# Estimating the Variance of the Predictor in Stochastic Kriging

Kleijnen, J.P.C.; Mehdad, Ehsan

No. 2015-041

# ESTIMATING THE VARIANCE OF THE PREDICTOR IN STOCHASTIC KRIGING

By

Jack P.C. Kleijnen, Ehsan Mehdad

18 August 2015

ISSN 0924-7815
ISSN 2213-9532

# Estimating the variance of the predictor in stochastic Kriging

Jack P.C. Kleijnen and Ehsan Mehdad
Tilburg School of Economics and Management, Tilburg University

July 24, 2015

**Abstract**

We study the correct estimation of the true variance of the predictor in stochastic Kriging (SK). First, we obtain macroreplications for a SK metamodel that approximates a single-server simulation model; these macroreplications give independently and identically distributed predictions. This simulation may use common random numbers (CRN). From these macroreplications we conclude that the usual plug-in estimator of the variance significantly underestimates the true variance. Because macroreplications of practical simulation models are computationally expensive, we next formulate two bootstrap methods that use a single macroreplication: (i) a distribution-free method that resamples simulation replications (within the single macroreplication), and (ii) a parametric method that assumes a Gaussian distribution for the SK predictor, and estimates the (hyper)parameters of that distribution from the single macroreplication. Altogether we recommend distribution-free bootstrapping for the estimation of the SK predictor variance in practical simulation experiments.

**Keywords:** Kriging, Gaussian process, predictor variance, plug-in, bootstrap

**JEL**: C0, C1, C9, C15, C44

## 1 Introduction

In practice, the final goal of simulation is often sensitivity analysis and optimization of the simulated real system. For this goal, the simulation analysts often use a *metamodel*—also called an emulator or a surrogate—which is a model of the underlying simulation model; i.e., it is an explicit simple approximation of the input/output (I/O) function that is implicitly defined by the simulation model. There are many types of metamodels; see Kleijnen (2015, p. 10). We, however, focus on *Kriging*—also called *Gaussian process* (GP)—metamodels. These metamodels have already acquired a track record in deterministic simulation, and are becoming popular in random (stochastic) simulation—including

discrete-event simulation—which transforms a stream of *pseudo-random numbers* (PRN) into an output. Kriging in stochastic simulation is called *stochastic Kriging* (SK); see the classic article Ankenman et al. (2010), recent articles such as Barton et al. (2014), Bekki et al. (2014), Chen and Kim (2014), Plumlee and Tuo (2014), Qu and Fu (2014), and Sun et al. (2014), and the many more references in Kleijnen (2015, pp. 206–211).

*Common random numbers* (CRN) are a popular variance-reduction technique that may improve the statistical analysis of the simulation I/O data in sensitivity analysis and optimization; see Law (2015, pp. 586–604). However, CRN also complicate this analysis when using SK; see Chen et al. (2012).

In practice, the (hyper)parameters of the GP are unknown, so they are estimated. We focus on the problem of correctly estimating the true variance of the predictor that uses SK with *estimated* Kriging parameters. This problem has already been investigated for Kriging in deterministic simulation; see the recent survey in Thiart et al. (2014). In our paper we investigate the following three methods for estimating the variance of the SK predictor: (i) macroreplications, (ii) distribution-free bootstrapping (DB), and (iii) parametric bootstrapping (PB). We illustrate and evaluate these three methods through experiments with the a waiting-time simulation model; namely, a single-server discrete-event simulation model with independent exponential interarrival and service times, known as the *M/M/1* model. For details on the simulation of this model we refer to Law (2015, pp. 73–77, 102–108).

We organize and summarize the rest of this paper as follows. In Section 2 we summarize SK. In Section 3 we first describe an M/M/1 simulation experiment, and then explain how macroreplications provide an unbiased estimator of the predictor variance; we find that this estimator is significantly higher than the plug-in estimator—without or with CRN. In Section 4 we detail DB, and find that DB is a relatively quick and easy method for estimating the true variance when CRN are applied. In Section 5 we detail PB, and find that PB gives a higher estimate than DB does. In Section 6 we summarize our conclusions; namely, DB gives a fast and correct estimator of the predictor variance in practical simulation experiments.

# 2 Stochastic Kriging

SK assumes that the simulation output (say) $w$ is stochastic. Notice that unlike most authors on SK, we do distinguish between $w$ (simulation output) and $y$ (metamodel output). Furthermore, we use Greek letters to denote unknown parameters, and bold upper-case letters for matrixes and bold lower-case letters for vectors. The simplest type of SK assumes the following metamodel:

$$y_r(\mathbf{x}) = \mu + M(\mathbf{x}) + \varepsilon_r(\mathbf{x}) \text{ with } \mathbf{x} \in \mathbb{R}^k \text{ and } r = 1, \ldots, m_i, \tag{1}$$

where we use the following symbols. We let $\mu$ denote the constant mean $E[y(\mathbf{x})]$ where $\mathbf{x}$ is an input combinations or 'point' in the given $k$-dimensional experimental area $\mathbb{R}^k$. The *extrinsic noise* $M(\mathbf{x})$ is the additive noise that forms

