Bayesian Local Kriging

Luc Pronzato, Joaõ Rendas

Bayesian Local Kriging*

Luc Pronzato and Joaõ Rendas

CNRS, Laboratoire I3S, UMR 7271, Université de Nice-Sophia Antipolis/CNRS
Bât. Euclide, Les Algorithmes, 2000 route des lucioles,
06900 Sophia Antipolis cedex, France
pronzato@i3s.unice.fr
rendas@i3s.unice.fr

December 12, 2014

Abstract

We consider the problem of constructing metamodels for computationally expensive simulation codes; that is, we construct interpolation/prediction of functions values (responses) from a finite collection of evaluations (observations). We use Gaussian process modeling and Kriging, and combine a Bayesian approach, based on a finite set of covariance functions, with the use of localized models, indexed by the point where the prediction is made. Our approach does not yield a single generative model for the unknown function, but by letting the weights of the different covariance functions depend on the prediction site, it gives enough flexibility for predictions to accommodate to non-stationarity. Contrary to Kriging prediction with plug-in parameter estimates, the resulting Bayesian predictor is constructed explicitly, without requiring any numerical optimization. It inherits the smoothness properties of the covariance functions that are used and its superiority over the plug-in Kriging predictor (sometimes also called empirical-best-linear-unbiased predictor) is illustrated on various examples, including the reconstruction of an oceanographic field from a small number of observations.

keywords prediction; interpolation; computer experiments; random field; non-stationary process

1 Introduction and motivation

The usual approach to Gaussian Process (GP) modeling and Kriging prediction raises two major issues: (*i*) stationarity is often a too strong assumption but seems hardly avoidable when a single realization of the random field is observed; (*ii*) the estimation of the kernel parameters that specify the correlation between distant observations is problematic, taking the uncertainty on these estimates into account in the construction of predictions is difficult and requires either heavy Monte-Carlo calculations or relies on (sometimes crude) approximations based on the asymptotic behavior of the estimators.

The classical plug-in approach, with which we shall compare, consists in predicting with a correlation function where unknown parameters are replaced by their estimated values, usually by Maximum Likelihood (ML). This method is simple but may produce poor results, in particular (*a*) when the modeled phenomenon is strongly non-stationary, (*b*) when an unlucky poor sampling wrongly suggests that the process is exaggeratedly smooth (this corresponds to the

*This work was partly supported by the ANR project 2011-IS01-001-01 DESIRE (DESIgns for spatial Random fiElds), joint with the Satistics Department of the JKU Universität, Linz, Austria.

notion of deceptive function, see Jones (2001)), or (c) far away from the design points where the process is observed, because Kriging prediction is then given by the global trend and may be locally very inaccurate. Such situations will be considered in Section 3.

To address difficulties (i) and (ii) above, we propose a local-Kriging approach that combines two features.

First, the approach is Bayesian. It relies on a finite set of L GP models $\{Z_\ell(x)\}_{\ell=1}^L$, for x varying in a given set $\mathscr{X}$, $Z_\ell(\cdot)$ having covariance function C_ℓ. We then consider that $f(\cdot)$ is the sample path of a process $Z_s(\cdot)$ such that $s = \ell$ with some probability w^ℓ. Starting with prior weights w_0^ℓ (for instance uniform), we can update them into w_n^ℓ after n observations $\mathbf{z}_n = (Z(x_1), \ldots, Z(x_n))^\top$ have been collected, and hence construct a Bayesian predictor $x \in \mathscr{X} \longmapsto \hat{\eta}_n(x)$ based on the L models. This prediction and its posterior squared error can be constructed explicitly when we assume a linear parametric trend $\mathbf{g}^\top(x)\beta$, with $\mathbf{g}(\cdot)$ a known vector of functions (the usual framework for universal Kriging), and a hierarchical prior $\beta|\sigma^2 \sim \mathscr{N}(\beta_0, \sigma^2\mathbf{V}_0)$ and $\sigma^2 \sim$ inverse chi-squared, common to the L models. This is rather standard, see Santner et al. (2003, Chap. 4). Notice that, due to the dependence of the w_n^ℓ in $\mathbf{z}_n$, this Bayesian predictor depends non linearly in $\mathbf{z}_n$. The method differs from the approach based on mixture of kernels in (Ginsbourger et al., 2008), which is not Bayesian, and it extends (Benassi et al., 2011) by allowing the possible assignment of prior weights to different covariance structures. Several examples (Section 3) will illustrate that a small number L (typically $L \approx 4$) of isotropic covariances is enough to obtain satisfactory results, at least for the two-dimensional sets $\mathscr{X}$ considered.

The second feature of our construction is meant to account for non-stationarity. Instead of proposing a unique view of the process realization, or in other words, of the phenomenon to be modeled, we consider that observers at different locations t and t' may contemplate different models. We thus condition all process characteristics, in particular the covariances C_ℓ, by the location t where prediction is made. This yields L covariance functions $C_{\ell|t}$ for each t, and a Bayesian predictor $\hat{\eta}_n(\cdot|t)$. By construction, the prediction $\hat{\eta}_n(x_0|t)$ at x_0 is only valid locally, for x_0 close to t, but the predictor $t \in \mathscr{X} \longmapsto \hat{\eta}_{BLK,n}(t) = \hat{\eta}_n(t|t)$, which we call Bayesian Local Kriging (BLK), inherits continuity and interpolating properties from the properties of the covariances $C_{\ell|t}$. The method differs from previously proposed local Kriging approaches, see for instance Lam et al. (2004); Sun et al. (2006); Nguyen-Tuong et al. (2009). It does not yield a single generative model for the process, but by assigning different posterior weights $w_n^\ell(t)$ to different locations t, it gives enough flexibility for predictions to accommodate to non-stationarity.

2 Construction

The objective is to reconstruct/predict/interpolate a function $f(\cdot)$: $x \in \mathscr{X} \longmapsto f(x) \in \mathbb{R}$ from a collection of observations $f(x_1), \ldots, f(x_n)$ at sites $x_1, \ldots, x_n$ in $\mathscr{X}$, with $\mathscr{X}$ a compact subset of $\mathbb{R}^d$, $d \geq 1$. Real-valued GP models will be used to predict $f(x_0)$ at some unsampled x_0 by Kriging. Since there is no reason to consider $f(\cdot)$ as the sample path of a stationary GP, we shall use collections of local models, each of them defining a local prior representation for $f(\cdot)$. We shall write $Z(\cdot) \sim \mathsf{GP}(\mu(\cdot), \sigma^2 C(\cdot,\cdot))$ when $Z(\cdot)$ is a GP satisfying

$$\mathsf{E}\{Z(x)\} = \mu(x)\,,\ \forall x \in \mathscr{X}\,,\ \text{and } \mathsf{E}\{[Z(x) - \mu(x)][Z(x') - \mu(x')]\} = \sigma^2\, C(x, x')\,,\ \forall (x, x') \in \mathscr{X}^2\,.$$

2.1 Local models

For the sake of simplicity, for the moment we consider processes with zero mean; see the Appendix for the case where a linear parametric trend is present. At any site $t \in \mathscr{X}$, we shall use

a set of L local GP models $\mathsf{GP}(0, \sigma^2 C_{\ell|t}(\cdot,\cdot))$ for $f(\cdot)$. Non stationarity is introduced by letting $C_{\ell|t}(x,x')$ depend on t, which makes the model local. We shall use covariance functions of the form

$$C_{\ell|t}(x,x') = \frac{k_{0,\ell}(x-x')}{k_{1,\ell}^{1/2}(x-t)k_{1,\ell}^{1/2}(x'-t)}\,, \tag{1}$$

where the $k_{i,\ell}(\cdot)$, $i=0,1$, $\ell=1,\ldots,L$, are stationary kernel functions, continuous at 0 and such that $k_{i,\ell}(0)=1$. The covariance $C_{\ell|t}(x,x')$ satisfies $C_{\ell|t}(x,x) = k_{1,\ell}^{-1}(x-t)$ and $C_{\ell|t}(t,t)=1$. By taking $k_{1,\ell}(x-t)$ decreasing in $\|x-t\|$, we obtain a prior for $f(\cdot)$ which may become quite vague for x far enough from t (however, the correlation corresponding to $C_{\ell|t}(x,x')$ is $k_{0,\ell}(x-x')$ for any t and is thus stationary).

2.2 The Bayesian local Kriging predictor

Consider a given site $t \in \mathscr{X}$. With each model $\mathsf{GP}(0, \sigma^2 C_{\ell|t}(\cdot,\cdot))$, $\ell=1,\ldots,L$, with $C_{\ell|t}(\cdot,\cdot)$ given by (1), we associate a prior probability w_0^ℓ. That is, we consider that, seen from t, $f(\cdot)$ is the sample path of a stochastic process $Z(\cdot)$ such that

$$Z(\cdot)|s(t),\sigma^2,t \sim \mathsf{GP}(0, \sigma^2 C_{s(t)|t}(\cdot,\cdot))\,,\ \text{with Prob}\{s(t)=\ell\} = w_0^\ell\,,\ \ell=1,\ldots,L\,, \tag{2}$$

with some prior distribution on σ^2 (common to all components), having density $\varphi_0(\cdot)$ with respect to the Lebesgue measure on $\mathbb{R}^+$. We thus have a finite mixture of GP's at each t, and different mixtures at t and $t' \neq t$.

Remark 1 (Linear combination of GP) *Note the difference between this Bayesian construction and the more usual consideration of a linear combination of processes, that is, $Z(x) = \sum_{\ell=1}^L w^\ell(x)\, Z_\ell(x)$, where $Z_\ell(\cdot) \sim \mathsf{GP}(0, \sigma^2 C_\ell(\cdot,\cdot))$, see, e.g., Nott and Dunsmuir (2002). In that case, assuming that $\mathsf{E}\{Z_\ell(x)Z_{\ell'}(x')\} = 0$ for all $(x,x') \in \mathscr{X}^2$ when $\ell \neq \ell'$, we obtain that $\mathsf{E}\{Z(x)Z(x')\} = \sigma^2 \sum_{\ell=1}^L w^\ell(x)w^\ell(x')C_\ell(x,x')$, which makes $Z(\cdot)$ non stationary also when $C_\ell(x,x') = C_\ell(x-x')$ for all (x,x'). This model may seem simpler to handle than the Bayesian hierarchical one considered above. However, predictions require the estimation of the weight functions $w^\ell(\cdot)$, which is hardly possible when a single realization of $Z(\cdot)$ is available and no prior parametric model is available for $w^\ell(x)$.* □

After n evaluations of $f(\cdot)$ at sites $\xi_n = (x_1,\ldots,x_n)$ in $\mathscr{X}$, the likelihood $\mathscr{L}(\mathbf{z}_n|\ell,\sigma^2,t)$ of $\mathbf{z}_n = (f(x_1),\ldots,f(x_n))^\top$ for the model $\mathsf{GP}(0,\sigma^2 C_{\ell|t}(\cdot,\cdot))$ is given by

$$\mathscr{L}(\mathbf{z}_n|\ell,\sigma^2,t) = \frac{1}{\sigma^n(2\pi)^{n/2}\det^{1/2}\mathbf{K}_n(\ell|t)}\,\exp\left[-\frac{1}{2\sigma^2}\,\mathbf{z}_n^\top \mathbf{K}_n^{-1}(\ell|t)\mathbf{z}_n\right]\,,$$

where $\mathbf{K}_n(\ell|t)$ is the matrix with elements $\{\mathbf{K}_n(\ell|t)\}_{i,j} = C_{\ell|t}(x_i,x_j)$, $i,j=1,\ldots,n$. Notice that $\mathbf{K}_n(\ell|t) = \mathbf{D}_n(\ell|t)\mathbf{K}_{n,0}(\ell)\mathbf{D}_n(\ell|t)$, where

$$\{\mathbf{K}_{n,0}(\ell)\}_{i,j} = k_{0,\ell}(x_i-x_j) \text{ and } \mathbf{D}_n(\ell|t) = \text{diag}\{k_{1,\ell}^{-1/2}(x_i-t)\,,\ i=1,\ldots,n\}\,,$$

see (1). Therefore, as we stressed before, for suitable kernels $k_{1,\ell}(\cdot)$, the information carried by observation $z_i = f(x_i)$ decreases with the distance from t to x_i.

