Properties and comparison of some Kriging sub-model aggregation methods

François Bachoc, Nicolas Durrande, Didier Rullière, Clément Chevalier

Properties and comparison of some Kriging sub-model aggregation methods*

François Bachoc Nicolas Durrande Didier Rullière Clément Chevalier

François Bachoc (✉), Institut de Mathématiques de Toulouse, Université Paul Sabatier, Toulouse, France
francois.bachoc@math.univ-toulouse.fr

Nicolas Durrande, Secondmind, Cambridge, UK
nicolas@secondmind.ai

Didier Rullière, Mines Saint-Etienne, UCA, CNRS, UMR 6158 LIMOS, Institut Henri Fayol, Saint-Etienne, France
didier.rulliere@emse.fr

Clément Chevalier, Institute of Statistics, University of Neuchâtel, Neuchâtel, Switzerland.
clement.chevalier@unine.ch

Abstract

Kriging is a widely employed technique, in particular for computer experiments, in machine learning or in geostatistics. An important challenge for Kriging is the computational burden when the data set is large. This article focuses on a class of methods aiming at decreasing this computational cost, consisting in aggregating Kriging predictors based on smaller data subsets. It proves that aggregation methods that ignore the covariance between sub-models can yield an inconsistent final Kriging prediction. In contrast, a theoretical study of the nested Kriging method shows additional attractive properties for it: First, this predictor is consistent, second it can be interpreted as an exact conditional distribution for a modified process and third, the conditional covariances given the observations can be computed efficiently. This article also includes a theoretical and numerical analysis of how the assignment of the observation points to the sub-models can affect the prediction ability of the aggregated model. Finally, the nested Kriging method is extended to measurement errors and to universal Kriging.

Keywords: Gaussian processes, model aggregation, consistency, error bounds, Nested Pointwise Aggregation of Experts, NPAE.

1 Introduction

Kriging (Krige (1951), Matheron (1970), see also (Cressie 1993, Stein 2012, Santner et al. 2013)) consists in inferring the values of a Gaussian random field given observations at a finite set of observation points. It has become a popular method for a large range of applications, such as geostatistics (Matheron 1970), numerical code approximation (Sacks et al. 1989, Santner et al. 2013, Bachoc et al. 2016), global optimization (Jones et al. 1998) or machine learning (Rasmussen and Williams 2006).

Let Y be a centered Gaussian process on $D \subset \mathbb{R}^d$, with covariance function $k : D \times D \to \mathbb{R}$ (ie $k(x, x') = \mathrm{Cov}\left[Y(x), Y(x')\right]$), and let $x_1, ..., x_n \in D$ be n points in the input space where Y is observed (exactly). The assumptions that Y is zero-mean (simple Kriging, corresponding to a known mean) and that it is observed without observation noise are common in the literature and they will be used in throughout the paper for conciseness. These two assumptions will however be relaxed in Sects. 5.1 and 5.2 to ensure the method under study can be applied in more practical case studies.

Let X be the $n \times d$ matrix with row i equal to ${x_i}^t$. For any functions $f : D \to \mathbb{R}$, $g : D \times D \to \mathbb{R}$ and for any matrices $A = (a_1, \ldots, a_n)^t$ and $B = (b_1, \ldots, b_m)^t$, with $a_i \in D$ for $i = 1, ..., n$ and $b_i \in D$ for $i = 1, ..., m$, $f(A)$ denotes the $n \times 1$ real valued vector with components $f(a_i)$ and $g(A, B)$ denotes the $n \times m$ real valued matrix with

*Part of this research was conducted within the frame of the Chair in Applied Mathematics OQUAIDO, which gathers partners in technological research (BRGM, CEA, IFPEN, IRSN, Safran, Storengy) and academia (Ecole Centrale de Lyon, Mines Saint-Étienne, University of Nice, University of Toulouse and CNRS) around advanced methods for Computer Experiments. The authors F.Bachoc and D.Rullière acknowledge support from the regional MATH-AmSud program, grant number 20-MATH-03. The authors are grateful to the Editor-in-Chief, an Associate Editor and a referee for their constructive suggestions that lead to an improvement of the manuscript.

components $g(a_i, b_j)$, $i = 1, \ldots, n$, $j = 1, \ldots, m$. With this notation, the conditional distribution of Y given the $n \times 1$ vector of observations $Y(X)$ is Gaussian with mean, covariance and variance:

$$\begin{cases} M_{full}(x) = \mathrm{E}\left[Y(x)|Y(X)\right] = k(x,X)k(X,X)^{-1}Y(X)\,, \\ c_{full}(x,x') = \mathrm{Cov}\left[Y(x), Y(x')|Y(X)\right] = k(x,x') - k(x,X)k(X,X)^{-1}k(X,x')\,, \\ v_{full}(x) = c_{full}(x,x)\,. \end{cases} \tag{1}$$

Computing the terms on the right hand side of (1) requires to invert the $n \times n$ covariance matrix $k(X,X)$, which leads to a $O(n^2)$ storage requirement and $O(n^3)$ computational complexity. In practice, this posterior distribution is hence difficult to compute when the number of observation points exceeds a few thousands. The challenge of a large number of observation points for Kriging is for instance acknowledged in Section 5 of Davis and Curriero (2019).

Many methods have been proposed in the literature to approximate the conditional distribution (1), without incurring a large computational cost. These methods include low rank approximations (see Stein 2014, and the references therein for a review), sparse methods (Hensman and Fusi 2013), covariance tapering (Furrer et al. 2006, Kaufman et al. 2008), Gaussian Markov Random Fields (Rue and Held 2005, Datta et al. 2016), and aggregation-based approximations (Deisenroth and Ng 2015). This paper focuses on the later approach, which consists of building sub-models based on subsets of the data before aggregating their predictions. More precisely, these methods first construct p sub-models $M_1, ..., M_p : D \to \mathbb{R}$, where $M_i(x)$ is a predictor of $Y(x)$ built from a subset X_i of size $n_i \times d$ of the observation points in X. The rationale is that when n_i is small compared to n, $M_i(x)$ can be obtained efficiently with a small computational cost. The sub-models $M_1, ..., M_p$ are then combined to obtain the aggregated predictor $M_{\mathcal{A}} : D \to \mathbb{R}$. Examples of aggregation techniques for Gaussian processes are (generalized) products of experts and (robust) Bayesian committee machines (Hinton 2002, Tresp 2000, Cao and Fleet 2014, Deisenroth and Ng 2015, van Stein et al. 2015), as well as the nested Kriging predictor (or Nested Pointwise Aggregation of Experts (NPAE)) (Rullière et al. 2018). It must be noted that nested Kriging relies on a particular aggregation of several predictors; a review of probability aggregation methods in Geoscience can be found in Allard et al. (2012).

Aggregation methods can be of particular interest in a geoscience context. As it is well known, the origins of Kriging are directly linked to mining and geostatistics (Cressie 1990, Chilès and Desassis 2018), and it is common to encounter large datasets in such context. As an example, imagine a measurement (say, radioactivity) at several locations on the ground. The measures can be done with simple movable devices at many locations, eventually repeated at several times, so that the number of measurements can be important, and each measure may come with a measurement error. There is a necessity to handle a large amount of potentially noisy measures, which is not possible with classical Kriging techniques, but becomes possible with aggregation techniques such as nested Kriging. Despite its novelty, the nested Kriging predictor has already been used in several application fields, including earth and geostatistical sciences: see Sun et al. (2019) for the study of air pollution, Bacchi et al. (2020) for tsunami analysis and Krityakierne and Baowan (2020) for contaminant source localization in the ground. This also emphasizes the importance of aggregation methods in a geostatistical context.

Benchmarks of different spatial interpolation methods are also of importance, in particular when dealing with big data. Among recent ones, a general benchmark on some other methods applicable with big data can be found in Heaton et al. (2019). It is worth noting that nested Kriging often appears among the two or three best competitors (typically among around 12 methods) in the numerical studies that include it into the benchmark (Rullière et al. 2018, Liu et al. 2018, He et al. 2019, Liu et al. 2020, Van Stein et al. 2020). These good empirical performances are supported by the theoretical properties of nested Kriging which guaranty some optimal performances under a correct estimation of the underlying hyperparameters and under stationarity.

This paper provides additional theoretical insights into aggregation methods, with an emphasis on nested Kriging (Rullière et al. 2018). A distinction is introduced between aggregation techniques that only rely on the conditional variances $v_1(x), ..., v_p(x)$ of the sub-models (such as products of expert and Bayesian committee machines), and the ones, like nested Kriging, where the aggregation accounts for the covariance between the sub-models. As shown in Proposition 1, techniques based only on the sub-model variances can lead to inconsistent estimators of $Y(x)$ in the infill (fixed-domain) asymptotic setting (Cressie 1993, Stein 2012). On the other hand, Proposition 2 guaranties the consistency, again in the infill (fixed-domain) asymptotic setting, of the nested Kriging predictor. In addition, the nested Kriging predictor can be interpreted as an exact conditional expectation, for a slightly different Gaussian process prior.

Furthermore, the paper introduces two extensions of the nested Kriging methodology which broaden the use cases where the approach can be applied. The first one is to make nested Kriging amenable to observation noise corrupting the measurements (the initial exposition in Rullière et al. (2018) focused on the noiseless setting). The second is to generalise the method to universal Kriging, where the Gaussian process prior includes an unknown

mean function that must be estimated. Note that both generalisations result in similar storage or computational requirements as the original approach.

The structure of the article is as follows. Section 2 introduces covariance-free aggregation techniques and present the non-consistency result. Section 3 summarizes the aggregation method of Rullière et al. (2018), gives its consistency property, shows how it can be interpreted as an exact conditional expectation and provides some error bounds for the nested Kriging approximation. It also provides a numerical illustration of the consistency and inconsistency properties shown in this paper. Section 4 studies the impact of the assignment of the observation points to the sub-models. Finally, Sect. 5 provides the extensions to measurement errors and universal Kriging and concluding remarks are given in Sect. 6. For the sake of the clarity of the exposition, most of the proofs are postponed to the appendix.

2 Covariance-free aggregation techniques

For $i = 1, ..., p$, let X_i be a $n_i \times d$ matrix composed of a subset of the lines of X, such that $n_1 + \cdots + n_p = n$ and $X_1, ..., X_p$ constitute a partition of X. For $i = 1, ..., p$, let $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1}Y(X_i)$ and $v_i(x) = k(x, x) - k(x, X_i)k(X_i, X_i)^{-1}k(X_i, x)$ be the conditional mean and variance of $Y(x)$ given $Y(X_i)$. This section focuses on aggregated predictors that only depend on (predicted) variances

$$M_{\mathcal{A}}(x) = \sum_{k=1}^{p} \alpha_k(v_1(x), ..., v_p(x), v_{prior}(x))M_k(x), \tag{2}$$

where $v_{prior}(x) = k(x, x)$ and with $\alpha_k : [0, \infty)^{p+1} \to \mathbb{R}$. Several aggregation techniques, such as product of expert (POE), generalized product of expert (GPOE), Bayesian committee machines (BCM) and robust Bayesian committee machines (RBCM), can be written under the form of (2). For POE (Hinton 2002, Deisenroth and Ng 2015) and GPOE (Cao and Fleet 2014) the weights associated to each sub-model are

$$\alpha_k(v_1, ..., v_p, v_{prior}) = \frac{\beta_k(x)\frac{1}{v_k}}{\sum_{i=1}^{p} \beta_i(x)\frac{1}{v_i}}$$

with $\beta_i(x) = 1$ for POE and $\beta_i(x) = (1/2)[\log(v_{prior}(x)) - \log(v_i(x))]$ for GPOE. For BCM (Tresp 2000) and RBCM (Deisenroth and Ng 2015) they are

$$\alpha_k(v_1, ..., v_p, v_{prior}) = \frac{\beta_k(x)\frac{1}{v_k}}{\sum_{i=1}^{p} \beta_i(x)\frac{1}{v_i} + (1 - \sum_{i=1}^{p} \beta_i(x))\frac{1}{v_{prior}}}$$

with $\beta_i(x) = 1$ for BCM and $\beta_i(x) = (1/2)[\log(v_{prior}(x)) - \log(v_i(x))]$ for RBCM.

The next proposition shows that aggregations given by (2) can lead to mean square prediction errors that do not go to zero as $n \to \infty$, when considering triangular arrays of observation points that become dense in a compact set D (which is the infill asymptotic setting, Cressie (1993), Stein (2012)). This proposition thus provides a counter-example, but does not prove that aggregation procedures given by (2) are inconsistent in general. This inconsistency will however be confirmed in some further simple numerical experiments. The property relies on Gaussian processes satisfying the no-empty ball (NEB) property, which has been introduced in Vazquez and Bect (2010a).

Definition 1. *A Gaussian process Y on D has the NEB property if for any $x_0 \in D$ and for any sequence $(x_i)_{i\geq 1}$ of points in D, the following two assertions are equivalent.*

1. $\mathbb{V}[Y(x_0)|Y(x_1), ..., Y(x_n)]$ *goes to* 0 *as* $n \to \infty$,
2. x_0 *is an adherent point of the sequence* $(x_i)_{i\geq 1}$.

Proposition 1 (Non-consistency of some covariance-free aggregations)**.** *Let D be a compact subset of $\mathbb{R}^d$ with non-empty interior. Let Y be a Gaussian process on D with mean zero and covariance function k. Assume that k is defined on $\mathbb{R}^d$, continuous and satisfies $k(x, y) > 0$ for two distinct points x, y in the interior of D. Assume also that Y has the NEB property. For $n \in \mathbb{N}$ and for any triangular array of observation points $(x_{ni})_{1\leq i\leq n; n\in\mathbb{N}}$, let p_n be a number of Kriging predictors, X be the $n \times d$ matrix with row i equal to x_{ni}^t, and $X_1, ..., X_{p_n}$ be a partition of X. For $n \in \mathbb{N}$ let $M_{\mathcal{A},n}$ be defined as in (2) with p replaced by p_n. Finally, assume that*

$$\alpha_k(v_1(x), ..., v_{p_n}(x), v_{prior}(x)) \leq \frac{a(v_k(x), v_{prior}(x))}{\sum_{l=1}^{p_n} b(v_l(x), v_{prior}(x))}, \tag{3}$$

where a and b are given deterministic continuous functions from $\Delta = \{(x,y) \in (0,\infty)^2; x \leq y\}$ to $[0,\infty)$, with a and b positive on $\mathring{\Delta} = \{(x,y) \in (0,\infty)^2; x < y\}$.

Then, there exists a triangular array of observation points $(x_{ni})_{1\leq i\leq n; n\in\mathbb{N}}$ such that $\lim_{n\to\infty} \sup_{x\in D} \min_{i=1,...,n} ||x_{ni} - x|| = 0$, a triangular array of submatrices $X_1, ..., X_{p_n}$ forming a partition of X, with $p_n \to_{n\to\infty} \infty$ and $p_n/n \to_{n\to\infty} 0$, and such that

$$\liminf_{n\to\infty} \int_D \mathrm{E}\left[\left(Y(x) - M_{\mathcal{A},n}(x)\right)^2\right] dx > 0. \tag{4}$$

As a consequence, there exists a subset C of D with strictly positive Lebesgue measure so that, for all $x_0 \in C$,

$$\mathrm{E}\left[\left(Y(x_0) - M_{\mathcal{A},n}(x_0)\right)^2\right] \not\to_{n\to\infty} 0. \tag{5}$$

It is easy to see that the proposition applies to the POE, GPOE, BCM, RBCM methods introduced above. Hence, Proposition 1 constitutes a significant theoretical drawback for an important class of aggregation techniques in the literature, which are based solely on conditional variances.

The detailed proof is given in Appendix A. The intuitive explanation is that the aggregation methods for which the proposition applies ignore the correlations between the different Kriging predictors. Hence, for prediction points around which the density of observation points is smaller than on average, too much weight can be given to Kriging predictors based on distant observation points. It is worth noting that, in the proof of Proposition 1, the density of observation points in the subset of D where the inconsistency occurs is asymptotically negligible compared to the average density of observation points. Hence, this proof does not apply to triangular arrays of observation points for which the density is uniform (for instance grids of points or uniformly distributed random points). Thus, Proposition 1 does not preclude the consistency of the POE, GPOE, BCM, RBCM methods for uniformly dense observation points. It should be noted that when doing optimization, or when looking for optimal designs for parameter estimation, see Fig. 4 in Zhu and Zhang (2006), one may naturally end up with strongly non-uniform densities of observation points, so that unbalanced designs leading to non-consistency are not purely theoretical.

