ANNALES
DE LA FACULTÉ
DES SCIENCES
DE
TOULOUSE
Mathématiques

SÉBASTIEN DA VEIGA, AMANDINE MARREL

Gaussian process modeling with inequality constraints

Tome XXI, n° 3 (2012), p. 529-555.

<http://afst.cedram.org/item?id=AFST_2012_6_21_3_529_0>

Annales de la Faculté des Sciences de Toulouse Vol. XXI, n° 3, 2012
pp. 529–555

Gaussian process modeling with inequality constraints

Sébastien Da Veiga[1], Amandine Marrel[2]

Abstract. — Gaussian process modeling is one of the most popular approaches for building a metamodel in the case of expensive numerical simulators. Frequently, the code outputs correspond to physical quantities with a behavior which is known a priori: Chemical concentrations lie between 0 and 1, the output is increasing with respect to some parameter, etc. Several approaches have been proposed to deal with such information. In this paper, we introduce a new framework for incorporating constraints in Gaussian process modeling, including bound, monotonicity and convexity constraints. We also extend this framework to any type of linear constraint. This new methodology mainly relies on conditional expectations of the truncated multinormal distribution. We propose several approximations based on correlation-free assumptions, numerical integration tools and sampling techniques. From a practical point of view, we illustrate how accuracy of Gaussian process predictions can be enhanced with such constraint knowledge. We finally compare all approximate predictors on bound, monotonicity and convexity examples.

Résumé. — La modélisation par processus Gaussiens est une des approches les plus utilisées pour construire un métamodèle dans le cas de simulateurs numériques coûteux. Souvent, les sorties du code correspondent à des quantités physiques dont le comportement est connu à l'avance: les concentrations chimiques sont comprises entre 0 et 1, la sortie est croissante par rapport à un des paramètres, etc. Plusieurs approches ont été proposées pour prendre en compte de telles informations. Dans cet article, nous introduisons un nouveau cadre théorique pour inclure des contraintes dans la modélisation par processus Gaussiens, qui englobe les contraintes de bornes, de monotonie et de convexité. Nous étendons également ce cadre à tous les types de contraintes linéaires. Cette nouvelle méthodologie fait appel aux moments conditionnels de lois normales multivariées tronquées. Nous proposons plusieurs approximations basées sur une hypothèse de décorrélation, des outils d'intégration numérique et des

(*) Reçu le 27/02/2012, accepté le 16/04/2012

(1) IFP Energies nouvelles, 1 & 4 avenue de Bois Préau, 92852 Rueil-Malmaison, France
sebastien.da-veiga@ifpen.fr
(2) CEA, DEN, DER, F-13108 Saint-Paul-lez-Durance, France
amandine.marrel@cea.fr

techniques d'échantillonnage. D'un point de vue pratique, nous illustrons l'amélioration des performances de prédiction par processus Gaussiens lorque l'on inclut des contraintes. Nous comparons finalement les différents prédicteurs approchés sur des exemples avec contraintes de bornes, monotonie et convexité.

1. Introduction

To provide guidance to a better understanding of numerical simulators and to reduce the response uncertainties most efficiently, sensitivity measures of the input importance on the output variability are highly informative indicators [36]. For models that require a reasonable computational time, direct sampling methods (Monte Carlo) can be used to conduct uncertainty propagation studies or sensitivity analysis. In the case of simulators that take several hours or days for a single run, such direct sampling methods are impractical. To deal with these expensive models, several metamodel-based methods have been proposed in the past few years [28], [7]. Their potential was clearly illustrated when dealing with expensive simulators, since they outperform standard Monte-Carlo techniques. All these methods consist of the following steps. At first, an approximating metamodel is built from a small number of simulations so as to reproduce the expensive model. Then, this proxy model is used to compute sensitivity indices, through analytical or Monte-Carlo formulations. Among them, the Gaussian process approach is one of the most popular due to the wide range of applications where it was successfully used [25]. Frequently, the expensive model is built upon physical equations that involve symmetries, positiveness or monotonicity constraints. Accounting for such knowledge, while building the proxy model, could greatly improve the quality of approximation. Consequently, the estimation of sensitivity indices is more accurate while reducing the number of simulations required to build the proxy model. However, so far, very few of the metamodel-based methods are able to use this *a priori* information. In a one-dimensional setting, monotonicity constraints were investigated by [17], [31] and [3]. In a general kernel regression framework with no dimensionality restriction, monotonicity has been studied by [8] and [30]. In [17], monotonicity is incorporated through several optimization constraints in kernel weights identification and involves, in practice, quadratic programming techniques. [30] further propose to include convexity constraints with a sequential quadratic programming point of view. In the kriging community, [1] proposed a data-augmentation based simulation algorithm to account for bound constraints on a Gaussian process. Later, [41] applied the optimization constraint idea of [17] to the identification

of the kriging weights in the case of bound constraints. A different point of view is taken by [26], where bounded kriging simulations are generated by means of a Gibbs sampler. Recently, [20] examined monotonicity constraint in the kriging setting. The authors develop a bootstrap approach in order to select only monotonic kriging realizations, which are then used to build a new predictor through averaging. Besides, let us point out the theoretical contribution to symmetry constraints by [15]. In this paper, we focus mainly on linear constraints for Gaussian process modeling. In Section 2, we develop first a theoretical framework capable of accounting for such constraints. This work mainly relies on the literature related to truncated multinormal variables ([38], [39], [23], [24], [27]). Starting from analytical expressions of truncated moments, we show how it is possible to incorporate bound, monotonicity and convexity prior information on the expensive model. Available formulas involve the computation of integrals with dimensionality directly linked to the number of constraints. When this number is large (or even infinite, when monotonicity is imposed on a given interval for example), it is then necessary to provide efficient approximations. We propose in Section 3 a list of powerful approximation methods, which consist of numerical integral approximations and sampling techniques. Section 4 is dedicated to several illustrations of our methodology on prediction examples. Finally, we conclude with a brief summary and outline questions for future research.

2. Theoretical formulation

In this section, we first briefly introduce the Gaussian process modeling framework. We then detail the theoretical setting for incorporating constraints. In what follows, we assume that the complex computer code is represented by a function $f : \mathbb{R}^D \to \mathbb{R}$ which is assumed to be continuous. For a given value of the vector of inputs $\mathbf{x} = \left(x^1, \ldots, x^D\right) \in \mathbb{R}^D$, a simulation run of the code yields a real value $y = f(\mathbf{x})$, which corresponds to the output of interest. In practice, one evaluation of the function f can take several hours or even days. As a result, we make use here of response surface methods. The idea is the following. For a given set of input vectors $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, we compute the corresponding outputs $y_1, \ldots, y_n$. The goal is to build an approximating model of f using the sample $(\mathbf{x}_i, y_i)_{i=1,\ldots,n}$. We let $X_s = \left[\mathbf{x}_1^T, \ldots, \mathbf{x}_n^T\right]^T$ and $Y_s = [y_1, \ldots, y_n]^T$ denote the matrix of sample locations and the vector of responses, respectively.

2.1. Standard Gaussian process modeling

Gaussian process modeling [35] considers the deterministic response $y = f(\mathbf{x})$ as a realization of a random field $Y(\mathbf{x})$ given by the following decomposition:

$$Y(\mathbf{x}) = f_0(\mathbf{x}) + W(\mathbf{x})$$

where $f_0(\mathbf{x})$ is the mean function (*e.g.* a polynomial) and $W(\mathbf{x})$ is a stationary centered Gaussian field with variance σ^2 and correlation function R. Note that stationarity implies that its covariance function $C(\mathbf{x}, \mathbf{x}')$ can be written as $C(\tau) = \sigma^2 R(\tau)$ with $\tau = \mathbf{x} - \mathbf{x}'$. In this setting, the conditional distribution of the response at a new location $\mathbf{x}^*$ is a Gaussian distribution with moments given by

$$\mathbb{E}\left(Y(\mathbf{x}^*)|Y(X_s) = Y_s\right) = f_0(\mathbf{x}^*) + k(\mathbf{x}^*)^T \Sigma_S^{-1}\left(Y_s - F_s\right) \tag{2.1}$$

$$\mathrm{Var}\left(Y(\mathbf{x}^*)|Y(X_s) = Y_s\right) = \sigma^2 - k(\mathbf{x}^*)^T \Sigma_S^{-1} k(\mathbf{x}^*) \tag{2.2}$$

where $F_s = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]^T$ is the vector of the mean function at sample locations, $k(\mathbf{x}^*)$ is the covariance vector between $\mathbf{x}$ and sample locations X_s and Σ_S is the covariance matrix at sample locations. The conditional mean (2.1) serves as the predictor at location $\mathbf{x}$, and the prediction variance is given by (2.2). In practice, the mean function $f_0(\mathbf{x})$ has a parametric form $f_0(\mathbf{x}) = \sum_{j=1}^{J} \beta_j f_j(\mathbf{x}) = F(x)\beta$ where functions $F(x) = [f_1(\mathbf{x}), \ldots, f_J(\mathbf{x})]$ are known and $\beta = [\beta_1, \ldots, \beta_J]^T$ are regression parameters to be estimated. Moreover, R is chosen among a class of standard correlation functions (Gaussian, Matérn, ...) given up to some unknown hyperparameters ψ, corresponding to correlation lengths for example, see [32]. R is then denoted by R_ψ. As a result, in order to use the conditional expectation as a predictor, these parameters need to be estimated. Maximum likelihood estimators are usually preferred. For example, provided that ψ is known, regression parameters are obtained with the generalized least square estimator

$$\hat{\beta} = (F_s R_\psi^{-1} F_s)^{-1} F_s^T R_\psi^{-1} Y_s$$

and the maximum likelihood estimator of σ^2 is

$$\widehat{\sigma^2} = \frac{1}{n}(Y_s - F_s\hat{\beta})^T R_\psi^{-1}(Y_s - F_s\hat{\beta}).$$

In addition, estimation of hyperparameters consists in solving the following minimization problem

$$\psi^* = \arg\min_{\psi} \widehat{\sigma^2} \det(R_\psi)^{\frac{1}{n}}.$$

Consequently, the conditional field $\tilde{Y}(\mathbf{x}^*) = Y(\mathbf{x}^*)|Y(X_s) = Y_s$ with estimated parameters is a Gaussian field with mean $\tilde{\mu}(\mathbf{x}^*)$ given by

$$\tilde{\mu}(\mathbf{x}^*) = F(\mathbf{x}^*)\hat{\beta} + k(\mathbf{x}^*)^T\Sigma_S^{-1}\left(Y_s - F_s\hat{\beta}\right)$$

and covariance function equal to

$$\tilde{C}(\mathbf{x}, \mathbf{x}') = C(\mathbf{x}, \mathbf{x}') - k(\mathbf{x})^T\Sigma_S^{-1}k(\mathbf{x}').$$

Note that the covariance function has an additional component if the variance estimation on $\hat{\beta}$ is accounted for ([37]), but this case will not be considered here.

