Two applications of functional data

Matthieu Saumard

ATER au CNAM

Séminaire CNAM
18-11-2016

Outline

Outline

- This first part of the talk is already published.
- This is a joint work with Robbiano, S. and Curé, M. [4].

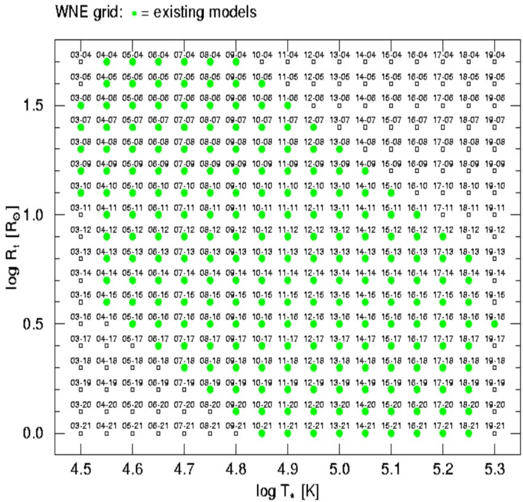
Problem

(Astrophysicist) Goal : determine the temperature (T) and the radius of a star (R)

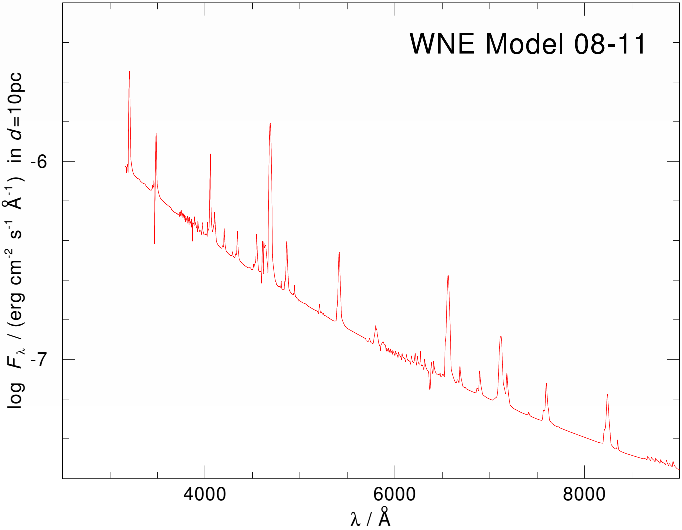
Popular method

- Create a physic model depending of T and R to build spectrum
- Generate a grid of simulated spectrum
- Compute the closest spectrum of the grid to the real data
- Say that they have the same T and R

