Application de la méthode SIR

Modélisation et prédiction de courbes de charge individuelles

Frédéric Proïa

(avec Bernard Bercu et Jérôme Saracco)

Université de Bordeaux

(Ex-)Équipe ALEA (INRIA Bordeaux Sud-Ouest)
Équipe Proba-Stat (Institut de Mathématiques de Bordeaux)

Journées MAS
Toulouse
27-29 août 2014

Sommaire

La méthode SIR

Le modèle semi-paramétrique

- Méthode introduite par Li en 1991, qui repose sur un argument géométrique.
- On s'intéresse au modèle de régression semi-paramétrique donné, pour tout $t \geq 1$, par

$$Y_t = f_\theta(X_t) + \varepsilon_t = f\left(\theta_1' X_t, \theta_2' X_t, \ldots, \theta_q' X_t\right) + \varepsilon_t$$

où

- (Y_t) est une suite de variables aléatoires de $\mathbb{R}$ à expliquer.
- (X_t) est une suite de vecteurs aléatoires de $\mathbb{R}^p$ explicatifs,
  - indépendants,
  - de même loi, de moyenne μ et de covariance Σ,
  - vérifiant la condition fondamentale.
- (ε_t) est une suite de variables aléatoires de $\mathbb{R}$,
  - iid, centrées, de variance σ^2,
  - indépendantes de (X_t).
- Le paramètre inconnu θ est une matrice de $\mathcal{M}_{p,q}(\mathbb{R})$ à estimer, avec $q < p$, dont les colonnes $\theta_1, \ldots, \theta_q$ sont linéairement indépendantes.
- La fonction de lien $f : \mathbb{R}^q \to \mathbb{R}$ est inconnue et à estimer.

La méthode SIR

Le modèle semi-paramétrique

- (X_t) vérifie la condition fondamentale :
  - Il existe $\alpha, \beta_1, \ldots, \beta_q \in \mathbb{R}^p$ tels que

$$\mathbb{E}[X_t \mid \theta_1' X_t, \ldots, \theta_q' X_t] = \alpha + \sum_{k=1}^{q} (\theta_k' X_t)\beta_k.$$

  - En particulier, la condition est vérifiée lorsque la loi de (X_t) est elliptique. C'est-à-dire que la densité f_X de X_t est de la forme

$$f_X(x \mid m, \Phi) = \frac{1}{\sqrt{\det(\Phi)}} h\big((x-m)' \Phi^{-1} (x-m)\big).$$

- Les vecteurs de θ ne sont pas totalement identifiables.
  - Pour A matrice carrée d'ordre q et inversible,

$$Y_t = f(\theta' X_t) + \varepsilon_t = f\big((A')^{-1}(\theta A)' X_t\big) + \varepsilon_t.$$

  - Ainsi, f absorbe $(A')^{-1}$ et l'algorithme estime θA.
  - Seul le sous-espace vectoriel engendré par θ est identifiable.

La méthode SIR

Le modèle semi-paramétrique

- L'espace EDR, pour *Effective Dimension Reduction*, est le sous-espace vectoriel E de $\mathbb{R}^p$ engendré par $\theta_1, \ldots, \theta_q$,

$$E = \text{Vect}(\theta) = \left\{ \sum_{k=1}^{q} \gamma_k \theta_k \text{ avec } \gamma_1, \ldots, \gamma_q \in \mathbb{R} \right\}.$$

- Étape 1-SIR : estimation d'une base de l'espace EDR.
- Étape 2-NWR : estimation non paramétrique de la fonction de lien f.

La méthode SIR

Descriptif de la méthode SIR

- On préférera travailler avec des variables explicatives standardisées,

$$Z_t = \Sigma^{-1/2}(X_t - \mu) \qquad \text{et} \qquad v_k = \Sigma^{1/2}\,\theta_k \text{ direction EDR standardisée } (1 \leq k \leq q).$$

- Le découpage en H tranches du support de (Y_t) permet l'estimation d'une matrice $\widehat{\Gamma}_n$ aisée à manipuler et possédant les mêmes propriétés que $\Gamma = \mathrm{Var}(\mathbb{E}[Z_t|Y_t])$.
- Calcul de la moyenne $\overline{X}_n$ et de la covariance Σ_n empiriques de l'échantillon (X_n).

- Pour chacune des H tranches,
  - calcul de la proportion empirique des Y_t appartenant à la tranche,
  - calcul de la moyenne empirique des Z_t de la tranche.

- Calcul de la covariance $\widehat{\Gamma}_n$ des moyennes pondérées par tranche.

- La k-ème direction EDR est obtenue par $\Sigma_n^{-1/2}\widehat{v}_k$, où $\widehat{v}_k$ est le vecteur propre associé à la k-ème plus grande valeur propre de $\widehat{\Gamma}_n$.

- Choix de H ?
  - $H = 1$ et $H = n$ n'ont aucun intérêt.
  - $H > q$ pour ne pas faire de réduction de dimension artificielle.
  - En pratique, H impacte peu la qualité des estimations lorsque n est grand.

Sommaire

La méthode SIR pour les séries chronologiques

Le modèle semi-paramétrique dans un cadre chronologique

- Dans le cadre chronologique, (X_t) n'est pas formé de vecteurs indépendants.
- Il n'existe aucun résultat théorique relatif à la méthode SIR dans un cadre chronologique.
- Observation du comportement de SIR sur des données simulées,
  - modèle purement autorégressif,
  - modèle semi-paramétrique autorégressif.

La méthode SIR pour les séries chronologiques

Le modèle semi-paramétrique dans un cadre chronologique

- On considère un signal (Y_t) de taille N, ainsi qu'un couple de paramètres (p, q).
  - p est la taille du vecteur explicatif,
  - q est le nombre de directions EDR significatives.