Using Bayes rule, the marginal likelihood $\mathscr{L}(\mathbf{z}_n|\ell,t) = \int_{\mathbb{R}^+} \mathscr{L}(\mathbf{z}_n|\ell,\sigma^2,t)\,\varphi_0(\sigma^2)\,\mathrm{d}\sigma^2$ is used to compute the posterior weights

$$w_n^\ell(t) = \frac{w_0^\ell \mathscr{L}(\mathbf{z}_n|\ell,t)}{\sum_{\ell'=1}^L w_0^{\ell'}\mathscr{L}(\mathbf{z}_n|\ell',t)}\,,\ \ell=1,\ldots,L\,; \tag{3}$$

that is, the posterior distribution of $s(t)$.

Remark 2 *The different covariance functions $C_{\ell|t}(\cdot,\cdot)$, $\ell = 1,\ldots,L$, may correspond to different parameter values θ_l in a given parameterized kernel $C_{\ell|t}(\cdot,\cdot|\theta)$, typically different length scales. The method can then be extended straightforwardly to the infinite mixture case, with a continuous limit that gives some prior density $\pi_0(\cdot)$ to θ. The posterior is then*

$$\pi_n(\theta) = \frac{\pi_0(\theta)\mathscr{L}(\mathbf{z}_n|\theta,t)}{\int \pi_0(\theta)\mathscr{L}(\mathbf{z}_n|\theta,t)\,\mathrm{d}\theta}\,.$$

However, in general the integral on the denominator cannot be calculated analytically, and one must resort to MCMC methods or use the Laplace approximation. Note that, besides being more easily tractable, the finite mixture model considered here allows one to let the different $C_{\ell|t}(\cdot,\cdot)$ represent different correlation characteristics (isotropy, smoothness, etc.). □

Let η_0 denote the prediction of the unobserved value $Z(x_0)$ for the local model at site t. Its posterior squared prediction error is

$$\mathsf{PSPE}(\eta_0) = \mathsf{E}\{[Z(x_0)-\eta_0]^2|\mathbf{z}_n,t\}\,,$$

where the expectation is with respect to $Z(x_0)$ given $\mathbf{z}_n$ and t, with $Z(x_0)|s(t),\sigma^2,\mathbf{z}_n,t \sim \mathsf{GP}(0,\sigma^2 C_{s(t)|t}(\cdot,\cdot)|\mathbf{z}_n)$, $\mathrm{Prob}\{s(t)=\ell\} = w_0^\ell$, $l = 1,\ldots,L$ (which gives again a finite mixture of GP), and $\sigma^2 \sim \varphi_0(\cdot)$. Denote

$$\hat{\eta}_n(x_0|t) = \mathsf{E}\{Z(x_0)|\mathbf{z}_n,t\}\,. \tag{4}$$

Then, $\mathsf{PSPE}(\eta_0) = \mathsf{var}\{Z(x_0)|\mathbf{z}_n,t\} + [\hat{\eta}_n(x_0|t)-\eta_0]^2$, which is minimum for $\eta_0 = \hat{\eta}_n(x_0|t)$. The associated posterior squared prediction error equals the posterior variance

$$\begin{aligned}\mathsf{PSPE}(\hat{\eta}_n(x_0|t)) &= \mathsf{var}\{Z(x_0)|\mathbf{z}_n,t\} = \sum_{\ell=1}^{L} w_n^\ell(t)\,\mathsf{E}\{[Z(x_0)-\hat{\eta}_n(x_0|t)]^2|\mathbf{z}_n,\ell,t\} \\ &= \sum_{\ell=1}^{L} w_n^\ell(t)\,\mathsf{E}\{[Z(x_0)-\hat{\eta}_n(x_0|\ell,t)]^2|\mathbf{z}_n,\ell,t\} + \sum_{\ell=1}^{L} w_n^\ell(t)\,[\hat{\eta}_n(x_0|\ell,t)-\hat{\eta}_n(x_0|t)]^2\,,\end{aligned}$$

where we have denoted $\hat{\eta}_n(x_0|\ell,t) = \mathsf{E}\{Z(x_0)|\mathbf{z}_n,\ell,t\}$. Since $\hat{\eta}_n(x_0|t) = \sum_{\ell=1}^{L} w_n^\ell(t)\hat{\eta}_n(x_0|\ell,t)$, we obtain

$$\mathsf{PSPE}(\hat{\eta}_n(x_0|t)) = \sum_{\ell=1}^{L} w_n^\ell(t)\,\mathsf{var}\{Z(x_0)|\mathbf{z}_n,\ell,t\} + \mathsf{var}\{d_n(x_0|t)\}\,, \tag{5}$$

where $d_n(x_0|t)$ denotes the discrete distribution that allocates weight $w_n^\ell(t)$ to $\hat{\eta}_n(x_0|\ell,t)$.

Due to the use of localized covariance functions, see (1), the prediction $\hat{\eta}_n(x_0|t)$ only makes sense for x_0 close to t. We call *Bayesian Local Kriging* (BLK) the predictor

$$t \in \mathscr{X} \longmapsto \hat{\eta}_{BLK,n}(t) = \hat{\eta}_n(t|t)\,.$$

Note that it depends nonlinearly on $\mathbf{z}_n$ even in situations where each prediction $\hat{\eta}_n(t|\ell,t)$ is linear in $\mathbf{z}_n$, since the weights w_n^ℓ depend on $\mathbf{z}_n$ through the marginal likelihood $\mathscr{L}(\mathbf{z}_n|\ell,t)$, see (3). Also note that $\hat{\eta}_n(t|t)$ inherits the smoothness properties of $k_{0,\ell}(\cdot)$ and $k_{1,\ell}(\cdot)$. The posterior squared prediction errors, or posterior variances, $\mathsf{PSPE}(\hat{\eta}_{BLK,n}(t)) = \mathsf{var}\{Z(t)|\mathbf{z}_n,t\}$, $t \in \mathscr{X}$, can be used to measure the precision of predictions made after the collection of observations $\mathbf{z}_n$. The preposterior variances, or mean-squared prediction errors,

$$\mathsf{MSPE}(\hat{\eta}_{BLK,n}(t)) = \mathsf{E}\{\mathsf{var}\{Z(t)|\mathbf{z}_n,t\}|t\}\,,\ t \in \mathscr{X}\,, \tag{6}$$

can be used for experimental design: one may choose a set of locations $\xi_n = (x_1, \ldots, x_n) \in \mathscr{X}^n$ that ensures a precise prediction of the behavior of $f(\cdot)$ over $\mathscr{X}$ by minimizing

$$\Phi_M(\xi_n) = \max_{t\in\mathscr{X}} \mathsf{MSPE}(\hat{\eta}_{BLK,n}(t)) \quad \text{or} \quad \Phi_I(\xi_n) = \int_{\mathscr{X}} \mathsf{MSPE}(\hat{\eta}_{BLK,n}(t))\zeta(\mathrm{d}t)\,, \tag{7}$$

with $\zeta(\cdot)$ some suitable measure on $\mathscr{X}$.

As shown in Section 2.3, the predictor (4) and its prediction error (5) can be expressed explicitly when the prior density $\varphi_0(\cdot)$ on σ^2 is suitably chosen (conjugate prior).

2.3 An inverse chi-square prior for the process variance

Suppose that $Z(\cdot)$ satisfies (2) with σ^2 having the inverse chi-squared distribution (common to all components),

$$\varphi_0(\sigma^2) = \varphi_{\sigma_0^2,\nu_0}(\sigma^2) = \frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\,\frac{\exp[-\nu_0\sigma_0^2/(2\sigma^2)]}{\sigma^{2(1+\nu_0/2)}}\,, \tag{8}$$

so that $\mathsf{E}\{1/\sigma^2\} = 1/\sigma_0^2$, $\mathsf{var}\{1/\sigma^2\} = 2/(\nu_0\sigma_0^4)$, $\mathsf{E}\{\sigma^2\} = \sigma_0^2\nu_0/(\nu_0-2)$ (for $\nu_0 > 2$) and $\mathsf{var}\{\sigma^2\} = 2\sigma_0^4\nu_0^2/[(\nu_0-2)^2(\nu_0-4)]$ (for $\nu_0 > 4$). Direct calculations give

$$\int_0^\infty \mathscr{L}(\mathbf{z}_n|\ell,t)\varphi_0(\sigma^2)\,\mathrm{d}\sigma^2 = \frac{1}{(2\pi)^{n/2}\det^{1/2}\mathbf{K}_n(\ell|t)}\,\frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\,\frac{\Gamma(\nu_n/2)}{(\sigma_{n|\ell,t}^2\nu_n/2)^{\nu_n/2}}\,,$$

where $\nu_n = \nu_0 + n$ and

$$\sigma_{n|\ell,t}^2 = \frac{\nu_0\sigma_0^2 + n\hat{\sigma}_n^2(\ell|t)}{\nu_0+n} \tag{9}$$

with $\hat{\sigma}_n^2(\ell|t) = \mathbf{z}_n^\top\mathbf{K}_n^{-1}(\ell|t)\mathbf{z}_n/n$, the Maximum-Likelihood (ML) estimator of σ^2 given $\mathbf{z}_n, \ell, t$. Also, given $\mathbf{z}_n, \ell$ and t, σ^2 has the inverse chi-squared distribution $\varphi_{n|\ell,t}(\cdot) = \varphi_{\sigma^2_{n|\ell,t},\nu_n}(\cdot)$, see (8).

