ESAIM: PS 23 (2019) 271–309
https://doi.org/10.1051/ps/2018023

ESAIM: Probability and Statistics
www.esaim-ps.org

OPTIMAL COMPROMISE BETWEEN INCOMPATIBLE CONDITIONAL PROBABILITY DISTRIBUTIONS, WITH APPLICATION TO OBJECTIVE BAYESIAN KRIGING

Joseph Muré[1,2,*]

Abstract. Models are often defined through conditional rather than joint distributions, but it can be difficult to check whether the conditional distributions are compatible, *i.e.* whether there exists a joint probability distribution which generates them. When they are compatible, a Gibbs sampler can be used to sample from this joint distribution. When they are not, the Gibbs sampling algorithm may still be applied, resulting in a "pseudo-Gibbs sampler". We show its stationary probability distribution to be the optimal compromise between the conditional distributions, in the sense that it minimizes a mean squared misfit between them and its own conditional distributions. This allows us to perform Objective Bayesian analysis of correlation parameters in Kriging models by using univariate conditional Jeffreys-rule posterior distributions instead of the widely used multivariate Jeffreys-rule posterior. This strategy makes the full-Bayesian procedure tractable. Numerical examples show it has near-optimal frequentist performance in terms of prediction interval coverage.

Mathematics Subject Classification. 62F15, 62M30, 60G15.

Received October 11, 2017. Accepted October 31, 2018.

1. Introduction

Generally speaking, there are two ways to create statistical models for multiple random variables. One can either consider them simultaneously and directly define their joint distribution, or one can define a system of conditional distributions. The first approach is conceptually easier and often (but not always) leads to models with well understood properties because a closed-form expression is available. The second one allows for more flexibility in modeling but makes theoretical analysis more difficult.

The main problem with the second approach is that conditional distributions may not be compatible. In this context, it means that there exists no joint distribution from which the conditional distributions can all be derived. Other definitions of compatibility exist in the literature. For example, in the context of a model with a given prior distribution, Dawid and Lauritzen [11] examine the problem of eliciting a compatible prior distribution for a submodel. In the domain of Bayesian Networks, a probability distribution can be compatible or not with a given Directed Acyclic Graph (DAG) [30]. Moreover, an abstraction (simplification) of a DAG

Keywords and phrases: Incompatibility, conditional distribution, Markov kernel, optimal compromise, Kriging, reference prior, integrated likelihood, Gibbs sampling, posterior propriety, frequentist coverage.

[1] EDF R&D, Dpt. PRISME, 6 quai Watier, 78401 Chatou, France.
[2] Université Paris Diderot, Laboratoire de Probabilités, Statistique et Modélisation, France.

* Corresponding author: joseph.mure@edf.fr

can be compatible or not [34]. A concept of compatibility of two prior distributions also exists in the field of Bayesian model selection. It is based on the Kullback–Leibler divergence of the corresponding marginal (predictive) distributions [9]. In this paper however, the notion of compatibility concerns families of conditional distributions. A family of conditional distributions is called compatible if there exists a joint distribution that agrees with them all [16]. Uniqueness of this joint distribution is desirable but not included in the requirements of compatibility.

To illustrate this definition of compatibility, consider the following random-effect model ([16], cited by Robert [29] page 41), which shows that "in general, reasonable-seeming conditional models will not be compatible with any single joint distribution" [12]. Define $y_{ij} = \beta + u_i + \epsilon_{ij}$, $i \in [\![1, I]\!]$ and $j \in [\![1, J]\!]$ where $u_i \sim \mathcal{N}(0, \sigma^2)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \tau^2)$. The parameters of the model being $(\beta, \sigma^2, \tau^2)$, let us consider the prior distribution $\pi(\beta, \sigma^2, \tau^2) \propto (\sigma^2\tau^2)^{-1}$: the corresponding posterior is improper. The conditional posterior distributions cannot therefore be compatible. Nevertheless, they are all proper. With $\bar{y}$ and $\bar{u}$ being the empirical means of $\boldsymbol{y} = (y_{ij})_{i\in[\![1,I]\!],j\in[\![1,J]\!]}$ and $\boldsymbol{u} = (u_i)_{i\in[\![1,I]\!]}$, $\beta|\boldsymbol{u}, \boldsymbol{y}, \sigma^2, \tau^2 \sim \mathcal{N}(\bar{y} - \bar{u}, \tau^2/IJ)$, $\sigma^2|\boldsymbol{u}, \boldsymbol{y}, \beta, \tau^2 \sim \mathcal{IG}\left(I/2, \sum_i u_i^2/2\right)$, $\tau^2|\boldsymbol{u}, \beta, \boldsymbol{y}, \sigma^2 \sim \mathcal{IG}\left(IJ/2, \sum_{ij}(y_{ij} - u_i - \beta)^2/2\right)$. If the posterior distribution was presented only through these conditionals, one might miss the fact that Gibbs sampling is impossible in this case due to the corresponding Markov chain being null recurrent.

For finite state spaces, null recurrent Markov chains are impossible, but attempting to define a joint probability distribution through its conditional distributions may still lead to incompatibility. Accordingly, the problem of efficiently determining whether a given system of conditionals is compatible has received considerable attention over the years. Kuo *et al.* [22], after listing previous attempts, provide probably the best solution to date. Their idea relies on the Structural Ratio Matrix which contains ratios between conditional distributions.

However, even if a system contains incompatible conditional probability distributions, it does not follow that it is useless. Since Heckerman *et al.* [15], practitioners have been using systems of conditional probability distributions without reference to compatibility. Indeed, providing the Markov chain is positive recurrent, it is always possible to fire up Gibbs samplers to deal with a system of conditional distributions. Some authors use the colorful acronym PIGS for "Potentially Incompatible Gibbs Sampler" to describe such a procedure. When the conditionals are definitely known to be incompatible, the most widely used term seems to be "pseudo-Gibbs sampler" (PGS).

Behind the practice of PIGS is the intuition that the Gibbs sampler should converge to the joint distribution that best represents the system of conditionals. Kuo and Wang [21] provide a detailed analysis and geometrical interpretation of the behavior of PGSs for discrete conditional distributions. In particular, they show how the scanning order determines its stationary distribution. In Section 2 of the present paper, we provide some theoretical foundation for the intuition that the stationary distribution of a PGS with random scanning order is, in case of uniqueness, the best "compromise" between incompatible conditionals. Section 3 provides further discussion of this theory by considering alterations to the main definitions and showing them to lead to undesirable results.

In Section 4, we use the theory of optimal compromise to derive Objective Bayesian inference on correlation parameters of Kriging models. Kriging models are widely used in spatial statistics, but they are more complex than standard models. Their parameters are numerous, and their interactions complex, which makes eliciting a joint objective prior difficult. This makes an approach resting on conditional priors attractive despite the risk of incompatibility of the associated conditional posteriors. It is to deal with such situations that the theory of optimal compromise was developed.

The Objective Bayesian paradigm as explained by Berger [4] consists in eliciting for every model a "default", reasonable prior distribution that could be used when no explicit prior information is available. In particular, the Berger–Bernardo reference prior [8], hereafter simply named "reference prior", can be algorithmically computed with minimal user intervention.

For models with a single scalar parameter, the reference prior rewards parameter values that are easily discriminated by the likelihood function. Its definition is related to the Kullback–Leibler divergence between posterior and prior [8]. For usual continuous models – essentially models where the Fisher information matrix

is equal to the opposite of the expectancy of the second derivative of the log-likelihood, see Clarke and Barron [10] for an exhaustive list of conditions – it coincides with the Jeffreys-rule prior.

For models with multiple parameters, the reference prior algorithm requires the user to specify an ordering on the parameters and then iteratively compute the reference prior on each parameter conditionally to all subsequent parameters. The only user input is therefore this ordering, and common sense arguments often make one more sensible than others. Yang and Berger [33] list different reference priors obtained with different parameter orderings for a large number of statistical models. Of course, one could also group several parameters and treat them as one single multi-dimensional parameter, but doing so tends to produce less satisfactory inference [5]. In particular, for usual continuous models Berger *et al.* [7] state "We actually know of no multivariable example in which we would recommend the Jeffreys-rule prior. In higher dimensions, the prior always seems to be either 'too diffuse' [...] or 'too concentrated' ".

Berger *et al.* [6] were the first to derive a reference prior for the parameters of a Gaussian process regression model. This model contained only one correlation parameter, however. When several correlation parameters are involved, there is no reasonable way to order them. Even if one were arbitrarily picked, computation of the prior would be analytically intractable. Several authors [13, 19, 25, 27, 28] have therefore resolved to treat all correlation parameters as a single multi-dimensional parameter. It is in order to avoid having to do this that we make use of PIGS.

The idea is simple: for every correlation parameter, it is possible to analytically derive the reference prior for this parameter conditionally to all others. Each of the corresponding posterior distributions can be seen as a conditional probability distribution on one correlation parameter when all others are known. These conditional distributions then serve as input to a PIGS.

Theorem 4.2 is the main result with respect to the application.

First, under reasonable assumptions, the PIGS admits one single stationary probability distribution. Second, the Markov kernel defined by the PIGS is uniformly ergodic. Since this Markov kernel is defined over an uncountable state space, the latter fact is significant. The stationary distribution, which we call the Gibbs reference posterior distribution, can be used to improve prediction of the value taken by the Gaussian process at unobserved points. Sections 5 and 6 illustrate the inferential and predictive performance of the stationary distribution, respectively.

2. Optimal compromise: a general theory

2.1. Definitions and notations

In this section, we introduce the concepts necessary to define the optimal compromise between potentially incompatible conditional distributions. For the sake of readability, all proofs are provided in Appendix A.

First, note that in this context, "conditional distribution" is really an informal way of referring to a Markov kernel.

Definition 2.1. Let $(A, \mathcal{A})$ and $(B, \mathcal{B})$ be measurable sets. A mapping $\pi : A \times \mathcal{B} \to [0,1]$ is called a Markov kernel if:

(1) for all $x \in A$, $\pi(x, \cdot) : \mathcal{B} \to [0,1]$ is a probability distribution and
(2) for all $S \in \mathcal{B}$, $\pi(\cdot, S) : A \to [0,1]$ is $\mathcal{A}$-measurable.

We use the following notation: for every $(x, S) \in A \times \mathcal{B}$, $\pi(S|x) := \pi(x, S)$.

Let r be a positive integer and let $(\Omega_1, \mathcal{A}_1)$,...,$(\Omega_r, \mathcal{A}_r)$ be measurable sets. Define $\Omega = \times_{i=1}^{r} \Omega_i = \Omega_1 \times ... \times \Omega_r$ and $\mathcal{A} := \bigotimes_{i=1}^{r} \mathcal{A}_r = \mathcal{A}_1 \otimes ... \otimes \mathcal{A}_r$.

For every $i \in [\![1, r]\!]$, let π_i be a Markov kernel $(\times_{j \neq i} \Omega_j) \times \mathcal{A}_i \to [0,1]$.

Intuitively (we formalize this below), every π_i should be assembled with a distribution $m_{\neq i}$ on $\bigotimes_{j \neq i} \mathcal{A}_j$ to create a "joint" distribution, that is a probability distribution on $\mathcal{A}$. We refer to every $m_{\neq i}$ ($i \in [\![1, r]\!]$) as an $(r-1)$-dimensional distribution. If the $m_{\neq i}$ can be chosen in such a way as to make all joint distributions equal,

then the Markov kernels in the sequence $(\pi_i)_{i\in[\![1,r]\!]}$ are called compatible. And if no choice of $(m_{\neq i})_{i\in[\![1,r]\!]}$ can make all joint distributions equal, we have to look for a "compromise" between the Markov kernels.

Consider the following example [3] with $r = 2$ and $\Omega_1 = \Omega_2 = (0, +\infty)$ are endowed with their Borel σ-algebra: let π_1 and π_2 be Markov kernels such that for all $x, y > 0$ $\pi_1(\cdot|y)$ and $\pi_2(\cdot|x)$ are absolutely continuous with respect to the Lebesgue measure on $(0, +\infty)$. Let λ be this measure and let the densities of π_1 and π_2 respectively be

$$p_1(x|y) := \frac{\mathrm{d}\pi_1(\cdot|y)}{\mathrm{d}\lambda}(x) = (y+2)\exp(-(y+2)x); \tag{2.1}$$

$$p_2(y|x) := \frac{\mathrm{d}\pi_2(\cdot|x)}{\mathrm{d}\lambda}(y) = (x+3)\exp(-(x+3)y). \tag{2.2}$$

Let us define $m_{\neq 1}$ and $m_{\neq 2}$ as the probability distributions with the following densities with respect to the Lebesgue measure on $(0, +\infty)$:

$$\frac{\mathrm{d}m_{\neq 1}}{\mathrm{d}\lambda}(y) = \frac{(y+2)^{-1}\exp(-3y)}{\int_0^{+\infty}(t+6)^{-1}\exp(-t)\mathrm{d}t}; \tag{2.3}$$

$$\frac{\mathrm{d}m_{\neq 2}}{\mathrm{d}\lambda}(x) = \frac{(x+3)^{-1}\exp(-2x)}{\int_0^{+\infty}(t+6)^{-1}\exp(-t)\mathrm{d}t}. \tag{2.4}$$

Then the joint probability distributions $\pi_1 m_{\neq 1}$ and $\pi_2 m_{\neq 2}$ are equal. Now denoting by λ the Lebesgue measure on $(0, +\infty) \times (0, +\infty)$, the density of $\pi_1 m_{\neq 1} = \pi_2 m_{\neq 2}$ is

$$\frac{\mathrm{d}\pi_1 m_{\neq 1}}{\mathrm{d}\lambda}(x, y) = \frac{\mathrm{d}\pi_2 m_{\neq 2}}{\mathrm{d}\lambda}(x, y) = \frac{\exp(-xy - 2x - 3y)}{\int_0^{+\infty}(t+6)^{-1}\exp(-t)\mathrm{d}t}. \tag{2.5}$$

Remark 2.2 (Producing incompatibility is easy)**.** Take $r = 2$ and let $\Omega_1 = \Omega_2$ be a Borel subset of $\mathbb{R}$. Assume that for all $x, y \in \Omega_1$, $\pi_1(\cdot|y)$ and $\pi_2(\cdot|x)$ are absolutely continuous with respect to λ, which denotes here the Lebesgue measure on Ω_1. Let $p_1(\cdot|y)$ and $p_2(\cdot|x)$ be their respective density functions and further assume that for λ-almost all real numbers x and y, $p_1(x|y) > 0$ and $p_2(y|x) > 0$. A necessary condition [3] for the compatibility of π_1 and π_2 can be derived from Bayes' rule: there must exist two mappings u and v defined on Ω_1 such that for λ-almost all real numbers x and y

$$\frac{\mathrm{d}\pi_1(\cdot|y)}{\mathrm{d}\lambda}(x) \Big/ \frac{\mathrm{d}\pi_2(\cdot|x)}{\mathrm{d}\lambda}(y) = u(x)v(y). \tag{2.6}$$

In the previous example, $\Omega_1 = \Omega_2 = (0, +\infty)$ and this necessary condition is fulfilled: for all $x, y > 0$,

$$\frac{p_1(x|y)}{p_2(y|x)} = \frac{(x+3)^{-1}\exp(-2x)}{(y+2)^{-1}\exp(-3y)}. \tag{2.7}$$

Taking $p_1(x|y) = (y+2)\exp(-(y+2)x)$ and $p_2(y|x) = \exp(-y)$ makes π_1 and π_2 fail the necessary condition for compatibility.

Definition 2.3. Let ϕ be a probability distribution on $\mathcal{A}$. For every $i \in [\![1, r]\!]$, denote by ϕ_{-i} the probability distribution on $\bigotimes_{j\neq i}\mathcal{A}_j$ defined as follows. For every set S_{-i} that can be decomposed as $S_{-i} = \times_{j\neq i}S_j$ (with

$S_j \in \mathcal{A}_j$ for every $j \neq i$),

$$\phi_{-i}(S_{-i}) = \phi(\times_{j<i} S_j \times \Omega_i \times \times_{k>i} S_k). \tag{2.8}$$

ϕ_{-i} is called the ith $(r-1)$-marginal distribution of ϕ.

Remark 2.4. The above definition is valid because any probability distribution on $\mathcal{A}$ can be characterized by its values on "rectangles" $\times_{i=1}^{r} S_i$ (where for every $i \in [\![1, r]\!]$, $S_i \in \mathcal{A}_i$).

For every $i \in [\![1, r]\!]$ and every probability distribution $m_{\neq i}$ on $\bigotimes_{j\neq i} \mathcal{A}_j$, denote by $\pi_i m_{\neq i}$ the distribution on $\mathcal{A}$ defined as follows. For every $i \in [\![1, r]\!]$, for every set $S_{<i} \in \bigotimes_{j<i} \mathcal{A}_j$, every set $S_{>i} \in \bigotimes_{k>i} \mathcal{A}_k$ and every set $S_i \in \mathcal{A}_i$,

$$\pi_i m_{\neq i}(S_{<i} \times S_i \times S_{>i}) = \int_{S_{<i}\times S_{>i}} \pi_i(S_i|\omega_{-i}) \mathrm{d}m_{\neq i}(\omega_{-i}). \tag{2.9}$$

Naturally, for $i = 1$ (resp. $i = r$), remove $S_{<i}$ (resp. $S_{>i}$) from the formula above. In the following, do this kind of operation when $i = 1$ or $i = r$.

Notice that for every $i \in [\![1, r]\!]$, $m_{\neq i}$ is the ith $(r-1)$-marginal distribution of $\pi_i m_{\neq i}$:

$$(\pi_i m_{\neq i})_{-i} = m_{\neq i}. \tag{2.10}$$

If there exists a sequence of $(r-1)$-dimensional distributions $(m_{\neq i})_{i\in[\![1,r]\!]}$ such that all distributions $\pi_i m_{\neq i}$ are equal, then the Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ are compatible. If no such sequence $(m_{\neq i})_{i\in[\![1,r]\!]}$ exists, then we wish to find a sequence $(m_{\neq i})_{i\in[\![1,r]\!]}$ that makes the $\pi_i m_{\neq i}$ share some "common ground". The following definition expresses this constraint formally.

Definition 2.5. A sequence of $(r-1)$-dimensional distributions $(m_{\neq i})_{i\in[\![1,r]\!]}$ (each $m_{\neq i}$ being a probability distribution on $\bigotimes_{j\neq i} \mathcal{A}_j$) is said to be *compatible* with the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if for every $i \in [\![1, r]\!]$

$$m_{\neq i} = \frac{1}{r} \sum_{j=1}^{r} (\pi_j m_{\neq j})_{-i}. \tag{2.11}$$

So the "common ground" we require for the sequence of distributions $(\pi_i m_{\neq i})_{i\in[\![1,r]\!]}$ is that their $(r-1)$-marginal distributions should be the same on average. Other constraints would have been possible, and we discuss some of them in Section 3.1 below.

Consider the case where $r = 2$, $\Omega_1 = \Omega_2 = \mathbb{R}$ and for all $x, y \in \mathbb{R}$, $\pi_1(\cdot|y) = \mathcal{N}(y/4, 1/8)$ and $\pi_2(\cdot|x) = \mathcal{N}(x, 1)$, the second argument of $\mathcal{N}(\cdot,\cdot)$ being the variance. Now define $m_{\neq 1} = \mathcal{N}(0, 2)$ and $m_{\neq 2} = \mathcal{N}(0, 1)$. We have

$$\pi_1 m_{\neq 1} = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 1/2\\ 1/2 & 2\end{pmatrix}\right) \quad \text{and} \quad \pi_2 m_{\neq 2} = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & 1\\ 1 & 2\end{pmatrix}\right) \tag{2.12}$$

Since $(\pi_1 m_{\neq 1})_{-2} = \mathcal{N}(0, 1) = m_{\neq 2}$ and $(\pi_2 m_{\neq 2})_{-1} = \mathcal{N}(0, 2) = m_{\neq 1}$, we have *a fortiori* $m_{\neq 1} = 1/2(\pi_1 m_{\neq 1})_{-1} + 1/2(\pi_2 m_{\neq 2})_{-1}$ and $m_{\neq 2} = 1/2\ (\pi_1 m_{\neq 1})_{-2} + 1/2\ (\pi_2 m_{\neq 2})_{-2}$, so $(m_{\neq 1}, m_{\neq 2})$ is compatible with (π_1, π_2) in the sense of Definition 2.5.

The definition of a compromise follows from this new definition of compatibility.

Definition 2.6. A probability distribution P on $\mathcal{A}$ is called a *compromise* between the Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if these two conditions are verified:

(1) for every $i \in [\![1, r]\!]$, $\pi_i P_{-i}$ is absolutely continuous with respect to P;
(2) the sequence $(P_{-i})_{i\in[\![1,r]\!]}$ of P's $(r-1)$-marginal distributions is compatible with $(\pi_i)_{i\in[\![1,r]\!]}$.

In the definition of a compromise, the first condition exists to give meaning to the definition of an optimal compromise below. It is reasonable on its own though: a compromise should not deem events impossible if they are considered possible by the Markov kernels.

Returning to the previous example, both $\pi_1 m_{\neq 1}$ and $\pi_2 m_{\neq 2}$ are compromises, and any convex combination of these two distributions is one as well. The following definition introduces a cost function for compromises in order to determine which compromises are optimal.

Definition 2.7. Let λ be a positive measure on $\mathcal{A}$. Let P be a compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ that is absolutely continuous with respect to λ. P is called an *optimal compromise* with respect to λ between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if it minimizes the functional E_λ over all compromises between $(\pi_i)_{i\in[\![1,r]\!]}$ that are absolutely continuous with respect to λ. E_λ is defined by:

$$E_\lambda(P) = \sum_{i=1}^{r} \int_{\mathcal{A}} \left[\frac{\mathrm{d}(\pi_i P_{-i})}{\mathrm{d}\lambda}(\omega) - \frac{\mathrm{d}P}{\mathrm{d}\lambda}(\omega) \right]^2 \mathrm{d}\lambda(\omega). \tag{2.13}$$

In the previous example, the optimal compromise with respect to the Lebesgue measure on $\mathbb{R} \times \mathbb{R}$ can be shown (*cf.* Sect. 2.2) to be $1/2\ \pi_1 m_{-1} + 1/2\ \pi_2 m_{-2}$.