- Le modèle semi-paramétrique autorégressif s'écrit, pour $p+1 \leq t \leq N$,

$$Y_t = f\left(\theta' \, Y_{t-1}^p\right) + \varepsilon_t$$

avec le vecteur Y_{t-1}^p composé de la $(t-p)$-ème colonne de X', et

$$Y = \begin{pmatrix} Y_{p+1} \\ Y_{p+2} \\ \vdots \\ Y_N \end{pmatrix}, \qquad X = \begin{pmatrix} Y_p & Y_{p-1} & \cdots & Y_1 \\ Y_{p+1} & Y_p & \cdots & Y_2 \\ \vdots & \vdots & & \vdots \\ Y_{N-1} & Y_{N-2} & \cdots & Y_{N-p} \end{pmatrix}.$$

La méthode SIR pour les séries chronologiques

Le modèle semi-paramétrique dans un cadre chronologique

- Étape 1 : identification du paramétrage.
  - (H, q, p) sont les paramètres de l'algorithme SIR,
  - $\alpha \in [1/3, 1[$ est utilisé dans la fenêtre de NWR.

- Étape 2 : estimation du paramétrage.
  - Déterminer l'algorithme à utiliser, puis le couple $(\widehat{H}, \widehat{q})$ optimal sur un cube de qualité.
  - Déterminer le couple $(\widehat{p}, \widehat{\alpha})$ optimal par validation croisée ou étude d'un sous-ensemble de courbes, pour un critère lié à la prédiction (ou à la qualité de la modélisation).

- Étape 3 : estimation de θ et de f.
  - $\widehat{\theta}$ par application de SIR$(\widehat{H}, \widehat{q}, \widehat{p})$.
  - $\widehat{f}$ par estimation fonctionnelle NWR$(\widehat{\alpha})$.

- Étape 4 : prédiction à l'instant $N + 1$,

$$\widehat{Y}_{N+1} = \widehat{f}\left(\widehat{\theta}' \, Y_N^{\widehat{p}}\right).$$

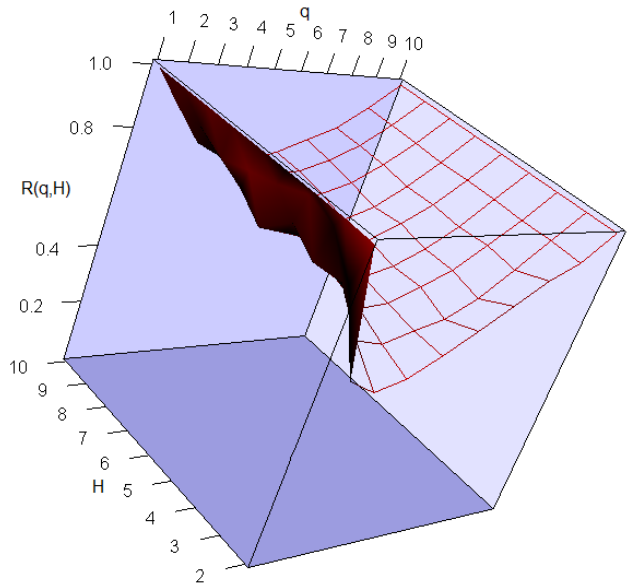
- Étape 5 : éventuellement, tests de pertinence de l'estimation.
  - Test de blancheur des résidus d'estimation.
  - Test de normalité résiduelle (?).

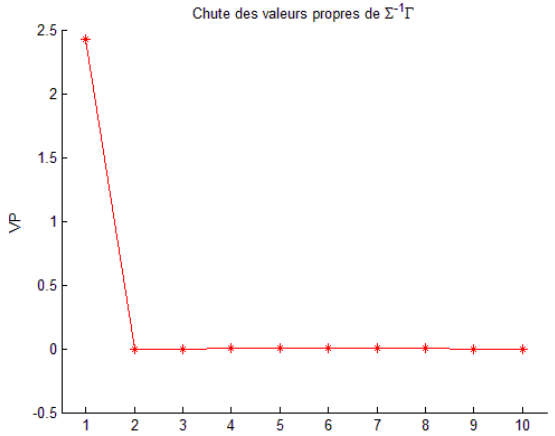
La méthode SIR pour les séries chronologiques

Données simulées

- Génération d'un signal purement autorégressif d'ordre $p = 10$.
- Soit $(Y_1, \ldots, Y_p) \overset{\text{iid}}{\sim} \mathcal{U}([9, 11])$, $\theta \in \mathbb{R}^p$ avec $\sum_{k=1}^{p} \theta_k = 1$, et (ε_t) un bruit gaussien de moyenne nulle et de variance σ^2. Le signal est défini par

