Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

CONDITIONAL SIMULATION FOR EFFICIENT GLOBAL OPTIMIZATION

Jack P.C. Kleijnen
Ehsan Mehdad

Tilburg University
Warandelaan 2
5037 AB Tilburg, NETHERLANDS

ABSTRACT

A classic Kriging or Gaussian process (GP) metamodel estimates the variance of its predictor by plugging-in the estimated GP (hyper)parameters; namely, the mean, variance, and covariances. The problem is that this predictor variance is biased. To solve this problem for deterministic simulations, we propose "conditional simulation" (CS), which gives predictions at an old point that in all bootstrap samples equal the observed value. CS accounts for the randomness of the estimated GP parameters. We use the CS predictor variance in the "expected improvement" criterion of "efficient global optimization" (EGO). To quantify the resulting small-sample performance, we experiment with multi-modal test functions. Our main conclusion is that EGO with classic Kriging seems quite robust; EGO with CS only tends to perform better in expensive simulation with small samples.

1 INTRODUCTION

The goals of *metamodels* may be sensitivity analysis of simulation models and optimization of real systems being simulated. There are several types of metamodels, but most popular are linear regression analysis and Kriging or Gaussian process (GP) models; see the many references to various types of metamodels in Kleijnen 2008, p. 8. In this paper, however, we focus on *Kriging*—which is gaining popularity in the WSC community.

To estimate a Kriging metamodel, we simulate (say) k "points" $\mathbf{x}_i$ ($i = 1, ..., k$) or combinations of the $d \geq 1$ simulation inputs. In this paper we limit our research to *deterministic* simulation, which is popular in engineering (this paper will be a building block for future research on random simulation). We assume that the simulation model is "expensive"; i.e., a single simulation run requires so much computer time to obtain the output (simulation response) w_i that the set of input/output (I/O) data $(\mathbf{X}, \mathbf{w})$ is relatively small—obviously, $\mathbf{X}$ denotes the $k \times d$ matrix with rows $\mathbf{x}_i$, and $\mathbf{w} = (w_1, \ldots, w_k)^\top$. A rule-of-thumb for k (number of points to be simulated) states that a valid Kriging metamodel requires $k = 10d$ points when using (popular) Latin hypercube sampling (LHS) to select these points; see Loeppky et al. (2009).

Classic Kriging (CK) estimates the variance of its predictor by plugging-in the estimated (hyper)parameters of the assumed stationary GP; these parameters are the constant mean β_0, the constant variance τ^2, and the covariance matrix that is determined by the distances among the k points $\mathbf{x}_i$ and the assumed correlation function. For this correlation function we assume the popular so-called Gaussian function with parameters $\boldsymbol{\theta}$; see the definition in (2). Let $\boldsymbol{\psi} = (\beta_0, \tau^2, \boldsymbol{\theta}^\top)^\top$ denote the $(2+d)$-dimensional vector of Kriging parameters. Plugging-in the estimator $\widehat{\boldsymbol{\psi}}$ makes the classic variance of the Kriging predictor *biased*. This bias arises because the resulting Kriging predictor is nonlinear. To study nonlinear statistics, we may apply *parametric bootstrapping*, which is a kind of Monte Carlo sampling with parameters estimated from the "original" data—in our case $(\mathbf{X}, \mathbf{w})$—so it is data-driven; see the general textbook on bootstrapping by Efron and Tibshirani 1993, p. 52 and the additional recent references in Kleijnen 2008,

pp. 81, 84. In this paper we study "conditional simulation" (CS), which improves bootstrapped Kriging (BK) that was originally studied by Den Hertog et al. (2006). Both CS and BK resample old and new points, but CS gives a prediction at an old point w_i that in all bootstrap samples equals the observed value w_i; this property is attractive in deterministic simulation. BK and CS give estimates of the variance of $\widehat{y}(\mathbf{x}_0)$ where $\widehat{y}(\mathbf{x}_0)$ denotes the Kriging predictor of the output at the new point $\mathbf{x}_0$. We use this variance estimator $\widehat{\sigma}^2[\widehat{y}(\mathbf{x}_0)]$ in *efficient global optimization* (EGO), which is a sequential algorithm that balances local and global search for expensive black-box functions; i.e., EGO combines exploitation and exploration (see the classic EGO article Jones et al. (1998)). EGO with BK has already been studied by Kleijnen et al. (2012); now we also combine EGO with CS, and compare EGO combined with CK, BK, or CS.

To select the new point $\mathbf{x}_0$, EGO uses the *expected improvement* (EI) criterion, defined in Section 5. As we shall see, this criterion implies that if two candidate new points (say) $\mathbf{x}_{0;1}$ and $\mathbf{x}_{0;2}$ have the same predicted outputs $\widehat{y}(\mathbf{x}_{0;1}) = \widehat{y}(\mathbf{x}_{0;2})$ but different estimated predictor variances (say) $\widehat{\sigma}^2[\widehat{y}(\mathbf{x}_{0;1})] > \widehat{\sigma}^2[\widehat{y}(\mathbf{x}_{0;2})]$, then EGO selects $\mathbf{x}_{0;1}$, the point with the bigger variance (more uncertainty in the predictor). Consequently, bias in the variance estimates $\widehat{\sigma}^2[\widehat{y}(\mathbf{x})]$ would not matter in EGO, if CK, BK, and CS gave estimates $\widehat{\sigma}^2[\widehat{y}(\mathbf{x})]$ that would reach its maximum for the same new point $\mathbf{x}_0$.

Note: EGO is popular in mathematics and engineering with its deterministic simulation models (often called "computer codes"), but not yet in the "simulation optimization" domain of Management Science/Operations Research. Obviously, bootstrapping is computationally inexpensive compared with expensive simulation, which may take days. Bootstrapping (including CS) can be easily implemented on a parallel computer system.

We numerically illustrate EGO with CK, BK, or CS, and estimate their relative performance. Our experiment uses popular multi-modal test functions for which GPs are only approximations of the true I/O functions. We detail one example with a single input (d = 1). Based on this example and additional examples our main conclusion is that EGO with CK seems quite robust; i.e., EGO with CS or BK only tends to perform better in expensive simulation with small samples.

Besides this introductory section, our paper comprises the following sections. Section 2 summarizes CK. To make this paper stand on its own, Section 3 summarizes BK devised by Den Hertog et al. (2006) and combined with EGO by Kleijnen et al. (2012). Section 4 details CS. Section 5 summarizes EGO. Section 6 details one numerical example for EGO, and summarizes additional experiments. Section 7 presents conclusions and topics for further research.

2 CLASSIC KRIGING (CK)

The basics of Kriging are discussed in many publications, in several disciplines such as geostatistics, engineering, and operations research. Most of our terminology and symbols come from Ankenman et al. (2010).

In deterministic simulation, Kriging is an *exact interpolator*; i.e., the Kriging predictions $y(\mathbf{x}_i) = y_i$ equal the observed simulation outputs $w(\mathbf{x}_i) = w_i$ for the k old input combinations $\mathbf{x}_i$: $y_i = w_i$ (i = 1, ..., k). These "old" I/O data are called the "training sample" in some Kriging publications.

