Discrete Mixtures of Kernels for Kriging-based Optimization

David Ginsbourger, Céline Helbert, Laurent Carraro
Département 3MI, Ecole Nationale Supérieure des Mines
158 cours Fauriel, 42023 Saint-Etienne, France
"last name"@emse.fr

February 4, 2008

Abstract: Kriging-based exploration strategies often rely on a single Ordinary Kriging model which parametric covariance kernel is selected *a priori* or on the basis of an initial data set. Since choosing an unadapted kernel can radically harm the results, we wish to reduce the risk of model misspecification. Here we consider the simultaneous use of multiple kernels within Kriging. We give the equations of discrete mixtures of Ordinary Krigings, and derive a multikernel version of the *expected improvement* optimization criterion. We finally provide an illustration of the *Efficient Global Optimization* algorithm with mixed exponential and Gaussian kernels, where the parameters are estimated by *Maximum Likelihood* and the mixing weights are likelihood ratios.

Key words: *Gaussian Processes, Global Optimization, Kernel Selection, Mixture of Experts*

The global optimization of numerical simulators is a challenging problem since the number of runs is severely limited by computation time. Furthermore, the derivatives are generally not available. For the past decade, Kriging-based derivative-free algorithms like EGO ([5]) have been developed to address this issue. Kriging metamodels are indeed convenient for building exploration strategies since they provide for every potential input vector both a mean predicted response value (Kriging mean) and an associated measure of accuracy (Kriging variance). Along this paper, the simulator is seen as a determinist numerical black-box function y with d-dimensional input:

$$y : \mathbf{x} \in D \subset \mathbb{R}^d \rightarrow y(\mathbf{x}) \in \mathbb{R} \tag{1}$$

y is known at first on the initial Design of Experiments $\mathbf{X} = \{\mathbf{x}^1, ..., \mathbf{x}^{n_0}\}$, where $n_0 \in \mathbb{N}$ is the number of initial runs. We denote by $\mathbf{Y} = \{y(\mathbf{x}^1), ..., y(\mathbf{x}^{n_0})\}$ the set of observations made by evaluating y at the points of $\mathbf{X}$. In almost all Kriging models, the starting point is to make the assumption that y is one realization of a random process of the form

$$Y(x) = \mu(x) + \varepsilon(x) \tag{2}$$

where $\mu(x)$ is a deterministic trend function, $\varepsilon(x)$ is a centered stationary random field with covariance function k, here assumed to belong to a set of positive definite stationary kernels:

$$\mathcal{K} = \{k_{(r,\sigma^2,\psi)} : h \in D - D \longrightarrow \sigma^2 r(h;\psi),\ r \in \mathcal{R},\ \sigma^2 \in \mathbb{R}^+,\ \psi \in \Psi_r\} \tag{3}$$

$\mathcal{K}$ is indexed by a finite set $\mathcal{R}$ of correlation kernel parametric families, by their respective continous hyperparameters $\psi \in \Psi_r$(e.g. correlation lengths), and by a positive parameter σ^2 (the process variance). In the following, $\chi = (r, \sigma^2, \psi)$ denotes the tree-structured covariance

parameters. In many industrial applications, r is arbitrarily chosen to belong to a parametric family (exponential, Gaussian, Matèrn, etc...), (σ^2, ψ) are then fitted to the data using automatic estimation procedures, and χ is finally plugged in as if it were known. Our particular concern here is to review and extend the EGO Algorithm ([5]) in taking the risk of model into account. After recalling some basics about Gaussian processes and metamodel-based optimization, we propose an adaptation of Ordinary Kriging with mixed kernels. We then derive an optimization criterion based on a discrete mixture of Kriging, and finally illustrate its efficiency by applying EGO with two simultaneous kernels to a classical test case function.

1 Gaussian Processes and Metamodel-based Optimization

Ordinary Kriging (OK) is a spatial interpolator developed by G. Matheron and named after the mining engineer D.G. Krige. It provides at each point $\mathbf{x} \in D$ a prediction of Y as a linear combination of the observed values $\mathbf{Y}$. The weights depend on the distance between the prediction point $\mathbf{x}$ and the design of experiments $\mathbf{X}$ through the chosen covariance kernel. Here we give the equations of Kriging for a fixed kernel $k_\chi(.) = \sigma^2 r(.;\psi)$. The Kriging mean m_χ and mean squared error (or variance) s^2_χ at $\mathbf{x}$ are the following functions (see [9] for pointwise derivation):

