Nested Kriging models for large data-sets

Didier Rullière

joint work with Nicolas Durrande, François Bachoc and Clément Chevalier

Journées MAS 2016 – Grenoble

Introduction

Some notations :

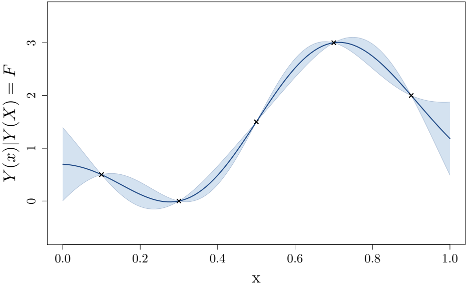
One prediction point (site, location) : $x \in D$ with $D \subset \mathbb{R}^d$
Unknown response at this point : $Y(x) \in \mathbb{R}$
n observations sites : $X \in D^n$
n observed responses : $Y(X) \in \mathbb{R}^n$

The conditional distribution of Gaussian Process Y having covariance function $k(.,.)$ is

$$m_{full}(x) = \mathrm{E}\left[Y(x)|Y(X)=F\right] = k(x,X)k(X,X)^{-1}F$$

$$c_{full}(x,x') = \mathrm{Cov}\left[Y(x),Y(x')|Y(X)=F\right] = k(x,x') - k(x,X)k(X,X)^{-1}k(X,x')$$

It can be represented as a mean function with confidence intervals.



If we denote by n the number of observation points, the complexity of building such models is

- $O(n^2)$ in space (storing $k(X, X)$)
- $O(n^3)$ in time (inverting $k(X, X)^{-1}$)

Furthermore, hyperparameter estimation requires to do this many times...

In practice,

- space complexity is often more limiting than time complexity
- the **maximum number of observations** that can be handled **lies in the range [1000, 10000]**.

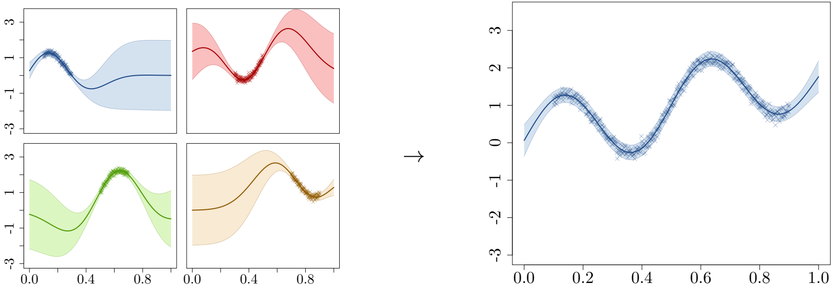
Various methods have been introduced to deal with a large number of observations :

- methods based on inducing points (sparse GPs)
- methods based on aggregating sub-models
- low rank approximations
- kernels with compact support
- ...

See Rasmussen and Williams, GPML, Chap. 8.

Aggregation of sub-models

In this talk, we focus on the aggregation of sub-models : make many sub-models based on subset of data, and then to find a way to merge these models together



Nested Kriging Models

Framework - Sub-models

Inputs :
One prediction point : $\mathbf{x} \in \mathbf{D}$.
Response random field : $\mathbf{Y}(\mathbf{x}) \in \mathbb{R}$.
Sub-models vector : $\mathbf{M}(\mathbf{x}) = (\mathbf{M_1}(\mathbf{x}), \dots, \mathbf{M_p}(\mathbf{x})) \in \mathbb{R}^p$.
– *Sub-models are typically functions of* **random** *vector $Y(X)$ at observation points X.*
– *We consider only* **one** *prediction point x here.*

Known covariances :
we assume that $(Y(x), M(x))$ is centred with $(1+p) \times (1+p)$ covariance matrix :

$$\mathrm{Cov}\left[(Y(x), M(x))\right] = \begin{pmatrix} k(x,x) & k_M(x)^t \\ k_M(x) & K_M(x) \end{pmatrix} \tag{1}$$

$k_M(x)$ is a $p \times 1$ vector with entries $k_M(x)_i = \mathrm{Cov}\left[Y(x), M_i(x)\right]$,
$K_M(x)$ is a $p \times p$ matrix with entries $(K_M(x))_{i,j} = \mathrm{Cov}\left[M_i(x), M_j(x)\right]$.

