A note on estimation in Hilbertian linear models

Łukasz Kidziński

Département de Mathématique
Université Libre de Bruxelles

Based on joint work with Siegfried Hörmann (Université Libre de Bruxelles).

Outline

- Some examples
- The functional regression
- Weakly dependent setup
- PCA based estimation
- A data-driven order selection
- Simulations

Geophysical data

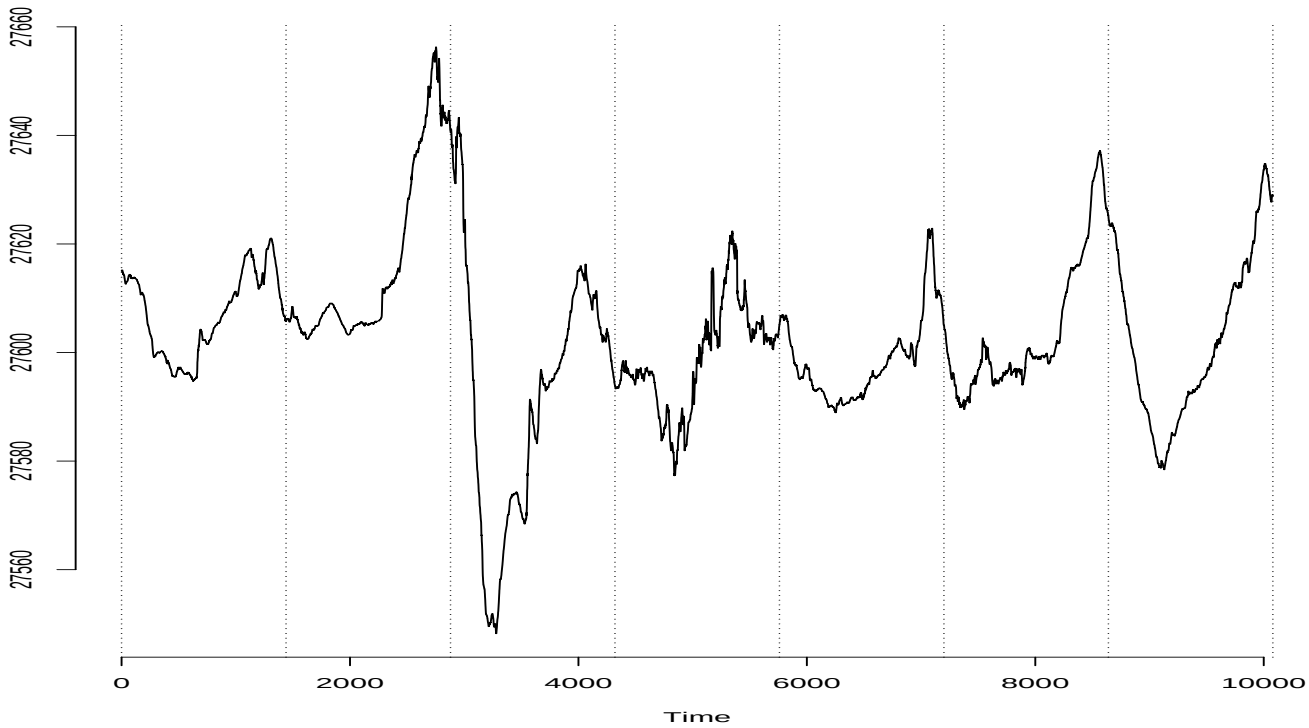

Figure 1: Horizontal component of the magnetic field measured in one minute resolution at Honolulu magnetic observatory from 1/1/2001 00:00 UT to 1/7/2001 24:00 UT. 1440 measurements per day.

