The Correct Kriging Variance Estimated by Bootstrapping

den Hertog, D.; Kleijnen, J.P.C.; Siem, A.Y.D.

Publication date:
2004

No. 2004–46

THE CORRECT KRIGING VARIANCE ESTIMATED BY BOOTSTRAPPING

By D. den Hertog, J.P.C. Kleijnen, A.Y.D. Siem

May 2004

ISSN 0924-7815

The correct Kriging variance estimated by bootstrapping

D. den Hertog* J.P.C. Kleijnen‡ A.Y.D. Siem§

4th May 2004

Abstract

The classic Kriging variance formula is widely used in geostatistics and in the design and analysis of computer experiments. This paper proves that this formula is wrong. Furthermore, it shows that the formula underestimates the Kriging variance in expectation. The paper develops parametric bootstrapping to estimate the Kriging variance. The new method is tested on several artificial examples and a real-life case study. These results demonstrate that the classic formula underestimates the true Kriging variance.

Keywords: Kriging, Kriging variance, bootstrapping, design and analysis of computer experiments (DACE), Monte Carlo, global optimization, black-box optimization

1 Introduction

Kriging is an interpolation technique that was originally invented in the field of geostatistics; see Cressie (1991). Next, Sacks, Welch, Mitchell and Wynn (1989) applied Kriging to the design and analysis of computer experiments (DACE). Since then, many others followed; see Jones, Schonlau and Welch (1997), Jones (2001), Koehler and Owen

*Department of Econometrics and Operations Research/ Center for Economic Research (CentER), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, Phone:+31 13 4662122, Fax:+31 13 4663280, E-mail: d.denhertog@uvt.nl

‡Department of Information Systems and Management/ Center for Economic Research (CentER), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, Phone:+31 13 4662029, Fax:+31 13 4663377, E-mail: kleijnen@uvt.nl

§Department of Econometrics and Operations Research/ Center for Economic Research (CentER), Tilburg University, P.O. Box 90153, 5000 LE Tilburg, The Netherlands, Phone:+31 13 4663254, Fax:+31 13 4663280, E-mail: a.y.d.siem@uvt.nl

(1996), Santner, Williams and Notz (2003) and Stehouwer and den Hertog (1999). In DACE, Kriging models are used as *response surface models*, which are also called *meta-models*, *compact models* or *surrogates*; i.e. Kriging models the input/output behavior of the underlying simulation model, which is treated as a black box.

The classic Kriging variance formula is used for three different goals. First, it is used to select new input design points to obtain better Kriging models; see Sacks, Welch, Mitchell and Wynn (1989). In Kleijnen and van Beers (2004) this approach is called application-driven sequential design of experiments, but they use a type of cross-validation, instead of the classic Kriging variance. Also in Jin and Chen (2002) such an approach is followed. Second, the formula is used for the global optimization of time-consuming computer simulations (black-box functions), namely to select new input design points to find the global optimum of the underlying computer simulation model; see Booker, Dennis, Frank, Serafini, Torczon and Trosset (1999), Cox and John (1997), Sasena, Papalambros and Goovaerts (2002) and Schonlau, Welch and Jones (1998). An overview of these methods is given in Jones (2001). Third, the Kriging variance can be used as a quality measure of a Kriging model since it quantifies the accuracy of the prediction. This can be used e.g. as a criterion for the number of design points.

In this paper we show that the Kriging variance formula used in the literature (see e.g. Cressie (1991), Jones (2001) and Sacks, Welch, Mitchell and Wynn (1989)) is wrong, because it neglects the fact that certain correlation parameters (discussed in Section 2) are *estimated*. Cressie (1991, p. 127) mentions that the classic variance formula is expected to underestimate the true variance. Indeed, we show that it is an underestimator in expectation. Furthermore, we present a bootstrap method to estimate the correct Kriging variance. For a general discussion of bootstrapping we refer to Efron and Tibshirani (1993). We apply our bootstrap method to both some artificial examples and a real-life case study. We will see that the difference between the classic and the bootstrapped Kriging variance can be very large. This is especially the case where the classic Kriging variance is large. Because of the wide application of the Kriging variance, we expect that our method may have substantial impact on the methods mentioned above.

This paper is organized as follows. In Section 2 we summarize some classic theory on Kriging models, including the Kriging variance formula. In Section 3 we show what is wrong with this formula. In Section 4 we present our new (bootstrap) method to estimate the correct Kriging variance. In Section 5 we apply this method to several artificial, academic examples. In Section 6 we treat a practical real-life case study from Sacks, Welch, Mitchell and Wynn (1989). Finally, in Section 7 we summarize our conclusions

and give recommendations for further research.

2 Kriging models

In this section we summarize some Kriging theory according to Sacks, Welch, Mitchell and Wynn (1989). The response function $y(x)$ is treated as a realization of a stochastic process $Y(x)$, where x denotes the d-dimensional input variable. This stochastic process is assumed to consist of a regression part and a stochastic part:

$$Y(x) = \sum_{j=0}^{k} \beta_j f_j(x) + Z(x), \tag{1}$$

where $k+1$ is the number of regression functions including $f_0(x) \equiv 1$. Often, the regression functions f_j are left out except for $f_0(x)$, because they do not yield better Kriging models. The stochastic part $Z(x)$ is assumed to have zero mean and constant process variance (say) σ^2. The covariance between $Z(w)$ and $Z(x)$, with w and x elements of the input space, is given by

$$V(w, x) = \sigma^2 R(w, x),$$

where $R(w, x)$ denotes the correlation between $Z(w)$ and $Z(x)$. Given is a set of computer simulation input data $S = [x^1, ..., x^n]^T$ and a set of corresponding output data $y_s = [y(x^1), ..., y(x^n)]^T$. We assume y_s is a realization of the stochastic vector $Y_s = [Y(x^1), ..., Y(x^n)]^T$, defined by (1). Further, we assume a scalar output, as most of the Kriging literature does.

Now consider the linear predictor

$$\hat{y}(x) = c^T(x) y_s.$$

Kriging chooses these weights $c(x)$ such that they minimize

$$\text{MSE}[\hat{y}(x)] = E[c^T(x) Y_s - Y(x)]^2 \tag{2}$$

under the constraint

$$E[c^T(x) Y_s] = E[Y(x)]; \tag{3}$$

in other words, $c(x)$ gives the so-called "Best Linear Unbiased Predictor" (BLUP).