Quite general setting :

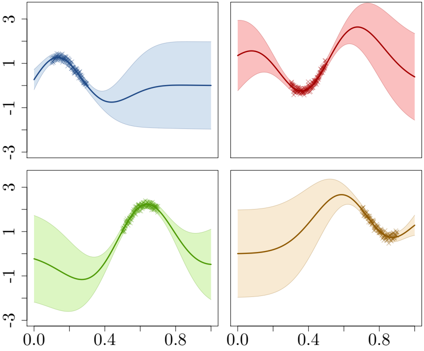
- We assume the existence of the first two moments of $(Y(x), M_1(x), \dots, M_p(x))$
- No other assumption on the joint distribution of $(Y(x), M_1(x), \dots, M_p(x))$
- $M(x)$ are covariates that are not necessarily a linear combinations of $Y(X)$

Framework - Case of Kriging submodels

Let $X_1, \ldots, X_p$ be matrices corresponding to subsets of observation points X.
Define p associated Kriging sub-models (or *experts*) :

$$\begin{cases} \mathbf{M_i(x)} & = & \mathrm{E}\left[Y(x)|Y(X_i)\right] = k(x,X_i)k(X_i,X_i)^{-1}Y(X_i) \\ \left(\mathbf{k_M(x)}\right)_\mathbf{i} & = & \mathrm{Cov}\left[Y(x),M_i(x)\right] = k(x,X_i)k(X_i,X_i)^{-1}k(X_i,x) \\ \left(\mathbf{K_M(x)}\right)_\mathbf{i,j} & = & \mathrm{Cov}\left[M_i(x),M_j(x)\right] = k(x,X_i)k(X_i,X_i)^{-1}k(X_i,X_j)k(X_j,X_j)^{-1}k(X_j,x). \end{cases}$$

Here, $M_i(x)$ are linear combinations of components of **random** vector $Y(X)$



Main questions

Classical kriging outputs (Gaussian case) :

pointwise :

- Kriging mean $\mathrm{E}\left[Y(x)|Y(X)\right]$
- Kriging variance $\mathrm{V}\left[Y(x)|Y(X)\right]$

cross-points :

- Kriging covariances $\mathrm{Cov}\left[Y(x), Y(x')|Y(X)\right]$
- Conditional sample paths

Corresponding questions when aggregating models :

pointwise :

- Aggregation $M_{1\oplus\ldots\oplus p}(x)$ of $M_1(x), \ldots, M_p(x)$, in order to estimate $Y(x)$?
- Variance $v_{1\oplus\ldots\oplus p}(x)$ of the error $M_{1\oplus\ldots\oplus p}(x) - Y(x)$?

cross-points :

- Covariances between $M_{1\oplus\ldots\oplus p}(x)$, $M_{1\oplus\ldots\oplus p}(x')$?
- Conditional sample paths (Gaussian case) ?

Proposed pointwise aggregation

Definition (sub-models aggregation)

For a given point $x \in D$, we define the aggregation of the sub-models (or mixture of experts) by

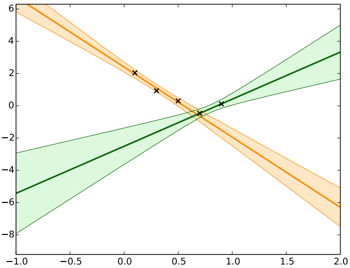
$$\mathbf{M}_{\mathbf{1}\oplus\ldots\oplus\mathbf{p}}(\mathbf{x}) = \mathbf{k}_{\mathbf{M}}(\mathbf{x})^{\mathbf{t}}\mathbf{K}_{\mathbf{M}}(\mathbf{x})^{-\mathbf{1}}\mathbf{M}(\mathbf{x}). \tag{2}$$

Basic properties for pointwise estimation

- **Optimal** : $M_{1\oplus\ldots\oplus p}(x)$ is the BLUE of $Y(x)$ that writes $\sum_i \alpha_i(x)M_i(x)$.
- **Square error** : $v_{1\oplus\ldots\oplus p}(x) = \mathrm{E}\left[\left(Y(x) - M_{1\oplus\ldots\oplus p}(x)\right)^2\right] = k(x,x) - k_M(x)^t K_M(x)^{-1} k_M(x)$
- **Conditional distribution** : If $(Y(x), M(x))$ is a Gaussian random vector, then the conditional distribution of $Y(x)$ given $M(x)$ is normal with moments

$$\mathrm{E}\left[Y(x)|M_1(x),\ldots,M_p(x)\right] = M_{1\oplus\ldots\oplus p}(x)$$
$$\mathrm{V}\left[Y(x)|M_1(x),\ldots,M_p(x)\right] = v_{1\oplus\ldots\oplus p}(x).$$