As stated in the introduction, our goal is to incorporate constraints on the so-called kriging predictor $\tilde{\mu}(\mathbf{x}^*) = \mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)\right)$. In contrast to previous work on constrained kriging ([41], [20]), we propose to keep the conditional expectation framework. For instance, if we know that predictions must lie between two real values a and b for all points in a subset I of $\mathbb{R}^D$, the predictor is naturally replaced with the conditional expectation

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall \mathbf{x} \in I, a \leqslant \tilde{Y}(\mathbf{x}) \leqslant b\right). \tag{2.3}$$

Similarly, in a one-dimensional setting, a monotonicity condition on I would lead to some conditional expectation

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall \mathbf{x} \in I, \tilde{Y}'(\mathbf{x}) \geqslant 0\right). \tag{2.4}$$

Let us remark that computation of such expectations is strongly linked to the theory of extrema of random fields. Indeed, equation (2.3) can be rewritten as

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\min_{\mathbf{x}\in I}\tilde{Y}(\mathbf{x}) \geqslant a, \max_{\mathbf{x}\in I}\tilde{Y}(\mathbf{x}) \leqslant b\right). \tag{2.5}$$

Explicit formulation of these quantities is not available in general. Indeed, distribution functions of extrema of random fields can be approximated through Rice formulas [2]. But here, equation (2.5) implies that we also need the joint distribution of the field $\tilde{Y}(\mathbf{x})$ and its extrema, which is not available. Hence, we suggest a discrete-location approximation. Instead of imposing a constraint on a given subset, we discretize it into a (large) number of conditioning points. As an illustration, consider a set of N points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ chosen in a subset I. Then, the conditional expectation in equation (2.3) is approximated by

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i = 1, \ldots, N, a \leqslant \tilde{Y}(\mathbf{x}_i) \leqslant b\right). \tag{2.6}$$

Note that this conditioning does not imply that the computer code f must be evaluated at points $\mathbf{x}_1, \ldots, \mathbf{x}_N$. More generally, in this work, any constraint defined on a subset will be replaced with its discrete-location counterpart. We will discuss the quality of this approximation according to the value of N in Section 4. In the following subsections, we will examine common constraints and the corresponding conditional expectation predictors that we propose.

2.2. Bound constraints

The most common type of constraints concerns bounds on the predictor. Indeed, it is highly frequent that computer code outputs correspond to physical quantities that are known to lie in a given interval. In chemical science, such quantities would be species concentration (between 0 and 1), or, in reservoir engineering, they would be oil production (strictly positive) for instance. Incorporation of these physical constraints should provide better predictions, and, in worst cases, neglecting them can even yield non-physical estimations. Let $I \subset \mathbb{R}^D$ denote a subset of $\mathbb{R}^D$ and $\mathbf{x}_1, \ldots, \mathbf{x}_N$ be a set of N points chosen in I. We assume that for all points in I, predictions must lie between two functions $a(\mathbf{x})$ and $b(\mathbf{x})$. Then, denoting $a_i = a(\mathbf{x}_i)$ and $b_i = b(\mathbf{x}_i)$ for all $i = 1, \ldots, N$, our constrained predictor is given by

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*) | \forall i = 1, \ldots, N, a_i \leqslant \tilde{Y}(\mathbf{x}_i) \leqslant b_i\right). \tag{2.7}$$

This is a straightforward generalization of equation (2.6). Since the random vector $\left(\tilde{Y}(\mathbf{x}^*), \tilde{Y}(\mathbf{x}_1), \ldots, \tilde{Y}(\mathbf{x}_N)\right)$ follows a multivariate Gaussian distribution, the conditional expectation in equation (2.7) actually corresponds to the mean of the so-called truncated multinormal distribution [38], see Section 3 for details.

2.3. Derivative constraints

Another usual available constraint is related to monotonicity. The increasing or decreasing behavior of the output with respect to some variables is often governed by physical equations. In this case, differentiability of f is generally first assumed. Consequently, the covariance function C is chosen among functions yielding differentiable trajectories. Here, we focus on fields which are mean-square differentiable, *i.e.* $\partial^2 C(\tau)/\partial \tau_j^2$ exists and is finite for all $1 \leqslant j \leqslant D$ ([6]). Assume that predictions must be increasing with respect to some variable x^j, $1 \leqslant j \leqslant D$, on some subset $I \subset \mathbb{R}^D$ and denote $\mathbf{x}_1, \ldots, \mathbf{x}_N$ a set of N points chosen in I. Therefore, the constrained

predictor is equal to

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i=1,\ldots,N,\frac{\partial\tilde{Y}}{\partial x^j}(\mathbf{x}_i)\geqslant 0\right). \tag{2.8}$$

Derivative constraints related to other variables on other subsets can be readily incorporated in this conditional expectation. An important result is that all vectors encompassing $\tilde{Y}(\mathbf{x}^*)$ and any of its partial derivatives is Gaussian, since the differential operator is linear. As a result, equation (2.8) is also the mean of a truncated multinormal distribution. Note that we have for example the following relations [6]:

$$\mathrm{Cov}\left(\tilde{Y}(\mathbf{x}),\frac{\partial\tilde{Y}}{\partial x'^j}(\mathbf{x}')\right)=\frac{\partial}{\partial x'^j}\tilde{C}(\mathbf{x},\mathbf{x}'). \tag{2.9}$$

$$\mathrm{Cov}\left(\frac{\partial\tilde{Y}}{\partial x^i}(\mathbf{x}),\frac{\partial\tilde{Y}}{\partial x'^j}(\mathbf{x}')\right)=\frac{\partial^2}{\partial x^i\partial x'^j}\tilde{C}(\mathbf{x},\mathbf{x}'). \tag{2.10}$$

2.4. Convexity constraints

Sometimes, the practitioner further knows that f is convex at some locations, due to physical insight. Assuming that f is twice-differentiable imposes to choose a covariance function C such that $\partial^4 C(\tau)/\partial\tau_j^4$ exists and is finite for all $1\leqslant j\leqslant D$. In a one-dimensional setting, if we require that predictions must be convex at some locations $x_1,\ldots,x_n$, the constrained predictor is

$$\mathbb{E}\left(\tilde{Y}(x^*)|\forall i=1,\ldots,N,\frac{\partial^2\tilde{Y}}{\partial x^2}(x_i)\geqslant 0\right). \tag{2.11}$$

In a two-dimensional setting, it becomes

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i=1,\ldots,N,\nabla^2_{\mathbf{x}^*\mathbf{x}^*}\tilde{Y}(x_i,x_i')\geqslant 0\right) \tag{2.12}$$

for locations $(x_i,x_i')\in\mathbb{R}^2$, $i=1,\ldots,N$. Equation (2.11) corresponds to a standard truncated multinormal distribution because a vector with components consisting of $\tilde{Y}(\mathbf{x}^*)$ and any of its second-order derivatives is Gaussian. In a two-dimensional setting, the predictor (2.12) involves the mean of the so-called elliptically truncated multinormal distribution [39]. Convexity in higher dimensions can also be incorporated if we impose that the Hessian matrix of the Gaussian field $\tilde{Y}(\mathbf{x}^*)$ is positive-semidefinite at each constraint point. By Sylvester's criterion, this is equivalent to impose that each leading principal minor of the Hessian is positive. However, since this involves

computing determinants, such a constraint leads to polynomial constraints, which have not been treated yet from a truncation perspective in the literature. Even if it is possible to deal with them through the Gibbs sampler which will be presented in Section 3.2, note that at each constraint point there are D minors. As a result, the number of constraints is equal to $N \times D$ and will increase with dimension D. For simplicity, we will only investigate convexity constraints for $D = 1$ and $D = 2$ in the following sections. Let us remark that more generally, if we consider $\tilde{Y}(\mathbf{x}^*)$ derivatives:

$$\tilde{Y}^{(\kappa)}(\mathbf{x}) = \frac{\partial^{|\kappa|}}{\partial(x^1)^{\kappa_1} \cdots \partial(x^D)^{\kappa_D}} \tilde{Y}(\mathbf{x}), \tag{2.13}$$

we have:

$$\begin{aligned} &\mathrm{Cov}\left(\tilde{Y}^{(\kappa)}(\mathbf{x}), \tilde{Y}^{(\lambda)}(\mathbf{x}')\right) \\ &= \frac{\partial^{|\kappa|+|\lambda|}}{\partial(x^1)^{\kappa_1} \cdots \partial(x^D)^{\kappa_D} \partial(x'^1)^{\lambda_1} \cdots \partial(x'^D)^{\lambda_D}} \tilde{C}(\mathbf{x}, \mathbf{x}') \end{aligned} \tag{2.14}$$

where $\kappa = (\kappa_1, \ldots, \kappa_D)$ and $\lambda = (\lambda_1, \ldots, \lambda_D)$ are non-negative integers with $|\kappa| = \sum_i \kappa_i$ and $|\lambda| = \sum_i \lambda_i$, see [6].