Before we proceed, we introduce some further notation. We write

$$f(x) = [f_0(x), \ldots, f_k(x)]^T$$

for the $k+1$ regression functions in (1), and

$$F = \begin{bmatrix} f^T(x^1) \\ \vdots \\ f^T(x^n) \end{bmatrix} \tag{4}$$

for the values of these regression functions in the n design points. Furthermore, let R be the correlation matrix with elements

$$R_{ij} = R(x^i, x^j), \text{ for } i = 1, \ldots, n \text{ and } j = 1, \ldots, n;$$

i.e., R_{ij} is the correlation between $Z(x^i)$ and $Z(x^j)$. Let

$$r(x) = \left[R(x^1, x), \ldots, R(x^n, x)\right]^T \tag{5}$$

be the vector with correlations between $Z(x^i)$ and $Z(x)$.

Classic Kriging assumes that $c(x)$ is *independent* of the output data. Then we can rewrite the MSE in (2) as (see Santner, Williams and Notz (2003))

$$\text{MSE}[\hat{y}(x)] = \sigma^2[1 + c^T(x)Rc(x) - 2c^T(x)r(x)]. \tag{6}$$

Under the same assumption the constraint (3) can be rewritten as (see again Santner, Williams and Notz (2003))

$$F^T c(x) = f(x). \tag{7}$$

To minimize the MSE in (6) with respect to $c(x)$ under the constraint (7), Lagrange multipliers $\lambda(x)$ are used. This gives the following system of equations:

$$\begin{bmatrix} 0 & F^T \\ F & R \end{bmatrix} \begin{bmatrix} \lambda(x) \\ c(x) \end{bmatrix} = \begin{bmatrix} f(x) \\ r(x) \end{bmatrix}.$$

Solving this system of equations for $c(x)$ and $\lambda(x)$ gives

$$\begin{aligned} \lambda(x) &= (F^T R^{-1} F)^{-1}(F^T R^{-1} r(x) - f(x)) \\ c(x) &= R^{-1}(r(x) - F\lambda(x)), \end{aligned} \tag{8}$$

which yields the Kriging predictor:

$$\begin{aligned} \hat{y}(x) &= c^T(x) y_s \\ &= f^T(x)\hat{\beta} + r^T(x)R^{-1}(y_s - F\hat{\beta}), \end{aligned} \tag{9}$$

where

$$\hat{\beta} = (F^T R^{-1} F)^{-1} F^T R^{-1} y_s \tag{10}$$

is the generalized least-squares (GLS) estimate of β in (1).

The MSE of the predictor–also known as the *Kriging variance*–becomes (see also Lophaven, Nielsen and Sondergaard (2002)):

$$\mathrm{MSE}[\hat{y}(x)] = \sigma^2(1 + u^T(x)(F^T R^{-1} F)^{-1} u(x) - r^T(x) R^{-1} r(x)), \tag{11}$$

where $u(x) = F^T R^{-1} r(x) - f(x)$. Note that the Kriging variance is in fact a Mean Squared Error.

Until now, we have not discussed the form of the correlation function $R(w, x)$. Most publications assume that the correlation structure is *stationary*; i.e. $R(w, x) = R(w - x)$. Usually a parametric family of correlation functions is chosen. A popular choice is the exponential family

$$R^{\theta,p}(w, x) = \prod_{j=1}^{d} \exp(-\theta_j |w_j - x_j|^{p_j}). \tag{12}$$

In this paper, we will use (12) with $p_j = 2$, as done in Sacks, Welch, Mitchell and Wynn (1989); then (12) is called the Gaussian correlation function.

Furthermore, we assume that the stochastic process $Z(x)$ is Gaussian. Then, its log likelihood is a function of the process variance σ^2, the regression parameters β, and the correlation parameters θ. The maximum likelihood estimator (MLE) $\hat{\beta}$ of β equals the GLS estimator, and is given by (10); the MLE $\hat{\sigma}^2$ of σ^2 is given by

$$\hat{\sigma}^2 = \frac{1}{n}(y_s - F\hat{\beta})^T R^{-1}(y_s - F\hat{\beta}). \tag{13}$$

To find the MLE $\hat{\theta}$ of θ, we should solve (see Sacks, Welch, Mitchell and Wynn (1989))

$$\min_{\theta} |R|^{1/n} \hat{\sigma}^2. \tag{14}$$

Solving (14) is achieved by some numerical optimization procedure; we use the Matlab toolbox DACE provided by Lophaven, Nielsen and Sondergaard (2002).

3 The classic Kriging variance formula

The derivation of the MSE in (6) assumed that the weight vector $c(x)$ does not depend on the output data vector. Actually, this assumption is false: $c(x)$ *does* depend on the data, namely on R, see (8). Given the chosen Gaussian correlation family $R^{\theta,p}(w, x)$–see (12) with $p_j = 2$–this correlation structure is parameterized by θ. This θ is estimated by $\hat{\theta}$ via (14), so it depends on the output data.

Because $c(x)$ depends on y_s, the reasoning in Section 2 fails at equations (6) and (7). We do not know the accuracy of the approximation in (6). In the literature, (9) is called the 'Best Linear Unbiased Predictor'. However, this predictor is neither linear nor

unbiased. Therefore, Santner, Williams and Notz (2003) calls (9) the Empirical Best Unbiased Linear Predictor (EBLUP). Also the final expression for the Kriging variance (11) does not hold anymore.

It seems difficult to evaluate the magnitude of the approximation error. As we said in Section 1, Cressie (1991) expects that (11) is a lower bound of the true Kriging variance, but no arguments are given. We present the following explanation. A well-known equation in mathematical statistics is

$$\mathrm{var}(W) = \mathrm{E}_V\left[\mathrm{var}(W|V)\right] + \mathrm{var}_V\left[\mathrm{E}(W|V)\right], \tag{15}$$

where V and W are stochastic variables. Now we take $V = \hat{\theta}$ and $W = \hat{y} - Y(x)$. Substitution into (15) gives

$$\mathrm{var}(\hat{y} - Y(x)) = \mathrm{E}_{\hat{\theta}}\left[\mathrm{var}(\hat{y} - Y(x)|\hat{\theta})\right] + \mathrm{var}_{\hat{\theta}}\left[\mathrm{E}(\hat{y} - Y(x)|\hat{\theta})\right]. \tag{16}$$

Note that $\mathrm{MSE}[\hat{y}(x)] = \mathrm{var}(\hat{y} - Y(x)) + (E[\hat{y} - Y(x)])^2$ is the true Kriging variance and that $\mathrm{MSE}[\hat{y}(x)|\hat{\theta}] = \mathrm{var}(\hat{y} - Y(x)|\hat{\theta}) + \left(E[\hat{y} - Y(x)|\hat{\theta}]\right)^2$ is the classic Kriging variance. Since the second term on the righthand side in (16) is positive, we get

$$\begin{aligned} \mathrm{MSE}[\hat{y}(x)] &= \mathrm{var}(\hat{y} - Y(x)) + (\mathrm{E}[\hat{y} - Y(x)])^2 \\ &\geq \mathrm{E}_{\hat{\theta}}\left[\mathrm{var}(\hat{y} - Y(x)|\hat{\theta})\right] \\ &= \mathrm{E}_{\hat{\theta}}\left[\mathrm{MSE}[\hat{y}(x)|\hat{\theta}]\right], \end{aligned}$$

where in the last step we used the fact that $\hat{y}(x)$ is unbiased if $\hat{\theta}$ is known. So, the "average" Kriging variance is indeed an underestimator of the true Kriging variance. Note that there may be realizations of $\hat{\theta}$ such that the Kriging variance is not an underestimator for the true Kriging variance.