'Tick'-data

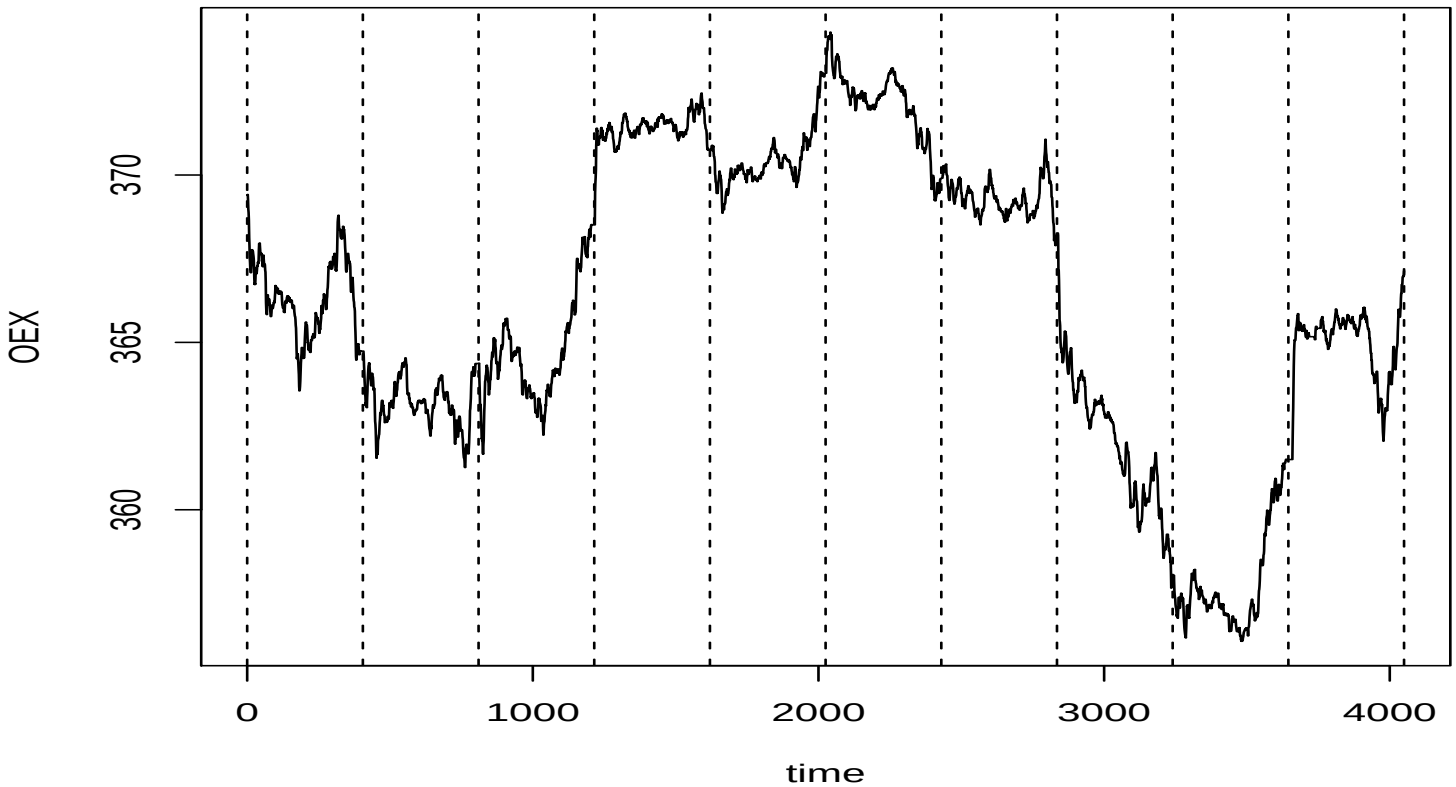

Figure 2: S&P100 market index plotted for 10 consecutive trading days. 405 minutely measurements per day.

Functional time series

The previous examples can be casted in the framework of functional time series.

Every observation $X_n = \{X_n(t),\ t \in [0,1]\}$ is a curve.

Usually the intraday process $\{X_n(t),\ t \in [0,1]\}$ is not stationary, while the process $\{X_n,\ n \geq 1\}$ is stationary.

Objective of this talk: Discuss the regression problem

$$Y_n = \Psi(X_n) + \varepsilon_n$$

when the $\{X_n\}$ process forms a functional time series.

