Proceedings of the
44th IEEE Conference on Decision and Control, and
the European Control Conference 2005
Seville, Spain, December 12-15, 2005

Estimating derivatives and integrals with Kriging

Emmanuel Vazquez
Department of Signal and Electronic Systems
Supélec
91192 Gif-sur-Yvette, France
emmanuel.vazquez@supelec.fr

Eric Walter
Laboratoire des Signaux et Systèmes
CNRS, Supélec, Univ. Paris-Sud
91192 Gif-sur-Yvette, France
eric.walter@lss.supelec.fr

***Abstract*— This paper formalizes a methodology based on Kriging, a technique developped by geostatisticians, for estimating derivatives and integrals of signals that are only known via possibly irregularly spaced and noisy observations. This finds direct applications, e.g., in system identification when differential algebra is used to express parameters as nonlinear functions of the inputs and outputs and their derivatives. The procedure is quite simple to implement, and allows confidence intervals on the predicted values to be derived.**

I. Introduction

Let $f^*(\boldsymbol{x})$ be a real function defined on some compact set $\mathbb{X} \subset \mathbb{R}^d$. The problem to be considered here is the estimation of the derivative (or the integral) of f^* at any given $\boldsymbol{x} \in \mathbb{X}$ from a finite set of observations $S = \{(\boldsymbol{x}_i, f^{\rm obs}_{\boldsymbol{x}_i}),\, i = 1, \ldots, n\}$. These observations may be corrupted by (not necessarily white) noise, so $f^{\rm obs}_{\boldsymbol{x}_i}$ is not equal to $f(\boldsymbol{x}_i)$ in general. Moreover, the observations need not be regularly sampled. Such a problem is frequently encountered in system identification and control, for instance when algebraic differential methods are used to express parameters as nonlinear functions of the input and outputs and their derivatives, see, e.g., [1], [4].

The methodology to be presented is based upon techniques developed by Geostatisticians and known as *Kriging* and *Intrinsic Kriging*. The possibility of estimating derivatives and integrals via Kriging has already been suggested in the context of Geostatistics [2] and we would like to call the attention of the control community on its simplicity, pertinence and performances. This paper will consist of two parts. The first one will briefly recall the theory of Kriging and Intrinsic Kriging, before considering the estimation of derivatives and integrals with these methods. To the best of our knowledge, the mathematical formalization of the prediction of derivatives and integrals with intrinsic Kriging had never been published. The second part of the paper provides illustrative examples.

II. From linear prediction to the estimation of derivatives and integrals

A. Random Processes and Kriging

Kriging can be used to approximate or interpolate data, just as splines do. See, for instance, [2], [3], [6], [9]. As splines, it is a kernel regression method [10], [13]. Its specificity is that the kernel is chosen after a statistical analysis of the data. Let $(\Omega, \mathcal{A}, \mathsf{P})$ be a probability space and $L^2(\Omega, \mathcal{A}, \mathsf{P})$ be the set of second-order real-valued random variables defined on $(\Omega, \mathcal{A}, \mathsf{P})$.

Definition 1: The set of all random variables $F(\omega, \boldsymbol{x}) \in L^2(\Omega, \mathcal{A}, \mathsf{P})$ obtained when $\boldsymbol{x}$ runs through $\mathbb{X}$ is called a *second-order random process* with parameter $\boldsymbol{x} \in \mathbb{X}$.

Let $m(\boldsymbol{x}) = \mathrm{E}[F(\boldsymbol{x})]$ be the *mean* of $F(\boldsymbol{x})$, and $k(\boldsymbol{x}, \boldsymbol{y}) = \mathrm{cov}(F(\boldsymbol{x}), F(\boldsymbol{y})) = \mathrm{E}[(F(\boldsymbol{x}) - m(\boldsymbol{x}))(F(\boldsymbol{y}) - m(y))]$ be its *covariance function*. The covariance function is *positive* since

$$\mathrm{var}\left[\sum_{i=1}^{n} \lambda_i F(\boldsymbol{x}_i)\right] = \sum_{i,j=1}^{n} \lambda_i \lambda_j k(\boldsymbol{x}_i, \boldsymbol{x}_j) \geq 0\,,$$

for all $\lambda_i \in \mathbb{R}$, $\boldsymbol{x}_i \in \mathbb{X}$ and $n > 0$.

Let $F_1(\boldsymbol{x}), \ldots, F_q(\boldsymbol{x})$ be random processes defined on the same probability space and parameter space $\mathbb{X}$. Let $m_\alpha(\boldsymbol{x})$ be the mean of $F_\alpha(\boldsymbol{x})$ and $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y})$ be the covariance $\mathrm{E}[(F_\alpha(\boldsymbol{x}) - m_\alpha(\boldsymbol{x}))(F_\beta(\boldsymbol{y}) - m_\beta(y))]$, $\alpha, \beta \in \{1, \ldots, q\}$. Note that $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y}) = k_{\beta,\alpha}(\boldsymbol{y}, \boldsymbol{x})$ but in general $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y}) \neq k_{\alpha,\beta}(\boldsymbol{y}, \boldsymbol{x})$.

Assume, for the time being, that $f^*(\boldsymbol{x})$ is a *trajectory* of $F(\boldsymbol{x})$ (*i.e.*, there exists $\omega \in \Omega$ such that $F(\omega, \boldsymbol{x}) = f^*(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathbb{X}$) and that the observations are noise-free. Each $f^{\rm obs}_{\boldsymbol{x}_i} = f^*(\boldsymbol{x}_i)$ is then a *realization* of the random variable $F(\boldsymbol{x}_i)$. The data S can then be interpolated by predicting $F(\boldsymbol{x})$ given the random variables $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$, which amounts to finding a function $\hat{F}(\boldsymbol{x})$ of $F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)$ that minimizes $F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x})$ in some sense. A predicted value is obtained by replacing the random variables $F(\boldsymbol{x}_1), \ldots F(\boldsymbol{x}_n)$ by their realizations $f^{\rm obs}_1, \ldots, f^{\rm obs}_n$ in the expression of $\hat{F}(\boldsymbol{x})$ (ω remaining uncertain). Consider the class of *linear predictors*, which can be written as

$$\hat{F}(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F(\boldsymbol{x}_i)\,.$$