a Gaussian (multivariate normal) stationary process with zero mean and covariance matrix $\mathbf{\Sigma}_M$ (by definition, a stationary process has a constant mean, a constant variance, and covariances that depend only on the distance between the points $\mathbf{x}$ and $\mathbf{x}'$). Because stochastic simulation has noisy outputs, we should obtain *replications*; i.e., *identically independently distributed* (IID) simulation outputs. The *intrinsic noise* $\varepsilon_r(\mathbf{x})$ has a Gaussian distribution with zero mean and variance $\text{Var}[\varepsilon(\mathbf{x})]$ and is independent of the extrinsic noise $M(\mathbf{x})$. We let $m_i$ denote the number of replications if $\mathbf{x} = \mathbf{x}_i$ so the subscript $r$ runs from 1 to $m_i$. We let $\mathbf{\Sigma}_\varepsilon$ denote the covariance matrix of $\varepsilon$. So, if the simulation does not use CRN, then $\mathbf{\Sigma}_\varepsilon$ is diagonal with the elements $\text{Var}[\varepsilon(\mathbf{x})]$ on the main diagonal. If the simulation does use CRN, then $\mathbf{\Sigma}_\varepsilon$ is not diagonal; obviously, $\mathbf{\Sigma}_\varepsilon$ should still be symmetric and positive definite.

Note: The SK model defined in Eq. (1)—but without intrinsic noise—is known as *ordinary Kriging* (OK). A more general model replaces the constant $\mu$ in Eq. (1) by a prespecified low-order polynomial in $\mathbf{x}$. General wisdom, however, claims that in practice the latter model usually does not give better predictions; seeTajbakhsh et al. (2014). So we focus on estimating the predictor variance in SK with a constant mean $\mu$.

Averaging the $m_i$ replications gives the average metamodel output $\overline{y}(\mathbf{x}_i)$ and average intrinsic noise $\overline{\varepsilon}(\mathbf{x}_i)$, so Eq. (1) is replaced by

$$\overline{y}(\mathbf{x}_i) = \mu + M(\mathbf{x}_i) + \overline{\varepsilon}(\mathbf{x}_i) \text{ with } \mathbf{x}_i \in \mathbb{R}^k \text{ and } i = 1, \ldots, n. \tag{2}$$

where $n$ denotes the number of so-called *old* simulated points. If we do not use CRN, then $\mathbf{\Sigma}_{\overline{\varepsilon}}$ is a diagonal matrix with main-diagonal elements $\text{Var}[\varepsilon(\mathbf{x}_i)]/m_i$. If we do use CRN and $m_i$ is a constant $m$, then $\mathbf{\Sigma}_{\overline{\varepsilon}} = \mathbf{\Sigma}_\varepsilon/m$. Notice that $\mathbf{x}_i = (x_{i;j})$ with $j = 1, \ldots, k$ and $k$ defined below Eq. (1), so $\mathbf{X}' = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ is a $k \times n$ matrix; $\mathbf{X}$ is the so-called design matrix.

To estimate $\text{Var}[\varepsilon(\mathbf{x}_i)]$, SK uses the following classic estimator that assumes $m_i > 1$:

$$s_i^2 = \frac{\sum_{r=1}^{m_i} (w_{i;r} - \overline{w}_i)^2}{m_i - 1} \text{ with } \overline{w}_i = \frac{\sum_{r=1}^{m_i} w_{i;r}}{m_i}. \tag{3}$$

Note: Goldberg et al. (1998) uses $\log s_i^2$, which may be normally (instead of $\chi^2$) distributed; also see Kamiński (2015).

Assuming CRN with a constant number of replications $m$, we may estimate $\mathbf{\Sigma}_\varepsilon$ analogously to Eq. (3):

$$s_{i;i'} = \frac{\sum_{r=1}^{m} (w_{i;r} - \overline{w}_i)(w_{i';r} - \overline{w}_{i'})}{(m-1)}, \tag{4}$$

where $i = i'$ gives $s_{i;i'} = s_i^2$. Dykstra (1970) shows that the estimated covariance matrix $\widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}} = (s_{i;i'})$ is *singular* if $m \leq n$.

We define $\overline{\mathbf{w}} = (\overline{w}_1, \ldots, \overline{w}_n)'$. SK uses a *linear* predictor $\widehat{y}(\mathbf{x}_0)$ for a *new* point $\mathbf{x}_0$:

$$\widehat{y}(\mathbf{x}_0) = \sum_{i=1}^{n} \lambda_i \overline{w}_i = \boldsymbol{\lambda}' \overline{\mathbf{w}} \tag{5}$$

with $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_n)'$.To determine the *optimal* weights $\lambda_o$, SK uses the *best linear unbiased predictor (BLUP)* criterion where 'best' means that SK minimizes the *mean squared predictor error* (MSPE) of the predictor so

$$\min_{\lambda} \text{MSPE}[\widehat{y}(\mathbf{x}_0)] = \min_{\boldsymbol{\lambda}} E[\widehat{y}(\mathbf{x}_0) - y(\mathbf{x}_0)]^2, \tag{6}$$

and *unbiased* means

$$E[\widehat{y}(\mathbf{x}_0)] = E[y(\mathbf{x}_0)]. \tag{7}$$

It can be proven that the solution of the constrained minimization problem defined by Eq. (6) and Eq. (7) implies that $\boldsymbol{\lambda}$ must satisfy the following condition with the $n$-dimensional vector $\mathbf{1} = (1, \ldots, 1)'$ :

$$\sum_{i=1}^{n} \lambda_i = \mathbf{1}'\lambda = 1. \tag{8}$$

Furthermore, it can be proven that

$$\lambda_o' = [\sigma(\mathbf{x}_0) + \mathbf{1}\frac{1 - \mathbf{1}'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1}\sigma(\mathbf{x}_0)}{\mathbf{1}'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1}\mathbf{1}}]'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1} \tag{9}$$

where $\sigma(\mathbf{x}_0)$ $(\text{cov}(y_i, y_0))$ denotes the $n$-dimensional vector with the covariances between the metamodel's $n$ *old* outputs $y_i$ and the metamodel's *new* output $y_0$. In general, the element $i$ in $\lambda_o$ decreases as the distance between the new input combination $\mathbf{x}_0$ and the old combination $\mathbf{x}_i$ increases.