4 Bootstrap Kriging variance

Parametric bootstrapping is a well-known method to estimate the distribution of intricate functions of stochastic variables or functions with parameterized distribution; see Efron and Tibshirani (1993). In our case, we want to estimate the distribution of the prediction error $\hat{y} - Y(x)$ to estimate the value of $\mathrm{MSE}[\hat{y}(x)]$.

Note that the distribution type of Y_s and $Y(x)$ are known: Y_s is multivariate Gaussian, and $Y(x)$ is a Gaussian process. Therefore we apply parametric bootstrapping–not distribution-free bootstrapping. However, the parameters (namely the means and covariances of $Y(x)$) are unknown.

To estimate these parameters, we first select a parametric family of correlation functions; in our case this is the Gaussian family given by (12) with $p_j = 2$. Next, we estimate the family's parameters θ, the mean β, and the process variance σ^2 from the input/output data $(x^i, y(x^i))$. In Kriging this is usually done by using the maximum likelihood criterion, which gives $\hat{\theta}$, $\hat{\beta}$ and $\hat{\sigma}^2$; see (10), (13) and (14).

Parametric bootstrapping assumes that (12) with $\theta = \hat{\theta}$ is the correct correlation function, that $\hat{\beta}$ is the true mean, and that $\hat{\sigma}^2$ is the true variance of the stochastic process $Y(x)$. Given these estimated mean and covariance of the Gaussian process $Y(x)$, the distribution from which the bootstrap draws values (say) y^*, is known. This bootstrap is repeated B times, which gives y_b^* with $b = 1, \ldots, B$.

Since MSE$[\hat{y}(x)]$ is a function of x, which is a continuous variable, we cannot simulate MSE$[\hat{y}(x)]$ for *all* x in the design space. We can proceed in three ways, which we present in the next three subsections.

4.1 A fixed test set

Suppose we know beforehand that we want to estimate the value of MSE in a finite set of test values $x_t^1, ..., x_t^{n_t}$, for which we will estimate the value of the MSE. Then, we sample y^* from a multivariate normal distribution

$$y^* \sim \mathcal{N}_{n+n_t}(F_M\hat{\beta}, \hat{\sigma}^2 R(\hat{\theta})), \tag{17}$$

where

$$F_M = \begin{bmatrix} f^T(x^1) \\ \vdots \\ f^T(x^n) \\ f^T(x_t^1) \\ \vdots \\ f^T(x_t^{n_t}) \end{bmatrix},$$

which extends (4). In other words, we sample in the n "old" points $x^1, \ldots, x^n$ and in the n_t "new" points $x_t^1, \ldots, x_t^{n_t}$ *simultaneously*, because all "old" and "new" data points are assumed to be a realization of the *same* Gaussian stochastic process (1) so they are correlated.

Next, we repeat the sampling from (17) B times (as mentioned above), where the b-th sample consists of "old" input/output data $y_{s;b}^* = [y_b^*(x^1), \ldots, y_b^*(x^n)]^T$ and "new" input/output data $y_{t;b}^* = [y_b^*(x_t^1), \ldots, y_b^*(x_t^{n_t})]^T$. Based on each "old" dataset $y_{s;b}^*$, we estimate β, σ^2 and θ. The estimates $\hat{\beta}$, $\hat{\sigma}^2$ and $\hat{\theta}$ determine a Kriging model based on

$y^*_{s;b}$. Using this model, we predict the output values in the "new" input data points, and calculate the squared errors in these test points. The average, based on B bootstrap realizations of these squared errors per input point x, is an estimator of MSE$[\hat{y}(x)]$. Obviously, in the "old" points MSE$[\hat{y}(x)]$ is zero. More formally, from the B samples we compute the estimate $\widehat{\text{MSE}}[\hat{y}^*(x)]$ for $x = x_t^1,...,x_t^{n_t}$ as follows:

$$\widehat{\text{MSE}}[\hat{y}^*(x_t^j)] = \frac{1}{B}\sum_{b=1}^{B}\left(\hat{y}_b^*(x_t^j) - y_b^*(x_t^j)\right)^2,$$

where $\hat{y}_b^*(x_t^j)$ is the value of the Kriging model in point x_t^j, fitted with the "old" input/output dataset of the b-th sample: $y^*_{s;b} = [y_b^*(x^1), \ldots, y_b^*(x^n)]^T$ and $y_b^*(x_t^j)$ is the j-th element of the "new" input/output dataset of the b-th sample: $y^*_{t;b} = [y_b^*(x_t^1), \ldots, y_b^*(x_t^{n_t})]^T$. We summarize our bootstrap procedure as follows:

Algorithm:

Estimate the distribution of $Y(x)$ from the n original data points.

Repeat B times

Sample $Y^*(x)$ in the n "old" data points and in the n_t "new" test points simultaneously.

Fit a Kriging model from the n bootstrapped "old" data points.

Calculate the Kriging predictions in the n_t "new" test points.

Calculate the squared prediction error in the test points.

End

For all n_t test points

Calculate the sample mean of the squared 'prediction errors' in the test point.

End

In practice, we might omit finding the MLE's $\hat{\beta}_b^*$, $\hat{\sigma}_b^{2*}$, and $\hat{\theta}_b^*$ $(b = 1, \ldots, B)$ of the B Kriging models, and simply take the MLE's of the original data, $\hat{\beta}$, $\hat{\sigma}^2$, and $\hat{\theta}$, instead. The bootstrap MLE's will not differ much from the original MLE. This saves computation time.

