Sequential design of Gaussian process surrogates using pre-posterior analysis and Bayesian model averaging

W. Fauriat

▶ **To cite this version:**

W. Fauriat. Sequential design of Gaussian process surrogates using pre-posterior analysis and Bayesian model averaging. ICASP 14 - 14th International Conference on Applications of Statistics and Probability in Civil Engineering, Jul 2023, Dublin, Ireland. cea-04170299

HAL Id: cea-04170299
https://cea.hal.science/cea-04170299

Submitted on 25 Jul 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

14th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP14
Dublin, Ireland, July 9-13, 2023

Sequential design of Gaussian process surrogates using pre-posterior analysis and Bayesian model averaging

W. Fauriat
Research Engineer, CEA, DAM, DIF, F-91297, Arpajon, France

ABSTRACT:
To build a surrogate model from experimental data – from tests or computer simulations – numerous options may exist when choosing a mathematical form. This is true for Gaussian Processes (GP) models, which may or may not include a regression basis – or mean trend – and be built on different correlation structures – through the selected covariance kernel. When data is scarce and prior information on the modeled phenomena is poor, it may be difficult to come to a conclusion on important decisions such as model selection or model improvement. If additional experimental information can be collected, often at significant cost, it is interesting to carry out model selection sequentially and efficiently. In this paper, we propose to leverage the ability of GPs to provide probabilistic descriptions and use it to look for the next "best" point in the design space using a pre-posterior analysis scheme, or Value Of Information (VoI) evaluation. At such point, we expect to get the most relevant information, when the aim is to reduce expected prediction error, given a previous state of knowledge on the likelihood of various modeling options, *e.g.* using the idea of Bayesian Model Averaging (BMA). With successive queried points, we update our respective beliefs in these options through an "information-optimal" exploration of the design space – given current expectations according to priors. Hence, we attempt to learn efficiently both model structure and parameters.

1. INTRODUCTION

In many branches of engineering, information on a quantity of interest is generally obtained either from testing or from simulation, for various experimental conditions or settings, *i.e.* across a particular design space. Whether the studied system is actually tested or rather simulated, relevant data may be costly to collect.

When Uncertainty Quantification (UQ) techniques are considered, *e.g.* for the study of the sensitivity to a particular design parameter, for robust or "chance-constrained" design and optimization, or for parameter identification and inverse problems, sampling on a large scale – much larger than the size of most Designs of Experiments (DoEs) – is generally required. In this context, the available experimental data should be used to build a cost-effective surrogate model of the physical system. The latter may then be used to perform heavy sampling.

So-called optimal design for computer experiments or surrogate model building has been a fruitful research topic, see *e.g* Sacks et al. (1989); Bates et al. (1996); Picheny et al. (2010) or Liu et al. (2018) for a review. When data points are queried sequentially in the design space, the approach is often referred to as "active learning", "adaptive learning" or "sequential design", see the works of *e.g* Osio and Amon (1996); Jones et al. (1998); Kleijnen and van Beers (2004).

Efficient collection of information during the sequential process constitutes the main objective of this paper. Here, it is pursued using Gaussian Process (GP) surrogate models assembled together with Bayesian Model Averaging (BMA) and sequentially enriched through the means of pre-

posterior analysis. In section 2, a goal-oriented overview is given about the key features of the proposed approach, namely GPs, BMA and pre-posterior analysis. In section 3 formulation details about the surrogate candidates are introduced. The proposed application of pre-posterior analysis for sequential design is presented in section 4. The methodology is illustrated on a simple academic example in section 5. Specific implementation choices are also discussed in this section.

2. SEQUENTIAL DESIGN USING GPS AND PRE-POSTERIOR ANLYSIS

GPs are an interesting choice for surrogate modeling in the context of UQ as they naturally embed a probabilistic description, thus facilitating the evaluation of the quality of model predictions – or equivalently, enabling a tailored control on model error – as well as allowing improvement of the model through additional training. So-called sequential design, often based on GP models, is involved in numerous algorithmic schemes for various UQ-related or robust optimization applications, see *e.g.* Santner et al. (2003); Jones et al. (1998). Additionally, the underlying correlation structure of GP models offers a particularly relevant framework for thinking in terms of extracting information from available samples – or future samples – in the design space. The choice of a particular correlation structure for a specific model is of major importance and a central question in this paper.

