MULTI-OUTPUT SUPPPORT VECTOR REGRESSION

Emmanuel Vazquez * **and Eric Walter** *

* *Laboratoire des Signaux et Systèmes, 91192, Gif-sur-Yvette, France*
CNRS – Supélec – Université Paris-Sud

Abstract: Support vector regression builds a model of a process that depends on a set of factors. It traditionally considers one output at a time, which means that advantage cannot be taken of the correlations that may exist between outputs. The purpose of this paper is to show how the body of knowledge accumulated by geostatisticians on Kriging and its extensions over the last 40 years can help extend support vector regression to the multi-output case and provides guidance for the choice of a suitable kernel for a given application, a recurrent, fundamental and largely open question.

Keywords: Black-box modeling, Kriging, MIMO systems, Nonlinear regression, Regularization, Reproducing kernels, Robust estimation, Support vector machines

1. INTRODUCTION

Support Vector Regression (SVR) belongs to the category of reproducing-kernel methods, just as *splines* (Wahba, 1990), *Kriging* (Chiles and Delfiner, 1999) (also known under the name of Gaussian processes (Williams, 1997)) and methods involving *radial basis functions* (Schaback, 1999). Based on the theory of Support Vector Machines, SVR is now a well-established method for designing *black-box* models in engineering. A frequently felt difficulty with SVR, however, is the lack of guiding rules to choose a kernel for a given application. Kriging is seldom mentioned in the literature in this context, although it turns out to be very useful for this task (the kernel is viewed as the covariance of a second-order Gaussian process).

The aim of SVR is to build a model f_{sv} of the output of a process or system that depends on a set of factors. The values taken by these factors will be stored in a vector $\mathbf{x} \in \mathbb{R}^d$. SVR is traditionally used with only one output, and the multi-output case is then dealt with by modeling each output independently of the others. The purpose of this paper is to extend SVR to multi-output systems by considering the so-called *Cokriging* method (Chiles and Delfiner, 1999), which is the multi-variable version of Kriging. This will make it possible to take advantage of the possible correlations between the outputs to improve the quality of the predictions provided by the model.

First, the SVR framework will be presented in order to make notation clear. The link between a given kernel and the covariance of the corresponding Gaussian process will then be recalled in the case of a single-output system. In order to extend SVR to the multi-output case we shall then present Cokriging and show how it can be adapted to the framework of SVR.

2. CLASSICAL SVR

Support vector regression is based on n elementary experiments during which the vector of factors $\mathbf{x}$ is assigned values specified by the user – within an admissible domain $\mathscr{X}$ of $\mathbb{R}^d$ – and the resulting system output is measured. The value taken by $\mathbf{x}$ during the i-th elementary experiment, $i \in \{1, \cdots, n\}$, will be denoted by $\mathbf{x}_i$, and the corresponding (scalar) system output by $f(\mathbf{x}_i)$.

SVR builds a function

$$f_{\mathsf{sv}} : \begin{array}{ccc} \mathbb{R}^d & \to & \mathbb{R} \\ \mathbf{x} & \mapsto & (w, \psi(\mathbf{x}))_{\mathscr{F}} + b \end{array} \tag{1}$$

minimizing the functional

$$\frac{1}{2}||w||^2_{\mathscr{F}} + C\sum_{i=1}^{n}[f_{\mathsf{sv}}(\mathbf{x}_i) - f(\mathbf{x}_i)]_{\varepsilon}\,. \qquad (2)$$

In these equations, b is a *bias* term and $C \in \mathbb{R}_+$ is a parameter chosen by the user. $\psi(.) : \mathbb{R}^d \to \mathscr{F}$ is an application mapping the space of factors into a so-called *feature space* $\mathscr{F}$, typically a space of functions $\mathbb{R}^d \to \mathbb{R}$. Finally, $[\cdots]_\varepsilon$ is the ε-*insensitive* loss function. In practice, $\mathscr{F}$ and the functions $w \in \mathscr{F}$ and $\psi(\mathbf{x}) \in \mathscr{F}$ are deduced from a kernel $k(.,.) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ involved in the computation of a scalar product:

$$w = \sum_i w_i k(\mathbf{x}_i, .)\,, \qquad (3)$$

$$\psi(\mathbf{x}) = k(\mathbf{x}, .)\,, \qquad (4)$$

$$(w, \psi(\mathbf{x}))_{\mathscr{F}} = \sum_i w_i k(\mathbf{x}_i, \mathbf{x})\,. \qquad (5)$$