Ordinary Kriging assumes that its output $y(\mathbf{x})$ is a realization of the random process

$$Y(\mathbf{x}) = \beta_0 + M(\mathbf{x}) \tag{1}$$

with the constant mean β_0 and the following stochastic process $M(\mathbf{x})$ with covariance matrix $\mathbf{\Sigma}_M$ ("universal" Kriging does not assume a constant mean, but a linear regression model $\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$). The covariance between $M(\mathbf{x})$ and $M(\mathbf{x}')$ is $\Sigma_M(\mathbf{x}, \mathbf{x}') = \tau^2 R_M(\mathbf{x}, \mathbf{x}')$ where $\mathbf{R}_M$ is a correlation matrix and τ^2 is the constant process variance. More precisely, $M(\mathbf{x})$ is a zero-mean second-order stationary stochastic process so $E[M(\mathbf{x})] = 0$ and the correlation between two points $\mathbf{x}$ and $\mathbf{x}'$ depends only on the distance $|\mathbf{x} - \mathbf{x}'|$. In this paper we use the correlation function that is most popular in simulation; namely, the *Gaussian* correlation

function in product form:

$$R_M(\mathbf{x}, \mathbf{x}', \boldsymbol{\theta}) = \prod_{j=1}^{d} \exp[-\theta_j (x_j - x'_j)^2] \ (\theta_j > 0) \tag{2}$$

where θ_j measures the importance of input j (j = 1, ..., d).

To select $\widehat{Y}(\mathbf{x}_0)$—the *predictor* of the simulation output at a new point $\mathbf{x}_0$—Kriging minimizes the mean squared prediction error (MSPE) criterion:

$$\text{MSPE}[\widehat{Y}(\mathbf{x}_0)] = E[\widehat{Y}(\mathbf{x}_0) - w(\mathbf{x}_0)]^2. \tag{3}$$

The minimum of (3) is determined by the following $(1 + k)$-dimensional Gaussian or Normal distribution:

$$\begin{pmatrix} Y(\mathbf{x}_0) \\ Y(\mathbf{x}) \end{pmatrix} \sim \text{N}_{1+k} \left[\beta_0 \mathbf{1}_{1+k}, \begin{pmatrix} \tau^2 & \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)^\top \\ \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot) & \boldsymbol{\Sigma}_M \end{pmatrix} \right] \tag{4}$$

where $\mathbf{1}_{1+k}$ denotes the vector with all its $(1 + k)$ elements equal to 1 and $\boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)$ denotes the k-dimensional vector with $\text{Cov}[M(\mathbf{x}_0), M(\mathbf{x}_i)]$, which denotes the covariance between the output of the "new" point $\mathbf{x}_0$ and the output of the old point $\mathbf{x}_i$. The predictor is required to be linear (say) $\widehat{Y}(\mathbf{x}_0) = \mathbf{a}^\top Y(\mathbf{x})$ and unbiased so $E[\widehat{Y}(\mathbf{x}_0)|Y(\mathbf{x})] = E[Y(\mathbf{x}_0)|Y(\mathbf{x})]$. The *best linear unbiased predictor* (BLUP) can then be derived to be

$$\widehat{Y}(\mathbf{x}_0, \boldsymbol{\psi}) = \beta_0 + \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_M^{-1} [Y(\mathbf{x}) - \beta_0 \mathbf{1}_k] \tag{5}$$

where we introduce the symbol $\widehat{Y}(\mathbf{x}_0, \boldsymbol{\psi})$ to emphasize that the predictor depends on $\boldsymbol{\psi}$, the vector of GP parameters. Together, (3) and (5) give the $\text{MSPE}[\widehat{Y}(\mathbf{x}_0, \boldsymbol{\psi})]$. Because $\widehat{Y}(\mathbf{x}_0, \boldsymbol{\psi})$ is unbiased, this $\text{MSPE}[\widehat{Y}(\mathbf{x}_0, \boldsymbol{\psi})]$ equals $\sigma^2[\widehat{Y}(\mathbf{x}_0, \psi)]$. It can be derived that

$$\sigma^2[\widehat{Y}(\mathbf{x}_0, \boldsymbol{\psi})] = \tau^2 - \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)^\top \boldsymbol{\Sigma}_M^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot) + \frac{[1 - \mathbf{1}_k^\top \boldsymbol{\Sigma}_M^{-1} \boldsymbol{\Sigma}_M(\mathbf{x}_0, \cdot)]^2}{\mathbf{1}_k^\top \boldsymbol{\Sigma}_M^{-1} \mathbf{1}_k}. \tag{6}$$

In practice, however, $\boldsymbol{\psi}$ is unknown and must be *estimated*. Typically, Kriging uses the maximum likelihood estimators (MLEs), denoted by a hat so $\widehat{\boldsymbol{\psi}} = (\widehat{\beta}_0, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}}^\top)^\top$. These MLEs follow from the log-likelihood function, which follows from the distribution (4). Because this log-likelihood function is rather complicated—possibly with ridges and local maxima—Kriging computes these MLEs numerically through a constrained maximization algorithm. Different Kriging packages use different algorithms. We use the free MATLAB Kriging toolbox called *DACE*, which is well documented by Lophaven et al. (2002) and is often applied in practice; DACE applies the Hooke-Jeeves algorithm.

Note: Bachoc (2013) studies both MLEs and cross-validation estimators and finds that MLEs give more bias for misspecified $\boldsymbol{\Sigma}_M$. We shall use a correctly specified $\boldsymbol{\Sigma}_M$ in Sections 3 and 4, and a misspecified $\boldsymbol{\Sigma}_M$ in Section 6.

The predictor with plugged-in MLE $\widehat{\boldsymbol{\psi}}$ follows from (5):

$$\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}) = \widehat{\beta}_0 + \widehat{\boldsymbol{\Sigma}}_M(\mathbf{x}_0, \cdot)^\top \widehat{\boldsymbol{\Sigma}}_M^{-1} [Y(\mathbf{x}) - \widehat{\beta}_0 \mathbf{1}_k]. \tag{7}$$

Obviously, this predictor is *nonlinear*. Its MSPE and variance are unknown. We define $\widehat{\sigma}^2_{\text{CK}}[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})]$ as the CK estimator of the predictor variance (6) with plugged-in estimators:

$$\widehat{\sigma}^2_{\text{CK}}[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})] = \widehat{\tau}^2 - \widehat{\boldsymbol{\Sigma}}_M(\mathbf{x}_0, \cdot)^\top \widehat{\boldsymbol{\Sigma}}_M^{-1} \widehat{\boldsymbol{\Sigma}}_M(\mathbf{x}_0, \cdot) + \frac{[1 - \mathbf{1}_k^\top \widehat{\boldsymbol{\Sigma}}_M^{-1} \widehat{\boldsymbol{\Sigma}}_M(\mathbf{x}_0, \cdot)]^2}{\mathbf{1}_k^\top \widehat{\boldsymbol{\Sigma}}_M^{-1} \mathbf{1}_k}. \tag{8}$$

We conjecture that this estimator underestimates the true variance, because it ignores the randomness of the MLEs. Above, we introduced the term "classic Kriging" (CK) for all Kriging methods (e.g., ordinary Kriging) that ignore this randomness. To derive the alternative estimators BK and CS, we shall use bootstrapping in the next sections.