Example 1 - linear regressions



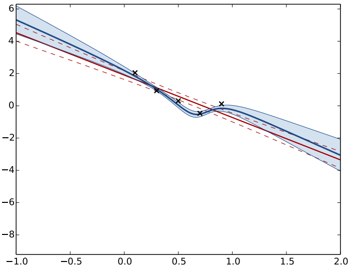


FIGURE : Example 1 : aggregation of two linear regression models. The left panel shows the sub-models and the right one the merged one in blue as well as the full model in red lines. Exhibited confidence bands corresponds to a difference to mean value of two standard deviation.

Example 2 - kriging submodels “3+2”

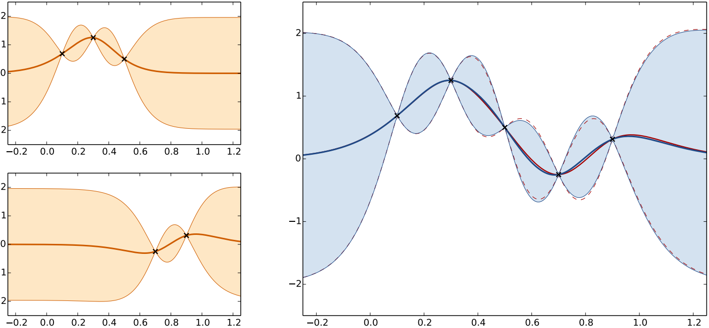


FIGURE : Example 2 : aggregation of two Gaussian process regression models. The left panel shows the sub-models and the right one the merged one in blue as well as the full model in red lines.

Example 3 - fully informative submodels

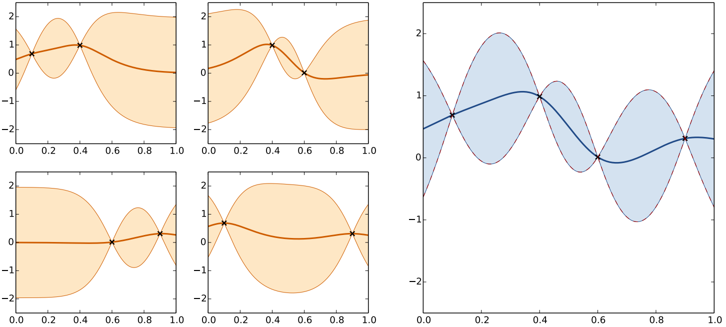


FIGURE : Example of merging sub-models without loss of information. The four submodels are shown on the left panels. As it can be seen on the right panel, the merged model (blue lines and shaded area) as well as the full model (red dashed lines) cannot be distinguished.

Aggregated process

We now focus on the case where (Y, M) is a centred Gaussian process with given covariances

$$\mathrm{Cov}\left[(Y(x), M(x)), (Y(x'), M(x'))\right] = \begin{pmatrix} k(x,x') & k_M(x,x')^t \\ k_M(x',x) & K_M(x,x') \end{pmatrix}. \qquad (3)$$

Definition (Aggregated process)

We define the process $Y_{1\oplus\ldots\oplus p}$ as

$$\mathbf{Y}_{\mathbf{1}\oplus\ldots\oplus\mathbf{p}} = \mathbf{M}_{\mathbf{1}\oplus\ldots\oplus\mathbf{p}} + \varepsilon'_{\mathbf{1}\oplus\ldots\oplus\mathbf{p}} \qquad (4)$$

where $\varepsilon'_{1\oplus\ldots\oplus p}$ is an independent replicate of $Y - M_{1\oplus\ldots\oplus p}$.

