Sélection de modèles pour la prédiction de séries temporelles faiblement dépendantes

Avec O. Wintenberger, Université Paris Dauphine

Pierre Alquier

Université Paris 7, LPMA
& CREST, Laboratoire de Statistique

Journées MAS de la SMAI, Bordeaux, le 1er septembre 2010

Contexte général

On observe $(X_1, \ldots, X_n)$ d'une série stationnaire à valeurs réelles, dont la loi de probabilité (inconnue) est P.

Objectif : fabriquer p et f tels que $f(X_{t-1}, ..., X_{t-p})$ soit une bonne prédiction de X_t (à toute date t).

Paramétrisation

Pour un $p \in \{1, ..., \lfloor n/2 \rfloor\}$, on choisit une famille de «prédicteurs» $f_\theta : \mathbb{R}^p \to \mathbb{R}$ avec $\theta \in \Theta_p$.

Exemple

Soit $\Theta_p \subset \mathbb{R}^p$ *avec* $f_\theta(x_{-1}, ..., x_{-p}) = \sum_{i=1}^p x_{-i}\theta_i$.

En général, on découpera l'espace des paramètres $\Theta_p = \bigcup_{\ell=1}^L \Theta_{p,\ell}$.

Exemple

Un exemple, $\Theta_{p,\ell} \subset \mathbb{R}^{p\times\ell}$ *avec*
$f_\theta(x_{-1}, ..., x_{-p}) = \sum_{i=1}^p \sum_{j=1}^\ell \theta_{i,j} g_j(x_{-i})$.

Une condition technique sur les modèles

On suppose que l'on connait une constante $\mathcal{L}$ telle que

$$|f_\theta(x_{-1}, ..., x_{-p}) - f_\theta(x'_{-1}, ..., x'_{-p})| \leq \sum_{j=1}^{p} a_j(\theta)|x_{-j} - x'_{-j}|$$

où

$$\sum_{j=1}^{p} a_j(\theta) \leq \mathcal{L}.$$

Risque (ou erreur de prédiction) d'un paramètre

On propose l'erreur moyenne de prédiction comme évaluation d'un paramètre.

Définition

Pour $\theta \in \Theta_{p,\ell}$ on définit

$$R(\theta) = \mathbb{E}\left[\left|X_{p+1} - f_\theta(X_p, ..., X_1)\right|\right].$$

Formalisation de l'objectif : trouver un $\hat{p}$ et un $\hat{\theta} \in \Theta_p$ (dépendant des données) tels que $R(\theta)$ soit aussi petit que possible.

Le meilleur prédicteur dans chaque modèle

Définition

On définit $\overline{\theta}_{p,\ell}$ *pour tout* (p, ℓ) *donné par*

$$R(\overline{\theta}_{p,\ell}) = \inf_{\theta \in \Theta_{p,\ell}} R(\theta).$$

Rmq : pas forcément unique.

Bien sûr, $\overline{\theta}_{p,\ell}$ est inconnu, il dépend de la loi inconnue P du processus X.

Spoiler : la fin de l'histoire

Ce que l'on va obtenir à la fin de l'exposé.

Pour un estimateur $\hat{\theta}$ (défini dans la Section 2), sous certaines hypothèses sur P (données dans la Section 3), on a

$$\mathbb{E}\left(R(\hat{\theta})\right) \leq \inf_{p,\ell}\left\{R(\overline{\theta}_{p,\ell}) + \Delta_{p,\ell}(n)\right\},$$

où $\Delta_{p,\ell}(n) \xrightarrow[n\to\infty]{} 0$.

Risque empirique

On définit un estimateur de $R(\theta)$.

Définition

Pour p et $\theta \in \Theta_p$ on définit

$$r_n(\theta) = \frac{1}{n-p} \sum_{i=p+1}^{n} \left| X_i - f_\theta(X_{i-1}, \ldots, X_{i-p}) \right|.$$

Noter que $\mathbb{E}(r_n(\theta)) = R(\theta)$ pour un θ fixé.