Assume, moreover, that $F(\boldsymbol{x})$ is a zero-mean process ($m(\boldsymbol{x}) = 0$, $\forall \boldsymbol{x} \in \mathbb{X}$), an hypothesis that will be relaxed in Section II-B. The *best predictor* or *Kriging predictor* $\hat{F}(\boldsymbol{x})$ of $F(\boldsymbol{x})$ is the orthogonal projection of $F(\boldsymbol{x})$ onto the subspace

$$\mathcal{H}_S = \mathrm{span}\{F(\boldsymbol{x}_i), i = 1, \ldots, n\}\,.$$

0-7803-9568-9/05/$20.00 ©2005 IEEE

Since

$$(F(\boldsymbol{x}) - \hat{F}(\boldsymbol{x}), F(\boldsymbol{x}_i))_{L^2(\Omega,\mathcal{A},\mathsf{P})} = 0 \Rightarrow$$
$$k(\boldsymbol{x}, \boldsymbol{x}_i) - \sum_{j=1}^{n} \widehat{\lambda}_{j,\boldsymbol{x}} k(\boldsymbol{x}_j, \boldsymbol{x}_i) = 0\,, \quad (1)$$

for $i = 1, \ldots, n$, the scalars $\widehat{\lambda}_{i,\boldsymbol{x}}$, $i = 1, \ldots, n$, are obtained by solving a linear system of equations.

Proposition 1 (Kriging): Let $F_1(\boldsymbol{x})$, $\ldots$, $F_q(\boldsymbol{x})$ be second-order random processes, with zero mean, and covariance functions $k_{\alpha,\beta}(\boldsymbol{x}, \boldsymbol{y})$. The best predictor of $F_\alpha(\boldsymbol{x})$ from $F_{\alpha_i}(\boldsymbol{x}_i)$, $i = 1, \ldots, n$, is the orthogonal projection of $F_\alpha(\boldsymbol{x})$ onto $\mathcal{H}_S = \mathrm{span}\{F_{\alpha_i}(\boldsymbol{x}_i), i = 1, \ldots, n\}$, written as

$$\hat{F}_\alpha(\boldsymbol{x}) = \sum_{i=1}^{n} \widehat{\lambda}_{i,\boldsymbol{x}} F_{\alpha_i}(\boldsymbol{x}_i)\,, \quad (2)$$

where the $\widehat{\lambda}_{i,\boldsymbol{x}}$s are the solution of the linear system[1] of equations

$$\boldsymbol{K}\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = \boldsymbol{k}_{\boldsymbol{x}}\,. \quad (3)$$

In (3), $\boldsymbol{K}$ is the $n \times n$ covariance matrix of the random vector $\boldsymbol{F}_S = (F_{\alpha_1}(\boldsymbol{x}_1), \ldots, F_{\alpha_n}(\boldsymbol{x}_n))^\mathsf{T}$, $\widehat{\boldsymbol{\lambda}}_{\boldsymbol{x}} = (\widehat{\lambda}_{1,\boldsymbol{x}}, \ldots, \widehat{\lambda}_{n,\boldsymbol{x}})^\mathsf{T}$ is the vector of the Kriging coefficients, and $\boldsymbol{k}_{\boldsymbol{x}} = (k_{\alpha,\alpha_1}(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, k_{\alpha,\alpha_n}(\boldsymbol{x}, \boldsymbol{x}_n))^\mathsf{T}$ is the covariance vector of $\boldsymbol{F}_S$ and $F_\alpha(\boldsymbol{x})$. Confidence intervals are obtained by evaluating the variance of $F_\alpha(\boldsymbol{x}) - \hat{F}_\alpha(\boldsymbol{x})$.

B. Intrinsic Kriging and Intrinsic Random Functions

Intrinsic Kriging (IK) [8] extends linear prediction to the case where the mean of $F(\boldsymbol{x})$ is unknown. In this framework, the function f^* generating the data is assumed to fluctuate around $m(\boldsymbol{x})$, which can be written as a linear parametric function $\boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x})$, where $\boldsymbol{b}$ and $\boldsymbol{r}(\boldsymbol{x})$ are l-dimensional vectors. Let $\mathcal{N}$ be the vector space $\{\boldsymbol{b}^\mathsf{T}\boldsymbol{r}(\boldsymbol{x}), \boldsymbol{b} \in \mathbb{R}^l\}$ and $F(\boldsymbol{x})$ be a random process with mean $m(\boldsymbol{x}) \in \mathcal{N}$. The main idea of IK is to find some linear transformations of $F(\boldsymbol{x})$ *filtering out* the mean so as to consider a zero-mean process again.

We first recall the notion of generalized random processes. Let $\tilde{\Lambda}$ be the vector space of *finite-support measures*, *i.e.* the space of linear combinations $\sum_{i=1}^n \lambda_i \delta_{\boldsymbol{x}_i}$, where $\delta_{\boldsymbol{x}}$ stands for the Dirac measure, such that for any $B \subset \mathbb{X}$, $\delta_{\boldsymbol{x}}(B)$ equals one if $\boldsymbol{x} \in B$ and zero otherwise. Let $\tilde{\Lambda}_{\mathcal{N}^\perp}$ be the subset of the elements of $\tilde{\Lambda}$ that vanish on $\mathcal{N}$. Thus, $\lambda \in \tilde{\Lambda}$ implies

$$\langle \lambda, f \rangle := \sum_{i=1}^{n} \lambda_i f(\boldsymbol{x}_i) = 0\,, \quad \forall\, f \in \mathcal{N}\,.$$

Definition 2: A symmetric function $k : \mathbb{X} \times \mathbb{X} \to \mathbb{R}$ is *conditionally positive* with respect to $\mathcal{N}$ if, for all $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, $k(\lambda, \lambda) \geq 0$, where $k(\lambda, \mu)$, $\lambda, \mu \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, is defined by

$$k(\lambda, \mu) := \sum_{i,j=1}^{n,m} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j)\,.$$

[1] $\boldsymbol{K}$ in (3) is generally a full rank matrix since covariances are most often *positive definite* functions. However, adapted solving techniques such as rank reduction must be used when the condition number of $\boldsymbol{K}$ is large, which may happen for instance when two observations are close in the space of factors.

