Objective Bayesian analysis of Kriging models with anisotropic correlation kernel

Joseph Muré

Université Sorbonne
Paris Cité

École Doctorale Paris Centre

Thèse de doctorat

Discipline : Mathématiques

Thèse de l'Université Sorbonne Paris Cité
préparée à l'Université Paris Diderot

présentée par

Joseph Muré

Objective Bayesian analysis of Kriging models with anisotropic correlation kernel

---

dirigée par Josselin Garnier

Soutenue publiquement à Paris le 5 octobre 2018 devant le jury composé de :

Pr Clémentine Prieur	Université Grenoble Alpes	présidente
Pr Kerrie Mengersen	Queensland University of Technology	rapporteure
Pr Jean-Michel Marin	Université de Montpellier	rapporteur
Pr Josselin Garnier	École Polytechnique	directeur
Pr Stéphane Boucheron	Université Paris Diderot	examinateur
Dr Loïc Le Gratiet	EDF R&D Chatou	encadrant
Mme Anne Dutfoy	EDF R&D Saclay	encadrante

Thèse réalisée dans le cadre d'un Contrat de Formation par la Recherche à EDF.

Remerciements

Je remercie en premier lieu Josselin Garnier, mon directeur de thèse. Tout au long de ces trois années de travail, il a veillé à la rigueur de mes travaux, en étant pleinement disponible pour suivre leurs progrès, semaine après semaine. Il m'a laissé conduire mes recherches dans les directions qui m'intéressaient, ce dont je lui suis profondément reconnaissant.

Je remercie aussi Loïc Le Gratiet, mon premier encadrant à EDF, qui m'a fait confiance par deux fois, d'abord au cours du stage préliminaire que j'ai réalisé à EDF, puis pour effectuer ce travail de thèse. Je lui suis reconnaissant pour ses précieux conseils et ses nombreux encouragements.

Je remercie de même Anne Dutfoy qui, ayant rejoint l'encadrement de ma thèse en cours de route, a immédiatement saisi ses enjeux, m'a apporté son aide et m'a témoigné une grande bienveillance.

Je suis honoré que Kerrie Mengersen et Jean-Michel Marin soient rapporteurs de ma thèse. Leurs rapports très détaillés m'ont permis d'améliorer la clarté de ce travail, et leurs encouragements me poussent à le faire connaître. Je les remercie pour cela. *I am honored that Kerrie Mengersen and Jean-Michel Marin reviewed this dissertation. Their detailed recommentations allowed me to make this work clearer, and I will keep their encouragement in mind when disseminating it. I thank them for that.* Je remercie aussi Clémentine Prieur et Stéphane Boucheron, qui ont accepté d'être membres de mon jury de thèse.

Je veux remercier tous mes collègues et ex-collègues du département PRISME d'EDF R&D, pour l'ambiance de travail chaleureuse qu'ils font régner et qui est si essentielle pour mener à bien avec sérénité un travail de longue haleine comme une thèse de doctorat. Je tiens à remercier particulièrement Merlin Keller et Nicolas Bousquet, qui m'ont souvent donné de judicieux conseils en me faisant profiter de leur expérience.

Je remercie le Laboratoire de Probabilités, Statistique et Modélisation de Sorbonne Université pour son accueil et la mise à disposition de ses ressources.

Je veux remercier aussi les doctorants des départements PRISME et PERICLES d'EDF R&D, dont l'entraide est vraiment un atout précieux. Je pense tout particulièrement à ceux qui sont déjà partis, Nazih Benoumechiara, Xavier Yau, Tuan Dinh Trong, Thomas Browne, et notamment à nos vacances inoubliables à Aussois. Et à ceux qui sont toujours là, Thomas Galtier, Jérôme Stenger, Thomas Bittar, Pablo Pereira-Alvarez, Mouna Rifi, Léna Masson, Artémis Drakos, Aurélie Daanen, Déborah Fontaine.

Enfin, je tiens à rmercier mes parents pour leur soutien constant, ainsi que Jérémie Genty, Arthur Soulié, Guillaume Metzler, amis fidèles sans qui je ne serais pas arrivé jusqu'à ce point.

Abstract

A recurring problem in surrogate modelling is the scarcity of available data which hinders efforts to estimate model parameters. The Bayesian paradigm offers an elegant way to circumvent the problem by describing knowledge of the parameters by a posterior probability distribution instead of a pointwise estimate. However, it involves defining a prior distribution on the parameter. In the absence of expert opinion, finding an adequate prior can be a trying exercise. The Objective Bayesian school proposes default priors for such situations, like the Berger-Bernardo reference prior. Such a prior was derived by Berger et al. [2001] for the Kriging surrogate model with isotropic covariance kernel. Directly extending it to anisotropic kernels poses theoretical as well as practical problems because the reference prior framework requires ordering the parameters. Any ordering would in this case be arbitrary. Instead, we propose an Objective Bayesian solution for Kriging models with anisotropic covariance kernels based on conditional reference posterior distributions. This solution is made possible by a theory of compromise between incompatible conditional distributions. The work is then shown to be compatible with Trans-Gaussian Kriging. It is applied to an industrial case with nonstationary data in order to derive Probability Of defect Detection (POD) by non-destructive tests in steam generator tubes of nuclear power plants.

Keywords: Incompatibility, Conditional distribution, Markov kernel, Optimal compromise, Kriging, Reference prior.

Résumé

Les métamodèles statistiques sont régulièrement confrontés au manque de données qui engendre des difficultés à estimer les paramètres. Le paradigme bayésien fournit un moyen élégant de contourner le problème en décrivant la connaissance que nous avons des paramètres par une loi de probabilité *a posteriori* au lieu de la résumer par une estimation ponctuelle. Cependant, ce paradigme nécessite de définir un loi *a priori* adéquate, ce qui est un exercice difficile en l'absence de jugement d'expert. L'école bayésienne objective propose des priors par défaut dans ce genre de situation tels que le prior de référence de Berger-Bernardo. Un tel prior a été calculé par Berger et al. [2001] pour le modèle de krigeage avec noyau de covariance isotrope. Une extension directe au cas des noyaux anisotropes poserait des problèmes théoriques aussi bien que pratiques car la théorie de Berger-Bernardo ne peut s'appliquer qu'à un jeu de paramètres ordonnés. Or dans ce cas de figure, tout ordre serait nécessairement arbitraire. Nous y substituons une solution bayésienne objective fondée sur les posteriors de référence conditionnels. Cette solution est rendue possible par une théorie du compromis entre lois conditionnelles incompatibles. Nous montrons en outre qu'elle est compatible avec le krigeage trans-gaussien. Elle est appliquée à un cas industriel avec des données non-stationnaires afin de calculer des Probabilités de Détection de défauts (POD de l'anglais *Probability Of Detection*) par tests non-destructifs dans les tubes de générateur de vapeur de centrales nucléaires.

Mots-clés : Incompatibilité, Loi conditionnelle, Noyau markovien, Compromis optimal, Krigeage, Prior de référence.

Contents

Contents **5**

Introduction **7**

I Tools **11**

1 Kriging Overview **13**
1.1 Introduction . . . 13
1.2 Gaussian random processes . . . 14
1.3 Mean square continuity and differentiability of Gaussian processes . . . 20
1.4 Spectral representation . . . 25
1.5 Examples of correlation kernels . . . 31
1.6 Current Kriging-related research . . . 33

2 Reference Prior Theory **35**
2.1 Introduction . . . 35
2.2 Basic idea . . . 36
2.3 Full definition of the reference prior . . . 41
2.4 Regular continuous case . . . 49
2.5 Properties of reference priors . . . 55
2.6 Examples . . . 56
2.7 Reference priors for multiparametric models . . . 59

3 Propriety of the reference posterior distribution in Gaussian Process regression **65**
3.1 Introduction . . . 65
3.2 Setting . . . 66
3.3 Smoothness of the correlation kernel . . . 68
3.4 Propriety of the reference posterior distribution . . . 70
3.5 Conclusion . . . 72
Appendices . . . 74
3.A Algebraic facts . . . 74
3.B Maclaurin series . . . 77
3.C Spectral decomposition . . . 81
3.D Asymptotic study of the correlation matrix $\mathbf{\Sigma}_\theta$ . . . 85
3.E Details of the proof of Theorem 3.9 . . . 88

II Compromise **93**

4 Optimal compromise between incompatible conditional probability distributions **95**
4.1 Introduction . . . 95
4.2 Optimal compromise: a general theory . . . 96
4.3 Discussion of the notion of compromise . . . 102
4.4 Conclusion . . . 108

III Application **109**

5 Application of the Optimal Compromise to Simple Kriging models with Matérn correlation kernels **111**
5.1 Introduction . . . 111
5.2 Optimal compromise between Objective Posterior conditional distributions in Gaussian Process regression . . . 113
5.3 Comparisons between the MLE and MAP estimators . . . 117
5.4 Comparison of the predictive distributions associated with the estimators (MLE and MAP) and the full posterior distribution . . . 120
5.5 Conclusion and Perspectives . . . 124
Appendices . . . 126
5.A Proofs of Section 5.2 . . . 126

6 A Comprehensive Bayesian Treatment of the Universal Kriging model with Matérn correlation kernels **139**
6.1 Introduction . . . 140
6.2 Analytical treatment of the location-scale parameters . . . 141
6.3 Reference prior on a one-dimensional $\boldsymbol{\theta}$ . . . 145
6.4 The Gibbs reference posterior on a multi-dimensional $\boldsymbol{\theta}$ . . . 145
6.5 Comparison of the predictive performance of the full-Bayesian approach versus MLE and MAP plug-in approaches . . . 148
6.6 Conclusion . . . 156
Appendices . . . 159
6.A Matérn kernels . . . 159
6.B Proofs of the existence of the Gibbs reference posterior . . . 159

7 Trans-Gaussian Kriging in a Bayesian framework: a case study **167**
7.1 Introduction . . . 167
7.2 Probability Of Detection (POD) . . . 169
7.3 An Objective Bayesian outlook to Trans-Gaussian Kriging . . . 170
7.4 Industrial Application . . . 177
7.5 Conclusion . . . 181

Conclusion **183**

Bibliography **185**

Introduction

Objective

This thesis was funded by EDF and the ANR in order to solve a specific problem of emulation of a computer experiment in a situation with drastic budgetary constraints.

Emulation of computer codes has become in the last decades an object of considerable attention from the statistical community [Santner et al., 2003]. Surrogate models are statistical models designed to represent uncertainty about the result of experiments when the computational budget is limited. Linear regression, neural networks, polynomial chaos and Kriging are examples of surrogate models.

The problem EDF was facing was computing Probability Of defect Detection (POD) curves for a non-destructive test based on eddy-currents. A computer code was available to simulate such a procedure under various parameter choices. Because of the cost involved in making this code run, it was desirable to design a surrogate model that could be used in its stead. Because of the intrinsic Uncertainty Quantification capability of the Kriging model, which can easily provide prediction intervals or – critically for this application – probabilities of reaching a threshold, it was a natural choice. An apparent difficulty, however, was that the data collected seemed to defy the traditional Kriging assumption that a stationary Gaussian process could adequately represent them. There seemed to be a need to define non-Gaussian, non-stationary Kriging procedures.

Contributions of the thesis

It turned out, however, that a simple transformation of the data could tackle these problems, so there was no need to try any of the sophisticated procedures available in the literature. The true issue, which was not so easily solved, was that the relative lack of data (100 points dispersed in a 9-dimensional space) made estimating the Kriging hyperparameters difficult. This might not have been significant if they had not had a tremendous impact on prediction.

In the absence of a convincing estimator, it seemed reasonable to try to embed hyperparameter uncertainty into the prediction. And in this regard, the Bayesian approach seemed adequate. This framework treats the parameter as a random variable rather than as an unknown quantity. The associated probability distribution does not express actual randomness but is a quantification of an opinion about the value of the parameter.

Although philosophically coherent, this subjectivist view could not be used in this particular instance because there was no prior opinion that could be encoded by a Bayesian prior dis-

tribution. For the sake of convenience, it was necessary to find a generic, default approach in the absence of relevant information. The Berger-Bernardo [Bernardo, 1979a, Berger and Bernardo, 1992, Bernardo, 2005, Berger et al., 2009] "reference prior" theory provided a quasi-systematic approach that was quite appealing. It relies on a formally defined criterion of noninformativity aiming to "let the data speak for themselves".

The reference prior was derived by Berger et al. [2001] for the Kriging model with isotropic correlation kernels – which require only one correlation parameter. They provided a proof, which we show is valid for a subset of all considered correlation kernels. Moreover, we contribute a proof for the complementary subset.

Aside from this technical point, reference prior theory does leave some wiggle room for models with several parameters. The user is asked to provide an ordering from least to most important. Such an ordering is not always easy to determine, especially when the ultimate goal of the study is predictive rather than inferential. For the model at hand, a partial, but not total, order on the parameters could be derived.

This situation led to the development of a new technique designed to bypass this requirement and resulting in the "Gibbs reference posterior distribution" as a way to quantify parameter uncertainty after having observed the available data in the absence of parameter ordering.

This technique relies on the novel notion of optimal compromise between potentially incompatible conditional distributions. Thus reference priors for every single parameter conditional to all others can be computed. The "Gibbs reference posterior distribution" distribution is defined as the optimal compromise between the resulting conditional reference posterior distributions. We prove that it is proper and therefore usable for Bayesian inference and prediction.

Combining the data transformation and the Gibbs reference posterior distribution produced a full solution to the surrogate model problem by naturally including uncertainty about the parameters of the model in the computation of Probability Of defect Detection curves.

Organization of the text

The dissertation is organized in three parts.

The first part, "Tools", is a description of mathematical objects that were needed to solve the problem. It contains the "state-of-the-art" chapters.

Chapter 1 presents Kriging. Gaussian processes are defined and characterized in terms of mean function and correlation kernel. The crucial notion of stationarity is presented, and the link between stationarity of Gaussian process and stationarity of its covariance kernel is explained. The definition of mean square continuity and differentiability of Gaussian processes is then provided. Special attention is devoted to the way stationarity grounds this theory in standard differential calculus. The chapter ends with a presentation of standard covariance kernels and a short discussion of their properties.

Chapter 2 is independent from Chapter 1 and can be read first if desired. It presents the parts of reference analysis that are relevant to this thesis. It endeavors to make the formally difficult definition of a reference prior as intuitive as possible. The chapter is mainly devoted to the

one-parameter setting, because this is the setting where the theory is the most developed and yields the most interesting results. Although mainly a review of previous results, a small original contribution is added in the form of a nice uniqueness result. The final part of the chapter deals with the murkier multi-parameter setting. The approach presented is a departure from the theory established in Berger and Bernardo [1992], but it is nothing more than a formalization of the concept of "exact marginalization" prescribed by Berger et al. [2001] for Kriging models.

Chapter 3 combines notions from Chapters 1 and 2 insofar as it concerns reference analysis of Kriging models with isotropic covariance kernels. With regard to its result, it is a state-of-the-art chapter: it relates results from Berger et al. [2001]. However, it shows that Berger et al. [2001]'s proof is adequate for rough Gaussian processes only and gives an original proof for smoother processes.

The second part, "Compromise" is the shortest part – it contains a single chapter – but it plays a crucial role in this thesis.

Chapter 4 proposes a theory of compromise between possibly incompatible conditional distributions. It could be read independently from the rest of the dissertation because it does not specifically concern spatial or even Bayesian statistics. It tackles a problem that arises from the popularity of Gibbs samplers: sometimes, there is no guarantee that the conditional distributions truly define a joint distribution. However, the chapter shows that if it exists and is unique, the stationary distribution of the Markov chain associated with the Gibbs sampler can be viewed as the optimal compromise between the input conditional distributions.

The third part, "Application" details the consequences of the aforementioned theory of compromise for Objective Bayesian analysis of Kriging models with anisotropic covariance kernels.

Chapter 5 details how the theory of compromise makes it possible, in a Simple Kriging setting, not to order correlation parameters and instead use the optimal compromise between the posterior distributions derived from conditional reference priors. Numerical simulations show that this optimal compromise, the Gibbs reference posterior distribution, has good frequentist coverage properties.

Chapter 6 takes the results from Chapter 5 and extends them to the Universal Kriging setting, where the mean function of the Gaussian process is unknown. Comparable simulations are run in this setting, which is much wider than the Simple Kriging one.

Chapter 7 uses the Gibbs reference posterior distribution to tackle the industrial problem of EDF. It provides a heuristic for integrating covariance parameters out of the model thanks to the Gibbs reference posterior distribution. After a transformation family for the observation data has been defined, it makes estimating the transformation parameter easier, since all others have been integrated out. Finally, it shows how covariance parameter uncertainty can be embedded in POD curves.

Part I

Tools

Chapter 1

Kriging Overview

Abstract

The Kriging or Gaussian process regression model is introduced and its basic properties highlighted. The chapter covers Gaussian random processes, stationarity, mean square continuity and differentiability. Properties of Gaussian processes are linked to properties of their covariance kernels. Spectral analysis of Gaussian processes, beginning with Bochner's theorem, is introduced. Important examples of covariance kernels are provided, along with their spectral representations.

Résumé

Le modèle de krigeage ou régression par processus gaussiens est introduit et ses propriétés fondamentales mises en évidence. Le chapitre couvre les processus aléatoires gaussiens et les notions de stationarité, continuité en moyenne quadratique et différentiabilité en moyenne quadratique. Les propriétés des processus gaussiens sont reliées à celles de leurs noyaux de covariance. Le chapitre introduit aussi l'analyse spectrale des processus gaussiens, à commencer par le théorème de Bochner. Des exemples importants de noyaux de covariance ainsi que leur représentation spectrale sont fournis.

1.1 Introduction

This chapter provides a short introduction to Kriging from a theoretical rather than practical perspective. Its goal is to introduce tools necessary to understand the notions used in this thesis. It draws heavily on Stein [1999], Rasmussen and Williams [2006] and Bachoc [2013a]. The only prerequisites are standard probability and statistical theory.

Gaussian Stochastic Processes offer a convenient way of expressing the uncertainty about the value of some real-valued quantity on a given spatial domain $\mathcal{D}$ [Stein, 1999] when said quantity is only observed on a finite set of points in $\mathcal{D}$. This is why Gaussian Process Regression is used as a supervised learning method [Rasmussen and Williams, 2006, chapter 2], although it originally appeared in the geostatistical literature [Matheron, 1960]. Throughout the dissertation, we follow the geostatistical naming convention for this model: Kriging.

1.2 Gaussian random processes

General random processes

Most random processes considered in this thesis are Gaussian. It is however useful to keep in mind the more general framework for random processes. Detailed information about this topic can be found in Billingsley [1995], Dudley [1989] or Bass [1995].

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{D}$ be a spatial or temporal domain – typically $\mathcal{D} \subset \mathbb{R}^r$ for some positive integer r. Let $\mathcal{S}$ be a metric space – typically $\mathcal{S} = \mathbb{R}$ – and let $\mathcal{B}(\mathcal{S})$ be its Borel σ-algebra.

Definition 1.1. *A random process or random field is a mapping from the domain $\mathcal{D}$ to the set of all random variables $(\Omega, \mathcal{F}) \to (\mathcal{S}, \mathcal{B}(\mathcal{S}))$.*

No assumption is made as to whether the domain $\mathcal{D}$ or the metric space $\mathcal{S}$ is discrete or continuous. Markov chains are random processes with discrete domain $\mathcal{D}$ and discrete or continuous metric space $\mathcal{S}$. Poisson processes have continuous domain $\mathcal{D}$ and discrete metric space $\mathcal{S}$. Wiener processes (Brownian motion) have continuous domain $\mathcal{D}$ and continuous metric space $\mathcal{S}$.

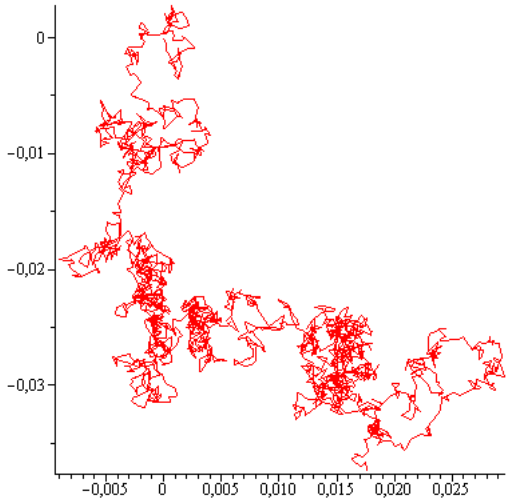


Figure 1.1 – *A trajectory of a 2-dimensional Brownian motion. Approximation through a 3000-step Markov chain where each step follows a Gaussian bi-variate distribution. Credit goes to Wikipedia contribution "Ipipipourax" who put this work in the public domain.*

An important property of Gaussian processes is that they can themselves be viewed as random variables. To see this, a few additional notions need to be defined.

Definition 1.2. *For all $\boldsymbol{x} \in \mathcal{D}$, define the mapping $\pi_{\boldsymbol{x}} : \mathbb{R}^{\mathcal{D}} \to \mathbb{R}; f \mapsto f(\boldsymbol{x})$. Denote by $\mathcal{B}(\mathbb{R})^{\otimes \mathcal{D}}$ the σ-algebra of $\mathbb{R}^{\mathcal{D}}$ spanned by $\{\pi_{\boldsymbol{x}}^{-1}(B) : \boldsymbol{x} \in \mathcal{D}, B \in \mathcal{B}(\mathbb{R})\}$.*

Although a random process is defined as a collection of random variables, the following proposition shows that it can also be represented as a random variable with functional value [Billingsley, 1995, chapter 7].

Proposition 1.3. *Let Y be a random process. The mapping $\tilde{Y} : \Omega \to \mathbb{R}^{\mathcal{D}}; \omega \mapsto [\boldsymbol{x} \mapsto Y(\boldsymbol{x})(\omega)]$ is $(\mathcal{F}, \mathcal{B}(\mathbb{R})^{\otimes \mathcal{D}})$-measurable.*

Proof. Let $B \in \mathcal{B}(\mathbb{R})$. Then for all $\boldsymbol{x} \in \mathcal{D}$, $\tilde{Y}^{-1}(\pi_{\boldsymbol{x}}^{-1}(B)) = (\pi_{\boldsymbol{x}} \circ \tilde{Y})^{-1}(B) = Y(\boldsymbol{x})^{-1}(B) \in \mathcal{F}$. □

The probability distribution of a random process is characterized by its finite-dimensional distributions [Khoshnevisan, 2002, chapter 3]:

Proposition 1.4. *Let Y_1 and Y_2 be two random processes and let $\tilde{Y}_1$ and $\tilde{Y}_2$ be the mappings $\Omega \to \mathbb{R}^{\mathcal{D}}$ defined by $\tilde{Y}_1(\omega)(\boldsymbol{x}) = Y_1(\boldsymbol{x})(\omega)$ and $\tilde{Y}_2(\omega)(\boldsymbol{x}) = Y_2(\boldsymbol{x})(\omega)$. If, for every positive integer n and every family $(\boldsymbol{x}^{(i)})_{i \in [\![1,n]\!]}$ of points in $\mathcal{D}$, the Gaussian vectors $(Y_1(\boldsymbol{x}^{(1)}), ..., Y_1(\boldsymbol{x}^{(n)}))^\top$ and $(Y_2(\boldsymbol{x}^{(1)}), ..., Y_2(\boldsymbol{x}^{(n)}))^\top$ have the same probability distribution, then $\tilde{Y}_1$ and $\tilde{Y}_2$ have the same probability distribution.*

Proof. For every positive integer n, for every family $(\boldsymbol{x}^{(i)})_{i \in [\![1,n]\!]}$ of elements of $\mathcal{D}$ and every family $(B_i)_{i \in [\![1,n]\!]}$ of elements of $\mathcal{B}(\mathbb{R})$,

$$\begin{aligned}\tilde{Y}_1^{-1}\left(\pi_{\boldsymbol{x}^{(1)}}^{-1}(B_1) \cap ... \cap \pi_{\boldsymbol{x}^{(n)}}^{-1}(B_n)\right) &= \tilde{Y}_1^{-1}\left(\pi_{\boldsymbol{x}^{(1)}}^{-1}(B_1)\right) \cap ... \cap \tilde{Y}_1^{-1}\left(\pi_{\boldsymbol{x}^{(n)}}^{-1}(B_n)\right) \\ &= (\pi_{\boldsymbol{x}^{(1)}} \circ \tilde{Y}_1)^{-1}(B_1) \cap ... \cap (\pi_{\boldsymbol{x}^{(n)}} \circ \tilde{Y}_1)^{-1}(B_n) \\ &= Y_1(\boldsymbol{x}^{(1)})^{-1}(B_1) \cap ... \cap Y_1(\boldsymbol{x}^{(n)})^{-1}(B_n). \end{aligned} \tag{1.1}$$

Similarly,

$$\tilde{Y}_2^{-1}\left(\pi_{\boldsymbol{x}^{(1)}}^{-1}(B_1) \cap ... \cap \pi_{\boldsymbol{x}^{(n)}}^{-1}(B_n)\right) = Y_2(\boldsymbol{x}^{(1)})^{-1}(B_1) \cap ... \cap Y_2(\boldsymbol{x}^{(n)})^{-1}(B_n). \tag{1.2}$$

Using the assumption,

$$\mathbb{P}\left(Y_1(\boldsymbol{x}^{(1)})^{-1}(B_1) \cap ... \cap Y_1(\boldsymbol{x}^{(n)})^{-1}(B_n)\right) = \mathbb{P}\left(Y_2(\boldsymbol{x}^{(1)})^{-1}(B_1) \cap ... \cap Y_2(\boldsymbol{x}^{(n)})^{-1}(B_n)\right). \tag{1.3}$$

Combining these equations, we obtain

$$\mathbb{P}\left(\tilde{Y}_1^{-1}\left(\pi_{\boldsymbol{x}^{(1)}}^{-1}(B_1) \cap ... \cap \pi_{\boldsymbol{x}^{(n)}}^{-1}(B_n)\right)\right) = \mathbb{P}\left(\tilde{Y}_2^{-1}\left(\pi_{\boldsymbol{x}^{(1)}}^{-1}(B_1) \cap ... \cap \pi_{\boldsymbol{x}^{(n)}}^{-1}(B_n)\right)\right). \tag{1.4}$$

Consider the following set:

$$\left\{\pi_{\boldsymbol{x}^{(1)}}^{-1}(B_1) \cap ... \cap \pi_{\boldsymbol{x}^{(n)}}^{-1}(B_n) : n \in \mathbb{Z}_+, (\boldsymbol{x}^{(i)})_{i \in [\![1,n]\!]} \in \mathcal{D}^n, (B_i)_{i \in [\![1,n]\!]} \in \mathcal{B}(\mathbb{R})^n\right\}.$$

It is a π-system which includes $\{\pi_{\boldsymbol{x}}^{-1}(B) : \boldsymbol{x} \in \mathcal{D}, B \in \mathcal{B}(\mathbb{R})\}$, the set spanning $\mathcal{B}(\mathbb{R})^{\otimes \mathcal{D}}$. The distributions of $\tilde{Y}_1$ and $\tilde{Y}_2$ agree on this π-system and can therefore only be equal.

□

From this point onwards, for any random process Y, we will abuse notations and also denote by Y the mapping $\tilde{Y}$ defined in Proposition 1.3.

Gaussian random processes

The theory of Gaussian vectors, i.e. of the multivariate Normal distribution, is a prerequisite of the following. Please refer to Appendix B.1 of Santner et al. [2003] or to Appendix A.2 of Rasmussen and Williams [2006].
Gaussian random processes or Gaussian processes for short are a particular type of random processes. The most prominent members of this class are Wiener processes. For details, see for example Khoshnevisan [2002, chapter 5].

Definition 1.5. *A Gaussian process is a random process Y with metric space $\mathcal{S} = \mathbb{R}$ such that for every positive integer n and every family $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$, $(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))^\top$ is a Gaussian vector.*

The class of Gaussian processes is, among all classes of random processes, the one that has been most studied. Many nice properties are available, and we only use some of them. For a larger view, please refer to Adler [1990] or Stein [1999].