BMA consists in combining predictions from multiple models in order to provide a global answer that properly integrates all available knowledge on the different models' ability to represent the "true" system. When multiple modeling options are considered, especially when data is scarce – at early stages of the sequential design – BMA is an interesting device to account for all these options. As new data is collected, Bayes' rule can be used to update the respective beliefs – or weights – associated to the different options that were originally considered. The latter may then be either validated or discredited depending on new observations.

Pre-posterior analysis or Value of Information (VoI) evaluation is an approach that originated in the field of optimal decision theory and is deeply rooted into a Bayesian view on information collection and exploitation. In a nutshell, its objective is to quantify the interest of collecting a given piece of information by comparing, for optimal decisions, the outcomes that may be expected with and without this piece of information see *e.g* Raiffa and Schaifler (1961); Howard (1966). The underlying rationale is that a decision taken in accordance with a more "precise" state of knowledge can generally, though not systematically, be tailored more finely in order to obtain a desired outcome. In practice, a piece of information is worth collecting when the gain expected from it does not exceed the cost of collecting it. In the context of sequential design, VoI is an interesting tool to compare various alternatives for the gathering of new data in the design space.

In this paper, it is proposed to use BMA with GP surrogate models built on different correlation structures – different kernels and mean trends. Pre-posterior analysis is employed to carry out sequential design and select points in the design space that are expected to maximize error reduction, given the current state of knowledge in the form of a BMA description. With new data collected from the "true" process, prior weights are updated and model selection is progressively performed. In this context, information is particularly collected in order to help discriminating between competing model options that provide predictions with strong relative discrepancy. Hence, the goal is to learn the surrogate model's structure along with the sequential exploration of the design space, while optimizing the information collection process – *i.e.* limiting simulation or testing cost.

3. GAUSSIAN PROCESS MODELS AND BAYESIAN AVERAGING

A GP is a collection of random variables, each of them following a Gaussian distribution with mean:

$$\mu_Y(x^*) = g(x^*) + k(x^*, \mathbf{X})^T \mathbb{K}^{-1}(\mathbf{Y} - g(\mathbf{X})) \quad (1)$$

and variance:

$$\sigma_Y^2(x^*) = k(x^*, x^*) - k(x^*, \mathbf{X})^T \mathbb{K}^{-1} k(x^*, \mathbf{X}) \quad (2)$$

for any point x^* in the design space $\mathbb{R}^d$, where $\mu_Y \in \mathbb{R}$, $\sigma_Y \in \mathbb{R}$, $k(x,x')$ is a kernel function quantifying the covariance between two observations $y(x)$ and $y(x')$ and g is a given function, hereinafter denoted as trend – often a polynomial decomposition. $\mathbf{X}$ is a set of available training points, with their associated "observed values from the true system" stored in $\mathbf{Y}$. $\mathbb{K}$ is the covariance matrix computed from the dataset $\mathbf{X}$ using the kernel function k.

This mathematical structure, used as a supervised regression tool, offers a good balance between a globally defined behavior of the model – through the trend function g – and locally extracted information – through the "correlation strength" attributed by the kernel function k in the vicinity of observed design sites. The previous interpretation is only one among many views that may be adopted with respect to GPs, arguably a highly versatile tool, see *e.g.* Rasmussen and Williams (2006) for technical details. Regardless of interpretation, the GP model naturally provides predictions in the probabilistic form of a Gaussian random variable:

$$\Pr(Y(x^*) \leq y) = \Phi\left(y|\mu_Y(x^*|g,k), \sigma^2_Y(x^*|g,k)\right) \quad (3)$$

where Φ is the Gaussian CDF, non-bold capital notations represent random variables and g and k functions are stressed to emphasize their crucial influence on the predictions.

