High-dimensional Bayesian Optimization with a Combination of Kriging models

Tanguy Appriou, Didier Rullière, David Gaudrie

High-dimensional Bayesian Optimization with a Combination of Kriging models

Tanguy Appriou*†‡, Didier Rullière†, David Gaudrie‡

Monday 26th February, 2024

Abstract

In Bayesian optimization (BO), Kriging surrogate models reducing the number of function evaluations to reach the optimum. This method has been successfully applied to many real-world applications in low dimensions (less than 30 design parameters). However, in high dimension, building an accurate Kriging model is difficult, especially when the number of samples is limited as is the case when dealing with numerical simulators. This is due to the inner optimization of the Kriging length-scale hyper-parameters which can lead to inaccurate models and impacts the performances of the optimization. In this paper, we introduce a new method for high-dimensional BO which bypasses the length-scales optimization by combining sub-models with random length-scales, and whose expression, obtained in closed-form, avoids any inner optimization. We also describe how to sample suitable length-scales for the sub-models using an entropy-based criterion, in order to avoid degenerated sub-models having either too large or too small length-scales. Finally, the variance of the combination being not directly available, we present a method to compute the prediction variance for any weighting method. We apply our combined Kriging model to high-dimensional BO for analytical test functions and for the design of an electric machine. We show that our method builds more accurate surrogate models than ordinary Kriging when the number of samples is small. This results in faster convergence for BO using the combination.

Keywords— Bayesian Optimization, Kriging, Gaussian Process Regression, High Dimension, Maximum Likelihood Estimation, Model Aggregation.

1 Introduction

In engineering design optimization, surrogate models (also called metamodels) are widely used to emulate black-box functions we want to optimize (Forrester et al., 2008) because such functions are often obtained via a computationally expensive computer simulation (e.g. computational fluid dynamics or finite-elements solvers). This makes the optimization prohibitively expensive to perform with usual optimization methods, such as evolutionary algorithmsor gradient-based approaches, due to the large number of function evaluations required. Global Bayesian optimization (BO) strategies hinging on surrogate models were developed to solve this type of optimization problems. In particular, Efficient Global

* *Corresponding author*, tanguy.appriou@emse.fr

† Mines Saint-Etienne, Univ Clermont Auvergne, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, F - 42023 Saint-Etienne France

‡ Stellantis, Centre d'Expertise R&D Carrières-sous-Poissy, F-78955 Carrières-sous-Poissy, France

Optimization (EGO) introduced by Jones et al. (1998) has been applied to a wide variety of design optimization problems (e.g. Candelieri et al., 2018; Meliani et al., 2019; Picheny et al., 2019). This algorithm uses a non-parametric class of statistical models called Kriging models (Cressie, 1993; Stein, 1999) as a surrogate. Kriging models have proven to be effective in modelling black-box functions, not only for the approximation of numerical experiments (Sacks et al., 1989; Santner et al., 2003), but also in other fields such as machine learning (Rasmussen and Williams, 2006), where it is known as Gaussian Process regression.

Although EGO has been applied successfully to a number of low-dimensional applications (dimension $d < 20$), one of its main drawback is that Kriging scales poorly to high-dimensional cases. Originating from geostatistics (Krige, 1951; Matheron, 1963) where it was used as an interpolator for spatial fields, Kriging was initially devised for problems with dimension limited to either $d = 2$ or 3. However in real-world design optimization, engineering designs may commonly be parametrized by 50 or more parameters (Shan and Wang, 2010; Gaudrie et al., 2020). In higher dimensions, Kriging suffers from the curse of dimensionality (Bellman, 1966) and building an accurate surrogate model is met with various setbacks. This, in turn, is problematic for design optimization since the convergence speed of EGO is related to the accuracy of the surrogate. Thus, using accurate models is instrumental in order to reduce the number of function evaluations to reach the optimum. One of the main challenges in Kriging is the inner optimization of the covariance length-scale hyperparameters. The latter regulate the decay of the correlation between observations when their distance increases. Usually Kriging models implement anisotropy by considering one length-scale per dimension. Estimating these hyperparameters correctly is essential to obtain a model with a good accuracy and they are typically determined by Maximum Likelihood Estimation (MLE), an extremely popular method for fitting Kriging models whose theoretical properties have been studied extensively (e.g. Stein, 1999; Zhang, 2004; Van Der Vaart and Van Zanten, 2011; Kaufman and Shaby, 2013; Karvonen et al., 2020). However, most of these results are based on asymptotic considerations on the number of samples, and fewer works study practical cases where the number of samples is limited. For example, Karvonen and Oates (2023) showed that, in certain conditions, MLE is ill-posed in the sense that it leads to an infinite estimate of the length-scales. Practical MLE of the Kriging hyperparameters is difficult, especially for high-dimensional problems, in particular because the size of the search space grows exponentially with the dimension. Since the optimization is typically solved using gradient-based methods (e.g BFGS) with multi-start or evolutionary algorithms (Roustant et al., 2012), there is no better solution than increasing the number of iterations which might result in an increased computational effort as the cost of the likelihood and its gradient scale with $O(n^3)$. However, most of the times in design optimization, the number of samples is very limited and the cost of the length-scale optimization remains negligible compared to the cost of obtaining the samples. In these cases where the number of samples is small, Ginsbourger et al. (2009) and Mohammed and Cawley (2017) showed empirically that MLE can lead to very dispersed results, and Appriou et al. (2023) that it can fail to recover the true hyperparameters. While not as frequently used as MLE, other methods exist to determine the length-scales. For example, Bachoc (2013) suggested that cross-validation is more adapted than MLE when the model is ill-specified. Li and Sudjianto (2005) and Yi et al. (2011) proposed penalized versions of the likelihood to reduce the variance of the length-scale estimator, and Gu et al. (2018) used maximum a posteriori estimation with a reference prior to obtain a robust estimator avoiding the setting of lower and upper bounds which can also be a tricky point of the length-scale optimization. However, all these length-scale estimation

procedures also face difficulties in high-dimension.

High-dimensional Kriging and Bayesian optimization has recently gained attention, see Binois and Wycoff (2022) for a review. One classical approach is to reduce the problem's dimension by embedding the design space into a lower dimension space (e.g. Constantine, 2015; Bouhlel et al., 2016), and by building the Kriging model in this low-dimension space. Other methods consider simplifying hypotheses such as additive models Durrande et al. (2012), where the function is assumed to decompose into a sum of one-dimensional components, enabling a sequential optimization of the length-scales. However, all these additional assumptions (low dimension representation or additive structure) are not necessarily satisfied in practice and may not generalize to any design engineering problem. Appriou et al. (2023) introduced another method for high-dimensional Kriging based on a combination of Kriging sub-models, each one having fixed random length-scales, thus bypassing the difficult hyperparameter optimization, with no additional assumption on the black-box. However in this paper, the focus was on comparing different weighting methods for combining the sub-models, and on studying the influence of the number of sub-models.

In this paper, we apply this method to high-dimensional Bayesian optimization. Notation and main concepts of Kriging and Bayesian optimization are introduced in Section 2. Then, in Section 3, we present a motivating example illustrating how classical MLE fails to produce a good model when only few observations are available. In Section 4, we detail our new model for high-dimensional Bayesian optimization: in Section 4.1, we present a new entropy-based criterion to sample suitable length-scales for the sub-models to avoid degenerated cases with either too large or too small length-scales; in Section 4.2 we describe the weighting method to combine the sub-models; and in Section 4.3 and 4.4, we explain how we compute the prediction variance of the combination, required for Bayesian optimization, and which is not readily available as the correlation between the sub-models is unknown. Finally, Section 5 discusses the performances of our method for the optimization of two high-dimensional analytical test functions, and for one real-world application.

2 Kriging surrogate models and Bayesian optimization

2.1 Ordinary Kriging

Originally, Ordinary Kriging (OK) was developed for interpolation of spatial data by Matheron (1963) who named the method after the South African mining engineer D.G. Krige. The method was later applied to the approximation of computer experiments (Sacks et al., 1989; Santner et al., 2003). This method is also known as Gaussian Process Regression (Rasmussen and Williams, 2006). This section briefly covers the basics of the method and introduces the notation used throughout this paper. We denote by $y : \boldsymbol{x} \in \mathcal{X} \subset \mathbb{R}^d \to y(\boldsymbol{x}) \in \mathbb{R}$ the d-dimensional black-box function that we want to approximate by a surrogate model. We have n training points $\boldsymbol{X} = (\boldsymbol{x}_1, \dots, \boldsymbol{x}_n)^\top$ where the function y is known, and we denote $\boldsymbol{Y} = (y(\boldsymbol{x}_1), \dots, y(\boldsymbol{x}_n))^\top$ the function values at these locations. In Kriging, y is assumed to be the realization of a Gaussian process (GP) on $\mathcal{X}$:

$$Y(.) \sim \mathcal{GP}\left(\mu(.), \sigma^2 k_{\boldsymbol{\theta}}(.,.)\right).$$

$\mu(.)$ is the mean function of the GP which we consider constant in this paper: $\mu(.) = \mu$. This is a standard choice for the approximation of computer codes where no we have prior knowledge about the true function (Forrester et al., 2008; Ginsbourger et al., 2009) known as *ordinary* Kriging. $k_{\boldsymbol{\theta}} : \mathcal{X} \times \mathcal{X} \to [-1, 1]$ is the positive definite correlation function

indexed by the hyperparameters $\boldsymbol{\theta} \in \mathbb{R}^d$, called the correlation length-scales vector (also range or scale parameters), with one length-scale per dimension of the input space. Finally, $\sigma^2 \in \mathbb{R}^+$ is another hyperparameter calibrating the amplitude of the variance, and $\sigma^2 k(.,.)$ is called the covariance function or kernel. A stationary GP with a Matérn-class covariance function is often recommended (Stein, 1999; Rasmussen and Williams, 2006). One example of which, used throughout this paper, is the radial Matérn 5/2 correlation defined as:

$$k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}') := \left(1 + \sqrt{5}\left\|\frac{\boldsymbol{x}-\boldsymbol{x}'}{\boldsymbol{\theta}}\right\| + \frac{5}{3}\left\|\frac{\boldsymbol{x}-\boldsymbol{x}'}{\boldsymbol{\theta}}\right\|^2\right) \exp\left(-\sqrt{5}\left\|\frac{\boldsymbol{x}-\boldsymbol{x}'}{\boldsymbol{\theta}}\right\|\right), \tag{1}$$

where $\left\|\frac{\boldsymbol{x}-\boldsymbol{x}'}{\boldsymbol{\theta}}\right\|$ is the scaled distance between two points $\boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$ using component-wise division: $\left\|\frac{\boldsymbol{x}-\boldsymbol{x}'}{\boldsymbol{\theta}}\right\|^2 := \sum_{\ell=1}^{d}\left(\frac{x^{(\ell)}-x'^{(\ell)}}{\theta^{(\ell)}}\right)^2$. This is a typical choice for design optimization (Roustant et al., 2012) when there is no prior information on the function. This is because GP trajectories with this correlation are twice differentiable (Abrahamsen, 1997), which may be better suited than the Gaussian kernel (infinitely differentiable trajectories) or than the exponential correlation (trajectories not differentiable). Other covariance functions can be used if prior knowledge about the unknown function is available (e.g. cylindrical kernels), and, notice that even when the covariance is misspecified, a proper estimation of the hyperparameters can still yield a model with good predictive capacities (Bachoc, 2013).

The ordinary Kriging predictor is a linear combination of the observations which is obtained by conditioning the Gaussian process Y over $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$:

$$\hat{y}(\boldsymbol{x}) := \mathbf{E}(Y(\boldsymbol{x})|\mathcal{D}) = \mu + k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{X})^{-1}(\boldsymbol{Y} - \mu), \tag{2}$$

where $k(\boldsymbol{x}, \boldsymbol{X})$ is the vector of correlations between the prediction point $\boldsymbol{x}$ and the sample points $\boldsymbol{X}$, and $k(\boldsymbol{X}, \boldsymbol{X})$ is the $n \times n$ matrix of correlations between the components of $\boldsymbol{X}$. Note that this predictor does not depend on σ^2. One of the main advantage of Kriging is that it not only provides a prediction, but also the prediction variance, interpreted as a model uncertainty. This is particularly important in Bayesian optimization where, alongside with the prediction, this model uncertainty helps exploring under-visited parts of the design space. The prediction variance is:

$$\hat{s}^2(\boldsymbol{x}) := \mathbf{Var}(Y(\boldsymbol{x})|\mathcal{D}) = \sigma^2\left(k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{x}) - k_{\boldsymbol{\theta}}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{X})^{-1} k_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{x})\right). \tag{3}$$

In the following, we will sometimes denote the correlation matrix as $\boldsymbol{K}_{\boldsymbol{\theta}} := k_{\boldsymbol{\theta}}(\boldsymbol{X}, \boldsymbol{X})$.

