curves

Piotr Kokoszka, Colorado State University

Collaborators:

Matt Reimherr, Penn State University

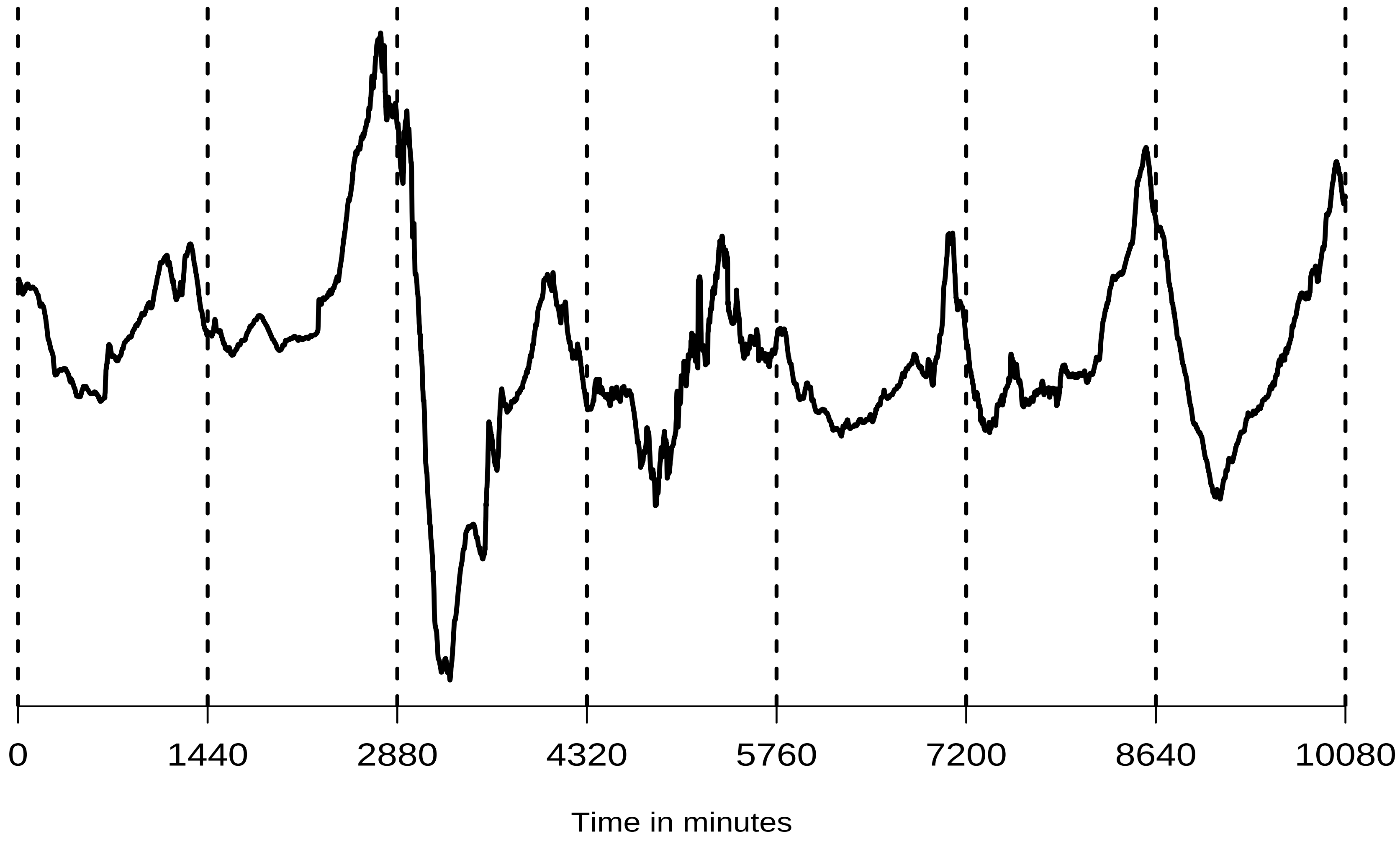
Xi Zhang, PNC Bank

Hong Miao, Colorado State University

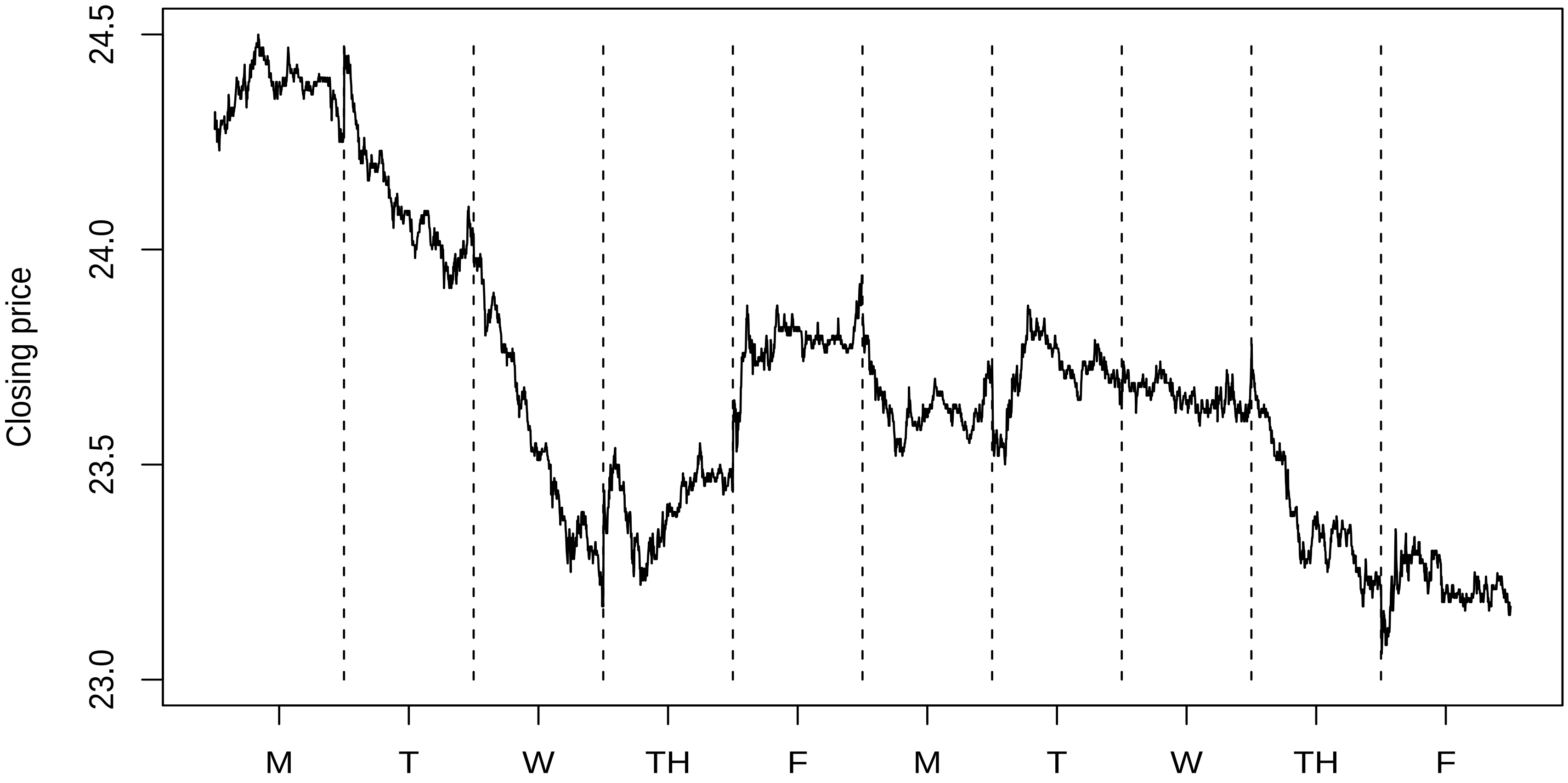
Outline

- Functional time series
- Some applications to financial data
- Cumulative intraday returns
- Functional factor model
- Tests for predictability