Remark 1. *The NEB property holds for many Gaussian processes defined on $D \subset \mathbb{R}^d$ with zero mean function and covariance function k. In particular, assume that k has a positive spectral density (defined by $\hat{k}(\omega) = \int_{\mathbb{R}^d} k(x)\exp(-Jx^t\omega)dx$ with $J^2 = -1$ and for $\omega \in \mathbb{R}^d$). Assume that there exist $0 \leq A < \infty$ and $0 \leq T < \infty$ such that $1/\hat{k}(\omega) \leq A(1 + ||\omega||^t)$, with $||.||$ the Euclidean norm. Then Y has the NEB property (Vazquez and Bect 2010a;b). These assumptions are satisfied by many stationary covariance functions, such as Matérn ones, but a notable exception is the Gaussian covariance function (Proposition 1 in Vazquez and Bect 2010b).*

Remark 2. *The partitions $X_1, \ldots, X_{p_n}$ for which the inconsistency occurs in Proposition 1 can typically be representative of outputs of clustering algorithms, in the sense that points in the same group X_i would be close to each other. This is further discussed in Remark 6 in Appendix A.*

Remark 3. *Proposition 1 does not imply that all the aggregation methods based only on the conditional variances are inconsistent. In particular, consider the aggregation consisting in predicting from the subset of observations yielding the smallest conditional variance, defined by $M_{\mathcal{A}}(x) = M_{i(x)}(x)$ where $i(x) = \mathrm{argmin}_{j=1,...,p} v_j(x)$. Then, the aggregated predictor $M_{\mathcal{A}}(x)$ can be seen to be consistent from the proof of Proposition 2 below.*

3 The nested Kriging prediction

This section assumes that $M_1(x), ..., M_p(x)$ have mean zero and finite variance, but not necessarily that they can be written as $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1}Y(X_i)$. Let $M(x) = (M_1(x), ..., M_p(x))^t$ be the vector of sub-models, $K_M(x)$ be the $p \times p$ covariance matrix of $(M_1(x), ..., M_p(x))$, and $k_M(x)$ be the $p \times 1$ vector with component i equal to $\mathrm{Cov}\left[M_i(x), Y(x)\right]$. The main assumption that will be required hereafter is:

Assumption 1 (Assumptions on sub-models)**.** *For all $x \in D$, the random variables $Y(x), M_1(x), ..., M_p(x)$ have mean zero and finite variance, and the matrix $K_M(x) = \mathrm{Cov}\left[M(x), M(x)\right]$ is invertible. Furthermore, the following assumptions may be considered separately:*

(H1) M is linear in $Y(X)$: for all $x \in D$, there exists a deterministic $p \times n$ matrix $\Lambda(x)$ such that $M(x) = \Lambda(x)Y(X)$, i.e. each sub-model is a linear combination of observations $Y(X)$.

(H2) *M interpolates Y at X: for any component x_k of X there is at least one index $i_k \in \{1, \dots, p\}$ such that $M_{i_k}(x_k) = Y(x_k)$, i.e. any observation is interpolated by at least one sub-model.*

(H3) *(M, Y) is Gaussian: the joint process $(M_1(x), ..., M_p(x), Y(x))_{x \in D}$ is multivariate Gaussian.*

These assumptions are not particularly restrictive and they are satisfied in the classical situation where the sub-models are given by interpolating Kriging models $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1}Y(X_i)$, $i \in \{1, \dots, p\}$. Note that the relaxation of (H2) is takled in Section 5 and that several results presented in this section can be extended to the case where (H3) is not satisfied by using matrix pseudo-inverses.

In Rullière et al. (2018), the aggregated predictor $M_{\mathcal{A}}(x)$ is defined as the best linear predictor of $Y(x)$ from $M_1(x), ..., M_p(x)$, which implies

$$\left\{ \begin{array}{lcl} M_{\mathcal{A}}(x) & = & k_M(x)^t K_M(x)^{-1} M(x)\,, \\ v_{\mathcal{A}}(x) & = & \mathrm{E}\left[(Y(x) - M_{\mathcal{A}}(x))^2\right] = k(x,x) - k_M(x)^t K_M(x)^{-1} k_M(x)\,. \end{array} \right. \tag{6}$$

Under assumptions (H1), (H2) and (H3), the aggregated predictor $M_{\mathcal{A}}$ preserves the linearity, the interpolation properties, and the conditional Gaussianity. Furthermore, using (H1) one easily gets the expressions of $k_M(x)$ and $K_M(x)$

$$\left\{ \begin{array}{lcl} k_M(x) & = & \Lambda(x) k(X, x), \\ K_M(x) & = & \Lambda(x) k(X, X) \Lambda(x)^t. \end{array} \right. \tag{7}$$

The aggregated predictor is straightforward to compute in this case, which occurs for example when the submodels $M_1(x), ..., M_p(x)$ are simple Kriging predictors.

Rullière et al. (2018) show that, for n observation points and q prediction points, the complexity of the aggregation procedure $M_{\mathcal{A}}$ can reach simultaneously $O(n)$ in storage requirement and $O(n^2 q)$ in computational complexity when $q = o(n)$. This computational complexity is larger than the one of covariance-free aggregation procedures but much smaller than the standard Kriging complexity (see Sect. 5.1 for more details). This makes possible the use of this aggregation method with a large number of observations (up to one million points in Rullière et al. (2018)). The calculation of the nested Kriging predictor can also benefit from parallel computing, both for building the sub-models and for predicting at different prediction points. A public implementation using parallel computation and allowing measurement errors (see Sect. 5.1) and universal Kriging (see Sect. 5.2) is available at `https://github.com/drulliere/nestedKriging`.

Although this article focuses on the case where the covariance function k of the Gaussian process Y is known, the parameters of the covariance function often need to be estimated in practice (Roustant et al. 2012, Abrahamsen 1997, Stein 2012). In a big data context where the aggregated predictor of (6) is relevant, classical parameter estimation methods like maximum likelihood (Stein 2012) or cross validation (Bachoc 2013, Bachoc et al. 2017, Zhang and Wang 2010) are too computationally prohibitive to be carried out directly. Rullière et al. (2018) suggest to apply cross validation to the aggregated predictor in (6) rather than to the full Kriging predictor in (1), and to use stochastic gradient for optimization with respect to the covariance parameters. This results in a procedure that is applicable to a large data set. One could also use a smaller subset of a large data set specifically for covariance parameter estimation by classical maximum likelihood or cross validation. Finally, one could also optimize, with respect to the covariance parameters, the sum of the logarithms of the likelihoods (or of cross validation scores) from each of the subsets $X_1, Y(X_1), \dots, X_n, Y(X_n)$. This enables to exploit the entire data set for covariance parameter estimation, while keeping a manageable computational complexity.

The rest of the section focuses on the theoretical properties of this particular aggregation method: it contains the consistency results under infill asymptotics, reinterprets the nested Kriging approximation as the exact conditional expectation for a modified Gaussian process, and provides bounds on the errors $M_{\mathcal{A}}(x) - M_{full}(x)$ and $v_{\mathcal{A}}(x) - v_{full}(x)$.

3.1 Consistency

The next proposition provides the consistency result in the case where Y is a Gaussian process on D with mean zero and $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1}Y(X_i)$, which implies (H1), (H2), (H3). The proof is given in Appendix B.

Proposition 2 (Consistency)**.** *Let D be a compact nonempty subset of $\mathbb{R}^d$. Let Y be a Gaussian process on D with mean zero and continuous covariance function k. Let $(x_{ni})_{1 \leq i \leq n, n \in \mathbb{N}}$ be a triangular array of observation points so that $x_{ni} \in D$ for all $1 \leq i \leq n, n \in \mathbb{N}$ and so that for all $x \in D$, $\lim_{n \to \infty} \min_{i=1,\dots,n} ||x_{ni} - x|| = 0$.*

For $n \in \mathbb{N}$, let $X = (x_{n1}, ..., x_{nn})^t$, let $M_1(x), ..., M_{p_n}(x)$ be any collection of p_n Kriging predictors based on respective design points $X_1, \ldots, X_{p_n}$, where X_i is a subset of X, with $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1}Y(X_i)$ for $i = 1, ..., p_n$. Assume that each row of X is a row of at least one X_i, $1 \leq i \leq p_n$. Then, for $M_{\mathcal{A}}(x)$ defined as in (6):

$$\sup_{x \in D} \mathbb{E}\left[\left(Y(x) - M_{\mathcal{A}}(x)\right)^2\right] \to_{n \to \infty} 0. \tag{8}$$

Proposition 2 shows that, contrary to several aggregation techniques, taking into account the correlations between the predictors enables the aggregation method of Rullière et al. (2018) to have a guaranteed consistency.

Numerical illustration of the consistency results. Propositions 1 and 2 are now illustrated on simple examples where the test functions are given by random samples of a centered Gaussian Process Y with Matérn 3/2 covariance (see Rasmussen and Williams 2006). The observation points $x_1, \ldots, x_n \in [0, 1]$ are ordered and gathered into groups of $\sqrt{n}$ consecutive points to build $\sqrt{n}$ sub-models. These sub-models are then aggregated following the various methods presented earlier in order to make predictions $M_{\mathcal{A}}(x_t)$ at $x_t = 0.8$. The criterion used to assess the quality of a prediction is the mean square error: $MSE = \mathrm{E}\left[(Y(x_t) - M_{\mathcal{A}}(x_t))^2\right]$. Since the prediction methods that are benchmarked all correspond to linear combinations of the observed values, this expectation can be computed analytically and there is no need to generate actual samples from the test functions.

Two different settings are considered for the input point distribution and the kernel parameters: (A) a uniform distribution and a lengthscale equal to 0.1 (Fig. 1.a) and (B), a beta distribution $\beta(10, 10)$ and a lengthscale of 0.2 (Fig. 1.b). In both case the variance of Y is set to one. A small nugget effect (10^{-9} for A and 10^{-10} for B) is also included in the sub-models to ensure their computations are numerical stable. Finally, the experiments are repeated 100 times with different input locations $x_1, \ldots, x_n$.

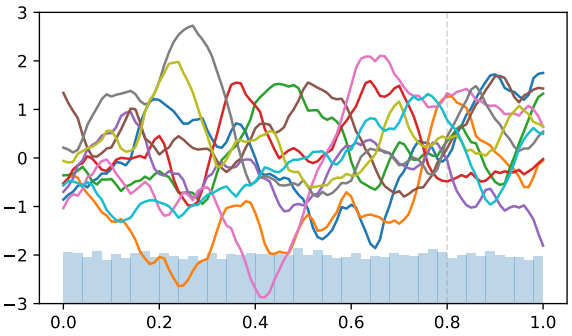
The results of the experiments are shown in panels (c) and (d) of Fig. 1. First of all, the non-consistency of the methods POE, GPOE, BCM and RBCM is striking: the MSE does not only fail to converge to zero but it actually increases when the number of observation points is greater than $5.\,10^3$ (Exp. A) or $20.\,10^4$ (Exp. B). Note that Proposition 1 only shows the existence of a training set where the variance based aggregation methods under study are non consistent: it is thus of significant practical interest to observe this behavior on these simple examples with reasonable settings. On the other hand, the nested Kriging aggregation does converge toward zero as guaranteed by Proposition 2.

3.2 The Gaussian process perspective

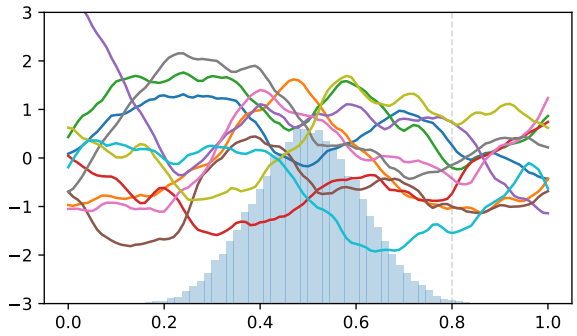
This section introduces an alternative construction of the aggregated predictor where the prior process Y is replaced by another process $Y_{\mathcal{A}}$ for which $M_{\mathcal{A}}(x)$ and $v_{\mathcal{A}}(x)$ correspond exactly to the conditional expectation and variance of $Y_{\mathcal{A}}(x)$ given $Y_{\mathcal{A}}(X)$. As discussed in Quinonero-Candela and Rasmussen (2005), this construction implies that the proposed aggregation is not only as an approximation of the full model but also as an exact method for a slightly different prior (as illustrated in the further commented Fig. 3). This type of decomposition can naturally occur in the context of predictive processes or low-rank Kriging models, see Finley et al. (2009), Banerjee et al. (2008), Cressie and Johannesson (2008). As a consequence, it also provides conditional cross-covariances and samples for the aggregated models. In particular, all the methods developed in the literature based on Kriging predicted covariances, such as Marrel et al. (2009) for sensitivity analysis and Chevalier and Ginsbourger (2013) for optimization, may hence be applied to the aggregated model in Rullière et al. (2018). Recall that $(M_1, \ldots, M_p, Y)^t$ is a centered process with finite variance on the whole input space D. The $p \times 1$ cross-covariance vector is defined as $k_M(x, x') = \mathrm{Cov}\left[M(x), Y(x')\right]$ and the $p \times p$ cross-covariance matrix $K_M(x, x') = \mathrm{Cov}\left[M(x), M(x')\right]$, for all $x, x' \in D$. These definitions result in a minor notation overloading with the definitions introduced in Sect. 3 ($K_M(x, x) = K_M(x)$ and $k_M(x, x) = k_M(x)$), but context should be sufficient to avoid confusion. The following definition introduces $Y_{\mathcal{A}}$ which is a Gaussian process for which $M_{\mathcal{A}}$ and $v_{\mathcal{A}}$ are the conditional mean and variance of $Y_{\mathcal{A}}$ given $Y_{\mathcal{A}}(X)$:

Definition 2 (Aggregated process)**.** *$Y_{\mathcal{A}}$ is defined as $Y_{\mathcal{A}} = M_{\mathcal{A}} + \varepsilon'_{\mathcal{A}}$ where $\varepsilon'_{\mathcal{A}}$ is an independent replicate of $Y - M_{\mathcal{A}}$ and with $M_{\mathcal{A}}$ as in* (6)*.*

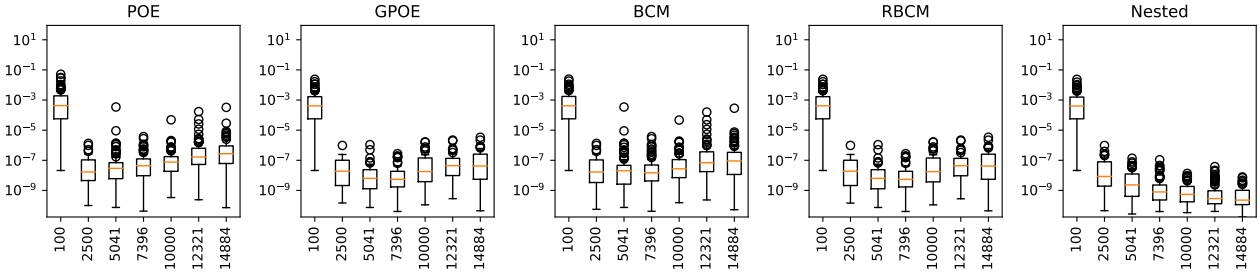
As $Y = M_{\mathcal{A}} + (Y - M_{\mathcal{A}})$, the difference between Y and $Y_{\mathcal{A}}$ is that $Y_{\mathcal{A}}$ neglects the covariances between $M_{\mathcal{A}}$ and the residual $Y - M_{\mathcal{A}}$.



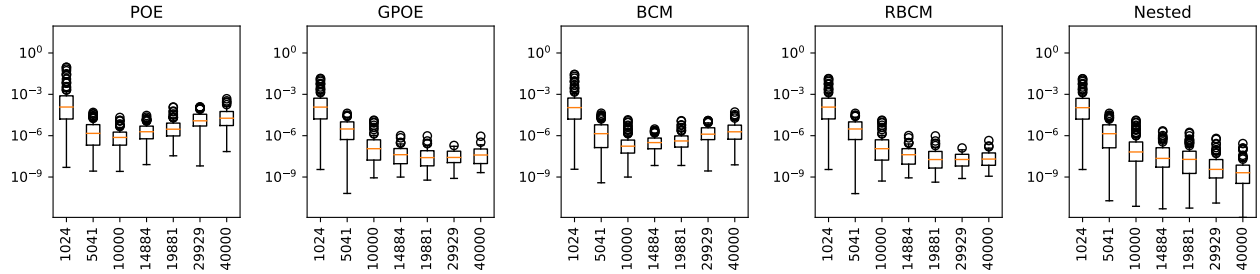
(a) Experiment A settings.



(b) Experiment B settings.



(c) Experiment A: MSE as a function of the number of observation points n.



(d) Experiment B: MSE as a function of the number of observation points n.

Figure 1: Illustration of the (non)-consistency of the various methods discussed in this paper. (a, b): Details of the experiment settings: Samples of the test functions and distribution of the input points (histogram in the background). The vertical dashed line denotes the test point where MSE is computed. (c, d) Prediction accuracy (MSE) versus the number of observation points.

