HAL
archives-ouvertes.fr

Model selection and randomization for weakly dependent time series forecasting

Pierre Alquier, Olivier Wintenberger

▶ **To cite this version:**

Pierre Alquier, Olivier Wintenberger. Model selection and randomization for weakly dependent time series forecasting. 41èmes Journées de Statistique, SFdS, Bordeaux, 2009, Bordeaux, France, France. inria-00386733

HAL Id: inria-00386733
https://hal.inria.fr/inria-00386733

Submitted on 22 May 2009

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Prédiction de séries chronologiques par sélection de modèles et randomisation.

Pierre Alquier & Olivier Wintenberger *Laboratoire de Probabilités et Modèles Aléatoires, Université Paris 7 (Diderot), 175, rue du Chevaleret, 75252 Paris CEDEX 05, FRANCE, et CREST-LS.*
&
CEREMADE Place du Maréchal De Lattre De Tassigny 75775 PARIS CEDEX 16 FRANCE

Abstract

Observing a stationary time series, we present in this presentation new two-steps procedures for predicting the next value of the time series. Following machine learning theory paradigm, the first step consists in determining randomized estimators, or "experts", in (possibly numerous) different predictive models. In the second step estimators are obtained by model selection or randomization associated with exponential weights of these experts. We prove Oracle inequalities for both estimators and provide some applications for linear, artificial Neural Networks and additive non-parametric predictors.

Résumé

A partir d'observations d'une série stationnaire nous proposons deux nouvelles méthodes de prédictions de la prochaine valeur de cette série. La méthode en deux étapes employée utilisent les outils de l'apprentissage statistique. La première étape consiste proposer des estimateurs randomisés dans de nombreux modèles prédictifs différents. La seconde étape fournit la prédiction soit par sélection parmi tous les modèles de la première étape, soit par randomisation sur l'ensemble des indices des modèles. Nous prouvons des inégalités type Oracle pour ces deux estimateurs et nous donnons des exemples de modèles prédictifs tels que les modèles AR(p), les réseaux de neurones artificiels et les modèles non-paramétriques additifs.

Mots clés

Principal : Statistique des processus - séries temporelles, Secondaire : Choix de modèles.

When observing a time series, one crucial issue is to predict first future value with the observed past values. Since the seminal works of Akaike, see for example [1], different model selection procedures have been studied for inferring how many observed past values are needed for predicting the next value. Efficiency of different penalized empirical risk minimizers such that AIC, BIC, Mallows, APE's predictors have been proved when the observations satisfy a linear auto-regressive model, see for instance Ing [2]. The main issue in this context is to determine the order of an efficient predictive linear autoregressive model and then to estimate its coefficients. There the model fitted by the observations is assumed to belong into the same class than the predictive models.

In the same time, model selection procedure have been hugely improved using learning theory in the independent and identically distributed (iid for short) case, see Vapnik [3] and Massart [4] among others. Results such that Oracle inequalities have been settled in very extended context. Even if the true model does not belong into one of the models proposed by the experts recent procedures ensure that the risk is as small as possible. However, few works have been done for dependent observations, principally in two direction: penalized lest square and randomization techniques. Baraud *et al.* [5] proved Oracle inequalities with respect to the quadratic loss and under β-mixing condition. Their penalized empirical risk minimizers select an efficient predictive model when the number of useful past values is known. Recently, the theory of individual sequences leads also to Oracle inequalities for risk of prediction. Randomization with exponential weights of experts advices predicts the observations as if it was a deterministic sequence. We refer the reader to Lugosi and Cesa-Bianchi [6] for more details. Good predictors are then obtained given the expert devices. But the form of the expert devices given the observations is not given and then the form of the predictors is not tractable.

In this paper, we give Oracle inequalities for the $\mathbb{L}^1$-risk of prediction of some stationary time series. We introduce two new procedures that find an efficient predictive model associated with an efficient number of past values. To prove this we use the PAC-Bayesian approach introduced by McAllester [7]. This general theoretical framework has proved to efficiently give Oracle inequalities in many iid frameworks, see Catoni [8, 9, 10], Audibert [11] and Alquier [12]. There exist procedures and Oracle inequalities in the dependent cases, see Baraud *et al.* [5] and Modha and Masry [13]. In Modha and Masry [13], their procedure use the α-mixing coefficients of the observations. To our knowledge, there is no efficient estimation of this coefficients and their procedure is not implementable in practice. In Baraud *et al.* [5], the Oracle inequality holds only if the β-mixing coefficients and the prediction procedure satisfy together intricate conditions. Here again, as β-mixing coefficients are not estimable there is no way to check those conditions. In this paper, the prediction procedures are for the first time completely free of the dependence properties of the observations. It represents an important progress for learning theory applications with dependent observations.