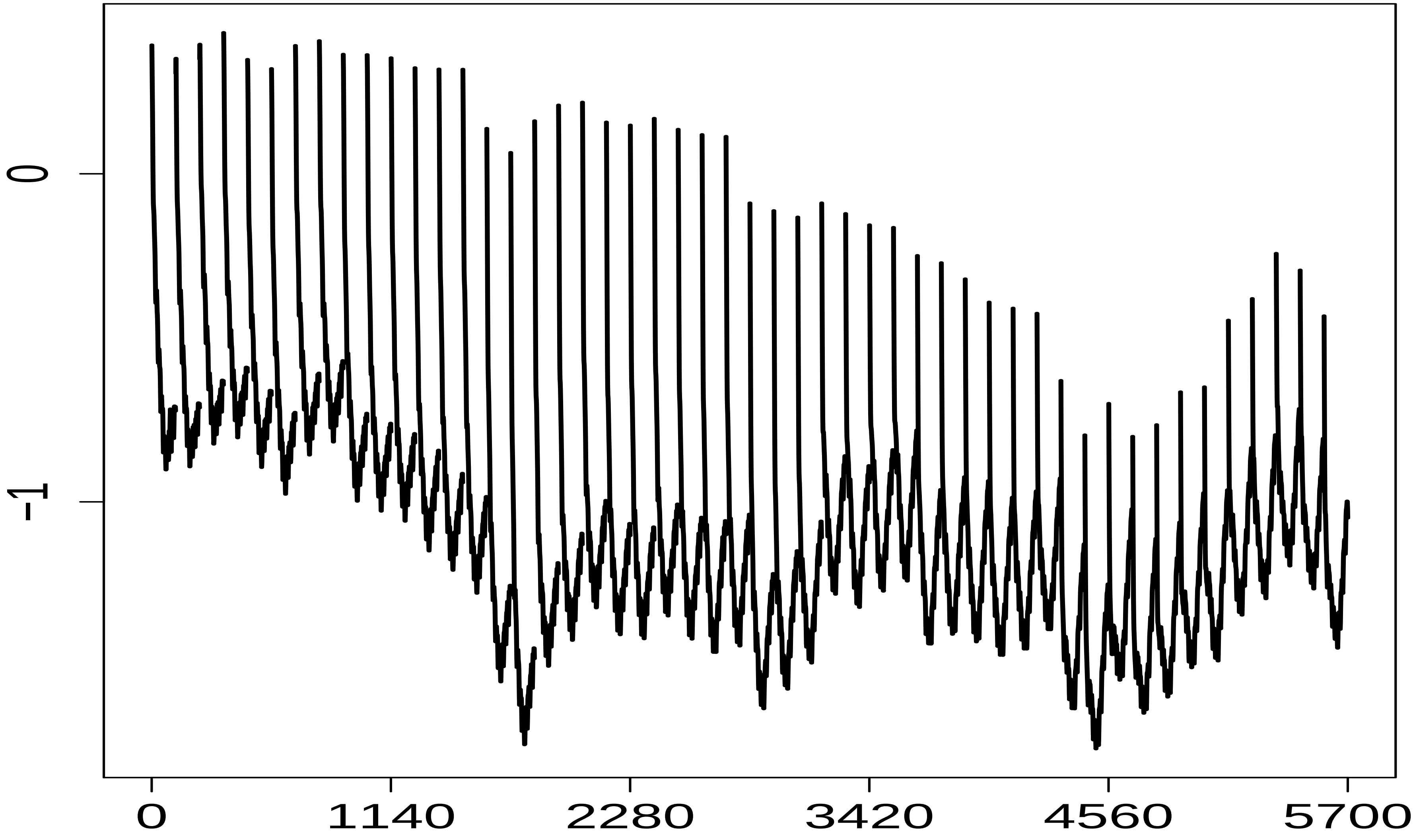
Examples of functional time series



Microsoft stock prices in one-minute resolution



Centered Eurodollar futures curves over a 50 day period.



Recent research

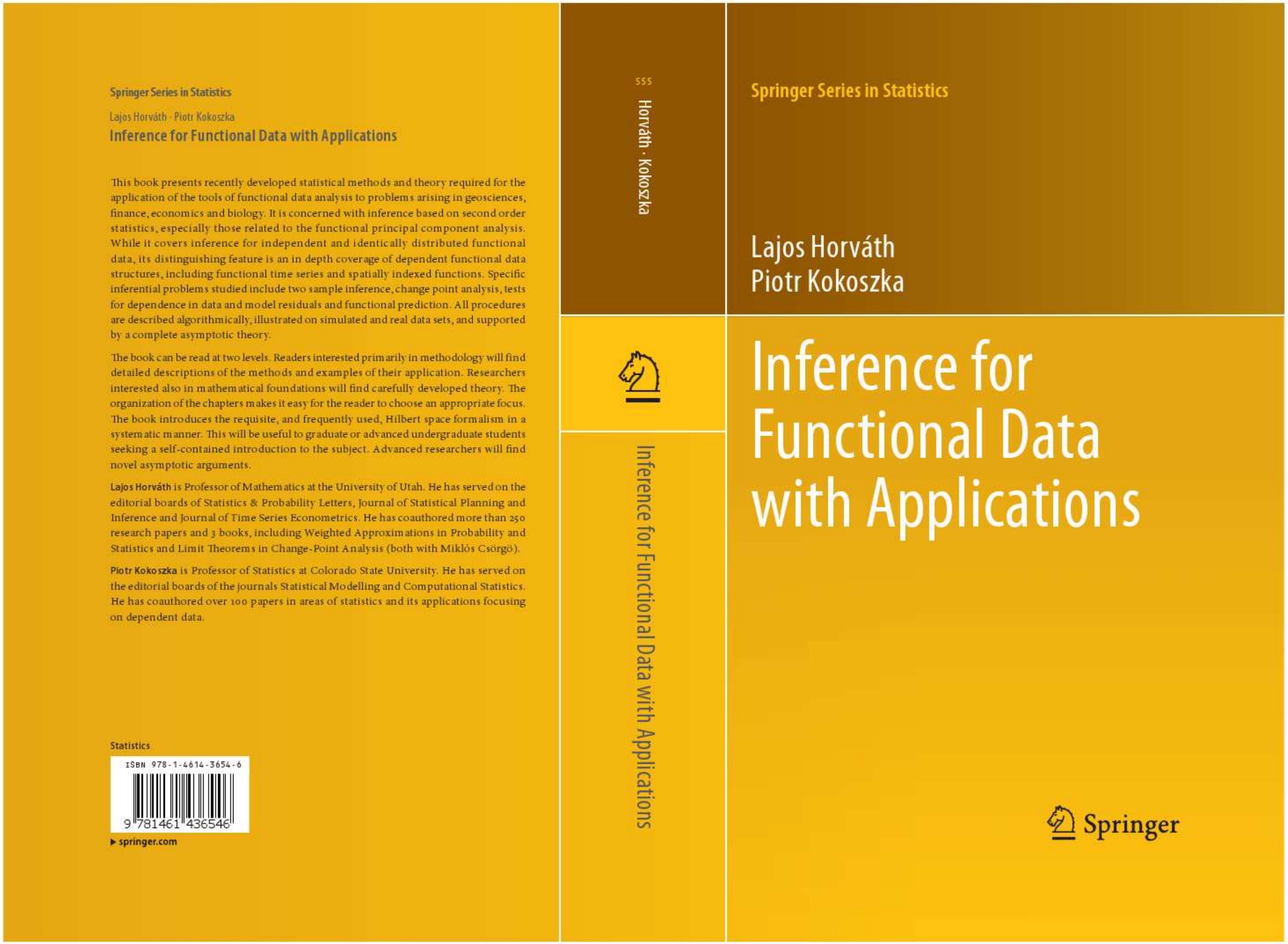
Functional autoregressive (FAR, ARH) process