To demonstrate one iteration of this algorithm, we present an example that shows how one bootstrap sample may look. We take the test function $f_1 : [0, 10] \mapsto \mathbb{R}$ and $f_1(x) = -0.0579x^4 + 1.11x^3 - 6.845x^2 + 14.1071x + 2$; see Figure 3. From this function we generate a dataset of $n = 4$ equidistant input points and the corresponding output values. We compute the MLE of the parameters β, σ^2, and θ of the Gaussian process

with the Matlab Toolbox DACE, which gives $\hat{\beta} = 1.2114$, $\hat{\sigma}^2 = 48.7939$ and $\hat{\theta} = 1.0800$. These MLE estimates fix the parameters of the underlying Gaussian process, which is to be bootstrapped. Then we sample $Y^*(x)$. This sample is represented by the balls in Figure 1. Furthermore, we estimate a Kriging model from the bootstrapped n "old" data points; see the solid line in Figure 1. With this information we can calculate the prediction error in each of the test points.

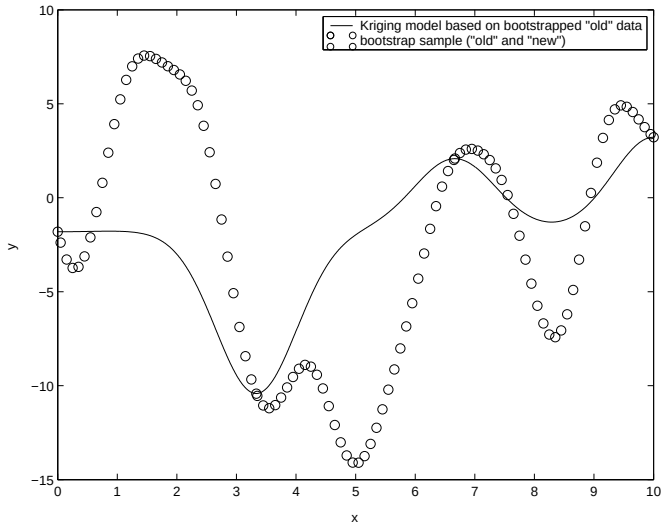

Figure 1: Example of one Bootstrap sample and the corresponding Kriging model for f_1.

4.2 A variable test set

Suppose that we do not know beforehand in which points we shall estimate the Kriging variance. Or suppose that we know the estimated Kriging variance for some input data, and later on we want to estimate the variance in other points too. Then, it is still possible to bootstrap the Kriging variance in these points provided that the bootstrapped data is saved in the computer's memory. This is necessary because the values that we wish to bootstrap are correlated with the data already bootstrapped (both the "old" and the "new"). More precisely, if we want to bootstrap from $Y_2 = [Y(x^{n_{t_1}+1}), \ldots, Y(x^{n_{t_2}})]^T$ and we already have bootstrapped the values $y^*(x^1), \ldots, y^*(x^{n_{t_1}})$ from $Y_1 = [Y(x^1), \ldots, Y(x^{n_{t_1}})]^T$, then we must take these realizations into account when bootstrapping Y_2.

Let $[Y_1^T, Y_2^T]^T$ be multivariate Gaussian distributed with mean:

$$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$$

and covariance:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}.$$

Let $y_1^* = [y^*(x^1), \ldots, y^*(x^{n_{t_1}})]^T$ be a bootstrapped realization of Y_1. Then (see e.g. Mittelhammer (1996)), the *conditional* distribution of Y_2 given y_1^* is as follows:

$$Y_2|y^*(x^1), \ldots, y^*(x^{n_{t_1}}) \sim \mathcal{N}(\mu_2 + \Sigma_{12}^T \Sigma_{11}^{-1}(y_1^* - \mu_1), \Sigma_{22} - \Sigma_{12}^T \Sigma_{11}^{-1} \Sigma_{12}) \qquad (18)$$

We summarize our procedure for this test set as follows:

Algorithm:

Estimate the distribution of $Y(x)$ from the n original data points.

Repeat B times

Sample $Y^*(x)$ in $n_{t_2} - n_{t_1}$ "new" input data points, given y_b^* in the n_{t_1} bootstrapped points and the n "old" data points.

Fit a Kriging model from the n "old" bootstrapped data points.

Calculate the Kriging predictions in the $n_{t_2} - n_{t_1}$ "new" input test points.

Calculate the squared prediction error in the new input test points.

End

For all $n_{t_2} - n_{t_1}$ new test points

Calculate the sample mean of the 'prediction errors' in the test point.

End

4.3 Adding new points one-at-a-time

Suppose we do not know beforehand in which points we want to estimate the Kriging variance and we want to add points one at a time: This happens e.g. if we are interested in finding the x for which the Kriging variance is maximal. Then we could use the approach in Subsection 4.2. However, this method becomes very time-consuming as the number of test points gets large, because of the calculation of Σ_{11}^{-1} in (18), which grows with every iteration. Therefore, we now estimate the Kriging variance in an arbitrary test point–independently of estimated Kriging variances in other test points.

Let x_t^0 be a test point, in which we want to estimate the Kriging variance. We sample $Y^*(x_t^0)|y(x^1),\ldots,y(x^n)$ (see (18)); i.e. we sample from

$$y^*(x_t^0)|y(x^1),\ldots,y(x^n) \sim \mathcal{N}(f^T(x_t^0)\hat{\beta}+r_{\hat{\theta}}^T(x_t^0)R^{-1}(\hat{\theta})(y_s^*-F\hat{\beta}), \hat{\sigma}^2-\hat{\sigma}^2 r_{\hat{\theta}}^T(x_t^0)R^{-1}(\hat{\theta})r_{\hat{\theta}}(x_t^0))), \tag{19}$$

where $r_{\hat{\theta}}$ is as in (5).

We summarize this procedure as follows:

Algorithm:

Estimate the distribution of $Y(x)$ from the n original data points.

Repeat B times

Sample $Y^*(x)$ in the n "old" input points.

Fit a Kriging model from these n "old" input points.

End

For all test points x_t^0 (not necessarily fixed beforehand)

Repeat B times

Sample $Y^*(x_t^0)$, given y_b^* in the n bootstrapped "old" input points using (19).

Calculate the Kriging prediction in x_t^0.

Calculate the prediction error in x_t^0.

End

Calculate the sample mean of the 'prediction errors' in the test point.

End

This algorithm has the advantage that we do not have to save the information on other test points. Furthermore, we do not have to calculate Σ_{11}^{-1} repeatedly. This saves computation time, and makes our procedure more applicable in practice.

