Some properties of nested Kriging predictors

François Bachoc, Nicolas Durrande, Didier Rullière, Clément Chevalier

Some properties of nested Kriging predictors

François Bachoc[*], Nicolas Durrande[†], Didier Rullière[‡] and Clément Chevalier[§]

Thursday 13th July, 2017

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. We focus on a class of methods aiming at decreasing this computational cost, consisting in aggregating Kriging predictors based on smaller data subsets. We prove that aggregations based solely on the conditional variances provided by the different Kriging predictors can yield an inconsistent final Kriging prediction. In contrasts, we study theoretically the recent proposal by [Rullière et al., 2017] and obtain additional attractive properties for it. We prove that this predictor is consistent, we show that it can be interpreted as an exact conditional distribution for a modified process and we provide error bounds for it.

1 Introduction

Kriging [Stein, 2012, Santner et al., 2013, Williams and Rasmussen, 2006] 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., 1016], global optimization [Jones et al., 1998] or machine learning.

We let Y be a centered Gaussian process on $D \subset \mathbb{R}^d$ with covariance function $k : D \times D \to \mathbb{R}$, where $k(x, x') = \mathrm{Cov}\left[Y(x), Y(x')\right]$. We consider Y to be observed at n observation points $x_1, ..., x_n \in D$. We let X be the $n \times d$ matrix with line 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 $i = 1, ..., m$, we denote by $f(A)$ the $n \times 1$ real valued vector with components $f(a_i)$ and by $g(A, B)$ the $n \times m$ real valued matrix with 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:

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

In (1), one observes that, in order to compute the exact conditional distribution of Y, it is required to invert the $n \times n$ covariance matrix $k(X, X)$, which lead to a $O(n^2)$ complexity in space and $O(n^3)$ in time. In practice, this exact distribution is hence difficult to compute when the number of observation points is in the range $[10^3, 10^4]$ or greater.

Many methods have been proposed in the literature to approximate the conditional distribution (1), with a smaller computational cost. These methods include low rank approximation (see [Stein, 2014]

[*]Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118 route de Narbonne, 31062 Toulouse, France, francois.bachoc@math.univ-toulouse.fr.

[†]Institut Fayol—LIMOS, Mines Saint-Étienne, 158 cours Fauriel, Saint-Étienne, France, durrande@emse.fr.

[‡]Université de Lyon, Université Claude Bernard Lyon 1, ISFA, Laboratoire SAF, EA2429, 50 avenue Tony Garnier, 69366 Lyon, France, didier.rulliere@univ-lyon1.fr.

[§]Institute of Statistics, University of Neuchâtel, avenue de Bellevaux 51, 2000 Neuchâtel, Switzerland, clement.chevalier@unine.ch.

and the references therein for a review), sparse approximation [Hensman et al., 2013], covariance tapering [Furrer et al., 2006, Kaufman et al., 2008], Gaussian Markov Random Fields approximation [Rue and Held, 2005, Datta et al., 2016].

In this paper, we focus on aggregation-based approximations of $M_{full}(x)$ in (1). The principle of these methods is to first construct p submodels $M_1, ..., M_p : D \to \mathbb{R}$, 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 with a small computational cost. Then, the submodels $M_1, ..., M_p$ are 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 recent proposal [Rullière et al., 2017].

In this paper, we first consider a large class of methods, including products of experts and Bayesian committee machines, for which the predictors $M_1(x), ..., M_p(x)$ are aggregated based on their conditional variances $v_1(x), ..., v_p(x)$. We prove that the aggregated predictor $M_{\mathcal{A}}(x)$ can be inconsistent for predicting $Y(x)$. The interpretation is that these methods neglect the correlation between the predictors $M_1(x), ..., M_p(x)$.

Then, we address the aggregated predictor recently proposed by [Rullière et al., 2017], where these correlations are explicitly taken into account. In [Rullière et al., 2017], it is shown that this predictor is a drastic improvement of products of experts and Bayesian committee machines, and remain computationally tractable for data sets of size 10^6. We provide additional theoretical insights for the aggregated predictor in [Rullière et al., 2017]. We show that this predictor is always consistent. In addition, we prove that it can be interpreted as an exact conditional expectation, for a slightly different Gaussian process. Finally, we show several bounds on the difference between the aggregated predictor $M_{\mathcal{A}}(x)$ and the exact predictor $M_{full}(x)$.