Conditional Gaussian processes

Recall the following result from Gaussian vector theory:

Theorem 1.6. *Let $(\boldsymbol{V}_1^\top, \boldsymbol{V}_2^\top)^\top$ be a Gaussian vector. Assume that $\boldsymbol{V}_1$ is nondegenerate. Then its covariance matrix $\mathrm{Var}[\boldsymbol{V}_1]$ is nonsingular and the distribution of $\boldsymbol{V}_2$ conditionnally to $\boldsymbol{V}_1$ is Gaussian with mean vector*

$$\mathbb{E}[\boldsymbol{V}_2|\boldsymbol{V}_1] = \mathbb{E}[\boldsymbol{V}_2] - \mathrm{Cov}[\boldsymbol{V}_2, \boldsymbol{V}_1]\,\mathrm{Var}[\boldsymbol{V}_1]^{-1}(\boldsymbol{V}_1 - \mathbb{E}[\boldsymbol{V}_1]) \tag{1.5}$$

and covariance matrix

$$\mathrm{Var}[\boldsymbol{V}_2|\boldsymbol{V}_1] = \mathrm{Cov}[\boldsymbol{V}_2, \boldsymbol{V}_1]\,\mathrm{Var}[\boldsymbol{V}_1]^{-1}\,\mathrm{Cov}[\boldsymbol{V}_1, \boldsymbol{V}_2]. \tag{1.6}$$

This theorem is of great practical importance in the study of Gaussian processes. It can be generalized as follows:

Corollary 1.7. *Let $(\boldsymbol{V}_1^\top, \boldsymbol{V}_2^\top)^\top$ be a Gaussian vector. If n is the size of $\boldsymbol{V}$ and w is the rank of $\mathrm{Var}[\boldsymbol{V}_1]$, let $\boldsymbol{W}$ be an $n \times w$ matrix representing an isometry from the subspace of $\mathbb{R}^n$ spanned by $\mathrm{Var}[\boldsymbol{V}_1]$ to $\mathbb{R}^w$. Then the distribution of $\boldsymbol{V}_2$ conditionnally to $\boldsymbol{V}_1$ is Gaussian with mean vector*

$$\mathbb{E}[\boldsymbol{V}_2|\boldsymbol{V}_1] = \mathbb{E}[\boldsymbol{V}_2] - \mathrm{Cov}[\boldsymbol{V}_2, \boldsymbol{V}_1]\boldsymbol{W}\left(\boldsymbol{W}^\top \mathrm{Var}[\boldsymbol{V}_1]\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top(\boldsymbol{V}_1 - \mathbb{E}[\boldsymbol{V}_1]) \tag{1.7}$$

and covariance matrix

$$\mathrm{Var}[\boldsymbol{V}_2|\boldsymbol{V}_1] = \mathrm{Cov}[\boldsymbol{V}_2, \boldsymbol{V}_1]\boldsymbol{W}\left(\boldsymbol{W}^\top \mathrm{Var}[\boldsymbol{V}_1]\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\,\mathrm{Cov}[\boldsymbol{V}_1, \boldsymbol{V}_2]. \tag{1.8}$$

Proof. The $n \times n$ matrix $\boldsymbol{W}\boldsymbol{W}^\top$ represents an orthogonal projection on the subspace of $\mathbb{R}^n$ spanned by $\mathrm{Var}[\boldsymbol{V}_1]$. So $\boldsymbol{V}_1 = \boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{V}_1$, which implies that the distribution of $\boldsymbol{V}_2$ conditionnally to $\boldsymbol{V}_1$ is the distribution of $\boldsymbol{V}_2$ conditionnally to $\boldsymbol{W}^\top\boldsymbol{V}_1$.

Since $(\boldsymbol{V}_1^\top, \boldsymbol{V}_2^\top)^\top$ is a Gaussian vector, any linear transformation remains a Gaussian vector. In particular, $((\boldsymbol{W}\boldsymbol{V}_1)^\top), \boldsymbol{V}_2^\top)^\top$ is a Gaussian vector. Moreover, $\boldsymbol{W}^\top \boldsymbol{V}_1$ is nondegenerate so Theorem 1.6 is applicable.

The distribution of $\boldsymbol{V}_2$ conditionnally to $\boldsymbol{V}_1$ (or equivalently of $\boldsymbol{W}^\top \boldsymbol{V}_1$) is Gaussian with mean vector

$$\begin{aligned}\mathbb{E}[\boldsymbol{V}_2|\boldsymbol{V}_1] &= \mathbb{E}[\boldsymbol{V}_2] - \operatorname{Cov}[\boldsymbol{V}_2, \boldsymbol{W}\boldsymbol{V}_1] \operatorname{Var}[\boldsymbol{W}^\top \boldsymbol{V_1} \boldsymbol{W}]^{-1} \operatorname{Cov}[\boldsymbol{W}\boldsymbol{V}_1, \boldsymbol{V}_2] \\ &= \mathbb{E}[\boldsymbol{V}_2] - \operatorname{Cov}[\boldsymbol{V}_2, \boldsymbol{V}_1]\boldsymbol{W} \left(\boldsymbol{W}^\top \operatorname{Var}[\boldsymbol{V}_1]\boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top (\boldsymbol{V}_1 - \mathbb{E}[\boldsymbol{V}_1]) \qquad (1.9)\end{aligned}$$

and with covariance matrix

$$\begin{aligned}\operatorname{Var}[\boldsymbol{V}_2|\boldsymbol{V}_1] &= \operatorname{Cov}[\boldsymbol{V}_2, \boldsymbol{W}\boldsymbol{V}_1] \operatorname{Var}[\boldsymbol{W}\boldsymbol{V}_1]^{-1} \boldsymbol{W}^\top \operatorname{Cov}[\boldsymbol{W}\boldsymbol{V}_1, \boldsymbol{V}_2] \\ &= \operatorname{Cov}[\boldsymbol{V}_2, \boldsymbol{V}_1]\boldsymbol{W} \left(\boldsymbol{W}^\top \operatorname{Var}[\boldsymbol{V}_1]\boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \operatorname{Cov}[\boldsymbol{V}_1, \boldsymbol{V}_2]. \qquad (1.10)\end{aligned}$$

□

With this result, we can consider conditional Gaussian processes [Stein, 1999, chapter 1].

Definition 1.8. *Let Y be a Gaussian process. Let n be a positive integer and $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$ be a family of n points in $\mathcal{D}$. A Gaussian process Y_{cond} such that for every positive integer n' and every family $(\boldsymbol{x}^{(i)\prime})_{i\in[\![1,n']\!]}$ of points in $\mathcal{D}$, the Gaussian vector $(Y_{cond}(\boldsymbol{x}^{(1)\prime}), ..., Y_{cond}(\boldsymbol{x}^{(n')\prime}))^\top$ has the same distribution as the Gaussian vector $(Y(\boldsymbol{x}^{(1)\prime}), ..., Y(\boldsymbol{x}^{(n')\prime}))^\top$ conditionally to $Y(\boldsymbol{x}^{(1)})$,...,$Y(\boldsymbol{x}^{(n)})$ is called a version of the conditional process Y knowing $Y(\boldsymbol{x}^{(1)})$, ... , $Y(\boldsymbol{x}^{(n)})$.*

It is easy to see that all versions of the conditional process have the same distribution.
This notion of conditional Gaussian process is at the heart of the Kriging procedure.

Definition 1.9. *Let f be an unknown mapping $\mathcal{D} \to \mathbb{R}$. Assume that for some positive integer n, there exists a family of n points in $\mathcal{D}$ $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$ on which f was observed, so $f(\boldsymbol{x}^{(1)}), ..., f(\boldsymbol{x}^{(n)})$ are known. A Kriging model is a version of the conditional Gaussian process Y knowing $Y(\boldsymbol{x}^{(1)}) = f(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}) = f(\boldsymbol{x}^{(n)})$.*

Since all versions of the conditional Gaussian process have the same distribution, distinguishing them makes no sense from a statistical perspective. In the following, we abuse denominations by referring to any version of the conditional process as *the* conditional process.

In this construction, the distribution of the conditional Gaussian process Y is supposed to represent the uncertainty on f. It does *not* mean that f is assumed to be a realization of a Gaussian process, which would be a meaningless statement.

Characterizing a Gaussian process

Just like mean vector and covariance matrix characterize the distribution of a Gaussian vector, mean function and covariance function (or kernel) characterize a Gaussian process [Bachoc, 2013a, chapter 2].

Definition 1.10. *Let Y be a Gaussian process. The mapping $m : \mathcal{D} \to \mathbb{R}$ defined by $\boldsymbol{x} \mapsto \mathbb{E}[Y(\boldsymbol{x})]$ is called the mean function of Y.*

Definition 1.11. *Let Y be a Gaussian process. The mapping $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ defined by $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \mapsto \mathrm{Cov}[Y(\boldsymbol{x}^{(1)}), Y(\boldsymbol{x}^{(2)})]$ is called the covariance function or covariance kernel of Y.*

Proposition 1.12. *Let Y_1 and Y_2 be Gaussian processes with the same mean function and covariance kernel. Then they have the same distribution.*

Proof. For any positive integer n and any family $(\boldsymbol{x}^{(i)})_{i \in [\![1,n]\!]}$ of n points in $\mathcal{D}$, the random vectors $(Y_1(\boldsymbol{x}^{(1)}), ..., Y_1(\boldsymbol{x}^{(n)}))^\top$ and $(Y_2(\boldsymbol{x}^{(1)}), ..., Y_2(\boldsymbol{x}^{(n)}))^\top$ are Gaussian vectors with the same mean vector and covariance matrix, so their distributions are equal. The result then follows from Proposition 1.4. □

If mean function and covariance kernel are enough to characterize the distribution of a Gaussian process, it raises the question of which functions are admissible as mean function and covariance kernel. That is, for which mean function and covariance kernel candidates is it possible to construct a Gaussian process that admits them as mean function and covariance kernel respectively?

Provided a Gaussian process with null mean function can be constructed, a Gaussian process with any mean function can be obtained simply by adding said function. So there is no condition on mean functions.

The picture is different for covariance kernels.

Definition 1.13. *A mapping $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is called positive definite if for every positive integer n and any family $(\boldsymbol{x}^{(i)})_{i \in [\![1,n]\!]}$ of n points in $\mathcal{D}$, the $n \times n$ matrix with (i, j)-th element $K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$ is positive semi-definite.*

A positive definite mapping is sometimes called a kernel, hence the name "covariance kernel" for "covariance function". See the study of the positive-definiteness of bivariate mappings conducted in Schölkopf and Smola [2012][chapter 13].

Proposition 1.14. *For any mapping $m : \mathcal{D} \to \mathbb{R}$ and any mapping $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$, there exists a Gaussian process Y with mean function m and covariance kernel K if and only if K is positive definite.*

In the context of the proposition above, the phrase "there exists a Gaussian process" means that there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a corresponding random process $Y : \mathcal{D} \mapsto \mathbb{R}$ which is a Gaussian process. In all preceding and folowing results, the existence of $(\Omega, \mathcal{F}, \mathbb{P})$ is taken for granted. This is the only one in which this is not the case. Obviously, for K to be the covariance kernel of some Gaussian process, it needs to be positive definite, so the condition given in the proposition is necessary. It is its sufficiency – i.e. the existence of $(\Omega, \mathcal{F}, \mathbb{P})$ and Y – that is difficult to show. It requires Kolmogorov's existence theorem, of which Billingsley [1995, chapter 7] provides two different proofs.

Stationarity

General Gaussian processes can be hard to handle. Practically speaking, most Kriging models used in the literature consider a particular class of Gaussian processes: stationary Gaussian processes [Rasmussen and Williams, 2006, chapter 4]. To properly define them, we need to assume that $\mathcal{D}$ is a subset of a real affine space. Let $\mathcal{V}$ be the corresponding real vector space.

Definition 1.15. *A Gaussian process Y is said to be stationary if for any positive integer n, any family $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$ of n points in $\mathcal{D}$ and any vector $\boldsymbol{h} \in \mathcal{V}$, the following holds: if for every integer $i \in [\![1,n]\!]$, $\boldsymbol{x}^{(i)} + \boldsymbol{h} \in \mathcal{D}$, then the Gaussian vectors $(Y(\boldsymbol{x}^{(1)} + \boldsymbol{h}), ..., Y(\boldsymbol{x}^{(n)} + \boldsymbol{h}))^\top$ and $(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))^\top$ share the same distribution.*

Stationarity can be characterized in terms of mean function and covariance kernel [Rasmussen and Williams, 2006, chapter 4].

Definition 1.16. *A positive definite mapping $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is said to be stationary if there exists a mapping $\tilde{K} : \mathcal{V} \to \mathbb{R}$ such that for every $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathcal{D}^2$, $K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \tilde{K}(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})$.*

Proposition 1.17. *A Gaussian process is stationary if and only if its mean function is constant and its covariance kernel stationary.*

Proof. Let Y be a Gaussian process with mean function m and covariance kernel K.

Assume that m is constant and that K is stationary. Then there exists a mapping $\tilde{K} : \mathcal{V} \to \mathbb{R}$ such that for every $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathcal{D}^2$, $K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) = \tilde{K}(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})$.
Let n be a positive integer and let $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$ be a family of n points in $\mathcal{D}$. Let $\boldsymbol{h}$ be a vector of $\mathcal{V}$ such that for every integer $i \in [\![1,n]\!]$, $\boldsymbol{x}^{(i)} + \boldsymbol{h} \in \mathcal{D}$. Then the Gaussian vector $(Y(\boldsymbol{x}^{(1)} + \boldsymbol{h}), ..., Y(\boldsymbol{x}^{(n)} + \boldsymbol{h}))^\top$ has constant mean vector of value $m(\boldsymbol{x}^{(1)})$ and its covariance matrix has (i,j)-th element $K(\boldsymbol{x}^{(i)} + \boldsymbol{h}, \boldsymbol{x}^{(j)} + \boldsymbol{h}) = \tilde{K}(\boldsymbol{x}^{(i)} + \boldsymbol{h} - (\boldsymbol{x}^{(j)} + \boldsymbol{h})) = \tilde{K}(\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}) = K(\boldsymbol{x}^{(i)}, \boldsymbol{x}^{(j)})$. So its mean vector and covariance matrix are respectively equal to the mean vector and covariance matrix of the Gaussian vector $(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)}))^\top$. Therefore Y is stationary.

Now assume that Y is stationary. Let $\boldsymbol{x}^{(1)} \in \mathcal{D}$. Then, for all $\boldsymbol{x}^{(2)} \in \mathcal{D}$, $Y(\boldsymbol{x}^{(2)}) = Y(\boldsymbol{x}^{(1)} + (\boldsymbol{x}^{(2)} - \boldsymbol{x}^{(1)}))$ has the same distribution as $Y(\boldsymbol{x}^{(1)})$. In particular, it has the same mean. Therefore the mean function m is constant.
Let $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathcal{D}^2$ and $(\boldsymbol{x}^{(1)\prime}, \boldsymbol{x}^{(2)\prime}) \in \mathcal{D}^2$ such that $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} = \boldsymbol{x}^{(1)\prime} - \boldsymbol{x}^{(2)\prime}$. Then the Gaussian vector $(Y(\boldsymbol{x}^{(1)\prime}), Y(\boldsymbol{x}^{(2)\prime}))^\top = (Y(\boldsymbol{x}^{(1)} + (\boldsymbol{x}^{(1)\prime} - \boldsymbol{x}^{(1)})), Y(\boldsymbol{x}^{(2)} + (\boldsymbol{x}^{(2)\prime} - \boldsymbol{x}^{(2)}))^\top = (Y(\boldsymbol{x}^{(1)} + (\boldsymbol{x}^{(1)\prime} - \boldsymbol{x}^{(1)})), Y(\boldsymbol{x}^{(2)} + (\boldsymbol{x}^{(1)\prime} - \boldsymbol{x}^{(1)}))^\top$ has the same distribution as the Gaussian vector $(Y(\boldsymbol{x}^{(1)}), Y(\boldsymbol{x}^{(2)}))^\top$. In particular, the covariance $K(\boldsymbol{x}^{(1)\prime}, \boldsymbol{x}^{(2)\prime})$ between $Y(\boldsymbol{x}^{(1)\prime})$ and $Y(\boldsymbol{x}^{(2)\prime})$ is equal to the covariance $K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$ between $Y(\boldsymbol{x}^{(1)})$ and $Y(\boldsymbol{x}^{(2)})$. So the mapping $\tilde{K}$ can be defined as follows: for every $v \in \mathcal{V}$, if there exists a pair $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathcal{D}^2$ such that $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} = v$, then $\tilde{K}(v) := K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$. Else $\tilde{K}(v) := 0$.

□

From this point onwards, we abuse denominations and call $\tilde{K} : \mathcal{V} \to \mathbb{R}$ the covariance kernel instead of $K : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$. This is a twofold abuse because $\tilde{K}$ is not necessarily unique, as the proof above shows. It is of no matter, because $\tilde{K}$ is uniquely defined at all points that can be written $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}$ with $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathcal{D}^2$ and thus K can be computed from $\tilde{K}$.

Since positive definite mappings are covariance kernels, this abuse leads to an extension of the definition of positive definiteness to mappings $\mathbb{R}^r \to \mathbb{R}$.

Definition 1.18. *A mapping $f : \mathbb{R}^r \to \mathbb{R}$ is called positive definite if there exists a positive definite mapping $K : \mathbb{R}^r \times \mathbb{R}^r \to \mathbb{R}$ such that for all $(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)}) \in \mathbb{R}^r \times \mathbb{R}^r$, $f(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}) = K(\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)})$.*

1.3 Mean square continuity and differentiability of Gaussian processes

The L^2 norm allows a natural definition of continuity and differentiability for Gaussian processes that is linked to the continuity and stationarity of its covariance kernel.
As before, we assume that $\mathcal{D}$ is a subset of a real affine space with corresponding vector space $\mathcal{V}$. Let us further assume that $\mathcal{V} = \mathbb{R}^r$ for a given $r \in \mathbb{N}$. We endow the affine space with the metric induced by the Euclidean norm on $\mathbb{R}^r$ and assume that $\mathcal{D}$ is an open subset.

Mean square continuity and differentiability

Let us consider mean square properties of Gaussian processes [Stein, 1999, chapter 2].

Definition 1.19. *A Gaussian process Y is mean square continuous on $\mathcal{D}$ if for any $\boldsymbol{x}^{(0)} \in \mathcal{D}$ and for all $\epsilon > 0$, there exists $\delta > 0$ such that for every $\boldsymbol{x} \in \mathcal{D}$ such that $\|\boldsymbol{x} - \boldsymbol{x}^{(0)}\| \leqslant \delta$,*

$$\mathbb{E}\left[\left(Y(\boldsymbol{x}) - Y(\boldsymbol{x}^{(0)})\right)^2\right] \leqslant \epsilon^2 \tag{1.11}$$

Remark. For a stationary process, being mean square continuous is equivalent to being mean square uniformly continuous. This means that in the above definition, δ does not depend on $\boldsymbol{x}^{(0)}$.

Definition 1.20. *A Gaussian process Y is mean square continuously differentiable on $\mathcal{D}$ if there exist r mean square continuous Gaussian processes $\partial_{\boldsymbol{e}_1} Y, ..., \partial_{\boldsymbol{e}_r} Y$ such that, defining $(\boldsymbol{e}_k)_{k \in [\![1,r]\!]}$ as the canonical basis of $\mathbb{R}^r$, for any integer $k \in [\![1, r]\!]$, the following property holds. For any $\boldsymbol{x}^{(0)} \in \mathcal{D}$ and for all $\epsilon > 0$, there exists $\delta > 0$ such that for all $0 < h \leqslant \delta$,*

$$\mathbb{E}\left[\left(\frac{Y(\boldsymbol{x}^{(0)} + h\boldsymbol{e}_k) - Y(\boldsymbol{x}^{(0)})}{h} - \partial_{\boldsymbol{e}_k} Y(\boldsymbol{x}^{(0)})\right)^2\right] \leqslant \epsilon^2 \tag{1.12}$$

In the definition above, for every $k \in [\![1, r]\!]$, the "partial derivative" $\partial_{\boldsymbol{e}_k} Y$ is unique in the sense that any other Gaussian process fitting the requirement would necessarily be a modification of $\partial_{\boldsymbol{e}_k} Y$.

Remark. If Y is a stationary Gaussian process that is continuously differentiable, then all $\partial \boldsymbol{e}_k Y$ are necessarily stationary Gaussian processes.

The following result provides an alternate view of continuous mean square differentiability [Bachoc, 2013a, chapter 2].

Proposition 1.21. *A stationary Gaussian process Y is mean square continuously differentiable on $\mathcal{D}$ if, and only if, $Y : \mathcal{D} \to L^2(\Omega)$ is continuously Fréchet differentiable on $\mathcal{D}$.*

Proof. If a stationary Gaussian process Y is Fréchet continuously differentiable on $\mathcal{D}$, then for all $\boldsymbol{x}^{(0)} \in \mathcal{D}$ and every integer $k \in [\![1, r]\!]$, consider the value taken by the Fréchet derivative $DY(\boldsymbol{x}^{(0)})$ at the vector $\boldsymbol{e}_k$.
Equation (1.12) holds with $\partial_{\boldsymbol{e}_k} Y(\boldsymbol{x}^{(0)}) := DY(\boldsymbol{x}^{(0)})(\boldsymbol{e}_k)$ by definition of $DY(\boldsymbol{x}^{(0)})(\boldsymbol{e}_k)$. Moreover, the Fréchet derivative $DY(\boldsymbol{x}^{(0)})(\boldsymbol{e}_k)$ is continuous. To prove that Y is mean square continuously differentiable, we only need to show that the mapping $\boldsymbol{x}^{(0)} \in \mathcal{D} \mapsto DY(\boldsymbol{x}^{(0)})(\boldsymbol{e}_k)$ is a Gaussian process. This follows from the fact that for any positive integer n and any points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$, $(DY(\boldsymbol{x}^{(1)})(\boldsymbol{e}_k), ..., DY(\boldsymbol{x}^{(n)})(\boldsymbol{e}_k))^\top$ is an L^2 limit of Gaussian vectors and is therefore a Gaussian vector.

Conversely, if Y is a stationary Gaussian process that is mean square differentiable, define for every point $\boldsymbol{x}^{(0)} \in \mathcal{D}$ the linear application $DY(\boldsymbol{x}^0) : \mathbb{R}^r \to L^2(\Omega)$ such that for every $k \in [\![1, r]\!]$, $DY(\boldsymbol{x}^0))(\boldsymbol{e}_k) = \partial_{\boldsymbol{e}_k} Y(\boldsymbol{x}^{(0)})$.

For every $\boldsymbol{x}^{(0)} \in \mathcal{D}$ and every $\boldsymbol{u} \in \mathbb{R}^r$ and $\boldsymbol{v} \in \mathbb{R}^r$ such that $\boldsymbol{x}^{(0)} + \boldsymbol{u} \in \mathcal{D}$ and $\boldsymbol{x}^{(0)} + \boldsymbol{u} + \boldsymbol{v} \in \mathcal{D}$, because $DY(\boldsymbol{x}^{(0)})$ is a linear mapping, we have

$$
\begin{aligned}
&Y(\boldsymbol{x}^{(0)} + \boldsymbol{u} + \boldsymbol{v}) - Y(\boldsymbol{x}^{(0)}) - DY(\boldsymbol{x}^{(0)})(\boldsymbol{u} + \boldsymbol{v}) && (1.13)\\
&= Y(\boldsymbol{x}^{(0)} + \boldsymbol{u} + \boldsymbol{v}) - Y(\boldsymbol{x}^{(0)} + \boldsymbol{u}) - DY(\boldsymbol{x}^{(0)} + \boldsymbol{u})(\boldsymbol{v})\\
&\quad + Y(\boldsymbol{x}^{(0)} + \boldsymbol{u}) - Y(\boldsymbol{x}^{(0)}) - DY(\boldsymbol{x}^{(0)})(\boldsymbol{u})\\
&\quad + DY(\boldsymbol{x}^{(0)} + \boldsymbol{u})(\boldsymbol{v}) - DY(\boldsymbol{x}^{(0)})(\boldsymbol{v}). && (1.14)
\end{aligned}
$$

Because D is open, for every $\boldsymbol{x}^{(0)} \in \mathcal{D}$, there exists $t > 0$ such that for every $\boldsymbol{h} \in \mathbb{R}^r$ such that $\|\boldsymbol{h}\| < t$, $\boldsymbol{x}^{(0)} + \boldsymbol{h} \in \mathcal{D}$. Let $h_1\boldsymbol{e}_1 + ... + h_r\boldsymbol{e}_r$ $(h_1, ..., h_r \in \mathbb{R})$ be the unique decomposition of $\boldsymbol{h}$ in the canonical basis of $\mathbb{R}^r$. Applying the result above, we have

$$
\begin{aligned}
&Y(\boldsymbol{x}^{(0)} + \boldsymbol{h}) - Y(\boldsymbol{x}^{(0)}) - DY(\boldsymbol{x}^{(0)})(\boldsymbol{u})\\
=&Y\left(\boldsymbol{x}^{(0)} + \sum_{k=1}^{r} h_k\boldsymbol{e}_k\right) - Y(\boldsymbol{x}^{(0)}) - DY(\boldsymbol{x}^{(0)})\left(\sum_{k=1}^{r} h_k\boldsymbol{e}_k\right)\\
=&\sum_{k=1}^{r} Y\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k} h_l\boldsymbol{e}_l\right) - Y\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k-1} h_l\boldsymbol{e}_l\right) - DY\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k-1} h_l\boldsymbol{e}_l\right)(h_k\boldsymbol{e}_k)\\
&+ \sum_{k=1}^{r} DY\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k-1} h_l\boldsymbol{e}_l\right)(h_k\boldsymbol{e}_k) - DY(\boldsymbol{x}^{(0)})(h_k\boldsymbol{e}_k)\\
=&\sum_{k=1}^{r} Y\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k} h_l\boldsymbol{e}_l\right) - Y\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k-1} h_l\boldsymbol{e}_l\right) - h_k\partial_{\boldsymbol{e}_k} Y\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k-1} h_l\boldsymbol{e}_l\right)\\
&+ \sum_{k=1}^{r} h_k\partial_{\boldsymbol{e}_k} Y\left(\boldsymbol{x}^{(0)} + \sum_{l=1}^{k-1} h_l\boldsymbol{e}_l\right) - h_k\partial_{\boldsymbol{e}_k} Y(\boldsymbol{x}^{(0)}) \qquad (1.15)
\end{aligned}
$$

In the following, $\|\cdot\|$ denotes the Euclidean norm if applied to a vector of $\mathbb{R}^r$ and the L^2 norm if applied to a random variable.

For every $k \in [\![1, r]\!]$, due to stationarity,

$$\begin{aligned}&\left\|Y\left(\boldsymbol{x}^{(0)}+\sum_{l=1}^{k}h_l\boldsymbol{e}_l\right)-Y\left(\boldsymbol{x}^{(0)}+\sum_{l=1}^{k-1}h_l\boldsymbol{e}_l\right)-h_k\partial_{\boldsymbol{e}_k}Y\left(\boldsymbol{x}^{(0)}+\sum_{l=1}^{k-1}h_l\boldsymbol{e}_l\right)\right\|\\ =&\left\|Y\left(\boldsymbol{x}^{(0)}+h_k\boldsymbol{e}_k\right)-Y\left(\boldsymbol{x}^{(0)}\right)-h_k\partial_{\boldsymbol{e}_k}Y\left(\boldsymbol{x}^{(0)}\right)\right\|. \qquad (1.16)\end{aligned}$$

So combining both equations above yields

$$\begin{aligned}&\frac{1}{\|\boldsymbol{h}\|}\left\|Y(\boldsymbol{x}^{(0)}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)})-DY(\boldsymbol{x}^{(0)})(\boldsymbol{u})\right\|\\ \leqslant&\sum_{k=1}^{r}\frac{|h_k|}{\|\boldsymbol{h}\|}\left\|\frac{Y\left(\boldsymbol{x}^{(0)}+h_k\boldsymbol{e}_k\right)-Y\left(\boldsymbol{x}^{(0)}\right)}{h_k}-\partial_{\boldsymbol{e}_k}Y\left(\boldsymbol{x}^{(0)}\right)\right\|\\ &+\sum_{k=1}^{r}\frac{|h_k|}{\|\boldsymbol{h}\|}\left\|\partial_{\boldsymbol{e}_k}Y\left(\boldsymbol{x}^{(0)}+\sum_{l=1}^{k-1}h_l\boldsymbol{e}_l\right)-\partial_{\boldsymbol{e}_k}Y(\boldsymbol{x}^{(0)})\right\|. \qquad (1.17)\end{aligned}$$

This leads to

$$\begin{aligned}&\frac{1}{\|\boldsymbol{h}\|}\left\|Y(\boldsymbol{x}^{(0)}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)})-DY(\boldsymbol{x}^{(0)})(\boldsymbol{u})\right\|\\ \leqslant&\sum_{k=1}^{r}\left\|\frac{Y\left(\boldsymbol{x}^{(0)}+h_k\boldsymbol{e}_k\right)-Y\left(\boldsymbol{x}^{(0)}\right)}{h_k}-\partial_{\boldsymbol{e}_k}Y\left(\boldsymbol{x}^{(0)}\right)\right\|\\ &+\sum_{k=1}^{r}\left\|\partial_{\boldsymbol{e}_k}Y\left(\boldsymbol{x}^{(0)}+\sum_{l=1}^{k-1}h_l\boldsymbol{e}_l\right)-\partial_{\boldsymbol{e}_k}Y(\boldsymbol{x}^{(0)})\right\|. \qquad (1.18)\end{aligned}$$

Because Y is continuously mean square differentiable, when $\|\boldsymbol{h}\| \to 0$ (and therefore every $h_k \to 0$), the right side of the inequality has null limit. The left side has therefore the same limit and Y is Fréchet differentiable at $\boldsymbol{x}^{(0)}$. Continuity of the Fréchet derivative follows once again from stationarity.

□

This Proposition roots the notion of mean square continuous differentiability for stationary Gaussian proceses in standard differential calculus. We may therefore write that a stationary Gaussian process is of class C^0 if it is mean square continuous and of class C^1 if it is mean square continuously differentiable. Moreover, the class C^n is well defined for every nonnegative integer n.