A drawback of this approach is that the bootstrapped Kriging variances are computed separately. Consequently, we obtain bumpy plots for the bootstrapped Kriging variance; see Figure 2. But, by using confidence intervals (see Subsection 5.2), we can still control the accuracy of the bootstrapped Kriging variances.

5 Artificial Examples

5.1 Selecting four examples

We perform bootstrap procedures for some artificial test functions. The advantage of these functions is that we know everything about them, so these experiments may give

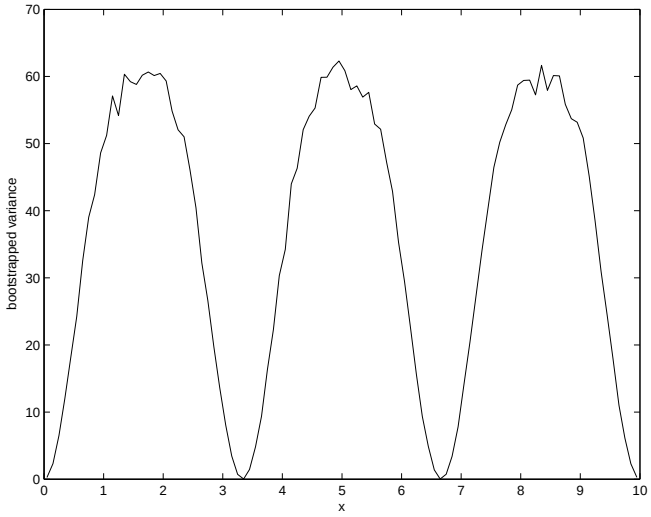

Figure 2: Example bootstrapped Kriging variance by calculating the variances one-at-a-time.

more insight. Also, we do not have to wait hours for a computer run evaluating the function.

We select the following functions:

- $f_1 : [0, 10] \mapsto \mathbb{R}$ and $f_1(x) = -0.0579x^4 + 1.11x^3 - 6.845x^2 + 14.1071x + 2$; see Figure 3.
- $f_2 : [0.1, 0.9] \mapsto \mathbb{R}$ and $f_2(x) = \frac{x}{1-x}$; see Figure 4.
- $f_3 : [-1, 1] \mapsto \mathbb{R}$ and $f_3(x) = \frac{3}{10} + \sin\left(\frac{16}{15}x - 1\right) + \sin^2\left(\frac{16}{15}x - 1\right) + \frac{2}{100}\sin\left(40(\frac{16}{15}x - 1)\right)$; see Figure 5.
- $f_4 : [-2, 2] \times [-1, 1] \mapsto \mathbb{R}$ and $f_4(x, y) = x^2(4 - 2.1x^2 + x^4/3) + xy + y^2(-4 + 4y^2)$; see Figure 6.

The one-dimensional functions f_1 and f_2 are also used in Kleijnen and van Beers (2004); f_1 is a multimodal function and f_2 equals the expected waiting time in the steady state of an M/M/1 queue. The function f_3 is also used in Giunta and Watson (1998); it consists of a 'smooth' part and a 'noisy' part where the 'smooth' part is given by the first two terms and the 'noisy' part by the last term; the 'noisy' part represents the numerical noise often encountered in practice. Finally, f_4 is a two-dimensional function with six local minima, of which two are global minima; see Dixon and Szego (1978).

To perform the bootstrap experiments, we use the multivariate normal distribution sampling routine in Matlab (used Matlab 6.5.).

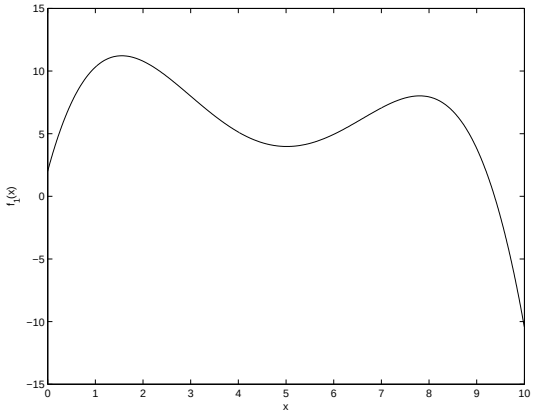

Figure 3: $f_1(x) = -0.0579x^4 + 1.11x^3 - 6.845x^2 + 14.1071x + 2$ (see Kleijnen and van Beers (2004)).

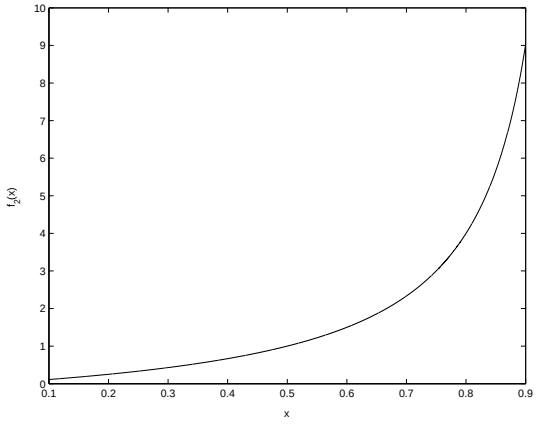

Figure 4: $f_2(x) = \frac{x}{1-x}$ (see Kleijnen and van Beers (2004)).

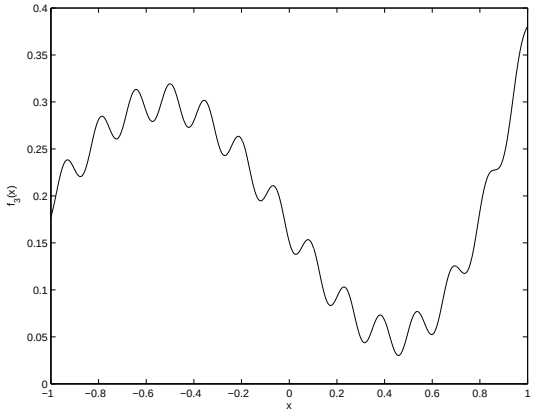

Figure 5: $f_3(x) = \frac{3}{10} + \sin\left(\frac{16}{15}x - 1\right) + \sin^2\left(\frac{16}{15}x - 1\right) + \frac{2}{100}\sin\left(40(\frac{16}{15}x - 1)\right)$ (see Giunta and Watson (1998)).