2.2 Hyperparameter estimation

The estimation of the covariance hyperparameters of a Kriging model drastically affects its precision. As illustrated in Figure 1, an appropriate choice of length-scales yields much better accuracy than arbitrary length-scales. As often in parametric statistic models, the usual approach to estimate the hyperparameters is to use maximum likelihood estimation (MLE) (Rasmussen and Williams, 2006), which consists in maximizing the marginal likelihood of the model:

$$\mathcal{L}(\sigma, \boldsymbol{\theta}) := \frac{1}{(2\pi)^{n/2}\,\det(\sigma^2 \boldsymbol{K}_{\boldsymbol{\theta}})^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{Y} - \mu)^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{Y} - \mu)\right). \tag{4}$$

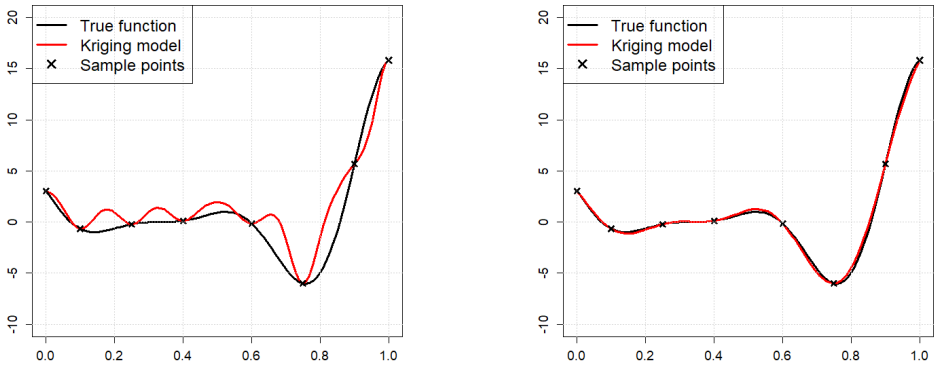
This is equivalent to minimizing $-\log(\mathcal{L}(\sigma, \boldsymbol{\theta}))$. Given a fixed $\boldsymbol{\theta}$, the maximum likelihood estimator for μ and σ^2 are:

$$\hat{\mu} = \frac{\mathbf{1}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \boldsymbol{Y}}{\mathbf{1}^\top \boldsymbol{K}_{\boldsymbol{\theta}}^{-1} \mathbf{1}}, \quad \text{and} \quad \hat{\sigma}^2_{MLE} = \frac{(\boldsymbol{Y} - \hat{\mu})^\top \boldsymbol{K}_{\boldsymbol{\theta}}{}^{-1}(\boldsymbol{Y} - \hat{\mu})}{n}, \tag{5}$$

The length-scales $\boldsymbol{\theta}$ are estimated by minimizing the concentrated log-likelihood obtained by injecting (5) into (4):

$$\hat{\boldsymbol{\theta}}_{MLE} = \arg\min_{\boldsymbol{\theta}} \ \frac{n}{2}\log\left(\hat{\sigma}^2_{MLE}\right) + \frac{1}{2}\log\left(\det(\boldsymbol{K}_{\boldsymbol{\theta}})\right). \tag{6}$$

In this expression, the length-scales are involved in both terms $\hat{\sigma}^2_{MLE}$ and $\det(\boldsymbol{K}_{\boldsymbol{\theta}})$ through the correlation matrix. The inner optimization (6) is solved numerically, typically using a gradient-based method (e.g BFGS) with multi-start as the gradient of the likelihood with respect to the length-scales can be in closed-form, or using evolutionary algorithms (Roustant et al., 2012).



(a) Random length-scale (b) MLE length-scale

Figure 1: Example of an ordinary Kriging model in 1D. Left: model with an arbitrary length-scale $\theta = 0.03$, right: model with length-scale obtained by MLE.

2.3 Bayesian optimization

As presented in the previous section, Kriging gives a way to build a surrogate model to approximate a black-box function based on some observations. Bayesian Optimization (BO) aims at finding the global optimum of this function $\boldsymbol{x}^* = \arg\min_{\boldsymbol{x}} y(\boldsymbol{x})$ in as few evaluations of the black-box as possible. One common Bayesian optimization framework is the Efficient Global Optimization (EGO) algorithm introduced by Jones et al. (1998), summarized in Algorithm 1.

Algorithm 1 Efficient Global Optimization (EGO) algorithm

Create an initial design plan: $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$.
Compute the associated values: $\boldsymbol{Y} = (y(\boldsymbol{x}_1), \ldots, y(\boldsymbol{x}_n))^\top$.
Fit the Kriging model to the data $\mathcal{D} = (\boldsymbol{X}, \boldsymbol{Y})$: find $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\theta}}$.
repeat
  $\boldsymbol{x}_{n+1} = \arg\max_{\boldsymbol{x}} EI(\boldsymbol{x})$.
  $\boldsymbol{X} \leftarrow \boldsymbol{X} \cup \boldsymbol{x}_{n+1}$.
  $\boldsymbol{Y} \leftarrow \boldsymbol{Y} \cup y(\boldsymbol{x}_{n+1})$.
  Re-estimate the hyperparameters $\hat{\mu}$, $\hat{\sigma}^2$, and $\hat{\boldsymbol{\theta}}$.
until convergence or budget exhaust

The EGO algorithm begins by evaluating a small set of designs $\boldsymbol{x}_i$ called the initial design of experiments (DoE), usually space-filling, and sequentially adds new observations by maximizing a so-called acquisition criterion which quantifies the worth of any unevaluated

$\boldsymbol{x}$. The most popular one is the Expected Improvement (EI), which is the expectation of the improvement $I(\boldsymbol{x}) = \max(0, y_{min} - Y(\boldsymbol{x}))$ over $y_{min} = \min(\boldsymbol{Y})$, the best value observed so far. For $\boldsymbol{x} \in \mathcal{X}$, the EI can be easily computed as:

$$EI(\boldsymbol{x}) := \mathbf{E}\left[I(\boldsymbol{x})\right] = (y_{min} - \hat{y}(\boldsymbol{x}))\Phi\left(\frac{y_{min} - \hat{y}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right) + \hat{s}(\boldsymbol{x})\phi\left(\frac{y_{min} - \hat{y}(\boldsymbol{x})}{\hat{s}(\boldsymbol{x})}\right), \quad (7)$$

where which $\Phi(.)$ and $\phi(.)$ are respectively the cumulative distribution function and the density of a standard normal distribution. Note that the EI is computed using both the mean predictor $\hat{y}$ and the prediction variance $\hat{s}$. It allows for a trade-off between exploitation and exploration by selecting new points near the current optimum (where $\hat{y}$ is small) and far from any observation (where $\hat{s}$ is large).

However, in high-dimensional problems, EGO suffers from some drawbacks related to the geometry of high-dimensional spaces. With a quick calculation, one can see that the volume of a thin layer of thickness ε in the border of the hypercube $[0,1]^d$ tends to 1 when d grows to infinity. In other words, in high-dimensional spaces, almost all the volume is located on the boundary of the hypercube. Added the fact that the EI is a highly multi-modal function, its optimization can also be difficult and often leads to new points located on one face of the hypercube, resulting in a optimizer too explorative. To avoid this behavior, it is common practice in high-dimensional Bayesian optimization to use trust regions in order to reduce the size of the search space (see e.g. Eriksson et al., 2019; Diouane et al., 2023; Binois and Wycoff, 2022). In these methods, the surrogate model and the optimization are restricted to a local neighborhood of the current best solution whose size increases if a new best solution is discovered, or decreases otherwise.

3 An example of Kriging failing in high dimension

In this section, we illustrate the inability of the classical Kriging approach to correctly estimate the length-scale hyperparameters for high-dimensional problems when not enough observations are available. While having too few observations can occur in any dimension, it is especially prevalent in higher dimensions where the thumb rule of $10d$ observations to build the model (Forrester et al., 2008) is often not affordable. In the context of Bayesian optimization, it is also common practice to start with a small initial DoE (about 2-3d) (Gaudrie, 2019; Garnett, 2023) and to allocate most of the budget for the acquisition points.

For this example, we consider two test functions for which we fit a Kriging model. First, the sphere test function defined as:

$$f_{sphere}(x_1, \ldots, x_d) := \sum_{\ell=1}^{d} (x_\ell - 0.5)^2, \quad 0 \leq x_\ell \leq 1, \quad \ell = 1, \ldots, d. \quad (8)$$

This function is simply a parabola centered in the middle of the hypercube $[0,1]^d$, and it is thus reasonable to expect a Kriging model to accurately approximate this simple smooth and convex function. Second, we consider trajectory samples of an isotropic GP:

$$Y_{iso}(.) \sim \mathcal{GP}\left(0, k_{iso}(.,.)\right), \quad (9)$$

where k_{iso} is a Matérn 5/2 covariance with isotropic length-scale $\theta_{true} = 3$ and amplitude $\sigma^2_{true} = 1$. These GP trajectories correspond to functions where the Kriging prior

assumption is exactly satisfied. For the numerical experiment, we consider a dimension $d = 50$, and we build an anisotropic Ordinary Kriging model using a varying number of samples $n \in [100, 1000]$. The d hyperparameters are optimized using the `DiceKriging R` package (Roustant et al., 2012) by MLE using a maximum of 500 L-BFGS-B iterations and 3 restarts. We compare the estimated hyperparameters with "real" ones, the latter being obtained after fitting a "reference GP" to 5000 samples of the Sphere function. For the GP trajectories, these reference hyperparameters are simply those used for sampling the trajectory, $\theta_{true} = 3$. To measure the global accuracy of the models, we compute the Q^2 coefficient based on $n_{test} = 10000$ random test points $\boldsymbol{x}_1^{(t)}, \ldots, \boldsymbol{x}_{n_{test}}^{(t)} \in [0, 1]^d$:

$$Q^2 := 1 - \frac{\sum_{k=1}^{n_{test}} \left(\hat{y}(\boldsymbol{x}_k^{(t)}) - f(\boldsymbol{x}_k^{(t)}) \right)^2}{\sum_{k=1}^{n_{test}} \left(f(\boldsymbol{x}_k^{(t)}) - \frac{1}{n_{test}} \sum_{l=1}^{n_{test}} f(\boldsymbol{x}_l^{(t)}) \right)^2}, \tag{10}$$

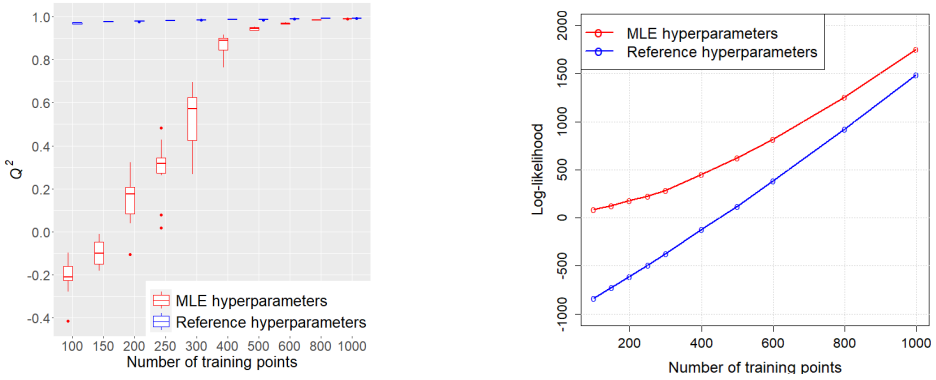
in which f is either of the test functions.

The results for 10 different initial DoEs are shown in Figure 2. Figures 2a and 2c show that the Ordinary Kriging fails in approximating the simple sphere function in the small data regime as highlighted by the very poor global accuracy, the Q^2 being even negative for less than 200 observations. This is especially problematic in early stages of Bayesian optimization since first surrogates with a poor global accuracy will negatively impact the discovery of areas of interest and decrease the convergence speed of the optimization. The good performance of the reference hyperparameters (blue boxplots), able to achieve a very good accuracy with only 100 observations, shows that the issue does not lie in the modelling capability of Kriging surrogates but rather in the hyperparameter values. Moreover, the log-likelihood plots in Figure 2b and 2d inform us that the issue does not reside in the convergence of the 50 dimensional hyperparameter optimization as the log-likelihood of the estimated hyperparameters is superior to the log-likelihood of the reference hyperparameters as expected. It rather seems that with few points, maximum likelihood is no longer a relevant metric to select good hyperparameters, since high likelihoods do not correspond to accurate models.