Some properties for aggregated process

- **known distribution** : $Y_{1\oplus\ldots\oplus p}$ is centred with known covariances

$$\begin{aligned} k_{1\oplus\ldots\oplus p}(x,x') = k(x,x') &+ 2k_M(x)^t k_M^{-1}(x) k_M^{-1}(x,x') k_M^{-1}(x') k_M(x') \\ &- k_M(x)^t k_M^{-1}(x) k_M(x',x) - k_M(x')^t k_M^{-1}(x') k_M(x,x'). \end{aligned} \tag{5}$$

- **optimality** : If $M_{1\oplus\ldots\oplus p}(x)$ writes $M_{1\oplus\ldots\oplus p}(x) = \lambda_{1\oplus\ldots\oplus p}(x)^t Y(X)$ and if $M_{1\oplus\ldots\oplus p}(X) = Y(X)$ then

$$\begin{aligned} M_{1\oplus\ldots\oplus p}(x) &= \mathrm{E}\left[Y_{1\oplus\ldots\oplus p}(x) \middle| Y_{1\oplus\ldots\oplus p}(X)\right] \\ v_{1\oplus\ldots\oplus p}(x) &= \mathrm{V}\left[Y_{1\oplus\ldots\oplus p}(x) \middle| Y_{1\oplus\ldots\oplus p}(X)\right]. \end{aligned}$$

- One can calculate $\mathrm{Cov}\left[Y_{1\oplus\ldots\oplus p}(x), Y_{1\oplus\ldots\oplus p}(x') \middle| Y_{1\oplus\ldots\oplus p}(X)\right]$.
- One can get conditional sample paths of $Y_{1\oplus\ldots\oplus p}$ given $Y_{1\oplus\ldots\oplus p}(X)$.
- On can get interpretations and bounds on the errors $M_{1\oplus\ldots\oplus p} - M_{full}$ and $v_{1\oplus\ldots\oplus p} - v_{full}$.

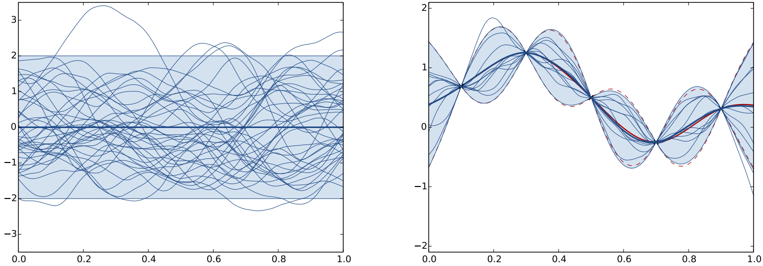
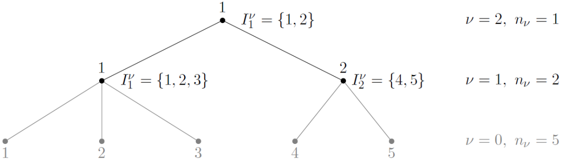


FIGURE : Interpretation of the results from Example 2 as a posterior Gaussian process distribution. The left panel shows the prior $Y_{1\oplus\ldots\oplus p}$ and the right one the conditional distribution given $Y_{1\oplus\ldots\oplus p}(X) = Y(X)$.

The approximated model is equivalent to an exact model based on a modified process.

Iterative model



Aggregation and covariance structure propagation at each level ν :

$$\text{From} \left\{ \begin{aligned} (M^\nu(x))_i &= M_i^\nu(x) \\ (k^\nu(x))_i &= \mathrm{Cov}\left[Y(x), M_i^\nu(x)\right] \\ (K^\nu(x))_{ij} &= \mathrm{Cov}\left[M_i^\nu(x), M_j^\nu(x)\right] \end{aligned} \right. \quad \text{get} \left\{ \begin{aligned} (M^{\nu+1}(x))_i &= \alpha_i^{\nu+1}(x)^t \left(M^\nu(x)_{[I_i^{\nu+1}]} \right) \\ (k^{\nu+1}(x))_i &= \alpha_i^{\nu+1}(x)^t \left(k^\nu(x)_{[I_i^{\nu+1}]} \right) \\ (K^{\nu+1}(x))_{ij} &= \alpha_i^{\nu+1}(x)^t \left(K^\nu_{[I_i^{\nu+1}, I_j^{\nu+1}]} \right) \alpha_j^{\nu+1}(x) \end{aligned} \right.$$

with vectors of optimal weights $\alpha_i^{\nu+1}(x) = \left(K^\nu_{[I_i^{\nu+1}, I_i^{\nu+1}]} \right)^{-1} \left(k^\nu(x)_{[I_i^{\nu+1}]} \right)$.

