Kriging-based sequential design strategies for computer codes

Loic Le Gratiet
Université Paris Diderot 75205 Paris Cedex 13
and
Claire Cannamela
CEA, DAM, DIF, F-91297 Arpajon, France

April 29, 2014

1 Kriging models and sequential designs

In this Section, we briefly introduce the kriging model and some classical sequential design criteria. Then, we will present our sequential strategies to enhance kriging models considering the region with large Leave-One-Out Cross-Validation (LOO-CV) errors.

1.1 The Kriging model

The kriging model is a widely used method to surrogate the output of a computer code from few simulations (Sacks et al. (1989)). Let us denote by $y(x)$ the output of the code at point $x \in Q$. Here, $y(x)$ is a scalar and $Q \subset \mathbb{R}^d$ is a compact. Furthermore, we denote by $\mathbf{D} = \{x_1, \dots, x_n\}$ the experimental design set and $\mathbf{y}_n = y(\mathbf{D})$ the value of $y(x)$ at points in $\mathbf{D}$.

In the kriging framework, we assume that the prior knowledge about the code can be modeled by a Gaussian process $Y_0(x)$ indexed by $x \in Q$ and with values in $\mathbb{R}$. Usually, we consider a Gaussian process with mean of the form $m_0(x) = \mathbf{f}'(x)\beta$, with $\mathbf{f}'(x) = (f_1(x), \dots, f_p(x))$ and with covariance function $k_0(x, \tilde{x}) = \sigma^2 r\left(x, \tilde{x}\right)$ where $r\left(x, \tilde{x}\right)$ is a symmetric positive definite kernel such that $r(x, x) = 1$ for all $x \in Q$. Then, the kriging equations are given by the Gaussian process $Y_0(x)$ conditioned by its known values $\mathbf{y}_n$ at points in $\mathbf{D}$

$$Y_n(x) \sim [Y_0(x)|Y_0(\mathbf{D}) = \mathbf{y}_n] = \mathrm{GP}\left(m_n(x), k_n(x, \tilde{x})\right), \tag{1}$$

where

$$m_n(x) = \mathbf{f}'(x)\hat{\beta} + \mathbf{r}'(x)\mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{F}\hat{\beta}), \tag{2}$$

and

$$k_n(x, \tilde{x}) = \sigma^2 \left(r(x, \tilde{x}) - \begin{pmatrix} \mathbf{f}'(x) & \mathbf{r}'(x) \end{pmatrix} \begin{pmatrix} 0 & \mathbf{F}' \\ \mathbf{F} & \mathbf{R} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{f}(\tilde{x}) \\ \mathbf{r}(\tilde{x}) \end{pmatrix} \right), \tag{3}$$

where $'$ stands for the transpose, GP denotes a Gaussian process, $\mathbf{F}$ are the values of $\mathbf{f}'(x)$ at points in $\mathbf{D}$, $\mathbf{r}(x)$ is the correlation vector between $\mathbf{D}$ and x with respect to the correlation function $r(x, \tilde{x})$, $\mathbf{R}$ is the correlation matrix of $\mathbf{D}$ with respect to $r(x, \tilde{x})$ and $\hat{\beta} = (\mathbf{F}'\mathbf{R}^{-1}\mathbf{F})^{-1}\mathbf{F}'\mathbf{R}^{-1}\mathbf{y}_n$ is the usual least-squares estimate of β. We note that $x, \tilde{x} \in Q$ can be either the same point or different points since $Y_n(x)$ is a Gaussian process. For $x = \tilde{x}$, $k_n(x, x)$ is the variance of $Y_n(x)$ and it represents the uncertainty on the predictive mean $m_n(x)$. Furthermore, the restricted Maximum Likelihood Estimate (MLE) of σ^2 is given by $\hat{\sigma}^2 = (\mathbf{y}_n - \mathbf{F}\hat{\beta})'\mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{F}\hat{\beta})/(n - p)$ (see Santner et al. (2003)).

1.1.1 One point at-a-time Sequential design

Now, let us suppose that we want to add a new point x_{n+1} in $\mathbf{D}$ in order to enhance the accuracy of the kriging model. From the kriging variance $k_n(x, x)$ - representing the model MSE - sequential design methods have been derived Sacks et al. (1989), Bates et al. (1996) and Picheny et al. (2010). A first one consists of adding x_{n+1} where the kriging variance is the largest (see Sacks et al. (1989))

$$x_{n+1} = \arg\max_x k_n(x, x). \tag{4}$$

However, as presented in Kleijnen and van Beers (2004), its performance is poor. Then, it has been improved with a criterion which consists of adding the new point which gives the most important Integrated Mean Squared Error (IMSE) reduction (see Bates et al. (1996) and Picheny et al. (2010))

$$x_{n+1} = \arg\max_x \int_{u \in Q} k_n(u, u) - k_{n+1}(u, u; x)\, du, \tag{5}$$

where

$$k_{n+1}(u, \tilde{u}; x) = \sigma^2 \left(r(u, \tilde{u}) - \begin{pmatrix} \mathbf{f}(u) \\ \mathbf{r}(u) \\ r(u, x) \end{pmatrix}' \begin{pmatrix} 0 & \mathbf{F}' & \mathbf{f}(x) \\ \mathbf{F} & \mathbf{R} & \mathbf{r}(x) \\ \mathbf{f}'(x) & \mathbf{r}'(x) & 1 \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{f}(\tilde{u}) \\ \mathbf{r}(\tilde{u}) \\ r(\tilde{u}, x) \end{pmatrix} \right).$$