Finally, although the experiment was performed for high-dimensional test functions ($d = 50$), the issue is not directly the dimension itself, but rather the low number of observations relative to the dimension. In fact, we retrieve in Figure 2 the usual empirical rule of $10d$ observations to obtain a precision close to that of the model with reference hyperparameters. As such, similar results can be obtained in lower dimension by scaling down the number of samples. However, while it seems reasonable to obtain more than $10d$ samples to build the model in lower dimension, it is more difficult in higher dimensions with limited observations. Therefore, the described lack-of-accuracy issue is more prevalent for high-dimensional problems.

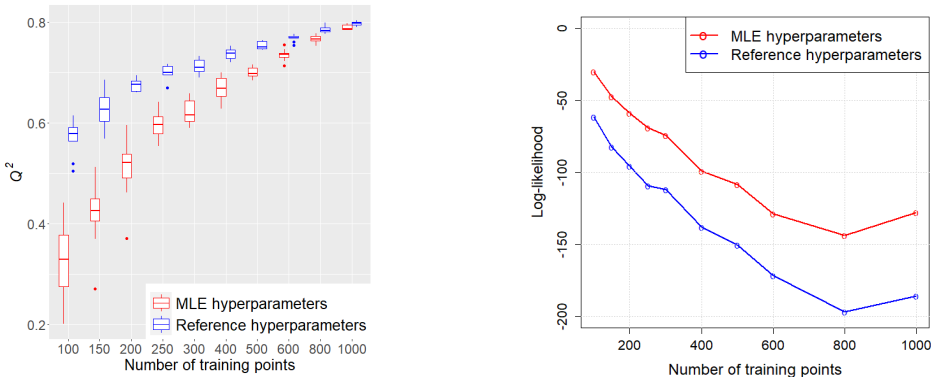
To avoid this issue, a usual solution is to reduce the dimension of the problem. While this facilitates the hyperparameter optimization, it raises the question of the accurate representation of the true function in a lower dimension space and of the information loss incurred. In the next section, an alternative approach based on a combination of Kriging models with random length-scales is presented. This method keeps the information about correlations among all variables and generalizes to any problem since it does not require the low-dimensional representation hypothesis.

Sphere function



(a) Q^2 (b) Log-likelihood

GP trajectories



(c) Q^2 (d) Log-likelihood

Figure 2: Q^2 of Kriging models (the higher the better) and corresponding log-likelihood for different number of training points over 10 independent runs. The top two figures give the results for the Sphere test function and the bottom two figures the results for GP trajectories. On the left are the accuracies of the models measured by the Q^2, on the right are the corresponding log-likelihood. The results in red correspond to the Kriging model with hyperparameters estimated by maximum likelihood, those in blue to the model with reference hyperparameters.

4 Combination of Kriging models with random length-scales

To address the issues raised in the previous section, Appriou et al. (2023) introduced a method to bypass the hyperparameter optimization by using a combination of Kriging models with random length-scales. The idea was to replace the difficult and sometimes costly length-scale estimation by the easier optimization of the weights in the combination. This results in an easier-to-build model, and which is more accurate in such cases where the length-scales are wrongly estimated by maximum likelihood estimation. In the aforementioned paper, the focus was on the weighting methods and on studying the impact of the number of sub-models. In the present paper, we develop a more effective methodology to sample the random sub-models length-scales, and the combination is further developed to accommodate to Bayesian optimization by providing a method to obtain the prediction variance.

The combined model M_{tot} proposed in Appriou et al. (2023) writes as:

$$M_{tot}(\boldsymbol{x}) := \sum_{i=1}^{p} w_i(\boldsymbol{x}) M_i(\boldsymbol{x}), \tag{11}$$

where w_i, $i = 1, \ldots, p$, are the weights of the p sub-models M_i. In the following sub-sections, we first discuss the choice of the sub-models, and in particular how to sample their random length-scales. Then, the second sub-section describes how the sub-models are weighted, and finally, the last two sub-sections detail how the prediction variance of the combined model is obtained, by modeling the correlation between sub-models and discussing the estimation of the variance amplitude hyperparameter.

4.1 Choice of the sub-models

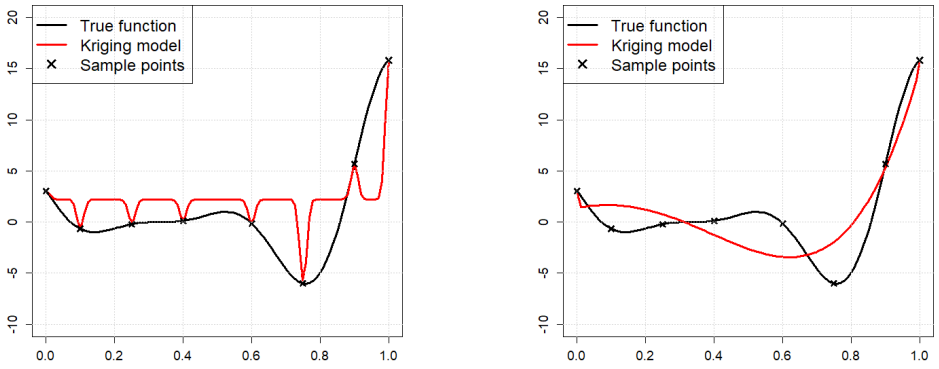
The sub-models we consider are ordinary Kriging models with length-scales $\boldsymbol{\theta}_i$ chosen randomly, hence not as the result of a maximum likelihood estimation:

$$M_i(\boldsymbol{x}) := \mathbf{E}(Y_{\boldsymbol{\theta}_i}(\boldsymbol{x})|\mathcal{D}_i) = \mu_i + k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}_i) k_{\boldsymbol{\theta}_i}(\boldsymbol{X}_i, \boldsymbol{X}_i)^{-1}(\boldsymbol{Y}_i - \mu_i),$$

In the following, we will assume that the training data set of each sub-model is the entire data set: $\mathcal{D}_i = (\boldsymbol{X}_i, \boldsymbol{Y}_i) = (\boldsymbol{X}, \boldsymbol{Y})$ though each sub-model could consider a specific subset of $(\boldsymbol{X}, \boldsymbol{Y})$. The sampling of the length-scales $\boldsymbol{\theta}_i$ is therefore critical as it is the only source of variability between the sub-models.

The sampling of length-scales should enable a wide diversity of sub-model behaviors so that the combination can select the appropriate ones through their weights w_i. To this aim, the length-scales must be dissimilar enough, but at the same time, they need to lie in an appropriate range to avoid degenerated sub-models occurring when the length-scales are either too large or too small compared to observed distances between samples. In particular, small length-scales will result in all cross-correlations being close to 0. In this case, as illustrated in Figure 3a, the Kriging prediction equals to its mean function $\mu(.)$ almost everywhere with spikes to interpolate the observations. In the contrary, too large length-scales will result in all correlation being close to 1 resulting in a correlation matrix which can be ill-conditioned and inverting it - when possible - can cause numerical instabilities. This issue can be resolved by adding a small nugget term in the diagonal of $\boldsymbol{K}_{\boldsymbol{\theta}}$ at the cost of loosing the interpolating property as illustrated in Figure 3b. The sampling procedure must not produce such non-informative sub-models. Note that this corresponds to only accepting robust sub-models as defined in Gu et al. (2018). In the original random length-scales combination paper (Appriou et al., 2023), to avoid the degenerated cases, the length-scales were sampled uniformly in a bounded interval computed according to their influence on the correlations. However, this method tended to produce too many length-scales on the smaller side of the acceptable range which tends to worsen the sub-models. Instead of trying to design a more suited distribution to sample from, we propose an other non-parametric approach to better sample the length-scales based on the entropy of the correlation.

First, we will derive the entropy for a Gaussian correlation and detail the sampling strategy. Then we will extend the method to any other correlation. Let us consider that design points are distributed as a random vector $\mathbf{X} = (X^{(1)}, \ldots, X^{(d)})$, with i.i.d components with common variance $\sigma^2_{\mathbf{X}}$ and kurtosis $\kappa_{\mathbf{X}}$. Let D^2 be the random squared distance



(a) Small length-scale (b) Large length-scale

Figure 3: Examples of the two degenerated cases in 1D. Left: ordinary Kriging with a small length-scale $\theta = 10^{-2}$, right ordinary Kriging with a large length-scale $\theta = 3$ and a small nugget term $\varepsilon^2 = 10^{-8}$.

between two independent points $\mathbf{X}$ and $\mathbf{X}'$. For a large enough dimension d:

$$D^2 := \sum_{\ell=1}^{d} \left(X_\ell - X'_\ell\right)^2 \sim \mathcal{N}\left(2d\sigma_{\mathbf{X}}^2,\ 2d\sigma_{\mathbf{X}}^4(\kappa_{\mathbf{X}}+1)\right). \tag{12}$$

For a Gaussian correlation function, the random correlation between these two points for a given length-scale θ is then:

$$R_\theta := \exp\left(-\frac{D^2}{2\theta^2}\right) \sim \log\mathcal{N}\left(-\frac{\sigma_{\mathbf{X}}^2}{\theta^2}d,\ \frac{\sigma_{\mathbf{X}}^4}{2\theta^4}(\kappa_{\mathbf{X}}+1)d\right). \tag{13}$$

Finally, the entropy of the correlation isthe entropy of the log-normal distribution in (13):

$$H(R_\theta) := \mathbf{E}[-\log f_{R_\theta}(R_\theta)] = -\frac{\sigma_{\mathbf{X}}^2}{\theta^2}d + \frac{1}{2}\log\left(2\pi\frac{\sigma_{\mathbf{X}}^4}{2\theta^4}(\kappa_{\mathbf{X}}+1)d\right) + \frac{1}{2}, \tag{14}$$

where f_{R_θ} is the density of R_θ. Figure 4 shows the entropy as a function of the length-scale θ for a dimension $d = 50$. The entropy decreases rapidly for length-scales below a critical value, while its drop is less pronounced for large length-scales. It is therefore especially important to avoid too small length-scales. Additionally, we can remark on the length-scale θ^* which maximizes the entropy:

$$\theta^* := \arg\max_\theta H(R_\theta) = \sigma_{\mathbf{X}}\sqrt{d}.$$

This is the length-scale which maximizes the potential to capture the information of the training set, prior to any observation. We also note that this length-scale is such that the square of the length-scale is of the same order as the squared-distances between samples: $2\theta^{*2} = \mathbf{E}[D^2]$. Obrezanova et al. (2007) used this definition as an initial value prior to any observation.

The entropy obtained in equation (14) measures the variability of the correlations. The two degenerated cases described above (correlation is always 0 for small length-scales or always 1 for large length-scales) correspond to cases where there is no variability and where the entropy is small:

$$\lim_{\boldsymbol{\theta}\to 0} H(R_\theta) = -\infty, \quad \text{and} \quad \lim_{\boldsymbol{\theta}\to\infty} H(R_\theta) = -\infty.$$

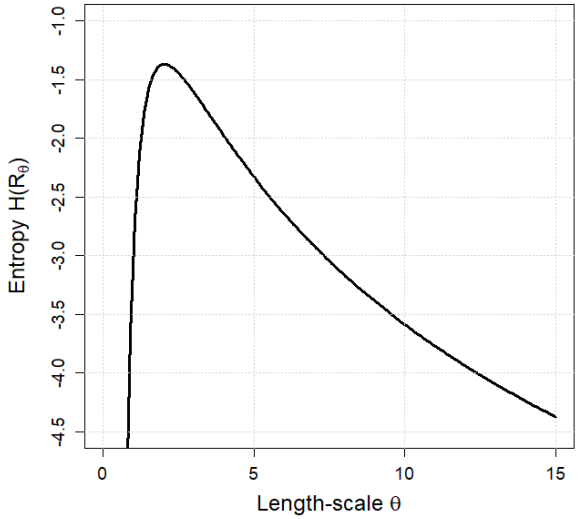


Figure 4: Entropy of a Gaussian correlation as a function of the length-scale θ. In this example, $d = 50$ and $\mathbf{X}$ are uniformly distributed ($\sigma^2_{\mathbf{X}} = \frac{1}{12}$ and $\kappa_{\mathbf{X}} = \frac{9}{5}$).