Let us assume that we observe $(X_1, \ldots, X_n)$ from a stationary time series $X = (X_t)_{t\in Z}$ distributed as π_0 on X^Z where X is an Hilbert space equipped with its usual norm $\|\cdot\|$. For each θ in the set of parameter Θ we associate a $p(\theta)$-autoregressive function f_θ from $X^{p(\theta)}$ to X that represents a predictive model. Then each $\theta \in \Theta$ is associated with a predictor $f_\theta(X_{n-1}, ..., X_{n-p(\theta)})$. The risk of prediction is the absolute loss $R(\theta)$ defined as:

$$R(\theta) = \pi_0 \left[\left\| f_\theta(X_{p(\theta)}, ..., X_1) - X_{p(\theta)+1} \right\| \right],$$

where here and all along the paper $\pi[h] = \int h d\pi$ for any measure π and any integrable function h. The choice of this risk instead of the classic quadratic loss is due to its Lipschitzian property, very well suited with the dependence context here. The main objective of this paper is to determine two different procedures that give estimators $\widehat{\theta}_n$ with associated risk $R(\widehat{\theta}_n)$ satisfying an Oracle inequality - in other words, $R(\widehat{\theta}_n)$ is not far from $\inf_\Theta R$.

As we have to deal with different models and different delays in the same time, it is convenient to split the set Θ in subsets of the form:

$$\Theta = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \Theta_p \text{ with } \Theta_p = \bigcup_{\ell=1}^{m_p} \Theta_{p,\ell},$$

where $m_p > 0$ has to be fixed carefully. The set Θ_p consists in different predictive models that need the same number of past values. To fix the idea, let us give the simple example additive non parametric predictive models when $X = R$. Let us define

$$\widehat{X}_{n+1} = \sum_{i=0}^{\hat{p}} \hat{f}_i(X_{n-i}).$$

Then we fix $\hat{\theta}_n = ((\hat{f}_i)_{0\le i\le \hat{p}})$. We split

$$\Theta = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \Theta_p = \bigcup_{p=1}^{\lfloor \frac{n}{2} \rfloor} \{(f_i)_{0\le i\le p} \in A_p\}$$

where C is a compact subset of R and A_p is a compact subset of F^{p+1} for F the set of integrable functions from R to R. Under suitable conditions on F, there exists an ordered functional basis $(\varphi_i)_{i\ge 1}$. Then the index ℓ corresponds to the number of the firsts functionals in the basis that we consider. Then $f_i = \sum_{j=1}^{\ell} a_{i,j}\varphi_j$ for each i and $\Theta_{p,\ell} = \{(a_{i,j})_{0\le i\le p, 1\le j\le \ell}\}$.

The common first step of our two prediction procedures consists on proposing a randomized estimator $\tilde{\theta}_{p,\ell}$ for each subset $\Theta_{p,\ell}$. Then we propose two different estimators $\hat{\theta}$ and $\tilde{\theta}$ of a parameter θ associated with an efficient predictive model. The first procedure is a model selection that provides $(\hat{p}, \hat{\ell})$. It leads to the natural choice $\hat{\theta} = \tilde{\theta}_{\hat{p},\hat{\ell}}$. Our model selection criterion for each indices (p, ℓ) is close to the following penalized empirical risk criterion

$$r_n(\hat{\theta}_{p,\ell}) + \sqrt{\frac{K_n d_{p,\ell}}{n-p}} \ln(d_{p,\ell} n),$$

where $r_n(\theta)$ is the empirical risk, $d_{p,\ell}$ is a measure of the complexity of $\Theta_{p,\ell}$, highly related to its dimension, and $K_n > 0$ is independent of p, ℓ. The second procedure is a second randomization step on the indexes (p, ℓ) that gives $(\tilde{p}, \tilde{\ell})$ and then leads to the corresponding estimator $\tilde{\theta} = \tilde{\theta}_{\tilde{p},\tilde{\ell}}$. The exponential weights associated to each indices (p, ℓ) have the same form than the ones used for randomizing expert devices in the theory of individual sequence. They deeply depends on a parameter $K_n > 0$.

The value of K_n has to be fixed arbitrarily and it has lot of consequences on the sharpness of the Oracle inequalities we obtained. For bounded observations, the best is to fix it larger than some constant depending on the (non-estimable) dependence properties of the observations. If we fail, remark that a less good Oracle inequality still holds, see the results in Section **??**. For possibly unbounded observations, we can fix it proportional to $\ln(n)$ independently on the observations. Such choice leads to an additional logarithmic term in the rate of convergence. But remark that even for K_n fixed as a constant we over-penalized the expected risk there is always additional logarithmic terms in the rates of the Oracle inequalities, see below. So we can fix as a rule of thumb $K_n = C \ln(n)$ for some known C and our procedure is free of the dependence properties of the observations.

