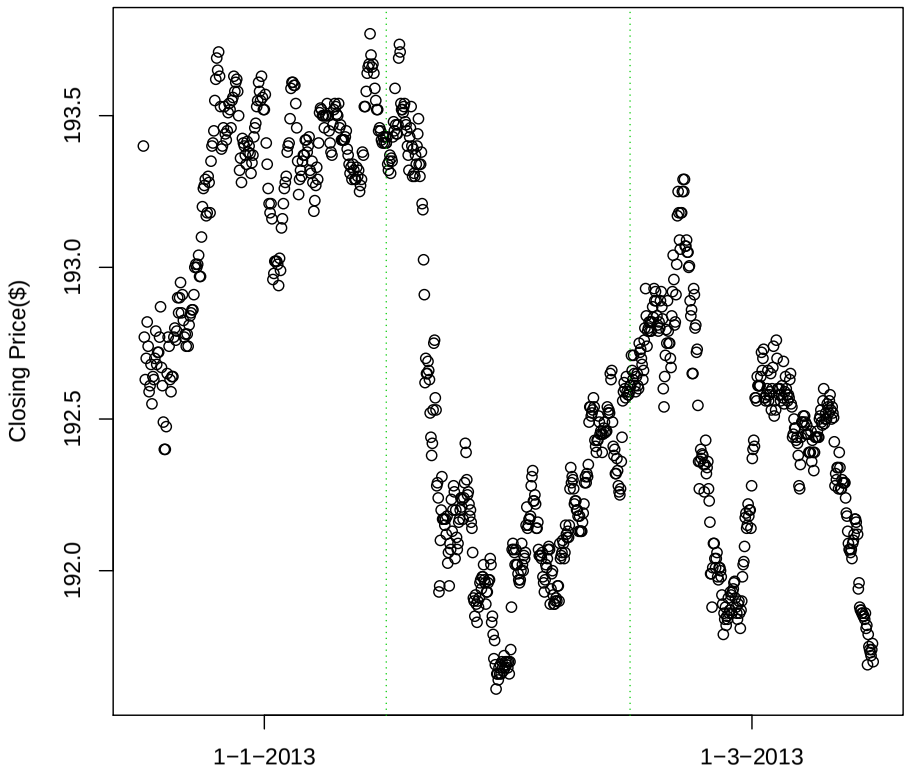
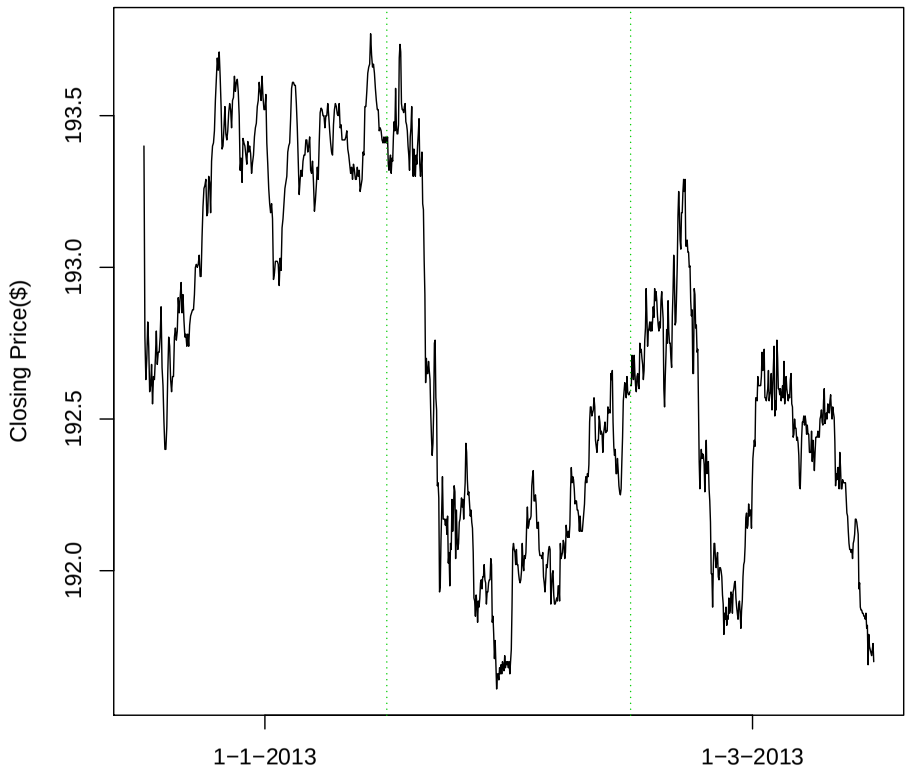
University Carlos III de Madrid June 2018

Functional data analysis

Lajos Horváth

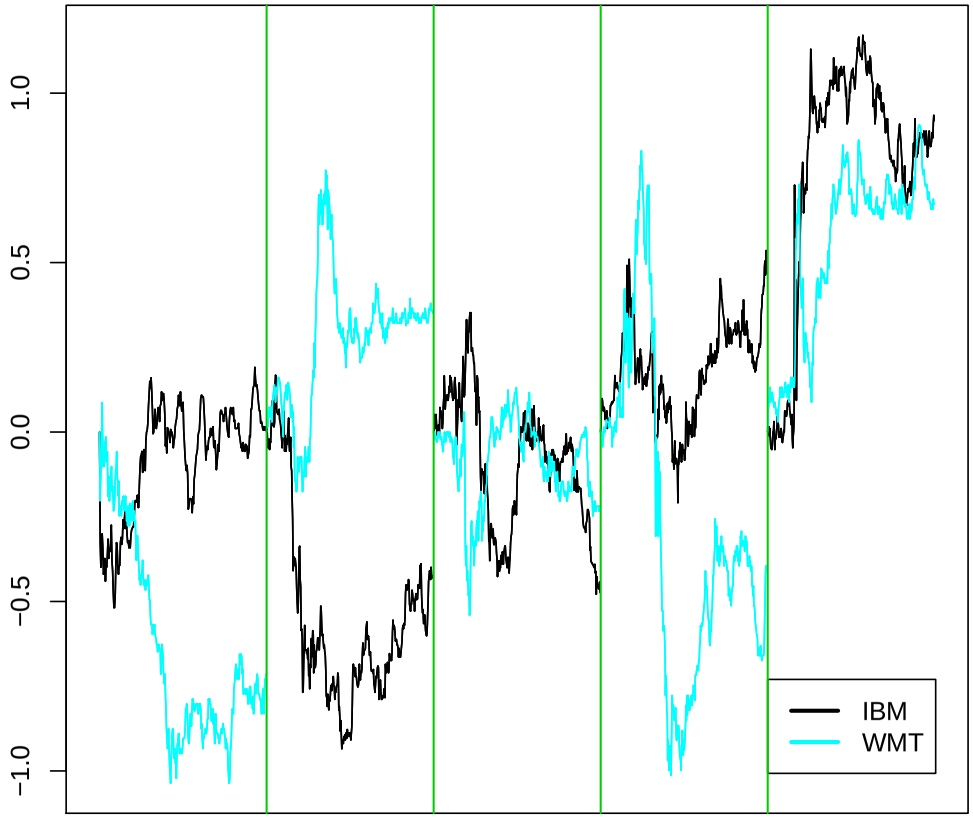
University of Utah, Salt Lake City





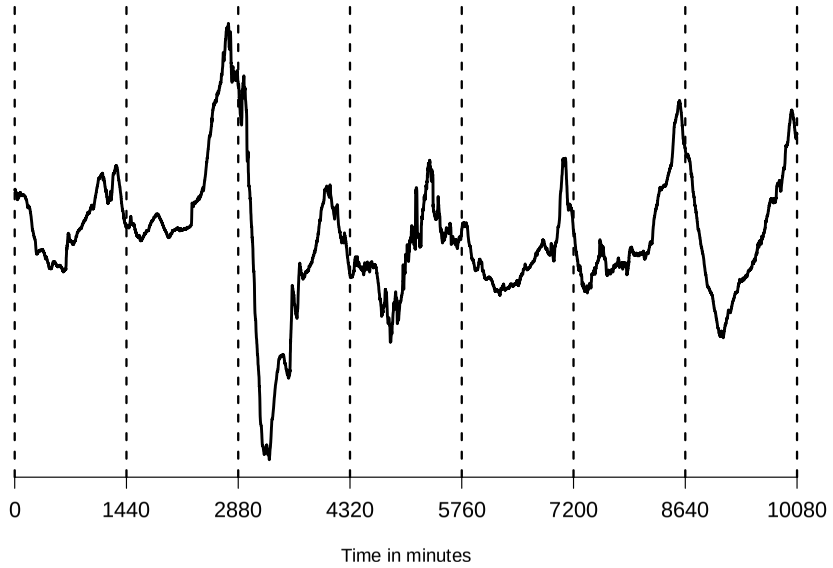
IBM stock price curves

IBM/Walmart stock price curves



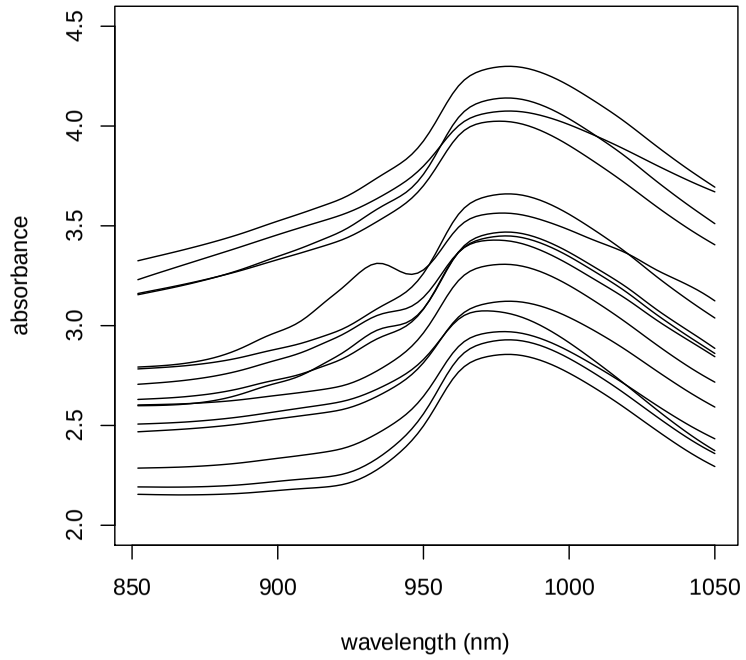
$X_n(t)$ magnetometer readings on day n at time t

Daglis, I. A., Kozyra, J. U., Kamide, Y., Vassiliadis, D., Sharma, A. S., Liemohn, M.W., Gonzalez, W. D., Tsurutani, B. T. and Lu, G. (2003). Intense space storms: Critical issues and open disputes. *Journal of Geophysical Research*, 108.



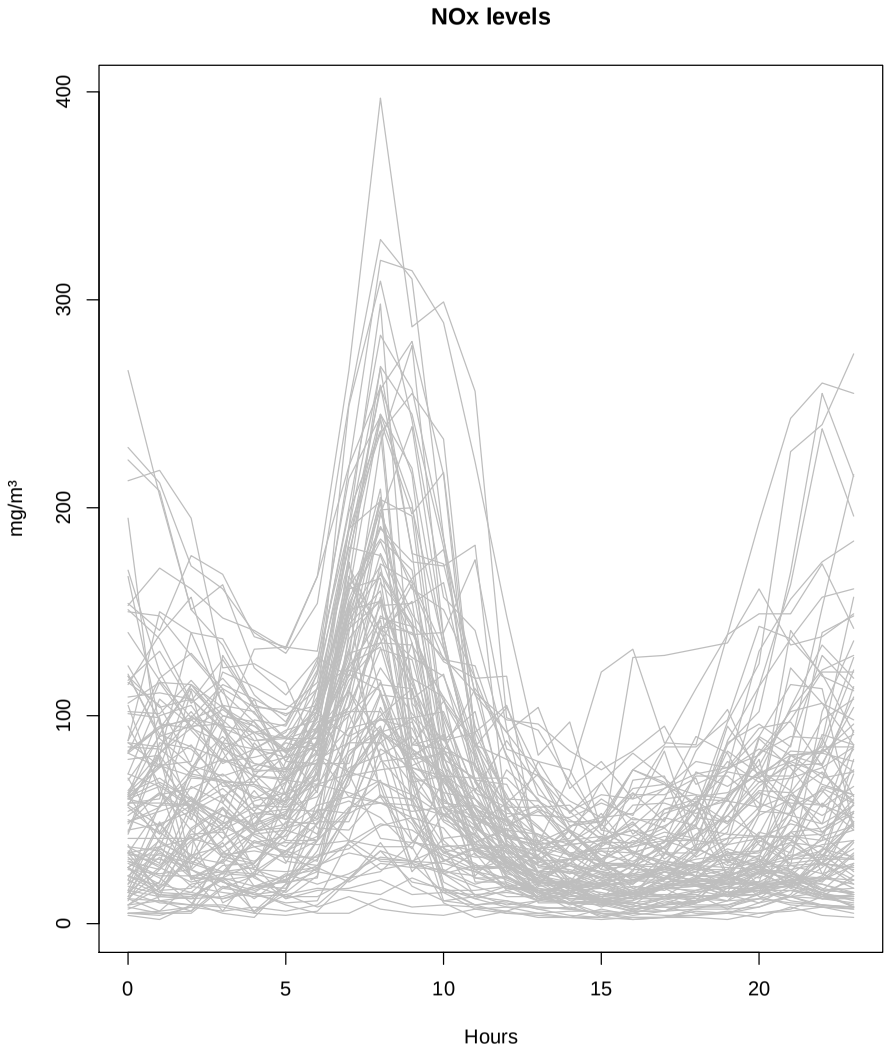
$X_n(t)$ is reading from the Tecator Infratec food and feed analyzer

Yao, F. and Müller, H-G. (2010). Functional quadratic regression. *Biometrika* 97, 49–64.