In particular, standard differential calculus theory yields the following characterization of the class C^n.

Proposition 1.22. *A stationary Gaussian process Y is of class C^n on $\mathcal{D}$ (i.e. n times mean square continuously differentiable) if, and only if, all partial derivatives $\partial_{\boldsymbol{e}_{k_1}}\partial_{\boldsymbol{e}_{k_2}}...\partial_{\boldsymbol{e}_{k_n}}Y$ $(k_1, k_2, ..., k_n \in [\![1, r]\!])$ exist and are mean square continuous on $\mathcal{D}$.*

Mean square continuity of stationary Gaussian processes can be characterized in terms of their covariance kernel, as the two following results show [Bachoc, 2013a, chapter 2].

Proposition 1.23. *A stationary Gaussian process Y is mean square continuous on $\mathcal{D}$ if, and only if, its covariance kernel $K : \mathbb{R}^r \to [0, +\infty)$ is continuous at $\mathbf{0}$.*

Proof. Without loss of generality, we may assume the mean function of Y to be null. Then

$$\begin{aligned}\mathbb{E}\left[\left(Y(\boldsymbol{x})-Y(\boldsymbol{x}^{(0)})\right)^2\right] &= \mathbb{E}\left[Y(\boldsymbol{x})^2\right]+\mathbb{E}\left[Y(\boldsymbol{x}^{(0)})^2\right]-2\mathbb{E}\left[Y(\boldsymbol{x})Y(\boldsymbol{x}^{(0)})\right]\\ &= 2K(\mathbf{0})-2K(\boldsymbol{x}-\boldsymbol{x}^{(0)}).\end{aligned} \tag{1.19}$$

This equality implies the result. □

Continuity of the covariance kernel at $\mathbf{0}$ is determinant, because it actually implies continuity everywhere (that matters). To make this notion precise, define $\mathbb{R}^r_{\mathcal{D}} := \{\boldsymbol{t}\in\mathbb{R}^r : (\exists\, \boldsymbol{x}^{(0)}\in\mathcal{D})\ \boldsymbol{x}^{(0)}+\boldsymbol{t}\in\mathcal{D}\}$. $\mathbb{R}^r_{\mathcal{D}}$ is a non-empty ($\mathbf{0}\in\mathbb{R}^r_{\mathcal{D}}$) open set because $\mathcal{D}$ is open.

Proposition 1.24. *A stationary Gaussian process Y is mean square continuous on $\mathcal{D}$ if, and only if, its covariance kernel $K:\mathbb{R}^r\to[0,+\infty)$ is continuous on $\mathbb{R}^r_{\mathcal{D}}$.*

Proof. Because of Proposition 1.23, we only need to prove that if a stationary Gaussian process Y is mean square continuous, then its covariance kernel is continuous on $\mathbb{R}^r_{\mathcal{D}}$.

Let $\boldsymbol{t}\in\mathbb{R}^r_{\mathcal{D}}$. Then, because $\mathcal{D}$ is open, for all $\boldsymbol{h}\in\mathbb{R}^r$ such that $\|\boldsymbol{h}\|$ is small enough, there exists $\boldsymbol{x}^{(0)}\in\mathcal{D}$ such that $\boldsymbol{x}^{(0)}+\boldsymbol{t}+\boldsymbol{h}\in\mathcal{D}$ and $\boldsymbol{x}^{(0)}+\boldsymbol{h}\in\mathcal{D}$.

$$\begin{aligned}&2(K(\boldsymbol{t}+\boldsymbol{h})-K(\boldsymbol{t}))\\ =&2K(\mathbf{0})-2K(\boldsymbol{t})-(2K(\mathbf{0})-2K(\boldsymbol{t}+\boldsymbol{h}))\\ =&\|Y(\boldsymbol{x}^{(0)}+\boldsymbol{t})-Y(\boldsymbol{x}^{(0)})\|^2-\|Y(\boldsymbol{x}^{(0)}+\boldsymbol{t}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)})\|^2\\ =&-\|Y(\boldsymbol{x}^{(0)}+\boldsymbol{t}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)}+\boldsymbol{t})\|^2\\ &\quad+2\mathbb{E}\left[\left(Y(\boldsymbol{x}^{(0)}+\boldsymbol{t}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)}+\boldsymbol{t})\right)\left(Y(\boldsymbol{x}^{(0)}+\boldsymbol{t})-Y(\boldsymbol{x}^{(0)})\right)\right]\\ =&-\|Y(\boldsymbol{x}^{(0)}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)})\|^2\\ &\quad+2\mathbb{E}\left[\left(Y(\boldsymbol{x}^{(0)}+\boldsymbol{t}+\boldsymbol{h})-Y(\boldsymbol{x}^{(0)}+\boldsymbol{t})\right)\left(Y(\boldsymbol{x}^{(0)}+\boldsymbol{t})-Y(\boldsymbol{x}^{(0)})\right)\right].\end{aligned} \tag{1.20}$$

The last equality holds because of Y's stationarity.

The definition of mean square continuity yields the result.

□

Mean square continuous differentiability of stationary Gaussian processes can also be characterized in terms of the covariance kernel [Bachoc, 2013a, chapter 2].

Proposition 1.25. *A stationary Gaussian process Y is mean square continuously differentiable if, and only if, its covariance kernel K is twice continuously differentiable on $\mathbb{R}^r_{\mathcal{D}}$. If this is the case, for all vectors $\boldsymbol{u}\in\mathbb{R}^r$, $\partial_{\boldsymbol{u}}\partial_{\boldsymbol{u}}K$ is the covariance kernel of the stationary mean square continuous Gaussian process $\partial_{\boldsymbol{u}}Y$.*

Proof. Assume the stationary process Y is mean square continuously differentiable. We prove that for every couple of vectors $(\boldsymbol{u},\boldsymbol{v})\in\mathbb{R}^r\times\mathbb{R}^r$, the partial derivative $\partial_{\boldsymbol{u}}\partial_{\boldsymbol{v}}K$ exists and is continuous over $\mathbb{R}^k_{\mathcal{D}}$. This is enough to establish that K is twice continuously differentiable on $\mathbb{R}^r_{\mathcal{D}}$.

Equation (1.20) shows that for every $\boldsymbol{t} \in \mathbb{R}^r_{\mathcal{D}}$

$$\partial_{\boldsymbol{v}} K(\boldsymbol{t}) := \lim_{h \to 0} \frac{K(\boldsymbol{t} + h\boldsymbol{v}) - K(\boldsymbol{t})}{h} = \mathbb{E}\left[\partial_{\boldsymbol{v}} Y(\boldsymbol{x}^{(0)}) \left(Y(\boldsymbol{x}^{(0)} + \boldsymbol{t}) - Y(\boldsymbol{x}^{(0)})\right)\right]. \tag{1.21}$$

So $\partial_{\boldsymbol{v}} K$ exists and, because Y is mean square continuous on $\mathcal{D}$, is continuous on $\mathbb{R}^r_{\mathcal{D}}$.

Now, for any $\boldsymbol{t} \in \mathbb{R}^r_{\mathcal{D}}$, as long as $h > 0$ is small enough, $\boldsymbol{t} + h\boldsymbol{u} \in \mathbb{R}^r_{\mathcal{D}}$. We have

$$\begin{aligned}
&\partial_{\boldsymbol{v}} K(\boldsymbol{t} + h\boldsymbol{u}) - \partial_{\boldsymbol{v}} K(\boldsymbol{t}) \\
=&\mathbb{E}\left[\partial_{\boldsymbol{v}} Y(\boldsymbol{x}^{(0)}) \left(Y(\boldsymbol{x}^{(0)} + \boldsymbol{t} + h\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})\right)\right] - \mathbb{E}\left[\partial_{\boldsymbol{v}} Y(\boldsymbol{x}^{(0)}) \left(Y(\boldsymbol{x}^{(0)} + \boldsymbol{t}) - Y(\boldsymbol{x}^{(0)})\right)\right] \\
=&\mathbb{E}\left[\partial_{\boldsymbol{v}} Y(\boldsymbol{x}^{(0)}) \left(Y(\boldsymbol{x}^{(0)} + \boldsymbol{t} + h\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)} + \boldsymbol{t})\right)\right].
\end{aligned} \tag{1.22}$$

$$\partial_{\boldsymbol{u}} \partial_{\boldsymbol{v}} K(\boldsymbol{t}) := \lim_{h \to 0} \frac{\partial_{\boldsymbol{v}} K(\boldsymbol{t} + h\boldsymbol{u}) - \partial_{\boldsymbol{v}} K(\boldsymbol{t})}{h} = \mathbb{E}\left[\partial_{\boldsymbol{u}} Y(\boldsymbol{x}^{(0)} + \boldsymbol{t}) \partial_{\boldsymbol{v}} Y(\boldsymbol{x}^{(0)})\right]. \tag{1.23}$$

So $\partial_{\boldsymbol{u}} \partial_{\boldsymbol{v}} K$ exists and, because $\partial_{\boldsymbol{v}} Y$ is mean square continuous on $\mathcal{D}$, is continuous on $\mathbb{R}^r_{\mathcal{D}}$. Notice that Equation (1.23) implies $\partial_{\boldsymbol{u}} \partial_{\boldsymbol{u}} K$ is the covariance kernel of the stationary Gaussian process with null mean function $\partial_{\boldsymbol{u}} Y$.

We now prove the converse. Assume the covariance kernel K of the stationary Gaussian process Y is twice continuously differentiable. Without loss of generality, we assume Y has null mean function. Let $\boldsymbol{u} \in \mathbb{R}^r$. We prove that $\partial_{\boldsymbol{u}} Y$ exists and is continuous on $\mathcal{D}$.

Let $(h_m)_{m \in \mathbb{N}}$ be a sequence of real numbers that converges to 0. First, we show that for all $\boldsymbol{x}^{(0)} \in \mathcal{D}$ $\left(h_m^{-1}(Y(\boldsymbol{x}^{(0)} + h_m\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)}))\right)_{m \in \mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega)$.
Let $m, n \in \mathbb{N}$.

$$\begin{aligned}
&\left\| \frac{Y(\boldsymbol{x}^{(0)} + h_m\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})}{h_m} - \frac{Y(\boldsymbol{x}^{(0)} + h_n\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})}{h_n} \right\|^2 \\
=&\left\| \frac{Y(\boldsymbol{x}^{(0)} + h_m\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})}{h_m} \right\|^2 + \left\| \frac{Y(\boldsymbol{x}^{(0)} + h_n\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})}{h_n} \right\|^2 \\
&\quad - \frac{2}{h_m h_n} \mathbb{E}\left[\left(Y(\boldsymbol{x}^{(0)} + h_m\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})\right) \left(Y(\boldsymbol{x}^{(0)} + h_n\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)})\right)\right] \\
=&\frac{2}{h_m^2}(K(\boldsymbol{0}) - K(h_m\boldsymbol{u})) + \frac{2}{h_n^2}(K(\boldsymbol{0}) - K(h_n\boldsymbol{u})) \\
&\quad - \frac{2}{h_m h_n}\left[K((h_m - h_n)\boldsymbol{u}) + K(\boldsymbol{0}) - K(h_m\boldsymbol{u}) - K(-h_n\boldsymbol{u})\right].
\end{aligned} \tag{1.24}$$

When $m, n \to \infty$, the third term converges to $2\partial_{\boldsymbol{u}} \partial_{\boldsymbol{u}} K(\boldsymbol{0})$. The first two terms must be examined more closely.

Because K is an even mapping, we have for every $\boldsymbol{v} \in \mathbb{R}^r$ $K(-\boldsymbol{v}) = K(\boldsymbol{v})$. So any partial derivative at $\boldsymbol{0}$ is necessarily null. Given K is twice continuously differentiable along $\boldsymbol{u}$, it admits a Taylor expansion of order 2 in this direction. $K(h_m\boldsymbol{u}) - K(\boldsymbol{0}) \underset{m \to \infty}{\sim} h_m^2 \partial_{\boldsymbol{u}} \partial_{\boldsymbol{u}} K(\boldsymbol{0})/2$, so each of the first two terms converges when $m, n \to \infty$ to $-\partial_{\boldsymbol{u}} \partial_{\boldsymbol{u}} K(\boldsymbol{0})$. The sum therefore converges to 0 and $\left(h_m^{-1}(Y(\boldsymbol{x}^{(0)} + h_m\boldsymbol{u}) - Y(\boldsymbol{x}^{(0)}))\right)_{m \in \mathbb{N}}$ is a Cauchy sequence in $L^2(\Omega)$.

$L^2(\Omega)$ is a Banach space so the sequence admits a limit. Let us denote this limit by $\partial_{\boldsymbol{u}} Y(\boldsymbol{x}^{(0)})$. It is Gaussian because it is an L^2 limit of Gaussian random variables. Furthermore, the mapping $\partial_{\boldsymbol{u}} Y : \mathcal{D} \to L^2(\Omega)$ $\boldsymbol{x}^{(0)} \mapsto \partial_{\boldsymbol{u}} Y(\boldsymbol{x}^{(0)})$ is a Gaussian process because for every

positive integer n and any points $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathcal{D}$, the vector $\left(\partial_{\boldsymbol{u}} Y(\boldsymbol{x}^{(1)}), ..., \partial_{\boldsymbol{u}} Y(\boldsymbol{x}^{(n)})\right)^\top$ is an L^2 limit of Gaussian vectors and is therefore Gaussian. For the same reason, $\partial_{\boldsymbol{u}} Y$ is stationary. All that remains is to show that it is mean square continuous. This follows from Equation (1.23), which is still valid because it was obtained using nothing more than Y's stationarity, Y's mean square continuity and the existence of partial derivatives of Y (but not their continuity). Taking $\boldsymbol{v} = \boldsymbol{u}$, it shows that $\partial_{\boldsymbol{u}}\partial_{\boldsymbol{u}} K$ is the covariance kernel of $\partial_{\boldsymbol{u}} Y$. Since it is continuous at $\mathbf{0}$, $\partial_{\boldsymbol{u}} Y$ is mean square continuous on $\mathcal{D}$.

□

Corollary 1.26. *For every nonnegative integer n, a stationary Gaussian process Y is of class C^n on $\mathcal{D}$ if, and only if, its covariance kernel K is of class C^{2n} on $\mathbb{R}^r_{\mathcal{D}}$.*

1.4 Spectral representation

Stein [1999] showed how fruitful the spectral representation of Gaussian processes could be in order to understand its properties. In particular, it provides a useful characterization of means square continuity and differentiability.

Bochner [1932]'s theorem is the cornerstone of all spectral results. A few preliminary results, which are by themselves of interest because they provide methods for constructing covariance kernels, are useful for showing it. These results are drawn from Rasmussen and Williams [2006].

In all that follows, $\mathcal{D} = \mathcal{V} = \mathbb{R}^r$ for some positive integer r.
The set of positive definite mappings is stable for several operations listed below.

Proposition 1.27. *The sum of two positive definite mappings $\mathbb{R}^r \to \mathbb{R}$ is positive definite.*

Proof. This is the translation of the fact that the sum of two positive semi-definite matrices is positive semi-definite. □

Proposition 1.28. *The product of two positive definite mappings $\mathbb{R}^r \to \mathbb{R}$ is positive definite.*

Proof. Let K and K' be two positive definite mappings $\mathbb{R}^r \to \mathbb{R}$. Let Y and Y' be independent stationary Gaussian processes with null mean function and covariance kernel K and K' respectively. Let Z be the random process defined by $Z(\boldsymbol{x}) = Y(\boldsymbol{x})Y'(\boldsymbol{x})$ $(\boldsymbol{x} \in \mathbb{R}^r)$.

$$\mathbb{E}[Z(\boldsymbol{x}^{(1)})Z(\boldsymbol{x}^{(2)})] = \mathbb{E}[Y(\boldsymbol{x}^{(1)})Y(\boldsymbol{x}^{(2)})]\mathbb{E}[Y'(\boldsymbol{x}^{(1)})Y'(\boldsymbol{x}^{(2)})] = K(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)})K'(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}) \tag{1.25}$$

So the mapping $\boldsymbol{x} \mapsto K(\boldsymbol{x})K'(\boldsymbol{x})$ is a covariance kernel: it is positive definite.

□

Proposition 1.29. *Let $K_1 : \mathbb{R}^r \to \mathbb{R}$ be a continuous function that is integrable with respect to the Lebesgue measure on $\mathbb{R}^r$ and let K_2 be a continuous positive definite mapping $\mathbb{R}^r \to \mathbb{R}$. Then the "double-convolution" of K_1 and K_2 (defined hereafter) is a positive definite mapping $\mathbb{R}^r \to \mathbb{R}$.*
For all $\boldsymbol{x} \in \mathbb{R}^r$,

$$K_1 * K_2 * K_1(\boldsymbol{x}) = \int_{\mathbb{R}^r}\int_{\mathbb{R}^r} K_1(\boldsymbol{x} - \boldsymbol{z})K_2(\boldsymbol{z} - \boldsymbol{z}')K_1(\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}' \tag{1.26}$$

Proof. Let Y be a stationary Gaussian process with null mean function and covariance kernel K_2. First, we need to properly define $Z(\boldsymbol{x}) := \int_{\mathbb{R}^r} K_1(\boldsymbol{x}-\boldsymbol{z})Y(\boldsymbol{z})d\boldsymbol{z}$. Since K_2 is continuous, Y is mean square continuous. Since it is stationary, $\|Y(\boldsymbol{x})\|$ is constant as a function of $\boldsymbol{x}$. Moreover, K_1 is integrable. This means that $Z(\boldsymbol{x}) := \int_{\mathbb{R}^r} K_1(\boldsymbol{x}-\boldsymbol{z})Y(\boldsymbol{z})d\boldsymbol{z}$ can be defined in the Riemann sense as an L^2 limit of Gaussian random variables and is therefore Gaussian. Similarly, for any positive integer n and any $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$, $(Z(\boldsymbol{x}^{(1)}), ..., Z(\boldsymbol{x}^{(n)}))^\top$ is an L^2 limit of Gaussian vectors and therefore a Gaussian vector, so $Z : \boldsymbol{x} \mapsto Z(\boldsymbol{x})$ is a Gaussian process with null mean function.
For all $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \in \mathbb{R}^r$,

$$\begin{aligned}
\mathbb{E}[Z(\boldsymbol{x}^{(1)})Z(\boldsymbol{x}^{(2)})] &= \mathbb{E}[\int_{\mathbb{R}^r} K_1(\boldsymbol{x}^{(1)}-\boldsymbol{z})Y(\boldsymbol{z})d\boldsymbol{z} \int_{\mathbb{R}^r} K_1(\boldsymbol{x}^{(2)}-\boldsymbol{z}')Y(\boldsymbol{z}')d\boldsymbol{z}'] \\
&= \int_{\mathbb{R}^r}\int_{\mathbb{R}^r} K_1(\boldsymbol{x}^{(1)}-\boldsymbol{z})\mathbb{E}[Y(\boldsymbol{z})Y(\boldsymbol{z}')]K_1(\boldsymbol{x}^{(2)}-\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}' \qquad (1.27) \\
&= \int_{\mathbb{R}^r}\int_{\mathbb{R}^r} K_1(\boldsymbol{x}^{(1)}-\boldsymbol{z})K_2(\boldsymbol{z}-\boldsymbol{z}')K_1(\boldsymbol{x}^{(2)}-\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}' \\
&= \int_{\mathbb{R}^r}\int_{\mathbb{R}^r} K_1(\boldsymbol{x}^{(1)}-\boldsymbol{x}^{(2)}-\boldsymbol{z})K_2(\boldsymbol{z}-\boldsymbol{z}')K_1(\boldsymbol{z}')d\boldsymbol{z}d\boldsymbol{z}'. \qquad (1.28)
\end{aligned}$$

In the above computation, Equation (1.27) holds because the scalar product of the limits is the limit of the scalar products and Equation (1.28) is obtained by the changes of variable $\boldsymbol{z} \leftarrow \boldsymbol{z}-\boldsymbol{x}^{(2)}$ and $\boldsymbol{z}' \leftarrow \boldsymbol{z}'-\boldsymbol{x}^{(2)}$ and because $K_1(-\boldsymbol{z}') = K_1(\boldsymbol{z})$.

Equation (1.28) shows that $K_1 * K_2 * K_1$ is the covariance kernel of a stationary Gaussian process, so it is a positive definite mapping.

□

Bochner's theorem is a characterization of positive definite mappings $\mathbb{R}^r \to \mathbb{R}$ in terms of their spectrum.

Theorem 1.30 (Bochner's theorem)**.** *A mapping $K : \mathbb{R}^r \to \mathbb{R}$ is continuous and positive definite if, and only if, there exists a finite positive measure μ such that for all $\boldsymbol{x} \in \mathbb{R}^r$*

$$K(\boldsymbol{x}) = \int_{\mathbb{R}^r} e^{i\langle\boldsymbol{\omega}|\boldsymbol{x}\rangle} d\mu(\boldsymbol{\omega}). \qquad (1.29)$$

Proof. Let μ be a finite positive measure. For all $\boldsymbol{x} \in \mathbb{R}^r$, define

$$K(\boldsymbol{x}) = \int_{\mathbb{R}^r} \exp(i\,\langle\boldsymbol{\omega}|\boldsymbol{x}\rangle) d\mu(\boldsymbol{\omega}). \qquad (1.30)$$

K is, due to the Dominated Convergence theorem, a continuous mapping $\mathbb{R}^r \to \mathbb{R}$. We now show it is positive definite.

For every positive integer n and all $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^\top \in \mathbb{R}^n$,

$$\begin{aligned}
\sum_{i,j\in[\![1,n]\!]} \xi_i\xi_j K(\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(j)}) &= \int_{\mathbb{R}^r} \sum_{i,j\in[\![1,n]\!]} \left(\xi_i \exp(i\langle\boldsymbol{\omega}|\boldsymbol{x}^{(i)}\rangle)\overline{\xi_j \exp(i\langle\boldsymbol{\omega}|\boldsymbol{x}^{(j)}\rangle)}\right) d\mu(\boldsymbol{\omega}) \\
&= \int_{\mathbb{R}^r} \sum_{i=1}^{n} \left|\xi_i \exp(i\langle\boldsymbol{\omega}|\boldsymbol{x}^{(i)}\rangle)\right|^2 d\mu(\boldsymbol{\omega}) \\
&\geqslant 0. \qquad (1.31)
\end{aligned}$$

To prove the converse result, a step-by-step approach is required. First, we prove the result for positive definite functions that are 1) integrable with respect to the Lebesgue measure on $\mathbb{R}^r$ and 2) whose Fourier transform is also integrable with respect to the Lebesgue measure on $\mathbb{R}^r$. We then successively relax 2) and 1).

Let K be a positive definite function that satisfies 1) and 2). Then define $\hat{K} : \mathbb{R}^r \to \mathbb{R}$ by

$$\hat{K}(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} K(\boldsymbol{x}) e^{-i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{x}. \tag{1.32}$$

By assumption, $\hat{K}$ is integrable. Notice that it is also continuous.

For all $\boldsymbol{x} \in \mathbb{R}^k$, denote by $\phi_{\boldsymbol{x}}$ the mapping $\mathbb{R}^k \to \mathbb{C}$ defined by $\phi_{\boldsymbol{x}}(\boldsymbol{\omega}) = \exp(i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle)$.

For any $\boldsymbol{x}^{(1)}, \boldsymbol{x}^{(2)} \in \mathbb{R}^k$,

$$K(\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}) = \int_{\mathbb{R}^r} \hat{K}(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)} \rangle} d\boldsymbol{\omega} = \int_{\mathbb{R}^r} \phi_{\boldsymbol{x}^{(1)}} \overline{\phi_{\boldsymbol{x}^{(2)}}} \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{1.33}$$

So, for any positive integer n, any $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^\top \in \mathbb{R}^n$ and any $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)} \in \mathbb{R}^k$,

$$\begin{aligned} \int_{\mathbb{R}^r} \left| \sum_{k=1}^{n} \xi_k \phi_{\boldsymbol{x}^{(k)}}(\boldsymbol{\omega}) \right|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} &= \sum_{k,l \in [\![1,n]\!]} \xi_k \xi_l \int_{\mathbb{R}^r} \phi_{\boldsymbol{x}^{(k)}}(\boldsymbol{\omega}) \overline{\phi_{\boldsymbol{x}^{(l)}}(\boldsymbol{\omega})} \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \sum_{k,l \in [\![1,n]\!]} \xi_k \xi_l K(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)}) \\ &\geqslant 0. \end{aligned} \tag{1.34}$$

Equation (1.34) shows that for any element f of the vector space spanned by $\{\phi_{\boldsymbol{x}} : \boldsymbol{x} \in \mathbb{R}^k\}$,

$$\int_{\mathbb{R}^r} |f(\boldsymbol{\omega})|^2 d\boldsymbol{\omega} \in [0, +\infty). \tag{1.35}$$

For all r-tuples $\boldsymbol{a}$ and $\boldsymbol{b}$ such that $\boldsymbol{a} \leqslant \boldsymbol{b}$ (in the sense that for every $i \in [\![1, r]\!]$, $a_i \leqslant b_i$), define the set $I_{\boldsymbol{a},\boldsymbol{b}} := [a_1, b_1] \times ... \times [a_r, b_r]$.

For all r-tuples $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{a}'$, $\boldsymbol{b}'$ such that $\boldsymbol{a}' \leqslant \boldsymbol{a} \leqslant \boldsymbol{b} \leqslant \boldsymbol{b}'$, define $p_{\boldsymbol{a},\boldsymbol{b}} : \mathbb{R}^k \to \mathbb{R}^k$ as the orthogonal projection on $I_{\boldsymbol{a},\boldsymbol{b}}$ and $p^c_{\boldsymbol{a}',\boldsymbol{b}'} : \mathbb{R}^k \to \mathbb{R}^k$ as the orthogonal projection on the closure of $\mathbb{R}^k \setminus I_{\boldsymbol{a}',\boldsymbol{b}'}$. Let $f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}'} : \mathbb{R}^k \to \mathbb{R}$ be the continuous mapping defined by

$$f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}'}(\boldsymbol{\omega}) = \frac{\|\boldsymbol{\omega} - p^c_{\boldsymbol{a}',\boldsymbol{b}'}(\boldsymbol{\omega})\|}{\|p_{\boldsymbol{a},\boldsymbol{b}}(\boldsymbol{\omega}) - p^c_{\boldsymbol{a}',\boldsymbol{b}'}(\boldsymbol{\omega})\|}. \tag{1.36}$$

For all r-tuples $\boldsymbol{a}$, $\boldsymbol{b}$, $\boldsymbol{a}'$, $\boldsymbol{b}'$, $\boldsymbol{a}''$, $\boldsymbol{b}''$ such that $\boldsymbol{a}'' \leqslant \boldsymbol{a}' \leqslant \boldsymbol{a} \leqslant \boldsymbol{b} \leqslant \boldsymbol{b}' \leqslant \boldsymbol{b}''$, define the continuous periodic mapping $f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''} : \mathbb{R}^k \to \mathbb{R}$ as follows. For every $\boldsymbol{\omega} \in \mathbb{R}^k$, there exists an r-tuple $\boldsymbol{n} = (n_1, ..., n_r) \in \mathbb{Z}^r$ such that $\boldsymbol{\omega} \in I_{\boldsymbol{a}'' + \boldsymbol{n} \cdot (\boldsymbol{b}'' - \boldsymbol{a}''), \boldsymbol{b}'' + \boldsymbol{n} \cdot (\boldsymbol{b}'' - \boldsymbol{a}'')}$ (here $\cdot$ denotes the dot product). Then

$$f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}(\boldsymbol{\omega}) = f_{\boldsymbol{a}+\boldsymbol{n}\cdot(\boldsymbol{b}''-\boldsymbol{a}''),\boldsymbol{b}+\boldsymbol{n}\cdot(\boldsymbol{b}''-\boldsymbol{a}''),\boldsymbol{a}'+\boldsymbol{n}\cdot(\boldsymbol{b}''-\boldsymbol{a}''),\boldsymbol{b}'+\boldsymbol{n}\cdot(\boldsymbol{b}''-\boldsymbol{a}'')}(\boldsymbol{\omega}). \tag{1.37}$$

Because $f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}$ is a continuous periodic function, there exists a sequence $(f_n)_{n \in \mathbb{N}}$ of functions belonging to the vector space spanned by $\{\phi_{\boldsymbol{x}} : \boldsymbol{x} \in \mathbb{R}^k\}$ that uniformly converges to $f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}$. Because the convergence is uniform, there exists a constant $C > 0$ such

that for all $n \in \mathbb{N}$ and all $\boldsymbol{\omega} \in \mathbb{R}^r$, $|f_n(\boldsymbol{\omega})| \leqslant C$. This fact, along with pointwise convergence makes the Dominated Convergence theorem applicable.

$$\begin{aligned}\int_{\mathbb{R}^r} |f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}(\boldsymbol{\omega})|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} &= \int_{\mathbb{R}^r} \left|\lim_{n\to\infty} f_n(\boldsymbol{\omega})\right|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \lim_{n\to\infty} \int_{\mathbb{R}^r} |f_n(\boldsymbol{\omega})|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &\geqslant 0. \end{aligned} \tag{1.38}$$

Clearly, for every $\boldsymbol{\omega} \in \mathbb{R}^r$, $|f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}(\boldsymbol{\omega})| \leqslant 1$. Let us make the period $\boldsymbol{b}'' - \boldsymbol{a}''$ increase to infinity and apply the Dominated Convergence theorem again.

$$\begin{aligned}\int_{\mathbb{R}^r} |f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}'}(\boldsymbol{\omega})|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} &= \int_{\mathbb{R}^r} \left|\lim_{\substack{\boldsymbol{a}''\to(-\infty)^r \\ \boldsymbol{b}''\to(+\infty)^r}} f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}(\boldsymbol{\omega})\right|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \lim_{\substack{\boldsymbol{a}''\to(-\infty)^r \\ \boldsymbol{b}''\to(+\infty)^r}} \int_{\mathbb{R}^r} |f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}',\boldsymbol{a}'',\boldsymbol{b}''}(\boldsymbol{\omega})|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &\geqslant 0. \end{aligned} \tag{1.39}$$

Finally, let us make $\boldsymbol{a}' \to \boldsymbol{a}$ ($\boldsymbol{a}' \leqslant \boldsymbol{a}$) and $\boldsymbol{b}' \to \boldsymbol{b}$ ($\boldsymbol{b}' \geqslant \boldsymbol{b}$) and apply the Dominated Convergence theorem one last time.

$$\begin{aligned}\int_{I_{\boldsymbol{a},\boldsymbol{b}}} \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} = \int_{\mathbb{R}^r} |\mathbf{1}_{I_{\boldsymbol{a},\boldsymbol{b}}}(\boldsymbol{\omega})|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} &= \int_{\mathbb{R}^r} \left|\lim_{\substack{\boldsymbol{a}'\to\boldsymbol{a} \\ \boldsymbol{b}'\to\boldsymbol{b}}} f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}'}(\boldsymbol{\omega})\right|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &= \lim_{\substack{\boldsymbol{a}'\to\boldsymbol{a} \\ \boldsymbol{b}'\to\boldsymbol{b}}} \int_{\mathbb{R}^r} |f_{\boldsymbol{a},\boldsymbol{b},\boldsymbol{a}',\boldsymbol{b}'}(\boldsymbol{\omega})|^2 \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} \\ &\geqslant 0. \end{aligned} \tag{1.40}$$

Because $\hat{K}$ is continuous, this implies that $\hat{K}$ is a nonnegative function. It is therefore the density with respect to the Lebesgue measure of a positive measure. Moreover, $\hat{K}$ is by assumption integrable with respect to the Lebesgue measure, so it is the density of a finite positive measure.

