PREDICTION OF TIME SERIES USING YULE-WALKER EQUATIONS WITH KERNELS

Maya Kallas [1,2], *Paul Honeine* [1], *Cédric Richard* [3], *Clovis Francis* [2] *and Hassan Amoud* [4]

[1] Institut Charles Delaunay (CNRS), LM2S, Université de Technologie de Troyes, France
[2] Laboratoire d'analyse des systèmes (LASYS), Lebanese University, Lebanon
[3] Observatoire de la Côte d'Azur (CNRS), Université de Nice Sophia-Antipolis, France
[4] Azm Center for Research in Biotechnology and its Applications, Lebanese University, Lebanon

ABSTRACT

The autoregressive (AR) model is a well-known technique to analyze time series. The Yule-Walker equations provide a straightforward connection between the AR model parameters and the covariance function of the process. In this paper, we propose a nonlinear extension of the AR model using kernel machines. To this end, we explore the Yule-Walker equations in the feature space, and show that the model parameters can be estimated using the concept of expected kernels. Finally, in order to predict once the model identified, we solve a pre-image problem by getting back from the feature space to the input space. We also give new insights into the convexity of the pre-image problem. The relevance of the proposed method is evaluated on several time series.

Index Terms— autoregressive model, Yule-Walker equations, expected kernels, pre-image problem, nonlinear model

1. INTRODUCTION

Many applications involve the analysis of time series data. One of the simplest, yet efficient, way to model time series is the autoregressive (AR) model. It states that each sample is given as a linear combination of a small number of previous samples. A prediction scheme is therefore inherent to this model, once the model parameters determined, i.e., the coefficients in the linear combination. These parameters are tightly linked to the covariance function of the process. The Yule-Walker equations explore this direct correspondence in order to estimate the parameters from the covariances of the time series. As a linear prediction model, the AR model is not adapted to treat nonlinear systems.

An elegant way to extend linear models into nonlinear ones is given by the concept of kernel methods in machine learning. The main idea is to map the data, using a nonlinear function, from the input space into a feature space, usually of a higher dimension. By using the *kernel trick* [1], it turns out that one can transform linear techniques into nonlinear ones, without the need to explicitly exhibit the mapped space. This principle has shown its capacity in many applications, initially with Vapnik's Support Vector Machines (SVM) [2], and now includes kernel principal component analysis and SVM novelty detection, only to name a few. In the same spirit, some kernel-based methods were considered for the analysis and prediction of time series data [3], including the SVM for regression and kernel Kalman filter [4].

In this paper, we explore the concept of kernel methods in order to provide a nonlinear extension of the AR model. To this end, we propose to take full advantage of the Yule-Walker equations in the feature space. We show that the parameters are estimated using the (lagged) expected kernels. The concept of expected kernels has shown its capacity in recent research [5, 6]. Finally, to provide a prediction, one needs to map the result back into the input space. This is the pre-image problem. We give in this paper new insights into the convexity of this problem, and propose a technique to solve the problem. See [7] for a recent review, with several applications in signal processing.

The rest of the paper is organized as follows: Next, we present the classical AR model with Yule-Walker equations. In Section 2, we present the kernel-based AR model and detail the solution of Yule-Walker equations in feature space. In Section 3, we provide a prediction scheme by solving the pre-image problem. Section 4 covers the experiments done to evaluate the relevance of the proposed technique on "MG$_{30}$" and "Lorenz attractor" time series.

AR model: the Yule-Walker equations

The AR model is widely used to analyze stationary and nonstationary time series [8]. It gives each sample as a linear combination of previous samples. The AR model is defined by the fixed weights α_i, for $i = 1, 2, \ldots, p$, where p defines the order of the model. Let $x_1, x_2, \ldots, x_n$ be a time series, a p-order AR model is described by

$$x_i = \sum_{j=1}^{p} \alpha_j x_{i-j} + \epsilon_i,$$

This work is partly supported by the Lebanese University and the French-Lebanese research program CEDRE No. 10 SCI F15/L5.