Proposition 2.8. *The set of all compromises between $(\pi_i)_{i\in[\![1,r]\!]}$ is convex, as is the subset of all compromises absolutely continuous with respect to λ.*

If the Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ are compatible and there exists a joint distribution π on $\mathcal{A}$ that agrees with them all, then for every positive measure λ on $\mathcal{A}$ such that π is absolutely continuous with respect to λ, $E_\lambda(\pi) = 0$ and π is an optimal compromise with respect to λ.

Even though Definition 2.7 makes it seem like the notion of optimal compromise is tied to a reference measure λ, it turns out that in many situations there exists a compromise that is optimal with respect to all possible reference measures – *cf.* Theorem 2.13 below. In the previous example, it is given by $1/2\pi_1 m_{-1} + 1/2\pi_2 m_{-2}$.

2.2. Deriving the optimal compromise

The concepts of compromise and optimal compromise defined in Section 2.1 are intimately linked to Gibbs sampling, or (because the Markov kernels are incompatible) pseudo-Gibbs sampling (PGS). Let us therefore recall the Gibbs sampling algorithm.

Algorithm 1: Gibbs sampling

input: Variable x_i for $i \in [\![1, r]\!]$ and Markov kernels π_i for $i \in [\![1, r]\!]$
Initialize x_1,...,x_r ;
while *additional samples are desired* **do**
  Select index i from $[\![1, r]\!]$;
  Sample x_i from $\pi_i(\cdot|x_1, ..., x_{i-1}, x_{i+1}, ..., x_r)$;
end

The method used to sample the index i is called *scanning order*. A *systematic* scan moves through each index in turn using a deterministic pattern while an *equiprobable random* scan selects for each step an index within $[\![1, r]\!]$ with probability $1/r$.

If the Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ are compatible, then the scanning order influences only the rate of convergence of the algorithm. He *et al.* [14] provide theoretical results on the subject and Mitliagkas and Mackey [24] propose a measure for the quality of any scan order for Gibbs sampling on finite state spaces.

If the Markov kernels are not compatible, then every scan order may produce a different target distribution. Kuo and Wang [21] thoroughly study the links between all possible systematic scan orders for Markov kernels on finite state spaces. The equiprobable random scan order has yet another target distribution.

The notion of Gibbs compromise is central to this theory of compromises between incompatible Markov kernels.

Definition 2.9. A probability distribution P_G on $\mathcal{A}$ is called a *Gibbs compromise* between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if it satisfies:

$$P_G = \frac{1}{r}\sum_{i=1}^{r} \pi_i(P_G)_{-i}. \tag{2.14}$$

If it exists, a Gibbs compromise is a stationary distribution for the Gibbs sampler with equiprobable random scan order. This fact makes it practical from a sampling standpoint. In the following, we show how it relates to the concepts of compromise and optimal compromise in the sense of Definitions 2.6 and 2.7.

Proposition 2.10. *A Gibbs compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ is also a compromise between this sequence of Markov kernels in the sense of Definition 2.6.*

The denomination "Gibbs compromise" is justified because it is a stationary distribution for the Gibbs sampler with random equiprobable scanning order.

The proposition below shows that all compromises are tied to Gibbs compromises.

Proposition 2.11. *If a sequence of $(r-1)$-dimensional probability distributions $(m_{\neq i})_{i\in[\![1,r]\!]}$ (each $m_{\neq i}$ being a probability distribution on $\bigotimes_{j\neq i}\mathcal{A}_j$) is compatible with the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$, then it is the sequence of $(r-1)$-dimensional distributions of a Gibbs compromise between the Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$.*

Remark 2.12 (Equivalence relation and convexity)**.** Let us say that two compromises are equivalent if they share the same sequence of $(r-1)$-marginal distributions. This is obviously an equivalence relation and every class of equivalence can be represented by a single Gibbs compromise. Moreover, each class of equivalence is a convex subset of the set of all compromises. And for any positive measure λ on $\mathcal{A}$, its intersection with the set of all compromises absolutely continuous with respect to λ is also convex. Finally, the functional E_λ is convex over this intersection.

2.3. A theoretical justification of pseudo-Gibbs sampling

At this stage, all necessary tools are available to derive the two most important results of this theory of compromise between incompatible Markov kernels.

Theorem 2.13. *If there exists a unique Gibbs compromise P_G between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$, then the following statements holds:*

(1) P_G is absolutely continuous with respect to any compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$;
(2) for any positive measure λ on $\mathcal{A}$ such that P_G is absolutely continuous with respect to λ, P_G is the unique optimal compromise with respect to λ.

Because of these two properties, we call P_G the *optimal compromise.*

Remark 2.14 (Equivalence class and convexity – continued)**.** The arguments used in the proof of Theorem 2.13 (*cf.* Appendix A) can also be used to deal with the case where there exist several different Gibbs compromises. First, one can show that each Gibbs compromise is absolutely continuous with respect to any equivalent compromise. Now let $\mathcal{C}$ be an equivalence class and let $\pi_{\mathcal{C}}$ be the unique Gibbs compromise in this class of equivalence.

Let λ be a positive measure on $\mathcal{A}$ such that $\pi_{\mathcal{C}}$ is absolutely continuous with respect to λ. π_C (uniquely) minimizes E_λ over the intersection of $\mathcal{C}$ and the set of all compromises absolutely continuous with respect to λ. This implies that for any positive measure λ on $\mathcal{A}$, any optimal compromise with respect to λ is a Gibbs compromise. Finally, note that the set of all Gibbs compromises is convex (however there is no reason E_λ should be convex over this set!).

Theorem 2.13 has important practical implications. It opens the possibility of using Gibbs sampling to find the optimal compromise between incompatible Markov kernels and justifies using PIGS.

The next result shows that, under the conditions of Theorem 2.13, the optimal compromise remains the same for a fairly large class of reparametrizations. This result is key to the application of PIGS in an Objective Bayesian framework, where some degree of invariance by reparametrization of priors and posteriors is usually expected.

For every $i \in [\![1, r]\!]$, let $(\tilde{\Omega}_i, \tilde{\mathcal{A}}_i)$ be a measurable space and let f_i be bijective measurable mapping $\Omega_i \to \tilde{\Omega}_i$ whose inverse f_i^{-1} is also measurable. Define $f = (f_1, ..., f_r) : \times_{i\in[\![1,r]\!]}\Omega_i \to \times_{i\in[\![1,r]\!]}\tilde{\Omega}_i$ and for every $i \in [\![1, r]\!]$ $f_{-i} = (f_1, .., f_{i-1}, f_{i+1}, ..., f_r) : \times_{j\neq i}\Omega_j \to \times_{j\neq i}\tilde{\Omega}_j$.

Also let $\tilde{\pi}_i$ be the Markov kernel $\left(\times_{j\neq i}\tilde{\Omega}_j\right) \times \tilde{\mathcal{A}}_i \to [0, 1]$ such that for every $\omega_{-i} \in \times_{j\neq i}\Omega_j$ and every $S_i \in \mathcal{A}_i$, $\tilde{\pi}_i(f_i(S_i)|f_{-i}(\omega_{-i})) = \pi_i(S_i|\omega_{-i})$.

Proposition 2.15. *Assume there exists a unique Gibbs compromise P_G between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$. Then the push-forward measure of P_G by f $\tilde{P}_G := P_G * f$ is the unique Gibbs compromise between the sequence of Markov kernels $(\tilde{\pi}_i)_{i\in[\![1,r]\!]}$.*

3. Testing the definitions of compatibility

While the previous section presented the theory of compromise, this section aims to test its foundations, namely Definitions 2.5 and 2.6. Section 3.1 shows that strengthening their requirements is not possible since it would in many cases threaten the very existence of a compromise. Section 3.2 on the other hand shows that these requirements cannot be weakened at a small price.

3.1. Stronger definitions of compromises are not possible

While the definition of the *optimal compromise* is straightforward, as it involves minimizing some measure of distance between the "targeted" conditionals and the conditionals of the compromise, the definition of *a compromise* may seem arbitrary. To motivate this definition, let us focus on the two-dimensional case.

Suppose that $r = 2$ and that π_1 and π_2 are incompatible. This means there exists no joint distribution π which agrees with both Markov kernels. This being the case, it seems sensible to weaken the definition of compatibility by applying it to the "marginals" instead of the "joint" distribution. The following definition makes this idea precise.

Definition 3.1. A pair of probability distributions $m_{\neq 1}$ (resp. $m_{\neq 2}$) on $\mathcal{A}_2$ (resp. $\mathcal{A}_1$) is *compatible* with the pair of Markov kernels π_1 and π_2 if the distributions $\pi_1 m_{\neq 1}$ and $\pi_2 m_{\neq 2}$ verify

$$(\pi_1 m_{\neq 1})_{-2} = m_{\neq 2} \text{ and } (\pi_2 m_{\neq 2})_{-1} = m_{\neq 1}. \tag{3.1}$$

While this definition may seem more restrictive at first glance than Definition 2.5, both definitions are in fact equivalent when applied to a pair of Markov kernels, because $(r-1)$-dimensional distributions are simply one-dimensional distributions in this case. Indeed, following directly from Definition 2.5, we have this result which holds for any r:

Proposition 3.2. *If a sequence of $(r-1)$-dimensional probability distributions $(m_{\neq i})_{i\in[\![1,r]\!]}$ (each $m_{\neq i}$ being a probability distribution on $\bigotimes_{j\neq i}\mathcal{A}_j$) is compatible (in the sense of Def. 2.5) with the sequence of Markov kernels*

$(\pi_i)_{i\in[\![1,r]\!]}$, then all joint distributions in the sequence $(\pi_i m_{\neq i})_{i\in[\![1,r]\!]}$ share the same marginals, that is

$$\forall i,j,k\in[\![1,r]\!],\quad \forall S_k\in\mathcal{A}_k,\quad \pi_i m_{\neq i}\left(\times_{k'<k}\Omega_{k'}\times S_k\times\times_{k''>k}\Omega_{k''}\right)=\pi_j m_{\neq j}\left(\times_{k'<k}\Omega_{k'}\times S_k\times\times_{k''>k}\Omega_{k''}\right). \tag{3.2}$$

Now let us consider the three-dimensional case ($r=3$). Because the aim of this section is merely to motivate the definitions of compromises and optimal compromises, there is no need for the discussion to be fully general. Let us therefore restrict the discussion to an important particular case. Assume that Ω_1, Ω_2 and Ω_3 are finite sets and that $\mathcal{A}_1$, $\mathcal{A}_2$ and $\mathcal{A}_3$ are respectively the sets of all their subsets. This has an important consequence: any mapping from a subset of Ω to a subset of Ω is measurable.

Also assume that the Markov kernels π_1, π_2 and π_3 are positive mappings. For π_1, this means that for every $(\omega_1,\omega_2,\omega_3)\in\Omega_1\times\Omega_2\times\Omega_3$, $\pi_1(\{\omega_1\}|\omega_2,\omega_3)>0$.

If we consider ω_3 known, then the situation is reduced to the two-dimensional case. Because π_1 and π_2 are positive mappings and both Ω_1 and Ω_2 are finite sets, Markov chain theory ensures there exists a unique Gibbs compromise $P(\cdot|\omega_3)$. For every $(\omega_1,\omega_2)\in\Omega_1\times\Omega_2$,

$$P(\{\omega_1\}\times\{\omega_2\}|\omega_3)=\frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3)P(\Omega_1\times\{\omega_2\}|\omega_3)+\frac{1}{2}\pi_1(\omega_2|\omega_1,\omega_3)P(\{\omega_1\}\times\Omega_2|\omega_3). \tag{3.3}$$

Thanks to Theorem 2.13, $P(\cdot|\omega_3)$ is the optimal compromise. Moreover, notice that equation (3.3) defines a Markov kernel on $\Omega_3\times(\mathcal{A}_1\otimes\mathcal{A}_2)$.

We may similarly derive Markov kernels $Q:\Omega_1\times(\mathcal{A}_2\otimes\mathcal{A}_3)$ and $R:\Omega_2\times(\mathcal{A}_1\otimes\mathcal{A}_3)$.

Once again, because π_1, π_2 and π_3 are positive mappings, it follows from Markov chain theory that P, Q and R are also positive mappings. We now show it using only elementary arguments, because these arguments will be useful again later.

Assume P is not a positive mapping. Then there exists $(\omega_1^{(0)},\omega_2^{(0)},\omega_3^{(0)})\in\Omega_1\times\Omega_2\times\Omega_2$ such that $P(\{\omega_1^{(0)}\}\times\{\omega_2^{(0)}\}|\omega_3^{(0)})=0$. Equation (3.3) then implies that $P(\Omega_1\times\{\omega_2^{(0)}\}|\omega_3^{(0)})=0$. So for every $\omega_1\in\Omega_1$, $P(\{\omega_1\}\times\{\omega_2^{(0)}\}|\omega_3^{(0)})=0$. But then equation (3.3) implies that $P(\{\omega_1\}\times\Omega_2|\omega_3^{(0)})=0$. Since this holds for every $\omega_1\in\Omega_1$, $P(\Omega_1\times\Omega_2|\omega_3^{(0)})=0$, which is absurd since $P(\cdot|\omega_3^{(0)})$ is supposed to be a probability distribution. So P is a positive mapping.

Ideally, we would wish to define the optimal compromise between π_1, π_2 and π_3 as the joint distribution ϕ such that for every $(\omega_1,\omega_2,\omega_3)\in\Omega_1\times\Omega_2\times\Omega_3$

$$\begin{aligned}\phi(\{\omega_1\}\times\{\omega_2\}\times\{\omega_3\})&=P(\{\omega_1\}\times\{\omega_2\}|\omega_3)\phi(\Omega_1\times\Omega_2\times\{\omega_3\}) &(3.4)\\ &=Q(\{\omega_2\}\times\{\omega_3\}|\omega_1)\phi(\{\omega_1\}\times\Omega_2\times\Omega_3) &(3.5)\\ &=R(\{\omega_1\}\times\{\omega_3\}|\omega_2)\phi(\Omega_2\times\{\omega_2\}\times\Omega_3). &(3.6)\end{aligned}$$

Unfortunately, the existence of such an "optimal compromise" ϕ implies that π_1, π_2 and π_3 are compatible. First, one can show that for every $(\omega_1,\omega_2,\omega_3)\in\Omega_1\times\Omega_2\times\Omega_3$, $\phi(\{\omega_1\}\times\{\omega_2\}\times\{\omega_3\})>0$. The arguments are similar to those used above to show that P is a positive mapping. Rewriting equations (3.4) and (3.6) yields

$$\begin{aligned}\frac{\phi(\{\omega_1\}\times\{\omega_2\}\times\{\omega_3\})}{\phi(\Omega_1\times\{\omega_2\}\times\{\omega_3\})}&=\frac{P(\{\omega_1\}\times\{\omega_2\}|\omega_3)}{P(\Omega_1\times\{\omega_2\}|\omega_3)}=\frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3)+\frac{1}{2}\pi_2(\{\omega_2\}|\omega_1,\omega_3)\frac{P(\{\omega_1\}\times\Omega_2|\omega_3)}{P(\Omega_1\times\{\omega_2\}|\omega_3)} &(3.7)\\ &=\frac{R(\{\omega_1\}\times\{\omega_3\}|\omega_2)}{R(\Omega_1\times\{\omega_2\}|\omega_3)}=\frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3)+\frac{1}{2}\pi_3(\{\omega_3\}|\omega_1,\omega_2)\frac{R(\{\omega_1\}\times\Omega_3|\omega_2)}{R(\Omega_1\times\{\omega_3\}|\omega_2)}. &(3.8)\end{aligned}$$

Now combine equations (3.7) and (3.8):

$$\frac{1}{2}\pi_2(\{\omega_2\}|\omega_1,\omega_3)\frac{\phi(\{\omega_1\}\times\Omega_2\times\{\omega_3\})}{\phi(\Omega_1\times\{\omega_2\}\times\{\omega_3\})}=\frac{1}{2}\pi_3(\{\omega_3\}|\omega_1,\omega_2)\frac{\phi(\{\omega_1\}\times\{\omega_2\}\times\Omega_3)}{\phi(\Omega_1\times\{\omega_2\}\times\{\omega_3\})}. \tag{3.9}$$

As this holds for every $(\omega_1,\omega_2,\omega_3)\in\Omega_1\times\Omega_2\times\Omega_3$, it implies that $\pi_2\phi_{-2}=\pi_3\phi_{-3}$. A similar proof then shows that $\pi_3\phi_{-3}=\pi_1\phi_{-1}$. This means that if an "optimal compromise" ϕ exists, then π_1, π_2 and π_3 are compatible and no compromise was needed.

Similarly to what was done in the two-dimensional case, we avoid this difficulty by weakening the compatibility requirements: we no longer require $P(\cdot|\omega_3)$ to be the optimal compromise between π_1 and π_2 *for every* $\omega_3\in\Omega_3$ (as expressed by Eq. (3.3)), but only *on average* over $\omega_3\in\Omega_3$. So a compromise ϕ should still verify equation (3.4), but it would only need to verify this weakened version of equation (3.3) for every $(\omega_1,\omega_2)\in\Omega_1\times\Omega_2$:

$$\begin{aligned}&\sum_{\omega_3\in\Omega_3}P(\{\omega_1\}\times\{\omega_2\}|\omega_3)\phi(\Omega_1\times\Omega_2\times\{\omega_3\})\\&\quad=\sum_{\omega_3\in\Omega_3}\left[\frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3)P(\Omega_1\times\{\omega_2\}|\omega_3)+\frac{1}{2}\pi_2(\{\omega_2\}|\omega_1,\omega_3)P(\{\omega_1\}\times\Omega_2|\omega_3)\right]\phi(\Omega_1\times\Omega_2\times\{\omega_3\}).\end{aligned} \tag{3.10}$$

Because equation (3.4) is still expected to hold, equation (3.10) is equivalent to

$$\phi_{-3}(\{\omega_1\}\times\{\omega_2\})=\frac{1}{2}\sum_{\omega_3\in\Omega_3}\pi_1(\{\omega_1\}|\omega_2,\omega_3)\phi_{-1}(\{\omega_2\}\times\{\omega_3\})+\frac{1}{2}\sum_{\omega_3\in\Omega_3}\pi_2(\{\omega_2\}|\omega_1,\omega_3)\phi_{-2}(\{\omega_1\}\times\{\omega_3\}). \tag{3.11}$$

So the requirement boils down to $2\phi_{-3}=(\pi_1\phi_{-1})_{-3}+(\pi_2\phi_{-2})_{-3}$. Of course, we symmetrically require $2\phi_{-1}=(\pi_2\phi_{-2})_{-1}+(\pi_3\phi_{-3})_{-3}$ and $2\phi_{-2}=(\pi_1\phi_{-1})_{-2}+(\pi_3\phi_{-3})_{-2}$ as well.

To sum this part of the discussion up, the necessity of weakening our "ideal" requirements for "optimal compatibility" made us downgrade from a requirement about a "joint" three-dimensional distribution ϕ to requirements about its $(3-1)$-marginal distributions ϕ_{-1}, ϕ_{-2} and ϕ_{-3}. Definition 2.5 is just another formulation of these requirements, which are taken to define a compatible sequence of $(r-1)$-dimensional distributions.

Proposition 2.11 shows that in cases where there exists a unique Gibbs compromise, even with this weakened set of requirements for compatibility, there exists only one compatible sequence of $(r-1)$-dimensional distributions, so we may not strenghten it if there is to be any solution. Indeed, in cases where no Gibbs compromise exists, no compatible set of $(r-1)$-dimensional distributions exists either!

3.2. Weaker definitions of compromises are inconvenient

As was shown in the previous subsection, the requirements given by Definition 2.5 for the compatibility of a sequence of $(r-1)$-dimensional distributions with a given sequence of Markov kernels cannot be strengthened. The following shows they cannot be weakened either.

3.2.1. Why we need some notion of compatibility: example in 2 dimensions

Why bother with the compatibility of $(r-1)$-dimensional distributions and not simply minimize the functional E_λ of Definition 2.7 over all distributions absolutely continuous with respect to λ ? The following two-dimensional example shows that doing so yields unsatisfactory results.

Consider the following situation: $\Omega_1=\Omega_2=\{0,1\}$ and $\mathcal{A}_1=\mathcal{A}_2=\{\emptyset,\{0\},\{1\},\{0,1\}\}$. Let X_1 (resp. X_2) be the identity function on Ω_1 (resp. Ω_2). Both X_1 and X_2 are measurable functions (and thus random variables

when $\mathcal{A}_1 \otimes \mathcal{A}_2$ is endowed with a probability measure). Define the following Markov kernels:

$$\pi_1(X_1 = 1|\omega_2) = \mathbf{1}_{\{\omega_2=0\}} + 1/2\,\mathbf{1}_{\{\omega_2=1\}}; \tag{3.12}$$
$$\pi_2(X_2 = 1|\omega_1) = 1/2\,\mathbf{1}_{\{\omega_1=0\}} + \mathbf{1}_{\{\omega_1=1\}}. \tag{3.13}$$

Let λ be the counting measure. Denote by π_C the optimal compromise with respect to λ (minimizing E_λ over all compromises) and π_E the distribution on $\mathcal{A}_1 \otimes \mathcal{A}_2$ that minimizes E_λ over all distributions on $\mathcal{A}_1 \otimes \mathcal{A}_2$. We have $E_\lambda(\pi_C) = 2/25 > E_\lambda(\pi_E) = 1/15$ and

$$\pi_C(\{\omega_1\} \times \{\omega_2\}) = 1/10\left(\mathbf{1}_{\{(\omega_1,\omega_2)=(0,0)\}} + \mathbf{1}_{\{(\omega_1,\omega_2)=(1,0)\}} + 3\,\mathbf{1}_{\{(\omega_1,\omega_2)=(0,1)\}} + 5\,\mathbf{1}_{\{(\omega_1,\omega_2)=(1,1)\}}\right); \tag{3.14}$$
$$\pi_E(\{\omega_1\} \times \{\omega_2\}) = 1/30\left(3\,\mathbf{1}_{\{(\omega_1,\omega_2)=(0,0)\}} + \mathbf{1}_{\{(\omega_1,\omega_2)=(1,0)\}} + 11\,\mathbf{1}_{\{(\omega_1,\omega_2)=(0,1)\}} + 15\,\mathbf{1}_{\{(\omega_1,\omega_2)=(1,1)\}}\right). \tag{3.15}$$

As π_C is the unique Gibbs compromise between π_1 and π_2, its marginals are the only compatible marginals: $(\pi_C)_{-1}(X_2 = 1) = 4/5$ and $(\pi_C)_{-2}(X_1 = 1) = 3/5$. The marginals of π_E are noticeably different: $(\pi_E)_{-1}(X_2 = 1) = 13/15$ and $(\pi_E)_{-2}(X_1 = 1) = 8/15$.