The posterior mean $\hat{\eta}_n(x_0|\ell,t)$ corresponds to the ordinary-Kriging predictor, also called Best Linear Unbiased Predictor (BLUP), $\mathbf{c}_n^\top(x_0|\ell,t)\mathbf{z}_n$ for the model $\mathsf{GP}(0, \sigma^2 C_{\ell|t}(\cdot,\cdot)|\mathbf{z}_n)$, with,

$$\mathbf{c}_n(x_0|\ell,t) = \mathbf{K}_n^{-1}(\ell|t)\mathbf{k}_n(x_0,\ell|t) \tag{10}$$

and $\{\mathbf{k}_n(x_0,\ell|t)\}_i = C_{\ell|t}(x_0,x_i)$, $i = 1,\ldots,n$. Therefore,

$$\hat{\eta}_n(x_0|t) = \sum_{\ell=1}^L w_n^\ell(t)\mathbf{c}_n^\top(x_0|\ell,t)\mathbf{z}_n\,. \tag{11}$$

The variance $\mathsf{var}\{Z(x_0)|\mathbf{z}_n,\ell,\sigma^2,t\}$ is the ordinary-Kriging variance for $\mathsf{GP}(0,\sigma^2C_{\ell|t}(\cdot,\cdot)|\mathbf{z}_n)$, $\mathsf{var}\{Z(x_0)|\mathbf{z}_n,\ell,\sigma^2,t\} = \sigma^2\,\rho_n^2(x_0|\ell,t)$, with

$$\rho_n^2(x_0|\ell,t) = C_\ell(x_0,x_0|t) - \mathbf{k}_n^\top(x_0,\ell|t)\mathbf{K}_n^{-1}(\ell|t)\mathbf{k}_n(x_0,\ell|t) \tag{12}$$

(note that it depends on the design $\xi_n = (x_1,\ldots,x_n)$ but not on $\mathbf{z}_n$). Since $\mathsf{E}\{Z(x_0)|\mathbf{z}_n,\ell,\sigma^2,t\} = \hat{\eta}_n(x_0|\ell,t)$ does noes not depend on σ^2,

$$\mathsf{var}\{Z(x_0)|\mathbf{z}_n,\ell,t\} = \mathsf{E}\{\sigma^2|\mathbf{z}_n,\ell,t\}\,\rho_n^2(x_0|\ell,t) = \sigma_{n|\ell,t}^2\nu_n/(\nu_n-2)\,\rho_n^2(x_0|\ell,t)\,.$$

From (5) and (11), the posterior squared error of the prediction $\hat{\eta}_n(x_0|t)$ at x_0 equals

$$\mathsf{PSPE}(\hat{\eta}_n(x_0|t)) = \sum_{\ell=1}^L w_n^\ell(t)\,\sigma_{n|\ell,t}^2\,\nu_n/(\nu_n-2)\,\rho_n^2(x_0|\ell,t) + \mathbf{z}_n^\top\Omega_n(x_0|t)\mathbf{z}_n\,, \tag{13}$$

where $\Omega_n(x_0|t)$ is the variance-covariance matrix $\mathsf{Var}\{D_n(x_0|t)\}$ with $D_n(x_0|t)$ the discrete distribution that allocates weight $w_n^\ell(t)$ to the vector $\mathbf{c}_n(x_0|\ell,t)$, $\ell = 1,\ldots,L$.

The preposterior variance (mean-squared prediction error) of BLK at t is (for $\nu_0 > 2$)

$$\begin{aligned}\mathsf{MSPE}(\hat{\eta}_{BLK,n}(t)) &= \sum_{\ell=1}^{L} \mathsf{E}\{w_n^\ell(t)\,\sigma^2_{n|\ell,t}\nu_n/(\nu_n-2)\}\,\rho_n^2(t|\ell,t) + \mathsf{E}\{\mathbf{z}_n^\top \Omega_n(t|t)\mathbf{z}_n|t\} \\ &= \sigma_0^2\,\nu_0/(\nu_0-2)\sum_{\ell=1}^{L} w_0^\ell(t)\,\rho_n^2(t|\ell,t) + \mathsf{E}\{\mathbf{z}_n^\top \Omega_n(t|t)\mathbf{z}_n|t\}\,. \end{aligned} \tag{14}$$

Note that $\mathsf{MSPE}(\hat{\eta}_{BLK,n}(t)) = \mathsf{E}\{\sigma^2\}\,\rho_n^2(t|t)$ when $L = 1$. The term

$$\mathsf{E}\{\mathbf{z}_n^\top \Omega_n(t|t)\mathbf{z}_n|t\} = \sum_{\ell'=1}^{L} w_0^{\ell'}\,\mathsf{E}\{\mathsf{var}\{d_n(x_0|t)\}|\ell',t\}\,,$$

see (5), plays a role similar to that of the correcting term added to the Kriging variance in (Harville and Jeske, 1992; Zimmerman and Cressie, 1992; Abt, 1999; Zhu and Zhang, 2006). The construction of an experimental design optimal in terms of $\Phi_M(\cdot)$ or $\Phi_I(\cdot)$, see (7), would require the evaluation of $\mathsf{E}\{\mathbf{z}_n^\top \Omega_n(t|t)\mathbf{z}_n|t\}$ and deserves further investigations. A sequential approach facilitates the construction: suppose that n observations have already been collected, with $\hat{\eta}_{BLK,n}(t)$ the prediction at t, see (11), and $\mathsf{PSPE}(\hat{\eta}_{BLK,n}(t))$ the associated posterior squared prediction error, see (13); a reasonable choice then places next observation at x_{n+1} where $\mathsf{PSPE}(\hat{\eta}_{BLK,n}(x))$ is maximum.

In this section we have considered GP with zero mean. The presence of a linear parametric trend (universal Kriging) is considered in Section 5: expressions (11), (13) and (14) remain valid, but with different values for ν_n, $\sigma^2_{n|\ell,t}$, $\mathbf{c}_n(x_0|\ell,t)$ and $\rho_n^2(x_0|\ell,t)$.

3 Examples

We shall consider isotropic kernels $k_{i,\ell}(\cdot)$ of the form $k_{i,\ell}(x-x') = K_{\gamma,\theta}(\|x-x'\|)$, $i = 0,1$, where $K_{\gamma,\theta}(\tau)$ denotes a Matérn covariance function, with

$$K_{\gamma,\theta}(\tau) = \begin{cases} (\theta\tau+1)\,\exp(-\theta\tau) & \text{for } \gamma = 3/2 \\ (\theta^2\tau^2/3+\theta\tau+1)\,\exp(-\theta\tau) & \text{for } \gamma = 5/2 \end{cases} \tag{15}$$

as special cases, see, e.g., Stein (1999, pp. 31, 48). For fixed γ and θ, $K_{\gamma,\theta}(\tau)$ is a decreasing function of $\tau \in \mathbb{R}^+$, with a correlation $K_{\gamma,\theta}(\tau) = 20\%$ for $\tau \simeq 2.9943/\theta$ when $\gamma = 3/2$ and for $\tau \simeq 3.9141/\theta$ when $\gamma = 5/2$. The design space $\mathscr{X}$ is often renormalized to $[0,1]^d$, so that prior guesses on reasonable inverse correlation lengths θ can be set depending on the assumed smoothness of the function $f(\cdot)$. One dimensional processes with covariance $K_{\gamma,\theta}(\tau)$ are m times mean-square (and almost surely) differentiable if and only if $\gamma > m$, see Cramér and Leadbetter (1967, p. 185), Stein (1999, p. 32), and $K_{3/2,\theta}(\cdot)$ (respectively $K_{5/2,\theta}(\cdot)$) yields one-time (resp. two-times) isotropic differentiable processes on $\mathbb{R}^d$.

Example 1: average performance. We simulate various realizations of stochastic processes on $\mathscr{X} = [0,1]^2$, stationary or not, using a linear combination of processes as mentioned in Remark 1,

$$Z(x) = [1-w(x)]Z_1(x) + w(x)Z_2(x)\,, \tag{16}$$

where $Z_1(\cdot)$ and $Z_2(\cdot)$ are stationary, with mean, variance and covariance functions $\bar{\beta}_i$, σ_i^2 and $K_{\gamma_i,\theta_i}(\cdot)$, $i = 1,2$, see (15). For the stationary case we shall use $w(x) \equiv 1$, so that $Z(x) = Z_2(x)$

for all x; realizations of a non-stationary process will be generated with $w(x) = w_{\mathrm{n-s}}(x)$ given by the product of two sigmoid functions,

$$w_{\mathrm{n-s}}(x) = \frac{\exp[a(x_1 - 1/2)]}{\exp[a(x_1 - 1/2)] + \exp[-a(x_1 - 1/2)]} \frac{\exp[a(x_2 - 1/2)]}{\exp[a(x_2 - 1/2)] + \exp[-a(x_2 - 1/2)]},$$

$x = (x_1, x_2) \in \mathscr{X}$. For a large enough (we take $a = 30$ in the numerical experiments below), $Z(x)$ is approximately equal to $Z_2(x)$ for x_1 and x_2 larger than $1/2$ and is approximately equal to $Z_1(x)$ otherwise. The experimental design $\xi_n = (x_1, \ldots, x_n)$ is a random Latin hypercube with $n = 20$ points in $\mathscr{X}$, see Figure 1.

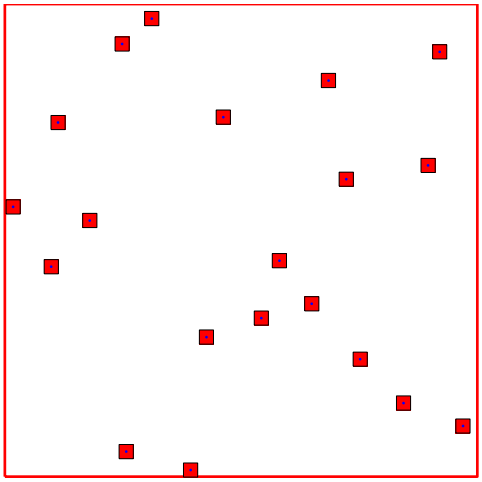

Figure 1: Experimental design ξ_n in Example 1.

Two predictors are compared. The first one is the Empirical BLUP (EBLUP). A stationary GP model is assumed, with unknown constant mean β, unknown variance σ^2 and covariance function $K_{\gamma_1,\theta}(\cdot)$ with unknown θ. The parameters β, σ and θ are estimated by Restricted ML, see Section 5.1; these estimated values are then plugged in the ordinary Kriging predictor (the BLUP) — it corresponds to (21) in the Appendix, with $\mathbf{g}(x) \equiv 1$, $\ell = 1$ and no conditioning on t. We denote by $\hat{\eta}_{MV,n}(t)$ this prediction at t.

The second one is the BLK predictor $\hat{\eta}_{BLK,n}(t)$ of Section 2.2, with $L = 4$ and covariance functions given by (1) with $k_{0,\ell}(x-x') = K_{3/2,\theta_{0,\ell}}(\|x-x'\|)$ and $k_{1,\ell}(x-x') = K_{3/2,\theta_{1,\ell}}(\|x-x'\|)$, $\ell = 1, \ldots, 4$. We suppose that $\mathbf{g}(x) \equiv 1$ and that β has a uniform prior, see Section 5.1; we take $\gamma_0 = 2$, $\sigma_0 = 1$ which corresponds to a very vague prior on σ^2.

We construct predictions on a regular grid of $21 \times 21 = 441$ points s_i in $\mathscr{X}$ and consider different values for the parameters of the process (16): $\theta_1 = 1$, $\bar{\beta}_1 = 0$, $\sigma_1 = 1$, $\gamma_1 = 3/2$ or $5/2$, $\gamma_2 = 3/2$, $\theta_2 = 7$, $\sigma_2 = 2$, $\bar{\beta}_2 = 0$ or 10. Note that the EBLUP assumes that the covariance is $K_{\gamma_1,\theta}(\cdot)$. BLK uses $\theta_{0,\ell} = 1, 5, 10, 20$ and either $\theta_{1,\ell} = 0$ (i.e., it assumes stationarity) or $\theta_{1,\ell} = \theta_* > 0$ for all $\ell = 1, \ldots, 4$ (non-stationarity). The value θ_* is chosen according to the following rule. Assuming that the design ξ_n is space-filling, we wish to ensure that, for any $x \in \mathscr{X}$, $k_{1,\ell}(x - x_i)$ is large enough (say, larger than 20%) for a significant set of design points $x_i \in \xi_n$ (say, $10d/4$ points). For the covariance Matérn 3/2, we obtain that the hypercube with side length $\tau_* = 2.9943/\theta_*$ should contain $10d/4$ points, which yields $\tau_* = \min\{1, [10d/(4n)]^{1/d}\}$. For $n = 10d$ and $d = 2$, this gives $\tau^* = 0.5$ and $\theta_* \simeq 5.9886$. The different configurations considered are indicated in Table 1. In columns A-C the process $Z(x)$ is stationary: in A and B the EBLUP uses the correct covariance function, BLK does not assume stationarity in A but does in B and C; the EBLUP has a wrong covariance function in C. The process $Z(x)$ is non-stationary in D and E, with $Z_1(x)$ and $Z_2(x)$ having different covariances in D and also different

means in E. Figure 2 presents realizations of the process $Z(x)$ for configurations corresponding to columns D and E of Table 1.