The functional (2) may be interpreted as a classical regularization approach to solving an ill-posed inverse problem, with its first term a *regularization term* (a functional of w). This kernel-based regularization has received a number of interpretations, within the theory of regularization itself (Smola *et al.*, 1998), and in the context of the theories of splines and Gaussian processes (Wahba, 1998; Jaakkola and Haussler, 1999). The second part of (2) is a *data fidelity term*, involving the ε-insensitive loss function. This loss function comes from the principle of best discrimination and the notion of optimal hyperplane (Vapnik, 1995) and a more direct interpretation is provided by the theory of robust regression (see, for instance, (Huber, 1981; Schölkopf and Smola, 2002)). The value given to $\varepsilon \in \mathbb{R}_+$ can be used to control the complexity of the model.

It is well known that the feature space $\mathscr{F}$ is a reproducing kernel Hilbert space (rkhs) where $k(.,.)$ plays the role of the reproducing kernel, with the following properties:

(1) $k(\mathbf{x},.) \in \mathscr{F}$,
(2) $f(\mathbf{x}) = (f, k(\mathbf{x},.))_{\mathscr{F}}, \forall f \in \mathscr{F}$,
(3) $k(\mathbf{x},\mathbf{y}) = \langle \delta_{\mathbf{x}}, k(.,\mathbf{y})\rangle_{\mathscr{F}^*,\mathscr{F}} = (\delta_{\mathbf{x}}, \delta_{\mathbf{y}})_{\mathscr{F}^*}$,

where $\mathscr{F}^*$ is the dual space of $\mathscr{F}$, and $\delta_{\mathbf{x}} \in \mathscr{F}^*$ the evaluation functional at $\mathbf{x}$.

The SVR feature space $\mathscr{F}$ will now be reinterpreted as a space of random variables in order to establish the link between SVR and Kriging.

3. THE LINK WITH KRIGING

3.1 *Kernels and covariances*

The connection between rkhs and random processes is now well documented. See (Wahba and Kimeldorf, 1970) in the context of splines, (Matheron, 1981) in the context of Kriging, and (Wahba, 1998) in the context of SVM.

Kriging (Chiles and Delfiner, 1999; Matheron, 1963) is a probabilistic method that can be used to approximate or interpolate data, just as SVR does but without the robust regression property provided by the ε-insensitive loss function. The observed data are considered as realizations of a second-order random process $F(\mathbf{x})$. Thus, for each experiment i, $f(\mathbf{x}_i)$ is a realization of the random variable $F(\mathbf{x}_i)$. The Kriging estimate is the *best linear unbiased prediction* (BLUP) of $F(\mathbf{x})$ in the space $\mathscr{H}_S$ generated by the random variables $F(\mathbf{x}_1), \cdots, F(\mathbf{x}_n)$. This estimate can be written as a linear combination

$$\hat{F}(\mathbf{x}) = \sum_{i=1}^{n} \lambda_{i,\mathbf{x}} F(\mathbf{x}_i)\,, \qquad (6)$$

and is obtained by orthogonal projection of $F(\mathbf{x})$ onto $\mathscr{H}_S$ (in an appropriate Hilbert space).

Equation (6) can also be written as a stochastic integral

$$\hat{F}(\mathbf{x}) = \int_{\mathbb{R}^d} F(\mathbf{x}) d\lambda(\mathbf{x})\,, \qquad (7)$$

where λ is a finite-support measure that can be formally written as $\lambda = \sum_{i=1}^{n} \lambda_i \delta_{\mathbf{x}_i}$, using the Dirac measure.

An rkhs $\mathscr{F}$ is a Hilbert space of functions in which the evaluation functional $\delta_{\mathbf{x}} : f \in \mathscr{F} \to f(\mathbf{x}) \in \mathbb{R}$ is continuous. A given element of $\mathscr{F}$ can be identified to an element λ of its dual, as well as to a random variable obtained as in (7), *i.e.*, an element of the Hilbert space $\mathscr{H}$ generated by linear combinations of $F(\mathbf{x})$.

