A COMPARISON BETWEEN KRIGING AND RADIAL BASIS FUNCTION NETWORKS FOR NONLINEAR PREDICTION

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
{*costa, pronzato, et*}*@i3s.unice.fr*

ABSTRACT

Predictions by Kriging and radial basis function (RBF) networks with gaussian Kernels are compared. Kriging is a semi–parametric approach that does not rely on any specific model structure, which makes it much more flexible than approaches based on parametric behavioural models. On the other hand, accurate predictions are obtained for short training sequences, which is not the case for nonparametric prediction methods based on neural networks. Examples are presented to illustrate the effectiveness of the method.

1. INTRODUCTION

We consider the situation where the relationship beetween the input and output sequences $\{x_k\}$ and $\{y_k\}$ of a SISO system S is dominated by nonlinear characteristics. When using a parametric nonlinear model, first one has to choose a suitable model structure [1]. Second, once a structure has been chosen, one has to estimate its parameters. Traditional parametric representations for nonlinear unknown structures are the Volterra, Wiener or NARMAX (Nonlinear AutoRegressive Moving Average model with eXogenous inputs) models. They generally involve a very large number of unknown parameters [2, 3], so that one has to collect a large amount of data (training sequence) to be able to estimate these parameters.

Linear prediction by Kriging can be considered as a general statistical tool for modeling spatial observations, with or without observation errors [4, 5, 6]. Kriging is based on a semi–parametric model which allows much more flexibility than parametric models, since no specific model structure is used : the model contains a linear regression part (parametric) and a non–parametric part considered as the realization of a random process. Assuming that the process is Gaussian, the parameters of its covariance matrix can be estimated by maximum likelihood. It happens that the choice of the linear regression has little influence on the predictive properties of the model see, e.g., [7]. The memory length m_x of the input thus corresponds to the only important prior choice concerning the unknown structure, and a prior over–estimation of m_x only results in heavier computations.

An alternative approach to parametric, or semi–parametric, models is to use a nonparametric model. Nonparametric approaches based on neural networs, for instance Radial Basis Function (RBF) networks, are becoming more and more popular. A reason is that, in principle, neural networks can approximate any continuous behaviour with arbitrary precision [8]. However, besides the problem of having to choose the structure of the network, remains the major issue of choosing a suitable (and long enough) training sequence [9].

If the linear part of the Kriging model is reduced to a constant term, Kriging corresponds to a RBF network with gaussian kernels, with the centers of the kernels corresponding to the inputs in the training data set, see section 3.2.

2. PREDICTION BY KRIGING

Let $\{\mathbf{x}_k\}$ and $\{y_k\}$ be the input and output sequences of a system S, which are observed for $k = 1, \ldots, n$. Prediction by Kriging consists in interpolating these data by constructing the best linear unbiased predictor at new unsampled values of $\mathbf{x}$, that is $\mathbf{x}_{n+1}, \mathbf{x}_{n+2}, \ldots$

2.1. Prediction

When there are no observation errors, the observations y_k are modelled by

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

where $\mathbf{x}_k$ denotes the vector formed by lagged scalar inputs

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

$\mathbf{f}(\mathbf{x}_k)$ is the linear regressor with $\beta \in I\!R^p$ the vector of unknown parameters, and $z(\mathbf{x}_k)$ is a realization of a stochastic process. When $\mathbf{f}(\mathbf{x}) = 1$, which is often suitable, the method is usually called *simple Kriging* [10]. We shall see in section 4 that the choice of the regressor is not crucial.

The process $z(\cdot)$ is assumed to have zero mean and covariance defined by

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

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 $R(\mathbf{x}) = R(-\mathbf{x})$. A typical choice is

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

The function $R(.)$ is continous at $\mathbf{0}$, which corresponds to a process continuous in the mean–square sense. The choice of the functional form of the covariance is important, since it influences the predictive ability of the method. The form (1) allows enough flexibility through the parameters θ_i and γ_i, which correspond respectively to a correlation and smoothness parameter (see [11]). Other covariance functions are considered, e.g., in [12]. Let $\mathbf{y}_n$ denote the vector of observations in the training sample,