The rest of the paper is organized as follows. In Section 2, we introduce aggregation techniques based solely on the conditional variances, and present the non-consistency result. In Section 3 we introduce the aggregation method of [Rullière et al., 2017], give its consistency property, show how it can be interpreted as an exact conditional expectation and provide the bounds discussed above. Concluding remarks are given in Section 4. The proofs of the consistency and non-consistency results are postponed to the appendix.

2 Aggregation techniques based solely on conditional variances

For $i = 1, ..., p$, let X_i be a $n_i \times d$ matrix composed of a subset of the lines of X. In this section, we assume that $n_1 + ... + n_p = n$ and that $X_1, ..., X_p$ constitute a partition of X. We let, for $i = 1, ..., p$, $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1}Y(X_i)$ and $v_i(x) = k(x, x) - k(x, X_i)k(X_i, X_i)^{-1}k(X_i, x)$. Then, conditionally to $Y(X_i)$, $Y(x)$ is Gaussian with mean $M_i(x)$ and variance $v_i(x)$. We consider aggregated predictors of the form

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

where $v_{prior}(x) = k(x, x)$ and with $\alpha_k : \mathbb{R}^{p+1} \to \mathbb{R}$. The aggregation techniques Product of Expert (PoE), generalized Product or Expert (gPoE), Bayesian Committee Machines (BCM) and robust Bayesian Committee machines (rBCM) satisfy (2). For PoE [Hinton, 2002, Deisenroth and Ng, 2015] we have

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

with $\beta_i(x) = 1$. For gPoE, the previous display holds with $\beta_i(x) = 1$ replaced by $\beta_i(x) = (1/2)[\log(v_{prior}(x)) - \log(v_i(x))]$ [Cao and Fleet, 2014]. For BCM [Tresp, 2000, Deisenroth and Ng, 2015] we have

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

with $\beta_i(x) = 1$. For rBCM, the previous display holds with $\beta_i(x) = 1$ replaced by $\beta_i(x) = (1/2)[\log(v_{prior}(x)) - \log(v_i(x))]$ [Deisenroth and Ng, 2015].

In the next proposition, we show 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 array of observation points that are dense in a compact set D.

Proposition 1 (Non-consistency of variance based aggregations)**.** *Let D be a compact nonempty subset of $\mathbb{R}^d$. Let Y be a Gaussian process on D with mean zero and stationary covariance function k. Assume that k is defined on $\mathbb{R}^d$, continuous and satisfies $k(x, y) > 0$ for two distinct points $x, y \in D$ such that D contains two open balls with strictly positive radii and centers x and y. Assume also 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 exists $0 \leq A < \infty$ and $0 \leq T < \infty$ such that $1/\hat{k}(\omega) \leq A||\omega||^T$, with $||.||$ the Euclidean norm.*

For any triangular array of observation points $(x_{ni})_{1 \leq i \leq n; n \in \mathbb{N}}$, for $n \in \mathbb{N}$ we let p_n be a number of Kriging predictors, we let X be the $n \times d$ matrix with line i equal to ${x_{ni}}^t$, we let $X_1, ..., X_{p_n}$ constitute a partition of X. Finally, for $n \in \mathbb{N}$ we let $M_{\mathcal{A},n}$ be obtained from (2) with p replaced by p_n. We also assume that

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

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 subvectors $X_1, ..., X_{p_n}$ 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 there exists $x_0 \in D$ such that

$$\liminf_{n\to\infty} \mathrm{E}\left[(Y(x_0) - M_{\mathcal{A},n}(x_0))^2\right] > 0. \tag{3}$$

The detailed proof is given in Appendix B. Its 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.

We remark that Proposition 1 applies to the PoE, gPoE, BCM, rBCM methods introduced above. Furthermore, the assumptions made on k in this proposition are satisfied by many stationary covariance functions, including those of the Matérn model, with the notable exception of the Gaussian covariance function (Proposition 1 in [Vazquez and Bect, 2010]).