3 BOOTSTRAPPED KRIGING (BK)

BK was developed by Den Hertog et al. (2006) to estimate the predictor variance as a function of the location of the new point $\mathbf{x}_0$. It is well-known that as the new point $\mathbf{x}_0$ is closer to an old point $\mathbf{x}_i$, its predictor variance decreases and becomes zero when the new point coincides with an old point. Den Hertog et al. (2006) derive algorithms for (i) a fixed set of new points, (ii) a variable set of new points, and (iii) adding new points one-at-a-time. We use algorithm (iii), because in EGO we shall use a fixed set of candidate points in our search for the one candidate point that maximizes the EI and we ignore the correlation between the outputs of two new points (also see Den Hertog et al. 2006, p. 404). Algorithm (iii) uses the property that N_{1+k} defined in (4) implies that the distribution of the new output given the k old outputs is a conditional normal distribution (also see Den Hertog et al. (2006)'s equation 19). Now we give the steps of their algorithm (iii).

1. We use $\mathrm{N}_k\left(\widehat{\beta}_0\mathbf{1}_k, \widehat{\mathbf{\Sigma}}_M\right)$ B times (B denotes the bootstrap sample size) to sample the k old points $\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}) = (w_{1;b}^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}), \ldots, w_{k;b}^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}))^\top$ where we compute $\widehat{\boldsymbol{\psi}}$ from $(\mathbf{X}, \mathbf{w})$. For each new point $\mathbf{x}_0$ we repeat steps 2 through 4 B times.
2. Given the k old points $\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}})$ of Step 1, we sample $w_b^*(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})$ from

$$\mathrm{N}\left[\widehat{\beta}_0 + \widehat{\mathbf{\Sigma}}_M(\mathbf{x}_0, \cdot)^\top \widehat{\mathbf{\Sigma}}_M^{-1}[Y(\mathbf{x}) - \widehat{\beta}_0\mathbf{1}_k], \widehat{\tau}^2 - \widehat{\mathbf{\Sigma}}_M(\mathbf{x}_0, \cdot)^\top \widehat{\mathbf{\Sigma}}_M^{-1} \widehat{\mathbf{\Sigma}}_M(\mathbf{x}_0, \cdot)\right]. \tag{9}$$

3. Using $\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}})$ of Step 1, we compute $\widehat{\boldsymbol{\psi}}_b^*$. Next we calculate:

$$\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}_b^*) = \widehat{\beta}_{0;b}^* + \widehat{\mathbf{\Sigma}}_M(\mathbf{x}_0, \cdot, \widehat{\boldsymbol{\psi}}_b^*)^\top \widehat{\mathbf{\Sigma}}_M^{-1}(\widehat{\boldsymbol{\psi}}_b^*)[\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}) - \widehat{\beta}_{0;b}^* \mathbf{1}_k].$$

4. This $\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}_b^*)$ together with $w_b^*(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})$ (the bootstrapped new output of Step 2) gives SPE_b = $\mathrm{SPE}[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}_b^*)]$ = $[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}_b^*) - w_b^*(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})]^2$, which is the bootstrap estimator of the squared prediction error (SPE).
5. The B bootstrap samples give the following bootstrap estimator of $\mathrm{MSPE}[\widehat{Y}(\mathbf{x}_0)]$, defined in (3):

$$\mathrm{MSPE}[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}^*)] = \frac{\sum_{b=1}^B \mathrm{SPE}_b}{B}. \tag{10}$$

We ignore the fact that the BK predictor $\widehat{Y}(\mathbf{x}_0, \widehat{\psi}^*)$ is biased, like we assumed that the CK predictor $\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})$ is unbiased—even though the true parameters $\boldsymbol{\psi}$ are replaced by plug-in estimators. Obviously, this makes (10) also give $\widehat{\sigma}^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}^*)]$—abbreviated to $\widehat{\sigma}^2_{\mathrm{BK}}$—which is the bootstrap estimator of $\sigma^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})]$.

The standard error (SE) of $\widehat{\sigma}^2_{\mathrm{BK}}$ is computed straightforwardly, because the B random variables in (10) are independently sampled from the same distribution so they are independently and identically distributed (IID): $\mathrm{SE}\{\widehat{\sigma}^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}^*)]\}$ = $[\sum_{b=1}^B (\mathrm{SPE}_b - \mathrm{MSPE})^2/\{(B-1)B\}]^{1/2}$. So this SE decreases with $B^{1/2}$. We use t_{B-1} (Student t-statistic with $B-1$ degrees of freedom) to compute a two-sided symmetric $(1-\alpha)$ CI:

$$P\{\sigma^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})] \in \widehat{\sigma}^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}^*)] \pm t_{B-1;\alpha/2}\, \mathrm{SE}\{\widehat{\sigma}^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}^*)]\}\} = 1 - \alpha. \tag{11}$$

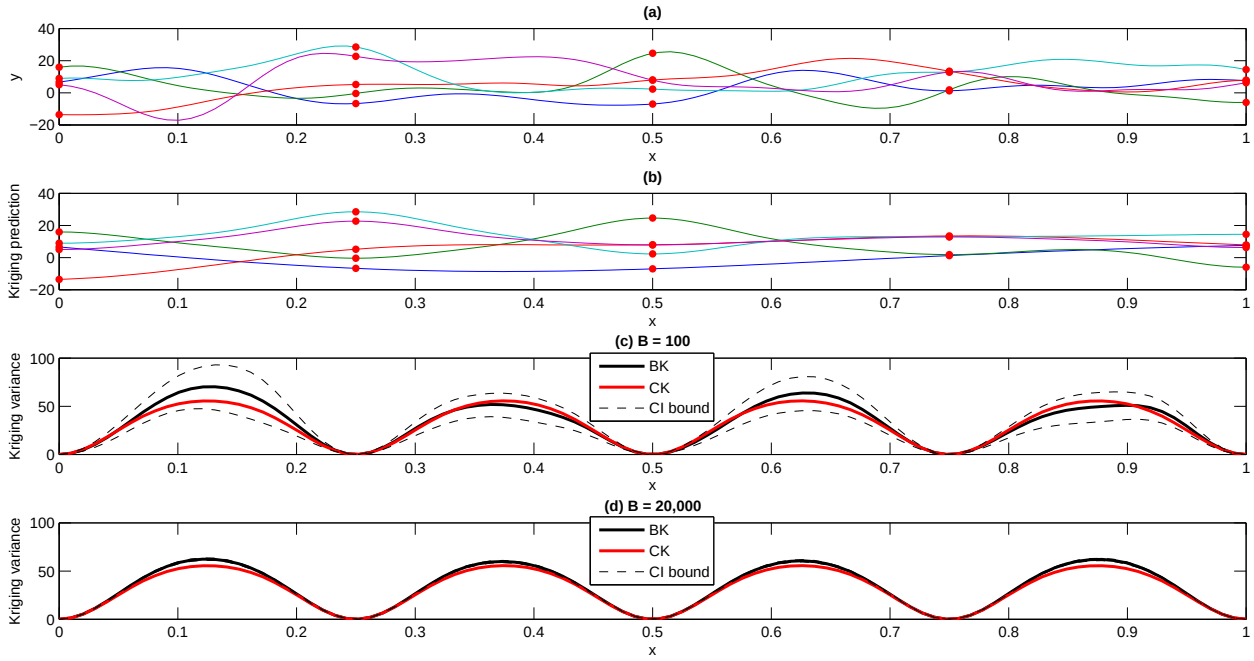


Figure 1: Illustration of BK; (a): jointly sampled outputs at 5 equi-spaced old and 98 equi-spaced new points, for B = 5; (b): Kriging predictions for 98 new points based on 5 old points sampled in (a); (c): estimated predictor variances and their 95% CIs with B = 100, and CK's predictor variances; (d): same as (c) but with B = 20,000