Observe that according to π_1, $X_2 = 0$ implies $X_1 = 1$ but that according to π_2, $X_1 = 1$ implies that $X_2 = 1 \neq 0$. This discrepancy is a major source of incompatibility between the two Markov kernels. So, as π_E makes both $X_1 = 1$ and $X_2 = 0$ less likely than π_C, it "ignores the inconsistent parts" of π_1 and π_2 to some extent. Therefore, if the marginals are not set in advance (say, by imposing compatibility with the conditionals in the sense of Def. 2.5), one may "cheat" by having the marginals disadvantage inconvenient values for the parameters.

3.2.2. Why the compatibility requirements can hardly be weakened: example in 3 dimensions

In the two-dimensional case, because of Proposition 3.2, Definitions 3.1 and 2.5 give the same meaning to the concept of compatibility of marginals, so Definition 2.5 may be thought of as a generalization of Definition 3.1 to cases with more than two dimensions. However, another generalization of the latter definition is possible. To avoid confusion, this other generalization will be called *weak compatibility*. In the following, the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ is defined as in Section 2.1.

Definition 3.3. A sequence of $(r-1)$-dimensional distributions $(m_{\neq i})_{i\in[\![1,r]\!]}$ is *weakly compatible* with a sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if equation (3.2) holds.

Proposition 3.2 means that compatibility in the sense of Definition 2.5 implies weak compatibility in the sense of Definition 3.3, hence its denomination as "weak".

Using the concept of weak compatibility of a sequence of $(r-1)$-marginal distributions, we define weak compromises and the optimal weak compromise as analogues to compromises and optimal compromises, respectively.

Definition 3.4. A probability distribution P on $\bigotimes_{i\in[\![1,r]\!]} \mathcal{A}_i$ is called a *weak compromise* between the Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if these two conditions are verified:

(1) for every $i \in [\![1, r]\!]$, $\pi_i P_{-i}$ is absolutely continuous with respect to P;
(2) the sequence $(P_{-i})_{i\in[\![1,r]\!]}$ of P's $(r-1)$-marginal distributions is weakly compatible with $(\pi_i)_{i\in[\![1,r]\!]}$.

Definition 3.5. Let λ be a positive measure on $\mathcal{A}$. Let P be a weak compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ that is absolutely continuous with respect to λ. P is called an *optimal weak compromise* with respect to λ between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if it minimizes the functional E_λ over all compromises between $(\pi_i)_{i\in[\![1,r]\!]}$ that are absolutely continuous with respect to λ. E_λ is defined by equation (2.13).

Because, for any positive measure λ on $\mathcal{A}$, the set of all weak compromises absolutely continuous with respect to λ includes the set of all compromises absolutely continuous with respect to λ, an optimal weak compromise

with respect to λ makes the functional E_λ no greater than an optimal compromise with respect to λ. However, as shown in the following example with $r = 3$, optimal weak compromises may have undesirable behavior.

Assume $\Omega_1 = \Omega_2 = \Omega_3 = \{0, 1\}$ and $\mathcal{A}_1 = \mathcal{A}_2 = \mathcal{A}_3 = \{\emptyset, \{0\}, \{1\}, \{0, 1\}\}$. Let X_1 (resp. X_2, X_3) be the identity function on Ω_1 (resp. Ω_2, Ω_3).

Consider the following Markov kernels. For every $(\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$

$$\pi_1(X_1 = 1|\omega_2, \omega_3) = 1/2; \tag{3.16}$$
$$\pi_2(X_2 = 1|\omega_1, \omega_3) = 1/2; \tag{3.17}$$
$$\pi_3(X_3 = 1|\omega_1, \omega_2) = \mathbf{1}_{\{\omega_1=\omega_2\}} + 1/2\,\mathbf{1}_{\{\omega_1\neq\omega_2\}}. \tag{3.18}$$

Notice that, provided ω_3 is known, π_1 and π_2 are compatible Markov kernels. The unique probability distribution on $\mathcal{A}_1 \otimes \mathcal{A}_2$ that fits both π_1 and π_2 (conditional to ω_3) verifies for every $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$

$$P(\{\omega_1\} \times \{\omega_2\}|\omega_3) = 1/4. \tag{3.19}$$

Thus, any joint distribution fitting the Markov kernel P would make X_1, X_2 and X_3 mutually independent. Unfortunately, no such joint distribution could fit π_3, but we may expect compromises between π_1, π_2 and π_3 to retain the independence of X_1 and X_2.

Let λ be the counting measure on $\mathcal{A}$.

Denote by π_C the (unique) optimal compromise between π_1, π_2 and π_3 with respect to λ: we have $E_\lambda(\pi_C) = 1/48 \approx 0.021$. For every $(\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$

$$\pi_C(\{\omega_1\} \times \{\omega_2\} \times \{\omega_3\}) = \frac{1}{24}\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=0\}} + \frac{1}{12}\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=0\}} + \frac{5}{24}\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=1\}} + \frac{1}{6}\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=1\}}. \tag{3.20}$$

Notably, its third 2-marginal distribution $(\pi_C)_{-3}$ verifies for every $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$

$$(\pi_C)_{-3}(\{\omega_1\} \times \{\omega_2\}) = 1/4. \tag{3.21}$$

So, as expected, π_C retains the independence of X_1 and X_2. Because π_C is the unique Gibbs compromise between π_1, π_2 and π_3, Proposition 2.11 implies any other compromise between π_1, π_2 and π_3 also retains this property.

Let us now consider an optimal weak compromise π_W between π_1, π_2 and π_3 with respect to λ. Numerical computation gives us the following approximation, with $E_\lambda(\pi_W) \approx 0.019$. For every $(\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$,

$$\pi_W(\{\omega_1\} \times \{\omega_2\} \times \{\omega_3\}) \approx 0.04\,\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=0\}} + 0.10\,\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=0\}} + 0.19\,\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=1\}} + 0.17\,\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=1\}}. \tag{3.22}$$

Its third 2-marginal distribution $(\pi_W)_{-3}$ is approximately

$$(\pi_W)_{-3}(\{\omega_1\} \times \{\omega_2\}) \approx 0.23\,\mathbf{1}_{\{\omega_1=\omega_2\}} + 0.27\,\mathbf{1}_{\{\omega_1\neq\omega_2\}}. \tag{3.23}$$

Thus the independence of X_1 and X_2 is lost. Therefore, weak compatibility is no adequate notion of compatibility. As a matter of fact, although $(\pi_W)_{-3}$ and $(\pi_C)_{-3}$ share the same marginals, that is

$$(\pi_W)_{-3}(\{\omega_1\} \times \Omega_2) = 1/2, \tag{3.24}$$
$$(\pi_W)_{-3}(\Omega_1 \times \{\omega_2\}) = 1/2, \tag{3.25}$$

$(\pi_W)_{-3}$ slightly disadvantages the event $X_1 = X_2$, which is where the incompatibility between π_3 and the pair (π_1, π_2) is most obvious: according to π_3, $X_1 = X_2$ implies $X_3 = 1$, so conversely, $X_3 = 0$ should imply

$X_1 \neq X_2$, when in fact π_1 and π_2 state that even given $\omega_3 = 0$, $\{X_1 \neq X_2\}$ only happens with probability 1/2. On the other hand, according to π_3, if $X_1 \neq X_2$, then X_3 can with equal probability be 0 or 1, which matches π_1 and π_2 better.

4. Optimal compromise between objective posterior conditional distributions in Gaussian process regression

Kriging is a surrogate model used to emulate a real-valued function on a spatial domain $\mathcal{D}$ when said function can only be evaluated on a finite subset of $\mathcal{D}$ called "design set". The "Kriging prediction" is the mean function of the process taken conditionally to all known values of the emulated function, *i.e.* the values at the points in the design set. The main advantage of the framework is its natural way of representing uncertainty about the value of the function at unobserved points [31]. Prediction does not consist of a single value but of a complete Normal distribution. "Kriging" is the name given to the framework in the geostatistical literature [18], but is also frequently used in the context of computer experiments and machine learning under the label "Gaussian process regression" [26]. In this work, we focus on Simple Kriging, where the Gaussian process is assumed to be stationary with known mean, as opposed to Universal Kriging, which incorporates an unknown mean function.

The probability distribution of a stationary Gaussian process is characterized by a variance parameter and a correlation function (also known as "correlation kernel") which itself depends on parameters. So one should deal with uncertainty about model parameters.

The problem is "notoriously difficult", as highlighted by Kennedy and O'Hagan [20], because the likelihood function may often be quite flat [23]. In a Bayesian framework, this uncertainty is represented by a prior distribution on the parameters.

4.1. Issues raised by objective prior elicitation for Gaussian processes

Let $Y(\boldsymbol{x})$, $\boldsymbol{x} \in \mathcal{D}$ be a real-valued random field on a bounded subset $\mathcal{D}$ of $\mathbb{R}^r$. We assume Y is Gaussian with zero mean (or known mean) and with covariance of the form $\mathrm{Cov}(Y(\boldsymbol{x}), Y(\boldsymbol{x}')) = \sigma^2 K_{\boldsymbol{\theta}}(\boldsymbol{x} - \boldsymbol{x}')$. σ^2 thus denotes the variance of the Gaussian process and $\boldsymbol{\theta} \in (0, +\infty)^r$, hereafter named the "vector of correlation lengths", is the vector of scaling parameters used by the chosen class of correlation kernels $K_{\boldsymbol{\theta}}$.

Consider a set of $n \in \mathbb{N}$ points $(\boldsymbol{x}^{(i)})_{i \in [\![1,n]\!]}$ belonging to the domain $\mathcal{D}$. This set is called the design set and Y is observed at all points of this set. Let $\boldsymbol{Y}$ be the Gaussian vector $(Y(\boldsymbol{x}^{(i)}))_{i \in [\![1,n]\!]}$ and let $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ be its correlation matrix: the distribution of $\boldsymbol{Y}$ is therefore $\mathcal{N}(\boldsymbol{0}_n, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$.

Let $\boldsymbol{y}$ be the vector of observations. When applied to a matrix, $|\cdot|$ refers to its determinant.

With these notations, the likelihood of the parameters σ^2 and $\boldsymbol{\theta}$ is

$$L(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\theta}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} \boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}\right\} . \tag{4.1}$$

The reference prior with parameter ordering $\sigma^2 \prec \boldsymbol{\theta}$ is given by Berger *et al.* [6]:

$$\pi(\sigma^2, \boldsymbol{\theta}) = \pi(\sigma^2 | \boldsymbol{\theta}) \pi(\boldsymbol{\theta}) \quad \text{with} \quad \pi(\sigma^2 | \boldsymbol{\theta}) \propto 1/\sigma^2. \tag{4.2}$$

The distribution $\pi(\sigma^2 | \boldsymbol{\theta})$ has infinite mass: it is an improper prior.

Let us integrate σ^2 out of the likelihood (4.1):

$$L^1(\boldsymbol{y} \mid \boldsymbol{\theta}) \propto \int_0^\infty L(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2 \mid \boldsymbol{\theta})\, d(\sigma^2) = \left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right)^{-1} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}\right)^{-\frac{n}{2}} . \tag{4.3}$$

Ren *et al.* [27] provide the reference prior $\pi(\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is regarded as a single multi-dimensional parameter. It is proportional to the square root of the determinant of the $r \times r$ matrix with (i,j)th entry

$$\operatorname{Tr}\left[\left(\frac{\partial}{\partial\theta_i}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\left(\frac{\partial}{\partial\theta_j}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right] - \frac{1}{n}\operatorname{Tr}\left[\frac{\partial}{\partial\theta_i}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]\operatorname{Tr}\left[\frac{\partial}{\partial\theta_j}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]. \tag{4.4}$$

However, this method has the disadvantage of requiring the use of a multi-dimensional Jeffreys-rule prior distribution, which may show the sort of undesirable behavior mentioned in the introduction.

Alternatively, we could draw inspiration from the one-dimensional case in the following way. Suppose that we know every entry of $\boldsymbol{\theta}$ except one, θ_i. Then, according to equation (4.4), the prior density on θ_i knowing all entries θ_j $(j \neq i)$ would be

$$\pi_i(\theta_i \mid \theta_j\ \forall j \neq i) \propto \sqrt{\operatorname{Tr}\left[\left(\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]^2}. \tag{4.5}$$

The density functions $\pi_i(\theta_i \mid \theta_j\ \forall j \neq i)$ $(i \in [\![1,r]\!])$ define Markov kernels. Indeed, they are continuous with respect to the θ_j $(j \neq i)$ and are probability densities with respect to the Lebesgue measure. They are unfortunately likely to violate the necessary condition for compatibility given by equation (2.6).

Let us now consider the corresponding posterior conditional densities $\pi_i(\theta_i \mid \boldsymbol{y},\ \theta_j\ \forall j \neq i)$ $(i \in [\![1,r]\!])$. Just like their prior counterparts, they are likely to violate the necessary condition for compatibility. However, each of them represents our opinion about one parameter if all others were known. This is a setting where the results of Section 2 can be applied in order to find the optimal compromise between the Markov kernels $\mathbb{R}^{r-1} \times \mathcal{B}(\mathbb{R})$ they define. This optimal compromise will then be taken as posterior probability of the vector $\boldsymbol{\theta}$. In the following, we describe settings in which there exists a single Gibbs compromise between these Markov kernels. Theorem 2.13 then asserts it is the optimal compromise. We call this compromise the Gibbs reference posterior distribution because of its link to the reference posterior distribution in settings with a one-dimensional parameter $\boldsymbol{\theta}$.

However, even though we call it a "posterior" distribution, it is unclear whether there exists a prior distribution from which it could be derived using Bayes' rule. Denote by $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$ the probability density with respect to the Lebesgue measure of the Gibbs reference posterior distribution. Bayes' rule requires that in case a (proper or improper) prior density $\pi_G(\boldsymbol{\theta})$ exists, there should also exist a function $\tilde{L}(\boldsymbol{y})$ such that, for almost every $\boldsymbol{\theta} \in \mathbb{R}^r$ in the sense of the Lebesgue measure,

$$\frac{\pi_G(\boldsymbol{\theta}|\boldsymbol{y})}{L^1(\boldsymbol{y}|\boldsymbol{\theta})} = \frac{\pi_G(\boldsymbol{\theta})}{\tilde{L}(\boldsymbol{y})}. \tag{4.6}$$

As we have no explicit expression of $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$, we have no way to check whether equation (4.6) holds or not.

In this section, we establish that, whenever Matérn anisotropic geometric or tensorized kernels with known smoothness parameter ν are used, under certain conditions to be detailed later, there exists a unique Gibbs compromise between the reference posterior conditionals, which thanks to Theorem 2.13 is the optimal compromise. Henceforth, it will be called "Gibbs reference posterior distribution", even though this "posterior" has not been derived from a prior distribution using the Bayes rule.

All proofs for this section can be found in Appendix B.

4.2. Definitions

In this work, we use the following convention for the Fourier transform: the Fourier transform $\widehat{g}$ of a smooth function $g : \mathbb{R}^r \to \mathbb{R}$ verifies $g(\boldsymbol{x}) = \int_{\mathbb{R}^r} \widehat{g}(\boldsymbol{\omega}) e^{i\langle\boldsymbol{\omega}|\boldsymbol{x}\rangle} \mathrm{d}\boldsymbol{\omega}$ and $\widehat{g}(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} g(\boldsymbol{x}) e^{-i\langle\boldsymbol{\omega}|\boldsymbol{x}\rangle} \mathrm{d}\boldsymbol{x}$.

Let us set up a few notations.

(a) $\mathcal{K}_\nu$ is the modified Bessel function of second kind with parameter ν;

(b) $K_{r,\nu}$ is the r-dimensional Matérn isotropic covariance kernel with variance 1, correlation length 1 and smoothness $\nu \in (0, +\infty)$ and $\widehat{K}_{r,\nu}$ is its Fourier transform:
(i) $\forall \boldsymbol{x} \in \mathbb{R}^r$,

$$K_{r,\nu}(\boldsymbol{x}) = \frac{1}{\Gamma(\nu)2^{\nu-1}} \left(2\sqrt{\nu}\|\boldsymbol{x}\|\right)^{\nu} \mathcal{K}_{\nu}\left(2\sqrt{\nu}\|\boldsymbol{x}\|\right); \tag{4.7}$$

(ii) $\forall \boldsymbol{\omega} \in \mathbb{R}^r$,

$$\widehat{K}_{r,\nu}(\boldsymbol{\omega}) = \frac{M_r(\nu)}{(\|\boldsymbol{\omega}\|^2 + 4\nu)^{\nu+\frac{r}{2}}} \text{ with } M_r(\nu) = \frac{\Gamma(\nu + \frac{r}{2})(2\sqrt{\nu})^{2\nu}}{\pi^{\frac{r}{2}}\Gamma(\nu)}. \tag{4.8}$$

(c) $K^{\text{tens}}_{r,\nu}$ is the r-dimensional Matérn tensorized covariance kernel with variance 1, correlation length 1 and smoothness $\nu \in \mathbb{R}_+$ and $\widehat{K}^{\text{tens}}_{r,\nu}$ is its Fourier transform:
(i) $\forall \boldsymbol{x} \in \mathbb{R}^r$,

$$K^{\text{tens}}_{r,\nu}(\boldsymbol{x}) = \prod_{j=1}^{r} K_{1,\nu}(\boldsymbol{x}_j); \tag{4.9}$$

(ii) $\forall \boldsymbol{\omega} \in \mathbb{R}^r$,

$$\widehat{K}^{\text{tens}}_{r,\nu}(\boldsymbol{\omega}) = \prod_{j=1}^{r} \widehat{K_{1,\nu}}(\boldsymbol{\omega}_j). \tag{4.10}$$

(d) if $\boldsymbol{t} \in \mathbb{R}^r$, $\frac{\boldsymbol{t}}{\boldsymbol{\theta}} = \left(\frac{t_1}{\theta_1}, ..., \frac{t_r}{\theta_r}\right)$ and $\boldsymbol{t}\,\boldsymbol{\mu} = (t_1\mu_1, ..., t_r\mu_r)$.

We define the Matérn geometric anisotropic covariance kernel with variance parameter σ^2, correlation lengths $\boldsymbol{\theta}$ (resp. inverse correlation lengths $\boldsymbol{\mu}$) and smoothness ν as the function $\boldsymbol{x} \mapsto \sigma^2 K_{r,\nu}\left(\frac{\boldsymbol{x}}{\boldsymbol{\theta}}\right)$ (resp. $\boldsymbol{x} \mapsto \sigma^2 K_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})$).

Similarly, we define the Matérn tensorized covariance kernel with variance parameter σ^2, correlation lengths $\boldsymbol{\theta}$ (resp. inverse correlation lengths $\boldsymbol{\mu}$) and smoothness ν as the function $\boldsymbol{x} \mapsto \sigma^2 K^{\text{tens}}_{r,\nu}\left(\frac{\boldsymbol{x}}{\boldsymbol{\theta}}\right)$ (resp. $\boldsymbol{x} \mapsto \sigma^2 K^{\text{tens}}_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})$).

Thanks to Proposition 2.15, we may choose any parametrization we wish for the Matérn correlation kernels. We have found that the parametrization involving inverse correlation lengths makes proofs easier.

Several key passages in the proofs (to be found in Appendix B) involve a technical assumption on the design set:

Definition 4.1. A design set is said to have coordinate-distinct points, or simply to be coordinate-distinct, if for any distinct points in the set $\boldsymbol{x}$ and $\boldsymbol{x'}$, every component of the vector $\boldsymbol{x} - \boldsymbol{x'}$ differs from 0.

Most randomly sampled design sets almost surely have coordinate-distinct points – for instance Latin Hypercube Sampling (LHS). Cartesian product design sets, however, do not.

4.3. Main result

The result is valid for Simple Kriging models with the following characteristics:

(a) the design set contains n coordinate-distinct points in $\mathbb{R}^r$ (n and r are positive integers);

(b) the covariance function is Matérn anisotropic geometric or tensorized with variance parameter $\sigma^2 > 0$, smoothness parameter ν and vector of correlation lengths (resp. inverse correlation lengths) $\boldsymbol{\theta} \in (0,+\infty)^r$ (resp. $\boldsymbol{\mu} \in (0,+\infty)^r$);
(c) one of the following conditions is verified:
  (i) $\nu \in (0,1)$ and $n > 1$ and the Matérn kernel is tensorized;
  (ii) $\nu \in (1,2)$ and $n > r+2$;
  (iii) $\nu \in (2,3)$ and $n > r(r+1)/2 + 2r + 3$.

Theorem 4.2. *In a Simple Kriging model with the characteristics described above, there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that, for any $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, there exists a unique Gibbs compromise $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$ (resp. $\pi_G(\boldsymbol{\mu}|\boldsymbol{y})$) between the reference posterior conditionals $\pi_i(\theta_i|\boldsymbol{y},\boldsymbol{\theta}_{-i})$ (resp. $\pi_i(\mu_i|\boldsymbol{y},\boldsymbol{\mu}_{-i})$). It is the unique stationary distribution of the Markov kernel $P_{\boldsymbol{y}} : (0,+\infty)^r \times \mathcal{B}((0,+\infty)^r) \to [0,1]$ defined by*

$$P_{\boldsymbol{y}}(\boldsymbol{\theta}^{(0)}, \mathrm{d}\boldsymbol{\theta}) = \frac{1}{r}\sum_{i=1}^{r} \pi_i(\theta_i|\boldsymbol{y}, \boldsymbol{\theta}_{-i}^{(0)})\mathrm{d}\theta_i \; \delta_{\boldsymbol{\theta}_{-i}^{(0)}}(\mathrm{d}\boldsymbol{\theta}_{-i})$$

$$\textit{(resp. } P_{\boldsymbol{y}}(\boldsymbol{\mu}^{(0)}, \mathrm{d}\boldsymbol{\mu}) = \frac{1}{r}\sum_{i=1}^{r} \pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}^{(0)})\mathrm{d}\mu_i \; \delta_{\boldsymbol{\mu}_{-i}^{(0)}}(\mathrm{d}\boldsymbol{\mu}_{-i})\textit{).}$$

The Markov kernel $P_{\boldsymbol{y}}$ is uniformly ergodic: $\lim_{n\to\infty} \sup_{\boldsymbol{\theta}^{(0)} \in (0,+\infty)^r} \|P_{\boldsymbol{y}}^n(\boldsymbol{\theta}^{(0)}, \cdot) - \pi_G(\cdot|\boldsymbol{y})\|_{TV} = 0$ (resp. $\lim_{n\to\infty} \sup_{\boldsymbol{\mu}^{(0)} \in (0,+\infty)^r} \|P_{\boldsymbol{y}}^n(\boldsymbol{\mu}^{(0)}, \cdot) - \pi_G(\cdot|\boldsymbol{y})\|_{TV} = 0$), where $\|\cdot\|_{TV}$ is the total variation norm.