Altogether, Eqs. (2), (5), and (9) imply that the optimal predictor is

$$\widehat{y}(\mathbf{x}_0) = \mu + \sigma(\mathbf{x}_0)'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1}(\overline{\mathbf{w}} - \mu\mathbf{1}). \tag{10}$$

If $\tau^2$ denotes the variance of $y$, then the MSPE of this $\widehat{y}(\mathbf{x}_0)$ can be proven to be

$$\begin{aligned} \text{MSPE}[\widehat{y}(\mathbf{x}_0)] = \tau^2 &- \sigma(\mathbf{x}_0)'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1}\sigma(\mathbf{x}_0) \\ &+ \frac{[1 - \mathbf{1}'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1}\sigma(\mathbf{x}_0)]^2}{\mathbf{1}'(\boldsymbol{\Sigma}_M + \boldsymbol{\Sigma}_{\overline{\varepsilon}})^{-1}\mathbf{1}}. \end{aligned} \tag{11}$$

Because $\widehat{y}(\mathbf{x}_0)$ is unbiased, this MSPE equals the predictor variance—which is often called the *Kriging variance*. We denote this variance at $\mathbf{x}$ by $\sigma^2(\mathbf{x})$. Obviously, Kriging gives a high MSPE in extrapolation—compared with interpolation.

We have already mentioned that $\boldsymbol{\Sigma}_M(\mathbf{x})$ in Eq. (2) must be symmetric and positive definite. Now we add that this requirement may be satisfied by many specifications of $\boldsymbol{\Sigma}_M(\mathbf{x})$. We, however, focus on the specification that is most popular in SK; namely, the *Gaussian correlation function*

$$\rho(\theta, \mathbf{h}) = \prod_{j=1}^{k} \exp\left(-\theta_j h_j^2\right) = \exp(-\sum_{j=1}^{k} \theta_j h_j^2) \tag{12}$$

where $\theta = (\theta_1, \ldots, \theta_k)'$ with $\theta_j \geq 0$, $\mathbf{h} = (h_1, \ldots, h_k)'$ and $h_j = |x_{g;j} - x_{g';j}|$ with $j = 1, \ldots, k$ and $g, g' = 0, 1, \ldots, n$ (the subscript 0 refers to the new point).

In practice, we must estimate the *Kriging parameters* $\mu$, $\mathbf{\Sigma}_M(\tau^2, \theta)$, and $\mathbf{\Sigma}_{\overline{\varepsilon}}$. For this estimation we usually select the *maximum likelihood* (ML) criterion, which gives the ML estimators (MLEs). We denote these MLEs by a 'hat'; e.g., $\widehat{\mu}$. We denote the MLEs of all the SK parameters by the vector $\widehat{\psi}$; this notation implies that the elements of $\widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}}$ are arranged into a vector. *Plugging* $\widehat{\psi}$ into Eq. (10), we obtain

$$\widehat{y}(\mathbf{x}_0, \widehat{\psi}) = \widehat{\mu} + \widehat{\sigma}(\mathbf{x}_0)'(\widehat{\mathbf{\Sigma}}_M + \widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}})^{-1}(\overline{\mathbf{w}} - \widehat{\mu}\mathbf{1}). \tag{13}$$

Obviously, this predictor is *nonlinear*. However, most publications on Kriging—including SK—ignore possible bias of this predictor and compute the *estimated MSPE*—denoted by EMSPE—by simply plugging $\widehat{\psi}$ into Eq. (11):

$$\begin{aligned} \text{EMSPE}[\widehat{y}(\mathbf{x}_0, \widehat{\psi})] = \widehat{\tau}^2 - \widehat{\sigma}(\mathbf{x}_0)'(\widehat{\mathbf{\Sigma}}_M + \widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}})^{-1}\widehat{\sigma}(\mathbf{x}_0) \\ + \frac{[1 - \mathbf{1}'(\widehat{\mathbf{\Sigma}}_M + \widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}})^{-1}\widehat{\sigma}(\mathbf{x}_0)]^2}{\mathbf{1}'(\widehat{\mathbf{\Sigma}}_M + \widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}})^{-1}\mathbf{1}}; \end{aligned} \tag{14}$$

also seeAnkenman et al. (2010, Eq. 25). For brevity's sake we denote EMSPE[$\widehat{y}(\mathbf{x}_0, \widehat{\psi})$] by $s_P^2(\mathbf{x}_0)$ where the subscript $P$ stands for plug-in.

We shall discuss various estimators of the true MSPE of $\widehat{y}(\mathbf{x}_0, \widehat{\psi})$, in the next sections. We conjecture that ignoring the randomness of $\widehat{\psi}$ tends to underestimate $\sigma^2(\mathbf{x}_0)$. Notice that Yin et al. (2009) also studies the effects that the estimation of the Kriging parameters has on the predictor variance.

There is much *software* for Kriging; see Kleijnen (2015). In our experiments we use the free R package `DiceKriging`—which is well documented in Roustant et al. (2012). However, this software does not accept CRN, so for our experiment with CRN we use the SK algorithm in Ankenman et al. (2010) for the plug-in Kriging predictor $\widehat{y}(\mathbf{x}_0, \widehat{\psi})$.

