Tests of independence for functional observations

Piotr Kokoszka

Utah State University

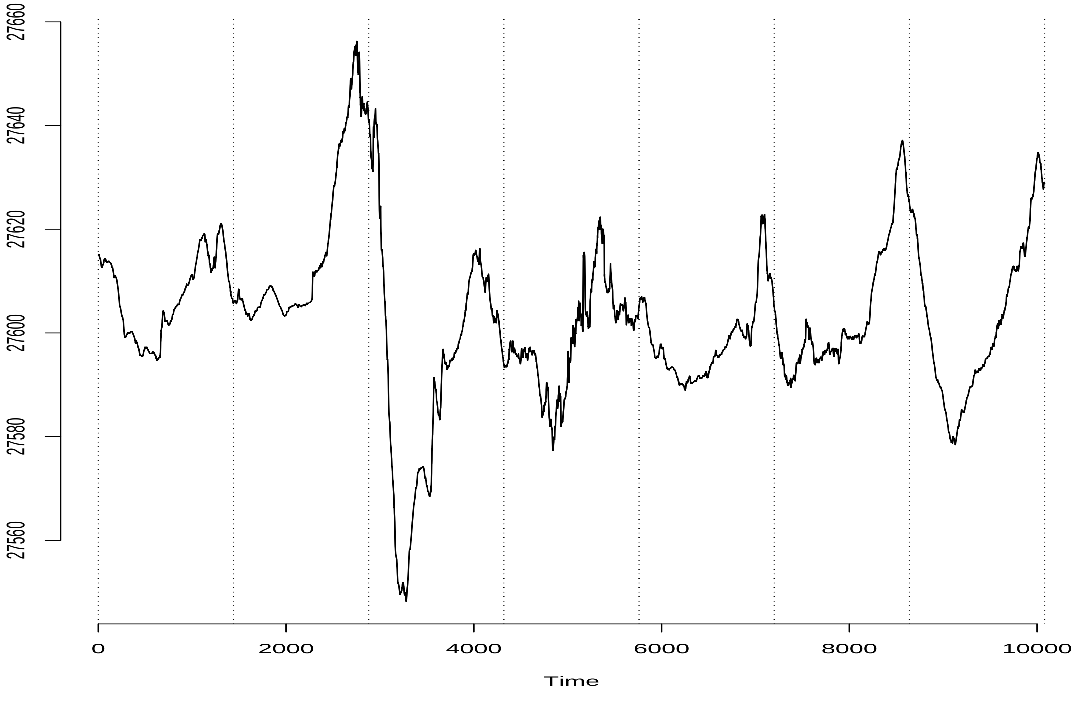
Joint work with

Robertas Gabrys and Inga Maslova

Seven functional observations

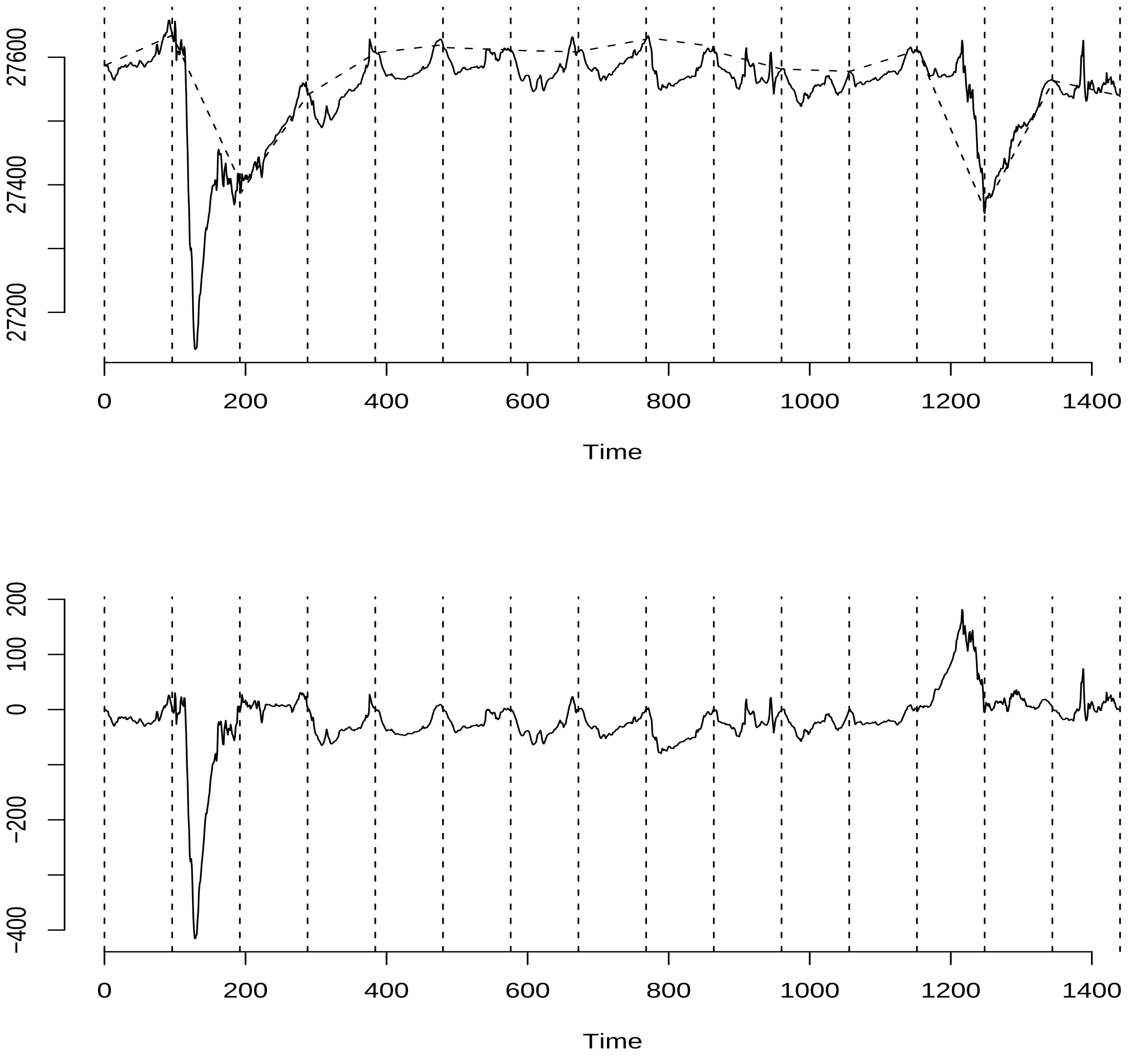
(1440 measurements per day)

The horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT.



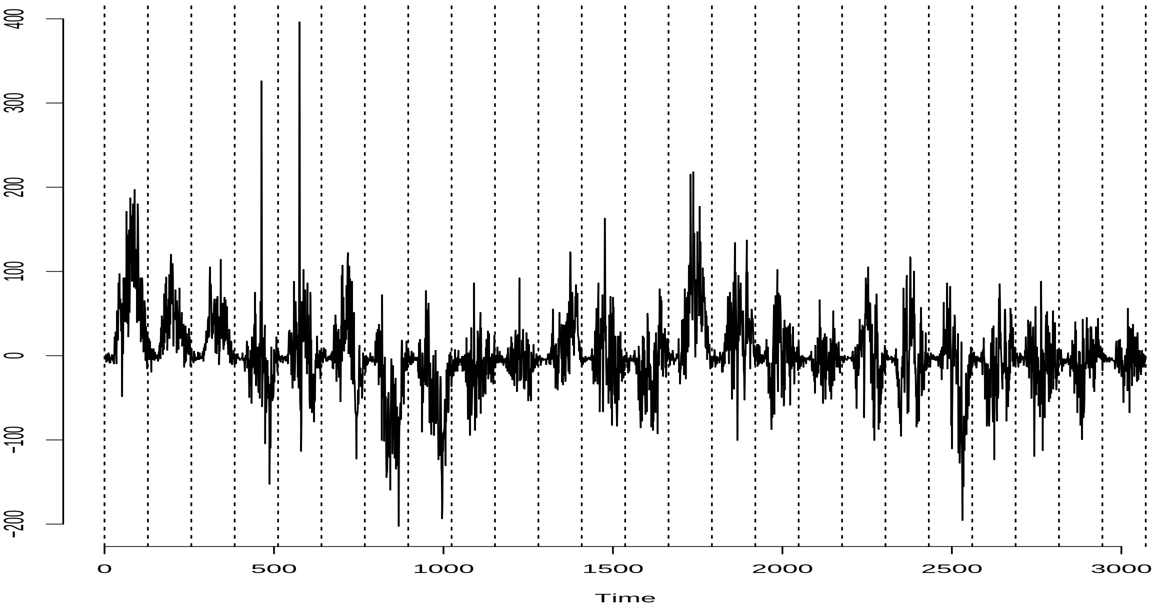
Usual tools of TSA not always useful

Top: horizontal intensity (nT) measured at Honolulu 30/3/2001 - 13/4/2001. Bottom: the same after subtracting the lines.



Functional observations often have a dependence structure

Three weeks of a time series derived from credit card transaction data. The vertical dotted lines separate days.



Why do we want to test independence and identical distribution of functional observations?

Most statistical procedures require a random sample

Errors of functional time series models should be iid random functions

Two settings:

1) Check if functions $X_1, X_2, \ldots, X_N$ derived from functional data are iid

2) If (X_i, Y_i) are iid functional random vectors satisfying

$$Y_i = \Psi X_i + \varepsilon_i$$

Test $H_0: \ \Psi = 0$
(no effect, Y_i independent of X_i)

Main idea of FDA: dimension reduction

Principal components:

We want to express each observation X_n as a linear combination of p functions v_k, $k = 1, 2, \ldots, p$.

$$X_n(t) \approx \sum_{k=1}^{p} X_{kn} v_k(t), \quad 1 \le n \le N.$$

The functions v_k (principal components) must be chosen so that most information is preserved by the approximation.

The random weights X_{kn} are called scores.

Idea: p is small, like 3 or 4. If $p = 4$, we replace 1440 measurements in day n by 4 scores $X_{1n}, X_{2n}, X_{3n}, X_{4n}$.

Idea of the test of independence and identical distribution of $X_1, X_2, \ldots, X_N$:

The (unknown) principal components v_k in

$$X_n(t) \approx \sum_{k=1}^{p} X_{kn} v_k(t), \quad 1 \leq n \leq N.$$

are deterministic orthonormal functions.

$$X_{kn} \approx \int X_n(t) v_k(t) dt.$$

If the functions X_n are iid, then the vectors

$$\mathbf{X}_n = [X_{1n}, X_{2n}, \ldots, X_{pN}]'$$

are also iid.