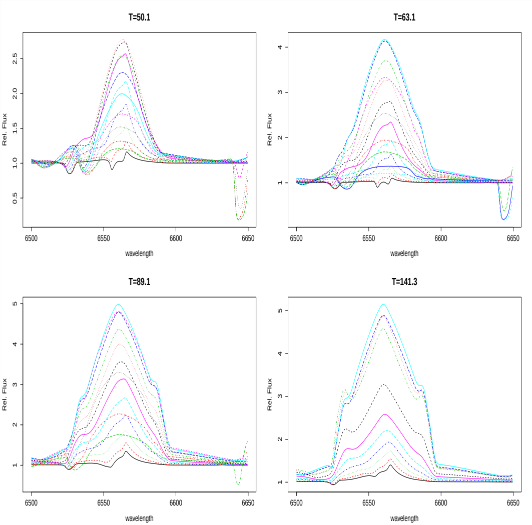
Grid

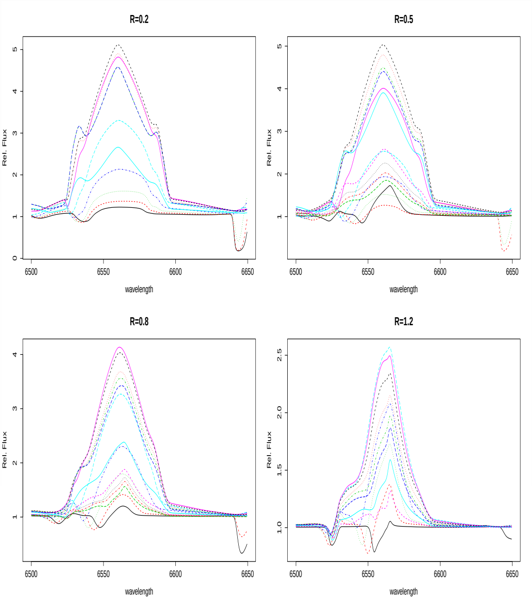

WNE grid: • = existing models
03-04 04-04 05-04 06-04 07-04 08-04 09-04 10-04 11-04 12-04 13-04 14-04 15-04 16-04 17-04 18-04 19-04
03-05 04-05 05-05 06-05 07-05 08-05 09-05 10-05 11-05 12-05 13-05 14-05 15-05 16-05 17-05 18-05 19-05
03-06 04-06 05-06 06-06 07-06 08-06 09-06 10-06 11-06 12-06 13-06 14-06 15-06 16-06 17-06 18-06 19-06
03-07 04-07 05-07 06-07 07-07 08-07 09-07 10-07 11-07 12-07 13-07 14-07 15-07 16-07 17-07 18-07 19-07
03-08 04-08 05-08 06-08 07-08 08-08 09-08 10-08 11-08 12-08 13-08 14-08 15-08 16-08 17-08 18-08 19-08
03-09 04-09 05-09 06-09 07-09 08-09 09-09 10-09 11-09 12-09 13-09 14-09 15-09 16-09 17-09 18-09 19-09
03-10 04-10 05-10 06-10 07-10 08-10 09-10 10-10 11-10 12-10 13-10 14-10 15-10 16-10 17-10 18-10 19-10
03-11 04-11 05-11 06-11 07-11 08-11 09-11 10-11 11-11 12-11 13-11 14-11 15-11 16-11 17-11 18-11 19-11
03-12 04-12 05-12 06-12 07-12 08-12 09-12 10-12 11-12 12-12 13-12 14-12 15-12 16-12 17-12 18-12 19-12
03-13 04-13 05-13 06-13 07-13 08-13 09-13 10-13 11-13 12-13 13-13 14-13 15-13 16-13 17-13 18-13 19-13
03-14 04-14 05-14 06-14 07-14 08-14 09-14 10-14 11-14 12-14 13-14 14-14 15-14 16-14 17-14 18-14 19-14
03-15 04-15 05-15 06-15 07-15 08-15 09-15 10-15 11-15 12-15 13-15 14-15 15-15 16-15 17-15 18-15 19-15
03-16 04-16 05-16 06-16 07-16 08-16 09-16 10-16 11-16 12-16 13-16 14-16 15-16 16-16 17-16 18-16 19-16
03-17 04-17 05-17 06-17 07-17 08-17 09-17 10-17 11-17 12-17 13-17 14-17 15-17 16-17 17-17 18-17 19-17
03-18 04-18 05-18 06-18 07-18 08-18 09-18 10-18 11-18 12-18 13-18 14-18 15-18 16-18 17-18 18-18 19-18
03-19 04-19 05-19 06-19 07-19 08-19 09-19 10-19 11-19 12-19 13-19 14-19 15-19 16-19 17-19 18-19 19-19
03-20 04-20 05-20 06-20 07-20 08-20 09-20 10-20 11-20 12-20 13-20 14-20 15-20 16-20 17-20 18-20 19-20
03-21 04-21 05-21 06-21 07-21 08-21 09-21 10-21 11-21 12-21 13-21 14-21 15-21 16-21 17-21 18-21 19-21
1.5
1.0
0.5
0.0
log R$_t$ [R$_\odot$]
4.5 4.6 4.7 4.8 4.9 5.0 5.1 5.2 5.3
log T$_*$ [K]

Example


WNE Model 08-11
log F_λ / (erg cm^{-2} s^{-1} Å^{-1}) in d=10pc
-6
-7
4000
6000
8000
λ / Å

Data


T=50.1
Rel. Flux
wavelength
T=63.1
Rel. Flux
wavelength
T=89.1
Rel. Flux
wavelength
T=141.3
Rel. Flux
wavelength


R=0.2
R=0.5
R=0.8
R=1.2
Rel. Flux
wavelength
6500
6550
6600
6650

Functional Model

Functional linear model

$$Y = \alpha_1 + \langle \beta, X \rangle + \varepsilon \tag{1.1}$$

Approximation of X

$$X(t) = \sum_{k=1}^{p} x_k \rho_k(t) + r_p,$$

where $x_k = \int_K X(t)\rho_k(t)dt$ and r_p is an error.