Here, the covariance kernel $k_{n+1}(u, \tilde{u}; x)$ corresponds to the one of the Gaussian process $Y_n(u)$ (1) conditioned on a new observation at x. Furthermore, (3) shows that the kriging variance does not depend on the observations if we consider that the parameter σ^2 is known. Therefore, in that case, $k_{n+1}(u, u; x)$ can be computed without evaluating $y(x)$ at the new point x. We denote by MinIMSE this criterion. Finally, we also consider the criterion presented by Kleijnen and van Beers (2004) using a Jackknife estimator for the predictor's variance. Its principle is the following one. Let us consider $m_{n,-i}(x)$ the kriging mean built without the ith observation, the jackknife variance is given by

$$s^2_{jack}(x) = \frac{1}{n(n-1)} \sum_{i=1}^{n} (\tilde{y}_i - \bar{\tilde{y}})^2, \tag{6}$$

where $\tilde{y}_i = n m_n(x) - (n-1) m_{n,-i}(x)$ and $\bar{\tilde{y}} = \sum_{i=1}^{n} \tilde{y}_i / n$. Then, we consider candidate points coming from a maximin LHS Design (Fang et al. (2006)) and we add those which maximize the jackknife variance. We denote by KleiCrit this criterion.

1.1.2 q points at-a-time Sequential design

There is a natural way to extend these algorithms when the simulations can be performed simultaneously. Indeed, the covariance kernel $k_{n+1}(x, \tilde{x}; x_{n+1})$ of the Gaussian process $Y_n(x)$ conditioned by the new observation at point x_{n+1} can be computed without knowing $y(x_{n+1})$ when we consider the model parameter σ^2 as known. Then, from $k_{n+1}(x, \tilde{x}; x_{n+1})$, we can find a new point x_{n+2} where to perform a new simulation (i.e. a new evaluation of $y(\dot{)})$ using the same criterion as in (5) and the kernel $k_{n+2}(x, \tilde{x}; x_{n+1}, x_{n+2})$. Thus, considering the parameter σ^2 as known (they are fixed to their estimated values), we can determine with this procedure q good locations where to perform simulations. We call this method the "liar" sequential kriging. We highlight that one can also decide to perform the optimization by adding the q points simultaneously. However, it is not relevant for the criterion (4) since they will be concentrated around the maximum of $k_n(x, x)$. Furthermore, it is extremely complex and time-consuming for the criterion (5) since for each set of q points we have to compute $k_{n+1}(u, u; x)$ and integrate it over Q with a d-dimensional integration.

1.2 LOO-CV based strategies for kriging sequential design

We present in this subsection original sequential-kriging strategies. The main difference between these new strategies and the previous ones is that they take into account the model errors through the Leave-One-Out Cross-Validation (LOO-CV) equations. First, let us introduce some notations.

Notations: $\mathbf{A}_{i,i}$ is the ith element of the main diagonal of $\mathbf{A}$, $\mathbf{A}_i$ is the ith row of the matrix $\mathbf{A}$, $\mathbf{A}_{-i}$ is the matrix $\mathbf{A}$ without its ith row, $\mathbf{A}_{-i,i}$ is the ith column of $\mathbf{A}$ without its ith element, $\mathbf{A}_{i,-i} = \mathbf{A}'_{-i,i}$ and $\mathbf{A}_{-i,-i}$ is the matrix $\mathbf{A}$ without the ith row and column.

Let us denote by $Y_{n,-i}(x)$ the Gaussian process $Y_0(x)$ conditioned by the values $\mathbf{y}_{n,-i} = y(\mathbf{D}) \setminus y(x_i)$. Then, the predictive mean of $Y_{n,-i}(x)$ at point x_i is given by

$$m_{n,-i}(x_i) = y(x_i) - \left[\mathbf{R}^{-1}(\mathbf{y}_n - \mathbf{F}\hat{\beta}_{-i})\right]_i / \left[\mathbf{R}^{-1}\right]_{i,i}, \tag{7}$$

(see Dubrule (1983) and Fasshauer and Zhang (2007)) where $\hat{\beta}_{-i} = (\mathbf{F}'_{-i}\mathbf{K}_i\mathbf{F}_{-i})^{-1}\mathbf{F}'_{-i}\mathbf{K}_i\mathbf{y}_{n,-i}$ and $\mathbf{K}_i = [\mathbf{R}^{-1}]_{-i,-i} - [\mathbf{R}^{-1}]_{-i,i}[\mathbf{R}^{-1}]_{i,-i} / [\mathbf{R}^{-1}]_{i,i}$. This result is presented in Fasshauer and Zhang (2007)) to estimate the shape parameter of the correlation kernel. These equations allow for avoiding the computation of $[\mathbf{R}_{-i,-i}]^{-1}$ corresponding to the correlation matrix of $\mathbf{y}_{n,-i}$. Since the inverse $\mathbf{R}^{-1}$ has been already computed during the model building, the computation only requires matrix products.