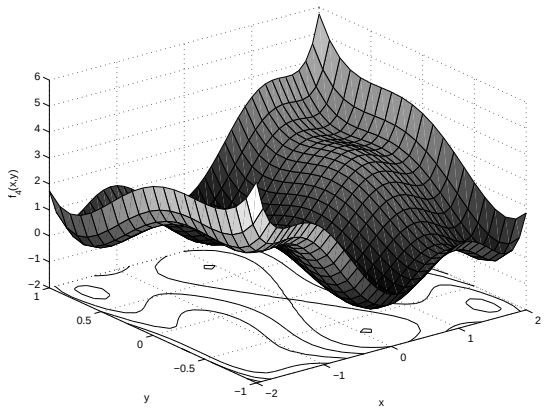

Figure 6: $f_4(x, y) = x^2(4 - 2.1x^2 + x^4/3) + xy + y^2(-4 + 4y^2)$ (see Dixon and Szego (1978)).

5.2 Analysis of bootstrap experiments

For test function f_1 we generate a dataset of four equidistant input points and calculate the corresponding output values. We compute the MLE of the parameters β, σ^2 and θ of the Gaussian process with the Matlab Toolbox DACE, which gives: $\hat{\beta} = 1.2114$, $\hat{\sigma}^2 = 48.7939$ and $\hat{\theta} = 1.0800$. Next, we calculate both the classic Kriging variance and the bootstrapped Kriging variance. Furthermore, we construct a 95% confidence interval for the bootstrapped Kriging variance by using the Central Limit Theorem as follows:

$$\left(\widehat{\mathrm{MSE}}[\hat{y}^*(x)] - 1.96 \frac{\hat{\sigma}_{SE}[\hat{y}^*(x)]}{\sqrt{B}}, \widehat{\mathrm{MSE}}[\hat{y}^*(x)] + 1.96 \frac{\hat{\sigma}_{SE}[\hat{y}^*(x)]}{\sqrt{B}} \right)$$

where $\hat{\sigma}_{SE}[\hat{y}^*(x)]$ is the estimated standard deviation of the bootstrapped squared errors:

$$\hat{\sigma}_{SE}[\hat{y}^*(x)] = \frac{1}{B-1} \sum_{b=1}^{B} \left((\hat{y}_b^*(x) - y_b^*(x))^2 - \widehat{\mathrm{MSE}}[\hat{y}^*(x)] \right)^2 .$$

In Figure 7 the solid line shows the bootstrap Kriging variance for $B = 50$. The dotted lines show the upperbound and the lowerbound of the pointwise 95% confidence interval of this variance. The dashed line shows the classic Kriging variance (11).

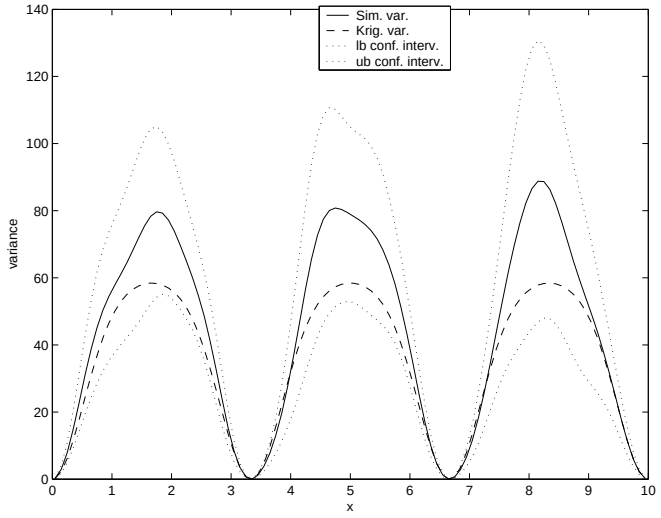

Figure 7: Bootstrap and classic Kriging variances for f_1 and $B = 50$.

We see that the bootstrap Kriging variance is larger than the classic variance almost everywhere. However, the lowerbound of the confidence interval is smaller than the classic Kriging variance. Therefore we cannot conclude that the bootstrapped Kriging variance is *significantly* larger than the classic variance.

Therefore we carry out the same experiment with a larger number of bootstrap samples, namely $B = 24000$; see Figure 8. Now the bootstrapped Kriging variance is signifi-

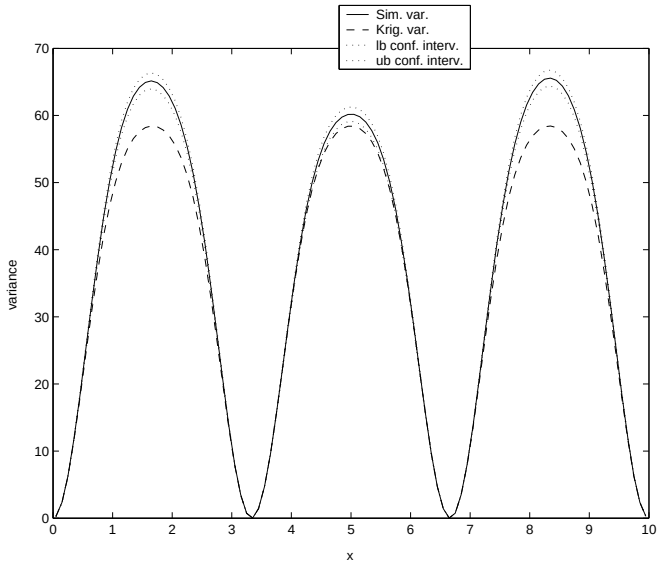

Figure 8: Bootstrap and Kriging variance for f_1 and $B = 24000$.

cantly larger than the classic Kriging variance. Furthermore the peaks of the bootstrapped Kriging variance are not all equally high, whereas the peaks of the classic Kriging variance are.

In this example, the difference between the classic and the bootstrap Kriging variances is not so big. In the next example, however, we will see that this difference can be much bigger.

Figure 9 shows the results for f_2. Again, we choose four points equidistant. Now we choose $B = 5000$. We get $\hat{\beta} = 3.0692$, $\hat{\sigma}^2 = 13.2505$, and $\hat{\theta} = 21.1043$. The figure shows that the bootstrap Kriging variance is again significantly larger than the classic Kriging variance.

For f_3 we again select equidistant input data, which gives $\hat{\beta} = 0.2244$, $\hat{\sigma}^2 = 0.0155$ and $\hat{\theta} = 27.0005$. The results for $B = 25000$ are shown in Figure 10. This figure again shows that the bootstrap Kriging variance is significantly larger than the classic variance.

For f_4 we choose a dataset of 20 input points. We choose a "maximin non-collapsing" Latin Hypercube Design (LHD); see van Dam, den Hertog, Husslage and Melissen (2004). This gives $\hat{\beta} = 1.5316$, $\hat{\sigma}^2 = 2.1994$, $\hat{\theta}_1 = 0.8058$, and $\hat{\theta}_2 = 3.2232$. The bootstrap variance for $B = 8000$ is given in Figure 11. Figure 12 shows the difference between the lowerbound of the confidence interval of the bootstrapped variance and the classic variance, which shows that the bootstrapped variance is significantly larger than the classic variance.