For a general non-Gaussian correlation function, e.g. a Matérn one, the entropy cannot be obtained analytically. In this case, we use instead a non-parametric estimate of the entropy (Ahmad and Lin, 1976; Beirlant et al., 1997) given by:

$$\hat{H}(R_\theta) = -\frac{1}{n_{corr}} \sum_{k=1}^{n_{corr}} \log \hat{f}_{R_\theta}\left(k_\theta^{(k)}\right), \tag{15}$$

where $n_{corr} = \frac{n(n-1)}{2}$ is the number of observed correlations $k_\theta^{(k)}$, $k = 1, \ldots, n_{corr}$ between distinct samples, and $\hat{f}_{R_\theta}$ is a kernel density estimate of correlations density f_{R_θ}.

To obtain varied length-scales in an appropriate range for the sub-models, we sample length-scales corresponding to high entropies:

$$f(\theta^{(\ell)}) \propto \exp\left(H(R_{\theta^{(\ell)}})\right), \quad \ell = 1, \ldots, d, \tag{16}$$

where $f(\theta^{(\ell)})$ is the density from which $\theta^{(\ell)}$ is sampled. Note that each of the d length-scales of the vector $\boldsymbol{\theta} = (\theta^{(1)}, \ldots, \theta^{(d)})$ are sampled independently. Note also that since the differential entropy can be negative (contrarily to the entropy for discrete random variables), a positive transformation of the entropy (exponential) is used in equation (16).

4.2 Weighting scheme

To obtain the weights of the sub-models, Appriou et al. (2023) compared different methods coming from the model combination literature (see e.g. Ginsbourger et al., 2008; Cao and Fleet, 2014; Deisenroth and Ng, 2015; Rullière et al., 2018). In this paper, we follow the definition proposed in Acar and Rais-Rohani (2009) and Viana et al. (2009) which aim at minimizing the global mean-square error (MSE) of the combination given by:

$$\mathbf{E}\left[\left(M_{tot}(\boldsymbol{X}) - y(\boldsymbol{X})\right)^2\right]. \tag{17}$$

As the MSE is a global measure, we keep the weights constant throughout the design space: $\boldsymbol{w}(\boldsymbol{x}) = \boldsymbol{w}$. Since we only have access to a finite number of observations, a discrete

approximation of the MSE is obtained using leave-one-out cross-validation (LOOCV):

$$e_{LOOCV}(M_{tot}) := \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i=1}^{p} w_i M_{i_{-k}}(\boldsymbol{x}_k) - y(\boldsymbol{x}_k)\right)^2 = \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w}, \tag{18}$$

where $M_{i_{-k}}$ is the ith Kriging sub-model built by removing the kth sample $\boldsymbol{x}_k$. The components of the matrix $\boldsymbol{C} \in \mathbb{R}^{p\times p}$ are $c_{ij} = \frac{1}{n}\boldsymbol{e}_i^\top \boldsymbol{e}_j$ with $\boldsymbol{e}_i$ the LOOCV vector for the ith sub-model: $\boldsymbol{e}_i = (\boldsymbol{e}_i^{(1)}, \dots, \boldsymbol{e}_i^{(n)})$. For Kriging models, these residuals can be computed easily without the need to build n distinct models using the conditional Gaussian and block-matrix inversion formulas (Dubrule, 1983; Ginsbourger and Schärer, 2021):

$$\boldsymbol{e}_i^{(k)} = \frac{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)]_k}{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{k,k}}. \tag{19}$$

The weights $\boldsymbol{w}$ are obtained by minimizing (18):

$$\boldsymbol{w}^* = \arg\min_{\boldsymbol{w}} \ \boldsymbol{w}^\top \boldsymbol{C}\boldsymbol{w}, \quad \text{subject to } \sum_{i=1}^{p} w_i = 1.$$

Using a Lagrange multiplier and setting the derivatives to zero:

$$\boldsymbol{w}^* = \frac{\boldsymbol{C}^{-1}\mathbf{1}}{\mathbf{1}^\top \boldsymbol{C}^{-1}\mathbf{1}}. \tag{20}$$

One issue pointed out in Viana et al. (2009) and Appriou et al. (2023) is that w can be negative or greater than one, which can lead to over-fitting, especially when the number of sub-models p is large. To solve this issue, instead of combining all sub-models at once, we combine them two-by-two, following a binary tree structure illustrated in Figure 5. With this definition, we can enforce $w \in [0, 1]$ by imposing $w = 0$ when $w < 0$ and $w = 1$ when $w > 1$, while keeping the sum of weights equal to one. In practice, it is also possible to implement a threshold to induce more sparsity in the weights. In addition, this two-by-two combination scheme will also prove itself convenient in the following subsection for computing the variance of the combination.

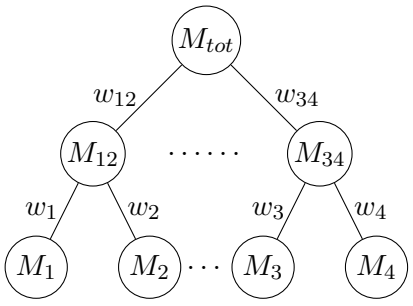


Figure 5: Illustration of the sub-models combination following a binary tree structure (3 levels for 4 sub-models here).

4.3 Variance of the combination

As noted in section 2.3, the prediction error is essential in Bayesian optimization as it is used inside most acquisition functions. For Kriging models, it is obtained naturally in

analytical form through the variance of the prediction as in equation (3). Each individual model in the combination is a Kriging model with $Y_i \sim \mathcal{GP}\left(\mu_i, \sigma^2 k_{\boldsymbol{\theta}_i}(.,.)\right)$ and has a prediction variance:

$$\hat{s}_i^2(\boldsymbol{x}) = \mathbf{E}\left[(M_i(\boldsymbol{x}) - Y_i(\boldsymbol{x}))^2\right] = \sigma^2\left(k_{\boldsymbol{\theta}_i}(\boldsymbol{x},\boldsymbol{x}) - k_{\boldsymbol{\theta}_i}(\boldsymbol{x},\boldsymbol{X})k_{\boldsymbol{\theta}_i}(\boldsymbol{X},\boldsymbol{X})^{-1}k_{\boldsymbol{\theta}_i}(\boldsymbol{X},\boldsymbol{x})\right).$$

However, since the correlation structure between the sub-models is unspecified, we cannot access directly to the variance of the combination. In the literature, several works aim at estimating prediction errors for surrogate models where the latter is not readily available. For example, Viana et al. (2013) and Viana and Haftka (2009) import the uncertainty estimate of a Kriging sub-model to sub-models not equipped with a measure of uncertainty (e.g. RBFs) to perform Bayesian optimization. Den Hertog et al. (2006) use parametric bootstrapping instead of the classical formula to estimate the variance of a Kriging model. Bootstrapping is also used in Kleijnen et al. (2012) and Kleijnen (2014) to perform EGO using this corrected error estimate. In the field of neural networks, various methods were also developed to estimate the uncertainty of the prediction (see for instance Papadopoulos et al., 2001; Khosravi et al., 2010; Pearce et al., 2018; Abdar et al., 2021). Conformal prediction (Lei et al., 2018; Romano et al., 2019) is an other empirical approach to build prediction interval without making distributional assumptions. Berk et al. (2013) and Bachoc et al. (2020) propose a procedure to construct valid confidence intervals post model selection for linear models. Acharki et al. (2023) use cross-validation to calibrate prediction intervals for GPs in the case of model misspecification. Another related class of method are the mixture models (see Yuksel et al., 2012, for a review) which is a Bayesian approach to combining models in which the posterior predictive distribution of the combination is given as a mixture of posterior predictive distributions of several sub-models. In the context of Kriging, Pronzato and Rendas (2017) builds a fully-Bayesian mixture of sub-models with different covariances to obtain a non-stationary model, and Ginsbourger et al. (2008) uses a mixture of Kriging models in the context of Bayesian optimization. While the mixture models give the same mean prediction as the linear combination of models used in this paper, their variance differs. The variance of a mixture of Kriging models can be obtained analytically and the mixed expected improvement can be expressed as the convex combination of the individual expected improvements. However, the variance of a mixture is larger than that of a linear combination of models. In addition, the weights of the mixture can be interpreted as the probability to randomly select one model among all others, even in high-dimension where no individual sub-model is expected to be clearly better than all others. It is preferable to have a combination as it outperforms the best sub-model as shown in Appriou et al. (2023). In this section, we introduce a method to obtain the variance of the linear combination of Kriging models introduced in the previous section. This approach is based on a LOOCV strategy and can be seen as a natural extension of the linear combination of models hypothesis.

The variance of the combination depends on the global covariance structure between sub-models. We make the hypothesis that the underlying Gaussian process Y_{tot} is also a linear combination of independent Gaussian processes Y_i, $i = 1, \ldots, p$ with the same fixed length-scales as those of the M_i:

$$Y_{tot} = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i Y_i, \quad \text{with } Y_i \sim \mathcal{GP}\left(\mu_{tot}, k_{\boldsymbol{\theta}_i}(.,.)\right), \quad \sum_{i=1}^{p} \alpha_i = 1. \tag{21}$$

Here, σ_{tot}^2 is the amplitude of the variance whose estimation is discussed in the next sub-section. Without loss of generality, we consider centered GPs ($\mu_{tot} = 0$) as we are only

interested in the associated covariance:

$$\mathbf{Cov}(Y_{tot}(\boldsymbol{x}), Y_{tot}(\boldsymbol{x}') := \sigma_{tot}^2 k_{tot}(\boldsymbol{x}, \boldsymbol{x}') = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i^2 k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{x}'). \tag{22}$$

Note that the weights α_i in the combination of GPs (21) are different from the weights w_i of the combination of models in (11). Similarly to the combination of models in Figure 5, we also assume that the GPs are combined two-by-two as illustrated in Figure 6. This enables the derivation of an analytical expression for the weights $\boldsymbol{\alpha}$ using a LOOCV procedure. Notice the link between this approach and a mixture of models. Indeed, if $\boldsymbol{\alpha}^2 = \boldsymbol{w}$, the MSE of a mixture equals the MSE of the combination:

$$\mathbf{E}\left[(M_{mix}(\boldsymbol{x}) - Y_{mix}(\boldsymbol{x}))^2\right] = \mathbf{E}\left[(M_{tot}(\boldsymbol{x}) - Y_{tot}(\boldsymbol{x}))^2\right], \quad \text{when } k_{tot}(\boldsymbol{x}, \boldsymbol{x}') = \sum_{i=1}^{p} w_i k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{x}'). \tag{23}$$

Details on the proof of this proposition are given in Appendix A.1.

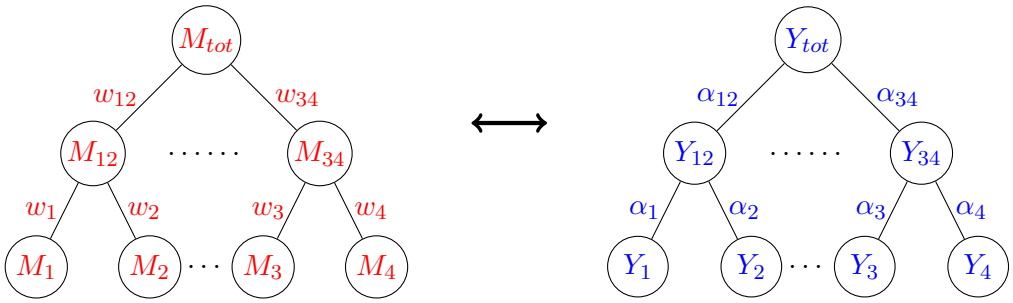


Figure 6: Illustration of the two-by-two GPs combination following a binary tree structure (3 levels for 4 individuals GPs here) with weights $\boldsymbol{\alpha}$ (right) associated to the two-by-two combination of sub-models with weights $\boldsymbol{w}$ (left).