Approximation of β

$$\beta(t) = \sum_{k=1}^{p} b_k \rho_k(t) + rb_p$$

Functional Model multi-lines

$$\begin{cases} Y_i^{(1)} = \alpha_1 + \sum_{j=1}^{M} \langle \beta_{1,j}, X_{i,j} \rangle + \varepsilon_i \\ Y_i^{(2)} = \alpha_2 + \sum_{j=1}^{M} \langle \beta_{2,j}, X_{i,j} \rangle + \varepsilon_i' \end{cases} \tag{1.3}$$

$$X_{i,j} = \sum_{k=1}^{p} x_{ijk} \rho_k^j + r_{ijp},$$

where $x_{ijk} = \int_K X_{i,j}(t)\rho_k^j(t)dt$ and r_{ijp} is an error. We decompose the parameters $\beta_{1,j}$ and $\beta_{2,j}$ in the same basis with the same notation. Then, we have

$$\begin{cases} Y_i^{(1)} = \alpha_1 + \sum_{j=1}^{M} \sum_{k=1}^{p} \beta_{1jk} x_{ijk} + \epsilon_i \\ Y_i^{(2)} = \alpha_2 + \sum_{j=1}^{M} \sum_{k=1}^{p} \beta_{2jk} x_{ijk} + \epsilon_i'. \end{cases} \tag{1.4}$$

If we note $\boldsymbol{X_j}$ the matrix $(x_{ijk})_{i=1\ldots n,k=1,\ldots,p}$ we have

$$\begin{cases} \boldsymbol{Y^{(1)}} = \alpha_1 + \sum_{j=1}^{M} \boldsymbol{X_j} \boldsymbol{\beta_{1,j}} + \epsilon_i \\ \boldsymbol{Y^{(2)}} = \alpha_2 + \sum_{j=1}^{M} \boldsymbol{X_j} \boldsymbol{\beta_{2,j}} + \epsilon'_i. \end{cases} \quad (1.5)$$

Prediction Methods

- Problem : Ordinary least squares estimator is very unstable
- Robust linear regression
- Ridge regression

Others models

- Nonparametrics methods namely Ferraty et Vieu [1]
- Lasso
- Elastic net

Evaluation of the prediction

$$\begin{cases} \widehat{Y^{(1)}} = \widehat{\alpha}_1 + \sum_{j=1}^{M} \langle \widehat{\beta}_{1,j}, x_j \rangle \\ \widehat{Y^{(2)}} = \widehat{\alpha}_2 + \sum_{j=1}^{M} \langle \widehat{\beta}_{2,j}, x_j \rangle \end{cases}$$

$$RMSE(\beta_l) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} |Y_i^{(l)} - \hat{Y}_i^{(l)}|^2}.$$

$$\Gamma ME(\beta) = \sqrt{\frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i, Y_i - \hat{Y}_i)_\Gamma}.$$

Prediction procedure

- We split the database in 5 parts of equal size.
- We use 3 parts (the learning set) to compute the parameter β_1^p and β_2^p for several parameter p (in practice we choose $p = 1 + 4k$ for $k = 1, \ldots, 10$).
- We choose the parameter p that minimize the mean Γ error on the validation.
- We evaluate the prediction accuracy of our method on the last part of the data set (testing set).

1. (Input.) A training dataset $\mathcal{D} = (\mathbf{X}, \boldsymbol{Y})$ of size n, $\mathcal{A}$ a prediction algorithm, X_f an observation.

2. (Estimation) Compute $\widehat{\boldsymbol{Y}} = \mathbf{X}\widehat{\beta}$ using $\mathcal{D}$ and $\mathcal{A}$. Set $\widehat{r} = (\widehat{r}_1, \ldots, \widehat{r}_n) = \boldsymbol{Y} - \widehat{\boldsymbol{Y}}$ the residues.

3. (Iterations.) For $t = 1, \ldots, T$,
   1. (Sample.) Draw with replacement form $\widehat{r}$ a sample $r^* = (r_1^*, \ldots, r_n^*)$ and r_f^*. Create $\mathcal{D}^* = (\mathbf{X}, \boldsymbol{Y}^*)$ where $\boldsymbol{Y}^* = \mathbf{X}\widehat{\beta} + r^*$
   2. (Learning.) Use $\mathcal{A}$ and $\mathcal{D}^*$ to compute $\widehat{\beta}^*$. Keep $B_t^* = X_f\widehat{\beta} - X_f\widehat{\beta}^* + r_f^*$.

4. (Output) A prediction interval $\tilde{I} = [X_f\widehat{\beta} + q^*; X_f\widehat{\beta} + Q^*]$ where q^* (resp. Q^*) is the empirical quantile at level $\alpha/2$ (resp. $1 - \alpha/2$) of $B^* = (B_1^*, \ldots, B_T^*)$.