for $i = p+1, \ldots, n$, where ϵ_i is the unfitness error, assumed white Gaussian with zero mean. The parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ are directly connected with the covariance function of the process. One can therefore determine these parameters from the autocorrelation function. This is the essence of the Yule-Walker equations: Let r be the autocorrelation function of the time series, then $r(\tau) = \sum_{j=1}^{p} \alpha_j r(\tau - j)$, for a lag $\tau \geq 1$. Since $r(-\tau) = r(\tau)$, we obtain the matrix form of the Yule Walker equations

$$\boldsymbol{r} = \boldsymbol{R}\boldsymbol{\alpha},$$

where $\boldsymbol{r} = [r(1) \; \cdots \; r(p)]^\top$, $\boldsymbol{\alpha} = [\alpha_1 \; \cdots \; \alpha_p]^\top$, and

$$\boldsymbol{R} = \begin{bmatrix} 1 & r(1) & \ldots & r(p-1) \\ r(1) & 1 & \ldots & r(p-2) \\ \vdots & & \ddots & \vdots \\ r(p-1) & r(p-2) & \ldots & 1 \end{bmatrix},$$

where $r(0) = 1$ without loss of generality. Assuming that the $p \times p$ matrix $\boldsymbol{R}$ is invertible, the coefficients $\boldsymbol{\alpha}$ are estimated by $\boldsymbol{\alpha} = \boldsymbol{R}^{-1}\boldsymbol{r}$. Once the coefficients are estimated, the AR model can be applied to predict future samples.

2. KERNEL AUTOREGRESSIVE MODEL USING YULE-WALKER EQUATIONS

In order to derive a nonlinear extension of the Yule-Walker equations for autoregressive models, we use the principle of kernel machines. Let $\mathcal{X}$ be the input space, and let the kernel $\kappa\colon \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ be positive semi-definite, namely $\sum_{i,j} \alpha_i \alpha_j \kappa(x_i, x_j) \geq 0$ for all $x_i, x_j \in \mathcal{X}$ and any $\alpha_i, \alpha_j \in \mathbb{R}$. The Moore-Aronszajn theorem [9] states that a positive semi-definite kernel corresponds to an inner product in some arbitrary feature space. Let $\Phi(\cdot)$ denotes the mapping function from the input space $\mathcal{X}$ into the feature space $\mathcal{H}$, then

$$\kappa(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle_{\mathcal{H}},$$

for any $x_i, x_j \in \mathcal{X}$, where $\langle \cdot\,, \cdot \rangle_{\mathcal{H}}$ denotes the corresponding inner product in $\mathcal{H}$, and $\| \cdot \|_{\mathcal{H}}$ its norm.

Using this concept, each sample of the time series $x_1, x_2, \ldots, x_n$ is mapped from the input space into some feature space, yielding $\Phi(x_1), \Phi(x_2), \ldots, \Phi(x_n)$. Thus, an AR model in the feature space is defined by

$$\Phi(x_i) = \sum_{j=1}^{p} \alpha_j \, \Phi(x_{i-j}) + \varepsilon_i^{\Phi},$$

which belongs to the feature space span by the images of the samples with the map $\Phi(\cdot)$. While the samples x_i are assumed zero-mean, this is not often the case for the mapped data $\Phi(x_i)$. Let μ be the mean of the mapped time series, namely

$$\mu = \mathbb{E}[\Phi(x_i)],$$

where $\mathbb{E}[\cdot]$ is the expectation, with $\sum_i \Phi(x_i)$. Then,

$$\begin{aligned} \Phi(x_i) - \mu &= \sum_{j=1}^{p} \alpha_j \Phi(x_{i-j}) + \varepsilon_i^{\Phi} - \mu \\ &= \sum_{j=1}^{p} \alpha_j \big(\Phi(x_{i-j}) - \mu\big) + \varepsilon_i^{\Phi} - \Big(1 - \sum_{j=1}^{p} \alpha_j\Big)\mu. \end{aligned}$$

With the expectation of both sides, the times series being assumed stationary, we get $(1 - \sum_{j=1}^{p} \alpha_j)\mu = \mathbb{E}[\varepsilon_i^{\Phi}]$. By considering the inner product (in the feature space) of both sides of the above equation with $(\Phi(x_{i-\tau}) - \mu)$, for some positive lag τ, we get

$$\begin{aligned} \langle \Phi(x_i) - \mu,\, \Phi(x_{i-\tau}) - \mu \rangle_{\mathcal{H}} &= \langle \varepsilon_i^{\Phi} - \mathbb{E}[\varepsilon_i^{\Phi}],\, \Phi(x_{i-\tau}) - \mu \rangle_{\mathcal{H}} \\ &\quad + \sum_{j=1}^{p} \alpha_j \langle \Phi(x_{i-j}) - \mu,\, \Phi(x_{i-\tau}) - \mu \rangle_{\mathcal{H}}. \end{aligned} \tag{1}$$