Estimateurs randomisés

Pour chaque

- $p \in \{1, ..., \lfloor n/2 \rfloor\}$,
- $1 \leq \ell \leq L$, et
- $\lambda \in \Lambda = \{2, 4, 8, 16, ...\} \cap \{1, ..., n^2\}$,

on tire au hasard

$$\hat{\theta}^{\lambda}_{p,\ell} \sim e^{-\lambda r_n(.)} d\pi_{p,\ell}(.)$$

où $\pi_{p,\ell}(.)$ est la mesure uniforme sur $\Theta_{p,\ell}$.

Définition de $\hat{R}_{p,\ell,\lambda}$

Pour

- $K_n > 0$ donné,
- des poids *a priori* sur les modèles $\Theta_{p,\ell}$: $w_{p,\ell} \geq 0$ avec $\sum w_{p,\ell} \leq 1$,

on définit

$$\hat{R}(p,\ell,\lambda) = -\frac{1}{\lambda}\log\int_{\Theta_{p,\ell}} e^{-\lambda r_n(\theta)} d\pi_{p,\ell}(\theta) + \frac{1}{\lambda}\log\frac{\log n}{w_{p,\ell}} + \frac{\lambda K_n^2}{n(1-p/n)^2}$$

Sélection d'un modèle

Sélection d'un modèle $\Theta_{p,\ell}$ et d'un paramètre de «température» λ.

Définition

$$\hat{\theta} = \hat{\theta}^{\hat{\lambda}}_{\hat{p},\hat{\ell}}$$

où

$$(\hat{p}, \hat{\ell}, \hat{\lambda}) \in \arg\min_{p,\ell,\lambda} \hat{R}_{p,\ell,\lambda}.$$

On peut maintenant utiliser $f_{\hat{\theta}}(.)$ pour faire de la prédiction.

Hypothèses sur le processus $(X_t)_t$

On suppose que pour tout $c > 0$, $\mathbb{E}(e^{c|\xi_0|}) = \Psi(c) < \infty$ et que $X = (X_n)_{n\in\mathbb{Z}}$ est une solution de l'équation

$$\forall n, \quad X_n = F(X_{n-1}, X_{n-2}, \ldots; \xi_n)$$

avec les ξ_i i.i.d., et F telle que

$$|F(x;y) - F(x';y')| \leq \sum_{j=1}^{\infty} a_j \left|x_j - x'_j\right| + u|y - y'|,$$

$$\sum_{j=1}^{\infty} a_j < 1 \text{ et } \sum_{j=1}^{\infty} j\log(j)a_j < \infty.$$

Rmk : Pour tout F satisfaisant cette conditon, il existe un tel processus X, voir Doukhan et Wintenberger [DW08].

Contrôle de la complexité

On suppose qu'il y a une constante $1 \leq d_{p,\ell} < \infty$ telle que

$$\sup_{\gamma > e} \frac{\gamma}{\log \gamma} \frac{\int_{\Theta_{p,\ell}} \left[R(\theta) - R(\overline{\theta}_{p,\ell}) \right] e^{-\gamma R(\theta)} d\pi_{p,\ell}(\theta)}{\int_{\Theta_{p,\ell}} e^{-\gamma R(\theta)} d\pi_{p,\ell}(\theta)} \leq d_{p,\ell}$$

Dans la plupart des cas usuels $d_{p,\ell}$ est $C \times$ «dimension du modèle».

Exemple

Prédicteurs autoregressifs $f_\theta(x_{-1}, \ldots, x_{-p}) = \sum_1^p \theta_i X_{-i}$.
$\Theta_p = \{\theta : \|\theta\|_1 \leq L\}$. *Alors pour une constante* c *qui ne dépend pas de* p *où* n, $d_p = c.p$.