2.5. Linear constraints

The last type of constraints that we study here concerns linear inequality constraints. For illustration, imagine that the output of the computer code is a physical quantity subject to some conservation-type constraint, *e.g.* $\int_{\mathbf{x}\in\Omega} f(\mathbf{x})d\mathbf{x} \leqslant M$ for some constant M and a subset $\Omega \subset \mathbb{R}^D$. If we seek predictions satisfying this equation, the constrained predictor should be

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*) | \int_{\mathbf{x}\in\Omega} \tilde{Y}(\mathbf{x})d\mathbf{x} \leqslant M\right). \tag{2.15}$$

Unfortunately, this integral has no analytical formulation. In practice, it can be approximated through numerical integration. In the following, we will assume that it is approximated linearly, *i.e.*

$$\int_{\mathbf{x}\in\Omega} \tilde{Y}(\mathbf{x})d\mathbf{x} \approx \sum_{i=1}^{N} w_i \tilde{Y}(\mathbf{x}_i) \tag{2.16}$$

for some locations $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and associated weights $w_1, \ldots, w_N$. This includes trapezoidal and Gaussian quadrature rules for $D = 1$, and Monte-Carlo approximation for any value of D. Hence, the approximate constrained predictor is

$$\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*) | \sum_{i=1}^{N} w_i \tilde{Y}(\mathbf{x}_i) \leqslant M\right). \tag{2.17}$$

This is the mean of a Gaussian vector subject to a plane truncation, which has been studied by [40].

2.6. Parameter estimation

As mentioned in Section 2.1, regression parameters β , variance parameter σ^2 and covariance hyperparameters ψ are preliminarily estimated by maximizing the unconditional likelihood of the observations. In other words, all the above constraints are not used in this estimation process and are included in subsequent predictions exclusively. In the context of bound truncation of multinormal variables, mean and covariance estimation can be performed via Gibbs sampling, see [16]. To the best of our knowledge, no extension was developed for general linear constraints and covariance hyperparameters. Since there is no successful stragey currently available, we propose to investigate conditional maximum likelihood estimation in future research and to work only with unconditional estimations in what follows.

3. Approximations of truncated moments

As mentioned in the previous section, all constrained predictors exhibited so far fall into the framework of truncated multinormal distributions. There is a large amount of literature dedicated to truncation of Gaussian vectors, the pioneering work being that of Tallis ([38], [39], [40]). In this section, we first detail the analytical derivations which are available in the literature. Since they consist of several integral computations, we then present adapted numerical integration tools. Alternatively, we discuss convenient simulation algorithms capable of generating samples from a truncated Gaussian vector. This makes it possible to estimate the constrained predictors through averaging. For simplicity, we work in this section with a D-dimensional Gaussian vector denoted $\mathbf{Z} = (Z_1, \ldots, Z_D)$ with mean $\mu \in \mathbb{R}^D$ and covariance matrix Σ. Its probability density function (pdf) is then given by

$$\phi_{\mu,\Sigma}(\mathbf{z}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{z}-\mu)^T\Sigma^{-1}(\mathbf{z}-\mu)\right), \mathbf{z} \in \mathbb{R}^D.$$

For example, $\mathbf{Z}$ will stand for the vector $\left(\tilde{Y}(\mathbf{x}^*), \tilde{Y}(\mathbf{x}_1), \ldots, \tilde{Y}(\mathbf{x}_N), \tilde{Y}'(\mathbf{x}'_1), \ldots, \tilde{Y}'(\mathbf{x}'_N)\right)$ if one wishes to build the kriging predictor at some location $\mathbf{x}^*$ with bound constraints on both the predictor and its derivative imposed at points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and $\mathbf{x}'_1, \ldots, \mathbf{x}'_N$ respectively.

3.1. Available formulas

Let us study first the case of bound constraints, *i.e.* $\mathbf{a} \leqslant \mathbf{Z} \leqslant \mathbf{b}$ for two vectors $\mathbf{a}$ and $\mathbf{b}$ in $\mathbb{R}^D$. The pdf of the truncated Gaussian vector $\mathbf{Z}$, or equivalently the conditional pdf of $\mathbf{Z}$ knowing $\mathbf{a} \leqslant \mathbf{Z} \leqslant \mathbf{b}$, is given by:

$$\phi_{\mu,\Sigma,\mathbf{a},\mathbf{b}}(\mathbf{z}) = \begin{cases} \frac{\phi_{\mu,\Sigma}(\mathbf{z})}{\mathbb{P}(\mathbf{a} \leqslant \mathbf{Z} \leqslant \mathbf{b})}, & \text{for } \mathbf{a} \leqslant \mathbf{z} \leqslant \mathbf{b}, \\ 0, & \text{otherwise.} \end{cases}$$

In what follows, we denote $\alpha = \mathbb{P}(\mathbf{a} \leqslant \mathbf{Z} \leqslant \mathbf{b})$ and assume without loss of generality that $\mu = 0$ (all variables and truncation points are translated by $-\mu$). Following [38], a direct computation shows that the moment generating function of the truncated distribution is given by

$$m(\mathbf{t}) = \mathbb{E}\left(\exp(\langle t, \mathbf{Z}\rangle)\right) = \frac{\exp(T)}{\alpha}\Phi(\mathbf{a} - \xi, \mathbf{b} - \xi; \Sigma) \tag{3.1}$$

where $T = \frac{1}{2}\mathbf{t}^T\Sigma\mathbf{t}$, $\xi = \Sigma\mathbf{t}$ and function Φ is defined by

$$\Phi(\mathbf{u}, \mathbf{v}; \Sigma) = \int_{u_1}^{v_1} \dots \int_{u_D}^{v_D} \phi_{0,\Sigma}(\mathbf{z}) d\mathbf{z}.$$

[38] then uses equation (3.1) to derive the first- and second-order moments for the truncated distributions by computing the partial derivatives of $m(\mathbf{t})$ at the origin. More precisely, coming back to a general μ and denoting

$$f_i(z) =$$

$$\int_{a_1}^{b_1} \dots \int_{a_{i-1}}^{b_{i-1}} \int_{a_{i+1}}^{b_{i+1}} \dots \int_{a_D}^{b_D} \phi_{\mu,\Sigma,\mathbf{a},\mathbf{b}}(z_1, \dots, z_{i-1}, z, z_{i+1}, \dots, z_D) dz_1 \dots d\mathbf{z}_{-i} \tag{3.2}$$

the i-th marginal density of the truncated distribution, we have

$$\mathbb{E}(Z_i | \mathbf{a} \leqslant \mathbf{Z} \leqslant \mathbf{b}) = \mu + \sum_{j=1}^{D} \sigma_{ij} \left(f_j(a_j) - f_j(b_j)\right) \tag{3.3}$$

where $\sigma_{ij} = (\Sigma)_{ij}$. Other contributions include recursive formulas in [23], [24], as well as explicit ones for the two-dimensional case in [27] and other properties related to marginalization and conditioning in [19]. In the case $D = 1$, truncated moments are then given by the following formulas:

$$\mathbb{E}(Z_1 | a_1 \leqslant Z_1 \leqslant b_1) = \mu_1 + \sigma_1 \frac{\phi(\tilde{a}_1) - \phi(\tilde{b}_1)}{\Phi(\tilde{b}_1) - \Phi(\tilde{a}_1)} \tag{3.4}$$

$$\mathrm{Var}(Z_1 | a_1 \leqslant Z_1 \leqslant b_1) = \sigma_1^2 \left[1 + \frac{\tilde{a}_1\phi(\tilde{a}_1) - \tilde{b}_1\phi(\tilde{b}_1)}{\Phi(\tilde{b}_1) - \Phi(\tilde{a}_1)} - \left(\frac{\phi(\tilde{a}_1) - \phi(\tilde{b}_1)}{\Phi(\tilde{b}_1) - \Phi(\tilde{a}_1)}\right)^2\right] \tag{3.5}$$

where $\tilde{a}_1 = (a_1 - \mu_1)/\sigma_1$ and $\tilde{b}_1 = (b_1 - \mu_1)/\sigma_1$ with $\sigma_1 = \sqrt{\sigma_{11}}$. These monodimensional results also correspond to equation (3.3) when all correlations are neglected for $D > 1$, *i.e.* they can be used to compute all expectations and variances of the truncated distribution if Σ is assumed to be diagonal. Equation (3.3) yields an analytical expression for the constrained predictors in equations (2.3), (2.8) and (2.11). Similar results were derived by [40] for the case of linear inequality constraints which, after linear transformation, involve the same formulas as for bound constraints. This yields an analytical expression for the constrained predictor of equation (2.17) as well. In addition, [39] computed the moment generating function of the elliptically truncated Gaussian distribution for standardized variables and ellipses centered at the origin. [5] extended this result to general variables and ellipses. In their work, approximations of first- and second-order moments are obtained through Laguerre expansions. We do not report the calculations here.