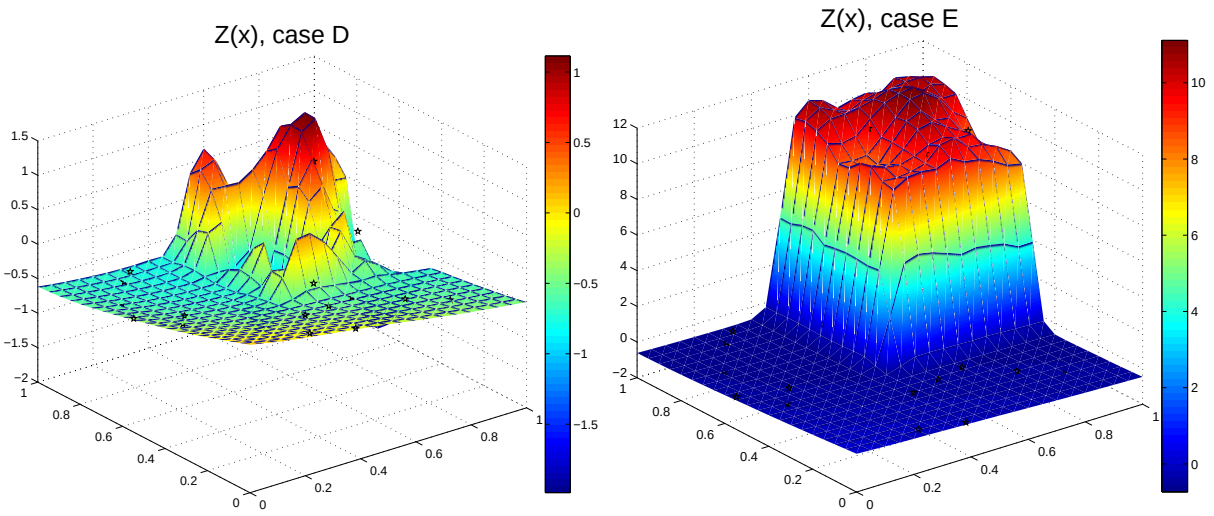

Figure 2: Realizations of a non-stationary process $Z(x)$ with $w(x) = w_{\mathrm{n-s}}(x)$, for configurations D (left) and E (right) of Table 1.

For each choice of covariance structure for the process $Z(\cdot)$ we repeated 1,000 independent simulations of the 20-dimensional vector of observations $\mathbf{z}_n$. To compare the performances of the two predictors without simulating realizations of $Z(s_i)$ for the 21×21 grid points, one may notice that, for any predictor $\hat{z}_n(x_0)$ which is a function of $\mathbf{z}_n$, we have

$$E^2[\hat{z}_n(x_0)] = \mathsf{E}\{[\hat{z}_n(x_0) - Z(x_0)]^2 | \mathbf{z}_n\} = \Delta_n^2[\hat{z}_n(x_0)] + V_n(x_0)\,,$$

with $\Delta_n^2[\hat{z}_n(x_0)] = [\hat{z}_n(x_0) - \mathsf{E}\{Z(x_0)|\mathbf{z}_n\}]^2$ and $V_n(x_0) = \mathsf{E}\{[Z(x_0) - \mathsf{E}\{Z(x_0)|\mathbf{z}_n\}]^2|\mathbf{z}_n\}$, where $\mathsf{E}\{Z(x_0)|\mathbf{z}_n\}$ and $V_n(x_0)$ can easily be calculated using the characteristics of $Z(\cdot)$, see Remark 1. The values of the integrated squared errors

$$\{IE^2[\hat{z}_n]\}_k = \frac{1}{441}\sum_{i=1}^{441}\left\{E^2[\hat{z}_n(s_i|\mathbf{z}_n^{(k)})]\right\}_k$$

are then calculated for each vector of simulated data $\mathbf{z}_n^{(k)}$, for both predictors $\hat{\eta}_{MV,n}$ and $\hat{\eta}_{BLK,n}$. The empirical means

$$\widetilde{IE^2}[\hat{z}_n] = \frac{1}{1,000}\sum_{k=1}^{1,000}\{IE^2[\hat{z}_n]\}_k$$

of the integrated squared errors are indicated in Table 1, together with the results of paired-comparisons tests T_{pc} for the differences $\{IE^2[\hat{\eta}_{MV,n}]\}_k - \{IE^2[\hat{\eta}_{BLK,n}]\}_k$ and their corresponding p-values (see, e.g., Kanji (1993, p. 30)). A box-plot of these differences is presented in Figure 3, indicating that BLK produces more stable predictions than the EBLUP. We also indicate in Table 1 the values $\widetilde{MedE^2}[\hat{z}_n]$ and $\widetilde{MaxE^2}[\hat{z}_n]$ of the empirical means, over the 1,000 simulations, of respectively the median and maximum values of $\left\{E^2[\hat{z}_n(s_i|\mathbf{z}_n^{(k)})]\right\}_k$ on the 441 grid points.

The values of the paired-comparisons tests T_{PC} and associated p-values in Table 1 indicate that conclusions about the best predictor (in terms of integrated squared errors) between $\hat{\eta}_{MV,n}$ and $\hat{\eta}_{BLK,n}$ are highly significant. The EBLUP, that is, ordinary Kriging with ML estimation of the process parameters, appears to yield more precise predictions (on average) than BLK in

	A	B	C	D	E
$Z(\cdot)$	stationary	stationary	stationary	non-stationary	non-stationary
$w(x)$	$\equiv 1$	$\equiv 1$	$\equiv 1$	$w_{\mathrm{n-s}}(x)$	$w_{\mathrm{n-s}}(x)$
γ_1	3/2	3/2	5/2	5/2	5/2
$\bar{\beta}_2$	10	10	10	0	10
$\theta_{1,\ell}$	θ_*	0	0	θ_*	θ_*
$\widetilde{IE^2}[\hat{\eta}_{MV,n}]$	**0.971**	0.971	0.975	0.301	2.591
$\widetilde{IE^2}[\hat{\eta}_{BLK,n}]$	0.989	**0.961**	**0.961**	**0.295**	**2.307**
T_{pc}	-6.49	4.96	7.33	6.24	99.7
p-value	$< 10^{-10}$	$3.6\ 10^{-7}$	$< 10^{-10}$	$2.2\ 10^{-10}$	$< 10^{-10}$
$\widetilde{MedE^2}[\hat{\eta}_{MV,n}]$	**0.787**	0.787	0.786	0.0109	0.119
$\widetilde{MedE^2}[\hat{\eta}_{BLK,n}]$	0.808	0.787	0.787	**0.0025**	**0.0118**
$\widetilde{MaxE^2}[\hat{\eta}_{MV,n}]$	**3.795**	3.795	3.797	4.412	53.60
$\widetilde{MaxE^2}[\hat{\eta}_{BLK,n}]$	4.234	**3.578**	**3.578**	**4.354**	**52.40**
$\widehat{IE^2}[\hat{\eta}_{MV,n}]$	0.937	0.937	0.893	0.202	1.325
$\widehat{IE^2}[\hat{\eta}_{BLK,n}]$	1.032	1.001	1.001	0.266	1.668

Table 1: Simulation results in Example 1. Best performances (smallest values) among $\hat{\eta}_{MV,n}$ and $\hat{\eta}_{BLK,n}$ are indicated in bold face.

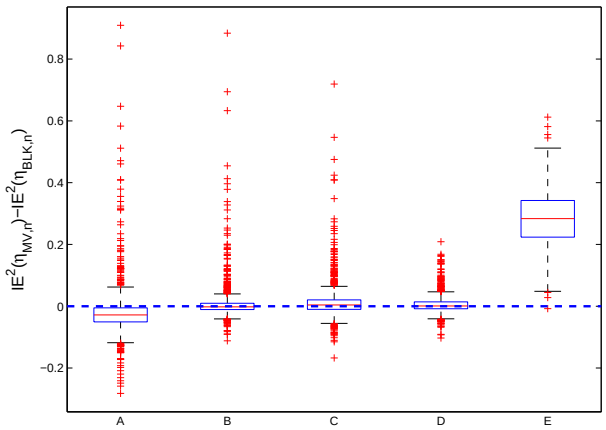

Figure 3: Box-plot of the differences $\{IE^2[\hat{\eta}_{MV,n}]\}_k - \{IE^2[\hat{\eta}_{BLK,n}]\}_k$ for the 5 cases considered in Table 1.

one situation only among those considered (column A of Table 1): the process is stationary, the EBLUP has the correct covariance function and BLK does not assume stationarity. For the same random process, the situation is reversed when BLK makes use of the stationarity assumption (column B); the superiority of BLK over the EBLUP increases when $\hat{\eta}_{MV,n}$ wrongly assumes a Matérn 5/2 covariance (column C). BLK was always found to be superior to the EBLUP when $Z(\cdot)$ is non-stationary, as illustrated by columns D and E, where $Z_1(\cdot)$ and $Z_2(\cdot)$ differ by their regularity (column D) and also by their mean (column E). Note that predictions are much more precise in column D than in the others, due to the fact that that the process is quite smooth ($\gamma_1 = 5/2$) and has a rather large correlation length ($\theta_1 = 1$) on a big part of $\mathscr{X}$; see in particular the small values of $\widetilde{MedE^2}[\hat{\eta}_{MV,n}]$ and $\widetilde{MedE^2}[\hat{\eta}_{BLK,n}]$ in columns D and E.

We also indicate in the table the values of the empirical errors predicted by the two modeling approaches. For the EBLUP, the squared prediction error at x_0 is given by $\hat{\sigma}_n^2(x_0)\,\rho_n^2(x_0)$, see (19) and (23). For BLK with $\theta_{1,\ell} > 0$, observations that are far away from x_0 have negligible influence on the prediction of $Z(x_0)$. We thus construct an "equivalent number of observations" $n'(x_0)$, given by $n'(x_0) = \sum_{i=1}^n k_{1,\ell}(x_0 - x_i)$. This value is substituted for n in (18) and (19)

for the evaluation of the posterior squared prediction error (13) (but not for the evaluation of the marginal likelihood $\mathscr{L}(\mathbf{z}_n|\ell,t)$). We then compute the empirical means (over the 1,000 simulations and 441 grid points) of these posterior squared prediction errors, which we denote by $\widetilde{IE^2}[\hat{\eta}_{MV,n}]$ and $\widetilde{IE^2}[\hat{\eta}_{BLK,n}]$, to be compared respectively with $\widetilde{IE^2}[\hat{\eta}_{MV,n}]$ and $\widetilde{IE^2}[\hat{\eta}_{BLK,n}]$.