Feasible approach: Compute the vectors $\mathbf{X}_n$ using estimated principal components

$$v_{kN} = v_{kN}(X_1, X_2, \ldots, X_N).$$

The difference between v_k and v_{kN} is asymptotically negligible, and small in finite samples.

The test

$\boldsymbol{C}_h$ is the sample autocovariance matrix with entries

$$c_h(k,l) = \frac{1}{N} \sum_{t=1}^{N-h} X_{kt} X_{l,t+h}, \quad 0 \le h < N.$$

Denote by $r_{f,h}(i,j)$ and $r_{b,h}(i,j)$ the (i,j) entries of $\boldsymbol{C}_0^{-1}\boldsymbol{C}_h$ and $\boldsymbol{C}_h\boldsymbol{C}_0^{-1}$, respectively.

Test statistics:

$$Q_N = N \sum_{h=1}^{H} \sum_{i,j=1}^{p} r_{f,h}(i,j) r_{b,h}(i,j).$$

If the X_n are iid with finite fourth moment (in L^2), then

$$Q_N \xrightarrow{d} \chi^2_{p^2 H}.$$

Idea of consistency: If there is some dependence, there is $h \geq 1$ and coordinates k and l such that

$$c_h(k,l) = \frac{1}{N} \sum_{t=1}^{N-h} X_{kt} X_{l,t+h} \xrightarrow{P} K(k,l) \neq 0.$$

Then

$$N^{-1} Q_N = \sum_{h=1}^{H} \sum_{i,j=1}^{p} r_{f,h}(i,j) r_{b,h}(i,j) \xrightarrow{P} K > 0.$$

Example: Functional AR(1) process (ARH(1))

$$X_{n+1} = \Psi X_n + \varepsilon_n$$

Empirical size and power

Empirical size; Brownian bridges.

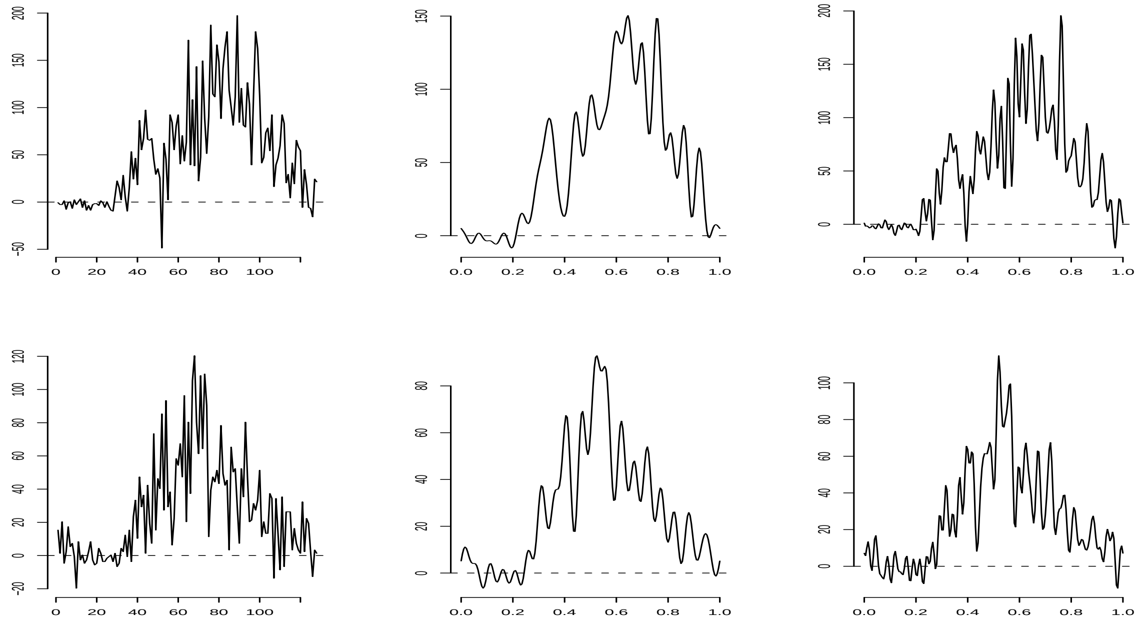
Lag	p=3			p=4			p=5		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
N=50									
1	7.7	2.5	0.6	7.4	2.8	0.3	7.9	3.5	0.4
3	6.8	2.5	0.3	6.7	3.3	0.6	4.9	2.0	0.3
5	4.9	2.0	0.0	3.6	1.4	0.2	4.0	1.7	0.2
N=100									
1	9.0	5.1	0.4	8.9	3.9	0.6	10.0	3.9	0.8
3	8.1	3.5	0.6	8.3	4.0	0.9	7.5	3.2	0.4
5	8.8	3.6	0.6	6.6	2.7	0.3	6.7	2.4	0.3
N=300									
1	9.8	4.6	1.2	9.4	4.0	0.9	10.1	4.7	0.6
3	9.3	4.8	1.0	9.1	4.7	0.9	10.0	5.4	0.8
5	7.2	3.7	1.0	8.2	3.8	0.7	10.6	5.5	1.2

Empirical power against ARH(1), $||\Psi|| = 0.5$.