$X_n(t)$ is the pollution level on day n at time t

Fernández de Castro, B., Guillas, S. and Gonzáles Manteiga, W. (2005). Functional samples and bootstrap for predicting sulfur dioxide levels. *Technometrics*, 47, 212–222.



The space of square integrable functions

L^2 is the space of all functions$\{f\}$ such that $\int f(t)dt < \infty$ ($\int$ mean $\int_0^1$)

inner product $\langle f, g\rangle = \int f(t)g(t)dt$

the norm is induced by the inner product $\|f\| = \sqrt{\langle f, f\rangle}$

We assume that $E\int X^2(t)dt < \infty$ and therefore $P\{\omega : X(t;\omega) \in L^2\} = 1$.

$\{\varphi_i, i \geq 1\}$ is an orthonormal basis of L^2

Karhunen–Loéve expansion: $$X(t) = \sum_{i=1}^{\infty}\langle X, \varphi_i\rangle\varphi_i(t)$$

We can approximate $X(t)$ with the final dimensional process $\sum_{i=1}^{d}\langle X, \varphi_i\rangle\varphi_i(t)$ and

$$E\left\|X - \sum_{i=1}^{d}\langle X, \varphi_i\rangle\varphi_i\right\|^2 \to 0 \quad \text{as} \quad d \to \infty$$

Dimension reduction–replace X (infinite dimensional) with $\langle X, \varphi_1\rangle, \langle X, \varphi_2\rangle, \ldots, \langle X, \varphi_d\rangle$.

How to choose the bases? The best mean squared error:

$\inf_{f\in L^2} E\|X - \xi_1(f)f\|^2 = E\|X - \xi_1(f_1)f_1\|^2$,
$\inf_{f\in L^2, \langle f, f_1\rangle = 0} E\|X - (\xi_1(f_1)f_1 + \xi(f)f)\|^2 = E\|X - (\xi_1(f_1)f_1 + \xi_2(f_2)f_2\|^2$

Covariance operator

Solution to the minimization problem

Covariance kernel: $EX(t) = 0$ and $E \int X^2(t)dt < \infty$

$$C(t, s) = EX(t)X(s)$$

$C(t, s)$ is a symmetric, positive definite function–general Hilbert space theory (spectral theorem) gives

there are $\lambda_1 \geq \lambda_2 \geq \ldots \geq 0$ and orthonormal functions $\{\varphi_i, i \geq 1\}$ such that

$$\lambda_i \varphi_i(t) = \int C(t, s)\varphi_i(s)ds \quad 1 \leq i < \infty$$

$$C(t, s) = \sum_{i=1}^{\infty} \lambda_i \varphi_i(t)\varphi)i(s)$$

(eigenfunctions and eigenvalues)

Technical comment:

If $\lambda_d > 0$ and $\lambda_{d+1} = 0$ then C has only d eigenfunctions. These can be extended into an orthogonal bases and $\int C(t, s)\varphi_i(s)ds = 0$ for all $i > d$

Partial sums in Hilbert spaces

Partial sums process: $S_N(x,t) = N^{-1/2} \sum_{i=1}^{\lfloor Nx \rfloor} X_i(t).$

Theorem 2: If $X_1, X_2, \ldots, X_N$ are iid with $EX_1(t) = 0$ and $E\int X_1^2(t)dt < \infty$, then we can define Gaussian processes $\Gamma_N(x,t)$ such that such that

$$\sup_{0\le x\le 1} \int_0^1 \left(S_N(x,t) - \Gamma_N(x,t)\right)^2 dt \quad \xrightarrow{P} \quad 0,$$

and $E\Gamma_N(x,t) = 0$ and $E\Gamma_N(x,t)\Gamma_N(y,s) = \min(x,y)C(t,s)$.

Note

$$\Gamma_N(x,t) \stackrel{\mathcal{D}}{=} \sum_{i=1}^{\infty} \lambda_i^{1/2} \mathcal{W}_i(x)\varphi_i(t),$$

where $\mathcal{W}_i(x), 1 \le i < \infty$ are independent Wiener processes (standard Brownian motion), i.e. Gaussian processes with $E\mathcal{W}_i(x) = 0$ and $E\mathcal{W}_i(x)\mathcal{W}_i(y) = \min(x,y)$.

Estimation in functional models

We need to estimate the eigenvalues and eigenfunctions–we need to estimate C

$$\hat{C}_N(t,s) = \frac{1}{N}\sum_{i=1}^{N}(X_i(t) - \bar{X}_N(t))(X_i(s) - \bar{X}_N(s)) \quad \text{with} \quad \bar{X}_N(t) = \frac{1}{N}\sum_{i=1}^{N} X_i(t).$$

By the law of large numbers in Hilbert spaces we have

$$\|\hat{C}_N - C\| \xrightarrow{P} 0.$$

Empirical eigenvalues and orthonormal eigenfunctions:

$$\hat{\lambda}_{i,N}\hat{\varphi}_{i,N}(t) = \int \hat{C}_N(t,s)\hat{\varphi}_{i,N}(s)ds, \quad 1 \le i < \infty.$$

Theorem 3: $\lambda_1 > \lambda_2 > \ldots > \lambda_d > \lambda_{d+1} > 0$

$$\max_{1\le i\le d} |\hat{\lambda}_{i,N} - \lambda_i| \xrightarrow{P} 0$$

and

$$\max_{1\le i\le d} \|\hat{\varphi}_{i,N} - \hat{c}_{i,N}\varphi_i\| \xrightarrow{P} 0,$$

where $\hat{c}_{i,N}, 1 \le i \le d$ are random signs.

The rate of convergence is exactly $N^{-1/2}$.

Change point detection in the mean curve

Model: $Y_i(t) = \mu_i(t) + X_i(t), \quad 1 \le i \le N$

$$H_0: \quad \mu_1(t) = \mu_2(t) = \ldots = \mu_N(t) \quad \text{in the } L^2 \text{ sense}$$

against the alternative that there is an integer k^* such that

$$\mu_1(t) = \ldots = \mu_{k^*}(t) \neq \mu_{k^*+1}(t) = \ldots = \mu_N(t) \quad \text{in the } L^2 \text{ sense.}$$

The CUSUM (CUmulative SUM) process is

$$S_N^\circ(x,t) = N^{-1/2}\left(\sum_{i=1}^{\lfloor Nx \rfloor} Y_i - \frac{\lfloor Nx \rfloor}{N}\sum_{i=1}^{N} Y_i\right), \quad 0 \le x,t \le 1.$$

By the previous result $S_N^\circ(x,t) \approx \Gamma^\circ(x,t) = \Gamma(x,t) - x\Gamma(1,t)$. Clearly, $\Gamma^\circ(x,t)$ is Gaussian with zero mean and $E\Gamma^\circ(x,t)\Gamma^\circ(y,s) = (\min(x,y) - xy)C(t,s)$

We have under the no change H_0:

$$\sup_{0\le x\le 1}\int (S_N^\circ(x,t))^2 dt \xrightarrow{\mathcal{D}} \sup_{0\le x\le 1}\int (\Gamma^\circ(x,t))^2 dt \quad \text{and} \quad \int\int (S_N^\circ(x,t))^2 dtdx \xrightarrow{\mathcal{D}} \int\int (\Gamma^\circ(x,t))^2 dtdx.$$

Representation:

$$\Gamma^\circ(x,t) = \sum_{i=1}^{\infty} \lambda_i^{1/2}\mathcal{B}_i(x)\varphi_i(t), \quad \mathcal{B}_1, \mathcal{B}_2, \ldots \text{ are iid Brownian bridges}$$

Approximations for the limit distributions

$$\sup_{0\le x\le 1}\int(\Gamma^\circ(x,t))^2dt=\sum_{i=1}^{\infty}\lambda_i\mathcal{B}_i^2(x)\quad\text{and}\quad\int\int(\Gamma^\circ(x,t))^2dtdx=\sum_{i=1}^{\infty}\int\lambda_i\mathcal{B}_i^2(x)dx$$

Approximation:

$$\sup_{0\le x\le 1}\int(\Gamma^\circ(x,t))^2dt\approx\sum_{i=1}^{d}\hat\lambda_{i,N}\mathcal{B}_i^2(x)\quad\text{and}\quad\int\int(\Gamma^\circ(x,t))^2dtdx\approx\sum_{i=1}^{d}\hat\lambda_{i,N}\int\mathcal{B}_{i,N}^2(x)dx$$

How to choose d ?
small λ_i is estimated with large errors
crow's feet

$$\frac{\hat\lambda_{1,N}+\hat\lambda_{2,N}+\ldots+\hat\lambda_{d,N}}{\hat\lambda_{1,N}+\hat\lambda_{2,N}+\hat\lambda_{3,N}\ldots}\approx 0.9\quad(\text{or}\ \ .95,0.99)$$

What happens to these approximations under the alternative ?
Under the alternative there is a symmetric positive function C^* such that $\|\hat C_N-C^*\|\xrightarrow{P}0$. Hence

$$\sum_{i=1}^{d}\hat\lambda_{i,N}\mathcal{B}_i^2(x)\approx\sum_{i=1}^{d}\lambda_i^*\mathcal{B}_i^2(x)\quad\text{and}\quad\sum_{i=1}^{d}\hat\lambda_{i,N}\int\mathcal{B}_{i,N}^2(x)dx\approx\sum_{i=1}^{d}\lambda_i^*\int\mathcal{B}_i^2(x)dx,$$

where $\lambda_1^*\ge\lambda_2^*\ge\ldots\ge\lambda_d^*$ are the eigenvalues of C^*.

Finite critical values under H_0 as well as under H_0 using the method above.

The functionals of $S_N(x,t)$ converge to ∞ in probability under H_A.

Projection method

We use empirical projections using $\hat{\varphi}_{1,N}, \hat{\varphi}_{2,N}, \ldots, \hat{\varphi}_{d,N}$, the eigenfunctions associated with the d largest eigenvalues $\hat{\lambda}_{1,N} \geq \hat{\lambda}_{2,N} \geq \ldots \geq \hat{\lambda}_{d,N}$ of $\hat{C}_N$

Projected CUSUM

$$P_N(x) = \frac{1}{N} \sum_{\ell=1}^{d} \frac{1}{\hat{\lambda}_{\ell,N}} \sum_{i=1}^{\lfloor Nx \rfloor} \langle Y_i - \bar{Y}_N, \hat{\varphi}_{\ell,N} \rangle^2$$
$$= \sum_{\ell=1}^{d} \frac{1}{\hat{\lambda}_{\ell,N}} \langle S^\circ(x, \cdot), \hat{\varphi}_{\ell,N}(\cdot) \rangle^2.$$

Theorem 4: If $X_1, X_2, \ldots, X_N$ are iid (H_0 holds) with $EX_1(t) = 0$ and $E \int X_1^2(t)dt < \infty$, then we have

$$P_N(x) \quad \xrightarrow{D[0,1]} \sum_{i=1}^{d} \mathcal{B}_i^2(x), \quad \text{where } \mathcal{B}_1, \mathcal{B}_2, \ldots \text{ are iid Brownian bridges.}$$

Condition for consistency:

Under H_A we project the differences $Y_i - \bar{Y}_N$ to the space spanned by $\varphi_1^*, \varphi_2^*, \ldots, \varphi_d^*$, the eigenfunctions associated with the d largest eigenvalues of C^*. So the procedure is consistent if $\langle E(Y_{k^*} - Y_{k^*+1}), \varphi_\ell^* \rangle \neq 0$ for at least one $\ell \in \{1, 2, \ldots, d\}$.