3 The nested Kriging prediction of [Rullière et al., 2017]

In this section, we only assume that $M_1(x), ..., M_p(x)$ have mean zero and finite variance, and do not necessarily assume that $M_i(x) = k(x, X_i)k(x_i, X_i)^{-1}Y(X_i)$. In [Rullière et al., 2017], it is proposed to define the aggregated predictor $M_{\mathcal{A}}(x)$ as the best linear predictor of $Y(x)$ from $M_1(x), ..., M_p(x)$. Let $K_M(x)$ be the $p \times p$ covariance matrix of $(M_1(x), ..., M_p(x))$, let $k_M(x)$ be the $p \times 1$ vector with component i equal to $\mathrm{Cov}\left[Y(x), M_i(x)\right]$ and let $M(x) = (M_1(x), ..., M_p(x))^t$. Then, we have

$$M_{\mathcal{A}}(x) = k_M(x)^t K_M(x)^{-1} M(x) \tag{4}$$

and

$$v_{\mathcal{A}}(x) = \mathrm{E}\left[(Y(x) - M_{\mathcal{A}}(x))^2\right] = k(x, x) - k_M(x)^t K_M(x)^{-1} k_M(x).$$

As shown in [Rullière et al., 2017], the aggregated predictor $M_{\mathcal{A}}$ preserves the interpolation properties, the linearity, and the conditional Gaussianity of the predictors $M_1, ..., M_p$. Furthermore, the practical benefit of $M_{\mathcal{A}}$ is demonstrated in the simulations and real data examples in [Rullière et al., 2017]. In the rest of the section, we show the consistency of $M_{\mathcal{A}}$, show that it can be interpreted as a conditional expectation for a modified Gaussian process, and provide bounds on the errors $M_{\mathcal{A}}(x) - M_{full}(x)$.

3.1 Consistency

In the next proposition, we provide the consistency result in the case where $M_i(x) = k(x, X_i)k(X_i, X_i)^{-1} Y(X_i)$. The proof is given in Appendix A.

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 $X_1, \dots, X_{p_n}$, 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 line of X is a line of at least one X_i, $1 \leq i \leq p_n$. Then we have, with $M_{\mathcal{A}}(x)$ as in (4)*,*

$$\sup_{x\in D} \mathbb{E}\left((Y(x) - M_{\mathcal{A}}(x))^2\right) \to_{n\to\infty} 0. \tag{5}$$

3.2 The Gaussian process perspective

In this section, we develop an alternative construction where the process Y is replaced by an alternative 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 point of view allows us to see the proposed aggregation not only as an approximation of the full model but also as an exact method for a slightly different process (as illustrated in the further commented Figure 1). 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., 2017].

Recall that we consider here that $(M_1, \dots, M_p, Y)^t$ is a centered process with finite variance on the whole input space D. We define the $p \times 1$ cross-covariance vector $k_M(x, x') = \mathrm{Cov}\left[M(x), Y(x')\right]$ and the $p \times p$ cross-covariance matrix $K_M(x, x') = \mathrm{Cov}\left[M(x), M(x')\right]$, for all $x, x' \in D$. Notice that we hence have $K_M(x) = K_M(x, x)$ and $k_M(x) = k_M(x, x)$. We now define an aggregated process $Y_{\mathcal{A}}$ based on $M_{\mathcal{A}}$ that aims at reproducing the behavior of the process Y:

Definition 1 (Aggregated process)**.** *We define the process $Y_{\mathcal{A}}$ 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* (4)*.*

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 $Y - M_{\mathcal{A}}$. The process $Y_{\mathcal{A}}$ is centered with a covariance function given for all $x, x' \in D$ by

$$\begin{aligned} k_{\mathcal{A}}(x, x') = k(x, x') &+ 2k_M(x)^t K_M^{-1}(x) K_M(x, x') K_M^{-1}(x') k_M(x') \\ &- k_M(x)^t K_M^{-1}(x) k_M(x, x') - k_M(x')^t K_M^{-1}(x') k_M(x', x)\,, \end{aligned} \tag{6}$$

The main interest of introducing $Y_{\mathcal{A}}$ is that it corresponds to a Gaussian process for which $M_{\mathcal{A}}$ and $v_{\mathcal{A}}$ are the conditional mean and variance of $Y_{\mathcal{A}}$ given $Y_{\mathcal{A}}(X)$:

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 $M_{\mathcal{A}}(X) = Y(X)$, then*

$$\begin{cases} M_{\mathcal{A}}(x) = \mathrm{E}\left[Y_{\mathcal{A}}(x) | Y_{\mathcal{A}}(X)\right] , \\ v_{\mathcal{A}}(x) = \mathrm{V}\left[Y_{\mathcal{A}}(x) | Y_{\mathcal{A}}(X)\right] . \end{cases} \tag{7}$$