Lag	p=3			p=4			p=5		
	10%	5%	1%	10%	5%	1%	10%	5%	1%
N=50									
1	44.7	33.8	17.7	41.9	29.4	12.6	38.5	26.1	9.2
3	35.2	27.0	13.3	34.0	24.7	10.8	33.2	21.6	8.7
5	26.7	20.0	11.0	24.4	15.8	8.1	21.5	14.3	6.0
N=100									
1	71.2	64.2	51.4	74.4	66.5	48.1	77.7	68.0	46.1
3	67.9	61.0	44.9	67.5	58.6	42.8	68.4	56.9	38.1
5	62.3	54.6	38.6	59.0	49.9	32.3	55.1	45.5	27.9
N=300									
1	98.7	98.2	96.7	99.2	98.9	97.2	99.8	99.5	98.5
3	97.6	97.1	95.5	99.0	98.4	96.8	99.2	98.3	96.6
5	96.8	95.9	92.8	98.1	97.0	93.8	98.4	97.3	94.4

Application to credit card transactions

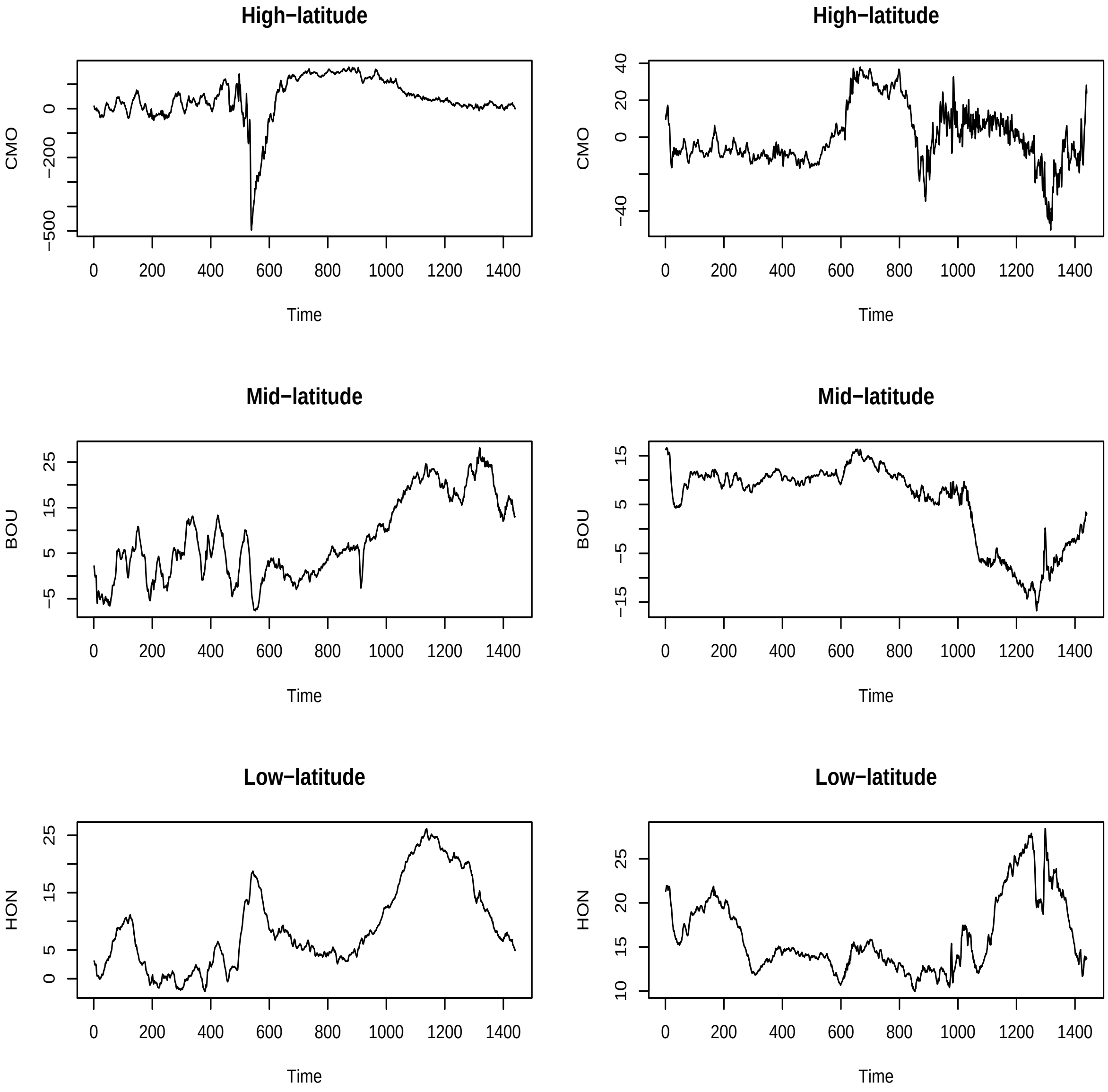
Two functional observations derived from the credit card transactions (left–most panel) together with smooths obtained by projection on 40 and 80 Fourier basis functions.



P-values for the functional AR(1) residuals of the credit card data.

Lag, H	p=1	p=2	p=3	p=4	p=5	p=6	p=7
BF=40							
1	69.54	22.03	13.60	46.29	80.35	96.70	99.20
2	35.57	38.28	7.75	47.16	64.92	95.00	99.04
3	54.44	53.63	25.28	52.61	71.33	86.84	94.93
BF=80							
1	57.42	18.35	53.30	89.90	88.33	95.40	99.19
2	35.97	23.25	23.83	45.07	55.79	46.39	70.65
3	36.16	36.02	26.79	30.21	56.81	34.51	47.00

Testing for linear effect



Horizontal intensities of the magnetic field measured at a high-, mid- and low-latitude stations during a sub-storm (left column) and a quiet day (right column). Note the different vertical scales for high-latitude records.