The choice of the trend and kernel functions, often specified through the use of hyper-parameters such that $g = g(\cdot|\beta)$ and $k = k(\cdot,\cdot|\theta)$, the latter generally determined through maximum likelihood or cross validation procedures, is a difficult question. Yet, its answer will define the quality of the surrogate model's predictions. From a model selection perspective the following remark, naive but quite illustrative, should be kept in mind. Models that favor a strong "fit" to the data rather than an ability to "generalize" will appear more "wiggly" rather than "smooth".

Here, a finite number – a pool – of model alternatives, using particular trend functions and constraints on the kernel function's hyper-parameters, will be considered. At early stages of the sequential design, when data is scarce and it is difficult to identify a preferred option, this pool of models will represent possible candidates, *e.g.* a model with no trend, a "wiggly" model with a trend, a "smooth" model with a trend, etc. BMA is then used to express the prediction at any given point $x^* \in \mathbb{R}^d$:

$$\Pr(Y(x^*) = y) = \sum_{j=1}^{q} \Pr(Y(x^*) = y|M_j) \cdot \Pr(M_j) \quad (4)$$

for q considered models M_j with prior beliefs $\Pr(M_j)$.

4. PRE-POSTERIOR ANALYSIS SCHEME AND BAYESIAN UPDATING

At any given stage of the sequential design, one holds a probabilistic description, through the GPs and $Y(x^*)$, of the "anticipated value" of the modeled system. This constitutes a prior state of knowledge for the pre-posterior analysis.

In general, this type of analysis consists in computing the difference between two quantities: on the one side, the expected outcome associated to the unconditional optimal decision and on the other side, the expected outcomes associated to the optimal decisions that are conditioned on specific pieces of information. The value attached to the act of collecting a particular piece of information z is computed as follows:

$$\text{VoI} = \min_a \mathbb{E}_Y[L(y,a)] - \mathbb{E}_Z\left[\min_a \mathbb{E}_{Y|Z}[L(y,a)]\right] \quad (5)$$

where $L(y,a)$ is a cost function, whose outcome depends on the selected alternative a and on a random variable Y. $\mathbb{E}_Y$ is the expectation with respect to Y. Z is a particular piece of information that influences the state of possible values of Y, namely from Y to $Y|Z$. If conditional optimal decisions tend to lead to lower expected cost, it is interesting – on average, over possibly collected z values – to wait and collect the piece of information and then decide after obtaining it, rather than deciding unconditionally without it. As the purpose is to evaluate the outcomes of decisions that might be taken with a collected piece of information – thus conditioned on a posterior state of knowledge – before actually

seeing such piece, this approach is often called pre-posterior analysis.

In this paper, pre-posterior analysis is applied for model identification, with the objective of reaching good overall prediction quality – other objectives could be pursued through different definitions of L. Hence, it is implemented in the following way. The cost function $L(y,a)$ is chosen as the squared error:

$$L(y,a) = |y-a|^2 \tag{6}$$

such that y represents the "true" value of the process, which could be observed, simulated or tested, and a is the prediction given by the surrogate. The objective of model identification is to select a in order to minimize the expected prediction error (EPE):

$$\min \text{EPE} = \min_a \mathbb{E}_Y[|y-a|^2] \tag{7}$$

or in a more detailed form:

$$\min \text{EPE}(x^*) = \min_a \int |y-a|^2 \Pr(Y(x^*)=y)dy \tag{8}$$

where the knowledge on the "true" value of the process is given here by the BMA prediction $\Pr(Y(x^*)=y)$, conditioned on the current priors $\Pr(M_j)$.