$$\left\{ \begin{array}{c} m_\chi(\mathbf{x}) = \hat{\mu}_\chi + \mathbf{k}_\chi(\mathbf{x})^T \mathbf{K}_\chi^{-1} \left(\mathbf{Y} - \hat{\mu}_\chi \mathbb{1}_n\right) \\ s^2_\chi(x) = \sigma^2 - \mathbf{k}_\chi(\mathbf{x})^T \mathbf{K}_\chi^{-1} \mathbf{k}_\chi(\mathbf{x}) + \left(\mathbb{1}_n^T \mathbf{K}_\chi^{-1} \mathbb{1}_n\right)^{-1} \left(1 - \mathbb{1}_n^T \mathbf{K}_\chi^{-1} \mathbf{k}_\chi(\mathbf{x})\right)^2 \end{array} \right. \tag{4}$$

where we recall that $\chi = (k, \sigma^2, \psi)$. $\mathbf{K}_\chi$ and $\mathbf{k}_\chi(\mathbf{x})$ are the matrices [1] :

$$\mathbf{K}_\chi = \begin{pmatrix} k_\chi(0) & k_\chi(\mathbf{x}_1 - \mathbf{x}_2) & ... & k_\chi(\mathbf{x}_1 - \mathbf{x}_n) \\ k_\chi(\mathbf{x}_2 - \mathbf{x}_1) & k_\chi(0) & ... & k_\chi(\mathbf{x}_2 - \mathbf{x}_n) \\ ... & ... & .... & .... \\ k_\chi(\mathbf{x}_n - \mathbf{x}_1) & ... & .... & k_\chi(0) \end{pmatrix} \text{ and } \mathbf{k}_\chi(\mathbf{x}) = \begin{pmatrix} k_\chi(\mathbf{x} - \mathbf{x}_1) \\ k_\chi(\mathbf{x} - \mathbf{x}_2) \\ ... \\ k_\chi(\mathbf{x} - \mathbf{x}_n) \end{pmatrix}$$

and $\hat{\mu}_\chi$ is given by: $\hat{\mu}_\chi = (\mathbb{1}^T \mathbf{K}_\chi^{-1} \mathbb{1})^{-1} \mathbb{1}^T \mathbf{K}_\chi^{-1} \mathbf{Y}$. The classical geostatistical interpretation of eq. (4) is to see $m_\chi(\mathbf{x})$ as the Best Linear Unbiased Predictor (BLUP) of $Y(\mathbf{x})$, under the hypothesis that eq. (2) holds with $\mu(\mathbf{x})$ being an unknown constant μ estimated by maximum likelihood. Here we find more convenient to consider an interpretation of OK in terms of Gaussian processes, in the flavour of ([1]). Assuming that $\varepsilon(\mathbf{x})$ is a centered stationary Gaussian process with known covariance function $k_\chi(.)$ and that μ is an unknown constant with improper uniform prior distribution $\mu \sim \mathcal{U}(\mathbb{R})$, we obtain the following conditional distribution for $Y(\mathbf{x})$

$$Y^{OK}_\chi(\mathbf{x}) := [Y(\mathbf{x}) / Y(\mathbf{X}) = \mathbf{Y}] \sim \mathcal{N}\left(m_\chi(\mathbf{x}), s^2_\chi(\mathbf{x})\right) \tag{5}$$

This approach allows the analytical calculation of various quantities involving $Y(\mathbf{x})$ knowing the observations, as well as conditional simulations[2] of Y, which ensures for instance that the expectation of any fonction involving $Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}$ can be estimated by Monte-Carlo. In the practice, one chooses a parametric correlation kernel r, then estimates the parameters (σ^2, ψ) , and finally plugs in the estimated values in the formulas. It seems however that such a sketch forgets the uncertainty associated with the choice of r and the estimation of (σ^2, ψ) and hence underestimates the modeling uncertainty. Assessing uncertainty with a variance s^2_χ obtained by plugging in χ in the Kriging equations entails a hidden risk of trusting too much a "bad" model.

[1] In the deterministic case, these equations are often written in an equivalent way using correlation matrices.

[2] A conditional simulation on a fine grid covering D can only be performed when D is low-dimensional (typically up to $d = 3$). However, conditional simulations at a small set of points are affordable whatever the dimension d.