Special case: the FAR(1) model

One of the most widely emloyed time series models is the functional AR (FAR) model, studied intensively by Bosq (2000).

$$X_{n+1}(t) = \int_0^1 \psi(t,s) X_n(s) ds + \varepsilon_{n+1}(t).$$

More conveniently

$$X_{n+1} = \Psi(X_n) + \varepsilon_{n+1}.$$

To forecast X_{n+1} we may set

$$\hat{X}_{n+1} = \hat{\Psi}(X_n),$$

and thus we need accurate estimator for Ψ.

Setup

We consider estimation of the operator Ψ when

$$Y_n = \Psi(X_n) + \varepsilon_n$$

for *fully observed data*. Let us collect our main assumptions.

- $\Psi : H_1 \to H_2$ is a bounded linear operator.
- $\{\varepsilon_k\}$ and $\{X_k\}$ are zero mean variables and are assumed to be L^4–m–approximable (see below).
- In addition $\{\varepsilon_k\}$ is Hilbertian white noise. For any $k \geq 1$ we have $X_k \perp \varepsilon_k$.

Weak dependence

A random sequence $\{X_n\}_{n\geq 1}$ with values in H is called L^p–m–approximable if it can be represented as

$$X_n = f(\delta_n, \delta_{n-1}, \delta_{n-2}, \ldots)$$

where the δ_i are iid elements taking values in a measurable space S and f is a measurable function $f : S^\infty \to H$. Moreover if δ_i' are independent copies of δ_i defined on the same probability space, then for

$$X_n^{(m)} = f(\delta_n, \delta_{n-1}, \delta_{n-2}, \ldots, \delta_{n-m+1}, \delta'_{n-m}, \delta'_{n-m-1}, \ldots)$$

we have

$$\sum_{m=1}^{\infty} (E\|X_m - X_m^{(m)}\|^p)^{1/p} < \infty.$$

Remarks

- L^p–m–approximability implicitely assumes stationarity.
- Trivial example: iid sequences.
- Setup can be shown to cover the FAR(1).
- Setup also covers many linear and non-linear processes.
- One can show that

$$\sqrt{n}\|\bar{X} - \mu\|_{H_1} = O_P(1).$$

Thus the mean can be accurately removed in a preprocessing step and that $EX_k = 0$ is not a stringent assumption. Similar argument works for intercept.

Estimation of Ψ

We define the covariance operator

$$C = E[X_1 \otimes X_1] \quad (\Longrightarrow C(x) = E[\langle X_1, x \rangle X_1] \,)$$

and the cross-covariance operator

$$\Delta = E[X_1 \otimes Y_1] \quad (\Longrightarrow \Delta(x) = E[\langle X_1, x \rangle \, Y_1] \,).$$

By framework assumptions both of them are Hilbert-Schmidt operators.

Let $(\lambda_i, v_i)_{i \geq 1}$ be the eigenvalues and corresponding eigenfunctions of the operator C, such that $\lambda_1 \geq \lambda_2 \geq ....$ The eigenfunctions are orthonormal and those belonging to a non-zero eigenvalue form an orthonormal basis of $C(H_1)$.

Estimation of Ψ

Using linearity of Ψ and the requirement $X_k \perp \varepsilon_k$ we obtain

$$\begin{aligned}\Delta(v_j) &= E\langle X_1, v_j\rangle Y_1 \\ &= E\langle X_1, v_j\rangle \Psi(X_1) + E\langle X_1, v_j\rangle \varepsilon_1 \\ &= E\Psi(\langle X_1, v_j\rangle X_1) \\ &= \Psi(C(v_j)) \\ &= \lambda_j \Psi(v_j).\end{aligned}$$

Then for any $x \in H_1$ the derived equation leads us formally to

$$\Psi(x) = \Psi\left(\sum_{j=1}^{\infty}\langle v_j, x\rangle v_j\right) = \sum_{j=1}^{\infty}\frac{\Delta(v_j)}{\lambda_j}\langle v_j, x\rangle. \qquad (1)$$

Estimation of Ψ

We will estimate Δ, v_j and λ_j from our sample

$$X_1, \ldots, X_n, Y_1, \ldots, Y_n$$

and substitute the estimators into formula (1).