Temperatures in Central England 1780–2007 (228 years)

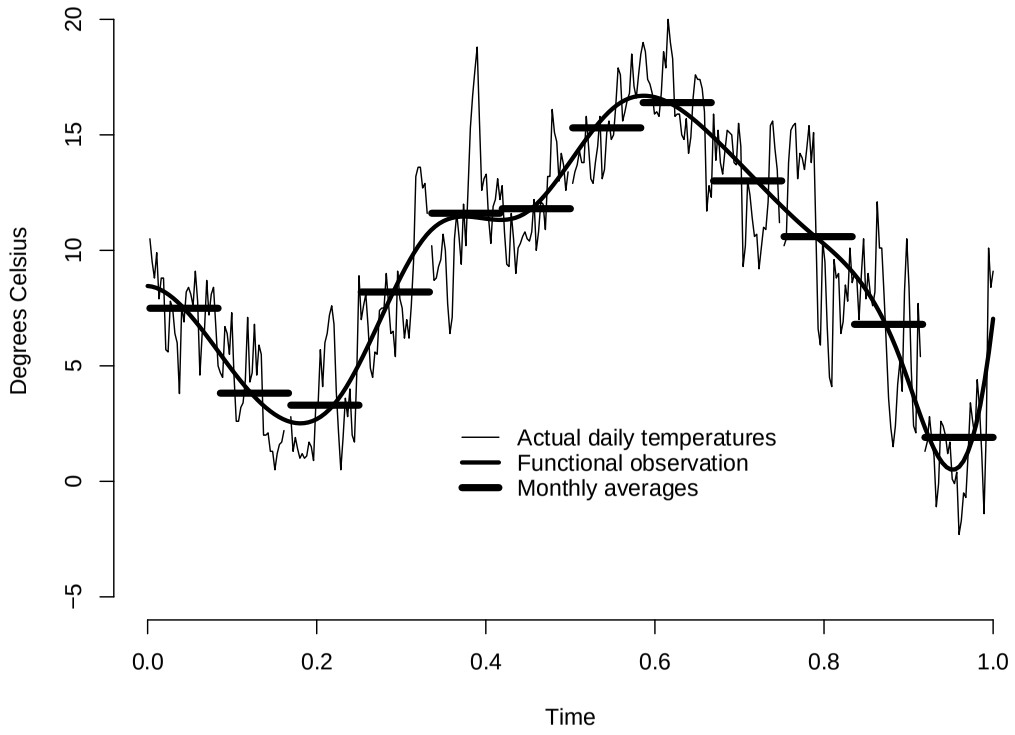


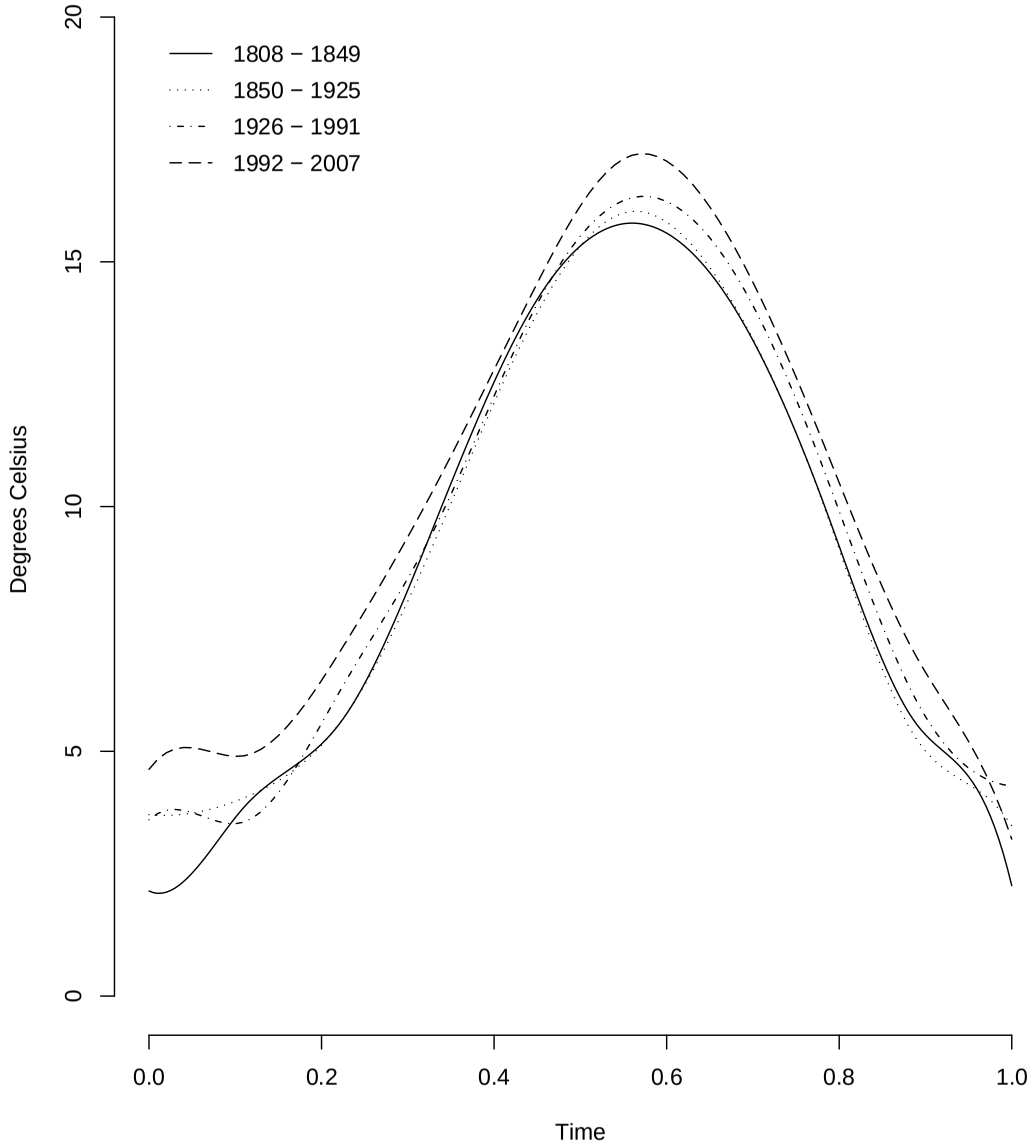
Table 1: Summary and comparison of segmentation. Beginning and end of data period in bold.

Approach	**Change points**					
FDA	**1780**	1808	1850	1926	1992	**2007**
MDA	**1780**	1815		1926		**2007**

Results

The traditional MDA uses $d = 12$ dimensional data, FDA uses $d = 8$ projections.

Average curves



Dependent process in L^2

Autoregressive(1) process in L^2

$$X_k(t) = \int_0^1 \Psi(s,t) X_{k-1}(s) ds + \varepsilon_k(t), \quad -\infty < k < \infty$$

Condition for stationarity:

$$\varepsilon_k(t) \quad \text{are independent and identically distributed}$$

$$E\varepsilon_0(t) = 0 \quad \text{and} \quad \int_0^1 E\varepsilon_0^2(t) dt < \infty,$$

$$\int_0^1 \int_0^1 |\Psi(s,t)|^m \, ds \, dt < 1 \quad \text{with some} \quad m \geq 2.$$

Explicit expression for the stationarity solution $X_k(t)$ using iterations of $\Psi(s,t)$ and the errors $\varepsilon_j, j \leq k$.

More dependent process in L^2

Functional ARMA(p, q)

$$X_k(t) = \sum_{\ell=1}^{p} \int_0^1 \Psi_\ell(s,t) X_{k-\ell}(s)\, ds + \varepsilon_k(t) + \sum_{\ell=1}^{q} \int_0^1 \Theta_\ell(s,t) \varepsilon_{k-\ell}(s)\, ds, \quad t \in [0,1], \quad -\infty < k < \infty.$$

There is no necessary and sufficient condition for the existence of the stationary solution.

Functional linear processes

$$X_k(t) = \sum_{\ell=0}^{\infty} \int \Psi_\ell(s,t) \varepsilon_{k-\ell}(s) ds.$$

Condition for stationarity:

$$\varepsilon_k(t) \quad \text{are independent and identically distributed}$$

$$E\varepsilon_0(t) = 0 \quad \text{and} \quad \int_0^1 E\varepsilon_0^2(t) dt < \infty,$$

$$\sum_{\ell=1}^{\infty} \|\Psi_\ell\| < \infty.$$

Volatility process in L^2

Functional ARCH(1)

$$X_k(t) = \varepsilon_k(t)\sigma_k(t), \quad \text{and} \quad \sigma_k^2(t) = \delta(t) + \int_0^1 \Psi(s,t)X_{k-1}^2(s)\,ds$$

$$\delta \geq 0 \quad \text{and} \quad \Psi \geq 0$$

Condition for consistency:

$$E\left\{\int_0^1 \int_0^1 \Psi^2(t,s)\varepsilon_0^4(s)dtds\right\}^{\tau} < 1 \quad \text{with some } \tau > 0$$

constant conditional correlation

Functional GARCH(1,1)

$$X_k(t) = \varepsilon_k(t)\sigma_k(t),$$

$$\sigma_k^2(t) = \delta(t) + \int_0^1 \Psi_1(s,t)X_{k-1}^2(s)\,ds + \int_0^1 \Psi_2(s,t)Y\sigma_{k-1}^2(s)$$

$$\delta \geq 0 \quad \text{and} \quad \Psi_1 \geq 0 \quad \Psi_2 \geq 0$$

Bernoulli processes in Hilbert spaces

there is a functional $a\colon S^\infty \to L^2$ such that $X_k = a(\epsilon_k, \epsilon_{k-1}, \ldots)$

a is a measurable functional

$(\epsilon_k\colon k \in \mathbb{Z})$ are iid sequences of random elements with values in some measurable spaces S

$\omega_0(t;\omega)$ is jointly measurable in (t,ω)

Weak Dependence

$$E\|X_k\|^\kappa < \infty, \quad \text{and} \quad \sum_{\ell=1}^{\infty} (E\|X_k - X_k^{(\ell)}\|^\kappa)^{1/\kappa} < \infty \quad \text{with some} \quad \kappa > 2,$$

where

$$X_k^{(\ell)} = a(\epsilon_k, \epsilon_{k-1}, \ldots, \epsilon_{k-\ell+1}, \epsilon_{k-\ell}^{(\ell)}, \epsilon_{k-\ell-1}^{(\ell)}, \ldots),$$

$(\epsilon_k^{(\ell)}\colon k, \ell \in \mathbb{Z})$ are iid copies of ϵ_0

dependence on previous innovations is "dying our" fast enough

Sums of weakly dependent functional time series

$$S_N(x,t) = \frac{1}{N^{1/2}} \sum_{i=1}^{\lfloor Nx \rfloor} X_i(t)$$

Theorem 5: Under the weak dependence assumptions, there is a sequence of Gaussian processes $\bar{\Gamma}_N(x,t)$ with

$$E\bar{\Gamma}_N(x,t) = 0 \quad \text{and} \quad E\bar{\Gamma}_N(x,t)\bar{\Gamma}_N(y,s) = \min(x,y)D(t,s)$$

such that

$$\sup_{0 \le x \le 1} \|\alpha_N(x,t) - \Gamma_N(x,t)\| \xrightarrow{P} 0,$$

as $N \to \infty$, where

$$D(t,s) = \sum_{\ell=-\infty}^{\infty} EX_0(t)EX_\ell(s).$$

What is the best basis? the eigenfunctions of $C(t,s) = EX_1(t)X_1(s)$ or the eigenfunctions of the long run covariance function $D(t,s)$?

We project $S_N(x,t)$ so we need to preserve the variability in $S_N(x,t)$ – use the eigenfunctions of the long run covariance function $D(t,s)$.

Estimation of the long run covariance function

We estimate $D(t,s)$ with

$$\hat{D}_N(t,s) = \sum_{i=-\infty}^{\infty} K\left(\frac{i}{h}\right) \hat{\gamma}_i(t,s)$$

with the empirical covariances

$$\hat{\gamma}_i(t,s) = \hat{\gamma}_{i,N}(t,s) = \begin{cases} \frac{1}{N}\sum_{j=1}^{N-i}(X_j(t) - \bar{X}_N(t))(X_{j+i}(s) - \bar{X}_N(s)), & i \geq 0 \\ \frac{1}{N}\sum_{j=1-i}^{N}(X_j(t) - \bar{X}_N(t))(X_{j+i}(s) - \bar{X}_N(s)), & i < 0 \end{cases}$$

and the sample mean

$$\bar{X}_N(t) = \frac{1}{N}\sum_{\ell=1}^{N} X_\ell(t).$$

Assumption on the window size:

$$h = h(N) \to \infty \quad \text{and} \quad \frac{h(N)}{N} \to 0, \quad \text{as} \quad N \to \infty$$

Assumptions on the kernel

$$K(0) = 1$$

$$K \text{ is symmetric around } 0 \text{ and } K(u) = 0 \text{ if } u > c \text{ with some } c > 0$$

$$K \text{ is Lipschitz continuous on } [-c, c]$$

Examples: Bartlett kernel, Parzen kernel, flat–top kernel of Politis and so on

Theorem 6: We assume that the weak Bernoulli assumption, and the conditions on h and K hold, then we have

$$\|\hat{D}_N - D\| \xrightarrow{P} 0.$$

How to choose h ? Minimize the mean–squared error, i.e. minimize $E\|\hat{D}_N - D\|^2$. Using standard arguments we need that in case of the "optimal" h for large N (so large h) we have $E\|\hat{D}_N - E\hat{D}_N\|^2 = \|E\hat{D}_N - D\|^2$. The "optimal" h depends on the unknown D–practical (data driven ways) to choose h.