From a computational perspective, a result on general truncation given in [22] will be useful. In equation (3.3), if not all the variables are truncated, we can naively replace the corresponding bounds by infinite ones and compute the corresponding $D-1$ dimensional integrals. However, it is possible to reduce the problem complexity. Up to permutation of its components, we can assume that $\mathbf{Z}$ consists of truncated variables $Z_1, \ldots, Z_k$ and non-truncated ones $Z_{k+1}, \ldots, Z_D$ for $k < D$. Truncation here is to be understood in a general way, *i.e.* bound, linear or even elliptical. Without loss of generality, we take $\mu = 0$ and let the non-truncated covariance matrix Σ be partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},$$

where Σ_{11} corresponds to the covariance of the non-truncated version of $Z_1, \ldots, Z_k$, Σ_{22} to the non-truncated version of $Z_{k+1}, \ldots, Z_D$ and Σ_{12} to their non-truncated cross-covariance. Assume that the mean ν_1 and the covariance matrix Γ_{11} of $Z_1, \ldots, Z_k$ after truncation are available, by using for example the expansions of [38]. Then, the mean ν_2 and covariance matrix Γ_{22} of the remaining variables $Z_{k+1}, \ldots, Z_D$ are given by

$$\nu_2 = \Sigma_{21}\Sigma_{11}^{-1}\nu_1 \tag{3.6}$$

and

$$\Gamma_{22} = \Sigma_{22} - \Sigma_{21}\left(\Sigma_{11}^{-1} - \Sigma_{11}^{-1}\Gamma_{11}\Sigma_{11}^{-1}\right)\Sigma_{12}. \tag{3.7}$$

In the case of bound constraints, equation (3.6) is easily obtained from (3.3). Indeed, non-truncated variables $Z_{k+1}, \ldots, Z_D$ correspond to infinite bounds in (3.2), meaning that their mean only involves the marginals f_i for $i = 1, \ldots, k$ in equation (3.3). This yields $\nu_2 = \Sigma_{21}w$ with

$w := (f_j(a_j) - f_j(b_j))_{1 \leqslant j \leqslant k}$. But, in this case, the f_i for $i = 1, \ldots, k$ degenerate into the marginals of the truncated distribution of variables $Z_1, \ldots, Z_k$ only. This means that we also have $\nu_1 = \Sigma_{11} w$. Eliminating w in these equations finally gives (3.6). A similar reasoning yields (3.7). In other words, computation of truncated moments will only be performed in a space of dimensionality $k-1$, where $k < D$ is the number of constrained components. Once the truncated moments are computed, moments of non-truncated variables are obtained by standard Gaussian regression formulas. From a practical point of view, let us consider for instance that we want to incorporate bound constraints on $\tilde{Y}(\mathbf{x})$. Since we impose constraints at points $\mathbf{x}_1, \ldots, \mathbf{x}_N$ only, the procedure for building predictions at a new location $\mathbf{x}^*$ is the following. First, we compute the truncated mean of $\left(\tilde{Y}(\mathbf{x}_1), \ldots, \tilde{Y}(\mathbf{x}_N)\right)$ with equation (3.3). This yields ν_1 in equation (3.6). Second, we solve the linear system $\Sigma_{11}^{-1}\Sigma_{12}$ of equation (3.6). Note that it only requires information on the non-truncated covariance matrix of the field $\tilde{Y}(\mathbf{x})$. Finally, the prediction at $\mathbf{x}^*$ is given by (3.6). Generalization to other types of constraints is straightforward, as long as all constraint points for each constrained field (bounds, monotonicity, linear inequality) are included in the vector $Z_1, \ldots, Z_k$ which is subject to truncation. Note also that when considering several kind of constraints, it is possible to impose them at different locations.

In general, we will mainly rely on equation (3.3) for the evaluation of the truncated mean ν_1 at constraint points. Remark that it involves computation of normal integrals of dimension $D-1$ (or $k-1$ as explained above). When D is large, as should be the case in the discrete-location method we propose, it is necessary to have powerful algorithms capable of producing accurate approximations. In this work, we will use the latest versions of the algorithms based on the methodology introduced by Genz ([11], [12]). It relies on a preliminary Cholesky decomposition of Σ and successive mono-dimensional integrations with a Quasi Monte-Carlo procedure. Algorithms are available on Allan Genz's webpage (http://www.math.wsu.edu/faculty/genz/homepage). In practice, we have been able to use this methodology for D around 1000. Interested readers can find an introduction and details on available algorithms in [13]. The algorithm complexity depends on the dimension D (the number of constraints) and on the size of the Quasi Monte-Carlo sample denoted by q. More precisely, the complexity is given by $O(D^3 + Dq)$, where the $O(D^3)$ term correponds to the Cholesky decomposition and $O(Dq)$ to mono-dimensional integrations.

3.2. Sampling techniques

Another approach for approximating truncated moments involves sampling techniques. Practically, if we were able to generate samples from the truncated vector $\mathbf{Z}$, empirical moments calculated on the basis of this sample would yield approximations of the constrained predictors. Literature on simulation of truncated multinormal variables is abundant. In the case of linear truncations, most algorithms are based on MCMC methods. Some call for a Gibbs sampler ([14], [33], [21], [34]), while others are rooted in the perfect simulation framework ([29], [10]). All MCMC methods for multivariate sampling involve numerous calls to a sampling algorithm for the univariate truncated normal. Inversion of the cumulative distribution function is possible, but approximation errors occur when the probability of the normal variable to be inside the truncation bounds is low. Similarly, crude rejection sampling from the normal distribution will be inefficient in the same case. [14] and [33] proposed rejection samplers with a high acceptance rate. Since we will use the algorithm of [33], we detail its implementation in what follows. Assume without loss of generality that we sample from a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$ in the interval $[a, b]$. The rejection algorithm is

1. Generate a uniform random number z in $[a, b]$
2. Compute

$$g(z) = \begin{cases} \exp(-z^2/2) & \text{if} \quad 0 \in [a, b] \\ \exp((b^2 - z^2)/2) & \text{if} \quad b < 0 \\ \exp((a^2 - z^2)/2) & \text{if} \quad 0 < a \end{cases}$$

3. Generate a uniform random number u in $[0, 1]$ and accept z if $u \leqslant g(z)$; otherwise go back to step 1.

Note that [4] also designed a univariate algorithm based on tables, which seems to be the fastest algorithm so far. The complete algorithm (Gibbs sampler with the previous rejection technique) has a complexity equal to $O(D^3 + Dq)$, where q is the size of the requested truncated sample. Once again, the $O(D^3)$ term comes from a preliminary Cholesky decomposition used for the Gibbs sampler, while calls to the univariate sampler have complexity $O(Dq)$.

In the case of elliptically truncated multinormal vectors, it is still possible to use the Gibbs sampler of [33] since it is actually designed for general convex truncation sets, provided that one-dimensional slices of the subset are easily computable. However, let us mention that an efficient multidimensional rejection sampler has been developed by [9]. Their idea is based

on the following result concerning truncated simulation outside a sphere for zero-mean normal vectors. Let $\mathbf{X} \sim N_D(0, \mathbf{I})$ and let Y be independent of $\mathbf{X}$ and distributed as

$$f_Y(y) \propto \exp(-y/2) y^{\frac{D}{2}-1} \mathbf{1}(y > r).$$

Then, $\mathbf{Z} = \sqrt{Y} \frac{\mathbf{X}}{\|\mathbf{X}\|}$ has density given by

$$f_{\mathbf{Z}}(\mathbf{z}) \propto \exp\left(-\frac{\mathbf{z}^T\mathbf{z}}{2}\right) \mathbf{1}(\mathbf{z}^T\mathbf{z} > r),$$

i.e. $\mathbf{Z}$ follows a truncated normal distribution outside the sphere of radius r. Sampling from the univariate variable Y is straightforward, which implies that sampling from $\mathbf{Z}$ is very fast. The authors then develop a method for computing the largest origin-centered sphere contained in the truncation ellipsoid, as well as the smallest origin-centered sphere containing the ellipsoid. These two spheres are finally used in a rejection algorithm.

4. Numerical studies

Let us illustrate the behavior of the constrained predictors on several analytical examples with bound, monotonicity and convexity constraints. For all examples with bound and monotonicity, we will compare the constrained predictors given by three different techniques:

- Genz approximation,
- Gibbs sampling technique,
- Correlation-free approximation (equation (3.4)).

4.1. Examples on 1-D functions

Consider first the one-dimensional function f_1 given by

$$f_1(x) = \frac{\sin(10\pi x)}{10\pi x}$$

for $x \in [0, 1]$. We assume that n observations $(x_i, y_i = f_1(x_i))_{i=1,\dots,n}$ are available, where the x_i are sampled according to the uniform distribution on $[0, 1]$. These observations are used to build the conditional field $\tilde{Y}(\mathbf{x})$ and the corresponding unconstrained kriging predictor $\tilde{\mu}(\mathbf{x}^*) = \mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)\right)$ introduced in Section 2.1. The mean function $f_0(\mathbf{x})$ is assumed to be constant and the hyperparameters are estimated by maximum likelihood. Now, we suppose further that f_1 is known to be positive on every interval $\left[\frac{2m}{10}, \frac{2m+1}{10}\right]$ and negative on $\left[\frac{2m+1}{10}, \frac{2m+2}{10}\right]$ for $m = 0, \dots, 4$. Consequently, we build the

constrained predictor $\mathbb{E}\left(\tilde{Y}(\mathbf{x}^*)|\forall i = 1,\ldots,N, a_i \leqslant \tilde{Y}(\mathbf{x}_i) \leqslant b_i\right)$ of equation (7) by choosing N locations uniformly on $[0,1]$ and by setting a_i and b_i according to the above intervals. We then compute its approximation by Genz algorithm, Gibbs sampling and the correlation-free formula. To evaluate the accuracy of the metamodels, we use the predictivity coefficient Q^2. It is the determination coefficient R^2 computed from a test sample (composed here by $n_{\text{test}} = 100$ uniformly chosen points):

$$Q^2(Y,\hat{Y}) = 1 - \frac{\sum_{i=1}^{n_{\text{test}}}\left(Y_i - \hat{Y}_i\right)^2}{\sum_{i=1}^{n_{\text{test}}}\left(Y_i - \bar{Y}\right)^2},$$

where Y denotes the n_{test} true observations (or exact values) of the test set, $\bar{Y}$ their empirical mean and $\hat{Y}$ the metamodel predicted values. Results for a random sample of x_i with $n = 10$ observations and $N = 20$ constraint locations are given in Figure 1, top. The predictor variance (equation (3.7)) is depicted in Figure 1, bottom. Here, the covariance function is the Gaussian one and predictions are performed on a set of 100 points chosen uniformly on $[0,1]$.