Proposition 3 (Gaussian process perspective)**.** *If $M_{\mathcal{A}}$ is a deterministic and interpolating function of $Y(X)$, i.e. if for any $x \in D$ there exists a deterministic function $g_x : \mathbb{R}^n \to \mathbb{R}$ such that $M_{\mathcal{A}}(x) = g_x(Y(X))$ and if $M_{\mathcal{A}}(X) = Y(X)$, or in particular under linearity and interpolation assumptions (H1) and (H2) then*

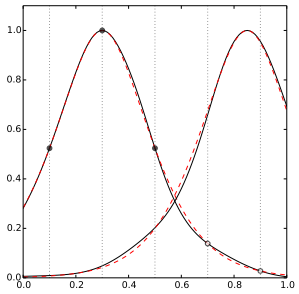
$$\begin{cases} M_{\mathcal{A}}(x) = \mathrm{E}\left[Y_{\mathcal{A}}(x)|Y_{\mathcal{A}}(X)\right] , \\ v_{\mathcal{A}}(x) = \mathrm{V}\left[Y_{\mathcal{A}}(x)|Y_{\mathcal{A}}(X)\right] . \end{cases} \tag{9}$$

As already stated, given the observations $Y_{\mathcal{A}}(X)$, the conditional process $Y_{\mathcal{A}}$ is interesting since its conditional mean and variance, at any point x, correspond to the approximated conditional mean and variance of the process Y obtained by the nested Kriging technique. It is thus natural to consider sample paths of this conditional process $Y_{\mathcal{A}}$. In the Gaussian setting, studying the unconditional (prior) distribution of the centred process $Y_{\mathcal{A}}$ and the conditional (posterior) distribution of $Y_{\mathcal{A}}$ given the observations $Y_{\mathcal{A}}(X)$ boils down to studying the prior and posterior covariances of $Y_{\mathcal{A}}$. The covariance $k_{\mathcal{A}}$ of the process $Y_{\mathcal{A}}$ can be calculated and shown to coincide with the one of the process Y at several locations, in particular, denoting $k_{\mathcal{A}}(x, x') = \mathrm{Cov}\left[Y_{\mathcal{A}}(x), Y_{\mathcal{A}}(x')\right]$, one can show that for all $x \in D$, $Y(x)$ and $Y_{\mathcal{A}}(x)$ have the same variance: $k_{\mathcal{A}}(x, x) = k(x, x)$. Furthermore, under the interpolation assumption (H2), $k_{\mathcal{A}}(X, X) = k(X, X)$. Figure 2 illustrates the difference between the covariance functions k and $k_{\mathcal{A}}$, using the same settings as in Fig. 3. Each panel of the figure deserves some specific comments:

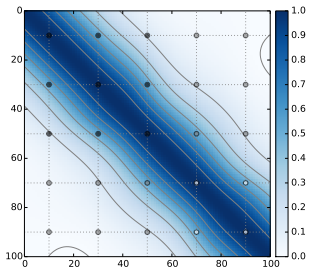
(a) the absolute difference between the two covariance functions k and $k_{\mathcal{A}}$ is small. Furthermore, the identity $k_{\mathcal{A}}(X, X) = k(X, X)$ is illustrated : as 0.3 is a component of X, $k_{\mathcal{A}}(0.3, x_k) = k(0.3, x_k)$ for any of the five components x_k of X.

(b) the contour lines for $k_{\mathcal{A}}$ are not straight lines, as it is the case for stationary processes. In this example, Y is stationary whereas $Y_{\mathcal{A}}$ is not. However, the latter only departs slightly from the stationary assumption.

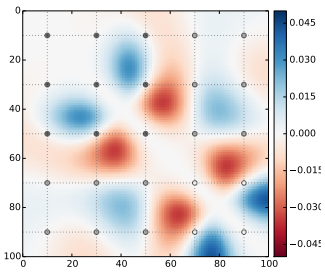
(c) the difference $k_{\mathcal{A}} - k$ vanishes at some places, among which are the places of the bullets points and the diagonal which correspond respectively to $k_{\mathcal{A}}(X, X) = k(X, X)$ and $k_{\mathcal{A}}(x, x) = k(x, x)$. Furthermore, the absolute differences between the two covariances functions are again quite small. It also shows that the pattern of the differences is quite complex.



(a) Covariance functions $k_{\mathcal{A}}(a, \cdot)$ (solid lines) and $k(a, \cdot)$ (dashed lines) with $a = 0.3 \in X$ and $a = 0.85 \notin X$.



(b) Contour plot of the modified covariance function $k_{\mathcal{A}}$.



(c) Image plot of the difference between covariance functions $k_{\mathcal{A}} - k$.

Figure 2: Comparisons of the modified covariance $k_{\mathcal{A}}$ and the initial covariance k. The horizontal and vertical dotted lines correspond to the locations of the observation points x_i for $i \in \{1, \ldots, 5\}$. The bullets indicate locations where $k_{\mathcal{A}}(x_i, x_j) = k(x_i, x_j)$.

The previous considerations may help understanding the differences between Y and $Y_{\mathcal{A}}$, and thus the approximation that is made by the nested Kriging technique. Another interest of $Y_{\mathcal{A}}$ is that one can introduce conditional cross-covariances and sample paths. The following proposition shows that conditional (posterior) cross-covariances of $Y_{\mathcal{A}}$ can be easily computed. In particular, it enables the computation of conditional sample paths of $Y_{\mathcal{A}}$. The proposition also gives some simplifications that make computations tractable even in the case where the number of observations is large.

Proposition 4 (Posterior covariances of $Y_{\mathcal{A}}$)**.** *Define the conditional (posterior) cross-covariances of the process $Y_{\mathcal{A}}$ given $Y_{\mathcal{A}}(X)$ as*

$$c_{\mathcal{A}}(x,x') = \mathrm{Cov}\left[Y_{\mathcal{A}}(x), Y_{\mathcal{A}}(x')|Y_{\mathcal{A}}(X)\right] , \tag{10}$$

with $x, x' \in D$. Assume that (M, Y) is Gaussian, then $Y_{\mathcal{A}}$ is also Gaussian and the following results hold:

(i) The posterior covariance function $c_{\mathcal{A}}$ writes, for all $x, x' \in D$,

$$c_{\mathcal{A}}(x,x') = k_{\mathcal{A}}(x,x') - k_{\mathcal{A}}(x,X)k_{\mathcal{A}}(X,X)^{-1}k_{\mathcal{A}}(X,x'). \tag{11}$$

(ii) Denote $\alpha_{\mathcal{A}}(x) = K_M(x)^{-1}k_M(x)$. Under linear and interpolation assumptions (H1) and (H2),

$$c_{\mathcal{A}}(x,x') \quad = \quad k(x,x') - \alpha_{\mathcal{A}}(x)^t k_M(x,x') - k_M(x',x)\alpha_{\mathcal{A}}(x') + \alpha_{\mathcal{A}}(x)^t K_M(x,x')\alpha_{\mathcal{A}}(x') , \tag{12}$$

(iii) Under linear and interpolation assumptions (H1) and (H2),

$$c_{\mathcal{A}}(x,x') \quad = \quad \mathrm{E}\left[\left(Y(x) - M_{\mathcal{A}}(x)\right)\left(Y(x') - M_{\mathcal{A}}(x')\right)\right] . \tag{13}$$

In other words, conditional covariances can be understood as prior covariances between residuals.

It should be noted that computing $c_{\mathcal{A}}$ or generating conditional samples of $Y_{\mathcal{A}}$ by using Eq. (11) requires to inverse the $n \times n$ matrix $k_{\mathcal{A}}(X,X)$ which is computationally costly for large n. On the contrary, computing $c_{\mathcal{A}}$ by using Eq. (12) does only require the computation of covariances between predictors and is tractable even with large datasets. Consider the prediction problem with n observation points and q prediction points where both n and q can be large, with $q = o(n)$. Consider a reasonable dimension $d \leq O(q)$ and a typical number of sub-models $p = \sqrt{n}$. The complexity for obtaining the nested Kriging mean and variance $\{M_{\mathcal{A}}(x), v_{\mathcal{A}}(x)\}$ for all prediction points is $\mathcal{C} = O(qn^2)$ in computational complexity and $\mathcal{S} = O(nq)$ in storage requirement for the fastest implementations (see Rullière et al. 2018). This storage requirement can be reduced to $\mathcal{S} = O(n)$ when recalculating some quantities. When computing $\{M_{\mathcal{A}}(x), v_{\mathcal{A}}(x)\}$ together with output covariances $\{c_{\mathcal{A}}(x,x')\}$ for all prediction points, using Eq. (12), one can show that the reachable computational complexity is unchanged and is $\mathcal{C} = O(qn^2)$ when $pq \leq n$, or becomes $\mathcal{C} = O(q^2pn)$ otherwise. The associated storage requirement becomes $\mathcal{S} = O(nq^2)$. At last, in the more general case where $O(n^{1/2}) \leq p \leq O(n^{2/3})$ and $q \leq O(n^{2/3})$, one can show that computational complexity is $\mathcal{C} = O(qn^2)$ without computing the covariances $c_{\mathcal{A}}(x,x')$ or $\mathcal{C} = O(q^2p^2)$ when computing these covariances.

This Gaussian Process perspective and Proposition 4 can be combined to define unconditional and conditional sample paths. This is illustrated in Fig. 3 which displays prior and posterior samples of a process $Y_{\mathcal{A}}$ based on a process Y with squared exponential covariance $k(x,x') = \exp\left(-12.5(x-x')^2\right)$. In this example, the test function is $f(x) = \sin(2\pi x) + x$, and the input $X = (0.1, 0.3, 0.5, 0.7, 0.9)^t$ is divided into $p = 2$ subgroups $X_1 = (0.1, 0.3, 0.5)^t$ and $X_2 = (0.7, 0.9)^t$.

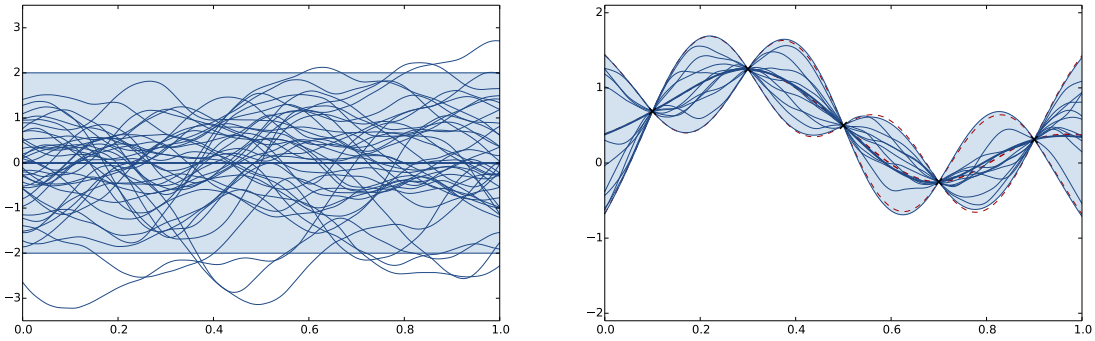
Proposition 4 can be used to draw similarities between nested Kriging and low-rank Kriging (see Stein 2014, and references therein). In both cases, the predictions can be seen as a tractable approximation of an initial model, but they also correspond to an exact posterior for their stated covariance models (which is not stationary, in general, see Figure 2). The main difference between the two methods is that contrarily to low-rank Kriging, nested Kriging remains a non-parametric approach. This however comes with an additional computational cost which comes from matrix inverse that need to be computed at prediction time.

Knowing that the predictor $M_{\mathcal{A}}$ is a conditional expectation for the process $Y_{\mathcal{A}}$ can be used to analyze its error for predicting $Y(x)$, by studying the differences between the distributions of Y and $Y_{\mathcal{A}}$, in the same vein as in Stein (2012) or Putter et al. (2001). The next section provides more details on the prediction errors made by choosing $Y_{\mathcal{A}}$ in place of Y as a prior.

3.3 Bounds on aggregation errors

This section aims at studying the differences between the aggregated model $M_{\mathcal{A}}, v_{\mathcal{A}}$ and the full one M_{full}, v_{full}. This section focuses on the case where $M(x)$ is linear in $Y(X)$, i.e. there exists a $p \times n$ deterministic matrix $\Lambda(x)$ such that $M(x) = \Lambda(x)Y(X)$. This results in

$$\left\{ \begin{array}{lcl} M_{\mathcal{A}}(x) - M_{full}(x) & = & -k(x,X)\Delta(x)Y(X) , \\ v_{\mathcal{A}}(x) - v_{full}(x) & = & k(x,X)\Delta(x)k(X,x) , \end{array} \right. \tag{14}$$



(a) Unconditional samples.

(b) Conditional samples.

Figure 3: Illustration of the modified process $Y_{\mathcal{A}}$. (a) Unconditional sample paths from the modified Gaussian process $Y_{\mathcal{A}}$, with mean 0 and covariance $k_{\mathcal{A}}$. (b) Conditional sample paths of $Y_{\mathcal{A}}$ given $Y_{\mathcal{A}}(X) = f(X)$, with mean $M_{\mathcal{A}}$ and covariance $c_{\mathcal{A}}$. The thick lines and the blue areas correspond to the means and 95% confidence intervals for $Y_{\mathcal{A}}$. The dashed red lines are the mean and 95% confidence intervals for the full model with the original prior Y.

where $\Delta(x) = k(X,X)^{-1} - \Lambda(x)^t\big(\Lambda(x)k(X,X)\Lambda(x)^t\big)^{-1}\Lambda(x)$, as soon as $\Lambda(x)k(X,X)\Lambda(x)^t$ is invertible.

As illustrated in Fig. 2, the covariance functions k and $k_{\mathcal{A}}$ are very similar. The following proposition shows that the difference between these covariances can be linked to the aggregation error and can provide a bounds for the absolute errors.

Proposition 5 (Errors using covariance differences)**.** *Under the linear and interpolation assumptions (H1) and (H2), the differences between the full and aggregated models write as differences between covariance functions:*

$$\begin{cases} \mathrm{E}\left[(M_{\mathcal{A}}(x) - M_{full}(x))^2\right] = \|k(X,x) - k_{\mathcal{A}}(X,x)\|_K^2 \,, \\ v_{\mathcal{A}}(x) - v_{full}(x) = \|k(X,x)\|_K^2 - \|k_{\mathcal{A}}(X,x)\|_K^2 \,. \end{cases} \tag{15}$$

The absolute differences can be bounded:

$$\begin{cases} |M_{\mathcal{A}}(x) - M_{full}(x)| & \leq & \|k(X,x) - k_{\mathcal{A}}(X,x)\|_K \|Y(X)\|_K \,, \\ |v_{\mathcal{A}}(x) - v_{full}(x)| & \leq & \|k(X,x)\|_K^2 \,, \end{cases} \tag{16}$$

where $\|u\|_K^2 = u^t k(X,X)^{-1} u$. *Assuming that the smallest eigenvalue* $\lambda_{\min}$ *of* $k(X,X)$ *is non zero, this norm can be bounded by* $\|u\|_K^2 \leq \frac{1}{\lambda_{\min}}\|u\|^2$ *where* $\|u\|$ *denotes the Euclidean norm. Furthermore, since* $v_{full}(x) = \mathrm{E}\left[(Y(x) - M_{full}(x))^2\right]$, *then*

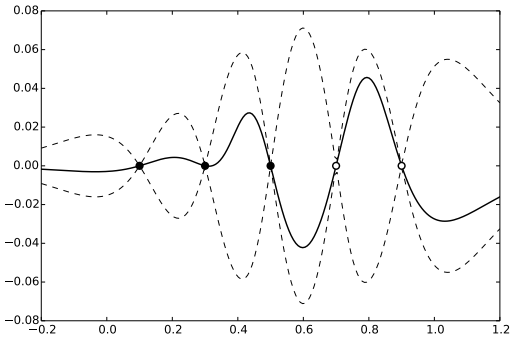
$$0 \leq v_{\mathcal{A}}(x) - v_{full}(x) \leq \min_{k\in\{1,\dots,p\}} \mathrm{E}\left[(Y(x) - M_k(x))^2\right] - v_{full}(x) \,. \tag{17}$$

Note that previous result is provided for a given number n of observations, for a finite a dimensional $n \times n$ matrix X. The asymptotic of the bounds as n grows to infinity depends on the design sequence and the nature of the asymptotic setting (e.g., expansion domain or fixed domain). It would require further developments that are not considered here.

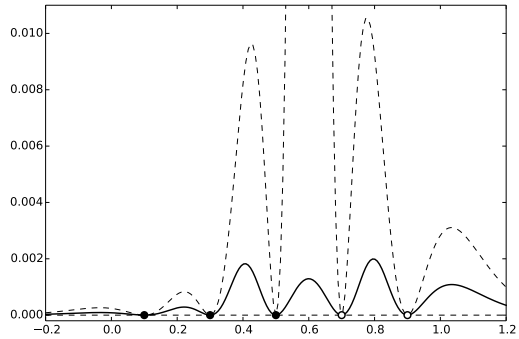
Proposition 5 implies that the nested Kriging aggregation has two desirable properties that are detailed in Remarks 4 and 5 (with proofs in Appendix).

Remark 4 (Far prediction points)**.** *For a given number of observations* n *and a given design* X*, if one can choose a prediction point* x *far enough from the observation points in* X*, in the sense* $\|k(X,x)\| \leq \epsilon$ *for any given* $\epsilon > 0$*, then* $|M_{\mathcal{A}}(x) - M_{full}(x)|$ *and* $|v_{\mathcal{A}}(x) - v_{full}(x)|$ *can be as small as desired.*

One consequence of the previous remark is that when the covariances between the prediction point x and the observed ones X become small, both models tend to predict the unconditional distribution of $Y(x)$. This is a natural property that is desirable for any aggregation method but it is not always fulfilled. For example, aggregating two sub-models with POE leads to overconfident models with wrong variance as discussed in Deisenroth and Ng (2015).