Note: We adhere to a frequentist view, but there are many publications that interpret Kriging models in a *Bayesian* way; see Yuan and Ng (2014). Our major problem with the Bayesian approach to Kriging is that we find it hard to come up with prior distributions for $\psi$, because we have little intuition about this $\psi$.

# 3 Macroreplications for M/M/1 simulation

The M/M/1 simulation model is a basic building block of many *practical* simulation models. One example of such a practical model is a model with the exponential service times replaced by a lognormal distribution so we get a M/G/1 model where the G stands for 'general'. Furthermore, we might experiment with the priority rule, the waiting-room capacity, and the balking probability. Another example is a series of single-server models such that the customers of

a preceding server move to the next server. Instead of a single server we might simulate multiple parallel servers, etc. Altogether, the M/M/1 model is the building block of supply chain models. The M/M/1 model looks misleadingly simple; actually, its I/O function is highly nonlinear for the mean steady-state waiting time—and even more nonlinear for the variance of this waiting time.

As 'the' output of the M/M/1 simulation model, we select an output that has a known analytical solution so we can easily verify the correctness of our three methods. We therefore select the *expected steady-state waiting time* (say) $E(w)$. Furthermore, we experiment only with the *traffic rate* $x$. We simulate $n = 11$ traffic rates $x_i$ $(i = 1, \ldots, 11)$. For the experimental area we select $0.10 \leq x \leq 0.90$. Furthermore, we select $x_i$ at fixed subintervals in this interval such that these subintervals have the constant length $(0.90 - 0.10)/10 = 0.08$; so $x_i$ is 0.10, 0.18, ..., 0.82, and 0.90. Notice that Kriging software normalizes the input values such that $0 \leq x_i \leq 1$; obviously, we make the normalized and the original values coincide. We obtain $m_i = m = 20$ replications—each with a runlength of 1000 customers—not starting in the empty state with zero customers waiting, but starting with the number of waiting customers equal to the steady-state mean. First we simulate the M/M/1 system without CRN; next we simulate it with CRN. We fit a SK metamodel to the resulting simulation I/O data $(x_i, w_{i;j})$ with $i = 1, \ldots, 11$, and $j = 1, \ldots, 20$. Because we experiment only with the traffic rate $x$, the number of inputs is $k = 1$ and the $k$-dimensional vector $\theta$ reduces to the single scalar $\theta_1$. Instead of a single new point $\mathbf{x}_0$, we select $T$ new traffic rates $x_t$ $(t = 1, \ldots, T)$. Actually, we select $T = 3$ new non-extrapolating values $x_t$; namely, 0.20, 0.46, and 0.80.

By definition, *macroreplications* imply that the whole experiment with the M/M/1 simulation model with its $m_i$ replications is repeated (say) $M$ times, so these macroreplications provide $M$ IID predictors $\widehat{y}(\mathbf{x}_0, \widehat{\psi}_r)$ with $r = 1, \ldots, M$; we use the subscript (local variable) $r$, which we also use for the replications; see Eq. (1). In each macroreplication we fit a SK metamodel to the I/O data of the M/M/1 simulation experiment. To generate IID predictors $\widehat{y}(\mathbf{x}_0, \widehat{\psi}_r)$, the macroreplications use nonoverlapping PRN streams.

The $M$ macroreplications give the *unbiased* estimator $s^2_{\mathrm{M}}(x_t)$ (where the subscript M—not to be confused with $M$—stands for macroreplications) of $\sigma^2(x_t)$:

$$s^2_{\mathrm{M}}(x_t) = \frac{\sum_{r=1}^{M} [\widehat{y}(x_t, \widehat{\psi}_r) - \overline{\widehat{y}}(x_t, \widehat{\psi})]^2}{M - 1} \text{ with } \overline{\widehat{y}}(x_t, \widehat{\psi}) = \frac{\sum_{r=1}^{M} \widehat{y}(x_t, \widehat{\psi}_r)}{M}. \quad (15)$$

Assuming that $\widehat{y}(x_t, \widehat{\psi}_r)$ is normally distributed, we get a $\chi^2_{M-1}$-distribution for $(M - 1)s^2_{\mathrm{M}}(x_t)/\sigma^2(x_t)$. Let $\chi^2_{M-1;\alpha}$ denote the $a$-quantile of the $\chi^2_{M-1}$-distribution, so $P[\chi^2_{M-1} \leq \chi^2_{M-1;\alpha}] = \alpha$. This gives a two-sided $(1 - \alpha)$ asymmetric *confidence interval* (CI) for $\sigma^2(x_t)$:

$$P[\frac{(M-1)s^2_{\mathrm{M}}(x_t)}{\chi^2_{M-1;1-\alpha/2}} \leq \sigma^2(x_t) \leq \frac{(M-1)s^2_{\mathrm{MC}}(x_t)}{\chi^2_{M-1;\alpha/2}}] = 1 - \alpha. \quad (16)$$

Macroreplication $r$ also gives $s^2_{\mathrm{P}}(x_t)_r$, which denotes the plug-in (symbol P)

estimator of the predictor variance in macroreplication $r$; see Eq. (14). So an unbiased estimator of $E[s^2_{\rm P}(x_t)]$ is

$$\overline{s}^2_{\rm P}(x_t) = \frac{\sum_{r=1}^{M} s^2_{\rm P}(x_t)_r}{M}.$$

The *standard error* (SE) of this $\overline{s}^2_{\rm P}(x_t)$ is

$$\text{SE}[\overline{s}^2_{\rm P}(x_t)] = \sqrt{\frac{\sum_{r=1}^{M}[s^2_{\rm P}(x_t)_r - \overline{s}^2_{\rm P}(x_t)]^2}{(M-1)M}}.$$