In the context of sequential design, information collection consists in querying at a new site e to obtain a new data point $z=y(e)$. Through their correlation structures, all the GP models taking part in the BMA prediction will be updated when this new point is integrated, thus yielding the conditional $\Pr(Y(x^*)=y|y(e)=z)$ for any desired point x^*. Then, VoI evaluation can be carried out using expressions (4), (5) and (6), in order to quantify and weight the potential for gain when querying at various alternative sites e of the design space:

$$\text{VoI}_e(x^*) = \min_a \int |y-a|^2 \Pr(Y(x^*)=y)dy - \int \left(\min_a \int |y-a|^2 \Pr(Y(x^*)=y|y(e)=z)dy \right) dz \tag{9}$$

where $VoI_e(x^*)$ represents the expected gain, in terms of prediction error reduction, at point x^*, when a new collection at site e is considered.

Let us note right here that for a squared error cost function and when $Y(x^*)$ is Gaussian, then the minimum EPE is equivalent to the so-called mean-square error (MSE) or variance $\sigma_Y^2(x^*)$ and it is reached by the mean of the Gaussian variable, *i.e.* $\text{arg}min_a \text{EPE}(x^*) = \mu_Y(x^*)$. In this situation (9) becomes:

$$\text{VoI}_e(x^*) = \sigma_Y^2(x^*) - \int \sigma_Y^2(x^*|y(e)=z)dz \tag{10}$$

Let us also remark that, for a single – stationary – GP process, $\sigma_Y^2(x^*|y(e)=z)$ only depends on the distance between x^* and e, but not actually on z, through the covariance kernel and (2). Consequently, the most interesting site from the perspective of error reduction is the one that is the "most distant" from available points in the original DoE, namely $\mathbf{X}$. This is a well-known feature in the use of GP for optimal and sequential designs, yet it is derived here from the general framework of pre-posterior analysis. The previous feature is no longer valid for a BMA combination of GPs, and the collected value z will actively come into play during the sequential design.

Identifying the most interesting site e, for the whole design and prediction space, with the help of (9), involves significant computation effort and will be discussed hereinafter. Yet, once this identification has been performed, the effective collection of the new data point $z=y(e)$ can be done and the BMA priors be updated using Bayes' rule:

$$\Pr(M_j|y(e)=z) = \frac{\Pr(Y(e)=z|M_j)\cdot\Pr(M_j)}{\sum_j \Pr(Y(e)=z|M_j)\cdot\Pr(M_j)} \tag{11}$$

where $\Pr(Y(e)=z|M_j)$ is computed before integrating z in the training set of model M_j, *i.e.* as if it was predicted by model M_j.

The aforementioned approach, as described by (9), is generic and constitutes a sound theoretical framework for "optimal" sequential design . Here, its application for a BMA combination of GPs is, as far as the author knows, an original proposal.

5. ILLUSTRATION ON A SIMPLE EXAMPLE AND TECHNICAL DETAILS

In this paper, the following practical implementation choices are made in order to render the computation of VoI affordable. At any given stage of the sequential design, the most interesting next candidate point e is identified using (9) and also:

- Integrating "over dy" and optimizing "over a" numerically – on a regular grid.
- Integrating "over dz" using samples of "reasonable size", where z is sampled from the prior state of knowledge, namely the BMA combination of GPs, *i.e.* $\Pr(Y(e) = z)$ computed using (4).
- Generating a random grid of N_c candidates in the design space. Here, this is done using Latin Hypercube Sampling (LHS), so as to explore evenly without making assumptions.
- Averaging the error reduction, potentially obtained from the collection of the data point $z = y(e_l)$, over the sampled LHS grid, rather than only at point x^* or over the complete design space, for obvious reasons in terms of computation effort, *i.e.*:

$$\mathrm{VoI}_{e_l} = (1/N_c) \sum_{k \in [1,N_c]} \mathrm{VoI}_{e_l}(x_k) \qquad (12)$$

  for any $l \in [1,N_c]$. Hence, one may speak of integrated error reduction, to be compared with IMSE in sequential design literature, see *e.g.* Sacks et al. (1989).
- Picking the candidate with highest expected value, *i.e.*:

$$e_{\text{next}} = \arg \max_{l \in [1,N_c]} VoI_{e_l} \qquad (13)$$

Once e_{next} is identified, $y(e_{\text{next}})$ is collected and BMA weights $\Pr(M_j)$ are updated using (11). At the end of this process, $y(e_{\text{next}})$ is added to the current DoE, *i.e.* $\{\mathbf{X}, \mathbf{Y}\} \leftarrow \{\mathbf{X}, \mathbf{Y}, e_{\text{next}}, y(e_{\text{next}})\}$. Then, at the following step of the sequential design, all the GPs' hyper-parameters will be updated using maximum likelihood estimation and this enriched dataset as a new DoE.