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

$$\begin{aligned} \mathrm{E}\left[Y_{\mathcal{A}}(x)|Y_{\mathcal{A}}(X)\right] &= \mathrm{E}\left[Y_{\mathcal{A}}(x)|M_{\mathcal{A}}(X) + 0\right] \\ &= \mathrm{E}\left[M_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] + \mathrm{E}\left[\varepsilon'_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] \\ &= \mathrm{E}\left[g_x(Y(X))|Y(X)\right] + 0 \\ &= M_{\mathcal{A}}(x). \end{aligned} \tag{8}$$

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

$$\begin{aligned} \mathrm{V}\left[Y_{\mathcal{A}}(x)|Y_{\mathcal{A}}(X)\right] &= \mathrm{V}\left[Y_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] \\ &= \mathrm{V}\left[M_{\mathcal{A}}(x)|M_{\mathcal{A}}(X)\right] + \mathrm{V}\left[\varepsilon'_{\mathcal{A}}(x)\right] \\ &= v_{\mathcal{A}}(x). \end{aligned} \tag{9}$$

□

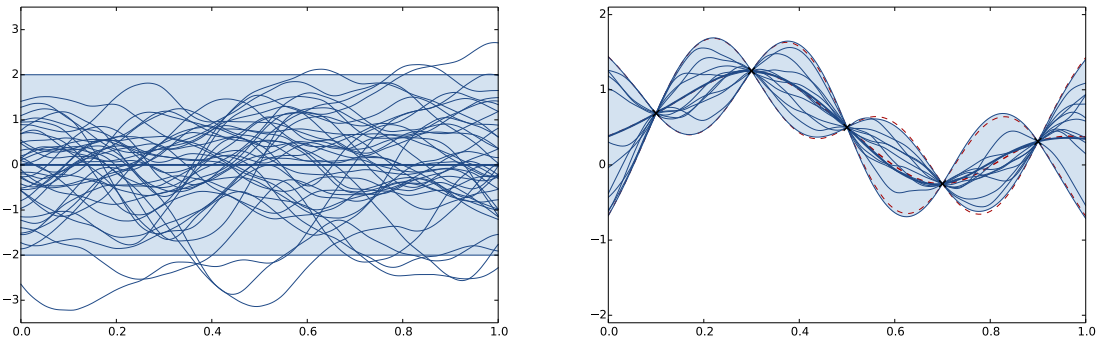
One great advantage of Proposition 3 is to introduce the conditional covariance:

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

In the case where (M, Y) is Gaussian, then $Y_{\mathcal{A}}$ is also Gaussian and (10) writes

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

This point of view thus enables us to define conditional sample paths. As an illustration, in Figure 1, we set $D = \mathbb{R}$ and assume the function $f(x) = \sin(2\pi x) + x$ to be a sample path of a centered Gaussian process Y with squared exponential covariance $k(x, x') = \exp\left(-12.5(x - x')^2\right)$. We consider the five observation points in $X = (0.1, 0.3, 0.5, 0.7, 0.9)^t$, which we divide into the $p = 2$ subgroups $X_1 = (0.1, 0.3, 0.5)^t$ and $X_2 = (0.7, 0.9)^t$. In Figure 1, we show unconditional realizations of $Y_{\mathcal{A}}$ and conditional realizations given $Y_{\mathcal{A}}(X) = f(X)$.



(a) Unconditional samples.

(b) Conditional samples.

Figure 1: 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 new covariance $k_{\mathcal{A}}$ of the process $Y_{\mathcal{A}}$ can be shown to coincide with the one of the process Y at several locations, as detailed in the following proposition.

Proposition 4 (Covariance interpolation)**.** *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, if $M_{\mathcal{A}}$ is interpolating Y at X, i.e. if $M_{\mathcal{A}}(X) = Y(X)$ then $k_{\mathcal{A}}(X, X) = k(X, X)$.*

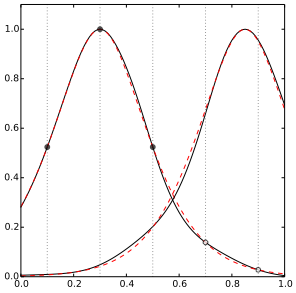
Proof. The first property of this proposition is a direct consequence of (6). The second one relies on the fact that $Y_{\mathcal{A}}(X) = Y(X)$ under the interpolation assumption. □

In Figure 2 we illustrate the difference between the covariance functions k and $k_{\mathcal{A}}$, using the settings of Figure 1. We observe that

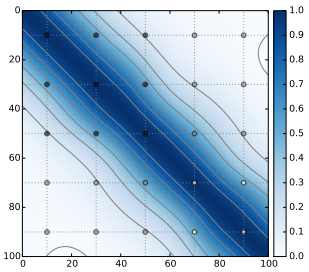
(a) the absolute difference between the two covariance functions k and $k_{\mathcal{A}}$ is quite small. Furthermore, the identity $k_{\mathcal{A}}(X, X) = k(X, X)$ of Proposition 4 is illustrated : as 0.3 is a component of X, $k_{\mathcal{A}}(0.3, x_k) = k(0.3, x_k)$ for any of the five components x_k 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.

