NONLINEAR FILTERING BY KRIGING, WITH APPLICATION TO SYSTEM INVERSION

J.-P. Costa, L. Pronzato and E. Thierry

Laboratoire I3S, CNRS-UNSA, Les Algorithmes/Bât. Euclide,
2000 route des Lucioles, Sophia-Antipolis
06410 Biot, FRANCE

ABSTRACT

Prediction by kriging does not rely on any specific model structure, and is thus much more flexible than approaches based on parametric behavioural models. Since accurate predictions are obtained for extremely short training sequences, it generally performs better than prediction methods using parametric models. Application to nonlinear system inversion is considered

1. INTRODUCTION

We consider the situation where the input and output sequences $\{x_k\}$ and $\{y_k\}$ of a SISO nonlinear system S are observed on a given horizon n, and the response of the system to future inputs must be predicted. Such a situation is rather common, and finds applications in many signal processing problems.

Usually, one builds an input/output model for S and uses this model to predict its response for new values of the input. When no prior knowledge on S is available, a behavioural model is used.

The Volterra and Wiener functional series [6], or the NARMAX (Nonlinear AutoRegressive Moving Average model with eXogenous inputs) model proposed in [4], are traditional parametric representations for nonlinear systems of unknown structure. The inclusion of information from both lagged inputs and outputs in the NARMAX model gives a lot of flexibility, but the choice of the characteristics of the structure (polynomial degree D, memory lengths m_x and m_y of the input and output signals) is rather difficult. Moreover, the number p of parameters of the model, determined by D, m_x and m_y, is generally very large, which makes their estimation difficult (long data sets are required).

The parametric approaches mentioned above rely on *a priori* choices, which strongly influence the quality of the model obtained. Kriging, which originated in geostatistics, see e.g., [3, 2], is a statistical tool for modeling spatial observations, with or without observation errors, and does not rely explicitly on any specific structure. To the best of our knowledge, it has never been used in a signal processing context to solve problems such as system inversion.

The kriging approach is described in Section 2. It can be called semi–parametric since the model contains a linear regression part (parametric), and a non–parametric part considered as the realization of a random process. The covariance matrix of this process is parameterized and, assuming that the process is Gaussian, the parameters of the covariance are estimated by maximum likelihood. The memory length m_x of the input is the only important prior choice concerning the structure, and a prior over–estimation of m_x only results in heavier computations [1]. This makes the approach especially attractive when the structure of nonlinear system involved is totally unknown.

2. SEMI–PARAMETRIC MODELING BY KRIGING

Let S be a system with output y depending in an unknown (possibly nonlinear) manner on a vector $\mathbf{x}$ of inputs. Once inputs $\{\mathbf{x}_k\}$ and associated outputs $\{y_k\}$ are observed, $k = 1, \ldots, n$, we predict the value of y at new unsampled values of $\mathbf{x}$, that is $\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots$, by interpolating previous data by the best linear unbiased predictor. Note that although we are predicting *future* responses $y_{n+1}, y_{n+2}, \ldots$, we are not approximating a function of the time index k from observed values at $k = 1, \ldots, n$. This is crucial, since the approach has rather bad *extrapolating* properties, see, e.g., [8]. On the other hand, prediction by kriging has nice *interpolating* properties: we only need the training sample $\{\mathbf{x}_k, y_k\}$, $k = 1, \ldots, n$, to be representative of the data to be predicted. Kriging then yields accurate predictions even for extremely short training sequences, which does not seem to be the case for neural networks.

Consider first the case of a deterministic system (no observation errors), where

$$y_k = y(\mathbf{x}_k) = F(\mathbf{x}_k)\,, \tag{1}$$

with $F(\cdot)$ an unknown nonlinear function, and $\mathbf{x}_k$ the vector formed by lagged scalar inputs, that is:

$$\mathbf{x}_k = (x_k, x_{k-1}, \ldots, x_{k-m_x+1})^T\,.$$

The observations y_k are modeled by

$$y_k = \mathbf{f}^T(\mathbf{x}_k)\beta + Z(\mathbf{x}_k)\,, \qquad (2)$$

where the regressor $\mathbf{f}(\mathbf{x}_k)$ is function of $\mathbf{x}_k$, $\beta \in I\!R^p$ is a vector of unknown parameters and $Z(\mathbf{x}_k)$ is a realization of a stochastic process. *Bayesian kriging* corresponds to the case where a prior distribution is put on β, see [5], and will not be considered here. In practise, it is generally enough to take $\mathbf{f}(\mathbf{x}_k) = 1$ and β scalar. The process $Z(\cdot)$ is assumed to have zero mean and covariance