Likelihood maximization (ML) is one of the standard ways, along with cross-validation, to estimate μ and (σ^2, ψ) on the basis of observations (see [1] for instance). It relies on the hypothesis that $\mathbf{Y}$ is a gaussian vector with mean μ and covariance matrix K_χ (this can be seen as a direct consequence of the Gaussian process interpretation of Kriging). Considering that r is known, one searches for the parameters $(\hat{\mu}, \hat{\psi}, \hat{\sigma^2})$ that give the largest density value to $\mathbf{Y}$. Noting [3] $R_\chi = \frac{1}{\sigma^2} K_\chi$, MLE then relies on the maximization of the Gaussian likelihood function:

$$L(\sigma^2, \psi, \mu; \mathbf{Y}) = f(\mathbf{Y}/\sigma^2, \psi, \mu) = \frac{1}{(2\pi)^{\frac{d}{2}} (\sigma^2)^{\frac{d}{2}} det(R_\chi)^{\frac{1}{2}}} e^{-\left[\frac{(\mathbf{Y}-\mu\mathbb{1})' R_\chi^{-1} (\mathbf{Y}-\mu\mathbb{1})}{2\sigma^2}\right]} \tag{6}$$

or equivalently on the minimization of $-2 \times log(L(\sigma^2, \psi, \mu; \mathbf{Y}))$. It can be shown that for every fixed ψ, the optimal μ and σ^2 are given by $\mu = \hat{\mu_\chi}$ [4] and:

$$\hat{\sigma^2}(\psi) = \frac{(\mathbf{Y} - \hat{\mu_\psi}\mathbb{1})^T R_\chi^{-1} (\mathbf{Y} - \hat{\mu_\psi}\mathbb{1})}{d} \tag{7}$$

After some direct calculations, ML can be restricted to the p_r-dimensional minimization problem:

$$\min_{\psi \in \Psi_r} \{log(|K_{(r, \hat{\sigma^2}(\psi), \psi)}|)\} \tag{8}$$

This optimization problem is generally solved numerically, which adds both computational complexity and randomness in the result. The last point has been studied in detail in the theory of likelihood, and more recently discussed in this particular framework in ([3]).

Once a kriging interpolator and its associated uncertainty are computed, one dispose of a metamodel that may be used to explore the simulator with a design dedicated to optimize y. One way to derive such a design is by iteratively maximizing a figure of merit based on the Kriging metamodel. ([4]) and ([7]) provide a review of most used Kriging-based optimization criteria.

The expected improvement (EI) is a broadly used optimization criterion that makes a trade-off between promizing (with low predictions, for minimization) and uncertain zones. Let $y_{min} = \min\{y(\mathbf{x}_1), ..., y(\mathbf{x}_n)\}$ be the minimum of the currently known observation values. Let $\mathbf{x} \in D$ be a candidate point for a next evaluation of y. In the end, evaluating y at $\mathbf{x}$ would bring an improvement of $y_{min} - y(\mathbf{x})$ if $y(\mathbf{x})$ is below y_{min} and no improvement if $y(\mathbf{x})$ is above y_{min}. Of course, this improvement $(y_{min} - y(\mathbf{x}))^+$ cannot be known without evaluating y (else, we could directly find the minimum). But eq. (5) makes it possible to know the statistical distribution of the random variable improvement $(y_{min} - Y(\mathbf{x}))^+$ conditionally on the observations $Y(\mathbf{X}) = \mathbf{Y}$. In particular, the expected improvement is defined as the following function of $\mathbf{x}$:

$$C_\chi^{EI}(\mathbf{x}) = \mathbb{E}\left[\left(y_{min} - Y_\chi^{OK}(\mathbf{x})\right)^+\right] = \mathbb{E}\left[\left(y_{min} - Y(\mathbf{x})\right)^+ / Y(\mathbf{X}) = \mathbf{Y}\right] \tag{9}$$

Note that the expression $C_\chi^{EI}(\mathbf{x})$ can be calculated analytically:

$$C_\chi^{EI}(\mathbf{x}) = (y_{min} - m_\chi(\mathbf{x}))\, \Phi\left(\frac{y_{min} - m_\chi(\mathbf{x})}{s_\chi(\mathbf{x})}\right) + s_\chi(\mathbf{x}) \phi\left(\frac{y_{min} - m_\chi(\mathbf{x})}{s_\chi(\mathbf{x})}\right) \tag{10}$$

where ϕ and Φ are the probability density and the cumulative distribution function of the standard Gaussian law. This expression sheds light on the trade-off between promising and uncertain

[3] There is implicitely no "nugget" effect since we work here with deterministic experiments

[4] Directly maximizing the likelihood with respect to μ and (σ^2, ψ) delivers the same value of μ as in the frame of OK, $\hat{\mu} = \hat{\mu_\chi}$. Note however that OK includes the variability due to μ's estimation in its variance term.

zones: the first term of the sum enhances local search via the mean prediction $m_\chi(\mathbf{x})$, whereas the second term puts more emphasis on global search via the prediction variance. ([4]), and ([7]) intensively commented the criterion of expected improvement.