To obtain the weights $\boldsymbol{\alpha}$, we aim at minimizing the discrepancy between the combination of sub-models that we obtained, and the new hypothesis we introduced about the combination of GPs. This discrepancy can be measured as the expected MSE of the combined model with respect to Y_{tot}. For a combination of two models, with $Y_{tot} = \alpha Y_1 + (1-\alpha)Y_2$, the discrepancy is minimal for:

$$\alpha^* = \arg\min_{\alpha} \mathbf{E}\left[\mathbf{E}\left[(wM_1(\boldsymbol{x}) + (1-w)M_2(\boldsymbol{x}) - \alpha Y_1(\boldsymbol{x}) - (1-\alpha)Y_2(\boldsymbol{x}))^2 \,|Y_1, Y_2\right]\right]. \tag{24}$$

Finally, by approximating the global MSE using the LOOCV error as in equation (18), we obtain an analytical expression for α^*:

$$\alpha^* = \frac{a_1(w)}{a_1(w) + a_2(w)}, \tag{25}$$

where $a_1(w)$ and $a_2(w)$ depend on the weight w of the combination of models and on the LOOCV vector of the sub-models, and are given by:

$$a_1(w) = w^2 \mathbf{E}(e_{LOOCV}(M_1)|Y_2) + (1-w^2)\mathbf{E}(e_{LOOCV}(M_2)|Y_2), \tag{26}$$

$$a_2(w) = (1-w^2)\mathbf{E}(e_{LOOCV}(M_1)|Y_1) + (1-(1-w^2))\mathbf{E}(e_{LOOCV}(M_2)|Y_1). \tag{27}$$

Here, $e_{LOOCV}(M_i)|Y_j, \quad i,j \in \{1,2\}$, is the LOOCV error of the model M_i when the global GP is $Y_j(.) \sim \mathcal{GP}(0, k_{\boldsymbol{\theta}_j}(.,.))$, which can be obtained with the Kriging LOO formula (19):

$$e_{LOOCV}(M_i)|Y_j = \sum_{k=1}^{n} \left(\frac{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{k,}}{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{k,k}} Y_j(\boldsymbol{X}) \right)^2, \quad \text{with } Y_j(\boldsymbol{X}) \sim \mathcal{N}(0, k_{\boldsymbol{\theta}_j}(\boldsymbol{X}, \boldsymbol{X})). \tag{28}$$

Details of the proof are given in Appendix A.2.

Finally, the variance of the combination is the standard Kriging variance using the global covariance in (22) with the weights $\boldsymbol{\alpha}$ computed with (25):

$$\hat{s}^2_{tot}(\boldsymbol{x}) = \mathbf{Var}(Y_{tot}(\boldsymbol{x})|\mathcal{D}) = \sigma^2_{tot} \left(k_{tot}(\boldsymbol{x}, \boldsymbol{x}) - k_{tot}(\boldsymbol{x}, \boldsymbol{X}) k_{tot}(\boldsymbol{X}, \boldsymbol{X})^{-1} k_{tot}(\boldsymbol{X}, \boldsymbol{x}) \right). \tag{29}$$

Notice that if the combined GP hypothesis in (21) was used from the start to obtain the mean prediction, we would have recovered an additive Kriging model (Durrande et al., 2012):

$$\tilde{M}(\boldsymbol{x}) := \mathbf{E}(Y_{tot}(\boldsymbol{x})|\mathcal{D}) = \mu_{tot} + k_{tot}(\boldsymbol{x}, \boldsymbol{X}) k_{tot}(\boldsymbol{X}, \boldsymbol{X})^{-1} (Y - \mu_{tot}). \tag{30}$$

However, the issue in (30) is that it involves $k_{tot}(\boldsymbol{X}, \boldsymbol{X})^{-1}$, the inverse of a sum of correlation matrices. As no direct formula exists for the inverse of a sum of matrices, the optimization of the weights for this alternative model (using a LOOCV procedure for instance) would involve a large number of matrix inversions and an inner optimization loop similarly to the classical hyperparameter optimization in ordinary Kriging. This is precisely what we aim at avoiding with the proposed method, in which both the weights of the combination in (20) and the weights for the variance in (25) are obtained analytically.

4.4 Amplitude of the variance

The final step to completely obtain the variance of the combination is to set the hyperparameter σ^2_{tot} in (21) which is used to calibrate the amplitude of the variance. In leave-one-out strategies, this hyperparameter is typically set by observing that the normalized LOO residuals should be normally distributed if the model is well-specified. Based on this, the usual method is to set the empirical variance of the normalized LOO residuals to 1 (see for instance Cressie, 1993, p. 102; Bachoc, 2013):

$$\hat{\sigma}^2_{LOO} = \frac{1}{n} \sum_{k=1}^{n} \left(\frac{e^{(k)}_{LOO}}{s^{(k)}_{LOO}} \right)^2, \tag{31}$$

where, for $k = 1, \ldots, n$, $e^{(k)}_{LOO}$ are the LOO residuals of the model that can be obtained using the LOO formula (19) for the combination of Kriging models. $s^{(k)}_{LOO}$ are the LOO standard deviations, obtained similarly:

$$e^{(k)}_{LOO} := y(\boldsymbol{x}_k) - M_{-k}(\boldsymbol{x}_k) = \sum_{i=1}^{p} w_i \frac{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}(\boldsymbol{Y} - \mu_i)]_k}{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{k,k}}, \quad \text{and } s^{(k)}_{LOO} := \hat{s}_{-k}(\boldsymbol{x}_k) = \frac{1}{\sqrt{[\boldsymbol{K}_{tot}^{-1}]_{k,k}}}.$$

However, when the model is ill-defined (which is often the case in practice), the LOO residuals are not normally distributed and in particular, the presence of outliers can lead to an over-estimation the variance amplitude. Thus, we propose a similar approach more

robust to outliers by fitting the amplitude hyperparameter using the empirical interquartile distance instead of the empirical variance:

$$\hat{\sigma}_{tot} = \frac{IQ\left(\frac{e_{LOO}}{s_{LOO}}\right)}{IQ_{norm}}, \tag{32}$$

where $IQ(.)$ designates the empirical interquartile distance, and IQ_{norm} is the interquartile distance of a standard normal distribution.

5 Numerical results

In this section, we investigate the performances of the combination of Kriging models introduced in the previous section, and we compare it to ordinary Kriging. The focus is on high-dimensional problems where the number of samples used to build the model is limited, since we showed in section 3 that ordinary Kriging can fail to provide accurate models when the length-scales hyperparameters are estimated using MLE. As in section 3, we consider two analytical test functions in dimension 50 which are the sphere function (equation (8)) and trajectory samples from an isotropic GP (equation (9)). The sphere function is an example of a simple convex function which should be simple to model and to optimize. The GP trajectories are more complex multimodal functions, more difficult to optimize, and closer to practical usecases. In addition to these analytical functions, to validate the methods on a more realistic problem, they are also compared on a real-world application: the design of an electrical machine. The shape of the machine is parameterized by $d = 37$ design variables representing the position and size of air holes and magnets as well as the radius of the machine. The layout of the machine is illustrated in Figure 7. The performance of a machine is assessed by two objective functions which are its consumption and its cost. In addition, a valid machine must satisfy 10 constraints related to the dynamic of the car (e.g. maximum speed, acceleration, ...), to the dynamic of the machine (e.g. oscillation amplitudes, ...), and to the dimensioning of the reducer.

First, we will compare the global accuracy of the different models on these functions for a fixed number of observations. We will also assess the quality of the prediction intervals constructed. These two aspects are important in Bayesian optimization as an accurate surrogate model with correct error estimation should result in faster convergence, thus requiring less function evaluations to reach the optimum. Finally, since we are rather interested in faster convergence towards optimal designs than in obtaining a better global accuracy, we compare the performance of Bayesian optimization (EGO procedure) using the combination of Kriging models to that of Bayesian optimization equipped with standard ordinary Kriging.

5.1 Model precision and confidence intervals assessment

The global accuracy and confidence interval precision of the different methods are evaluated for the three test functions described above. The reference method is ordinary Kriging using a Matérn 5/2 covariance with hyperparameters estimated by MLE. The optimization is performed using the package `DiceKriging` in the `R` language (Roustant et al., 2012) with 300 maximum iterations of L-BFGS-B, and the length-scales lower and upper bounds are fixed to 0.1 and 20 respectively. The second method is the combination introduced in this paper using $p = 16$ sub-models (5 levels in the tree structure) with random length-scales sampled following the procedure in section 4.1, with weights given by equation (20),

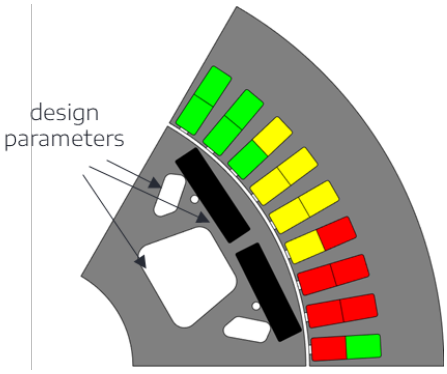


Figure 7: Layout of the electrical machine. The 37 design parameters are the size and position of the air holes (in white) and of the magnets (in black), as well as the radius of the machine.

and with variance obtained using equations (25), (22), (29) and (32). Finally, for the GP trajectories where the true length-scale $\theta_{true} = 3$ is known, we also compare both methods to a Kriging model with length-scales θ_{true} as a reference. Each model is built using the same space-filling set of training points $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{n_{train}} \in [0,1]^d$ obtained with Latin Hypercube Sampling (LHS). For the sphere function, the number of training points is $n_{train} = 250$, and for the GP trajectories and the real-world application which are more complex we have $n_{train} = 500$. The accuracy of each model is evaluated by computing the Q^2 (10) on a set of $n_{test} = 5000$ test points $\boldsymbol{x}_1^{(t)}, \dots, \boldsymbol{x}_{n_{test}}^{(t)} \in [0,1]^d$ sampled uniformly on the design space. To assess the precision of the confidence intervals, the coverage probabilities, corresponding to the proportion of test data that lies in the $\alpha\%$ confidence interval, are computed on this same test set. For different levels of confidence α, we plot the coverage probabilities which should be equal to their theoretical level if the confidence intervals are well-calibrated. Each experiment is repeated for 10 different random seeds and theresults are presented in boxplots. For the electrical machine, as there are 12 functions to model (2 objectives and 10 constraints), the boxplot shows the performance averaged over the 12 functions.

The results for the three test functions are given in Figure 8. In each case, the combination of Kriging models is globally more accurate than ordinary Kriging with hyperparameters estimated by MLE as measured by the Q^2 indicator. In Figure 8d, the combination even manages to achieve a precision similar to that of the reference model (with the true length-scales $\boldsymbol{\theta}_{true}$). The poor performances of ordinary Kriging can be explained by what was observed in Figure 2: MLE-optimal hyperparameters do not necessarily correspond to good Q^2 for high-dimensional problems with a limited number of observations. This is verified explicitly for the GP trajectories test case where the MLE-estimated hyperparameters differ from their true values resulting in a worse accuracy. In addition, unlike in Section 3 where the parameters of the hyperparameters optimizer were set to ensure convergence with 500 iterations and several restarts, here we take a lighter though more common setting in BO, with 300 iterations and no restart. This can lead to additional variability in the results for ordinary Kriging as the difficult length-scale optimization may not have fully converged in every cases. Still, for the GP trajectories, the log-likelihood at the end of the inner optimization is superior to the log-likelihood of the reference hyperparameters, even if the maximum likelihood optimization may not have fully converged. Overall, this shows that estimating the hyperparameters by MLE may not guarantee a good model.