Together, these last two equations give a two-sided symmetric $(1-\alpha)$ CI for $E[s^2_{\rm P}(x_t)]$:

$$P[E[s^2_{\rm P}(x_t)] \in \overline{s}^2_{\rm P}(x_t) \mp t_{M-1;\alpha/2}\text{SE}[\overline{s}^2_{\rm P}(x_t)] = 1-\alpha \tag{17}$$

where we assume that the $t$-statistic is insensitive to the nonnormality of $s^2_{\rm P}(x_t)$. Because $M = 50$ we replace $t_{M-1;\alpha}$ by $z_\alpha$.

Eqs. (16) and (17) give the following results.

- If $t = 1$ so $x_t = 0.20$, then the unbiased estimator $s^2_{\rm M}(x_t)$ has the following 95% CI: [0.1470, 0.3271]×1.0E-05, whereas the CI for the biased plug-in estimator $s^2_{\rm P}(x_t)$ is 2.5843E-06 ± 1.3150E-13; these two CIs imply that $s^2_{\rm P}(x_t)$ is significantly lower.
- If $t = 2$ so $x_t = 0.46$, then $s^2_{\rm M}(x_t)$ has the CI [0.0827, 0.1840]×1.0E-03, whereas $s^2_{\rm P}(x_t)$ has 8.6835E-05 ± 2.1809E-11. So $s^2_{\rm P}(x_t)$ is again significantly lower. Notice that a higher traffic rate $x$ gives higher intrinsic noise, so the CI for $s^2_{\rm M}(x_t)$ is wider.
- If $t = 3$ so $x_t = 0.80$, then $s^2_{\rm M}(x_t)$ has the CI [0.0466, 0.1038], whereas $s^2_{\rm P}(x_t)$ has 0.0208 ± 2.2694E-06 so $s^2_{\rm P}(x_t)$ is again significantly lower.

From these three results we conclude that the plug-in estimator $s^2_{\rm P}(x_t)$ significantly underestimates the true variance.

Note: Whereas we fix the 11 'old' traffic rates $x_i$ at fixed values in all MC macroreplications, we might use *Latin hypercube sampling* (LHS) to sample 11 traffic rates $x_i$—within the ten subintervals with length 0.08—such that these traffic rates are not constant across the macroreplications in the MC experiment. This gives the following results, which suggest that this LHS increases $s^2_M(x_t)$ compared with fixed $x_i$:

1. $s^2_P(x_1)$: 6.2794E-06 ± 4.6522E-12 and $s^2_{\rm M}(x_1)$: 1.0E-04×[0.0657, 0.1462]
2. $s^2_P(x_2)$: 9.2900E-05 ± 4.5362E-11 and $s^2_{\rm M}(x_2)$: 1.0E-03×[0.1302, 0.2897]
3. $s^2_P(x_3)$: 0.0185 ± 1.4459E-06 and $s^2_{\rm M}(x_3)$: [0.0320, 0.0711].

Next we repeat the experiment *with CRN* when simulating the $n$ traffic rates $x_i$. Obviously, CRN create correlations between $w_{i;j}$ and $w_{i';j}$ with $i$, $i' = 1, \ldots, n$. These correlations are estimated through $\mathrm{s}_{i;i'}$ defined in in Eq. (4). These $\mathrm{s}_{i;i'}$ define $\widehat{\boldsymbol{\Sigma}}_{\overline{\varepsilon}}$ and affect the SK predictor $\widehat{y}_r(x_t)$ defined in Eq. (13). Substituting this $\widehat{y}_r(x_t)$ into Eq. (15) gives the estimator (say) $s^2_{\mathrm{M}}(x_t)$, which we abbreviate to $s^2_{\mathrm{CRN}}(x_t)$. We test whether the estimated predictor variances with and without CRN differ significantly. We ensure that the PRN streams in this experiment do not overlap with the PRN streams in the preceding experiment without CRN, so that we may apply an $F$-statistic (this statistic assumes that its numerator and denominator are independent). More precisely, we use $F_{M-1,M-1}(x_t) = s^2_{\mathrm{M}}(x_t)/s^2_{\mathrm{CRN}}(x_t)$, which gives a two-sided nonsymmetric CI; e.g., $M = 50$ and $\alpha = 0.05$ give a CI for $F_{M-1,M-1}$ from 0.5675 to 1.7622. This gives the following results.

- If $t = 1$, then $F_{M-1,M-1}(x_t) = s^2_{\mathrm{M}}(x_t)/s^2_{\mathrm{CRN}}(x_t) = 0.4764$ so the estimated predictor variance without CRN is significantly lower. Our explanation is that CRN requires the estimation of more parameters; namely, $s_{i;i'}$ with $i \neq i'$ (see Eq. (4)).
- If $t = 2$, then $F_{M-1,M-1}(x_t) = 0.3404$ so the estimated variance without CRN is again significantly lower.
- If $t = 3$, then $F_{M-1,M-1}(x_t) = 0.4602$ so the estimated variance without CRN is still significantly lower.

From these three results we conclude that *CRN* increases the predictor variance in SK; this conclusion agrees with Chen et al. (2012). (Nevertheless, CRN may decrease the variance of the gradient, but we do not investigate gradients.)

