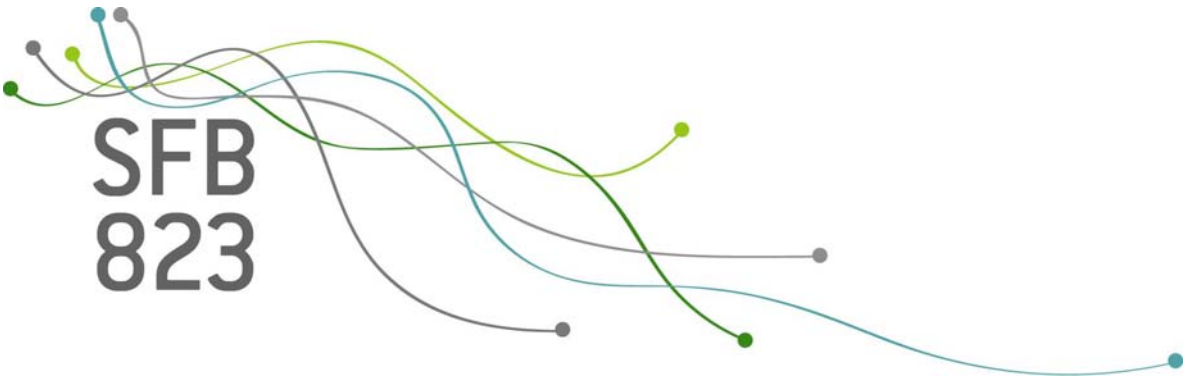
SFB 823

Efficient global optimization: Motivation, variations and applications

Claus Weihs, Swetlana Herbrandt, Nadja Bauer, Klaus Friedrichs, Daniel Horn

Nr. 64/2016

Discussion Paper



Efficient Global Optimization: Motivation, Variations, and Applications

Claus Weihs, Swetlana Herbrandt, Nadja Bauer, Klaus Friedrichs, and Daniel Horn

Abstract A popular optimization method of a black box objective function is Efficient Global Optimization (EGO), also known as Sequential Model Based Optimization, SMBO, with kriging and expected improvement. EGO is a sequential design of experiments aiming at gaining as much information as possible from as few experiments as feasible by a skillful choice of the factor settings in a sequential way. In this paper we will introduce the standard procedure and some of its variants. In particular, we will propose some new variants like regression as a modeling alternative to kriging and two simple methods for the handling of categorical variables, and we will discuss focus search for the optimization of the infill criterion. Finally, we will give relevant examples for the application of the method. Moreover, in our group, we implemented all the described methods in the publicly available R package mlrMBO.

1 Introduction

The overall goal of this paper is global optimization of factor settings by minimizing an objective function. Exemplarily, such objective function may be a measure of a product or algorithm property such as stability, yield, or predictive power. Influencing factors may be ingredients, process settings, or algorithm's parameters. Typical challenges are that the evaluation of the (black box) ob-

Department of Statistics, TU Dortmund University, D-44221 Dortmund, Germany
✉ claus.weihs,swetlana.herbrandt,nadja.bauer,klaus.friedrichs,daniel.horn@tu-dortmund.de

ARCHIVES OF DATA SCIENCE (ONLINE FIRST)
KIT SCIENTIFIC PUBLISHING ISSN 2363-9881
Vol. -, No. -, -

jective function is expensive (in money, storage, animal lives, etc.) and/or time consuming and that many different factor settings have to be studied to find the best settings. The solution studied in this paper is sequential design of experiments in order to gain as much information as possible from as few experiments as feasible by a skillful choice of the factor settings in a sequential way.

Let the design matrix X consist of n settings in the rows for p factors (ingredients/parameters) in the columns: $\mathbf{X} = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} \end{pmatrix}$. Let the objective function $f(\mathbf{x})$ be a real-valued black box function with values $f(\mathbf{x}_i) = y_i$ for the ith setting $\mathbf{x}_i = (x_{i1} \ldots x_{ip})^T$ of the p factors. Let the vector of values of the objective be $\mathbf{y} = (y_1 \ldots y_n)^T$.

In classical design of experiments, an optimization design might be a bigger response surface design like a central-composite (Myers et al, 2009) or a space filling design (McKay et al, 1979; Cioppa and Lucas, 2007), or the like, directly evaluated by experiments or simulations. The results would be evaluated by means of a regression model, estimating the unknown parameters. Finally, optimum factor levels would be determined regarding the estimated model (Myers et al, 2009). Unfortunately, this procedure has at least two very important drawbacks: the time for carrying-out the, typically, many experiments and the need to repeat the whole procedure in case of a poor fit.

The iterative approach called Efficient Global Optimization (EGO) (Jones et al, 1998) also starts with an initial design which is evaluated directly by experiments or simulations. However, this design is typically a space filling design, which is much smaller than the classical response surface designs. Then, the experimental results are analyzed by means of an approximate (so-called surrogate) model, which can be evaluated much faster than an experiment or a simulation. A new design point is found by means of the optimization of a (so-called infill) criterion evaluating the estimated approximate model. This new design point is the next parameter setting directly evaluated by an experiment or a simulation. This step of model estimation and generation of a new design point is iterated until a stop criterion is fulfilled. The selection of surrogate models and infill criteria will be discussed in detail in this paper.

Example 1. Consider only one factor x (i.e., $p = 1$), $x \in [0,7]$, and the true (in practice unknown) objective function $y = f(x) = sin(x) + 2sin(2x) + sin(3x)$. The optimal factor value is $x^* \approx 5.549$. In the classical approach we typically assume a quadratic regression model. Here, we even assume a higher order

model, namely of 8th (!) order, to improve approximation: $y = \beta_0 + \sum_{k=1}^{8} \beta_k x^k + \varepsilon$. As a design we use 16 equidistant points (see Figure 1). Then, the best evaluated setting is $x = 5.6875$ (minimum of the design points) and the best predicted setting $\hat{x}^* \approx 5.586$ (calculated minimum of estimated red curve in Figure 1). Therefore, the predicted optimal argument is very close to the true optimal factor value $x^* \approx 5.549$ (see above, minimum of true black curve in Figure 1). The question is: Can EGO produce even better results?

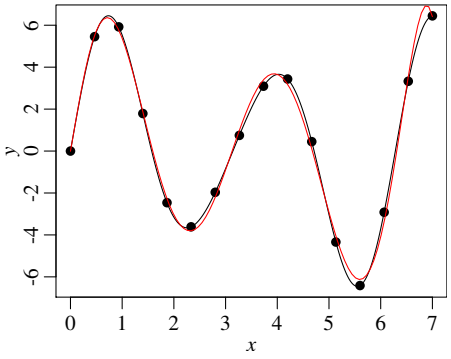


Fig. 1 Example function with 16 equidistant design points (black dots), black line = true function, red line = estimated function. Note, how near the minima of the true and the estimated curves are.

The paper is structured as follows. In Section 2, the EGO procedure will be introduced. In Section 3, variations of this procedure will be discussed. Here, we will especially introduce our new variations of the EGO procedure, namely regression as a modeling alternative to kriging, focus search for the optimization of the infill criterion, and two simple methods for the handling of categorical variables. Section 4 presents applications to parameter tuning, onset detection, and cutting optimization. The paper finishes with a conclusion.

2 Efficient Global Optimization

In this section, the EGO procedure will be described in detail. See Figure 2 for an overview over the whole procedure. The steps of this procedure will be discussed in the following subsections.

2.1 Initial Design

In step 1 of the EGO procedure, the initial design should be chosen small but nevertheless covering the whole region of interest from the outset. Also, if possible, all factors should be considered which might improve the objective. From

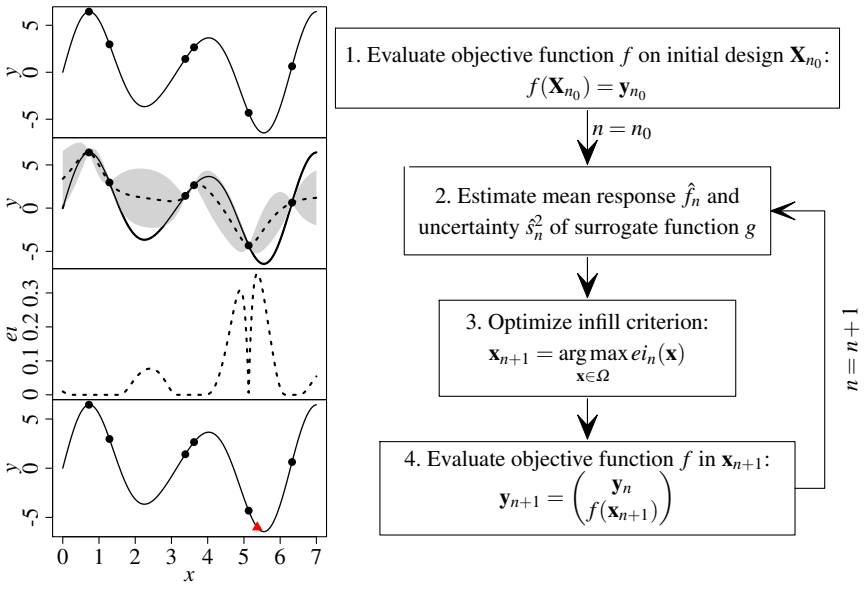


Fig. 2 Global optimization steps: On the left, an example is given (predetermined design points as black dots, new design point as a red triangle); on the right, the steps are briefly described (see the main text for a detailed explanation).

now on, we assume that the region of interest $\Omega = \Omega_1 \times \ldots \times \Omega_p$ of the factor vector $\mathbf{x} = (x_1 \ldots x_p)^T$ is restricted by box constraints $\Omega_j = [x_{j,\text{lower}}, x_{j,\text{upper}}] \subset \mathbb{R}$.