Functional Linear Model:

$$Y_n = \Psi X_n + \varepsilon_n, \quad n = 1, 2, \ldots, N.$$

$$H_0 : \ \Psi = 0$$

Introduce the operators:

$$\Gamma x = \mathsf{E}[\langle X_1, x\rangle \, X_1], \quad \Lambda x = \mathsf{E}[\langle Y_1, x\rangle \, Y_1],$$

$$\Delta x = \mathsf{E}[\langle X_1, x\rangle \, Y_1].$$

Denote their empirical counterparts by $\Gamma_N, \Lambda_N, \Delta_N$, e.g.

$$\Gamma_N x = \frac{1}{N} \sum_{n=1}^{N} \langle X_n, x\rangle \, X_n.$$

Define the eigenelements by

$$\Gamma v_k = \gamma_k v_k, \qquad \Lambda u_j = \lambda_j u_j.$$

Emprical eigenelements:

$$(\widehat{\gamma}_k, \widehat{v}_k), (\widehat{\lambda}_j, \widehat{u}_j)$$

.

$$\Delta = \Psi\Gamma$$

implies

$$\Psi v_k = \lambda_k^{-1}\Delta v_k$$

Ψ vanishes on sp$\{v_1, \ldots, v_p\}$

if and only if

$\Delta v_k = 0$ for each $k = 1, \ldots, p$.

$\Delta v_k \approx \Delta_N v_k = \frac{1}{N}\sum_{n=1}^{N} \langle X_n, v_k \rangle Y_n$.

Δv_k is practically contained in

sp$\{Y_1, \ldots, Y_N\} \approx$ sp$\{u_1, \ldots, u_q\}$

Ψ practically vanishes if

$\left\langle \Delta_N v_k, u_j \right\rangle = 0, \quad k = 1, \ldots, p,\ j = 1, \ldots, q.$

Ψ practically vanishes if

$$\left\langle \Delta_N v_k, u_j \right\rangle = 0, \quad k = 1, \ldots, p,\ j = 1, \ldots, q.$$

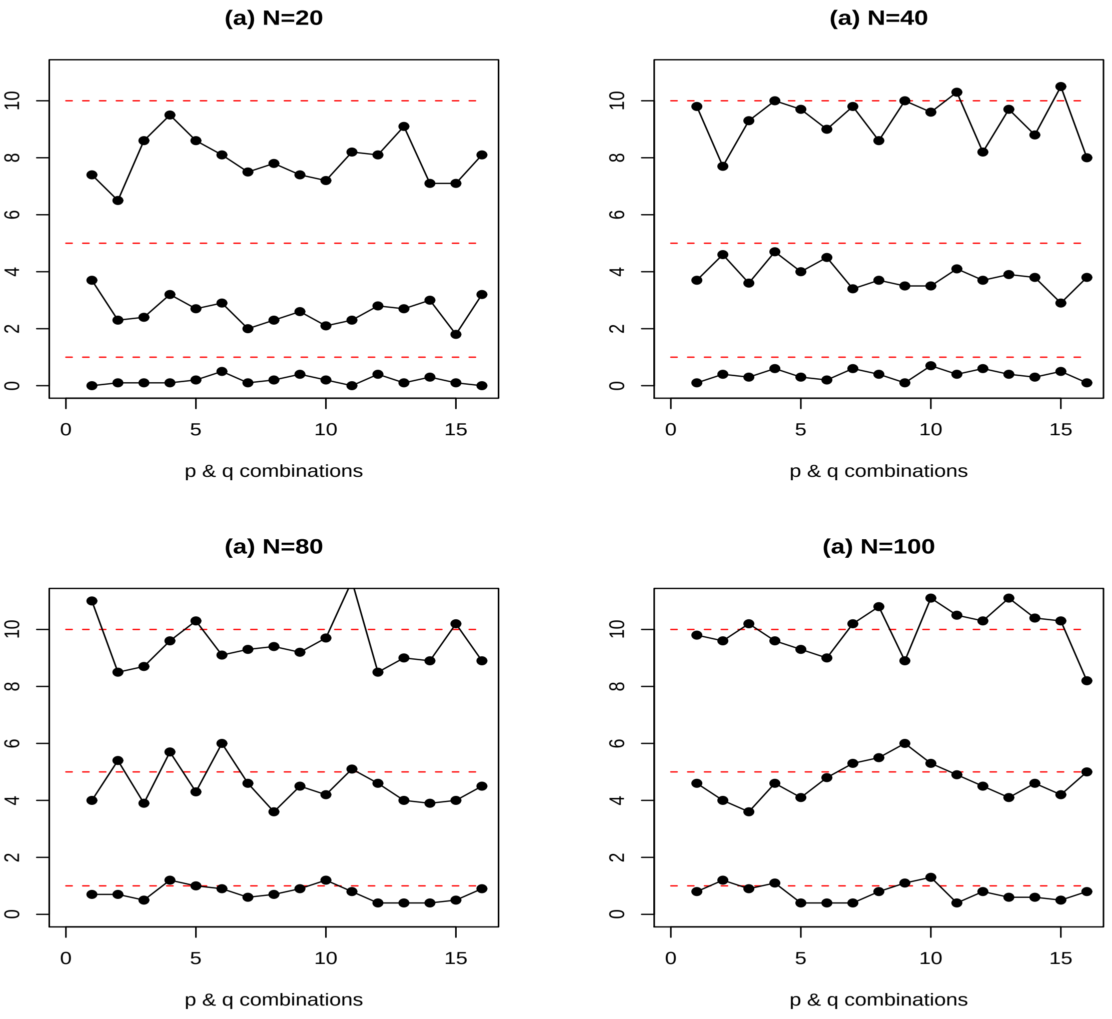
$$\widehat{T}_N(p,q) = N \sum_{k=1}^{p} \sum_{j=1}^{q} \widehat{\gamma}_k^{-1} \widehat{\lambda}_j^{-1} \left\langle \Delta_N \widehat{v}_k, \widehat{u}_j \right\rangle^2$$