$$Z_i = \Phi(Z_{i-1}) + \varepsilon_i,$$

$$Z_i(t) = \int \phi(t,s) Z_{i-1}(s) ds + \varepsilon_i(t), \quad 0 \leq t \leq 1.$$

$$Z_i = \sum_{j=1}^{p} \Phi_j(Z_{i-j}) + \varepsilon_i.$$

- Order selection in the FAR(p) model.
- Optimal prediction using the FAR model (Eurodollar futures).
- Change point analysis in the FAR model
- General weakly dependent functional time series, including ARCH type models.
- Two sample problems
- Limit theory for the mean function, estimation of the LRV.
- Spectral analysis of functional time series



Springer Series in Statistics

Lajos Horváth · Piotr Kokoszka

Inference for Functional Data with Applications

This book presents recently developed statistical methods and theory required for the application of the tools of functional data analysis to problems arising in geosciences, finance, economics and biology. It is concerned with inference based on second order statistics, especially those related to the functional principal component analysis. While it covers inference for independent and identically distributed functional data, its distinguishing feature is an in depth coverage of dependent functional data structures, including functional time series and spatially indexed functions. Specific inferential problems studied include two sample inference, change point analysis, tests for dependence in data and model residuals and functional prediction. All procedures are described algorithmically, illustrated on simulated and real data sets, and supported by a complete asymptotic theory.

The book can be read at two levels. Readers interested primarily in methodology will find detailed descriptions of the methods and examples of their application. Researchers interested also in mathematical foundations will find carefully developed theory. The organization of the chapters makes it easy for the reader to choose an appropriate focus. The book introduces the requisite, and frequently used, Hilbert space formalism in a systematic manner. This will be useful to graduate or advanced undergraduate students seeking a self-contained introduction to the subject. Advanced researchers will find novel asymptotic arguments.

Lajos Horváth is Professor of Mathematics at the University of Utah. He has served on the editorial boards of Statistics & Probability Letters, Journal of Statistical Planning and Inference and Journal of Time Series Econometrics. He has coauthored more than 250 research papers and 3 books, including Weighted Approximations in Probability and Statistics and Limit Theorems in Change-Point Analysis (both with Miklós Csörgö).

Piotr Kokoszka is Professor of Statistics at Colorado State University. He has served on the editorial boards of the journals Statistical Modelling and Computational Statistics. He has coauthored over 100 papers in areas of statistics and its applications focusing on dependent data.

Statistics

ISBN 978-1-4614-3654-6

9 781461 436546

▸ springer.com

SSS

Horváth · Kokoszka

Inference for Functional Data with Applications

Springer Series in Statistics

Lajos Horváth
Piotr Kokoszka

Inference for Functional Data with Applications

Springer

Cumulative intraday returns

$P_n(t)$ - price of an asset on day n at time t within that day (trading hours).

Normalize the trading day to be the interval $[0, 1]$.

Cumulative intraday return (CIDR) on day n:

(0.1) $$r_n(t) = 100(p_n(t) - p_n(0)), \quad p_n(t) = \log P_n(t).$$

Compare:
daily log return is $p_n(1) - p_{n-1}(1)$.

Comments:

CIDR’s are not directly comparable to daily returns. They do not include the overnight price change $p_n(0) - p_{n-1}(1)$.