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

and define $\mathbf{F}_n$ as

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

We predict $y(\mathbf{x})$ at a given value of $\mathbf{x}$ by $\hat{y}(\mathbf{x}) = \mathbf{c}^T(\mathbf{x})\mathbf{y}_n$. One can show [12] that minimizing the mean–square error of this linear predictor under the unbiasedness condition

$$\mathbf{f}^T(\mathbf{x}) = \mathbf{c}^T(\mathbf{x})\mathbf{F}_n \,,$$

one gets

$$\hat{y}(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\hat{\beta} + \mathbf{r}^T(\mathbf{x})\mathbf{V}_n^{-1}(\mathbf{y}_n - \mathbf{F}_n\hat{\beta}) \,, \qquad (2)$$

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

$$[\mathbf{R}_n]_{ij} = R(\mathbf{x}_i - \mathbf{x}_j) \,, \qquad (3)$$

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

$$\hat{\beta} = (\mathbf{F}_n^T \mathbf{R}_n^{-1} \mathbf{F}_n)^{-1} \mathbf{F}_n^T \mathbf{R}_n^{-1} \mathbf{y}_n \qquad (4)$$

is the Least–Squares estimator for β. This predictor is a perfect interpolator: $\hat{y}(\mathbf{x}_k) = y_k, k = 1, \ldots, n$. The mean–square error for the prediction is

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

It satisfies $\sigma^2(\mathbf{x}_k) = 0$, $k = 1, \ldots, n$. Assuming a normal distribution for the process $z(\mathbf{x})$, confidence intervals can be constructed for the prediction. For instance, a 95% confidence interval is given by

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

When observation errors are present, the observations are modelled as

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

with $\{\epsilon_k\}$ an i.i.d. sequence of errors with zero mean and variance σ_ϵ^2 and $z(\cdot)$ 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 (3). The prediction at $\mathbf{x}$ is then still given by (2). When $\sigma_\epsilon^2 \neq 0$, this predictor is not a perfect interpolator.

We assume in the rest of this paper that observation errors are negligible.

2.2. Estimation

The prediction $\hat{y}(\mathbf{x})$ depends on the parameters θ_i and γ_i in the covariance function (1). 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 $\gamma_i \in [1, 2]$, $i = 1, \cdots, m_x$. Assuming that the stochastic process $z(\cdot)$ is Gaussian, one can estimate the θ_i's and γ_i's by maximum likelihood, together with β and σ_z^2. Elementary calculations give:

$$\{\hat{\theta},\ \hat{\gamma}\} = \arg \min_{\{\theta \in I\!R^{+\,m_x},\ \gamma \in [1,2]^{m_x}\}} [n \ln(\hat{\sigma}_z^2) + \ln\det(\mathbf{R}_n)] \,, \qquad (5)$$

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 (4) respectively correspond to the maximum likelihood estimators of σ_z^2 and β.

Numerical optimization methods are required for the solution of (5). The problem is sometimes difficult (see e.g. [13]), but numerical simulations show that a precise determination of the estimates is not necessary to get an accurate prediction. In particular, local optima are generally acceptable. It is recommended in practice to impose constraints on θ, such as $\theta_i \geq \delta > 0$ to preserve the positive–definite character of $\mathbf{R}_n$ during the optimization. Freezing the γ_i's at 2 is often acceptable.

3. RADIAL BASIS FUNCTION NETWORK

3.1. Structure of RBF network

A radial basis function network consists of an input layer of source nodes, a single hidden layer of nonlinear processing units, and an output layer of linear weights, as depicted in Fig. 1.

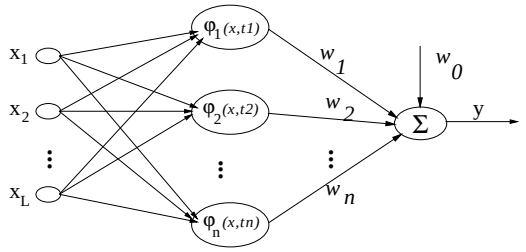

Figure 1: RBF network

Using the terminology of this figure, we may describe the input–output mapping performed by the RBF network as follows:

$$y(x) = w_0 + \sum_{i=1}^{n} w_i \varphi_i(\mathbf{x}; \mathbf{t}_i)\,, \tag{6}$$