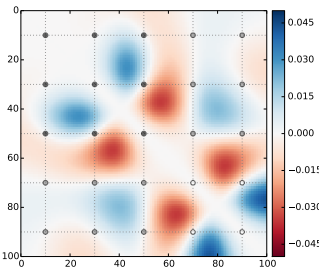
(c) the difference $k_{\mathcal{A}} - k$ vanishes at some places, among which are the places of the bullets points and the diagonal which correspond respectively to $k_{\mathcal{A}}(X, X) = k(X, X)$ and $k_{\mathcal{A}}(x, x) = k(x, x)$. Furthermore, the absolute differences between the two covariances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



(a) covariance functions $k_{\mathcal{A}}$ (solid lines) and k (dashed lines) with one variable fixed to $0.3 \in X$ and $0.85 \notin X$.



(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 locations of observed points x_i for $i \in \{1, \ldots, 5\}$. The bullets indicate locations where $k_{\mathcal{A}}(x_i, x_j) = k(x_i, x_j)$.

Finally, it should be noted that computing $c_{\mathcal{A}}$ or generating conditional samples of $Y_{\mathcal{A}}$ requires to inverse the $n \times n$ matrix $k_{\mathcal{A}}(X, X)$ which is computationally costly for large n. Hence, we see the benefit of the process $Y_{\mathcal{A}}$ more for theoretical analysis and interpretation than for computational gain. Notably, knowing that the predictor $M_{\mathcal{A}}$ is a conditional expectation for the process $Y_{\mathcal{A}}$ could 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].

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}. In this section we focus 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 $M(x) = \Lambda(x)Y(X)$. Hence we have

$$\left\{ \begin{array}{rcl} M_{\mathcal{A}}(x) - M_{full}(x) & = & -k(x, X)\Delta(x)Y(X)\,, \\ v_{\mathcal{A}}(x) - v_{full}(x) & = & k(x, X)\Delta(x)k(X, x)\,. \end{array} \right. \tag{12}$$

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

Proposition 5 (Bounds for maximal errors)**.** *Let $x \in D$. If $M(x)$ is linear in $Y(X)$, then for any norm $\|.\|$, there exists some constants $\lambda, \mu \in \mathbb{R}^+$ such that*

$$\left\{ \begin{array}{rcl} |M_{\mathcal{A}}(x) - M_{full}(x)| & \leq & \lambda\|k(X, x)\|\|Y(X)\|\,, \\ |v_{\mathcal{A}}(x) - v_{full}(x)| & \leq & \mu\|k(X, x)\|^2\,. \end{array} \right. \tag{13}$$

This implies that, if one can choose a prediction point x far enough from the observations points in X, in the sense $\|k(X,x)\| \leq \epsilon$ for any given $\epsilon > 0$, $|M_{\mathcal{A}}(x) - M_{full}(x)|$ and $|v_{\mathcal{A}}(x) - v_{full}(x)|$ can be as small as desired. Furthermore, since $v_{full}(x) = \mathrm{E}\left[(Y(x) - M_{full}(x))^2\right]$, we have:

$$0 \leq v_{\mathcal{A}}(x) - v_{full}(x) \leq \min_{k \in \{1,\dots,p\}} \mathrm{E}\left[(Y(x) - M_k(x))^2\right] - v_{full}(x)\,. \tag{14}$$

Proof. For the first part of the proposition, $\Delta(x)$ is the difference of two positive semi-definite matrices. After expanding (12) both terms can thus be interpreted as differences of inner products. We can thus conclude using successive application of triangular inequality, Cauchy-Schwartz inequality, and equivalence of norms for finite-dimensional real vector spaces. Regarding the second part, 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. □

One consequence of previous proposition 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 (see for instance the methods addressed in Section 2).