By substituting the inner product in the feature space with the corresponding kernel function, we define the *centered version* of a kernel $\kappa(\cdot, \cdot)$ with

$$\kappa_c(x_i, x_j) = \langle \Phi(x_i) - \mu,\, \Phi(x_j) - \mu \rangle_{\mathcal{H}}.$$

By analogy with the linear AR case, we assume that the noise ε_i^{Φ} and $\Phi(x_{i-\tau})$ are uncorrelated for every positive lag τ. Therefore, by taking the expectations of expression (1) and assuming the stationarity, we get for any $\tau \geq 1$:

$$\mathbb{E}[\kappa_c(x_i, x_{i-\tau})] = \sum_{j=1}^{p} \alpha_j \, \mathbb{E}[\kappa_c(x_{i-j}, x_{i-\tau})], \tag{2}$$

where the notion of expected kernels is equivalent to the one recently studied in [6]. By considering all the lag values, expression (2) is written in matrix form

$$\boldsymbol{r}_{\kappa} = \boldsymbol{R}_{\kappa} \, \boldsymbol{\alpha},$$

where

$$\boldsymbol{r}_{\kappa} = \Big[\mathbb{E}[\kappa_c(x_i, x_{i-1})] \;\; \mathbb{E}[\kappa_c(x_i, x_{i-2})] \; \cdots \; \mathbb{E}[\kappa_c(x_i, x_{i-p})]\Big]^\top,$$

and $\boldsymbol{R}_{\kappa}$ is the matrix described by the expected kernels with

$$\begin{bmatrix} \mathbb{E}[\kappa_c(x_i, x_i)] & \mathbb{E}[\kappa_c(x_{i-2}, x_{i-1})] & \cdots & \mathbb{E}[\kappa_c(x_{i-p}, x_{i-1})] \\ \mathbb{E}[\kappa_c(x_{i-1}, x_{i-2})] & \mathbb{E}[\kappa_c(x_i, x_i)] & \cdots & \mathbb{E}[\kappa_c(x_{i-p}, x_{i-2})] \\ \vdots & \vdots & \ddots & \vdots \\ \mathbb{E}[\kappa_c(x_{i-1}, x_{i-p})] & \mathbb{E}[\kappa_c(x_{i-2}, x_{i-p})] & \cdots & \mathbb{E}[\kappa_c(x_i, x_i)] \end{bmatrix}$$

The vector of coefficients $\boldsymbol{\alpha}$ is obtained by inverting the matrix $\boldsymbol{R}_{\kappa}$, with

$$\boldsymbol{\alpha} = \boldsymbol{R}_{\kappa}^{-1} \boldsymbol{r}_{\kappa}.$$

In practice, the expectations are estimated over a set of n available samples. The centered version of the kernel is evaluated using

$$\kappa_c(x_i, x_j) = \kappa(x_i, x_j) - \frac{1}{n}\sum_{k=1}^{n} \kappa(x_i, x_k) - \frac{1}{n}\sum_{k=1}^{n} \kappa(x_j, x_k) + \frac{1}{n^2}\sum_{k,k'=1}^{n} \kappa(x_{k'}, x_k).$$

3. PREDICTION BY SOLVING THE PRE-IMAGE PROBLEM

Once the model parameters are determined on a set of n available samples, one may predict some future sample by

$$\psi_i = \sum_{j=1}^{p} \alpha_j \, \Phi(x_{i-j}),$$

starting with $i = n+1$. Being a linear combination of p images, ψ_i belongs to the feature space. In order to predict a sample, one needs to map back ψ_i from the feature space into the input space. This is the pre-image problem. Several techniques have been proposed to solve this ill-posed problem. This includes a multidimensional scaling technique, a conformal map, and a learning scheme. See [7] for a formal definition of the pre-image problem, and a recent survey of the literature.