(a) differences between predicted means $m_{\mathcal{A}}(x) - m_{full}(x)$.



(b) differences between predicted variances $v_{\mathcal{A}}(x) - v_{full}(x)$.

Figure 4: Comparisons of the full and aggregated model. The dashed lines correspond to the bounds given in Proposition 5: $\pm\lambda_{\min}^{-1/2}\|k(X,x) - k_{\mathcal{A}}(X,x)\|$ on panel (a) and bounds of (17) on panel (b).

The difference between the full model and the aggregated one of Fig. 3 is illustrated in Fig. 4. Various remarks can be made on this figure. First, the difference between the aggregated and full model is small, both on the predicted mean and variance. Second, the error tends toward 0 when the prediction point x is far away from the observations X. This illustrates Proposition 5 in the case where $\|k(X,x)\|$ is small. Third, it can be seen that the bounds on the left panel are relatively tight on this example, and that both the errors and their bounds vanish at observation points. At last, the right panel shows $v_{\mathcal{A}}(x) \geq v_{full}(x)$. This is because the estimator $M_{\mathcal{A}}$ is expressed as successive optimal linear combinations of $Y(X)$, which have a quadratic error necessarily greater or equal than M_{full} which is the optimal linear combination of $Y(X)$. Panel (b) also illustrates that the bounds given in (17) are relatively loose. This means that the nested aggregation is more informative than the most accurate sub-model.

At last, the following remark gives another very natural optimality property that is however not satisfied by other aggregation methods such as POE, GPOE, BCM and RBCM (see Sect. 2): if the sub-models contain enough information, the aggregated model corresponds to the full one.

Remark 5 (Fully informative sub-models)**.** *Assume (H1) that $M(x)$ is linear in $Y(X)$: $M(x) = \Lambda(x)Y(X)$ and that $\Lambda(x)$ is a $n \times n$ matrix with full rank, then*

$$\begin{cases} M_{\mathcal{A}}(x) &= M_{full}(x)\,, \\ v_{\mathcal{A}}(x) &= v_{full}(x)\,. \end{cases} \tag{18}$$

Furthermore, if (H3) also holds,

$$Y_{\mathcal{A}} \overset{law}{=} Y \quad \text{and thus} \quad Y_{\mathcal{A}}|Y_{\mathcal{A}}(X) \overset{law}{=} Y|Y(X). \tag{19}$$

In other words, there is no difference between the full and the approximated models when $\Lambda(x)$ is invertible.

Note that there is of course no computational interest in building and merging fully informative sub-models since it requires computing and inverting a matrix that has the same size as $k(X,X)$ so there is no complexity gain compared to the full model.

4 Analysis of the impact of the group choice

This section studies the impact of the choice of the partition $X_1, \ldots, X_p$ of a set of n two-by-two distinct observation points $\{x_1, \ldots, x_n\}$, on the quality of the predictor obtained by aggregating p Gaussian process models based on $X_1, \ldots, X_p$.

4.1 Theoretical results in dimension 1

This section focuses on the univariate case $d = 1$ where with the input locations $x_1, \ldots, x_n \in \mathbb{R}$ are fixed and distinct points and where Y is a centered Gaussian process with exponential covariance function k defined by

$$k(t_1, t_2) = \sigma^2 \exp(-|t_1 - t_2|/\theta), \quad t_1, t_2 \in \mathbb{R} \tag{20}$$

for fixed $(\sigma^2, \theta) \in (0, \infty)^2$. This choice of covariance function makes Y a Markovian GP (Ying 1991), which will prove useful to derive theoretical properties on the influence of clustering. More precisely, the idea is to assess whether selecting the groups $X_1, \ldots, X_p$ based on distances (i.e., placing observation points close to each other in the same group) is beneficial for the approximation accuracy or not. In dimension 1, the concept of perfect clustering can be defined as follows.

Definition 3 (Perfect clustering)**.** *A partition $X_1, \ldots, X_p$ of $\{x_1, \ldots, x_n\}$, composed of non-empty groups, is a perfect clustering if there does not exist any triplet u, v, w, with $u, v \in X_i$ and $w \in X_j$ with $i, j \in \{1 \ldots, p\}$, $i \neq j$, and so that $u < w < v$.*

The above definition means that the groups $X_1, \ldots, X_p$ are constituted of consecutive points. A partition $X_1, \ldots, X_p$ is a perfect clustering if and only if it can be reordered as $X_{i_1}, \ldots, X_{i_p}$ with $\{i_1, \ldots, i_p\} = \{1, \ldots, p\}$ and so that for any $u_1 \in X_{i_1}, \ldots, u_p \in X_{i_p}$, the u_i are ordered $u_1 < \ldots < u_p$.

The next proposition shows that the nested Kriging predictor coincides with the predictor based on the full Gaussian process model, if and only if $X_1, \ldots, X_p$ is a perfect clustering. It thus provides a theoretical confirmation that placing observations points close to each other in the same group is beneficial to the nested Kriging procedure.

Proposition 6 (Nested Kriging and perfect clustering)**.** *Consider an exponential covariance function k in dimension $d = 1$, as in* (20)*. Let $M_{full}(x)$ be the full predictor as in* (1) *and let $M_{\mathcal{A}}(x)$ be the nested Kriging predictor as in* (6)*, where $M_1, \ldots, M_p$ are the Gaussian process predictors based on the individual groups $X_1, \ldots, X_p$, as in Sect. 2, that is assumed to be non-empty. Then, $\mathbb{E}[(M_{full}(x) - M_{\mathcal{A}}(x))^2] = 0$ for all $x \in \mathbb{R}$ if and only if $X_1, \ldots, X_p$ is a perfect clustering.*

Proof. Let $(x, v_1, \ldots, v_r)$ be $r + 1$ two-by-two distinct real numbers. If $x < \min(v_1, \ldots, v_r)$, then the conditional expectation of $Y(x)$ given $Y(v_1), \ldots, Y(v_r)$ is equal to $\exp(-|min(v_1, \ldots, v_r) - x|/\theta)Y(min(v_1, \ldots, v_r))$ (Ying 1991). Similarly, if $x > \max(v_1, \ldots, v_r)$, then the conditional expectation of $Y(x)$ given $Y(v_1), \ldots, Y(v_r)$ is equal to $\exp(-|max(v_1, \ldots, v_r) - x|/\theta)Y(max(v_1, \ldots, v_r))$. If $\min(v_1, \ldots, v_r) < x < \max(v_1, \ldots, v_r)$, then the conditional expectation of $Y(x)$ given $Y(v_1), \ldots, Y(v_r)$ is equal to $aY(x_<) + bY(x_>)$ where $x_<$ and $x_>$ are the left-most and right-most neighbors of x in $\{v_1, \ldots, v_r\}$ and where a, b are non-zero real numbers (Bachoc et al. 2017). Finally, because the covariance matrix of $Y(v_1), \ldots, Y(v_r)$ is invertible, two linear combinations $\sum_{i=1}^r a_i Y(v_i)$ and $\sum_{i=1}^r b_i Y(v_i)$ are equal almost surely if and only if $(a_1, \ldots, a_r) = (b_1, \ldots, b_r)$.

Assume that $X_1, \ldots, X_p$ is a perfect clustering and let $x \in \mathbb{R}$. It is known from Rullière et al. (2018) that $M_{\mathcal{A}}(x) = M_{full}(x)$ almost surely if $x \in \{x_1, \ldots, x_n\}$. Consider now that $x \notin \{x_1, \ldots, x_n\}$.

If $x < \min(x_1, \ldots, x_n)$, then for $i = 1, \ldots, p$, $M_i(x) = \exp(-|x_{j_i} - x|/\theta)Y(x_{j_i})$ with $x_{j_i} = \min\{x; x \in X_i\}$. Let $i^* \in \{1, \ldots, p\}$ be so that $\min(x_1, \ldots, x_n) \in X_{i^*}$. Then $M_{full}(x) = \exp(-|x_{j_{i^*}} - x|/\theta)Y(x_{j_{i^*}})$. As a consequence, the linear combination $\lambda_x^t M(x)$ minimizing $\mathbb{E}[(\lambda^t M(x) - Y(x))^2]$ over $\lambda \in \mathbb{R}^p$ is given by $\lambda_x = e_{i^*}$ with e_r the r-th base column vector of $\mathbb{R}^p$. This implies that $M_{full}(x) = M_{\mathcal{A}}(x)$ almost surely. Similarly, if $x > \max(x_1, \ldots, x_n)$, then $M_{full}(x) = M_{\mathcal{A}}(x)$ almost surely.

Consider now that there exists $u \in X_i$ and $v \in X_j$ so that $u < x < v$ and (u, v) does not intersect with $\{x_1, \ldots, x_n\}$. If $i = j$, then $M_i(x) = M_{full}(x)$ almost surely because the left-most and right-most neighbors of x are both in X_i. Hence, also $M_{\mathcal{A}}(x) = M_{full}(x)$ almost surely in this case. If $i \neq j$, then $u = \max\{t; t \in X_i\}$ and $v = \min\{t; t \in X_j\}$ because $X_1, \ldots, X_p$ is a perfect clustering. Hence, $M_i(x) = \exp(-|x - u|)Y(u)$, $M_j(x) = \exp(-|x - v|)Y(v)$ and $M_{full}(x) = aY(u) + bY(v)$ with $a, b \in \mathbb{R}$. Hence, there exists a linear combination $\lambda_i M_i(x) + \lambda_j M_j(x)$ that equals $M_{full}(x)$ almost surely. As a consequence, the linear combination $\lambda_x^t M(x)$ minimizing $\mathbb{E}[(\lambda^t M(x) - Y(x)]^2)$ over $\lambda \in \mathbb{R}^p$ is given by $\lambda_x = \lambda_i e_i + \lambda_j e_j$. Hence $M_{full}(x) = M_{\mathcal{A}}(x)$ almost surely. All the possible sub-cases have now been treated, which proves the first implication of the proposition.

Assume now that $X_1, \ldots, X_p$ is not a perfect clustering. Then there exists a triplet u, v, w, with $u, v \in X_i$ and $w \in X_j$ with $i, j = 1 \ldots, p$, $i \neq j$, and so that $u < w < v$. Without loss of generality it can further be assumed that there does not exits $z \in X_i$ satisfying $u < z < v$.

Let x satisfies $u < x < w$ and so that (u, x) does not intersect $\{x_1, \ldots, x_n\}$. Then $M_{full}(x) = aY(u) + bY(z)$ with $a, b \in \mathbb{R}\backslash\{0\}$ and $z \in \{x_1, \ldots, x_n\}$, $z \neq v$. Also, $M_i(x) = \alpha Y(u) + \beta Y(v)$ with $\alpha, \beta \in \mathbb{R}\backslash\{0\}$. As a consequence, there can not exist a linear combination $\lambda^t M(x)$ with $\lambda \in \mathbb{R}^p$ so that $\lambda^t M(x) = aY(u) + bY(w)$. Indeed a linear combination $\lambda^t M(x)$ is a linear combination of $Y(x_1), \ldots, Y(x_n)$ where the coefficients for $Y(u)$ and $Y(v)$ are $\lambda_i \alpha$ and $\lambda_i \beta$, which are either simultaneously zero or simultaneously non-zero. Hence, $M_{\mathcal{A}}(x)$ is not equal to $M_{full}(x)$ almost surely. This concludes the proof. □

The next proposition shows that the aggregation techniques that ignore the covariances between sub-models can never recover the full Gaussian process predictor, even in the case of a perfect clustering. This again highlights the additional quality guarantees brought by the nested Kriging procedure.

Proposition 7 (Non-perfect other aggregation methods)**.** *Consider an exponential covariance function k in dimension $d = 1$, as in* (20)*. Let $p \geq 3$ and let $X_1, \ldots, X_p$ be non-empty. Let $M_{\mathcal{A}}$ be a covariance-free aggregation method defined as in* (2)*, with $\alpha_k(v_1, \ldots, v_p, v) \in \mathbb{R} \setminus \{0\}$ for $v_1, \ldots, v_p, v \in (0, \infty)$ and $v_1 < v, \ldots, v_p < v$. Then, for all $x \in \mathbb{R} \setminus X$, $\mathbb{E}[(M_{full}(x) - M_{\mathcal{A}}(x))^2] > 0$.*

Proof. Let $x \in \mathbb{R} \setminus X$. For $i = 1, \ldots, p$, $0 < v_i(x) < v_{prior}(x)$, so that $\alpha_i(v_1(x), \ldots, v_p(x), v_{prior}(x)) \in \mathbb{R} \setminus \{0\}$. Hence, the linear combination $M_{\mathcal{A}}(x) = \sum_{k=1}^{p} \alpha_k(v_1(x), ..., v_p(x), v_{prior}(x)) M_k(x)$ is a linear combination of $Y(x_1), \ldots, Y(x_n)$ with at least p non-zero coefficients (since each $M_k(x)$ is a linear combination of one or two elements of $Y(x_1), \ldots, Y(x_n)$, all these elements being two-by-two distinct, see the beginning of the proof of Proposition 6). Hence, because the covariance matrix of $Y(x_1), \ldots, Y(x_n)$ is invertible, $M_{\mathcal{A}}(x)$ can not be equal to $M_{full}(x)$ almost surely, since $M_{full}(x)$ is a linear combination of $Y(x_1), \ldots, Y(x_n)$ with one or two non-zero coefficients. □

The above Proposition applies to the POE, GPOE, BCM and RBCM procedures presented in Sect. 2.

4.2 Empirical results

The aim of this section is to illustrate Proposition 6, and to study the influence of the allocation of the observation points to the sub-models. Two opposite strategies are indeed possible: the first one consists in allocating all the points in one region of the input space to the same sub-model (which is then accurate in this region but not informative elsewhere). The second is to have, for each sub-model, points that are uniformly distributed among the set of observations which leads to having a lot of sub-models that are weekly informative. This section illustrates the impact of this choice on the nested Kriging MSE.

The experiment is as follow. A set of $32^2 = 1024$ observation points are distributed on a regular grid in one dimension and two methods are considered for creating 32 subsets of points: a k-means clustering and the optimal clustering which consists in grouping together sets of 32 consecutive points. These initial grouping of points can be used to build sub-models that are experts in their region of the space. In order to study the influence of the quality of the clustering is, the clusters are perturbed by looping over all observations points and for each of them the group is swapped with another random observation point with probability p. The value p can then be used as a measure of the disorder in the group assignment: for $p = 0$ the groups are perfect clusters and for $p = 1$, each observation is assigned a group at random.

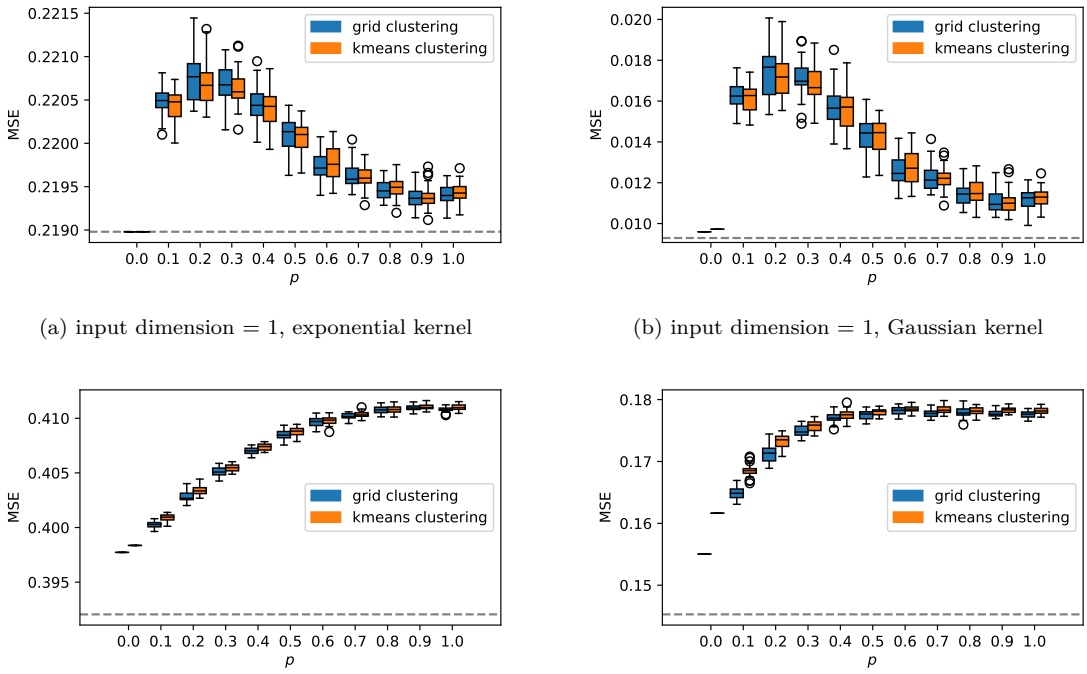
Figure 5 (top) shows the MSE of one dimensional nested Kriging models as a function of p, for test functions that correspond to samples of Gaussian processes and a test set of 200 uniformly distributed points. The covariance functions of the Gaussian processes are either the exponential or the Gaussian (i.e. squared exponential) kernels, with unit variance and a lengthscale such that the covariance between two neighbour points is 0.5. As predicted by Proposition 6, the error is null for $p = 0$ (which corresponds to a perfect clustering) when using an exponential kernel. Although this is not supported by theoretical guaranties, one can see that the prediction error is also the smallest for at $p = 0$ for a Gaussian kernel. Finally, one can note that the choice of the initial clustering method does not have a strong influence on the MSE. This can probably be explained by the good performance of the k-means algorithm in one dimension.