It is well known that using the plug-in mean-squared prediction error of the BLUP underestimates the true prediction error, the reason being that the uncertainty due to the estimation of the covariance parameters is not accounted for, see Stein (1999, Section 6.8). The table corroborates this result. On the other hand, the error predicted by BLK slightly overestimates (columns A and B,C) or slightly underestimates (column D) the true empirical error, with an exception for column E where the abrupt change in the mean β yields large and hardly predictable errors (see the values of $\widetilde{MaxE^2}[\hat{\eta}_{BLK,n}]$).

Example 2: expected improvement and deceptive function. Kriging prediction can be used for the global optimization of a function $f(\cdot) : x \in \mathscr{X} \subset \mathbb{R}^d \longmapsto f(x) \in \mathbb{R}$, see Mockus et al. (1978); Mockus (1989), a method which has been popularized under the name of Expected Improvement (EI), see Jones et al. (1998). The function is considered as the realization of a GP, whose characteristics (in particular the parameters θ of the chosen covariance function) are estimated from observations that correspond to evaluations of the function, generally by ML. However, when the estimated parameters are plugged into the Kriging predictor and its associated mean-squared prediction error, the performance of the method may be rather disappointing since evaluation results may not contain enough information to estimate the covariance parameters in a satisfactory manner. This may wrongly provide the sensation that the function is extremely flat in some areas, that will thus not be explored, or on the opposite extremely wiggly, so that all $\mathscr{X}$ would seem to deserve a close exploration. This phenomenon is well described in (Benassi et al., 2011) through the concept of deceptive function, an example of which is presented below.

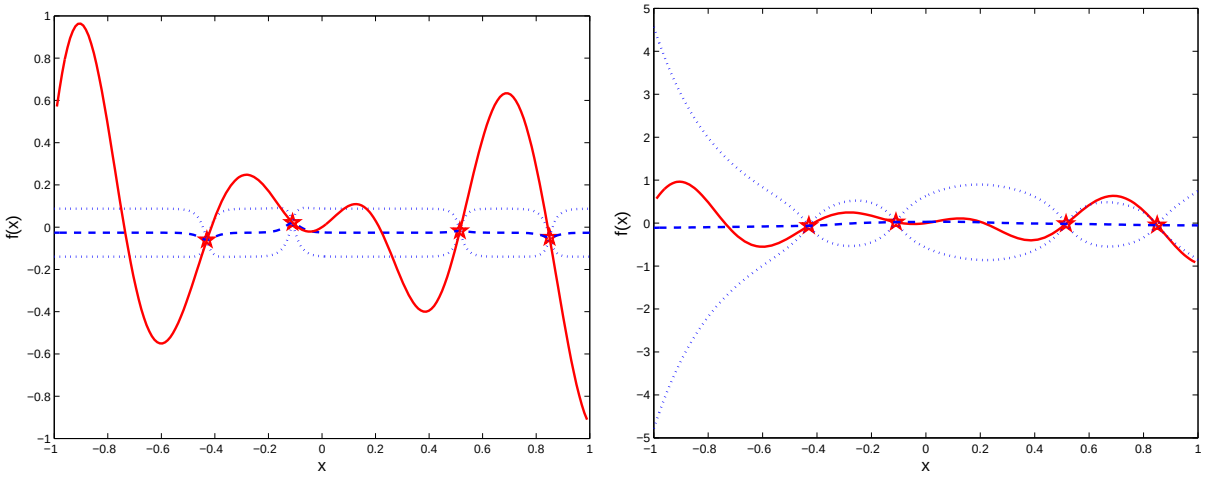

Figure 4: Left: a deceptive function (solid line), with the EBLUP (dashed line) and 95% credible intervals (dotted lines). Right: the same function with BLK (dashed line) and 95% credible intervals (dotted lines).

Consider the function in (Benassi et al., 2011, Section 5.1), $f(x) = x\,[\sin(10x+1) + 0.1\sin(15x)]$, $x \in \mathscr{X} = [-1,1]$, plotted in solid line in Figure 4. When the function is evaluated at the design $\xi_n = (-0.43,\ -0.11,\ 0.515,\ 0.85)$ all observations are nearly zero, see the stars in Figure 4. We assume that $f(\cdot)$ can be represented by a GP with unknown mean β and Matérn 3/2 covariance function, see (15). The EBLUP (blue solid line) and associated (approx-

imate) 95% credible intervals (dashed lines), corresponding to $\hat{\eta}_{MV,n}(x_0) \pm 3.182\,\hat{\sigma}_n(x_0)\rho_n(x_0)$, with 3.182 the critical value of the student t-distribution with $n-p=3$ degrees of freedom, see Santner et al. (2003, p. 95), are plotted in Figure 4-left. The most promising region in terms of maximization of $f(\cdot)$ is around zero and several additional evaluations of $f(\cdot)$ are required before the method starts exploring the neighborhood of the maximizer of $f(\cdot)$, at $x^* \simeq -0.9052$, see Benassi et al. (2011).

Figure 4-right presents the BLK predictor and approximate 95% credible intervals given by $\hat{\eta}_{BLK,n}(x_0) \pm 2.571\,[PSPE(\hat{\eta}_{BLK,n}(x_0))]^{1/2}$ (with 2.571 the critical value of the student t-distribution with $n-p+\nu_0 = 5$ degrees of freedom), under the same setting as in Example 1 and $\theta_{0,\ell} = 1, 5, 10, 20$, $\theta_{1,\ell} = 5$ for $\ell = 1, \ldots, L = 4$. The large uncertainty on the behavior of $f(\cdot)$ in the left-hand side of the domain is an incitation to put observations there. The EI algorithm can be expected to perform much more efficiently when based on BLK, even with a small L, than when based on the EBLUP.

Example 3: behaviour of BLK far from design points. This example illustrates the behavior of the BLK predictor $\hat{\eta}_{BLK,n}(t)$ when t is far enough from design points so that the influence of the closest point x_i dominates all others through $k_{1,\ell}(t-x_i)$, see (1).

We consider a process $Z(x)$ with zero mean and covariance $K_{3/2,2}(\cdot)$, see (15), which is observed at the 7 design points $(-1,\ -0.9,\ -0.8,\ -0.5,\ -0.2,\ 0,\ 1)$, indicated by stars. Figure 5 shows the predictions obtained by ordinary Kriging (with covariance $K_{3/2,15}(\cdot)$) and BLK (with $L=4$, $k_{0,\ell}(\delta) = K_{3/2,\theta_{0,\ell}}(|\delta|)$ and $\theta_{0,\ell} = 1,\ 5,\ 10,\ 20$, $k_{1,\ell}(\delta) = K_{3/2,5}(|\delta|)$ for all ℓ), both assuming that $\mathbf{g}(x) \equiv 1$. The prediction by limit-Kriging (Joseph, 2006), with covariance $K_{3/2,15}(\cdot)$, is also presented on the figure.

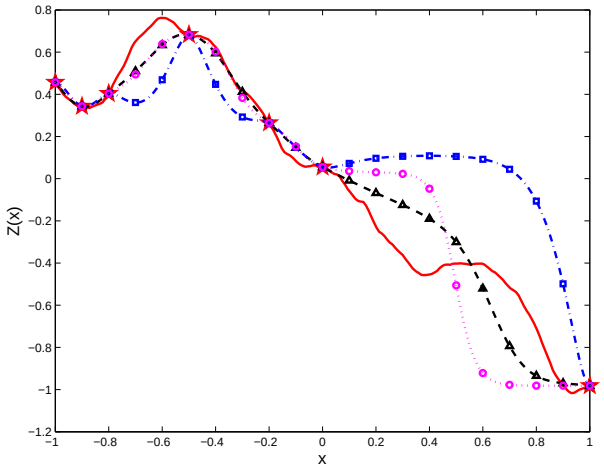

Figure 5: Predictions for BLK (dashed line with triangles), ordinary Kriging (dash-dotted line with squares) and limit Kriging (dotted line with circles) with underestimated correlation length; $Z(x)$ is in solid line, observations are indicated by stars.

The correlation length θ being underestimated (1/15 instead of 1/2), the ordinary-Kriging prediction $\hat{\eta}_n(t)$ is close to the estimated mean $\hat{\beta}_n$ given by (20) for t far enough from design points, see the right part of the figure. On the other hand, in the same region the BLK predictor is mainly influenced by the closest design point x_i, so that $\hat{\eta}_n(t)$ is close to $Z(x_i)$, a reasonable behavior that resembles that of limit-Kriging with underestimated θ, see Joseph (2006, Section 3).

Example 4: prediction of an oceanographic field, using outputs of a formal (numerical) model. The data used in this study were made available through a collaboration with

the institute MUMM, a department of the Royal Belgian Institute of Natural Sciences. It consists of snapshots of the output of the biogeochemical oceanographic model MIRO&CO (Lacroix et al., 2007), run to simulate the evolution of inorganic and organic carbon and nutrients, phytoplankton, bacteria and zooplankton with realistic forcing conditions. The model covers the entire water column of the Southern Bight of the North Sea. In the study presented here we concentrate on a surface grid $\mathscr{G}$ of 49×21 points, 879 of which corresponding to sea surface and form our design space $\mathscr{X}$. The objective is to assess the possibility of predicting chlorophyll concentration over $\mathscr{X}$ from observations at a small number of sites. The simulation model is used as a substitute for the real phenomenon; it provides pseudo-observations at n design points $\xi_n = \{x_1, \ldots, x_n\}$ in $\mathscr{X}$ ($n = 25$) and allows the computation of empirical prediction squared errors.

Experimental design. Figure 6-Left presents the model response $f(x)$, $x \in \mathscr{X}$. It is manifest that variability is stronger along the French coast, so that obtaining precise predictions there would require a denser concentration of observation sites than in other areas where the response is more flat. However, in a realistic situation the true response is not available, and this information cannot be used to choose ξ_n.

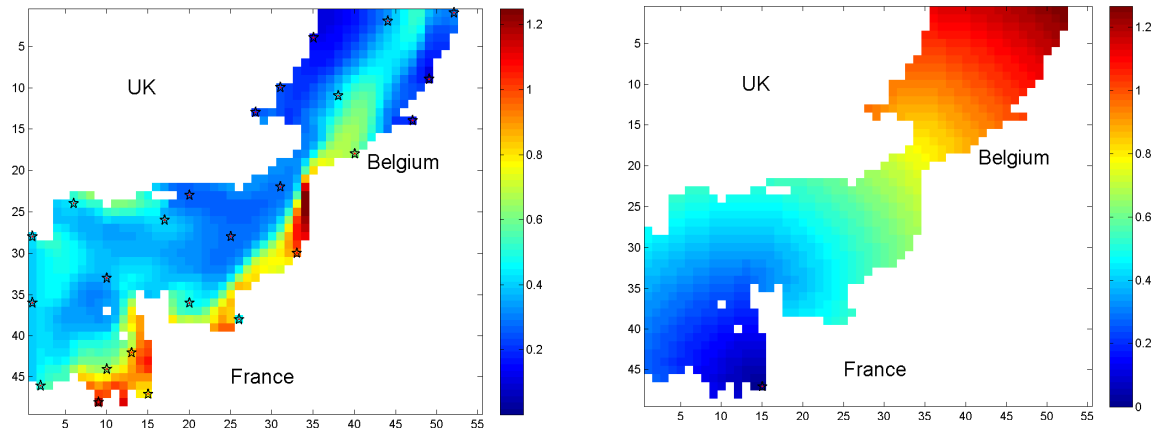

Figure 6: Left: model response over $\mathscr{X}$ and design points (stars). Right: shortest maritime distance $\Delta(x, x_{21})$ between $x \in \mathscr{X}$ and the 21st design point x_{21}.