Under H_0 ($\Psi = 0$)

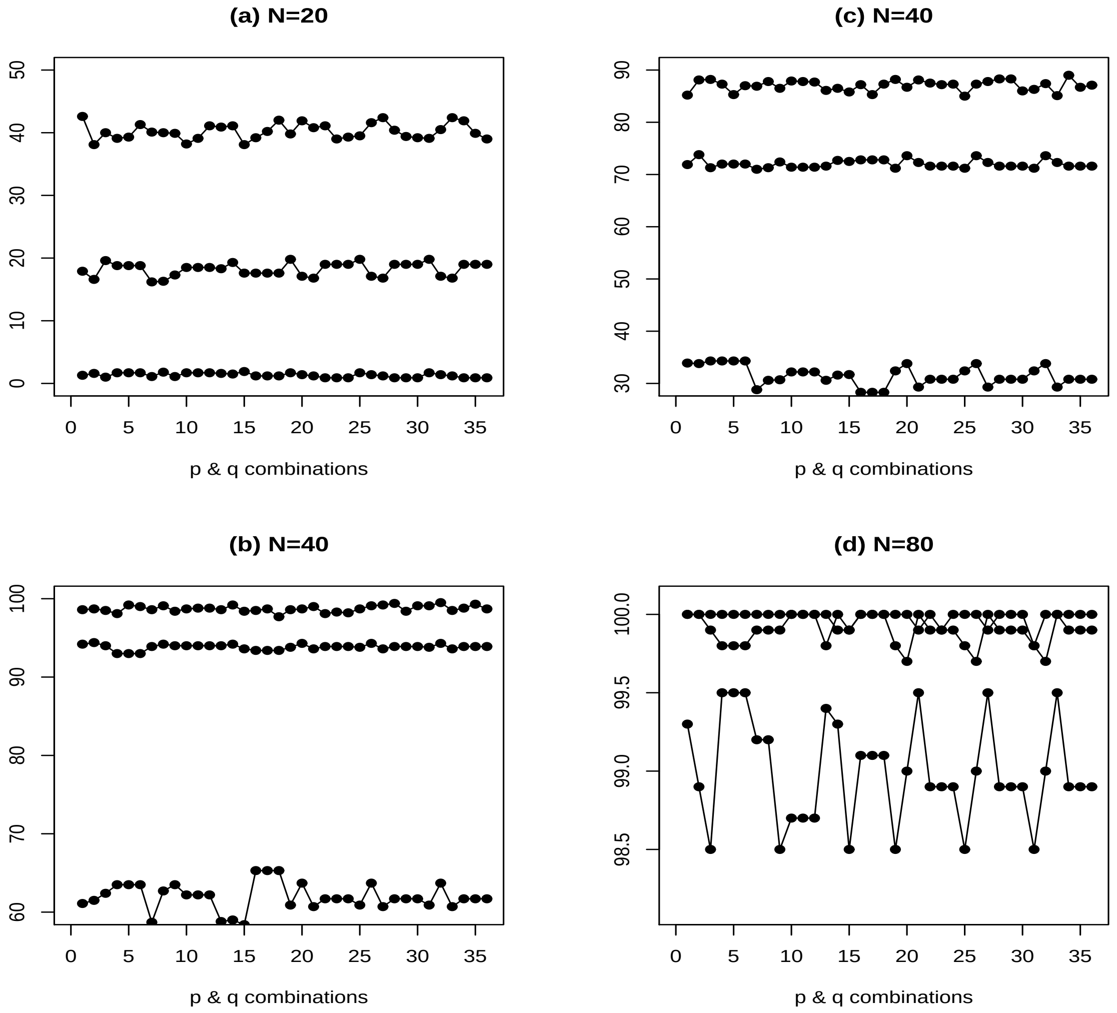
$$\widehat{T}_N(p,q) \xrightarrow{d} \chi^2_{pq}.$$