We have seen in Figure 2 that the covariance functions k and $k_{\mathcal{A}}$ are very similar. The following proposition gives a link between the aggregation errors and the covariance differences.

Proposition 6 (Errors as covariance differences)**.** *Assume that for all $x \in D$, $M(x)$ is a linear function of $Y(X)$ and that M interpolates Y at X, i.e. if for any component x_k of the vector X there is at least one index $i_k \in \mathcal{A}$ such that $M_{i_k}(x_k) = Y(x_k)$, then the differences between the full and aggregated models write as differences between covariance functions :*

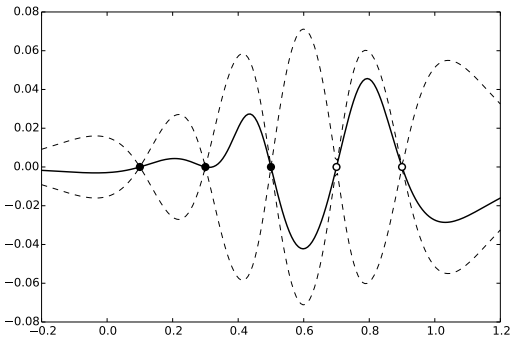
$$\begin{cases} \mathrm{E}\left[(M_{\mathcal{A}}(x) - M_{full}(x))^2\right] = \|k(X,x) - k_{\mathcal{A}}(X,x)\|_K^2\,, \\ v_{\mathcal{A}}(x) - v_{full}(x) = \|k(X,x)\|_K^2 - \|k_{\mathcal{A}}(X,x)\|_K^2\,, \end{cases} \tag{15}$$

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.

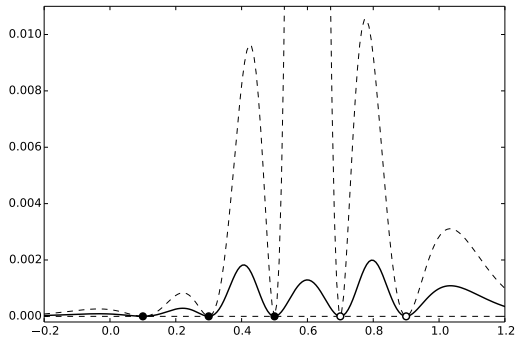
Proof. The first equality comes from $k(X,X) = k_{\mathcal{A}}(X,X)$ from Proposition 2 in [Rullière et al., 2017] and Proposition 4, which leads to $M_{\mathcal{A}}(x) - M_{full}(x) = (k(x,X) - k_{\mathcal{A}}(x,X))k(X,X)^{-1}Y(X)$. The second equality uses both $k(X,X) = k_{\mathcal{A}}(X,X)$ and $k(x,x) = k_{\mathcal{A}}(x,x)$ which leads to $v_{\mathcal{A}}(x) = k(x,x) - k_{\mathcal{A}}(x,X)k(X,X)^{-1}k_{\mathcal{A}}(X,x)$. The result is then obtained by subtracting $v_{full}(x)$. Finally, the classical inequality between $\|.\|_K$ and $\|.\|$ derives from the diagonalization of $k(X,X)$. □

The difference between the full model and the aggregated one of Figure 1 is illustrated in Figure 3. Various remarks can be made on this figure. First, the difference between the aggregated and full model is small, both on the predicted means and variances. 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 (14) are relatively loose. This means that the nested aggregation is more informative than the most accurate sub-model.

At last, the following result gives another optimality property that is often not satisfied by other aggregation methods (for instance these of Section 2): if the sub-models contain enough information, the aggregated model corresponds to the full one.



(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 3: Comparisons of the full and aggregated model. The dashed lines correspond to the bounds given in Proposition 6: $\pm\lambda_{\min}^{-1/2}\|k(X,x) - k_{\mathcal{A}}(X,x)\|$ on panel (a) and bounds of (14) on panel (b).