For the sake of completeness, the experiment is repeated with the same settings as above except for the input space dimension that is changed from one to five, and the locations of the $1024 = 4^5$ input points that are now given by the grid $\{1/8,\ 3/8,\ 5/8,\ 7/8\}^5$. With such settings, the optimal clustering of the observations can be obtained analytically with the $32 = 2^5$ cluster centers located at $\{1/4,\ 3/4\}^5$. As previously, the models that perform the best are obtained with $p = 0$. Furthermore, the difference between the two clustering methods is now more pronounced and the MSE obtained with k-means is always higher than the one with the optimal clustering.

These two experiments, together with the theoretical result in dimension one, suggest it is good practice to apply a clustering algorithm to decide how to assign the observation points to the sub-models.

5 Extensions of the nested Kriging prediction

This section extends nested Kriging to the cases where the Gaussian process Y is observed with measurement errors, and where Y has a parametric mean function. It also provides theoretical guaranties similar to Propositions 1 and 2 on the (non-)consistency of various aggregation methods in the noisy setting.



(a) input dimension = 1, exponential kernel

(b) input dimension = 1, Gaussian kernel

(c) input dimension = 5, exponential kernel

(d) input dimension = 5, Gaussian kernel

Figure 5: Nested Kriging MSE as a function of how clustered the inputs of the sub-models are. For small values of p, sub-models are built on points that tend to form clusters whereas they do not for large values of p. The horizontal dashed line corresponds to the optimal MSE, which is obtained with full Kriging model.

5.1 Measurement errors

This section assumes that the vector of observation is given by $Y(X) + \xi_X$ where the vector of measurement errors $\xi_X = (\xi_1, \ldots, \xi_n)^t$ has independent components, independent of Y, and where $\xi_i \sim \mathcal{N}(0, \eta_i)$ for $i = 1, \ldots, n$, with the error variances $\eta_1 > 0, \ldots, \eta_n > 0$. Consider a partition $X_1, \ldots, X_p$ of X, where $X_1, \ldots, X_p$ have cardinalities $n_1, \ldots, n_p$. For $i = 1, \ldots, p$, write ξ_{X_i} as the subvector of ξ_X corresponding to the submatrix X_i. Write also $D_i = \mathrm{Cov}\left[\xi_{X_i}\right]$. Then, for $x \in D$, the Kriging sub-model based on the noisy observations of $Y(X_i)$ is

$$M_{\eta,i}(x) = k(x, X_i)(k(X_i, X_i) + D_i)^{-1}(Y(X_i) + \xi_{X_i}). \tag{21}$$

Note that $M_{\eta,i}(x)$ is the best linear unbiased predictor of $Y(x)$ from $Y(X_i) + \xi_{X_i}$. Then, the best linear unbiased predictor of $Y(x)$ from $M_{\eta,1}(x), \ldots, M_{\eta,p}(x)$ is

$$M_{\mathcal{A},\eta}(x) = k_{M,\eta}(x)^t K_{M,\eta}(x)^{-1} M_\eta(x). \tag{22}$$

In (22), $M_\eta(x) = (M_{\eta,1}(x), \ldots, M_{\eta,p}(x))^t$ is the vector of sub-models, $K_{M,\eta}(x)$ is the covariance matrix of $M_\eta(x)$ and $k_{M,\eta}(x)$ is the $p \times 1$ covariance vector between $M_\eta(x)$ and $Y(x)$. For $i, j = 1, \ldots, p$, their entries are

$$(k_{M,\eta}(x))_i = k(x, X_i)(k(X_i, X_i) + D_i)^{-1} k(X_i, x)\,, \tag{23}$$

$$(K_{M,\eta}(x))_{i,j} = k(x, X_i)(k(X_i, X_i) + D_i)^{-1} k(X_i, X_j)(k(X_j, X_j) + D_j)^{-1} k(X_j, x)\,. \tag{24}$$

The mean square error can also be computed analytically

$$v_{\mathcal{A},\eta}(x) = \mathrm{E}\left[(Y(x) - M_{\mathcal{A},\eta}(x))^2\right] = k(x, x) - k_{M,\eta}(x)^t K_{M,\eta}(x)^{-1} k_{M,\eta}(x). \tag{25}$$

Equations (22) and (25) follow from the same standard proof as in the case where there are no measurement errors, see for instance Rullière et al. (2018). The computational complexity and storage requirement of these expressions are the same as their counterpart without measurement errors. In order to analyse more finely the cost of computing $M_{\mathcal{A},\eta}(x)$ and $v_{\mathcal{A},\eta}(x)$, the computations is broken down in four steps: (1) computing and storing the

vectors $(k(x,X_i)(k(X_i,X_i)+D_i)^{-1})_{i=1,\dots,p}$, (2) computing and storing $(M_{\eta,i}(x))_{i=1,\dots,p}$, (3) computing and storing $K_{M,\eta}(x)$ and $(K_{M,\eta}(x))^{-1}$ and (4) computing $M_{\mathcal{A},\eta}(x)$ and $v_{\mathcal{A},\eta}(x)$.

Assume that $X_1,\dots,X_p$ have cardinalities of order n/p for simplicity. Then the computational complexity of steps (1-4) are respectively $O(p(n/p)^3)$, $O(p(n/p))$, $O(p^2(n/p)^2+p^3)$ and $O(p(n/p)+p^2)$. The total computational cost is thus $O(n^3/p^2+n^2+p^3)$, which boils down to $O(n^2)$ by taking p of order n^β with $\beta\in[1/2,2/3]$ (as opposed to $O(n^3)$ for the full Kriging predictor). The storage requirement (including intermediary storage) of steps (1) to (4) is $O((n/p)^2+p(n/p)+p^2)$. This cost becomes minimal for p of order $n^{1/2}$, reaching $O(n)$ (as opposed to $O(n^2)$ for the full Kriging predictor).

The remaining of this section focuses on consistency results when observations are corrupted with noise. Similarly to Prop. 2, the following results considers infill asymptotics and a triangular array of observation points.

Proposition 8 (Sufficient condition for nested Kriging consistency with measurement errors)**.** *Let D be a fixed nonempty subset of $\mathbb{R}^d$. Let Y be a Gaussian process on D with mean zero and continuous covariance function k. Let $(x_{ni})_{1\leq i\leq n,n\in\mathbb{N}}$ be a triangular array of observation points such that $x_{ni}\in D$ for all $1\leq i\leq n, n\in\mathbb{N}$. For $n\in\mathbb{N}$, let $X=(x_{n1},\dots,x_{nn})^t$ and let $Y(X)+\xi_X$ be observed, where $\xi_X=(\xi_1,\dots,\xi_n)^t$ has independent components, with $\xi_i\sim\mathcal{N}(0,\eta_i)$, where $(\eta_i)_{i\in\mathbb{N}}$ is a bounded sequence. Let also ξ_X be independent of Y. Let $x\in D$ be fixed. For $n\in\mathbb{N}$, let $M_{\eta,1}(x),...,M_{\eta,p_n}(x)$ be defined from (21), for a partition $X_1,\ \dots,\ X_{p_n}$ of X.*

Assume the following sufficient condition: for all $\epsilon>0$, there exists a sequence $(i_n)_{n\in\mathbb{N}}$, such that $i_n\in\{1,\dots,p_n\}$ for $n\in\mathbb{N}$ and such that the number of points in X_{i_n} at Euclidean distance less than ϵ from x goes to infinity as $n\to\infty$.

Then for $M_{\mathcal{A},\eta}(x)$ defined as in (22),

$$\mathbb{E}\left[\left(Y(x)-M_{\mathcal{A},\eta}(x)\right)^2\right]\to_{n\to\infty}0.$$

In Proposition 8, the interpretation of the sufficient condition for consistency is that, when n is large, at least one of the subsets $X_1,\dots,X_{p_n}$ contains a large number of observation points close to the prediction point x. If the minimal size of the subsets $X_1,\dots,X_{p_n}$ goes to infinity, and if these subsets are obtained from a clustering algorithm, that is the points in a subset are close to each other, then the sufficient condition in Proposition 8 typically holds. This can be seen as an additional argument in favor of selecting the subsets from a clustering algorithm. This is in agreement with Sect. 4, which conclusions also support clustering algorithms.

A particular case where the condition of Proposition 8 always holds (regardless of how the partition into subsets is made) is when the triangular array of observation points is a sequence of randomly sampled points, with a strictly positive sampling density, and when the number of subsets is asymptotically smaller than n.

Lemma 1. *Let D be fixed, bounded with non-empty interior. Let x in the interior of D be fixed. Consider a triangular array of observation points $(x_{ni})_{1\leq i\leq n,n\in\mathbb{N}}$ that is obtained from a sequence $(x_i)_{i\in\mathbb{N}}$, that is $x_{ni}=x_i$ for $1\leq i\leq n, n\in\mathbb{N}$. Assume that $(x_i)_{i\in\mathbb{N}}$ are independently sampled from a distribution with strictly positive density g on D. Consider any sequence of partitions $(X_1,\dots,X_{p_n})_{n\in\mathbb{N}}$ of $x_1,\dots,x_n$. Assume that $p_n=o(n)$ as $n\to\infty$. Then, almost surely with respect to the randomness of $(x_i)_{i\in\mathbb{N}}$, the sufficient condition of Proposition 8 holds.*

The theoretical setting of Lemma 1 is realistic with respect to situations where the observation points are not too irregulalry spaced over D. The setting $p_n=o(n)$ is particularly relevant for the nested Kriging predictor, since this setting is necessary to obtain a smaller order of computational complexity than the full Kriging predictor.

The following Proposition shows that there are situations with measurement errors where the nested Kriging predictor is consistent whereas other aggregation methods that do not use the covariances between the sub-models are inconsistent. These situations are constructed similarly as in Proposition 1. In particular, Proposition 9 applies to the extensions of POE, GPOE, BCM and RBCM methods to the case of measurement errors (see the references given in Sect. 2).

Proposition 9 (non-consistency of some covariance-free aggregations with measurement errors)**.** *Consider D, Y and k satisfying the same conditions as in Proposition 1. Let $(\eta_i)_{i\in\mathbb{N}}$ be a bounded sequence. For any triangular array of observation points $(x_{ni})_{1\leq i\leq n,n\in\mathbb{N}}$, let, for $n\in\mathbb{N}$, X be the $n\times d$ matrix with row i equal to $x_{n,i}^t$. Let ξ_X be as in Proposition 8. Then, for any partition $X_1,\dots,X_{p_n}$ of X, for $i=1,\dots,p_n$, let n_i be the cardinality of X_i, let $M_{\eta,i}(x)$ be defined as in (21) and $v_{\eta,i}(x)=k(x,x)-k(x,X_i)(k(X_i,X_i)+D_i)^{-1}k(X_i,x)$, with D_i also as in (21). Let then $M_{\mathcal{A},\eta,n}(x)$ be defined as $M_{\mathcal{A},n}(x)$ in Proposition 1, with the same assumption (3), with $v_i(x)$ replaced by $v_{\eta,i}(x)$.*

Then there exist a fixed $x\in D$, a triangular array of observation points $(x_{ni})_{1\leq i\leq n,n\in\mathbb{N}}$, and a sequence of partitions $X_1,\dots,X_{p_n}$ of X, that satisfy the sufficient condition of Proposition 8, and such that

$$\liminf_{n\to\infty}\mathbb{E}\left[\left(M_{\mathcal{A},\eta,n}(x)-Y(x)\right)^2\right]>0.$$

5.2 Universal Kriging

Consider here the case where the Gaussian process Z defined with a trend, for $x \in D$,

$$Z(x) = \sum_{i=1}^{m} \beta_i h_i(x) + Y(x),$$

where Y is, as above, a centered Gaussian process on D with mean zero and covariance function k, where the functions $h_1, \ldots, h_m : D \to \mathbb{R}$ are known and where the vector $\beta = (\beta_1, \ldots, \beta_m)^t$ is unknown. This is the setting of universal Kriging (Chiles and Delfiner 2009).

Consider the partition $X_1, \ldots, X_p$ of X with cardinalities $n_1, \ldots, n_p$. For $i = 1, \ldots, p$, let H_i be the $n_i \times m$ matrix $(h_1(X_i), \ldots, h_m(X_i))$. Then the best linear unbiased predictor of $Z(x)$ given $Z(X_i)$ is (Sacks et al. 1989)

$$M_{\mathrm{UK},i}(x) = h(x)^t \hat{\beta}_i + k(x, X_i) k(X_i, X_i)^{-1} \left(Z(X_i) - H_i \hat{\beta}_i \right),$$

with $h(x) = (h_1(x), \ldots, h_m(x))^t$ and

$$\hat{\beta}_i = \left(H_i^t k(X_i, X_i)^{-1} H_i \right)^{-1} H_i^t k(X_i, X_i)^{-1} Z(X_i).$$

The predictor $M_{\mathrm{UK},i}(x)$ is a linear function of $Z(X_i)$, satisfies $\mathbb{E}[M_{\mathrm{UK},i}(x)] = \mathbb{E}[Z(x)]$ (for all the possible values of β in $\mathbb{R}^m$) and has smallest mean square prediction error among all the predictors with these two properties.

The next proposition provides the linear aggregation of $M_{\mathrm{UK},1}(x), \ldots, M_{\mathrm{UK},p}(x)$ that is unbiased and has the smallest mean square prediction error. It thus gives an extension of nested Kriging to the universal Kriging case.

Proposition 10. *Let $M_{UK}(x) = (M_{UK,1}(x), \ldots, M_{UK,p}(x))^t$. For $i = 1, \ldots, p$ let*

$$\begin{aligned} w_i(x)^t = {} & h(x)^t \left(H_i^t k(X_i, X_i)^{-1} H_i \right)^{-1} H_i^t k(X_i, X_i)^{-1} - k(x, X_i) k(X_i, X_i)^{-1} H_i \left(H_i^t k(X_i, X_i)^{-1} H_i \right)^{-1} H_i^t k(X_i, X_i)^{-1} \\ & + k(x, X_i) k(X_i, X_i)^{-1}. \end{aligned}$$

Let $K_{UK,M}(x)$ be the $p \times p$ matrix defined by, for $i, j = 1, \ldots, p$,

$$(K_{UK,M}(x))_{i,j} = w_i(x)^t k(X_i, X_j) w_j(x).$$

Let $k_{UK,M}(x)$ be the $p \times 1$ vector defined by, for $i = 1, \ldots, p$,

$$(k_{UK,M}(x))_i = w_i(x)^t k(X_i, x).$$

Let

$$\hat{m}_{UK,M}(x) = (1_p^t K_{UK,M}(x)^{-1} 1_p)^{-1} 1_p^t K_{UK,M}(x)^{-1} M_{UK}(x),$$

with 1_p the $p \times 1$ vector with entries equal to one. Let then

$$M_{\mathcal{A},UK}(x) = \hat{m}_{UK,M}(x) + k_{UK,M}(x)^t K_{UK,M}(x)^{-1} \left(M_{UK}(x) - \hat{m}_{UK,M}(x) 1_p \right). \tag{26}$$

Then $M_{\mathcal{A},UK}(x)$ is a linear function of $M_{UK,1}(x), \ldots, M_{UK,p}(x)$, satisfies $\mathbb{E}[M_{\mathcal{A},UK}(x)] = \mathbb{E}[Z(x)]$ (for all the possible values of β in $\mathbb{R}^m$) and has smallest mean square prediction error among all the predictors with these two properties. The vector of aggregation weights is

$$\begin{aligned} \alpha_{\mathcal{A},UK}(x)^t = {} & (1_p^t K_{UK,M}(x)^{-1} 1_p)^{-1} 1_p^t K_{UK,M}(x)^{-1} \\ & - k_{UK,M}(x)^t K_{UK,M}(x)^{-1} 1_p (1_p^t K_{UK,M}(x)^{-1} 1_p)^{-1} 1_p^t K_{UK,M}(x)^{-1} \\ & + k_{UK,M}(x)^t K_{UK,M}(x)^{-1}. \end{aligned}$$

Then $M_{\mathcal{A},UK}(x) = \alpha_{\mathcal{A},UK}(x)^t M_{UK}(x)$ and the mean square error is given by

$$v_{\mathcal{A},UK}(x) = \mathbb{E}\left[\left(M_{\mathcal{A},UK}(x) - Z(x) \right)^2 \right] = k(x,x) + \alpha_{\mathcal{A},UK}(x)^t K_{UK,M}(x) \alpha_{\mathcal{A},UK}(x) - 2 \alpha_{\mathcal{A},UK}(x)^t k_{UK,M}(x). \tag{27}$$

The aggregated predictor $M_{\mathcal{A},\mathrm{UK}}(x)$ can be interpreted as a universal Kriging predictor of $Z(x)$, with the "observations" $M_{\mathrm{UK},1}(x), \ldots, M_{\mathrm{UK},p}(x)$, and with a constant unknown mean. This is particularly apparent in (26), and can be further understood in the proof of Proposition 10. It is worth noting that the "observations"

$M_{\mathrm{UK},1}(x), \ldots, M_{\mathrm{UK},p}(x)$ are already themselves universal Kriging predictors. Hence, it turns out that there are two nested steps of universal Kriging predictions when extending the nested Kriging predictor to universal Kriging.