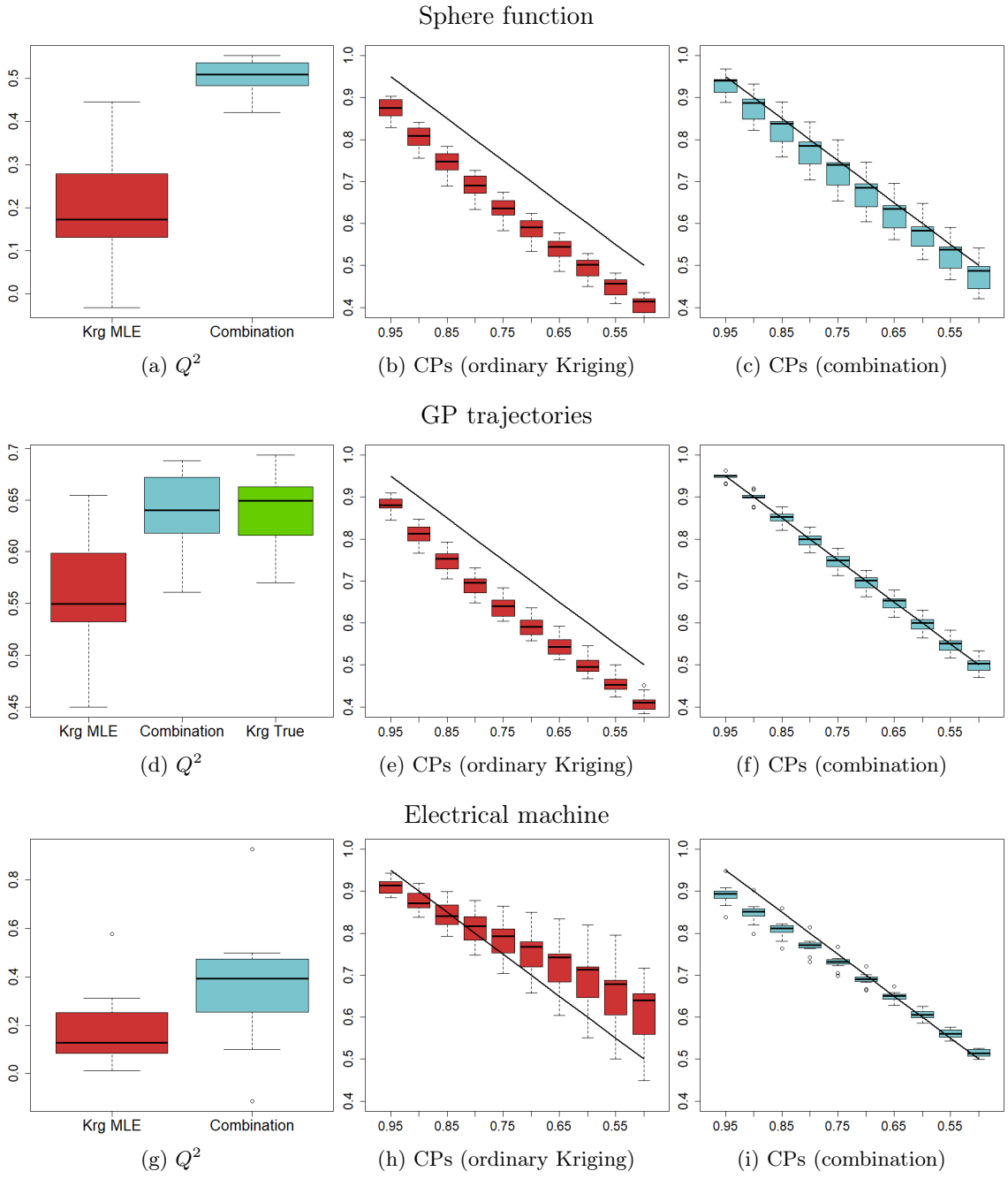
Regarding the prediction intervals, we see in Figures 8b and 8e that the ordinary Kriging is

overconfident in its predictions as all probability levels CPs are lower than their theoretical values. This phenomenon was expected as the Kriging variance (3) usually underestimates the true prediction variance since it is obtained using an estimated covariance whose uncertainty on the estimation is not considered (Cressie, 1993, p. 127; Den Hertog et al., 2006). In contrast, the prediction intervals for the combination in Figures 8c and 8f are well calibrated as the CPs are aligned with the line $y = x$. For the electrical machine in Figure 8h and 8i, we observe a different behavior: the CPs are higher than their theoretical values at low probability levels and lower at high probability levels. One possible explanation is that the true prediction distribution is better approximated by a heavier-tailed distribution rather than by a Gaussian one. Still, we see that the confidence intervals for the combination are better calibrated than the original ones.

5.2 Bayesian optimization

The previous results show that the combination of Kriging models produces more accurate surrogates than ordinary Kriging with few points in high-dimension. In this section, we investigate whether more accurate models translate into faster convergence in BO. We perform EGO for the same three test functions: the sphere function given in equation 8, GP trajectories in equation 9, and the real-world application of an electrical machine. Note that in this paper, we only consider single-objective problems. As such, for the optimization of the electrical machine, we will only maximize on the first constraint which gives the maximum speed of the vehicle, as high values of this constraint are well correlated with good performances of the machine in general. However, as for ordinary Kriging, the combination can also be employed for constrained multi-objective Bayesian optimization using adapted acquisition criterion such as the Expected Hypervolume Improvement (EHVI) (see for instance Forrester and Keane, 2009). For the sphere function and the GP trajectories, we consider 3 cases with varying dimension $d = 15$, 30, and 50. Each time, an initial design plan with $n_{init} = 2d$ is built by LHS, then n_{iter} iterations of EGO are performed to add new samples. For the dimensions $d = 15$ and 30, we take $n_{iter} = 10d$ and for $d = 50$, we take $n_{iter} = 8d$. At each iteration, the ordinary Kriging model and the combination are built using the same settings as described in the Section 5.1. The acquisition criterion is the Expected Improvement given in equation (7), and the EI maximization is performed with the `R` package `DiceOptim` (Roustant et al., 2012). To improve the performances, especially for high dimensions, we also use the TREGO (Diouane et al., 2023) algorithm to implement trust regions. Note that while the global minimum of the sphere function is known ($\min_{\boldsymbol{x}} f_{sphere}(\boldsymbol{x}) = 0$), the global optimum for the GP trajectories and for the electrical machine are unknown, hence we only compare the performances of the two methods. The optimization is repeated for 10 different random seeds, with different initial design of experiments each time.

Figure 9 shows the evolution of the current best solution during the EGO iterations for the two methods. For the sphere function (Figure 9a, 9b and 9c) both methods are able to approach the global optimum of the function. However, for all dimensions, EGO with the combination converges faster than with ordinary Kriging, especially at the start of the optimization. The solution found after n_{iter} iterations with ordinary Kriging is obtained within 2 to 3 times less iterations with the combination because the combination is globally more accurate than ordinary Kriging when very few samples are present, as is the case at the beginning of the optimization. This is confirmed by Figure 10 giving the global accuracy of both models during the EGO optimization for the sphere function with $d = 30$. After only a few iterations, the combination is able to achieve a good global accuracy in contrast



(a) Q^2 (b) CPs (ordinary Kriging) (c) CPs (combination)

(d) Q^2 (e) CPs (ordinary Kriging) (f) CPs (combination)

(g) Q^2 (h) CPs (ordinary Kriging) (i) CPs (combination)

Figure 8: Left column: accuracy of the models measured by the Q^2. Central and right column: precision of the confidence intervals for the ordinary Kriging (center) and the combination (right), the closest to the $y = x$ line the better. First row: sphere function, middle row: GP trajectories, bottom row: electrical machine. Red boxes stand for Kriging models with hyperparameters obtained by MLE and blue boxes for the combination of Kriging models.

to ordinary Kriging where the Q^2 is still negative. Because it is more globally accurate at the start of the optimization, the combination of Kriging models is able to identify the interesting areas of the design space to explore, near the true global optimum of the function, more rapidly. Similarly, the convergence is improved with the combination on the GP trajectories with $d = 15$ (Figure 9d). For the cases with $d = 30$ and 50 in Figures 9e and 9f, in addition to a faster convergence, we also observe that the final solution is significantly

better than the one obtained with ordinary Kriging because, since the GP trajectories are multimodal functions, the less globally accurate ordinary Kriging models get trapped by a local optimum and miss the better one found by the combination. The same behavior is observed on the real-world application in Figure 9g (maximization problem in this case), where the EGO with the combination discovers a better solution faster.

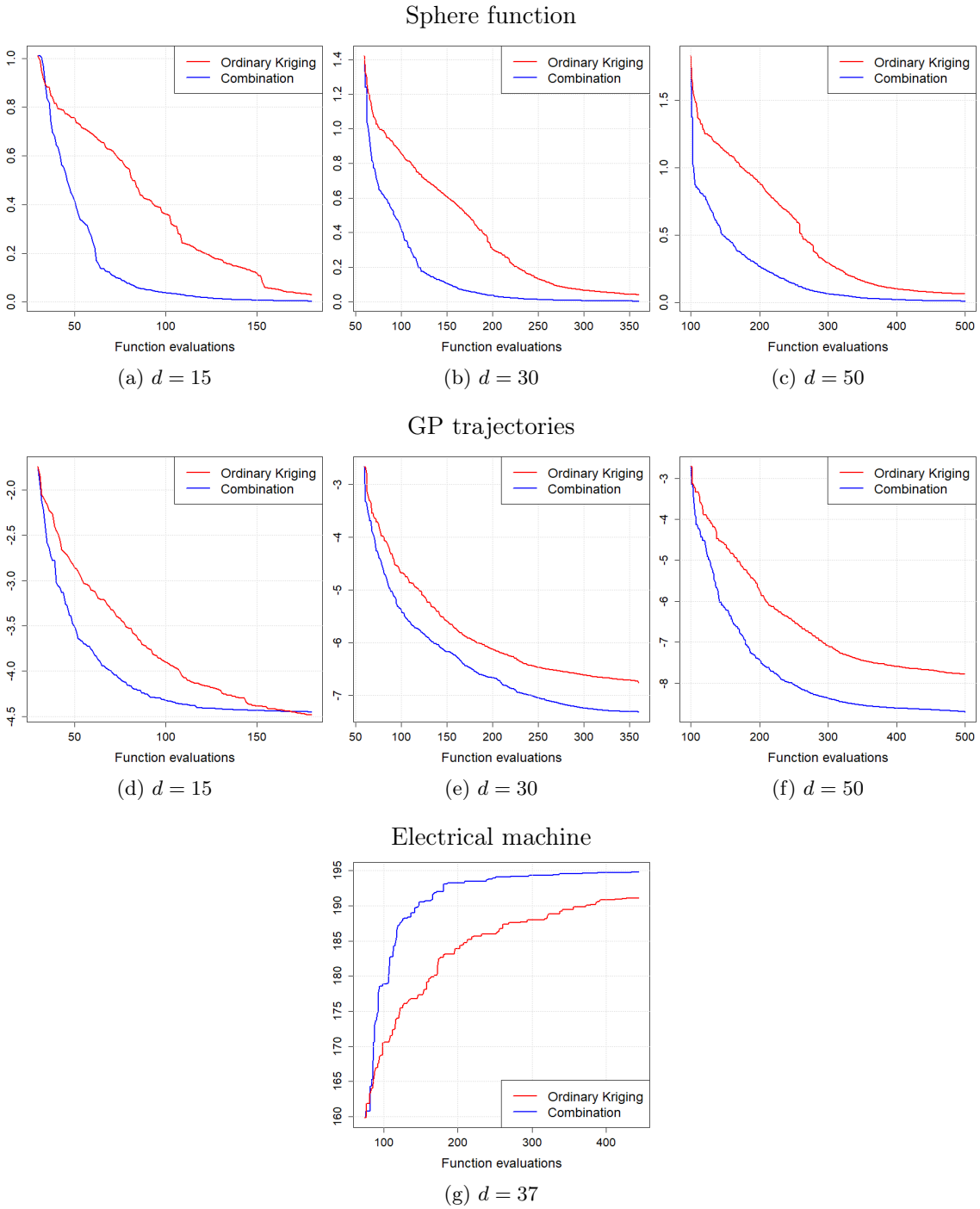


Figure 9: Evolution of the current best solution during the EGO iterations. The x-axis corresponds to the number of function evaluations. Top row: sphere function, middle row: GP trajectories, bottom row: electrical machine. Red: EGO using an ordinary Kriging model, blue: EGO using the combination of Kriging models.

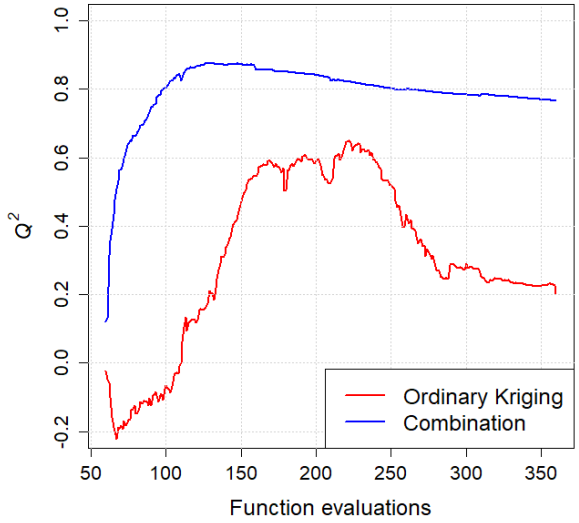


Figure 10: Q^2 of the ordinary Kriging model (red) and the combination of Kriging models (blue) during the optimization (sphere function with $d = 30$, Figure 9b).