This CI assumes that t_{B-1} is not very sensitive to possible non-normality of SPE, which features in (10). For large B, we have $t_{B-1;\alpha/2} \downarrow z_{\alpha/2}$ where $z_{\alpha/2}$ denotes the $\alpha/2$ quantile of the standard normal variable z so (4) implies $z \sim \mathrm{N}(0,1)$ and $P(z < z_{\alpha/2}) = \alpha/2$.

Figure 1 illustrates BK. Part (a) displays only B = 5 samples, to avoid cluttering-up the plot; notice that each of these B samples has its own output values at the old points. Part (b) shows less "wiggling" than part (a); the predictions at old points coincide with the values sampled in part (a). Part (c) uses B = 100; it also displays CK's $\widehat{\sigma}^2_{\mathrm{CK}}$ computed from (8). Part (d) uses B = 20,000 to confirm our conjecture that BK tends to give bigger estimates $\widehat{\sigma}^2_{\mathrm{BK}}$ than $\widehat{\sigma}^2_{\mathrm{CK}}$.

4 CONDITIONAL SIMULATION (CS)

CS is popular in the French literature on Kriging; see the references in Wackernagel 2003, p. 188. We formulate the *basic idea* of CS in Chilès and Delfiner 1999, pp. 465-469 as follows. Let $S(\cdot)$ be a non-conditional simulation (or bootstrap sample) of $Y(\cdot)$ independent of $Y(\cdot)$ and with the same covariance as $Y(\cdot)$. When "conditioning", we pass from $S(\cdot)$ to a simulation $Y_{\mathrm{CS}}(\cdot)$ that is equal to $Y(\cdot)$ in the old points. Let $\widehat{Y}(\mathbf{x}_0)$ be the Kriging predictor of $Y(\mathbf{x}_0)$ based on the old I/O data $(\mathbf{X}, \mathbf{w})$. Obviously, we have $Y(\mathbf{x}_0) = \widehat{Y}(\mathbf{x}_0) + [Y(\mathbf{x}_0) - \widehat{Y}(\mathbf{x}_0)]$ where the Kriging error $Y(\mathbf{x}_0) - \widehat{Y}(\mathbf{x}_0)$ is unknown because $Y(\mathbf{x}_0)$ is unknown. Analogously, we have $S(\mathbf{x}_0) = \widehat{S}(\mathbf{x}_0) + [S(\mathbf{x}_0) - \widehat{S}(\mathbf{x}_0)]$, but now $S(\mathbf{x}_0)$ is known and so is the error term. Substituting the simulated error into the decomposition of $Y(\mathbf{x}_0)$, we obtain $Y_{\mathrm{CS}}(\mathbf{x}_0) = \widehat{Y}(\mathbf{x}_0) + [S(\mathbf{x}_0) - \widehat{S}(\mathbf{x}_0)]$. Because Kriging is an exact interpolator at an old point $\mathbf{x}$, we have $\widehat{Y}(\mathbf{x}) = Y(\mathbf{x})$ and $\widehat{S}(\mathbf{x}) = S(\mathbf{x})$ so $Y_{\mathrm{CS}}(\mathbf{x}) = Y(\mathbf{x})$. Notice that Journel and Huijbregts 2003, pp. 496-498 prove that $Y_{\mathrm{CS}}(\cdot)$ preserves the covariance of $Y(\cdot)$.

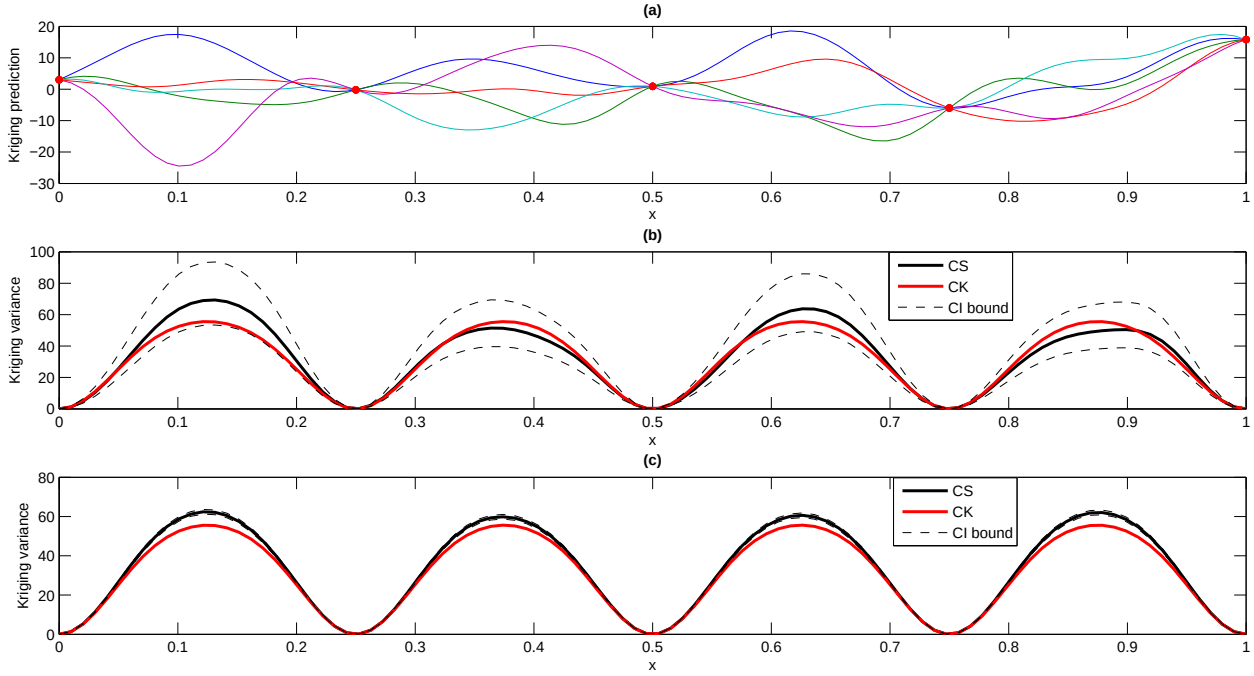


Figure 2: Illustration of CS; (a): predictions at 98 new points, for $B = 5$; (b): estimated predictor variances and their 95% CIs for $B = 100$, and CK's predictor variances; (c): same as (b) but for $B = 20{,}000$

Whereas Chilès and Delfiner (1999) focus on spatial data in geostatistics, we focus on simulation models. Whereas we see CS as a type of parametric bootstrapping, Chilès and Delfiner 1999, p. 453 call CS a "Monte Carlo method". We detail our CS algorithm as follows.