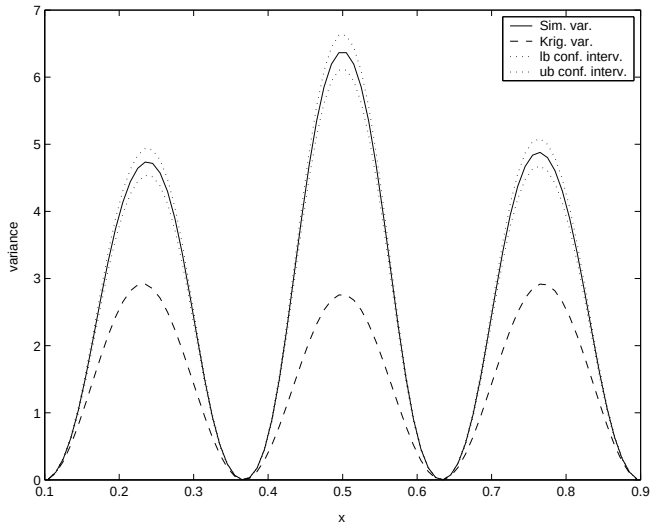

Figure 9: Bootstrap and classic Kriging variances for f_2 and $B = 5000$.

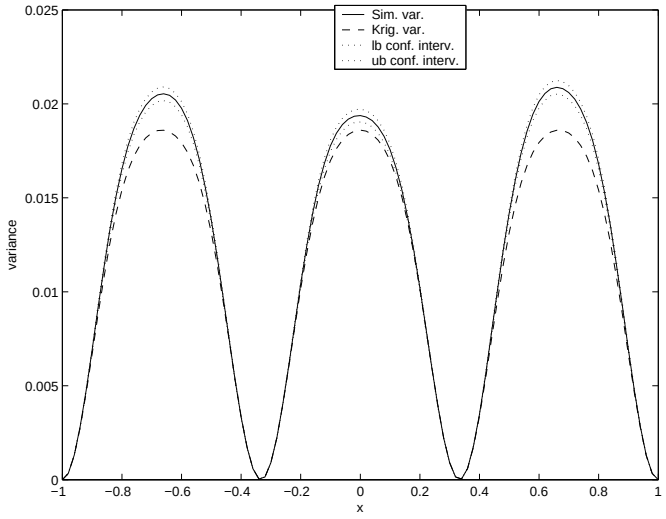

Figure 10: Bootstrap and classic Kriging variances for f_3 and $B = 25000$.

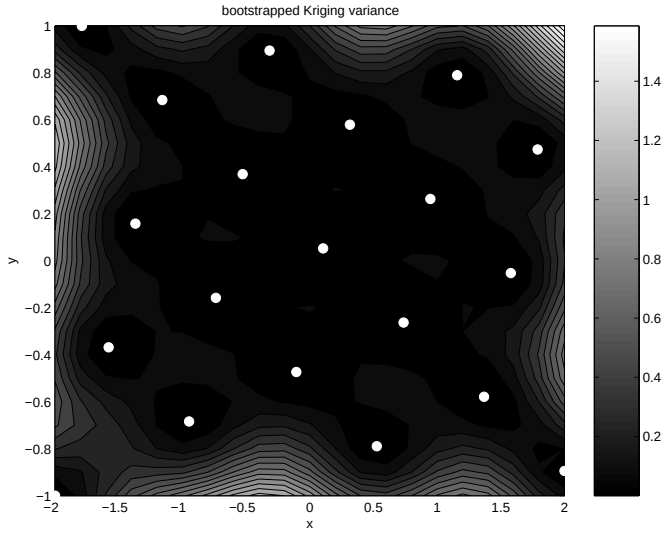

Figure 11: Bootstrap Kriging variance for f_4 and $B = 8000$.

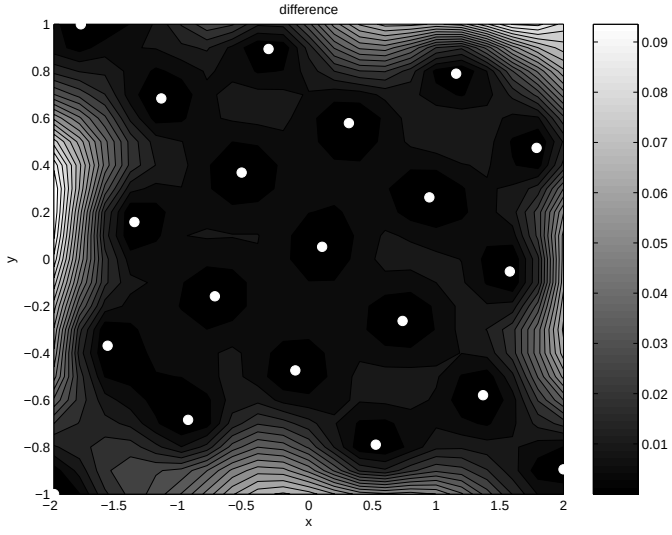

Figure 12: Difference between the lowerbound of the confidence interval of the bootstrap Kriging variance and the classic Kriging variance for f_4 and $B = 8000$.

6 Case study: a circuit-simulator

The real-life dataset taken from Sacks, Welch, Mitchell and Wynn (1989) consists of data of a circuit-simulator. The dataset consists of $n = 32$ runs. The dataset has $d = 6$ input variables. In Sacks, Welch, Mitchell and Wynn (1989), the experimental region is the unit cube $[-0.5, 0.5]^6$. We, however, want to avoid "extrapolation" as much as possible, so we take $[-0.46, 0.31] \times [-0.39, 0.45] \times [-0.47, 0.38] \times [-0.43, 0.46] \times [-0.47, 0.47] \times [-0.49, 0.41]$, which is determined by the minimum and maximum values of every original input variable in the 32 data points. All 32 input values still fall inside our reduced experimental region.