If, moreover, $k(\lambda, \lambda) = 0$ implies $\lambda = 0$, for all $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, then $k(\boldsymbol{x}, \boldsymbol{y})$ is *conditionally positive definite*.

Let $F_\mathrm{G}(\lambda)$ be a linear application defined on $\tilde{\Lambda}_{\mathcal{N}^\perp}$, with values in $L^2(\Omega, \mathcal{A}, \mathsf{P})$. Assume that $F_\mathrm{G}(\lambda)$ is zero-mean for all λ and that $\mathrm{cov}[F_\mathrm{G}(\lambda), F_\mathrm{G}(\mu)] = k(\lambda, \mu)$, where $k(\boldsymbol{x}, \boldsymbol{y})$ is a conditionally positive definite function. Then, $F_\mathrm{G}(\lambda)$ is called a *generalized random process*. Such a random process is no longer defined on $\mathbb{X}$ but on a space of measures, and $k(\boldsymbol{x}, \boldsymbol{y})$ is called a *generalized covariance* (see Section II-E). Let $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ be the subspace of $L^2(\Omega, \mathcal{A}, \mathsf{P})$ generated by $F_\mathrm{G}(\lambda)$, $\lambda \in \tilde{\Lambda}_{\mathcal{N}^\perp}$. Since random variables in $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ are zero-mean, the inner product of $L^2(\Omega, \mathcal{A}, \mathsf{P})$ can be expressed in $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ as

$$(F_\mathrm{G}(\lambda), F_\mathrm{G}(\mu))_{L^2(\Omega,\mathcal{A},\mathsf{P})} = k(\lambda, \mu) = \sum_{i,j} \lambda_i \mu_j k(\boldsymbol{x}_i, \boldsymbol{y}_j)\,,$$

where $\lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i}$ and $\mu = \sum_j \mu_j \delta_{\boldsymbol{y}_j}$ are in $\tilde{\Lambda}_{\mathcal{N}^\perp}$. Thus, the bilinear form $k(\lambda, \mu)$ endows $\tilde{\Lambda}_{\mathcal{N}^\perp}$ and $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ with a structure of pre-Hilbert space. The completions $\mathcal{H}_{\mathcal{N}^\perp}$ and $\Lambda_{\mathcal{N}^\perp}$ of $\tilde{\mathcal{H}}_{\mathcal{N}^\perp}$ and $\tilde{\Lambda}_{\mathcal{N}^\perp}$ under this inner product define isomorphic Hilbert spaces. $F_\mathrm{G}(\lambda)$ can be extended on $\Lambda_{\mathcal{N}^\perp}$ by continuity. The generalized random process $F_\mathrm{G}(\lambda)$ is used as a random model. Simplifying hypotheses are introduced in the next paragraph.

Intrinsic Random Functions (IRF) are obtained when generalized random processes are endowed with a stationarity property. IRF are flexible models to use since unknown means can be conveniently dealt with and stationarity makes the inference of the parameters of their (generalized) covariance function feasible. Let $\tau_{\boldsymbol{h}} : \tilde{\Lambda}_{\mathcal{N}^\perp} \to \tilde{\Lambda}$ be the translation operator such that for $\lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i} \in \tilde{\Lambda}_{\mathcal{N}^\perp}$, $\tau_{\boldsymbol{h}}\lambda = \sum_i \lambda_i \delta_{\boldsymbol{x}_i + \boldsymbol{h}}$. Assume that $\tilde{\Lambda}_{\mathcal{N}^\perp}$ is stable under translation. $\mathcal{N}$ must therefore be itself a stable space of functions under $\tau_{\boldsymbol{h}}$. Assume further that the generalized covariance $k(\boldsymbol{x}, \boldsymbol{y})$ is invariant by translation. In the following, we shall write $k(\boldsymbol{h})$ with $\boldsymbol{h} = \boldsymbol{x} - \boldsymbol{y}$ instead of $k(\boldsymbol{x}, \boldsymbol{y})$, when the covariance is assumed to be stationary. Then $\tau_{\boldsymbol{h}}$ is continuous and can be uniquely extended on $\Lambda_{\mathcal{N}^\perp}$.

Definition 3: Let $F_\mathrm{G}(\lambda)$ be a zero-mean generalized random process defined on $\Lambda_{\mathcal{N}^\perp}$, with stationary generalized covariance $k(\boldsymbol{h})$. The random process $\boldsymbol{h} \mapsto F(\tau_{\boldsymbol{h}}\lambda)$, $\lambda \in \Lambda_{\mathcal{N}^\perp}$, is therefore weakly stationary. $F_\mathrm{G}(\lambda)$, $\lambda \in \Lambda_{\mathcal{N}^\perp}$, is then an *Intrinsic Random Function*.

The stability of $\mathcal{N}$ under the group of translations implies that $\mathcal{N}$ is necessarily a vector space of *exponential–polynomial* functions [7]. Such a space is generated by functions that can be written as $\boldsymbol{x}^{\boldsymbol{l}} e^{\boldsymbol{a}^\mathsf{T}\boldsymbol{x}}$, where $\boldsymbol{a}$ is a real or complex vector, $\boldsymbol{l}$ is the vector-valued index $(l_1, \ldots, l_d)$ and $\boldsymbol{x}^{\boldsymbol{l}} = x_{[1]}^{l_1} \cdots x_{[d]}^{l_d}$. (For a vector-valued index $\boldsymbol{l}$, we shall write $|\boldsymbol{l}| = l_1 + \cdots + l_q$.) For $\mathcal{N}$ to be stable by linear combinations and translations, the monomials $\boldsymbol{x}^{\boldsymbol{l}}$ must form a complete basis. We restrict ourselves to the case where $\mathcal{N}$ is a vector space of polynomials of degree at most equal to l. Let $\mathcal{N}_l = \mathrm{span}\{\boldsymbol{x}^{\boldsymbol{l}},\ \forall\, \boldsymbol{l} \text{ such that } |\boldsymbol{l}| \leq l\}$ and $\tilde{\Lambda}_l = \tilde{\Lambda}_{\mathcal{N}_l^\perp}$. Let Λ_l be a completion of $\tilde{\Lambda}_l$ under the inner product $k(\lambda, \mu)$. If the IRF $F_\mathrm{G}(\lambda)$ is defined on Λ_l, $F_\mathrm{G}(\lambda)$ is an IRF of order l,

or IRF(l) in short.