# 4 Distribution-free bootstrapping for M/M/1 simulation

Using $M$ macroreplications—discussed in the preceding section—is expensive in practical simulations that require much computer time for the total simulation experiment with its $\sum_{i=1}^{n} m_i$ simulation runs. We may therefore replace the method using macroreplications by *distribution-free bootstrapping* (DB)—which resamples the $m_i$ replications in a single macroreplication, as we explain next. (if there is only a single macroreplication, then the whole concept of macroreplications makes little sense; likewise, if there is a single replication so $m_i = 1$, then we may say that there are no replications—as is the case in deterministic simulation).

Like in our preceding experiments with macroreplications, we first simulate *without CRN*. Again we select $E(w)$ as the performance measure, and simulate the same $n = 11$ traffic rates $x_i$ ($i = 1, \ldots, n$), and obtain $m_i = m = 20$ replications per simulated traffic rate. Now, however, we apply DB; i.e., per simulated traffic rate $x_i$ we resample with replacement $w_{i;j}$ ($j = 1, \ldots, 20$), which gives

the bootstrapped output $w^*_{i;j}$. We use these bootstrapped $w^*_{i;r}$ to compute $\overline{w}^*_i$ and $s^{2*}_i$, which are the classic estimators of the mean and the variance; see Eq. (3) with $w$ replaced by $w^*$. We use these bootstrap means $\overline{w}^*_i$ to estimate the Kriging parameters $\mu$ and $\mathbf{\Sigma}_M$, and we use the bootstrap variances $s^{2*}_i$ to estimate the diagonal matrix $\mathbf{\Sigma}_{\overline{\varepsilon}}$. Next we compute the bootstrapped SK predictor $\widehat{y}^*(x_t)$ for the same new $x_t$ values as in the macroreplications; namely, the values 0.20, 0.46, and 0.80. We repeat this bootstrapping $B$ times; we select $B = 50$ (so $B = M$). This gives $\widehat{y}^*_b(x_t)$ with $b = 1, \ldots, B$. This gives the following DB estimator of $\sigma^2(x_t)$:

$$s^{2*}_{\mathrm{DB}}(x_t) = \frac{\sum_{b=1}^{B}[\widehat{y}^*_b(x_t) - \overline{\widehat{y}}^*(x_t)]^2}{B-1} \text{ with } \overline{\widehat{y}}^*(x_t) = \frac{\sum_{b=1}^{B}\widehat{y}^*_b(x_t)}{B}, \tag{18}$$

which is the analogue of Eq. (15).

Now we test whether there is a significant difference between $s^{2*}_{\mathrm{DB}}(x_t)$ and $s^2_{\mathrm{M}}(x_t)$ (the DB and the macroreplications estimates without CRN).

- If $t = 1$, then $F_{M-1,B-1}(x_t) = s^2_{\mathrm{M}}(x_t)/s^{2*}_{\mathrm{DB}}(x_t) = 0.2499$, which is significantly low.
- If $t = 2$, then $F_{M-1,B-1}(x_t) = 0.3566$, which is also significantly low.
- If $t = 3$, then $F_{M-1,B-1}(x_t) = 1.6748$, which is not significant; an explanation may be that the high traffic rate $x_3$ creates high intrinsic noise so the resulting $F$-statistic becomes nonsignificant.

Altogether, in two of the three new points, DB gives significantly higher estimated variances than MC does, so DB tends to give a conservative estimator.

Next we repeat the experiment *with CRN*. In the preceding section we have already observed that CRN create correlation between $w_{i;j}$ and $w_{i';j}$ with $i$, $i' = 1, \ldots, n$ and $j = 1, \ldots, m$. Because of these correlations we now resample with replacement the *vectors* $\mathbf{w}_j$ where $\mathbf{w}_j$ denotes the $n$-dimensional vector with the simulation outputs that use the CRN of replication $j$. This resampling gives $\mathbf{w}^*_j$. We use these $\mathbf{w}^*_j$ to compute the bootstrap averages $\overline{w}^*_i$ and the bootstrap covariances $s^*_{i;i'}$ where $i = i'$ gives $s^*_{i;i} = s^{2*}_i$; see $s_{i;i'}$ defined in Eq. (4). We use these $\overline{w}^*_i$ to estimate the Kriging parameters $\mu$ and $\mathbf{\Sigma}_M$, and we use $s^{2*}_{i;i'}$ to estimate $\mathbf{\Sigma}_{\overline{\varepsilon}}$; the resulting estimated bootstrap covariance matrix $\widehat{\mathbf{\Sigma}}^*_{\overline{\varepsilon}}$ is not singular as $m = 20 > n = 11$. We test whether $s^{2*}_{\mathrm{DB}}(x_t)$ (DB with CRN) and $s^2_{\mathrm{CRN}}(x_t)$ (macroreplications with CRN) differ significantly. We obtain the following results.

- If $t = 1$, then $F_{M-1,B-1}(x_t) = s^2_{\mathrm{M}}(x_t)/s^{2*}_{\mathrm{DB}}(x_t) = 0.6806$, which is not significant.
- If $t = 2$, then $F_{M-1,B-1}(x_t) = 0.5890$, which is again not significant.
- If $t = 3$, then $F_{M-1,B-1}(t) = 1.\ 1900$, which is also nonsignificant.

Altogether, all three new points gives nonsignificant differences between macroreplications and DB when CRN are applied. Actually, CRN are the default option in many discrete-event simulation software packages (e.g., Arena). So we conclude that DB is a useful method for estimating the variance of the SK predictor when CRN are applied.