$$E\{Z(\mathbf{x})Z(\mathbf{x}')\} = W(\mathbf{x}, \mathbf{x}')\,.$$

We assume spatial stationarity, that is

$$W(\mathbf{x}, \mathbf{x}') = V(\mathbf{x} - \mathbf{x}') = \sigma_Z^2 R(\mathbf{x} - \mathbf{x}')\,,$$

with $R(\mathbf{x}) = R(-\mathbf{x})$. We use below

$$R(\mathbf{x} - \mathbf{x}') = \exp\left(\sum_{i=1}^{m_x} -\theta_i |x_i - x_i'|^{\gamma_i}\right)\,, \qquad (3)$$

which is typical. The function $R(.)$ is continuous at $\mathbf{0}$, which corresponds to a process continuous in the mean–square sense. The case $\gamma_i = 1$, $i = 1, \ldots, m_x$, corresponds to the product of Ornstein–Uhlenbeck processes, which are continuous but not differentiable everywhere. When $\gamma_i = 2$, $i = 1, \ldots, m_x$, the process has infinitely differentiable paths (in the mean–square sense). A classical assumption is $\gamma_i \in [1, 2]$, $i = 1, \cdots, m_x$. The choice of the functional form of the covariance is important, since it influences the predictive capacity of the method. We found that the form (3) allows enough flexibility through the parameters θ_i and γ_i (see [1]) and generally gives satisfactory results. Let $\mathbf{y}_n$ denote the vector of observations in the training sample,

$$\mathbf{y}_n = (y_1, \cdots, y_n)^T\,,$$

and define $\mathbf{F}_n$ as

$$\mathbf{F}_n = \begin{pmatrix} \mathbf{f}^T(\mathbf{x}_1) \\ \vdots \\ \mathbf{f}^T(\mathbf{x}_n) \end{pmatrix}.$$

The prediction $y(\mathbf{x})$ at a given value $\mathbf{x}$ is $\hat{y}(\mathbf{x}) = \mathbf{c}^T(\mathbf{x})\mathbf{y}_n$. Minimizing the mean–square error of this linear predictor under the unbiasedness condition $\mathbf{f}^T(\mathbf{x}) = \mathbf{c}^T(\mathbf{x})\mathbf{F}_n$ gives:

$$\hat{y}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\beta} + \mathbf{r}^T(\mathbf{x})\mathbf{V}_n^{-1}(\mathbf{y}_n - \mathbf{F}_n\hat{\beta})\,, \qquad (4)$$

where $\mathbf{V}_n = \sigma_Z^2 \mathbf{R}_n$ is the covariance matrix for $\mathbf{Z}_n = (Z(\mathbf{x}_1), \ldots, Z(\mathbf{x}_n))^T$, with

$$[\mathbf{R}_n]_{ij} = R(\mathbf{x}_i - \mathbf{x}_j)\,, \qquad (5)$$

$\mathbf{r}(\mathbf{x}) = E\{Z(\mathbf{x})\mathbf{Z}_n\}$, that is $[\mathbf{r}(\mathbf{x})]_i = \sigma_Z^2 R(\mathbf{x} - \mathbf{x}_i)$, and where

$$\hat{\beta} = (\mathbf{F}_n^T \mathbf{R}_n^{-1} \mathbf{F}_n)^{-1} \mathbf{F}_n^T \mathbf{R}_n^{-1} \mathbf{y}_n \qquad (6)$$

is the Least–Squares estimator for β. The mean–square error for the prediction is then

$$\sigma^2(\mathbf{x}) = \sigma_Z^2 - [\mathbf{f}^T(\mathbf{x})\ \mathbf{r}^T(\mathbf{x})] \begin{bmatrix} \mathbf{O} & \mathbf{F}_n^T \\ \mathbf{F}_n & \mathbf{V}_n \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{f}(\mathbf{x}) \\ \mathbf{r}(\mathbf{x}) \end{bmatrix}, \qquad (7)$$

which satisfies $\sigma^2(\mathbf{x}_k) = 0$, $k = 1, \ldots, n$. This means that this predictor is a perfect interpolator: $\hat{y}(\mathbf{x}_k) = y_k$, $k = 1, \ldots, n$. Assuming a normal distribution for the process $z(\mathbf{x})$, confidence intervals can be constructed for the prediction. For instance:

$$\text{Prob}\{y(\mathbf{x}) \in [\hat{y}(\mathbf{x}) - 1.96\,\sigma(\mathbf{x}),\ \hat{y}(\mathbf{x}) + 1.96\,\sigma(\mathbf{x})]\} \simeq 0.95\,. \qquad (8)$$