Such process can be repeated as long as it is profitable to do so. Error reduction potential will generally decrease during the process, up to a point when the cost of information collection becomes too high in comparison with the expected return.

The method is illustrated hereinafter on a very simple analytical function $f : x \to x \cdot \sin(x) + \varepsilon$, which represents the "true" system's behavior – known up to "precision" ε. The notion of "precision" or "uncertainty" can be used to account for imperfect knowledge, or lack of experimental repeatability, on either conditions or measurements or both – either x or y or both.

Here, three GP models are combined through BMA – red, green and magenta on figures. Constraints on hyper-parameters are imposed such that:

- M_1 – red – includes a second-order polynomial trend function g and has a rather smooth kernel
- M_2 – green – uses the same form g for the trend function but its kernel correlation length is constrained to be shorter
- M_3 – magenta – has no trend and a possibly short correlation length
- So-called nugget or pure-noise effects are included in the kernel definitions, thus GPs are not necessarily interpolating

The fit of the three GP models and the initial DoE is illustrated on Figure 1.

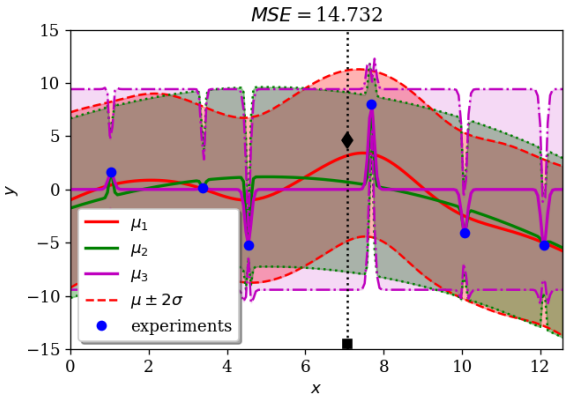

Figure 1: Initial DoE (6 points, step 0), fit of the three GP models, next site to be collected (black square) and obtained value (black diamond). Displayed MSE value is the average over the whole design space – as an indication of "global" precision of the surrogate.

The computation of VoI and the evolution of MSE or variance across the design space for the three GPs is illustrated on Figure 2. The next best

candidate site for information collection is identified on Figure 2 and the result of the collection, namely $z = y(e)$, is visible on Figure 1. Additional steps of the sequential design are displayed on figures 3 through 8. These figures demonstrates the proper implementation of the proposed methodology and its interest for "optimal" information collection – given initial hypotheses for BMA priors.

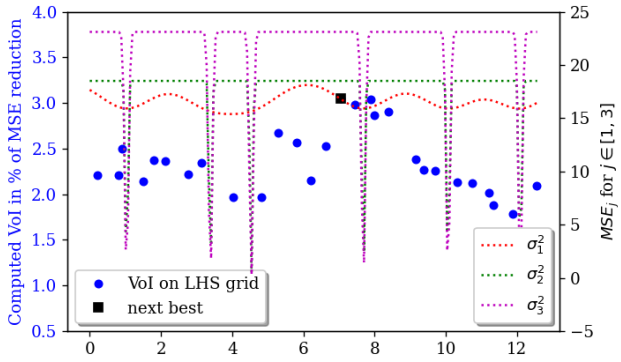

Figure 2: Computed VoI for step 0 and MSEs.