Furthermore, the predictive variance of $Y_{n,-i}(x)$ at point x_i is given by

$$k_{n,-i}(x_i) = \sigma^2 / \left[\mathbf{R}^{-1}\right]_{i,i} + \varsigma_{-i}(x_i), \tag{8}$$

where $\varsigma_{-i}(x_i) = \left([\mathbf{R}^{-1}\mathbf{F}]_i / [\mathbf{R}^{-1}]_{i,i}\right)' (\mathbf{F}'_{-i}\mathbf{K}_i\mathbf{F}_{-i})^{-1} \left([\mathbf{R}^{-1}\mathbf{F}]_i / [\mathbf{R}^{-1}]_{i,i}\right)$.

The variance parameter σ^2 in (8) is here considered as known. In fact, we can easily re-estimate it by noticing the equality $\mathbf{K}_i = (\mathbf{R}_{-i,-i})^{-1}$. Therefore, we have the following MLE of σ^2 when we do not consider the ith observation $y(x_i)$

$$\hat{\sigma}^2_{-i} = \left(\mathbf{y}_{n,-i} - \mathbf{F}_{-i}\hat{\beta}_{-i}\right)' \mathbf{K}_i \left(\mathbf{y}_{n,-i} - \mathbf{F}_{-i}\hat{\beta}_{-i}\right) / (n-p-1). \tag{9}$$

The previous results provide a powerful tool to compute the LOO-CV predictive means and variances. Indeed, the complexity for computing (7) and (8) for all $i = 1, \dots, n$ is $\mathcal{O}(n^3)$ whereas the one of a direct LOO-CV procedure with (2) and (3) is $\mathcal{O}(n^4)$. Consequently, the LOO-CV equations are fast to compute and can be easily recomputed at each step of the sequential strategy. We note that as the value of $k_{n,-i}(x_i)$ is strongly dependent on $\hat{\sigma}^2_{-i}$, in our forthcoming developments it is important to re-estimate it.

Now, let us denote by $\mathbf{e}^2_{\mathrm{LOO-CV}} = \left[((y(x_i) - m_{n,-i}(x_i))^2\right]_{i=1,\dots,n}$ the vector of the LOO-CV squared errors and $\mathbf{s}^2_{\mathrm{LOO-CV}} = [k_{n,-i}(x_i)]_{i=1,\dots,n}$ the vector of the LOO-CV variances. Furthermore, let us consider the Voronoi cells $(V_i)_{i=1,\dots,n}$ associated with the points $(x_i)_{i=1,\dots,n}$

$$V_i = \{x \in Q,\, ||x - x_i|| \le ||x - x_j||,\, \forall j \ne i\},\, i, j = 1, \dots, n. \tag{10}$$

In the remainder of this section, we present two strategies to sequentially add simulations which use $\mathbf{e}^2_{\mathrm{LOO-CV}}$, $\mathbf{s}^2_{\mathrm{LOO-CV}}$ and V_i. Their are based on the criterion presented in (11). The intuitive idea of the suggested criterion is to minimize the predictive variance in the locations where the LOO-CV errors are important. The implicit assumption is to consider that the LOO predictive error and variance at point x_i is a proxy for the actual error for all points in the Voronoi cells of x_i.

1.2.1 LOO-CV-based one point at-a-time Sequential design

Let us denote by x_{n+1} the new point that we want to add to $\mathbf{D}$. We consider the point solving the following problem

$$x_{n+1} = \arg\max_x k_n(x, x) \left(1 + \sum_{i=1}^{n} \frac{[\mathbf{e}^2_{\mathrm{LOO-CV}}]_i}{[\mathbf{s}^2_{\mathrm{LOO-CV}}]_i} \mathbf{1}_{x \in V_i}\right), \tag{11}$$

where $\mathbf{1}$ stands for the indicator function.

This criterion considers the predictor's Mean Squared Error (MSE) $k_n(x, x)$ adjusted with the LOO-CV errors and variances. For equivalent $k_n(x, x)$, the criterion favors the points close to an experimental design point with large LOO-CV errors. Furthermore, if two points are in the same Voronoi cell, the one with the largest predictor's MSE is considered. Therefore, a sequential strategy with this criterion focuses on the regions of Q where the LOO-CV errors are the largest. We note that the normalization with $\mathbf{s}^2_{\mathrm{LOO-CV}}$ is important since it is not necessary to enlarge the predictor's MSE in the regions where it is well or over estimated. As an example, $[\mathbf{e}^2_{\mathrm{LOO-CV}}]_i \ll [\mathbf{s}^2_{\mathrm{LOO-CV}}]_i$ means that the kriging variance is over-estimated around the point x_i, i.e. $k_n(x,x)$ is too large for $x \in V_i$. In that case, the normalization with $[\mathbf{s}^2_{\mathrm{LOO-CV}}]_i$ implies that $\sum_{i=1}^n \frac{[\mathbf{e}^2_{\mathrm{LOO-CV}}]_i}{[\mathbf{s}^2_{\mathrm{LOO-CV}}]_i} \mathbf{1}_{x \in V_i} \approx 0$ for $x \in V_i$ and thus the term in (11) is approximately equal to $k_n(x,x)$.

We illustrate in Figure 1 the adjusted variance presented in (11) and the classical kriging variance (3) in a 1-dimensional example. The considered function is $f(x) = (\sin(7x) + \cos(14x))x^2 \exp(-4x)$, $x \in [0,4]$. We use a kriging model with a 5/2-Matérn kernel with $\sigma^2 = 1.10^{-3}$ and $\theta = 1$ (see Rasmussen and Williams (2006) p.84) and the experimental design set is a regular grid of 8 points between 0 and 4. We see in Figure 1 that the kriging model is not accurate in the domain $[0,2]$ where the function variations are important and the adjusted kriging variance (11) focuses on that region.