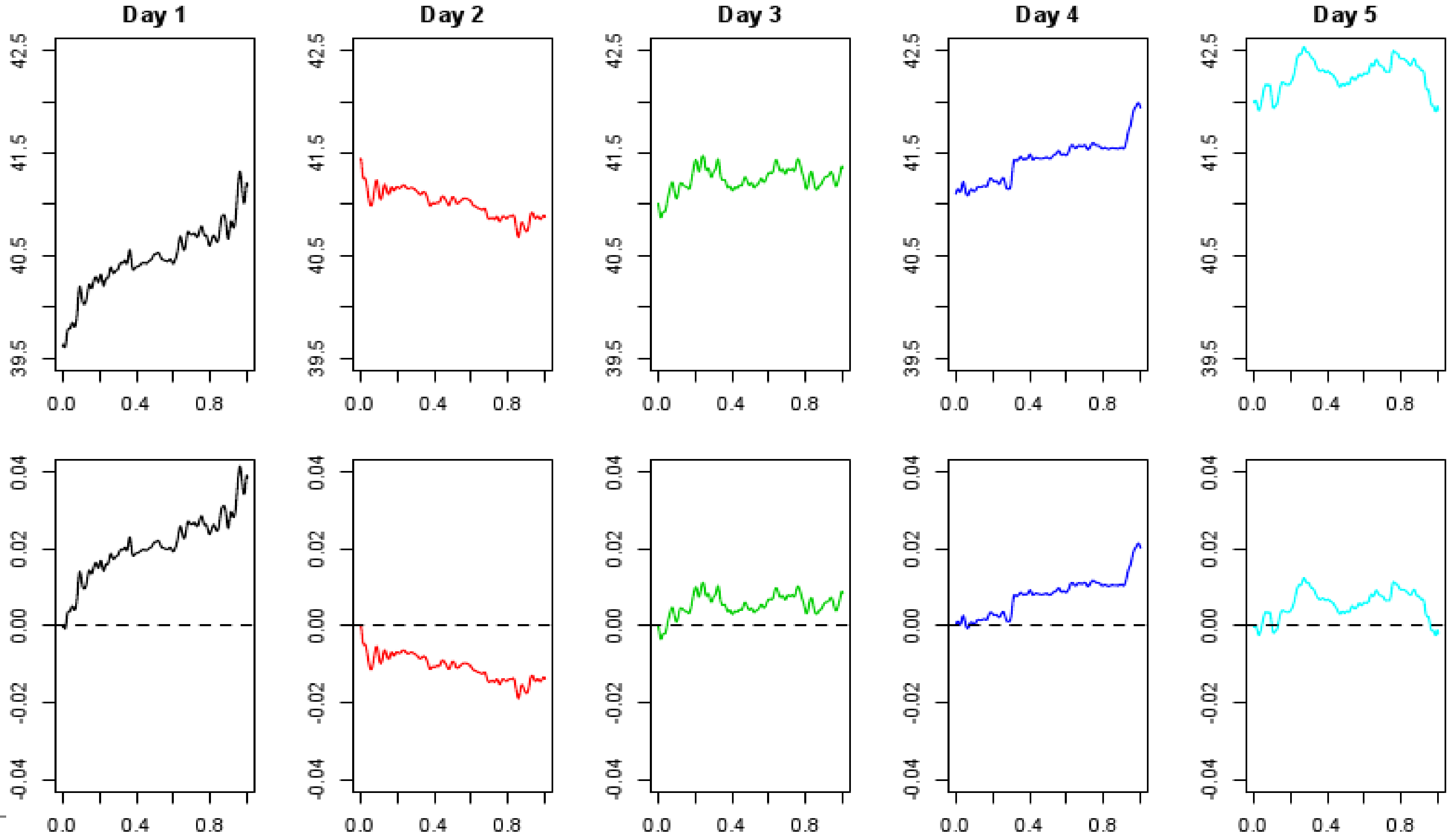
Since

$$r_n(t) \approx 100(P_n(t) - P_n(0))/P_n(0)$$

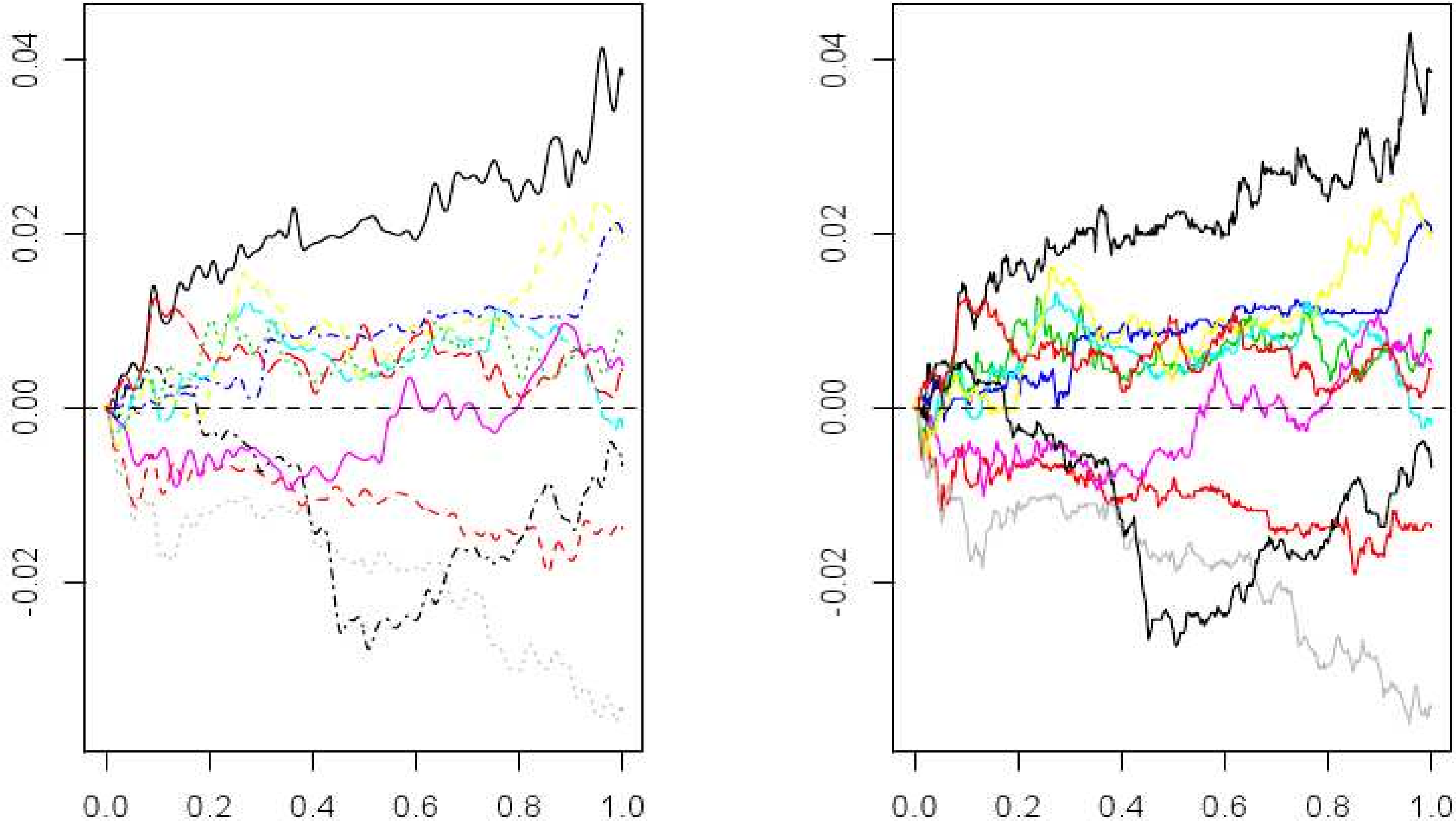
($P_n(0)$ constant for a given day n),
the curves $r_n(t)$ and $P_n(t)$ have similar shapes but different scales and origins.

Top panel: price curves on five consecutive days for XOM.

Bottom panel: cumulative intraday returns on the same days.



Ten CIDR’s on Exon Mobil (XOM). Left smoothed, Right raw.



Factor models for CIDR’s

R_n CIDR curve on day n (a single asset)
M_n CIDR curve for a market index

$$R_n(t) = \beta_0(t) + \beta_1 M_n(t) + \varepsilon_n(t).$$

S_n, H_n potential scalar risk factors

$$R_n(t) = \beta_0(t) + \beta_1 M_n(t) + \beta_2 S_n + \beta_3 H_n + \varepsilon_n(t).$$

C_n CIDR curves for oil futures

$$R_n(t) = \beta_0(t) + \beta_1 M_n(t) + \beta_2 C_n(t) + \varepsilon_n(t).$$

Combinations of scalar and curve factors are allowed.

A general model

$$R_n(t) = \beta_0(t) + \sum_{j=1}^{p} \beta_j F_{nj}(t) + \varepsilon_n(t).$$

Question: are any of the scalar coefficients β_i significant?

Statistical factor model of Hays, Shen and Huang (2012):

$$X_n(t) = \sum_{k=1}^{K} \gamma_{nk} F_k(t) + \varepsilon_n(t).$$

The factors F_k do not depend on n and are *orthonormal* functions to be estimated.

Estimation

Functional $\beta_0(\cdot)$, scalar β_i
hybrid estimators: method of moments / least squares

Consistency and asymptotic normality established under a weak dependence condition.

SE's of the $\hat{\beta}_i$ can be estimated using asymptotic variances.