In order to cover the whole region of interest to search for the best setting, typically, a space-filling design like a Latin hypercube design is used (McKay et al, 1979). A *Latin Hypercube Design* (LHD) is constructed so that each factor has the same number of levels. For each level of each factor there is exactly one design point. The idea is to divide the range of the p characteristics into n_0 equally probable intervals, taking the center as the design level. Then, n_0 design points are placed to satisfy the Latin hypercube requirements. Typically, the design is filled by a random permutation of levels factor by factor. Note that the number of divisions, n_0, is assumed equal for each factor. Also note that this design does not require more design points for more factors. This independence is one of the main advantages of this design. For the number n_0 of initial design points, typically $n_0 \in \{5p, 6p, \ldots, 10p\}$ is recommended.

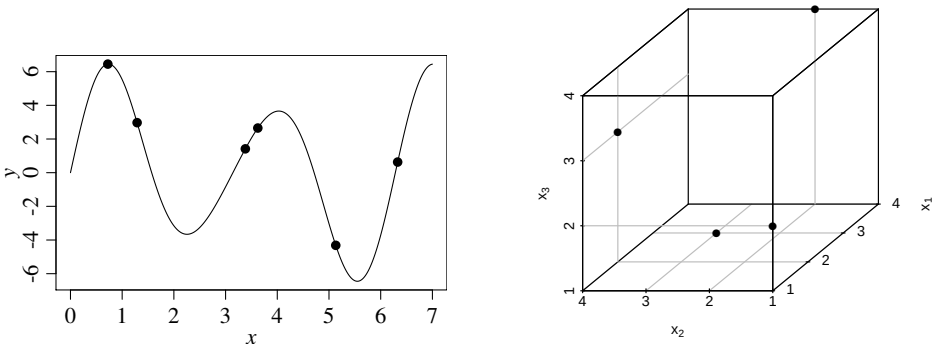


Fig. 3 Sine example function with 6 random design points (left); Latin hypercube design with 4 points (right).

Example 2 (Example 1 cont.). In the above example the region of interest is $\Omega = [0,7]$. Let the initial design include $n_0 = 6p = 6$ points from a random Latin hypercube with $\mathbf{X}_6^T = (5.13\ 3.38\ 1.29\ 3.62\ 6.33\ 0.72)$ and $\mathbf{y}_6^T = (-4.32\ 1.42\ 2.97\ 2.65\ 0.63\ 6.45)$.

Example 3. An LHD for $p = 3$ factors with $n = 4$ design points in $\Omega_j = \{1,2,3,4\}$ may look like this: $\mathbf{X}_4 = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \mathbf{x}_3^T \\ \mathbf{x}_4^T \end{pmatrix} = \begin{pmatrix} 1\ 1\ 2 \\ 2\ 4\ 3 \\ 4\ 2\ 4 \\ 3\ 3\ 1 \end{pmatrix}$ (cp. Figure 3 (right)). Obviously, each factor level appears once for each of the 3 factors.

2.2 Surrogate Function

In step 2 of the EGO procedure, a reasonable approximation $g(\mathbf{x})$ of the unknown objective function is looked for. With this approximation we predict the objective for each factor setting. In order to determine this approximation, we consider an approximation model, also called surrogate model or meta model, $g(\mathbf{x})$. For fast evaluation, we assume that this model is of simple form.

We first concentrate on *ordinary kriging* (Cressie, 1988) with the surrogate model $g(\mathbf{x}) = \mu + Z(\mathbf{x})$. The constant μ can be interpreted as a global mean, and $Z(\mathbf{x})$ is a Gaussian process with $E(Z(\mathbf{x})) = 0$ and a stationary covariance $Cov(Z(\mathbf{x}), Z(\mathbf{x}')) = \sigma^2 \rho(\mathbf{x} - \mathbf{x}', \psi)$ with, e.g., the *Matérn 3/2* kernel function

$\rho(\mathbf{x}-\mathbf{x}',\psi) = (1+\sqrt{3}\frac{||\mathbf{x}-\mathbf{x}'||}{\psi})\exp(-\sqrt{3}\frac{||\mathbf{x}-\mathbf{x}'||}{\psi})$ (Matérn, 1986). The constant σ^2 can be interpreted as the global variance, ψ as a scaling parameter, and the Matérn kernel as a correlation function. A Gaussian process is a statistical model for observations in a continuous input space like time or space, where every point is associated with a normally distributed random variable and every finite collection of those random variables has a multivariate normal distribution. As a measure of the similarity between points, their covariance is used, specified by a kernel function (we use the Matérn 3/2 kernel function). Prediction is not just a point estimate, but also includes uncertainty information being a one-dimensional Gaussian distribution itself (called the marginal distribution at that point). For more information on Gaussian processes see, e.g., Adler and Taylor (2007). Unknown parameters are μ, σ^2, and ψ. To estimate these parameters, we only rely on the observations in the n points already evaluated by the objective function: $\mathbf{y}_n = (y_1 \dots y_n)^T$.

Obviously, in ordinary kriging $\mathbf{g} = (g(\mathbf{x}_1)\dots g(\mathbf{x}_n))^T \sim \mathcal{N}(\mathbf{1}\mu, \sigma^2 R(\psi))$ is valid with correlation matrix $R(\psi) = (\rho(\mathbf{x}_i - \mathbf{x}_j, \psi))_{i,j=1,\dots,n}$, where $\mathbf{1} \in \mathbb{R}^n$ and $\mathbf{1}^T = (1\dots 1)$. Using normality and assuming *interpolation*, i.e., $(g(\mathbf{x}_1)\dots g(\mathbf{x}_n)) = (y_1 \dots y_n)$, the *likelihood function* looks like:

$$L(\mu, \sigma^2, \psi) = \frac{1}{\sqrt{(2\pi)^n \sigma^{2n} det(R)}} \exp(-\frac{1}{2\sigma^2}(\mathbf{y}-\mathbf{1}\mu)^T R^{-1}(\mathbf{y}-\mathbf{1}\mu)),$$

where $det(R)$ is the *determinant* of R. *Maximum likelihood estimation* of the unknown parameters (see, e.g., Mood (1950)) corresponds to:

$$\hat{\mu} = \arg\max_{\mu} L(\mu, \sigma^2, \psi) = \frac{\mathbf{1}^T R^{-1} \mathbf{y}_n}{\mathbf{1}^T R^{-1} \mathbf{1}},$$

$$\hat{\sigma}^2 = \arg\max_{\sigma^2} L(\hat{\mu}, \sigma^2, \psi) = \frac{1}{n}(\mathbf{y}_n - \mathbf{1}\hat{\mu})^T R^{-1}(\mathbf{y}_n - \mathbf{1}\hat{\mu}), \text{ and}$$

$$\hat{\psi} = \arg\max_{\psi} L(\hat{\mu}, \hat{\sigma}^2, \psi).$$

One problem is left, though, how to predict the objective function? The surrogate prediction function is realized as a *linear unbiased predictor*, namely $E(g(\mathbf{x})|g(\mathbf{x}_i) = f(\mathbf{x}_i) = y_i, i = 1,\dots,n) = \lambda(\mathbf{x})^T \mathbf{y}_n = \sum_{i=1}^n \lambda_i(\mathbf{x}) y_i$ with $E(\lambda(\mathbf{x})^T \mathbf{g}) = \lambda(\mathbf{x})^T \mathbf{1}\mu = \mu = E(g(\mathbf{x}))$, leading to $\hat{\lambda}(\mathbf{x})^T \mathbf{1} = 1$.