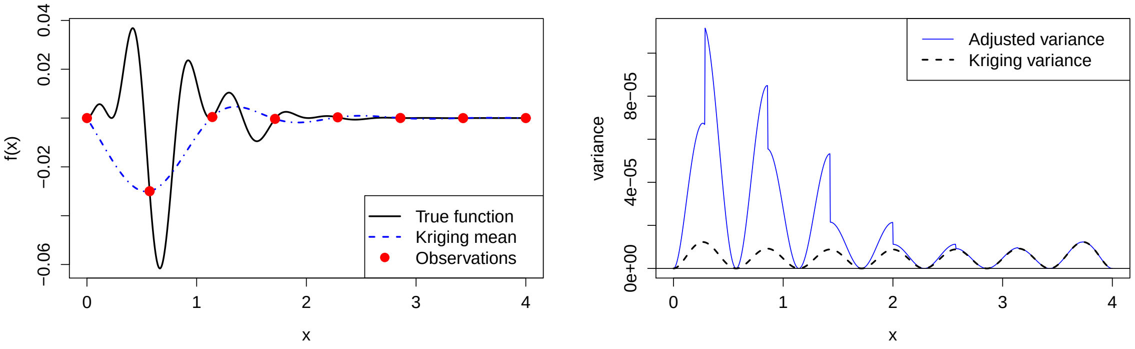

Figure 1: Illustration of the adjusted kriging variance in a 1-dimensional example. We see that the kriging variance is enlarged in the domain where the function variations are important.

As illustrated in Figure 1, the adjusted kriging variance allows for taking into account the LOO-CV error in a sequential procedure focusing on the large error domain. Nevertheless, it does not entirely fix the issue of the relevance of $k_n(x,x)$ to represent the model error. Indeed, our criterion enlarges the kriging variance around points where $k_n(x,x)$ is under-estimated but it does not reduce it at locations where it is over-estimated. However, it gives more information about the relevance of $m_n(x)$ since it highlights the regions where it is not accurate. Furthermore, it also helps in the interpretation of $k_n(x,x)$ since it emphasizes whether it is under-estimated or not.

An efficient method to solve the problem in (11) is to use an evolutionary algorithm coupled with a descent algorithm. To explore different Voronoi cells $(V_i)_{i=1,\dots,n}$, we can use a genetic algorithm or simply a Monte-Carlo sample for low-dimensional problems (i.e. $d < 10$). Then, for a given Voronoi cell, the criterion to optimize is continuous. Therefore, we can solve the problem with classical optimization methods (direct or simplex methods). Furthermore, it is common to have covariance kernel such that the criterion is once or twice continuously differentiable. In this case, we can use a gradient method, a conjugate method or a Newton method (for the twice differentiable case). We note that it is not necessary to compute the Voronoi tessellation since the criterion only requires to determine in which Voronoi cell a given point $x \in Q \subset \mathbb{R}^d$ lies. This is computationally simple and cheap even for high dimension d.

1.2.2 LOO-CV-based q points at-a-time Sequential design

We extend here the previous criterion for a q points at-a-time sequential design. First, we emphasize that the liar sequential kriging is not relevant for this new criterion. Indeed, conditioning on model parameters, with a liar method we can compute the kriging variances $(k_{n+i}(x,x))_{i=1,\dots,q}$ but not the LOO-CV (7) and (8). Therefore, we use another strategy to propose q new locations where to perform the simulations. This approach is proposed in Dubourg et al. (2011) in a different framework. The idea of the suggested method is to select the q best points with respect to the criterion (11) from N candidate points. These N candidate points are chosen with the following algorithm.

1. Generate N_{MCMC} samples with respect to the probability density function proportional to $k_n(x,x)$ with a suitable Markov Chain Monte Carlo (MCMC) technique Robert and Casella (2004).

2. Extract from these samples N representative points with a N-means clustering technique MacQueen (1967).

As presented in Dubourg et al. (2011) the use of this algorithm to select N candidate points in a kriging framework is efficient. Indeed, it allows us to concentrate the points at the modes of the kriging variance. In the proposed strategy, we always take $N \geq q$ and we choose from the N cluster centers $(C_i)_{i=1\dots,N}$ the q points where $k_{n,\text{adj}}(x,x) = k_n(x,x)\left(1 + \sum_{i=1}^{n} \frac{[\mathbf{e}^2_{\text{LOO-CV}}]_i}{[\mathbf{s}^2_{\text{LOO-CV}}]_i} \mathbf{1}_{x \in V_i}\right)$ is the largest. For the MCMC procedure, we use a Metropolis-Hastings (M-H) algorithm with a Gaussian proposal distribution. It is centered on the last sample point and has a standard deviation such that the acceptance rate is around 30% (see Robert and Casella (2004)). Furthermore, we set N_{MCMC} such that $N_{\text{MCMC}} \gg N$. For the N-means procedure, we choose the value of N with respect to the criterion

$$\max_{N \geq q} \min_{x \in (C_i)_i} k_n(x,x), \tag{12}$$

where $(C_i)_{i=1\dots,N}$ are the cluster centers. This criterion prevents from having a cluster center C_i in a region where the kriging variance is close to zero. Furthermore, if the number of clusters is too high, the cluster centers get away from the modes and consequently the value of $\min_{x \in (C_i)_{i=1\dots,N}} k_n(x,x)$ decreases. Therefore, this criterion also prevents the number of clusters from being too large. In practice, we choose N on a finite sequence from q to $2n$ where n is the number of observations and we run the N-means procedure several times for each N. Then, we select the cluster centers minimizing (12). We note that the MCMC plus N-means procedure requires careful implementation and appropriate diagnostics. For the N-means procedure, we use the algorithm suggested by Hartigan and Wong (1979) with complexity $\mathcal{O}(N N_{\text{MCMC}})$. For the M-H procedure we use the package mcmc.