$$Y_t = \theta' \, Y_{t-1}^p + \varepsilon_t.$$



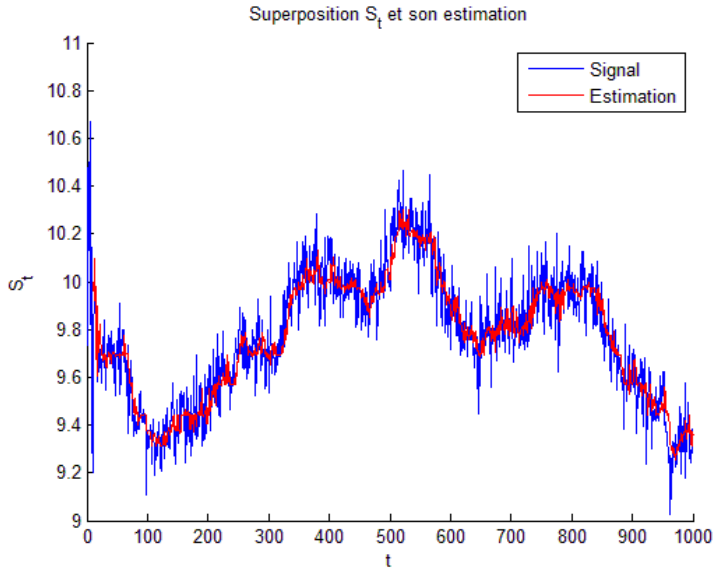


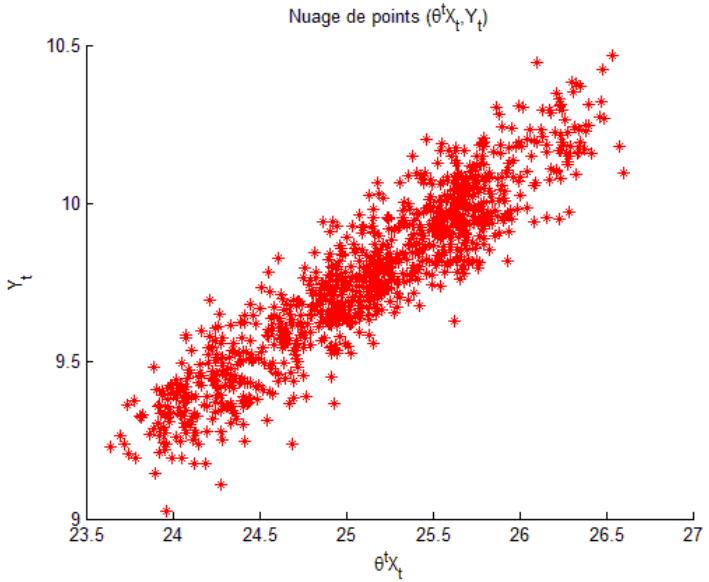
- SIR-I, $\widehat{q} = 1$, $\widehat{H} = 2$.

La méthode SIR pour les séries chronologiques

Données simulées

- Estimation SIR-I($\widehat{H} = 2, \widehat{q} = 1, p = 10$) pour $N = 1000$.



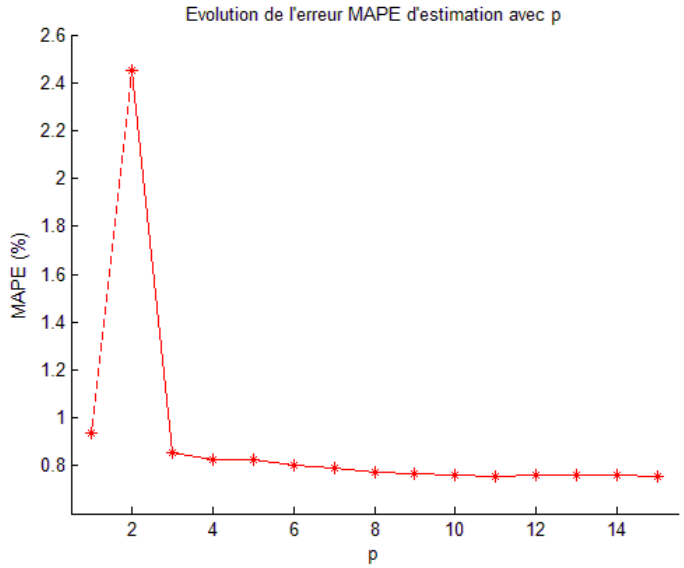


  - Tendance linéaire évidente.

La méthode SIR pour les séries chronologiques

Données simulées

- Erreur MAPE d'estimation lorsque p grandit.

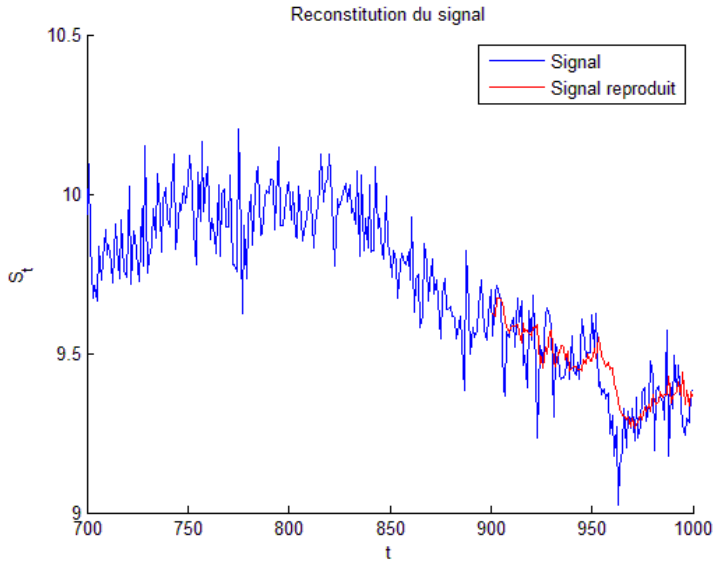


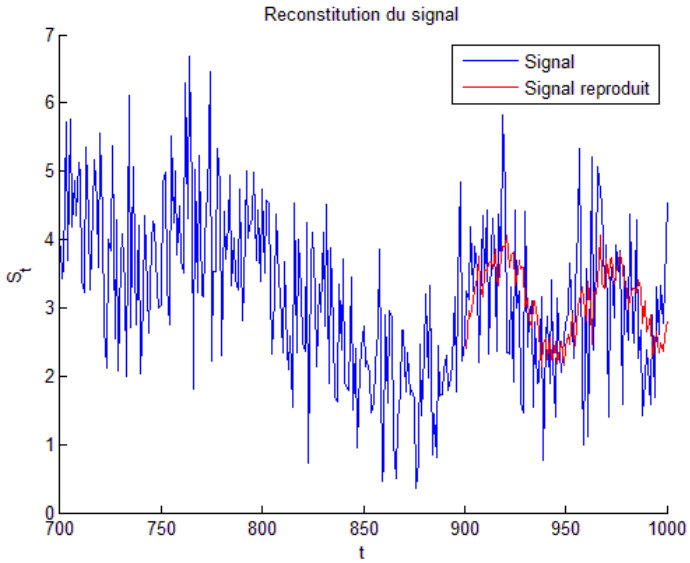
  - $\widehat{p} \geq 4$.

La méthode SIR pour les séries chronologiques

Données simulées

- Prédiction des 100 dernières valeurs par SIR-I($\widehat{H} = 2, \widehat{q} = 1, \widehat{p} = 10$) pour $\sigma^2 = 0.01$ puis $\sigma^2 = 1$, à horizon 1.