For all $R > 0$, define the mapping $G_R : \mathbb{R}^r \to \mathbb{R}$ by $G_R(\boldsymbol{x}) = (2\pi R^2)^{-r/2} \exp(-\|\boldsymbol{x}\|^2/(2R^2))$. G_R is continuous and integrable with respect to the Lebesgue measure on $\mathbb{R}^r$. Its Fourier transform is defined by

$$\hat{G}_R(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} G_R(\boldsymbol{x}) e^{-i\langle \boldsymbol{\omega}|\boldsymbol{x}\rangle} d\boldsymbol{x} = (2\pi)^{-r} e^{-\frac{R^2\|\boldsymbol{\omega}\|^2}{2}}. \tag{1.41}$$

$\hat{G}_R$ is also integrable with respect to the Lebesgue measure on $\mathbb{R}^r$. It is the density of a positive finite measure, so applying the result shown above, G_R is a positive definite function. It is a "Squared Exponential" kernel.
Squared Exponential kernels are the main tools we will use to relax Assumptions 1) and 2).

First, let us relax Assumption 2). Let K be a continuous positive definite mapping that satisfies Assumption 1): it is integrable with respect to the Lebesgue measure on $\mathbb{R}^r$.

Then, for all $R > 0$, $G_R * K * G_R$ is also a continuous positive definite mapping. And for all $\boldsymbol{\omega} \in \mathbb{R}^r$,

$$\widehat{G_R * K * G_R}(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} G_R * K * G_R(\boldsymbol{x}) e^{-i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{x} = \hat{K}(\boldsymbol{\omega}) e^{-R^2 \|\boldsymbol{\omega}\|^2} \tag{1.42}$$

$\widehat{G_R * K * G_R}$ is integrable with respect to the Lebesgue measure on $\mathbb{R}^r$. Therefore, $G_R * K * G_R$ is a continuous positive definite mapping that satisfies both Assumption 1) and 2). The result established previously can be applied: $\widehat{G_R * K * G_R}$ is the density with respect to the Lebesgue measure on $\mathbb{R}^r$ of a finite positive measure. Equation (1.42) then shows that $\hat{K}$ is a nonnegative mapping, so it is the density of a positive measure.

The Monotone Convergence theorem yields that

$$\lim_{R \to 0} \int_{\mathbb{R}^r} \hat{K}(\boldsymbol{\omega}) e^{-R^2 \|\boldsymbol{\omega}\|^2} d\boldsymbol{\omega} = \int_{\mathbb{R}^r} \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{1.43}$$

For all $R > 0$ $\int_{\mathbb{R}^r} \hat{K}(\boldsymbol{\omega}) e^{-R^2 \|\boldsymbol{\omega}\|^2} d\boldsymbol{\omega} = G_R * K * G_R(\mathbf{0})$. Moreover, because K is continuous and G_R is the probability density of the Normal distribution with null mean and standard deviation R,

$$\lim_{R \to 0} G_R * K * G_R(\mathbf{0}) = K(\mathbf{0}). \tag{1.44}$$

Equations (1.43) and (1.44) yield

$$\int_{\mathbb{R}^r} \hat{K}(\boldsymbol{\omega}) d\boldsymbol{\omega} = K(\mathbf{0}), \tag{1.45}$$

so $\hat{K}$ is the density of a finite measure with total mass $K(\mathbf{0})$.

All that remains to do is relax Assumption 1). Let K be a continuous positive definite mapping $\mathbb{R}^r \to \mathbb{R}$ such that $K(\mathbf{0}) \neq 0$ (because if $K(\mathbf{0}) = 0$, then K is the null function and there is nothing to prove). For all $R > 0$, the mapping $K_R : \boldsymbol{x} \mapsto (2\pi R)^{r/2} K(\boldsymbol{x}) G_R(\boldsymbol{x}) K(\mathbf{0})^{-1}$ is positive definite. Because K is necessarily bounded and G_R is integrable with respect to the Lebesgue measure on $\mathbb{R}^r$, K_R is integrable too. So, applying the previous result, it is the characteristic function of a probability measure μ_R. When $R \to +\infty$, K_R converges pointwise to the function $K_0 : \boldsymbol{x} \mapsto K(\boldsymbol{x}) K(\mathbf{0})^{-1}$. This implies that μ_R converges narrowly to a probability measure μ_0 and that K_0 is the characteristic function of μ_0. Therefore, for all $\boldsymbol{x} \in \mathbb{R}^r$,

$$K(\boldsymbol{x}) = K(\mathbf{0}) \int_{\mathbb{R}^r} e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\mu_0(\boldsymbol{\omega}). \tag{1.46}$$

□

An other proof can be found in Gihman and Skorohod [1974] page 208.

Sometimes, a more precise version of Bochner's theorem is required. The usual terminology is somewhat misleading. A positive definite kernel leads to positive **semi**-definite covariance matrices. But when does it lead to positive definite covariance matrices?

Proposition 1.31. *Let μ be a positive measure on $\mathbb{R}^r$ with finite non-null total mass that is absolutely continuous with respect to the Lebesgue measure. Then the mapping $K : \mathbb{R}^r \to \mathbb{R}$ defined by*

$$K(\boldsymbol{x}) = \int_{\mathbb{R}^r} e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\mu(\boldsymbol{\omega}) \tag{1.47}$$

is positive definite. Moreover, for any $\boldsymbol{\xi} \in \mathbb{R}^n \setminus \{\mathbf{0}_n\}$,

$$\sum_{k,l \in [\![1,n]\!]} \xi_k \xi_l K(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)}) > 0. \tag{1.48}$$

Proof. The first part results from Bochner's theorem. Let us show the second.

$$\begin{aligned} \sum_{k,l \in [\![1,n]\!]} \xi_k \xi_l K(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)}) &= \sum_{k,l \in [\![1,n]\!]} \xi_k \xi_l \int_{\mathbb{R}^r} e^{i\langle \boldsymbol{\omega} | \boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)} \rangle} d\mu(\boldsymbol{\omega}) \\ &= \int_{\mathbb{R}^r} \left| \sum_{k=1}^{n} \xi_k e^{i\langle \boldsymbol{\omega} | \boldsymbol{x}^{(k)} \rangle} \right|^2 d\mu(\boldsymbol{\omega}). \end{aligned} \tag{1.49}$$

Given $\boldsymbol{x}^{(1)}, ..., \boldsymbol{x}^{(n)}$ are all distinct, for almost all unitary vector $\boldsymbol{u}$ in the sense of the Lebesgue measure on the unit sphere S^{r-1}, the real numbers $\langle \boldsymbol{u} | \boldsymbol{x}^{(1)} \rangle, ..., \langle \boldsymbol{u} | \boldsymbol{x}^{(n)} \rangle$ are distinct. Indeed, if two of these numbers, say $\langle \boldsymbol{u} | \boldsymbol{x}^{(1)} \rangle$ and $\langle \boldsymbol{u} | \boldsymbol{x}^{(2)} \rangle$, were equal, then $\boldsymbol{u}$ would be orthogonal to $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}$. But the set of all vectors of $\mathbb{R}^r$ orthogonal to $\boldsymbol{x}^{(1)} - \boldsymbol{x}^{(2)}$ is a hyperplane. So there exists a finite number of hyperplanes of $\mathbb{R}^r$ such that, if $\boldsymbol{u}$ does not belong to any of them, the real numbers $\langle \boldsymbol{u} | \boldsymbol{x}^{(1)} \rangle, ..., \langle \boldsymbol{u} | \boldsymbol{x}^{(n)} \rangle$ are distinct.

Now, notice that the mapping $\mathbb{C} \to \mathbb{C}; z \mapsto \sum_{k=1}^{n} \xi_k e^{iz\langle \boldsymbol{u} | \boldsymbol{x}^{(k)} \rangle}$ is holomorphic. So either it is the null function or all its zeros are isolated. Given it clearly is not the null function, its zeros are isolated. So the set of all zeros that belong to $\mathbb{R}$ is countable and therefore of null Lebesgue measure.

Let $f : \mathbb{R}^r \to \{0, 1\}$ be the measurable mapping such that $f(\boldsymbol{\omega}) = 1$ if $\sum_{k=1}^{n} \xi_k e^{i\langle \boldsymbol{\omega} | \boldsymbol{x}^{(k)} \rangle} = 0$ and $f(\boldsymbol{\omega}) = 0$ if not.

$$\int_{\mathbb{R}^r} f(\boldsymbol{\omega}) d\boldsymbol{\omega} = \frac{2\pi^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)} \int_{S^{r-1}} \int_{(0,+\infty)} f(t\boldsymbol{u}) t^{r-1} dt d\boldsymbol{u} = \frac{2\pi^{\frac{r}{2}}}{\Gamma\left(\frac{r}{2}\right)} \int_{S^{r-1}} 0 \, d\boldsymbol{u} = 0. \tag{1.50}$$

Therefore the mapping $\mathbb{R}^r \to \mathbb{C}; \boldsymbol{\omega} \mapsto \sum_{k=1}^{n} \xi_k e^{i\langle \boldsymbol{\omega} | \boldsymbol{x}^{(k)} \rangle}$ takes null values on a Borel set that is negligible with respect to the Lebesgue measure. This set is therefore also negligible with respect to μ, which yields the conclusion. □

The following results give the link between the existence of moments of the spectral measure and the smoothness of any associated Gaussian process [Stein, 1999, chapter 2].

Proposition 1.32. *For every nonnegative integer n, a positive definite mapping $\mathbb{R}^r \to \mathbb{R}$ is of class C^n if, and only if, its Fourier transform μ verifies*

$$\int_{\mathbb{R}^r} \|\boldsymbol{\omega}\|^n d\mu(\boldsymbol{\omega}) < +\infty. \tag{1.51}$$

Corollary 1.33. *For every nonnegative integer n, a stationary Gaussian process $Y : \mathcal{D} \subset \mathbb{R}^k \to L^2(\Omega)$ whose covariance kernel K is positive definite on $\mathbb{R}^k$ is of class C^n on $\mathcal{D}$ if, and only if, K's Fourier transform μ verifies*

$$\int_{\mathbb{R}^r} \|\boldsymbol{\omega}\|^{2n} d\mu(\boldsymbol{\omega}) < +\infty. \tag{1.52}$$

To illustrate this, we provide in Figure 1.2 Kriging examples for differently smooth processes.

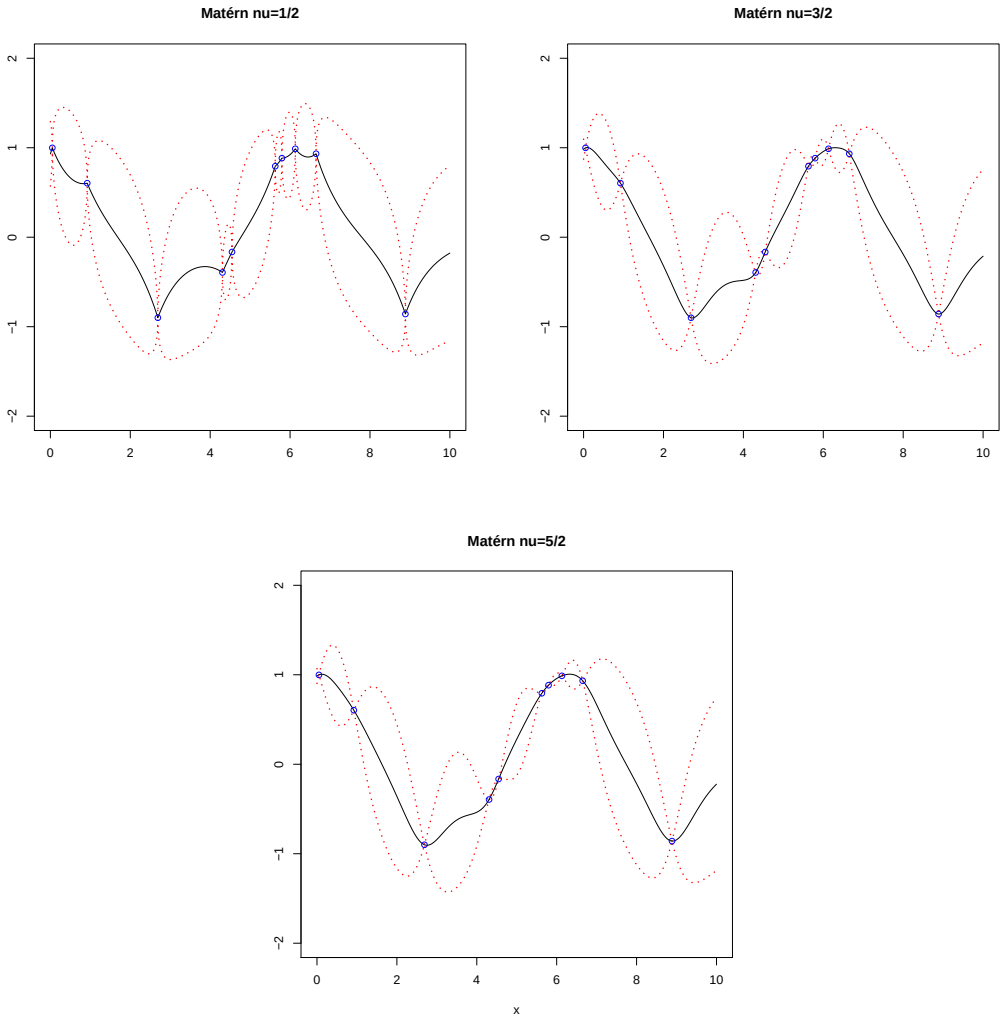


Figure 1.2 – *Kriging with differently smooth Gaussian processes. Top left: mean square continuous. Top right: mean square continuously differentiable. Bottom: twice mean squre continuously differentiable. Solid lines represent conditional means and dotted lines add or substract conditional standard deviations.*

1.5 Examples of correlation kernels

In this thesis, four families of correlation kernels are discussed (cf. Table 1.1). All corresponding spectral probability measures are absolutely continuous with respect to the Lebesgue measure.

Kernel	$K_\theta(\lvert x\rvert)$	parameter range
Spherical ($r = 1, 2, 3$)	$\left(1 - \frac{3}{2}\left(\frac{\lvert x\rvert}{\theta}\right) + \frac{1}{2}\left(\frac{\lvert x\rvert}{\theta}\right)^3\right)\mathbf{1}_{\{\lvert x\rvert \leqslant \theta\}}$	$\emptyset$
Power Exponential	$\exp\left\{-\left(\frac{\lvert x\rvert}{\theta}\right)^q\right\}$	$q \in (0, 2]$
Rational Quadratic	$\left(1 + \left(\frac{\lvert x\rvert}{\theta}\right)^2\right)^{-\nu}$	$\nu \in (0, +\infty)$
Matérn	$\Gamma(\nu)^{-1}2^{1-\nu}\left(2\sqrt{\nu}\frac{\lvert x\rvert}{\theta}\right)^{\nu}\mathcal{K}_\nu\left(2\sqrt{\nu}\frac{\lvert x\rvert}{\theta}\right)$	$\nu \in (0, +\infty)$

Table 1.1 – *Formulas for several correlation kernel families. The Squared Exponential kernel is the Power Exponential kernel with $q = 2$. $\mathcal{K}_\nu$ is the modified Bessel function of second kind with parameter ν [Abramowitz and Stegun, 1964](9.6.).*

The spherical kernel is only used in geostatistics ([Journel and Huijbregts, 1978], p.116, [Isaaks and Srivastava, 1989], p.374, [Bras and Rodriguez-Iturbe, 1985], p.418, [Christakos, 1992], p.71, [Wackernagel, 1995], p.42, [Kitadinis, 1997], p.56, [Goovaerts, 1997], p.88) mainly because it is limited to 1, 2 and 3-dimensional settings: Bochner's theorem can be used to show that its Fourier transform in 4 or higher-dimensional settings is no positive measure. Its main advantage is its simplicity. It is continuous, so Gaussian processes using it are mean square continuous. It is not differentiable at 0, so corresponding Gaussian processes are not mean square continuously differentiable. Moreover, the values of the Gaussian process at points with distance greater or equal to θ are independent. This is a peculiar assumption that is generally hard to justify, but can lead to sparsity in correlation matrices.

Spectral density depends on the dimension. In the following, J_0 and J_1 are respectively the Bessel function of first kind with parameter 0 and 1. H_0 and H_1 are respectively the Struve functions with parameter 0 and 1.

$$\hat{K}(\boldsymbol{\omega}) = \begin{cases} \theta\ \frac{3(\|\theta\boldsymbol{\omega}\|^2 - 2\|\theta\boldsymbol{\omega}\|\sin(\|\theta\boldsymbol{\omega}\|) - 2\cos(\|\theta\boldsymbol{\omega}\|) + 2)}{2\pi\theta\|\theta\boldsymbol{\omega}\|^4} & \text{for } r = 1. \\ \theta^2\ \left[\frac{H_0(\|\theta\boldsymbol{\omega}\|)J_1(\|\theta\boldsymbol{\omega}\|)}{4} + \frac{(2 - \pi H_1(\|\theta\boldsymbol{\omega}\|))J_0(\|\theta\boldsymbol{\omega}\|)}{4\pi} - \frac{J_1(\|\theta\boldsymbol{\omega}\|)}{\|\theta\boldsymbol{\omega}\|} - \frac{J_2(\|\theta\boldsymbol{\omega}\|)}{\|\theta\boldsymbol{\omega}\|^2}\right] & \text{for } r = 2. \\ \theta^3\ \frac{\|\theta\boldsymbol{\omega}\|^3 - 3\sin(\|\theta\boldsymbol{\omega}\|) + 3\|\theta\boldsymbol{\omega}\|\cos(\|\theta\boldsymbol{\omega}\|)}{\pi^2\|\theta\boldsymbol{\omega}\|^5} & \text{for } r = 3. \end{cases}$$

Power Exponential kernels, and especially the Exponential ($q = 1$) and Squared Exponential ($q = 2$) kernels are probably the most widely used family. They are as simple as the Spherical kernel and have the advantage of extending to dimensions greater than 3. They provide some flexibility with respect the smoothness of the associated Gaussian processes, although it is of the all-or-nothing kind. For $q < 2$, Gaussian processes are mean square continuous but not mean square differentiable. For $q = 2$, Gaussian processes are infinitely mean square continuously differentiable. In fact, Stein [1999] points out that in a 1-dimensional setting, letting $Y^{(i)}(0)$ be the i-th derivative of the Gaussian process at 0, $\sum_{i=0}^{n} Y^{(i)}(0)t^i/i!$ converges to $Y(t)$ in L^2. So observing Y on any neighborhood of 0 (and therefore of any point) is enough to almost surely know the value of Y at any other point! This behavior is usually considered physically unrealistic.

The Squared Exponential kernel has an other remarkable feature: it is left invariant (up to a multiplicative constant) by the Fourier transform:

$$\hat{K}(\boldsymbol{\omega}) = (2\pi)^{-r}(\pi\theta^2)^{r/2}\exp(-\theta^2\|\boldsymbol{\omega}\|^2/4). \tag{1.53}$$

Rational Quadratic kernels provide no flexibility with respect to smoothness: all lead to infinitely mean square continuously differentiable Gaussian processes. They share the Squared Exponential kernel's defect, but on a smaller scale: $\sum_{i=0}^{n} Y^{(i)}(0)t^i/i!$ converges to $Y(t)$ in L^2 only if $t < \theta$.

Rational Quadratic kernels have the following spectral density, with $\mathcal{K}_\nu$ being the modified Bessel function of second kind with parameter ν:

$$\hat{K}(\boldsymbol{\omega}) = \theta^r \frac{(2\pi)^{-r/2}\|\boldsymbol{\omega}\|^{\nu-r/2}\mathcal{K}_{\nu-r/2}(\|\boldsymbol{\omega}\|)}{2^{\nu-1}\Gamma(\nu)}. \tag{1.54}$$

Matérn kernels are recommended by Stein [1999] because they offer great flexibility regarding smoothness of the Gaussian process. The density with respect to the Lebesgue measure of the associated spectral probability measure is proportional to $\left(4\nu + \|\boldsymbol{\omega}\|^2/\theta^2\right)^{-\frac{r}{2}-\nu}$ [Rasmussen and Williams, 2006]. In fact, this kernel is specifically constructed in order to have a nice spectral density. The study of this measure is straightforward and leads to the conclusion that the associated Gaussian process is $\lfloor \nu \rfloor$ times mean square continuously differentiable if ν is not an integer and $\nu - 1$ continuously differentiable if it is. Moreover, it is mean square continuous for any $\nu > 0$. The Matérn family admits comfortable expressions when ν is a half-integer ($\nu = 1/2, 3/2, 5/2, ...$). Notably, it includes the Exponential kernel ($\nu = 1/2$) and admits the Squared Exponential kernel as its limit when $\nu \to +\infty$. Notice that Matérn and Rational Quadratic kernels are the Fourier transform of one another.

Parameter ν	Analytical expression
$\frac{1}{2}$	$\sigma^2 e^{-\sqrt{2}\frac{\lvert x\rvert}{\ell}}$
$\frac{3}{2}$	$\sigma^2\left(1+\sqrt{6}\frac{\lvert x\rvert}{\ell}\right)e^{-\sqrt{6}\frac{\lvert x\rvert}{\ell}}$
$\frac{5}{2}$	$\sigma^2\left(1+\sqrt{10}\frac{\lvert x\rvert}{\ell}+\frac{10}{3}\frac{\lvert x\rvert^2}{\ell^2}\right)e^{-\sqrt{10}\frac{\lvert x\rvert}{\ell}}$
∞	$\sigma^2 e^{-\frac{\lvert x\rvert^2}{\ell^2}}$

Figure 1.3 illustrates several kernels from all four families.
Estimating kernel parameters is often no simple task, and is one of the motivations of this thesis. Here, the matter is tackled from an Objective Bayesian standpoint. For a frequentist perspective, see Bachoc [2013b].

1.6 Current Kriging-related research

Although the aim of this chapter was only to provide an understanding of the Kriging tools used in this thesis, let us briefly mention several recent developments in the field. See also Bachoc et al. [2017a].

Apart from geostatistical applications, one of the main uses for Kriging methods is emulation of computer codes. Kriging is therefore used for both calibration and prediction purposes [Bachoc et al., 2014]. Gaussian processes can be used to perform uncertainty quantification when optimizing under constraints [Binois et al., 2015]. The Bayesian framework allows additional flexibility by considering several different covariance kernels at once [Pronzato and Rendas, 2017] – the present thesis makes such an approach more systematic.

An important issue for all spatial statistics is the design of experiments. One can for example wish for a "space-filling" design set. Although the concept is simple, practical implementation

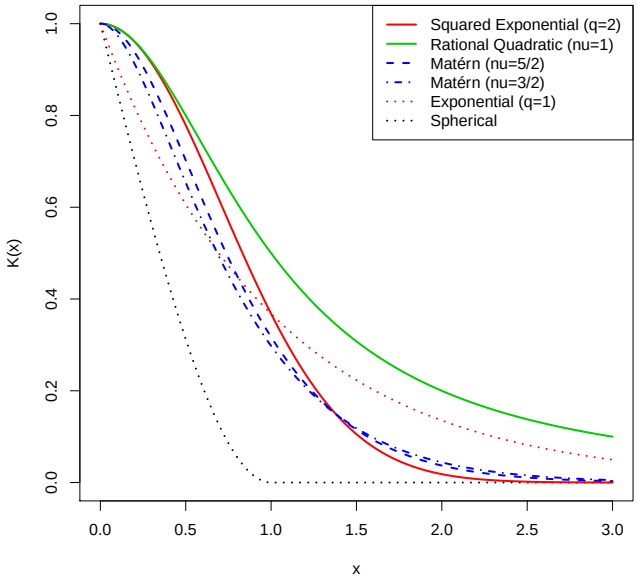


Figure 1.3 – *Correlation kernels with parameter* $\theta = 1$.

in higher-dimensional spaces is a difficult problem [Damblin et al., 2013, Pronzato, 2017]. Other approaches rely on the statistical properties of Gaussian processes in order to minimize global uncertainty [Chevalier et al., 2014, Bect et al., 2016]. In some applications, a spherical symmetry may be observed and Kriging methods would need to take this information into account [Padonou and Roustant, 2016].

Because of its versatility, Kriging can be used in varied industrial problems. The one which motivated this thesis is the computation of Functional Risk Curves. Although by no means the only possible solution, the Kriging framework is particularly well suited because it provides ready-to-use uncertainty quantifying objects like prediction intervals [Iooss and Le Gratiet, 2017, Le Gratiet et al., 2017].

Despite its original geostatistical purpose of quantifying quantities in 3-dimensional spaces, it is part of the solution to ever more complex problems. Co-Kriging, i.e. Kriging with multidimensional output, has been successfully used [Le Gratiet and Garnier, 2014] to emulate hierarchical multi-fidelity codes [Le Gratiet, 2013]. In a different context, methods have been developed to deal with nested codes, where the output of one is the input of the next [Perrin et al., 2017]. And when the number of data is large, small Kriging models can be aggregated [Rullière et al., 2018].

Qualitative limits of Kriging are also being pushed. Roustant et al. [2018] consider Kriging-based methods for categorical inputs, and Bachoc et al. [2017b] propose a theory for Kriging of distributional rather than numerical inputs.

Chapter 2

Reference Prior Theory

Abstract

Reference analysis is presented with inferential and predictive goals in mind (hypothesis testing is not considered). Formal definitions of reference priors are given for single-parameter and multi-parameter models. Reference priors are linked to both Laplace's insufficient reason argument and Jeffrey's rule for prior elicitation. Fundamental properties of reference priors – independence of sample size, compatibility with sufficient statistics, invariance under reparametrization – are highlighted. Although primarily a review on reference analysis in the vein of Bernardo [2005] and Berger et al. [2009], the chapter contributes a uniqueness result under some assumptions for single-parameter reference priors. Examples of single-parameter and multi-parameter models and associated reference priors are given.

Résumé

L'analyse bayésienne fondée sur les priors de référence est présentée dans une optique inférentielle et prédictive (le test d'hypothèses n'est pas traité). Les priors de référence pour modèles à un ou plusieurs paramètres sont formellement définis. La théorie est reliée à la fois au principe de raison insuffisante de Laplace et à la règle de Jeffreys pour éliciter des priors. Les propriétés fondamentales des priors de référence – indépendance vis-à-vis de la taille d'échantillon, compatibilité avec les statistiques exhaustives, invariance par reparamétrisation – sont mises en relief. Quoique se voulant avant tout un résumé de la théorie dans les pas de Bernardo [2005] et Berger et al. [2009], ce chapitre apporte une contribution sous la forme d'un résultat d'unicité sous conditions du prior de référence pour les modèles à un seul paramètre. Des exemples de modèles à un ou plusieurs paramètres sont fournis avec les priors de référence qui y sont associés.