The prediction $\hat{y}(\mathbf{x})$ depends on the parameters θ_i and γ_i of the covariance function (3). Assuming that the stochastic process $Z(\cdot)$ is Gaussian, they can be estimated by maximum likelihood, together with β and σ_Z^2. Elementary calculations give:

$$\{\hat{\theta},\ \hat{\gamma}\} = \arg \min_{\{\theta \in I\!R^{+m_x},\ \gamma \in [1,2]^{m_x}\}} [n \ln(\hat{\sigma}_Z^2) + \ln\det(\mathbf{R}_n)]\,, \qquad (9)$$

where $\hat{\sigma}_Z^2 = \frac{1}{n}(\mathbf{y}_n - \mathbf{F}_n\hat{\beta})^T \mathbf{R}_n^{-1}(\mathbf{y}_n - \mathbf{F}_n\hat{\beta})$ and $\hat{\beta}$ given by (6) respectively correspond to the maximum likelihood estimators of σ_Z^2 and β.

Assume now that

$$y_k = F(\mathbf{x}_k) + \epsilon_k\,,$$

with $\{\epsilon_k\}$ an i.i.d. sequence of errors with zero mean and variance σ_ϵ^2. The observations are modeled as

$$y_k = \mathbf{f}^T(\mathbf{x}_k)\beta + Z(\mathbf{x}_k) + \epsilon_k\,, \qquad (10)$$

where $Z(\cdot)$ is a stochastic process independent of $\{\epsilon_k\}$. Define $\mathbf{V}_n = \sigma_\epsilon^2 \mathbf{I}_n + \sigma_Z^2 \mathbf{R}_n$, with $\mathbf{I}_n$ the n-dimensional identity matrix and $\mathbf{R}_n$ given by (5). The prediction at $\mathbf{x}$ is then still given by (4), with now

$$\hat{\beta} = (\mathbf{F}_n^T \mathbf{V}_n^{-1} \mathbf{F}_n)^{-1} \mathbf{F}_n^T \mathbf{V}_n^{-1} \mathbf{y}_n\,, \qquad (11)$$

which coincides with (6) when $\sigma_\epsilon^2 = 0$. When $\sigma_\epsilon^2 \neq 0$, this predictor is not a perfect interpolator: the mean–square error for the prediction is given by (7) and, in general, $\sigma^2(\mathbf{x}_k) \neq 0$ for $k = 1, \ldots, n$. Assuming that ϵ_k is normal $\mathcal{N}(0, \sigma_\epsilon^2)$ and $Z(\cdot)$ is also Gaussian, one can still use maximum likelihood to estimate the parameters of $\mathbf{R}$ together, with β, σ_Z^2

and σ_ϵ^2. Define α as $\alpha = \frac{\sigma_\epsilon^2}{\sigma_Z^2}$, so that $\mathbf{V}_n = \sigma_Z^2(\mathbf{R}_n + \alpha \mathbf{I}_n)$. The maximum likelihood estimator of α and θ is given by

$$\begin{aligned}\{\hat\alpha,\ \hat\theta,\ \hat\gamma\} \quad = \quad & \arg \min_{\{\alpha>0, \theta\in \mathbb{R}^{+m_x}, \gamma\in[1,2]^{m_x}\}} [n \ln(\hat\sigma_Z^2) + \\ & \ln\det(\mathbf{R}_n + \alpha\mathbf{I}_n)] \,, \qquad (12)\end{aligned}$$

where

$$\hat\sigma_Z^2 = \frac{1}{n}(\mathbf{y}_n - \mathbf{F}_n\hat\beta)^T(\mathbf{R}_n + \alpha\mathbf{I}_n)^{-1}(\mathbf{y}_n - \mathbf{F}_n\hat\beta)\,,$$

and $\hat\beta$ given by (11) respectively correspond to the maximum likelihood estimators of σ_Z^2 and β.