1. We use $\mathrm{N}_k\left(\widehat{\beta}_0\mathbf{1}_k, \widehat{\mathbf{\Sigma}}_M\right)$ B times to sample the k old points $\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}) = (w_{1;b}^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}), \ldots, w_{k;b}^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}))^\top$ where we compute $\widehat{\boldsymbol{\psi}}$ from $(\mathbf{X}, \mathbf{w})$. For each new point $\mathbf{x}_0$ we repeat steps 2 through 4 B times.
2. We use the conditional normal distribution (9) to sample $w_b^*(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})$ given the k old points $\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}})$.
3. $\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}})$ from Step 1 gives $\widehat{\boldsymbol{\psi}}_b^*$. Next we calculate:

$$\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}_b^*) = \widehat{\beta}_{0;b}^* + \widehat{\mathbf{\Sigma}}_M(\mathbf{x}_0, \cdot, \widehat{\boldsymbol{\psi}}_b^*)^\top \widehat{\mathbf{\Sigma}}_M^{-1}(\widehat{\boldsymbol{\psi}}_b^*)[\mathbf{w}_b^*(\mathbf{X}, \widehat{\boldsymbol{\psi}}) - \widehat{\beta}_{0;b}^*\mathbf{1}_k]. \tag{12}$$

4. Combining CK and (12), we compute the CS output at the new point:

$$\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0, b) = \widehat{\beta}_0 + \widehat{\mathbf{\Sigma}}_M(\mathbf{x}_0, \cdot)^\top \widehat{\mathbf{\Sigma}}_M^{-1}(\mathbf{w} - \widehat{\beta}_0\mathbf{1}_k) + [w_b^*(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}) - \widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}}_b^*)]. \tag{13}$$

5. Finally, we use the B samples to compute

$$\widehat{\sigma}^2[\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0)] = \frac{\sum_{b=1}^B [\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0, b) - \overline{\widehat{Y}}_{\mathrm{CS}}(\mathbf{x}_0)]^2}{B-1} \text{ with } \overline{\widehat{Y}}_{\mathrm{CS}}(\mathbf{x}_0) = \frac{\sum_{b=1}^B \widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0, b)}{B}.$$

Figure 2 illustrates CS. Part (a) displays $\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0, b)$ for $b = 1, \cdots, 5$. To obtain a CI for $\sigma^2[\widehat{Y}(\mathbf{x}_0, \widehat{\boldsymbol{\psi}})]$, we replace t_{B-1} in (11) (which assumes B IID variables) by $\chi_{B-1}^2 = (B-1)\widehat{\sigma}^2[\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0)]/\sigma^2[\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0)]$ (which applies for the classic variance estimator $\widehat{\sigma}^2[\widehat{Y}_{\mathrm{CS}}(\mathbf{x}_0)]$); this gives a two-sided asymmetric $(1-\alpha)$ CI:

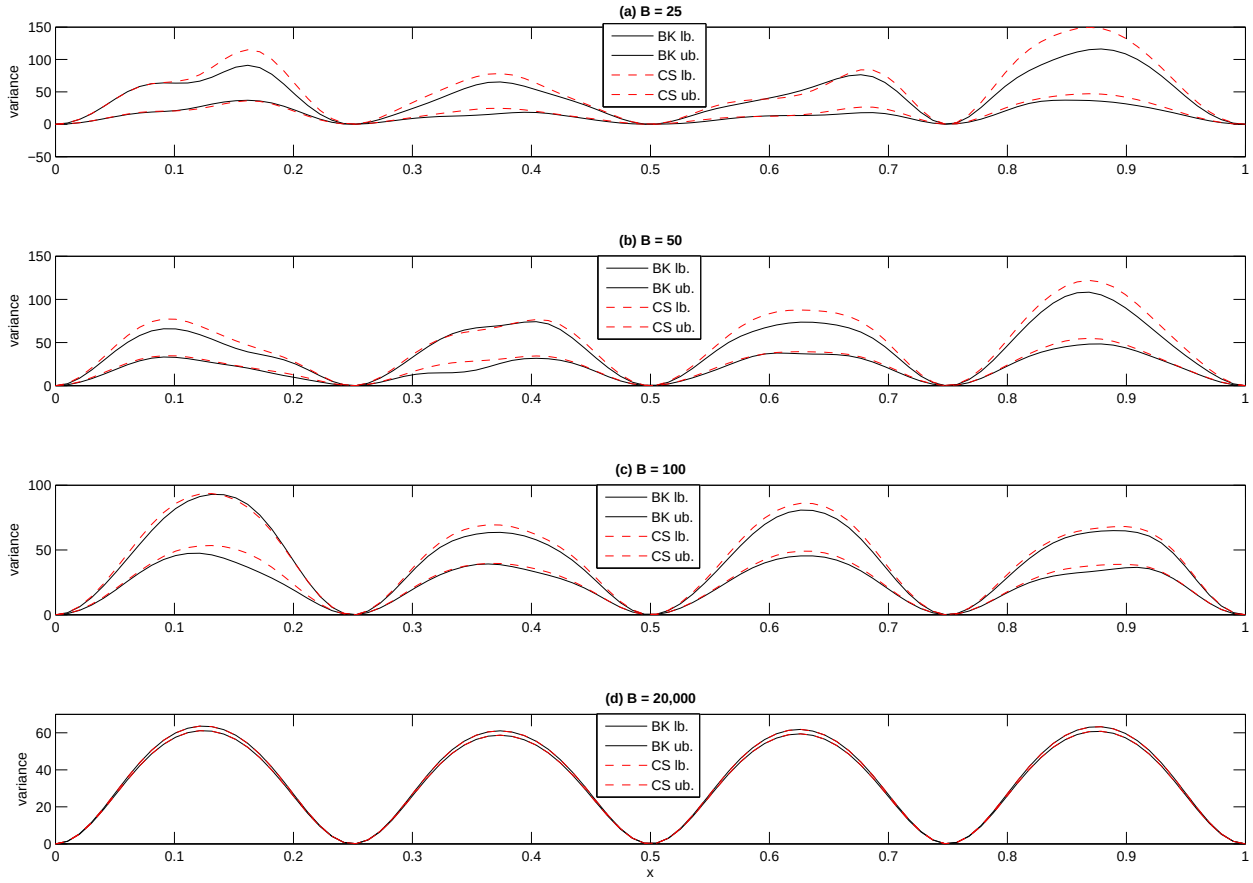


Figure 3: Illustration of CIs for BK versus CS for various B

$$P\{\frac{(B-1)\widehat{\sigma}^2[\widehat{Y}_{\text{CS}}(\mathbf{x}_0)]}{\chi^2_{B-1;1-\alpha/2}} \leq \sigma^2[\widehat{Y}(\mathbf{x}_0,\widehat{\boldsymbol{\psi}})] \leq \frac{(B-1)\widehat{\sigma}^2[\widehat{Y}_{\text{CS}}(\mathbf{x}_0)]}{\chi^2_{B-1;\alpha/2}}\} = 1-\alpha. \tag{14}$$

Part (b) of the figure displays $\widehat{\sigma}^2[\widehat{Y}_{\text{CS}}(\mathbf{x}_0)] = \widehat{\sigma}^2_{\text{CS}}$ and its 95% CIs if B = 100; it also displays $\widehat{\sigma}^2_{\text{CK}}$. Based on this part we conjecture that $\widehat{\sigma}^2_{\text{CS}}$ tends to exceed $\widehat{\sigma}^2_{\text{CK}}$; part (c) displays results if B = 20,000, which confirms our conjecture.