Conclusions: The coefficients of
the market CIDR's are significant
scalar factors are not significant (shape!)
oil futures CIDR's are sometimes significant
(oil companies)

A weak dependence condition

Bernoulli shifts

$$F_{nj} = f_j(\delta_n, \delta_{n-1}, \ldots), \qquad \varepsilon_n = e(\eta_n, \eta_{n-1}, \ldots).$$

A sequence $\{X_n\}$ is called L^p–m–approximable if each X_n is a Bernoulli shift $X_n = f(u_n, u_{n-1}, \ldots)$ and

$$\sum_{m=1}^{\infty} \nu_p(X_n - X_n^{(m)}) < \infty,$$

where $\nu_p(X) = (E\|X\|^p)^{1/p}$;

$$X_n^{(m)} = f(u_n, u_{n-1}, \ldots, u_{n-m+1}, u'_{n-m}, u'_{n-m-1}, \ldots).$$

Predictability

Related to Efficient Market Hypothesis

Many forms:
Introduction in Campbell, Lo, MacKinlay (1997)

Celebrated history:
Bachelier, Working, Cowles, Granger, Fama, Samuelson and Mandelbrot.

Still subject of research:
If frictions and nonstationarities are eliminated,
the direction of returns cannot be predicted.

Question: Can CIDR curves be predicted form the past curves of the same asset?

Functional Principal Components

Normalize the trading day to be the interval $[0,1]$.
Treat the IDCR's as random elements of $L^2([0,1])$.
Assume that the L^2–valued sequence $\{r_n\}$ is
strictly stationary (at least under H_0).

r – a random function with the same distribution as each r_n.

(0.2)
$$r(t) = \mu(t) + \sum_{k=1}^{\infty} \xi_k v_k(t),$$

$\mu(t) = Er(t)$ – mean function,
$v_k(\cdot), k \geq 1,$ – functional principal components
(optimal orthonormal factors),
$\xi_k = E\int (r(t) - \mu(t))v_k(t)dt$ – scores.

$\mu(\cdot)$ and $v_k(\cdot), k \geq 1,$ are population parameters.

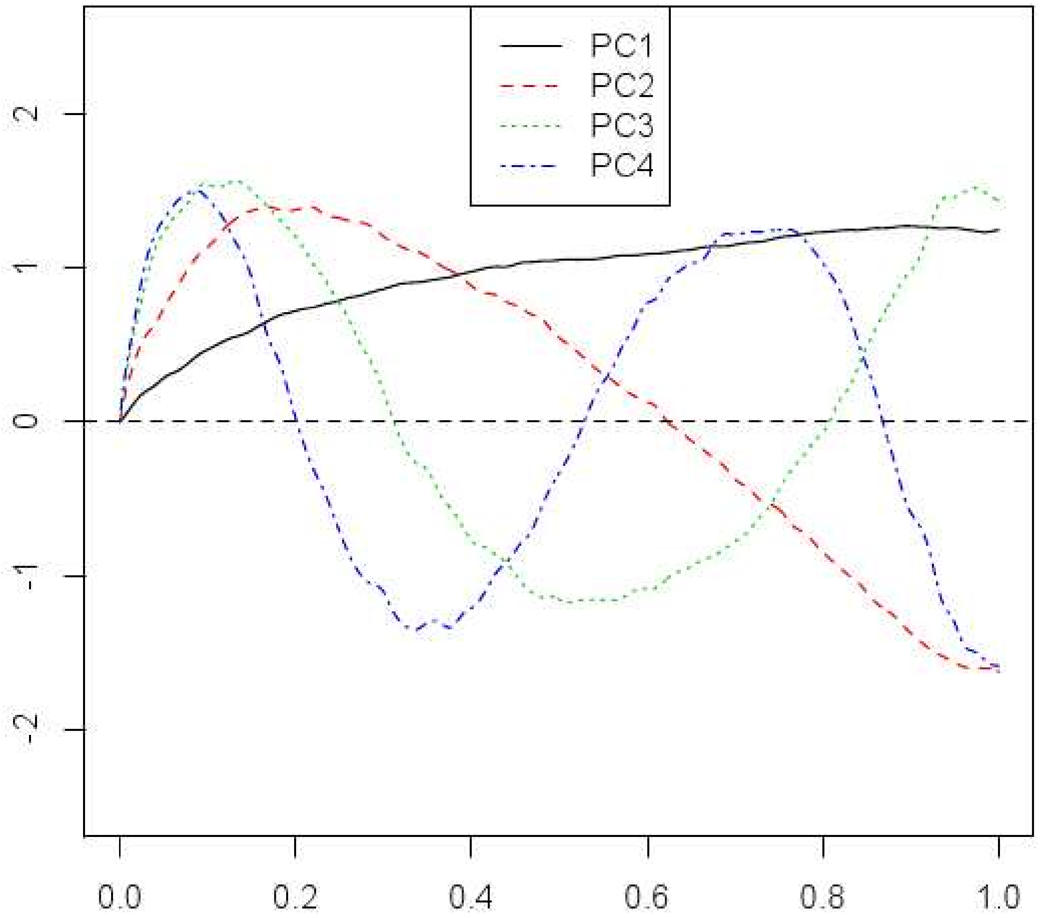
They are estimated by $\hat{\mu}(\cdot)$ and $\hat{v}_k(\cdot), 1 \leq k \leq p.$

All procedures of FDA that use FPC depend on p.

There are several ways of selecting optimal p,
but we feel most comfortable when conclusions do not depend strongly on p in a reasonable range.

The estimated mean function $\hat{\mu}(\cdot)$ is very close to zero, for most blue chip stocks statistically not significantly different from zero.

First four EFPC’s of XOM.



Idea of testing predictability of shapes

The shape of the curve r_n observed on day n is quantified by the vector of scores

$$[\hat{\xi}_{1n}, \hat{\xi}_{2n}, \dots, \hat{\xi}_{pn}]^T, \qquad \hat{\xi}_{kn} = \int \{r_n(t) - \hat{\mu}(t)\} \hat{v}_k(t) dt$$

These vectors describe the temporal evolution of the shapes.

Example of interpretation: The sequence of the scores $\hat{\xi}_{1n}$ shows "how much" component $\hat{v}_1$ is present on day n.

If the sign of $\hat{\xi}_{1n}$ is positive, the r_n is mostly increasing.

If $P(\xi_{1n} > 0 | \xi_{1,n-1} > 0) > 1/2$, there is some predictability of shapes: increasing curves tend to follow increasing curves.

Current approaches and difficulties

The portmanteau test of Gabrys and Kokoszka (2007).

$H_0:$ The r_n are iid.