Optimal weights $\lambda(\mathbf{x})$ are received by minimizing the prediction variance:

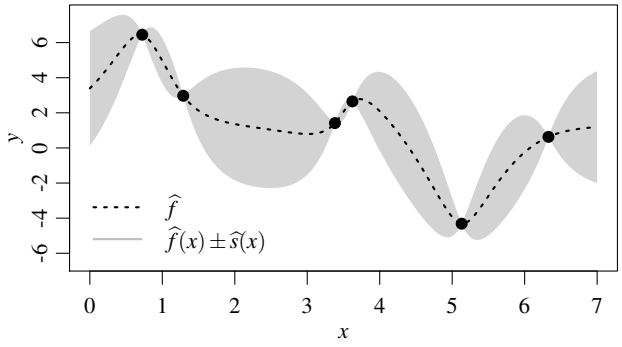


Fig. 4 Surrogate function with design points; predicted uncertainty region in grey.

$$\begin{aligned} s^2(\mathbf{x}) &= var(g(\mathbf{x})|g(\mathbf{x}_i) = y_i, i = 1,\ldots,n) \\ &= E((\lambda(\mathbf{x})^T\mathbf{g} - g(\mathbf{x}))^2|g(\mathbf{x}_i) = y_i, i = 1,\ldots,n) \\ &= \sigma^2(\lambda(\mathbf{x})^T R\lambda(\mathbf{x}) - 2\lambda(\mathbf{x})^T r(\mathbf{x}) + 1) = min! \end{aligned}$$

This yields $\hat{\lambda}(\mathbf{x}) = \hat{R}^{-1}\left(\hat{r}(\mathbf{x}) + \mathbf{1}\frac{1-\mathbf{1}^T\hat{R}^{-1}\hat{r}(\mathbf{x})}{\mathbf{1}^T\hat{R}^{-1}\mathbf{1}}\right)$ with $r(\mathbf{x}) = (\rho(\mathbf{x}_i - \mathbf{x};\psi))_{i=1,\ldots,n}$.

This leads to the following expressions for the predictor and its variance:

$$\begin{aligned} \hat{f}(\mathbf{x}) &= \hat{\lambda}(\mathbf{x})^T\mathbf{y}_n = \hat{\mu} + \hat{r}(\mathbf{x})^T\hat{R}^{-1}(\mathbf{y}_n - \mathbf{1}\hat{\mu}) \text{ and} \\ \hat{s}^2(\mathbf{x}) &= \hat{\sigma}^2\left(1 - \hat{r}(\mathbf{x})^T\hat{R}^{-1}\hat{r}(\mathbf{x}) + \frac{(1-\hat{r}(\mathbf{x})^T\hat{R}^{-1}\mathbf{1})^2}{\mathbf{1}^T\hat{R}^{-1}\mathbf{1}}\right). \end{aligned}$$

Notice that kriging uses interpolation so that $\hat{f}(\mathbf{x}_i) = y_i$ and $\hat{s}^2(\mathbf{x}_i) = 0, i = 1,\ldots,n$. A graphical illustration of the kriging situation can be seen in Figure 4 showing the prediction $\hat{f}(\mathbf{x})$ together with the grey uncertainty region $\hat{f}(\mathbf{x}) \pm \hat{s}(\mathbf{x})$.

2.3 Next Design Point

In step 3, the surrogate function based prediction and its uncertainty are constructed to select the next point for the evaluation. This is realized by the optimization of an infill criterion using information of the current model balancing between regions of low mean prediction (exploitation) and of high standard error (exploration). The most typical criterion for the choice of the next design

point is the maximization of the *expected improvement* conditionally on the observations, as explained now.

Let $\hat{f}_n(\mathbf{x})$ be the surrogate prediction and $\hat{s}_n^2(\mathbf{x})$ the prediction uncertainty based on the first n evaluations of f. By construction, the estimated surrogate function $\hat{g}(\mathbf{x})$ follows a normal distribution: $\hat{g}(\mathbf{x}) \sim \mathcal{N}(\hat{f}_n(\mathbf{x}), \hat{s}_n^2(\mathbf{x}))$. Let ϕ_g be the corresponding probability density function and Φ_g the cumulative distribution function.

Let the actual best value be $y_{\min} = \min\limits_{i=1,\ldots,n} y_i = \min\limits_{i=1,\ldots,n} f(\mathbf{x}_i)$. For a point $\mathbf{x}$ and the estimated surrogate $\hat{g}(\mathbf{x})$ an improvement is then given by: $I_n(\mathbf{x}) = \max(y_{\min} - \hat{g}(\mathbf{x}), 0)$. Since $I_n(\mathbf{x})$ is stochastic, consider the expected improvement

$$\begin{aligned} ei(\mathbf{x}) = E(I_n(\mathbf{x})) &= \int_{-\infty}^{+\infty} \max(y_{\min} - y, 0)\phi_g(y)dy \\ &= y_{\min}\Phi_g(y_{\min}) - E(\hat{g}(\mathbf{x})|\hat{g}(\mathbf{x}) \leq y_{\min})\Phi_g(y_{\min}) \\ &= (y_{\min} - \hat{f}_n(\mathbf{x}))\Phi\left(\frac{y_{\min} - \hat{f}_n(\mathbf{x})}{\hat{s}_n(\mathbf{x})}\right) + \hat{s}_n(\mathbf{x})\phi\left(\frac{y_{\min} - \hat{f}_n(\mathbf{x})}{\hat{s}_n(\mathbf{x})}\right), \\ &\quad \text{if } \hat{s}_n(\mathbf{x}) > 0, \text{and } E(I_n(\mathbf{x})) = 0 \text{ otherwise,} \end{aligned}$$

where $\phi(y)$ and $\Phi(y)$ are the density function and the distribution function of standard normal distribution. The next point for evaluation is then defined as: $\mathbf{x}_{n+1} = \arg\max_{\mathbf{x}\in\Omega} E(I_n(\mathbf{x}))$. In order to identify this point, the expected improvement has to be optimized numerically, by means of an Evolutionary Algorithm (EA) (see, e.g., Simon (2013)), as originally, or, e.g., by the search method introduced in Section 3.3.

2.4 Iteration

In step 4 of the EGO procedure, the objective function is calculated in the new design point. Then, the next iteration step is initiated.

Example 4 (Example 1 cont.). In our example, let us now look at the iterations generating the next design points after the initial design of 6 random points. In Figure 5 iterations $1, 2, 7, 10$ are visualized. For each iteration the graphic shows in the upper part the original function (solid line) and its approximation $\hat{f}(x)$ (dashed line) together with its uncertainty region $\pm\hat{s}(\mathbf{x})$ (grey) and in the lower part the expected improvement $ei(x)$. In iteration 1, corresponding to

the first new design point, a point near the present minimum is selected (red triangle). This is also true for iteration 2. This *exploitation* of the same region continues up to iteration 6, whereas in iteration 7 a point far off of this region is selected, leading to an *exploration* for other minima. It is this property that enables the method to leave local minima and find global ones. In iteration 10 the search is stopped since also the competing 'classical' method was based on a *budget* of overall 16 design points.

The true optimal setting is $x_{\text{opt}} \approx 5.549$. With the classical regression model of order 8, the best evaluated setting was $x = 5.6$ with an absolute difference to the optimum x_{opt} of 0.051 and the best predicted setting is $x = 5.586\,(0.036)$.

With our EGO model the best evaluated setting is $x = 5.550\,(0.001)$ and the best predicted setting is $x = 5.515\,(0.034)$ which is even somewhat better than for the 'classical' performance.

2.5 Stopping

Finally, we should discuss the problem 'when to stop iteration?'. Typically, a budget is pre-fixed, i.e., the number of iterations. This might lead to early or late stopping, i.e., to stopping without convergence or to stopping far beyond convergence.

Alternatives can be, e.g., based on expected improvement (Huang et al, 2006). We might stop, if $ei(\mathbf{x})$ is small after n iterations, i.e., $\max\limits_{\mathbf{x}\in\{\mathbf{x}_1,\ldots\mathbf{x}_n\}} ei(\mathbf{x}) < \Delta_s$, where Δ_s = stopping tolerance, or if relative expected improvement is small, e.g., if the magnitude of y is not known: $\frac{\max\limits_{\mathbf{x}\in\{\mathbf{x}_1,\ldots\mathbf{x}_n\}} ei(\mathbf{x})}{\max\{y_1,\ldots,y_n\}-\min\{y_1,\ldots,y_n\}} < \Delta_r$, where Δ_r = relative stopping tolerance.

This is the first variation of the standard method we consider in this paper. In the next section, we will discuss other variations.