To avoid computational issues, one could instead extract the q-points from candidates generated with space-filling design techniques Fang et al. (2006). However, with this technique, the candidate points will not anymore be concentrated in the regions of high mean square error and the method will be less efficient.

2 Numerical study

We compare in this Section the MinIMSE, KleiCrit and AdjMMSE criteria on toy examples and on an application concerning a spherical tank under pressure. We present both the cases of 1 point at-a-time and q points at-a-time sequential kriging. The purpose of this section is to emphasize the efficiency of the LOO-CV-based criteria for kriging models compared to the ones presented by Bates et al. (1996), Kleijnen and van Beers (2004) and Picheny et al. (2010).

2.1 Comparison between sequential kriging criteria

The 1 point at-a-time sequential kriging criteria (MinIMSE, KleiCrit, AdjMMSE) are compared on three tabulated functions:

- Ackley's function on $[-2,2]^2$:

$$f(x_1,x_2) = -20\mathrm{exp}\left(-0.2\sqrt{\frac{x_1^2+x_2^2}{2}}\right) - \exp\left(\frac{\cos(2\pi x_1)+\cos(2\pi x_2)}{2}\right) + 20 + \exp(1).$$

- Shubert's function on $[-2,2]^2$:

$$f(x_1,x_2) = \left(\sum_{k=1}^{5} k\cos\left((k+1)x_1+k\right)\right)\left(\sum_{k=1}^{5} k\cos\left((k+1)x_2+k\right)\right).$$

- Michalewicz's function on $[0,\pi]^2$ (Michalewicz (1992)):

$$f(x_1,x_2) = -\sin(x_1)\left(\sin\left(\frac{x_1^2}{\pi}\right)\right)^{20} - \sin(x_2)\left(\sin\left(\frac{x_2^2}{\pi}\right)\right)^{20}.$$

The comparison is performed on a test set $\mathbf{D}_{\mathrm{test}}$ composed of $n_{\mathrm{test}} = 1,000$ points uniformly spread on the input parameter space and from 50 different initial experimental design sets. We compare the different methods with respect to the Normalized RMSE

$$\mathrm{Norm\,RMSE} = \frac{\sqrt{\sum_{i=1}^{n_{\mathrm{test}}}\left(y_{\mathrm{real}}(x_{\mathrm{test}}^i) - y_{\mathrm{pred}}(x)\right)^2/n_{\mathrm{test}}}}{\max_{x\in\mathbf{D}_{\mathrm{test}}} y_{\mathrm{real}}(x) - \min_{x\in\mathbf{D}_{\mathrm{test}}} y_{\mathrm{real}}(x)},$$

where $y_{\mathrm{real}}(x)$ is the real value of the output and $y_{\mathrm{pred}}(x)$ the predicted one. The 50 initial experimental design sets are LHS designs of 10 points optimized with respect to the S-optimality (Stocki (2005)). From these designs, 50 sequential krigings are performed and the convergence of the mean and the quantiles of the Normalized RMSE are computed for the three criteria. The mean and confidence intervals of the Normalized RMSE with respect to these 50 initial design sets are presented in Figure 2. We use for each kriging a tensorised 5/2-Matérn covariance function (see Rasmussen and Williams (2006)) and a constant trend. Furthermore, after each point addition, the parameters β, σ^2 and θ (see (1), (2) and (3)) of the kriging models are re-estimated with a maximum likelihood method.

These estimations are performed thanks to the R library 'DiceKriging' Roustant et al. (2012).

Figure 2 illustrates the efficiency of the criterion AdjMMSE. Indeed, for the Shubert's and the Michalewicz's functions, we see that the accuracy of the 1 point at-a-time kriging with this criterion is significantly better than the other two criteria (both in terms of mean and quantiles of the Normalized RMSE). In fact, these functions have the particularity to have important variations in some areas of the input parameter space. Thus, the errors are more important in these locations and the suggested criterion focuses the new points on it. Furthermore, the contrast of variations are particulary important for Schubert's function. For this reason, the IMSE criterion performed very poorly in that case. Indeed, this criterion is efficient for functions with homogeneous variations (i.e. when the predictor's MSE well predicts the model errors). In contrast, the jackknife predictor's MSE provided by the criterion KleiCrit manages to catch this heterogeneity and it performs better than the IMSE criterion. Moreover, we see that the performance of the AdjMMSE and IMSE criteria are equivalent for the Ackley's function. We note that the variations of the Ackley's function have the same order of magnitude over the input parameter space.

These examples illustrate the fact that our criterion is more efficient than the other criteria when the functions have important contrast variations and it remains efficient even in the cases where the functions have homogeneous variations (its efficiency is equivalent to the one of the IMSE criterion).