We solve the pre-image problem by seeking an approximate solution x_i^* whose counterpart $\Phi(x^*)$ is as close as possible to ψ_i, the latter being defined as above. The resulting optimization problem is:

$$x_i^* = \arg\min_x \frac{1}{2}\Big\|\Phi(x) - \sum_{j=1}^{p} \alpha_j \, \Phi(x_{i-j})\Big\|_{\mathcal{H}}^2.$$

This is equivalent to the optimization problem

$$x_i^* = \arg\min_x J_i(x),$$

with

$$J_i(x) = -\sum_{j=1}^{p} \alpha_j \, \kappa(x_{i-j}, x) + \frac{1}{2}\kappa(x, x), \qquad (3)$$

where the term independent of x has been removed.

Consider the case of the Radial Basis Functions, with kernels of the form

$$\kappa(x_j, x_{j'}) = f(\|x_j - x_{j'}\|^2). \qquad (4)$$

A sufficient condition for a function $f \in \mathcal{C}^\infty$ to be a valid positive-definite kernel is its complete monotonicity, i.e., its k-th derivative satisfies

$$(-1)^k f^{(k)}(\zeta) \geq 0 \qquad (5)$$

for any non-negative ζ [10]. The following Proposition gives insights on the convexity of the pre-image problem, as defined by (3):

Proposition 1. *A sufficient condition for the convexity of the cost function is given by the non-negativity of the coefficients* $\alpha_1, \ldots, \alpha_p$.

Proof. Taking the second derivative of the cost function (3) with respect to x, we get

$$\begin{aligned}\nabla_x^2 J_i(x) &= \nabla_x\Big[2\sum_{j=1}^{p} \alpha_j \, (x_{i-j} - x)\, f^{(1)}(\|x_{i-j} - x\|^2)\Big] \\ &= 2\sum_{j=1}^{p} \alpha_j \Big(- f^{(1)}(\|x_{i-j} - x\|^2) \\ &\qquad + 2(x_{i-j} - x)^2 \, f^{(2)}(\|x_{i-j} - x\|^2)\Big)\end{aligned}$$

The term between parentheses is positive, due to condition (5). Therefore, a sufficient condition for the second derivative to be positive, and thus for the convexity of (3), is that all the coefficients α_j are positive. □

Unfortunately, the cost function is not convex in the general case. However, it is reasonable to consider a local model, since the pre-image is estimated from its p previous samples. Therefore, we consider a local gradient descent approach with

$$x_{i,t+1}^* = x_{i,t}^* - \eta_t \nabla_x J_i(x_{i,t}^*),$$

where the index t denotes the iterative technique. The convergence is controlled by the step size η_t, in the opposite direction of the gradient of $J_i(x)$ with respect to x. Using the form (4), the gradient of the kernel with respect to x is given by

$$\nabla_x \kappa(x_{i-j}, x) = 2(x_{i-j} - x)\, f^{(1)}(\|x_{i-j} - x\|^2).$$

By combining this expression with the gradient of the cost function $J_i(x)$, we get

$$\nabla_x J_i(x) = 2\sum_{j=1}^{p} \alpha_j \, (x_{i-j} - x)\, f^{(1)}(\|x_{i-j} - x\|^2).$$

Such expression simplifies further for several kernel functions, such as the Gaussian kernel with

$$f(\zeta) = \exp\left(\tfrac{-1}{2\sigma^2}\zeta\right),$$

thus $f^{(1)}(\zeta) = -\frac{1}{2\sigma^2} f(\zeta)$.

4. EXPERIMENTS

We illustrate the efficiency of the proposed method with two well-known time series data: "Mackey-Glass" (MG_{30}) time series provides a model of the blood cells production evolution, and "Lorenz attractor" is the solution to a system of

	"MG$_{30}$"	"Lorenz"
Multilayer perceptron	0.0461	0.2837
Support vector regression	0.0313	0.1811
Nonlinear Kalman filter	0.0307	0.03183
Kernel AR model	**0.00006**	**0.1793**

Table 1. Mean square error for different nonlinear prediction approaches.