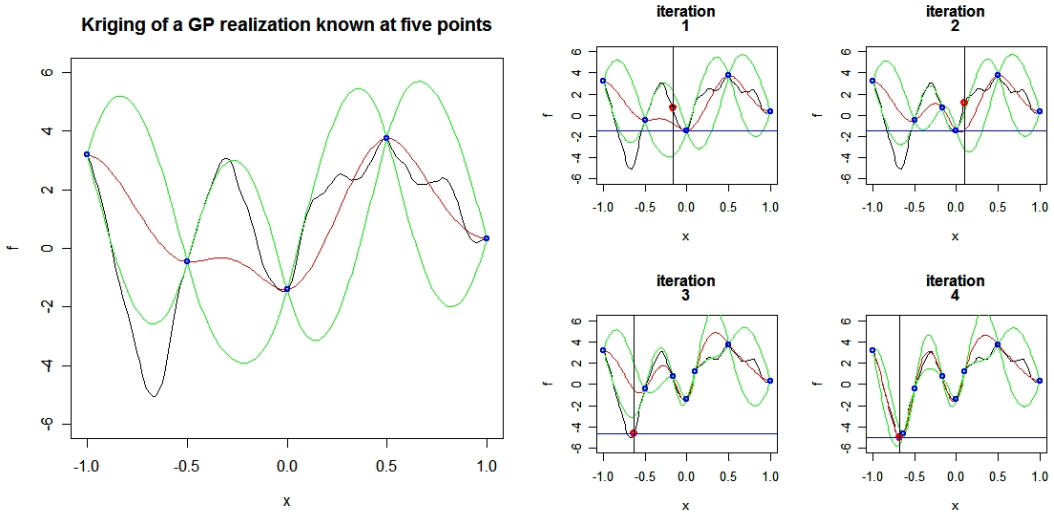

Figure 1: Left: OK (in red, with green 95% confidence intervals) of y (in black), a realization of a Gaussian process with known covariance function (cubic correlation, variance 4, scale 0.6). Right: 4 iterations of EGO applied to y. The blue points represent the visited sites, the red points (and their associated vertical lines) are the current maximizers of the expected improvement. The horizontal blue line represents the current y_{min} values. This example illustrates how an EGO sequences explores the objective function without getting trapped in the zones of local optimum.

Efficient Global Optimization (EGO) is an algorithm proposed by ([5]). It relies on a sequential exploration based on OK and on the maximization of the expected improvement criterion.

Algorithm 1 The E.G.O. Algorithm

function EGO($\mathbf{X}$, $\mathbf{Y}$, q)
  for $i \leftarrow 1, q$ **do**
    $(\sigma^{2*}, \psi^*) = argmax_{(\sigma^2,\psi)\in\mathbb{R}^+\times\Psi_r} L(\sigma^2, \psi; Y(\mathbf{X}) = \mathbf{Y})$ ▷ Estimating (σ^2, ψ) by ML
    $\mathbf{x}^* = argmax_{\mathbf{x}\in D}\{C^{EI}_{(r,\sigma^{2*},\psi^*)}(\mathbf{x})\}$ ▷ Maximizing the Expected Improvement
    $\mathbf{X} = \mathbf{X}\bigcup\{\mathbf{x}^*\}$ *and* $\mathbf{Y} = \mathbf{Y}\bigcup\{Y(\mathbf{x}^*)\}$ ▷ Updating the Design of Experiments
  end for
end function

In the exact version ([5]), the algorithms loops until $(max_{x\in D}\{C^{EI}_{\psi^*}(x)\} > \delta)$ for a δ fixed by the user (in the original E.G.O., δ is 0.01 times the value of $max_{x\in D}\{C^{EI}_{\psi^*}(x)\}$ at the previous iteration). Here we consider the case in which the number of runs $q \in \mathbb{N}$ is fixed in advance. EGO has been found to be a competitive global optimization algorithm, in particular in the fields of automotive and aerospace engineering, with objective functions having one to six inputs (see ([4]), ([7])). In other respects, ([8]) proposes a adaptation of EGO for physical experiments.