Another point of interest is to compare the gain of CPU-time by using the short cuts of Leave-One-Out Cross Validation presented in (7) and (8). For the three academic examples, the CPU-time of the sequential design using the criterion AdjMMSE with (7) and (8) is around 14s whereas the one without them is around 19s. Therefore, the gain is substancial (it is approximately 25%).

2.2 Spherical tank under internal pressure example

In this section, we deal with the example about a spherical tank under internal pressure. Figure 3 compares the different criteria of the 1 point at-a-time and the $q = 5$ points at-a-time sequential kriging. We see that the criteria MinIMSE and AdjMMSE give equivalent values for the MSE for the 1 point at-a-time procedure and they perform better than the KleiCrit criterion. There are equivalent since the output $y^2(x)$ has homogeneous variations. Nevertheless, the criterion AdjMMSE is the most efficient for the $q = 5$ points at-a-time procedure. We note that the $q = 5$ points for the MinIMSE criterion are provided by a 'liar' method whereas those for the AdjMMSE criterion are provided by the MCMC+N-means procedure suggested in the article. This comparison emphasize that the MCMC+N-means procedure can be worthwhile. Indeed, the MinIMSE criterion and the AdjMMSE one having equivalent performance for 1-step at-a-time design, we expect that they are still equivalent for a $q = 5$ points at-a-time procedure. The difference is hence explained by the point selection procedure.

Finally, we note that the 5 at-a-time approach with the AdjMMSE criterion appears to be as good as the 1 point at-a-time procedure. This highlights the relevance of the suggested point selection approach.

3 Proofs of equations (17), (18) and (19)

Let us consider x_i^l the i^{th} point of $\mathbf{D}^l$ and i_j the index of the element of $\mathbf{D}^j$ corresponding to the point x_i^l. Sorting the experimental design sets such that x_i^l corresponds to the last point of $\mathbf{D}^l$ and thanks to the block-wise inversion formula, we have the equality

$$\mathbf{R}_l^{-1} = \begin{pmatrix} \mathbf{A} & \mathbf{b} \\ \mathbf{b}' & Q^{-1} \end{pmatrix},$$

with $\mathbf{A} = [\mathbf{R}_l]^{-1}_{-i_l,-i_l} + [\mathbf{R}_l]^{-1}_{-i_l,-i_l} [\mathbf{R}_l]_{-i_l,i_l} [\mathbf{R}_l]_{i_l,-i_l} [\mathbf{R}_l]^{-1}_{-i_l,-i_l}/Q$, $\mathbf{b}' = -[\mathbf{R}_l]_{i_l,-i_l} [\mathbf{R}_l]^{-1}_{-i_l,-i_l}/Q$ and

$$Q = [\mathbf{R}_l]_{i_l,i_l} - [\mathbf{R}_l]_{i_l,-i_l} [\mathbf{R}_l]^{-1}_{-i_l,-i_l} [\mathbf{R}_l]_{-i_l,i_l}.$$

We note that

$$\sigma^2 Q = \frac{\sigma^2}{\left[\mathbf{R}_l^{-1}\right]_{i_l,i_l}}, \tag{13}$$

represents the variance at point x_i^l with respect to the covariance kernel of a Gaussian process of kernel $\sigma^2 r_l(x, \tilde{x})$ conditioned by the points $\mathbf{D}^l \setminus x_i^l$.

Furthermore, we have the equality

$$\begin{aligned}
\left(\left[\mathbf{R}_l^{-1}\right]_{i_l,i_l}\right)^{-1}\left[\mathbf{R}_l^{-1}\left(\mathbf{y}^l-\mathbf{H}_l\begin{pmatrix}\rho_{l-1}\\ \beta_l\end{pmatrix}\right)\right]_{i_l} &= y^l(x_i^l)-\rho_{l-1}y^{l-1}(x_i^l)-\mathbf{f}_l'(x_i^l)\beta_l \\
&- \left[\mathbf{R}_l\right]_{i_l,-i_l}\left[\mathbf{R}_l\right]^{-1}_{-i_l,-i_l} \\
&\times \left(y^l(\mathbf{D}^l_{-i_l})-[\mathbf{H}_l]_{[-i_l]}\begin{pmatrix}\rho_{l-1}\\ \beta_l\end{pmatrix}\right)
\end{aligned} \quad . \tag{14}$$

Now, let us consider $\hat{\sigma}^2_{l,-i_l}$ and $\begin{pmatrix}\hat{\rho}_{l-1,i_l}\\ \hat{\beta}_{l,i_l}\end{pmatrix}$. We have the equality

$$\begin{aligned}
[\mathbf{R}_l]^{-1}_{[-i_l,-i_l]} &= \mathbf{A}-\mathbf{b}Q\mathbf{b}' \\
&= [\mathbf{R}_l^{-1}]_{-i_l,-i_l}-[\mathbf{R}_l^{-1}]_{-i_l,i_l}[\mathbf{R}_l^{-1}]_{i_l,-i_l}/[\mathbf{R}_l^{-1}]_{i_l,i_l}.
\end{aligned}$$