Numerical optimization methods are required for the determination of $\hat\theta$ and $\hat\gamma$ in (9), or $\hat\theta$, $\hat\gamma$ and $\hat\alpha$ in (12). Although the problem is sometimes difficult (see e.g. [9]), numerical simulations show that a precise determination of the estimates is not necessary to get an accurate prediction, and local optima are generally acceptable. The derivatives of the likelihood functions in (9,12) are easily obtained, and local search methods (conjugate gradients or quasi–Newton), can be used efficiently. Imposing constraints on θ, such as $\theta_i \geq \delta > 0$, is recommended to preserve the positive–definite character of $\mathbf{R}_n$ during the optimization. Note that the value of θ_i indicates the importance of the ith input of the model, so that the method permits to screen out important input factors.

3. APPLICATION TO SYSTEM INVERSION

We consider the situation where observations $\mathbf{y}_k$ satisfy an input/ouput relationship of the form

$$\begin{aligned} y_k &= \varphi(z_k) + v_k \\ z_k &= \sum_{i=0}^{n_a} a_i x_{1_{k-i}} + \sum_{i=0}^{n_b} b_i x_{2_{k-i}} \end{aligned}$$

where $\varphi(.)$ is a static nonlinearity :

$$\varphi(z) = 2/[1 + \exp(-10z)] - 1\,, \qquad (13)$$

see Figure 1, $\{x_{1_k}\}$, $\{x_{2_k}\}$ are input sequences and $\{v_k\}$ is an i.i.d. sequence $\mathcal{N}(0, \sigma_v^2)$.

We assume that a training sequence $\{x_{1_k}\}$, $\{x_{2_k}\}$, $\{y_k\}$, $k = 1, 2, \ldots, n$, is available, and we wish to invert the system and predict x_{2_k} as a function of y_k, y_{k-1}, $\ldots$, x_{1_k}, $x_{1_{k-1}}, \ldots$ for $k > n$. We model x_{2_k} as

$$x_{2_k} = \beta + z(\underline{\mathbf{x}}) + \varepsilon_k \qquad (14)$$

with

$$\underline{\mathbf{x}} = (y_k, y_{k-1}, \ldots, y_{k-m_y+1}, x_{1_k}, x_{1_{k-1}}, \ldots, x_{k-m_x+1})\,,$$

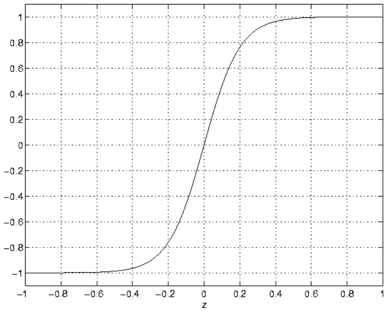

Figure 1: Static nonlinearity

Linear	Volterra	Kriging
-8.5	+3.2	-16.5

Table 1: Normalized mean–square error E_r (dB)

and predict $\hat x_{2_k}$ by kriging. Note that when $n_b > 0$, large values of m_y and m_{x_1} may be required since x_{2_k}, modeled as a function of z_k and x_{1_k}, contains an autoregressive part.

The performances are evaluated in terms of the normalized mean–square error E_r:

$$E_r = 10 \log \frac{(\mathbf{x}_{2_{n+1}}^N - \hat{\mathbf{x}}_{2_{n+1}}^N)^T(\mathbf{x}_{2_{n+1}}^N - \hat{\mathbf{x}}_{2_{n+1}}^N)}{(\mathbf{x}_{2_{n+1}}^N)^T\mathbf{x}_{2_{n+1}}^N}\,, \qquad (15)$$

where $\mathbf{x}_{2_{n+1}}^N$ denotes the vector of observations $(x_{2_{n+1}}, \ldots, x_{2_N})$ and $\hat{\mathbf{x}}_{2_{n+1}}^N$ denotes the vector of predictions $(\hat x_{2_{n+1}}, \ldots, \hat x_{2_N})$. Table 1 corresponds to the case where $\sigma_v^2 = 0$ (no observation errors), x_{1_k} and x_{2_k} are distributed $\mathcal{N}(0,1)$, $\sigma_\varepsilon^2 = 0$, $n = 50, N = 250, n_a = n_b = 2$,

$$\begin{aligned} a &= [0.08, -0.06]\,, \\ b &= [0.09, -0.02]\,, \end{aligned}$$

and $m_y = m_{x_1} = 4$. For the kriging predictor we use $f = 1$, $m_x = m_y + m_{x_1} = 8$ and $\gamma_i = 2$, $i = 1, \ldots, 8$. The results are averaged over 10 independent realisations.