Under H_A, there are p, q such that

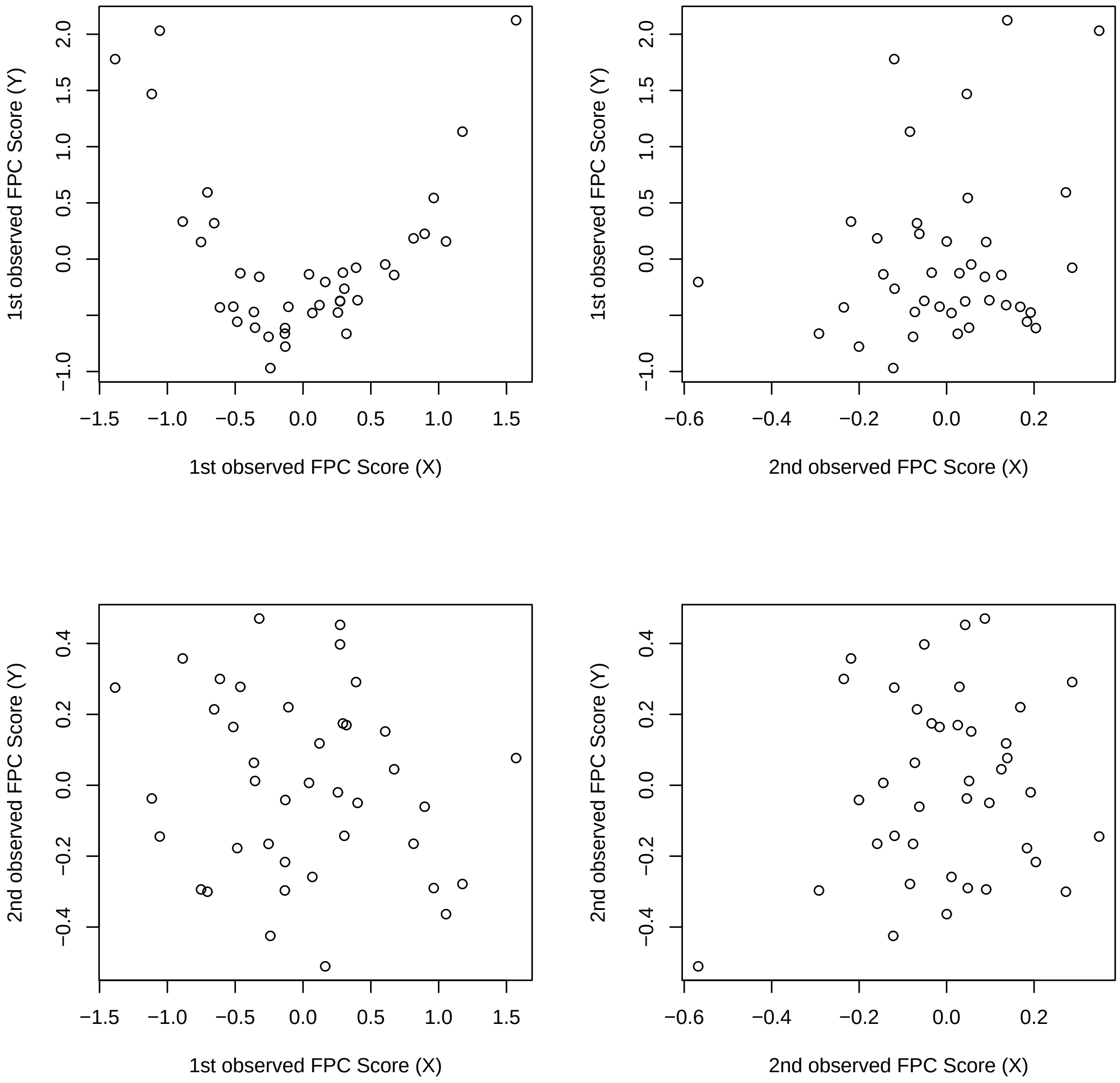
$$\widehat{T}_N(p,q) \xrightarrow{P} \infty$$