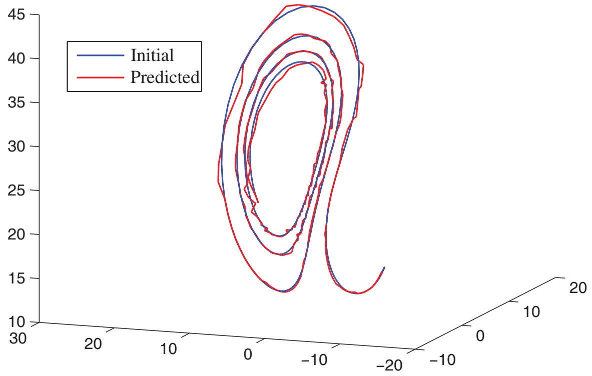

Fig. 1. Initial "Lorenz attractor" time series, with its prediction using the proposed approach.

differential equations. We considered the Gaussian kernel. A set of $n = 300$ samples was used to determine the order p, the parameters $\alpha_1, \alpha_2, \ldots, \alpha_p$ of the AR model, as well as the value of the bandwidth σ of the Gaussian kernel. The next 300 samples, for $i = n+1, \ldots, 2n$, were used to evaluate the performance of the proposed kernel-based AR model.

This configuration is identical to the one given in [4], from which we borrowed the comparative results given in Table 1. Depending on the bandwidth chosen for the Gaussian kernel, the error can be reduced to be approximately inconsiderable. As far as the bandwidth is getting smaller, the error will be decreased. The mean square error was estimated with

$$\epsilon = \frac{1}{n} \sum_{i=n+1}^{2n} \|x_i^* - x_i\|^2,$$

where x_i^* is the predicted value at instant i, and x_i is the true value of the time series at the same time. An illustration of the prediction performance is given in Figure 1 for the "Lorenz attractor". It is worth noting that the proposed approach is simpler to implement, with less parameters, and significantly lower computational complexity than all the other methods given in Table 1.

5. CONCLUSION

We presented a kernel-based AR model for the prediction of time series data. We showed that one can take advantage of the Yule-Walker equations in the feature space, by using the concept of expected kernels. The prediction was assured by solving a pre-image problem. The relevance of this proposed method was revealed by a comparison to well-known nonlinear prediction techniques.

There are many possibilities for future work: we are currently working on methods to estimate the optimal order of the AR model in the feature space, in the same spirit of the Akaike Information Criterion. We are also studying other parameter estimation techniques, such as the Levinson-Durbin method. Also, we are considering the use of Gaussian process methods where the covariance function is the kernel function. Moreover, it would be desirable to extend the proposed approach to the autoregressive moving average model.

6. REFERENCES

[1] M. A. Aizerman, E. A. Braverman, and L. Rozonoer, "Theoretical foundations of the potential function method in pattern recognition learning," in *Automation and Remote Control*, 1964, number 25, pp. 821–837.

[2] V. N. Vapnik, *Statistical Learning Theory*, Wiley-Interscience, September 1998.

[3] C. Richard, J. C. M. Bermudez, and P. Honeine, "Online prediction of time series data with kernels," *IEEE trans. on Signal Processing*, vol. 57, no. 3, pp. 1058–1067, March 2009.

[4] L. Ralaivola and F. D'alche-Buc, "Time series filtering, smoothing and learning using the kernel kalman filter," in *Proc. IEEE International Joint Conference on Neural Networks*, 2005, vol. 3, pp. 1449–1454.

[5] H. S. Anderson, M. R. Gupta, E. Swanson, and K. Jamieson, "Channel-robust classifiers," *IEEE Transactions on Signal Processing*, vol. 59, no. 4, pp. 1421–1434, 2011.

[6] H. S. Anderson and M. R. Gupta, "Expected kernel for missing features in support vector machines," in *IEEE Workshop on Statistical Signal Processing*, Nice, France, 28-30 June 2011.

[7] P. Honeine and C. Richard, "The pre-image problem in kernel-based machine learning," *IEEE Signal Processing Magazine, special issue on "dimensionality reduction via subspace and manifold learnin"*, vol. 28 (2), March 2011.

[8] M. Chevalier and Y. Grenier, "Autoregressive models with time-dependent log area ratios," in *Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP '85.*, apr 1985, vol. 10, pp. 1049 – 1052.

[9] N. Aronszajn, "Theory of reproducing kernels," *Trans. Amer. Math. Soc.*, vol. 68, pp. 337–404, 1950.

[10] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bulletin of the American Mathematical Society*, vol. 39, pp. 1–49, 2002.