Proposition 7 (Fully informative sub-models)**.** *Assume $M(x)$ 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*

$$\left\{ \begin{array}{rcl} M_{\mathcal{A}}(x) & = & M_{full}(x)\,, \\ v_{\mathcal{A}}(x) & = & v_{full}(x)\,. \end{array} \right. \tag{16}$$

Furthermore,

$$Y_{\mathcal{A}} \overset{law}{=} Y \quad \textit{and thus} \quad Y_{\mathcal{A}}|Y_{\mathcal{A}}(X) \overset{law}{=} Y|Y(X). \tag{17}$$

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

Proof. 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 $Y_{\mathcal{A}} \overset{law}{=} Y$. □

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 Concluding remarks

We have analyzed theoretically several procedures, recently proposed in the literature, aiming at aggregating Kriging submodels constructed separately from subsets of a large data set of observations. We have shown that aggregating the submodels based only on their conditional variances can yield inconsistent aggregated Kriging predictors. In contrasts, we have shown the consistency of the procedure in [Rullière et al., 2017], which explicitly takes into account the correlations between the submodel predictors. We have also shed some light on this procedure, by showing that it provides an exact conditional distribution, for a different Gaussian process distribution, and by obtaining bounds on the differences with the exact full Kriging model.

Some perspectives remain open. It would be interesting to see whether the consistency of the procedure of [Rullière et al., 2017] always carries over in the case of noisy observations. It would also be beneficial to improve the aggregation methods of Section 2, in order to guarantee their consistency while keeping their low computational costs. Finally, the interpretation of the predictor in [Rullière et al., 2017] as an exact conditional expectation could be the basis of further asymptotic studies, as discussed in Section 3.2.

A Proof of Proposition 2

Because D is compact we have $\lim_{n\to\infty} \sup_{x\in D} \min_{i=1,...,n} ||x_{ni} - x|| = 0$. Indeed, if this does not hold, there exists $\epsilon > 0$ and a subsequence $\phi(n)$ such that $\sup_{x\in D} \min_{i=1,...,\phi(n)} ||x_{\phi(n)i} - x|| \geq 2\epsilon$. Hence, there exists a sequence, $x_{\phi(n)} \in D$ such that $\min_{i=1,...,\phi(n)} ||x_{\phi(n)i} - x_{\phi(n)}|| \geq \epsilon$. 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, $\min_{i=1,...,\phi(n)} ||x_{\phi(n)i} - x_{lim}|| \geq \epsilon/2$, 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,

$$\begin{aligned} \sup_{x\in D} \mathrm{E}\left[(Y(x) - M_{\mathcal{A}}(x))^2\right] &\leq \sup_{x\in D} \mathrm{E}\left[\left(Y(x) - M_{j_n(x)}(x)\right)^2\right] \\ &\leq \sup_{x\in D} \mathrm{E}\left[\left(Y(x) - \mathrm{E}\left[Y(x)|Y(x_{ni_n(x)})\right]\right)^2\right]. \end{aligned} \tag{18}$$

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 (18),

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

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.

B Proof of Proposition 1

Because of the assumptions on k, Y has the no-empty-ball property (Definition 1 and Proposition 1 in [Vazquez and Bect, 2010]). Hence for $\delta > 0$, letting

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

we have that $V(\delta) > 0$.

Consider a 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, $\inf_{1\leq i\leq n} ||u_i - x_0|| \geq \delta_n$.

Let x_0 and $\bar{x}$ be such that $k(x_0, \bar{x}) > 0$ and D contains two open balls with strictly positive radii and centers x_0 and $\bar{x}$ (the existence is assumed in the proposition). We can find $0 < r_1 < ||x_0 - \bar{x}||/4$ such that $B(\bar{x}, r_1) \subset D$. Then, by continuity of k, we can find $\epsilon_2 > 0$, $0 < r \leq r_1$ and $0 < \delta_1 \leq r_1$ such that $B(\bar{x}, r) \subset D$ and for all $x \in B(\bar{x}, r)$, $||x - x_0|| \geq \delta_1$ and

$$k(x_0, x_0) - \frac{k(x, x_0)^2}{k(x, x)} \leq k(x_0, x_0) - \epsilon_2.$$

Consider then the sequence $(w_n)_{n\in\mathbb{N}} \in D^{\mathbb{N}}$ such that for all n, $w_n = \bar{x} - (r/(1+n))e_1$ 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.

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.

Observing that $\inf_{i\in\mathbb{N}} ||w_i - x_0|| \geq \delta_1$ and letting $\epsilon_1 = V(\delta_1) > 0$, we have for all $n \in \mathbb{N}$ and for all $k = k_n + 1, ..., p_n$, since then X_k is nonempty and only contains elements $w_i \in B(\bar{x}, r)$,

$$\epsilon_1 \leq v_k(x_0) \leq k(x_0, x_0) - \epsilon_2. \tag{19}$$

From (19), and since $\hat{k}$ is a positive function and x_0 is not a component of X, we have $v_k(x_0) > 0$ for all k, and $v_{p_n}(x_0) < k(x_0, x_0)$. Hence, $M_{\mathcal{A},n}$ is well-defined, at least for n large enough.