The estimated eigenelements are obtained from empirical covariance operator

$$\hat{C}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \otimes X_k.$$

In a similar straightforward manner we set

$$\hat{\Delta}_n = \frac{1}{n} \sum_{k=1}^{n} X_k \otimes Y_k.$$

Estimation of Ψ

Apparently, from the finite sample we cannot estimate the entire sequence (λ_j, v_j), rather we have to work with a truncated version. This leads to

$$\hat{\Psi}_K(x) = \sum_{j=1}^{K} \frac{\hat{\Delta}(\hat{v}_j)}{\hat{\lambda}_j} \langle \hat{v}_j, x \rangle, \tag{2}$$

where the choice of $K = K_n$ is crucial.

- K_n has to grow with the sample size to infinity for consistency.
- Since $\lambda_j \to 0$ it will be a delicate issue to control the behavior of $\frac{1}{\hat{\lambda}_j}$.

A practical possibility is to use cross-validation.

For theoretical purposes it may be interesting to have a choice $K = K_n \to \infty$ such that $\hat{\Psi}_K \to \Psi$.

This has been established in special cases (e.g. Bosq for FAR(1)) but requires delicate assumptions:

- Ψ is Hilbert-Schmidt (excludes $\Psi = \mathrm{Id}$).
- Specific decay of eigenvalues and gaps

$$\lambda_j \quad \text{and} \quad \alpha_j = \min\{\lambda_j - \lambda_{j+1}, \lambda_{j-1} - \lambda_j\}$$

  which are impossible to check.

A data-driven selection of K.

Theorem: If we impose the following:

(A): Ψ is Hilbert-Schmidt and λ_j are distinct.

(B): Let $K_n = \min(B_n, E_n, m_n)$ where

$$B_n = \arg\max\{j \geq 1 | \frac{1}{\hat{\lambda}_j} \leq m_n\}$$

and

$$E_n = \arg\max\{k \geq 1 | \max_{1 \leq j \leq k} \frac{1}{\hat{\alpha}_j} \leq m_n\}$$

for some sequence $\{m_n\}$ such $m_n^6 = o(n)$.

Then $\|\hat{\Psi} - \Psi\| \to 0$ in probability.

A data-driven selection of K.

The technical Assumption **(A)** appears still unsatisfactory.
Unfortunately it cannot be completely avoided.

Assume e.g. that

$$\Psi = \mathrm{Id} = \sum_{k \geq 1} v_k \otimes v_k.$$

Even if we perfectly estimate the first K terms of Ψ by

$$\hat{\Psi}_K = \sum_{k=1}^{K} v_k \otimes v_k,$$

we have that

$$\|\Psi - \hat{\Psi}_K\| = 1.$$

A data-driven selection of K.

A way to overcome such difficulties it to argue that in practice we will be satisfied if the estimator $\hat{\Psi}$ is such that $\|\Psi(X) - \hat{\Psi}(X)\|$ is small if $X \stackrel{d}{=} X_1$.

Theorem: We define the same estimator $\hat{\Psi}$ as before with $K_n = \arg\max\{j \geq 1 | \frac{\hat{\lambda}_1}{\hat{\lambda}_j} \leq m_n\}$, where $m_n = o(\sqrt{n})$. Then

$$\|\Psi(X) - \hat{\Psi}(X)\| \to 0,$$

in probability.

Simuation

Let $H_1 = H_2 = \mathrm{span}\{v_j : 0 \leq j \leq 34\}$, where $v_k(t)$ are the first 35 elements of a Fourier basis on $[0, 1]$.

We work with Gaussian curves $X_i(t)$ by setting

$$X_i(t) = \sum_{j=1}^{35} A_i^{(j)} v_{j-1}(t), \tag{3}$$

where $(A_i^{(1)}, A_i^{(2)}, \ldots, A_i^{(35)})'$ are independent Gaussian random vectors with mean zero and covariance Σ.

We test three setups

- $\Lambda_1 : (1, e^{-1/5}, e^{-2/5}, \ldots, e^{-35/5})$ [fast decay],
- $\Lambda_2 : (1, \frac{34}{35}, \ldots, \frac{1}{35})$ [slow decay],
- $\Lambda_3 : (1, 1, \ldots, 1)$ [no decay].