3 Variations

Let us now consider variations of the EGO method, particularly for surrogate modeling and the infill criterion. Moreover, details will be discussed on parts

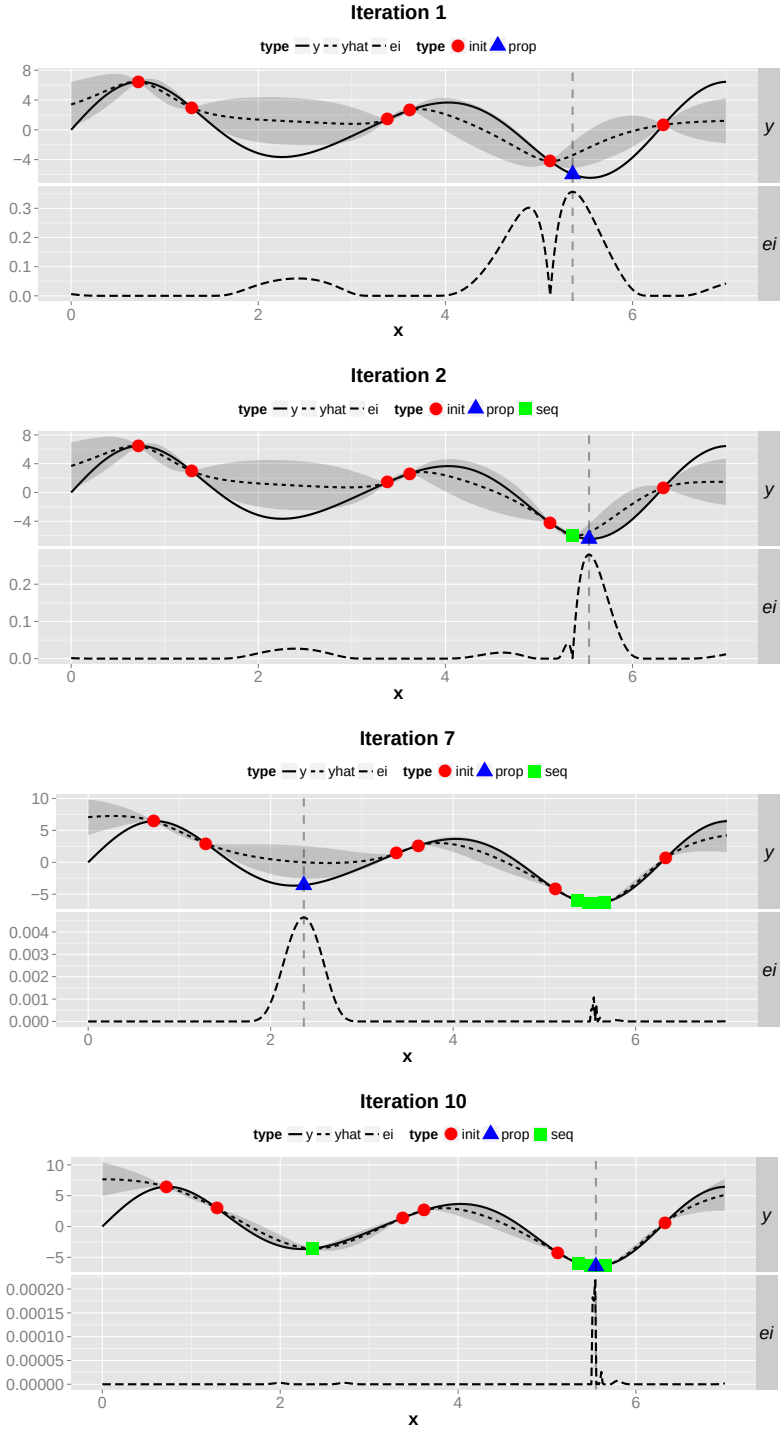


Fig. 5 Iteration steps 1, 2 (exploitation), 7 (exploration) and 10 (last additional point before stopping); upper parts: objective $f(x)$ (solid line), predicted objective $\hat{f}(x)$ (dashed line), design points (red dots = initial design, blue triangle = proposed point, green squares = sequential design points after initial design), predicted uncertainty region in grey; lower parts: $ei(x)$.

of the method not even mentioned yet, namely on the optimization of the infill criterion, and the handling of categorical features.

3.1 Surrogate Models

The choice of the correlation function is one decisive factor for a kriging model. Alternatives to Matérn 3/2 detailed above are, e.g., linear, exponential, Gaussian, and *Matérn 5/2* correlations functions (Rasmussen and Williams, 2006, p. 79ff.). Note that these functions should be as flexible as possible to be able to model any shape of the data.

There are also correlation functions with more than one unknown parameter, e.g., one scaling parameter ψ_j for each influencing factor in a *product correlation rule*: $R(\psi_1,\ldots,\psi_p)=\prod_{j=1}^{p}R(\psi_j)$ (Sasena, 2002).

Also, the ordinary kriging model is generalized, e.g., to the *universal kriging* model, built by a polynomial instead of a constant trend: $g(\mathbf{x})=\sum_{j=1}^{k}\beta_j h_j(\mathbf{x})+Z(\mathbf{x})$ with arbitrary functions h_j and corresponding coefficients β_j (Sasena, 2002). However, a constant term is generally sufficient because of the flexibility of the correlation functions.

Another variation is the dropping of the interpolation property of kriging. Instead, the kriging model is modeled to be noisy, leading to *augmented kriging* (cp. Huang et al (2006)): $g(\mathbf{x})=\mu+Z(\mathbf{x})+\varepsilon$ with $\varepsilon\sim$ i.i.$\mathcal{N}(0,\tau^2)$ assumed to be stochastically independent of $Z(\mathbf{x})$. The overall covariance can then be written as $Cov(Z(\mathbf{x}_i)+\varepsilon_i,Z(\mathbf{x}_j)+\varepsilon_j)=\sigma^2\rho(\mathbf{x}_i-\mathbf{x}_j,\psi)+\tau^2 I(i=j), i,j=1,\ldots,n$, where $I(cond)$ is the indicator function of condition $cond$. Such a model is used in Section 4.3.

As a further modeling alternative, we propose the use of a classical quadratic regression model instead of a kriging model. Then, variation is mainly modeled explicitly as a function of $\mathbf{x}$: $g(\mathbf{x}_i)=\beta_0+\sum_{j=1}^{p}\beta_j x_{ij}+\sum_{j=1}^{p}\sum_{k>j}^{p}\beta_{jk}x_{ij}x_{ik}+\sum_{j=1}^{p}\beta_{jj}x_{ij}^2+\varepsilon_i,\ \varepsilon_i$ i.i.$\mathcal{N}(0,\sigma^2)$. The unknown coefficients $\beta_j,\beta_{jk},\beta_{jj}, j=1,\ldots,p, k>j$, and σ^2 are estimated by means of least squares. Prediction is then given by $\hat{f}(\mathbf{x})=\hat{\beta}_0+\sum_{j=1}^{p}\hat{\beta}_j x_j+\sum_{j=1}^{p}\sum_{k>j}^{p}\hat{\beta}_{jk}x_j x_k+\sum_{j=1}^{p}\hat{\beta}_{jj}x_j^2$ and prediction variance by $\hat{s}^2(\mathbf{x})=\hat{\sigma}^2(1+\mathbf{x}^T(\mathbf{X}_n^T\mathbf{X}_n)^{-1}\mathbf{x})$. Note that this leads to smoothing, not to interpolation. Further note the relationship to the universal kriging model (see above), where the $h_j(\mathbf{x})$ stand for the linear, interaction, and quadratic terms. In universal kriging, for the model residuals a quite general correlation matrix is assumed. In the regression model, the errors are assumed

independent, instead. This model is compared to the standard procedure in Section 4.1.

3.2 Infill Criteria

In case of the augmented kriging model the expected improvement criterion is extended to the *Augmented Expected Improvement* $aei(\mathbf{x})$ (cp. Huang et al (2006)):

$$\begin{aligned} aei(\mathbf{x}) = & \left[(T - \hat{f}_n(\mathbf{x}))\Phi\left(\frac{T-\hat{f}_n(\mathbf{x})}{\hat{s}_n(\mathbf{x})}\right) + \hat{s}_n(\mathbf{x})\phi\left(\frac{T-\hat{f}_n(\mathbf{x})}{\hat{s}_n(\mathbf{x})}\right)\right] \\ & \cdot \ (1 - \frac{\hat{\tau}}{\sqrt{\hat{s}_n(\mathbf{x})^2 + \hat{\tau}^2}}), \text{where } T = \hat{f}_n(\mathbf{x}^*), \\ & \mathbf{x}^* = \arg\min_{\mathbf{x}} (\hat{f}_n(\mathbf{x}) + \kappa \hat{s}_n(\mathbf{x})), \kappa = 1, \text{ typically} \end{aligned}$$

Note that τ^2 is the variance of the error term ε in the augmented kriging model. The term $(1 - \frac{\hat{\tau}}{\sqrt{\hat{s}_n(\mathbf{x})^2+\hat{\tau}^2}})$ is interpreted as a multiplicative penalty term for the expected improvement caused by the extra error term ε in the augmented model. For $\tau = 0$, the penalty term is 1 (no penalty). The bigger τ is, the smaller is the penalty term and the smaller is $aei(\mathbf{x})$.