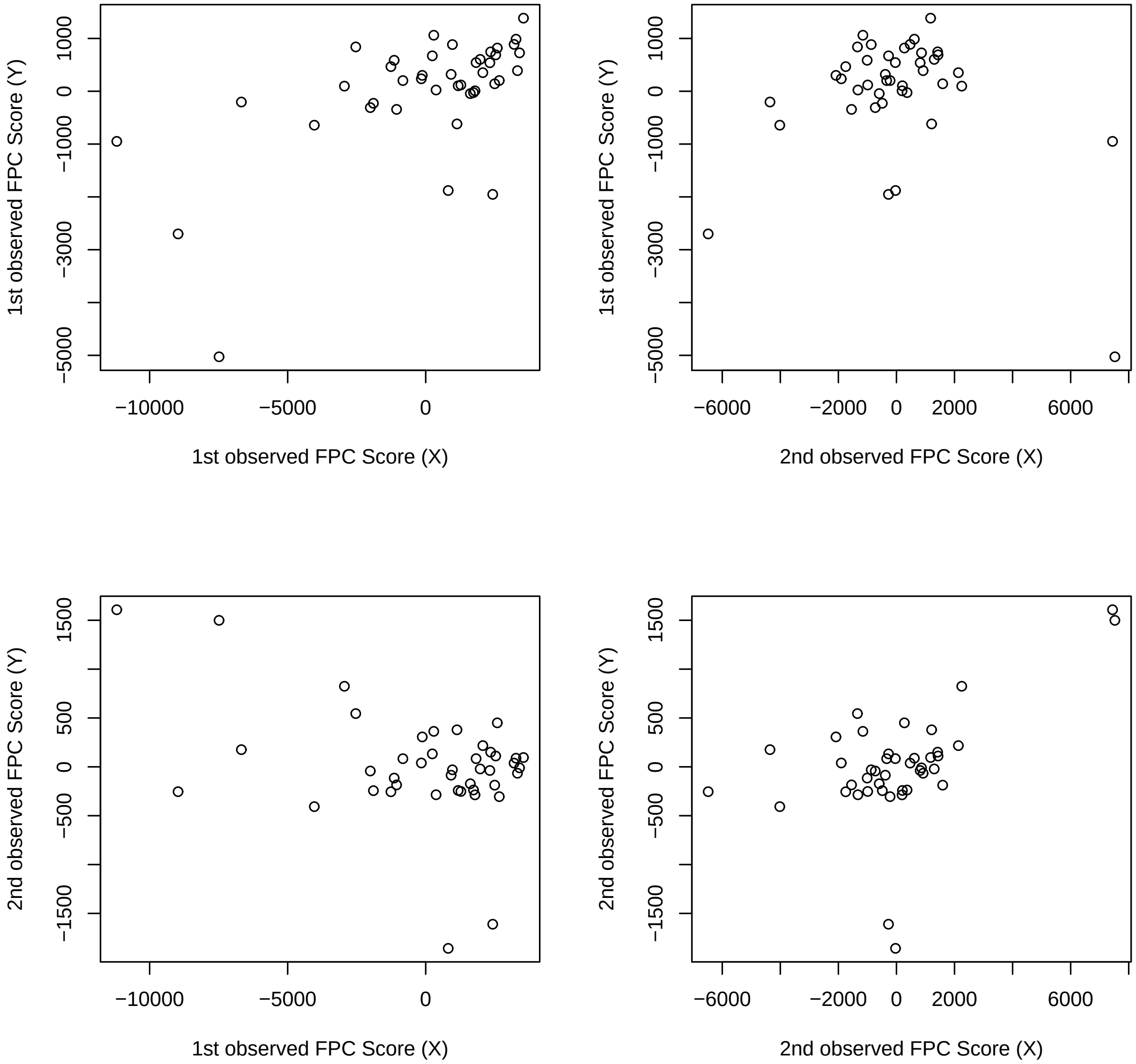
Empirical size of the test for $\alpha = 1\%, 5\%, 10\%$ (indicated by dotted lines) for different combinations of p and q. Here ε_n and Y_n, $n = 1, 2, \ldots, N$ are two independent Brownian Bridges.



Empirical power of the test for different combinations of principal components and different sample sizes N. Here X_n and ε_n are Brownian Bridges. In panels (a), (b) $||\Psi|| = 0.75$; in panels (c), (d) $||\Psi|| = 0.5$.



Functional predictor-response plots of FPC scores of response functions versus FPC scores of predictor functions for $Y_n(t) = H_2(X_n(t)) + \varepsilon_n(t)$, where $H_2(x) = x^2 - 1$, $n = 1, \ldots, 40$.

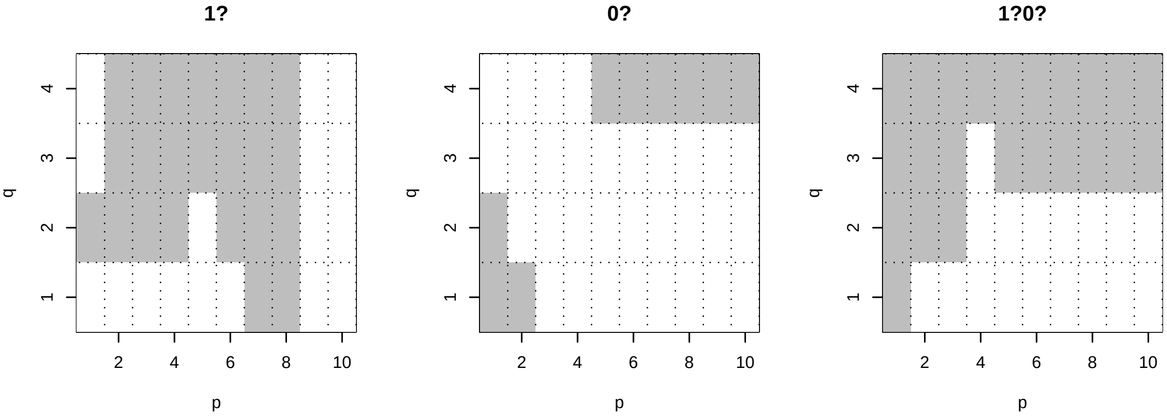


Functional predictor-response plots of FPC scores of response functions versus FPC of explanatory functions for magnetometer data (CMO vs THY0)

Results of the test for (lagged) sub-storm days in March–May, 2001; 0 indicates acceptance, 1 rejection. Number following station code denotes lag in days.

CMO							
BOU0	BOU1	BOU2	BOU3	HON0	HON1	HON2	HON3
1	1	0	0	1	1	0	0
THL							
FRD0	FRD1	FRD2	FRD3	SJG0	SJG1	SJG2	SJG3
1	1	0	0	1	1	0	0
ABK							
THY0	THY1	THY2	THY3	HER0	HER1	HER2	HER3
1	1	0	0	1?	1?	0	0
IRT							
MMB0	MMB1	MMB2	MMB3	KAK0	KAK1	KAK2	KAK3
1	1	0?	0	1	1	0?	0

Examples of rejection/acceptance plots at 5% level which are difficult to interpret. Grey area – reject H_0, white – fail to reject H_0.


1?
0?
1?0?
q
1
2
3
4
p
2
4
6
8
10