First note that Genz algorithm and Gibbs sampling yield the same results for the predictor and its variance. The unconstrained predictor cannot reproduce the shape of f_1 on the left part, since no observation points fall in this region. The behavior is similar on the far right part, leading to an overall Q^2 equal to 0.021, which is very low. Incorporation of positivity constraints makes it possible to greatly improve the prediction in the regions with few observation points. The correlation-free approximation still exhibits incorrect variations on the far left and the far right, but it yields a Q^2 equal to 0.83, which is much higher than the unconstrained predictor. Genz and sampling approximations both produce extremely well-behaved predictors, with similar Q^2 equal to 0.98. Concerning the predictor variance, we can observe that it is heavily reduced when accounting for constraints, especially with Genz and sampling approximations. This example, with an unfortunate sample, clearly illustrates the added value of constraint incorporation in the predictions.

In order to validate our approach on a larger set of possible locations, we repeat this procedure 100 times with both the Gaussian and the Matérn 3/2 covariance functions. We selected the Matérn 3/2 covariance function because it usually gives good predictions in many of our numerical experiments. Note however that it produces trajectories which are only differentiable at order one (in the mean square sense). We also study several choices for n and N. Results on the Q^2 for each method are given in Table 1

and Table 2 for the Gaussian and the Matérn covariance function, respectively. When the number of observations is small (10 or 15), accounting for constraints yield better prediction in mean, and the standard deviation is smaller. Also note that Genz and sampling approximations yield equivalent results and outperform the correlation-free formula. As expected, when the number of observations increases, incorporation of constraints is less interesting. In addition, the number of location points is not influential, meaning that 20 constraints are already sufficient for a good approximation. Besides, in this example, the Gaussian covariance function is superior to the Matérn 3/2 one.

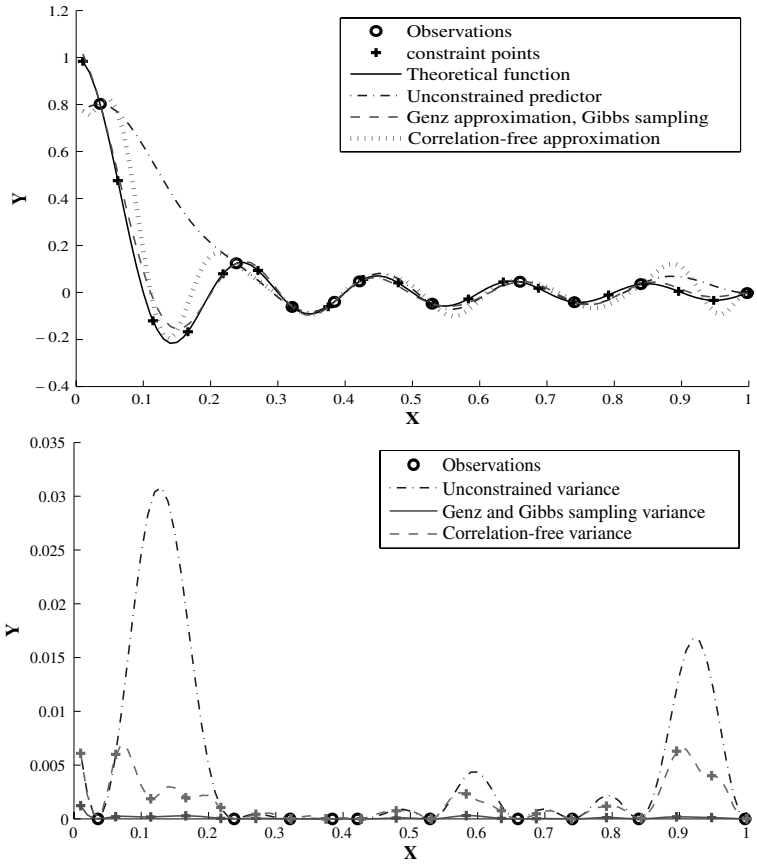

Figure 1. — Unconstrained and bound-constrained predictor (top) and predictor variance (bottom) for function f_1. Constraint points are marked with a cross.

Table 1. — Mean and standard deviation of Q^2 for different values of n and N with a Gaussian covariance for function f_1.

n	N	Unconstrained	Genz	Sampling	Correlation-free
n = 10	N = 20	0.30 +/- 0.46	0.53 +/- 0.36	0.47 +/- 0.39	0.42 +/- 0.29
n = 10	N = 30	0.24 +/- 0.43	0.41 +/- 0.39	0.39 +/- 0.43	0.41 +/- 0.30
n = 15	N = 20	0.58 +/- 0.53	0.72 +/- 0.36	0.72 +/- 0.36	0.69 +/- 0.34
n = 15	N = 30	0.62 +/- 0.56	0.80 +/- 0.34	0.77 +/- 0.37	0.74 +/- 0.33
n = 20	N = 20	0.93 +/- 0.28	0.96 +/- 0.12	0.96 +/- 0.12	0.91 +/- 0.28
n = 20	N = 30	0.91 +/- 0.34	0.95 +/- 0.14	0.95 +/- 0.15	0.93 +/- 0.18
n = 25	N = 20	0.99 +/- 0.03	0.99 +/- 0.03	0.99 +/- 0.03	0.99 +/- 0.03
n = 25	N = 30	0.96 +/- 0.29	0.99 +/- 0.05	0.99 +/- 0.05	0.98 +/- 0.10

Table 2. — Mean and standard deviation of Q^2 for different values of n and N with a Matérn 3/2 covariance for function f_1.

n	N	Unconstrained	Genz	Sampling	Correlation-free
n = 10	N = 20	0.46 +/- 0.61	0.60 +/- 0.41	0.46 +/- 0.58	0.45 +/- 0.33
n = 10	N = 30	0.29 +/- 0.46	0.52 +/- 0.33	0.32 +/- 0.39	0.47 +/- 0.28
n = 15	N = 20	0.60 +/- 0.40	0.72 +/- 0.32	0.70 +/- 0.34	0.68 +/- 0.31
n = 15	N = 30	0.55 +/- 0.45	0.74 +/- 0.34	0.65 +/- 0.42	0.69 +/- 0.31
n = 20	N = 20	0.73 +/- 0.39	0.83 +/- 0.29	0.80 +/- 0.33	0.81 +/- 0.30
n = 20	N = 30	0.79 +/- 0.33	0.85 +/- 0.25	0.84 +/- 0.26	0.83 +/- 0.25
n = 25	N = 20	0.82 +/- 0.38	0.87 +/- 0.27	0.87 +/- 0.27	0.86 +/- 0.28
n = 25	N = 30	0.89 +/- 0.28	0.92 +/- 0.20	0.92 +/- 0.20	0.91 +/- 0.21

Our second example involves function f_2 given by

$$f_2(x) = \frac{\sin(10\pi x^{5/2})}{10\pi x}$$

for $x \in [0,1]$. This function is more difficult to approximate because it has a frequency which exhibits strong variations on $[0,1]$. As before, we select n observations $(x_i, y_i = f_2(x_i))_{i=1,\ldots,n}$ where the x_i are sampled according to the uniform distribution on $[0,1]$ and we build the unconstrained kriging predictor. Again, the mean function $f_0(\mathbf{x})$ is assumed to be constant and the hyperparameters are estimated by maximum likelihood. We first include bound constraints as in previous example; *i.e.* we choose N locations uniformly on $[0,1]$ and impose positivity or negativity according to f_2. The so-obtained constrained predictor is approximated by Genz algorithm, Gibbs sampling and the correlation-free formula. By repeating this procedure 100 times, we obtain the mean and the standard deviation of the Q^2 for each method, for different values of n and N, as illustrated in Table 3 and Table 4 for the Gaussian and the Matérn 3/2 covariance function, respectively. Once again, the number of constraints does not seem to impact the results. However, when the number of observations is small (10, 20 and 30), accounting for constraints improves much more the unconstrained predictor than in the previous example. This can be explained due

to the non-stationary behavior of function f. Moreover, in this example, the correlation-free formula performs very well. Note also that this time, the Matérn 3/2 covariance function gives slightly better results than the Gaussian one.

Let us investigate now the incorporation of derivative constraints for the approximation of f_2. In what follows, we will focus on the case $n = 15$ and fix $N = 20$ locations where we will impose constraints. We then build constrained predictors with three different types of constraints:

- Bound constraints only at the N locations (positive / negative);
- Derivative constraints only at the N locations (increasing / decreasing);
- Bound and derivative constraints at the N locations;
- Bound, derivative and convexity constraints at the N locations (only for the Gaussian covariance function, which is sufficiently differentiable).

In each case, we use Genz approximation to evaluate the constrained predictor. The mean and the standard deviation of the Q^2 for a Matérn 3/2 and a Gaussian covariance function are given in Table 5.