Remark 4.3. The reference posterior conditionals are invariant by reparametrization, so the Markov kernel $P_{\boldsymbol{y}}$ does not depend on whether the chosen parametrization uses correlation lengths $\boldsymbol{\theta}$ or inverse correlation lengths $\boldsymbol{\mu}$. Due to Proposition 2.15, the Gibbs compromise does not either. The parametrization using inverse correlation lengths $\boldsymbol{\mu}$ is more convenient for proving this theorem, however.

Notice that in such a Kriging model, the vector of observations $\boldsymbol{y}$ almost surely belongs to $\mathbb{R}^n \setminus \mathcal{H}$, so this assumption is of no practical consequence. Theorem 4.2 therefore asserts that the Gibbs compromise between the incompatible conditionals $\pi_i(\mu_i|\boldsymbol{y},\boldsymbol{\mu}_{-i})$ exists, is unique, and can be sampled from using Potentially Incompatible Gibbs Sampling (PIGS). In the following, it is called "Gibbs reference posterior distribution".

4.4. Using the Gibbs reference posterior distribution

Let $\boldsymbol{x}_0$ be a point in the domain $\mathcal{D}$ that does not belong to the design set. Denote by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}$ the correlation matrix between $Y(\boldsymbol{x}_0)$ and $\boldsymbol{Y}$, and by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0}$ its transpose the correlation matrix between $\boldsymbol{Y}$ and $Y(\boldsymbol{x}_0)$.

Theorem 4.1.2. (case 4) of Santner *et al.* [31] provides this useful result for prediction:

Proposition 4.4. *Conditionally to $\boldsymbol{Y} = \boldsymbol{y}$ and assuming $\boldsymbol{\theta}$ is known, the random variable Z_0 defined below follows the Student t-distribution with n degrees of freedom.*

$$Z_0 := \sqrt{\frac{n}{\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}}} \frac{Y(\boldsymbol{x}_0) - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{y}}{\sqrt{1 - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0}}}.$$

Remark 4.5. If n exceeds 30, it is usually accepted that the Student t-distribution with n degrees of freedom can be approximated by the standard Normal distribution. As this threshold should be exceeded in practical cases, we would recommend performing all computations as though the Student t-distribution were Normal.

This proposition implicitly gives the distribution of $y_0 := Y(\boldsymbol{x}_0)$ conditionally to $\boldsymbol{Y} = \boldsymbol{y}$ and $\boldsymbol{\theta}$. For later reference, denote it by $L^1(y_0|\boldsymbol{y},\boldsymbol{\theta})$. In practice, when $\boldsymbol{\theta}$ is unknown, the distribution of $y_0 = Y(\boldsymbol{x}_0)$ conditionally

to $\boldsymbol{Y} = \boldsymbol{y}$ can be obtained once $\boldsymbol{\theta}$ has been sampled from the Gibbs reference posterior distribution:

$$P(y_0|\boldsymbol{y}) := \int L^1(y_0|\boldsymbol{y}, \boldsymbol{\theta})\pi_G(\boldsymbol{\theta}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}.$$

Its cdf can be approximated by averaging the cdfs of the Student t-distributions (or their Normal approximations) corresponding to every point in the sample.

5. Comparisons between the MLE and MAP estimators

To illustrate the inferential performance of the Gibbs reference posterior distribution, let us introduce the Maximum A Posteriori estimator (MAP). It takes the value $\hat{\boldsymbol{\theta}}_{\mathrm{MAP}}$ of $\boldsymbol{\theta}$ where the density with respect to the Lebesgue measure of the Gibbs reference posterior distribution is largest. We contrast it with the Maximum Likelihood Estimator (MLE) $\hat{\boldsymbol{\theta}}_{\mathrm{MLE}}$ which does the same with the likelihood function.

5.1. Methodology

In this section, we compare the MLE and MAP estimators for accuracy and robustness.

Our test cases are three-dimensional Gaussian processes with Matérn anisotropic geometric correlation kernels with smoothness 5/2. Their mean is the null function, which only leaves us with the matter of estimating their correlation length for each dimension.

We use uniform designs: our observation points are randomly generated according to the uniform distribution on a cube with side length 1.

In order to measure the performance of our estimators, we define a suitable distance between two vectors of correlation lengths. Then the error of an estimator is defined as its distance to the "true" vector of correlation lengths.

Let g be the function such that for any t in $(-1, 1)$, $g(t) = \operatorname{argtanh}(t)$ and $g(-1) = g(1) = 0$. We use the convention that, for any matrix $\boldsymbol{M}$ with elements in $[0, 1]$, $g(\boldsymbol{M})$ is the matrix resulting from applying g to every element of $\boldsymbol{M}$.

Definition 5.1. For a given design set, the distance between two vectors of correlation lengths $\boldsymbol{\theta}^1$ and $\boldsymbol{\theta}^2$ is $\|g(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^1}) - g(\boldsymbol{\Sigma}_{\boldsymbol{\theta}^2})\|$, where $\|\cdot\|$ denotes the Frobenius norm.

This distance involves applying the Fisher transformation [17] (that is, the inverse hyperbolic tangent function) to every (non-unitary) correlation coefficient in both associated correlation matrices. This is a variance-stabilizing transformation. For any random variables U and V following the normal distribution with mean 0 and variance 1, let ρ denote the correlation coefficient between U and V $(-1 < \rho < 1)$. If (U_i, V_i) $(1 \leqslant i \leqslant N)$ are independent copies of (U, V), then $\hat{\rho} = \sum_{i=1}^{N} U_i V_i / n$ is a random variable and $\operatorname{argtanh}(\hat{\rho})$ follows the normal distribution with mean $\operatorname{argtanh}(\rho)$ and variance $1/(N-3)$. So the variance of $\operatorname{argtanh}(\hat{\rho})$ does not depend on ρ, whereas the variance of $\hat{\rho}$ does and goes to zero for $|\rho| \to 1$. Involving the Fisher transformation in the definition of the distance between two vectors of correlation lengths is therefore a way to assert that vectors of correlation lengths can be far apart even if they both lead to highly correlated observations.

This allows us to make sure errors made when estimating near-1 correlation coefficients are no less taken into account than errors made when estimating near-0 correlation coefficients.

Let us choose a "true" vector of correlation lengths (and also a variance parameter, but this parameter has no effect on either the MLE or the MAP). Then we need to:

(1) Sample n points randomly according to the uniform distribution on the unit cube (in the following, $n = 30$).
(2) Generate the observations of the Gaussian process at the sampled points according to the selected "true" variance and correlation lengths.

TABLE 1. RMSE (where the error is measured in terms of the distance in Def. 5.1) of the MLE and MAP estimators for several "true" vectors of correlation lengths. The last column displays in percents the decrease of the RMSE of the MAP estimator with respect to the MLE.

Corr. lengths	MLE	MAP	−(%)
0.4 – 0.8 – 0.2	3.49	2.97	15
0.5 – 0.5 – 0.5	4.00	3.46	13
0.7 – 1.3 – 0.4	4.02	3.64	9
0.8 – 0.3 – 0.6	3.75	3.26	13
0.8 – 1.0 – 0.9	4.65	4.18	10

(3) Sample the vector of correlation lengths according to the Gibbs reference posterior distribution $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$ through PIGS.
(4) Compute the MLE and the MAP of the vector of correlation lengths and their errors.
(5) Repeat steps 1 to 4 $m-1$ times (in the following, $m = 500$).

This method allows us to derive an approximate distribution of the errors of both estimators when both the realization of the Gaussian process and the design set vary. Thus, we get to test the robustness of both estimators *versus* the variability of both the Gaussian process and the choice of design set.

5.2. Results

This subsection provides results obtained on three-dimensional Gaussian processes with null mean function and Matérn anisotropic geometric correlation kernels with smoothness 5/2. The results are divided by "true" vectors of correlation lengths. In each case, we give in Table 1 the empirical Root Mean Square Errors (RMSEs) of both MLE and MAP estimators as functions of varying instances of the Gaussian process and uniform design sets on the unit cube.

Most of the "true" vectors of correlation lengths featured in Table 1 were selected in a way to showcase the behavior of both estimators in strongly anisotropic cases, but one $(0.5-0.5-0.5)$ also showcases their behavior if the true kernel is actually isotropic. And the final one $(0.8-1-0.9)$ is used to illustrate the performance in the case of a strongly correlated Gaussian process: this case is fundamentally different from all others, because the Matérn anisotropic geometric family of correlation kernels is designed in such a way that the correlation length with greatest influence is the lowest. Informally speaking, it is enough for one correlation length to be near zero to make the whole process very uncorrelated, even should all other correlation lengths be very high.

In all studied cases, the MAP estimator was more robust than the MLE estimator: its RMSE was between 9% and 15% lower, as showcased in Table 1.

To get a better sense of the distribution of the error when the design set and the realization of the Gaussian process vary, we give in Figure 1 violin plots of the errors in the two most extreme case: very low correlation (0.4 – 0.8 – 0.2) and very high correlation (0.8 – 1.0 – 0.9).

6. COMPARISON OF THE MLE AND MAP PLUG-IN DISTRIBUTIONS AND THE FULL POSTERIOR PREDICTIVE DISTRIBUTION

6.1. Methodology

We use the same test cases as before. In this section, our goal is to assess the accuracy of prediction intervals associated with both estimators and with the full posterior distribution (FPD). We consider 95% intervals: the lower bound is the 2.5% quantile and the upper bound the 97.5% quantile of plug-in distributions $\hat{P}_{\mathrm{MLE}}(y_0|\boldsymbol{y}) = L^1(y_0|\boldsymbol{y}, \hat{\boldsymbol{\theta}}_{\mathrm{MLE}})$, $\hat{P}_{\mathrm{MAP}}(\boldsymbol{y}_0|\boldsymbol{y}) = L^1(y_0|\boldsymbol{y}, \hat{\boldsymbol{\theta}}_{\mathrm{MAP}})$ and $P(\boldsymbol{y}_0|\boldsymbol{y}) = \int L^1(y_0|\boldsymbol{y}, \boldsymbol{\theta})\, \pi_G(\boldsymbol{\theta}|\boldsymbol{y})\, \mathrm{d}\boldsymbol{\theta}$. For the sake of comprehensiveness, we also consider predictive intervals associated with the "true" predictive

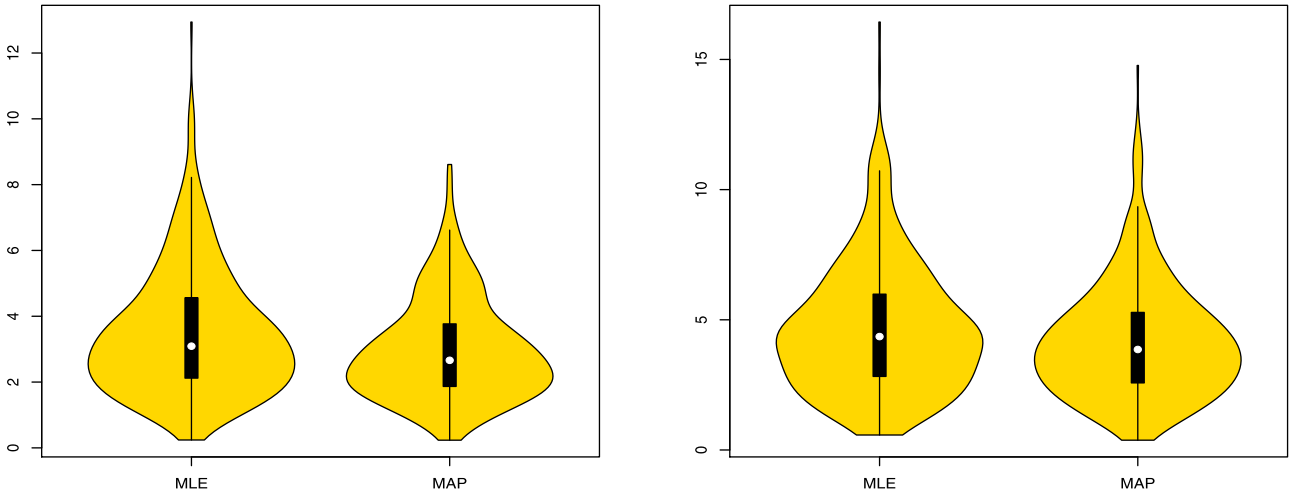

FIGURE 1. Violin plots of the error of the MLE and MAP estimators with respect to a design set following the uniform distribution and a Gaussian process with correlation lengths 0.4 – 0.8 – 0.2 (*left*) and 0.8 – 1.0 – 0.9 (*right*).

distribution $L(\boldsymbol{y}_0 \mid \boldsymbol{y}, \sigma^2, \boldsymbol{\theta})$, which is the predictive distribution we would use if we knew the correct values of the parameters σ^2 and $\boldsymbol{\theta}$.

Let us choose a "true" vector of correlation lengths $\boldsymbol{\theta}$ (and also a variance parameter σ^2, but this parameter has no effect on predictive accuracy). Then we do the following:

(1) Sample n observation points randomly according to the uniform distribution on the unit cube (in the following, $n = 30$).
(2) Generate the observations of the Gaussian process at the sampled points according to the selected "true" variance and correlation lengths.
(3) Sample the vector of correlation lengths according to the Gibbs reference posterior distribution $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$ through PIGS.
(4) Compute the MLE and the MAP of the vector of correlation lengths.
(5) Sample n_0 test points randomly according to the uniform distribution on the unit cube (in the following, $n_0 = 100$).
(6) At each point, determine the 95% prediction intervals derived from $L(\boldsymbol{y}_0 \mid \boldsymbol{y}, \sigma^2, \boldsymbol{\theta})$ (σ^2 and $\boldsymbol{\theta}$ being the "true" parameters), $\hat{P}_{\mathrm{MLE}}(\boldsymbol{y}_0 \mid \boldsymbol{y})$, $\hat{P}_{\mathrm{MAP}}(\boldsymbol{y}_0 \mid \boldsymbol{y})$ and $P(\boldsymbol{y}_0 \mid \boldsymbol{y})$.
(7) Generate the values of the Gaussian process at the newly sampled points (naturally, do this conditionally to the previously generated observations).
(8) Count the number of points within the prediction intervals derived from each of the four distributions. Divide the counts by n_0: this yields four *coverages* corresponding to each type of predictive intervals. Also compute the *mean length* of every type of prediction interval.
(9) Repeat steps 1 to 8 $m - 1$ times (in the following, $m = 500$).

6.2. Results

There is no reason for individual coverages of 95% predictive intervals given by the predictive distribution to be equal to 95%. Recall that any coverage is given for a unique realization of the Gaussian process, and that the values of this process at different points are correlated. If the predictive interval at some point fails to cover the true value at this point, it is likely that predictive intervals at neighboring points will also fail to cover the true values at those points, even though the nominal value is 95% everywhere. Conversely, if it actually covers

TABLE 2. Average with respect to randomly sampled design sets and realizations of the Gaussian process (with variance parameter 1 and smoothness parameter 5/2) of the coverage of 95% Prediction Intervals across the sample space. "True" stands for the prediction based on the knowledge of the true variance parameter and the true vector of correlation lengths.

Corr. lengths	True	MLE	MAP	FPD
0.4 – 0.8 – 0.2	0.95	0.88	0.91	0.95
0.5 – 0.5 – 0.5	0.95	0.89	0.90	0.94
0.7 – 1.3 – 0.4	0.95	0.90	0.92	0.95
0.8 – 0.3 – 0.6	0.95	0.89	0.91	0.95
0.8 – 1.0 – 0.9	0.95	0.90	0.92	0.94

TABLE 3. Average with respect to randomly sampled design sets and realizations of the Gaussian process (with variance parameter 1 and smoothness parameter 5/2) of the mean length of 95% Prediction Intervals across the sample space. The numbers in parentheses represent in percents the increase when using the MLE/MAP/FPD instead of the "true" vector of correlation lengths and variance parameter.

Corr. lengths	True	MLE	MAP	FPD
0.4 – 0.8 – 0.2	2.23	2.05 (−8)	2.13 (−4)	2.59 (+16)
0.5 – 0.5 – 0.5	1.69	1.55 (−8)	1.58 (−6)	1.84 (+9)
0.7 – 1.3 – 0.4	1.09	1.02 (−6)	1.07 (−2)	1.21 (+11)
0.8 – 0.3 – 0.6	1.63	1.51 (−7)	1.56 (−4)	1.82 (+12)
0.8 – 1.0 – 0.9	0.71	0.66 (−7)	0.69 (−3)	0.76 (+8)

the true value, then prediction intervals at neighboring points are more than 95% likely to cover their true values.

In short, prediction intervals give information that is only valid if understood to refer to what can be guessed on the sole basis of the observations made at the design points, which is why coverages for individual realizations of the Gaussian process are not necessarily 95% *even if the predictive distribution is perfectly accurate* (*i.e.* based on the true values of σ^2 and $\boldsymbol{\theta}$).

However, *if the predictive distribution is perfectly accurate*, then the average of the coverages is the nominal value: 95%. It is thus interesting to compute the average of the coverages for all distributions, whether they are plug-in distributions based on the MLE or MAP estimator, or the predictive distribution based on the full posterior distribution (hereafter noted FPD). In the above described methodology, the average was taken over the realizations of the Gaussian process with the chosen true parameters and over all design sets with n design points. The results below are obtained in this way.

The results given in Table 2 show that using the FPD to derive the predictive distribution is the best possible choice from a frequentist point of view as the nominal value is nearly matched by the average coverage. Predictive intervals derived from the MAP estimator do not perform as well, and predictive intervals derived from the MLE perform even worse.

Let us now focus on the average (with respect to the uniform design sets and realizations of the Gaussian process) of the mean (over the test set for a given realization of the Gaussian process and a given uniform design set) length of prediction intervals. The results are given in Table 3, where the figures between parentheses give the increase or decrease (in percents) of the average mean length when compared to the average mean length of prediction intervals obtained using the true values of the parameters.

Predictive intervals derived from the FPD are on average the largest, but not much larger than predictive intervals derived using the true parameters. In the tests we conducted, they seemed on average to be larger by about one fifth at worst. Predictive intervals derived from the MLE and MAP estimators are on average shorter

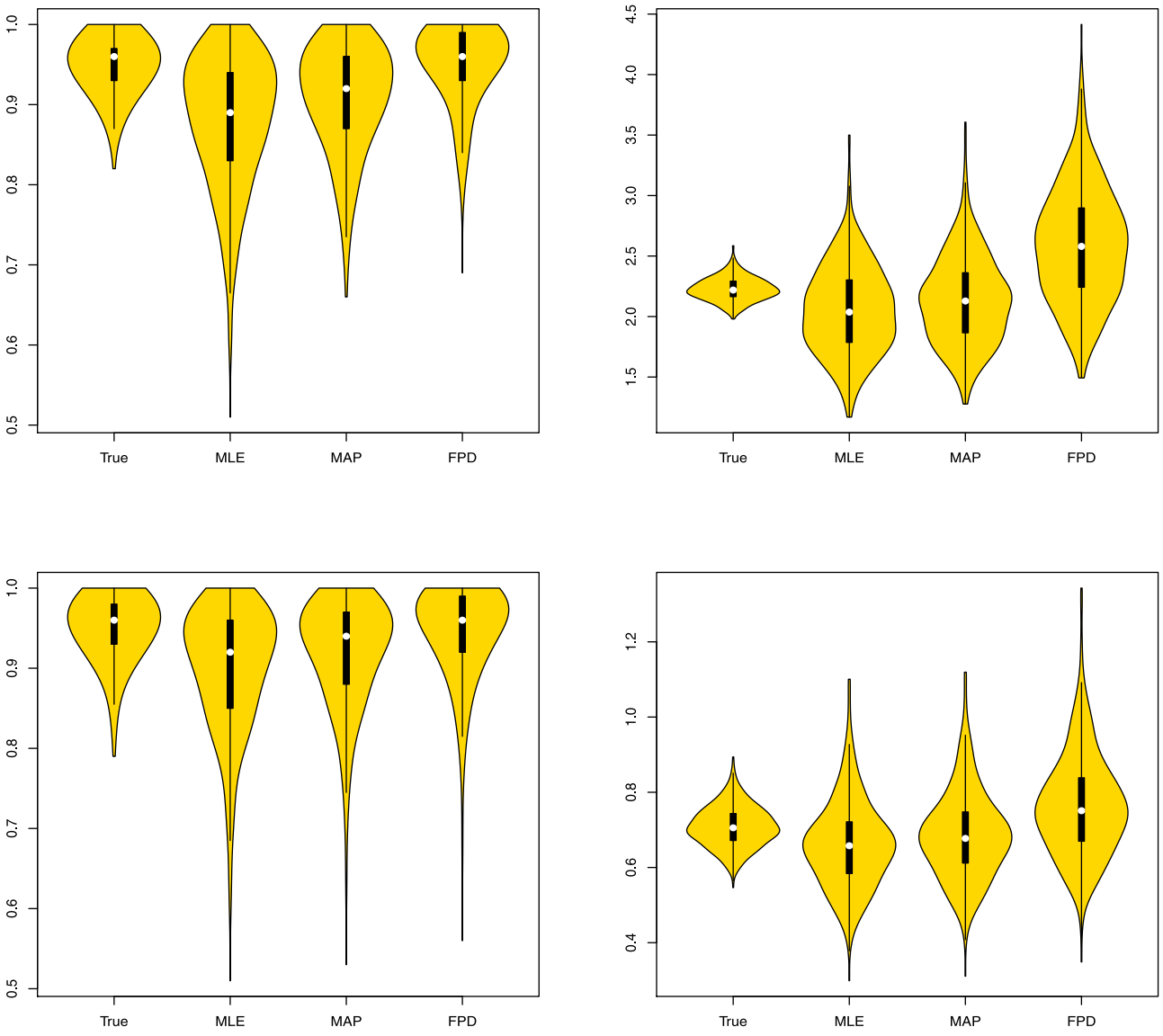

FIGURE 2. Violin plots of the coverage (*left*) and mean length (*right*) of Prediction Intervals with respect to a design set following the uniform distribution and a Gaussian process with correlation lengths 0.4 – 0.8 – 0.2 (*top*) and 0.8 – 1.0 – 0.9 (*bottom*).

than those derived from the true parameters. This can be interpreted as an under-estimation of the uncertainty of the prediction when fixing the vector of correlation lengths to the most likely value given the observations, and this can explain the low observed coverage in Table 2.