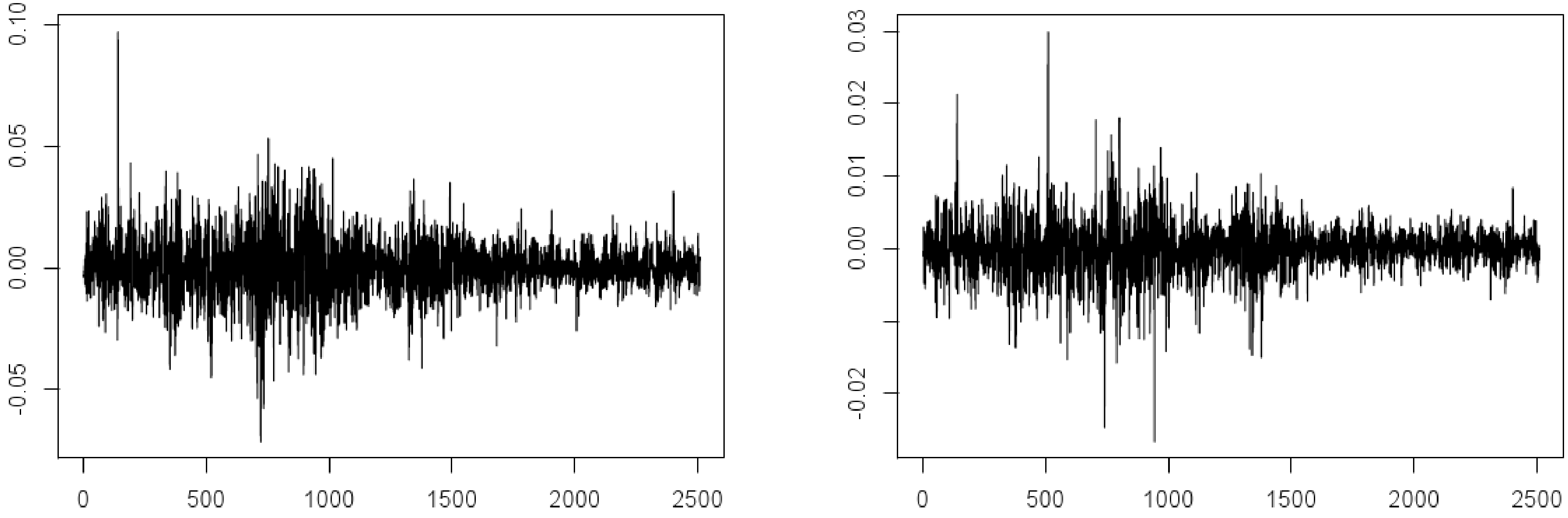
When applied to IDCR's, the results depend very strongly on p.

Next slide:

P–values of the portmanteau test as a function p.

p	1	2	3	4
BOA	0.167	0.000	0.000	0.000
CITI	0.082	0.000	0.000	0.000
COCA	0.021	0.013	0.008	0.000
CVX	0.488	0.093	0.025	0.022
DIS	0.099	0.206	0.140	0.003
IBM	0.511	0.000	0.000	0.000
MCD	0.164	0.110	0.368	0.055
MSFT	0.227	0.145	0.005	0.000
WMT	0.032	0.008	0.000	0.000
XOM	0.054	0.091	0.367	0.000

Time series of scores for WMT: left ξ_{1n}, right ξ_{2n}, $1 \leq n \leq 2500$.



Time series of scores for IDCR's are heteroskedastic and heavy–tailed.

The portmanteau test is very sensitive to both.
(It is a functional version of the Ljung–Box–Pierce test.)

We want to construct the test of predictability of signs of the scores ξ_{kn}.

It will not be sensitive to heteroskedasticity and heavy–tails.

It will have a clear interpretation.

X_n – centered r_n.

Null Hypothesis:
The sequence $\{X_n\}$ is conditionally symmetric:

$$\mathcal{L}(X_n|X_{n-1}, X_{n-2}, \ldots) = \mathcal{L}(-X_n|X_{n-1}, X_{n-2}, \ldots) \quad a.s.$$

Define the triangular array

$$I_{N,n}^{(k)} = \text{sign}\{\langle X_n, \hat{v}_k \rangle\}.$$

Unlike the sign of a scalar return, sign$\{\langle X_n, v_k \rangle\}$ is not observable.
Notice that $I_{N,n}^{(k)} I_{N,n+1}^{(k)}$ is positive for a sequence and negative for a reversal.

Alternative Hypotheses:

H_A : null hypothesis does not hold.

$$H^-_{A,j} : E[I_1^{(j)} I_2^{(j)}] < 0 \quad \text{or} \quad H^+_{A,j} : E[I_1^{(j)} I_2^{(j)}] > 0.$$

1. under $H^+_{A,j}$, $P(\xi_{j,n} > 0 | \xi_{j,n-1} > 0) > 1/2$ and $P(\xi_{j,n} < 0 | \xi_{j,n-1} < 0) > 1/2$,
2. under $H^-_{A,j}$, $P(\xi_{j,n} > 0 | \xi_{j,n-1} > 0) < 1/2$ and $P(\xi_{j,n} < 0 | \xi_{j,n-1} < 0) < 1/2$.

The tests

To test against $H_{A,j}^{+}$ and $H_{A,j}^{-}$ we use statistics $\Lambda^{(j)}$.

To test against the general H_A we use a statistics Λ_p based on the first p EFPC's $\hat{v}_j$.
(Λ_p weighs the $\Lambda^{(j)}$ according to their importance.)

Second part of the talk:

Definitions of these statistics and their asymptotic theory.

Simulation study assessing their finite sample performance.

Empirical evidence

P–values for the test based on the statistic Λ_p for $p = 1, \ldots, 4$.

We denote with * the P–values under 10% favoring sequences.

p	1	2	3	4
BOA	0.255	0.243	0.232	0.240
CITI	0.194	0.213	0.229	0.223
COCA	*0.053	*0.062	*0.063	*0.060
CVX	0.106	0.093	0.086	0.085
DIS	0.181	0.226	0.204	0.193
IBM	0.590	0.656	0.653	0.642
MCD	0.239	0.218	0.212	0.209
MSFT	0.952	0.841	0.887	0.896
WMT	0.984	0.927	0.916	0.897
XOM	0.016	0.015	0.015	0.015

P–values of the tests based on statistics $\Lambda^{(j)}$ for $j = 1, \ldots, 4$.

We denote with * the P–values under 10% favoring sequences.

j	1	2	3	4
BOA	0.255	0.646	0.309	0.194
CITI	0.194	0.618	0.223	0.290
COCA	*0.053	0.413	0.857	0.168
CVX	0.106	0.419	0.196	0.628
DIS	0.181	0.098	*0.040	*0.033
IBM	0.590	0.290	0.857	0.369
MCD	0.239	0.391	0.536	0.536
MSFT	0.952	0.069	*0.040	0.413
WMT	0.984	0.194	0.646	0.115
XOM	0.016	0.592	0.834	0.625

Summary

- Unlike the Portmaneteau test, the new statistic Λ_p yields consistent conclusions accross various p.
- The auxiliary statistics $\Lambda^{(j)}$ allow us to identify the components with predicatability and its direction.
- For most blue chip stocks (2000-2007) there is no evidence of predictability.
- If overall predictability exists, it is in the the first FPC (increasing/decreasing shape).
- These conclusions remain valid for the period of the financial cricis 2008/03/18 to 2009/03/31
  - MCD becomes predictable (sequences). P-value drops from 0.2 to 0.02.
  - XOM becomes unpredictable. P-value increses to 0.6 from 0.15.

Asymptotic Theory

Difficulty: The usual bound

$$E\int\left(\hat{v}_k(t)-v_k(t)\right)^2 dt = O\left(N^{-1}\right)$$

cannot be used when working with signs (not continuous).

We used a martingale limit theorem due to D. L. McLeish.
(The Annals of Probability, 1974, **2**, 620–628.)