2.1 Introduction

This chapter aims to provide the basics necessary to gain a working understanding of reference prior theory. It draws mainly on Bernardo [2005] and Berger et al. [2009]. Reference prior theory is an attempt to create "objective" priors. Such priors should be sensible defaults in cases where no prior information is available.

The original noninformative prior was the uniform prior. For finite states problems, it derives from the insufficient reason principle ascribed to Laplace [Gelman et al., 2013]. Its extension

to continuous state spaces in not defendable, however, because it leads to widely different posterior distributions for different parametrizations. Subsequent approaches to the problems have relied on invariance arguments (see Hartigan [1964], Jaynes [1968], Dawid [1983]), resulting in left and right Haar measures [Nachbin, 1965]. Jaynes based the notion of noninformativity on Shannon entropy [Jaynes, 1982]. Despite great successes in many problems like spectral analysis and image processing, this approach was shown to lead to paradoxical results [Seidenfeld, 1986]. There have also been efforts to find priors that make Bayesian credible intervals fit frequentist confidence intervals, starting with Lindley [1958] and Welch and Peers [1963]. For a review of efforts to find noninformative priors, see Kass and Wasserman [1996] or [Ghosh, 2011]. The most fruitful approach however was the Jeffreys-rule prior [Jeffreys, 1961], and many methods end up justifying its use in several cases. As we will see, the reference prior method is one of them.

Because of the vague nature of the notion of noninformative prior – each method is a formalization of this idea – it is the subject of an ongoing controversy. We do not discuss it in this chapter, but the interested reader may for example read Fienberg [2006] and Berger [2006], or more recently Seaman III et al. [2012] and Kamary and Robert [2014].

Apart from inference and prediction, hypothesis testing is yet another statistical problem where noninformative priors are desired. Although no such problem is considered in this thesis, it is a very active research field. See Bernardo [2011] for solutions based on reference prior theory, and Kamary et al. [2014] for a general modern perspective.
This chapter details the intuition behind the formal definition of the reference prior in order to explain both its strengths and shortcomings. The formal definition is not given straight away, because it is quite complex and is of little help if the ideas behind it are not well understood. Instead, it is presented as the end result of a development process where "naive" conceptions are put forward, criticized, and then corrected.

Even though intuition is emphasized, formal proofs are provided for all results except Theorem 2.16, for which heuristic arguments are provided. The reader is referred to Clarke and Barron [1994] for formal arguments in this instance.

2.2 Basic idea

Our main tool for defining the reference prior is the Kullback-Leibler divergence.

Definition 2.1 (Kullback-Leibler divergence)**.** *Let P and Q be two probability measures absolutely continuous with respect to a measure μ. Let p and q be their respective Radon-Nikodym derivatives with respect to μ. The Kullback-Leibler divergence $D(P||Q)$ of Q with respect to P is defined by*

$$D(P||Q) = \int p \log\left(\frac{p}{q}\right) d\mu. \tag{2.1}$$

The Kullback-Leibler divergence has several useful properties [Lindley, 1956]. Among them:

- It does not depend on the dominating measure μ.
- Jensen's inequality can be used to show that it is nonnegative.
- It is equal to 0 if, and only if, $Q = P$.

— $\int p \left|\log\left(\frac{p}{q}\right)\right| d\mu \leqslant D(P||Q) + 1$.

The Kullback-Leibler divergence is usually interpreted as a way to quantify how different Q is from P. The smaller $D(P||Q)$ is, the better Q approximates P. In the discrete case, it is axiomatically justified in the context of Shannon's information theory [Shannon, 1948, Lee, 1964]. Good [1966] gives a probabilistic interpretation of information. Bernardo [1979b] discusses it from a decision-theoretic standpoint.

The Kullback-Leibler divergence is not symmetric, so it does not qualify as a distance between probability distributions.

This interpretation can be used to determine what an objective prior distribution should be. Let $\mathcal{Y}$ be the observation space and endow it with a σ-algebra $\mathbb{Y}$. Let Θ be the parameter space. A model is a collection of probability distributions $(P_\theta)_{\theta\in\Theta}$. In a Bayesian setting, it is coupled with a prior distribution Π on the parameter space Θ. For ease of use, we assume that Θ is a metric space (its distance is denoted by $dist$) endowed with its Borel σ-algebra. Every prior distribution Π is a probability distribution on this measurable space. The predictive distribution based on prior knowledge [Robert, 2007] is

$$P_\Pi := \int_\Theta P_\theta d\Pi(\theta). \tag{2.2}$$

This is to say that for any measurable subset T of $\mathcal{Y}$,

$$P_\Pi(T) = \int_\Theta P_\theta(T) d\Pi(\theta). \tag{2.3}$$

Naturally, this definition assumes that for any measurable set $A \in \mathbb{Y}$, the mapping $\Theta \to [0,1]; \theta \mapsto P_\theta(A)$ is measurable. In other words, the mapping $\Theta \times \mathbb{Y}; (\theta, A) \mapsto P_\theta(A)$ is assumed to be a Markov kernel. This assumption is technical and has no practical relevance. Typically, $P_\theta(A)$ is continuous as a function of θ, or else Θ is a countable set, or some other reason makes this requirement hold.

If an observation $y \in \mathcal{Y}$ from the model $(P_\theta)_{\theta\in\Theta}$ (with unknown θ) has been made, let $\Pi(\cdot|y)$ be the posterior distribution resulting from the prior Π.
The posterior distribution [Robert, 2007] is defined as a Markov kernel $\mathcal{Y} \times \mathcal{B}(\Theta) : (y, B) \mapsto \Pi(B|y)$ such that for any measurable set $A \subset \mathcal{Y}$ and any measurable set $B \subset \Theta$

$$\int_A \Pi(B|y) dP_\Pi(y) = \int_B P_\theta(A) d\Pi(\theta). \tag{2.4}$$

Remark. If the model $(P_\theta)_{\theta\in\Theta}$ is dominated by some measure $\mu_\mathcal{Y}$ (i.e. if for any $\theta \in \Theta$, P_θ is absolutely continuous with respect to $\mu_\mathcal{Y}$ – in other words, if the model admits a likelihood function), then let p_θ be the density of P_θ and p_Π be the density of P_Π (for $\mu_\mathcal{Y}$-almost any y, $p_\Pi(y) = \int_{\theta\in\Theta} p_\theta(y) d\Pi(\theta)$). Then, for $\mu_\mathcal{Y}$-almost any y, the posterior distribution $\Pi(\cdot|y)$ is defined as having density $\theta \mapsto p_\theta(y)/p_\Pi(y)$ with respect to the prior Π.

Intuitively, a prior Π is noninformative if it plays a small role in the formation the posterior $\Pi(\cdot|y)$. This way, the amount of information conveyed by the data will be maximized. In practice, we use the Kullback-Leibler divergence $D(\Pi(\cdot|y) \,||\, \Pi)$, which says how badly Π approximates $\Pi(\cdot|y)$ to measure this [Berger et al., 2009]. Given that y is actually unknown at the time the prior is defined, $D(\Pi(\cdot|y) \,||\, \Pi)$ should be maximized on average over $y \in \mathcal{Y}$. We therefore want to maximize over all priors Π the quantity $\int_\mathcal{Y} dP_\Pi(y) D(\Pi(\cdot|y)||\Pi)$.

If the model is dominated (cf. previous remark), then

$$\begin{aligned}\int_{\mathcal{Y}} dP_\Pi(y) D(\Pi(\cdot|y)||\Pi) &= \int_{\mathcal{Y}} dP_\Pi(y) \int_\Theta d\Pi(\theta|y) \log\left(\frac{d\Pi(\cdot|y)}{d\Pi}(\theta)\right) \\ &= \int_{\mathcal{Y}} dP_\Pi(y) \int_\Theta d\Pi(\theta|y) \log \frac{p_\theta(y)}{p_\Pi(y)}. \end{aligned} \tag{2.5}$$

If this quantity is finite, then the Fubini-Lebesgue theorem is applicable, because

$$\int_{\mathcal{Y}} dP_\Pi(y) \int_\Theta d\Pi(\theta|y) \left|\log \frac{p_\theta(y)}{p_\Pi(y)}\right| \leqslant \int_{\mathcal{Y}} dP_\Pi(y) \int_\Theta d\Pi(\theta|y) \log \frac{p_\theta(y)}{p_\Pi(y)} + 1. \tag{2.6}$$

Therefore, if $\int_{\mathcal{Y}} dP_\Pi(y) D(\Pi(\cdot|y)||\Pi) < +\infty$, then the two integral signs can be switched.

$$\int_{\mathcal{Y}} dP_\Pi(y) D(\Pi(\cdot|y)||\Pi) = \int_\Theta d\Pi(\theta) \int_{\mathcal{Y}} p_\theta(y) \log \frac{p_\theta(y)}{p_\Pi(y)} d\mu_{\mathcal{Y}}(y) = \int_\Theta d\Pi(\theta) D(P_\theta||P_\Pi). \tag{2.7}$$

The same reasoning could be applied in reverse. So either both $\int_{\mathcal{Y}} dP_\Pi(y) D(\Pi(\cdot|y)||\Pi)$ and $\int_\Theta d\Pi(\theta) D(P_\theta||P_\Pi)$ are finite, in which case they are equal, or both are infinite. In all cases, they are equal.

This identity, which holds for all dominated models $(P_\theta)_{\theta\in\Theta}$, leads to a new intuition. A prior Π is noninformative if there is a great difference between knowing it and knowing the true value θ, that is, if P_Π is generally a poor substitute to P_θ. So $D(P_\theta||P_\Pi)$ should be large on average over Θ.

Of course, "on average over Θ" means in practice "on average over the prior distribution". Averaging over the prior whose noninformativity is being evaluated makes the noninformativity criterion somewhat self-referential. One would perhaps wish to average over some other measure, but which one? Note that the original idea of maximizing $\int_{\mathcal{Y}} dP_\Pi(y) D(\Pi(\cdot|y)||\Pi)$ runs into the same problem. In that formulation, the average is being made over the predictive distribution, but this distribution still depends on the prior! The formulation (2.7) can however be defended from a game-theoretic standpoint [Clarke and Barron, 1994]. Imagine a game between two players, Nature and the Statistician. Nature first picks the parameter θ randomly using a prior distribution Π. Then, using the picked θ, Nature draws y randomly using the distribution P_θ. Then, knowing only the prior distribution Π, the Statistician attempts to guess y. For Nature, the optimal play is to maximize (2.7), because on average, this is what makes the Statistician's job most difficult.

Example 1. *Let $\Theta = \{1, 2\}$, P_1 be the uniform distribution on $[0, 1]$ and P_2 be the uniform distribution on $[0, 2]$. Let Π be a prior distribution and let $x := \Pi(1)$ (so that $\Pi(2) = 1 - x$). Then*

$$\int_\Theta D(P_\theta||P_\Pi) d\Pi(\theta) = x \log\left(\frac{2}{1+x}\right) + \frac{1-x}{2} \log\left(\frac{1}{1-x^2}\right), \tag{2.8}$$

which reaches its maximum at $x = 3/5$.

A problem with this definition is that P_θ can be changed simply by taking several independent observations instead of just one. Such a change cannot be said to alter the model, so one would expect the optimum not to change. Unfortunately, it does.

For any distribution Q, let $Q^{\otimes n}$ ($n \in \mathbb{N}$) be the distribution of a sample of n independent draws from Q. And let $P_{\Pi,n}$ be the distribution on $\mathcal{Y}^n$ (with σ-algebra $\mathbb{Y}^{\otimes n}$) such that for all $T_1, ..., T_n \in \mathcal{Y}$,

$$P_{\Pi,n}(\times_{i=1}^n T_i) = \int_\Theta \prod_{i=1}^n P_\theta(T_i) d\Pi(\theta). \tag{2.9}$$

Example 2 (Previous example–continued)**.**

$$\int_\Theta D(P_\theta^{\otimes 2}||P_{\Pi,2}) d\Pi(\theta) = x \log\left(\frac{4}{1+3x}\right) + (1-x)\left[\frac{1}{4}\log\left(\frac{1}{1+3x}\right) + \frac{3}{4}\log\left(\frac{1}{1-x}\right)\right], \tag{2.10}$$

which reaches its maximum at $x = 1/283(247 - 128\,2^{1/3} + 48\,2^{2/3}) \approx 0.57 < 3/5$.

A solution to this problem is to consider the asymptotic case ($n \to +\infty$).

Let $\mathcal{P}(\Theta)$ be the set all proper prior distributions supported on Θ, that is all probability distributions supported on Θ for which a posterior distribution is well defined. In the case of dominated models, all probability distributions supported on Θ fit this requirement.

We give a first definition of a reference prior, that is a prior which maximizes a noninformativity criterion and can therefore be used as a default prior when little prior information is available.

Definition 2.2 (Naive definition of a reference prior)**.** *A reference prior* $\Pi^\star$ *is an element of* $\mathcal{P}(\Theta)$ *such that*

$$\lim_{n\to+\infty} \int_\Theta D(P_\theta^{\otimes n}||P_{\Pi^\star,n}) d\Pi^\star(\theta) = \lim_{n\to+\infty} \sup_{\Pi\in\mathcal{P}(\Theta)} \int_\Theta D(P_\theta^{\otimes n}||P_{\Pi,n}) d\Pi(\theta) < +\infty. \tag{2.11}$$

Remark. The definition assumes that both limits exist and are finite. As will be seen, this assumption is too strong for practical purposes [Berger et al., 2009]. Nevertheless, there are cases where this definition works [Bernardo, 2005].

Proposition 2.3. *If* Θ *is a finite set and its topology is the set of all its parts, and if for all* $\theta_1, \theta_2 \in \Theta$, $P_{\theta_1} \neq P_{\theta_2}$, *then the reference prior is given by*

$$\forall \theta \in \Theta, \quad \Pi^\star(\theta) = \frac{1}{\#\Theta}. \tag{2.12}$$

Proof. For any prior Π and any $\theta \in \Theta$ such that $\Pi(\theta) > 0$,

$$\begin{aligned} D(P_\theta^{\otimes n}||P_{\Pi,n}) &= \int_{\mathcal{Y}^n} \log\left(\frac{dP_\theta^{\otimes n}(y)}{\sum_{\theta'\in\Theta}\Pi(\theta')dP_{\theta'}^{\otimes n}(y)}\right) dP_\theta^{\otimes n}(y) \\ &= -\int_{\mathcal{Y}^n} \log\left(\Pi(\theta) + \sum_{\theta'\neq\theta}\Pi(\theta')\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)\right) dP_\theta^{\otimes n}(y) \\ &= -\log(\Pi(\theta)) - \int_{\mathcal{Y}^n} \log\left(1 + \sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)\right) dP_\theta^{\otimes n}(y). \end{aligned} \tag{2.13}$$

Now, since Θ is finite and for all $\theta_1, \theta_2 \in \Theta$, $P_{\theta_1} \neq P_{\theta_2}$, there exists a weakly consistent estimator $\hat{\theta}_n$ for the model.

$$
\begin{aligned}
&\int_{\mathcal{Y}^n} \log\left(1+\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)\right) dP_\theta^{\otimes n}(y)\\
&= \int_{\hat{\theta}_n=\theta} \log\left(1+\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)\right) dP_\theta^{\otimes n}(y)\\
&\quad + \frac{P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)}{P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)}\int_{\hat{\theta}_n\neq\theta} \log\left(1+\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)\right) dP_\theta^{\otimes n}(y)\\
&\leqslant \int_{\hat{\theta}_n=\theta}\left(\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)\right) dP_\theta^{\otimes n}(y)\\
&\quad + P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)\log\left(1+\frac{1}{P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)}\int_{\hat{\theta}_n\neq\theta}\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{dP_{\theta'}^{\otimes n}}{dP_\theta^{\otimes n}}(y)dP_\theta^{\otimes n}(y)\right)\\
&= \sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}P_{\theta'}^{\otimes n}(\hat{\theta}_n=\theta) + P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)\log\left(1+\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{P_{\theta'}^{\otimes n}(\hat{\theta}_n\neq\theta)}{P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)}\right)\\
&\leqslant \sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}P_{\theta'}^{\otimes n}(\hat{\theta}_n=\theta) + P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)\log\left(1+\sum_{\theta'\neq\theta}\frac{\Pi(\theta')}{\Pi(\theta)}\frac{1}{P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)}\right) \qquad (2.14)
\end{aligned}
$$

The first equality holds unless $P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)=0$, but Equation (2.14) trivially holds in that case. The first inequality results from applying the Jensen inequality to the convex mapping $\log(1+x)$ and from the inequality $\log(1+x)\leqslant x$, which holds for all $x\geqslant 0$.

Since $\hat{\theta}_n$ is weakly consistent, we have both $\lim_{n\to+\infty} P_{\theta'}^{\otimes n}(\hat{\theta}_n=\theta)=0$ for all $\theta'\neq\theta$ and $\lim_{n\to+\infty} P_\theta^{\otimes n}(\hat{\theta}_n\neq\theta)=0$. Therefore

$$\lim_{n\to+\infty} D(P_\theta^{\otimes n}||P_{\Pi,n}) = -\log(\Pi(\theta)) \qquad (2.15)$$

and

$$\lim_{n\to+\infty}\sum_{\theta\in\Theta} D(P_\theta^{\otimes n}||P_{\Pi,n})\Pi(\theta) = -\sum_{\theta\in\Theta}\log(\Pi(\theta))\Pi(\theta). \qquad (2.16)$$

Because the logarithm is convex, for any prior Π,

$$\sum_{\theta\in\Theta}\log\left(\frac{1}{\Pi(\theta)}\right)\Pi(\theta) \leqslant \log\left(\sum_{\theta\in\Theta}\frac{1}{\Pi(\theta)}\Pi(\theta)\right) = \log(\#\Theta). \qquad (2.17)$$

This inequality becomes an equality if, and only if, $\Pi(\theta)=1/\#\Theta$ for every $\theta\in\Theta$.

□

Remark. The right member of (2.16) is the Shannon entropy of Π.

Interestingly, the reference prior here fits the Laplace [1814] insufficient reason principle, which consists in assuming all elementary events to be equiprobable. This is the first example of the reference prior naturally fitting earlier conceptions of noninformativity.

Although Definition 2.2 intuitively makes sense, it is too restrictive. The limit in (2.11) is often infinite, which means no reference prior exists. Nevertheless, it is sufficient for cases where $\mathcal{Y}$ is a finite set.

2.3 Full definition of the reference prior

The definition of a reference prior below is a more restrictive version of the definition given by Bernardo [2005], because it requires such a prior to be proper.

Definition 2.4 (Reference prior: proper case)**.** *A reference prior* $\Pi^\star$ *is an element of* $\mathcal{P}(\Theta)$ *such that for any other element* Π *of* $\mathcal{P}(\Theta)$*,*

$$\liminf_{n\to+\infty}\left[\int_\Theta D(P_\theta^{\otimes n}||P_{\Pi^\star,n})d\Pi^\star(\theta)-\int_\Theta D(P_\theta^{\otimes n}||P_{\Pi,n})d\Pi(\theta)\right]\geqslant 0. \tag{2.18}$$

Definition 2.4 is an extension of Definition 2.2. Any reference prior in the sense of Definition 2.2 is one in the sense of Definition 2.4.

Remark. If $(P_\theta)_{\theta\in\Theta}$ is a dominated model, the following alternate definition is equivalent: a reference prior $\Pi^\star$ is an element of $\mathcal{P}(\Theta)$ such that for any other element Π of $\mathcal{P}(\Theta)$,

$$\liminf_{n\to+\infty}\left[\int_{\mathcal{Y}^n} D(\Pi^\star(\cdot|y)||\Pi^\star)dP_{\Pi^\star,n}(y)-\int_{\mathcal{Y}^n} D(\Pi(\cdot|y)||\Pi)dP_{\Pi,n}(y)\right]\geqslant 0. \tag{2.19}$$

This definition is more in line with the first intuition presented in this chapter, which can now be refined. When $n\to+\infty$, the posterior distribution approaches perfect knowledge of the parameter. The Kullback-Leibler divergence between the posterior and prior can thus be seen as the difference between (almost) perfect knowledge and knowledge due to the prior. It is the knowledge "missing" from the prior [Berger et al., 2009].

One question that naturally arises is whether the reference prior, in case it exists, is unique. This part of the discussion is to our knowledge new, in the sense that uniqueness does not seem to have been studied in a general setting before. Bernardo [2005] and Berger et al. [2009] do not care about it, which is understandable, since two reference priors would necessarily be equally noninformative, so any of them could be chosen.

Lemma 2.5. *Let* $t\in(0,1)$ *and let* Π_t*,* Π_0 *and* Π_1 *be proper priors such that* $\Pi_t=(1-t)\Pi_0+t\Pi_1$*. For every positive integer* n*,*

$$\int_\Theta d\Pi_1(\theta)D(P_\theta^{\otimes n}||P_{\Pi_t,n})-\int_\Theta d\Pi_1(\theta)D(P_\theta^{\otimes n}||P_{\Pi_1,n})=D(P_{\Pi_1,n}||P_{\Pi_t,n})\leqslant-\log(t). \tag{2.20}$$

Proof. The inequality is due to the fact that $P_{\Pi_t,n}=(1-t)P_{\Pi_0,n}+tP_{\Pi_1,n}$. So Π_1-almost surely, $dP_{\Pi_1,n}/dP_{\Pi_t,n}\leqslant 1/t$, so

$$D(P_{\Pi_1,n}||P_{\Pi_t,n})=\int_{\mathcal{Y}^n} dP_{\Pi_1,n}(y)\log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)\leqslant\log\left(\frac{1}{t}\right). \tag{2.21}$$

To obtain the equality, notice that

$$\begin{aligned}
&\int_\Theta d\Pi_1(\theta)D(P_\theta^{\otimes n}||P_{\Pi_t,n})\\
&=\int_\Theta d\Pi_1(\theta)\int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(\boldsymbol{y})\log\left(\frac{dP_\theta^{\otimes n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)\\
&=\int_\Theta d\Pi_1(\theta)\int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(\boldsymbol{y})\log\left(\frac{dP_\theta^{\otimes n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right)+\int_\Theta d\Pi_1(\theta)\int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(\boldsymbol{y})\log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)\\
&=\int_\Theta d\Pi_1(\theta)D(P_\theta^{\otimes n}||P_{\Pi_1,n})+D(P_{\Pi_1,n}||P_{\Pi_t,n}).
\end{aligned} \tag{2.22}$$

The last step holds due to the Fubini-Lebesgue theorem, which can be used because

$$\begin{aligned}\int_\Theta d\Pi_1(\theta)\int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(\boldsymbol{y})\left|\log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)\right| &\leqslant \int_\Theta d\Pi_1(\theta)\int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(\boldsymbol{y})\log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)+1\\ &\leqslant \int_\Theta d\Pi_1(\theta)\int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(\boldsymbol{y})\log\left(\frac{1}{t}\right)+1\\ &= -\log(t)+1<+\infty. \end{aligned} \tag{2.23}$$

□

Lemma 2.6. *For any $t\in(0,1)$, for any proper priors Π_t, Π_0 and Π_1 such that $\Pi_t=(1-t)\Pi_0+t\Pi_1$,*

$$\begin{aligned}\int_\Theta d\Pi_t(\theta)D(P_\theta^{\otimes n}||P_{\Pi_t,n}) = (1-t)\int_\Theta d\Pi_0(\theta)D(P_\theta^{\otimes n}||P_{\Pi_0,n})+t\int_\Theta d\Pi_1(\theta)D(P_\theta^{\otimes n}||P_{\Pi_1,n})\\ +(1-t)D(P_{\Pi_0,n}||P_{\Pi_t,n})+tD(P_{\Pi_1,n}||P_{\Pi_t,n}).\end{aligned} \tag{2.24}$$

Proof. The result follows from this decomposition:

$$\int_\Theta d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi_t,n}) = (1-t)\int_\Theta d\Pi_0(\theta)D(P_\theta^{\otimes n}||P_{\Pi_t,n})+t\int_\Theta d\Pi_1(\theta)D(P_\theta^{\otimes n}||P_{\Pi_t,n}). \tag{2.25}$$

Applying Lemma 2.5 to both Π_0 and Π_1 yields the result.

□

Lemma 2.7. *For any $t\in(0,1)$, for any proper priors Π_t, Π_0 and Π_1 such that $\Pi_t=(1-t)\Pi_0+t\Pi_1$, the sequences $(D(P_{\Pi_0,n}||P_{\Pi_t,n}))_{n\in\mathbb{Z}_+}$ and $(D(P_{\Pi_1,n}||P_{\Pi_t,n}))_{n\in\mathbb{Z}_+}$ are nondecreasing.*

Proof. We only prove that $(D(P_{\Pi_1,n}||P_{\Pi_t,n}))_{n\in\mathbb{Z}_+}$ is nondecreasing because the proof that $(D(P_{\Pi_0,n}||P_{\Pi_t,n}))_{n\in\mathbb{Z}_+}$ is nondecreasing is similar.

Recall that for every positive integer n, $P_{\Pi_t,n}=(1-t)P_{\Pi_0,n}+tP_{\Pi_1,n}$ and therefore $P_{\Pi_0,n}$-almost surely, $dP_{\Pi_1,n}/dP_{\Pi_t,n}\leqslant 1/t$.

$$\begin{aligned}D(P_{\Pi_1,n}||P_{\Pi_t,n}) &= \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y})\log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)\\ &\leqslant \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y})\log\left(\frac{1}{1-t}\right)=-\log(t)<+\infty.\end{aligned} \tag{2.26}$$

Therefore

$$\int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y})\left|\log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right)\right|\leqslant -\log(t)+1<+\infty. \tag{2.27}$$

Now, define $\Pi_{0,n}(\cdot|\boldsymbol{y})$ (resp. $\Pi_{1,n}(\cdot|\boldsymbol{y})$, $\Pi_{t,n}(\cdot|\boldsymbol{y})$) as the posterior distribution resulting from the prior Π_1 (resp. Π_0, Π_t) after having made n independent observations $(\boldsymbol{y})$. Then define the "conditional" predictive distribution $P^{\boldsymbol{y}}_{\Pi_1|n}=\int_\Theta d\Pi_{1,n}(\theta|\boldsymbol{y})P_\theta$ (resp. $P^{\boldsymbol{y}}_{\Pi_0|n}=\int_\Theta d\Pi_{0,n}(\theta|\boldsymbol{y})P_\theta$, $P^{\boldsymbol{y}}_{\Pi_t|n}=\int_\Theta d\Pi_{t,n}(\theta|\boldsymbol{y})P_\theta$). Applying the Fubini-Lebesgue theorem,

$$\begin{aligned}
D(P_{\Pi_1,n+1}||P_{\Pi_t,n+1}) &= \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \int_{\mathcal{Y}} dP^{\boldsymbol{y}}_{\Pi_1|n}(y') \log\left(\frac{dP_{\Pi_1,n+1}}{dP_{\Pi_t,n+1}}(\boldsymbol{y},y')\right) \\
&= \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \int_{\mathcal{Y}} dP^{\boldsymbol{y}}_{\Pi_1|n}(y') \log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\frac{dP^{\boldsymbol{y}}_{\Pi_1|n}}{dP^{\boldsymbol{y}}_{\Pi_t|n}}(y')\right) \\
&= \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(\frac{dP_{\Pi_1,n}}{dP_{\Pi_t,n}}(\boldsymbol{y})\right) \\
&\quad + \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \int_{\mathcal{Y}} dP^{\boldsymbol{y}}_{\Pi_1|n}(y') \log\left(\frac{dP^{\boldsymbol{y}}_{\Pi_1|n}}{dP^{\boldsymbol{y}}_{\Pi_t|n}}(y')\right) \\
&= D(P_{\Pi_1,n}||P_{\Pi_t,n}) + \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) D(P^{\boldsymbol{y}}_{\Pi_1|n}||P^{\boldsymbol{y}}_{\Pi_t|n}) \\
&\geqslant D(P_{\Pi_1,n}||P_{\Pi_t,n}). \qquad (2.28)
\end{aligned}$$

The second equality holds because $P_{\Pi_t,n+1} = (1-t)P_{\Pi_1,n+1} + tP_{\Pi_1,n+1}$ and therefore

— $P_{\Pi_1,n}$-almost surely, $dP_{\Pi_1,n}/dP_{\Pi_t,n} \leqslant 1/t$;

— for $\Pi_{1,n}$-almost any $\boldsymbol{y} \in \mathcal{Y}^n$, $P^{\boldsymbol{y}}_{\Pi_1|n}$-almost surely, $dP^{\boldsymbol{y}}_{\Pi_1|n}/dP^{\boldsymbol{y}}_{\Pi_t|n} \leqslant 1/t$.