2 Mixtures of Kriging for prediction and optimization

Let us focus on the situation in which several kernels are in competition to model a sample of observed data. This is typically the case when different methods are available for the estimation of (σ^2, ψ) (e.g. ML with different initial values, ML and cross-validation, etc...), or when there is a choice to make between a set of functional forms for the correlation kernel r. Let us assume that the function y has already been evaluated at a finite set of points $\mathbf{X}_{obs}$, and denote by $\mathbf{Y}_{obs}$ the associated responses. We now consider that $M \in \mathbb{N}$ *experts* $\{\mathcal{E}_i,\ i \in [1, M]\}$ are at disposal to estimate χ on the basis of the observations. The $\mathcal{E}_i$'s are functional estimators, providing at the same time a correlation structure and its associated parameters. They are defined as follows:

$$\forall i \in [1, M],\ \mathcal{E}_i : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \in \bigcup_{k=n_0}^{+\infty} \left(D^k \times \mathbb{R}^k\right) \longrightarrow \chi_i = \mathcal{E}_i(\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \in \mathcal{K} \tag{11}$$

For the sake of convenience, we identify here $\mathcal{K}$ with the set of possible χ's (there is an obvious one-to-one mapping between both sets). Given the set of kernels $\{\chi_1, ..., \chi_M\}$ delivered by the experts $\{\mathcal{E}_i, i \in [1, M]\}$, and $\mathcal{W} = \{w_1, ..., w_M\}$ a set of weights meant to quantify the respective relevance levels of the M experts, we study the idea of replacing the classical approach of kernel selection by a mixture of kernels: instead of keeping the best kernel and dropping off the others, we propose to keep them all and integrate them within Ordinary Kriging in probabilizing χ.

Discrete mixture of Gaussian processes: the unknown function y is now seen as one path of a random field associated with an OK model, which underlying kernel is independently chosen at random following a discrete law supported by the set of kernels delivered by the M experts:

$$\begin{cases} [Y_{mix}^{OK}/\chi] = Y_{\chi}^{OK} \\ P(\chi = \chi_i) = w_i \end{cases} \tag{12}$$

Note that the proposed approach is not strictly bayesian: the *prior distribution* on χ (in the sense of a bayesian framework) would be here depending on the data (we focus only on M experts, who are defined on the basis of observations). Here we first select a set of kernels and then mix them. Following eq. (12), the conditional distribution of $Y(\mathbf{x})$ ($\mathbf{x} \in D$) is a mixture of Gaussians:

$$Y_{mix}^{OK}(\mathbf{x}) := [Y(\mathbf{x})/Y(\mathbf{X}) = \mathbf{Y}] \text{ with density function } \sum_{j=1}^{M} w_j p_{\mathcal{N}\left(m_{\chi_j}(\mathbf{x}), s^2_{\chi_j}(\mathbf{x})\right)}(.) \tag{13}$$

Ordinary Kriging with a mixed kernel: Following eq. (13), Y_{mix}^{OK} is a field of Gaussian mixtures. This entails the equations of the mixed mean [5] and variance [6]:

$$m_{mix}(\mathbf{x}) = \mathbb{E}[\mathbb{E}[Y_{mix}^{OK}(\mathbf{x})/\chi]] = \sum_{i=1}^{M} w_i m_{\chi_i}(\mathbf{x}) \tag{14}$$

Hence, the mean of the resulting metamodel is the weighted average of the means associated with the different krigings (which coincides with the concept of weighted average surrogate model developped in [10]). Furthermore, the corresponding variance is given by

$$\begin{aligned} s^2_{mix}(\mathbf{x}) = Var[Y_{mix}^{OK}(\mathbf{x})] &= \mathbb{E}[Var[Y_{mix}^{OK}(\mathbf{x})/\chi]] + Var[\mathbb{E}[Y_{mix}^{OK}(\mathbf{x})/\chi]] \\ &= \sum_{i=1}^{M} w_i s^2_{\chi_i}(\mathbf{x}) + \sum_{i=1}^{M} w_i [(m_{\chi_i}(\mathbf{x}) - m_{mix}(\mathbf{x}))^2] \end{aligned} \tag{15}$$

[5] Using the law of total expectation: $\mathbb{E}[X_1] = \mathbb{E}[\mathbb{E}[X_1/X_2]]$.

[6] Using the law of total variance: $Var[X_1] = \mathbb{E}[Var[X_1/X_2]] + Var[\mathbb{E}[X_1/X_2]]$

The first term is a linear combination of the model variances weigthed by the $w_i's$, whereas the second term reflects the dispersion between the different kriging means. The latter plays a capital role since it introduces heteroscedasticity in the Kriging variance: contrarily to the case of regular OK, the variance now depends on the observations $\mathbf{Y}$ through the second term.