Algorithm 1: Iterative kriging algorithm

```
inputs  : $M_1$, vector of length $n_1$ (sub-models evaluated at $x$)
          $k_1$, vector of length $n_1$ (covariance between $Y(x)$ and sub-models at $x$)
          $K_1$, matrix of size $n_1 \times n_1$ (covariance between sub-models at x)
          $I$, a list describing the tree structure
outputs: $M_{\nu_{\max}}$, $K_{\nu_{\max}}$

for $\nu = 2, \ldots, \nu_{\max}$ do
    for $i = 1, \ldots, n_\nu$ do
        $M \leftarrow$ subvector of $M_{\nu-1}$ on $I_i^\nu$
        $K \leftarrow$ submatrix of $K_{\nu-1}$ on $I_i^\nu$
        if $\nu = 2$ then $k \leftarrow k_1$ else $k \leftarrow \mathrm{Diag}(K)$
        $\alpha_i \leftarrow K^{-1}k$
        $M_\nu[i] \leftarrow (\alpha_i)^t M$
        $K_\nu[i,i] \leftarrow (\alpha_i)^t k$
        for $j = 1, \ldots, i-1$ do
            $K \leftarrow$ submatrix of $K_{\nu-1}$ on $I_i^\nu \times I_j^\nu$
            $K_\nu[i,j] \leftarrow (\alpha_i)^t K \alpha_j$
            $K_\nu[j,i] \leftarrow K_\nu[i,j]$
```

Reachable storage footprint $\mathbf{O(n)}$. Reachable algorithm complexity $\mathbf{O(qn^2)}$.
Possibility of parallel computing.

n number of observations, q number of prediction points.

Consistency

Deisenroth and Ng 2015, Cao and Fleet 2014 and Van Stein et al 2015 propose aggregations based on sub-models variances $v_i(x)$:

$$\bar{M}_{1\oplus\ldots\oplus p_n}(x) = \sum_{k=1}^{p_n} \alpha_{k,n}(v_1(x), \ldots, v_{p_n}(x), v_{prior}(x)) M_k(x)$$

where a, b are positive deterministic continuous functions and

$$\alpha_{k,n}(v_1(x), \ldots, v_{p_n}(x), v_{prior}(x)) \leq \frac{a(v_k(x), v_{prior}(x))}{\sum_{l=1}^{p_n} b(v_l(x), v_{prior}(x))},$$

Proposition (Non-Consistency of variance-based methods)

Let the observation domain $\mathcal{X}$ be fixed and bounded, let $x_0 \in \mathcal{X}$ be fixed and let $N, p \to \infty$. For a standard class of covariance functions, with the aggregation methods above, there exists a dense triangular array of observation points so that

$$\liminf_{N,p\to\infty} \mathbb{E}\left(\{Y(x_0) - M_{1\oplus\ldots\oplus p}(x_0)\}^2\right) > 0$$

Proposition (Consistency)

$\Longrightarrow$ On the contrary, our proposed aggregation method yields a consistent predictor.

$$\liminf_{N,p\to\infty} \mathbb{E}\left(\{Y(x_0) - M_{1\oplus\ldots\oplus p}(x_0)\}^2\right) = 0$$

Parameter estimation

Parametric covariance model

Set of covariance functions

$$\{\sigma^2 k_\theta, \sigma^2 \geq 0, \theta \in \Theta\}$$

with $\Theta \subset \mathbb{R}^m$
Yields predictors and predictive variances

$$m_{1\oplus\ldots\oplus p,\theta}(x)$$

and

$$v_{1\oplus\ldots\oplus p,\sigma^2,\theta}(x)$$

Goal : $\hat{\theta}$ and $\hat{\sigma}^2$

Stochastic gradient for $\hat{\theta}$

Let $M_{1\oplus\ldots\oplus p,\theta,-i}(x_i)$ be the Leave One Out prediction of y_i based on the $n-1$ remaining points

We want to use the Leave One Out estimator

$$\hat{\theta} \in \operatorname*{argmin}_{\theta\in\Theta} \frac{1}{n} \sum_{i=1}^{n} \left\{ M_{1\oplus\ldots\oplus p,\theta,-i}(x_i) - y_i \right\}^2$$

Computing q Leave One Out errors costs $O(qn^2)$ flops $\Longrightarrow$ stochastic gradient :

$$\theta_{k+1} = \theta_k - a_k h \left\{ \frac{1}{\epsilon_k} \left(\frac{1}{q} \sum_{i\in I} \left(M_{1\oplus\ldots\oplus p,\theta+\epsilon_k h,-i}(x_i) - y_i \right)^2 - \frac{1}{q} \sum_{i\in I} \left(M_{1\oplus\ldots\oplus p,\theta-\epsilon_k h,-i}(x_i) - y_i \right)^2 \right) \right\}$$

where I is a random sample of size q and h is a random direction

$\Longrightarrow$ Stochastic gradient is not worth it for the exact Gaussian process prediction ($O(n^3)$ cost for q error computations) but is useful with our aggregation method