Therefore, we can deduce the inverse of the correlation matrix of the observations at points in $\mathbf{D}^l_{-i_l}$ from the one of the observations at points in $\mathbf{D}^l$. Let us denote by $\mathbf{K}_l=[\mathbf{R}_l]^{-1}_{[-i_l,-i_l]}$, $\hat{\sigma}^2_{l,-i_l}$ and $\begin{pmatrix}\hat{\rho}_{l-1,i_l}\\ \hat{\beta}_{l,i_l}\end{pmatrix}$ are given by the equations

$$\begin{pmatrix}\hat{\rho}_{l-1,i_l}\\ \hat{\beta}_{l,i_l}\end{pmatrix}\left([\mathbf{H}_l']_{-i_l}\mathbf{K}_l[\mathbf{H}_l]_{-i_l}\right)=[\mathbf{H}_l']_{-i_l}\mathbf{K}_l y^l(\mathbf{D}^l_{-i_l}), \tag{15}$$

and

$$\hat{\sigma}^2_{l,-i_l}=\frac{\left(y^l(\mathbf{D}^l_{-i_l})-[\mathbf{H}_l]_{-i_l}\begin{pmatrix}\hat{\rho}_{l-1,i_l}\\ \hat{\beta}_{l,i_l}\end{pmatrix}\right)'\mathbf{K}_l\left(y^l(\mathbf{D}^l_{-i_l})-[\mathbf{H}_l]_{-i_l}\lambda_{l,-i_l}\right)}{n_l-p_l-2}. \tag{16}$$

The equation (17) is directly deduced from (14) and (15) and (19) comes from (16). Finally, we have the equality

$$\left([\mathbf{H}_s]_{i_l}-[\mathbf{R}_l]_{[i_l,-i_l]}\mathbf{K}_s[\mathbf{H}_s]_{-i_l}\right)\Sigma_s\left(\mathbf{H}_s]_{i_l}-[\mathbf{R}_l]_{[i_l,-i_l]}\mathbf{K}_s[\mathbf{H}_s]_{-i_l}\right)'=\varsigma_l, \tag{17}$$

with $\Sigma_s=([\mathbf{H}_s']_{-i_l}\mathbf{K}_s[\mathbf{H}_s]_{-i_l})^{-1}$, $\varsigma_l=\mathrm{u}_l^2\Sigma_s$ and $\mathrm{u}_l=\left[\mathbf{R}_l^{-1}\mathbf{H}_l\right]_{i_l}/[\mathbf{R}_l^{-1}]_{i_l,i_l}$. The equations (13) and (17) allow to obtain (18).

References

Bates, R. A., Buck, R., Riccomagno, E., and Wynn, H. (1996). Experimental design and observation for large systems. *Journal of the Royal Statistical Society, Series B (Statistical Methodology)*, 58 (1):77–94.

Dubourg, V., Sudret, B., and J-M, B. (2011). Reliability-based design optimization using kriging surrogates and subset simulation. *Structural and Multidisciplinary Optimization*, 44(5):673–690.

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

Fang, K.-T., Li, R., and Sudjianto, A. (2006). *Design and Modeling for Computer Experiments*. Chapman & Hall - Computer Science and Data Analysis Series, London.

Fasshauer, G. E. and Zhang, J. G. (2007). On choosing "optimal" shape parameters for rbf approximation. *Numerical Algorithms*, 45(1-4):345–368.

Hartigan, J. A. and Wong, M. A. (1979). Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society, Series C (Applied Statistics)*, 28(1):100–108.

Kleijnen, J. and van Beers, W. (2004). Application-driven sequential designs for simulation experiments: Kriging metamodelling. *Journal of the Operational Research Society*, 55:876–883.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics & Probability, Berkeley, University of California Press*, 1:281–297.

Michalewicz, Z. (1992). *Genetic Algorithms + Data Structures = Evolution Programs*. Springer, New York.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Nam-Ho, K. (2010). Adaptative designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132(7):071008–1 – 071008–9.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Cambridge.

Robert, C. and Casella, G. (2004). *Monte Carlo statistical methods (2nd edition)*. Springer Texts in Statistics, Springer Verlag, New York.

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.

Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Santner, T. J., Williams, B. J., and Notz, W. I. (2003). *The Design and Analysis of Computer Experiments*. Springer Verlag, New York.

Stocki, R. (2005). A method to improve design reliability using optimal latin hypercube sampling. *Computer Assisted Mechanics and Engineering Sciences*, 12:87–105.