Furthermore, we conjecture that $\widehat{\sigma}^2_{\text{BK}}$ tends to exceed $\widehat{\sigma}^2_{\text{CS}}$ because CS implies *conditional* sampling. Figure 3 displays $\widehat{\sigma}^2_{\text{CS}}$ and $\widehat{\sigma}^2_{\text{BK}}$ and their CIs, for various values of B. Actually, this figure suggests that for $B \uparrow \infty$ the two estimators tend to the same asymptotic value; for small samples, CS does not give a significantly smaller value. On hindsight, these results seem reasonable; i.e., both CS and BK use $\widehat{\boldsymbol{\psi}}$, which is the *sufficient* statistic of the GP computed from the same $(\mathbf{X}, \mathbf{w})$. We feel that CS is simpler than BK—both computationally and conceptually.

5 EFFICIENT GLOBAL OPTIMIZATION (EGO)

EGO searches for the global optimum, sequentially. To guide its search, EGO uses the *expected improvement (EI)* criterion. This EI is computed as follows, if the EGO goal is to minimize the simulation output w.

1. Fit a Kriging metamodel $Y(\mathbf{x})$ to the old I/O data $(\mathbf{X}, \mathbf{w})$.
2. Find the minimum output observed (simulated) so far: $f_{\min} = \min_{1 \leq i \leq k} w(\mathbf{x}_i)$.

3. Find $\widehat{\mathbf{x}}_{\text{opt}}$, which denotes the estimate of $\mathbf{x}_0$ that *maximizes* $\text{EI}(\mathbf{x}) = E\left[\max(f_{\min} - Y(\mathbf{x}), 0)\right]$. Assuming $Y(\mathbf{x}) \sim \text{N}\left(\widehat{Y}(\mathbf{x}), \widehat{\sigma}^2_{\text{CK}}(\mathbf{x})\right)$, Jones et al. (1998) derive

$$\text{EI}(\mathbf{x}) = \left(f_{\min} - \widehat{Y}(\mathbf{x})\right) \Phi\left(\frac{f_{\min} - \widehat{Y}(\mathbf{x})}{\widehat{\sigma}_{\text{CK}}(\mathbf{x})}\right) + \widehat{\sigma}_{\text{CK}}(\mathbf{x}) \phi\left(\frac{f_{\min} - \widehat{Y}(\mathbf{x})}{\widehat{\sigma}_{\text{CK}}(\mathbf{x})}\right) \tag{15}$$

   with $\widehat{Y}(\mathbf{x})$ and $\widehat{\sigma}_{\text{CK}}(\mathbf{x})$ defined in (7) and (8), and Φ and ϕ denoting the cumulative distribution function and the probability density function of the standard normal variable z.
4. Run the simulation model with $\widehat{\mathbf{x}}_{\text{opt}}$ found in step 3, and obtain $w(\widehat{\mathbf{x}}_{\text{opt}})$.
5. Fit a new Kriging metamodel to the old I/O data of step 1 and the new I/O of step 4. Update k and return to step 2 if the stopping criterion is not yet satisfied.

Like all sequential procedures, EGO needs to select an initial sample size k and a stopping criterion. If EGO starts with a "too small" k, then the Kriging metamodel is a poor approximation which gives poor guidance of the search for the optimum. As a stopping criterion Jones et al. (1998) use EI $<$ 0.01 $f_{\min}$; Kleijnen et al. (2012) use EI $< 10^{-20}$, which is scale dependent; Huang et al. (2006) require that their stopping criterion is satisfied $d + 1$ times in a row. We think that in expensive simulation, a practical stopping criterion may be exhaustion of the computer budget or meeting the deadline for reporting the estimated optimal I/O to the client. We shall report some numerical results, in the next section.

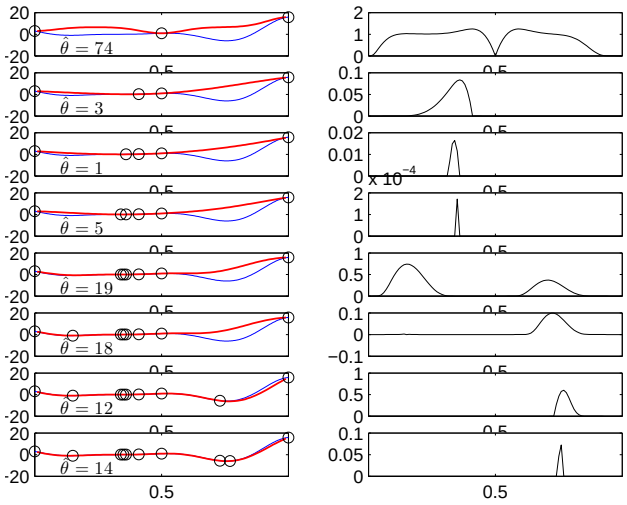


Figure 4: EGO for the example in Forrester et al. 2008, p. 83

Figure 4 illustrates EGO through the following test function that is also used by Forrester et al. 2008, p. 83:

$$w(x) = (6x - 2)^2 \sin(12x - 4) \text{ with } 0 \leq x \leq 1. \tag{16}$$

It is easy to derive that if x is continuous, then this function has one local minimum at $x = 0.01$ and one global minimum at $x_{\text{opt}} = 0.7572$ with output $w(x_{\text{opt}}) = -6.02074$; also see the curves in the left part of the figure, the blue curve is the true function and the red one is the Kriging metamodel. Forrester et al. 2008, p. 92 start with $k = 3$ old points, and stop after sequentially adding seven new points. After each new point they re-estimate θ; see $\widehat{\theta}$ in the left part. Below (6) we have already pointed out that the GP log-likelihood function is complicated, and uses some constrained maximization algorithm. Forrester et al. (2008) use their own genetic algorithm, and we use MATLAB's DACE; moreover, DACE needs an initial value for θ, which may affect $\widehat{\theta}$. Different algorithms may give different MLEs; e.g., Forrester et al. (2008)

display $\widehat{\theta}$ = 1 if k = 3 and $\widehat{\theta}$ = 12.51 when k = 10, whereas our figure displays $\widehat{\theta}$ = 74 when k = 3 and $\widehat{\theta}$ = 14 when k = 10. So, when k is "small", the Kriging metamodel is a "poor" approximation; i.e., $\widehat{\theta}$ shows much variation around the true θ. For k = 10 the difference between Forrester et al. (2008)'s $\widehat{\theta}$ and our $\widehat{\theta}$ is small. Obviously, $\widehat{\theta}$ determines EI in (15); the right-hand part of the figure displays EI as k increases; our results are similar to Forrester et al. (2008)'s results, except for k = 3 where we have two hills instead of one hill.