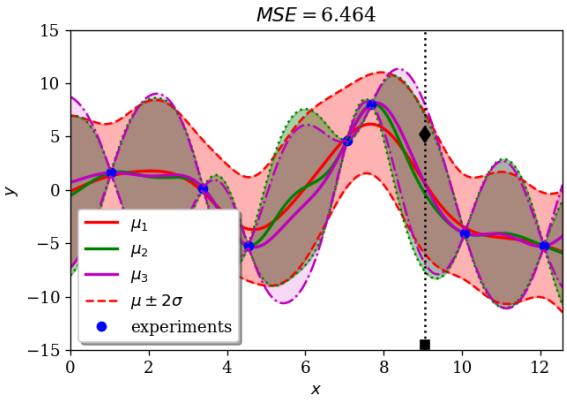

Figure 3: Step 1 of the sequential design (7 points) and next site and value (black square and diamond).

The result of the complete sequential design, stopped at step 15, is illustrated on Figure 9. One clearly sees MSE reduction along the process, not necessarily in a monotonous way, since pre-posterior analysis is based on "what is expected", but then the sequential design moves to the tune of "what is explored and encountered". Progressive model selection occurs as new points are collected

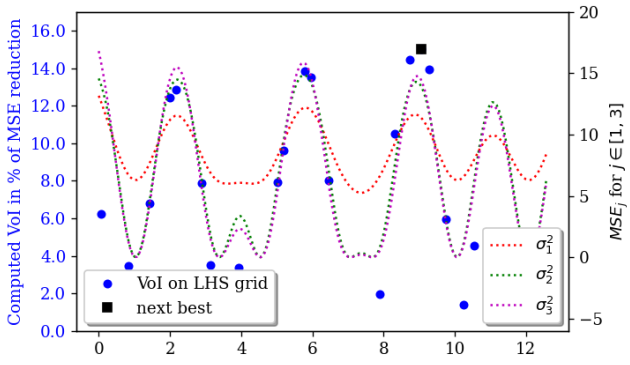

Figure 4: Computed VoI for step 1 and MSEs.

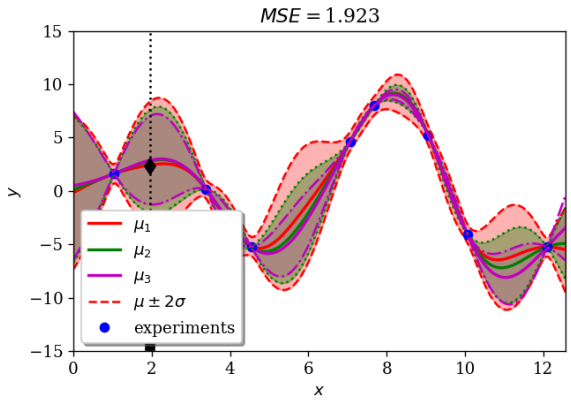

Figure 5: Step 2 of the sequential design (8 points) and next site and value (black square and diamond).

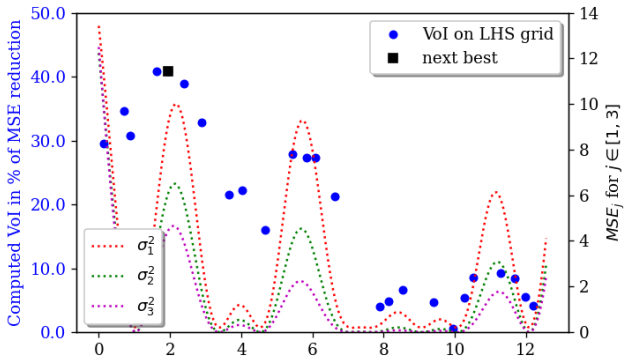

Figure 6: Computed VoI for step 2 and MSEs.

and BMA priors are updated through successive applications of Bayes' rule, as displayed on Figure 10.