Data driven choice of the window

Flat–top kernel:

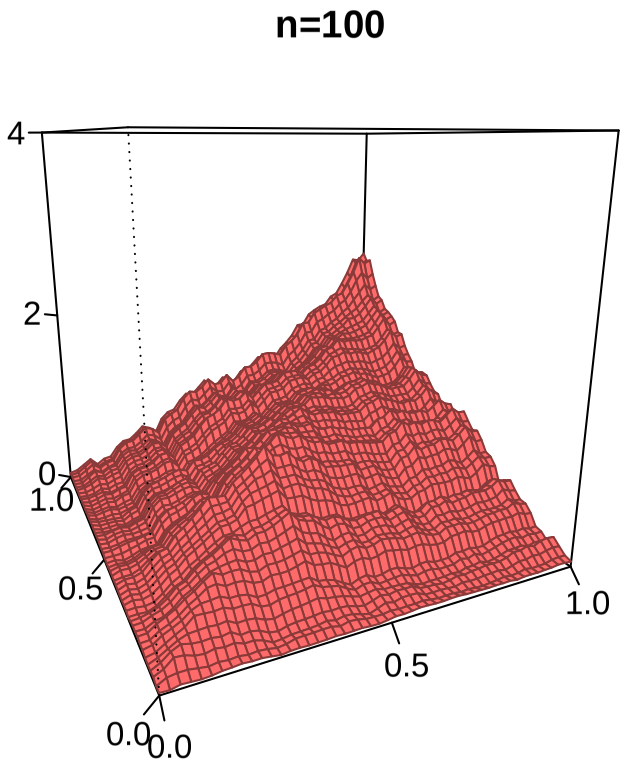
$$K_f(t;x) = \begin{cases} 1, & 0 \le |t| < x \\ (x-1)^{-1}(|t|-1), & x \le |t| < 1 \\ 0, & |t| \ge 1, \end{cases}$$

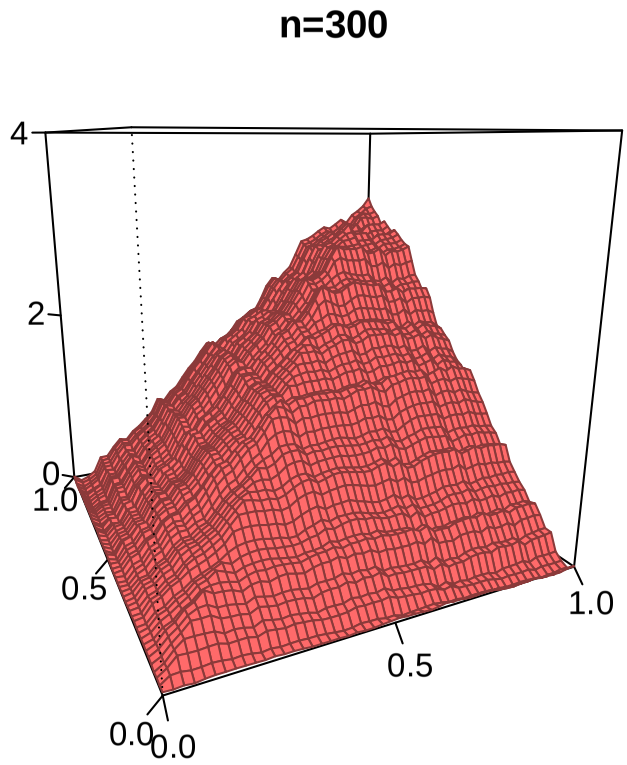
$$\hat{\rho}_i = \|\hat{\gamma}_i\| / \int \hat{\gamma}_0(t,t)dt$$

Procedure for choosing h**:** Find the first non–negative integer $\hat{m}$ such that $\hat{\rho}_{\hat{m}+r} < T\sqrt{\log N/N}$ for $r = 1, ..., H$, where $T > 0$, and H is a positive integer. Take $h = \hat{h}$ where $\hat{h} = \lceil \hat{m}/x \rceil$.

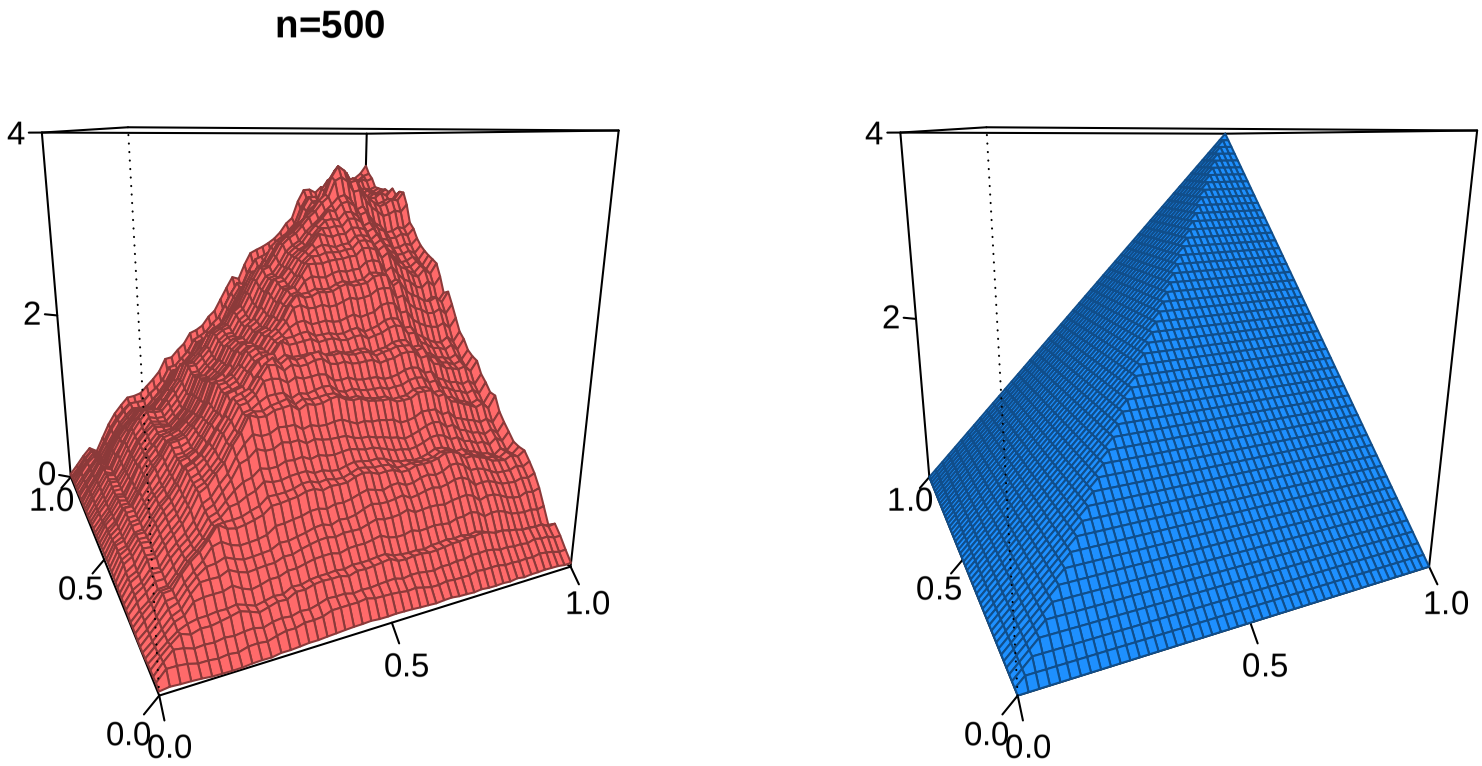
Simulations: $\mathrm{FAR}^*_{1/2}(1)$

$$X_i(t) = \frac{1}{2}X_{i-1}(t) + W_i(t), \quad \text{where } W_i, -\infty < i < \infty \text{ are iid Wiener processes}$$

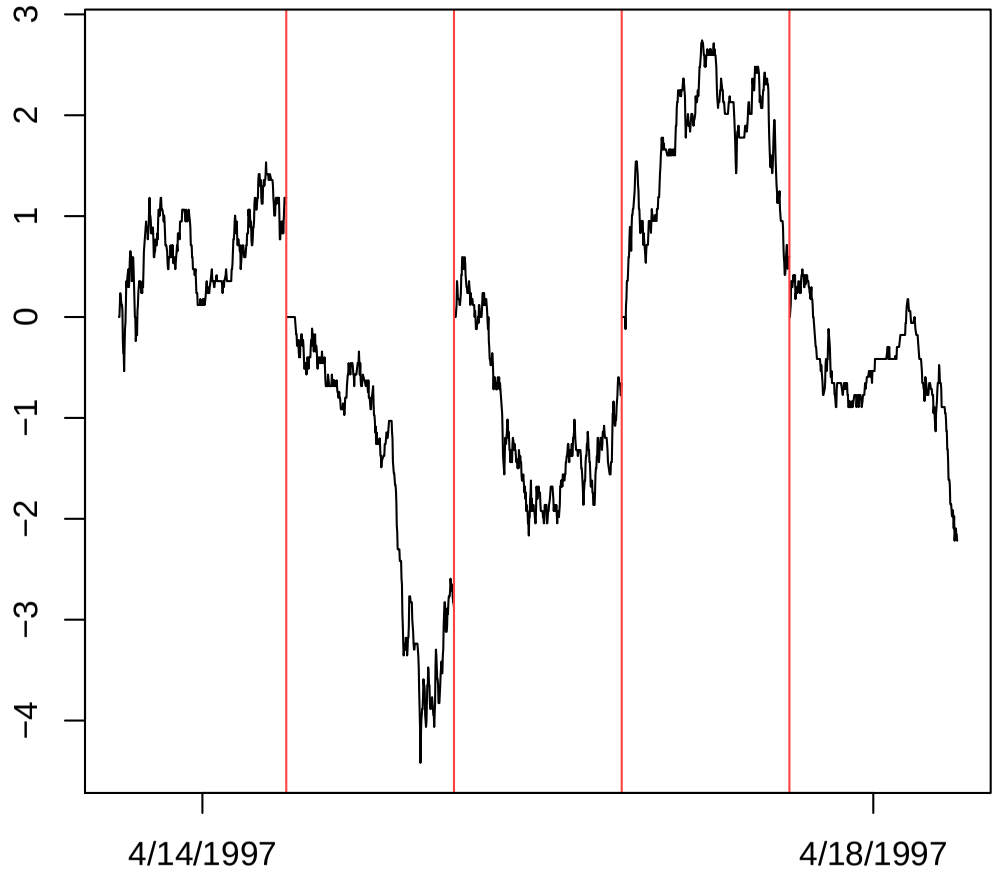


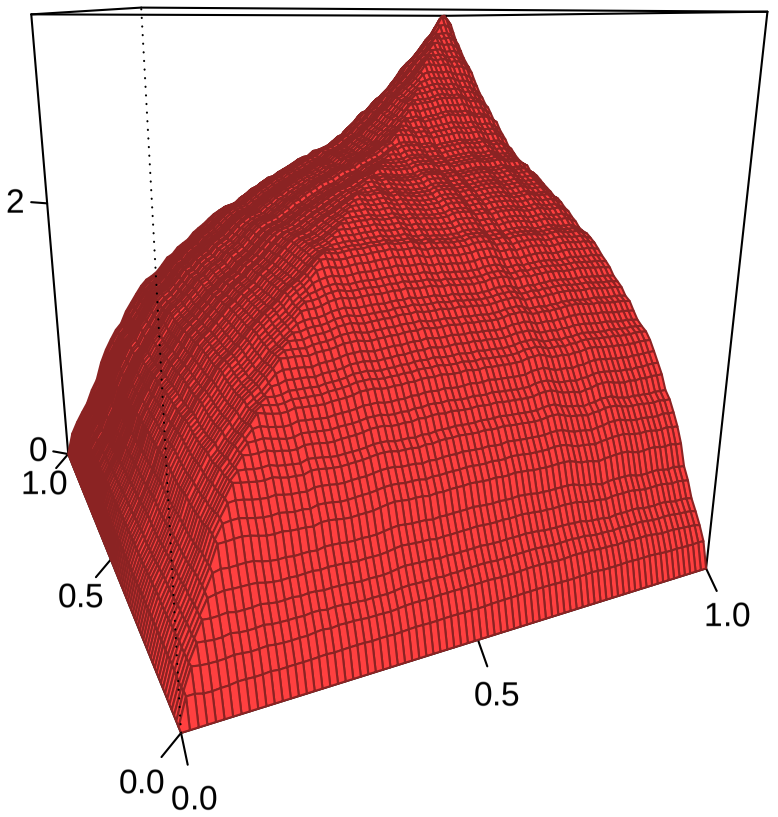


Simulations



The limit





Cumulative intraday returns

Functional analysis of variance

Model

$$X_{i,j}(t) = \mu_i(t) + \varepsilon_{i,j}(t), \quad t \in [0,1], \quad 1 \le i \le k \text{ and } 1 \le j \le N_i,$$