THEOREM 0.1 $\{Y_{n,N}\}$ – *array of martingale differences such that*

(a) $\max_{1\le n\le N}|Y_{n,N}|$ *is uniformly bounded in* $L_2(P)$ *norm,*

(b) $\max_{1\le n\le N}|Y_{n,N}|\to 0$*, in probability,*

(c) $\sum_{1\le n\le N} Y_{n,N}^2 \to 1$*, in probability.*

Then, $\sum_{1\le n\le N} Y_{n,N} \xrightarrow{d} N(0,1)$.

Set

$$\mathbf{I}_{N,n}^T = (I_{N,n}^{(1)}, \ldots, I_{N,n}^{(p)}), \quad \mathcal{F}_n = \sigma\{\mathbf{I}_{N,1}, \ldots, \mathbf{I}_{N,n}\}$$

and consider the pointwise (Hadamard) products

$$\mathbf{I}_{N,n} \circ \mathbf{I}_{N,n+1}.$$

Verify that the assumptions of the theorem of McLeish hold.

Conclude that

$$\Lambda^{(k)} := (N-1)^{-1/2} \sum_{i=1}^{N-1} I_{N,n}^{(k)} I_{N,n+1}^{(k)} \xrightarrow{d} N(0,1);$$

$$(\Lambda^{(1)}, \ldots, \Lambda^{(p)})^T \xrightarrow{d} N(\mathbf{0}, \mathbf{\Sigma}).$$

A simulation study

Many general functional ARCH type processes satisfy the assumptions.
(Their specific formulations have not been studied or applied.)

We work with the model

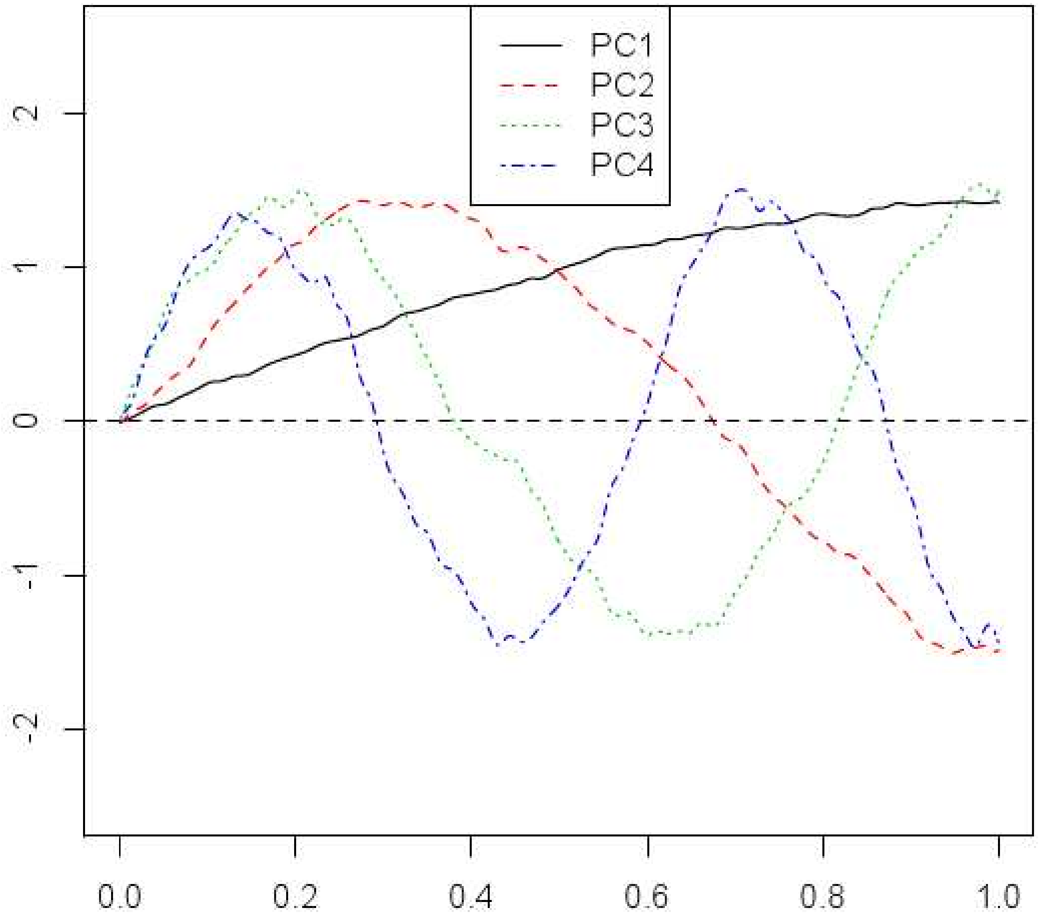
$$X_i(t) = \sigma_i W_i(t),$$

where W_i is a standard Brownian motion and σ_i is a univariate process defined recursively as

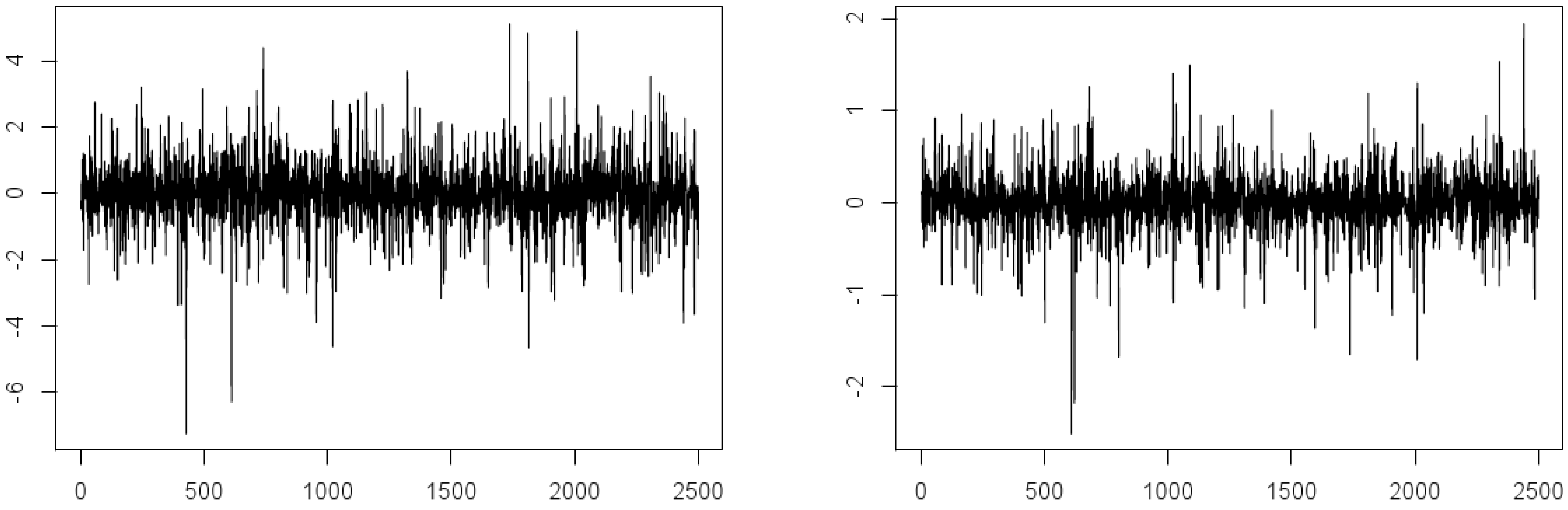
$$\log(\sigma_i) = a \log(\sigma_{i-1}) + 0.5\delta_i,$$

and the δ_i are iid standard normal random variables.