The design ξ_n is thus constructed as follows. First, $\mathscr{G}$ is renormalized to $[0,1]^2$ and $\mathscr{X}$ is renormalized accordingly. Then, we generate a low-discrepancy (Sobol') sequence in $\mathscr{G}$; the design points $x_1, \ldots, x_{10}$ are given by the first 10 points of the sequence that fall in $\mathscr{X}$. Finally, the next 15 points are generated sequentially, according to

$$x_{k+1} = \arg\max_{x \in \mathscr{X}} \rho_k(x)\,, \tag{17}$$

with $\rho_k(x)$ the universal Kriging variance (the mean-squared error of the BLUP) for a process with unknown constant mean β and covariance function $K_{3/2,\theta}(\Delta(x, x'))$, see (15), and observations $f(x_1), \ldots, f(x_k)$. We take $\theta = 10$ to ensure that ξ_n will be well dispersed over $\mathscr{X}$. In order to take the non-convexity of the design region into account, the "distance" $\Delta(x, x')$ corresponds to the shortest maritime route between x and x', computed by Dijkstra's algorithm (Dijkstra, 1959). The 25 design points x_i of ξ_n are indicated by stars in Figure 6-Left. Figure 6-Right presents the distance $\Delta(x, x_{21})$ from $x \in \mathscr{X}$ to the 21st design point x_{21} (with indices $(15, 47)$, corresponding to coordinates $(0.2593, 0.9583)$ in the renormalized space), showing a neat difference with respect to the Euclidean distance $\|x - x_{21}\|$. Notice that in order to compute

predictions and prediction errors over $\mathscr{X}$ we only need to compute distances $\Delta(x_i, x)$ from the design points x_i to the $x \in \mathscr{X}$ (and not all pairwise distances $\Delta(x, x')$, $(x, x') \in \mathscr{X}^2$).

Comparison between BLUP and BLK For this same design ξ_n, we compare the predictive performance of the EBLUP and BLK. The EBLUP uses the covariance function $K_{3/2,\hat{\theta}^n}(\Delta(x, x'))$, with $\hat{\theta}^n \simeq 19.2058$ estimated by ML from the observations $f(x_1), \ldots, f(x_n)$; BLK uses $K_{3/2,\theta}(\Delta(x, x'))$ both for $k_{0,\ell}$ and $k_{1,\ell}$, with $L = 4$ and $\theta_{0,\ell} = 5, 10, 20, 30$ in $k_{0,\ell}$, $\theta_{1,\ell} = \theta_* > 0$ for all $\ell = 1, \ldots, 4$ in $k_{1,\ell}$. The choice of θ_* should make a compromise between stepping the influence of distant points down (which means taking θ_* large enough to obtain local predictions) and maintaining a reasonable correlation with sufficient design points (which means taking θ_* not too large). Let $\phi_{mM}(\xi_n) = \max_{x \in \mathscr{X}} \min_{x_i \in \xi_n} \|x - x_i\|$ denote the value of the minimax distance criterion for ξ_n, and denote $\phi_{kNN}(\xi_n) = \max_{x_i \in \xi_n} \|x_i - x_{j^*(i)}\|$ with $x_{j^*(i)}$ the k-th nearest-neighbor of x_i in ξ_n. For all $x \in \mathscr{X}$, we can then guarantee that

$$\|x - x_i\| \leq \delta_k(\xi_n) = \phi_{mM}(\xi_n) + \phi_{kNN}(\xi_n)$$

for $k+1$ points x_i of ξ_n. We take $k = 19$ and $\theta_* = 2.99431/\delta_k(\xi_n)$, which ensures that, for each $x \in \mathscr{X}$, $k_{1,\ell}(\|x - x_i\|) \geq 20\%$ for at least 20 design points in ξ_n. For the design ξ_n plotted in Figure 6-Left, this gives $\theta_* \simeq 2.4202$, the value we use here for BLK.

Figure 7-Left shows the BLK predictions over $\mathscr{X}$, to be compared with the true responses on Figure 6-Left. Figure 7-Right shows the squared prediction errors $E^2[\hat{\eta}_{BLK,n}(x)]$, where for a predictor $\hat{z}_n(\cdot)$ we denote $E^2[\hat{z}_n(x)] = [\hat{z}_n(x) - f(x)]^2$. Taking into account that only 25 designs points have been used, predictions are fairly accurate, excepted at some areas along the French coast where the correlation structure strongly departs from the smoother variation in the open sea region.

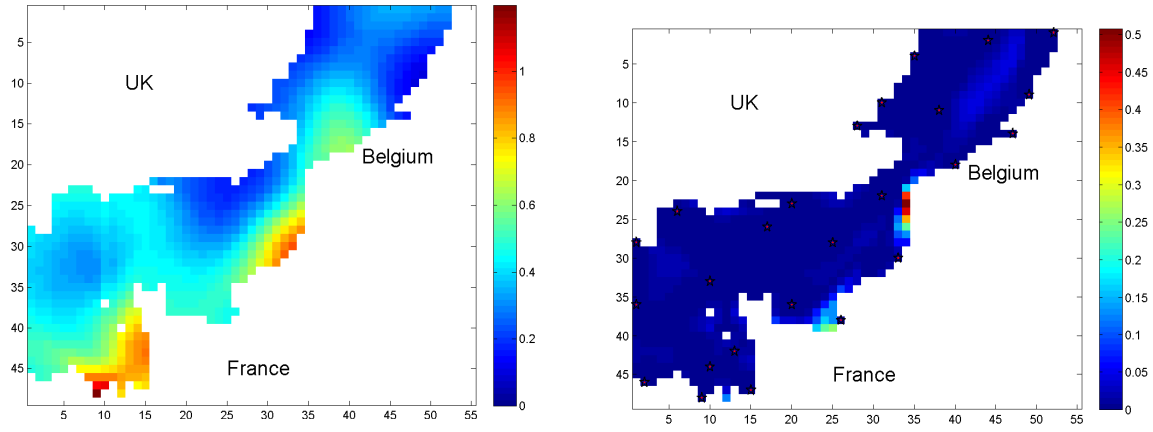

Figure 7: Left: BLK predictions $\hat{\eta}_{BLK,n}(x)$ over $\mathscr{X}$. Right: squared errors $E^2[\hat{\eta}_{BLK,n}(x)]$ for BLK.

Figure 8 presents the difference in squared prediction errors between the EBLUP and BLK, $E^2[\hat{\eta}_{MV,n}(x)] - E^2[\hat{\eta}_{BLK,n}(x)]$ for $x \in \mathscr{X}$. The accuracies of the BLK predictor and the BLUP are similar in most of the domain $\mathscr{X}$, but BLK is more accurate in a large portion of the French coast, precisely where good predictions are difficult to obtain. The mean, median and maximum values of $E^2[\hat{\eta}_{MV,n}(x)]$ over $\mathscr{X}$ are respectively 0.0144, 0.0035 and 0.6447; these values equal 0.0122, 0.0035 and 0.5070 for BLK; a paired-comparisons test for these squared errors gives approximately 5.52, with an associated p-value $\simeq 1.7\ 10^{-8}$ indicating that BLK is significantly more accurate than the EBLUP. Limit-Kriging (Joseph, 2006) with θ estimated by ML yields mean, median and maximum values of squared prediction errors respectively equal to 0.0140,

0.0035, 0.6527, and thus performs similarly to the EBLUP. The behavior of the EBLUP is marginally improved when the ML estimation of θ in $K_{3/2,\theta}(\Delta(x,x'))$ is replaced by leave-one-out cross validation, see Dubrule (1983): the estimated θ is then $\hat{\theta}^n \simeq 17.64$, the mean, median and maximum values of squared prediction errors over $\mathscr{X}$ become respectively 0.0142, 0.0036 and 0.6312; the paired-comparisons test with BLK gives $\simeq 5.82$, with a p-value $\simeq 3.0\ 10^{-9}$.

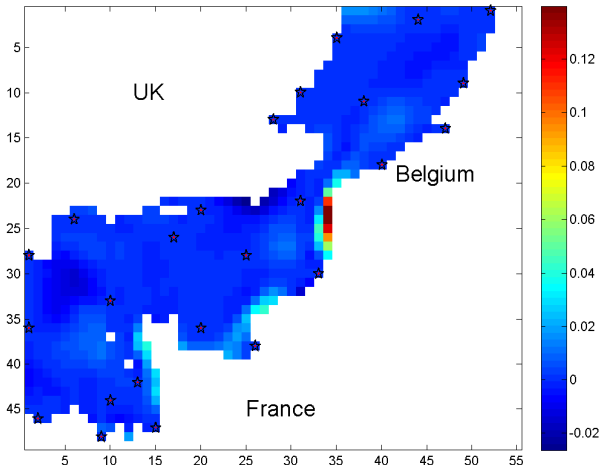

Figure 8: Difference in squared errors between the EBLUP and BLK, $E^2[\hat{\eta}_{MV,n}(x)] - E^2[\hat{\eta}_{BLK,n}(x)]$.

Choice of θ_*. The influence of the choice of θ_* on the performance of BLK (with $L = 4$ and $\theta_{0,\ell} = 5, 10, 10, 30$ in $k_{0,\ell}$) is illustrated in Figure 9, where the solid line (respectively dashed line) gives the mean (respectively $1/40\times$ the maximum) value of $E^2[\hat{\eta}_{BLK,n}(x)]$ over $\mathscr{X}$ as a function of θ_* varying between 0 (stationary model) and 10 (strong non-stationarity). The choice $\theta_* = 2.99431/\delta_{19}(\xi_n) \simeq 2.4202$ is not optimal but seems reasonable. Imposing that each $x \in \mathscr{X}$ has only at least 10 neighboring design points x_i such that $k_{1,\ell}(\|x - x_i\|) \geq 20\%$ would have been a better choice, since $2.99431/\delta_9(\xi_n) \simeq 4.2531$, closer to the optimum θ_* in Figure 9. Note that the errors $E^2[\hat{\eta}_{BLK,n}(x)]$ are normally not available, and therefore cannot be used to select θ_*. This indicates, however, that the choice of k in the construction $\theta_* = 2.99431/\delta_k(\xi_n)$ is not critical.

Influence of L. Increasing L does not necessarily improve the performance of BLK. For instance, taking $\theta_{0,\ell} = 5, 6, 7, \ldots, 30$ in $k_{0,\ell}$ ($L = 26$) and $\theta_{1,\ell} = 2.4202$ for all $\ell = 1, \ldots, 26$ in $k_{1,\ell}$, we obtain 0.0127, 0.0035, 0.5388, respectively for the mean, median and maximum values of $E^2[\hat{\eta}_{BLK,n}(x)]$ over $\mathscr{X}$. Figure 10-Left shows the posterior weights $w_n^\ell(x)$ associated with the correlations lengths $1/\theta_{1,\ell}$, $\ell = 1, \ldots, L = 4$, when $\theta_{0,\ell} = 5, 10, 20, 30$, pointing out areas where the response exhibits strong variability and those where it is fairly smooth. It should be stressed that the calculation of the $w_n^\ell(x)$ only uses the 25 response values $f(x_i)$ for $x_i \in \xi_n$. The similarity between the maps of $w_n^3(x)$ and $w_n^4(x)$ is an incitation to try reducing L to 3: when $\theta_{0,\ell} = 5, 10, 30$ in $k_{0,\ell}$ ($L = 3$), we obtain 0.0122, 0.0035 and 0.5040 for the mean, median and maximum values of $E^2[\hat{\eta}_{BLK,n}(x)]$; i.e., values that are marginally better than those with $L = 4$.