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

$$R = \left|\left| \sum_{k=1}^{k_n} \alpha_{k,n}(v_1(x_0), ..., v_{p_n}(x_0), v_{prior}(x_0)) M_k(x_0) \right|\right|.$$

Then, from the triangular inequality, and since, from the law of total variance, $||M_k(x_0)|| \leq ||Y(x_0)|| = v_{prior}(x_0)$ we have

$$\begin{aligned} R &\leq \frac{\sum_{k=1}^{k_n} a(v_k(x_0), v_{prior}(x_0)) \sqrt{v_{prior}(x_0)}}{\sum_{l=1}^{p_n} b(v_l(x_0), v_{prior}(x_0))} \\ &\leq \frac{k_n \sup_{s^2 \geq V(\delta_n)} a(s^2, v_{prior}(x_0)) \sqrt{v_{prior}(x_0)}}{(p_n - k_n) \inf_{\epsilon_1 \leq s^2 \leq v_{prior}(x_0) - \epsilon_2} b(s^2, v_{prior}(x_0))}, \end{aligned}$$

where the last inequality is obtained from (19) and the definition of δ_n and $V(\delta)$.

Let now for $\delta > 0$, $s(\delta) = \sup_{V(\delta) \leq s^2 \leq v_{prior}(x_0)} a(s^2, v_{prior}(x_0))$. 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 $\delta_n = \inf\{\delta \geq n^{-1/2}; V(\delta) \leq n^{1/2}\}$). 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

$$\frac{\sqrt{v_{prior}(x_0)}}{\inf_{\epsilon_1 \leq s^2 \leq v_{prior}(x_0) - \epsilon_2} b(s^2, v_{prior}(x_0))}$$

is a finite constant, as b is positive and continuous on $\mathring{D}$, we have that $R \to_{n\to\infty} 0$. As a consequence, we have from the triangular inequality

$$\begin{aligned}
&\left| ||Y(x_0) - M_{\mathcal{A},n}(x_0)|| - ||Y(x_0) - \sum_{k=k_n+1}^{p_n} \alpha_{k,n}(v_1(x_0), ..., v_{p_n}(x_0), v_{prior}(x_0))M_k(x_0)|| \right| \\
&\leq || \sum_{k=k_n+1}^{p_n} \alpha_{k,n}(v_1(x_0), ..., v_{p_n}(x_0), v_{prior}(x_0))M_k(x_0) - M_{\mathcal{A},n}(x_0)|| \\
&= R \\
&\to_{n\to\infty} 0.
\end{aligned}$$

Hence

$$\liminf_{n\to\infty} ||Y(x_0) - M_{\mathcal{A},n}(x_0)|| = \liminf_{n\to\infty} \left|\left| Y(x_0) - \sum_{k=k_n+1}^{p_n} \alpha_{k,n}(v_1(x_0), ..., v_{p_n}(x_0), v_{prior}(x_0))M_k(x_0) \right|\right|.$$

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

$$\liminf_{n\to\infty} ||Y(x_0) - M_{\mathcal{A},n}(x_0)|| \geq V(\delta_1) > 0.$$

References

[Bachoc et al., 1016] Bachoc, F., Ammar, K., and Martinez, J. (1016). Improvement of code behavior in a design of experiments by metamodeling. *Nuclear science and engineering*, 183(3):387–406.

[Cao and Fleet, 2014] Cao, Y. and Fleet, D. J. (2014). Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions. *ArXiv e-prints*.

[Chevalier and Ginsbourger, 2013] 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.

[Datta et al., 2016] 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*, accepted.

[Deisenroth and Ng, 2015] 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*.

[Furrer et al., 2006] Furrer, R., Genton, M. G., and Nychka, D. (2006). Covariance tapering for interpolation of large spatial datasets. *Journal of Computational and Graphical Statistics*.

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

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

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

[Kaufman et al., 2008] 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.

[Marrel et al., 2009] 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, 1970] 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 et al., 2001] 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 and Rasmussen, 2005] 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.

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

[Rullière et al., 2017] Rullière, D., Durrande, N., Bachoc, F., and Chevalier, C. (2017). Nested kriging predictions for datasets with a large number of observations.

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

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

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

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

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

[van Stein et al., 2015] 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.

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

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