One also sees on Figure 9 that the potential for gain decreases and becomes small after a few steps. Hence, VoI evaluation offers a way to control ex-

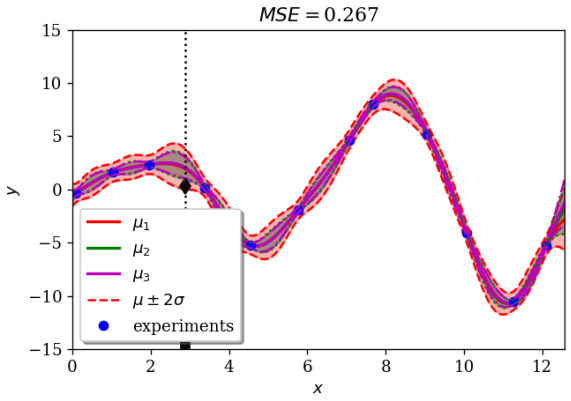

Figure 7: Step 6 of sequential design (12 points).

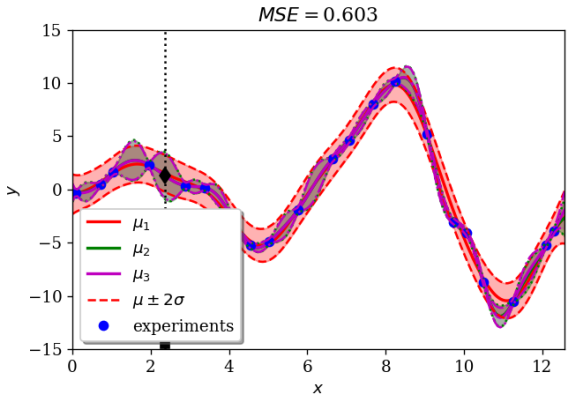

Figure 8: Step 15 of sequential design (21 points).

perimentation costs and benefits. Also, it may be observed on figures 1 through 6 that collected sites do not necessarily correspond to the location with highest variance – contrary to what often happens when trying to reduce expected error with GPs.

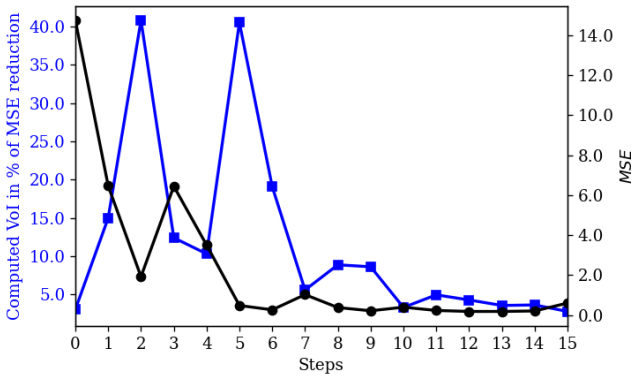

Figure 9: VoI and MSE during the sequential design

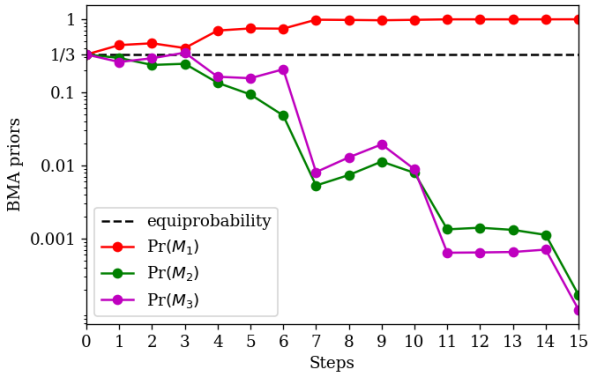

Figure 10: Evolution of BMA priors: progressive model selection

6. REMARKS

Successive application of Bayes' rule does not necessarily lead to an independent or absolute assessment of model quality, but rather to a comparative assessment, through the relative weights associated to the different considered options. When it comes to evaluating model quality, another criteria that appears a little bit more objective and arguably less framed in terms of alternative options, is the magnitude of Leave One Out (LOO) prediction errors – one of the possible choices for cross validation analyses. It is shown on Figure 11. Here, model M_1 is favored over M_2 and M_3, on Figure 10, due to significant noise in the "true" process – better described through M_1. Yet, it can be argued that prediction quality of M_2 and M_3, appears "acceptable" on Figure 11.