6 Conclusion

In this paper, we have proposed a method to build a combination of Kriging sub-models for Bayesian optimization. We have presented a way to sample the sub-models suitably to avoid degenerate cases. A simple method based on LOOCV provides weights for the sub-models in closed-form, avoiding any costly inner optimization. Finally, we presented an approach to compute the prediction variance of the model, which is not available directly, by introducing a global covariance whose weights are computed analytically from the weights of the sub-models using an LOOCV strategy. An approach based on the empirical interquartile distance of the normalized LOO residuals, found to be more robust to outliers than the commonly employed empirical variance, was used to provide the amplitude hyperparameter of the variance.

The proposed method avoids the cumbersome optimization of the Kriging length-scales hyperparameters, which often fails to provide correct values in high-dimensional cases or when the number of training samples is limited as was demonstrated in a first numerical experiment. This is of interest for Bayesian optimization as such low samples regimes are common in design optimization, especially at the start of the optimization. The numerical results we obtained show that the combination of Kriging models enables both the construction of more accurate models and the faster convergence towards optimal solutions in Bayesian optimization, especially in the start when the number of samples is very limited. 2 to 3 times less expensive function evaluations are required with our method and in multimodal cases, better solutions never found by the original method are obtained. Contrarily to other high-dimensional BO methods, such as dimension reduction or additive Kriging, no additional hypothesis on the underlying function, such as a low dimension representation or an additive structure, is required. Thus, the proposed method is easily generalizable to any design engineering problem.

Several aspects still need to be explored in further research. First, in this paper the proposed method was compared with a vanilla EGO using an ordinary Kriging model. It would be interesting to compare it with other existing high-dimensional BO approaches (e.g. Eriksson et al., 2019; Diouane et al., 2023; Amine Bouhlel et al., 2018; Antonov et al.,

2022; Binois et al., 2020) on a more complete benchmark. Furthermore, we have only considered single objective optimization here. Extending the method to multi-objective and/or constrained problems is a perspective that would allow it to tackle a wider range of realistic problems such as the electrical machine design optimization.

Acknowledgments

This research was conducted with the support of the consortium in Applied Mathematics CIROQUO (`https://doi.org/10.5281/zenodo.6581217`), gathering partners in technological and academia in the development of advanced methods for Computer Experiments. This research was partly funded by a CIFRE grant (convention #2021/1284) established between the ANRT and Stellantis for the doctoral work of Tanguy Appriou.

References

Abdar, M., Pourpanah, F., Hussain, S., Rezazadegan, D., Liu, L., Ghavamzadeh, M., Fieguth, P., Cao, X., Khosravi, A., Acharya, U. R., et al. (2021). A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information fusion*, 76:243–297.

Abrahamsen, P. (1997). A review of gaussian random fields and correlation functions.

Acar, E. and Rais-Rohani, M. (2009). Ensemble of metamodels with optimized weight factors. *Structural and Multidisciplinary Optimization*, 37:279–294.

Acharki, N., Bertoncello, A., and Garnier, J. (2023). Robust prediction interval estimation for gaussian processes by cross-validation method. *Computational Statistics & Data Analysis*, 178:107597.

Ahmad, I. and Lin, P.-E. (1976). A nonparametric estimation of the entropy for absolutely continuous distributions (corresp.). *IEEE Transactions on Information Theory*, 22(3):372–375.

Amine Bouhlel, M., Bartoli, N., Regis, R. G., Otsmane, A., and Morlier, J. (2018). Efficient global optimization for high-dimensional constrained problems by using the kriging models combined with the partial least squares method. *Engineering Optimization*, 50(12):2038–2053.

Antonov, K., Raponi, E., Wang, H., and Doerr, C. (2022). High dimensional bayesian optimization with kernel principal component analysis. In *International Conference on Parallel Problem Solving from Nature*, pages 118–131. Springer.

Appriou, T., Rullière, D., and Gaudrie, D. (2023). Combination of optimization-free kriging models for high-dimensional problems. *Computational Statistics*, pages 1–23.

Bachoc, F. (2013). Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2020). Uniformly valid confidence intervals post-model-selection.

Beirlant, J., Dudewicz, E. J., Györfi, L., Van der Meulen, E. C., et al. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6(1):17–39.

Bellman, R. (1966). Dynamic programming. *Science*, 153(3731):34–37.

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013). Valid post-selection inference. *The Annals of Statistics*, pages 802–837.

Binois, M., Ginsbourger, D., and Roustant, O. (2020). On the choice of the low-dimensional domain for global optimization via random embeddings. *Journal of global optimization*, 76:69–90.

Binois, M. and Wycoff, N. (2022). A survey on high-dimensional gaussian process modeling with application to bayesian optimization. *ACM Transactions on Evolutionary Learning and Optimization*, 2(2):1–26.

Bouhlel, M. A., Bartoli, N., Otsmane, A., and Morlier, J. (2016). Improving kriging surrogates of high-dimensional design models by partial least squares dimension reduction. *Structural and Multidisciplinary Optimization*, 53(5):935–952.

Candelieri, A., Perego, R., and Archetti, F. (2018). Bayesian optimization of pump operations in water distribution systems. *Journal of Global Optimization*, 71:213–235.

Cao, Y. and Fleet, D. J. (2014). Generalized product of experts for automatic and principled fusion of gaussian process predictions. *arXiv preprint arXiv:1410.7827*.

Constantine, P. G. (2015). *Active subspaces: Emerging ideas for dimension reduction in parameter studies*. SIAM.

Cressie, N. (1993). *Statistics for spatial data*. John Wiley & Sons.

Deisenroth, M. and Ng, J. W. (2015). Distributed gaussian processes. In *International Conference on Machine Learning*, pages 1481–1490. PMLR.

Den Hertog, D., Kleijnen, J. P., and Siem, A. Y. (2006). The correct kriging variance estimated by bootstrapping. *Journal of the Operational Research Society*, 57(4):400–409.

Diouane, Y., Picheny, V., Riche, R. L., and Perrotolo, A. S. D. (2023). Trego: a trust-region framework for efficient global optimization. *Journal of Global Optimization*, 86(1):1–23.

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

Durrande, N., Ginsbourger, D., and Roustant, O. (2012). Additive Covariance kernels for high-dimensional Gaussian Process modeling. *Annales de la Faculté des sciences de Toulouse : Mathématiques*, Ser. 6, 21(3):481–499.

Eriksson, D., Pearce, M., Gardner, J., Turner, R. D., and Poloczek, M. (2019). Scalable global optimization via local bayesian optimization. *Advances in neural information processing systems*, 32.

Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering design via surrogate modelling: a practical guide*. John Wiley & Sons.

Forrester, A. I. and Keane, A. J. (2009). Recent advances in surrogate-based optimization. *Progress in aerospace sciences*, 45(1-3):50–79.

Garnett, R. (2023). *Bayesian optimization*. Cambridge University Press.

Gaudrie, D. (2019). *High-Dimensional Bayesian Multi-Objective Optimization*. PhD thesis, Ecole des Mines de Saint-Etienne.

Gaudrie, D., Le Riche, R., Picheny, V., Enaux, B., and Herbert, V. (2020). Modeling and optimization with gaussian processes in reduced eigenbases. *Structural and Multidisciplinary Optimization*, 61(6):2343–2361.

Ginsbourger, D., Dupuy, D., Badea, A., Carraro, L., and Roustant, O. (2009). A note on the choice and the estimation of kriging models for the analysis of deterministic computer experiments. *Applied Stochastic Models in Business and Industry*, 25(2):115–131.

Ginsbourger, D., Helbert, C., and Carraro, L. (2008). Discrete mixtures of kernels for kriging-based optimization. *Quality and Reliability Engineering International*, 24(6):681–691.

Ginsbourger, D. and Schärer, C. (2021). Fast calculation of gaussian process multiple-fold cross-validation residuals and their covariances. *arXiv preprint arXiv:2101.03108*.

Gu, M., Wang, X., and Berger, J. O. (2018). Robust gaussian stochastic process emulation. *The Annals of Statistics*, 46(6A):3038–3066.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492.

Karvonen, T. and Oates, C. J. (2023). Maximum likelihood estimation in gaussian process regression is ill-posed. *Journal of Machine Learning Research*, 24(120):1–47.

Karvonen, T., Wynne, G., Tronarp, F., Oates, C., and Sarkka, S. (2020). Maximum likelihood estimation and uncertainty quantification for gaussian process approximation of deterministic functions. *SIAM/ASA Journal on Uncertainty Quantification*, 8(3):926–958.

Kaufman, C. and Shaby, B. A. (2013). The role of the range parameter for estimation and prediction in geostatistics. *Biometrika*, 100(2):473–484.

Khosravi, A., Nahavandi, S., Creighton, D., and Atiya, A. F. (2010). Lower upper bound estimation method for construction of neural network-based prediction intervals. *IEEE transactions on neural networks*, 22(3):337–346.

Kleijnen, J. P. (2014). Simulation-optimization via kriging and bootstrapping: a survey. *Journal of Simulation*, 8:241–250.

Kleijnen, J. P., Van Beers, W., and Van Nieuwenhuyse, I. (2012). Expected improvement in efficient global optimization through bootstrapped kriging. *Journal of global optimization*, 54:59–73.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the witwatersrand. *Journal of the Southern African Institute of Mining and Metallurgy*, 52(6):119–139.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., and Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523):1094–1111.

Li, R. and Sudjianto, A. (2005). Analysis of computer experiments using penalized likelihood in gaussian kriging models. *Technometrics*, 47(2):111–120.

Matheron, G. (1963). Principles of geostatistics. *Economic geology*, 58(8):1246–1266.

Meliani, M., Bartoli, N., Lefebvre, T., Bouhlel, M.-A., Martins, J. R., and Morlier, J. (2019). Multi-fidelity efficient global optimization: Methodology and application to airfoil shape design. In *AIAA aviation 2019 forum*, page 3236.

Mohammed, R. O. and Cawley, G. C. (2017). Over-fitting in model selection with gaussian process regression. In *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*, pages 192–205. Springer.

Obrezanova, O., Csányi, G., Gola, J. M., and Segall, M. D. (2007). Gaussian processes: a method for automatic qsar modeling of adme properties. *Journal of chemical information and modeling*, 47(5):1847–1857.

Papadopoulos, G., Edwards, P. J., and Murray, A. F. (2001). Confidence estimation methods for neural networks: A practical comparison. *IEEE transactions on neural networks*, 12(6):1278–1287.

Pearce, T., Brintrup, A., Zaki, M., and Neely, A. (2018). High-quality prediction intervals for deep learning: A distribution-free, ensembled approach. In *International conference on machine learning*, pages 4075–4084. PMLR.

Picheny, V., Binois, M., and Habbal, A. (2019). A bayesian optimization approach to find nash equilibria. *Journal of Global Optimization*, 73:171–192.

Pronzato, L. and Rendas, M.-J. (2017). Bayesian local kriging. *Technometrics*, 59(3):293–304.

Rasmussen, C. E. and Williams, C. K. (2006). *Gaussian processes for machine learning*. MIT press Cambridge, MA.

Romano, Y., Patterson, E., and Candes, E. (2019). Conformalized quantile regression. *Advances in neural information processing systems*, 32.

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*, 51:1–55.

Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2018). Nested kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28:849–867.

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., Notz, W. I., and Williams, B. J. (2003). *The design and analysis of computer experiments*, volume 1. Springer.

Shan, S. and Wang, G. G. (2010). Survey of modeling and optimization strategies to solve high-dimensional design problems with computationally-expensive black-box functions. *Structural and multidisciplinary optimization*, 41(2):219–241.

Stein, M. L. (1999). *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media.

Van Der Vaart, A. and Van Zanten, H. (2011). Information rates of nonparametric gaussian process methods. *Journal of Machine Learning Research*, 12(6).

Viana, F. and Haftka, R. (2009). Importing uncertainty estimates from one surrogate to another. In *50th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference 17th AIAA/ASME/AHS Adaptive Structures Conference 11th AIAA No*, page 2237.

Viana, F. A., Haftka, R. T., and Steffen, V. (2009). Multiple surrogates: how cross-validation errors can help us to obtain the best predictor. *Structural and Multidisciplinary Optimization*, 39:439–457.

Viana, F. A., Haftka, R. T., and Watson, L. T. (2013). Efficient global optimization algorithm assisted by multiple surrogate techniques. *Journal of Global Optimization*, 56:669–689.