The feature space $\mathscr{F}$ inherits the scalar product of $\mathscr{H}$. If the mean of $F(\mathbf{x})$ is zero, this scalar product is obtained from the covariance of $F(\mathbf{x})$ and $F(\mathbf{y})$, which is precisely the reproducing kernel of $\mathscr{F}$, *i.e.*,

$$k(\mathbf{x},\mathbf{y}) = \mathsf{cov}(F(\mathbf{x}), F(\mathbf{y}))\,. \qquad (8)$$

The feature space $\mathscr{F}$ can thus be viewed as a space of random variables endowed with a scalar product, *i.e.* a symmetric, strictly positive, bilinear functional equal to the covariance in this random space (remember that the mean of the random variables is assumed to be zero). The covariance structure (equivalently the kernel) $k(\mathbf{x},\mathbf{y})$ indicates how two values of the process output to be modeled are correlated depending on the corresponding values $\mathbf{x}$ and $\mathbf{y}$ taken by the factor vectors.

3.2 *Kernel choice*

3.2.1. *The Matérn class of kernels* Choosing a suitable kernel for a given application is a recurrent, fundamental and largely open question. The asymptotic theory of Kriging (Stein, 1999) stresses the importance of taking care of the behaviour of the covariance near the origin. This behaviour is indeed linked with the quadratic mean regularity of the random process. For instance, if the covariance is continuous at the

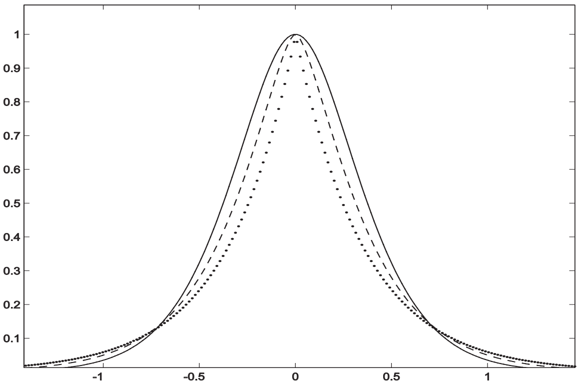

Figure 1. Matérn covariances for the parameterization (10) with $\rho = 0.5$, $\sigma^2 = 1$. Solid line corresponds to $\nu = 3$, dashed line to $\nu = 1$ and dotted line to $\nu = 0.5$.

origin, then the process will be continuous in quadratic mean. To address the issue of regularity, it is useful to consider a covariance model that is *isotropic*, *invariant by tranlation* and such that regularity can be adjusted with only one parameter. The *Matérn covariance* is a good candidate. Its Fourier transform is

$$\hat{k}(\omega) = \frac{\sigma^2}{(\omega_0^2 + \omega^2)^{\nu+1/2}}, \quad \omega \in \mathbb{R}, \tag{9}$$

where ν controls the decay of $\hat{k}(\omega)$ at infinity and therefore the regularity of the covariance. Stein (1999) advocates the use of the following parameterization of the Matérn model:

$$\begin{aligned} k(\mathbf{x}, \mathbf{y}) &= R(||\mathbf{x} - \mathbf{y}||) = R(h) \\ &= \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}h}{\rho}\right)^{\nu} \mathscr{K}_\nu \left(\frac{2\nu^{1/2}h}{\rho}\right), \end{aligned} \tag{10}$$

where $\mathscr{K}_\nu$ is the modified Bessel function of the second kind. This parameterization allows an easy interpretation of the parameters, as ν controls the regularity, σ^2 is the variance ($R(0) = \sigma^2$), and ρ represents the *range* of the covariance, *i.e.*, the characteristic correlation distance. Figure 1 shows the influence of the ν parameter. Using this kernel in applications turns out to be an improvement over the blind choice of a kernel among those classically used in SVR, such as Gaussian kernels. The next section addresses the estimation of the parameters of a given kernel from the data.

3.2.2. *Parameter estimation* In a probabilistic framework such as Kriging, the maximum-likelihood approach is one of the standard routes that can be followed to estimate the parameters of the covariance model. This approach can also be applied in the context of SVR. If the outputs are observed without noise, SVR can be interpreted as Kriging where the observed outputs $\mathbf{f} = [f(\mathbf{x}_1), \cdots, f(\mathbf{x}_n)]^\mathsf{T}$ are assumed to be realizations of a Gaussian vector $[F(\mathbf{x}_1), \cdots, F(\mathbf{x}_n)]^\mathsf{T}$ (a *high* value of C and $\varepsilon = 0$ is consistent with this hypothesis).