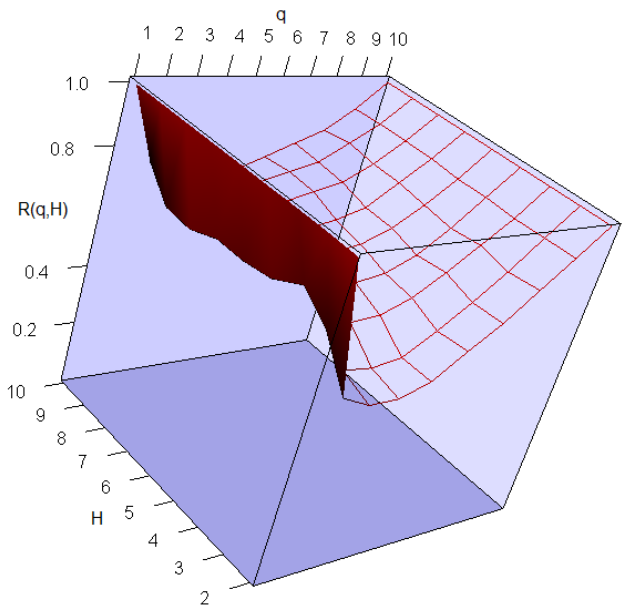
  - L'algorithme retrouve la tendance de la courbe.

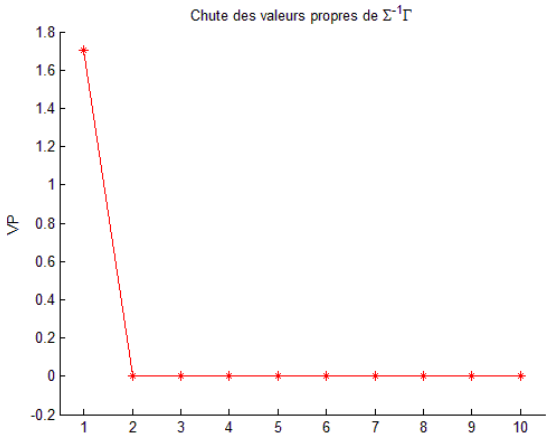
La méthode SIR pour les séries chronologiques

Données simulées

- Génération d'un signal semi-paramétrique autorégressif d'ordre $p = 10$.
- Soit $(Y_1, \ldots, Y_p) \overset{\text{iid}}{\sim} \mathcal{U}([9, 11])$, $\theta \in \mathbb{R}^p$ avec $\sum_{k=1}^{p} \theta_k = 1$, et (ε_t) un bruit gaussien de moyenne nulle et de variance σ^2. Le signal est défini par

$$Y_t = f\left(\theta' \, Y_{t-1}^p\right) + \varepsilon_t \qquad \text{où} \qquad f(x) = x(1 + \exp(-x)).$$



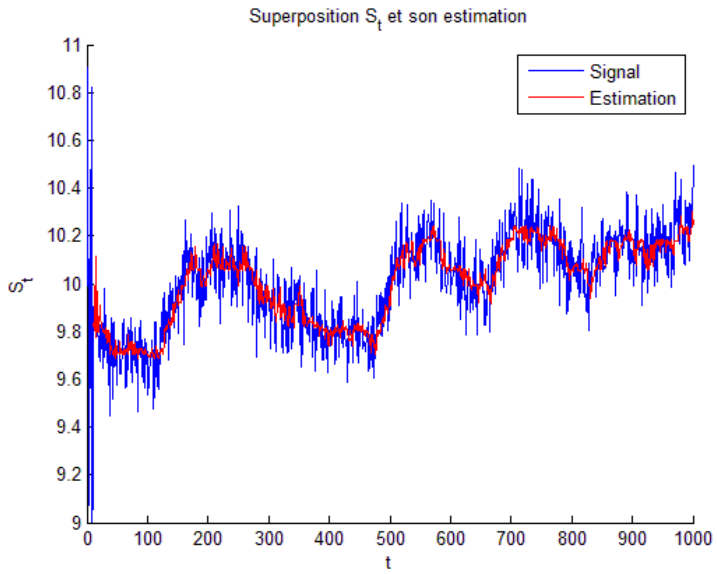


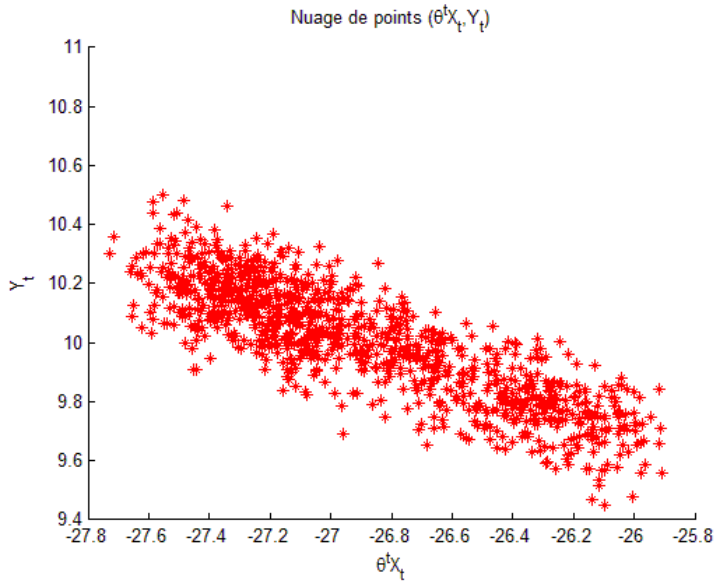
- SIR-I, $\widehat{q} = 1$, $\widehat{H} = 2$.

La méthode SIR pour les séries chronologiques

Données simulées

- Estimation SIR-I($\widehat{H} = 2, \widehat{q} = 1, p = 10$) pour $N = 1000$.



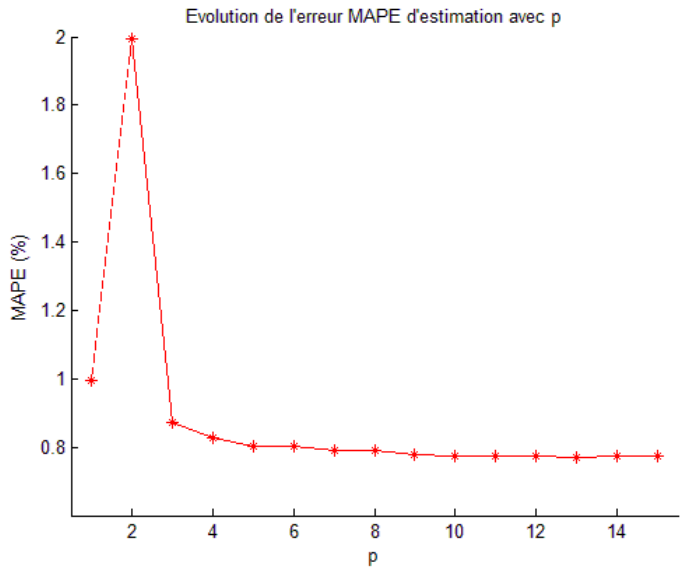


- Tendance quasi-linéaire, légèrement incurvée par la présence de l'exponentielle.