Estimation of σ^2

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i - m_{1\oplus\ldots\oplus p,-i,\hat{\theta}}(x_i)\right)^2}{v_{1\oplus\ldots\oplus p,-i,1,\hat{\theta}}(x_i),}$$

which is equivalent to

$$\frac{1}{n} \sum_{i=1}^{n} \frac{\left(y_i - m_{1\oplus\ldots\oplus p,-i,\hat{\theta}}(x_i)\right)^2}{v_{1\oplus\ldots\oplus p,-i,\hat{\sigma}^2,\hat{\theta}}(x_i)} = 1.$$

Numerical illustrations

Some alternatives methods

For a given x, let $f_{M_i}(y)$ be the predictive density of model i and $f_M(y)$ denote the aggregated prediction :

Various methods have been proposed in the literature :

- Product of Experts (PoE)

$$f_M(y) \propto \prod f_{M_l}(y)$$

- Generalised PoE

$$f_M(y) \propto \prod f_{M_l}^{\beta_i}(y)$$

- Bayesian Committee Machine (BCM)

$$f_M(y) \propto \frac{\prod f_{M_l}(y)}{f_Y^{(p-1)}(y)}$$

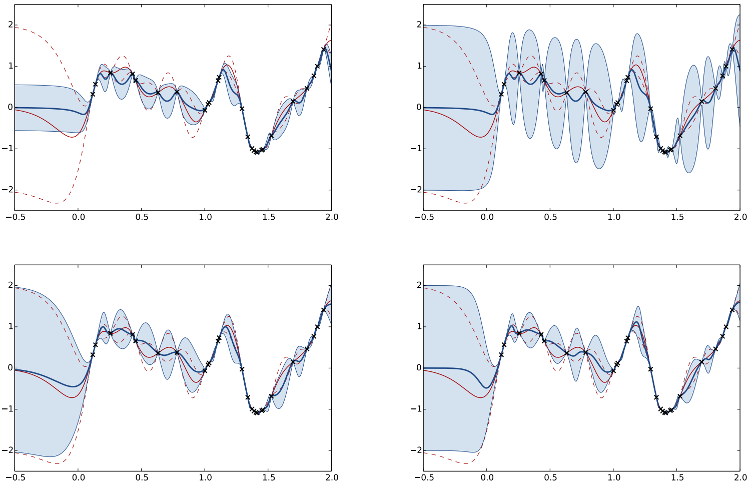
- Robust BCM

$$f_M(y) \propto \frac{\prod f_{M_l}^{\beta_i}(y)}{f_Y^{\left(\sum \beta_i - 1\right)}(y)}$$

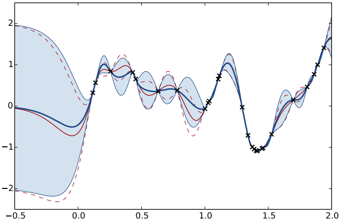
- Smallest prediction variance (SPV) :

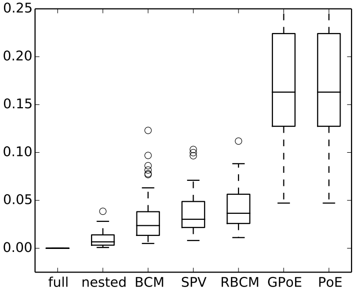
$$f_{spv}(y) = f_k(y) \qquad \text{with } k = \underset{i \in \{1,\dots,p\}}{\operatorname{argmin}} \; v_i(x).$$

Classical methods PoE, GPoE, BCM, RBCM (13 submodels based on two points each, n=26)



Proposed nested method - Average distance to full model boxplot.