□

Lemma 2.8. *For any proper prior Π, the mapping $i_\Pi : \mathcal{P}(\Theta) \to [0,+\infty]$ defined by*

$$i_\Pi(\Pi') := \liminf_{n\to+\infty} \left[\int_\Theta D(P_\theta^{\otimes n}||P_{\Pi',n}) d\Pi'(\theta) - \int_\Theta D(P_\theta^{\otimes n}||P_{\Pi,n}) d\Pi(\theta)\right] \qquad (2.29)$$

is essentially strictly convex in the following sense. For all $\Pi_0, \Pi_1 \in \mathcal{P}(\Theta)$ such that $i_\Pi(\Pi_0)$ and $i_\Pi(\Pi_1)$ are finite,

— *either for all $t \in (0,1)$, $i_\Pi((1-t)\Pi_0 + t\Pi_1) > (1-t)i_\Pi(\Pi_0) + ti_\Pi(\Pi_1)$;*

— *or for all $n \in \mathbb{Z}_+$, $P_{\Pi_0,n} = P_{\Pi_1,n}$.*

Proof. Define $\Pi_t := (1-t)\Pi_0 + t\Pi_1$. Lemmas 2.6 and 2.7 imply that for all $m \in \mathbb{Z}_+$,

$$\begin{aligned}
&\liminf_{n\to\infty} \left[\int_\Theta d\Pi_t(\theta) D(P_\theta^{\otimes n}||P_{\Pi_t,n}) \right. \\
&\qquad \left. -(1-t)\int_\Theta d\Pi_0(\theta) D(P_\theta^{\otimes n}||P_{\Pi_0,n}) - t\int_\Theta d\Pi_1(\theta) D(P_\theta^{\otimes n}||P_{\Pi_1,n})\right] \\
&\geqslant (1-t)D(P_{\Pi_0,m}||P_{\Pi_t,m}) + tD(P_{\Pi_1,m}||P_{\Pi_t,m}). \qquad (2.30)
\end{aligned}$$

For any proper prior Π such that $i_\Pi(\Pi_0)$ and $i_\Pi(\Pi_1)$ are finite,

$$\begin{aligned}
&\liminf_{n\to\infty} \int_\Theta d\Pi_t(\theta) D(P_\theta^{\otimes n}||P_{\Pi_t,n}) \\
&\qquad - (1-t)\int_\Theta d\Pi_0(\theta) D(P_\theta^{\otimes n}||P_{\Pi_0,n}) - t\int_\Theta d\Pi_1(\theta) D(P_\theta^{\otimes n}||P_{\Pi_1,n}) \\
&= \liminf_{n\to\infty} \left[\int_\Theta d\Pi_t(\theta) D(P_\theta^{\otimes n}||P_{\Pi_t,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n}) \right. \\
&\qquad -(1-t)\left(\int_\Theta d\Pi_0(\theta) D(P_\theta^{\otimes n}||P_{\Pi_0,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right) \\
&\qquad \left. -t\left(\int_\Theta d\Pi_1(\theta) D(P_\theta^{\otimes n}||P_{\Pi_1,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right)\right] \qquad (2.31)
\end{aligned}$$

$$
\begin{aligned}
&\liminf_{n\to\infty} \int_\Theta d\Pi_t(\theta) D(P_\theta^{\otimes n}||P_{\Pi_t,n}) \\
&\qquad - (1-t)\int_\Theta d\Pi_0(\theta) D(P_\theta^{\otimes n}||P_{\Pi_0,n}) - t\int_\Theta d\Pi_1(\theta) D(P_\theta^{\otimes n}||P_{\Pi_1,n}) \\
&\leqslant \liminf_{n\to\infty} \left[\int_\Theta d\Pi_t(\theta) D(P_\theta^{\otimes n}||P_{\Pi_t,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right] \\
&\qquad + \limsup_{n\to+\infty} \left[-(1-t)\left(\int_\Theta d\Pi_0(\theta) D(P_\theta^{\otimes n}||P_{\Pi_0,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right)\right] \\
&\qquad + \limsup_{n\to+\infty} \left[-t\left(\int_\Theta d\Pi_1(\theta) D(P_\theta^{\otimes n}||P_{\Pi_1,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right)\right]\Bigg] \\
&= \liminf_{n\to\infty} \left[\int_\Theta d\Pi_t(\theta) D(P_\theta^{\otimes n}||P_{\Pi_t,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right] \\
&\qquad -(1-t)\liminf_{n\to\infty} \left[\int_\Theta d\Pi_0(\theta) D(P_\theta^{\otimes n}||P_{\Pi_0,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right] \\
&\qquad -t\liminf_{n\to\infty} \left[\int_\Theta d\Pi_1(\theta) D(P_\theta^{\otimes n}||P_{\Pi_1,n}) - \int_\Theta d\Pi(\theta) D(P_\theta^{\otimes n}||P_{\Pi,n})\right] \\
&= i_\Pi(\Pi_t) - (1-t) i_\Pi(\Pi_0) - t i_\Pi(\Pi_1). \qquad (2.32)
\end{aligned}
$$

Combining Equations (2.30) and (2.32), we obtain that for all proper priors Π, Π_0 and Π_1 such that both $i_\Pi(\Pi_0)$ and $i_\Pi(\Pi_1)$ are finite and for every positive integer n,

$$
i_\Pi(\Pi_t) - (1-t) i_\Pi(\Pi_0) - t i_\Pi(\Pi_1) \geqslant (1-t) D(P_{\Pi_0,n}||P_{\Pi_t,n}) + t D(P_{\Pi_1,n}||P_{\Pi_t,n}). \qquad (2.33)
$$

Unless $P_{\Pi_0,n} = P_{\Pi_1,n}$ for every positive integer n, there exists a positive integer m such that

$$
(1-t) D(P_{\Pi_0,m}||P_{\Pi_t,m}) + t D(P_{\Pi_1,m}||P_{\Pi_t,m}) > 0. \qquad (2.34)
$$

Therefore

$$
i_\Pi(\Pi_t) - (1-t) i_\Pi(\Pi_0) - t i_\Pi(\Pi_1) > 0. \qquad (2.35)
$$

□

Let us make the following assumption:

Assumption 1. *For every pair of distinct proper priors Π_0 and Π_1 on Θ, there exists a positive integer n such that $P_{\Pi_0,n} \neq P_{\Pi_1,n}$.*

Proposition 2.9. *Under Assumption 1, if there exists a reference prior in the sense of Definition 2.4, then it is unique.*

Proof. Let $\Pi_0^\star$ and $\Pi_1^\star$ be reference priors. Then $i_{\Pi_0^\star}(\Pi_1^\star) = 0$. Besides, we also trivially have $i_{\Pi_0^\star}(\Pi_0^\star) = 0$. By Lemma 2.8,

— either for all $t \in (0,1)$, $i_{\Pi_0^\star}((1-t)\Pi_0^\star + t\Pi_1^\star) > 0$, which is impossible since it would contradict Definition 2.4;

— or for every positive integer n, $P_{\Pi_0^\star,n} = P_{\Pi_1^\star,n}$.

By Assumption 1, this implies that $\Pi_0^\star = \Pi_1^\star$. □

If the model $(P_\theta)_{\theta\in\Theta}$ is identifiable ($P_\theta = P_{\theta'} \Rightarrow \theta = \theta'$) and under some topological conditions, a de Finetti-type theorem may apply and make Assumption 1 hold (see Ressel [1985]). Sometimes more elementary arguments apply:

Proposition 2.10. *If there exists a weakly consistent estimator of θ, then Assumption 1 holds.*

Proof. For any subset $\Theta' \subset \Theta$ and any $\theta \in \Theta$, define $dist(\theta, \Theta') := \inf\{dist(\theta, \theta') : \theta' \in \Theta'\}$.

Let the sequence of mappings $\hat{\theta}_n : \mathcal{Y}^n \to \Theta$ ($n \in \mathbb{N}$) be a weakly consistent estimator. This means that for any $\theta \in \Theta$, for any $\epsilon > 0$, $\lim_{n\to+\infty} P_\theta^{\otimes n}(dist(\hat{\theta}_n, \theta) < \epsilon) = 1$. It implies that for any open set U, for any $\theta \in U$, $\lim_{n\to+\infty} P_\theta^{\otimes n}(\hat{\theta}_n \in U) = 1$ and that for any $\theta \in \Theta$ such that $dist(\theta, U) > 0$, $\lim_{n\to+\infty} P_\theta^{\otimes n}(\hat{\theta}_n \in U) = 0$.

Let U be an open subset of Θ. For all $\epsilon > 0$, define $U_\epsilon := \{\theta \in U : B(\theta, \epsilon) \subset U\}$, where $B(\theta, \epsilon)$ is the open ball centered around θ and with radius ϵ: $B(\theta, \epsilon) := \{\theta' \in \Theta : dist(\theta, \theta') < \epsilon\}$. U_ϵ is an open, possibly empty, set.

Now, let Π_0 and Π_1 be two proper prior distributions on Θ. Applying the dominated convergence theorem,

$$\Pi_0(U_\epsilon) = \int_{U_\epsilon} d\Pi_0(\theta) \lim_{n\to+\infty} P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon) = \lim_{n\to\infty} \int_{U_\epsilon} d\Pi_0(\theta) P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon). \tag{2.36}$$

For any $n \in \mathbb{N}$,

$$\int_\Theta d\Pi_0(\theta) P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon) \leqslant \Pi_0(U) + \int_{\Theta\setminus U} d\Pi_0(\theta) P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon). \tag{2.37}$$

Again, applying the dominated convergence theorem,

$$\lim_{n\to+\infty} \int_{\Theta\setminus U} d\Pi_0(\theta) P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon) = \int_{\Theta\setminus U} d\Pi_0(\theta) \lim_{n\to+\infty} P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon) = 0. \tag{2.38}$$

Gathering the two equations above,

$$\limsup_{n\to+\infty} \int_\Theta d\Pi_0(\theta) P_\theta^{\otimes n}(\hat{\theta}_n \in U_\epsilon) \leqslant \Pi_0(U). \tag{2.39}$$

Combining this with Equation (2.36),

$$\Pi_0(U_\epsilon) \leqslant \limsup_{n\to+\infty} P_{\Pi_0,n}(\hat{\theta}_n \in U_\epsilon) \leqslant \Pi_0(U). \tag{2.40}$$

Let $(\epsilon_m)_{m\in Z_+}$ be a decreasing sequence of positive real numbers with null limit. For example, $\epsilon_m = 1/m$. Then $U = \bigcup_{m\in\mathbb{Z}_+} U_{\epsilon_m}$. Since $(U_{\epsilon_m})_{m\in\mathbb{Z}_+}$ is an increasing sequence of open (hence measurable) sets, $\Pi_0(U) = \lim_{m\to+\infty} \Pi_0(U_{\epsilon_m})$.

In the end,

$$\Pi_0(U) = \lim_{m\to+\infty} \limsup_{n\to\infty} P_{\Pi_0,n}(\hat{\theta}_n \in U_{\epsilon_m}). \tag{2.41}$$

Similarly, we have

$$\Pi_1(U) = \lim_{m\to+\infty} \limsup_{n\to\infty} P_{\Pi_1,n}(\hat{\theta}_n \in U_{\epsilon_m}). \tag{2.42}$$

So, if for every positive integer n $P_{\Pi_0,n} = P_{\Pi_1,n}$, then for every open set U, $\Pi_0(U) = \Pi_1(U)$. Given a topology is a π-system that generates its Borel σ-algebra, this implies that $\Pi_0 = \Pi_1$.

As an aside, notice that a similar argument yields

$$\Pi_0(U) = \lim_{m\to+\infty} \liminf_{n\to\infty} P_{\Pi_0,n}(\hat{\theta}_n \in U_{\epsilon_m}) \tag{2.43}$$

and

$$\Pi_1(U) = \lim_{m\to+\infty} \liminf_{n\to\infty} P_{\Pi_1,n}(\hat{\theta}_n \in U_{\epsilon_m}). \tag{2.44}$$

□

The above served to prove that under a reasonable assumption – Assumption 1 – a proper reference prior is unique. Definition 2.13 below extends the definition of a reference prior in order to allow for improper reference priors. To ensure a similar uniqueness result holds with this extended definition, some preparatory work is needed.

First, a new assumption is needed.

Assumption 2. *If Π_0 and Π_1 are proper priors on Θ such that there exists an open subset $U \subset \Theta$ with $\Pi_0(U) = 0$ and $\Pi_1(U) = 1$, then there exists a sequence of sets $(T_n)_{n\in\mathbb{Z}_+}$ with the following properties:*

1. *For every positive integer n, $T_n \in \mathbb{Y}^{\otimes n}$.*
2. *$\lim_{n\to\infty} P_{\Pi_0,n}(T_n) = 0$ and $\lim_{n\to\infty} P_{\Pi_1,n}(T_n) = 1$.*

We later show that the existence of a weakly consistent estimator for the model $(P_\theta)_{\theta\in\Theta}$ implies Assumption 2.

Proposition 2.11. *Under Assumption 2, if a proper prior $\Pi^\star$ is a reference prior in the sense of Definition 2.4, then for any open subset U of Θ such that $\Pi^\star(U) > 0$, the renormalized restriction of $\Pi^\star$ to U is a reference prior on U in the sense of Definition 2.4.*

Proof. Let $\Pi^\star$ be a proper reference prior in the sense of Definition 2.4. Let U be an open subset of Θ such that $\Pi^\star(U) > 0$. Define $t := \Pi^\star(U)$.

If $t = 1$, then there is nothing to prove, so assume $t < 1$.

Define $\Pi_0^\star$ as the normalized restriction of $\Pi^\star$ to $\Theta \setminus U$ and $\Pi_1^\star$ as the normalized restriction of $\Pi^\star$ to U. Therefore $\Pi^\star = (1-t)\Pi_0^\star + t\Pi_1^\star$.

Now let Π_0 be any proper prior with support on $\Theta \setminus U$ and Π_1 be any proper prior with support on U. Define $\Pi_t := (1-t)\Pi_0 + t\Pi_1$.

Using Assumption 2, there exists a sequence $(T_n)_{n\in\mathbb{Z}_+}$ that has the properties 1. and 2.

For every positive integer n,

$$\begin{aligned} D(P_{\Pi_1,n}||P_{\Pi_t,n}) &= -\int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(\frac{dP_{\Pi_t,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right) \\ &= -\int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left((1-t)\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y}) + t\right) \\ &= -\log(t) - \int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(1 + \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right). \end{aligned} \tag{2.45}$$

$$\begin{aligned} &\int_{\mathcal{Y}^n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(1 + \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right) \\ =&\int_{T_n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(1 + \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right) + \int_{\mathcal{Y}^n \setminus T_n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(1 + \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right) \\ \leqslant&\int_{T_n} dP_{\Pi_1,n}(\boldsymbol{y}) \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y}) + \int_{\mathcal{Y}^n \setminus T_n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(1 + \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right) \\ =&\frac{1-t}{t} P_{\Pi_0,n}(T_n) + \frac{P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n)}{P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n)} \int_{\mathcal{Y}^n \setminus T_n} dP_{\Pi_1,n}(\boldsymbol{y}) \log\left(1 + \frac{1-t}{t}\frac{dP_{\Pi_0,n}}{dP_{\Pi_1,n}}(\boldsymbol{y})\right) \\ \leqslant&\frac{1-t}{t} P_{\Pi_0,n}(T_n) + P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n) \log\left(1 + \frac{1-t}{t}\frac{P_{\Pi_0,n}(\mathcal{Y}^n \setminus T_n)}{P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n)}\right) \\ \leqslant&\frac{1-t}{t} P_{\Pi_0,n}(T_n) + P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n) \log\left(1 + \frac{1-t}{t}\frac{1}{P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n)}\right). \end{aligned} \tag{2.46}$$

In the computation above, the first inequality holds because for any nonnegative real number x, $\log(1+x) \leqslant x$. The second equality only makes sense if $P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n) \neq 0$, but in case $P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n) = 0$, Equation 2.46 holds a fortiori. The second inequality results from Jensen's inequality applied to the concave logarithm function.

Due to Assumption 2, $\lim_{n\to+\infty} P_{\Pi_0,n}(T_n) = \lim_{n\to+\infty} P_{\Pi_1,n}(\mathcal{Y}^n \setminus T_n) = 0$. Therefore, combining Equations (2.45) and (2.46) yields

$$\lim_{n\to+\infty} D(P_{\Pi_1,n}||P_{\Pi_t,n}) = -\log(t). \tag{2.47}$$

A similar reasoning yields

$$\lim_{n\to+\infty} D(P_{\Pi_0,n}||P_{\Pi_t,n}) = -\log(1-t). \tag{2.48}$$

These two results do not actually depend on Π_0 and Π_1 but only on the fact that $\Pi_0(U) = 0$ and $\Pi_1(U) = 1$. Therefore they also hold for $\Pi_0^\star$ and $\Pi_1^\star$.

Let us take $\Pi_0 = \Pi_0^\star$.

Lemma 2.6, applied to $\Pi^\star$ and Π_t, implies that

$$i_{\Pi_t}(\Pi^\star) = t\, i_{\Pi_1}(\Pi_1^\star). \tag{2.49}$$

Because $\Pi^\star$ is a reference prior, this implies that $i_{\Pi_1}(\Pi_1^\star) \geqslant 0$. Given this holds for any proper prior Π_1 such that $\Pi_1(U) = 1$, $\Pi_1^\star$ is a reference prior on U. □

Intuitively, Assumption 2 states that for radically different proper priors (that do not even have the same support), prediction should tend to become radically different too when the number of observations increases. We only require this for "reasonable" supports, that is

supports that can be expressed in terms of open and closed sets. This is consistent with the general idea that the topology of Θ is important: P_θ and $P_{\theta'}$ should be close if $dist(\theta, \theta')$ is small.

Proposition 2.12. *If there exists a weakly consistent estimator of θ, then Assumption 2 holds.*

Proof. Let the sequence of mappings $\hat{\theta}_n : \mathcal{Y}^n \to \Theta$ $(n \in \mathbb{N})$ be a weakly consistent estimator. This means that for any $\theta \in \Theta$, for any $\epsilon > 0$, $\lim_{n\to+\infty} P_\theta^{\otimes n}(dist(\hat{\theta}_n, \theta) < \epsilon) = 1$.

Let Π_0 and Π_1 be two proper priors on Θ such that there exists an open subset $U \subset \Theta$ with $\Pi_0(U) = 0$ and $\Pi_1(U) = 1$. Notations are the same as in the proof of Proposition 2.10.

Define $N(0) := 0$.
For every integer $m = 1, 2, ...$, do the following:

1. Choose an integer $N_0(m) > N(m-1)$ such that for every integer $n \geqslant N_0(m)$, $P_{\Pi_0,n}(\hat{\theta}_n \in U_{\epsilon_m}) < \limsup_{n'\to\infty} P_{\Pi_0,n'}(\hat{\theta}_{n'} \in U_{1/m}) + 1/m$.
2. Choose an integer $N_1(m) > N(m-1)$ such that for every integer $n \geqslant N_1(m)$, $P_{\Pi_1,n}(\hat{\theta}_n \in U_{\epsilon_m}) > \liminf_{n'\to\infty} P_{\Pi_1,n'}(\hat{\theta}_{n'} \in U_{1/m}) - 1/m$.
3. Define $N(m) := \max(N_0(m), N_1(m))$ and $T_{N(m)}$ as the event $\{\hat{\theta}_{N(m)} \in U_{1/m}\} \in \mathbb{Y}^{\otimes N(m)}$.

For every positive integer n, there exists a nonnegative integer m such that $N(m) \leqslant n < N(m+1)$. Define $T_n := T_{N(m)} \times \mathcal{Y}^{n-N(m)}$. We have:

$$P_{\Pi_0,n}(T_n) = P_{\Pi_0,N(m)}(T_{N(m)}) < \limsup_{n'\to\infty} P_{\Pi_0,n'}(\hat{\theta}_{n'} \in U_{1/m}) + 1/m; \tag{2.50}$$

$$P_{\Pi_1,n}(T_n) = P_{\Pi_1,N(m)}(T_{N(m)}) > \liminf_{n'\to\infty} P_{\Pi_1,n'}(\hat{\theta}_{n'} \in U_{1/m}) - 1/m. \tag{2.51}$$

Following Equations (2.41) and (2.44),

$$0 = \Pi_0(U) = \lim_{m\to+\infty} \limsup_{n\to\infty} P_{\Pi_0,n}(\hat{\theta}_n \in U_{1/m}); \tag{2.52}$$

$$1 = \Pi_1(U) = \lim_{m\to+\infty} \liminf_{n\to\infty} P_{\Pi_1,n}(\hat{\theta}_n \in U_{1/m}). \tag{2.53}$$

Therefore $\lim_{n\to+\infty} P_{\Pi_0,n}(T_n) = 0$ and $\lim_{n\to+\infty} P_{\Pi_1,n}(T_n) = 1$.

□

The requirement that a reference prior should be proper (that is belong to $\mathcal{P}(\Theta)$) is counter-intuitive. On the contrary, one would rather expect it to be improper in some cases since it is supposed to have the least influence possible on inference [Berger and Bernardo, 1992]. The propriety requirement can be lifted in the following way:

Definition 2.13 (Reference prior: final definition)**.** *A reference prior $\Pi^\star$ is a measure such that there exists an increasing sequence of open subsets $(U_n)_{n\in\mathbb{N}}$ $(U_0 \subset U_1 \subset U_2 \subset ...)$ such that $\bigcup_{n\in\mathbb{N}} U_n = \Theta$ and for every nonnegative integer n, $\Pi^\star(U_n) < +\infty$ and which has the following property. For every nonnegative integer n, the probability distribution $\Pi_n^\star$ with support on U_n defined by $\Pi_n^\star(A) = \Pi^\star(A)/\Pi^\star(U_n)$ for every measurable set $A \subset U_n$ is a reference prior on U_n in the sense of Definition 2.4).*

Remark. This definition is slightly different from Definition 6 in Bernardo [2005] in that Bernardo only requires the increasing sequence to consist of measurable sets and not necessarily open sets. This restriction, which is of little practical relevance, is used to obtain uniqueness results.

Proposition 2.14. *Under Assumptions 1 and 2, if a reference prior in the sense of Definition 2.13 exists, it is unique up to a multiplicative constant.*

Proof. Let $\Pi_0^\star$ and $\Pi_1^\star$ be two reference priors in the sense of Definition 2.13. So there exists an increasing sequence of open subsets $(U_m^0)_{m\in\mathbb{N}}$ such that for each nonnegative integer m, the renormalized restriction of $\Pi_0^\star$ to U_m^0 is a reference prior on U_m^0. There also exists an increasing sequence of open subsets $(U_m^1)_{m\in\mathbb{N}}$ which plays the same role with respect to $\Pi_1^\star$.

Let us define for every nonnegative integer m the open subset of Θ $V_m := U_m^0 \cap U_m^1$. Then $(V_m)_{m\in\mathbb{N}}$ is an increasing sequence of open subsets of Θ. Moreover, Assumption 2 and Proposition 2.11 imply that for every nonnegative integer m, both the renormalized restriction of $\Pi_0^\star$ to V_m and the renormalized restriction of $\Pi_1^\star$ to V_m are reference priors on V_m in the sense of Definition 2.4. So Assumption 1 and Proposition 2.9 imply that both renormalized restrictions to V_m are equal. Given this holds for every V_m ($m \in \mathbb{N}$), $\Pi_0^\star$ and $\Pi_1^\star$ are proportional to one another.

□

This ends the original discussion about uniqueness.

2.4 Regular continuous case

In this section, we consider a particular case of great theoretical and practical importance. Everything here apart from Lemma 2.17 and Theorem 2.18 is taken from Clarke and Barron [1994].

Consider the following conditions.

Condition 0. Θ is a nonempty open subset of $\mathbb{R}^d$ (relative to the Euclidean norm $\|\cdot\|$) whose boundary has null d-dimensional Lebesgue measure. For any $\theta \in \Theta$, P_θ be absolutely continuous with respect to a measure μ. Let p_θ be the Radon-Nykodym derivative; for μ-almost any y, the mapping $\theta \mapsto p_\theta(y)$ is twice continuously differentiable over Θ.
Condition 1. There exists $\epsilon > 0$ such that for all $\theta \in \Theta$, there exists $\delta(\theta) > 0$ such that for any integers $j, k \in [\![1, d]\!]$, both

$$f(\theta) := \int_{\mathcal{Y}} \sup_{\theta' \in \Theta : \|\theta' - \theta\| < \delta(\theta)} \left| \frac{\partial^2}{\partial \theta'_j \partial \theta'_k} \log p_{\theta'}(y) \right|^2 p_\theta(y) d\mu(y)$$
$$\text{and } g_\epsilon(\theta) := \int_{\mathcal{Y}} \left| \frac{\partial}{\partial \theta_j} \log p_\theta(y) \right|^{2+\epsilon} p_\theta(y) d\mu(y)$$

are finite. Moreover, f and g_ϵ are continuous functions of $\theta \in \Theta$.

Condition 2. For all $\theta \in \Theta$, for any integers $j, k \in [\![1, d]\!]$,

$$\int_{\mathcal{Y}} \frac{\partial^2}{\partial \theta_j \partial \theta_k} p_\theta(y) d\mu(y) = 0.$$

Condition 3. For all $\theta, \theta' \in \Theta$, $P_\theta \neq P_{\theta'}$.
Under these conditions, the Jeffreys-rule prior is well defined on Θ.

Definition 2.15. *The Jeffreys-rule prior is the prior with density with respect to the Lebesgue measure proportional to the square root of the determinant of the Fisher information matrix:* $|\mathcal{I}(\theta)|^{1/2}$.

Remark. The Jeffreys-rule prior is traditionally viewed as noninformative because of its invariance by reparametrization and its reliance on Fisher information [Robert, 2007]. The Fisher information matrix, which is the variance of the score function and is linked to the curvature of the likelihood can be seen as representing the local discriminating power of the data [Robert et al., 2009].

The following theorem by Clarke and Barron [1994] concerns the asymptotic behavior of the noninformativity criterion and suggests that the Jeffreys-rule prior is a prime candidate for being a reference prior in this setting. This is the second case where asymptotic optimization of the criterion yields a prior that fits an earlier conception of noninformativity.

Theorem 2.16. *Under Conditions 0, 1, 2 and 3, for any compact set $K \subset \Theta$, we have*

$$\lim_{n\to+\infty}\left[\sup_{\Pi\in\mathcal{P}(K)}\int_K d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n}) - \frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right] = \int_K |\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta. \tag{2.54}$$

Moreover, if $\Pi^\star$ is the Jeffreys-rule prior renormalized so that $\Pi^\star(K) = 1$, then

$$\lim_{n\to+\infty}\left[\int_K d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n}) - \frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right] = \int_K |\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta. \tag{2.55}$$

Some additional work is needed to show that the Jeffreys-rule prior is, indeed, a reference prior. For now, let us detail Clarke and Barron [1994]'s heuristic arguments supporting the above theorem. For a proof, please refer to their paper.

Heuristic arguments in support of Theorem 2.16

To simplify matters, assume $\Theta = \mathbb{R}^d$. For all θ and $\theta' \in \mathbb{R}^d$, and for all $y_1, ..., y_n \in \mathcal{Y}$, there exists a point θ_n on the segment $[\theta, \theta']$ such that

$$\log\left(\frac{\prod_{i=1}^n p_{\theta'}(y_i)}{\prod_{i=1}^n p_\theta(y_i)}\right) = (\theta'-\theta)^\top\sum_{i=1}^n \nabla_\theta \log p_\theta(y_i) - \frac{1}{2}n(\theta'-\theta)^\top I_n(\theta_n)(\theta'-\theta), \tag{2.56}$$

where for every θ'', $I_n(\theta'')$ is the matrix with (j,k)-th entry $\frac{1}{n}\sum_{i=1}^n \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log(p_{\theta''}(y_i))$. The above is a Taylor series expansion of $\log(\prod_{i=1}^n p_\theta(y_i))$.

Now, assume that the $y_1, ..., y_n$ are independent realizations of P_θ.
Then, for all θ'', $I_n(\theta'')$ converges almost surely as $n \to \infty$ to the matrix with (j,k)-the entry $-\int \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_{\theta''}(y)p_\theta(y)d\mu(y)$. If θ'' is in the neighborhood of θ, provided continuity theorems are usable, this is an approximation of the quantity $-\int \frac{\partial^2}{\partial\theta_j\partial\theta_k}\log p_\theta(y)p_\theta(y)d\mu(y)$, which is the (j,k)-the entry of the Fisher information matrix.

Let us neglect the error resulting from replacing $I_n(\theta_n)$ by $\mathcal{I}(\theta)$, the Fisher information matrix at θ.