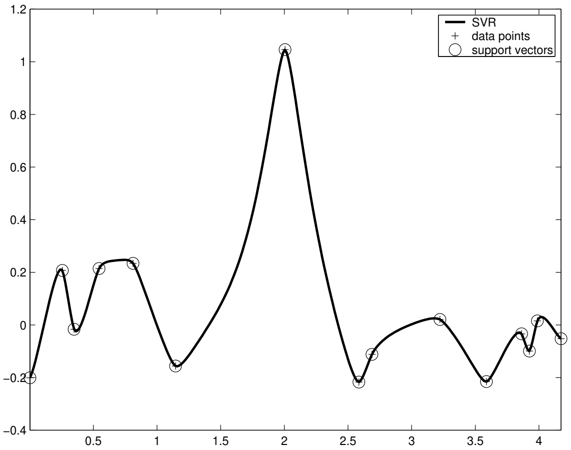

Figure 2. Plot of a SVR solution after maximum-likelihood estimation of the covariance parameters.

The maximum-likelihood estimate of the parameters of the covariance maximizes the probability density of the data. Using the joint probability density of the observed Gaussian vector, and assuming that the mean of $F(\mathbf{x})$ is zero, the maximum-likelihood estimate of the vector $\boldsymbol{\theta}$ of the covariance parameters is obtained (see for instance (Jones and Vecchia, 1993)) by minimizing the cost function

$$l(\boldsymbol{\theta}) = \frac{n}{2}\log 2\pi + \frac{1}{2}\log\det \mathbf{K}(\boldsymbol{\theta}) + \frac{1}{2}\mathbf{f}^\mathsf{T}\mathbf{K}(\boldsymbol{\theta})^{-1}\mathbf{f}, \tag{11}$$

where n is the number of observed output data and $\mathbf{K}(\boldsymbol{\theta})$ is the covariance matrix of the vector $[(F(\mathbf{x}_1), \cdots, F(\mathbf{x}_n)]^\mathsf{T}$.

Figure 2 illustrates an SVR with a Matérn kernel, the parameters of which have been estimated by maximum likelihood.

Assume now that the observations of the output are corrupted by noise. As a rough first approximation, assume moreover that the noise is white and Gaussian, although this hypothesis is not consistent with the use of the ε-insensitive function in SVR (see (Schölkopf and Smola, 2002)). For maximum-likelihood estimation, the observation noise must now be taken into account in the kernel structure as if it were part of the process. This simply means that a given kernel $k(\mathbf{h})$, where $\mathbf{h} = \mathbf{x} - \mathbf{y} \in \mathbb{R}^d$, is modified into:

$$k_\mathrm{n}(\mathbf{h}) = k(\mathbf{h}) + \sigma_\mathrm{n}^2 \delta_0(\mathbf{h}), \tag{12}$$

where $\delta_0(\mathbf{h})$ is the Dirac impulse at the origin and σ_n^2 is the noise variance. It is now possible to estimate the parameter σ_n^2 along with the other parameters of the covariance using the same maximum-likelihood formulation as above. Figures 3 and 4 illustrate the procedure. In this example, C is chosen by analogy with the splines–kriging equivalence where C is equal to $k(0)/\sigma_\mathrm{n}^2$ (Wahba, 1990). The choice of ε is related to the variance of the noise σ_n^2. We have set $\varepsilon = 0.3\sigma_\mathrm{n}$ ($0.3\sigma_\mathrm{n}$ corresponds to a confidence interval of approximatively 20% for a Gaussian variable).