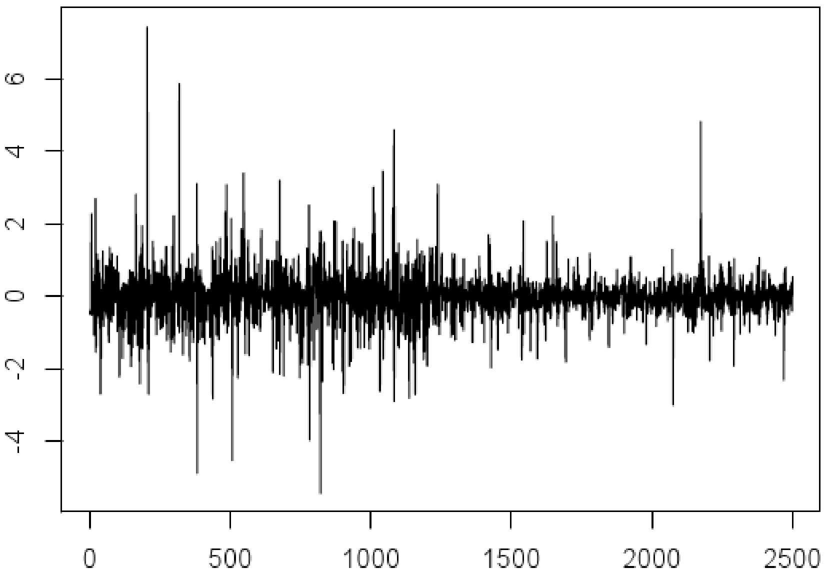
First four EFPC's of a simulated process; $N = 1000$, $a = 0.5$.

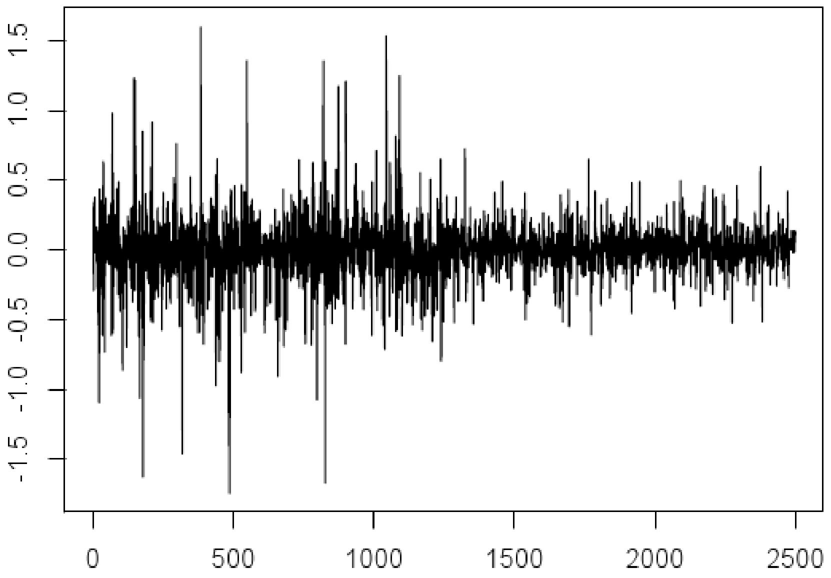


Plot of the first two PC scores for the simulated process with $a = 0.5$.



Plot of the first two PC scores for the simulated process with $a = 0.5$, and a 50% drop in variance after the 1250^{th} observation.





As the alternative, we consider the model

$$X_i(t) = \Phi(X_{i-1})(t) + \sigma_i W_i(t),$$

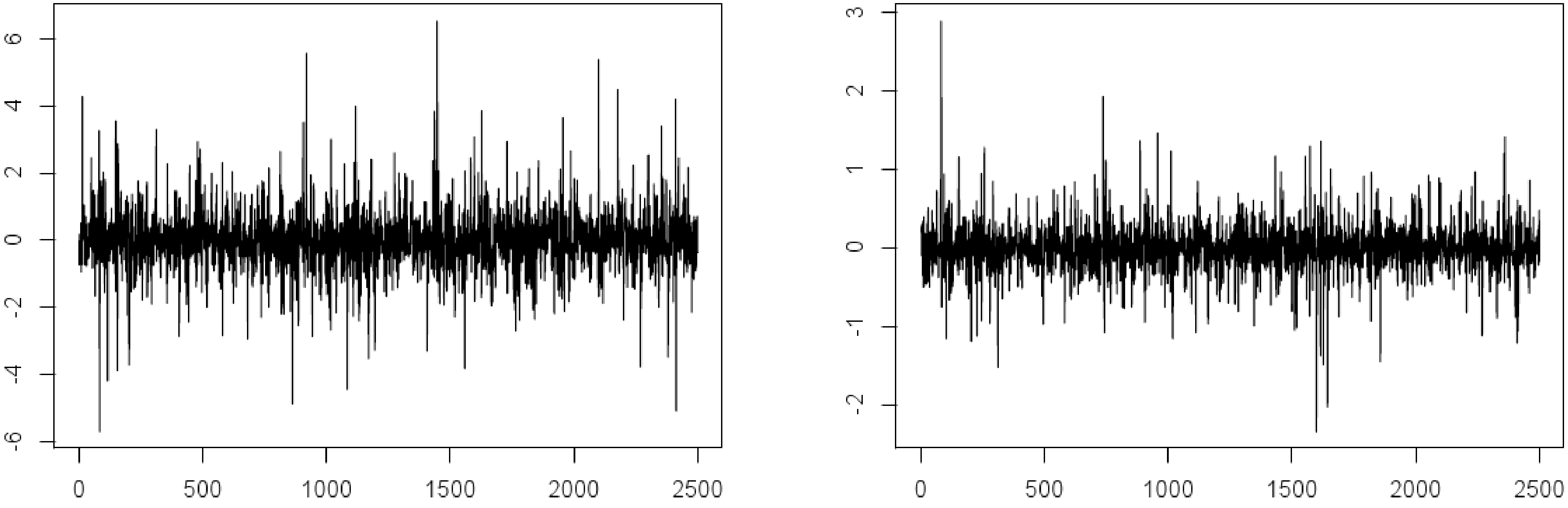
where W_i and σ_i are defined as before.
The operator Φ is an integral operator with the kernel

$$\phi(t, s) = c\frac{\exp(-(t-s)^2)}{0.8739}.$$

This is the FAR model with functional ARCH type errors.

The Hilbert–Schmidt norm of Φ is equal to c.

Plot of the first (left) and second (right) PC scores for the simulated process with $a = 0.5$ and $c = 0.15$ (**An alternative**).



Conclusions from numerical experiments:

- The new tests are robust to heavy tails and heteroskedasticity.
- In all scenarios they have almost perfect empirical size.
- For $N = 2500$ (size of real data) the new test has power of about 98% for $c = 0.15$.
- The portmanteau test can severely overrejects in some scenarios. (For $p = 4, N = 1000, a = 0.5$, empirical size is almost 60% for nominal size of 5%.)