In Figure 2, we give violin plots of coverage and mean length of Prediction Intervals in the two most extreme cases: correlation lengths 0.4 – 0.8 – 0.2 (very low correlation) and 0.8 – 1.0 – 0.9 (very high correlation). The results are similar and illustrate the fact that the FPD gives larger intervals in order to reach the derived coverage value.

TABLE 4. Coverage of 95% prediction intervals when emulating the Ackley function on the unit hypercube using a Gaussian process with null mean function and a Matérn anisotropic geometric covariance kernel with smoothness 5/2, unknown variance parameter and unknown vector of correlation lengths. The design sets contain 100 points.

Design set type	MLE	MAP	FPD
Unoptimized LHS	0.89	0.92	0.93
Optimized LHS	0.74	0.76	0.80
Random design	0.87	0.88	0.91

TABLE 5. Mean length of 95% prediction intervals when emulating the Ackley function on the unit hypercube using a Gaussian process with null mean function and a Matérn anisotropic geometric covariance kernel with smoothness 5/2, unknown variance parameter and unknown vector of correlation lengths. The design sets contain 100 points.

Design set type	MLE	MAP	FPD
Unoptimized LHS	0.31	0.33	0.35
Optimized LHS	0.24	0.24	0.28
Random design	0.28	0.29	0.32

6.3. A higher-dimensional case

In this subsection, we emulate using Simple Kriging the 10-dimensional Ackley function:

$$A(\boldsymbol{x}) = 20 + \exp(1) - 20\exp\left(-0.2\sqrt{\frac{1}{10}\sum_{i=1}^{10} x_i^2}\right) - \exp\left(\frac{1}{10}\sum_{i=1}^{10}\cos(2\pi x_i)\right). \tag{6.1}$$

The goal in this section is to emulate the Ackley function on the unit hypercube $[0,1]^{10}$ using design sets with 100 observation points. Although the impact of the design set type is not the focus of this study, we present the results with a randomly chosen design according to the Uniform distribution on the domain $[0,1]^{10}$, a design obtained through LHS, and a design obtained through LHS and subsequently optimized to maximize the minimum distance between two points. The Simple Kriging model uses the null function as mean function and the Matérn anisotropic geometric covariance kernel family with smoothness parameter 5/2. The Gibbs reference posterior distribution is accessed through a sample of 1000 points. The conditional densities are sampled using the Metropolis algorithm with normal instrumental density with standard deviation 0.4 and a 100-step burn-in period.

To evaluate the performance of prediction intervals, we follow steps 3, 4, 5, 6 and 8 of the method presented in this section (step 7 is skipped as the "values of the Gaussian process" are naturally the values of the Ackley function) with $n_0 = 1000$. The results are presented in Tables 4 and 5.

As is shown in Table 4, prediction intervals derived using the Full Posterior Distribution perform better than those derived from the MAP, which themselves perform better than those derived from the MLE. This order of performance is the same regardless of the type of design set, although the optimized design set leads to much worse performances on average for prediction intervals than unoptimized designs. The latter fact is not surprising since space-filling designs ensure than no two points can be very close to each other, which makes it harder to determine the correlation lengths.

As expected, prediction intervals derived from the FPD are on average longer than those derived from the MAP and *a fortiori* the MLE. Notice that prediction intervals are on average shorter with the optimized design set, which explains the poorer performances in terms of coverage.

7. Conclusion and perspectives

We provided theoretical foundation to the claim that the stationary distribution of the Markov chain underlying PIGS with random scanning order is the optimal compromise between the potentially incompatible conditional distributions.

This theory is mainly derived from intuitive conceptions of what a compromise should be. In places where such conceptions were inconclusive, we relied on concrete examples to precisely determine what was acceptable and what was not in a compromise and used it to complete the definition. One strength of this theory is that it can be applied to continuous as well as discrete probability distributions, whereas previous studies focused on the discrete, or even finite, case.

A question that remains open outside the finite-state case is how compatibility of conditional distributions is to be checked in practice. Although compromises are useful, not needing them is better.

Further investigation is needed to fully understand the properties of the optimal compromise. Nevertheless, its invariance by reparametrization and its respect of pairwise independence show that it preserves important features of the conditional distributions.

The theory of optimal compromise suggests a framework for deriving a new objective posterior distribution based on the conditionals yielded by the reference prior theory on Simple Kriging parameters. Applying this framework to Matérn anisotropic kernels, we showed prediction to have good frequentist properties.

Future work should investigate whether this posterior distribution formally corresponds to some joint prior distribution. And if it does, how the joint prior distribution could be accessed, and how it relates to the conditional prior distributions.

Regarding the specific Kriging application presented in this paper, the next step is to extend the framework to Universal Kriging, where instead of being known, the mean function is only assumed to be a linear combination of known functions $f_1, ..., f_p$. The linear coefficients $\beta_1, ..., \beta_p$ are then considered parameters of the model. This extension is of practical relevance, because the mean function can rarely be considered known. It can probably be done in the same way Berger *et al.* [6] extended the reference prior from the Simple Kriging to the Universal Kriging framework: they used the flat improper prior as joint prior on $\beta_1, ..., \beta_p$ conditional to σ^2 and $\boldsymbol{\theta}$ and used it to integrate $\beta_1, ..., \beta_p$ out of the likelihood function, and then proceeded to derive the reference prior on σ^2 and $\boldsymbol{\theta}$ with respect to the integrated likelihood.

A further extension would involve deriving an objective prior on the smoothness parameter ν. In this endeavor, one should take into account the relation between correlation length $\boldsymbol{\theta}$ and smoothness ν. Unfortunately, asymptotic theory is not of much help in this regard, as Anderes [2] shows that provided the spatial domain $\mathcal{D}$ is of dimension at least 5, then all parameters of the Matérn anisotropic geometric kernel are microergodic (Zhang [35] shows this to be untrue for spatial domains of dimension 1, 2 or 3, but the non-microergodic parameters are σ^2 and $\boldsymbol{\theta}$, not ν). This means that the Gaussian measures on $\mathcal{D}$ corresponding to Gaussian processes with two different smoothness parameters are orthogonal, which suggests that there exists a consistent estimator (the MLE possibly). Stein [32] (Sect. 6.6) considers the Fisher information on $\boldsymbol{\theta}$ and ν, and gives examples (with a one-dimensional sample space $\mathcal{D}$) showing that the Fisher information on these parameters depends a lot on the design set. Fisher information relative to the smoothness parameter ν increases when design points are chosen to be close to one another (relative to the "true" correlation length $\boldsymbol{\theta}$), whereas Fisher information relative to correlation length $\boldsymbol{\theta}$ is maximized for design points that are farther apart. This, according to him, is coherent with the fact that $\boldsymbol{\theta}$ has greater influence on the low frequency behavior of the Matérn kernel while ν has greater influence on its high frequency behavior. This also suggests to us that the smoothness parameter ν, like the variance parameter σ^2, can only be meaningfully estimated if the vector of correlation lengths $\boldsymbol{\theta}$ is known. Otherwise, the estimator could hardly tell which design points are close to each other, which intuitively seems a prerequisite to evaluating the smoothness of the process. If we wish to apply the reference prior algorithm to the case where ν is unknown, we should thus probably derive the reference prior on ν conditional to $\boldsymbol{\theta}$.

Appendix A. Proofs of Section 2

Proof of Proposition 2.8. Let $P^{(0)}$ and $P^{(1)}$ be two compromises between $(\pi_i)_{i\in[\![1,r]\!]}$ and let $t \in (0,1)$. Let us check that $(1-t)P^{(0)} + tP^{(1)}$ verifies conditions (1) and (2) from Definition 2.6.

For every $i \in [\![1,r]\!]$, $\pi_i \left((1-t)P^{(0)} + tP^{(1)}\right)_{-i} = (1-t)\pi_i P^{(0)}_{-i} + t\pi_i P^{(1)}_{-i}$.

Let A be a measurable set such that $\left((1-t)P^{(0)} + tP^{(1)}\right)(A) = 0$. Then $P^{(0)}(A) = 0$ and $P^{(1)}(A) = 0$. Because both $P^{(0)}$ and $P^{(1))}$ are compromises, for every $i \in [\![1,r]\!]$, $\pi_i P^{(0)}_{-i}(A) = 0$ and $\pi_i P^{(1)}_{-i}(A) = 0$. So $\left(\pi_i \left((1-t)P^{(0)} + tP^{(1)}\right)_{-i}\right)(A) = 0$ and condition (1) is verified.

Now, for every $i \in [\![1,r]\!]$,

$$\begin{aligned}\frac{1}{r}\sum_{j=1}^{r}\left(\pi_j\left((1-t)P^{(0)} + tP^{(1)}\right)_{-j}\right)_{-i} &= \frac{1-t}{r}\sum_{j=1}^{r}\left(\pi_j P^{(0)}_{-j}\right)_{-i} + \frac{t}{r}\sum_{j=1}^{r}\left(\pi_j P^{(1)}_{-j}\right)_{-i} \\ &= (1-t)P^{(0)}_{-i} + tP^{(1)}_{-i} = \left((1-t)P^{(0)} + tP^{(1)}\right)_{-i}. \end{aligned} \tag{A.1}$$

So condition (2) is also verified. □

Proof of Proposition 2.10. Let P_G be a Gibbs compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$. Then, for any measurable set A such that $P_G(A) = 0$,

$$P_G(A) = \frac{1}{r}\sum_{i=1}^{r}\pi_i(P_G)_{-i}(A) = 0. \tag{A.2}$$

So for every integer $i \in [\![1,r]\!]$, $\pi_i(P_G)_{-i}(A) = 0$. This fulfills the first condition of Definition 2.6.

Its second condition is also fulfilled because for every integer $i \in [\![1,r]\!]$,

$$(P_G)_{-i} = \left(\frac{1}{r}\sum_{j=1}^{r}\pi_j(P_G)_{-j}\right)_{-i} = \frac{1}{r}\sum_{j=1}^{r}\left(\pi_j(P_G)_{-j}\right)_{-i}. \tag{A.3}$$

□

Proof of Proposition 2.11. Define the probability distribution P on $\mathcal{A}$ as follows:

$$P = \frac{1}{r}\sum_{i=1}^{r}\pi_i m_{\neq i}. \tag{A.4}$$

Then for every $i \in [\![1,r]\!]$ the ith $(r-1)$-marginal distribution P_{-i} of P is given by

$$P_{-i} = \frac{1}{r}\sum_{j=1}^{r}\left(\pi_j m_{\neq j}\right)_{-i} = m_{\neq i}, \tag{A.5}$$

where the last equality is due to $(m_{\neq i})_{i\in[\![1,r]\!]}$ being compatible with $(\pi_i)_{i\in[\![1,r]\!]}$. Plugging this into equation (A.4), we obtain that P is a Gibbs compromise. Equation (A.5) then yields the result.

□

Proof of Theorem 2.13. If P_G is the unique Gibbs compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$, then Proposition 2.11 asserts that $((P_G)_{-i})_{i\in[\![1,r]\!]}$ is the only compatible sequence of $(r-1)$-dimensional distributions. So any compromise has the same sequence of $(r-1)$-marginal distributions. Let P be such a compromise.

For every $i \in [\![1,r]\!]$, $\pi_i(P_G)_{-i} = \pi_i P_{-i}$ is absolutely continuous with respect to P. So, the average P_G is also absolutely continuous with respect to P.

Let λ be a positive measure on $\mathcal{A}$ such that P is absolutely continuous with respect to λ. Because P_G and P share the same sequence of $(r-1)$-marginal distributions, for λ-almost any $\omega \in \Omega$,

$$\frac{\mathrm{d}(\pi_i P_{-i})}{\mathrm{d}\lambda}(\omega) = \frac{\mathrm{d}(\pi_i (P_G)_{-i})}{\mathrm{d}\lambda}(\omega). \tag{A.6}$$

Moreover, equation (A.4) implies that for λ-almost any $\omega \in \Omega$, $\frac{\mathrm{d}P_G}{\mathrm{d}\lambda}(\omega)$ is the arithmetic average between the $\frac{\mathrm{d}(\pi_i (P_G)_{-i})}{\mathrm{d}\lambda}(\omega)$ $(i \in [\![1,r]\!])$, so it minimizes the mean squared error. Together with equation (A.6), this implies that for λ-almost any $\omega \in \Omega$,

$$\sum_{i=1}^{r} \left[\frac{\mathrm{d}(\pi_i (P_G)_{-i})}{\mathrm{d}\lambda}(\omega) - \frac{\mathrm{d}P_G}{\mathrm{d}\lambda}(\omega) \right]^2 \leqslant \sum_{i=1}^{r} \left[\frac{\mathrm{d}(\pi_i P_{-i})}{\mathrm{d}\lambda}(\omega) - \frac{\mathrm{d}P}{\mathrm{d}\lambda}(\omega) \right]^2. \tag{A.7}$$

Consequently, $E_\lambda(P_G) \leqslant E_\lambda(P)$. Moreover, if $P \neq P_G$, then there exists $S \in \mathcal{A}$ such that $\lambda(S) > 0$ and for every $\omega \in S$,

$$\forall i \in [\![1,r]\!] \quad \frac{\mathrm{d}(\pi_i P_{-i})}{\mathrm{d}\lambda}(\omega) = \frac{\mathrm{d}(\pi_i (P_G)_{-i})}{\mathrm{d}\lambda}(\omega) \quad \text{and} \quad \frac{\mathrm{d}P}{\mathrm{d}\lambda}(\omega) \neq \frac{\mathrm{d}(P_G)}{\mathrm{d}\lambda}(\omega). \tag{A.8}$$

So for every $\omega \in S$, equation (A.7) is a strict inequality and thus $E_\lambda(P_G) < E_\lambda(P)$. P_G is therefore the unique optimal compromise with respect to λ. □

Proof of Proposition 2.15. For every $i \in [\![1,r]\!]$, for every $\tilde{S}_i \in \tilde{\mathcal{A}}_i$,

$$\begin{aligned}
\tilde{P}_G \left(\underset{i\in[\![1,r]\!]}{\times} \tilde{S}_i \right) &= P_G \left(f^{-1} \left(\underset{i\in[\![1,r]\!]}{\times} \tilde{S}_i \right) \right) = P_G \left(\underset{i\in[\![1,r]\!]}{\times} f_i^{-1}(\tilde{S}_i) \right) \\
&= \frac{1}{r} \sum_{i=1}^{r} \int_{\underset{j\neq i}{\times} f_j^{-1}(\tilde{S}_j)} \pi_i(f_i^{-1}(\tilde{S}_i)|\omega_{-i}) \mathrm{d}\left\{ (P_G)_{-i} \right\}(\omega_{-i}) \\
&= \frac{1}{r} \sum_{i=1}^{r} \int_{\underset{j\neq i}{\times} f_j^{-1}(\tilde{S}_j)} \tilde{\pi}_i(\tilde{S}_i|f_{-i}(\omega_{-i})) \mathrm{d}\left\{ (P_G)_{-i} \right\}(\omega_{-i}) \\
&= \frac{1}{r} \sum_{i=1}^{r} \int_{\underset{j\neq i}{\times} \tilde{S}_j} \tilde{\pi}_i(\tilde{S}_i|\tilde{\omega}_{-i}) \mathrm{d}\left\{ (P_G)_{-i} * f_{-i} \right\}(\tilde{\omega}_{-i}).
\end{aligned} \tag{A.9}$$

Now, for every $i \in [\![1,r]\!]$, for every $\tilde{T}_i \in \tilde{\mathcal{A}}_i$,

$$\begin{aligned}
(P_G)_{-i} * f_{-i} \left(\underset{j\neq i}{\times} \tilde{T}_j \right) &= (P_G)_{-i} \left(\underset{j\neq i}{\times} f_j^{-1}(\tilde{T}_j) \right) = P_G \left(\underset{j<i}{\times} f_j^{-1}(\tilde{T}_j) \times f_i^{-1}(\tilde{\Omega}_i) \times \underset{k>i}{\times} f_k^{-1}(\tilde{T}_k) \right) \\
&= P_G * f \left(\underset{j<i}{\times} \tilde{T}_j \times \tilde{\Omega}_i \underset{k>i}{\times} \tilde{T}_k \right) = (P_G * f)_{-i} \left(\underset{j\neq i}{\times} \tilde{T}_j \right).
\end{aligned} \tag{A.10}$$

So $(P_G)_{-i} * f_{-i} = (P_G * f)_{-i} = (\tilde{P}_G)_{-i}$. Then, returning to equation (A.9),

$$\tilde{P}_G\left(\underset{i\in[\![1,r]\!]}{\times}\tilde{S}_i\right) = \frac{1}{r}\sum_{i=1}^{r}\int_{\underset{j\neq i}{\times}\tilde{S}_j}\tilde{\pi}_i(\tilde{S}_i|\tilde{\omega}_{-i})\mathrm{d}\left\{(\tilde{P}_G)_{-i}\right\}(\tilde{\omega}_{-i}). \tag{A.11}$$

Finally, we obtain

$$\tilde{P}_G = \frac{1}{r}\sum_{i=1}^{r}\tilde{\pi}_i(\tilde{P}_G)_{-i}. \tag{A.12}$$

$\tilde{P}_G$ is therefore a Gibbs compromise between the sequence of Markov kernels $(\tilde{\pi}_i)_{i\in[\![1,r]\!]}$. We now prove its uniqueness. For every $i \in [\![1,r]\!]$, f_i is bijective and f_i^{-1} is measurable. So for any Gibbs compromise $\tilde{Q}_G$ between the sequence of Markov kernels $(\tilde{\pi}_i)_{i\in[\![1,r]\!]}$, $\tilde{Q}_G * f^{-1}$ is a Gibbs compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$. Given P_G is the unique Gibbs compromise between the Markov kernels in this sequence, $\tilde{Q}_G * f^{-1} = P_G$, so $\tilde{Q}_G = \tilde{Q}_G * f^{-1} * f = P_G * f = \tilde{P}_G$.

□

Appendix B. Proofs of Section 4

The following holds where there is no mention of the contrary. When applied to a vector, $\|\cdot\|$ denotes the Euclidean norm and when applied to a matrix, it denotes the Frobenius norm. The choice of norm does not matter much because in finite-dimensional vector spaces, all norms are equivalent.

B.1 Differentiating the Matérn correlation kernel

Lemma B.1. *The partial derivative with respect to μ_i of the Matérn tensorized kernel of variance σ^2, smoothness ν and inverse correlation length vector $\boldsymbol{\mu}$ is:*

$$\frac{\partial}{\partial\mu_i}\left(\sigma^2 K_{r,\nu}^{\mathrm{tens}}(\boldsymbol{x}\boldsymbol{\mu})\right) = -\frac{\sigma^2(2\sqrt{\nu})^2}{\Gamma(\nu)2^{\nu-1}}|x_i|^2\mu_i\left(2\sqrt{\nu}|x_i|\mu_i\right)^{\nu-1}\mathcal{K}_{\nu-1}\left(2\sqrt{\nu}|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right). \tag{B.1}$$

This can be rewritten as:

$$\frac{\partial}{\partial\mu_i}\left(\sigma^2 K_{r,\nu}^{\mathrm{tens}}(\boldsymbol{x}\boldsymbol{\mu})\right) = \begin{cases} \sigma^2\frac{2\nu}{\nu-1}|x_i|^2\mu_i K_{1,\nu-1}\left(|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) & \text{if } \nu>1 \\ \sigma^2 4|x_i|^2\mu_i\mathcal{K}_0\left(2|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) & \text{if } \nu=1 \\ \sigma^2 2\nu^\nu\frac{\Gamma(1-\nu)}{\Gamma(\nu)}|x_i|^{2\nu}\mu_i^{2\nu-1}K_{1,1-\nu}\left(|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) & \text{if } \nu<1. \end{cases} \tag{B.2}$$

Proof. The first assertion is a simple matter of differentiating equation (4.9). In the following calculation, the fourth line is given by formula 9.6.28 (p. 376) in Abramowitz and Stegun [1].

$$\begin{aligned}\frac{\partial}{\partial\mu_i}\left(\sigma^2 K^{\text{tens}}_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})\right) &= \sigma^2\frac{\partial}{\partial\mu_i}\left(K_{1,\nu}(x_i\mu_i)\right)\prod_{j\neq i}K_{1,\nu}(|x_j|\mu_j) = \sigma^2 x_i\left(K'_{1,\nu}(x_i\mu_i)\right)\prod_{j\neq i}K_{1,\nu}(|x_j|\mu_j)\\ &= \sigma^2 x_i\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}\left.\frac{\mathrm{d}}{\mathrm{d}y}\right|_{y=2\sqrt{\nu}x_i\mu_i}[y^\nu\mathcal{K}_\nu(y)]\right)\prod_{j\neq i}K_{1,\nu}(|x_j|\mu_j) \qquad \text{(B.3)}\\ &= \sigma^2 x_i\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}[-y\cdot y^{\nu-1}\mathcal{K}_{\nu-1}(y)]_{y=2\sqrt{\nu}x_i\mu_i}\right)\prod_{j\neq i}K_{1,\nu}(|x_j|\mu_j).\end{aligned}$$

From there, equation (B.1) follows immediately. Rewriting it in the form given in (B.2) only requires us to recall $\Gamma(\nu)=(\nu-1)\Gamma(\nu-1)$ (case $\nu>1$), $\Gamma(1)=1$ (case $\nu=1$) and $\mathcal{K}_{\nu-1}=\mathcal{K}_{1-\nu}$ (case $\nu<1$). □

Lemma B.2. *The partial derivative with respect to μ_i of the Matérn geometric anisotropic kernel of variance σ^2, smoothness ν and inverse correlation length vector $\boldsymbol{\mu}$ is:*

$$\frac{\partial}{\partial\mu_i}\left(\sigma^2 K_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})\right) = \frac{\sigma^2(2\sqrt{\nu})^2}{\Gamma(\nu)2^{\nu-1}}|x_i|^2\mu_i\left(2\sqrt{\nu}\,\|\boldsymbol{x}\boldsymbol{\mu}\|\right)^{\nu-1}\mathcal{K}_{\nu-1}\left(2\sqrt{\nu}\,\|\boldsymbol{x}\boldsymbol{\mu}\|\right). \qquad \text{(B.4)}$$

This can be rewritten as:

$$\frac{\partial}{\partial\mu_i}\left(\sigma^2 K_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})\right) = \begin{cases}\sigma^2\frac{2\nu}{\nu-1}|x_i|^2\mu_i K_{1,\nu-1}(\|\boldsymbol{x}\boldsymbol{\mu}\|) & \text{if } \nu>1\\ \sigma^2 4|x_i|^2\mu_i\mathcal{K}_0(2\|\boldsymbol{x}\boldsymbol{\mu}\|) & \text{if } \nu=1\\ \sigma^2 2\nu^\nu\frac{\Gamma(1-\nu)}{\Gamma(\nu)}\frac{1}{\mu_i}\left(\frac{|x_i|\mu_i}{\|\boldsymbol{x}\boldsymbol{\mu}\|^{1-\nu}}\right)^2 K_{1,1-\nu}(\|\boldsymbol{x}\boldsymbol{\mu}\|) & \text{if } \nu<1.\end{cases} \qquad \text{(B.5)}$$

Proof. The first assertion is a simple matter of differentiating equation (4.7). In the following calculation, the fourth line is given by formula 9.6.28 (page 376) in Abramowitz and Stegun [1].

$$\begin{aligned}\frac{\partial}{\partial\mu_i}\left(\sigma^2 K_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})\right) &= \sigma^2\frac{\partial}{\partial\mu_i}\left(K_{1,\nu}(\|\boldsymbol{x}\boldsymbol{\mu}\|)\right) = \sigma^2 x_i^2\mu_i\|\boldsymbol{x}\boldsymbol{\mu}\|^{-1}K'_{1,\nu}(\|\boldsymbol{x}\boldsymbol{\mu}\|)\\ &= \sigma^2 x_i^2\mu_i\|\boldsymbol{x}\boldsymbol{\mu}\|^{-1}\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}\left.\frac{\mathrm{d}}{\mathrm{d}y}\right|_{y=2\sqrt{\nu}\|\boldsymbol{x}\boldsymbol{\mu}\|}[y^\nu\mathcal{K}_\nu(y)]\right) \qquad \text{(B.6)}\\ &= \sigma^2 x_i^2\mu_i\|\boldsymbol{x}\boldsymbol{\mu}\|^{-1}\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}[-y\cdot y^{\nu-1}\mathcal{K}_{\nu-1}(y)]_{y=2\sqrt{\nu}\|\boldsymbol{x}\boldsymbol{\mu}\|}\right).\end{aligned}$$

From there, equation (B.4) follows immediately. Rewriting it in the form given in (B.5) only requires us to recall $\Gamma(\nu)=(\nu-1)\Gamma(\nu-1)$ (case $\nu>1$), $\Gamma(1)=1$ (case $\nu=1$) and $\mathcal{K}_{\nu-1}=\mathcal{K}_{1-\nu}$ (case $\nu<1$). □

B.2 Accounting for low correlation: $\|\boldsymbol{\mu}\|\to\infty$

In this subsection, we consider a fixed design set of n coordinate-distinct points $\boldsymbol{x}^{(k)}$ ($k\in[\![1,n]\!]$) in $\mathbb{R}^r$.