Proposition 2: Any IRF(l) is an IRF($l+1$).

Proof: This follows from the fact that the spaces Λ_l are nested:

$$\Lambda_{l+1} \subset \Lambda_l \,,$$

and that any IRF $F_{\rm G}(\lambda)$ defined on Λ_l will be stationary on Λ_{l+1}. ∎

Remark: Λ_l can be viewed as a set of finite-difference (increment) type operators. The condition for $\lambda = \sum_{i=1}^n \lambda_i \delta_{\boldsymbol{x}_i}$ to be in Λ_0 can be expressed as $\sum_{i=1}^n \lambda_i = 0$. Thus, $\lambda = \sum_{i=1}^n \lambda_i (\delta_{\boldsymbol{x}_i} - \delta_{\boldsymbol{x}_1})$, so λ is a linear combination of increment measures $\delta_{\boldsymbol{x}_i} - \delta_{\boldsymbol{x}_1}$. For $l > 0$, generalized increments are obtained. Note that if $\mathbb{X}$ is a space of dimension d, a minimum of $\binom{d+l}{l}$ points have to be taken to fully specify an element of Λ_l.

A generalized random process can be viewed as a class of equivalence of random processes defined on $\mathbb{X}$ with mean in $\mathcal{N}$. If $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, is a second-order random process, with mean in $\mathcal{N}$ and covariance $k(\boldsymbol{x}, \boldsymbol{y})$, the linear application

$$\begin{array}{rccl} F: & \tilde{\Lambda}_{\mathcal{N}^\perp} & \to & \mathcal{H} \\ & \lambda = \sum_{i=1}^n \lambda_i \delta_{\boldsymbol{x}_i} & \mapsto & F(\lambda) = \sum_{i=1}^n \lambda_i F(\boldsymbol{x}_i)\,, \end{array}$$

extends $F(\boldsymbol{x})$ on $\tilde{\Lambda}_{\mathcal{N}^\perp}$, where $\mathcal{H}$ stands for the Hilbert space generated by $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$. Since the mean of $F(\boldsymbol{x})$ is in $\mathcal{N}$, $F(\lambda)$, $\lambda \in \Lambda_{\mathcal{N}^\perp}$, is a zero-mean random variable, as λ filters out any function of $\mathcal{N}$. Assume that $k(\boldsymbol{x}, \boldsymbol{y})$ is positive definite. Then $(\lambda, \mu)_{\tilde{\Lambda}_{\mathcal{N}^\perp}} := (F(\lambda), F(\mu))_{\mathcal{H}}$ defines an inner product on $\tilde{\Lambda}_{\mathcal{N}^\perp}$. Let $\Lambda_{\mathcal{N}^\perp}$ be the completion of $\tilde{\Lambda}_{\mathcal{N}^\perp}$ under this inner product and extend $F(\lambda)$ on $\Lambda_{\mathcal{N}^\perp}$ by continuity. A generalized random process is thus obtained. The next paragraph indicates how the procedure may be reversed.

Definition 4: Let $F_{\rm G}(\lambda)$ be a generalized random process defined on $\Lambda_{\mathcal{N}^\perp}$. A second-order random process $F(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$, is a *representation* of $F_{\rm G}(\lambda)$ if

$$F_{\rm G}(\lambda) = F(\lambda), \quad \forall\, \lambda \in \Lambda_{\mathcal{N}^\perp}\,.$$

If $F_{\rm G}(\lambda)$ is an IRF(l), its representations can be written explicitly by taking appropriate measures $\lambda \in \Lambda_l$ [8]. If $F_0(\boldsymbol{x})$ is any representation of $F_{\rm G}(\lambda)$, other representations of $F_{\rm G}(\lambda)$ can be written as

$$F(\boldsymbol{x}) = F_0(\boldsymbol{x}) + \sum_{i=1}^q B_i p_i(\boldsymbol{x})\,, \tag{4}$$

where the p_is form a basis of $\mathcal{N}_l$ and the B_is are any second-order random variables [8]. Thus, the representations of an IRF(l) constitutes a class of random processes with mean in $\mathcal{N}_l$.

C. Derivation

In this section the notion of derivative of an IRF is developed. The aim is to estimate the derivative of a function f^* modeled by an IRF(l), which means that f^* comprises an unknown polynomial drift of degree at most equal to l. To simplify presentation, assume that $x \in \mathbb{R}$. Extension to the multi-dimensional case is straightforward.

Recall that a zero-mean stationary second-order random process $F(x)$ with covariance function $k(h)$ is mean-square differentiable at x if

$$F_h(x) = \frac{1}{h}(F(x+h) - F(x)) \tag{5}$$

converges in mean square when $h \to 0$. The limit exists if and only if $k^{(2)}(0)$ exists, and $F(x)$ is then mean-square differentiable at all x. The limit process is called the *derivative process* and denoted by $F^{(1)}(x)$. Higher-order derivatives are obtained by iteration, and it is straightforward to check that

$$\operatorname{cov}[F^{(q)}(x), F^{(r)}(y)] = (-1)^{(r)} k^{(q+r)}(x-y)\,. \tag{6}$$

Let $F_{\rm G}(\lambda)$ be an IRF(l), with generalized covariance $k(h)$. The difficulty for defining the derivative of $F_{\rm G}(\lambda)$ lies in the fact that neither $F_{\rm G}(\lambda)$ nor its derivatives can be defined point-wise. Thus, the notion of differentiability cannot be defined using the variance of an expression such as (5).

To define a derivative, we must use elements of Λ_l. Since for $\lambda = \sum_i \lambda_i \delta_{x_i} \in \Lambda_l$, $\tau_h \lambda \in \Lambda_l$, $\forall\, h \in \mathbb{R}$, define

$$\lambda_h = \frac{1}{h}(\tau_h \lambda - \lambda) \in \Lambda_l.$$

Definition 5: An IRF(l) $F_{\rm G}(\lambda)$ is *mean-square differentiable* at $\lambda \in \Lambda_l$ if $F_{\rm G}(\lambda_h)$ converges in mean square as $h \to 0$. When the limit exists, it is denoted by $F_{\rm G}^{(1)}(\lambda)$.