Table 3. — Mean and standard deviation of Q^2 for different values of n and N with a Gaussian covariance for function f_2.

n	N	Unconstrained	Genz	Sampling	Correlation-free
n = 10	N = 20	0.43 +/- 0.24	0.71 +/- 0.19	0.68 +/- 0.20	0.71 +/- 0.26
n = 10	N = 30	0.43 +/- 0.24	0.67 +/- 0.21	0.65 +/- 0.23	0.77 +/- 0.14
n = 10	N = 50	0.45 +/- 0.22	0.61 +/- 0.20	0.69 +/- 0.23	0.76 +/- 0.16
n = 10	N = 100	0.47 +/- 0.21	0.52 +/- 0.19	0.66 +/- 0.25	0.81 +/- 0.14
n = 20	N = 20	0.71 +/- 0.19	0.87 +/- 0.16	0.85 +/- 0.17	0.86 +/- 0.14
n = 20	N = 30	0.71 +/- 0.20	0.82 +/- 0.20	0.83 +/- 0.20	0.84 +/- 0.18
n = 20	N = 50	0.73 +/- 0.15	0.84 +/- 0.14	0.85 +/- 0.14	0.88 +/- 0.08
n = 20	N = 100	0.69 +/- 0.21	0.77 +/- 0.22	0.80 +/- 0.23	0.86 +/- 0.16
n = 30	N = 20	0.86 +/- 0.18	0.94 +/- 0.10	0.94 +/- 0.10	0.93 +/- 0.08
n = 30	N = 30	0.89 +/- 0.14	0.92 +/- 0.12	0.92 +/- 0.14	0.93 +/- 0.11
n = 30	N = 50	0.82 +/- 0.22	0.85 +/- 0.23	0.88 +/- 0.21	0.93 +/- 0.10
n = 30	N = 100	0.89 +/- 0.15	0.91 +/- 0.16	0.91 +/- 0.16	0.95 +/- 0.07
n = 40	N = 20	0.93 +/- 0.14	0.95 +/- 0.09	0.95 +/- 0.09	0.94 +/- 0.10
n = 40	N = 30	0.91 +/- 0.17	0.93 +/- 0.15	0.94 +/- 0.14	0.94 +/- 0.14
n = 40	N = 50	0.95 +/- 0.10	0.96 +/- 0.09	0.96 +/- 0.09	0.97 +/- 0.07
n = 40	N = 100	0.93 +/- 0.18	0.92 +/- 0.18	0.93 +/- 0.15	0.94 +/- 0.15
n = 50	N = 20	0.98 +/- 0.08	0.98 +/- 0.06	0.98 +/- 0.06	0.98 +/- 0.06
n = 50	N = 30	0.98 +/- 0.08	0.98 +/- 0.06	0.98 +/- 0.06	0.98 +/- 0.06
n = 50	N = 50	0.96 +/- 0.13	0.96 +/- 0.12	0.96 +/- 0.12	0.97 +/- 0.11
n = 50	N = 100	0.97 +/- 0.08	0.97 +/- 0.08	0.97 +/- 0.08	0.98 +/- 0.06

Table 4. — Mean and standard deviation of Q^2 for different values of n and N with a Matérn 3/2 covariance for function f_2.

n	N	Unconstrained	Genz	Sampling	Correlation-free
n = 10	N = 20	0.43 +/- 0.29	0.80 +/- 0.11	0.74 +/- 0.19	0.78 +/- 0.12
n = 10	N = 30	0.44 +/- 0.27	0.83 +/- 0.10	0.76 +/- 0.24	0.79 +/- 0.12
n = 10	N = 50	0.47 +/- 0.22	0.83 +/- 0.13	0.75 +/- 0.21	0.79 +/- 0.10
n = 10	N = 100	0.47 +/- 0.24	0.77 +/- 0.17	0.74 +/- 0.20	0.80 +/- 0.10
n = 20	N = 20	0.74 +/- 0.15	0.91 +/- 0.06	0.89 +/- 0.07	0.90 +/- 0.06
n = 20	N = 30	0.77 +/- 0.12	0.93 +/- 0.05	0.91 +/- 0.08	0.91 +/- 0.06
n = 20	N = 50	0.75 +/- 0.20	0.94 +/- 0.05	0.89 +/- 0.19	0.91 +/- 0.07
n = 20	N = 100	0.75 +/- 0.14	0.93 +/- 0.07	0.90 +/- 0.12	0.92 +/- 0.06
n = 30	N = 20	0.87 +/- 0.10	0.96 +/- 0.02	0.95 +/- 0.03	0.95 +/- 0.03
n = 30	N = 30	0.89 +/- 0.09	0.96 +/- 0.02	0.95 +/- 0.07	0.96 +/- 0.04
n = 30	N = 50	0.88 +/- 0.09	0.97 +/- 0.02	0.97 +/- 0.02	0.96 +/- 0.02
n = 30	N = 100	0.87 +/- 0.16	0.96 +/- 0.14	0.94 +/- 0.15	0.95 +/- 0.04
n = 40	N = 20	0.94 +/- 0.05	0.97 +/- 0.01	0.97 +/- 0.02	0.97 +/- 0.02
n = 40	N = 30	0.94 +/- 0.06	0.98 +/- 0.01	0.98 +/- 0.01	0.98 +/- 0.01
n = 40	N = 50	0.95 +/- 0.06	0.98 +/- 0.01	0.97 +/- 0.04	0.98 +/- 0.02
n = 40	N = 100	0.94 +/- 0.06	0.99 +/- 0.01	0.98 +/- 0.01	0.98 +/- 0.01
n = 50	N = 20	0.97 +/- 0.04	0.98 +/- 0.01	0.98 +/- 0.01	0.98 +/- 0.01
n = 50	N = 30	0.97 +/- 0.04	0.98 +/- 0.01	0.98 +/- 0.01	0.98 +/- 0.01
n = 50	N = 50	0.97 +/- 0.05	0.99 +/- 0.01	0.99 +/- 0.01	0.98 +/- 0.01
n = 50	N = 100	0.97 +/- 0.05	0.99 +/- 0.01	0.99 +/- 0.02	0.99 +/- 0.01

Table 5. — Mean and standard deviation of Q^2 when accounting for several constraints for function f_2.

Constraints	Gaussian	Matérn 3/2
No	0.62 +/- 0.20	0.63 +/- 0.19
Bounds only	0.79 +/- 0.19	0.88 +/- 0.06
Derivatives only	0.80 +/- 0.19	0.77 +/- 0.11
Bounds and derivatives	0.80 +/- 0.23	0.91 +/- 0.06
Bounds and derivatives and convexity	0.85 +/- 0.19	×

Observe first that the Matérn 3/2 covariance function is superior to the Gaussian one in terms of both the mean and the standard deviation of Q^2. When we account for bound and derivative constraints together, the predictions are improved, as expected. Note that in the Matérn 3/2 case, bound constraints alone yield better results than derivatives only.

Here, convexity can also be accounted for in the case of a Gaussian covariance function. We show in Figure 2 the cumulative incorporation of bounds, first derivatives and convexity for $n = 12$ and $N = 20$ on f_2, as well as the corresponding predictor variance. While the unconstrained predictor fails at retrieving information between observed points ($Q^2 = 0.43$), constraints greatly help reconstruct the function in these regions.

Bound constraints start by improving predictions on the right part ($Q^2 = 0.84$ in Figure 2, top). First-derivatives further enhance the approximation on the left part ($Q^2 = 0.95$ in Figure 2, middle), but deteriorate it on the right. This phenomenon is finally compensated by second-order derivatives ($Q^2 = 0.98$ in Figure 2, bottom). The predictor variance is first largely reduced with bound constraints and one can observe further reduction with derivatives. Convexity has a small impact on variance reduction, since there is only a slight improvement between predictions with first- and second-order derivatives.

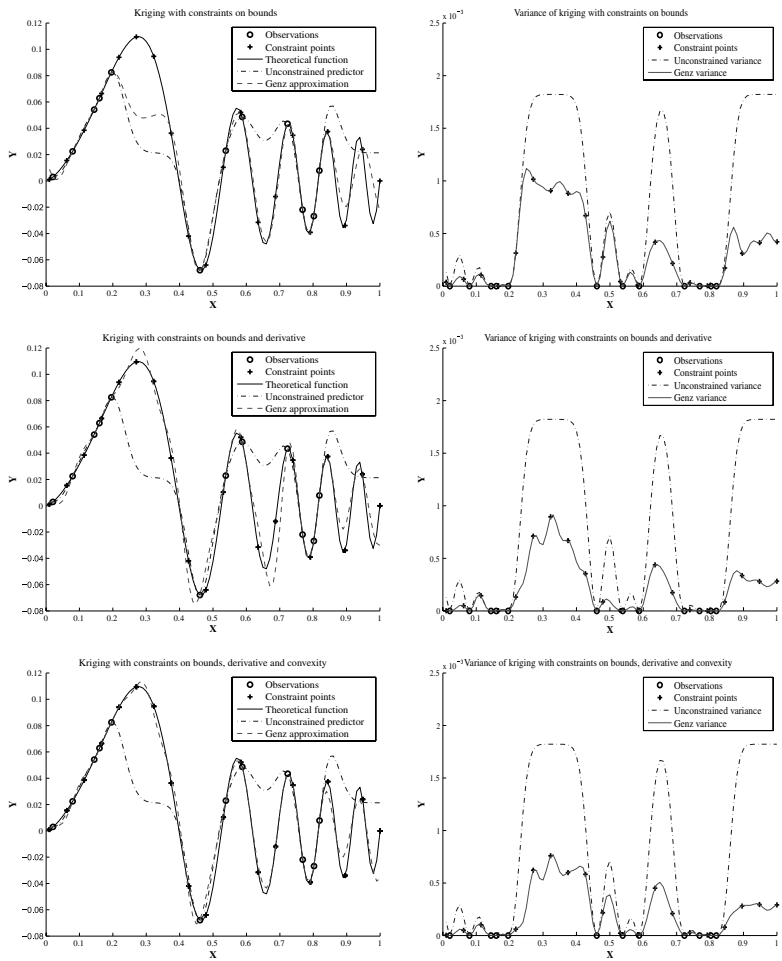

Figure 2. — Unconstrained and constrained predictor (left) and predictor variance (right) accounting for constraints on bounds only (top), on bounds and derivatives (middle) and on bounds, derivatives and convexity (bottom) on function f_2.