Model assumptions

$$E\eta_{i,j}(t) = 0, \ t \in [0,1], \quad 1 \le i \le k, \quad \text{and} \quad 1 \le j \le N_i$$

for each $i, 1 \le i \le k$, $\{\eta_{i,j}, 0 \le j < \infty\}$ is a weakly dependent Bernoulli sequence

the error sequences $\{\eta_{i,j}, 1 \le j \le N_i\}$ are independent

$$\lim_{N\to\infty} \frac{N_i}{N} = a_i > 0, \quad \text{where} \quad N = N_1 + N_2 + \cdots + N_k$$

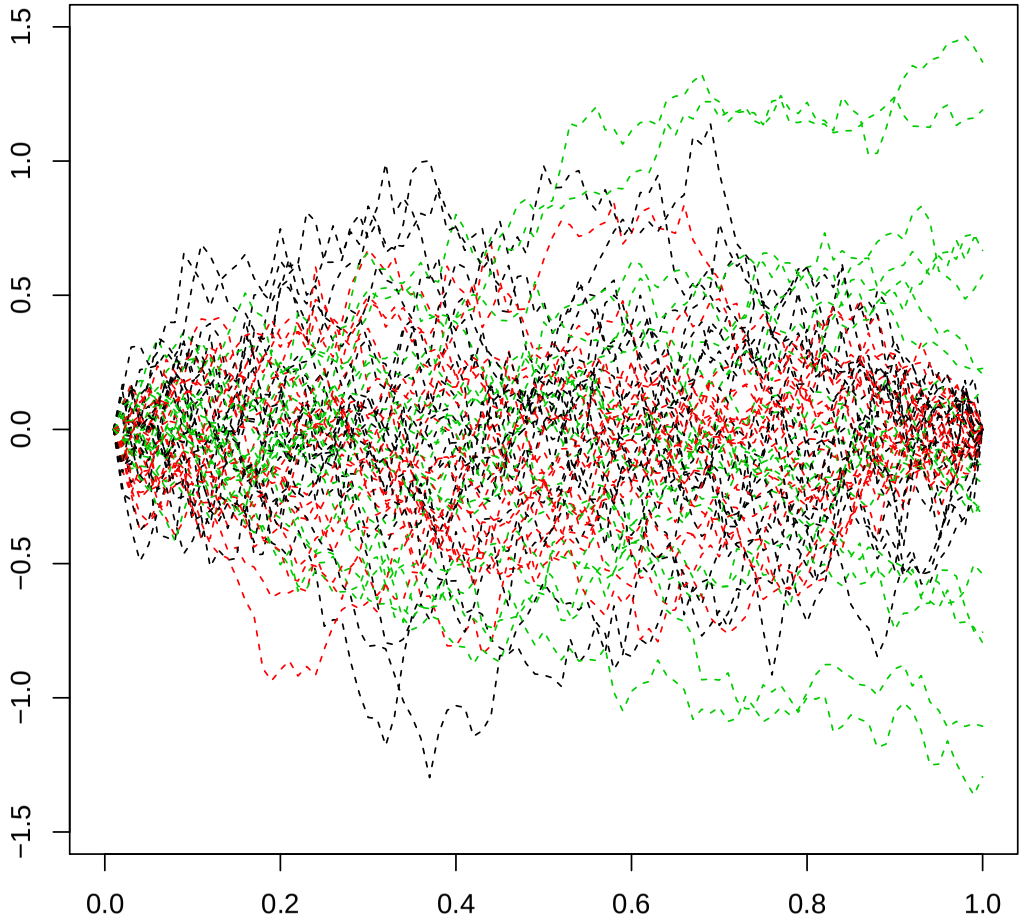
FANOVA null hypothesis

$$H_0 : \mu_1(\cdot) = \mu_2(\cdot) = \ldots = \mu_k(\cdot)$$

FANOVA alternative

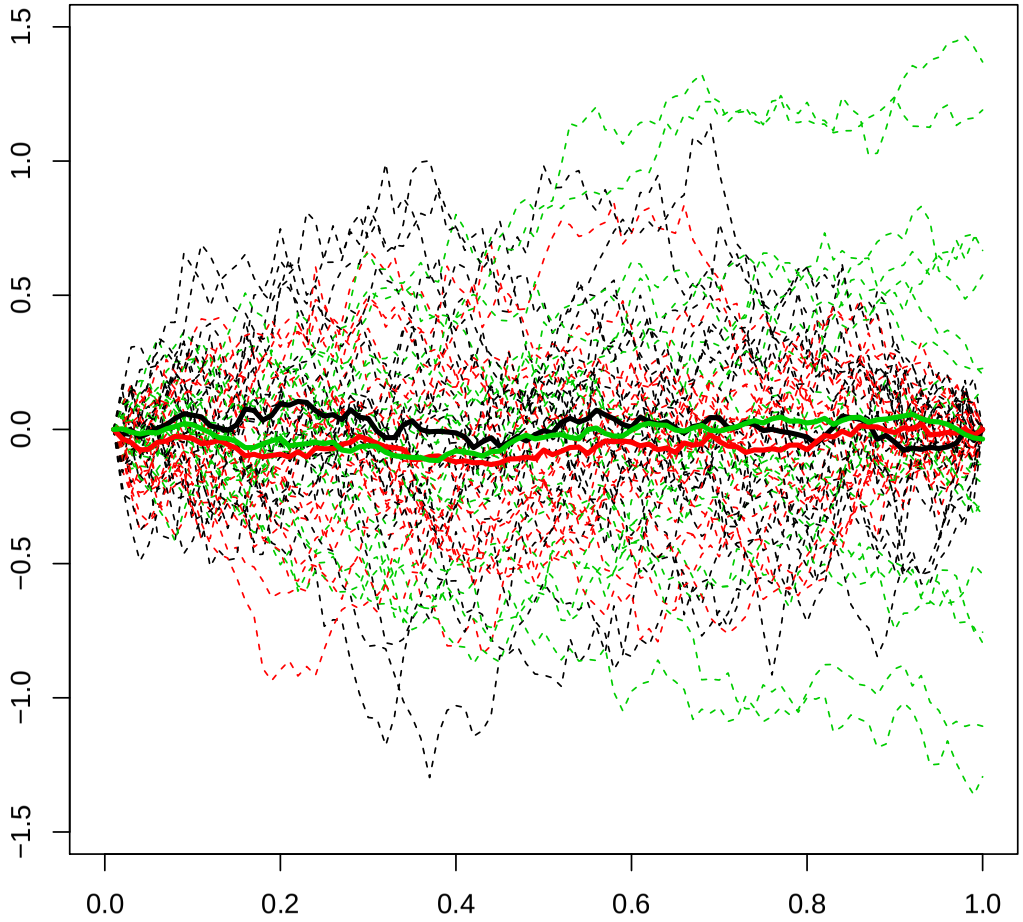
$$H_A : \quad H_0 \text{ does not hold}$$

Figure 1: Three populations



Three populations

Figure 2: Three populations with the sample means



Three populations with the sample means

Projection method

What basis to use?–We are comparing the means, hence averages must be compared. We will use a bases related to the long run covariances of the individual samples.

D_i is the long run covariance function of the i^{th} population,

$$D_i(t,s) = \sum_{\ell=-\infty}^{\infty} \text{cov}(X_{i,0}(t), X_{i,\ell}(s)).$$

We use the eigenfunctions of

$$D(t,s) = \sum_{i=1}^{k} \frac{N_i}{N} D_i(t,s), \qquad N = N_1 + N_2 + \cdots + N_k.$$

Estimation of D

$$\hat{D}_N(t,s) = \sum_{i=1}^{k} \frac{N_i}{N} \hat{D}_{i,N}(t,s)$$

Do we need the consistency of $\hat{D}_N(t,s)$? Under the null–definitely. Under the alternative?

Estimation of D–first method

$$\hat{D}_{i,N}(t,s) = \sum_{\ell=-\infty}^{\infty} K\left(\frac{\ell}{h}\right) \hat{\gamma}_{i,\ell}(t,s)$$

with the empirical covariances

$$\hat{\gamma}_{i,\ell}(t,s) = \hat{\gamma}_{i,N}(t,s) = \begin{cases} \frac{1}{N}\sum_{j=1}^{N-\ell}(X_{i,j}(t) - \bar{X}_N(t))(X_{i,j+i}(s) - \bar{X}_N(s)), & i \geq 0 \\ \frac{1}{N}\sum_{j=1-\ell}^{N}(X_{i,j}(t) - \bar{X}_N(t))(X_{i,j+i}(s) - \bar{X}_N(s)), & i < 0 \end{cases}$$

and the sample mean of the total population

$$\bar{X}_N(t) = \frac{1}{N}\sum_{i=1}^{k}\sum_{\ell=1}^{N_i} X_{i,\ell}(t).$$

Consistent only under the null hypothesis! Diverges to ∞ under the alternative.

Estimation of D–second method

$$\tilde{D}_{i,N}(t,s) = \sum_{\ell=-\infty}^{\infty} K\left(\frac{\ell}{h}\right) \tilde{\gamma}_{i,\ell}(t,s)$$

with the empirical covariances

$$\tilde{\gamma}_{i,\ell}(t,s) = \hat{\gamma}_{i,N}(t,s) = \begin{cases} \frac{1}{N}\sum_{j=1}^{N-\ell}(X_{i,j}(t) - \bar{X}_{i,N_i}(t))(X_{i,j+i}(s) - \bar{X}_{i,N_i}(s)), & i \geq 0 \\ \frac{1}{N}\sum_{j=1-\ell}^{N}(X_{i,j}(t) - \bar{X}_{i,N_i}(t))(X_{i,j+i}(s) - \bar{X}_{i,N_i}(s)), & i < 0 \end{cases}$$

and the sample mean of the i^{th} population

$$\bar{X}_{i,N_i}(t) = \frac{1}{N_i}\sum_{\ell=1}^{N_i} X_{i,\ell}(t)$$

Consistent under the null as well as under the alternative.

Projections based on the first method

Empirical eigenvalues/eigenfunctions:

$$\hat{\lambda}_i \hat{\varphi}_i(t) = \int \hat{D}_N(t,s)\hat{\varphi}_i(s)ds$$

Empirical projections of the i^{th} population:

$$\hat{\boldsymbol{\xi}}_{i,j} = (\langle X_{i,j}, \hat{\varphi}_1\rangle, \langle X_{i,j}, \hat{\varphi}_2\rangle, \ldots, \langle X_{i,j}, \hat{\varphi}_d\rangle)^T, \quad 1 \le j \le d$$

The averages of the empirical score vectors within each population are defined as

$$\hat{\boldsymbol{\xi}}_{i\cdot} = \frac{1}{N_i}\sum_{j=1}^{N_i} \hat{\boldsymbol{\xi}}_{i,j}, \quad 1 \le i \le k$$

Estimator for the common mean assuming that H_0 holds:

$$\hat{\boldsymbol{\xi}}_{\cdot\cdot} = \left(\sum_{i=1}^{k} N_i \hat{\boldsymbol{\Sigma}}_i^{-1}\right)^{-1} \sum_{i=1}^{k} N_i \hat{\boldsymbol{\Sigma}}_i^{-1} \hat{\boldsymbol{\xi}}_{i\cdot},$$

where

$$\hat{\boldsymbol{\Sigma}}_i = \left\{ \iint \hat{D}_{i,N_i}(t,s)\hat{\varphi}_\ell(t)\hat{\varphi}_j(s)dtds,\ 1 \le j,\ell \le d \right\},$$

Test statistic:

$$\hat{T}_N = \sum_{i=1}^{k} N_i \left(\hat{\boldsymbol{\xi}}_{i\cdot} - \hat{\boldsymbol{\xi}}_{\cdot\cdot}\right)^T \hat{\boldsymbol{\Sigma}}_i^{-1} \left(\hat{\boldsymbol{\xi}}_{i\cdot} - \hat{\boldsymbol{\xi}}_{\cdot\cdot}\right).$$

Distribution of $\hat{T}_N$ under H_0 and H_A

Theorem 7: We assume that the populations are independent weakly dependent Bernoulli shifts.

(H_0) If the means are the same, than we have

$$\hat{T}_N \quad \stackrel{\mathcal{D}}{\to} \quad \chi^2(d(k-1)),$$

where $\chi^2(d(k-1))$ stands for a χ^2 random variable with $d(k-1)$ degrees of freedom.

(H_A) If at least two means are different, then we have

$$\hat{T}_N \quad \stackrel{P}{\to} \quad \infty.$$

Electricity demand in Adelaide, Australia

Daily electricity demand curves constructed from half–hourly measurements of the electricity demand in Adelaide Australia from 7/6/1997 to 3/31/2007.