Proposition 3: Let $F_{\rm G}(\lambda)$ be an IRF(l), with generalized covariance $k(h)$. If $k^{(2)}(h)$ exists for all h, then $F_{\rm G}(\lambda)$ is mean-square differentiable for all λ in Λ_l, in which case $F_{\rm G}^{(1)}(\lambda)$ is an IRF(l) with generalized covariance $-k^{(2)}(h)$.

Proof: Let $\lambda = \sum_{i=1}^n \lambda_i \delta_{x_i}$ be in Λ_l. Then

$$\begin{aligned} \|F_{\rm G}(\lambda_h)\|^2 &= \frac{1}{h^2}\left\| F_{\rm G}\Big(\sum_{i=1}^n \lambda_i (\delta_{x_i+h} - \delta_{x_i}) \Big) \right\|^2 \\ &= \frac{1}{h^2} \sum_{i,j=1}^n \lambda_i \lambda_j (2k(x_i - x_j) \\ &\qquad - k(x_i - x_j + h) - k(x_i - x_j - h))\,. \end{aligned}$$

Assume further that $k(h)$ is twice differentiable for all $h \in \mathbb{R}$. Then $\|F_{\rm G}(\lambda_h)\|$ converges when $h \to 0$ and

$$\lim_{h \to 0} \|F_{\rm G}(\lambda_h)\|^2 = -\sum_{i,j=1}^n \lambda_i \lambda_j k^{(2)}(x_i - x_j)\,.$$

It follows that $F_{\rm G}^{(1)}(\lambda)$ is a generalized random process on Λ_l with zero-mean and generalized covariance $-k^{(2)}(h)$. ∎

Remark that the convergence of $F_{\rm G}(\lambda_h)$ in $L^2(\Omega, \mathcal{A}, \mathsf{P})$ when $h \to 0$ is equivalent to the convergence of λ_h in Λ_l. Let $\lambda^{(1)} = \lim_{h \to 0} \lambda_h$, if allowable. We shall then identify $F_{\rm G}^{(1)}(\lambda)$ and $F_{\rm G}(\lambda^{(1)})$.

Proposition 4: Let $F_{\rm G}(\lambda)$ be an IRF(l) and $F(\boldsymbol{x})$ be a representation. Then, $F^{(1)}(\boldsymbol{x})$ is a representation of $F_{\rm G}^{(1)}(\lambda)$.

Proof: For all $\lambda \in \Lambda_l$,

$$F_{\rm G}^{(1)}(\lambda) = \lim_{h \to 0} F_{\rm G}(\lambda_h) = \lim_{h \to 0} F(\lambda_h) = F(\lambda^{(1)}) = F^{(1)}(\lambda)\,.$$

∎

Order r derivatives are denoted by $F_{\mathrm{G}}^{(r)}(\lambda)$ and $\lambda^{(r)}$. Given $\lambda = \sum_i \lambda_i \delta_{x_i}^{(q_i)}$ and $\mu = \sum_j \mu_j \delta_{y_j}^{(r_j)}$ in Λ_l, it is easy to check that

$$\mathrm{cov}[F_{\mathrm{G}}(\lambda), F_{\mathrm{G}}(\mu)] = \sum_{i,j} (-1)^{r_j} \lambda_i \mu_j k^{(q_i + r_j)}(x_i - y_j)\,.$$

It becomes now possible to predict the derivatives of a representation of an IRF(l). The case where observations are corrupted by additive noise is directly studied. Observed values then correspond to realizations of the random variables $F^{\mathrm{obs}}(x_i) = F(x_i) + N_i$, $i = 1, \ldots, n$, where $F(x)$ is an unknown representation of $F_{\mathrm{G}}(\lambda)$, and the N_is are zero-mean random variables independent of $F(\boldsymbol{x})$, with covariance matrix $\boldsymbol{K}_N$. When the noise is white, $\boldsymbol{K}_N = \sigma_N^2 \boldsymbol{I}_n$. A linear predictor $\widehat{F^{(r)}}(x)$ of $F^{(r)}(x)$ can be written as

$$\widehat{F^{(r)}}(x) = \sum_i \widehat{\lambda}_{i,x} F^{\mathrm{obs}}(x_i)\,.$$

In IK the prediction error $F(x) - \widehat{F}(x)$ is minimized under the constraint $\delta_x - \sum \widehat{\lambda}_{i,x} \delta_{x_i} \in \Lambda_l$. To deal with derivatives, we similarly minimize $\mathrm{var}[F^{(r)}(x) - \widehat{F^{(r)}}(x)]$ under the constraint

$$\delta_x^{(r)} - \sum_i \widehat{\lambda}_{i,x} \delta_{x_i} \in \Lambda_l\,. \tag{7}$$

The solution can be obtained using $\mathrm{var}[F(\delta_x^{(r)} - \sum_i \widehat{\lambda}_{i,x}\delta_{x_i})] = \mathrm{var}[F_{\mathrm{G}}(\delta_x^{(r)} - \sum_i \widehat{\lambda}_{i,x}\delta_{x_i})]$. One can then check that the coefficients $\widehat{\lambda}_{i,x}$, $i = 1, \ldots, n$, are solutions of a system of linear equations, which can be written in matrix form as

$$\begin{pmatrix} \boldsymbol{K} + \boldsymbol{K}_N & \boldsymbol{P}^{\mathsf{T}} \\ \boldsymbol{P} & \boldsymbol{0} \end{pmatrix} \begin{pmatrix} \widehat{\boldsymbol{\lambda}}_x \\ \boldsymbol{\mu} \end{pmatrix} = \begin{pmatrix} \boldsymbol{k}_x^{(r)} \\ \boldsymbol{p}_x^{(r)} \end{pmatrix}, \tag{8}$$

where $\boldsymbol{K}$ is the $n \times n$ matrix of generalized covariances $k(x_i - x_j)$, $\boldsymbol{P} = (x_j{}^i)_{i=0,\,j=1}^{l,n}$ is a $(l+1) \times n$ matrix, $\boldsymbol{\mu}$ is a vector of Lagrange coefficients, $\boldsymbol{k}_x^{(r)}$ is a vector of size n with entries $k^{(r)}(x - x_i)$ and $\boldsymbol{p}_x^{(r)}$ is a vector of size $l+1$ with entries $(x^i)^{(r)}$, $i = 0, \ldots, l$. Note that

$$(x^i)^{(r)} = \begin{cases} 0 & \text{if } i < r \\ \frac{i!}{(i-r)!} x^{i-r} & \text{if } i \geq r \end{cases}$$

The variance of the prediction error is given by

$$\mathrm{var}[F^{(r)}(x) - \widehat{F^{(r)}}(x)] = -k^{(2r)}(0) - \begin{pmatrix} \widehat{\boldsymbol{\lambda}} \\ \boldsymbol{\mu} \end{pmatrix}^{\mathsf{T}} \begin{pmatrix} \boldsymbol{k}_x^{(r)} \\ \boldsymbol{p}_x^{(r)} \end{pmatrix}.$$

It can be used to assess confidence intervals, as illustrated in Section III.