Finally, we test the difference between $s^{2*}_{\text{DB}}(x_t)$ and $s^{2*}_{\text{CRN}}(x_t)$, which denote DB without and with CRN, respectively. For this test we use $F_{B-1,B-1}(x_t) = s^{2*}_{\text{DB}}(x_t)/s^{2*}_{\text{CRN}}(x_t)$. We obtain the following results.

- If $t = 1$, then $F_{B-1,B-1}(x_t) = 1.2975$ so the bootstrap SK predictor variance with and without CRN do not differ significantly (unlike we found in our macroreplications).
- If $t = 2$, then $F_{B-1,B-1}(x_t) = 0.5623$ so the bootstrap variance without CRN is significantly lower (like in our macroreplications).
- If $t = 3$, then $F_{B-1,B-1}(t) = 0.3270$ so the bootstrap variance without CRN is again significantly lower.

Altogether, we conclude that the variance of the SK predictor with CRN tends to be higher than without CRN, like Chen et al. (2012) also concluded.

# 5 Parametric bootstrapping for M/M/1 simulation

Next we replace DB by the following *parametric bootstrap* (PB). Again we first simulate the M/M/1 system without CRN, selecting the performance measure $E(w)$, $n = 11$ traffic rates $x_i$, and $m = 20$ replications per simulated traffic rate. This simulation experiment gives the simulation outputs $w_{i;j}$ with $i = 1, \ldots, 11$ and $j = 1, \ldots, 20$. PB means that we independently sample $w^*_{i;j}$ (the bootstrap output) $m$ times from the $n$-variate Normal distribution $\text{N}_n(\widehat{\mu}\mathbf{1}_n, \widehat{\mathbf{\Sigma}}_M + \widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}})$ where $\widehat{\mu}$ and $\widehat{\mathbf{\Sigma}}_M$ are computed through the SK algorithm using $\overline{w}_i$, and $\widehat{\mathbf{\Sigma}}_{\overline{\varepsilon}}$ denotes the $n \times n$ diagonal matrix with the elements $s^2_i$. We continue as in DB; i.e., we use these $w^*_{i;j}$ to compute $\overline{w}^*_i$ and $s^{2*}_i$, followed by $\widehat{\mu}^*$, $\widehat{\mathbf{\Sigma}}^*_M$, and $\widehat{\mathbf{\Sigma}}^*_{\overline{\varepsilon}}$. We use these $\widehat{\mu}^*$, $\widehat{\mathbf{\Sigma}}^*_M$, and $\widehat{\mathbf{\Sigma}}^*_{\overline{\varepsilon}}$ to compute the SK predictor $\widehat{y}^*(x_t)$ where $t = 1, 2, 3$ gives $x_t = 0.20, 0.46, 0.80$. We repeat this bootstrapping $B$ times. We compute $s^{2*}_{\text{PB}}(x_t)$—the PB estimator of $\sigma^2(x_t)$—through Eq. (18) where we replace the subscript DB by PB. First we test whether $s^{2*}_{\text{PB}}(x_t)$ and $s^{2*}_{\text{DB}}(x_t)$ (DB estimate without CRN) differ significantly. For this test we use $F_{B-1,B-1}(x_t) = s^{2*}_{\text{DB}}(x_t)/s^{2*}_{\text{PB}}(x_t)$. We obtain the following results.

- If $t = 1$, then $F_{B-1,B-1}(x_t) = 6.9489\text{E-}06$, which is an extremely low significant value; i.e., PB gives a much higher variance estimate.
- If $t = 2$, then $F_{B-1,B-1}(x_t) = 2.5460\text{E-}04$, which is also extremely low.

- If $t = 3$, then $F_{B-1,B-1}(x_t) = 0.0275$, which is significantly low.

So in all three new points, PB gives a much higher bootstrap variance than DB does. To explain this result, we point out that PB samples from a normal distribution for $\widehat{y}^*(x_t)$ with positive probabilities for values that are either lower or higher that the values sampled in DB (any normal distribution has tails that theoretically range from $-\infty$ to $\infty$).

Next we repeat the whole experiment *with CRN*; i.e., we sample $\mathbf{w}_j$ ($j = 1, \ldots, m$) from $\mathrm{N}_n(\widehat{\mu}\mathbf{1}_n, \widehat{\mathbf{\Sigma}}_M + \widehat{\mathbf{\Sigma}}_{\varepsilon})$ where CRN imply that $\widehat{\mathbf{\Sigma}}_{\varepsilon}$ is not a diagonal matrix but has the elements $s_{i;i'}$. This sampling gives the bootstrapped outputs $\mathbf{w}_j^*$, which give the averages $\overline{w}_i^*$ and the estimated covariances $s_{i;i'}^*$, etc. We test whether PB gives estimated predictor variances with CRN that differ significantly from the DB estimates with CRN. For this test we use $F_{B-1,B-1}(x_t) = s_{\mathrm{DB}}^{2*}(x_t)/s_{\mathrm{PB}}^{2*}(x_t)$. and obtain the following results.

- If $t = 1$, then $F_{B-1,B-1}(x_t) = 1.361\text{E-}05$, which is an extremely low significant value.
- If $t = 2$, then $F_{B-1,B-1}(x_t) = 0.0011$, which is again extremely low.
- If $t = 3$, then $F_{B-1,B-1}(x_t) = 0.2105$, which is again significantly low.

So in these three new points, PB again gives a much higher estimated predictor variances than DB does. (We gave a similar conclusion for simulation without CRN).