where the term $\varphi_i(\mathbf{x}; \mathbf{t}_i)$ is the ith radial basis function that computes the distance between the input vector $\mathbf{x}$ and the center $\mathbf{t}_i$. Gaussian kernels are the most commonly used in pratice. When the centers $\mathbf{t}_i$ correspond to the inputs $\mathbf{x}_i$ in the training data set, $i = 1, \ldots, n$, one gets

$$y(x) = w_0 + \sum_{i=1}^{n} w_i \exp(-\frac{1}{\sigma_i^2}||\mathbf{x} - \mathbf{x}_i||^2) \tag{7}$$

where σ_i is the width of the ith radial basis function [14], and is fixed by the user. The parameters w_i can be estimated by least squares.

3.2. Relation between RBF and Kriging

If the linear part of the Kriging model is reduced to a constant term $\mathbf{f}(.) = 1$,

$$\mathbf{F}_n = \mathbf{1} = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}.$$

$\hat{\beta} = \frac{\mathbf{1}^T \mathbf{V}_n^{-1}}{\mathbf{1}^T \mathbf{V}_n^{-1} \mathbf{1}} \mathbf{y}_n$, see (4), and the equation (2) can be rewritten as

$$\hat{y}(\mathbf{x}) = \hat{\beta} + \mathbf{r}^T(\mathbf{x})(\mathbf{I}_n - \frac{\mathbf{V}_n^{-1}\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{V}_n^{-1}\mathbf{1}})\mathbf{V}_n^{-1}\mathbf{y}_n$$

that is

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{i=1}^{n} \hat{w}_i R(\mathbf{x} - \mathbf{x}_i) \tag{8}$$

where $\hat{\mathbf{w}} = [\hat{w}_1 \ \ldots \ \hat{w}_n]^T = (\mathbf{I}_n - \frac{\mathbf{V}_n^{-1}\mathbf{1}\mathbf{1}^T}{\mathbf{1}^T\mathbf{V}_n^{-1}\mathbf{1}})\mathbf{V}_n^{-1}\mathbf{y}_n$, $\hat{w}_0 = \hat{\beta}$ and $\varphi(\mathbf{x}; \mathbf{x}_i) = R(\mathbf{x} - \mathbf{x}_i)$. Prediction by Kriging (8) has thus the same expression than the prediction by RBF with gaussian Kernels (7). However, the parameters σ_i^2 in (7) are fixed, whereas the θ_i's in (1) are estimated.

The examples of the next section compare the two approaches. As mentionned in [15], such a comparison between different approaches is important but intrinsically difficult. We consider, in particular, the influence of the training sequence on the quality of the prediction. For short training sequences Kriging shows better performance than RBF.

4. EXAMPLES

The examples presented correspond to simulated data. No measurement errors are added to the observations.

EXAMPLE 1. The intput–output relationship of the system is presented in Fig. 2 and given by

$$y_k = \text{sinc}(\sqrt{ax_{1k}^2 + bx_{2k}^2}\,),$$

with $a = 4$, $b = 2$.