Computing $M_{\mathcal{A},\mathrm{UK}}(x)$ and $v_{\mathcal{A},\mathrm{UK}}(x)$ can be done similarly to what has been proposed in Sect. 5.1. More precisely, the four computational steps are (1) to compute and store the vectors $(w_i(x))_{i=1,\ldots,p}$, (2) to compute and store $(M_{\mathrm{UK},i}(x))_{i=1,\ldots,p}$, with $M_{\mathrm{UK},i}(x) = w_i(x)^t Z(X_i)$ for $i = 1, \ldots, p$, (3) to compute and store $K_{\mathrm{UK},M}(x)$ and $(K_{\mathrm{UK},M}(x))^{-1}$ and (4) to compute $M_{\mathcal{A},\mathrm{UK}}(x)$ and $v_{\mathcal{A},\mathrm{UK}}(x)$.

To analyze the computational complexity and storage requirement, assume that $X_1, \ldots, X_p$ have cardinalities of order n/p for simplicity. Assume also that m is small compared to n/p and p, which is quite realistic in the framework of Kriging with big data, since the number of functions $h_1, \ldots, h_m$ is typically moderate. Then the computational cost of step (1) is $O(p(n/p)^3)$, the computational cost of step (2) is $O(p(n/p))$, the computational cost of step (3) is $O(p^2(n/p)^2 + p^3)$ and the computational cost of step (4) is $O(p(n/p) + p^2)$. As in Sect. 5.1, the total computational cost is $O(n^3/p^2 + n^2 + p^3)$ and can reach $O(n^2)$ by taking p of order n^β with $\beta \in [1/2, 2/3]$. Also as in Sect. 5.1, the storage cost (including intermediary storage) of steps (1) to (4) is $O((n/p)^2 + p(n/p) + p^2)$ and reaches $O(n)$ by taking p of order $n^{1/2}$.

6 Concluding remarks

This article proposes a theoretical analysis of several aggregation procedures recently proposed in the literature, aiming at combining predictions from Kriging sub-models constructed separately from subsets of a large data set of observations. It is shown that aggregating the sub-models based only on their conditional variances can yield inconsistent aggregated Kriging predictors. In contrasts, the consistency of the nested Kriging procedure (Rullière et al. 2018), which explicitly takes into account the correlations between the sub-model predictors, has been proved. The article also shed some light on this procedure, by showing that it provides an exact conditional distribution, for a different Gaussian process prior, and by obtaining bounds on the differences with the exact full Kriging model. Further results on the the efficient computation of conditional covariances have also been presented, which make possible sampling from the posterior distribution. The impact of the observation assignment to the sub-models has also been investigated, which resulted in some evidence that it is good practice to build them on clusters of observation points. Finally, the procedure of Rullière et al. (2018) has been extended to measurement errors and to universal Kriging, while retaining the same computational complexity and storage requirement.

Some perspectives remain open. It would be beneficial to improve the aggregation methods of Sect. 2, in order to guarantee their consistency while keeping their low computational costs. Finally, the interpretation of the predictor in Rullière et al. (2018) as an exact conditional expectation could be the basis of further asymptotic studies, as discussed in Sect. 3.2.

References

Abrahamsen, P. (1997). A review of Gaussian random fields and correlation functions. Technical report, Norwegian Computing Center.

Allard, D., Comunian, A., and Renard, P. (2012). Probability aggregation methods in geoscience. *Mathematical Geosciences*, 44(5):545–581.

Bacchi, V., Jomard, H., Scotti, O., Antoshchenkova, E., Bardet, L., Duluc, C.-M., and Hebert, H. (2020). Using meta-models for tsunami hazard analysis: An example of application for the French Atlantic coast. *Frontiers in Earth Science*, 8:41.

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

Bachoc, F., Ammar, K., and Martinez, J. (2016). Improvement of code behavior in a design of experiments by metamodeling. *Nuclear science and engineering*, 183(3):387–406.

Bachoc, F., Lagnoux, A., and Nguyen, T. M. N. (2017). Cross-validation estimation of covariance parameters under fixed-domain asymptotics. *Journal of Multivariate Analysis*, 160:42–67.

Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(4):825–848.

Cao, Y. and Fleet, D. J. (2014). Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions. In *Modern Nonparametrics 3: Automating the Learning Pipeline workshop at NIPS, Montreal*. arXiv preprint arXiv:1410.7827.

Chevalier, C. and Ginsbourger, D. (2013). Fast computation of the multi-points expected improvement with applications in batch selection. In *Learning and Intelligent Optimization*, pages 59–69. Springer.

Chiles, J.-P. and Delfiner, P. (2009). *Geostatistics: modeling spatial uncertainty*, volume 497. John Wiley & Sons.

Chilès, J.-P. and Desassis, N. (2018). Fifty years of Kriging. In *Handbook of mathematical geosciences*, pages 589–612. Springer, Cham.

Cressie, N. (1990). The origins of kriging. *Mathematical geology*, 22(3):239–252.

Cressie, N. (1993). *Statistics for spatial data*. J. Wiley.

Cressie, N. and Johannesson, G. (2008). Fixed rank Kriging for very large spatial data sets. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):209–226.

Datta, A., Banerjee, S., Finley, A. O., and Gelfand, A. E. (2016). Hierarchical nearest-neighbor Gaussian process models for large geostatistical datasets. *Journal of the American Statistical Association*, 111(514):800–812.

Davis, B. J. and Curriero, F. C. (2019). Development and evaluation of geostatistical methods for non-Euclidean-based spatial covariance matrices. *Mathematical Geosciences*, 51(6):767–791.

Deisenroth, M. P. and Ng, J. W. (2015). Distributed Gaussian processes. *Proceedings of the 32nd International Conference on Machine Learning, Lille, France. JMLR: W&CP volume 37*.

Finley, A. O., Sang, H., Banerjee, S., and Gelfand, A. E. (2009). Improving the performance of predictive process modeling for large datasets. *Computational statistics & data analysis*, 53(8):2873–2884.

Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*, 15(3):502–523.

He, J., Qi, J., and Ramamohanarao, K. (2019). Query-aware Bayesian committee machine for scalable Gaussian process regression. In *Proceedings of the 2019 SIAM International Conference on Data Mining*, pages 208–216. SIAM.

Heaton, M. J., Datta, A., Finley, A. O., Furrer, R., Guinness, J., Guhaniyogi, R., Gerber, F., Gramacy, R. B., Hammerling, D., and Katzfuss, M. (2019). A case study competition among methods for analyzing large spatial data. *Journal of Agricultural, Biological and Environmental Statistics*, 24(3):398–425.

Hensman, J. and Fusi, N. (2013). Gaussian processes for big data. *Uncertainty in Artificial Intelligence*, pages 282–290.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

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

Kaufman, C. G., Schervish, M. J., and Nychka, D. W. (2008). Covariance tapering for likelihood-based estimation in large spatial data sets. *Journal of the American Statistical Association*, 103(484):1545–1555.

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.

Krityakierne, T. and Baowan, D. (2020). Aggregated GP-based optimization for contaminant source localization. *Operations Research Perspectives*, 7:100151.

Liu, H., Cai, J., Wang, Y., and Ong, Y.-S. (2018). Generalized robust Bayesian committee machine for large-scale Gaussian process regression. *arXiv preprint arXiv:1806.00720*.

Liu, H., Ong, Y., Shen, X., and Cai, J. (2020). When Gaussian process meets big data: A review of scalable GPs. *IEEE Transactions on Neural Networks and Learning Systems*, pages 1–19.

Marrel, A., Iooss, B., Laurent, B., and Roustant, O. (2009). Calculations of Sobol indices for the Gaussian process metamodel. *Reliability Engineering & System Safety*, 94(3):742–751.

Matheron, G. (1970). *La Théorie des Variables Régionalisées et ses Applications.* Fasicule 5 in Les Cahiers du Centre de Morphologie Mathématique de Fontainebleau. Ecole Nationale Supérieure des Mines de Paris.

Putter, H., Young, G. A., et al. (2001). On the effect of covariance function estimation on the accuracy of Kriging predictors. *Bernoulli*, 7(3):421–438.

Quinonero-Candela, J. and Rasmussen, C. E. (2005). A unifying view of sparse approximate Gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959.

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

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).

Rue, H. and Held, L. (2005). *Gaussian Markov random fields, Theory and applications.* Chapman & Hall.

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(4):849–867.

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

Santner, T. J., Williams, B. J., and Notz, W. I. (2013). *The design and analysis of computer experiments.* Springer Science & Business Media.

Stein, M. L. (2012). *Interpolation of spatial data: some theory for Kriging.* Springer Science & Business Media.

Stein, M. L. (2014). Limitations on low rank approximations for covariance matrices of spatial data. *Spatial Statistics*, 8:1–19.

Sun, X., Luo, X.-S., Xu, J., Zhao, Z., Chen, Y., Wu, L., Chen, Q., and Zhang, D. (2019). Spatio-temporal variations and factors of a provincial pm 2.5 pollution in eastern china during 2013–2017 by geostatistics. *Scientific reports*, 9(1):1–10.

Tresp, V. (2000). A Bayesian committee machine. *Neural Computation*, 12(11):2719–2741.

van Stein, B., Wang, H., Kowalczyk, W., Bäck, T., and Emmerich, M. (2015). Optimally weighted cluster Kriging for big data regression. In *International Symposium on Intelligent Data Analysis*, pages 310–321. Springer.

Van Stein, B., Wang, H., Kowalczyk, W., Emmerich, M., and Bäck, T. (2020). Cluster-based Kriging approximation algorithms for complexity reduction. *Applied Intelligence*, 50(3):778–791.

Vazquez, E. and Bect, J. (2010a). Convergence properties of the expected improvement algorithm with fixed mean and covariance functions. *Journal of Statistical Planning and inference*, 140(11):3088–3095.

Vazquez, E. and Bect, J. (2010b). Pointwise consistency of the Kriging predictor with known mean and covariance functions. In *mODa 9 (Model-Oriented Data Analysis and Optimum Design) Springer*.

Ying, Z. (1991). Asymptotic properties of a maximum likelihood estimator with data from a Gaussian process. *Journal of Multivariate Analysis*, 36:280–296.

Zhang, H. and Wang, Y. (2010). Kriging and cross validation for massive spatial data. *Environmetrics*, 21:290–304.

Zhu, Z. and Zhang, H. (2006). Spatial sampling design under the infill asymptotic framework. *Environmetrics: The official journal of the International Environmetrics Society*, 17(4):323–337.

A Proof of Proposition 1

For $v \in \mathbb{R}^m$, we let $|v| = \max_{i=1,...,m} |v_i|$ and $B(v,r) = \{w \in \mathbb{R}^m, |v-w| \leq r\}$.

Let $x_0, \bar{x} \in D$, $r_{x_0} > 0$ and $r_{\bar{x}} > 0$ be fixed and satisfy $B(x_0, r_{x_0}) \subset D$, $B(\bar{x}, r_{\bar{x}}) \subset D$, $B(x_0, r_{x_0}) \cap B(\bar{x}, r_{\bar{x}}) = \varnothing$ and $k(x_0, \bar{x}) > 0$. [The existence is implied by the assumptions of the proposition.] By continuity of k, $r_{x_0} > 0$ and $r_{\bar{x}} > 0$ can be selected small enough so that, with some fixed $\epsilon_2 > 0$ and $\delta_1 > 0$, for $v \in B(x_0, r_{x_0})$ and $w \in B(\bar{x}, r_{\bar{x}})$, $|v-w| \geq \delta_1$, $k(x_0,x_0)/2 \leq k(v,v) \leq 2k(x_0,x_0)$, $k(\bar{x},\bar{x})/2 \leq k(w,w) \leq 2k(\bar{x},\bar{x})$ and

$$k(v,v) - \frac{k(v,w)^2}{k(w,w)} \leq k(v,v) - \epsilon_2. \tag{28}$$

For $\delta > 0$, let

$$V(\delta) = \inf_{n \in \mathbb{N}} \inf_{\substack{x_0, x_1, ..., x_n \in D; \\ \forall i = 1, ..., n, |x_i - x_0| \geq \delta}} \mathrm{V}\left[Y(x_0) | Y(x_1), ..., Y(x_n)\right].$$

Then $V(\delta) > 0$ because of the NEB, by continuity of k and by compacity.

Consider a decreasing sequence δ_n of non-negative numbers such that $\delta_n \to_{n\to\infty} 0$, and which will be specified below. There exists a sequence $(u_n)_{n\in\mathbb{N}} \in D^{\mathbb{N}}$, composed of pairwise distinct elements, such that $\lim_{n\to\infty} \sup_{x\in D} \min_{i=1,...,n} |u_i - x| = 0$, and such that for all n,

$$\inf_{\substack{1 \leq i,j \leq n \\ i \neq j \\ u_i, u_j \in B(x_0, r_{x_0})}} |u_i - u_j| \geq 4\delta_n.$$

Such a sequence indeed exists from Lemma 2 below.

Consider then the sequence $(w_n)_{n\in\mathbb{N}} \in D^{\mathbb{N}}$ such that for all n, $w_n = \bar{x} - (r_{\bar{x}}/(1+n))e_1$ with $e_1 = (1,0,...,0)$. We can assume furthermore that $\{u_n\}_{n\in\mathbb{N}}$ and $\{w_n\}_{n\in\mathbb{N}}$ are disjoint (this holds almost surely with the construction of Lemma 2 for (u_n)).

Let us now consider two sequences of integers p_n and k_n with $k_n \to \infty$ and $p_n \to \infty$ to be specified later. Let C_n be the largest natural number m satisfying $m(p_n - 1) < n$. Let $X = (X_1, ..., X_{p_n})$ be defined by, for $i = 1, ..., k_n$, $X_i = (u_j)_{j=(i-1)C_n+1,...,iC_n}$; for $i = k_n+1, ..., p_n - 1$, $X_i = (w_j)_{j=(i-k_n-1)C_n+1,...,(i-k_n)C_n}$; and $X_{p_n} = (w_j)_{j=(p_n-k_n-1)C_n+1,...,n-k_nC_n}$. With this construction, note that X_{p_n} is nonempty. Furthermore, the sequence of vectors $X = (X_1, ..., X_{p_n})$, indexed by $n \in \mathbb{N}$, defines a triangular array of observation points satisfying the conditions of the proposition.

Let us discuss the construction of $(u_n)_{n\in\mathbb{N}}$, $(w_n)_{n\in\mathbb{N}}$, k_n, C_n and p_n more informally. The sequence $(u_n)_{n\in\mathbb{N}}$ is dense in D, and $X_1, \ldots, X_{k_n}$ are composed by the k_nC_n first points of this sequence. Then, $X_{k_n+1}, \ldots, X_{p_n}$ are composed by the $n - C_nk_n$ first points of the sequence $(w_n)_{n\in\mathbb{N}}$, which is concentrated around $\bar{x}$. We will let $k_n/p_n \to 0$ so that the majority of the groups in X contain points of $(w_n)_{n\in\mathbb{N}}$, so that they do not contain relevant information on the values of Y on $B(x_0, r_{x_0})$ and yield an inconsistency of the aggregated predictor $M_{\mathcal{A},n}$ on $B(x_0, r_{x_0})$.

Coming back to the proof, observe that $\inf_{i\in\mathbb{N}} \inf_{x\in B(x_0,r_{x_0})} |w_i - x| \geq \delta_1$ and let $\epsilon_1 = V(\delta_1) > 0$. Then, we have for all $n \in \mathbb{N}$, for all $x \in B(x_0, r_{x_0})$, and for all $k = k_n + 1, ..., p_n$, since then X_k is nonempty and only contains elements $w_i \in B(\bar{x}, r)$, from (28),

$$\epsilon_1 \leq v_k(x) \leq k(x,x) - \epsilon_2. \tag{29}$$

Let $\mathcal{E}_n = \{x \in B(x_0, r_{x_0}); \min_{i=1,...,n} |x - u_i| \geq \delta_n\}$ and let $x \in \mathcal{E}_n$. Since x is not a component of X, we have $v_k(x) > 0$ for all k. Also $v_{p_n}(x) < k(x,x)$ from (29). Hence, $M_{\mathcal{A},n}(x)$ is well-defined.