Results

Table: Comparison of methods with Fourier decomposition : prediction.

		LM		Robust LM		Ridge	
	Astro	Brut	Norm	Brut	Norm	Brut	Norm
$Y^{(1)}$	0.171	0.088	0.068	0.062	0.060	0.080	0.066
$Y^{(2)}$	0.143	0.089	0.071	0.060	0.066	0.075	0.081
$\|\|\|\,\|\|\|_{\Gamma}$	0.437	0.280	0.211	0.189	0.189	0.231	0.233
Nbase	–	7	5	7	7	7	7

Results

Table: Comparison of methods with spline decomposition : prediction.

		LM		Robust LM		Ridge	
	Astro	Brut	Norm	Brut	Norm	Brut	Norm
$Y^{(1)}$	0.171	0.119	0.116	0.116	0.120	0.079	0.060
$Y^{(2)}$	0.143	0.102	0.102	0.096	0.112	0.075	0.067
$\|\|\|\|_\Gamma$	0.437	0.322	0.315	0.302	0.317	0.235	0.200
Nbase	–	5	5	5	5	21	13

Table: Comparison of methods with Fourier decomposition : coverage probabilities at 95 % .

		LM		Robust LM		Ridge	
	Astro	Brut	Norm	Brut	Norm	Brut	Norm
Coverage $Y^{(1)}$	0.871	0.918	0.896	0.908	0.919	0.944	0.957
Coverage $Y^{(2)}$	0.910	0.941	0.966	0.915	0.916	0.951	0.970

Table: Comparison of methods with spline decomposition : coverage probabilities at 95 % .

		LM		Robust LM		Ridge	
	Astro	Brut	Norm	Brut	Norm	Brut	Norm
Coverage $Y^{(1)}$	0.871	0.923	0.919	0.928	0.909	0.986	0.980
Coverage $Y^{(2)}$	0.910	0.949	0.946	0.919	0.885	0.986	0.986

Outline

- This second part is a joint work with Raissi, H.

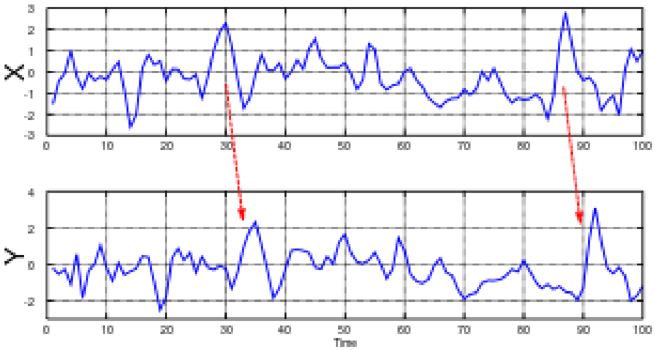


Figure: An example of causality with two time series. Ref : Wikipedia

Two principles

Granger [2] in 1969 propose the following definition for causality of two time series based on two principles :

1. The cause happens prior to its effect.
2. The cause has unique information about the future values of its effect.

As a consequence, the consideration of the cause allows to improve the prediction of the effect.

Mathematical definition

- Let $\{X_t, t \in \mathbb{Z}\}$ and $\{Y_t, t \in \mathbb{Z}\}$ be two time series.
- Let $A_t = \{Z_{k,t}, k \in I\}$, $I \subset \mathbb{Z}$, $\{X_t, Y_t\} \subseteq A_t$.
- Definitions : $\bar{X}_t = \{X_s, s \leq t\}$, $\bar{Y}_t = \{Y_s, s \leq t\}$, $\bar{A}_t = \bigcup_{s \leq t} A_s$.

Let B an information set and $\mathbb{P}(Y_t|B)$ the best linear predictor,

$$\varepsilon(Y_t|B) = Y_t - \mathbb{P}(Y_t|B), \tag{2.6}$$

$$\sigma^2(Y_t|B) = \mathbb{E}\left[\varepsilon(Y_t|B)^2\right]. \tag{2.7}$$

Mathematical definition

Definition (Granger, 1969)

The variable X causes the variable Y iff for at least one value of t :

$$\sigma^2(Y_{t+1}|\bar{A}_t) < \sigma^2(Y_{t+1}|\bar{A}_t \setminus \{\bar{X}_t\}). \tag{2.8}$$