Optimization under a mixture of kernels: when several values of χ are possible, finding the next most promizing point with a kriging-based optimization criterion (say C_{χ}^{EI}) becomes a multicriteria decisional problem. Our approach here is to combine all $C_{\chi_i}^{EI}$'s to provide a unified criterion that takes into account both sources of randomness. Using again the so-called law of total expectation, we derive the *mixed expected improvement*:

$$C_{mix}^{EI}(\mathbf{x}) := \mathbb{E}\left[\left(y_{min} - Y_{mix}^{OK}(\mathbf{x})\right)^{+}\right] \tag{16}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(y_{min} - Y_{mix}^{OK}(\mathbf{x})\right)^{+} / \chi\right]\right] \tag{17}$$

$$= \mathbb{E}\left[\mathbb{E}\left[\left(y_{min} - Y_{\chi}^{OK}(\mathbf{x})\right)^{+}\right]\right] = \sum_{i=1}^{M} w_i C_{\chi_i}^{EI}(\mathbf{x}) \tag{18}$$

and hence, the expected improvement function under a mixture of kernels is simply the convex combination of the M expected improvement functions weighted by the $\{w_i, \in [1, M]\}$. Note that any integral criterion under a mixture of Krigings can be calculated in the same way.

Selecting a benchmark of experts: replacing the step of model selection by a step dedicated at choosing a set of models may seem at first to create more problems than it solves indeed. For instance, if we consider several families of correlation kernels (e.g. a gaussian, an exponential, and even nonstationary correlation kernels -as in [2]-) and estimate each set of kernel parameters by ML, it naturally increases the computational amount. The price for mixing is in that case to multiply the time needed for model inference by the number of experts. In this flavour, possible approaches would be to consider simultaneously experts relying on the same correlation kernel but with hyperparameters inferred using different methods (e.g. mixing the ML and the LOO "best" models), or even getting several candidate hyperparameter sets by parametric bootstrap. Ideally, we would like to have all relevant classes of experts represented in a small set. One of the future issues to be addressed seem to be the selection of sets with dissimilar good experts.

Setting a probability measure over the set of experts: Once a set of experts is chosen, probability weights have to be defined. The most naïve way of probabilizing the models is to put a uniform distribution on them. This approach may be relevant when mixing models obtained by maximizing different criteria (e.g.: a $50\% - 50\%$ mix of the "best ML model" and the "best LOO model"). In an opposite fashion, it is also possible to consider the density of the mixture of models as a function of both the covariance parameters and parametric weights and then to perform likelihood maximization over all parameters including the $w_i's$. Such problems are typically numerically solved using an *Expectation-Maximization* (EM) algorithm. We propose a way inbetween, more informative than a raw uniform distribution and yet computationally cheaper than EM. Since we have a criterion of fit (the Gaussian likelihood), why not use it to weight models? At first, we propose a Kriging mixture with weights based upon the likelihood criterion. In what follows, we consider Akaike weights (see [6]):

$$\forall i \in [1, M],\ w_i = \frac{L(\psi_i, \hat{\sigma}_{\psi_i}, \hat{\mu_{\psi_i}}/\mathbf{Y}, k_i)}{\sum_{i=1}^{N} L(\psi_i, \hat{\sigma}_{\psi_i}, \hat{\mu_{\psi_i}}/\mathbf{Y}, k_i)} \tag{19}$$

Note that the w_i's are simply likelihood profile values divided by a normalization coefficient.

3 EGO with mixed kernels, applied to Branin's function

The Branin-Hoo function has been intensively studied in the litterature of global optimization of black-box functions ([5]). It is a smooth two-variable function defined by:

$$y(x_1, x_2) = \left(x_2 - \frac{5.1}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)cos(x_1) + 10, \ (x_1, x_2) \in [-5, 10] \times [0, 15]$$

y has three global minimizers $(-3.14, 12.27)$, $(3.14, 2.27)$, $(9.42, 2.47)$, and the global minimum is approximately equal to 0.4. We normalized the variables between 0 and 1. Now we wish to illustrate, and compare EGO with different kernels and kernel mixtures.