D. Integration

This problem can be viewed as the prediction of a function $f(x)$ from observations of its derivative. Formally, this is equivalent to the previous problem, with straightforward adaptation.

E. Choice and estimation of the covariance

Once the covariance function is chosen, the procedure of estimating a derivative is quite simple to implement, since it boils down to solving a linear system. The question of the choice of the covariance is now considered. Asymptotic results [11], [14] suggest that satisfactory performance may be obtained even if the covariance is chosen incorrectly. It could therefore be argued that covariance choice is not an important issue. However, for satisfactory performance with a finite and often relatively small number of samples, a proper choice of the covariance turns out to be very important.

Any classical parametrized covariance can be used as a generalized covariance. For instance, [11] advocates the *Matérn covariance*, which can be written as

$$k(\boldsymbol{h}) = \frac{\sigma^2}{2^{\nu-1}\Gamma(\nu)} \left(\frac{2\nu^{1/2}\|\boldsymbol{h}\|}{\rho} \right)^{\nu} \mathcal{K}_\nu \left(\frac{2\nu^{1/2}\|\boldsymbol{h}\|}{\rho} \right),$$

where $\mathcal{K}_\nu$ is the modified Bessel function of the second kind, $\nu > 0$ controls the regularity (the differentiability) of the covariance at the origin, $\sigma^2 > 0$ is the variance ($k(0) = \sigma^2$), and $\rho > 0$ is a scale parameter.

Polynomial covariances are also a useful class of generalized covariances [8]. Given an order l, they can be written as

$$k(\boldsymbol{h}) = \sum_{p=0}^{l} (-1)^{p+1} a_p \|\boldsymbol{h}\|^{2p+1} \quad \text{with } a_p > 0,\, \forall p\,.$$

Note that this expression is linear in its parameters a_p. For example, intrinsic Kriging based on a covariance written as $-a_0\|\boldsymbol{h}\|$ gives a piecewise-linear interpolation.

We use *Maximum Likelihood* to estimate the vector $\boldsymbol{\theta}$ of the parameters of a covariance $k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$ when the mean of the covariance is known [5]. Let $F(\boldsymbol{x})$ be a zero-mean Gaussian random process with covariance $k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{y})$. Assume also that the observation noise is Gaussian. Let $\boldsymbol{K}(\boldsymbol{\theta})$ be the covariance matrix of $\boldsymbol{F}_S = [F(\boldsymbol{x}_1), \ldots, F(\boldsymbol{x}_n)]^{\mathsf{T}}$ and $\boldsymbol{K}_N(\boldsymbol{\theta}')$ be the covariance matrix of the random vector $\boldsymbol{N}$ of the measurement noise, assumed Gaussian, with zero mean and a covariance depending on some parameter vector $\boldsymbol{\theta}'$.

To simplify presentation, take $\bar{\boldsymbol{\theta}} = [\boldsymbol{\theta}^{\mathsf{T}}, \boldsymbol{\theta}'^{\mathsf{T}}]^{\mathsf{T}}$ and $\boldsymbol{K}(\bar{\boldsymbol{\theta}}) = \boldsymbol{K}(\boldsymbol{\theta}) + \boldsymbol{K}_N(\boldsymbol{\theta}')$. The log-likelihood of the data can then be written as

$$\begin{aligned} L(\boldsymbol{f}^{\mathrm{obs}} \mid \bar{\boldsymbol{\theta}}) = &-\frac{n}{2}\log 2\pi - \frac{1}{2}\log\det \boldsymbol{K}(\bar{\boldsymbol{\theta}}) \\ &- \frac{1}{2}\boldsymbol{f}^{\mathrm{obs}\,\mathsf{T}} \boldsymbol{K}(\bar{\boldsymbol{\theta}})^{-1} \boldsymbol{f}^{\mathrm{obs}}\,. \end{aligned} \tag{9}$$

In the following paragraph we recall the Restricted Maximum Likelihood (REML) approach to estimating the covariance parameters of a random process with unknown mean. Instead of the likelihood function of the data, REML maximizes that of the increments (or generalized increments) of these data [11].

Let $F_{\mathrm{G}}(\lambda)$ be a Gaussian *IRF*(l). Let $\boldsymbol{F}^{\mathrm{obs}}$ be the random observation vector, the sum of $\boldsymbol{F}_S = [F(\boldsymbol{x}_1), \ldots, F(x_n)]^{\mathsf{T}}$, with $F(\boldsymbol{x})$ a representation of $F_{\mathrm{G}}(\lambda)$, and some zero-mean

measurement noise vector. Let $\boldsymbol{P} = (\boldsymbol{x}_j^{\boldsymbol{l}_i})_{i,j=1}^{q,n}$ be the $q \times n$ matrix of basis functions of $\mathcal{N}_l$ evaluated on $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$. Since the dimension of $\mathcal{N}_l$ is q, the dimension of the space of the measures with support $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$ that cancel out the functions of $\mathcal{N}_l$ is $n-q$. Assume an $n \times (n-q)$ matrix $\boldsymbol{W}$ with rank $n-q$ has been found, such that