6 EXPERIMENTS WITH THREE EGO VARIANTS

We use numerical experiments to evaluate three EGO variants; namely, EGO with $\widehat{\sigma}_{\text{CK}}(\mathbf{x})$ as in (8) and EGO with $\widehat{\sigma}_{\text{CK}}(\mathbf{x})$ replaced by either $\widehat{\sigma}_{\text{BK}}(\mathbf{x})$ or $\widehat{\sigma}_{\text{CS}}(\mathbf{x})$. We measure the performance of these variants through the number of simulated input combinations needed to estimate the true optimal input combination.

In the preceding sections we observed $\widehat{\sigma}^2_{\text{CK}} < \widehat{\sigma}^2_{\text{BK}} \approx \widehat{\sigma}^2_{\text{CS}}$. However, EGO searches for the point with the maximum predictor variance, if for simplicity we temporarily assume that the predicted values for the candidate points are the same. If (for example) $\widehat{\sigma}^2_{\text{CK}}(\mathbf{x}) = c(\mathbf{x})\widehat{\sigma}^2_{\text{CS}}(\mathbf{x})$ with $c(\mathbf{x}) = c > 1$, then the same combination (say) $\mathbf{x}_{\text{opt}}$ maximizes both $\widehat{\sigma}^2_{\text{CK}}(\mathbf{x})$ and $\widehat{\sigma}^2_{\text{CS}}(\mathbf{x})$. In general, however, $\widehat{\sigma}^2_{\text{CK}}(\mathbf{x})$ in (15) may or may not lead to a new point that differs from the new point selected through EGO with $\widehat{\sigma}_{\text{CS}}(\mathbf{x})$ or $\widehat{\sigma}_{\text{BK}}(\mathbf{x})$. Our experiments show that the three EGO variants may indeed select different points.

We detail one example—namely, (16), which implies d = 1—and we summarize three more examples with d is 2, 3, or 6. Because $\widehat{\sigma}^2_{\text{BK}}(\mathbf{x})$ and $\widehat{\sigma}^2_{\text{CS}}(\mathbf{x})$ imply sampling, we also obtain macroreplications, which use different non-overlapping PRN streams while fixing all other experimental factors (e.g., B). Altogether, we obtain 20 macroreplications; obviously, we do not need macroreplications for EGO with $\widehat{\sigma}^2_{\text{CK}}(\mathbf{x})$.

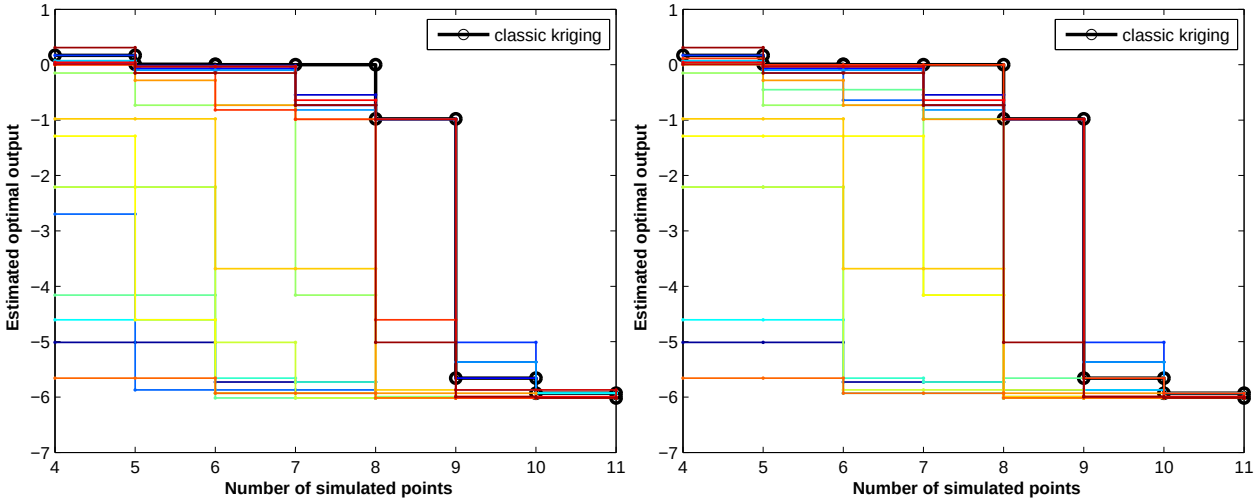
So we must select a value for B. Chilès and Delfiner 1999, p. 453 point out that selecting a specific value B of CS observations depends on the problem. In the general context of bootstrapping, Efron and Tibshirani 1993, p. 52 write: "B = 25 is usually informative; B = 50 is often enough to give a good estimate of the standard error ... very seldom B = 200 is needed". We select B = 100 (we could have started with a "smaller" B, and next add more bootstrap observations and observe how $\widehat{\sigma}_{\text{BK}}(\mathbf{x})$ or $\widehat{\sigma}_{\text{CS}}(\mathbf{x})$ converges).

We verify our computer code, comparing our results for EGO with $\widehat{\sigma}_{\text{CK}}(\mathbf{x})$ and $\widehat{\sigma}_{\text{BK}}(\mathbf{x})$ with Kleijnen et al. (2012). We apply Den Hertog et al. (2006)'s third algorithm, adding new points one-at-a-time (erroneously Kleijnen et al. (2012) state that the second algorithm was applied); we select this algorithm because we use EGO with a fixed set of *candidate points* $\mathbf{x}_{0;t}$ with t = 1, ..., k_0 when searching for $\widehat{\mathbf{x}}_{\text{opt}}$ in Step 3 of EGO. So we may ignore the correlation between the outputs of two new points $\mathbf{x}_{0;t}$ and $\mathbf{x}_{0;t'}$ (t, t' = 1, ..., k_0). This verification gives acceptable results.

In the example with d = 1, we select k_0 = 98; namely, 100 equi-spaced points, excluding the two extreme points, 0 and 1. We compare EGO with $\widehat{\sigma}_{\text{CK}}(\mathbf{x})$, $\widehat{\sigma}_{\text{BK}}(\mathbf{x})$, and $\widehat{\sigma}_{\text{CS}}(\mathbf{x})$. The latter two give similar (but not identical) estimates of θ and EI, for a given k. Of course, these estimates change as k increases, and vary among macroreplications. Altogether, these estimates are similar to the ones for $\widehat{\sigma}_{\text{CK}}(\mathbf{x})$ in Figure 4. To save space we do not detail these results, which are only intermediate; i.e., we now proceed to the final results.