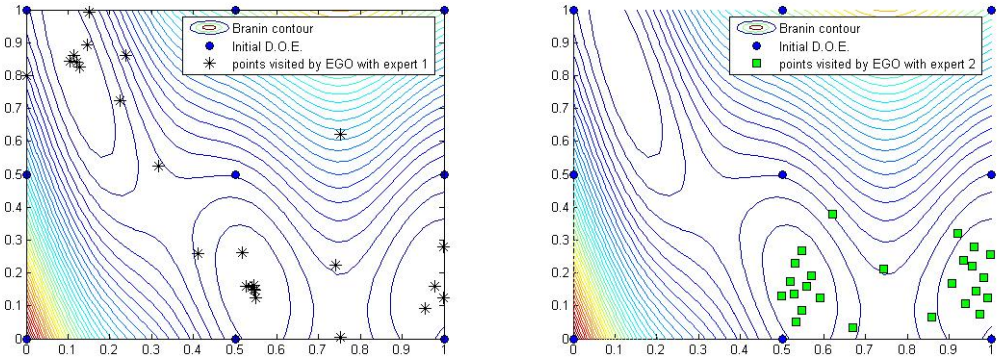

Figure 2: 25 iterations of the EGO algorithm applied to the Branin-Hoo function, with both experts $\mathcal{E}_1$ and $\mathcal{E}_2$ (see eq.(20)) and initial design $\mathbf{X}$ (in dark blue dots). Left: the path of EGO with expert $\mathcal{E}_1$ is represented by black thin stars. Right: the path of EGO with expert $\mathcal{E}_2$ is represented by light green squares. One of the three zones of minimum (upper left) is not visited.

The experimental set-up is the following: the initial design of experiments is a three-level full factorial design $\mathbf{X} \in ([0, 1] \times [0, 1])^{n_0}$ $(n_0 = 9)$. Two correlation kernels are selected:

Gaussian correlation	Exponential correlation
$r_1(h) = e^{-\sum_{i=1}^{2}\left\lvert\frac{h_i}{p_i}\right\rvert^2}$	$r_2(h) = e^{-\sum_{i=1}^{2}\left\lvert\frac{h_i}{p_i}\right\rvert}$

The experts considered here are the two parametric correlation kernels r_1 and r_2 with their respective parameters estimated by ML:

$$\begin{cases} \mathcal{E}_1 : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \rightarrow \chi_1 = (r_1, \sigma_1^*, \psi_1^*), \text{ where } (\sigma_1^*, \psi_1^*) = argmax_{\sigma,\psi}[L(\sigma, \psi; \mathbf{Y}_{obs}, r_1)] \\ \mathcal{E}_2 : (\mathbf{X}_{obs}, \mathbf{Y}_{obs}) \rightarrow \chi_2 = (r_2, \sigma_2^*, \psi_2^*), \text{ where } (\sigma_2^*, \psi_2^*) = argmax_{\sigma,\psi}[L(\sigma, \psi; \mathbf{Y}_{obs}, r_2)] \end{cases} \quad (20)$$

All the algorithms and computations have been implemented in the frame of the MatLab free toolbox "Gaussian Processes for Machine Learning" (illustrating the book [1]). In both cases, the kernel hyperparameters intitial values are fixed to $(1, 10, 0.5, 1)$.

The results are summarized on fig. (3). The left figure illustrates 25 iterations of EGO with mixed experts. The pattern of the visited points is close to the trajectory of EGO with Gaussian

Algorithm 2 The E.G.O. Algorithm with 2 mixed kernels weighted by their likelihood ratios

function EGO($\mathbf{X}$, $\mathbf{Y}$, q)
  for $i \leftarrow 1, q$ **do**
    for $j \leftarrow 1, 2$ **do**
      $(\sigma_j^{2*}, \psi_j^*) = argmax_{(\sigma_j^2, \psi_j) \in \mathbb{R}^+ \times \Psi_{r_j}} L(\sigma_j^2, \psi_j; Y(\mathbf{X}) = \mathbf{Y}, r_j)$ ▷ MLE
    end for
    for $j \leftarrow 1, 2$ **do**
      $w_j = \frac{L(\sigma_j^{2*}, \psi_j^*; Y(\mathbf{X})=\mathbf{Y}, r_j)}{\sum_{j=1}^2 L(\sigma_j^{2*}, \psi_j^*; Y(\mathbf{X})=\mathbf{Y}, r_j)}$ ▷ Computing the mixing weights
    end for
    $\mathbf{x}^* = argmax_{\mathbf{x} \in D} \sum_{j=1}^2 w_j C^{EI}_{(r_j, \sigma_j^{2*}, \psi_j^*)}(\mathbf{x})$ ▷ Maximizing the mixed EI
    $\mathbf{X} = \mathbf{X} \bigcup \{\mathbf{x}^*\}$ *and* $\mathbf{Y} = \mathbf{Y} \bigcup \{y(\mathbf{x}^*)\}$ ▷ Updating the Design of Experiments
  end for
end function

expert. In particular, the three zones of local optima are visited during the 25 first iterations. This similarities can be understood by looking at the sequences of weights plotted on the right figure. The two curves on the graphic below represent the log-likelihood associated with both experts as functions of the number of EGO iterations. Note that the likelihood of expert 2 is greater than the likelihood of expert 1 until the number of iterations reaches 6, and then becomes significantly lower than the other one. The likelihood ratios plotted on the graphic above show more precisely how the exponential kernel prevails at the beginning of the EGO algorithm, and is later dropped in favour of the Gaussian kernel. This kind of *automatic selection* seems due to an asymptotical steep decrease of the likelihood ratio. Using Akaike weights to mix kernels within a sequential exploration seems here to be a useful means to automatically select an expert without making a decision based on the initial design of experiments only. Hence, the proposed approach appears to be a sound option to increase EGO's robustness to modeling uncertainty.