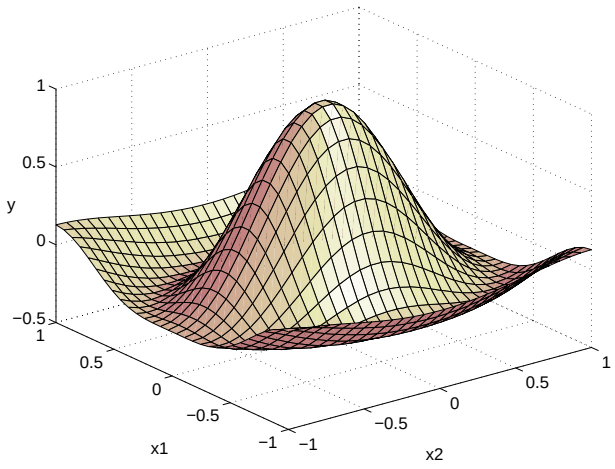

Figure 2: Nonlinear system

EXAMPLE 2. The observations are given by

$$y_k = \text{sinc}(\sqrt{ax_{1k}^2})\, G(x_{2k})$$

see Fig. 3, where $G(.)$ is a nonlinear static function which

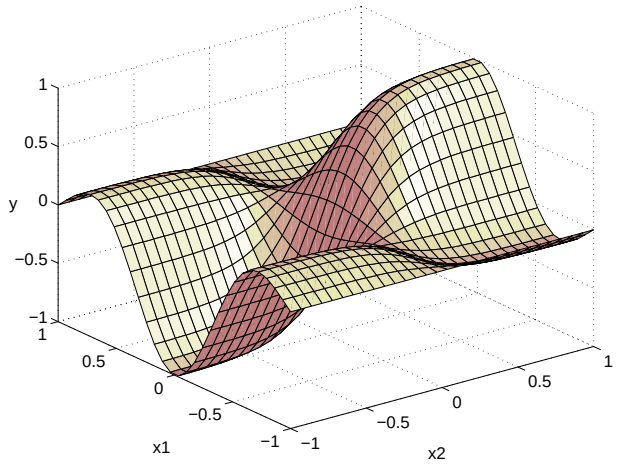

Figure 3: Nonlinear system

corresponds to a saturation, see Fig. 4,

$$G(x) = \frac{2}{1+\exp(-\alpha x)} - 1\,. \qquad (9)$$

We take $a = 4$, $\alpha = 7$.

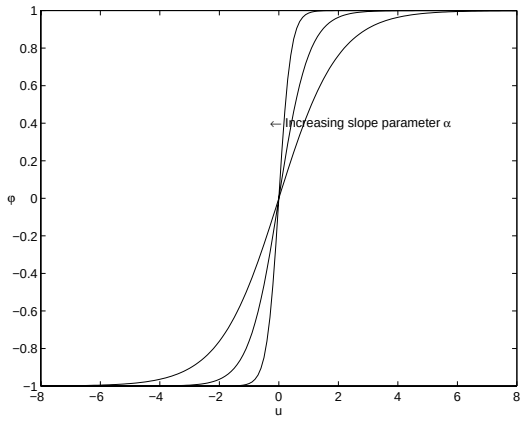

Figure 4: Sigmoid function

We consider a training sequence $\{\mathbf{x}_k = (x_{1k} = x_k, x_{2k} = x_{k-1}), y_k\}$ of length n, $15 \leq n \leq 50$, with $\{x_k\}$ i.i.d. uniformly in $[-1, 1]$.
The prediction is made over an horizon of 1000, that is for $y_{n+1}, \ldots, y_{n+1000}$. The same training sequence is used for the two approaches, Kriging and RBF.

For prediction by Kriging, the regressor $\mathbf{f}(.)$ is reduced to a constant term $\mathbf{f}(\mathbf{x}) = 1$. The correlation matrix is given by (1) with $\gamma_1 = \gamma_2 = 2$.

For prediction by RBF, the widths σ_i of the gaussian kernels are fixed to 1, $i = 1, \ldots, n$. The parameters w_i are estimated by least squares.

The results obtained with the two approaches are summarized in Table 1 and 2, which give the mean ($< \ . \ >$) and the standard deviation (std(.)) of the normalized mean–square error E_r over 10 independant repetitions:

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

with $\mathbf{y}_{n+1}^N$ the vector of observations $(y_{n+1}, \ldots, y_N)$ and $\hat{\mathbf{y}}_{n+1}^N$ the vector of predictions $(\hat{y}_{n+1}, \ldots, \hat{y}_N)$.

	Kriging		RBF	
n	$< Er >$	std(Er)	$< Er >$	std(Er)
50	-41.36	3.13	-30.15	4.37
45	-37.58	5.06	-28.20	4.27
40	-31.78	3.71	-24.99	5.31
35	-27.69	2.49	-17.52	4.28
30	-24.92	2.71	-13.57	2.07
25	-20.36	4.60	-11.17	2.48
20	-13.17	4.15	-7.78	3.59
15	-7.06	4.25	-1.65	3.32

Table 1: Example 1

	Kriging		RBF	
n	$< Er >$	std(Er)	$< Er >$	std(Er)
50	-22.37	6.26	-13.17	3.78
45	-20.24	5.49	-12.12	4.86
40	-20.04	4.09	-10.58	4.73
35	-17.24	4.34	-9.07	2.82
30	-16.39	3.20	-7.99	2.58
25	-14.90	2.74	-7.23	4.76
20	-11.40	3.07	-3.99	3.56
15	-6.93	3.48	-0.85	2.13

Table 2: Example 2

Fig. 5 and 6 give y as a function of $\hat{y}$, for $k = n + 1, \ldots, N$, for prediction by Kriging and RBF, for a typical realization of example 2 with $n = 50$.