La méthode SIR pour les séries chronologiques

Données simulées

- Erreur MAPE d'estimation lorsque p grandit.

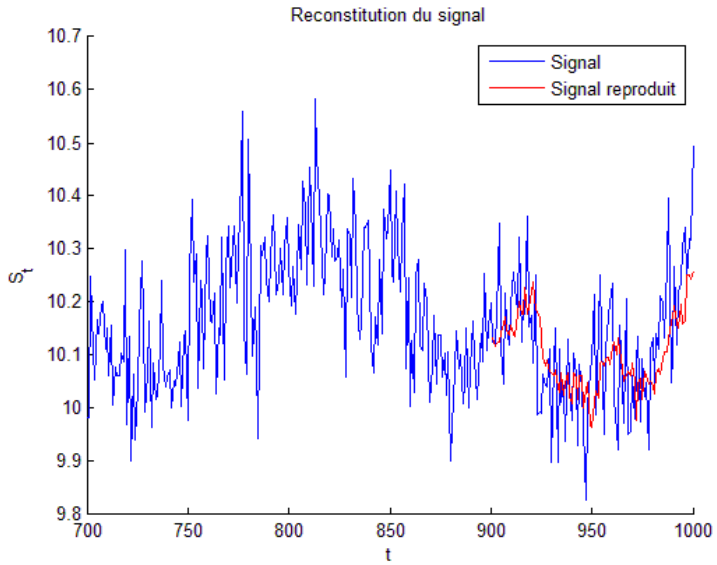


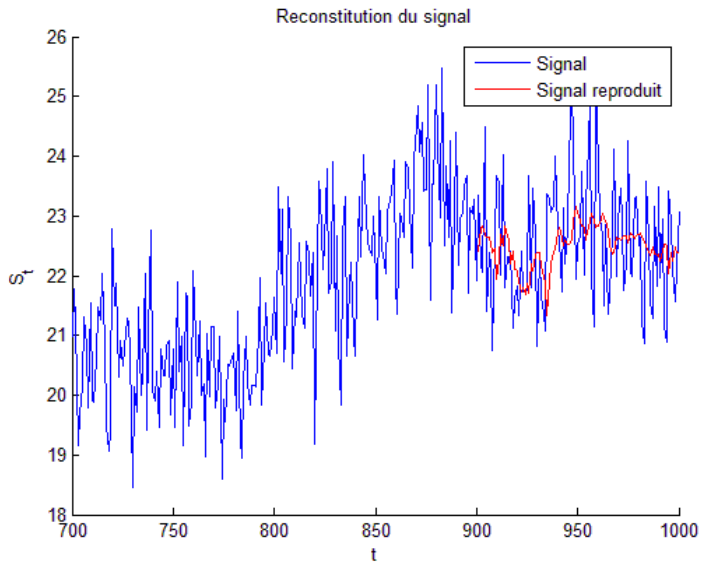
  - $\widehat{p} \geq 8$.

La méthode SIR pour les séries chronologiques

Données simulées

- Prédiction des 100 dernières valeurs par SIR-I($\widehat{H} = 2, \widehat{q} = 1, \widehat{p} = 10$) pour $\sigma^2 = 0.01$ puis $\sigma^2 = 1$, à horizon 1.



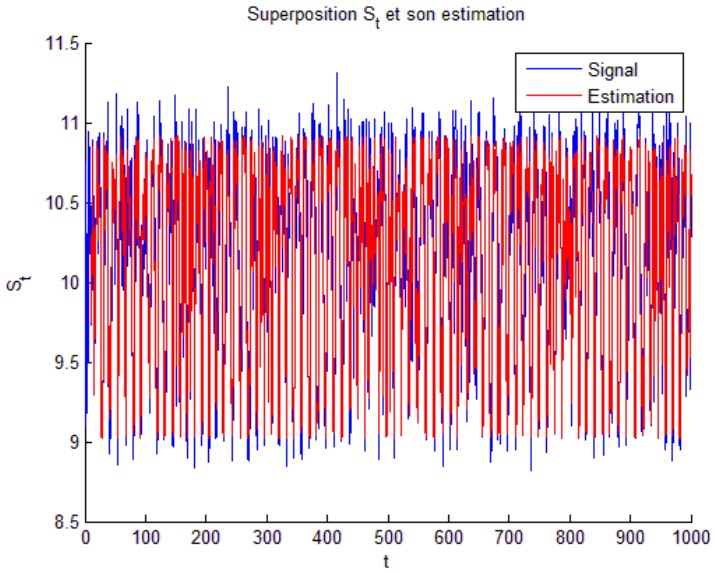


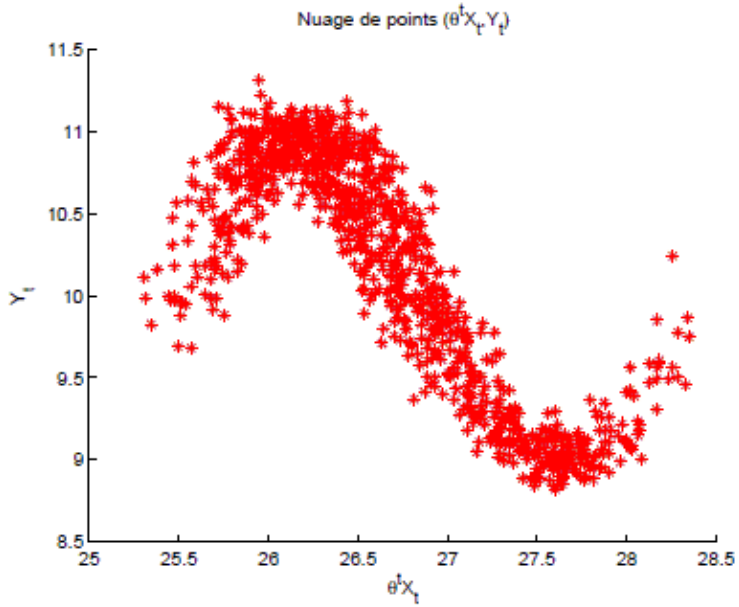
  - L'algorithme retrouve la tendance de la courbe.