Yi, G., Shi, J., and Choi, T. (2011). Penalized gaussian process regression and classification for high-dimensional nonlinear data. *Biometrics*, 67(4):1285–1294.

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty years of mixture of experts. *IEEE transactions on neural networks and learning systems*, 23(8):1177–1193.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

A More details on the variance of the combination

A.1 Relation between mixture and combination

In this appendix, we will detail the proof of the statement in equation 23 that the MSE of a mixture model is equal to the MSE of the linear combination of the same models when the weights $\boldsymbol{\alpha}^2$ of the global covariance in equation (21) are equal to the weights $\boldsymbol{w}$ of the combination. First, for the combination of Kriging models, we have:

$$M_{tot}(\boldsymbol{x}) := \sum_{i=1}^{p} w_i M_i(\boldsymbol{x}), \quad \text{with } M_i(\boldsymbol{x}) = k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}_i}(\boldsymbol{X}, \boldsymbol{X})^{-1} Y, \quad i = 1, \ldots, p.$$

Here, without loss of generality, we suppose the models centered. Using the definition for the global GP in equation (21), we obtain the following global covariance:

$$\mathbf{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}')) = \sigma_{tot}^2 \sum_{i=1}^{p} \alpha_i^2 k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{x}'). \tag{33}$$

Then the MSE of the combination is:

$$\begin{aligned} \mathbf{E}\left[(M_{tot}(\boldsymbol{x}) - Y(\boldsymbol{x}))^2\right] &= \mathbf{Var}\left(Y(\boldsymbol{x})\right) + \mathbf{Var}\left(M_{tot}(\boldsymbol{x})\right) - 2\mathbf{Cov}\left(M_{tot}(\boldsymbol{x}), Y(\boldsymbol{x})\right) \\ &= \sigma_{tot}^2 \left(\sum_{i=1}^{p} \alpha_i^2 k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{w}^\top \boldsymbol{K}_M(\boldsymbol{x}) \boldsymbol{w} - 2\boldsymbol{w}^\top k_M(\boldsymbol{x}) \right), \end{aligned} \tag{34}$$

with:

$$(\boldsymbol{K}_M(\boldsymbol{x}))_{i,j} = k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}_i}(\boldsymbol{X}, \boldsymbol{X})^{-1} \sum_{\ell=1}^{p} \alpha_\ell^2 k_{\boldsymbol{\theta}_\ell}(\boldsymbol{X}, \boldsymbol{X}) k_{\boldsymbol{\theta}_j}(\boldsymbol{X}, \boldsymbol{X})^{-1} k_{\boldsymbol{\theta}_j}(\boldsymbol{X}, \boldsymbol{x}), \quad i, j = 1, \ldots, p,$$

$$(k_M(\boldsymbol{x}))_i = k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}_i}(\boldsymbol{X}, \boldsymbol{X})^{-1} \sum_{\ell=1}^{p} \alpha_\ell^2 k_{\boldsymbol{\theta}_\ell}(\boldsymbol{X}, \boldsymbol{x}),, \quad i = 1, \ldots, p.$$

For the mixture now, we have a mixture of posterior GPs:

$$Y_I(\boldsymbol{x}) | Y_I(\boldsymbol{X}), \quad \text{with density } \sum_{i=1}^{p} w_i p(Y_i(\boldsymbol{x}) | Y_i(\boldsymbol{X})),$$

where I is the indicator such that $P(I = i) = w_i, \quad i = 1, \ldots, p$. The corresponding mixture model is equal to the combination:

$$M_{mix}(\boldsymbol{x}) = \mathbf{E}\left(Y_I(\boldsymbol{x}) | Y_I(\boldsymbol{X})\right) = \sum_{i=1}^{p} w_i M_i(\boldsymbol{x}).$$

The MSE of the mixture model is obtained by:

$$\mathbf{E}\left[(M_{mix}(\boldsymbol{x}) - Y_I(\boldsymbol{x}))^2\right] = \mathbf{Var}(Y_I(\boldsymbol{x})) + \mathbf{Var}(M_{mix}(\boldsymbol{x})) - 2\mathbf{Cov}(M_{mix}(\boldsymbol{x}), Y_I(\boldsymbol{x})). \tag{35}$$

Using the lax of total variance:

$$\begin{aligned} \mathbf{Var}(Y_I(\boldsymbol{x})) &= \mathbf{E}(\mathbf{Var}(Y_I(\boldsymbol{x})|I)) + \mathbf{Var}(\mathbf{E}(Y_I(\boldsymbol{x})|I)) \\ &= \sigma_{tot}^2 \sum_{i=1}^{p} w_i k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{x}) + 0, \end{aligned}$$

and:

$$\begin{aligned}\mathbf{Cov}(M_{mix}(\boldsymbol{x}), Y_I(\boldsymbol{x})) &= \mathbf{E}(\mathbf{Cov}(M_{mix}(\boldsymbol{x}), Y_I(\boldsymbol{x})|I)) + \mathbf{Cov}(\mathbf{E}(M_{mix}(\boldsymbol{x})|I), \mathbf{E}(Y_I(\boldsymbol{x})|I)) \\ &= \sum_{i=1}^{p} w_i \mathbf{Cov}(M_{mix}(\boldsymbol{x}), Y_i(\boldsymbol{x})) + 0 \\ &= \sum_{i=1}^{p}\sum_{j=1}^{p} w_i w_j \mathbf{Cov}(M_i(\boldsymbol{x}), Y_j(\boldsymbol{x})).\end{aligned}$$

Re-injecting these two expressions into (35):

$$\mathbf{E}\left[(M_{mix}(\boldsymbol{x}) - Y_I(\boldsymbol{x}))^2\right] = \sigma_{tot}^2 \left(\sum_{i=1}^{p} w_i k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{x}) + \boldsymbol{w}^\top \boldsymbol{K}_M(\boldsymbol{x})\boldsymbol{w} - 2\boldsymbol{w}^\top k_M(\boldsymbol{x})\right), \quad (36)$$

with:

$$(\boldsymbol{K}_M(\boldsymbol{x}))_{i,j} = k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}_i}(\boldsymbol{X}, \boldsymbol{X})^{-1} \sum_{\ell=1}^{p} w_\ell k_{\boldsymbol{\theta}_\ell}(\boldsymbol{X}, \boldsymbol{X}) k_{\boldsymbol{\theta}_j}(\boldsymbol{X}, \boldsymbol{X})^{-1} k_{\boldsymbol{\theta}_j}(\boldsymbol{X}, \boldsymbol{x}), \quad i, j = 1, \ldots, p,$$

$$(k_M(\boldsymbol{x}))_i = k_{\boldsymbol{\theta}_i}(\boldsymbol{x}, \boldsymbol{X}) k_{\boldsymbol{\theta}_i}(\boldsymbol{X}, \boldsymbol{X})^{-1} \sum_{\ell=1}^{p} w_\ell k_{\boldsymbol{\theta}_\ell}(\boldsymbol{X}, \boldsymbol{x}),, \quad i = 1, \ldots, p.$$

We have equality of the MSE of the combination in (34) and of the mixture model in (36) if $\boldsymbol{\alpha}^2 = \boldsymbol{w}$, hence the result announced.

A.2 Coefficients for the variance of the combination

In this appendix, we give the proof for the analytical expression of the optimal weights α of the global covariance given in equations (25), (26) and (27).

As defined in equation (24), the optimal weights are obtained by minimizing the expected MSE of the combined model with respect to Y_{tot}. For a combination of two models, with $Y_{tot} = \alpha Y_1 + (1-\alpha)Y_2$, we recall the expression:

$$\alpha^* = \arg\min_{\alpha} \mathbf{E}\left[\mathbf{E}\left[(wM_1(\boldsymbol{x}) + (1-w)M_2(\boldsymbol{x}) - \alpha Y_1(\boldsymbol{x}) - (1-\alpha)Y_2(\boldsymbol{x}))^2 | Y_1, Y_2\right]\right].$$

Similarly to the derivation of the combination weights $\boldsymbol{w}$ in equations (17) and (18), the global MSE is approximated using the LOOCV error:

$$\alpha^* = \arg\min_{\alpha} \mathbf{E}\left[e_{LOOCV}(wM_1 + (1-w)M_2) | Y_1, Y_2\right], \quad (37)$$

where the LOOCV error is obtained using the Kriging LOO formula (19):

$$\begin{aligned}e_{LOOCV}(wM_1 + (1-w)M_2) &= \sum_{k=1}^{n} \left(wM_{1_{-k}}(\boldsymbol{x}) + (1-w)M_{2_{-k}}(\boldsymbol{x}) - \alpha Y_1(\boldsymbol{x}_k) - (1-\alpha)Y_2(\boldsymbol{x}_k)\right)^2 \\ &= \sum_{k=1}^{n} \left(W_k Y_{tot}(\boldsymbol{X})\right)^2,\end{aligned}$$

where:

$$W_k := w\frac{[\boldsymbol{K}_{\boldsymbol{\theta}_1}^{-1}]_{k,}}{[\boldsymbol{K}_{\boldsymbol{\theta}_1}^{-1}]_{k,k}} + (1-w)\frac{[\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{k,}}{[\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{k,k}}. \quad (38)$$

Let $\boldsymbol{K}_{tot} := \alpha^2 \boldsymbol{K}_{\boldsymbol{\theta}_1} + (1-\alpha)^2 \boldsymbol{K}_{\boldsymbol{\theta}_2}$ be the global correlation matrix. Then, $Y_{tot}(\boldsymbol{X}) \sim \mathcal{N}(0, \boldsymbol{K}_{tot})$, and $W_k Y_{tot}(\boldsymbol{X}) \sim \mathcal{N}(0, W_k \boldsymbol{K}_{tot} W_k^\top)$. Thus, the expected LOOCV error is:

$$\mathbf{E}\left[e_{LOOCV}(wM_1 + (1-w)M_2)\right] = \sum_{k=1}^{n} W_k \boldsymbol{K}_{tot} W_k^\top.$$

Taking the derivative with respect to α:

$$\frac{\partial \mathbf{E}\left[e_{LOOCV}(wM_1 + (1-w)M_2)\right]}{\partial \alpha} = 2\alpha \sum_{k=1}^{n} W_k (\boldsymbol{K}_{\boldsymbol{\theta}_1} + \boldsymbol{K}_{\boldsymbol{\theta}_2}) W_k^\top - 2 \sum_{k=1}^{n} W_k \boldsymbol{K}_{\boldsymbol{\theta}_2} W_k^\top. \quad (39)$$

Plugging back the expression for W_k (38) in (39):

$$W_k \boldsymbol{K}_{\boldsymbol{\theta}_1} W_k^\top = (1-w)^2 \frac{[\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{k,} \boldsymbol{K}_{\boldsymbol{\theta}_1} [\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{,k}}{[\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{k,k}^2} + (1 - (1-w)^2) \frac{1}{[\boldsymbol{K}_{\boldsymbol{\theta}_1}^{-1}]_{k,k}},$$

and:

$$W_k \boldsymbol{K}_{\boldsymbol{\theta}_2} W_k^\top = w^2 \frac{[\boldsymbol{K}_{\boldsymbol{\theta}_1}^{-1}]_{k,} \boldsymbol{K}_{\boldsymbol{\theta}_2} [\boldsymbol{K}_{\boldsymbol{\theta}_1}^{-1}]_{,k}}{[\boldsymbol{K}_{\boldsymbol{\theta}_1}^{-1}]_{k,k}^2} + (1 - w^2) \frac{1}{[\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{k,k}}.$$

We can also rewrite these results using the expected LOOCV error seeing that:

$$\mathbf{E}(e_{LOOCV}(M_i)|Y_j) = \begin{cases} \sum_{k=1}^{n} \frac{1}{[\boldsymbol{K}_{\boldsymbol{\theta}_2}^{-1}]_{k,k}}, & \text{if } i = j \\ \sum_{k=1}^{n} \frac{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{k,} \boldsymbol{K}_{\boldsymbol{\theta}_j} [\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{,k}}{[\boldsymbol{K}_{\boldsymbol{\theta}_i}^{-1}]_{k,k}^2}, & \text{if } i \neq j \end{cases}, \quad i, j \in \{1, 2\}. \quad (40)$$

Finally, by setting the partial derivative equal to 0 in (39), and using the expression in (40), we obtain the result given in equations (24) for the optimal weights α^* solution of (37).