Lemma B.3. *For any Matérn anisotropic geometric or tensorized correlation kernel with smoothness $\nu>0$, for all $b<2\min(1,\nu)-1$ and $c>1$ (and if $\nu\neq1$, for all $b\leqslant 2\min(1,\nu)-1$),*

(a) $\forall\boldsymbol{\mu}\in(\mathbb{R}_+)^r$, $\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\|\leqslant M_{i,1}\,\mu_i^{-c}$.
(b) $\forall\boldsymbol{\mu}\in(\mathbb{R}_+)^r$, $\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\|\leqslant M_{i,2}\,\mu_i^{b}$.

Proof. This can be gathered from Lemma B.1 or B.2 after recalling that 1) a Matérn kernel is a bounded function, 2) $\forall \nu \geqslant 0$, as $z \to +\infty$, $\mathcal{K}_\nu(z) \sim \sqrt{\pi}\exp(-z)/\sqrt{2z}$ ([1] 9.7.2) and 3) as $z \to 0$, $\mathcal{K}_0(z) \sim -\log(z)$ ([1] 9.6.8). □

Let us define

$$f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := \sqrt{[\boldsymbol{\mathcal{I}}(\boldsymbol{\mu})]_{ii}}\,; \tag{B.7}$$

$$\pi_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) / \int_0^\infty f_i(\mu_i = t \mid \boldsymbol{\mu}_{-i})\mathrm{d}t. \tag{B.8}$$

Proposition B.4. *For any Matérn anisotropic geometric or tensorized correlation kernel with smoothness $\nu > 0$, for all $\mu_i \in (0, +\infty)$, $\pi(\mu_i|\boldsymbol{\mu}_{-i})$, seen as a function of $\boldsymbol{\mu}$, is well defined and continuous over $\{\boldsymbol{\mu} \in [0, +\infty)^r : \mu_i \neq 0,\ \boldsymbol{\mu}_{-i} \neq \mathbf{0}_{r-1}\}$.*

Proof. For any given $\tilde{\boldsymbol{\mu}} \in [0, +\infty)^r$ such that $\tilde{\mu}_i \neq 0$ and $\tilde{\boldsymbol{\mu}}_{-i} \neq \mathbf{0}_{r-1}$, we prove that $\pi(\mu_i|\boldsymbol{\mu}_{-i})$, seen as a function of $\boldsymbol{\mu}$, is well defined and continuous at $\boldsymbol{\mu} = \tilde{\boldsymbol{\mu}}$.

For a start, notice that if $\boldsymbol{\mu}$ is confined to a sufficiently small neighborhood of $\tilde{\boldsymbol{\mu}}$, then $\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\|$ remains bounded. Therefore, Lemma B.3 implies that $\int_0^\infty f_i(\mu_i = t \mid \boldsymbol{\mu}_{-i})\mathrm{d}t$ is finite and, thanks to the dominated convergence theorem, that it is continuous at $\boldsymbol{\mu}_{-i} = \tilde{\boldsymbol{\mu}}_{-i}$.

□

Definition B.5. An anisotropic geometric or tensorized correlation kernel is said to be "well-behaved" if its one-dimensional version is, for any set of parameters, a positive decreasing function on $[0, +\infty)$ that vanishes in the neighborhood of $+\infty$.

Lemma B.6. *Provided a coordinate-distinct design set is used, a well-behaved anisotropic geometric or tensorized correlation kernel parametrized by $\boldsymbol{\mu}$ has the following properties:*

(a) for any fixed $\boldsymbol{\mu}_{-i} \in [0, +\infty)^{r-1}$, it is a decreasing function of μ_i;
(b) as $\|\boldsymbol{\mu}\| \to \infty$, $\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}} - \boldsymbol{I}_n\| \to 0$.

Lemma B.7. *For any well-behaved correlation kernel, as $\|\boldsymbol{\mu}\| \to \infty$, $\mathrm{Tr}\left[\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right] = o\left(\left\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\|\right)$.*

Proof. This result is due to the fact that all $\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$'s diagonal coefficients are null and $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ goes to the identity matrix as $\|\boldsymbol{\mu}\| \to \infty$. □

Let us now define

$$h_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := \sqrt{\mathrm{Tr}\left[\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)^2\right]} = \left\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\|. \tag{B.9}$$

Lemma B.8. *For any well-behaved correlation kernel, as $\|\boldsymbol{\mu}\| \to \infty$, $f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \sim h_i(\mu_i \mid \boldsymbol{\mu}_{-i})$.*

Proof. Because $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ goes to the identity matrix, this is a direct consequence of Lemma B.7. □

Corollary B.9. *For any well-behaved correlation kernel, there exist $S > 0$, and $0 < a < b$ such that, whenever $\|\boldsymbol{\mu}\| \geqslant S$,*

$$a\, h_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \leqslant f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \leqslant b\, h_i(\mu_i \mid \boldsymbol{\mu}_{-i}). \tag{B.10}$$

In the following, $\boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}$ is the correlation matrix that would be obtained if μ_i were replaced by 0. Moreover, if $\boldsymbol{M}$ is a matrix, $\boldsymbol{M}^{(kl)}$ is its element in the kth row and lth column.

Lemma B.10. *If a well-behaved correlation kernel is used, there exist real constants $S > 0$ and $c > 0$ such that, for all $\mu_i \in (0, +\infty)$ and whenever $\|\boldsymbol{\mu}_{-i}\| \geqslant S$,*

$$\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \geqslant c\frac{\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\|}{\sum_{k\neq l}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}}. \tag{B.11}$$

Proof. If a well-behaved correlation kernel is used, then for any for any $\epsilon > 0$, Corollary B.9 implies that

$$\int_0^{+\infty} f_i(\mu_i = t|\boldsymbol{\mu}_{-i})\mathrm{d}t \leqslant b\int_0^{+\infty} h_i(\mu_i = t|\boldsymbol{\mu}_{-i})\mathrm{d}t \leqslant -b\sum_{k\neq l}\int_0^{+\infty}\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}}\mathrm{d}t. \tag{B.12}$$

The last inequality holds because the Frobenius norm of any matrix is smaller than or equal to the sum of the absolute values of its elements and the correlation kernel is a decreasing function of μ_i. Now, for all $k \neq l$, when $\mu_i \to +\infty$, $\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}} \to 0$ and when $\mu_i = 0$, $\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}} = \boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}$. From this, we gather that

$$\int_0^{+\infty} f_i(\mu_i = t|\boldsymbol{\mu}_{-i})\mathrm{d}t \leqslant b\sum_{k\neq l}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}. \tag{B.13}$$

From this, we deduce that

$$\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) = \frac{f_i(\mu_i|\boldsymbol{\mu}_{-i})}{\int_0^{+\infty} f_i(\mu_i = t|\boldsymbol{\mu}_{-i})\mathrm{d}t} \geqslant \frac{a}{b}\frac{\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\|}{\sum_{k\neq l}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}}. \tag{B.14}$$

□

This lemma has the following immediate consequence:

Proposition B.11. *If a well-behaved tensorized kernel is used, there exists $S > 0$ and for every $i \in [\![1, r]\!]$, there exists a function $M_i : (0, +\infty) \to (0, +\infty)$ such that for all $\|\boldsymbol{\mu}_{-i}\| \geqslant S$, $\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \geqslant M_i(\mu_i)$.*

Proof. If a tensorized correlation kernel is used, for every pair of integers $(k, l) \in [\![1, r]\!]^2$ such that $k \neq l$, define the function $M_i^{(kl)} : (0, +\infty) \to (0, +\infty)\ ;\ t \mapsto \left|\frac{\mathrm{d}}{\mathrm{d}t}\boldsymbol{\Sigma}^{(kl)}_{\mu_i=t,\boldsymbol{\mu}_{-i}=\mathbf{0}_{r-1}}\right|$.

$$\left\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\| \geqslant \frac{1}{n}\sum_{k\neq l}\left|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}}\right| = \frac{1}{n}\sum_{k\neq l}M_i^{(kl)}(\mu_i)\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}} \geqslant \frac{1}{n}\min_{k\neq l}M_i^{(kl)}(\mu_i)\sum_{k\neq l}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}. \tag{B.15}$$

This fact, joined with Lemma B.10, yields the result. □

Proposition B.12. *Assume a well-behaved anisotropic geometric correlation kernel is used. If the corresponding one-dimensional kernel K has the properties (P1) and (P2), then for every $i \in [\![1, r]\!]$, there exist positive functions s_i and m_i defined on $(0, +\infty)$ such that, for all $\|\boldsymbol{\mu}_{-i}\| \geqslant s_i(\mu_i)$, $\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \geqslant m_i(\mu_i)$.*

(P1): There exist $S_1 > 0$ and $M_1 > 0$ such that, for all $t \geqslant S_1$, $|K'(t)| \geqslant M_1 tK(t)$.

(P2): For any $a > 0$, there exist $S_2(a) > 0$ and $M_2(a) > 0$ such that, whenever $t \geqslant S_2(a)$, $K(t + a) \geqslant M_2(a)K(t)$.

Proof. From (P1), we gather that for all $a > 0$ and $t \geqslant S_1$, $|K'(\sqrt{t^2 + a^2})| \geqslant M_1\sqrt{t^2 + a^2}K(\sqrt{t^2 + a^2})$. Now, because the correlation kernel is well-behaved, K is a decreasing function. As $\sqrt{t^2 + a^2} \leqslant t + a$, $K(\sqrt{t^2 + a^2}) \geqslant K(t + a)$.

Plugging this into the previous inequality, we get $|K'(\sqrt{t^2 + a^2})| \geqslant M_1\sqrt{t^2 + a^2}K(t + a)$.

If $t \geqslant \max(S_1, S_2(a))$, we can then use (P2) to obtain

$$|K'(\sqrt{t^2+a^2})| \geqslant M_1 M_2(a)\sqrt{t^2+a^2}K(t). \tag{B.16}$$

Independently from this, we have the following algebraic fact:

$$\left\| \frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \right\| \geqslant \frac{1}{n} \sum_{k \neq l} \left| \frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(kl)} \right|. \tag{B.17}$$

Because we use a well-behaved anisotropic geometric kernel, defining the function $M_i^{(kl)} : (0,+\infty) \to (0,+\infty)$; $t \mapsto \left(x_j^{(k)} - x_j^{(l)}\right)^2$, we can write:

$$\left| \frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(kl)} \right| = -\frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(kl)} = \left(x_i^{(k)} - x_i^{(l)}\right)^2 \mu_i \frac{K'\left(\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|\right)}{\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|}. \tag{B.18}$$

Setting $a_{kl} := |x_i^{(k)} - x_i^{(l)}|\mu_i$ and $t_{kl} := \|(\boldsymbol{x}_{-i}^{(k)} - \boldsymbol{x}_{-i}^{(l)})\boldsymbol{\mu}_{-i}\|$ (and thus, naturally, $\sqrt{t_{kl}^2 + a_{kl}^2} = \|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|$), and provided $\|\boldsymbol{\mu}_{-i}\|$ is sufficiently large to make all t_{kl}s meet the conditions necessary to apply (P1) and (P2) (that depend in the case of (P2) on the a_{kl}s), equation (B.16) yields the existence of some number $m_i^{(kl)}(\mu_i) > 0$ such that

$$\frac{K'\left(\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|\right)}{\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|} \geqslant m_i^{(kl)}(\mu_i) K(\|(\boldsymbol{x}_{-i}^{(k)} - \boldsymbol{x}_{-i}^{(l)})\boldsymbol{\mu}_{-i}\|) = m_i^{(kl)}(\mu_i)\boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}. \tag{B.19}$$

Finally, setting $m_i(\mu_i) := \mu_i \min_{k \neq l}\left[\left(x_i^{(k)} - x_i^{(l)}\right)^2 m_i^{(kl)}(\mu_i)\right]$, we get

$$\left\| \frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \right\| \geqslant \frac{m_i(\mu_i)}{n} \sum_{k \neq l} \boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}. \tag{B.20}$$

Then, applying Lemma B.10 yields the result. □

Proposition B.13. *Matérn one-dimensional kernels with smoothness parameter $\nu > 1$ have the properties (P1) and (P2) of Proposition B.12.*

Proof. (P1) is given by Lemma B.2, after noticing that, denoting by K_ν the Matérn one-dimensional kernel of smoothness $\nu > 1$, provided t is sufficiently large, $K_\nu(t) \leqslant K_{\nu-1}(t)$.

This inequality ensues from the fact that $\forall \nu \geqslant 0$, as $t \to +\infty$, $\mathcal{K}_\nu(t) \sim \sqrt{\pi}\exp(-t)/\sqrt{2t}$ ([1] 9.7.2) and thus $K_\nu(t) \sim 2/\Gamma(\nu)(\sqrt{\nu}t)^\nu\sqrt{\pi/(4\sqrt{\nu}t)}\exp(-2\sqrt{\nu}t)$. Moreover, this last equivalence relation also implies (P2). □

Proposition B.14. *For Matérn anisotropic geometric kernels with smoothness $\nu > 1$ and Matérn tensorized correlation kernels with smoothness $\nu > 0$,*

for any $\delta > 0$, $i \in [\![1, r]\!]$ and $\mu_i \in (0,+\infty)$, there exists $b_{i,\delta}(\mu_i) > 0$ such that, if $\|\boldsymbol{\mu}_{-i}\| \geqslant \delta$, then $\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \geqslant b_{i,\delta}(\mu_i)$.

Proof. Matérn correlation kernels with such smoothness parameters make Proposition B.11 or B.12 applicable. Therefore, there exist $s_i(\mu_i) > 0$ and $m_i(\mu_i) > 0$ such that, if $\|\boldsymbol{\mu}_{-i}\| \geqslant s_i(\mu_i)$, $\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \geqslant m_i(\mu_i)$. Besides, we know from Proposition B.4 that $\pi_i(\mu_i|\boldsymbol{\mu}_{-i})$, seen as a function of $\boldsymbol{\mu}_{-i}$, is continuous and positive over the compact set $\{\boldsymbol{\mu}_{-i} : \delta \leqslant \|\boldsymbol{\mu}_{-i}\| \leqslant s_i(\mu_i)\}$. Thus its minimum $\tilde{m}_{i,\delta}(\mu_i)$ on this set is positive and we obtain the result by setting $b_{i,\delta}(\mu_i) := \min(m_i(\mu_i), \tilde{m}_{i,\delta}(\mu_i))$. □

Proposition B.15. *For Matérn anisotropic geometric correlation kernels with smoothness $\nu > 1$ and for Matérn tensorized correlation kernels with smoothness $\nu > 0$, for any $\boldsymbol{y} \in \mathbb{R}^n \setminus \{0\}^n$, any $\delta > 0$ and any $\mu_i \in (0, +\infty)$, there exists $b_{i,\delta,\boldsymbol{y}}(\mu_i) > 0$ such that, if $\|\boldsymbol{\mu}_{-i}\| \geqslant \delta$, then $\pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) \geqslant b_{i,\delta,\boldsymbol{y}}(\mu_i)$.*

Proof. Set $\delta > 0$ and $\boldsymbol{y} \in \mathbb{R}^n \setminus \{0\}^n$. There exist $m_\delta > 0$ and $M_\delta > 0$ s.t. $\forall \boldsymbol{\mu} \in (0, +\infty)^r$, $\|\boldsymbol{\mu}_{-i}\| \geqslant \delta \Rightarrow m_\delta \leqslant \|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\| \leqslant M_\delta$, so there also exist $m_{\delta,\boldsymbol{y}} > 0$ and $M_{\delta,\boldsymbol{y}} > 0$ s.t. $m_{\delta,\boldsymbol{y}} \leqslant L(\boldsymbol{y}|\boldsymbol{\mu}) \leqslant M_{\delta,\boldsymbol{y}}$. This, combined with Proposition B.14, yields the result. □

B.3 Accounting for high correlation: $\|\boldsymbol{\mu}\| \to 0$

This part of the proof relies on the combination of some spectral study of the Matérn kernels and on the study of the matrices that are part of the series expansion of the correlation matrix $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ when $\|\boldsymbol{\mu}\| \to 0$ for three types of Matérn kernels: isotropic, tensorized and anisotropic geometric.

Lemma B.16. *There exists a covariance kernel $\widetilde{K}_{r,\nu}$ such that for any design set $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, for all $\boldsymbol{\mu} \in (\mathbb{R}_+)^r$ and $\boldsymbol{\xi} = (\xi_1, ..., \xi_n) \in \mathbb{R}^n$,*

$$\sum_{j,k=1}^{n} \xi_j \xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right) \geqslant 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\|\boldsymbol{\mu}\|_\infty) \sum_{j,k=1}^{n} \xi_j \xi_k \widetilde{K}_{r,\nu}\left(\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty}\left(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right)\right) \tag{B.21}$$

where $f_{r,\nu}(t) = (2\sqrt{\nu})^{-r-2\nu} t^{-r}$ if $t \geqslant (2\sqrt{\nu})^{-1}$ and $f_{r,\nu}(t) = t^{2\nu}$ if $t \leqslant (2\sqrt{\nu})^{-1}$.

Proof. For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^r$, $K_{r,\nu}(\boldsymbol{x} - \boldsymbol{y}) = \int_{\mathbb{R}^r} \widehat{K}_{r,\nu}(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} - \boldsymbol{y} \rangle} \mathrm{d}\boldsymbol{\omega}$.

$$\begin{aligned}
\sum_{j,k=1}^{n} \xi_j \xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right) &= \int_{\mathbb{R}^r} \widehat{K_{r,\nu}}(\boldsymbol{\omega}) \left| \sum_{j=1}^{n} \xi_j e^{i\langle \boldsymbol{\omega} | \boldsymbol{x}^{(j)} \boldsymbol{\mu} \rangle} \right|^2 \mathrm{d}\boldsymbol{\omega} \\
&= M_r(\nu) \|\boldsymbol{\mu}\|_\infty^{-r} \int_{\mathbb{R}^r} \left(4\nu + \|\boldsymbol{\mu}\|_\infty^{-2} \|\boldsymbol{s}\|^2\right)^{-\frac{r}{2}-\nu} \left| \sum_{j=1}^{n} \xi_j e^{i\langle \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty} \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 \mathrm{d}\boldsymbol{s} \\
&\geqslant 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\|\boldsymbol{\mu}\|_\infty) \int_{\mathbb{R}^r \setminus B(0,1)} \|\boldsymbol{s}\|^{-r-2\nu} \left| \sum_{j=1}^{n} \xi_j e^{i\langle \boldsymbol{s} | \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty} \boldsymbol{x}^{(j)} \rangle} \right|^2 \mathrm{d}\boldsymbol{s}.
\end{aligned} \tag{B.22}$$

Now, let $\widetilde{K}_{r,\nu}$ be the function with Fourier transform $\widehat{\widetilde{K}}_{r,\nu}(\boldsymbol{\omega}) = \mathbf{1}_{\{\|\boldsymbol{\omega}\| \geqslant 1\}} \|\boldsymbol{\omega}\|^{-r-2\nu}$. According to Bochner's theorem, $\widetilde{K}_{r,\nu}$ is a correlation kernel, which leads to the conclusion. □

Lemma B.17. *For every design set with coordinate-distinct points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, there exists a constant $c_{\boldsymbol{x}} > 0$ such that for all $\boldsymbol{\mu} \in (\mathbb{R}_+)^r$,*

$$\forall \boldsymbol{\xi} = (\xi_1, ..., \xi_n) \in \mathbb{R}^n, \ \sum_{j,k=1}^{n} \xi_j \xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right) \geqslant c_{\boldsymbol{x}} \|\boldsymbol{\xi}\|^2 2^{-\frac{r}{2}-\nu} M_r(\nu) f_{r,\nu}(\|\boldsymbol{\mu}\|_\infty) \tag{B.23}$$

where $f_{r,\nu}(t) = (2\sqrt{\nu})^{-r-2\nu} t^{-r}$ if $t \geqslant (2\sqrt{\nu})^{-1}$ and $f_{r,\nu}(t) = t^{2\nu}$ if $t \leqslant (2\sqrt{\nu})^{-1}$.