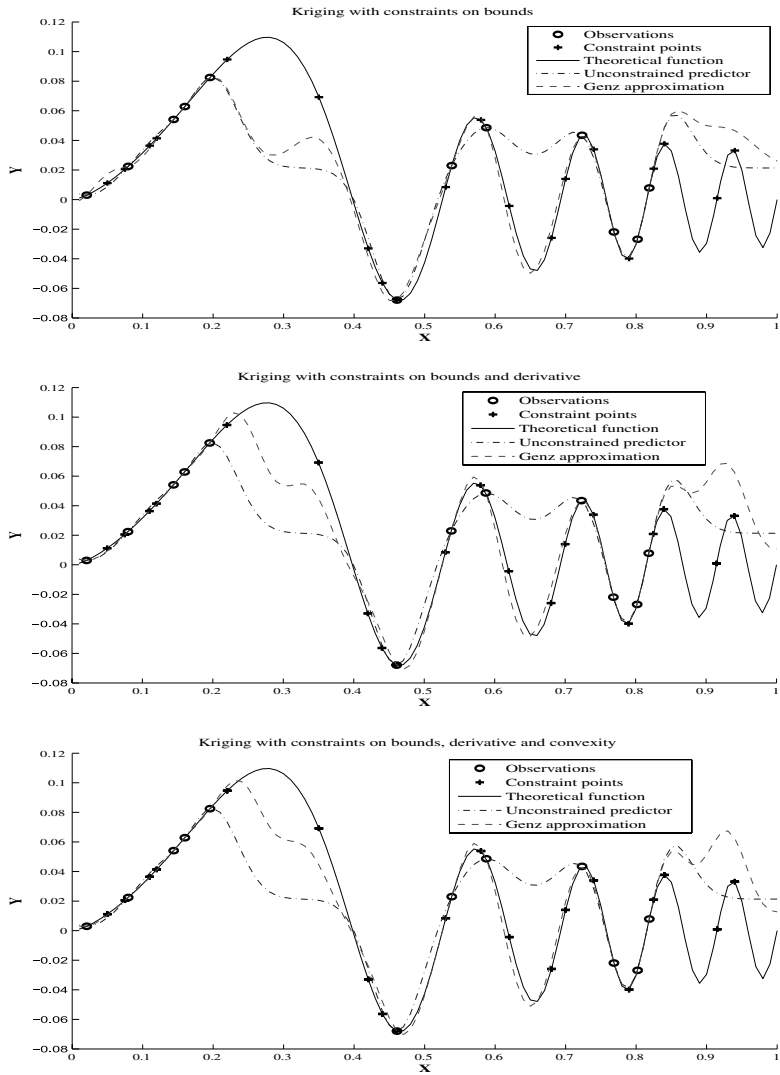

Figure 3. — Unconstrained and constrained predictor accounting for constraints on bounds only (top), on bounds and derivatives (middle) and on bounds, derivatives and convexity (bottom) on function f_2 with random constraint locations.

In this last example, constraint points were chosen equally spaced on $[0, 1]$. In order to examine the behavior of our predictor with respect to other constraint locations, we now generate them at random. Results are reported in Figure 3. Predictions are deteriorated in this case, with $Q^2 = 0.58$ for constraints on bounds, $Q^2 = 0.72$ for constraints on bounds and first derivatives

and $Q^2 = 0.76$ for constraints on bounds, first derivatives and convexity. The largest difference with the previous result in Figure 2 is observed on the right part, where constraints cannot reconstruct the true behavior of the unknown function. This can be explained by the unfortunate choice for the constraint points. Indeed, in the right part, there are no constraint points imposing a negative value to the predictor. On the left part, the wide peak is also ignored when derivative constraints are included. Once again, this is due to the absence of constraint points imposing a positive value to the first-derivative near the peak. As already illustrated in Table 5, incorporation of constraints improve predictions in average. However, specific sets of locations can lessen this improvement. We mention in the last section dedicated to discussion a possible approach for an efficient placement of these locations, which will be studied in future work.

4.2. Example on the 2-D Schwefel's function

Finally, we propose to investigate our constrained predictors on a two-dimensional example with bound constraints only. We consider the Schwefel's function defined by $f(x^1, x^2) = -x^1 \sin(\sqrt{|x^1|}) - x^2 \sin(\sqrt{|x^2|})$ on $[-200, 200]^2$. This function is displayed in Figure 4.

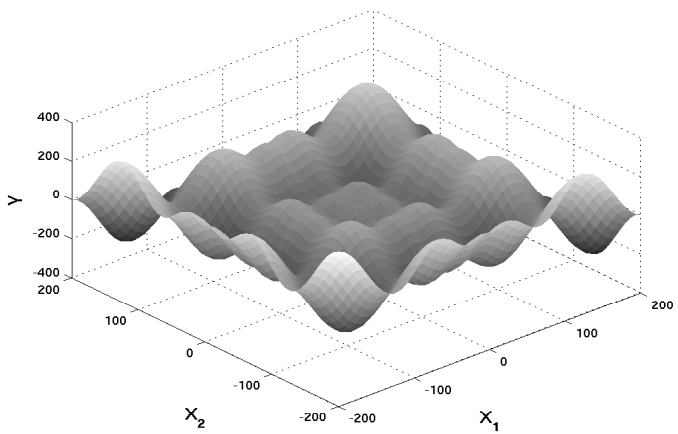

Figure 4. — 2-D Schwefel's function.

Once again, we choose at random n observations in $[-200, 200]^2$ and impose positivity and negativity according to f on N locations selected uniformly on $[-200, 200]^2$. We repeat this procedure 100 times with both the Gaussian and the Matérn 3/2 covariance functions, for several values of n and N. Results on the mean and standard deviation of the Q^2 are reported in Table 6 and Table 7 for the Gaussian and the Matérn 3/2 covariance function, respectively. In both cases, the correlation-free formula

and Genz approximation yield similar results. They both outperform the unconstrained predictor until $n = 200$, where the number of observation is sufficient enough to capture most of the information on f. Again, note that the number of constraint points does not seem to influence the quality of prediction in average.

Table 6. — Mean and standard deviation of Q^2 for different values of n and N with a Gaussian covariance function for the 2-D Schwefel's function.

n	N	Unconstrained	Genz	Correlation-free
n = 50	N = 100	0.29 +/- 0.08	0.57 +/- 0.07	0.56 +/- 0.06
n = 50	N = 225	0.31 +/- 0.09	0.58 +/- 0.14	0.61 +/- 0.19
n = 50	N = 400	0.29 +/- 0.09	0.41 +/- 0.16	0.60 +/- 0.21
n = 100	N = 100	0.52 +/- 0.07	0.68 +/- 0.04	0.68 +/- 0.04
n = 100	N = 225	0.52 +/- 0.08	0.69 +/- 0.09	0.70 +/- 0.11
n = 100	N = 400	0.52 +/- 0.07	0.61 +/- 0.10	0.64 +/- 0.44
n = 150	N = 100	0.66 +/- 0.06	0.76 +/- 0.04	0.76 +/- 0.04
n = 150	N = 225	0.67 +/- 0.05	0.74 +/- 0.07	0.76 +/- 0.06
n = 150	N = 400	0.66 +/- 0.05	0.69 +/- 0.06	0.77 +/- 0.05
n = 200	N = 100	0.75 +/- 0.04	0.81 +/- 0.03	0.81 +/- 0.03
n = 200	N = 225	0.75 +/- 0.04	0.80 +/- 0.05	0.81 +/- 0.05
n = 200	N = 400	0.74 +/- 0.05	0.76 +/- 0.05	0.81 +/- 0.05
n = 300	N = 100	0.84 +/- 0.05	0.87 +/- 0.04	0.87 +/- 0.04
n = 300	N = 225	0.84 +/- 0.05	0.85 +/- 0.03	0.86 +/- 0.03
n = 300	N = 400	0.84 +/- 0.04	0.85 +/- 0.04	0.88 +/- 0.04

Table 7. — Mean and standard deviation of Q^2 for different values of n and N with a Matérn 3/2 covariance function for the 2-D Schwefel's function.

n	N	Unconstrained	Genz	Correlation-free
n = 50	N = 100	0.32 +/- 0.09	0.60 +/- 0.10	0.60 +/- 0.07
n = 50	N = 225	0.32 +/- 0.09	0.69 +/- 0.10	0.69 +/- 0.06
n = 50	N = 400	0.32 +/- 0.10	0.63 +/- 0.11	0.68 +/- 0.04
n = 100	N = 100	0.64 +/- 0.07	0.78 +/- 0.03	0.77 +/- 0.03
n = 100	N = 225	0.65 +/- 0.07	0.81 +/- 0.05	0.79 +/- 0.05
n = 100	N = 400	0.64 +/- 0.08	0.78 +/- 0.05	0.80 +/- 0.03
n = 150	N = 100	0.82 +/- 0.05	0.88 +/- 0.03	0.87 +/- 0.03
n = 150	N = 225	0.83 +/- 0.06	0.88 +/- 0.03	0.87 +/- 0.03
n = 150	N = 400	0.83 +/- 0.04	0.87 +/- 0.03	0.88 +/- 0.02
n = 200	N = 100	0.92 +/- 0.03	0.94 +/- 0.02	0.94 +/- 0.03
n = 200	N = 225	0.92 +/- 0.03	0.93 +/- 0.02	0.93 +/- 0.02
n = 200	N = 400	0.92 +/- 0.03	0.93 +/- 0.03	0.94 +/- 0.02
n = 300	N = 100	0.98 +/- 0.02	0.99 +/- 0.01	0.99 +/- 0.01
n = 300	N = 225	0.98 +/- 0.02	0.98 +/- 0.02	0.98 +/- 0.02
n = 300	N = 400	0.98 +/- 0.02	0.98 +/- 0.02	0.99 +/- 0.01

5. Discussion and conclusion

In this paper, we introduced a new theoretical framework for incorporating constraints in Gaussian process modeling, including bound, monotonicity and convexity constraints. We also extended this framework to any type of linear constraint, such as bounds on some integral of the output. The final constrained predictor is based on a discrete-location approximation of conditional expectations. From a computational perspective, the main result which makes it possible to evaluate the proposed predictors is that a vector encompassing the kriging underlying Gaussian field and any of its derivatives is still Gaussian. Consequently, our constrained predictors can be written as expectations of the truncated multinormal distribution, which has been extensively studied in the past years. We then detailed the explicit formulas on such truncated expectations. Since they involve computation of integrals of dimensionality equal to the number of constraints, we proposed several numerical approximations. The first one is based on a simple correlation-free assumption, the second one on numerical integration tools and the last one calls for sampling techniques. All of them not only yield an approximation of the constrained predictor but also an estimation of the prediction error like in the standard Gaussian process modeling framework. Finally, we compared these predictors on bound, monotonicity and convexity examples. Results showed that incorporation of constraints greatly improve predictions. Future work is necessary to further improve the methodology.