La méthode SIR pour les séries chronologiques

Données simulées

- Lorsque $f(x) = 10 + \cos(2\pi x)$, avec $\widehat{H} = 5$.





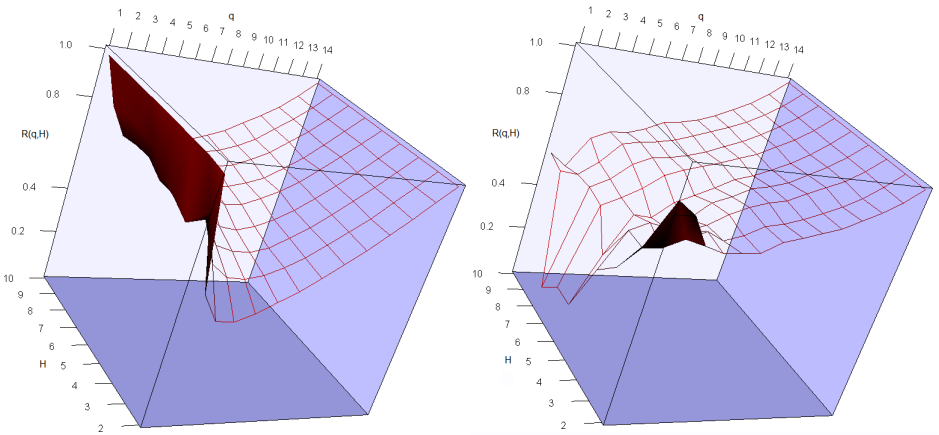
- Tendance trigonométrique.
- Nette séparation de θ et de f par l'algorithme.

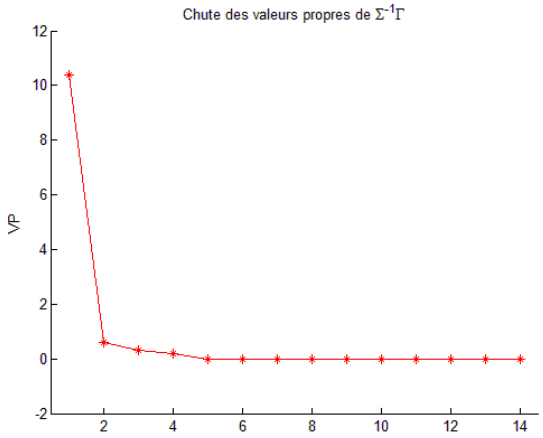
Sommaire

La méthode SIR sur les données EDF

Évaluation des paramètres

- Étude d'un sous-ensemble de 20 courbes hétérogènes, par nécessité.
- Majorité de courbes très réceptives à SIR.



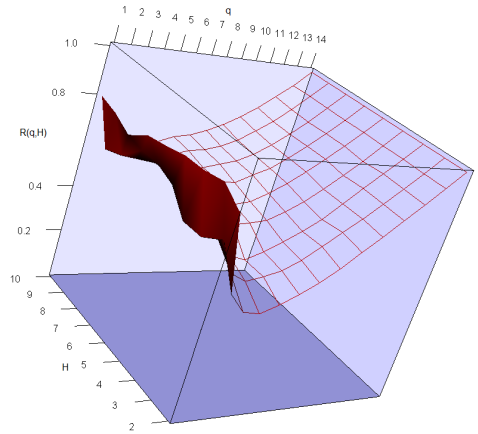


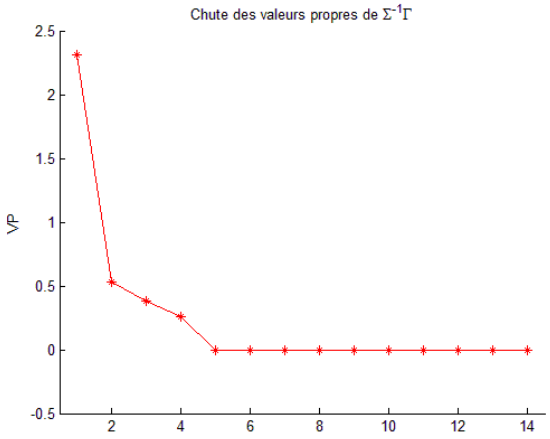
- SAVE parfois équivalent à SIR.
- Choix de SIR, $\widehat{q} = 1$, $\widehat{H} = 5$.

La méthode SIR sur les données EDF

Évaluation des paramètres

- Étude d'un sous-ensemble de 20 courbes hétérogènes, par nécessité.
- Certaines courbes moins réceptives à SIR.