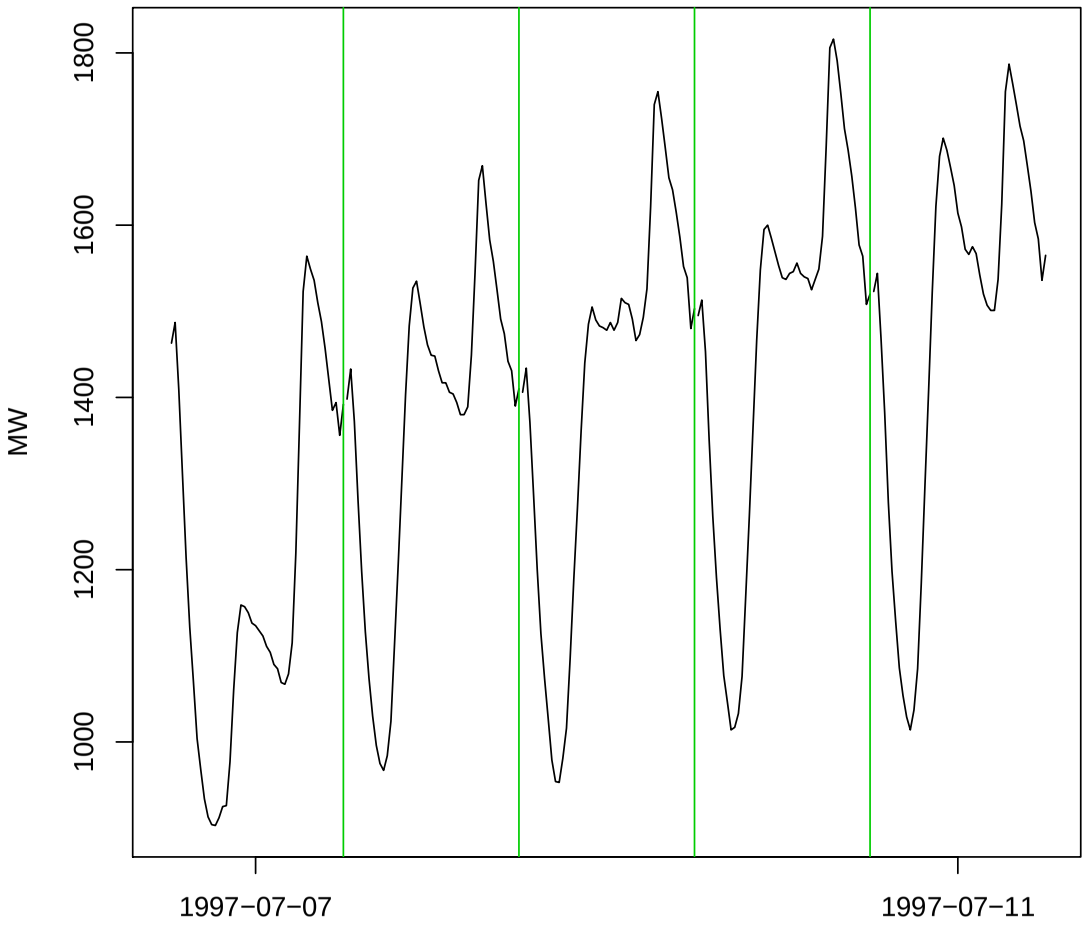
Each day is comprised of 48 observations

The cost of unserved energy can be valued at thousands of dollars per MWh, and hence there is an incentive to develop accurate models of the daily demand in order to reduce excess electricity generation.

Differing trends in the demand according to the season and the day of the weak

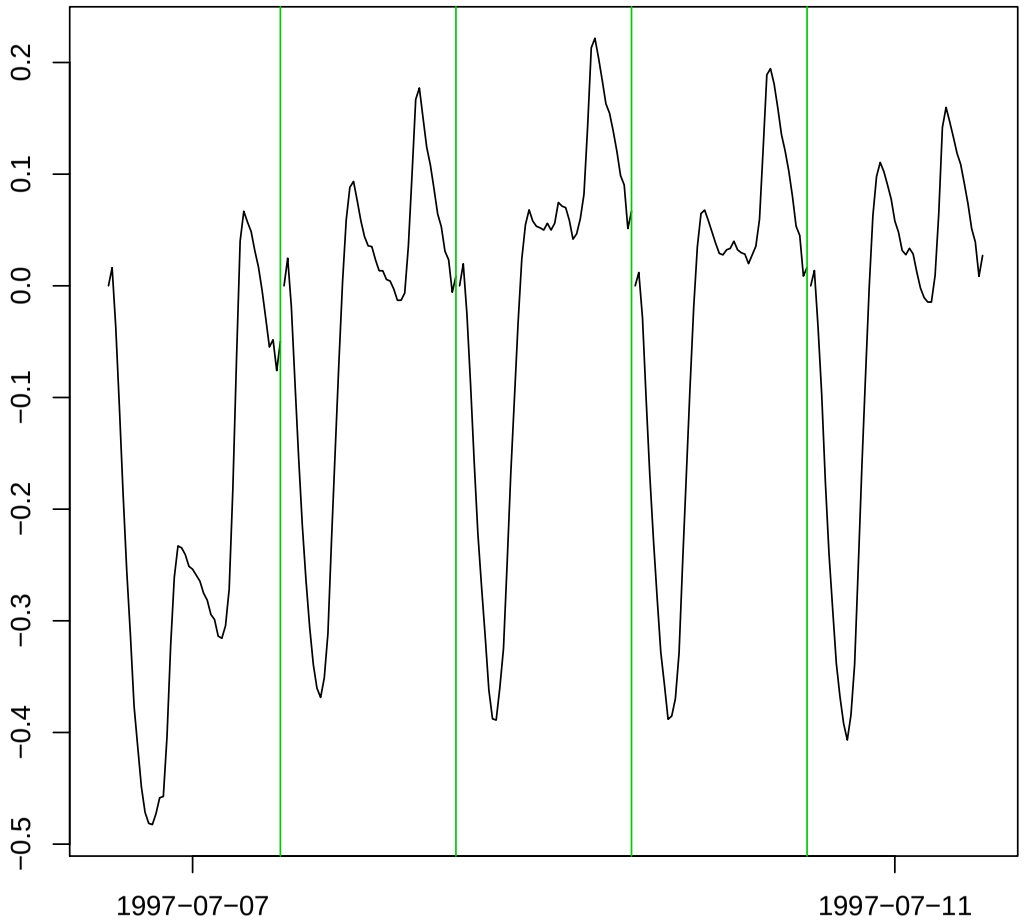
Suppose $U_n(t)$ is the electricity demand at time t on day n for $t \in [0,1],\ n = 1,\ldots,N$. The functions $R_n(t) = \ln D_n(t) - \ln D_n(0),\ t \in [0,1],\ n = 1,\ldots,N$, are called the *log differenced demand curves* (LDDC's).

Figure 3: Five functional data objects constructed from half-hourly measurements of the electricity demand in Adelaide Australia. The vertical lines separate the days.



Demand curves

Figure 4: Five LDDC's constructed from the curves in Figure 3.



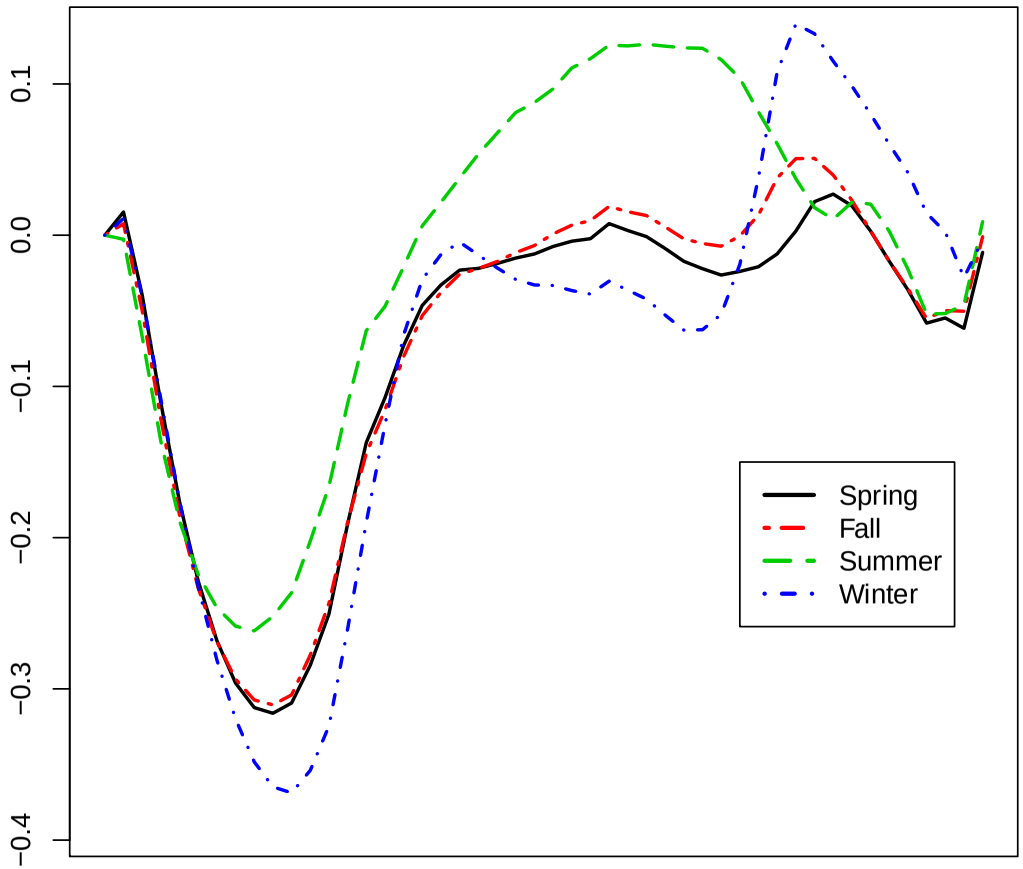
Differenced demand curves

Seasonal effect

First we consider the problem of testing if the mean of the LDDC's is homogeneous across the four predominant seasons in Adelaide: Summer (December, January, February), Fall (March, April, May), Winter (June July August), and Spring (September, November, December). We divided the data set consisting of 3556 daily curves into these four seasonal groups depending on the month in which the observation was taken. From this sample the observations corresponding to the weekends were removed since the demand behavior is vastly different on these days. After removing the weekends in total there are 642 observations from the Spring months, 628 from Fall, 630 from Summer, and 640 from Winter. The mean functions from these samples are shown in Figure 5 on the next slide.

When the FANOVA test is applied to these four populations the test rejects the null hypothesis with a p–value which is less than 10^{-6}. By examining Figure 5 on the next slide it appears that Spring and Fall have similar mean LDDC's. The approximate p–value of the test when applied to just the Spring and Fall samples is approximately .21, indicating that there is not sufficient evidence present in the data to reject the notion that Spring and Fall have the same demand patterns.

Figure 5: Mean Curves from each season constructed from the LDDC's taken from 7/6/1997 to 3/31/2007. The p-value of the FANOVA test applied to this sample was zero.



Seasonal mean curves

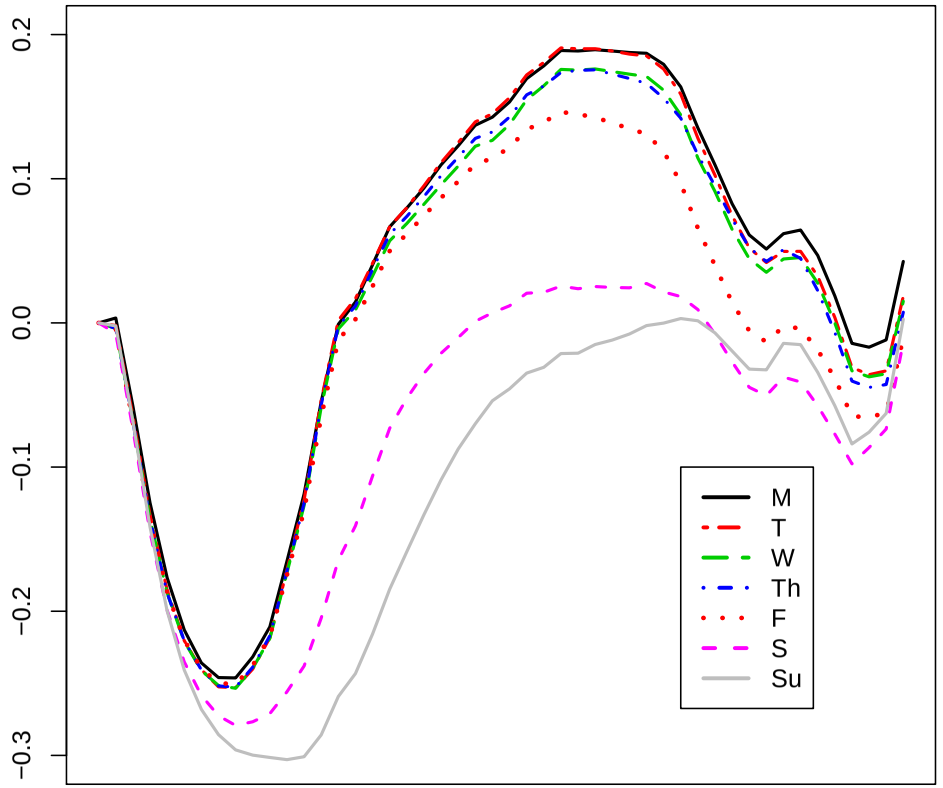
	Days of the week for which the test is applied					
Season	All	Weekdays	Weekends	TWTh	MTWTh	TWThF
Summer	.000	.035	.006	.750	.647	.056
Fall	.000	.003	.001	.886	.380	.023
Winter	.000	.000	.000	.257	.001	.000
Spring	.000	.002	.000	.582	.083	.001

Table 2: p–values of the FANOVA test when applied to samples of daily LDDC's organized according to the day of the week and season. Across the top of the table the days included in the sample are displayed. "All" denotes that all seven days were included ($k = 7$).