Let us resume the main results of this paper for K_n fixed to $\ln(n)$. For bounded observations, we prove a Probably Approximately Correct Oracle inequality: for n large enough, with probability at least $1 - \varepsilon$

$$R(\hat{\theta}_n) \leq \min_{p,\ell} \left\{ \inf_{\Theta_{p,\ell}} R(\theta) + C \sqrt{\frac{d_{p,\ell}}{n-p}} \ln(d_{p,\ell}/\varepsilon) \ln^2(n) \right\},$$

where C is a constant. For possibly unbounded observations, we obtain Oracle inequalities in expectation. More precisely, we obtain that for n sufficiently large

$$\pi_0[R(\hat{\theta}_n)] \leq \min_{p,\ell} \left\{ \inf_{\Theta_{p,\ell}} R(\theta) + C \sqrt{\frac{d_{p,\ell}}{n-p}} \ln(d_{p,\ell}) \ln^2(n) \right\},$$

where C is constant. This result can be compared with those of Baraud, Comte and Viennet [5] and Modha and Masry [13]. They achieve respectively Oracles inequalities of

the form

$$\begin{aligned}\pi_0[(R'(\hat{\theta}_{p,n})] &\leq (1+\frac{1}{C})^2 \min_{\ell}\left\{\inf_{\Theta_{p,\ell}} R'(\theta) + C^3\frac{d_{p,\ell}}{n-p}\right\} \text{ for each } p,\\ \pi_0[R'(\hat{\theta}_n)] &\leq (1+\frac{1}{C}) \min_{p,\ell}\left\{\inf_{\Theta_{p,\ell}} R'(\theta) + C\left(\frac{K_n d_{p,\ell}}{n-p}\right)^c \ln(d_{p,\ell})\right\}\end{aligned}$$

where R' is the excess quadratic risk, $0 < c < 1$ is a constant depending on the dependence structure of the observations and C is fixed by the statistician. Our Oracle inequalities are sharper than the ones of [13]. Baraud *et al.* [5] achieve the opitmal rates and we fail, but with a loss in the constant. Moreover, as already noticed, those authors are not fully adaptive in p.

To obtain such Oracle inequalities, sharp exponential inequalities are used in the dependent setting. For this, weakly dependence properties on the observations are assumed. This dependent setting might be more general than the mixing one, see the monograph of Dedecker *et al.* [14]. Here we use in the bounded cases the θ_∞-coefficients (also called γ-mixing coefficients) introduced in Rio [15] to derive a sharp Hoeffding inequality in the dependent framework. These coefficients generalize the uniform mixing ones. In the unbounded cases we use generic models called chains with infinite memory introduced by Doukhan and Wintenberger [16] that includes many classical econometric models such that ARMA, GARCH and LARCH. Here we work under restrictions of additive forms that unfortunately exclude unbounded volatility models. Our dependent framework is not comparable with the β- or α-mixing one as it deals with some dynamical systems that are not mixing, see Andrews [17] and Dedecker and Prieur [18] or details on these counter-examples.

Bibliographie

[1] Akaike, H. (1973) *Second International Symposium on Information Theory (Tsahkadsor, 1971)* pp. 267–281.

[2] Ing, C.-K. (2007) *Ann. Statist.* **35**, 1238–1277.

[3] Vapnik, V. N. (1998) The Nature of Statistical Learning Theory, Springer-Verlag, .

[4] Massart, P. (2006) Concentration Inequalities and Model Selection, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2003), Springer, .

[5] Baraud, Y., Comte, F., and Viennet, G. (2001) *The Annals of Statistics* **29(3)**, 839–875.

[6] Lugosi, G. and Cesa-Bianchi, N. (2006) Prediction, Learning and Games, Cambridge University Press, .

[7] McAllester, D. A. (1998) In Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT, Madison, WI, 1998) : ACM pp. 230–234.

[8] Catoni, O. (2003) *Preprint Laboratoire de Probabilités et Modèles Aéatoires*.

[9] Catoni, O. (2004) Statistical Learning Theory and Stochastic Optimization, Lecture Notes in Mathematics (Saint-Flour Summer School on Probability Theory 2001), Springer, .

[10] Catoni, O. (2007) PAC-Bayesian Supervised Classification (The Thermodynamics of Statistical Learning)volume **56**, of Lecture Notes-Monograph Series IMS, .

[11] Audibert, J.-Y. (2004) *Annales de l'Institut Henri Poincaré: Probability and Statistics* **40(6)**, 685–736.

[12] Alquier, P. (2008) *Mathematical Methods in Statistics* **17(4)**, 279–304.

[13] Modha, D. S. and Masry, E. (1998) *IEEE transactions on information theory* **44(1)**, 117–133.

[14] Dedecker, J., Doukhan, P., Lang, G., León, J. R., Louhichi, S., and Prieur, C. (2007) Weak Dependence, Examples and Applicationsvolume **190**, of Lecture Notes in Statistics Springer-Verlag, Berlin.

[15] Rio, E. (2000) *Comptes Rendus de l'Académie des Sciences de Paris, Srie I* **330**, 905–908.

[16] Doukhan, P. and Wintenberger, O. (2008) *Stochastic Processes and their Applications* **118**, 1997–2013.

[17] Andrews, D. W. K. (1984) *J. Appl. Probab.* **21(4)**, 930–934.

[18] Dedecker, J. and Prieur, C. (2005) *Probability Theory and Related Fields* **132**, 203–235.