Now let Π be a prior distribution. Let π be its Radon-Nykodym derivative with respect to the Lebesgue measure. Let us assume that π is continuous. Let $p_\Pi^{\otimes n}$ be defined by

$$p_\Pi(y_1, ..., y_n) = \int_{\mathbb{R}^d} \prod_{i=1}^n p_{\theta'}(y_i)\pi(\theta')d\theta'. \tag{2.57}$$

p_Π is the Radon-Nydkodym derivative with respect to μ of P_Π.

$$\begin{aligned}
&\frac{p_\Pi(y_1, ..., y_n)}{\prod_{i=1}^n p_\theta(y_i)} \\
&= \int_{\mathbb{R}^d} \frac{\prod_{i=1}^n p_{\theta'}(y_i)}{\prod_{i=1}^n p_\theta(y_i)}\pi(\theta')d\theta' \\
&= \int_{\mathbb{R}^d} \exp\left((\theta'-\theta)^\top \sum_{i=1}^n \nabla_\theta \log p_\theta(y_i)\right) \exp\left(-\frac{1}{2}n(\theta'-\theta)^\top I_n(\theta_n)(\theta'-\theta)\right)\pi(\theta')d\theta' \\
&\approx \int_{\mathbb{R}^d} \exp\left((\theta'-\theta)^\top \sum_{i=1}^n \nabla_\theta \log p_\theta(y_i)\right) \exp\left(-\frac{1}{2}n(\theta'-\theta)^\top \mathcal{I}(\theta)(\theta'-\theta)\right)\pi(\theta')d\theta'.
\end{aligned} \tag{2.58}$$

Now, $(2\pi)^{-d/2}|n\mathcal{I}(\theta)|^{1/2}\exp\left(-\frac{1}{2}n(\theta'-\theta)^\top\mathcal{I}(\theta)(\theta'-\theta)\right)$ is the density of the Normal distribution with mean θ and variance $n^{-1}\mathcal{I}(\theta)^{-1}$.

Let us define $S_n(\theta) = 1/\sqrt{n}\sum_{i=1}^n \nabla_\theta \log p_\theta(y_i)$. By the Central Limit theorem, if $y_1, ..., y_n$ are sampled according to P_θ, then the distribution of $S_n(\theta)$ is asymptotically Normal with mean $\mathbf{0}$ and variance $\mathcal{I}(\theta)$. So when $n \to \infty$, (2.58) can be approximated by averaging $\exp\left(\sqrt{n}(\theta'-\theta)^\top S_n(\theta)\right)\pi(\theta')$ over smaller and smaller neighborhoods of θ'. Due to π being continuous, we approximate further:

$$\begin{aligned}
&\int_{\mathbb{R}^d} \exp\left(\sqrt{n}(\theta'-\theta)^\top S_n(\theta)\right)\exp\left(-\frac{1}{2}n(\theta'-\theta)^\top\mathcal{I}(\theta)(\theta'-\theta)\right)\pi(\theta')d\theta' \\
\approx &\pi(\theta)\int_{\mathbb{R}^d} \exp\left(\sqrt{n}(\theta'-\theta)^\top S_n(\theta)\right)\exp\left(-\frac{1}{2}n(\theta'-\theta)^\top\mathcal{I}(\theta)(\theta'-\theta)\right)d\theta'.
\end{aligned} \tag{2.59}$$

Gathering all this,

$$\frac{p_\Pi(y_1, ..., y_n)}{\prod_{i=1}^n p_\theta(y_i)} \approx \pi(\theta)(2\pi)^{d/2}|n\mathcal{I}(\theta)|^{-1/2}\exp\left(\frac{1}{2}S_n(\theta)^\top\mathcal{I}(\theta)^{-1}S_n(\theta)\right), \tag{2.60}$$

so

$$\log\left(\frac{\prod_{i=1}^n p_\theta(y_i)}{p_\Pi(y_1, ..., y_n)}\right) \approx -\log(\pi(\theta)) + \frac{d}{2}\log\left(\frac{n}{2\pi}\right) + \frac{1}{2}\log|\mathcal{I}(\theta)| - \frac{1}{2}S_n(\theta)^\top\mathcal{I}(\theta)^{-1}S_n(\theta). \tag{2.61}$$

Given the distribution of $S_n(\theta)$ is asymptotically Normal with mean $\mathbf{0}$ and variance $\mathcal{I}(\theta)$ (provided $y_1, ..., y_n$ are independently sampled from P_θ), we have

$$\underbrace{\int_{\mathcal{Y}} ... \int_{\mathcal{Y}}}_{n\int_{\mathcal{Y}}} S_n(\theta)^\top\mathcal{I}(\theta)^{-1}S_n(\theta)\prod_{i=1}^n p_\theta(y_i)d\mu(y_i) = d. \tag{2.62}$$

Therefore,

$$D(P_\theta^{\otimes n}||P_{\Pi,n}) = \underbrace{\int_{\mathcal{Y}} ... \int_{\mathcal{Y}}}_{n \int_{\mathcal{Y}}} \log\left(\frac{\prod_{i=1}^n p_\theta(y_i)}{p_\Pi(y_1,...,y_n)}\right) \prod_{i=1}^n p_\theta(y_i) d\mu(y_i)$$
$$\approx -\log(\pi(\theta)) + \frac{d}{2}\log\left(\frac{n}{2\pi e}\right) + \frac{1}{2}\log|\mathcal{I}(\theta)|. \tag{2.63}$$

Define $c = \int_{\mathbb{R}^d} |\mathcal{I}(\theta)|^{\frac{1}{2}} d\theta$ and assume c is finite. Then

$$\int_\Theta D(P_\theta^{\otimes n}||P_{\Pi,n})\pi(\theta)d\theta \approx \frac{d}{2}\log\left(\frac{n}{2\pi e}\right) + \log(c) - \int_\Theta \log\left(\frac{\pi(\theta)}{|\mathcal{I}(\theta)|^{\frac{1}{2}}/c}\right)\pi(\theta)d\theta. \tag{2.64}$$

Remark. The Jeffreys-rule prior can be proper or not, and is in fact usually improper. Here, though, by requiring c to be finite, we are assuming it to be proper.

The last term is the opposite of the Kullback-Leibler divergence between Π and the Jeffreys-rule prior, which is proper. To maximize this quantity, Π must be the Jeffreys-rule prior.

Equation (2.64) suggests that all priors Π that have continuous density π with respect to the Lebesgue measure make $\int_\Theta D(P_\theta^{\otimes n}||P_{\Pi,n})\pi(\theta)d\theta$ go to infinity as $n \to +\infty$. Notice that the speed of increase is the same for all continuous priors: $\log(n)$.

The Jeffreys-rule prior as reference prior

Although the Jeffreys-rule prior being a reference prior is well-known [Bernardo, 2005, Berger et al., 2009], because of our unusual definition of a reference prior, we need to make sure that the Jeffreys-rule prior satisfies this definition. The discussion in this part of the section is therefore, to the best of our knowledge, original.

Lemma 2.17 (Corollary of Clarke and Barron's theorem)**.** *Under Conditions 0, 1, 2 and 3, let U be an open set and K a compact set such that $U \subset K \subset \Theta \subset \mathbb{R}^d$. Then for any proper prior Π with support on U,*

$$\limsup_{n\to+\infty}\left[\int_U d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n}) - \frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right] \leqslant \int_U |\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta < +\infty. \tag{2.65}$$

Moreover, if $\Pi^\star$ is the renormalized restriction of the Jeffreys-rule prior to U, then

$$\lim_{n\to+\infty}\left[\int_U d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n}) - \frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right] = \int_U |\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta < +\infty. \tag{2.66}$$

Proof. Let Π be a proper prior with support on U.

If there exists a compact set $K_U \subset U$ such that $\Pi(K_U) = 1$, then Theorem 2.16 is applicable:

$$\begin{aligned}
&\limsup_{n\to+\infty}\left[\int_U d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n}) - \frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right] \\
=&\limsup_{n\to+\infty}\left[\int_{K_U} d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n}) - \frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right] \\
\leqslant&\int_{K_U}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta \leqslant \int_U |\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta.
\end{aligned} \tag{2.67}$$

Else, let K_0 be a compact subset of U such that $\Pi(K_0) > 0$.
Let Π_0 be the renormalized restriction of Π to K_0 (i.e., for any measurable subset A of K_0, $\Pi_0(A) = \Pi(A)/\Pi(K_0)$ and let Π_1 be the renormalized restriction of Π to $U \setminus K_0$.

Lemma 2.6 yields that for every positive integer n,

$$\begin{aligned}
&\int_U d\Pi(\theta) D(P_\theta^{\otimes n} || P_{\Pi,n}) \\
=&\Pi(K_0) \int_{K_0} d\Pi_0(\theta) D(P_\theta^{\otimes n} || P_{\Pi_0,n}) + (1 - \Pi(K_0)) \int_{U \setminus K_0} d\Pi_1(\theta) D(P_\theta^{\otimes n} || P_{\Pi_1,n}) \\
&+ \Pi(K_0) D(P_{\Pi_0,n} || P_{\Pi,n}) + (1 - \Pi(K_0)) D(P_{\Pi_1,n} || P_{\Pi,n}).
\end{aligned} \tag{2.68}$$

By Lemma 2.5,

$$\begin{aligned}
&\int_U d\Pi(\theta) D(P_\theta^{\otimes n} || P_{\Pi,n}) \\
\leqslant&\Pi(K_0) \int_{K_0} d\Pi_0(\theta) D(P_\theta^{\otimes n} || P_{\Pi_0,n}) + (1 - \Pi(K_0)) \int_{U \setminus K_0} d\Pi_1(\theta) D(P_\theta^{\otimes n} || P_{\Pi_1,n}) \\
&- \Pi(K_0) \log(\Pi(K_0)) - (1 - \Pi(K_0)) \log(1 - \Pi(K_0)).
\end{aligned} \tag{2.69}$$

Therefore, defining $f_\Pi(K_0) := \Pi(K_0) \log(\Pi(K_0)) + (1 - \Pi(K_0)) \log(1 - \Pi(K_0))$,

$$\begin{aligned}
&\int_U d\Pi(\theta) D(P_\theta^{\otimes n} || P_{\Pi,n}) + f_\Pi(K_0) \\
\leqslant&\Pi(K_0) \int_{K_0} d\Pi_0(\theta) D(P_\theta^{\otimes n} || P_{\Pi_0,n}) + (1 - \Pi(K_0)) \sup_{\Pi' \in \mathcal{P}(K)} \int_K d\Pi'((\theta) D(P_\theta^{\otimes n} || P_{\Pi',n})
\end{aligned} \tag{2.70}$$

From this we obtain

$$\begin{aligned}
&\int_U d\Pi(\theta) D(P_\theta^{\otimes n} || P_{\Pi,n}) - \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) + f_\Pi(K_0) \\
\leqslant&\Pi(K_0) \left[\int_{K_0} d\Pi_0(\theta) D(P_\theta^{\otimes n} || P_{\Pi_0,n}) - \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) \right] \\
&+ (1 - \Pi(K_0)) \left[\sup_{\Pi' \in \mathcal{P}(K)} \int_K d\Pi'((\theta) D(P_\theta^{\otimes n} || P_{\Pi',n}) - \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) \right]
\end{aligned} \tag{2.71}$$

Theorem 2.16 implies that

$$\begin{aligned}
&\limsup_{n \to +\infty} \left[\int_U d\Pi(\theta) D(P_\theta^{\otimes n} || P_{\Pi,n}) - \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) \right] + f_\Pi(K_0) \\
\leqslant&\Pi(K_0) \limsup_{n \to +\infty} \left[\int_{K_0} d\Pi_0(\theta) D(P_\theta^{\otimes n} || P_{\Pi_0,n}) - \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) \right] \\
&+ (1 - \Pi(K_0)) \lim_{n \to +\infty} \left[\sup_{\Pi' \in \mathcal{P}(K)} \int_K d\Pi'((\theta) D(P_\theta^{\otimes n} || P_{\Pi',n}) - \frac{d}{2} \log\left(\frac{n}{2\pi e}\right) \right] \\
\leqslant&\Pi(K_0) \int_{K_0} |\mathcal{I}(\theta)|^{\frac{1}{2}} d\theta + (1 - \Pi(K_0)) \int_K |\mathcal{I}(\theta)|^{\frac{1}{2}} d\theta.
\end{aligned} \tag{2.72}$$

Now, let $(K_m)_{m\in\mathbb{Z}_+}$ be an increasing sequence of compact subsets of U with limit U.
For any positive integer m,

$$\begin{aligned}&\limsup_{n\to+\infty}\left[\int_U d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]+f_\Pi(K_m)\\ \leqslant&\Pi(K_m)\int_{K_m}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta+(1-\Pi(K_m))\int_K|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta.\end{aligned}\tag{2.73}$$

Because $\lim_{m\to+\infty}\Pi(K_m)=1$, we have $\lim_{m\to+\infty}f_\Pi(K_m)=0$ and thus

$$\limsup_{n\to+\infty}\left[\int_U d\Pi(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]\leqslant\lim_{m\to+\infty}\int_{K_m}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta=\int_U|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta.\tag{2.74}$$

So the first part of the lemma is proved.

Now, let $\Pi^\star$ be the Jeffreys-rule prior normalized so that $\Pi^\star(U)=1$. If there exists a compact subset $K_U^\star$ of U such that $\Pi^\star(K_U^\star)=1$, then by Theorem 2.16

$$\begin{aligned}&\lim_{n\to+\infty}\left[\int_U d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi^\star,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]\\ =&\lim_{n\to+\infty}\left[\int_{K_U} d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]\\ =&\int_{K_U}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta=\int_U|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta.\end{aligned}\tag{2.75}$$

Else, for any compact subset $K_0\subset U$ such that $\Pi^\star(K_0)>0$, let $\Pi_0^\star$ be the renormalized restriction of $\Pi^\star$ to K_0 Equation (2.68) implies

$$\int_U d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi^\star,n})\geqslant\Pi^\star(K_0)\int_{K_0}d\Pi_0(\theta)D(P_\theta^{\otimes n}||P_{\Pi_0,n}).\tag{2.76}$$

This yields

$$\begin{aligned}&\liminf_{n\to+\infty}\left[\int_U d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi^\star,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]\\ \geqslant&\Pi^\star(K_0)\lim_{n\to+\infty}\left[\int_{K_0} d\Pi_0(\theta)D(P_\theta^{\otimes n}||P_{\Pi_0,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]\\ =&\Pi^\star(K_0)\int_{K_0}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta.\end{aligned}\tag{2.77}$$

And from there

$$\begin{aligned}\liminf_{n\to+\infty}\left[\int_U d\Pi^\star(\theta)D(P_\theta^{\otimes n}||P_{\Pi^\star,n})-\frac{d}{2}\log\left(\frac{n}{2\pi e}\right)\right]&\geqslant\lim_{m\to+\infty}\Pi^\star(K_m)\int_{K_m}|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta\\ &=\int_U|\mathcal{I}(\theta)|^{\frac{1}{2}}d\theta.\end{aligned}\tag{2.78}$$

So the liminf is greater or equal to the limsup: this means that liminf and limsup are equal and yields the second part of the lemma.

□

Theorem 2.18. *Under Conditions 0, 1, 2 and 3, the Jeffreys-rule prior is a reference prior in the sense of Definition 2.13. If it is proper, then it is unique. If it is improper, then it is unique up to a multiplicative constant.*

Proof. Let $(U_m)_{m\in\mathbb{N}}$ be an increasing sequence of bounded open subsets of Θ with limit Θ. Therefore, for any nonnegative integer m, the closure of U_m is a compact set. This makes Lemma 2.17 applicable and proves that the Jeffreys-rule prior is a reference prior in the sense of Definition 2.13.

Both uniqueness results follow from the existence of a consistent estimator (the Maximum Likelihood estimator), which implies that Assumptions 1 and 2 hold and make Proposition 2.14 applicable. □

2.5 Properties of reference priors

Reference priors have several interesting properties which cement their status as sensible default priors [Bernardo, 2005, Berger et al., 2009].

Theorem 2.19 (Independence from sample size)**.** *A reference prior for a model $(P_\theta)_{\theta\in\Theta}$ remains the same for the model $(P_\theta^{\otimes n})_{\theta\in\Theta}$ regardless of $n \in \mathbb{Z}_+$.*

Proof. This is because the definition of a reference prior relies on asymptotics. □

Theorem 2.20 (Compatibility with sufficient statistics)**.** *If the observed data are restricted to a sufficient statistic, reference priors remain unchanged.*

Proof. Let S be a sufficient statistic. This means that there exists a family of probability distributions $(Q_s)_{s\in S(\mathcal{Y})}$ on $(\mathcal{Y}, \mathbb{Y})$ such that for all $\theta \in \Theta$, $dP_\theta(y) = d(P_\theta * S)(S(y))dQ_{S(y)}(y)$, where $P_\theta * S$ is the push-forward measure of P_θ by S (if Y is a random variable that follows the distribution P_θ, then $S(Y)$ follows $P_\theta * S$).
For every positive integer n,

$$\begin{aligned}
D(P_\theta^{\otimes n}||P_{\Pi_n}) &= \int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(y_1,...,y_n)\log\left(\frac{\prod_{i=1}^n dP_\theta(y_i)}{\int_\Theta d\Pi(\theta)\prod_{i=1}^n dP_\theta(y_i)}\right)\\
&= \int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(y_1,...,y_n)\log\left(\frac{\prod_{i=1}^n d(P_\theta * S)(S(y_i))dQ_{S(y_i)}(y_i)}{\int_\Theta d\Pi(\theta)\prod_{i=1}^n d(P_\theta * S)(S(y_i))dQ_{S(y_i)}(y_i)}\right)\\
&= \int_{\mathcal{Y}^n} dP_\theta^{\otimes n}(y_1,...,y_n)\log\left(\frac{d(P_\theta * S)^{\otimes n}(S(y_1),...,S(y_n))}{\int_\Theta d\Pi(\theta)d(P_\theta * S)^{\otimes n}(S(y_1),...,S(y_n))}\right)\\
&= \int_{S(\mathcal{Y})^n} d(P_\theta * S)^{\otimes n}(s_1,...,s_n)\log\left(\frac{d(P_\theta * S)^{\otimes n}(s_1,...,s_n)}{\int_\Theta d\Pi(\theta)d(P_\theta * S)^{\otimes n}(s_1,...,s_n)}\right).
\end{aligned} \tag{2.79}$$

□

Theorem 2.21 (Consistency under reparametrization)**.** *If Θ' is a metric space and $f : \Theta \to \Theta'$ is a measurable bijection whose inverse f^{-1} is also measurable, and if $\Pi^\star$ is a reference prior for the model $(P_\theta)_{\theta\in\Theta}$, then the push-forward measure $\Pi^\star * f$ is a reference prior for the model $(P_{f^{-1}(\theta')})_{\theta'\in\Theta'}$.*

Proof. This follows from the change-of-variable formula. For any measurable function $g : \Theta \to [0, +\infty)$,

$$\int_{\Theta} d\Pi(\theta) g(\theta) = \int_{\Theta'} d(\Pi * f)(\theta')\ g(f^{-1}(\theta')). \tag{2.80}$$

In particular, for any positive integer n and any measurable set $A \in \mathbb{Y}^{\otimes n}$,

$$P_{\Pi,n}(A) = \int_{\Theta} d\Pi(\theta) P_{\theta}^{\otimes n}(A) = \int_{\Theta'} d(\Pi * f)(\theta') P_{f^{-1}(\theta')}^{\otimes n}(A). \tag{2.81}$$

□

The latter two properties can be interpreted this way: if it exists, a reference prior only cares about the model as a family of probability distributions. The parameter is only an index with no intrinsic meaning (consistency under reparametrization). What matters is how the probability distributions in the model differ from each other and any common parts are disregarded (compatibility with sufficient statistics).

2.6 Examples

The first two examples – location and scale models – are drawn from Bernardo [2005] and Berger et al. [2009].

Proposition 2.22 (Location models)**.** *If $\mathbf{\Theta} = \mathcal{Y} = \mathbb{R}^d$ ($d \in \mathbb{Z}_+$), with $\mathbb{R}^d$ endowed with the Borel σ-algebre $\mathcal{B}(\mathbb{R}^d)$, if for any $\theta, \theta' \in \mathbb{R}^d$ and any measurable subset A of $\mathbb{R}^d$, $P_{\theta'}(A) = P_{\theta}(A + \theta - \theta')$, and if Assumption 2 holds, then any reference prior is proportional to the Lebesgue measure.*

Proof. Let $\Pi^\star$ be a reference prior for such a model. For any $\theta_0 \in \mathbb{R}^d$, define the statistic $S_{\theta_0} : \mathbb{R}^d \to \mathbb{R}^d; y \mapsto y + \theta_0$. S_{θ_0} is a sufficient statistic, so compatibility with sufficient statistics implies $\Pi^\star$ stays reference prior for the model $(P_\theta * S_{\theta_0})_{\theta \in \mathbb{R}^d}$. However, for all $\theta \in \mathbb{R}^d$, $P_\theta * S_{\theta_0} = P_{\theta+\theta_0}$. In other words, S defines a reparametrization. Therefore consistency under reparametrization implies the reference prior for $(P_{\theta+\theta_0})_{\theta \in \mathbb{R}^d}$ corresponding to $\Pi^\star$ is the prior $\Pi^\star_{\theta_0}$ defined for every measurable set A by $\Pi^\star_{\theta_0}(A) = \Pi^\star(A - \theta_0)$.

Gathering this, we obtain $\Pi^\star_{\theta_0} = \Pi^\star$, and this holds for any $\theta_0 \in \mathbb{R}^d$. This implies that $\Pi^\star$ is proportional to the Lebesgue measure.

□

Proposition 2.23 (Scale models)**.** *Assume $\mathcal{Y} = \mathbb{R}^d$ ($d \in \mathbb{Z}_+$) and $\Theta = (0, +\infty)$ are both endowed with the Euclidean or any equivalent distance. If for any $\theta, \theta' \in (0, +\infty)$ and any measurable subset A of $\mathbb{R}^d$, $P_{\theta'}(A) = P_\theta(\frac{\theta}{\theta'} A)$, and if Assumption 2 holds, then any reference prior is proportional to $\theta^{-1} d\theta$.*

Proof. In such a model, there must exist a sufficient statistic with value on $(0, +\infty)$ for which θ works as a scale parameter (to build it, first construct a statistic whose distribution under any value of θ is isotropic and then take its Euclidean norm). Reference priors are compatible with sufficient statistics, so let us replace the model with this one. This allows us to apply the logarithm to the observation space, mapping it to $\mathbb{R}$. Taking $\log(\theta)$ as the new parametrization, the model becomes a location model and any reference prior is proportional

to the Lebesgue measure. Given reference priors are consistent under reparametrization, any reference prior on θ is proportional to the push-forward of the Lebesgue measure by the exponential function, that is the measure $\theta^{-1}d\theta$. □

Location and scale models are remarkable because invariance properties yield the reference prior. We now study an example where no such property is available. This example appears in Ren et al. [2012], but the reference prior is derived differently.

Example 3. *Consider a model $(P_\theta)_{\theta\in(0,+\infty)}$ where each P_θ is absolutely continuous with respect to the Lebesgue measure on the unit sphere $S^{n-1} := \{\boldsymbol{y} \in \mathbb{R}^n : \boldsymbol{y}^\top \boldsymbol{y} = 1\}$. The Radon-Nykodym derivative p_θ of P_θ (i.e. the likelihood function) is given by:*

$$p_\theta(\boldsymbol{y}) = \left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right)^{-1} |\boldsymbol{\Sigma}_\theta|^{-\frac{1}{2}} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{y}\right)^{-\frac{n}{2}}, \tag{2.82}$$

where Σ_θ is a symmetric positive definite matrix and is twice continuously differentiable as a function of $\theta \in (0, +\infty)$. Therefore Condition 0 of the regular case holds. Further, let us assume that $\Sigma_\theta = \Sigma_{\theta'} \Rightarrow \theta = \theta'$, which means that Condition 3 holds. Since the sphere S^{n-1} is a compact set, Conditions 1 and 2 hold as well, so Theorem 2.18 is applicable. The reference prior is the Jeffreys-rule prior.

Define a matrix $\sqrt{\boldsymbol{\Sigma_\theta}}$ such that $\boldsymbol{\Sigma_\theta} = \sqrt{\boldsymbol{\Sigma_\theta}}\sqrt{\boldsymbol{\Sigma_\theta}}^\top$ (use for instance the Cholesky decomposition). Now let $\boldsymbol{Y}$ be a random variable following P_θ. Then $f_\theta(\boldsymbol{Y}) := \sqrt{\boldsymbol{\Sigma}_\theta}^{-1}\boldsymbol{Y} / \left\|\sqrt{\boldsymbol{\Sigma}_\theta}^{-1}\boldsymbol{Y}\right\|$ follows the Uniform distribution on the sphere S^{n-1}.

Besides,

$$\partial_\theta \left(\log p_\theta(\boldsymbol{y})\right) = -\frac{n}{2} f_\theta(\boldsymbol{y})^\top \boldsymbol{M}_\theta^\Sigma f_\theta(\boldsymbol{y}) + C_\theta, \tag{2.83}$$

where

$$C_\theta := -\log\left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right) - \frac{1}{2}\partial_\theta |\boldsymbol{\Sigma}_\theta|^{-\frac{1}{2}} \quad \text{and } \boldsymbol{M}_\theta^\Sigma := \left(\sqrt{\boldsymbol{\Sigma}_\theta}\right)^\top \partial_\theta\left(\boldsymbol{\Sigma}_\theta^{-1}\right)\sqrt{\boldsymbol{\Sigma}_\theta}. \tag{2.84}$$

Therefore

$$\operatorname{Var}\left[\partial_\theta \left(\log p_\theta(\boldsymbol{Y})\right)\right] = \frac{n^2}{4} \operatorname{Var}\left[f_\theta(\boldsymbol{Y})^\top \boldsymbol{M}_\theta^\Sigma f_\theta(\boldsymbol{Y})\right]. \tag{2.85}$$

As $\boldsymbol{M}_\theta^\Sigma$ is a symmetric matrix, the spectral theorem guarantees the existence of a diagonal matrix $\boldsymbol{\Lambda}_\theta^\Sigma$ and an orthogonal matrix $\boldsymbol{O}_\theta^\Sigma$ such that $\boldsymbol{M}_\theta^\Sigma = (\boldsymbol{O}_\theta^\Sigma)^\top \boldsymbol{\Lambda}_\theta^\Sigma (\boldsymbol{O}_\theta^\Sigma)$, with the diagonal coefficients of $\boldsymbol{\Lambda}_\theta^\Sigma$ being the eigenvalues of $\boldsymbol{M}_\theta^\Sigma$. Setting $\boldsymbol{U}_0 := (\boldsymbol{O}_\theta^\Sigma)\boldsymbol{U}$, we can now compute $\operatorname{Var}[\boldsymbol{U}_0^\top \boldsymbol{\Lambda}_\theta^\Sigma \boldsymbol{U}_0] = \operatorname{Var}[\boldsymbol{U}^\top \boldsymbol{M}_\theta^\Sigma \boldsymbol{U}]$, $\boldsymbol{U}_0$ following the uniform distribution on S^{n-1}.

Let $(\lambda_i)_{1\leqslant i\leqslant n}$ be the eigenvalues of $\boldsymbol{M}_\theta^\Sigma$. We can write $\operatorname{Var}[\boldsymbol{U}_0^\top \boldsymbol{\Lambda}_\theta^\Sigma \boldsymbol{U}_0] = \operatorname{Var}[\sum_{1\leqslant i\leqslant n} \lambda_i X_i]$, where X_i $(1 \leqslant i \leqslant n)$ are nonnegative identically distributed random variables such that $\sum_{1\leqslant i\leqslant n} X_i = 1$.

$$\operatorname{Var}\left[\sum_{i=1}^n \lambda_i X_i\right] = \operatorname{Var}[X_1] \sum_{i=1}^n \lambda_i^2 + 2\operatorname{Cov}(X_1, X_2) \sum_{1\leqslant i<j\leqslant n} \lambda_i \lambda_j. \tag{2.86}$$

Because $\mathbb{E}[X_1] = \frac{1}{n}$, $\operatorname{Cov}(X_1, X_2) = -1/(n-1)\operatorname{Var}[X_1]$.

$$\begin{aligned}\frac{\mathrm{Var}\left[\sum_{i=1}^n \lambda_i X_i\right]}{\mathrm{Var}[X_1]} &= \sum_{i=1}^n \lambda_i^2 - \frac{1}{n-1}\sum_{i=1}^n \lambda_i \sum_{j\neq i} \lambda_j \\ &= \left(1+\frac{1}{n-1}\right)\left(\mathrm{Tr}\left[(\boldsymbol{M}_\theta^\Sigma)^2\right] - \frac{1}{n}\mathrm{Tr}\left[\boldsymbol{M}_\theta^\Sigma\right]^2\right) \\ &= \left(1+\frac{1}{n-1}\right)\left(\mathrm{Tr}\left[\left(\left(\frac{\partial}{\partial_\theta}\boldsymbol{\Sigma}_\theta\right)\boldsymbol{\Sigma}_\theta^{-1}\right)^2\right] - \frac{1}{n}\mathrm{Tr}\left[\left(\frac{\partial}{\partial_\theta}\boldsymbol{\Sigma}_\theta\right)\boldsymbol{\Sigma}_\theta^{-1}\right]^2\right).\end{aligned} \tag{2.87}$$

The reference prior therefore has density π with respect to the Lebesgue measure on $(0,+\infty)$, where π is defined by:

$$\pi(\theta) \propto \sqrt{\mathrm{Tr}\left[\left(\left(\frac{\partial}{\partial_\theta}\boldsymbol{\Sigma}_\theta\right)\boldsymbol{\Sigma}_\theta^{-1}\right)^2\right] - \frac{1}{n}\mathrm{Tr}\left[\left(\frac{\partial}{\partial_\theta}\boldsymbol{\Sigma}_\theta\right)\boldsymbol{\Sigma}_\theta^{-1}\right]^2}. \tag{2.88}$$

Remark. Consider the case where $\theta \in (0,+\infty)^r$, r being a greater than 1 integer. To emphasize that θ is multidimensional, let us denote it by $\boldsymbol{\theta}$. The reference prior is still the Jeffreys rule prior: it is the square root of the determinant of the Fisher information matrix $\mathcal{I}(\boldsymbol{\theta})$ whose (i,j)-th element is

$$\mathrm{Tr}\left[\left(\frac{\partial}{\partial\theta_i}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\left(\frac{\partial}{\partial\theta_j}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right] - \frac{1}{n}\mathrm{Tr}\left[\frac{\partial}{\partial\theta_i}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]\mathrm{Tr}\left[\frac{\partial}{\partial\theta_j}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]. \tag{2.89}$$

This results from polarization formula $\mathrm{Cov}(A,B) = 1/4(\mathrm{Var}(A+B) - \mathrm{Var}(A-B))$.