Proof. For every design set $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, the set of all design sets that can be written $\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty} \boldsymbol{x}^{(1)}, ..., \frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty} \boldsymbol{x}^{(n)}$ ($\boldsymbol{\mu} \in (\mathbb{R}_+)^r$) is compact. If the design set $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ has coordinate-distinct points, then every design set in the aforementioned compact set has no overlapping points. Thus the conclusion follows from Lemma B.16. □

Proposition B.18. *With Matérn anisotropic geometric or tensorized kernels, for every design set with coordinate-distinct points* $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$*, as* $\|\boldsymbol{\mu}\| \to 0$*,* $\left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| = O(\|\boldsymbol{\mu}\|^{-2\nu})$*.*

Proof. For Matérn anisotropic geometric kernels, we need only apply Lemma B.17. In the case of tensorized Matérn kernels, analogous results to Lemma B.16 and then Lemma B.17 may be used. □

Abramowitz and Stegun [1] give the following results on the modified Bessel function of second kind (usually noted K_ν and which we note $\mathcal{K}_\nu$ in order to avoid confusion with the Matérn correlation kernel). If I_ν is the modified Bessel function of first kind and ψ is the function defined in (6.3.2) by $\psi : \mathbb{N} \setminus \{0\} \to \mathbb{R}$; $k \mapsto -\gamma + \sum_{i=1}^{k-1} i^{-1}$:

$$I_\nu(z) = \left(\frac{1}{2}z\right)^\nu \sum_{k=0}^{\infty} \frac{\left(\frac{1}{4}z^2\right)^k}{k!\Gamma(\nu+k+1)} \qquad \text{(9.6.10 in [1])}$$

$$\mathcal{K}_\nu(z) = \frac{1}{2}\pi \frac{I_{-\nu}(z) - I_\nu(z)}{\sin(\nu z)} \quad \text{if } \nu \notin \mathbb{Z}. \qquad \text{(9.6.2 in [1])}$$

This gives us the series expansion of $K_{1,\nu}(z)$ ($\nu \in [0, +\infty) \setminus \mathbb{N}$) when $z \to 0$:

$$\begin{aligned} K_{1,\nu}(z) &= \frac{\pi}{\Gamma(\nu)\sin(\nu\pi)} \left(\sum_{0 \leqslant k < \nu} \frac{\nu^k z^{2k}}{k!\Gamma(-\nu+k+1)} - \frac{\nu^\nu z^{2\nu}}{\Gamma(\nu+1)} + o\left(z^{2\nu}\right) \right) \\ &= \frac{\pi}{\Gamma(\nu)\sin(\nu\pi)\Gamma(-\nu+1)} \left(\sum_{0 \leqslant k < \nu} \frac{\Gamma(-\nu+1)}{k!\Gamma(-\nu+k+1)} \nu^k z^{2k} - \frac{\Gamma(-\nu+1)}{\Gamma(\nu+1)} \nu^\nu z^{2\nu} + o\left(z^{2\nu}\right) \right) \\ &= \sum_{0 \leqslant k < \nu} \frac{\Gamma(-\nu+1)}{k!\Gamma(-\nu+k+1)} \nu^k z^{2k} + \frac{\Gamma(-\nu)}{\Gamma(\nu)} \nu^\nu z^{2\nu} + o\left(z^{2\nu}\right) \\ &= \sum_{0 \leqslant k < \nu} (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)} \nu^k z^{2k} + \frac{\Gamma(-\nu)}{\Gamma(\nu)} \nu^\nu z^{2\nu} + o\left(z^{2\nu}\right). \end{aligned} \tag{B.24}$$

In the remainder of this subsection, we consider a fixed design set with n coordinate-distinct points $\boldsymbol{x}^{(k)}$ ($k \in [\![1, n]\!]$) in $\mathbb{R}^r$. Moreover, all Matérn kernels we consider are assumed to have non-integer smoothness parameter ν.

Let us now define, for every nonnegative integer $k < \nu$ the matrix $\boldsymbol{D}^k$ whose (i, j) element is

$$\boldsymbol{D}^k(i,j) := (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)} \nu^k \left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right\|^{2k}. \tag{B.25}$$

Let us also define the matrix $\boldsymbol{D}^\nu$ whose (i, j) element is

$$\boldsymbol{D}^\nu(i,j) := \frac{\Gamma(-\nu)}{\Gamma(\nu)} \nu^\nu \left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right\|^{2\nu} \qquad \text{if } \nu \in [0, +\infty) \setminus \mathbb{N}. \tag{B.26}$$

If the correlation kernel is Matérn isotropic, $\boldsymbol{\Sigma}_\mu$ has the following series expansion if ν is not an integer when $\mu \to 0+$:

$$\boldsymbol{\Sigma}_\mu = \sum_{0 \leqslant k < \nu} \mu^{2k} \boldsymbol{D}^k + \mu^{2\nu} \boldsymbol{D}^\nu + \boldsymbol{R}_\mu. \tag{B.27}$$

In this expansion, $\mu^{-2\nu}\|\boldsymbol{R}_\mu\| \to 0$.

For any integer $i \in [\![1, r]\!]$ and any nonnegative integer $k < \nu$ define the matrix $\boldsymbol{D}_i^k$ whose (m, p) element is

$$\boldsymbol{D}_i^k(m,p) := (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)} \nu^k \left| x_i^{(m)} - x_i^{(p)} \right|^{2k} \tag{B.28}$$

and also the matrix $\boldsymbol{D}_i^\nu$ whose (m, p) element is

$$\boldsymbol{D}_i^\nu(m,p) := \frac{\Gamma(-\nu)}{\Gamma(\nu)} \nu^\nu \left| x_i^{(m)} - x_i^{(p)} \right|^{2\nu}. \tag{B.29}$$

For every $i \in [\![1, r]\!]$, if the points in the design set differed only through their ith coordinate, the series expansion of the correlation matrix (using a Matérn anisotropic geometric or tensorized kernel) when $\|\boldsymbol{\mu}\| \to 0$ (and thus when $\mu_i \to 0$) would be

$$\boldsymbol{\Sigma}_{\mu_i} = \sum_{0 \leqslant k < \nu} \mu_i^{2k} \boldsymbol{D}_i^k + \mu_i^{2\nu} \boldsymbol{D}_i^\nu + \boldsymbol{R}_{\mu_i} \tag{B.30}$$

where $\mu_i^{-2\nu} \|\boldsymbol{R}_{\mu_i}\| \to 0$ as $\mu_i \to 0$.

Note the following identities:

$$\boldsymbol{D}_i^0 = \mathbf{1}\mathbf{1}^\top; \tag{B.31}$$

$$\boldsymbol{D}_i^1 = -\frac{\Gamma(\nu-1)}{\Gamma(\nu)} \nu \left\{ \mathbf{1} \left(\boldsymbol{X}_i^{\circ 2}\right)^\top + \left(\boldsymbol{X}_i^{\circ 2}\right) \mathbf{1}^\top - 2\boldsymbol{X}_i \boldsymbol{X}_i^\top \right\}; \tag{B.32}$$

$$\boldsymbol{D}_i^2 = \frac{\Gamma(\nu-2)}{\Gamma(\nu)} \nu^2 \left\{ \mathbf{1} \left(\boldsymbol{X}_i^{\circ 4}\right)^\top + \left(\boldsymbol{X}_i^{\circ 4}\right) \mathbf{1}^\top - 4\boldsymbol{X}_i \left(\boldsymbol{X}_i^{\circ 3}\right)^\top - 4 \left(\boldsymbol{X}_i^{\circ 3}\right) \boldsymbol{X}_i^\top + 6 \left(\boldsymbol{X}_i^{\circ 2}\right) \left(\boldsymbol{X}_i^{\circ 2}\right)^\top \right\}. \tag{B.33}$$

If a tensorized correlation kernel is used, the correlation matrix $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}}$ may be written

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}} = \overset{\circ}{\prod_{i \in [\![1,r]\!]}} \boldsymbol{\Sigma}_{\mu_i} \tag{B.34}$$

where the subscript $\circ$ above the symbol $\prod$ serves to denote the Hadamard product of matrices.

In case a Matérn anisotropic geometric kernel is used, then define for any nonnegative interger $k < \nu$ the matrix $\boldsymbol{D}^k(\boldsymbol{\mu})$ whose (m, p) element is

$$\boldsymbol{D}^k(\boldsymbol{\mu})(m,p) := (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)} \nu^k d_{m,p}(\boldsymbol{\mu})^{2k} \tag{B.35}$$

where $d_{m,p}(\boldsymbol{\mu}) = \left\| \left(\boldsymbol{x}^{(m)} - \boldsymbol{x}^{(p)}\right) \boldsymbol{\mu} \right\|$.

And, similarly, we may define the matrix $\boldsymbol{D}^\nu(\boldsymbol{\mu})$ whose (m, p) element is

$$\boldsymbol{D}^\nu(\boldsymbol{\mu})(m,p) := \frac{\Gamma(-\nu)}{\Gamma(\nu)} \nu^\nu d_{m,p}(\boldsymbol{\mu})^{2\nu} \qquad \text{if } \nu \in [0, +\infty) \setminus \mathbb{N}. \tag{B.36}$$

We thus have (if $\nu \in [0,+\infty) \setminus \mathbb{N}$)

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{geom}} = \sum_{0 \leqslant k < \nu} \boldsymbol{D}^k(\boldsymbol{\mu}) + \boldsymbol{D}^\nu(\boldsymbol{\mu}) + \boldsymbol{R}_{\boldsymbol{\mu}}^{\text{geom}} \tag{B.37}$$

where $\|\boldsymbol{\mu}\|^{-2\nu} \, \|\boldsymbol{R}_{\boldsymbol{\mu}}^{\text{geom}}\| \to 0$ as $\|\boldsymbol{\mu}\| \to 0$.

Similar identities to those of equation (B.31) can be derived to make equation (B.37) more explicit for small values of ν.

$$\boldsymbol{D}^0(\boldsymbol{\mu}) = \mathbf{1}\mathbf{1}^\top; \tag{B.38}$$

$$\boldsymbol{D}^1(\boldsymbol{\mu}) = -\frac{\nu}{\nu-1}\left\{\sum_{i=1}^r \mu_i^2 \left(\mathbf{1}\left(\boldsymbol{X}_i^{\circ 2}\right)^\top + \left(\boldsymbol{X}_i^{\circ 2}\right)\mathbf{1}^\top - 2\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right\}; \tag{B.39}$$

$$\begin{aligned}\boldsymbol{D}^2(\boldsymbol{\mu}) = \frac{\nu^2}{(\nu-1)(\nu-2)}\Bigg\{ & \sum_{i,j\in[\![1,r]\!]} \mu_i^2\mu_j^2 \Big(\mathbf{1}\left(\boldsymbol{X}_i^{\circ 2} \circ \boldsymbol{X}_j^{\circ 2}\right)^\top + \left(\boldsymbol{X}_i^{\circ 2} \circ \boldsymbol{X}_j^{\circ 2}\right)\mathbf{1}^\top \\ & - 2\boldsymbol{X}_i\left(\boldsymbol{X}_i \circ \boldsymbol{X}_j^{\circ 2}\right)^\top - 2\left(\boldsymbol{X}_i \circ \boldsymbol{X}_j^{\circ 2}\right)\boldsymbol{X}_i^\top - 2\boldsymbol{X}_j\left(\boldsymbol{X}_j \circ \boldsymbol{X}_i^{\circ 2}\right)^\top - 2\left(\boldsymbol{X}_j \circ \boldsymbol{X}_i^{\circ 2}\right)\boldsymbol{X}_j^\top \\ & + \left(\boldsymbol{X}_i^{\circ 2}\right)\left(\boldsymbol{X}_j^{\circ 2}\right)^\top + \left(\boldsymbol{X}_j^{\circ 2}\right)\left(\boldsymbol{X}_i^{\circ 2}\right)^\top + 4\left(\boldsymbol{X}_i \circ \boldsymbol{X}_j\right)\left(\boldsymbol{X}_i \circ \boldsymbol{X}_j\right)^\top\Big)\Bigg\}.\end{aligned} \tag{B.40}$$

Fortunately, for small values of ν, $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}}$ can also be simply written.

$$\text{For } \nu \in (0,1): \quad \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}} = \mathbf{1}\mathbf{1}^\top + \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{D}_i^\nu + \boldsymbol{R}_{\boldsymbol{\mu}}^{\text{tens}}. \tag{B.41}$$

$$\text{For } \nu \in (1,2): \quad \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}} = \mathbf{1}\mathbf{1}^\top + \boldsymbol{D}^1(\boldsymbol{\mu}) + \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{D}_i^\nu + \boldsymbol{R}_{\boldsymbol{\mu}}^{\text{tens}}. \tag{B.42}$$

$$\text{For } \nu \in (2,3): \quad \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}} = \mathbf{1}\mathbf{1}^\top + \boldsymbol{D}^1(\boldsymbol{\mu}) + \frac{\nu-2}{\nu-1}\boldsymbol{D}^2(\boldsymbol{\mu}) + \sum_{i=1}^r \mu_i^4\boldsymbol{D}_i^2 + \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{D}_i^\nu + \boldsymbol{R}_{\boldsymbol{\mu}}^{\text{tens}}. \tag{B.43}$$

In the three expressions above, $\|\boldsymbol{\mu}\|^{-2\nu} \, \|\boldsymbol{R}_{\boldsymbol{\mu}}^{\text{tens}}\| \to 0$ as $\|\boldsymbol{\mu}\| \to 0$.

Define k_ν as the orthogonal complement in $\mathbb{R}^n$ of the vector space spanned by:

(1) if $\nu \in (0,1)$: $\mathbf{1}$;
(2) if $\nu \in (1,2)$: $\mathbf{1}$ and $\boldsymbol{X}_i$ ($i \in [\![1,r]\!]$);
(3) if $\nu \in (2,3)$: $\mathbf{1}$ and $\boldsymbol{X}_i$ ($i \in [\![1,r]\!]$) and $\boldsymbol{X}_i \circ \boldsymbol{X}_j$ ($i,j \in [\![1,r]\!]$).

Clearly, for any $\nu \in (0,1) \cup (1,2) \cup (2,3)$, for any vector $\boldsymbol{v} \in k_\nu$,

$$\boldsymbol{v}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{geom}} \boldsymbol{v} = \boldsymbol{v}^\top \boldsymbol{D}^\nu(\boldsymbol{\mu})\boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{R}_{\boldsymbol{\mu}}^{\text{geom}} \boldsymbol{v}, \tag{B.44}$$

$$\boldsymbol{v}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}} \boldsymbol{v} = \sum_{i=1}^r \mu_i^{2\nu} \boldsymbol{v}\boldsymbol{D}_i^\nu\boldsymbol{v} + \boldsymbol{v}^\top \boldsymbol{R}_{\boldsymbol{\mu}}^{\text{tens}} \boldsymbol{v}. \tag{B.45}$$

Since when $\mu \to 0$ $\|\boldsymbol{D}^\nu(\boldsymbol{\mu})\| = O(\|\boldsymbol{\mu}\|^{2\nu})$, $\|\boldsymbol{R}_{\boldsymbol{\mu}}^{\text{geom}}\| = o(\|\boldsymbol{\mu}\|^{2\nu})$ and $\|\boldsymbol{R}_{\boldsymbol{\mu}}^{\text{tens}}\| = o(\|\boldsymbol{\mu}\|^{2\nu})$, for any $\boldsymbol{\mu} \in (0,+\infty)^r$ such that $\|\boldsymbol{\mu}\|$ is small enough, there exists $c > 0$ such that for any $\boldsymbol{v} \in k_\nu$,

$$\max(\boldsymbol{v}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{geom}} \boldsymbol{v}, \boldsymbol{v}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{\text{tens}} \boldsymbol{v}) \leqslant c\|\boldsymbol{\mu}\|^{2\nu} \boldsymbol{v}^\top \boldsymbol{v}. \tag{B.46}$$

Proposition B.19. *For a Matérn anisotropic geometric or tensorized correlation kernel with smoothness parameter $\nu \in (0,1) \cup (1,2) \cup (2,3)$, for any vector $\boldsymbol{y} \in \mathbb{R}^n$ not orthogonal to k_ν, when $\|\boldsymbol{\mu}\| \to 0$, $\|\boldsymbol{\mu}\|^{-2\nu} = O\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}\right)$.*

Proof. Let d_ν be the dimension of k_ν and let $\boldsymbol{O}_\nu$ be an orthogonal $n \times n$ matrix whose first $n - d_\nu$ columns form an orthonormal basis of $k_\nu^\perp$ and whose last d_ν columns form an orthonormal basis of k_ν.

Then $\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = \boldsymbol{O}_\nu \boldsymbol{O}_\nu^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{O}_\nu \boldsymbol{O}_\nu^\top$. Consider the following decomposition of $\boldsymbol{O}_\nu^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{O}_\nu$:

$$\boldsymbol{O}_\nu^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{O}_\nu = \begin{pmatrix} \boldsymbol{A}_{\boldsymbol{\mu}} & \boldsymbol{B}_{\boldsymbol{\mu}} \\ \boldsymbol{B}_{\boldsymbol{\mu}}^\top & \boldsymbol{C}_{\boldsymbol{\mu}} \end{pmatrix} \tag{B.47}$$

where the blocks $\boldsymbol{A}_{\boldsymbol{\mu}}$, $\boldsymbol{B}_{\boldsymbol{\mu}}$ and $\boldsymbol{C}_{\boldsymbol{\mu}}$ are respectively $(n - d_\nu) \times (n - d_\mu)$, $(n - d_\mu) \times d_\mu$ and $d_\mu \times d_\mu$ matrices. Note that $\boldsymbol{A}_{\boldsymbol{\mu}}$ and $\boldsymbol{C}_{\boldsymbol{\mu}}$ represent the restriction of the scalar product defined by $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ to $k_\nu^\perp$ and k_ν respectively. When $\|\boldsymbol{\mu}\|$ is small enough, defining $c > 0$ as in equation (B.46), $\|\boldsymbol{C}_{\boldsymbol{\mu}}\| \leqslant c\|\boldsymbol{\mu}\|^{2\nu}$.

$$\boldsymbol{O}_\nu^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{O}_\nu = \begin{pmatrix} \boldsymbol{I}_{n-d_\nu} & \boldsymbol{0} \\ -\boldsymbol{B}_{\boldsymbol{\mu}} \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} & \boldsymbol{I}_{d_\nu} \end{pmatrix} \begin{pmatrix} \left(\boldsymbol{A} - \boldsymbol{B}_{\boldsymbol{\mu}} \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{B}_{\boldsymbol{\mu}}^\top\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{n-d_\nu} & -\boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{B}_{\boldsymbol{\mu}}^\top \\ \boldsymbol{0} & \boldsymbol{I}_{d_\nu} \end{pmatrix}. \tag{B.48}$$

For any vector $\boldsymbol{y} \in \mathbb{R}^n$, there exist $\boldsymbol{y}_1 \in \mathbb{R}^{n-d_\nu}$ and $\boldsymbol{y}_2 \in \mathbb{R}^{d_\nu}$ such that

$$\boldsymbol{O}_\nu^\top \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \tag{B.49}$$

$$\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 - \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{B}_{\boldsymbol{\mu}} \boldsymbol{y}_2 \\ \boldsymbol{y}_2 \end{pmatrix}^\top \begin{pmatrix} \left(\boldsymbol{A} - \boldsymbol{B}_{\boldsymbol{\mu}} \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{B}_{\boldsymbol{\mu}}^\top\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 - \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{B}_{\boldsymbol{\mu}} \boldsymbol{y}_2 \\ \boldsymbol{y}_2 \end{pmatrix}. \tag{B.50}$$

Given $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}$ is positive definite, the diagonal block $\left(\boldsymbol{A} - \boldsymbol{B}_{\boldsymbol{\mu}} \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{B}_{\boldsymbol{\mu}}^\top\right)^{-1}$ is positive definite too. This implies $\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y} \geqslant \boldsymbol{y}_2^\top \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}_2$. When $\|\boldsymbol{\mu}\|$ is small enough, $\boldsymbol{y}_2^\top \boldsymbol{C}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}_2 \geqslant c^{-1} \|\boldsymbol{\mu}\|^{-2\nu} \|\boldsymbol{y}_2\|^2$.

If $\boldsymbol{y}$ is not orthogonal to k_ν, then $\|\boldsymbol{y}_2\| \neq 0$ and thus $\|\boldsymbol{\mu}\|^{-2\nu} = O(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y})$.

□

Proposition B.20. *Assume $\nu \in (1,2) \cup (2,3)$. For every $\boldsymbol{y} \in \mathbb{R}^n$ that is not orthogonal to the vector subspace k_ν, $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i})$ is a bounded function of $\boldsymbol{\mu}$.*

Proof. Let $v_1(\boldsymbol{\mu}) \geqslant v_2(\boldsymbol{\mu}) \geqslant ... \geqslant v_n(\boldsymbol{\mu})$ be the ordered eigenvalues of $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$. We can now rewrite $L(\boldsymbol{y}|\boldsymbol{\mu})$ as

$$L(\boldsymbol{y}|\boldsymbol{\mu})^2 \propto \prod_{k=1}^{n} \left[v_k(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}\right)^{-1} \right]. \tag{B.51}$$

Proposition B.19 asserts that for any $\boldsymbol{y} \in \mathbb{R}^n$ that is not orthogonal to k_ν, $\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}\right)^{-1} = O(\|\boldsymbol{\mu}\|^{2\nu})$ for $\|\boldsymbol{\mu}\| \to 0$. Besides, Proposition B.18 asserts that $\left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| = O(\|\boldsymbol{\mu}\|^{-2\nu})$, so $\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}\right)^{-1} = O\left(\left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\|^{-1}\right)$. This implies that for every integer $i \in [\![1, r]\!]$, $v_k(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \boldsymbol{y}\right)^{-1} = O(1)$.