$$\boldsymbol{P}\boldsymbol{W} = \boldsymbol{0}\,.$$

(The columns of $\boldsymbol{W}$ are in the kernel of $\boldsymbol{P}$.) The columns of $\boldsymbol{W}$ are therefore the coefficients of measures with support $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, $\sum_{j=1}^n w_{i,j}\delta_{\boldsymbol{x}_j} \in \Lambda_l$. Then $\boldsymbol{Z} = \boldsymbol{W}^\mathsf{T}\boldsymbol{F}^{\text{obs}}$ is a Gaussian random vector taking its values in $\mathbb{R}^{n-q}$, with zero mean and covariance matrix $\boldsymbol{W}^\mathsf{T}(\boldsymbol{K}(\boldsymbol{\theta}) + \boldsymbol{K}_N(\boldsymbol{\theta}'))\boldsymbol{W} = \boldsymbol{W}^\mathsf{T}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W}$, where $\boldsymbol{K}(\boldsymbol{\theta})$ is the generalized covariance matrix with entries $k_{\boldsymbol{\theta}}(\boldsymbol{x}_i - \boldsymbol{x}_j)$ and where $\boldsymbol{K}_N(\boldsymbol{\theta}')$ is the covariance matrix of the observation noise. The random vector $\boldsymbol{Z}$ is a *contrast vector*. The log-likelihood of the contrasts is given by

$$L(\boldsymbol{z} \mid \bar{\boldsymbol{\theta}}) = -\frac{n-q}{2}\log 2\pi - \frac{1}{2}\log\det(\boldsymbol{W}^\mathsf{T}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W}) - \frac{1}{2}\boldsymbol{z}^\mathsf{T}(\boldsymbol{W}^\mathsf{T}\boldsymbol{K}(\bar{\boldsymbol{\theta}})\boldsymbol{W})^{-1}\boldsymbol{z}\,. \tag{10}$$

Various methods may be employed to compute the matrix $\boldsymbol{W}$. We favor the QR decomposition of $\boldsymbol{P}^\mathsf{T}$

$$\boldsymbol{P}^\mathsf{T} = (\boldsymbol{Q}_1 \mid \boldsymbol{Q}_2)\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix},$$

where $(\boldsymbol{Q}_1 \mid \boldsymbol{Q}_2)$ is an $n \times n$ orthogonal matrix and $\boldsymbol{R}$ is a $q \times q$ upper triangular matrix. It is trivial to check that the columns of $\boldsymbol{Q}_2$ form a basis of the kernel of $\boldsymbol{P}$, so we may chose $\boldsymbol{W} = \boldsymbol{Q}_2$. Note that $\boldsymbol{W}^\mathsf{T}\boldsymbol{W} = \boldsymbol{I}_{n-q}$.

III. Examples

A. Estimation of derivatives

Figure 1 represents a system output and its derivative. We see that the prediction of the derivative from a number of irregularly spaced noise-free observations of this output is satisfactory. Confidence intervals for this prediction can be provided, which is one of the advantages of the methodology advocated here. As could be expected, the uncertainty intervals are narrower close to the locations of the observations but, may be more surprisingly, when two observations are close enough, prediction is best *between* these observations.

The experiment is repeated in Figure 2, now with the addition of noise on the observations. Again the prediction is quite satisfactory, and the potential applications of such predictors are numerous.

B. Black-box model with prior information on derivatives

The previous examples illustrate only partially the possibilities offered by Kriging for the prediction of the derivatives of a signal. In Figure 3, observations of both the function and its derivative are used. This makes it possible to improve prediction, for instance by taking into account the knowledge that the value of the derivative at time zero is zero. This is an opportunity for introducing some prior information in a black-box model [12].

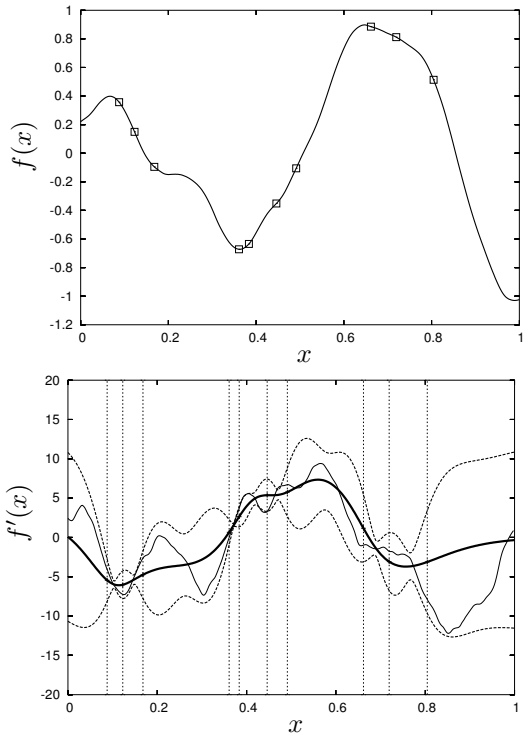

Fig. 1. Top: $f(x)$, $x \in [0,1]$, and 10 noise-free irregularly sampled observations (squares). Bottom: predicted value of the derivative of $f(x)$ (bold solid line) from previous observations. True derivative (plain solid line) and 95% confidence intervals (dashed lines) are shown. Vertical bars indicate the positions of the observations.

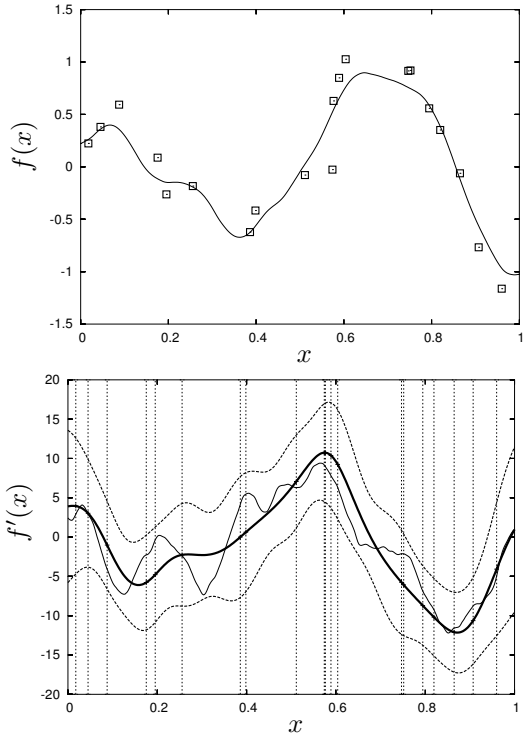

Fig. 2. Top: $f(x)$, $x \in [0,1]$, and 20 noisy irregularly sampled observations (squares). Standard deviation of the noise is 0.2. Bottom: predicted value of the derivative of $f(x)$ (bold solid line) from previous observations. True derivative (plain solid line) and 95% confidence intervals (dashed lines) are shown. Vertical bars indicate the positions of the observations.