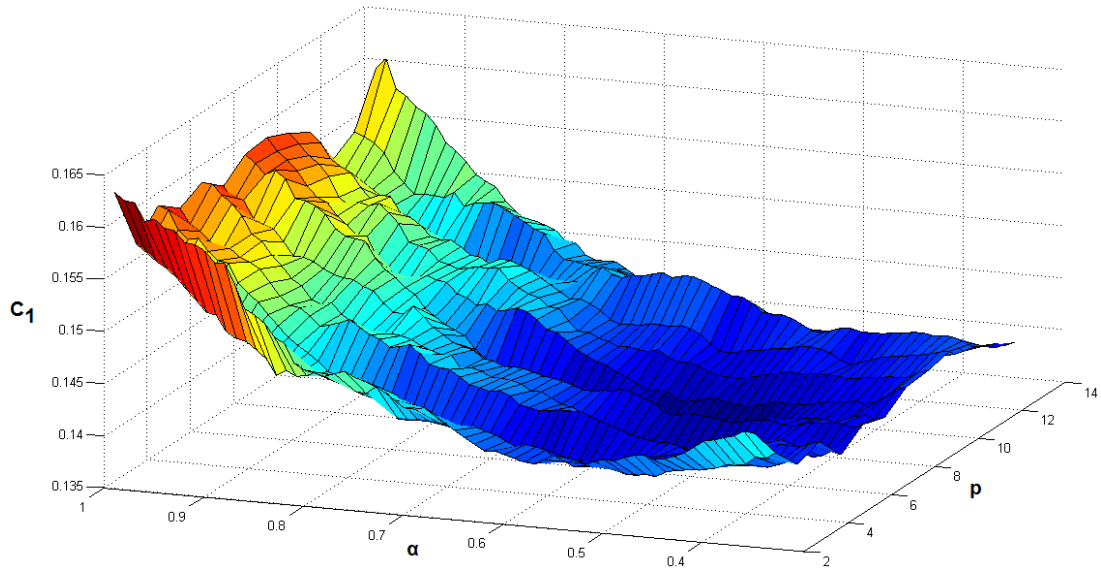


- Courbes peu réceptives à la modélisation semi-paramétrique.
- Choix de SIR malgré tout, $\widehat{q} = 1$, $\widehat{H} = 5$.

La méthode SIR sur les données EDF

Évaluation des paramètres

- Sélection de $(\widehat{p}, \widehat{\alpha})$ sur une grille bidimensionnelle, pour SIR$(\widehat{H} = 5, \widehat{q} = 1)$.

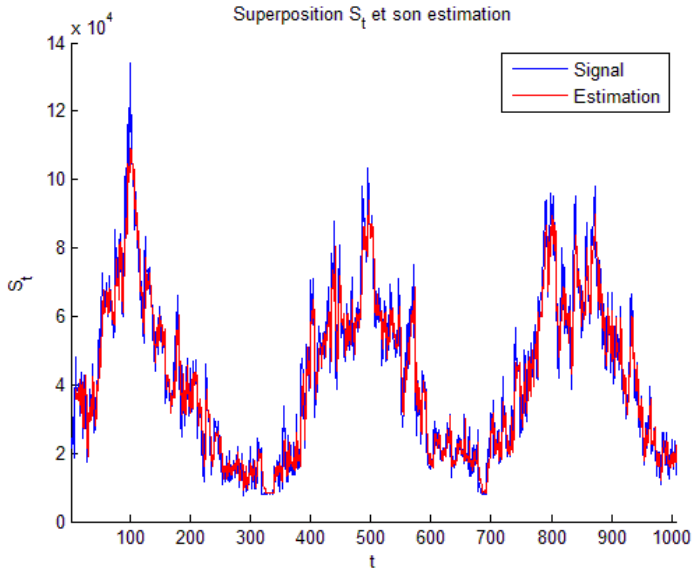


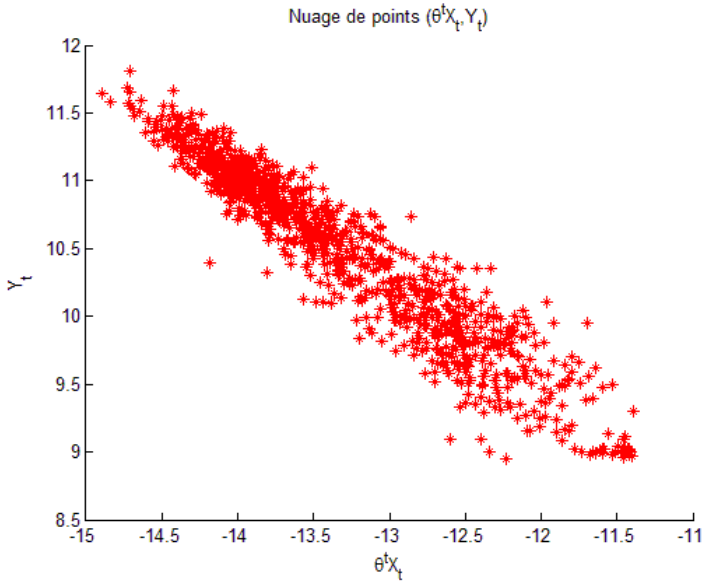
- Critère évalué sur une expérience de 183 prédictions.
- $\widehat{p} = 8$, $\widehat{\alpha} = 0.52$.
- Nécessité de faire varier ces valeurs dans leur voisinage.
- Attention : 2 valeurs possibles de $\widehat{\alpha}$, selon la finalité de l'étude.

La méthode SIR sur les données EDF

Évaluation des paramètres

- Courbe très réceptive, pour SIR($\widehat{H} = 5, \widehat{q} = 1, \widehat{p} = 8$) et $\widehat{\alpha} = 0.52$.