Besides the expected improvement, there are other proposals for infill criteria. The most popular alternative is to minimize the *lower confidence bound* $lcb(\mathbf{x}) = \hat{f}_n(\mathbf{x}) - \kappa \hat{s}_n(\mathbf{x})$ with a fixed $\kappa > 0$ (Cox and John, 1997).

3.3 Search Methods

The optimization of the infill criterion is typically implemented as an approximative search. In our group, we propose the *focus search* as a global-local optimization of the infill criterion to find the next point $\mathbf{x}_{n+1}$ to be evaluated by the objective black box function f. The focus search will repeatedly start with a coarse LHD with a sequence of refinements near the found optima. The search can be implemented as in Algorithm 1.

Algorithm 1 Focus Search

Global search: Repeat for $s = 1, \ldots, S$ (e.g., $S = 10$)

Local search: Repeat for $j = 1, \ldots, J$ (e.g., $J = 5$)
(a) Sample LHD: $D^j \subset \Omega^j (\Omega^1 = \Omega)$
(b) Infill Criterion, e.g., lcb with $\kappa = 1$:
Candidate point $\mathbf{x}_{*j} = \arg\min_{\mathbf{x} \in D^j} [\hat{f}(\mathbf{x}) - \hat{s}(\mathbf{x})]$
(c) Reduce parameter space: $\Omega^{j+1} \subset \Omega^j$ around $\mathbf{x}_{*j}$
Local candidates: $D_{*s} = \{\mathbf{x}_{*1}, \ldots, \mathbf{x}_{*J}\}$ and $\mathbf{x}_s = \arg\min_{\mathbf{x} \in D_{*s}} [\hat{f}(\mathbf{x}) - \hat{s}(\mathbf{x})]$ best point

Global candidates: Let $D_* = \{\mathbf{x}_1, \ldots, \mathbf{x}_S\}$ be the global candidate set, then $\mathbf{x}_{n+1} := argmin_{\mathbf{x} \in D_*} [\hat{f}(\mathbf{x}) - \hat{s}(\mathbf{x})]$ is the best point to be evaluated next

3.4 Categorical Attributes

One of the disadvantages of standard kriging is its limitation to numerical influential parameters. However, there are extensions to categorical attributes. In the literature, most of the time, extensions of the covariance function to qualitative variables are proposed (see, e.g., Qian et al (2008)). In contrast, we propose two very simple methods, indicated below:

Naïve kriging: Categorical attributes are handled as numerical ones by assigning an integer value to each level. This method is easy to implement, but the order of the levels and the distance between levels are typically artificial.

Dummy kriging: Dummy variables are built for each categorical attribute, e.g., for possible levels A, B, C of attribute x we take $x_A = I_{[x=A]}, x_B = I_{[x=B]}, x_C = I_{[x=C]}$. This method is statistically more correct, but time consuming if categorical attributes have a large number of levels.

For an application, where these two methods are compared, see Section 4.2.

4 Applications

4.1 Parameter Tuning

The first application will be parameter tuning for *Support Vector Machine* (SVM) classification (Schölkopf and Smola, 2002).

To illustrate the method, we use 5 publicly available classification benchmark problems from *OpenML* (Vanschoren et al, 2014) (`www.openml.org`) characterized in Table 1. We would like to compare kriging modeling with

regression modeling in EGO as well as *ei* and *lcb* as infill criteria. As the objective $f(cost, \gamma, \varepsilon)$ we use the error rate of the nonlinear SVM classifier using the *Radial Basis Function* (RBF) kernel (as provided in the library LIBSVM, cp. Chang and Lin (2011)) with the $p = 3$ parameters *cost*, penalizing errors, γ in the RBF kernel, and ε in the loss function. In this way, optimal tuning of the SVM parameters is realized.

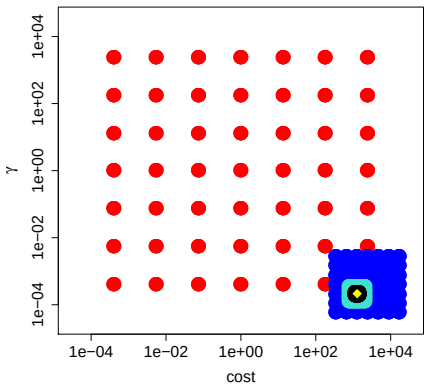


Fig. 6 Focused Grid Search for the 2 SVM parameters *cost* and γ. First, the focus is on the bottom right corner (blue region). Then, the focus is shifted somewhat to the left (light-blue region).

Since the ground truth is not available, we apply a somewhat naïve approach for generating a near optimum of the objective being the accuracy of the model. We use a Focused Grid Search over the 3 parameters $cost, \gamma, \varepsilon$. We consider $cost \in 2^{[-15,15]}$, $\gamma \in 2^{[-15,15]}$, $\varepsilon \in 2^{[-13,-1]}$, vary all parameters on a logarithmic scale, and use $7 \cdot 7 \cdot 5 = 245$ points, equidistant in each dimension, as the starting grid. We use J = 4 Focus Search iterations (and S = 1, i.e., only one fixed starting grid)(cp. Section 3.3) leading to 980 function evaluations. See Figure 6 for an illustration.

The performance corresponding to a found $(cost, \gamma, \varepsilon)$ setting is assessed via a 2:1 train-test-split. The accuracy of the optimum found by the Focused Grid Search is taken as the ground truth for a comparison with the best accuracies found by the different versions of the EGO procedure. The results of the Focused Grid Search can be found in Table 1.

Table 1 Data sets used in SVM parameter tuning characterized by the number of observations (obs), the number of variables (m), and the best accuracy reached with the focused grid search for the shown parameter settings.

Dataset	obs	m	Best Accuracy	*cost*	γ	ε
spambase	4 601	58	0.059	1.38e+03	2.06e-04	1.28e-04
wilt	4 829	6	0.007	2.34e+03	2.19e-02	3.29e-04
ada_agnostic	4 562	49	0.182	1.56e+00	3.31e-05	1.98e-01
eeg-eye-state	14 980	15	0.095	2.37e+04	7.23e-01	8.73e-02
MagicTelescope	19 020	12	0.001	1.71e+04	5.52e-02	1.94e-04

For the evaluation of the EGO procedures, we count the number of function evaluations until EGO reaches the same performance as the grid search. We start with a random LHS with $4p = 12$ points (in order to make sure that all parameters are estimable with quadratic regression), vary the infill criterion (*ei* and *lcb*) and the surrogate model (kriging and quadratic regression), and consider 5 replications. We say that EGO failed if the solution was not reached after 500 iterations since we expect that one EGO step is roughly a factor of two slower than one pure function evaluation.

In Table 2, the iteration characteristics of EGO are reported for the different data sets and method variants, namely the median of the iteration counts of successful EGO runs and the number of successful EGO replications in brackets. In Table 3, the differences of the best optimizations runs to the grid search performance are reported.

Table 2 Median iteration counts of SVM parameter tuning (number of successful KM or QM replications), KM = Kriging Model, QM = Quadratic regression Model

Dataset	ei + KM	lcb + KM	ei + QM	lcb + QM
spambase	30 (5)	48 (5)	- (0)	- (0)
wilt	18 (5)	15 (5)	17 (5)	18 (5)
ada_agnostic	35 (1)	80 (3)	- (0)	- (0)
eeg-eye-state	45 (5)	43 (5)	84 (2)	260 (2)
MagicTelescope	29 (5)	39 (5)	16 (1)	15 (2)

Table 3 Differences of the best optimization results to the grid search optimum, KM = Kriging Model, QM = Quadratic regression Model

Dataset	Grid Search	ei + KM	lcb + KM	ei + QM	lcb + QM
spambase	0.059	-6.52e-04	-6.52e-04	+6.52e-04	+6.52e+04
wilt	0.007	-6.20e-04	-6.20e-04	-6.20e-04	-6.20e-04
ada_agnostic	0.182	0	-6.57e-04	+1.31e-03	+1.31e-03
eeg-eye-state	0.095	-1.60e-03	-4.00e-04	-2.00e-04	+2.00e-4
MagicTelescope	0.001	0	-1.58e-04	0	0

Overall, the results can be interpreted as follows. When comparing EGO and Grid Search, EGO wins by a clear margin. If EGO is able to find the optimal target function value, it is around 20 times faster. If EGO does not reach the optimal target value, it is only slightly off-target.