A very simple academic example has been proposed here for illustration purposes. Yet the approach can be exploited for more realistic engineering applications in higher dimension, or for other objectives than error reduction over the whole design space. In such cases, the cost and the technical difficulties to compute VoI will obviously be more significant. Nonetheless, if information collection is expensive for the considered application, the proposed approach may allow large savings when both VoI computation and additional experimentation costs are considered in relation to one another. Validating this remark on a costly application constitutes one perspective of the work proposed in this paper.

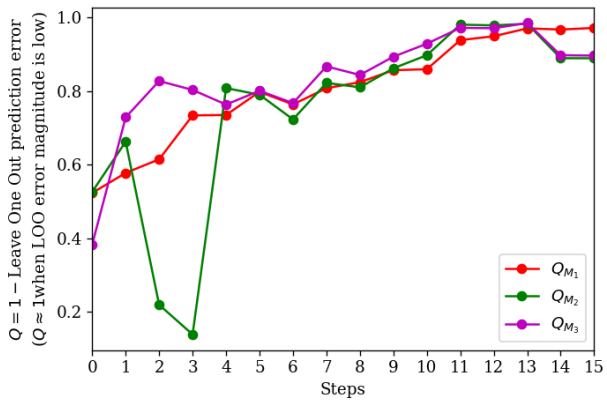

Figure 11: Model quality expressed using LOO prediction error magnitude. When $Q \approx 1$, LOO error magnitude is low and model – prediction – quality is high.

Also, specific implementation choices for the computation of VoI using (9), as detailed in the beginning of section 5, may seem quite crude in this paper. With BMA combinations of Gaussian PDF, more efficient analytical expressions can probably be derived, avoiding heavy numerical integration effort and possible boundary effects. This is another direction to explore.

7. CONCLUSIONS

Leveraging the ability of GPs to provide a full probabilistic representation, a sequential design approach based on pre-posterior analysis has been proposed in order to model – or learn the behavior of – any desired physical or engineering system from experiments, while seeking optimal collection of information, *i.e.* with controlled experimentation cost. It has been applied to a Bayesian Model Average of GPs, used as a means to carry out a progressive model selection, gathering knowledge as efficiently as possible – given priors.

A crucial perspective to keep in mind is that, with the framework of pre-posterior analysis, one may start with a relatively rough idea of where the "true" process lies and progressively "narrow" that state of knowledge by exploring "where one expects to find the most relevant information". Proceeding sequentially allows to integrate "what is encountered" and adjust, sometimes with some discrepancy compared to "what was originally expected". The underlying objective is to extract information from observations progressively and refrain from making too much assumptions or too early, especially when they are not supported by explicit reasons or evidence, and at the same time keep track of the confidence in model predictions.

8. REFERENCES

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

Howard, R. (1966). "Information value theory." *IEEE Transaction on System Science and Cybernetics*, 2, 22–26.

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

Kleijnen, J. and van Beers, W. (2004). "Application-driven sequential designs for simulation experiments: kriging metamodelling.." *Journal of the Operational Research Society*, 55(8), 876–883.

Liu, H., Cai, J., and Ong, Y. (2018). "A survey of adaptive sampling for global metamodeling in support of simulation-based complex engineering design." *Structural and Multidisciplinary Optimization*, 57(1), 393–416.

Osio, I. and Amon, C. (1996). "An engineering design methodology with multistage bayesian surrogates and optimal sampling." *Research in Engineering Design*, 8, 189–206.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R., and Nam-Ho, K. (2010). "Adaptative designs of experiments for accurate approximation of a target." *Journal of Mechanical Design*, 132.

Raiffa, H. and Schaifler, R. (1961). *Applied Statistical Decision Theory*.

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

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

Santner, T., Williams, B., and Notz, W. (2003). *The design and analysis of computer experiments*. Springer.