One Example

With an autoregressive model, we have that :

$$Y_t = \alpha + \sum_{k=1}^{K} \gamma_k Y_{t-k} + \sum_{k=1}^{L} \beta_k X_{t-k} + \epsilon_t$$

the variable X does not cause the variable Y iff

$$\beta_k = 0, \quad \forall k = 1, \ldots, L.$$

Operator of covariance

$$\forall u \in L^2([0,1]),\ \Gamma u = \mathbb{E}\left(\langle X_i - \mathbb{E}(X_i), u\rangle (X_i - \mathbb{E}(X_i))\right). \qquad (2.9)$$

Definition (Causality)

We say that Y *is causing* X *if* $\Gamma_{\varepsilon(X|\overline{U-Y})} - \Gamma_{\varepsilon(X|U)}$ *is a positive definite operator.*

The model

$$\begin{cases} X_t = \rho_{11}(X_{t-1}) + \rho_{12}(Y_{t-1}) + \varepsilon_{1t}, \\ Y_t = \rho_{21}(X_{t-1}) + \rho_{22}(Y_{t-1}) + \varepsilon_{2t}, \end{cases}$$

where $\varepsilon_1, \varepsilon_2$ are errors and the $\rho_{..}$ are operators.

The null hypothesis

The test we want to perform is then

$$H_0 : \ \rho_{12} = 0, \tag{2.10}$$

against the alternative

$$H_1 : \ \rho_{12} \neq 0.$$

1. Data X and Y
2. Estimation of the parameters ρ
3. Estimation of the errors ε
4. Test the equality of operators based on the errors

Zhang and Shao (2015) [5] recently have proposed :

- a test procedure to compare the covariance operators of two mean zero stationary functional time series.

The null hypothesis is

$$H_0 : \ \Gamma_X = \Gamma_Y, \tag{2.11}$$

against the alternative

$$H_1 : \ \Gamma_X \neq \Gamma_Y,$$

Define $\{\hat{\lambda}^j_{XY}\}$ and $\{\hat{\phi}^j_{XY}\}$ the eigenvalues and eigenfunctions of

$$\hat{\Gamma}_{XY} = \frac{1}{2N}\left(\sum_{i=1}^{N} X_i \otimes X_i + Y_i \otimes Y_i\right).$$

Let $\hat{\Gamma}_{X,m} = 1/m \sum_{i=1}^{m} X_i \otimes X_i$. Let $\{\hat{\lambda}^j_{X,m}\}$ and $\{\hat{\phi}^j_{X,m}\}$ be the eigenvalues and eigenfunctions of $\hat{\Gamma}_{X,m}$. Similar quantities are defined for the second sample.

Let K be a fixed user-chosen number and

$$c_k^{i,j} = \langle (\hat{\Gamma}_{X,\lfloor k/2 \rfloor} - \hat{\Gamma}_{Y,\lfloor k/2 \rfloor})(\hat{\phi}^i_{XY}), \hat{\phi}^j_{XY} \rangle, \quad 2 \leq k \leq 2N, \ 1 \leq i,j \leq K.$$

Denote by $\hat{\alpha}_k = \text{vech}(C_k)$, with $C_k = (c_k^{i,j})_{i,j=1}^K$. To take the dependence into account, they introduce a self-normalized matrix :

$$V = \frac{1}{4N^2} \sum_{k=1}^{2N} k^2 (\hat{\alpha}_k - \hat{\alpha}_{2N})(\hat{\alpha}_k - \hat{\alpha}_{2N})'.$$

The test statistic is then

$$G = 2N \hat{\alpha}_{2N}' V^{-1} \hat{\alpha}_{2N}.$$

Define :

- $B_q(r)$ as a q-dimensional vector of independent Brownian motion
- $W_q = B_q'(1)J_q^{-1}B_q(1)$, where

$$J_q = \int_0^1 (B_q(r) - rB_q(1))(B_q(r) - rB_q(1))' dr.$$

- The critical values of W_q have been tabulated by Lobato (2001) [3].

Outline

Frédéric Ferraty and Philippe Vieu.
Nonparametric functional data analysis : theory and practice.
Springer, 2006.

C. W. J. Granger.
Investigating causal relations by econometric models and cross-spectral methods.
Econometrica, 37(3) :pp. 424–438, 1969.

Ignacio N Lobato.
Testing that a dependent process is uncorrelated.
Journal of the American Statistical Association, 96(455) :1066–1076, 2001.

S. Robbiano, M. Saumard, and M. Curé.
Improving prediction performance of stellar parameters using functional models.
Journal of applied statistics, To appear.

Xianyang Zhang and Xiaofeng Shao.
Two sample inference for the second-order property of temporally dependent functional data.
Bernoulli, 21(2) :909–929, 2015.