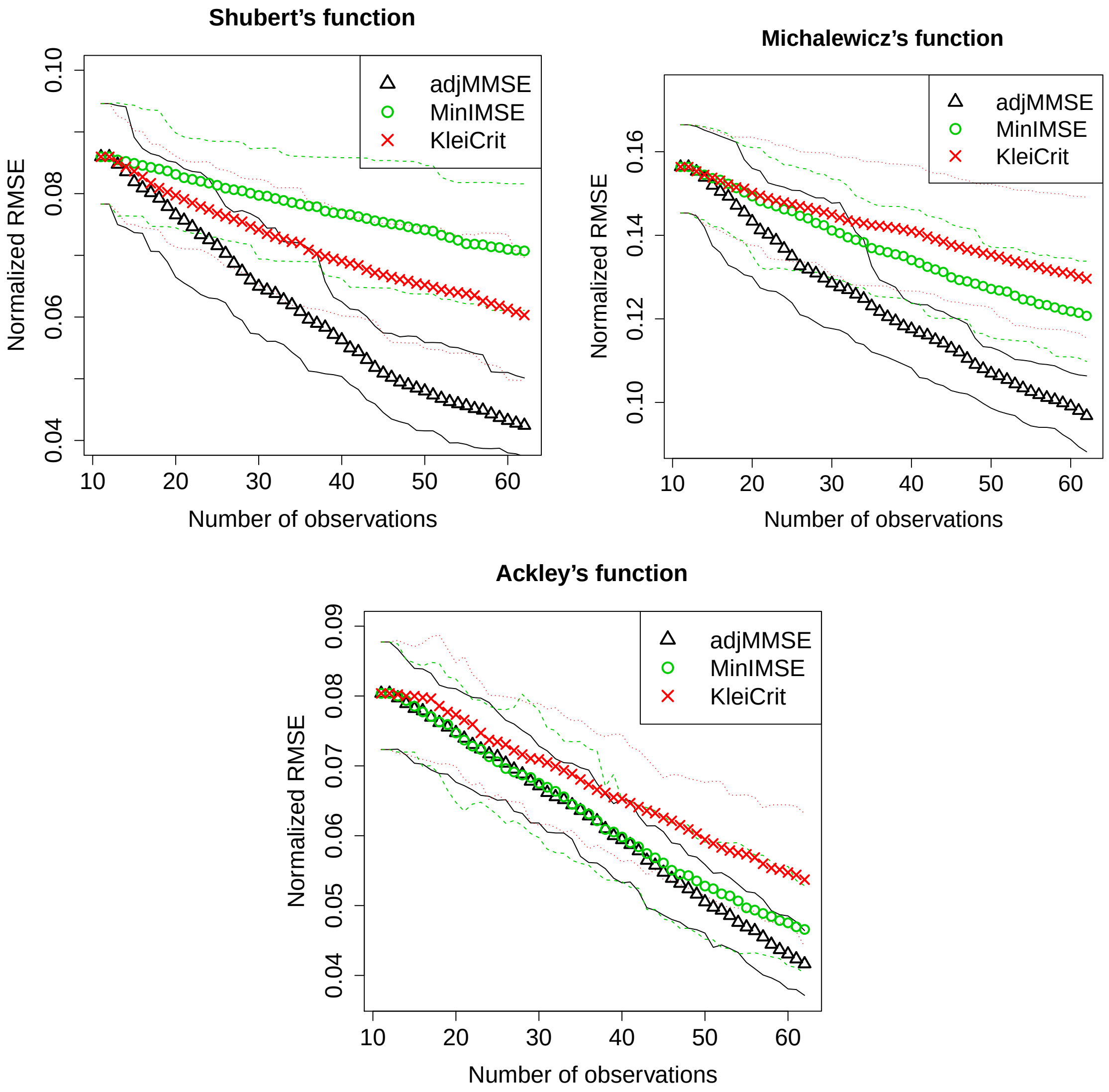

Figure 2: Comparison between 1 point at-a-time sequential kriging criteria on toy examples. The solid lines represent the quantiles of probabilities 10% and 90% of the Normalized RMSE, the dotted lines represent them for the MinIMSE criterion and the dotted lines represents them for the KleiCrit criterion. The means and confidence intervals are computed from 50 different sequential design procedures.

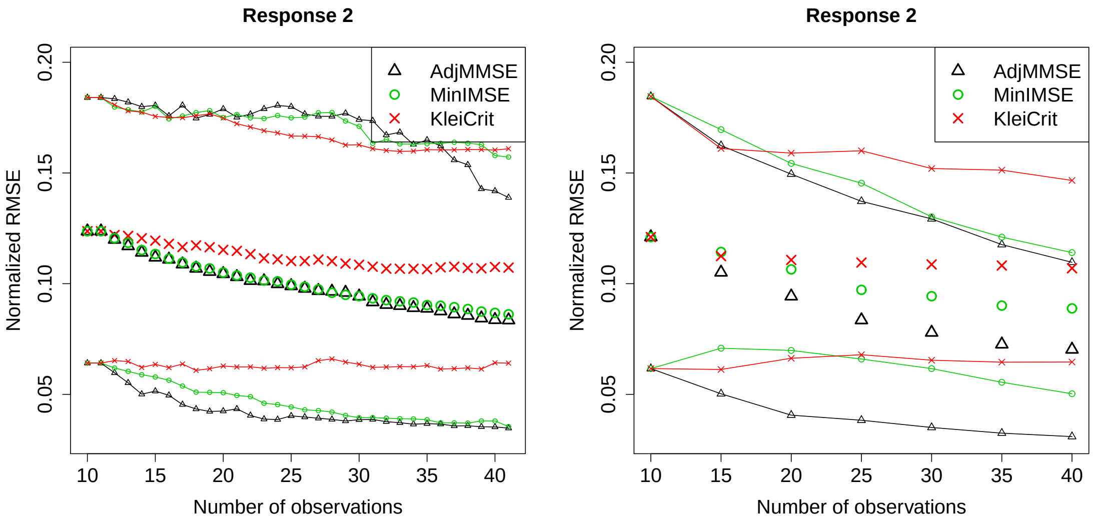

Figure 3: Comparison between 1 point at-a-time sequential kriging criteria (on left) and $q = 5$ points at-a-time sequential kriging criteria (on right) on the spherical tank example. The solid lines represent the quantiles of probabilities 10% and 90% of the Normalized RMSE, the dashed lines represent them for the MinIMSE criterion and the dotted lines represent them for the KleiCrit criterion.