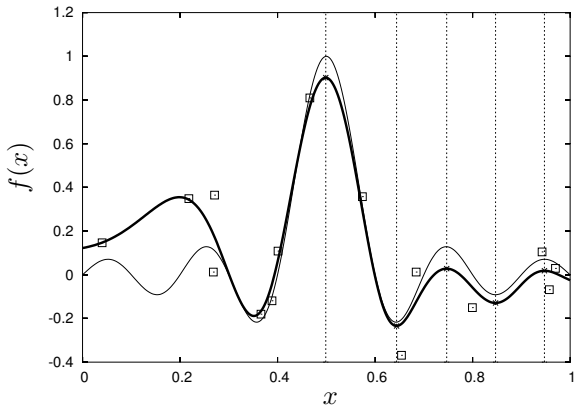

Fig. 3. Approximation (bold solid line) of the sinc function (plain solid line) from noisy observations (squares) taking into account the derivative of the sinc function at positions indicated by the vertical bars.

C. Approximation of a function from observations of its derivative — integration.

To conclude this section, let us apply Kriging to integration (Figure 4). Note that at least one observation of f is required in order to specify an initial condition. This approach may be extended to the numerical integration of partial differential equations. It may also be used to model nonlinear dynamical systems described by the state equation $\dot{\boldsymbol{x}} = \boldsymbol{f}(\boldsymbol{x}, \boldsymbol{u})$, which constitutes a promising perspective.

IV. Conclusions and perspectives

The methodology presented and formalized in this paper provides tools for solving two basic problems in control and signal processing, namely the differentiation and the integration of possibly multivariable signals that are only known via possibly noisy and irregularly sampled observations. It is based on intrinsic Kriging, a general-purpose technique for black-box modeling developed by geostatisticians during more than 50 years but still relatively ignored by the control community. As demonstrated by the examples treated, the resulting methodology is quite versatile and allows prior information, e.g., on boundary values of the derivative to be taken into account. Another distinctive feature of this statistically-based approach is that it allows confidence intervals on the predicted values of the derivative or integral to be provided. If the covariance function is chosen a priori, the technique requires only the solution of linear systems of equations. Better results, however, are to be expected if this covariance function is estimated from the data, for instance by maximum likelihood or Bayesian estimation. Numerical differentiation finds direct applications, e.g., in the framework of algebraic differential methods for parameter and state estimation, while the possibilities offered by numerical integration in the context of dynamical system modeling look very promising.

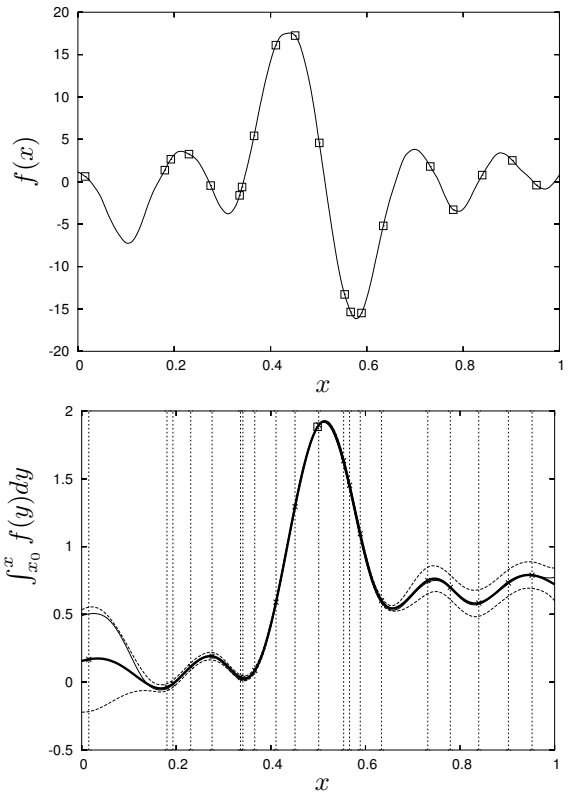

Fig. 4. Integration. Top: $f(x)$ and 20 noisy irregularly sampled observations (squares). Bottom: approximation of the integral of $f(x)$ (bold solid line) from previous observations knowing the value of the primitive at $x_0 = 0.5$ (indicated by a square). The true integral (plain solid line) and confidence intervals (dashed lines) are shown. Vertical bars indicate the positions of the observations.

References

[1] M. Braci and S. Diop. On numerical algorithms for nonlinear estimation. In *Proc. IEEE CDC 2003*, pages 2896–2901, Hawaii, USA, 2003.

[2] J.-P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty*. Wiley, New York, 1999.

[3] N. Cressie. *Statistics for Spatial Data*. Wiley, New York, 1993.

[4] S. Diop, V. Fromion, and J. W. Grizzle. A resettable extended Kalman filter based on numerical differentiation. In *Proc. IEEE CDC 2001*, New York, 2001.

[5] K. V. Mardia and R. J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, 73:135–146, 1984.

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

[7] G. Matheron. La théorie des fonctions aléatoires intrinsèques généralisées. Note Géostatistiques 117, Centre de Géostatistique de l'Ecole des Mines, Paris, 1971.

[8] G. Matheron. The intrinsic random functions, and their applications. *Adv. Appl. Prob.*, 5:439–468, 1973.

[9] 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.

[10] B. Schölkopf and A. Smola. *Learning with Kernels*. MIT Press, Cambridge, 2002.

[11] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York, 1999.

[12] E. Vazquez and E. Walter. Intrinsic Kriging and prior information. *Applied Stochastic Models in Business and Industry*, 2 (to appear), 2005.

[13] G. Wahba. *Spline Models for Observational Data*, volume 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. SIAM, Philadelphia, 1990.

[14] S. J. Yakowitz and F. Szidarovszky. A comparison of Kriging with nonparametric regression methods. *J. Multivariate Analysis*, 16:21–53, 1985.