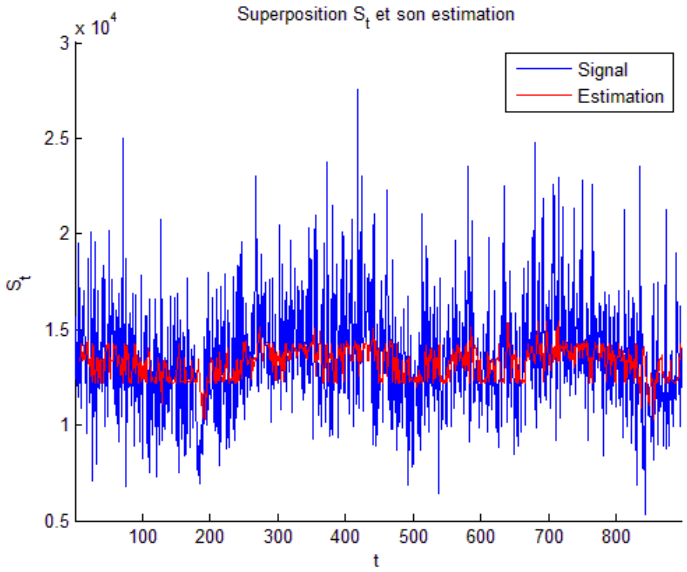


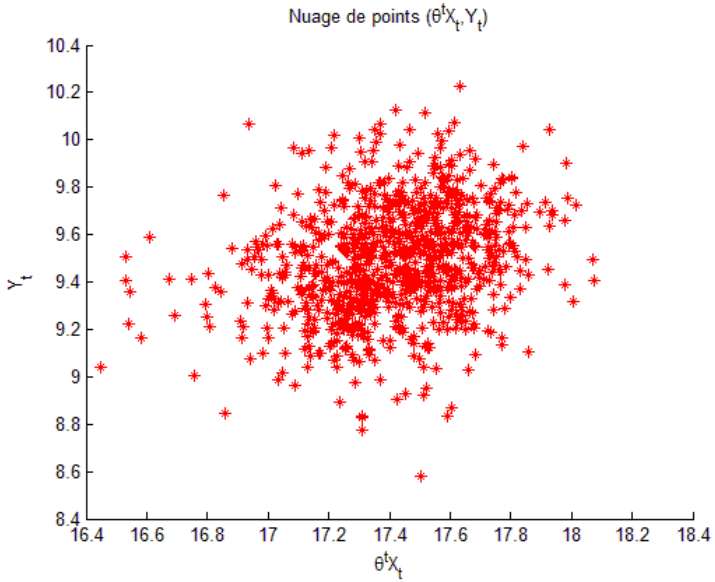
  - Tendance linéaire évidente.

La méthode SIR sur les données EDF

Évaluation des paramètres

- Courbe peu réceptive, pour SIR($\widehat{H} = 5, \widehat{q} = 1, \widehat{p} = 8$) et $\widehat{\alpha} = 0.52$.



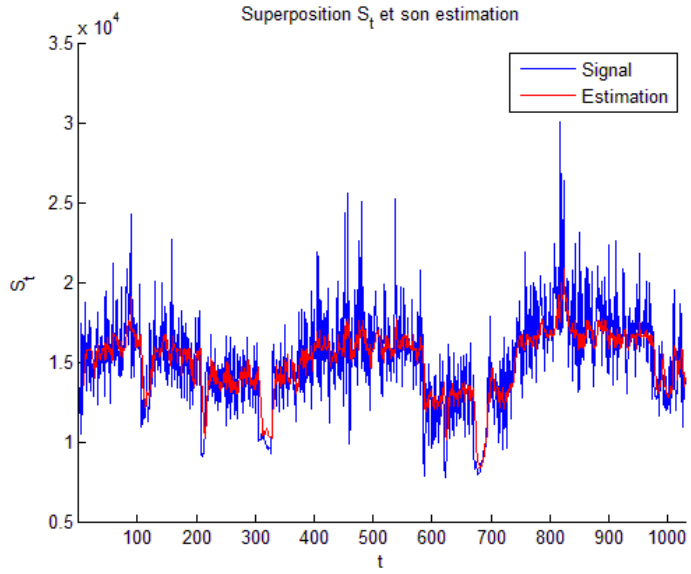


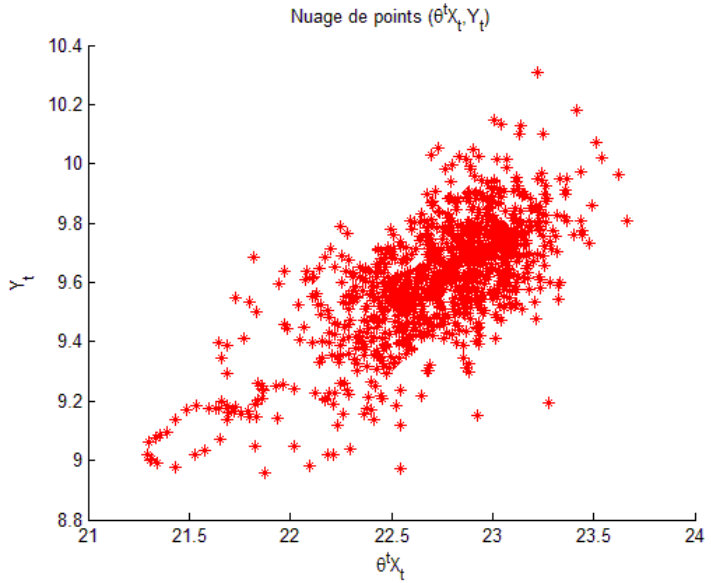
- Tendance floue, nuage dispersé.
- Courbe probablement trop bruitée.
- On ne peut pas faire mieux que retrouver la tendance de la courbe.

La méthode SIR sur les données EDF

Évaluation des paramètres

- Cas général, pour SIR($\widehat{H} = 5, \widehat{q} = 1, \widehat{p} = 8$) et $\widehat{\alpha} = 0.52$.



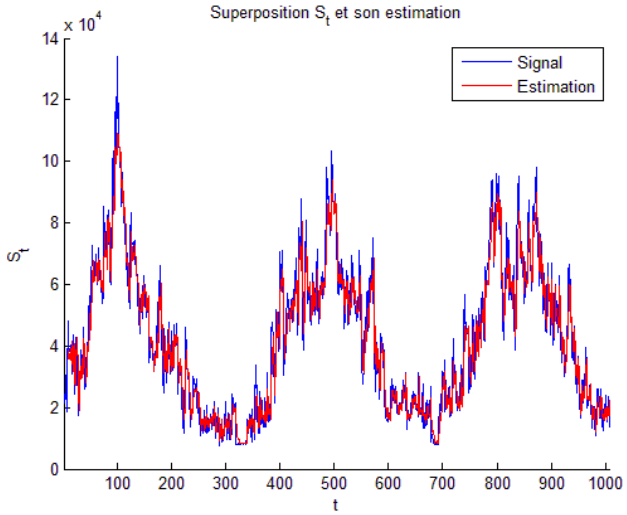


  - Tendance linéaire évidente.

Sommaire

Conclusion

- La stratégie semi-paramétrique donne de meilleurs résultats que la stratégie non paramétrique (NW, NWR).
- La stratégie paramétrique (SARIMA) donne de meilleurs résultats que la stratégie semi-paramétrique.
- Au pas horaire, une saisonnalité journalière évidente se dégage, ce qui rend les modèles SARIMA largement supérieurs : réduction de dimension obligatoire !
- Méthode SIR dans un cadre de dépendance des variables explicatives ?



Merci de votre attention !