We select as stopping criterion EI $< 10^{-20}$, so we do not stop "early"; i.e., we can observe possible *convergence*, as Figure 5 illustrates. Like Huang et al. (2006) we display $f_{\min}(k) = \min_{1 \le i \le k} w(\mathbf{x}_i)$, which denotes the estimated optimal simulation output after k simulated input combinations; horizontal lines mean that the most recent simulated point $\widehat{\mathbf{x}}_{\text{opt}}(k)$ does not give a lower output than a preceding point. This figure shows $f_{\min}(k)$ for BK and CS relative to CK. More specifically, the black step function with circles represents $f_{\min}(k)$ for CK. The colored step functions represent $f_{\min}(k)$ for BK (left-hand panel) or CS (right-hand panel). Actually, a colored step function may represent more than one macroreplication; e.g., for Forrester et al. (2008)'s function we obtain 20 macroreplications, but in the two panels we cannot distinguish 20 colored step functions. All three EGO variants give the same estimated optimal I/O for k =11; namely, $\widehat{x}_{\text{opt}}$ = 0.76 and $\widehat{w}_{\text{opt}} = w(\widehat{x}_{\text{opt}})$ = -6.017 (the true values for continuous x were listed

below (16): $x_{\text{opt}} = 0.7572$ and $w_{\text{opt}} = -6.02074$). For expensive simulations with small sample sizes, this asymptotic solution is not relevant; the detailed data behind the figure reveal that CK performs better than both BK and CS in one macroreplication when $k = 4$, three mecroreplications when $k = 9$, and two macroreplications when $k = 10$.



(a) EGO with $\widehat{\sigma}_{\text{BK}}(\mathbf{x})$ (b) EGO with $\widehat{\sigma}_{\text{CS}}(\mathbf{x})$

Figure 5: Estimated optimal output after k simulated input combinations

Finally, we perform additional experiments with three more popular test functions; namely, the six-humped camel-back with $d = 2$, and the Hartman-3 and Hartman-6 functions with $d = 3$ and $d = 6$. We shall report details in a next paper. Based on these experiments, we conclude that EGO with CK seems quite robust; i.e., EGO with CS or BK only tends (but does not guarantee) to perform better in expensive simulation with small samples.

7 CONCLUSIONS AND FUTURE RESEARCH

In this paper we studied the problem that CK gives estimates of the variance of its predictor for a new point by simply plugging-in the estimated GP parameters $\widehat{\psi}$ so this variance is biased. As a new solution we propose CS, which improves BK; i.e., CS is computationally and conceptually simpler. We find experimentally that CS gives predicted variances that do not differ significantly from BK, but that tend to exceed the classic estimate. We use CS in EGO's EI criterion, but CK seems quite robust.

In a next paper, we shall give details on examples with $d > 1$, and CIs for the Kriging predictors that are either parametric using the estimated variances of the Kriging predictors ($\widehat{\sigma}^2_{\text{CK}}$, $\widehat{\sigma}^2_{\text{BK}}$, $\widehat{\sigma}^2_{\text{CS}}$) or distribution-free using CS with the percentile method in Efron and Tibshirani 1993, p. 52. In future research, we shall also adapt EGO for random simulation with replications, using distribution-free bootstrapping. We may also apply mathematical methods that update $\widehat{\psi}$ when adding one new point, such that computations do not start from "scratch"; see Emery (2009) and Frazier et al. (2009).

ACKNOWLEDGMENTS

We thank David Ginsbourger (University of Bern, Bern, Switzerland) for suggesting conditional simulation (CS) as an alternative for bootstrapped Kriging (BK), Inneke van Nieuwenhuyse (K.U. Leuven, Leuven, Belgium) for sharing her MATLAB code for EGO using BK and her help with the implementation of that code, and Alex Siem (ORTEC, Gouda, Netherlands) for sharing his MATLAB code for BK. We thank two

anonymous referees and Dick den Hertog (Tilburg University) for their comments on an earlier version submitted to the WSC 2013 proceedings.

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for simulation metamodeling". *Operations Research* 58:371–382.

Bachoc, F. 2013. "Cross Validation and Maximum Likelihood estimations of hyper-parameters of Gaussian processes with model misspecification". *Computational Statistics & Data Analysis*. In Press.

Chilès, J., and P. Delfiner. 1999. *Geostatistics: Modeling Spatial Uncertainty*. 1 ed. Wiley.

Den Hertog, D., J. P. C. Kleijnen, and A. Y. D. Siem. 2006. "The correct Kriging variance estimated by bootstrapping". *Journal of The Operational Research Society* 57:400–409.

Efron, B., and R. Tibshirani. 1993. *An Introduction to the Bootstrap*. Monographs on Statistics and Applied Probability. Chapman & Hall.

Emery, X. 2009. "The Kriging update equations and their application to the selection of neighboring data". *Computational Geosciences* 13 (3): 269–280.

Forrester, A., A. Sóbester, and A. Keane. 2008. *Engineering Design via Surrogate Modelling: A Practical Guide*. 1 ed. Wiley.

Frazier, P., W. Powell, and S. Dayanik. 2009. "The Knowledge-Gradient Policy for Correlated Normal Beliefs". *INFORMS Journal on Computing* 21 (4): 599–613.

Huang, D., T. T. Allen, W. I. Notz, and N. Zeng. 2006. "Global Optimization of Stochastic Black-Box Systems via Sequential Kriging Meta-Models". *Journal of Global Optimization* 34:441–466.

Jones, D. R., M. Schonlau, and W. J. Welch. 1998. "Efficient Global Optimization of Expensive Black-Box Functions". *Journal of Global Optimization* 13:455–492.

Journel, A., and C. J. Huijbregts. 2003. *Mining Geostatistics*. The Blackburn Press.

Kleijnen, J. P., W. Van Beers, and I. Van Nieuwenhuyse. 2012. "Expected improvement in efficient global optimization through bootstrapped Kriging". *Journal of Global Optimization* 54:59–73.

Kleijnen, J. P. C. 2008. *Design and Analysis of Simulation Experiments*. Springer-Verlag.

Loeppky, J. L., J. Sacks, and W. J. Welch. 2009. "Choosing the Sample Size of a Computer Experiment: A Practical Guide". *Technometrics* 51:366–376.

Lophaven, S. N., H. B. Nielsen, and J. Sondergaard. 2002. *DACE: a MATLAB Kriging toolbox, version 2.0*. Lyngby, Denmark: IMM Technical University of Denmark.

Wackernagel, H. 2003. *Multivariate Geostatistics: An Introduction with Applications*. New York: Springer-Verlag.

AUTHOR BIOGRAPHIES

JACK P.C. KLEIJNEN is Professor of "Simulation and Information Systems" at Tilburg University, where he is a member of both the Department of Information Management and the Operations Research Group of the Center for Economic Research (CentER) in the Tilburg School of Economics and Management (TiSEM). His research concerns the statistical design and analysis of experiments with simulation models, in many scientific disciplines (e.g., management, economics, and engineering). He was a consultant for several organizations in the USA and Europe. He serves on many international editorial boards and scientific committees. He spent several years in the USA, at universities and private companies. He received a number of national and international awards; e.g., in 2008 he received a knighthood and in 2005 an LPAA. His e-mail address is kleijnen@tilburguniversity.edu and his web page is http://www.tilburguniversity.edu/webwijs/show/?uid=kleijnen.

EHSAN MEHDAD is a Ph.D. student at Tilburg University. His research interests are in discrete-event simulation, metamodels (Kriging) and simulation optimization. His email address is emehdad@gmail.com.