When comparing *ei* and *lcb*, no real winner can be identified. There are only small differences, each method is better in some cases. However, *lcb* could

be preferable since this criterion does not rely on the validity of the normal distribution.

When comparing the results of regression and kriging, kriging is distinctly superior. Regression fails very often (only 17 successful replications overall in contrast to 44 for kriging). Moreover, regression is not as flexible as kriging. Only if the optimum of the quadratic model is near to the real optimum, the real optimum is reliably found (e.g., for the data set "wilt"). Larger initial designs or the concentration on the best or the newest observations might help regression. This will be studied further.

Let us finish this discussion by a remark on parallel computing. Focused Grid Search can use many nodes in parallel. However, EGO suffers from its iterative behavior, i.e., its evaluation of the 500 points has to be realized one after the other. Therefore, in real time Focused Grid Search was much faster (not in computing time!). A possible way out of this problem is the development of efficient Multi Point Proposals, where multiple proposal points are proposed in parallel (cp. Bischl et al (2014)).

4.2 Onset Detection

The second application deals with the optimization of music onset detection. In this study, a tone onset is the time point of the beginning of a musical note played by a musical instrument. The ability of the detection of tone onsets depends on the instrument playing the tone, e.g., the beginning of piano tones can be much easier identified than the beginning of flute tones (see Figure 7).

The onset detection algorithm we use in this study is specified in Algorithm 2(cp. Figure 8 (left)); for more details see Bauer et al (2015).

As the objective, the F-measure $F = \frac{2c}{2c+f^+ +f^-}$ is used, where $c =$ number of correct detections, $f^+ =$ number of false positive detections, and $f^- =$ number of false negative detections. We use EGO to optimize F corresponding to the 15 parameters, 10 numerical and 5 categorical, of the onset detection in Algorithm 2, i.e., we optimize $f(\text{window size}, \ldots, \text{onset shifting time}) = F(\text{window size}, \ldots, \text{onset shifting time})$.

We use the following EGO settings: kriging with the covariance function Matérn 3/2 as the surrogate model, *ei* as the infill criterion, focus search with $S = 3$, $J = 5$, and size of $D_j = 10'000$ points as the infill optimizer, LHS with $5d = 75$ points as the initial design, and $20d = 300$ iterations. Note that 5

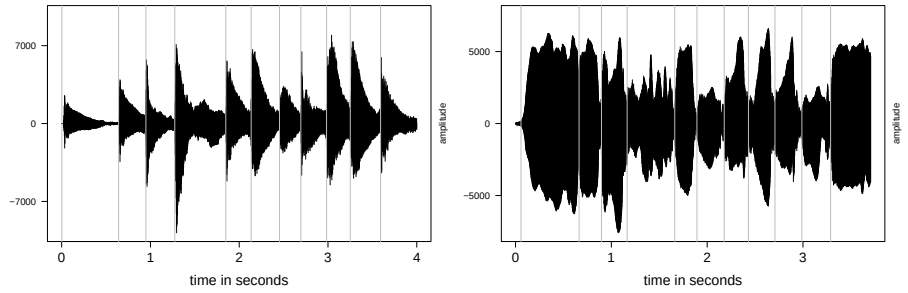


Fig. 7 Time series of the same music piece played by piano (left) and flute (right); tone beginnings are marked by vertical lines.

Algorithm 2 Onset detection algorithm

Input: Audio data in WAVE format with sampling rate of 44.1 kHz

1: Windowing: Split signal into small windows (parameters: window size 512-4096 samples and shift to next window (called hop size) (0–90% of the window size))
2: Pre-processing: Pre-process the data (parameters: windowing function (Hanning, Gauss, ...), spectral filter (yes /no), log-scale (yes /no), and log-parameter (1 – 20))
3: Detection function: Compute in each window an onset detection function (odf) (parameters: odf function (amplitude increase, spectral flux, ...) and exponential smoothing with a parameter in $[0, 1]$)
4: Peak-picking: Threshold the smoothed odf (parameters: moving function (mean, median, ...), threshold multiplier (1 – 3), threshold additive term (0 – 10), and threshold time back (0 – 0.5 seconds))
5: Onset localization: Localize the tone onsets (parameters: onset time back (0 – 0.5 seconds), min. distance between onsets (0 – 0.05 seconds), and onset shifting time (-0.01 – 0.02 seconds))

Output: F-value

categorical parameters have to be optimized. Therefore, we will compare naïve kriging and dummy kriging (cp. Section 3.4).

As the validation design, we use 321 hand labeled music data files in WAVE format (IBM and Microsoft, 1991) as the learning data set, 2/3 data for training and 1/3 for testing in the train-test-approach, and 30 replications (subsampling). The idea is to find the best setting of the parameters in Algorithm 2 on the training data and evaluate it on the test data. As the objective, F-values on the test data are used.

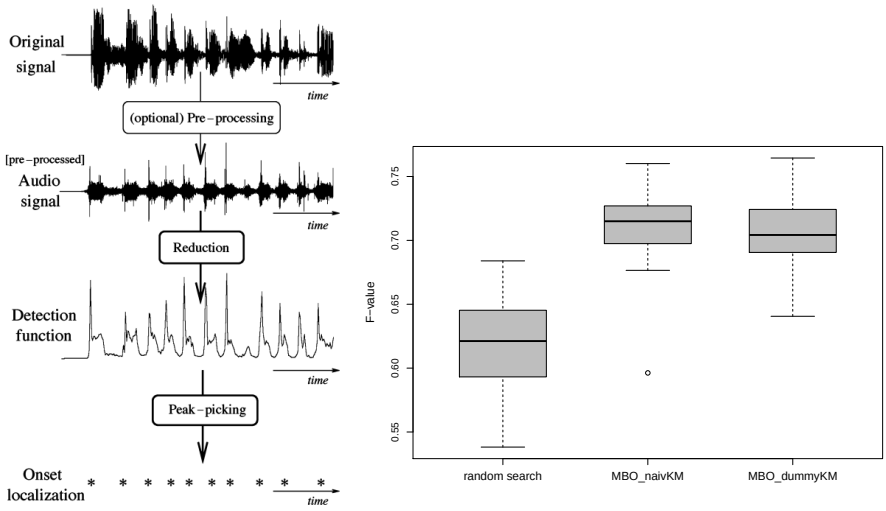


Fig. 8 Common onset detection procedure (left) (see Algorithm 2)(Bello et al, 2005); random search compared to EGO with naïve kriging and dummy kriging (right).

The resulting distribution of F-values can be found in Figure 8 (right). EGO variants clearly dominate random search based on an LHS design of size $25d$. Naïve kriging shows slightly better results than dummy kriging.

4.3 Cutting Optimization

The goal of the third application is the optimization of a cutting process with a diamond tipped drill core bit. We start with a single diamond scratch test, wanting to find settings of the process and production factors which minimize work to reach a certain drilling depth. For this, we first optimize a simulation model by means of parameter adjustment with EGO, and then optimize the process and production factors using the optimal simulation output (cp. Herbrandt et al (2016)).

In order to optimize a simulation model we first need to observe the real process. For this, we carry out a block design with 60 trials each for the 5 materials cement, basalt, concrete, steel, and reinforced concrete. We base on 5 samples per material. We aim to achieve $80\,\mu m$ total drilling depth. We compare 4 feed speed v_f (mm/min) settings and 4 cutting speed v_c (m/min)

settings. 12 drilling diameters (mm) per material sample are used, where the same speed settings (v_c, v_f) are treated in pairs of adjacent radii each. This leads to minimum 2 replicates per setting (v_c, v_f) (cp. Figure 9 for an example). Overall, we use 5 blocks (samples) with 6 trials (radii) each and speed settings from a full factorial design. The design is D-optimized, twice replicated in (v_c, v_f). The outputs are the forces (F_x, F_y, F_z) in all 3 dimensions. The normal force in Figure 9 corresponds to F_z. The sampling rate is 200'000 Hz.