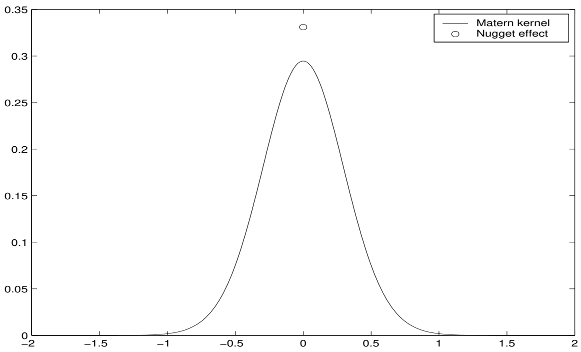

Figure 3. Matérn kernel with parameters estimated by maximum likelihood. The circle materializes the Dirac impulse at the origin corresponding to the white noise.

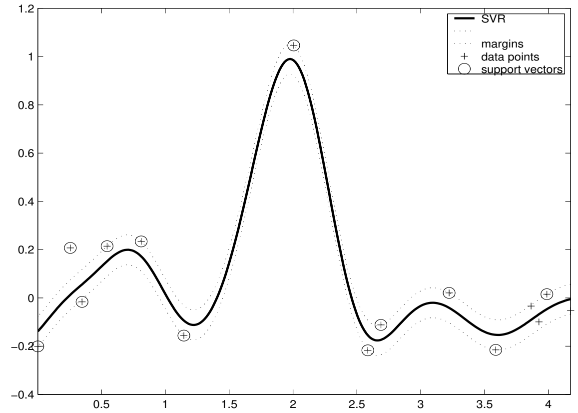

Figure 4. SVR solution after maximum likelihood estimation of the kernel parameters. The same data are used as in Figure 2.

To summarize, the properties of the reproducing kernel define the nature of the model. The Matérn covariance facilitates the tuning of the variance, regularity and range of correlation of the underlying random process describing the data. Moreover, these parameters can be estimated from the data in a unified framework by the maximum likelihood approach. Theses conclusions will be used again in the next sections when the multi-output case will be considered.

4. COKRIGING

Cokriging, the multi-output version of Kriging, exploits the correlations due to proximity in the space of factors (as Kriging does) but also the correlations between the outputs. Assume $F_1(\mathbf{x})$ is to be predicted from the observations of the random variables $F_\alpha(\mathbf{x}_i)$, $\alpha \in \Pi = \{1, \cdots, n\} \subset \mathbb{N}$ and $i \in S_\alpha \subset \mathbb{N}$ (we choose to index the random processes by Greek letters). With $n = 2$, prediction will based on the covariances and cross-covariances of $F_1(\mathbf{x})$ and $F_2(\mathbf{x})$, that is, on

$$k_{\alpha,\beta}(\mathbf{x}, \mathbf{y}) = \mathsf{cov}\, F_\alpha(\mathbf{x}) F_\beta(\mathbf{y}), \quad \alpha, \beta \in \{1, 2\}.$$

The estimator can again be written as a linear combination,

$$\hat{F}_1(\mathbf{x}) = \sum_{\alpha \in \Pi} \sum_{i \in S_\alpha} \lambda_{\alpha,i} F_\alpha(\mathbf{x}_i). \tag{13}$$

As for Kriging, $\hat{F}_1(\mathbf{x})$ is obtained as the best approximation of $F_1(\mathbf{x})$ on the space $\mathscr{H}_S$ generated by the variables $F_\alpha(\mathbf{x}_i)$. The random variables $F_\alpha(\mathbf{x})$ depend on $\alpha \in \Pi$, which can be considered as an additional factor. So with a slight change of notation we can write them as $F(\alpha, \mathbf{x})$ and write the covariance of $F(\alpha, \mathbf{x})$ and $F(\beta, \mathbf{y})$ as $k[(\alpha, \mathbf{x}), (\beta, \mathbf{y})]$. Thus, the multi-output case turns out to be formally equivalent to the single-output case. A single covariance function can then be used, since this function has been extended by means of the additional factor α to specify the nature of the random variable being considered.