Influence of the distance function. Finally, the interest of using the distance $\Delta(x,x')$ corresponding to the shortest maritime route between x and x' instead of the Euclidean distance

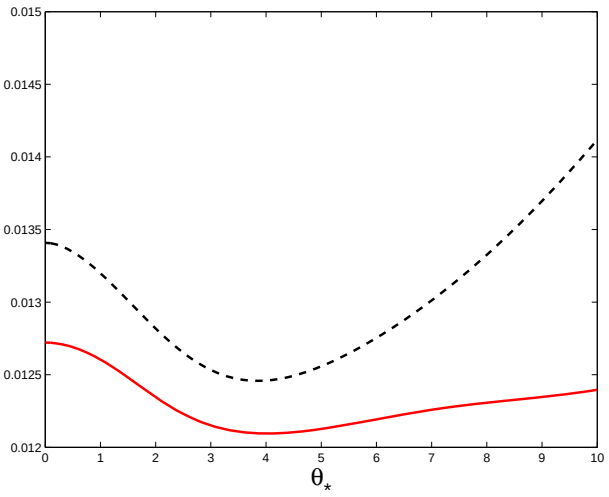

Figure 9: Mean value (solid line) and $1/40 \times$ maximum value (dashed line) of $E^2[\hat{\eta}_{BLK,n}(x)]$ over $\mathscr{X}$ as functions of θ_*.

$\|x - x'\|$ (in the renormalized space $[0,1]^2$) is illustrated for BLK in Figure 10-Right, which shows the differences between corresponding prediction squared errors. The distance $\Delta(x, x')$ yields more accurate predictions along the major part of the French coast; the mean, median and maximum values of $E^2[\hat{\eta}_{BLK,n}(x)]$ over $\mathscr{X}$ are respectively 0.0129, 0.0030 and 0.5257 when using Euclidean distance to compute correlations. The decrease of performance compared to the situation where $\Delta(x, x')$ is used is not caused to the fact that the sequential construction of design points x_{11} to x_{25} through (17) is based on $\Delta(x, x')$: replacing the design ξ_n by ξ'_n where x_{11} to x_{25} are constructed via (17) with Euclidean distance based covariances, we obtain mean, median and maximum squared prediction errors equal to 0.0142, 0.0034, 0.5175 when BLK uses Euclidean distance.

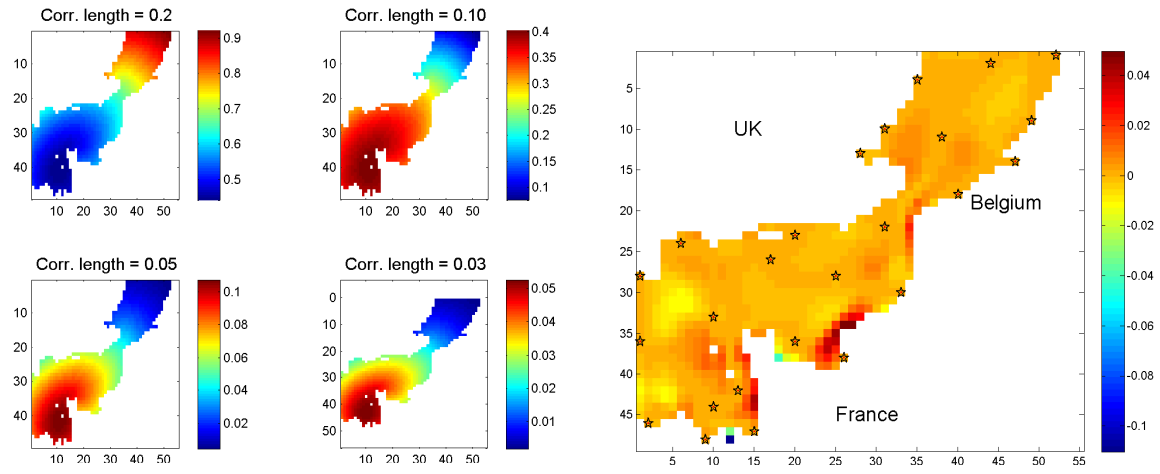

Figure 10: Left: posterior weights $w_n^\ell(x)$ for correlations lengths $1/\theta_{1,\ell}$. Right: differences between squared errors for BLK with Euclidean distance $\|x - x'\|$ and BLK with $\Delta(x, x')$.

4 Conclusions and further developments

We have presented a Bayesian local Kriging approach for the interpolation or prediction of random fields. The method uses localized covariance functions which allow us to account for non-stationarity. A finite set of L candidate covariance functions is used for each prediction point, that receive equal prior weights. Using a common hierarchical prior for the trend and variance of the process, posterior weights can easily be calculated to obtain posterior means and prediction squared errors. Numerical simulations indicate that, on average, the method performs slightly worse than universal Kriging with plug-in maximum-likelihood estimates for the covariance parameters when the true characteristics of the process satisfy the assumptions (stationarity, correct parametric trend, correct covariance functions), but performs significantly better when these assumptions are violated, even if the number L of concurrent covariance functions is very small ($L = 4$ in the examples considered). Also, BLK, which does not use any numerical optimization, is much faster and numerically stable than universal Kriging with plug-in estimates, which requires the estimation of covariance parameters.

To summarize, our feeling is that it seems illusory in many applications to try estimate covariance parameters from a few observations only, especially with a covariance structure not necessarily well-adapted to the variability of the modeled phenomenon. Using a small number of candidate processes able to reproduce a reasonable range of possible behaviors may be preferable: it is simpler to implement, numerically more stable, and seems to often yield better predictions.

Although these results are encouraging, further numerical experimentations (in particular for higher dimensional processes and different types and sizes of experimental designs) are needed to confirm these preliminary observations. We have restricted our attention to the situation where the regressor $\mathbf{g}(\cdot)$ in the parametric trend $\mathbf{g}^\top(x)\beta$ was identical for all L models (and, moreover, we only considered the case $\mathbf{g}(x) \equiv 1$ in all examples). The same approach could be used when different trends $\mathbf{g}_\ell(\cdot)$ are associated with different covariances $C_{\ell|t}(\cdot,\cdot)$, possibly allowing for a better consideration of uncertainty in the process trend.

The choice of the particular form of the kernels $k_{0,\ell}(\cdot)$ and $k_{1,\ell}(\cdot)$ seems to be less crucial than that of the correlation length for $k_{1,\ell}(\cdot)$, which depends on the assumed amount of non-stationarity. Using different correlations lengths for some of the L concurrent covariances is a possible option to investigate. Another one is to simply ensure that $k_{1,\ell}(x - x_i)$ be large enough for all points in x in $\mathscr{X}$ and enough design points x_i, with the motivation that the more dense the design ξ_n in $\mathscr{X}$, the more local the models can be and the stronger the non-stationarity that BLK can take into account. Two proposals have been made in this direction, see Examples 1 and 4. Further investigations are required to validate them, in particular concerning their asymptotic behavior, when the number n of design points tends to infinity and ξ_n is space-filling.

Finally, as mentioned in Section 2.3, designing experiments adapted to BLK requires the evaluation (approximation) of the second term in (14). This rather challenging problem is under current investigation.

5 Appendix: Universal BLK in presence of a parametric trend

Here we consider that, seen from t, $f(\cdot)$ is the sample path of a stochastic process $Z(\cdot)$ such that $Z(\cdot)|s(t), \beta, \sigma^2, t \sim \mathsf{GP}(\mathbf{g}^\top(\cdot)\beta, \sigma^2 C_{s(t)|t}(\cdot,\cdot))$, with $\mathrm{Prob}\{s(t) = \ell\} = w_0^\ell$, $\ell = 1, \ldots, L$, $\beta|\sigma^2 \sim \psi_0(\beta|\sigma^2)$, $\sigma^2 \sim \varphi_0(\sigma^2)$, see (8), and $\mathbf{g}(\cdot)$ a known p-dimensional vector of functions defined on $\mathscr{X}$. The likelihood of $\mathbf{z}_n = (f(x_1), \ldots, f(x_n))^\top$ for the model $\mathsf{GP}(\mathbf{g}^\top(\cdot)\beta, \sigma^2 C_{s(t)|t}(\cdot,\cdot))$ is

$$\mathscr{L}(\mathbf{z}_n|\beta, \sigma^2, \ell, t) = \frac{1}{\sigma^n (2\pi)^{n/2} \det{}^{1/2} \mathbf{K}_n(\ell|t)} \exp\left[-\frac{1}{2\sigma^2} (\mathbf{z}_n - \mathbf{G}_n\beta)^\top \mathbf{K}_n^{-1}(\ell|t)(\mathbf{z}_n - \mathbf{G}_n\beta)\right],$$

where the i-th row of the $n \times p$ matrix $\mathbf{G}_n$ equals $\mathbf{g}^\top(x_i)$, $i = 1, \ldots, n$. The weights w_0^l are updated according to (3), with now $\mathscr{L}(\mathbf{z}_n|\ell, t) = \int_{\mathbb{R}^p} \int_0^\infty \mathscr{L}(\mathbf{z}_n|\beta, \sigma^2, \ell, t)\, \psi_0(\beta|\sigma^2)\, \varphi_0(\sigma^2)\, \mathrm{d}\beta\, \mathrm{d}\sigma^2$.

5.1 Uniform prior for β

Suppose first that β has a uniform (improper) prior on $\mathbb{R}^p$. We obtain

$$\mathscr{L}(\mathbf{z}_n|\ell, t) = \frac{1}{(2\pi)^{(n-p)/2} \det^{1/2} \mathbf{K}_n(\ell|t)\ \det^{1/2}(\mathbf{G}_n^\top \mathbf{K}_n^{-1} \mathbf{G}_n)} \frac{(\sigma_0^2 \nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)} \frac{\Gamma(\nu_n/2)}{(\sigma_{n|\ell,t}^2 \nu_n/2)^{\nu_n/2}},$$

with $\nu_n = \nu_0 + n - p$ and

$$\sigma_{n|\ell,t}^2 = \frac{\nu_0 \sigma_0^2 + (n-p)\hat{\sigma}_n^2(\ell|t)}{\nu_0 + n - p}, \tag{18}$$

where

$$\hat{\sigma}_n^2(\ell|t) = \frac{1}{n-p} (\mathbf{z}_n - \mathbf{G}_n \hat{\beta}_n(\ell|t))^\top \mathbf{K}_n^{-1}(\ell|t)(\mathbf{z}_n - \mathbf{G}_n \hat{\beta}_n(\ell|t)) \tag{19}$$

is the Restricted Maximum-Likelihood (REML) estimator of σ^2, and

$$\hat{\beta}_n(\ell|t) = (\mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t) \mathbf{G}_n)^{-1} \mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t) \mathbf{z}_n \tag{20}$$

is the ML estimator of β, given $\mathbf{z}_n, \ell$ and t, see, e.g., Santner et al. (2003, p. 67, 95). Moreover, given $\mathbf{z}_n, \ell$ and t, σ^2 has the inverse chi-square distribution $\varphi_{n|\ell,t}(\cdot) = \varphi_{\sigma_{n|\ell,t}^2, \nu_n}(\cdot)$, see (8).