First, we provided examples on 1-D and 2-D functions only. Theoretically, generalization to more dimensions is straightforward. However, the discrete-location approximation for the constraints will require many points and subsequent integral approximations will suffer from the curse of dimensionality. At first glance, two remedies can be envisioned. Following the ideas of covariance tapering and of sequential Gaussian simulation in geostatistics, a solution would be to place a reasonable amount of constraints only in a neighborhood around the prediction point. Another point of view would consist in accomodating sequential strategies already used in Gaussian process modeling for adaptive designs, in the context of efficient global optimization for instance. Starting from given initial locations of constraint points, additional locations would be proposed where the predictor is more likely to violate the constraints, in a sequential way. Such relevant constraint points can be easily identified with the Gaussian process assumption.

Second, maximum-likelihood estimation of kriging hyperparameters should be investigated for consistency under constraint assumptions. In practice, this should also limit the number of constraint points needed for

an effective discrete-location approximation. Further, error bounds on this discrete-location approximation should be examined. Apart from quality control, they could also be used in order to propose suitable locations for the constraints.

Acknowledgements. — This work has been partially supported by the French National Research Agency (ANR) through COSINUS program (project COSTA-BRAVA ANR-09-COSI-015).

Bibliography

[1] Abrahamsen (P.) and Benth (F.E.). — Kriging with inequality constraints. Mathematical Geology, 33(6), p. 719-744 (2001).

[2] Azaïs (J.-M.) and Wschebor (M.). — Level sets and extrema of random processes and fields. New York: Wiley (2009).

[3] Bigot (J.) and Gadat (S.). — Smoothing under diffeomorphic constraints with homeomorphic splines. SIAM Journal on Numerical Analysis, 48(1), p. 224-243 (2010).

[4] Chopin (N.). — Fast simulation of truncated Gaussian distributions. Statistics and Computing, 21, p. 275-288 (2011).

[5] Cozman (F.) and Krotkov (E.). — Truncated Gaussians as Tolerance Sets. Fifth Workshop on Artificial Intelligence and Statistics, Fort Lauderdale Florida (1995).

[6] Cramér (H.) and Leadbetter (M.R.). — Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications. New York: Wiley (1967).

[7] Da Veiga (S.), Wahl (F.) and Gamboa (F.). — Local Polynomial Estimation for Sensitivity Analysis on Models With Correlated Inputs Technometrics, 51(4), p. 452-463 (2009).

[8] Dette (H.) and Scheder (R.). — Strictly monotone and smooth nonparametric regression for two or more variables. The Canadian Journal of Statistics, 34(44), p. 535-561 (2006).

[9] Ellis (N.) and Maitra (R.). — Multivariate Gaussian Simulation Outside Arbitrary Ellipsoids. Journal of Computational and Graphical Statistics, 16(3), p. 692-798 (2007).

[10] Fernandez (P.J.), Ferrari (P.A.) and Grynberg (S.P.). — Perfectly random sampling of truncated multinormal distributions. Adv. in Appl. Probab., 39(4), p. 973-990 (2007).

[11] Genz (A.). — Numerical Computation of Multivariate Normal Probabilities. J. Comp. Graph Stat., 1, p. 141-149 (1992).

[12] Genz (A.). — Comparison of Methods for the Computation of Multivariate Normal Probabilities. Computing Science and Statistics, 25, p. 400-405 (1993).

[13] Genz (A.) and Bretz (F.). — Computation of Multivariate Normal and t Probabilities. Lecture Notes in Statistics, Vol. 195, Springer-Verlag, Heidelberg (2009).

[14] Geweke (J.). — Efficient simulation from the multivariate normal and student t-distribution subject to linear constraints. Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface, p. 571-578 (1991).

[15] Ginsbourger (D.), Bay (X.) and Carraro (L.). — Noyaux de covariance pour le Krigeage de fonctions symétriques. submitted to C. R. Acad. Sci. Paris, section Maths (2009).

[16] Griffiths (W.). — A Gibbs' sampler for the parameters of a truncated multivariate normal distribution. Working Paper, http://ideas.repec.org/p/mlb/wpaper/856.html (2002).

[17] Hall (P.) and Huang (L.-S.). — Nonparametric kernel regression subject to monotonicity constraints. The Annals of Statistics, 29(3), p. 624-647 (2001).

[18] Hazelton (M.L.) and Turlach (B.A.). — Semiparametric regression with shape-constrained penalized slpines. Computational Statistics and Data Analysis, 55, p. 2871-2879 (2011).

[19] Horrace (W.C.). — Some results on the multivariate truncated normal distribution. Journal of Multivariate Analysis, 94, p. 209-221 (2005).

[20] Kleijnen (J.P.C.) and van Beers (W.C.M.). — Monotonicity-preserving bootstrapped Kriging metamodels for expensive simulations. Working Paper, http://www.tilburguniversity.edu/research/institutes-and-research-groups/center/staff/kleijnen/monotone_Kriging.pdf (2010).

[21] Kotecha (J.H.) and Djuric (P.M.). — Gibbs sampling approach for generation of truncated multivariate gaussian random variables. IEEE Computer Society, p. 1757-1760 (1999).

[22] Kotz (S.), Balakrishnan (N.) and Johnson (N.L.). — Continuous multivariate distributions, Volume 1: models and applications New York: Wiley (2000).

[23] Lee (L.-F.). — On the first and second moments of the truncated multi-normal distribution and a simple estimator. Economics Letters, 3, p. 165-169 (1979).

[24] Lee (L.-F.). — The determination of moments of the doubly truncated multivariate tobit model. Economics Letters, 11, p. 245-250 (1983).

[25] Marrel (A.), Iooss (B.), Van Dorpe (F.) and Volkova (E.). — An efficient methodology for modeling complex computer codes with Gaussian processes. Computational Statistics and Data Analysis, 52, p. 4731-4744 (2008).

[26] Michalak (A.M.). — A Gibbs sampler for inequality-constrained geostatistical interpolation and inverse modeling. Water Resour. Res., 44, W09437, doi:10.1029/2007WR006645 (2008).

[27] Muthén (B.). — Moments of the censored and truncated bivariate normal distribution. British Journal of Mathematical and Statistical Psychology, 43, p. 131-143 (1990).

[28] Oakley (JE.) and O'Hagan (A.). — Probabilistic sensitivity analysis of complex models: A Bayesian approach. Journal of the Royal Statistical Society, Series B, 66, p. 751-769 (2004).

[29] Philippe (A.) and Robert (C.). — Perfect simulation of positive Gaussian distributions. Statistics and Computing, 13(2), p. 179-186 (2003).

[30] Racine (J.S.), Parmeter (C.F.) and Du (P.). — Constrained nonparametric kernel regression: Estimation and inference. Working Paper, http:/economics.ucr.edu/spring09/Racine paper for 5 8 09.pdf (2009).

[31] Ramsay (J.O.) and Silverman (B.W.). — Functional Data Analysis. Springer Series in Statistics, Springer-Verlag (2005).

[32] Rasmussen (C.E.) and Williams (C.K.I.). — Gaussian Processes for Machine Learning (2006). The MIT Press.

[33] Robert (C.P.). — Simulation of truncated normal variables. Statistics and Computing, 5, p. 121-125 (1995).

[34] Rodriguez-Yam (G.), Davis (R.A.) and Scharf (L.). — Efficient Gibbs Sampling of Truncated Multivariate Normal with Application to Constrained Linear Regression. Working Paper, http://www.stat.columbia.edu/ rdavis/papers/CLR.pdf (2004).

[35] Sacks (J.), Welch (W.), Mitchell (T.) and Wynn (H.). — Design and analysis of computer experiments. Statistical Science, 4, p. 409-435 (1989).

[36] Saltelli (A.), Chan (K.) and Scott (E.M.) (Eds.). — Sensitivity Analysis. Wiley (2000).

[37] Santner (T.), Williams (B.) and Notz (W.). — The design and analysis of computer experiments. Springer (2003).

[38] Tallis (G.M.). — The moment generating function of the truncated multinormal distribution. Journal of the Royal Statistical Society, Series B, 23(1), p. 223-229 (1961).

[39] Tallis (G.M.). — Elliptical and radial truncation in normal populations. Ann. Math. Statist., 34, p. 940-944 (1963).

[40] Tallis (G.M.). — Plane truncation in normal populations. Journal of the Royal Statistical Society, Series B, 27(2), p. 301-307 (1965).

[41] Yoo (E.-H.) and Kyriakidis (P.C.). — Area-to-point Kriging with inequality-type data. Journal of Geographical Systems, 8(4), p. 357 (2006).