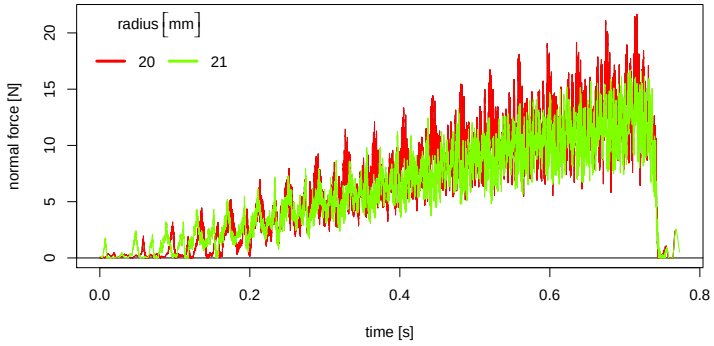


Fig. 9 Time series of normal forces on cement for the speed combination $v_c = 193.5\ \frac{m}{min}$, $v_f = 7\ \frac{mm}{min}$; red: radius = 20 mm, green: radius=21 mm

Let us now optimize a simulation model for the forces given as a weighted sum of removed volumes v and characteristic material values z:
$F(r, g_z, g_v, \mu_c, \sigma^2_{Yc}, \tau^2_c, \psi_c, p_c, q_c) = \frac{g_z}{r} z(\mu_c, \sigma^2_{Yc}, \tau^2_c, \psi_c) v(p_c, q_c) + \frac{g_v}{r} v(p_c, q_c) + \varepsilon$.

For the model of the removed volume v we assume that the cutting diamond has the shape of a pyramid and the cutting line has the profile of a triangle. The triangle size depends on the current observation, the cutting depth per revolution, and the stochastic brittleness of the material simulated by a restricted $Beta(p_c, q_c)$-distributed deviation from the perfect cutting profile. The removed volume of one observation is the volume between two consecutive triangles.

For the model of the characteristic material values z we assume that forces needed to remove material vary over the work piece. The characteristic material values (e.g. hardness) of the triangle points are sampled from a Gaussian random field $Z(\mu_c, \sigma^2_{Yc}, \tau^2_c, \psi_c)$. The material value of one observation is the mean value of two consecutive triangles.

As the objective function $f\left(F_z\left(r,\nu_c,\nu_f\right),F\left(r,\mathbf{x}\right)\right)$ for EGO optimization, we use the mean deviation between modeled forces $F\left(r,\mathbf{x}\right)$ and observed forces $F_z\left(r,\nu_c,\nu_f\right)$ corresponding to the three characteristics slope, range, and spectrum (Herbrandt et al, 2016). For force modeling, we use the same radii r as for the realization of observed forces. The objective is stochastic because of the stochastic character of the force model. The unknown parameters are $\mathbf{x}=\left(g_z\ g_v\ \mu_c\ \sigma_{Yc}^2\ \tau_c^2\ \psi_c\ p_c\ q_c\right)^T$ of the above force function. So we have $p=8$ unknown parameters. The initial design is taken to be a random Latin hypercube with $n_0=10p=80$ points, the number of iterations is set to be 720. This leads to a total number of function evaluations of $100p=800$. As the surrogate function we use the augmented (noisy) kriging model (cp. Section 3.1), as the infill criterion the augmented expected improvement (cp. Section 3.2). Please compare the pseudocode in Algorithm 3. — Please distinguish the Gaussian random field $Z(\mu_c,\sigma_{Yc}^2,\tau_c^2,\psi_c)$ for the characteristic material values (see above) from the Gaussian random field with parameters $\theta=\left(\mu\ \sigma^2\ \tau^2\ \psi_1\ \ldots\ \psi_8\right)^T$ used in kriging. — Model parameters are optimized individually for each of the 16 speed combinations (ν_c,ν_f) and each material (in our illustrations we take cement).

Algorithm 3 Optimization of force model

Input: Observed normal forces F_z with speed settings $\left(\nu_c,\nu_f\right)$

Initial Design:

(a) Sample 80 parameter settings $\mathbf{x}_1,\ldots,\mathbf{x}_{80}\in\mathbb{R}^8$ from a random Latin hypercube, $\mathbf{x}_i=\left(g_{zi}\ g_{vi}\ \mu_{ci}\ \sigma_{Yci}^2\ \tau_{ci}^2\ \psi_{ci}\ p_{ci}\ q_{ci}\right)^T$

(b) Evaluate objective function f in $\mathbf{x}_1,\ldots,\mathbf{x}_{80}$: $f\left(F_z\left(r,\nu_c,\nu_f\right),F\left(r,\mathbf{x}_i\right)\right)=y_i, i=1,\ldots,80$, where f is the mean deviation between modeled and observed forces corresponding to the three characteristics slope, range, and spectrum

EGO: Repeat for $i=80,\ldots,799$

Surrogate Function:

(1) Estimate the augmented kriging parameters $\theta=\left(\mu\ \sigma^2\ \tau^2\ \psi_1\ \ldots\ \psi_8\right)$ (see Section 3.1), where the ψ_j are the Matérn parameters for the individual dimensions

(2) Determine the augmented kriging prediction $\hat{f}_i\left(\hat{\theta}\right)$ conditional on $y_1,\ldots,y_i$

(3) Determine the augmented kriging uncertainty $\hat{s}_i^2\left(\hat{\theta}\right)$ conditional on $y_1,\ldots,y_i$

Infill Criterion: Maximize the augmented expected improvement $aei\left(\mathbf{x}\right)$ in Section 3.2 with the focus search in Algorithm 1: $\mathbf{x}_{i+1}=\arg\max_{\mathbf{x}\in\Omega} aei\left(\mathbf{x}\right)$

Evaluation: Evaluate objective function f in $\mathbf{x}_{i+1}$

Output: Optimal force model parameters $\mathbf{x}^\star=\left(g_z{}^*\ g_v{}^*\ \mu_c{}^*\ \sigma_{Yc}^{2\,*}\ \tau_c^{2*}\ \psi_c{}^*\ p_c^*\ q_c^*\right)^T$

Looking at the optimization results, the best fit is achieved for (v_c, v_f) = (270, 7), the worst fit for (v_c, v_f) = (270, 9.5) (cp. Table 4). The goodness of fit depends on the similarity of 'repetitions'. The iteration number corresponding to the optimal parameter setting is typically lower than the budget of 800 iterations. See Figure 10 for an example, where near optimal settings were already found very early, i.e., after less than 250 iterations. For each optimization we obtain the values of the evaluated objective function $y_1, \ldots, y_{800}$ and the optimal parameter vector $\mathbf{x}^* = (g_z{}^* \; g_v{}^* \; \mu_c{}^* \; \sigma^2_{Yc}{}^* \; \tau_c^{2*} \; \psi^* \; p_c^* \; q_c^*)^T$. Since the best model is stochastic, we only check how well an observed realization with this speed combination fits into the area of simulated realizations. See Figure 11 for an example.

Table 4 Optimal values $y_{\min}$ of objective for the different combinations (v_c, v_f) and corresponding iteration numbers $n_{\min}$

v_c	40.5	40.5	40.5	40.5	117	117	117	117	193.5	193.5	193.5	193.5	270	270	270	270
v_f	2	4.5	7	9.5	2	4.5	7	9.5	2	4.5	7	9.5	2	4.5	7	9.5
$y_{\min}$	3.25	3.04	5.64	4.32	1.49	3.81	7.08	4.41	1.66	3.73	1.75	6.95	2.09	4.75	1.45	7.61
$n_{\min}$	328	621	682	299	800	490	427	434	731	171	452	423	122	755	672	261

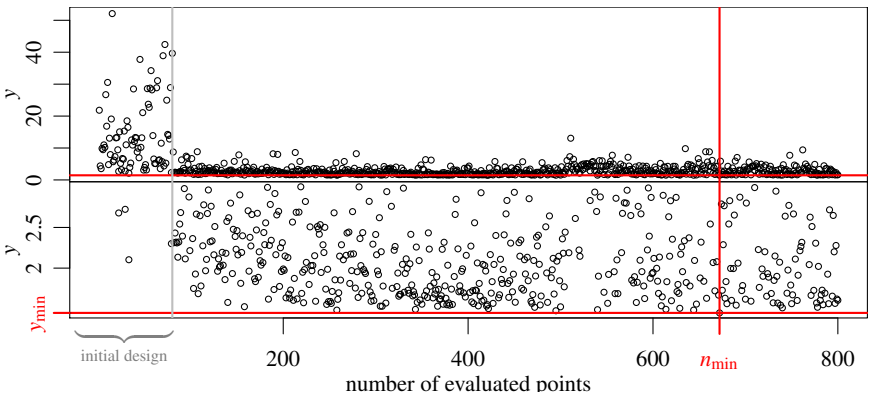


Fig. 10 Optimization path for speed combination $v_c = 270 \; \frac{m}{min}$ and $v_f = 7 \; \frac{mm}{min}$; the end of the initial design is indicated by a vertical line; the minimum y_{min} is reached for iteration number n_{min}. The lower part is a zoom-in of the upper part.