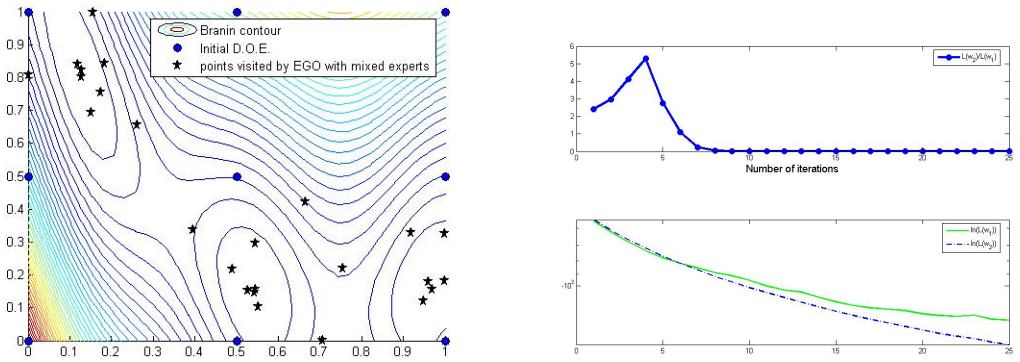

Figure 3: 25 iterations of EGO applied to the Branin-Hoo function, with a mixture of experts $\mathcal{E}_1$ and $\mathcal{E}_2$ weighted by Akaike weights. Left: the path of EGO with the mixture of experts is represented by black filled stars. Right: evolution of both sequences of log-weights (lower graphic), and the associated series of likelihood ratios (upper graphic).

4 Conclusions

We have derived and discussed an optimization criterion, the *mixed expected improvement*, relying on discrete mixtures of Ordinary Kriging metamodels with different covariance kernels. The presented framework of multiple experts allows one to handle several parametric correlation structures and/or different parameter estimation techniques within the same Kriging-based procedure. The application of the latter to the optimization of the Branin-Hoo function provided a first example, where mixing appears to be a succesful alternative to model selection based on initial data. It is illustrated that, possibly after an oscillating behaviour for a few iterations, mixing two *experts* within the EGO algorithm (Ordinary Krigings with Gaussian and Exponential correlation structures, with Akaike weights) may ultimately lead to an automatic selection of a unique metamodel. The issues of selecting parsimonious benchmarks of experts, and of using weighting methods dedicated to different purposes are to be addressed in forthcoming works.

Aknowledgements: This work was conducted within the frame of the DICE (Deep Inside Computer Experiments) Consortium between ARMINES, Renault, EDF, IRSN, ONERA, and Total S.A. We wish to thank Raphael T. Haftka and Victor Picheny for their help and useful comments. Special thanks to the R project people for developing such a useful freeware.

References

[1] Rasmussen C.E. and Williams K.I. *Gaussian Processes for Machine Learning*. M.I.T. Press, 2006.

[2] Paciorek C.J. *Nonstationary Gaussian Processes for Regression and Spatial Modelling*. PhD thesis, Carnegie Mellon University, 2003.

[3] Ginsbourger D., Dupuy D., Badea A., Roustant O., and Carraro L. On the selection and the estimation of kriging models for deterministic computer experiments. In *Joint ENBIS-DEINDE conference*, 2007.

[4] Jones D.R. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, (21):345–383, 2001.

[5] Jones D.R., Schonlau M., and Welch W.J. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492, 1998.

[6] Burnham K.P. and Anderson D.R. *Model Selection and Multimodel Inference*. Springer, 1998.

[7] Sasena Michael J., Papalambros Panos, and Goovaerts Pierre. Exploration of metamodeling sampling criteria for constrained global optimization. *Engineering Optimization*, 2001.

[8] Henkenjohann N., Göbel R., Kleiner M., and Kunert J. An adaptive sequential procedure for efficient optimization of the sheet metal spinning process. *Qual. Reliab. Engng. Int.*, 21:439–455, 2005.

[9] Cressie N.A.C. *Statistics for spatial data*. Wiley series in probability and mathematical statistics, 1993.

[10] Goel T., Haftka R.T., Shyy W., and Queipo N. Simultaneous use of multiple surrogates. *AIAA Journal*, 2006.