$Z(x_0)|\ell, \mathbf{z}_n, t$ has a non-central t-distribution, see Santner et al. (2003, p. 95), with

$$\hat{\eta}_n(x_0|\ell, t) = \mathsf{E}\{Z(x_0)|\ell, \mathbf{z}_n, t\} = \mathbf{g}^\top(x_0)\hat{\beta}_n(\ell|t) + \mathbf{k}_n^\top(x_0, \ell|t)\mathbf{K}_n^{-1}(\ell|t)(\mathbf{z}_n - \mathbf{G}_n\hat{\beta}_n(\ell|t)) \tag{21}$$

and $\hat{\beta}_n(\ell|t)$ given by (20). Therefore, $\hat{\eta}_n(x_0|t) = \mathbf{c}_n^\top(x_0|\ell, t)\mathbf{z}_n$ and $\hat{\eta}_n(x_0|t)$ is still given by (11), but with

$$\begin{aligned} \mathbf{c}_n(x_0|\ell, t) &= \mathbf{K}_n^{-1}(\ell|t)\mathbf{G}_n(\mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t)\mathbf{G}_n)^{-1}\mathbf{g}(x_0) \\ &\quad + \left[\mathbf{I}_n - \mathbf{K}_n^{-1}(\ell|t)\mathbf{G}_n(\mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t)\mathbf{G}_n)^{-1}\mathbf{G}_n^\top\right] \mathbf{K}_n^{-1}(\ell|t)\mathbf{k}_n(x_0, \ell|t), \end{aligned} \tag{22}$$

with $\mathbf{I}_n$ the n-dimensional identity matrix. Also,

$$\mathsf{var}\{Z(x_0)|\ell, \mathbf{z}_n, t\} = \frac{n - p + \nu_0}{n - p + \nu_0 - 2} \sigma_{n|\ell,t}^2 \rho_n^2(x_0|\ell, t)$$

see (18), with $\rho_n^2(x_0|\ell, t)$ the universal-Kriging variance

$$\rho_n^2(x_0|\ell, t) = C_\ell(x_0, x_0|t) - [\mathbf{g}^\top(x_0)\ \mathbf{k}_n^\top(x_0, \ell|t)] \begin{bmatrix} \mathbf{O} & \mathbf{G}_n^\top \\ \mathbf{G}_n & \mathbf{K}_n(\ell|t) \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{g}(x_0) \\ \mathbf{k}_n(x_0, \ell|t) \end{bmatrix},$$

or equivalently,

$$\begin{aligned} \rho_n^2(x_0|\ell, t) &= C_\ell(x_0, x_0|t) - \mathbf{k}_n^\top(x_0, \ell|t)\mathbf{K}_n^{-1}(\ell|t)\mathbf{k}_n(x_0, \ell|t) \\ &+ [\mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t)\mathbf{k}_n(x_0, \ell|t) - \mathbf{g}(x_0)]^\top (\mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t)\mathbf{G}_n)^{-1} [\mathbf{G}_n^\top \mathbf{K}_n^{-1}(\ell|t)\mathbf{k}_n(x_0, \ell|t) - \mathbf{g}(x_0)]. \end{aligned} \tag{23}$$

Therefore, the posterior squared prediction error is still given by (13), but with $\nu_n = \nu_0 + n - p$, $\sigma_{n|\ell,t}^2$ given by (18), $\rho_n^2(x_0|\ell, t)$ by (23) and $\mathbf{c}_n(x_0|\ell, t)$ by (22) in $\Omega_n(x_0|t)$. Similarly to Sect. 2.3, the preposterior variance at t is (for $\nu_0 > 2$) is given by (14).

5.2 Normal prior for β

Suppose now that $\psi_0(\beta|\sigma^2)$ is the density of the p-dimensional normal distribution with mean β_0 and variance-covariance matrix $\sigma^2\mathbf{V}_0$. Since $\mathbf{z}_n|\ell,\sigma^2,t \sim \mathscr{N}(\mathbf{G}_n\beta_0, \mathbf{K}_n(\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{G}_n^\top)$, we have

$$\begin{aligned}\mathscr{L}(\mathbf{z}_n|\sigma^2,\ell,t) &= \frac{1}{(2\pi)^{n/2}\sigma^n \det^{1/2}[\mathbf{K}_n(\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{G}_n^\top]} \\ &\quad \times \exp\left\{-\frac{1}{2\sigma^2}(\mathbf{z}_n-\mathbf{G}_n\beta_0)^\top[\mathbf{K}_n(\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{G}_n^\top]^{-1}(\mathbf{z}_n-\mathbf{G}_n\beta_0)\right\},\end{aligned}$$

so that, similarly to Sect. 2.3,

$$\mathscr{L}(\mathbf{z}_n|\ell,t) = \frac{1}{(2\pi)^{n/2}\det^{1/2}[\mathbf{K}_n(\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{G}_n^\top]}\,\frac{(\sigma_0^2\nu_0/2)^{\nu_0/2}}{\Gamma(\nu_0/2)}\,\frac{\Gamma(\nu_n/2)}{(\sigma^2_{n|\ell,t}\nu_n/2)^{\nu_n/2}},$$

where $\nu_n = \nu_0 + n$ and

$$\sigma^2_{n|\ell,t} = \frac{\nu_0\sigma_0^2 + n\hat\sigma_n^2(\ell|t)}{\nu_0+n} \tag{24}$$

with $\hat\sigma_n^2(\ell|t) = (1/n)\,(\mathbf{z}_n-\mathbf{G}_n\beta_0)^\top[\mathbf{K}_n(\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{G}_n^\top]^{-1}(\mathbf{z}_n-\mathbf{G}_n\beta_0)$. $Z(x_0)|\ell,\mathbf{z}_n,t$ has again a non-central t-distribution, with

$$\hat\eta_n(x_0|\ell,t) = \mathsf{E}\{Z(x_0)|\ell,\mathbf{z}_n,t\} = \mathbf{g}^\top(x_0)\hat\beta_n(\ell|t) + \mathbf{k}_n^\top(x_0,\ell|t)\mathbf{K}_n^{-1}(\ell|t)(\mathbf{z}_n-\mathbf{G}_n\hat\beta_n(\ell|t))$$

and

$$\hat\beta_n(\ell|t) = (\mathbf{G}_n^\top\mathbf{K}_n^{-1}(\ell|t)\mathbf{G}_n+\mathbf{V}_0^{-1})^{-1}[\mathbf{G}_n^\top\mathbf{K}_n^{-1}(\ell|t)\mathbf{z}_n+\mathbf{V}_0^{-1}\beta_0]\,.$$

The predictor $\hat\eta_n(x_0|t)$ is again $\sum_{\ell=1}^L w_n^\ell\hat\eta_n(x_0|\ell,t)$. Also,

$$\mathsf{var}\{Z(x_0)|\ell,\mathbf{z}_n,t\} = \nu_n/(\nu_n-2)\,\sigma^2_{n|\ell,t}\,\rho_n^2(x_0|\ell,t)$$

see (24), with now

$$\rho_n^2(x_0|\ell,t) = C_\ell(x_0,x_0|t) - [\mathbf{g}^\top(x_0)\ \mathbf{k}_n^\top(x_0,\ell|t)]\begin{bmatrix}-\mathbf{V}_0^{-1} & \mathbf{G}_n^\top\\ \mathbf{G}_n & \mathbf{K}_n\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{g}(x_0)\\ \mathbf{k}_n(x_0,\ell|t)\end{bmatrix},$$

or equivalently,

$$\begin{aligned}\rho_n^2(x_0|\ell,t) &= C_\ell(x_0,x_0|t) + \mathbf{g}^\top(x_0)\mathbf{V}_0\mathbf{g}(x_0)\\ &\quad -[\mathbf{k}_n(x_0,\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{g}(x_0)]^\top(\mathbf{K}_n(\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{G}_n^\top)^{-1}[\mathbf{k}_n(x_0,\ell|t)+\mathbf{G}_n\mathbf{V}_0\mathbf{g}(x_0)]\,.\end{aligned}$$

References

Abt, M. (1999). Estimating the prediction mean squared error in gaussian stochastic processes with exponential correlation structure. *Scandinavian Journal of Statistics*, 26(4):563–578.

Benassi, R., Bect, J., and Vazquez, E. (2011). Robust Gaussian process-based global optimization using a fully Bayesian expected improvement criterion. In Coello, C. C., editor, *Learning and Intelligent Optimization*, pages 176–190, Berlin. Springer-Verlag, LNCS 6683.

Cramér, H. and Leadbetter, M. (1967). *Stationary and Related Stochastic Processes*. Wiley, New York.

Dijkstra, E. (1959). A note on two problems in connexion with graphs. *Numerische mathematik*, 1(1):269–271.

Dubrule, O. (1983). Cross validation of kriging in a unique neighborhood. *Journal of the International Association for Mathematical Geology*, 15(6):687–699.

Ginsbourger, D., Helbert, C., and Carraro, L. (2008). Discrete mixture of kernels for kriging-based optimization. *Quality and Reliability Engineering International*, 24:681–691.

Harville, D. and Jeske, D. (1992). Mean squared error of estimation or prediction under a general linear model. *Journal of the American Statistical Association*, 87(419):724–731.

Jones, D. (2001). A taxonomy of global optimization methods based on response surfaces. *Journal of global optimization*, 21(4):345–383.

Jones, D., Schonlau, M., and Welch, W. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.

Joseph, V. (2006). Limit kriging. *Technometrics*, 48(4):458–466.

Kanji, G. (1993). *100 Statistical Tests*. Sage Pub., London.

Lacroix, G., Ruddick, K., Park, Y., Gypens, N., and Lancelot, C. (2007). Validation of the 3D biogeochemical model MIRO&CO with field nutrient and phytoplankton data and MERIS-derived surface chlorophyll α images. *Journal of Marine Systems*, 64:66–88.

Lam, K., Wang, Q., and Li, H. (2004). A novel meshless approach — Local kriging (LoKriging) method with two-dimensional structural analysis. *Computational Mechanics*, 33:235–244.

Mockus, J. (1989). *Bayesian Approach to Global Optimization, Theory and Applications*. Kluwer, Dordrecht.

Mockus, J., Tiesis, V., and Zilinskas, A. (1978). The application of Bayesian methods for seeking the extremum. In Dixon, L. and Szego, G., editors, *Towards Global Optimisation 2*, pages 117–129. North Holland, Amsterdam.

Nguyen-Tuong, D., Seeger, M., and Peters, J. (2009). Model learning with local Gaussian process regression. *Advanced Robotics*, 23(15):2015–2034.

Nott, D. and Dunsmuir, W. (2002). Estimation of nonstationary spatial covariance structure. *Biometrika*, 89(4):819–829.

Santner, T., Williams, B., and Notz, W. (2003). *The Design and Analysis of Computer Experiments*. Springer, Heidelberg.

Stein, M. (1999). *Interpolation of Spatial Data. Some Theory for Kriging*. Springer, Heidelberg.

Sun, W., Minasny, B., and McBratney, A. (2006). Analysis and prediction of soil properties using local regression-kriging. *Geoderma*, 171–172:16–23.

Zhu, Z. and Zhang, H. (2006). Spatial sampling design under the infill asymptotic framework. *Environmetrics*, 17(4):323–337.

Zimmerman, D. and Cressie, N. (1992). Mean squared prediction error in the spatial linear model with estimated covariance parameters. *Ann. Inst. Statist. Math.*, 44(1):27–43.