Based on the optimal force models we now would like to optimize the process and production factors. We use the modeled forces to predict a speed-combination with minimal work W_A needed to drill a desired depth A (here:

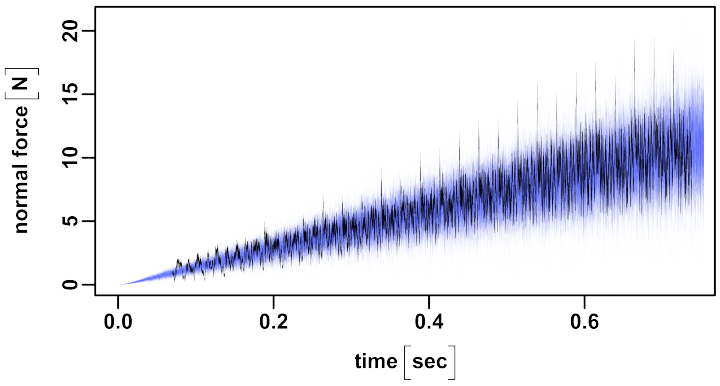


Fig. 11 Time series of 50 realizations of the force model with optimal parameter setting (blue) and observed force (black) for the speed combination $v_c = 270\ \frac{m}{min}$ and $v_f = 7\ \frac{mm}{min}$ on the radius 18 *mm*.

$A = 0.07mm$): $W_A = \int_0^{s_A} F(s)ds$, where $s_A = s_A(r) =$ total distance until reaching depth $A(mm)$, $F(s) =$ expected modeled forces for distance s. For an example, see Figure 12a).

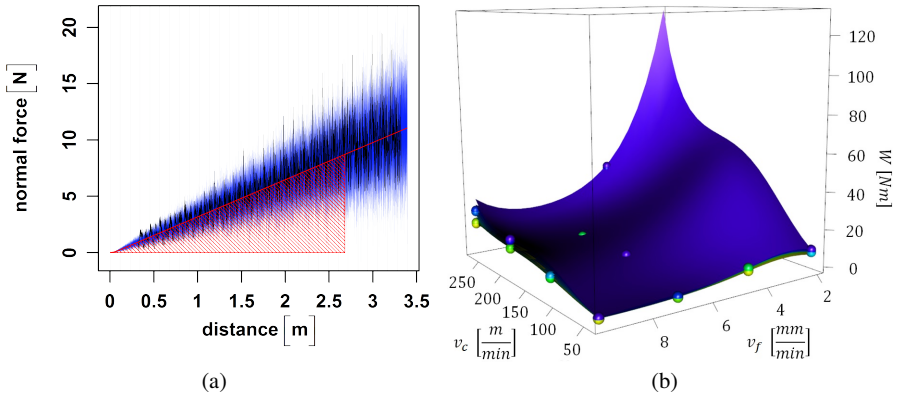


Fig. 12 (a) Realizations of the force model (blue), observed force (black), expected modeled forces $F(s)$ (red line), and the resulting work W_A (red hatched) to reach a total depth of $A = 0.07$ *mm* with $(v_c, v_f, r) = (270\ \frac{m}{min}, 7\ \frac{m}{min}, 18\ mm)$; (b) Surface of the generalized linear model of the resulting work, split by the radii $r = 16$ *mm* (yellow), ..., $r = 27$ *mm* (violet).

The independent variables are v_c, v_f, r. We use a *Generalized Linear Model* (GLM) (Nelder and Wedderburn, 1972) with Gamma distributed error, log link,

and with up to cubic terms and all interactions. We apply backwards variable selection based on the *Akaike Information Criterion* (AIC) (Akaike, 1974). The resulting model fit is nearly perfect (goodness of fit: pseudo $R^2 = 0.998$). This model results in optimal speeds which are nearly the same for each radius: $(v_c, v_f) \approx (40.5, 8.75)$. For the dependence of work on the speeds (v_c, v_f) please compare Figure 12b). The independence of optimal speeds (v_c, v_f) on the radius r is what we hoped for. Too small values of v_f lead to higher friction and lower material removal, i.e., to higher total work. Too high v_f values cause higher total work due to very high material removal rates.

5 Conclusions

EGO is especially appropriate for the optimization of expensive black box functions. If evaluation cost is low, cost of surrogate model estimation can exceed cost for objective function evaluation. EGO with kriging not only focuses on finding the optimum but the exploration during this process also leads to good model fit over the whole parameter region. Variations are diverse, concerning, e.g., the covariance structure, the infill criterion, the search structure, categorical factors, noise, the regression model, etc. We introduced some new variants like regression as a modeling alternative to kriging and two simple methods for the handling of categorical variables, and we discussed focus search for the optimization of the infill criterion. Applications are numerous. We looked at parameter tuning and optimization in practice. The R package mlrMBO, developed in our group (Bischl et al, 2015), provides all the described methods and further features like, e.g., optimization with more than one objective function.

Acknowledgements

This Technical Report is a preprint of a paper accepted for publication by "Archives of Data Science, Series A"
(url = http://www.archivesofdatascience.org/journals/series_a/).
We acknowledge support by the German Research Foundation (DFG) within the Collaborative Research Center SFB 823 *Statistical modeling of nonlinear dynamic processes*, Projects B3, B4, and C2.

References

Adler RJ, Taylor JE (2007) Random fields and geometry. Springer Monographs in Mathematics

Akaike H (1974) A new look at the statistical model identification. IEEE Transactions on Automatic Control 19(6):716–723, DOI: 10.1109/TAC.1974.1100705

Bauer B, Friedrichs K, Bischl B, Weihs C (2015) Fast model based optimization of tone onset detection by instance sampling. In: Wilhelm A, Kestler H (eds) Analysis of Large and Complex Data, Springer, Studies in Classification, Data Analysis, and Knowledge Organization, accepted

Bello J, Daudet L, Abdallah S, Duxbury C, Davies M, Sandler M (2005) A tutorial on onset detection in music signals. IEEE Transactions on Speech and Audio Processing 13:1035–1047

Bischl B, Wessing S, Bauer N, Friedrichs K, Weihs C (2014) MOI-MBO: Multiobjective infill for parallel model-based optimization. In: Pardalos PM, Resende MG, Vogiatzis C, Walteros JL (eds) Learning and Intelligent Optimization, Springer, Lecture Notes in Computer Science, pp 173–186

Bischl B, Bossek J, Richter J, Horn D, Lang M (2015) mlrMBO: mlr: Model-based optimization. https://github.com/berndbischl/mlrMBO, R package version 1.0

Chang CC, Lin CJ (2011) LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2:27:1–27:27

Cioppa TM, Lucas TW (2007) Efficient nearly orthogonal and space-filling latin hypercubes. Technometrics 49:45–55

Cox DD, John S (1997) SDO: A statistical method for global optimization. In: Alexandrov M, Hussaini M (eds) Multidisciplinary Design Optimization: State of the Art, SIAM, pp 315–329

Cressie N (1988) Spatial prediction and ordinary kriging. Mathematical Geology 20(4):405–421

Herbrandt S, Ligges U, Ferreira M, Kansteiner D, Biermann D, Tillmann W, Weihs C (2016) Model based optimization of a statistical simulation model for single diamond grinding. Computational Statistics, to appear

Huang D, Allen TT, Notz WI, Zheng N (2006) Global optimization of stochastic black-box systems via sequential kriging meta-models. Journal of Global Optimization 34:441–466

IBM, Microsoft (1991) Multimedia programming interface and data specifications 1.0. "http://www-mmsp.ece.mcgill.ca/documents/audioformats/wave/wave.html"

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

Matérn B (1986) Spatial variation, vol 36. Springer Lecture Notes in Statistics

McKay MD, Beckman RJ, Conover WJ (1979) A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. Technometrics 21(1):239–245

Mood AM (1950) Introduction to the Theory of Statistics. McGraw-Hill

Myers RH, Montgomery DC, Anderson-Cook CM (2009) Response Surface Methodology. Wiley

Nelder J, Wedderburn R (1972) Generalized linear models. Journal of the Royal Statistical Society Series A 135(3):370–384, DOI: 10.2307/2344614

Qian PZG, Wu H, Wu CFJ (2008) Gaussian process models for computer experiments with qualitative and quantitative factors. Technometrics 50(3):383–396, DOI: 10.1198/004017008000000262

Rasmussen CE, Williams CKI (2006) Gaussian processes for machine learning. The MIT Press

Sasena MJ (2002) Flexibility and efficiency enhancements for constrained global design optimization with kriging approximations. PhD thesis, General Motors

Schölkopf B, Smola AJ (2002) Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press

Simon D (2013) Evolutionary Optimization Algorithms. Wiley

Vanschoren J, van Rijn JN, Bischl B, Torgo L (2014) OpenML: Networked science in machine learning. SIGKDD Explor Newsl 15(2):49–60, DOI 10.1145/2641190.2641198, URL http://doi.acm.org/10.1145/2641190.2641198