Because this case study involves a six-dimensional input, it is not possible to make the type of plots we made in Section 5. Instead, we generate a test set of 200 input data points. We do this by generating a Latin Hypercube Sample (LHS), originated by McKay, Conover and Beckman (1979). We use the LHS procedure of the Matlab Toolbox DACE. In these 200 input data points we calculate the bootstrap Kriging variance for $B = 20000$, its 95% confidence interval and the classic Kriging variance. This gives $\hat{\beta} = -0.8207$, $\hat{\sigma}^2 = 0.2611$ and $\hat{\theta} = (0.0005, 0.2422, 9.5035, 0.6036, 1.1714, 1.9215)$. Then, we calculate the difference between the bootstrap and the classic Kriging variances, the difference between the lowerbound of the 95% confidence interval of the bootstrap and the classic Kriging variances, and the classic Kriging variance for every point of the test set. This gives the three boxplots in Figure 13. These plots show that in all test points the bootstrapped variance is significantly larger than the classic Kriging variance. We also made the same boxplots with a different test set originating from another realization of the same LHS; this gave similar results.

7 Conclusions and Further Research

We have proven that the "average" classic Kriging variance formula used in most of the literature underestimates of the true Kriging variance. To estimate the correct Kriging variance we introduced a parametric bootstrapping method. Several artificial examples and a real-life case study demonstrated that the classic Kriging variance formula often underestimates indeed.

The difference between the classic and the bootstrap Kriging variances can be rather big, as we saw for the second test function (f_2). This may have a substantial impact on the three types of applications of the Kriging variance formula that we discussed in Section 1, namely

- selecting new input design points to obtain better Kriging models.

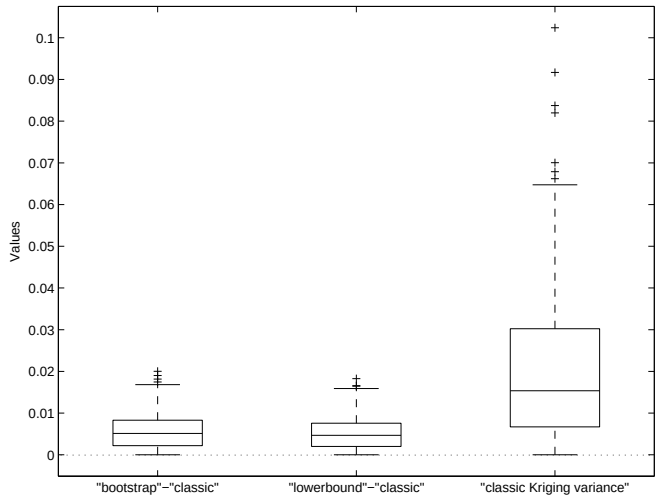

Figure 13: Boxplots of the difference between bootstrap and classic variance and the lower-bound of the 95% confidence interval of the bootstrap variance and the classic variance in 200 test points.

- selecting new input design points to find the global optimum of an underlying computer-simulation model (black-box function).
- measuring the quality of a Kriging model.

For further research we would therefore recommend to study the effect of using the bootstrap Kriging variance–instead of the classic Kriging variance formula–to these application areas of the Kriging variance.

References

Booker, A.J., Dennis J.E., Frank P.D., Serafini, D.B., Torczon, V., and Trosset, M.W. (1999), "A Rigorous Framework for Optimization of Expensive Functions by Surrogates," *Structural Optimization*, 17, 1–13.

Cox, D.D., and John, S. (1997), "SDO: A statistical method for global optimization," in *Multidisciplinary Design Optimization: State of the Art*, eds. N. Alexandrov and M.Y. Hussaini, Philadelphia: SIAM, pp. 315–329.

Cressie, N. (1991), *Statistics for spatial data*, New York: Wiley-Interscience.

Dam, E. van, Hertog, D. den, Husslage, B.G.M., Melissen, J.B.M. (2004), "Maximin Latin Hypercube designs in two dimensions," Working paper.

Dixon, L.C.W., and Szego, G.P. (1978), "The optimization problem: An introduction," in *Towards global optimization II*, eds. L.C.W. Dixon and G.P. Szego, New York: North-Holland.

Efron, B., and Tibshirani, R.J. (1993), *An introduction to the bootstrap*, New York: Chapman and Hall.

Giunta, A., and Watson, L. (1998), "A comparison of approximation modeling techniques: Polynomial versus interpolating models," in *7th AIAA/USAF/NASA/ISSMO symposium on multidisciplinary analysis and optimization*, AIAA-98-4758, vol. 1, St. Louis: AIAA.

Jin, R., Chen, W., and Sudjianto, A. (2002), "On sequential sampling for global metamodeling in engineering design," in *Proceedings of DETC'02 ASME 2002 design engineering technical conferences and computers and information in engineering conference.*

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

Jones, D. (2002), "A taxonomy of global optimization methods based on response surfaces," *Journal of global optimization*, 21, 345–383.

Kleijnen, J.P.C., and van Beers, W.C.M. (2004), "Application-driven sequential designs for simulation experiments: Kriging metamodeling," *Journal of the operational research society*, Preprint from: http://center.uvt.nl/staff/kleijnen/seqdesignjors2.pdf.

Koehler, J.R., and Owen, A.B. (1996), "Computer experiments," in *Handbook of statistics*, vol. 13, eds. S. Ghosh and C.R. Rao, New York: Elsevier Science B.V., pp. 261–308.

Lophaven, S.N., Nielsen, H.B., and Sondergaard, J. (2002), "DACE: A Matlab Kriging toolbox version 2.0," Technical Report IMM-TR-2002-12, Technical University of Denmark, Copenhagen.

McKay, M.D., Conover, W.J., and Beckman, R.J. (1979), "A comparison of three methods for selecting values of input variables in the analysis of output from a computer code," *Technometrics*, 21.

Mittelhammer, R.C. (1996), *Mathematical statistics for economics and business*, New York: Springer.

Sacks, J., Welch, W.J., Mitchell, T.J., and Wynn, H.P. (1989), "Design and analysis of computer experiments," *Statistical science*, 4, 409–435.

Santner, T.J., Williams, B.J., and Notz, W.I. (2003), *The design and analysis of computer experiments*, New York: Springer-Verlag.

Sasena, M.J., Papalambros, P., and Goovaerts, P. (2002), "Exploration of metamodeling sampling criteria for constrained global optimization", *Engineering optimization*, 34, 263–278.

Schonlau, M., Welch, W.J., and Jones, D. (1998), "Global versus local search in constrained optimization of computer models," in *New developments and applications in experimental design*, vol. 34, eds. N. Flournoy, W.F. Rosenberger and W.K. Wong, Hayward California: Institute of Mathematical Statistics, pp. 11–25.

Stehouwer, H.P., and Hertog, D. den (1999), "Simulation-based design optimization: Methodology and applications," in *Proceedings of the first ASMO UK/ISSMO conference on engineering design optimization*, Ilkly, UK.