Daily variation

To study whether the daily pattern in electricity demand is homogeneous across each day of the week we divided the data set into seven groups each of size 508 corresponding to the days of the week, Sunday through Saturday, and then computed their LDDC's. Due to the prior analysis of the seasonal trend above we further grouped the data into the four seasonal groups of Summer, Fall, Winter, and Spring; each subsample for each day contained at least 120 curves.

Figure 6: Mean curves for each day computed from the Summer months between 7/7/1997 to 7/5/1998 (52 curves for each day). The p-value of the FANOVA test was less than 10^{-4}.



Daily averages

Some further results on functional data

Tests for stationarity Is it true that the observations form a stationary sequence? If not what is the cause of the non stationarity? Changing mean? Changing covariance? Random walk errors (unit root problem)?

The asymptotic distribution of the estimator for the long run covariance function It is normally distributed, and therefore the corresponding eigenvalues/eigenfunction are normal too. Optimal choice of the window. Sample based selection of the window. Positive definite function.

Heteroscedastic errors The covariance function (long run covariance function) might depend on time.

Change point with heteroscedastic errors

Model: $X_i(t) = \sum_{k=1}^{K} \beta_{i,k} f_k(t) + \epsilon_i(t), \quad E\epsilon_i(t) = 0, \quad 0 \le t \le 1, 1 \le i \le N$

Motivation: Nelson–Siegel model for yield curves

$f_k(t)$ are given functions

$$\beta_{i,k} = \mu_{i,k} + b_{i,k} \quad Eb_{i,k} = 0$$

$$H_0 : \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \ldots = \boldsymbol{\mu}_N \qquad \boldsymbol{\mu}_i = (\mu_{i,1}, \mu_{i,2}, \ldots, \mu_{i,K})^\top.$$

$$H_A : \quad R \text{ possible changes in the means of the } \boldsymbol{\mu}_i' s$$

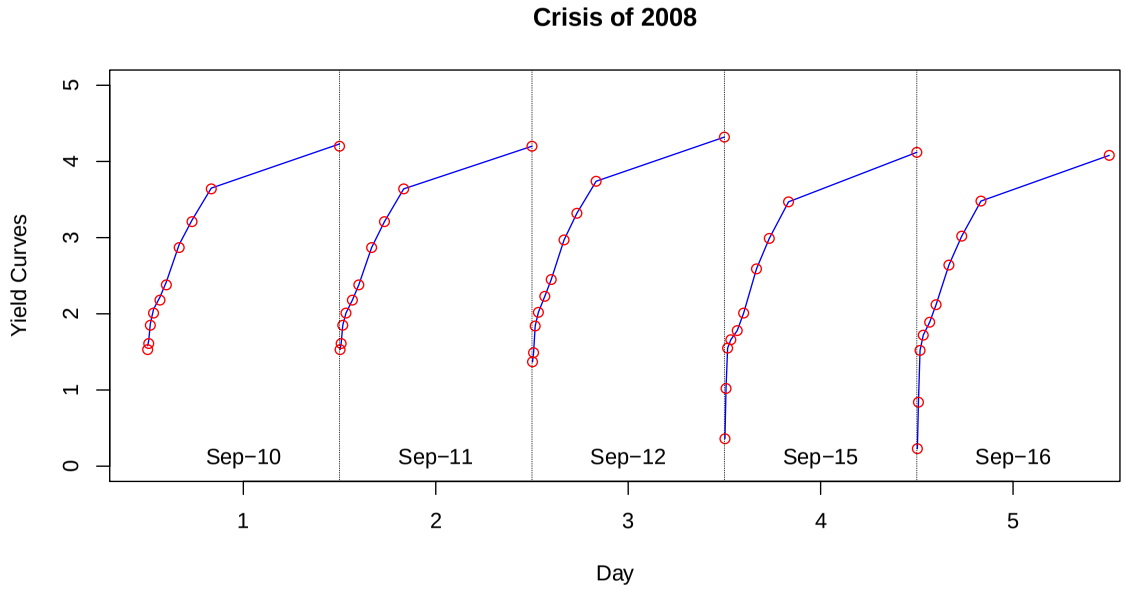
R and the times of the changes are unknown (change point problem in the mean)

Heteroscedasticity: the covariance structure becomes different at times $i_1 < i_2 < \ldots < i_M$ (these are known points) The error term is $\sum_{i=1}^{K} b_{i,k} f_k(t) + \epsilon_i(t), 1 \le i \le N$

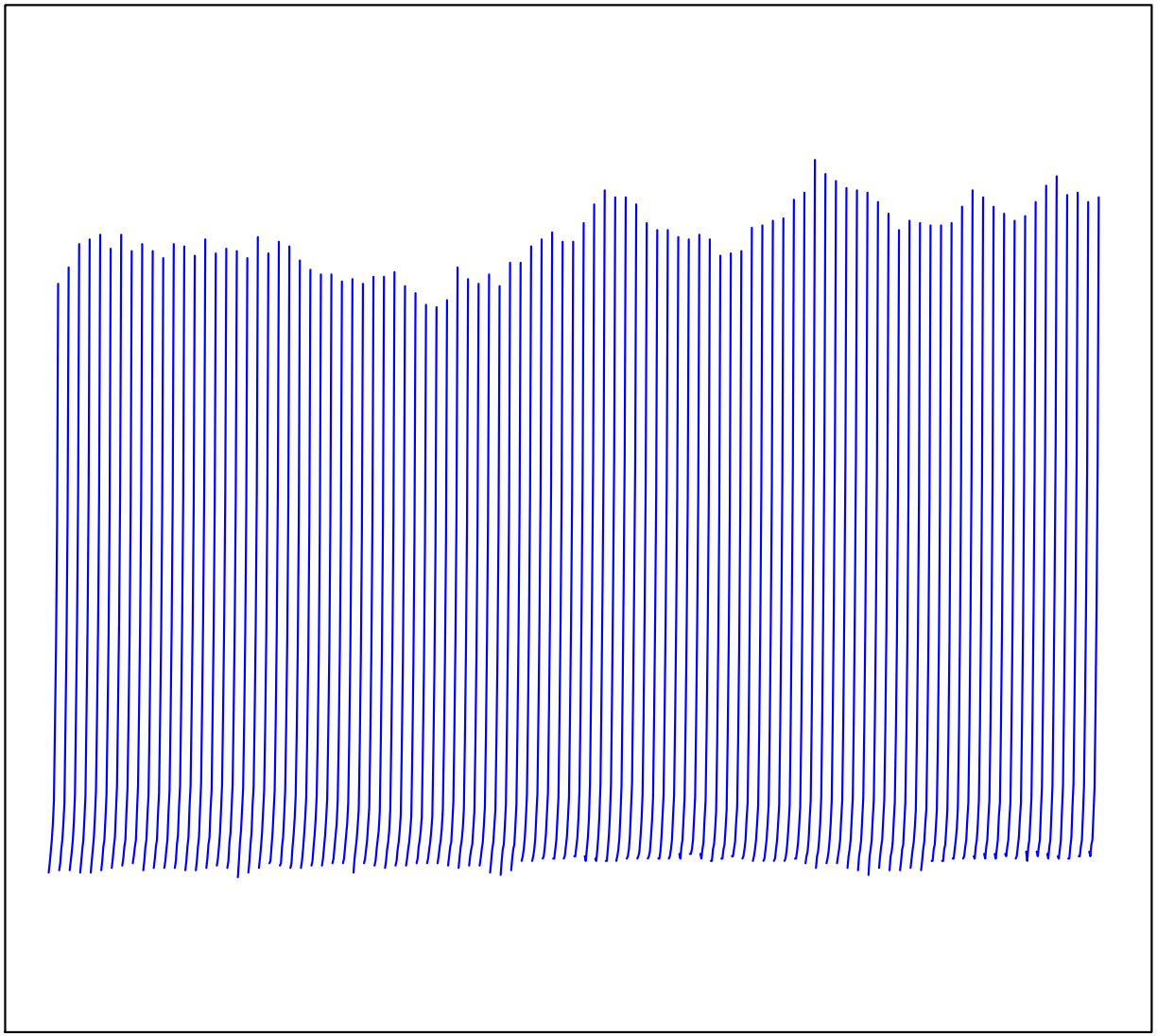
Error structure: Bernoulli shifts

$$(b_{i,1}, \ldots, b_{i,K}, \epsilon_i(t))^\top = g_m(\delta_i, \delta_{i-1} \ldots) \quad i_m < i_{m+1}, m = 0, 1, \ldots, M \ (i_0 = 0, i_{M+1} = N)$$

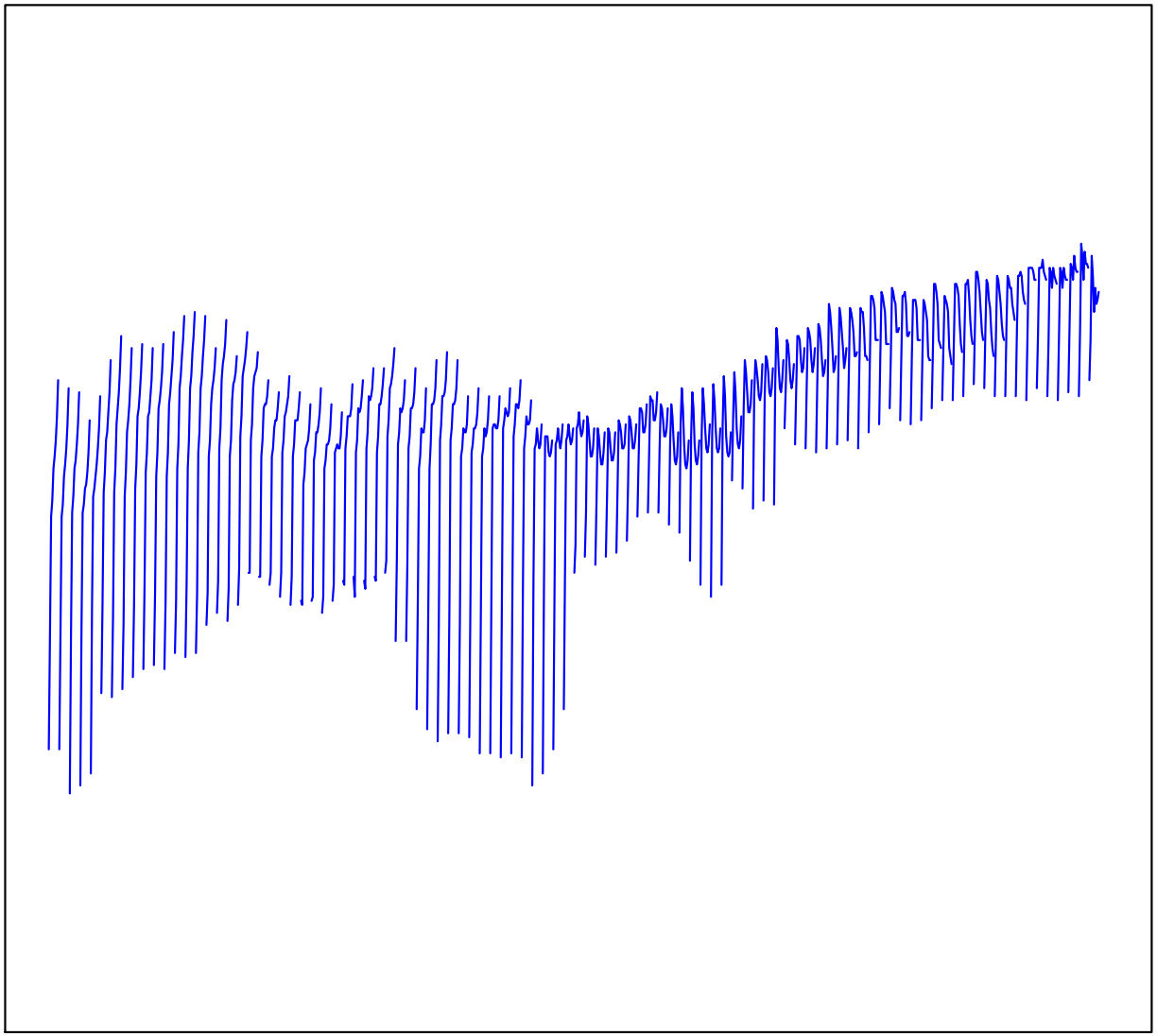
US yields curves for 5 days



US yields curves between June–4–2012 and October–24–2012



US yields curves between October–18–2005 and March–14–2006



Projection method

Projections into span$(f_1, \ldots, f_K)$ $\quad$ ($f_1, f_2, \ldots, f_K$ are linearly independent in L^2

$$\boldsymbol{z}_i = (\langle X_i, f_1\rangle, \langle X_i, f_2\rangle, \ldots, \langle X_K, f_K\rangle)^\top, \quad 1 \le i \le N$$