For two random variables A and B, we let $||A - B|| = (\mathrm{E}\left[(A-B)^2\right])^{1/2}$. Let, for $x \in \mathcal{E}_n$,

$$R(x) \quad = \quad \left|\left| \sum_{k=1}^{k_n} \alpha_{k,n}(v_1(x), ..., v_{p_n}(x), v_{prior}(x)) M_k(x) \right|\right|.$$

Then, from the triangular inequality, and since, from the law of total variance, $||M_k(x)|| \leq ||Y(x)|| = v_{prior}(x)^{1/2}$ we have, with $\mathcal{V} = \{k(x,x); x \in B(x_0, r(x_0))\}$,

$$\begin{aligned} R(x) \quad &\leq \quad \frac{\sum_{k=1}^{k_n} a(v_k(x), v_{prior}(x)) \sqrt{v_{prior}(x)}}{\sum_{l=1}^{p_n} b(v_l(x), v_{prior}(x))} \\ &\leq \quad \frac{k_n \sup_{v\in\mathcal{V}, V(\delta_n) \leq s^2 \leq v} a(s^2, v) \sqrt{v}}{(p_n - k_n) \inf_{v\in\mathcal{V}, \epsilon_1 \leq s^2 \leq v - \epsilon_2} b(s^2, v)}, \end{aligned}$$

where the last inequality is obtained from (29) and the definition of δ_n and $V(\delta)$.

Let now for $\delta > 0$, $s(\delta) = \sup_{v\in\mathcal{V}, V(\delta)\leq s^2\leq v} a(s^2, v)$. Since a is continuous and since $V(\delta) > 0$, we have that $s(\delta)$ is finite. Hence, we can choose a sequence δ_n of positive numbers such that $\delta_n \to_{n\to\infty} 0$ and $s(\delta_n) \leq \sqrt{n}$ (for instance, let $\delta_n = \inf\{\delta \geq n^{-1/2}; s(\delta) \leq n^{1/2}\}$). Then, we can choose $p_n = n^{4/5}$ and $k_n = n^{1/5}$. Then, for n large enough

$$\frac{k_n}{p_n - k_n} s(\delta_n) \leq 2n^{-3/5}\sqrt{n} \to_{n\to\infty} 0.$$

Hence, since

$$\frac{\sup_{v\in\mathcal{V}} \sqrt{v}}{\inf_{v\in\mathcal{V}, \epsilon_1\leq s^2\leq v-\epsilon_2} b(s^2, v)}$$

is a finite constant, as b is positive and continuous on $\mathring{\Delta}$, we have that $\sup_{x\in\mathcal{E}_n} R(x) \to_{n\to\infty} 0$. As a consequence, we have from the triangular inequality, for $x \in \mathcal{E}_n$

$$\begin{aligned}
||Y(x) - M_{\mathcal{A},n}(x)|| \geq & ||Y(x) - \sum_{k=k_n+1}^{p_n} \alpha_{k,n}(v_1(x), ..., v_{p_n}(x), v_{prior}(x))M_k(x)|| \\
& - || \sum_{k=k_n+1}^{p_n} \alpha_{k,n}(v_1(x), ..., v_{p_n}(x), v_{prior}(x))M_k(x) - M_{\mathcal{A},n}(x)|| \\
\geq & \inf_{x\in\mathcal{E}_n} \left|\left| Y(x) - \sum_{k=k_n+1}^{p_n} \alpha_{k,n}(v_1(x), ..., v_{p_n}(x), v_{prior}(x))M_k(x) \right|\right| \\
& - \sup_{x\in\mathcal{E}_n} R(x).
\end{aligned}$$

Since $X_{k_n+1}, ..., X_{p_n}$ are composed only of elements of $\{w_i\}_{i\in\mathbb{N}}$, we obtain

$$\liminf_{n\to\infty} \inf_{x\in\mathcal{E}_n} ||Y(x) - M_{\mathcal{A},n}(x)|| \geq V(\delta_1) > 0.$$

Hence, there exist fixed $n_0 \in \mathbb{N}$ and $A > 0$ so that for $n \geq n_0$, $||Y(x) - M_{\mathcal{A},n}(x)|| \geq A$. Hence, we have, for $n \geq n_0$

$$\begin{aligned}
\int_D \mathrm{E}\left[(Y(x) - M_{\mathcal{A},n}(x))^2\right] dx &\geq \int_{\mathcal{E}_n} \mathrm{E}\left[(Y(x) - M_{\mathcal{A},n}(x))^2\right] \\
&\geq \int_{\mathcal{E}_n} A^2 dx.
\end{aligned}$$

Hence, it remains to show that the limit inferior of the volume of $\mathcal{E}_n$ is not zero in order to show (4). Let N_n be the integer part of $r_{x_0}/4\delta_n$. Then, the ball $B(x_0, r_{x_0})$ contains $(2N_n)^d$ disjoint balls of the form $B(a, 4\delta_n)$ with $a \in B(x_0, r_{x_0})$. If one of these balls $B(a, 4\delta_n)$ does not intersect with $(u_i)_{i=1...,n}$, then we can associate to it a ball of the form $B(s_a, \delta_n) \subset B(a, 4\delta_n) \cap \mathcal{E}_n$. If one of these balls $B(a, 4\delta_n)$ does intersect with one $u_j \in \{u_i\}_{i=1...,n}$, then we can find a ball $B(s_a, \delta_n/2) \subset (B(u_j, 2\delta_n)\backslash B(u_j, \delta_n)) \cap B(a, 4\delta_n) \cap \mathcal{E}_n$. Hence, we have found $(2N_n)^d$ disjoint balls with radius $\delta_n/2$ in $\mathcal{E}_n$. Hence $\mathcal{E}_n$ has volume at least $2^d((r_{x_0}/4\delta_n) - 1)^d\delta_n^d$ which has a strictly positive limit inferior. Hence, (4) is proved.

Finally, if $\mathrm{E}\left[(Y(x_0) - M_{\mathcal{A},n}(x_0))^2\right] \to 0$ as $n \to \infty$ for almost all $x_0 \in D$, then

$$\int_D \max\left(\mathrm{E}\left[(Y(x_0) - M_{\mathcal{A},n}(x))^2\right], 1\right) dx \to_{n\to\infty} 0$$

from the dominated convergence theorem. This is contradictory with the proof of (4). Hence, (5) is proved.

Lemma 2. *There exists a sequence $(u_n)_{n\in\mathbb{N}} \in D^{\mathbb{N}}$, composed of pairwise distinct elements, such that*

$$\lim_{n\to\infty} \sup_{x\in D} \min_{i=1,...,n} |u_i - x| = 0, \tag{30}$$

and such that for all n,

$$\inf_{\substack{1\leq i,j\leq n \\ i\neq j \\ u_i, u_j \in B(x_0, r_{x_0})}} |u_i - u_j| \geq 4\delta_n. \tag{31}$$

Proof. Such a sequence can be constructed, for instance, by the following random procedure. Let $D \subset B(0,R)$ for $R>0$ large enough. Define $u_1 \in D$ arbitrarily. For $n=1,2,\ldots$: (1) if the set $\mathcal{S}_n = \{u \in B(x_0, r_{x_0}); \min_{i=1,\ldots,n} |u - u_i| > 4\delta_{n+1}\}$ is non-empty, sample u_{n+1} from the uniform distribution on $\mathcal{S}_n$. (2) If $\mathcal{S}_n$ is empty, sample $\tilde{u}_{n+1}$ from the uniform distribution on $B(0,R)\backslash B(x_0, r_{x_0})$, and set u_{n+1} as the projection of $\tilde{u}_{n+1}$ on $D\backslash B(x_0, r_{x_0})$. One can see that (31) is satisfied by definition. Furthermore, one can show that (30) holds almost surely. Indeed, let $x \in B(x_0, r_{x_0})$ and $\epsilon > 0$, and assume that with non-zero probability $B(x,\epsilon) \cap \{u_i\}_{i\in\mathbb{N}} = \varnothing$. Then, the case (1) occurs infinitely often and, for each i for which the case (1) occurs, there is a probability at least $\epsilon^d/(2r_{x_0})^d$ that $u_i \in B(x,\epsilon)$ (when $4\delta_n \leq \epsilon/2$). This yields a contradiction. Hence, for all $x \in B(x_0, r_{x_0})$ and $\epsilon > 0$, almost surely, $B(x,\epsilon) \cap \{u_i\}_{i\in\mathbb{N}} \neq \varnothing$. We show similarly, for all $x \in D\backslash B(x_0, r_{x_0})$ and $\epsilon > 0$, almost surely, $B(x,\epsilon) \cap \{u_i\}_{i\in\mathbb{N}} \neq \varnothing$. This show that (30) holds almost surely. Hence, a fortiori, there exists a sequence $(u_n)_{n\in\mathbb{N}} \in D^{\mathbb{N}}$ satisfying the conditions of the lemma. □

Remark 6. *Consider the case $d=1$. The proof of Proposition 1 can be modified so that the partition $X_1,\ldots,X_{p_n}$ also satisfies $x \leq x'$ for any $x \in X_i$, $x' \in X_j$, $1 \leq i < j \leq p_n$. To see this, consider the same X as in this proof. Let $X_1,\ldots,X_{p_n}$ have the same cardinality as in this proof, and let the C_n smallest elements of X be affected to X_1, the next C_n smallest be affected to X_2 and so on. Then, one can show that there are at most k_n+2 groups containing elements of $(u_i)_{i\in\mathbb{N}} \cap B(x_0, r_{x_0})$ and at least $p_n - k_n - 2$ groups containing only elements of $B(\bar{x}, r_{\bar{x}})$. From these observations,* (4) *and* (5) *can be proved similarly as in the proof of Proposition 1.*

B Proof of Proposition 2

Because D is compact we have $\lim_{n\to\infty} \sup_{x\in D} \min_{i=1,\ldots,n} ||x_{ni} - x|| = 0$. Indeed, if this does not hold, there exists $\epsilon > 0$ and a subsequence $\phi(n)$ such that $\sup_{x\in D} \min_{i=1,\ldots,\phi(n)} ||x_{\phi(n)i} - x|| \geq 2\epsilon$. Hence, there exists a sequence, $x_{\phi(n)} \in D$ such that $\min_{i=1,\ldots,\phi(n)} ||x_{\phi(n)i} - x_{\phi(n)}|| \geq \epsilon$. Since D is compact, up to extracting a further subsequence, we can also assume that $x_{\phi(n)} \to_{n\to\infty} x_{lim}$ with $x_{lim} \in D$. This implies that for all n large enough, $\min_{i=1,\ldots,\phi(n)} ||x_{\phi(n)i} - x_{lim}|| \geq \epsilon/2$, which is in contradiction with the assumptions of the proposition.

Hence there exists a sequence of positive numbers δ_n such that $\delta_n \to_{n\to\infty} 0$ and such that for all $x \in D$ there exists a sequence of indices $i_n(x)$ such that $i_n(x) \in \{1,...,n\}$ and $||x - x_{ni_n(x)}|| \leq \delta_n$. There also exists a sequence of indices $j_n(x)$ such that $x_{ni_n(x)}$ is a component of $X_{j_n(x)}$. With these notations we have, since $M_1(x)$,..., $M_{p_n}(x)$, $M_{\mathcal{A}}(x)$ are linear combinations with minimal square prediction errors,

$$\begin{aligned} \sup_{x\in D} \mathrm{E}\left[\left(Y(x) - M_{\mathcal{A}}(x)\right)^2\right] &\leq \sup_{x\in D} \mathrm{E}\left[\left(Y(x) - M_{j_n(x)}(x)\right)^2\right] \\ &\leq \sup_{x\in D} \mathrm{E}\left[\left(Y(x) - \mathrm{E}\left[Y(x)|Y(x_{ni_n(x)})\right]\right)^2\right]. \end{aligned} \tag{32}$$

In the rest of the proof we essentially show that, for a dense triangular array of observation points, the Kriging predictor that predicts $Y(x)$ based only on the nearest neighbor of x among the observation points has a mean square prediction error that goes to zero uniformly in x when k is continuous. We believe that this fact is somehow known, but we have not been able to find a precise result in the literature. We have from (32),

$$\begin{aligned} &\sup_{x\in D} \mathrm{E}\left[\left(Y(x) - M_{\mathcal{A}}(x)\right)^2\right] \\ &\leq \sup_{x\in D} \left[1\{k(x_{ni_n(x)}, x_{ni_n(x)}) = 0\} k(x,x) + 1\{k(x_{ni_n(x)}, x_{ni_n(x)}) > 0\} \left(k(x,x) - \frac{k(x, x_{ni_n(x)})^2}{k(x_{ni_n(x)}, x_{ni_n(x)})}\right)\right] \\ &\leq \sup_{\substack{x,t\in D; \\ ||x-t||\leq\delta_n}} \left[1\{k(t,t) = 0\} k(x,x) + 1\{k(t,t) > 0\} \left(k(x,x) - \frac{k(x,t)^2}{k(t,t)}\right)\right] \\ &= \sup_{\substack{x,t\in D; \\ ||x-t||\leq\delta_n}} F(x,t). \end{aligned}$$

Assume now that the above supremum does not go to zero as $n \to \infty$. Then there exists $\epsilon > 0$ and two sub-sequences $x_{\phi(n)}$ and $t_{\phi(n)}$ with values in D such that $x_{\phi(n)} \to_{n\to\infty} x_{lim}$ and $t_{\phi(n)} \to_{n\to\infty} x_{lim}$, with $x_{lim} \in D$ and such that $F(x_{\phi(n)}, t_{\phi(n)}) \geq \epsilon$. If $k(x_{lim}, x_{lim}) = 0$ then $F(x_{\phi(n)}, t_{\phi(n)}) \leq k(x_{\phi(n)}, x_{\phi(n)}) \to_{n\to\infty} 0$. If $k(x_{lim}, x_{lim}) > 0$ then for n large enough

$$F(x_{\phi(n)}, t_{\phi(n)}) = k(x_{\phi(n)}, x_{\phi(n)}) - \frac{k(x_{\phi(n)}, t_{\phi(n)})^2}{k(t_{\phi(n)}, t_{\phi(n)})}$$

which goes to zero as $n \to \infty$ since k is continuous. Hence we have a contradiction, which completes the proof.

C Proofs in Sect. 3.2

First notice that denoting $k_{\mathcal{A}}(x,x') = \mathrm{Cov}\left[Y_{\mathcal{A}}(x), Y_{\mathcal{A}}(x')\right]$, we easily get for all $x, x' \in D$,

$$\begin{aligned} k_{\mathcal{A}}(x,x') = k(x,x') &+ 2k_M(x)^t K_M^{-1}(x) K_M(x,x') K_M^{-1}(x') k_M(x') \\ &- k_M(x)^t K_M^{-1}(x) k_M(x,x') - k_M(x')^t K_M^{-1}(x') k_M(x',x). \end{aligned} \tag{33}$$

A direct consequence of (33) is $k_{\mathcal{A}}(x,x) = k(x,x)$, and under the interpolation assumption (H2), since $Y_{\mathcal{A}}(X) = Y(X)$, $k_{\mathcal{A}}(X,X) = k(X,X)$.