5. CONCLUSIONS

Predictions by Kriging and radial basis functions with gaussian Kernels have been compared. Although the predictors share the same structure, see (7) and (8), the results are

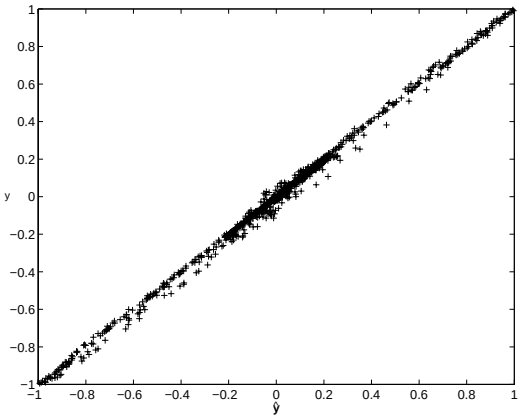

Figure 5: y versus $\hat{y}$ for Kriging

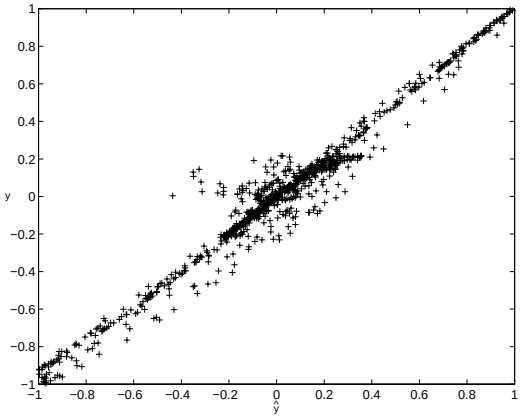

Figure 6: y versus $\hat{y}$ for RBF

much different.

The examples presented show that prediction by Kriging is much more accurate than prediction by RBF. This is due to the greater flexibility of Kriging, through the parameters of the covariance function of the stochastic process. The price for this increase of performance is the estimation of these parameters (the θ_i's in (1)), which requires the use of a nonlinear programming algorithm (sequential quadratic programming is used in the examples of section 4).

The construction of a *recursive* algorithm (recursive with respect to $k = 1, \ldots, n$ in the training sequence) for the solution of the estimation problem (5) will be the subject of further work.

6. REFERENCES

[1] S.A. Billings and S. Chen. Extented model set, global data and threshold model identification of severely nonlinear systems. *Int. J. Control*, 50(5):1897–1923, 1989.

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

[3] S.A. Billings and W. S. Voon. A prediction-error and stepwise-regression estimation algorithm for nonlinear systems. *Int. J. Control*, 44(3):803–822, 1986.

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

[5] G. Matheron. Principles of Geostatistics. *Economic Geology*, 58:1246–1266, 1963.

[6] N. Cressie. Kriging nonstationary data. *J. of the Amer. Statis. Assoc.*, 81:625-634, 1986.

[7] W.J. Welch, R.J. Buck, J. Sacks, H.P. Wynn, T.J. Mitchell, and M.D. Morris. Screening, predicting and computer experiments. *Technometrics*, 34(1):15–25, 1992.

[8] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[9] T. Poggio and F. Girosi. Networks for approximation and learning. *Neural Networks Proc. of the IEEE*, 78(9):1481–1497, 1990.

[10] N. Cressie. *Statistics for Spacial Data.* Wiley, 1993.

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

[12] J. Sacks, S.B. Schiller, and W.J. Welch. Designs for computer experiments. *Technometrics*, 31:41–47, 1989.

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

[14] S. Haykin. *Adaptive Filter Theory*. Third Edition, Prentice Hall, Inc, A Simon & Schuster Company, New Jersey, 1996.

[15] V. Cherkassky, D. Gehring, and F. Mulier. Comparison of adaptive methods for function estimation from samples. *IEEE Trans. on Neural Networks*, 7(4):969–984, 1996.