5. MULTI-OUTPUT SVR

Using the above covariance $k[(\alpha, \mathbf{x}), (\beta, \mathbf{y})]$ as a kernel, it becomes straightforward to extend the feature space to deal with multi-output processes. If the bias term b is the same for each output of the process, then, a multi-output SVR builds a function

$$\begin{aligned} f_{\mathsf{sv}} : \Pi \times \mathbb{R}^d &\to \mathbb{R} \\ \alpha, \mathbf{x} &\mapsto (w, \psi(\alpha, \mathbf{x}))_{\mathscr{F}} + b \end{aligned} \tag{14}$$

minimizing the functional

$$\frac{1}{2} ||w||^2_{\mathscr{F}} + C \sum_{i,\alpha} [f_{\mathsf{sv}}(\alpha, \mathbf{x}_i) - f(\alpha, \mathbf{x}_i)]_\varepsilon \tag{15}$$

Note that the function $w \in \mathscr{F}$ will now be written as

$$w = \sum_{\beta,i} w_{\beta,i} k[(\beta, \mathbf{x}_i), (.,.)] \tag{16}$$

and the scalar product in (14) will be written as

$$(w, \psi(\alpha, \mathbf{x}))_{\mathscr{F}} = \sum_{\beta,i} w_{\beta,i} k[(\beta, \mathbf{x}_i), (\alpha, \mathbf{x})]. \tag{17}$$

SVR is thus easily extended to the multi-output case and the selection of the output to be predicted is performed with α in f_{sv}. This formulation has the additional advantage that existing optimization algorithms for SVR need not be modified, since the formalism remains unchanged. It suffices to design an appropriate kernel of the type of $k[(\alpha, \mathbf{x}), (\beta, \mathbf{y})]$, which amounts to choosing models of covariance and cross-covariance.

Linear combinations of the f_{sv} are also easily predicted. For instance,

$$f_{\mathsf{sv}}(1, \mathbf{x}) - f_{\mathsf{sv}}(2, \mathbf{x}) = (w, \psi(1, \mathbf{x}) - \psi(2, \mathbf{x}))_{\mathscr{F}}. \tag{18}$$

Until now, we have assumed that the bias term b was the same for each output of the process. To deal with unknown and possibly different means, a semi-parametric formulation of SVR is necessary (Smola *et al.*, 1999). Note that because of the presence of the term b in the classical formulation of SVR, this classical formulation can already be seen as a semi-parametric prediction. The easiest case is when the different means are assumed not to be correlated. Then, SVR is extended with parametric terms, such that,

$$\begin{aligned} f_{\mathsf{sv}} : \Pi \times \mathbb{R}^d &\to \mathbb{R} \\ \alpha, \mathbf{x} &\mapsto (w, \psi(\alpha, \mathbf{x}))_{\mathscr{F}} + \sum_{\beta} b_\beta \delta_\beta(\alpha) \end{aligned} \quad (19)$$

where $\delta_\beta(\alpha)$ is equal to one if $\beta = \alpha$ and zero if $\beta \neq \alpha$.

If the means are correlated, the previous formulation can still be used if there are enough observations of each output to make it possible to estimate the means independently. Finally, in this multi-output context, it might be interesting to have different data fidelity terms in (15) because the nature of noise on the each output might be different. This means that the data fidelity functional could be written as

$$\sum_{i,\alpha} C_\alpha [f_{\mathsf{sv}}(\alpha, \mathbf{x}_i) - f(\alpha, \mathbf{x}_i)]_{\varepsilon_\alpha}. \quad (20)$$

This formulation will be exploited in the next section, which will show some examples of multi-output SVR and briefly discuss methods to choose the covariance structure.

6. CROSS-COVARIANCE CHOICE

Choosing a cross-covariance is even more difficult than choosing a covariance. Consider first the proportional cross-covariance model, assuming again invariance by translation. Then,

$$\begin{aligned} &k_{1,1}(\mathbf{h}) = k_{2,2}(\mathbf{h}) = k(\mathbf{h}) = k(-\mathbf{h}), \quad \mathbf{h} \in \mathbb{R}^d, \\ &k_{1,2}(\mathbf{h}) = k_{2,1}(\mathbf{h}) = \gamma k(\mathbf{h}), \end{aligned} \quad (21)$$

where the scalar γ must be chosen in $]-1,1[$ since the Cauchy-Schwarz inequality implies

$$|k_{\alpha,\beta}(\mathbf{h})| \leq \sqrt{k_{\alpha,\alpha}(0) k_{\beta,\beta}(0)}. \quad (22)$$

Figure 5 illustrates the use of the proportional model. We have considered two (arbitrary) functions, representing hypothetic outputs of a process. The prediction was performed by adding the proportional kernel to an existing SV software (Gunn, 1997).