Proof of Proposition 3. The interpolation hypothesis $M_{\mathcal{A}}(X) = Y(X)$ ensures $\varepsilon'_{\mathcal{A}}(X) = 0$ so we have

$$\begin{aligned} \mathrm{E}\left[Y_{\mathcal{A}}(x)|Y_{\mathcal{A}}(X)\right] &= \mathrm{E}\left[Y_{\mathcal{A}}(x)|M_{\mathcal{A}}(X) + 0\right] \\ &= \mathrm{E}\left[M_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] + \mathrm{E}\left[\varepsilon'_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] \\ &= \mathrm{E}\left[g_x(Y(X))|Y(X)\right] + 0 \\ &= M_{\mathcal{A}}(x). \end{aligned} \tag{34}$$

The proof that $v_{\mathcal{A}}$ is a conditional variance follows the same pattern:

$$\begin{aligned} \mathrm{V}\left[Y_{\mathcal{A}}(x)|Y_{\mathcal{A}}(X)\right] &= \mathrm{V}\left[Y_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] \\ &= \mathrm{V}\left[M_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] + \mathrm{V}\left[\varepsilon'_{\mathcal{A}}(x)\right] \\ &= v_{\mathcal{A}}(x). \end{aligned} \tag{35}$$

□

Proof of Proposition 4. Eq. (11) is the classical expression of Gaussian conditional covariances, based on the fact that $Y_{\mathcal{A}}$ is Gaussian. Let us now prove Eq. (12). For a component x_k of the vector of points X, using the interpolation assumption, we have $M_{\mathcal{A}}(x_k) = Y(x_k)$ and

$$\mathrm{Cov}\left[Y_{\mathcal{A}}(x), Y_{\mathcal{A}}(x_k)\right] = \mathrm{Cov}\left[M_{\mathcal{A}}(x) + \varepsilon'_{\mathcal{A}}(x), M_{\mathcal{A}}(x_k)\right] = \mathrm{Cov}\left[M_{\mathcal{A}}(x), Y(x_k)\right].$$

Remark that $\alpha_{\mathcal{A}}(x)$ is the $p \times 1$ vector of aggregation weights of different sub-models at point x, so that $M_{\mathcal{A}}(x) = \alpha_{\mathcal{A}}(x)^t M(x)$ and $k_{\mathcal{A}}(x,x_k) = \alpha_{\mathcal{A}}(x)^t \mathrm{Cov}\left[M(x), Y(x_k)\right]$. We thus get

$$k_{\mathcal{A}}(x,X) \quad = \quad \alpha_{\mathcal{A}}(x)^t \mathrm{Cov}\left[M(x), Y(X)\right]. \tag{36}$$

Under the linearity assumption, there exists a $p \times n$ deterministic matrix $\Lambda(x)$ such that $M(x) = \Lambda(x)Y(X)$. Thus $k_{\mathcal{A}}(x,X) = \alpha_{\mathcal{A}}(x)^t \Lambda(x) k(X,X)$. As remarked in Sect. 3, because of the interpolation condition, $k_{\mathcal{A}}(X,X) = k(X,X)$ and

$$k_{\mathcal{A}}(x,X) k_{\mathcal{A}}(X,X)^{-1} k_{\mathcal{A}}(X,x') = \alpha_{\mathcal{A}}(x)^t \Lambda(x) k(X,X) \Lambda(x')^t \alpha_{\mathcal{A}}(x'). \tag{37}$$

Using $K_M(x,x') = \mathrm{Cov}\left[M(x), M(x')\right] = \Lambda(x) k(X,X) \Lambda(x')^t$, we get

$$k_{\mathcal{A}}(x,X) k_{\mathcal{A}}(X,X)^{-1} k_{\mathcal{A}}(X,x') = \alpha_{\mathcal{A}}(x)^t K_M(x,x') \alpha_{\mathcal{A}}(x'). \tag{38}$$

At last, starting from Eq. (11) and using both Eqs (33) and (38), we get Eq. (12).
Finally, the development of $\mathrm{E}\left[\left(Y(x) - M_{\mathcal{A}}(x)\right)\left(Y(x') - M_{\mathcal{A}}(x')\right)\right]$ leads to the right hand side of Eq. (12) so that

$$\mathrm{E}\left[\left(Y(x) - M_{\mathcal{A}}(x)\right)\left(Y(x') - M_{\mathcal{A}}(x')\right)\right] = c_{\mathcal{A}}(x,x')$$

and Eq. (13) holds. □

D Proofs in Sect. 3.3

Proof of Proposition 5. Consider $\Delta(x)$ as defined in Eq. (14). From Eq. (36), using both the linear and the interpolation assumptions, we get $k(x,X)\Delta(x) = [k(x,X) - k_{\mathcal{A}}(x,X)]\, k(X,X)^{-1}$. Injecting this result in Eq. (14), we have

$$M_{\mathcal{A}}(x) - M_{full}(x) = [k_{\mathcal{A}}(x,X) - k(x,X)]k(X,X)^{-1}Y(X) \tag{39}$$

and the first equality holds. From (14), we also get $v_{\mathcal{A}}(x) - v_{full}(x) = k(x,X)k(X,X)^{-1}k(X,x) - k_{\mathcal{A}}(x,X)k(X,X)^{-1}k_{\mathcal{A}}(X,x)$ and the second equality holds. Note that under the same assumptions, we can also use $k_{\mathcal{A}}(X,X) = k(X,X)$ and $k_{\mathcal{A}}(x,x) = k(x,x)$ and start from $M_{\mathcal{A}} = k_{\mathcal{A}}(x,X)k_{\mathcal{A}}(X,X)^{-1}Y(X)$ and $v_{\mathcal{A}}(x) = k_{\mathcal{A}}(x,x) - k_{\mathcal{A}}(x,X)k_{\mathcal{A}}(X,X)^{-1}k_{\mathcal{A}}(X,x)$ to get the same results.

Let us now show Eq. (17). The upper bound comes from the fact that $M_{\mathcal{A}}(x)$ is the best linear combination of $M_k(x)$ for $k \in \{1,\dots,p\}$. The positivity of $v_{\mathcal{A}} - v_{full}$ can be proved similarly: $M_{\mathcal{A}}(x)$ is a linear combination of $Y(x_k)$, $k \in \{1,\dots,n\}$, whereas $M_{full}(x)$ is the best linear combination. Notice that $v_{\mathcal{A}}(x) - v_{full}(x) \geq 0$ implies, using Eq. (15), that $\|k_{\mathcal{A}}(X,x)\|_K \leq \|k(X,x)\|_K$. Let us show Eq. (16). We get the result starting from Eq. (39), applying Cauchy-Schwartz inequality. The bound on $v_{\mathcal{A}}(x) - v_{full}(x)$ directly derives from Eq. (15), using $\|k_{\mathcal{A}}(X,x)\|_K \leq \|k(X,x)\|_K$.

Finally, the classical inequality between $\|.\|_K$ and $\|.\|$ derives from the diagonalization of $k(X,X)$, one can notice that it depends on n and X, but it does not depend on the prediction point x. □

Proof of Remark 4. Using $\|k_{\mathcal{A}}(X,x)\|_K \leq \|k(X,x)\|_K$, using the equivalence of norms and triangular inequality, assuming that the smallest eigenvalue $\lambda_{\min}$ of $k(X,X)$ is non zero, bounds (16) in the previous Proposition 5 implies that

$$\begin{cases} |M_{\mathcal{A}}(x) - M_{full}(x)| & \leq & \frac{2}{\lambda_{\min}}\|k(X,x)\|\|Y(X)\|\,, \\ |v_{\mathcal{A}}(x) - v_{full}(x)| & \leq & \frac{1}{\lambda_{\min}}\|k(X,x)\|^2\,. \end{cases} \tag{40}$$

Noticing that the $\|.\|_K$ and $\lambda_{\min}$ do not depend on x (although they depend on X and n), the result holds. □

Proof of Remark 5. As $\Lambda(x)$ is $n \times n$ and invertible, we have

$$k_M(x)^t K_M(x)^{-1} M(x) = k(x,X)^t \Lambda(x)^t (\Lambda(x)k(X,X)\Lambda(x)^t)^{-1}\Lambda(x)Y(x) = M_{full}(x),$$

and similarly $v_{\mathcal{A}}(x) = v_{full}(x)$. As $M_{\mathcal{A}} = M_{full}$, we have $Y_{\mathcal{A}} = M_{full} + \varepsilon$ where ε is an independent copy of $Y - M_{full}$. Furthermore $Y = M_{full} + Y - M_{full}$ where M_{full} and $Y - M_{full}$ are independent, by Gaussianity, so $Y_{\mathcal{A}} \overset{law}{=} Y$. □

E Proofs in Sect. 5

Proof of Proposition 8. Because $M_{\mathcal{A},\eta}(x)$ is the best linear predictor of $Y(x)$, for $n \in \mathbb{N}$, we have

$$\mathbb{E}\left[(Y(x) - M_{\mathcal{A},\eta}(x))^2\right] \leq \mathbb{E}\left[(Y(x) - M_{\eta,i_n}(x))^2\right]. \tag{41}$$

Let $\epsilon > 0$. Let N_n be the number of points in X_{i_n} that are at Euclidean distance less than ϵ from x. By assumption, $N_n \to \infty$ as $n \to \infty$. Let us write these points as $x_{nj_1},\dots,x_{nj_{N_n}}$, with corresponding measurement errors $\xi_{j_1},\dots,\xi_{j_{N_n}}$. Since $M_{\eta,i_n}(x)$ is the best linear unbiased predictor of $Y(x)$ from the elements of $Y(x_{nj_1}) + \xi_{j_1},\dots,Y(x_{nj_{N_n}}) + \xi_{j_{N_n}}$, we have

$$\mathbb{E}\left[(Y(x) - M_{\eta,i_n}(x))^2\right] \leq \mathbb{E}\left[\left(Y(x) - \frac{1}{N_n}\sum_{a=1}^{N_n}(Y(x_{nj_a}) + \xi_{j_a})\right)^2\right]. \tag{42}$$

By independence of Y and ξ_X, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(Y(x) - \frac{1}{N_n}\sum_{a=1}^{N_n}(Y(x_{nj_a}) + \xi_{j_a})\right)^2\right] &= \mathbb{E}\left[\left(\frac{1}{N_n}\sum_{a=1}^{N_n}(Y(x) - Y(x_{nj_a}))\right)^2\right] + \mathbb{E}\left[\left(\frac{1}{N_n}\sum_{a=1}^{N_n}\xi_{j_a}\right)^2\right] \\ &\leq \left(\max_{a=1,\dots,N_n}\mathbb{E}\left[(Y(x) - Y(x_{nj_a}))^2\right]\right) + \frac{\sum_{a=1}^{N_n}\eta_a}{N_n^2}.\end{aligned}$$

The above inequality follows from Cauchy-Schwarz, the fact that Y has mean zero and the independence of $\xi_{j_1}, \ldots, \xi_{j_{N_n}}$. We then obtain, since $(\eta_a)_{a\in\mathbb{N}}$ is bounded,

$$\limsup_{n\to\infty} \mathbb{E}\left[\left(Y(x) - \frac{1}{N_n}\sum_{a=1}^{N_n}(Y(x_{nj_a}) + \xi_{j_a})\right)^2\right] \leq \sup_{\substack{u\in D \\ ||u-x||\leq\epsilon}} \mathbb{E}\left[(Y(x) - Y(u))^2\right]$$
$$= \sup_{\substack{u\in D \\ ||u-x||\leq\epsilon}} \left(k(x,x) + k(u,u) - 2k(x,u)\right).$$

From (41) and (42), we have, for any $\epsilon > 0$,

$$\limsup_{n\to\infty} \mathbb{E}\left[\left(Y(x) - M_{\mathcal{A},\eta}(x)\right)^2\right] \leq \sup_{\substack{u\in D \\ ||u-x||\leq\epsilon}} \left(k(x,x) + k(u,u) - 2k(x,u)\right). \tag{43}$$

The above display goes to zero as $\epsilon \to 0$ because k is continuous. Hence the lim sup in (43) is zero, which concludes the proof.

□

Proof of Lemma 1. Let $\epsilon > 0$. For $n \in \mathbb{N}$, let N_n be the number of points in $\{x_1, \ldots, x_n\}$ that are at Euclidean distance less than ϵ to x. Because x is in the interior of D and because $g > 0$ on D, we have $p_\epsilon = \mathbb{P}(||x_1 - x|| \leq \epsilon) > 0$. Hence from the law of large number, almost surely, for n large enough, $N_n \geq (p_\epsilon/2)n$. For each $n \in \mathbb{N}$, the N_n points in $\{x_1, \ldots, x_n\}$ that are at Euclidean distance less than ϵ to x are partitioned into p_n classes. Hence, one of these classes, say the class X_{i_n}, contains a number of points larger or equal to N_n/p_n. Since n/p_n goes to infinity by assumption, we conclude that the number of points in X_{i_n} at distance less than ϵ from x goes to infinity, almost surely. This concludes the proof. □

Proof of Proposition 9. The proof is based on the same construction of the triangular array of observation points and of the sequence of partitions as in the proof of Proposition 1. We take x as x_0 in this proof. Only a few comments are needed.

We let $V(\delta)$ be as in the proof of Proposition 1 and we remark that for any $\delta > 0$, for any $r \in \mathbb{N}$, for any Gaussian vector $(U_1, \ldots, U_r)$ independent of Y and for any $u_0, u_1, \ldots, u_r \in D$ with $||u_i - u_0|| \geq \delta$ for $i = 1, \ldots, r$, we have

$$V[Y(u_0)|Y(u_1) + U_1, \ldots, Y(u_r) + U_r] \geq V[Y(u_0)|Y(u_1), U_1, \ldots, Y(u_r), U_r] = V[Y(u_0)|Y(u_1), \ldots, Y(u_r)] \geq V(\delta).$$

We also remark that the triangular array and sequence of partitions of the proof of Proposition 1 do satisfy the condition of Proposition 8. Indeed, the first component X_1 of the partition, with cardinality $C_n \to \infty$, is dense in D.

We remark that for $k = k_n + 1, \ldots, p_n$ (notations of the proof of Proposition 1), for any row of X_k, of the form x_{nb} with $b \in \{1, \ldots, n\}$, we have $v_k(x) \leq V[Y(x)|Y(x_{nb}) + \xi_b] \leq k(x,x) - k(x,x_{nb})^2/(k(x_{nb},x_{nb}) + \eta_b)$. Hence, because $(\eta_i)_{i\in\mathbb{N}}$ is bounded, there is a fixed $\epsilon_2' > 0$ such that for $k = k_n + 1, \ldots, p_n$, $\epsilon_1 \leq v_k(x) \leq k(x,x) - \epsilon_2'$, with ϵ_1 as in the proof of Proposition 1.

With these comments, the arguments of the proof of Proposition 1 lead to the conclusion of Proposition 9. □

Proof of Proposition 10. We can see that $M_{\mathrm{UK},i}(x) = w_i(x)^t Z(X_i)$ for $i = 1, \ldots, p$. Hence, for $i, j = 1, \ldots, p$,

$$\mathrm{Cov}\left[M_{\mathrm{UK},i}(x), M_{\mathrm{UK},j}(x)\right] = w_i(x)^t \mathrm{Cov}\left[Z(X_i), Z(X_j)\right] w_j(x) = w_i(x)^t k(X_i, X_j) w_j(x).$$

Hence, $\mathrm{Cov}\left[M_{\mathrm{UK}}(x)\right] = K_{\mathrm{UK},M}(x)$. Furthermore, for $i = 1, \ldots, p$,

$$\mathrm{Cov}\left[M_{\mathrm{UK},i}(x), Z(x)\right] = w_i(x)^t \mathrm{Cov}\left[Z(X_i), Z(x)\right] = w_i(x)^t k(X_i, x).$$

Hence, $\mathrm{Cov}\left[M_{\mathrm{UK}}(x), Z(x)\right] = k_{\mathrm{UK},M}(x)$. Let

$$\alpha(x) = \underset{\substack{\gamma\in\mathbb{R}^p \\ \mathbb{E}[\gamma^t M_{\mathrm{UK}}(x)] = \mathbb{E}[Z(x)] \\ \text{for any value of } \beta \text{ in } \mathbb{R}^m}}{\operatorname{argmin}} \mathbb{E}\left[\left(\gamma^t M_{\mathrm{UK}}(x) - Z(x)\right)^2\right]. \tag{44}$$

Since $\mathbb{E}[M_{\mathrm{UK},i}(x)] = \mathbb{E}[Z(x)]$ for $i = 1, \ldots, p$ and for any value of $\beta \in \mathbb{R}^m$, the constraint in (44) can be written as $\gamma^t 1_p \mathbb{E}[Z(x)] = \mathbb{E}[Z(x)]$ that is $\gamma^t 1_p = 1$. The mean square prediction error in (44) can be written as

$$k(x,x) + \gamma^t K_{\mathrm{UK},M}(x)\gamma - 2\gamma^t k_{\mathrm{UK},M}(x).$$

Thus (44) becomes

$$\alpha(x) = \underset{\substack{\gamma \in \mathbb{R}^p \\ \gamma^t 1_p = 1}}{\operatorname{argmin}} \left(k(x,x) + \gamma^t K_{\mathrm{UK},M}(x)\gamma - 2\gamma^t k_{\mathrm{UK},M}(x) \right).$$

We recognize the optimization problem of ordinary Kriging which corresponds to universal Kriging with an unknown constant mean function (Sacks et al. 1989, Chiles and Delfiner 2009). Hence, we have

$$\alpha(x)^t M_{\mathrm{UK}}(x) = \hat{m}_{\mathrm{UK},M}(x) + k_{\mathrm{UK},M}(x)^t K_{\mathrm{UK},M}(x)^{-1} \left(M_{\mathrm{UK}}(x) - \hat{m}_{\mathrm{UK},M}(x) 1_p \right),$$

from for instance Sacks et al. (1989), Chiles and Delfiner (2009). Hence we have $\alpha(x)^t M_{\mathrm{UK}}(x) = M_{\mathcal{A},\mathrm{UK}}(x)$, the best linear predictor described in Proposition 10.

We can see that $M_{\mathcal{A},\mathrm{UK}}(x) = \alpha_{\mathcal{A},\mathrm{UK}}(x)^t M_{\mathrm{UK}}(x)$ and that $\alpha_{\mathcal{A},\mathrm{UK}}(x) = \alpha(x)$. Then since $\mathbb{E}[\alpha_{\mathcal{A},\mathrm{UK}}(x)^t M_{\mathrm{UK}}(x)] = Z(x)$, from $\mathrm{Cov}\left[M_{\mathrm{UK}}(x)\right] = K_{\mathrm{UK},M}(x)$ and from $\mathrm{Cov}\left[M_{\mathrm{UK}}(x), Z(x)\right] = k_{\mathrm{UK},M}(x)$, we obtain

$$\mathbb{E}\left[\left(M_{\mathcal{A},\mathrm{UK}}(x) - Z(x) \right)^2 \right] = k(x,x) + \alpha_{\mathcal{A},\mathrm{UK}}(x)^t K_{\mathrm{UK},M}(x) \alpha_{\mathcal{A},\mathrm{UK}}(x) - 2\alpha_{\mathcal{A},\mathrm{UK}}(x)^t k_{\mathrm{UK},M}(x).$$

This concludes the proof. □