Industrial case study

- Data provided by EDF (Geraud Blatman)
- $10,000$ input-outputs $(x_i, f(x_i))$, dimension $d = dim(x_i) = 6$

$$f(x_i) = \log \left[\sum_{j=1}^{m} (F(x_i, c_j) - m_j)^2 \right]$$

c_j : experimental condition, F : code, x_i : code parameter, m_j experimental value

- $n = 9000$ data points in the learning base
- $n_t = 1000$ data points in the test base
- One aggregation by our method or the "sum-based" aggregation methods
- $p = 20$ or $p = 90$ aggregated subsamples
- Subsamples chosen with K-means or randomly
- Covariance functions : exponential, Matérn 3/2, Matérn 5/2. Ordinary Kriging
- Covariance parameters chosen by our proposed stochastic gradient method or by minimizing the sum of the likelihoods over the subsamples

Prediction criteria

- MSE (should be minimal)

$$MSE = \frac{1}{n_t}\sum_{i=1}^{n_t}(m_{1\oplus\ldots\oplus p,\hat{\theta}}(x_{t,i}) - f(x_{t,i}))^2,$$

- MNLP (should be minimal)

$$= \frac{1}{n_t}\sum_{i=1}^{n_t}\left(\frac{1}{2}\log(2\pi v_{1\oplus\ldots\oplus p,\hat{\sigma}^2,\hat{\theta}}(x_{t,i})) + \frac{(m_{1\oplus\ldots\oplus p,\hat{\theta}}(x_{t,i}) - f(x_{t,i}))^2}{2v_{1\oplus\ldots\oplus p,\hat{\sigma}^2,\hat{\theta}}(x_{t,i})}\right),$$

(other criteria in the article)

Prediction results (a)

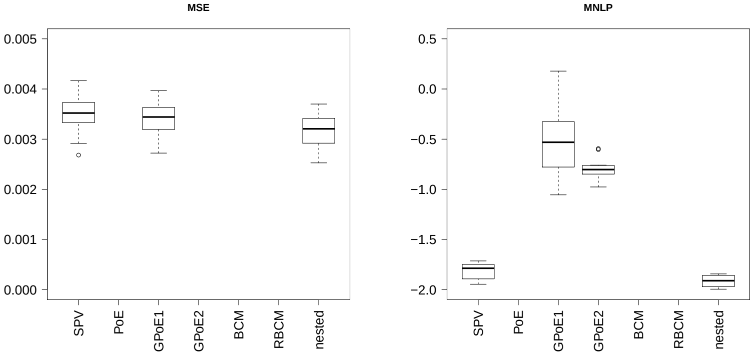


FIGURE : Box plots of 20 values of the mean square error (MSE) prediction criterion and of the logarithm of the mean negative log probability (MNLP) prediction criterion where the learning and test sets are randomly generated. ($p =$ **20 subsamples** obtained from the *k*-**means** algorithm ; Matérn 5/2 covariance function). Not represented if too large MSE or MLNP values. Covariance parameters estimated by log lik for SPV, PoE, gPoE1, gPoE2, BCM and rBCM and by LOO for our aggregation procedure.

Prediction results (b)

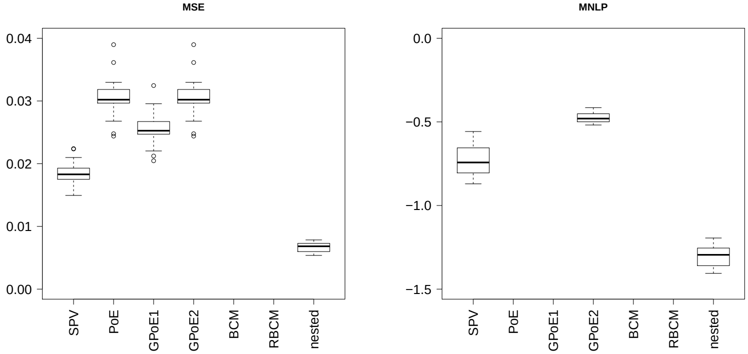


FIGURE : Same settings as in Figure 5 but with $p = \mathbf{90}$ **subsamples**, **randomly selected**. Too large values MSE or MLNP are not represented.

Conclusion

On proposed nested Kriging model : optimal linear weighting of submodels (or exact method on a modified process)

- proven to provide consistent predictors (some other classical aggregation techniques are shown inconsistent)
- bounds on errors compared to the full model (not presented here)
- dedicated covariance parameter estimation procedure
- encouraging numerical results but with increased computational cost (compared to other methods)

Perspectives

- The stochastic gradient algorithm could be further investigated
- Further reduction of the complexity (tree structure and approximations)

Thank you for your attention !

preprint available on Hal.

Nested Kriging estimations for datasets with large number of observations. 2016. <hal-01345959>