Simuation

The noise $\{\varepsilon_k\}$ is also assumed to be of the form (3) with coefficients $\{A_i^{(j)},\ i,j \geq 0\}$ i.i.d. $\mathcal{N}(0,1)$.

Finally we used the following 3 operators:

- Ψ_1 identity,
- $\Psi_2 = \Gamma_1 + \Gamma_2$, such that $\Gamma_1 : v_i \mapsto \frac{2}{3} v_{\pi_i}$ and $\Gamma_2 : v_i \mapsto \frac{1}{3} v_{\pi'_i}$, where $\pi_i = 1 + (i + 4 \mod 35)$ and $\pi'_i = 1 + (i \mod 35)$,
- $\Psi_3(x) = \sum_{i=1}^{35} \sum_{j=1}^{35} \psi_{ij} \langle x, v_i \rangle v_j$, where the coefficients ψ_{ij} have been generated as i.i.d. standard normal random variables (once generated, they were fixed for the entire simulation), normalized such that $\|\Psi_3\|_{\mathcal{L}_{12}} = 1$.

Simulation

As a performance measure for our procedure we

$$\texttt{NMSE} = \frac{\sum_{k=1}^{n} \|\Psi(X_k) - \hat{\Psi}(X_k)\|^2}{\sum_{k=1}^{n} \|\Psi(X_k)\|^2}.$$

We chose $m_n = \sqrt{n}$ with sample sizes $n = 10 \times 2^{\ell}$, $\ell = 0, \ldots, 11$.

Simulation

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
10	1	3.26	2	3.50	1	3.26	1	3.26
20	1	1.38	3	2.59	2	1.88	1	1.38
40	1	1.14	4	1.73	3	1.29	1	1.14
80	3	0.77	6	1.58	4	1.05	1	0.80
160	5	0.48	6	0.62	5	0.48	1	0.73
320	7	0.31	7	0.31	6	0.36	2	0.57
640	8	0.18	9	0.19	7	0.19	3	0.33
1280	11	0.11	9	0.11	8	0.12	4	0.25
2560	11	0.06	10	0.07	9	0.08	5	0.17
5120	15	0.03	11	0.04	9	0.05	6	0.10
10240	17	0.02	12	0.02	10	0.03	6	0.10
20480	17	0.01	13	0.01	11	0.02	7	0.06

Table 1: Ψ_1, Λ_1

Simulation

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
10	1	1.01	8	2	6	1.25	1	1.01
20	4	0.96	12	1.85	9	1.46	1	1.01
40	8	0.88	18	1.42	13	1.07	1	0.97
80	12	0.68	20	0.93	16	0.79	3	0.85
160	19	0.45	24	0.52	20	0.51	5	0.75
320	20	0.25	27	0.30	23	0.27	9	0.50
640	25	0.14	28	0.14	25	0.14	13	0.33
1280	29	0.07	30	0.08	27	0.08	17	0.19
2560	30	0.03	31	0.03	28	0.04	20	0.11
5120	35	0.02	31	0.02	30	0.02	23	0.06
10240	35	0.01	32	0.01	31	0.01	24	0.04
20480	34	0	33	0.01	31	0.01	26	0.02

Table 2: Ψ_1, Λ_2

Simulation

n	K_n^{OPT}	NMSE	K_n^1	NMSE	$K_n^{0.5}$	NMSE	$K_n^{0.1}$	NMSE
10	3	0.95	9	1.16	9	1.16	4	0.98
20	7	0.90	18	1.39	16	1.03	5	0.93
40	15	0.79	26	1.14	23	1	8	0.86
80	28	0.58	35	0.88	32	0.62	14	0.71
160	34	0.28	35	0.30	35	0.30	21	0.48
320	35	0.12	35	0.12	35	0.12	33	0.16
640	35	0.06	35	0.06	35	0.06	35	0.06
1280	35	0.03	35	0.03	35	0.03	35	0.03
2560	35	0.01	35	0.01	35	0.01	35	0.01
5120	35	0.01	35	0.01	35	0.01	35	0.01
10240	35	0	35	0	35	0	35	0
20480	35	0	35	0	35	0	35	0

Table 3: Ψ_1, Λ_3