The Volterra filter used for comparison is of degree 2 and contains 45 parameters to be estimated. This large number of parameters compared to the length of the training sequence ($n = 50$) explains the poor performances of this predictor compared to the linear one, with only 9 parameters to be estimated. Prediction by kriging clearly outperforms these two approaches. Note that the NARMAX model cannot be used here, due to the short length of the training sequence and the number of independent variables in the model ($m_y + m_{x_1} = 8$).

Figure 2 gives typical training sequences $\{x_{2_k}\}$ and $\{z_k\}$. Note that a large number of samples fall in the nonlinear part of $\varphi(.)$.

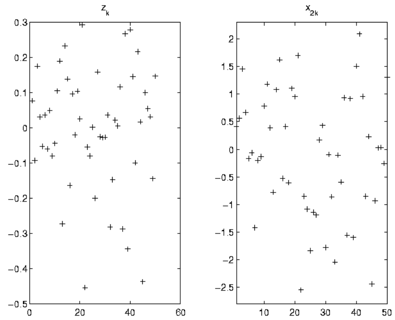

Figure 2: A typical training sequence

Figures 3 and 4 give $\hat{\mathbf{x}}_2$ as a function of $\mathbf{x}_2$, $k = n + 1, \ldots, N$, for prediction by kriging and a linear model.

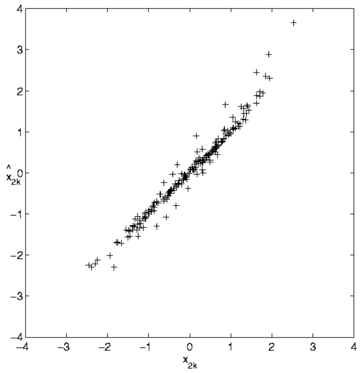

Figure 3: kriging

4. CONCLUSIONS

Kriging seems to be an attractive method in nonlinear filtering problems. The presence of a non–parametric part in the model allows a great flexibility, and choosing the parametric part as a simple constant generally gives satisfactory results, even in situations where the model is highly nonlinear and the training sequence is short. Inversion of a nonlinear system has been considered, with the ouput of the system depending on two inputs, one being known, the other to be reconstructed. Further developments will concern extension to multidimensional predictions, with application to simultaneous reconstruction of several inputs.

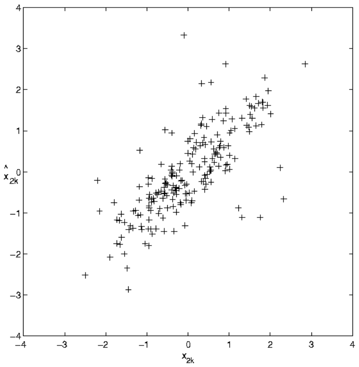

Figure 4: linear model

5. REFERENCES

[1] J.-P. Costa, L. Pronzato, and E. Thierry. Nonlinear predicting by kriging, with application to noise cancellation. *Internal report*, 1998.

[2] N. Cressie. Kriging nonstationary data. *Journal of the American Statistical Association*, 81:625-634, 1986.

[3] D. Krige. A statistical approach to some mine valuation and allied problems on the Witwatersrand. Master Thesis, University of Witwatersrand, 1951.

[4] I. Leontaritis and S. Billings. Input-Output parametric models for nonlinear systems part 2 : stochastic nonlinear systems. *International Journal of Control*, 41(2):329–344, 1985.

[5] R. Liebers. What can be done by Bayesian Kriging? *Tatra Mountains Mathematical Publications*, 7:275–282, 1996.

[6] W. J. Rugh. *Nonlinear System Theory : The Volterra / Wiener Approach.* The Johns Hopkins University Press, Baltimore, 1981.

[7] J. Sacks, W. Welch, T. Mitchell, and H. Wynn. Design and analysis of computer experiments. *Statistical Science*, 4(4):409–435, 1989.

[8] E. Walter and L. Pronzato. *Identification of Parametric Models from Experimental Data.* Springer, Heidelberg, 1997.

[9] J. Warnes and B. Ripley. Problems with likelihood estimation of covariance functions of spatial gaussian processes. *Biometrika*, 74(3):640–642, 1987.