Finally, we test whether PB gives estimates of the predictor variances with and without CRN that differ significantly. For this test we use $F_{B-1,B-1}(x_t) = s_{\mathrm{PB}}^{2*}(x_t)/s_{\mathrm{CRN}}^{2*}(x_t)$.

- If $t = 1$, then $F_{B-1,B-1}(x_t) = 2.5426$ so the estimated predictor variance without CRN is significantly higher.
- If $t = 2$, then $F_{B-1,B-1}(x_t) = 2.5278$ so the estimated variance without CRN is again significantly higher.
- If $t = 3$, then $F_{B-1,B-1}(x_t) = 2.5053$ so the estimated variance without CRN is still significantly higher.

Altogether, for all three new points the use of CRN reduces the PB estimates of the predictor variances. This result conflicts with our previous conclusions.

# 6 Conclusions

In this paper we investigated the plug-in estimator $s_P^2(\mathbf{x}_0)$ for the variance of $\widehat{y}(\mathbf{x}_0, \widehat{\psi})$, which denotes the SK predictor for the simulation output at the simulation input combination $\mathbf{x}_0$ when using $\widehat{\psi}$ (the plug-in estimates of the Kriging parameters $\psi$). For this investigation we experimented with the M/M/1 simulation

model—with and without CRN. Our macroreplications suggested that $s^2_{\mathrm{P}}(x_t)$ significantly underestimates the true variance; furthermore, CRN increase the true predictor variance. In practice, we might save on computer time, replacing macroreplications by DB. In our M/M/1 simulation experiment without CRN, DB gives significantly higher estimated variances than our macroreplications do; i.e., DB tends to give a conservative estimator. With CRN, however, DB and macroreplications do not give significantly different estimates. PB gives a much higher bootstrap variance than DB does—whether or not CRN are applied. So, altogether we recommend DB for the correct estimation of the predictor variance of SK in simulation experiments with CRN.

In future research we may investigate the consequences of the statistical properties of SK, when using SK for either optimization or sensitivity analysis. For deterministic simulation, Mehdad and Kleijnen (2015) has already investigated these consequences.

# Acknowledgment

We thank Barry Nelson (Northwestern University) for suggesting LHS to select the old input combinations.

# References

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

Barton, R., Nelson, B., and Xie, W. (2014). Quantifying input uncertainty via simulation confidence intervals. *INFORMS Journal on Computing*, 26(1):74–87.

Bekki, J., Chen, X., and Batur, D. (2014). Steady-state quantile parameter estimation: An empirical comparison of stochastic kriging and quantile regression. In *Proceedings of the 2014 Winter Simulation Conference (WSC)*, pages 3880–3891.

Chen, X., Ankenman, B., and Nelson, B. (2012). The effects of common random numbers on stochastic Kriging metamodels. *ACM Transactions on Modeling and Computer Simulation*, 22(2):7:1–7:20.

Chen, X. and Kim, K.-K. (2014). Stochastic Kriging with biased sample estimates. *ACM Transactions on Modeling and Computer Simulation*, 24(2):8:1–8:23.

Dykstra, R. (1970). Establishing the positive definiteness of the sample covariance matrix. *The Annals of Mathematical Statistics*, 41(6):2153–2154.

Goldberg, P., Williams, C., and Bishop, C. (1998). Regression with input-dependent noise: a Gaussian process treatment. In Jordan, M., Kearns, M.,

and Solla, S., editors, *Advances in Neural Information Processing Systems 10*, pages 493–499.

Kamiński, B. (2015). A method for updating of stochastic Kriging metamodels. *European Journal of Operational Research*. accepted.

Kleijnen, J. (2015). *Design and Analysis of Simulation Experiments*. Springer-Verlag, second edition.

Law, A. (2015). *Simulation Modeling and Analysis*. McGraw-Hill, 5th edition.

Mehdad, E. and Kleijnen, J. (2015). Classic Kriging versus Kriging with bootstrapping or conditional simulation: classic Krigings robust confidence intervals and optimization. *Journal of Operational Research Society*. Advanced Online Publication.

Plumlee, M. and Tuo, R. (2014). Building accurate emulators for stochastic simulations via quantile Kriging. *Technometrics*, 56(4):466–473.

Qu, H. and Fu, M. (2014). Gradient extrapolated stochastic Kriging. *ACM Transactions on Modeling and Computer Simulation*, 24(4):23:1–23:25.

Roustant, O., Ginsbourger, D., and Deville, Y. (2012). DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. *Journal of Statistical Software*, 51(1):1–55.

Sun, L., Hong, L., and Hu, Z. (2014). Balancing exploitation and exploration in discrete optimization via simulation through a Gaussian process-based search. *Operations Research*, 62(6):1416–1438.

Tajbakhsh, S., del Castillo, E., and Rosenberger, J. (2014). A bayesian approach to sequential optimization based on computer experiments. *Quality and Reliability Engineering International*. in press.

Thiart, C., Ngwenya, M., and Haines, L. (2014). Investigating 'optimal' kriging variance estimation using an analytic and a bootstrap approach. *Journal of the Southern African Institute of Mining and Metallurgy*, 114:613–619.

Yin, J., Ng, S., and Ng, K. (2009). A study on the effects of parameter estimation on kriging model's prediction error in stochastic simulations. In *Proceedings of the 2009 Winter Simulation Conference (WSC)*, pages 674–685.

Yuan, J. and Ng, S. (2014). An integrated approach to stochastic computer model calibration, validation and prediction. *ACM Transactions on Modeling and Computer Simulation*. in press.