Clearly, $\lim_{\|\boldsymbol{\mu}\|\to 0} |\mathbf{1}^\top \boldsymbol{v}_1(\boldsymbol{\mu})| = \|\mathbf{1}\|$ and $\lim_{\|\boldsymbol{\mu}\|\to 0} v_1(\boldsymbol{\mu}) = n$. The latter implies $\lim_{\|\boldsymbol{\mu}\|\to 0} v_1(\boldsymbol{\mu})^{-1} = n^{-1}$ and $v_1(\boldsymbol{\mu})^{-1}\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{y}\right)^{-1} = O\left(\|\boldsymbol{\mu}\|^{2\nu}\right)$. Now, for every $\boldsymbol{\mu} \in (0,+\infty)^r$, we may (thanks to the axiom of choice) choose a unit eigenvector $\boldsymbol{v}_1(\boldsymbol{\mu})$ corresponding to the largest eigenvalue $v_1(\boldsymbol{\mu})$ and $\boldsymbol{v}_2(\boldsymbol{\mu})$ corresponding to the second largest eigenvalue $v_2(\boldsymbol{\mu})$. Because $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$ is symmetric, $\boldsymbol{v}_1(\boldsymbol{\mu})^\top \boldsymbol{v}_2(\boldsymbol{\mu}) = 0$ for all $\boldsymbol{\mu} \in (0,+\infty)^r$, so $\lim_{\|\boldsymbol{\mu}\|\to 0} \mathbf{1}^\top \boldsymbol{v}_2(\boldsymbol{\mu}) = 0$.

$$\begin{aligned} v_2(\boldsymbol{\mu}) &= (\mathbf{1}^\top \boldsymbol{v}_2(\boldsymbol{\mu}))^2 + 2\nu(\nu-1)^{-1}\sum_{i=1}^r \mu_i^2 \left(\boldsymbol{X}_i^\top \boldsymbol{v}_2(\boldsymbol{\mu})\right)^2 - 2\mu_i^2 \left(\mathbf{1}^\top \boldsymbol{v}_2(\boldsymbol{\mu})\right)\left(\boldsymbol{X}_i^{\circ 2\top} \boldsymbol{v}_2(\boldsymbol{\mu})\right) + O\left(\|\boldsymbol{\mu}\|^4\right) \\ &\geqslant 2\nu(\nu-1)^{-1}\sum_{i=1}^r \mu_i^2 (\boldsymbol{X}_i^\top \boldsymbol{v}_2(\boldsymbol{\mu}))^2 + o\left(\|\boldsymbol{\mu}\|^2\right). \end{aligned} \tag{B.52}$$

For all $\boldsymbol{\mu} \in (0,+\infty)^r$, let $i(\boldsymbol{\mu})$ be the smallest integer $i \in [\![1,r]\!]$ such that $\mu_i = \max_{j=1}^r \mu_j$. Now for every integer $i \in [\![1,r]\!]$ let $\boldsymbol{w}_i(\boldsymbol{\mu})$ be the unit vector that belongs to the space spanned by $\boldsymbol{v}_1(\boldsymbol{\mu})$ and $\boldsymbol{X}_i$ that verifies $\boldsymbol{v}_1(\boldsymbol{\mu})^\top \boldsymbol{w}_i(\boldsymbol{\mu}) = 0$ and $\boldsymbol{X}_i^\top \boldsymbol{w}_i(\boldsymbol{\mu}) > 0$.

$$\boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu})\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \geqslant 2\nu(\nu-1)^{-1} r^{-1}\|\boldsymbol{\mu}\|^2 (\boldsymbol{X}_{i(\boldsymbol{\mu})}^\top \boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}))^2 + o\left(\|\boldsymbol{\mu}\|^2\right). \tag{B.53}$$

Because $\lim_{\|\boldsymbol{\mu}\|\to 0} |\mathbf{1}^\top \boldsymbol{v}_1(\boldsymbol{\mu})| = \|\mathbf{1}\|$, $\liminf_{\|\boldsymbol{\mu}\|\to 0} \boldsymbol{X}_{i(\boldsymbol{\mu})}^\top \boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \geqslant \min_{i=1}^r \lim_{\|\boldsymbol{\mu}\|\to 0} \boldsymbol{X}_i^\top \boldsymbol{w}_i(\boldsymbol{\mu}) > 0$, so there exists a constant $c_2 > 0$ such that when $\|\boldsymbol{\mu}\|$ is small enough

$$\boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu})\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \geqslant c_2\|\boldsymbol{\mu}\|^2. \tag{B.54}$$

Recall $v_2(\boldsymbol{\mu}) = \max\{\boldsymbol{\xi}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{\xi} | \boldsymbol{\xi} \in S^{n-1} \text{ and } \boldsymbol{\xi}^\top \boldsymbol{v}_1(\boldsymbol{\mu})\}$, so *a fortiori* $v_2(\boldsymbol{\mu}) \geqslant c_2\|\boldsymbol{\mu}\|^2$.

This implies $v_2(\boldsymbol{\mu})^{-1} = O(\|\boldsymbol{\mu}\|^{-2})$ and therefore $v_2(\boldsymbol{\mu})^{-1}\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{y}\right)^{-1} = O(\|\boldsymbol{\mu}\|^{2(\nu-1)})$.

Finally, $L(\boldsymbol{y}|\boldsymbol{\mu}) = O\left(\|\boldsymbol{\mu}\|^\nu\right) O\left(\|\boldsymbol{\mu}\|^{\nu-1}\right) = O\left(\|\boldsymbol{\mu}\|^{2\nu-1}\right)$. Given that $f_i(\mu_i|\boldsymbol{\mu}_{-i}) = O(\|\boldsymbol{\mu}\|^{1-2\nu})$, the product $L(\boldsymbol{y}|\boldsymbol{\mu})f_i(\mu_i|\boldsymbol{\mu}_{-i})$ is bounded when $\|\boldsymbol{\mu}\| \to 0$. □

Proposition B.21. *Assume $\nu \in (0,1)$. For every $\boldsymbol{y} \in \mathbb{R}^n$ that is not collinear to $\mathbf{1}$, when $\|\boldsymbol{\mu}\| \to 0$, $L(\boldsymbol{y}|\boldsymbol{\mu})f_i(\mu_i|\boldsymbol{\mu}_{-i}) = O(\mu_i^{-1+\nu})$.*

Proof. This proof is similar to the previous one, so we use the same notations. $v_1(\boldsymbol{\mu})^{-1} = O(1)$, so $v_1(\boldsymbol{\mu})^{-1}\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{y}\right)^{-1} = O(\|\boldsymbol{\mu}\|^{2\nu})$, which yields that $L(\boldsymbol{y}|\boldsymbol{\mu}) = O(\|\boldsymbol{\mu}\|^\nu)$.

This implies that $L(\boldsymbol{y}|\boldsymbol{\mu})\left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| = O(\|\boldsymbol{\mu}\|^{-\nu}) = O(\mu_i^{-\nu})$.

By Lemma B.3, $\left\|\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\| = O(\mu_i^{-1+2\nu})$. Putting all this together, $L(\boldsymbol{y}|\boldsymbol{\mu})f_i(\mu_i|\boldsymbol{\mu}_{-i}) = O(\mu_i^{-1+\nu})$. □

Proposition B.22. *For Matérn anisotropic geometric or tensorized kernels with smoothness parameter $\nu \in (0,1)\cup(1,2)\cup(2,3)$, if $\boldsymbol{y} \in \mathbb{R}^n$ is not orthogonal to k_ν, then the conditional posterior distribution $\pi_i(\mu_i|\boldsymbol{y},\boldsymbol{\mu}_{-i})$, seen as a function of $\boldsymbol{\mu}$, is continuous over $\{\boldsymbol{\mu} \in [0,+\infty)^r : \mu_i \neq 0\}$.*

Moreover,

$$\forall \mu_i > 0, \quad \pi_i(\mu_i|\boldsymbol{y},\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1}) = \frac{L(\boldsymbol{y}|\mu_i,\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1})f_i(\mu_i|\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1})}{\int_0^\infty L(\boldsymbol{y}|\mu_i = t,\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1})f_i(\mu_i = t|\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1})\mathrm{d}t} > 0.$$

Proof. Given Proposition B.4 and the fact that $\forall \boldsymbol{y} \in \mathbb{R}^n$, $\forall i \in [\![1, r]\!]$ and $\forall \mu_i \in (0, +\infty)$, as $\|\boldsymbol{\mu}_{-i}\| \to 0$, $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i})$ converges pointwise to $L(\boldsymbol{y}|\mu_i, \boldsymbol{\mu}_{-1} = \mathbf{0}_{r-1}) f_i(\mu_i|\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1})$, we only need to show that

$$\int_0^\infty L(\boldsymbol{y}|\mu_i = t, \boldsymbol{\mu}_{-i}) f_i(\mu_i = t|\boldsymbol{\mu}_{-i}) \mathrm{d}t \underset{\|\boldsymbol{\mu}_{-i}\| \to 0}{\longrightarrow} \int_0^\infty L(\boldsymbol{y}|\mu_i = t, \boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1}) f_i(\mu_i = t|\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1}) \mathrm{d}t < +\infty. \tag{B.55}$$

Lemma B.3 implies that there exists $M_i > 0$ such that

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i}) &= L(\boldsymbol{y}|\boldsymbol{\mu}) \left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| \left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\|^{-1} f_i(\mu_i|\boldsymbol{\mu}_{-i}) \\ &\leqslant L(\boldsymbol{y}|\boldsymbol{\mu}) \left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| M_i \mu_i^{-2}. \end{aligned} \tag{B.56}$$

Lemma B.6 then ensures $\left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} - \boldsymbol{I}_n\right\| \to 0$ as $\mu_i \to +\infty$, so the right member of the inequality is integrable in the neighborhood of $+\infty$. Let us now focus on the neighborhood of 0.

If $\nu > 1$, Proposition B.20 asserts that $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i})$ is bounded in the neighborhood of 0.

If $\nu < 1$, Proposition B.21 asserts that $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i}) \mu_i^{1-\nu}$ is bounded in the neighborhood of 0.

Therefore, there exists a function independent of $\boldsymbol{\mu}_{-i}$ that is both greater than $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i})$ and integrable over $\mu_i \in (0, +\infty)$, so the dominated convergence theorem is applicable.

□

B.4 Lower bound for conditional reference posterior densities

The following lemma provides the key to proving Theorem 4.2:

Lemma B.23. *In a Simple Kriging model with the characteristics described above, there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that, for any $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$ and any $i \in [\![1, r]\!]$, there exists a measurable function $m_{i,\boldsymbol{y}} : (0, +\infty) \to (0, +\infty)$ such that, for all $\boldsymbol{\mu}_{-i} \in (0, +\infty)^{r-1}$, the conditional reference posterior density verifies:*

$$\pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) \geqslant m_{i,\boldsymbol{y}}(\mu_i) > 0. \tag{B.57}$$

Proof. This proof consists in combining Propositions B.15 and B.22, which respectively deal with large and small values of $\|\boldsymbol{\mu}_{-i}\|$.

Proposition B.15 implies that for any $\boldsymbol{y} \in \mathbb{R}^n \setminus \{0\}^n$, for any $i \in [\![1, r]\!]$ and any $\mu_i \in (0, +\infty)$, there exists a compact neighborhood $N_i(\mu_i)$ of $\mathbf{0}_{r-1}$ within $[0, +\infty)^r$ such that

$$\inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in [0, +\infty)^{r-1} \setminus N_i(\mu_i)\} > 0. \tag{B.58}$$

The vector space $k_\nu \subset \mathbb{R}^n$ has dimension greater or equal to

(a) $n - 1$ if $\nu \in (0, 1)$;
(b) $n - (r + 1)$ if $\nu \in (1, 2)$;
(c) $n - (r + 1)(r + 2)/2$ if $\nu \in (2, 3)$.

For all Simple Kriging models tackled by this lemma, the dimension of k_ν is therefore greater or equal to 1. Its orthogonal complement $k_\nu^\perp$ is then included within a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$. Assuming $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, Proposition B.22 ensures that $\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i})$ is a continuous and positive function of $\boldsymbol{\mu}$ on $\{\boldsymbol{\mu} \in [0, +\infty)^r : \mu_i \neq 0\}$. In particular, this implies that for any $\mu_i \in (0, +\infty)$ and any compact neighborhood $N_i(\mu_i)$ of $\mathbf{0}_{r-1}$ within $[0, +\infty)^r$,

$$\inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in N_i(\mu_i)\} > 0. \tag{B.59}$$

Putting this together, if $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, for any $i \in [\![1, r]\!]$ and any $\mu_i \in (0, +\infty)$,

$$\inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in [0, +\infty)^{r-1}\} > 0. \tag{B.60}$$

The mapping $m_{i,\boldsymbol{y}} : \mu_i \mapsto \inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in [0, +\infty)^{r-1}\}$, which is measurable on $(0, +\infty)$ is therefore also positive on $(0, +\infty)$.

□

Proof of Theorem 4.2. Lemma B.23 implies that $\forall \boldsymbol{\mu}^{(0)} \in (0, +\infty)^r$,

$$P_{\boldsymbol{y}}(\boldsymbol{\mu}^{(0)}, \mathrm{d}\boldsymbol{\mu}) \geqslant \frac{1}{r} \sum_{i=1}^{r} m_{i,\boldsymbol{y}}(\mu_i)\mathrm{d}\mu_i \, \delta_{\boldsymbol{\mu}_{-i}^{(0)}}(\mathrm{d}\boldsymbol{\mu}_{-i}),$$

and thus $\forall \boldsymbol{\mu}^{(0)} \in (0, +\infty)^r$, $\forall n \geqslant r$,

$$P_{\boldsymbol{y}}^n(\boldsymbol{\mu}^{(0)}, \mathrm{d}\boldsymbol{\mu}) \geqslant \frac{1}{r^r} \prod_{i=1}^{r} m_{i,\boldsymbol{y}}(\mu_i)\mathrm{d}\mu_i.$$

Defining $f_{\boldsymbol{y}}(\boldsymbol{\mu}) := r^{-r} \prod_{i=1}^{r} m_{i,\boldsymbol{y}}(\mu_i)$, $f_{\boldsymbol{y}}$ is a measurable positive function. Therefore, $f_{\boldsymbol{y}}$ is the density with respect to the Lebesgue measure of a positive measure with mass $\epsilon_{\boldsymbol{y}} > 0$. So $\epsilon_{\boldsymbol{y}}^{-1} f_{\boldsymbol{y}}$ is a probability density with respect to the Lebesgue measure and the Markov kernel $P_{\boldsymbol{y}}$ thus satisfies the uniform $(n, \epsilon_{\boldsymbol{y}})$ Doeblin condition:

$$\forall \boldsymbol{\mu}^{(0)} \in (0, +\infty)^r \qquad P_{\boldsymbol{y}}^n(\boldsymbol{\mu}^{(0)}, \mathrm{d}\boldsymbol{\mu}) \geqslant \epsilon_{\boldsymbol{y}} \left(\frac{1}{\epsilon_{\boldsymbol{y}}} f_{\boldsymbol{y}}(\boldsymbol{\mu}) \right) \mathrm{d}\boldsymbol{\mu}. \tag{B.61}$$

This implies that $P_{\boldsymbol{y}}$ is uniformly ergodic: it has a unique invariant probability distribution $\pi_G(\cdot|\boldsymbol{y})$ and $\lim_{n\to\infty} \sup_{\boldsymbol{\mu}^{(0)} \in (0,+\infty)^r} \|P_{\boldsymbol{y}}^n(\boldsymbol{\mu}^{(0)}, \cdot) - \pi_G(\cdot|\boldsymbol{y})\|_{TV} = 0$, where $\|\cdot\|_{TV}$ is the total variation norm. By definition, $\pi_G(\cdot|\boldsymbol{y})$ is the Gibbs compromise between the incompatible posterior conditionals $\pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i})$.

□

Acknowledgements. The author would like to thank his PhD advisor Professor Josselin Garnier (École Polytechnique, Centre de Mathématiques Appliquées) for his guidance, Loic Le Gratiet (EDF R&D, Chatou) and Anne Dutfoy (EDF R&D, Saclay) for their advice and helpful suggestions. The author also thank the Associate Editor and both anonymous reviewers for their constructive criticism which considerably improved this paper. The author acknowledges the support of the French Agence Nationale de la Recherche (ANR), under grant ANR-13-MONU-0005 (project CHORUS).

References

[1] M. Abramowitz and I.A. Stegun, Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Vol. 55 of *Applied Mathematics Series*. National Bureau of Standards (1964).
[2] E. Anderes, On the consistent separation of scale and variance for Gaussian random fields. *Ann. Stat.* **38** (2010) 870–893.
[3] B.C. Arnold, E. Castillo and J.M. Sarabia, Conditionally specified distributions: an introduction. *Stat. Sci.* **16** (2001) 268–269.
[4] J. Berger, The case for objective bayesian analysis. *Bayesian Anal.* **1** (2006) 385–402.
[5] J.O. Berger and J.M. Bernardo, On the development of reference priors. *Bayesian Stat.* **4** (1992) 35–60.
[6] J.O. Berger, V. De Oliveira and B. Sansó, Objective Bayesian analysis of spatially correlated data. *J. Am. Stat. Assoc.* **96** (2001) 1361–1374.
[7] J.O. Berger, J.M. Bernardo and D. Sun. Overall objective priors. *Bayesian Anal.* **10** (2015) 189–221.
[8] J.M. Bernardo, Reference analysis, in Vol. 25 of Handbook of Statistics, edited by D. Dey and C. Rao. Elsevier (2005) 17–90.
[9] G. Celeux, J.-M. Marin and C. Robert, Sélection bayésienne de variables en régression linéaire. *J. Soc. Fr. Statistique* **147** (2005) 59–79.
[10] B.S. Clarke and A.R. Barron, Jeffreys' prior is asymptotically least favorable under entropy risk. *J. Stat. Plan. Inference* **41** (1994) 37–60.

[11] A.P. Dawid and S.L. Lauritzen, Compatible prior distributions. Bayesian Methods with Applications to Science, Policy and Official Statistics, Selected Papers from ISBA 2000: The Sixth World Meeting of the International Society for Bayesian Analysis, edited by E.I. George. Eurostat, Luxembourg (2001) 109–118.
[12] A. Gelman and T.E. Raghunathan, Comment on "Conditionally specified distributions: an introduction" by B.C. Arnold, E. Castillo and J.M. Sarabia. *Stat. Sci.* **16** (2001) 268–269.
[13] M. Gu, X. Wang and J.O. Berger, Robust Gaussian stochastic process emulation. *Ann. Stat.* **46** (2018) 3038–3066.
[14] B.D. He, C.M. De Sa, I. Mitliagkas and C. Ré, Scan order in Gibbs sampling: Models in which it matters and bounds on how much. *Adv. Neural Inf. Process. Syst.* **29** (2016) 1–9.
[15] D. Heckerman, D.M. Chickering, C. Meek, R. Rounthwaite and C. Kadie, Dependency networks for inference, collaborative filtering, and data visualization. *J. Mach. Learn. Res.* **1** (2000) 49–75.
[16] J.P. Hobert and G. Casella, Functional compatibility, Markov chains, and Gibbs sampling with improper posteriors. *J. Comput. Graph. Stat.* **7** (1998) 42–60.
[17] H. Hotelling, New light on the correlation coefficient and its transforms. *J. R. Stat. Soc. Ser. B (Methodol.)* **15** (1953) 193–232.
[18] A.G. Journel and Ch. J. Huijbregts, Mining Geostatistics. Academic Press, New York (1978).
[19] H. Kazianka and J. Pilz, Objective bayesian analysis of spatial data with uncertain nugget and range parameters. *Can. J. Stat.* **40** (2012) 304–327.
[20] M.C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *J. R. Stat. Soc., Ser. B (Stat. Methodol.)* **63** (2001) 425–464.
[21] K.L. Kuo and Y.J. Wang, Pseudo-Gibbs sampler for discrete conditional distributions. *Ann. Inst. Stat. Math.* (2017) 1–13.
[22] K.-L. Kuo, C.-C. Song and T.J. Jiang, Exactly and almost compatible joint distributions for high-dimensional discrete conditional distributions. *J. Multivar. Anal.* **157** (2017) 115–123.
[23] R. Li and A. Sudjianto, Analysis of computer experiments using penalized likelihood in Gaussian Kriging models. *Technometrics* **47** (2005) 111–120.
[24] I. Mitliagkas and L. Mackey, Improving Gibbs Sampler Scan Quality with DoGS, in Vol. 70 of *Proceedings of the 34th International Conference on Machine Learning*. (2017) 2469–2477.
[25] R. Paulo, Default priors for Gaussian processes. *Ann. Stat.* **33** (2005) 556–582.
[26] C.E. Rasmussen and C.K.I. Williams, *Gaussian Processes for Machine Learning*. MIT Press (2006).
[27] C. Ren, D. Sun and C. He, Objective bayesian analysis for a spatial model with nugget effects. *J. Stat. Plan. Inference* **142** (2012) 1933–1946.
[28] C. Ren, D. Sun and S.K. Sahu, Objective bayesian analysis of spatial models with separable correlation functions. *Can. J. Stat.* **41** (2013) 488–507.
[29] C.P. Robert, The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation. Springer-Verlag, New York (2007).
[30] A. Roverato and G. Consonni, Compatible prior distributions for directed acyclic graph models. *J. R. Stat. Soc., Ser. B (Stat. Methodol.)* **66** (2004) 47–61.
[31] T.J. Santner, B.J. Williams and W.I. Notz, The Design and Analysis of Computer Experiments. Springer-Verlag, New York (2003).
[32] M.L. Stein, Interpolation of Spatial Data. Some Theory for Kriging. *Springer Series in Statistics*. Springer-Verlag, New York (1999).
[33] R. Yang and J.O. Berger, A Catalog of Noninformative Priors. Institute of Statistics and Decision Sciences, Duke University (1996).
[34] B. Yet and W. Marsh, Compatible and incompatible abstractions in Bayesian networks. *Knowl.-Based Syst.* **62** (2014) 84–97.
[35] H. Zhang, Inconsistent Estimation and Asymptotically Equal Interpolations in Model-based Geostatistics. *J. Am. Stat. Assoc.* **99** (2004) 250–261.