It is not necessary to have $k_{\alpha,\beta}(-\mathbf{h}) = k_{\alpha,\beta}(\mathbf{h})$ as in (21) (whereas $k_{\alpha,\beta}(\mathbf{h}) = k_{\beta,\alpha}(-\mathbf{h})$ always holds). Many possible covariance structures can actually be considered to take into account complex relations between outputs. A simple example of non-symmetric cross-covariance is obtained when considering a delay between the outputs, for instance, with a single factor x, $F_2(x) = F_1(x - \tau) + G(x)$, assuming $G(x)$ and $F_1(x)$ independent and $\mathsf{cov}[F_1(x), F_1(y)] = k_{1,1}(|x - y|)$. In such a case, the cross-covariance is $k_{1,2}(x,y) = k_{1,1}(|x - y + \tau|) \neq k_{1,2}(y,x) = k_{1,1}(|x - y - \tau|)$.

Many other possibilities of covariance structures can be used (Chiles and Delfiner, 1999). Often, one has

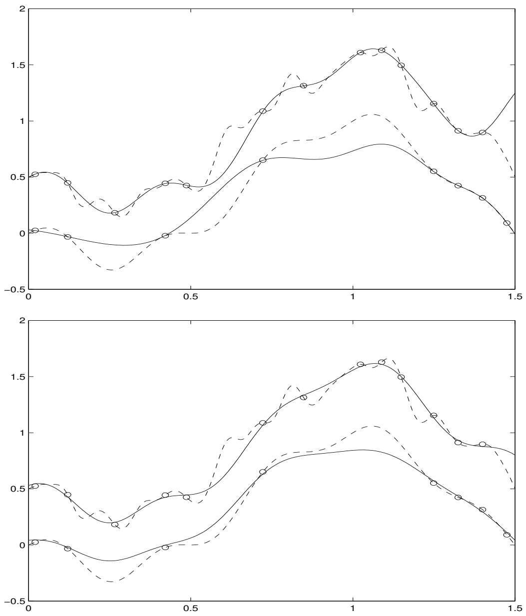

Figure 5. Multi-output SVR with a proportional model for the cross-covariance. The top plot shows independent predictions of the outputs (in solid lines). The circles materialize the data sampled from the dashed-line graphs. The bottom plot shows the improved results obtained with a cross-correlation coefficient $\gamma = 0.61$ estimated by maximum-likelihood.

some knowledge about the relations between the outputs. For instance the multi-output SVR framework can be used to take into account information about the derivatives of the process. This property is illustrated here in the simplest case, that is, when the space of factors is of dimension one, but generalization is straightforward. The inclusion of information about the derivatives is based on the consideration of two outputs, namely the one to be predicted, modeled by a random process $F(x)$, and its derivative. The derivative output is defined as

$$F'(x) = \lim_{h \to 0} \frac{F(x+h) - F(x)}{h} \quad (23)$$

From (23), and assuming that $F(x)$ is characterized by a twice differentiable radial covariance $k(h)$, $h \in \mathbb{R}_+$ the covariance structure of the multi-output model can be written as

$$\begin{aligned} &\mathsf{cov}[F(x_1), F'(x_2)] \\ &\qquad = \lim_{h \to 0} \frac{k(|x_1 - x_2 - h|) - k(|x_1 - x_2|)}{h} \\ &\qquad = -\,\mathsf{sgn}(x_1 - x_2)\, k'(|x_1 - x_2|) \end{aligned} \quad (24)$$

$$\mathsf{cov}[F'(x_1), F(x_2)] = \mathsf{sgn}(x_1 - x_2)\, k'(|x_1 - x_2|) \quad (25)$$

$$\mathsf{cov}[F'(x_1), F'(x_2)] = -k''(|x_1 - x_2|) \quad (26)$$

When Matérn covariance is used, the following property helps differentiate the modified Bessel function of the second kind :

$$\mathscr{K}'_\nu(x) = -\mathscr{K}_{\nu-1}(x) - \frac{\nu}{x} \mathscr{K}_\nu(x). \quad (27)$$