Résultat principal

Théorème

Prenons $K_n = \log(n)$. *On suppose que* $\sum_{p,\ell} \exp(-d_{p,\ell}) \leq 1$ *et on prend* $w_{p,\ell} \simeq \exp(-d_{p,\ell})$. *Alors*

$$\mathbb{E}\left(R(\hat{\theta})\right) \leq \inf_{p,\ell : d_{p,\ell} \leq n/2} \left\{ R(\overline{\theta}_{p,\ell}) + C(F, \Psi, \mathcal{L}) \sqrt{\frac{d_{p,\ell}}{n} \log^2(n)} \right\}$$

Remarque : $C(F, \Psi, \mathcal{L})$ est connue explicitement, un peu longue à écrire, mais cependant pas trop énorme si $\Psi(c) = \mathbb{E}(e^{c|\xi_0|})$ est raisonnable.

Commentaires (1/2) : complexité de la famille de modèles

On a supposé que $\sum_{p,\ell} \exp(-d_{p,\ell}) \leq 1$, donc le choix $w_{p,\ell} = \exp(-d_{p,\ell})$ est possible.

Sinon, on perdra le terme supplémentaire

$$\frac{\log\left(\frac{1}{w_{p,\ell}}\right)}{\sqrt{n}}.$$

Commentaires (2/2) : une autre version de ce théorème

Dans le preprint [AW09] on donne une autre version de ce théorème avec une hypothèse plus forte concernant des coefficients de dépendance faible du processus.

Dans ce cas, on obtient une borne vraie avec grande probabilité et plus seulement en espérance...

Quelques mots sur les preuves

Les preuves utilisent deux outils :

- les inégalités PAC-Bayesiennes, cf. Catoni [Cat07] (dans le contexte de la classification, étendues par Audibert [Aud04] et A. [Alq07] pour la régression) ;
- des inégalités de concentration : dans le cas i.i.d., Hoeffding, Bernstein, ... Ici une inégalité de Rio [Rio00].

Application : modèles additifs non paramétriques

On suppose que $X_n = f_1(X_{n-1}) + ... + f_q(X_{n-q}) + \xi_n$ avec f_1, ..., f_q et q inconnus. Soit s la régularité de la moins régulière des f_j.

Modèles : $\theta \in \Theta_{p,\ell} \subset \mathbb{R}^{p \times \ell}$ avec

$$f_\theta(x_{-1}, ..., x_{-p}) = \sum_{p'=1}^{p} \sum_{j=1}^{\ell} \theta_{p',j} e_j(x_{-p'})$$

et (e_j) est la base de Fourier. On a alors

$$\mathbb{E}\left(R(\hat{\theta})\right) \leq \mathbb{E}(|\xi_0|) + \frac{C \log^2(n)}{n^{\frac{s}{2s+1}}}.$$

Conclusion

Comparaison avec Baraud, Compte & Viennet [BCV01] (résultats fins sur la sélection de modèle avec risque quadratique) :

- ne perdent pas nécessairement le $\log(n)$;
- mais leur fonction de perte n'est pas reliée à une erreur de prévision.

Simulations :

- similaire à un estimateur bayésien : peut être implémenté par MCMC ;
- résultats corrects, mais peuvent être améliorés (travail en cours).

Références

P. Alquier.
Pac-bayesian bounds for empirical risk minimizers.
Mathematical methods of statistics, 2007.

J.-Y. Audibert.
Aggregated estimators and empirical complexity for least square regression.
Annales de l'IHP : P & S, 2004.

P. Alquier and O. Wintenberger.
Model selection and randomization for weakly dependent time series forecasting.
Preprint, 2009.

Y. Baraud, F. Comte, and G. Viennet.
Adaptative estimation in autoregression or β-mixing regression via model selection.
The Annals of Statistics, 2001.

O. Catoni.
PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning).
Lecture Notes-Monograph Series. IMS, 2007.

P. Doukhan and O. Wintenberger.
Weakly dependent chains with infinite memory.
Stochastic Processes and their Applications, 2008.

E. Rio.
Ingalites de hoeffding pour les fonctions lipschitziennes de suites dependantes.
Comptes Rendus de l'Academie des Sciences de Paris, Serie I, 2000.