Let

$$\boldsymbol{C} = \{\langle f_i, f_j\rangle, 1 \le i, j \le K\}, \quad \boldsymbol{\epsilon}_i = (\langle \epsilon_i, f_1\rangle, \langle \epsilon_i, f_2\rangle, \ldots \langle \epsilon_i, f_K\rangle)^\top, \quad \boldsymbol{b}_i = (b_{i,1}, b_{i,2}, \ldots, b_{i,K})^\top.$$

Projected Model:

$$\boldsymbol{z}_i = \boldsymbol{C}\boldsymbol{\mu}_i + \boldsymbol{\gamma}_i, \quad \text{with} \quad \boldsymbol{\gamma}_i = \boldsymbol{C}\boldsymbol{b}_i + \boldsymbol{\epsilon}_i$$

CUSUM process

$$\boldsymbol{\alpha}_N(x) = N^{-1/2}\left(\sum_{i=1}^{\lfloor Nx\rfloor} \boldsymbol{z}_i - \sum_{i=1}^{N} \boldsymbol{z}_i\right)$$

Theorem 1 If H_0 and the Bernoulli assumption holds, then

$$\boldsymbol{\alpha}_N(x) \xrightarrow{\mathcal{D}} \boldsymbol{G}^0(x) \quad \text{in} \quad \mathcal{D}^K([0,1]),$$

where $\boldsymbol{G}^0$ is a Gaussian process with $E\boldsymbol{G}^0(x) = 0$ and $E\boldsymbol{G}^0(x)(\boldsymbol{G}^0(y))^\top = \boldsymbol{R}(x,y)$ =explicitly computed (difficult looking).

Two test statistics

$$\mathcal{C}_N = \int_0^1 \|\boldsymbol{\alpha}_N(x)\|^2 dx \quad \text{and} \quad \mathcal{K}_N = \sup_{0\le x\le 1} \|\boldsymbol{\alpha}_N(x)\|^2.$$

Cramér–von Mises statistic under H_0

$$\mathcal{C}_N \xrightarrow{\mathcal{D}} \int_0^1 \|\boldsymbol{G}^0(x)\|^2 dx$$

Karhunen–Loéve expansion

$$\int_0^1 \|\boldsymbol{G}^0(x)\|^2 dx = \sum_{\ell=1}^{\infty} \lambda_\ell Z_\ell^2,$$

where $Z_1, Z_2, \ldots$ are independent standard normals, $\lambda_1 \geq \lambda_2 \geq \ldots$ satisfying

$$\lambda_\ell \boldsymbol{\phi}_\ell(x) = \int_0^1 \boldsymbol{R}(x,y)\boldsymbol{\phi}_\ell(y)dy, \quad 1 \leq \ell < \infty,$$

where the $\boldsymbol{\phi}_\ell$'s are orthonormal functions (defined on $[0,1]$, with values in R^K).

Approximation:

$$\sum_{\ell=1}^{\infty} \lambda_\ell Z_\ell^2 \approx \sum_{\ell=1}^{d} \lambda_\ell Z_\ell^2 \quad \text{where } d \text{ is sufficiently large}$$

Issue: $\boldsymbol{R}$ and therefore $\lambda_1, \lambda_2, \ldots$ are unknown

Critical values for the Cramér–von Mises statistic

Let $\hat{\boldsymbol{R}}_N(x,y)$ be a consistent estimator for $\boldsymbol{R}$ under H_0, i.e.

$$\int_0^1\int_0^1 \|\hat{\boldsymbol{R}}_N(x,y)-\boldsymbol{R}(x,y)\|^2 dxdy \quad \xrightarrow{P} \quad 0.$$

If $\hat{\lambda}_{1,N} \geq \hat{\lambda}_{2,N} \ldots$ satisfy

$$\hat{\lambda}_{\ell,N}\hat{\boldsymbol{\phi}}_{\ell,N}(x) = \int_0^1 \hat{\boldsymbol{R}}_N(x,y)\hat{\boldsymbol{\phi}}_{\ell,N}(y)dy, \quad 1 \leq \ell < \infty,$$

then under H_0 we have

$$\hat{\lambda}_{\ell,N} \quad \xrightarrow{P} \quad \lambda_\ell \ \text{ for all } \ell.$$

Hence under H_0

$$\sum_{\ell=1}^{d} \lambda_\ell Z_\ell^2 \approx \sum_{\ell=1}^{d} \hat{\lambda}_{\ell,N} Z_\ell^2$$

Construction of $\hat{\boldsymbol{R}}_N(x,y)$: if $i_{m-1} < \lfloor Nx \rfloor \leq i_m$, then

$$\sum_{i=1}^{\lfloor Nx \rfloor} \boldsymbol{z}_i = \sum_{j=1}^{m-1} \sum_{\ell=i_{j-1}+1}^{i_j} \boldsymbol{z}_i + \sum_{i=i_{m-1}+1}^{\lfloor Nx \rfloor} \boldsymbol{z}_i$$

so

$$E\left(\sum_{i=1}^{\lfloor Nx \rfloor} \boldsymbol{z}_i\right)\left(\sum_{i=1}^{\lfloor Nx \rfloor} \boldsymbol{z}_i\right)^\top \approx \sum_{j=1}^{m-1}(i_j - i_{j-1})\boldsymbol{D}_j + (\lfloor Nx \rfloor - i_{m-1})\boldsymbol{D}_m,$$

where $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_M$ are log run covariance matrices. Similar formula for the covariances

kernel type estimators for $\boldsymbol{D}_1, \boldsymbol{D}_2, \ldots, \boldsymbol{D}_M$ can be used to get $\hat{\boldsymbol{R}}_N(x,y)$

The behavior of critical values for the Cramér–von Mises statistic under H_A

Observation: there is $\boldsymbol{R}^*(x,y)$ such that

$$\int_0^1\int_0^1 \|\hat{\boldsymbol{R}}_N(x,y) - \boldsymbol{R}^*(x,y)\|^2 dxdy \quad \xrightarrow{P} \quad 0$$

and therefore

$$\hat{\lambda}_{i,N} \xrightarrow{P} \lambda_i^*,$$

where $\lambda_1^* \geq \lambda_2^* \geq \ldots$ are the eigenvalues of $\boldsymbol{R}^*(x,y)$. Hence under H_0

$$\sum_{\ell=1}^{d} \lambda_\ell Z_\ell^2 \approx \sum_{\ell=1}^{d} \hat{\lambda}_{\ell,N}^* Z_\ell^2.$$

Theorem 2 If there is a change in the $\boldsymbol{\mu}_i$'s (means), then

$$\frac{1}{N}\mathcal{C}_N \quad \xrightarrow{P} \quad c_0 > 0.$$

The test is consistent

Functional approach

Model

$$X_i(t) = \mu_i(t) + \eta_i(t), \quad \text{with} \quad \mu_i(t) = \sum_{k=1}^{K} \mu_{i,k} f_k(t) \quad \text{and} \quad \eta_i(t) = \sum_{k=1}^{K} b_{i,k} f_k(t) + \epsilon_i(t)$$

H_0 $\quad \mu_1(t) = \mu_2(t) = \ldots = \mu_N(t)$ (in L^2 sense)

Functional CUSUM

$$\alpha_N(x,t) = N^{-1/2} \left(\sum_{i=1}^{\lfloor Nx \rfloor} X_i(t) - \frac{\lfloor Nx \rfloor}{N} \sum_{i=1}^{N} X_i(t) \right).$$

Theorem 3 If H_0 and the Bernoulli assumption hold, then we can define a sequence of Gaussian processes $\Gamma_N^0(x,t)$ such that

$$\int_0^1 \int_0^1 (\alpha_N(x,t) - \Gamma_N^0(x,t))^2 dxdt \quad \xrightarrow{P} \quad 0$$

$E\Gamma_N^0(x,t) = 0$, $E\Gamma_N^0(x,t)\Gamma_N^0(y,s) = U(x,y;t,s)$ =explicit formula.

Consequence:

$$\mathcal{V}_N = \int_0^1 \int_0^1 \alpha_N^2(x,t) dxdt \quad \xrightarrow{\mathcal{D}} \quad \int_0^1 \int_0^1 (\Gamma^0(x,t))^2 dxdt,$$

where $\Gamma^0(x,t)$ is Gaussian with $E\Gamma^0(x,t) = 0$, $E\Gamma^0(x,t)\Gamma^0(y,s) = U(x,y;t,s)$.

Karhunen–Loéve expansion

$$\int_0^1 \int_0^1 (\Gamma^0(x,t))^2 dxdt = \sum_{\ell=1}^{\infty} \lambda_j Z_j^2,$$

where $Z_1, Z_2, \ldots$ are standard normals, $\lambda_1 \geq \lambda_2 \geq \ldots$

$$\lambda_i \phi_i(x,t) = \int_0^1 \int_0^1 U(x,y,t,s)\phi_i(y,s)dyds, \quad 1 \leq i < \infty.$$

Estimate $U(x,y,t,s)$ with $\hat{U}_N(x,y,t,s)$ (estimation of M long run covariance functions) satisfying $\|\hat{U}_N - U\| \xrightarrow{P} 0$.
If $\hat{\lambda}_{1,N} \geq \hat{\lambda}_{2,N} \geq \ldots$ are the eigenvalues of $\hat{U}_N(x,y,t,s)$, then

$$\int_0^1 \int_0^1 (\Gamma^0(x,t))^2 dxdt \approx \sum_{\ell=1}^{d} \hat{\lambda}_{\ell,N} Z_\ell^2$$

Dynamic Nelson–Siegel model for yield curves

K=3,

$$f_1(t) = 1, \quad f_2(t,\lambda) = \frac{1-e^{-\lambda t}}{\lambda t} \text{ and } f_2(t,\lambda) = \frac{1-e^{-\lambda t}}{\lambda t} - e^{-\lambda t}$$

$\lambda = 3.59$ (Diebold and Li (2003) and Diebold and Rudebusch (2013))

Outcome

Table 3: Application of the test procedures to yield curves over the six sampling periods. We expect small P–Values in periods (1)–(4), large in periods (5)–(6).

Sampling Period	Sample Size	Method	Break Point	P–value
(1) 07/08/2008 – 11/28/2008	$N = 100$	ProjSim	yes	1.8%
		ProjEigen	yes	1.7%
		NFEigen	yes	0.0%
		ProjSim	no	99.7%
		ProjEigen	no	99.9%
		NFEigen	no	85.3%
(2) 03/20/2008 – 03/19/2009	$N = 250$	ProjSim	yes	0.2%
		ProjEigen	yes	0.1%
		NFEigen	yes	0.0%
		ProjSim	no	92.9%
		ProjEigen	no	92.5%
		NFEigen	no	5.1%
(3) 10/18/2005 - 03/14/2006	$N = 100$	ProjSim	yes	0.6%
		ProjEigen	yes	0.1%
		NFEigen	yes	0.0%
		ProjSim	no	57.6%
		ProjEigen	no	52.9%
		NFEigen	no	21.7%
(4) 06/30/2005 - 06/29/2006	$N = 250$	ProjSim	yes	0.2%
		ProjEigen	yes	0.0%
		NFEigen	yes	0.1%
		ProjSim	no	53.3%
		ProjEigen	no	47.7%
		NFEigen	no	73.0 %
(5) 06/05/2012 – 10/24/2012	$N = 100$	ProjSim	yes	14.6%
		ProjEigen	yes	10.8%
		NFEigen	yes	3.0%
		ProjSim	no	42.7%
		ProjEigen	no	37.6%
		NFEigen	no	29.1%
(6) 02/16/2012 – 02/14/2014	$N = 250$	ProjSim	yes	78.0%
		ProjEigen	yes	74.6%
		NFEigen	yes	50.4%
		ProjSim	no	70.3%
		ProjEigen	no	65.6%
		NFEigen	no	62.6%

Acknowledgements

The talk was based on joint research with

Patrick Bardsley (University of Utah)
István Berkes (Rényi Institute)
Piotr Kokoszka (Colorado State)
Gregory Rice (University of Waterloo)
Gabriel Young (Colorado State)