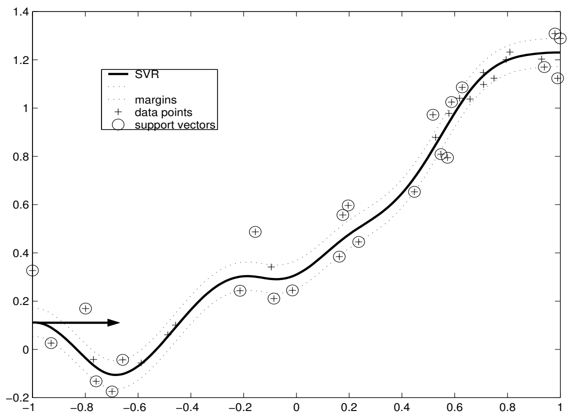

Figure 6. SVR with incorporation of prior information about the derivative of F. It is known that tangent at $x = -1$ should be horizontal.

This approach may use two types of data fidelity term, as in (20). Assume, for instance, that the available observations data on the output to be predicted are corrupted by noise, which leads to choosing $C \approx k(0)/\sigma_n^2$, but that it is known for physical reasons that the gradient must be strictly equal to zero at some boundary points. The prediction of the derivative of the output must then interpolate the constraints data, which implies $C \to \infty$, $\varepsilon = 0$. Figure 6 illustrates this approach.

7. CONCLUSIONS

The main contributions of this paper lie in the confrontation of the points of view of support vector machines and Kriging, and in the presentation of a multi-output support vector regression based on ideas that are now classical in the community of Kriging but perhaps less well known in that of SVMs.

We hope to have convinced the reader that the theory of Kriging provides useful guidelines for the solution of such fundamental problems as the choice of a kernel structure for support vector regression and the estimation of the parameters of this structure via a probabilistic interpretation of a given kernel as a correlation structure.

8. REFERENCES

Chiles, J.-P. and P. Delfiner (1999). *Geostatistics: Modeling Spatial Uncertainty*. Wiley. New York.

Gunn, S.R. (1997). Support vector machines for classification and regression. Technical report. Image Speech and Intelligent Systems Research Group. University of Southampton.

Huber, P. J. (1981). *Robust Statistics*. Wiley. New York.

Jaakkola, T. and D. Haussler (1999). Probabilistic kernel regression models. In: *Proceedings of the 1999 Conference on AI and Statistics*.

Jones, R. and A. Vecchia (1993). Fitting continuous ARMA models to unequally spaces spatial data. *J. of the Amer. Stat. Assoc.* **88**(423), 947–954.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology* **58**, 1246–1266.

Matheron, G. (1981). Splines and Kriging: their formal equivalence. In: *Down-to-Earth Statistics: Solutions Looking for Geological Problems* (Merriam D.F., Ed.). Syracuse Univ. of geology contributions ed.. pp. 77–95.

Schaback, R. (1999). Native Hilbert spaces for radial basis functions I. *Inter. Series of Numerical Mathematics* **132**, 255–282.

Schölkopf, B. and A. Smola (2002). *Learning with Kernels*. MIT Press. Cambridge.

Smola, A. J., B. Schölkopf and K.-R. Müller (1998). The connection between regularization operators and support vector kernels. *Neural Networks* **11**(4), 637–649.

Smola, A., T. Friess and B. Schölkopf (1999). Semiparametric support vector and linear programming machines. In: *Advances in Neural Information Processing Systems* (M. Kearns, S. Solla and D. Cohn, Eds.). pp. 585 – 591. 11. MIT Press. Cambridge.

Stein, M.L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer. New York.

Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer. Heidelberg.

Wahba, G. (1990). *Spline Models for Observational Data*. Vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM. Philadelphia.

Wahba, G. (1998). SVM, RKHS, and randomized GACV. In: *Advances in Kernel Methods – Support Vector Learning* (B. Schölkopf, C.J.C. Burges and A.J. Smola, Eds.). Chap. 6, pp. 69–87. MIT Press. Boston.

Wahba, G. and G.S. Kimeldorf (1970). Spline functions and stochastic processes. *Sankhyä: the Indian Journal of Statistics: Series A* **32**(2), 173–180.

Williams, C.K.I (1997). Regression with Gaussian processes. In: *Mathematics of Neural Networks: Models, Algorithms and Applications* (S.W. Ellacott, J.C. Mason and I.J. Anderson, Eds.). Kluwer. Presented at the Mathematics of Neural Networks and Applications Conference, Oxford, 1995.