Let us now consider a particular example that is a hybrid between location and scale models. It is often referenced in the literature. See for instance Robert [2007, section 3.5].

Example 4. *Take $\theta = (\beta,\sigma)$ and let $P_{\beta,\sigma} = \mathcal{N}(\beta,\sigma^2)$ be the one-dimensional Normal model. This is the sort of regular case where the reference prior is the Jeffreys-rule prior. The Fisher information matrix is diagonal with both diagonal elements equal to σ^{-2}. The Jeffreys-rule prior has therefore density σ^{-2}. Let us examine this a little further. If σ were known, then this would be a location model and the reference prior on β would have density $\propto 1$. If β were known, then this would be a scalar model and the reference prior would have density $\propto \sigma^{-1}$.*

Consider n independent real-valued observations $y_1,...,y_n$ of the model.
If β were known but not σ, then the posterior distribution of $\sum_{y=1}^n (y_i-\beta)^2/\sigma^2$ (i.e. its distribution knowing all y_i ($1 \leqslant i \leqslant n$) and β) would be the chi-squared distribution with n degrees of freedom.
In our case, both β and σ^2 are unknown. Let us define $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$. Then the posterior distribution of $\sum_{y=1}^n (y_i-\bar{y})^2/\sigma^2$ (i.e. its distribution knowing all y_i ($1 \leqslant i \leqslant n$) but not β) is also the chi-squared distribution with n degrees of freedom. In other words, using the reference prior in both situations implies that by simply substituting the empirical mean $\bar{y}$ to the actual mean β, we are able to reach the same state of knowledge about σ as if we actually knew β.

Jeffreys [1961] suggested using what is now called the independence Jeffreys prior distribution, which is the joint prior distribution on β and σ obtained by taking the product of the Jeffreys-rule (reference) prior on β when σ is known and of the Jeffreys-rule (reference) prior on σ

when β is known, thus yielding the joint prior distribution σ^{-1}. Then the posterior distribution of $\sum_{y=1}^{n}(y_i - \bar{y})^2/\sigma^2$ is the chi-squared distribution with $n-1$ degrees of freedom, which acknowledges the loss of information on σ^2 when β is unknown.

This example suggests that reference priors, as they are currently defined, are not necessarily adequate for dealing with multiparametric settings. It also suggests a method for solving the problem.

2.7 Reference priors for multiparametric models

Previously, we dealt with the case of one-parameter models. Naturally, in the case of multi-parameter models, it is always possible to view the list of parameters as one big multi-dimensional parameter, so as to reduce the problem to the one already tackled. Such a choice often leads to unfortunate results however, in the sense that the obtained reference prior seems intuitively unsatisfactory. More precisely, there exist several multi-parameter models in the literature in which the reference prior, as defined above, has undesirable statistical properties. See Berger et al. [2015] for a large review of such situations. In fact, the authors 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' ". And, as mentioned before, the reference prior is the Jeffreys-rule prior in regular cases.
Let $\Theta_1 \times ... \times \Theta_r$ ($r \in \mathbb{Z}_+$) be the parametric space. The "reference prior algorithm", which was first developed by Bernardo [1979a], first requires an ordering of the parameters. Let us consider them ordered in the following way:

$$\theta_1 \in \Theta_1 \prec ... \prec \theta_r \in \Theta_r$$

We denote by $\boldsymbol{\theta}_{[j]}$ the collection $(\theta_j, \theta_{j+1}, ..., \theta_r)$. Define $\mathbb{Y}_1 := \mathbb{Y}$.

Before delving into the details, let us explain the heuristic behind the algorithm. The idea is quite simple.

The reference prior is only defined in one-parameter settings. Let us therefore fix all parameters except one. Then we can compute the reference prior on the remaining parameter conditionally to all fixed parameters. Then, having defined a probability distribution on one of the parameters, we can integrate it out of the model.

We now find ourselves with a model with one less parameter. The same procedure can be applied again and again until none is left. The reference prior is then defined to be the product of the priors derived at every stage.

Proper reference prior

If possible, follow this algorithm: for every integer $j \in [\![1, r]\!]$,

1. For all $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$, compute $\Pi_j(\cdot|\boldsymbol{\theta}_{[j+1]})$. It is defined as the reference prior on θ_j with respect to the model $(P_{\boldsymbol{\theta}_{[j]}})_{\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r}$ when $\boldsymbol{\theta}_{[j+1]}$ is assumed known. If, for some $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$, it does not exist, abort.
2. If there exists $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$ such that $\Pi_j(\cdot|\boldsymbol{\theta}_{[j+1]})$ is improper, abort.

3. For all $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$, compute $P_{\boldsymbol{\theta}_{[j+1]}} := \int_{\Theta_j} P_{\boldsymbol{\theta}_{[j]}} d\Pi_j(\theta_j|\boldsymbol{\theta}_{[j+1]})$.

If Assumption 1 holds at every stage, no choice is required from the user aside from the initial ordering of the parameters.
The reference prior based on the chosen ordering is then defined as the product $d\Pi(\boldsymbol{\theta}_{[1]}) = \prod_{j=1}^{r} d\Pi_j(\theta_j|\boldsymbol{\theta}_{[j+1]})$.

The general idea behind this algorithm is to emulate the success of the one-dimensional reference prior by creating such a situation for the parameter that is of "greatest interest". To achieve this, "nuisance" parameters are integrated out of the model. To simplify, consider the 2-parameter case, in which there is only one nuisance parameter. In order to create a one-parameter model for the parameter of interest, the nuisance parameter must be integrated out of the original model. But to achieve this, a prior distribution on the nuisance parameter conditional to the parameter of interest must be defined. Hence the idea to define it as the reference prior for the model where the parameter of interest is known.

It must be stressed that this reference prior for multiparametric models is not a further extension of the concept of reference prior like Definition 2.13 was to Definition 2.4 and Definition 2.4 to Definition 2.2. It is an altogether different notion because it introduces structure in the model in the form or parameter ordering. Yang and Berger [1996] list different reference priors obtained with different parameter orderings for a large number of statistical models.

Applying this "reference prior algorithm" results in a "marginal reference prior" on the parameter of interest that is in fact the reference prior (in the sense of Definition 2.13) of a specific mixture model [Marin et al., 2005, Mengersen et al., 2011]. So the reference prior resulting from application of the reference prior algorithm is the solution of a very different optimization problem than the one that is solved by the reference prior in the sense of Definition 2.13.

Notice that the algorithm contains two abortive checks per step. The first is intrinsically linked to reference prior theory, which does not provide guarantees of existence. The second is due to the fact that a valid statistical model can only contain probability distributions: it does not tolerate infinite measures! In other words, a prior can well be improper, but a likelihood function cannot.

Note however that aborting at stage 2 is not a problem if $j = r$, because the last step has been reached. So all "conditional" reference priors must be proper, but the last, "marginal" one can be improper. As will be seen in the next subsection, this propriety requirement on the conditionals can be removed, making stage 2 of the algorithm irrelevant.

Improper reference prior

The algorithm above can be altered to remove one of the two abortive checks. Changes are written in **bold font**. First, **define** $\mathbb{Y}_1 := \mathbb{Y}$. Then, for every integer $j \in [\![1, r]\!]$,

1. For all $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$, compute $\Pi_j(\cdot|\boldsymbol{\theta}_{[j+1]})$. It is defined as the reference prior on θ_j with respect to the model $(P_{\boldsymbol{\theta}_{[j]}})_{\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r}$ when $\boldsymbol{\theta}_{[j+1]}$ is assumed known. If, for some $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$, it does not exist, abort.

2. If there exists $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$ such that $\Pi_j(\cdot|\boldsymbol{\theta}_{[j+1]})$ is improper, **do the following. Choose a sub-σ-algebra $\mathbb{Y}_{j+1} \subset \mathbb{Y}_j$ such that for every $A \in \mathbb{Y}_{j+1}$, $P_{\boldsymbol{\theta}_{[j]}}(A)$ only depends on $\boldsymbol{\theta}_{[j+1]}$. Define $P_{\boldsymbol{\theta}_{[j+1]}}$ on $\mathbb{Y}_{j+1} \subset \mathbb{Y}_j$ as the restriction of $P_{\boldsymbol{\theta}_{[j]}}$ to $\mathbb{Y}_{j+1}$.** Else, for all $\boldsymbol{\theta}_{[j+1]} \in \Theta_{j+1} \times ... \times \Theta_r$, compute $P_{\boldsymbol{\theta}_{[j+1]}} := \int_{\Theta_j} P_{\boldsymbol{\theta}_{[j]}} d\Pi_j(\theta_j|\boldsymbol{\theta}_{[j+1]})$.

As written above, this altered algorithm – henceforth called "reference prior algorithm" – has the advantage of removing one aborting check. In cases where the first algorithm works, this one is functionally identical and produces the same result. In cases where improper priors are encountered before the last stage ($j = r$) however, it requires a choice. Therefore, even if Assumption 1 holds at every stage, the reference prior is only defined with reference to a given filtering $\mathbb{Y} = \mathbb{Y}_1 \supset ... \supset \mathbb{Y}_r$.

Remark. The main problem of this algorithm, besides the lack of uniqueness of the reference prior, is the risk that at some stage, the only possible choice for $\mathbb{Y}_j$ might be the trivial σ-algebra. One possible solution is to consider the model with n independent observations instead of one, in order to manipulate richer σ-algebras. An example is shown below.

The altered algorithm does not fit the approach that is most often used to deal with impropriety. The idea recommended by Berger and Bernardo [1992] is to choose for every $j \in [\![1, r]\!]$ an increasing sequence of compact subspaces $(\Theta_j^{(k)})_{k\in\mathbb{N}}$ such that $\bigcup_{k\in\mathbb{N}} \Theta_j^{(k)} = \Theta_j$ with the understanding that the reference prior on each compact is proper. This happens when the reference prior is the Jeffreys-rule prior, for example. Then, it is possible to use the first algorithm on each $\Theta_1^{(k)} \times ... \times \Theta_r^{(k)}$, compute the associated reference posterior, take the limit of the posterior when $k \to +\infty$ and then define the reference prior as the prior which, using Bayes' rule, would yield the reference posterior. This increasing-compact-sequence approach is what Berger et al. [2001] call *asymptotic marginalization*, whereas the independent-sub-σ-algebra approach detailed above is a formalization of the *exact marginalization* they recommend for Kriging models.

Unfortunately, asymptotic marginalization does not guarantee uniqueness of the reference prior (even for a given parameter ordering) any more than exact marginalization does. The choice of sequences of compact subspaces can influence the result. Because this thesis deals with Kriging and Kriging-derived models, all reference priors discussed in the dissertation are obtained through exact marginalization.

Example 5. *Let us come back to the earlier Normal $\mathcal{N}(\beta, \sigma^2)$ example. To apply the reference prior algorithm, we need to define an ordering on the parameters β and σ. Let us choose the ordering $\beta \prec \sigma$.*

At the first stage of the algorithm, the reference prior on β (knowing σ) is, up to a multiplicative constant, the Lebesgue measure (cf. earlier example). This is an improper prior, so we need to find a sub-σ-algebra of $\mathcal{B}(\mathbb{R})$ on which the model does not depend on β. To do this, let us consider the model with $n \geqslant 2$ observations. We are looking for a sub-σ-algebra of $\mathcal{B}(\mathbb{R})^{\otimes n} = \mathcal{B}(\mathbb{R}^n)$.

Let us consider the statistic $S_n : \mathbb{R}^n \to \mathbb{R}^{n-1}$ defined by $S_n((y_1, ..., y_n)^\top) = (y_2 - y_1, ..., y_n - y_1)^\top$. If $\boldsymbol{Y}$ is a random variable following $\mathcal{N}(\beta, \sigma^2)^{\otimes n}$, $S_n(\boldsymbol{Y})$ follows the $(n-1)$-variate Normal distribution $\mathcal{N}\left(\mathbf{0}_{n-1}, \sigma^2(\mathbf{1}\mathbf{1}^\top + \boldsymbol{I}_{n-1})\right)$, with $\mathbf{1} = (1, 1, ..., 1)^\top$ and with $\boldsymbol{I}_{n-1}$ being the identity matrix. Therefore the sub-σ-algebra of $\mathcal{B}(\mathbb{R}^n)$ spanned by S_n does not depend

on β. Moreover, the restriction of the model to this sub-σ-algebra is a simple scale model. Therefore the reference prior on σ has density proportional $1/\sigma$ with respect to the Lebesgue measure on $(0, +\infty)$.

Gathering this, the reference prior associated to the decomposition we chose is $1/\sigma\, d\beta\, d\sigma$. Interestingly, this is the "independence Jeffreys prior" that was recommended by Jeffreys [1961] for this situation.

Finally, note that the same reference prior could have been obtained by choosing the other ordering: $\sigma \prec \beta$. At the first stage of the algorithm, when β is taken to be known, the model is a scale model so the reference prior on σ knowing β has density proportional to $1/\sigma$ with respect to the Lebesgue measure. This is an improper prior, so we need to find a sub-σ-algebra on which the model does not depend on σ. The statistic $T_n : (y_1, ..., y_n)^\top \mapsto \sqrt{n}\bar{y}_n/\hat{\sigma}_n$ (with $\bar{y}_n = n^{-1}\sum_{i=1}^n y_i$ and $\hat{\sigma}_n^2 = n^{-1}\sum_{i=1}^n (y_i - \bar{y}_n)^2$) spans such a sub-$\sigma$-algebra. If $\boldsymbol{Y}$ follows $\mathcal{N}(\beta, \sigma^2)$, then $T_n(\boldsymbol{y})$ follows a noncentral Student t-distribution with $n-1$ degrees of freedom and noncentrality parameter β. The resulting model is a location model, so the reference prior on β is proportional to the Lebesgue measure.

This example is important, but it should be noted that cases where the same reference prior is obtained regardless of the ordering of the parameters are the exception rather than the norm. Berger et al. [2015] explores several methods to find an "overall reference priors" which would be adequate regardless of the ordering of the parameters.

Some final thoughts

In the Normal example, the Jeffreys-rule prior is considered inadequate because it has undesirable inferential properties. It is, though, the reference prior for the model if $(\beta, \sigma)^\top$ is viewed as a single multidimensional parameter. So why does the reference prior behave "badly" in this instance? First, let us observe that it does not differ from the independence Jeffreys prior regarding the mean parameter β. The difference between the two concerns the standard deviation σ. The Jeffreys-rule prior favors small standard deviations, which lead to concentrated samples. Recalling the game-theoretic interpretation from Clarke and Barron [1994], Nature is giving a hint to the Statistician – the standard deviation is likely to be small. Why is it in Nature's interest to do so? Because this forces the Statistician to predict a concentrated sample, even though the Statistician does not know around which value it should concentrate. If Nature's prior favored greater standard deviations, the Statistician could "spread the prediction" around and would be more likely to cover the true mean. The Jeffreys-rule prior is truly noninformative in the sense that it makes prediction most difficult *before* viewing the data, even though the Statistician would have less trouble inferring σ *after*.

The Jeffreys independence prior is better not because it is less but because it is *more* informative than the Jeffreys-rule prior. It contains information about the structure of the model by clearly separating mean from standard deviation. It turns out that being invariant under reparametrization is not necessarily an advantage, because it can take away useful information that is contained in the structure of the parametrization.

Instead of being invariant under any reparametrization, a reference prior obtained by application of the reference prior algorithm is invarariant under "sensible" reparametrizations, i.e.

reparametrizations that respect the structure. Consider a reference prior for $\theta_1 \prec ... \prec \theta_r$ (with associated parameter space $\Theta_1 \times ... \times \Theta_r$). If for every integer $i \in [\![1, r]\!]$, f_i is a measurable bijection $\Theta_i \to \Theta'_i$ with measurable inverse, then this reference prior is invariant under the reparametrization $(\theta_1, ..., \theta_r) \mapsto (f_1(\theta_1), ..., f_r(\theta_r))$.

The claim in the introduction to this chapter that a reference prior should be used when no prior information is available is therefore not wholly correct. If the user can define a parametrization, it implies that some information is available and should be used.

Deriving a reference prior is not about producing the least informative prior, but rather about controlling the information contained in the prior. From this point onward, following Berger and Bernardo, we call it "objective" rather than noninformative.

However, this word should not lead to the conclusion that a reference prior is "better" than another because of its objectivity. Depending on whether the goal is prediction or inference, and on what is being inferred, other priors may produce better results. Tuyl et al. [2016] show reference (or Jeffreys-rule) priors are not adequate for the study of rare events. Indeed, this is not what they were designed for: by definition, rare events play a small role in the optimization criterion of reference priors.

Currently, research focuses on choosing the prior in order to obtain a specific, desired, effect on the posterior. Xueou et al. [2018] propose an elicitation mechanism for the prior Π based on history matching in order to make relevant summary statistics [Marin et al., 2014] of data produced by P_Π fit an expected behavior. Computation complexity is dealt with by using Approximate Bayesian Computation [Marin et al., 2012]. Simpson et al. [2017] introduce a new framework called "Penalized Complexity" priors as an attempt to formalize the principle of Occam's razor for prior elicitation – again with reference to P_Π; Robert and Rousseau [2017] note that the authors attempt to circumvent the subjectivity inherent to the ordering of parameters by dividing them into independent components. The original developers of the reference prior paradigm are also concerned with this issue and attempt to address it through the notion of "overall reference priors" [Berger et al., 2015]. Another direction for research of reasonable default priors is the Robust Bayesian framework; for a recent review of this field, see Watson and Holmes [2016].

Chapter 3

Propriety of the reference posterior distribution in Gaussian Process regression

This chapter draws on the article Muré [2018a].

Abstract

In a seminal article, Berger et al. [2001] compare several objective prior distributions for the parameters of Gaussian Process regression models with isotropic correlation kernel. The reference prior distribution stands out among them insofar as it always leads to a proper posterior. They prove this result for rough correlation kernels - Spherical, Exponential with power $q < 2$, Matérn with smoothness $\nu < 1$. This chapter provides a proof for smooth correlation kernels - Exponential with power $q = 2$, Matérn with smoothness $\nu \geqslant 1$, Rational Quadratic.

Résumé

Dans un article fondamental, Berger et al. [2001] comparent plusieurs lois *a priori* objectives sur les paramètres de modèles de régression par processus gaussiens avec noyau de corrélation isotrope. Le prior de référence se distingue parmi eux en cela qu'il mène systématiquement à un posterior propre. Ils démontrent ce résultat pour des noyaux de corrélation rugueux – noyau sphérique, exponentiel avec puissance $q < 2$, Matérn de régularité $\nu < 1$. Ce chapitre fournit une preuve valable pour des noyaux de corrélation réguliers – exponentiel de puissance $q = 2$, Matérn de régularité $\nu \geqslant 1$, rationnel quadratique.

3.1 Introduction

In a very influential paper, Berger et al. [2001] pioneered the field of Objective Bayesian analysis of spatial models. Previous works [De Oliveira et al., 1997, Stein, 1999] had noted that commonly used noninformative priors sometimes failed to yield proper posteriors, but Berger et al. [2001] were the first to thoroughly investigate the issue. Among several prior distributions – truncated priors, vague priors, Jeffreys-rule and independence Jeffreys prior – they showed that the reference prior (i.e., the reference prior with the parameter ordering

they chose) is the most satisfying choice for a default prior distribution. This is due in no small part to the fact that, in the wide variety of cases studied by Berger et al. [2001], it systematically yields a proper posterior distribution. In this article, we complete their proof of this property.

Interestingly, Berger et al. [2001] found that the corresponding reference prior obtained through asymptotic marginalization does not share this property. In this chapter and in the rest of this dissertation, all reference priors discussed are obtained through exact marginalization.

Section 3.2 describes the Gaussian Process models studied by Berger et al. [2001]. Section 3.3 shows that the proof of reference posterior propriety provided by Berger et al. [2001] only applies to those with rough correlation kernels – Spherical, Exponential with power $q < 2$, Matérn with smoothness $\nu < 1$. Section 3.4 contains the core of this chapter: a proof of Theorem 3.9 which asserts that the reference prior leads to a proper posterior for models with smoother correlation kernels – Exponential with power $q = 2$, Matérn with smoothness $\nu \geqslant 1$, Rational Quadratic.

Because it is difficult to obtain a satisfying default prior distribution which consistently yields a proper posterior, it is important to ascertain that the reference prior actually does. Indeed, a vast literature [Paulo, 2005, Ren et al., 2012, Kazianka and Pilz, 2012, Ren et al., 2013, Gu et al., 2018] builds upon Berger et al. [2001]'s result and depends on it.

3.2 Setting

Berger et al. [2001] consider models of Gaussian Process regression, also known as Universal Kriging, with isotropic autocorrelation kernels. Because isotropy is key, define $\|\cdot\|$ as the usual Euclidean norm if applied to a vector and as the Frobenius norm if applied to a matrix. In Universal Kriging, an unknown mapping from a spatial domain $\mathcal{D} \subset \mathbb{R}^r$ ($r \in \mathbb{Z}_+$) to $\mathbb{R}$ is assumed to be a realization of a Gaussian process Y. The mean function f of the Gaussian process is assumed to belong to some known vector space $\mathcal{F}_p$ of dimension $p \in \mathbb{N}$. If p is non-zero, once a basis $(f_j)_{j\in[\![1,p]\!]}$ of $\mathcal{F}_p$ has been set, f can be parametrized by $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top \in \mathbb{R}^p$ such that $f = \sum_{j=1}^p \beta_j f_j$.

$Y - f$ is assumed in the model to be an isotropic Gaussian process based on an autocorrelation kernel K. K is a mapping $[0, +\infty) \to \mathbb{R}$ such that for any positive integer n and any collection of n distinct points $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$ within $\mathcal{D}$, the symmetric $n \times n$ matrix $\boldsymbol{\Sigma}$ with (i, i')-th element $K(\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|)$ is a positive definite correlation matrix. Necessarily, $K(0) = 1$.

The autocovariance function of the Gaussian process Y is $\sigma^2 K_\theta$, where K_θ is the autocorrelation kernel parametrized by $\theta \in (0, +\infty)$ and defined by $K_\theta(d) = K(d/\theta)$, making $\sigma^2 \in (0, +\infty)$ the variance of $Y(\boldsymbol{x})$ for all $\boldsymbol{x} \in \mathcal{D}$.

Fix $n \in \mathbb{Z}_+$ and a collection of n distinct points $(\boldsymbol{x}^{(i)})_{i\in[\![1,n]\!]}$. Let this collection be the design set, i.e. the set of points where Y is observed. $\big(Y(\boldsymbol{x}^{(1)}), ..., Y(\boldsymbol{x}^{(n)})\big)^\top$ is a Gaussian vector with mean vector $\big(f(\boldsymbol{x}^{(1)}), ..., f(\boldsymbol{x}^{(n)})\big)^\top$ and covariance matrix $\sigma^2 \boldsymbol{\Sigma}_\theta$, where $\boldsymbol{\Sigma}_\theta$ denotes the $n \times n$ matrix with (i, i')-th element $K_\theta(\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|)$. Table 3.1 recalls Table 1.1, which provided the definition of several correlation kernels.

Kernel	$K_\theta(\lvert x\rvert)$	parameter range
Spherical ($r=1,2,3$)	$\left(1-\frac{3}{2}\left(\frac{\lvert x\rvert}{\theta}\right)+\frac{1}{2}\left(\frac{\lvert x\rvert}{\theta}\right)^3\right)\mathbf{1}_{\{\lvert x\rvert\leqslant\theta\}}$	$\emptyset$
Power Exponential	$\exp\left\{-\left(\frac{\lvert x\rvert}{\theta}\right)^q\right\}$	$q\in(0,2]$
Rational Quadratic	$\left(1+\left(\frac{\lvert x\rvert}{\theta}\right)^2\right)^{-\nu}$	$\nu\in(0,+\infty)$
Matérn	$\Gamma(\nu)^{-1}2^{1-\nu}\left(2\sqrt{\nu}\frac{\lvert x\rvert}{\theta}\right)^{\nu}\mathcal{K}_\nu\left(2\sqrt{\nu}\frac{\lvert x\rvert}{\theta}\right)$	$\nu\in(0,+\infty)$

Table 3.1 – *Formulas for several correlation kernel families. The Squared Exponential kernel is the Power Exponential kernel with $q=2$. $\mathcal{K}_\nu$ is the modified Bessel function of second kind with parameter ν [Abramowitz and Stegun, 1964](9.6.). This parametrization of the Matérn family is recommended by Handcock and Wallis [1994]. To recover the one used by Berger et al. [2001], simply replace $2\sqrt{\nu}\lvert x\rvert$ by $\lvert x\rvert$.*

If p is non-zero, let $\boldsymbol{H}$ denote the $n\times p$ matrix with (i,j)-th element $f_j(\boldsymbol{x}^{(i)})$. [Note: if $p=0$, then we adopt the convention that any term involving $\boldsymbol{H}$ can be ignored.] Then $(f(\boldsymbol{x}^{(1)},...,f(\boldsymbol{x}^{(n)}))^\top=\boldsymbol{H}\boldsymbol{\beta}$. Denote by $\boldsymbol{y}=(y_1,...y_n)^\top$ the observed value of the random vector $\left(Y(\boldsymbol{x}^{(1)}),...,Y(\boldsymbol{x}^{(n)})\right)^\top$. The likelihood function of the parameter triplet $(\boldsymbol{\beta},\sigma^2,\theta)$ has the following expression:

$$L(\boldsymbol{y}\mid\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta})=\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}}|\boldsymbol{\Sigma_\theta}|^{-\frac{1}{2}}\exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\beta})^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\beta})\right\}. \tag{3.1}$$

In order for the model to be identifiable, assume that $p<n$ and that $\boldsymbol{H}$ has rank p.
Berger et al. [2001] derive the reference prior corresponding to the parameter ordering $\boldsymbol{\beta}\prec(\sigma^2,\theta)$ [if $p=0$, $\boldsymbol{\beta}$ is meaningless, so the ordering is (σ^2,θ)]. One can see [Ren et al., 2012] that the reference prior corresponding to the ordering $\boldsymbol{\beta}\prec\sigma^2\prec\theta$ [if $p=0$, $\sigma^2\prec\theta$] is the same.
To express it conveniently, denote by $\boldsymbol{Q}_\theta$ the matrix $\boldsymbol{I}_n-\boldsymbol{H}\left(\boldsymbol{H}^\top\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^\top\boldsymbol{\Sigma}_\theta^{-1}$ [if $p=0$, $\boldsymbol{Q}_\theta=\boldsymbol{I}_n$]. Also fix $\boldsymbol{W}$, an $n\times(n-p)$ matrix such that $\boldsymbol{W}^\top\boldsymbol{W}=\boldsymbol{I}_{n-p}$ and $\boldsymbol{H}^\top\boldsymbol{W}$ is the $p\times(n-p)$ null matrix. $\boldsymbol{W}$'s columns form an orthonormal basis of the orthogonal complement of the subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{H}$ [if $p=0$, fix $\boldsymbol{W}$ as an orthogonal matrix, for instance $\boldsymbol{I}_n$].

$\boldsymbol{\beta}$ is a location parameter and $\sigma:=\sqrt{\sigma^2}$ a scale parameter. Therefore, conditional on θ, the reference prior with ordering $\boldsymbol{\beta}\prec\sigma$ is proportional to $d\boldsymbol{\beta}\sigma^{-1}d\sigma$. Because the multi-parameter reference prior is invariant under parameter-by-parameter reparametrizations and $d\sigma^2/d\sigma=2\sigma$, the reference prior with ordering $\boldsymbol{\beta}\prec\sigma^2$ is proportional to $d\boldsymbol{\beta}(\sigma^2)^{-1}d(\sigma^2)$.

Proposition 3.1. *If $p\geqslant 1$, after marginalizing $\boldsymbol{\beta}$ and σ^2 out, we have*

$$\begin{aligned}L(\boldsymbol{y}|\theta)&=\iint L(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\theta)/\sigma^2 d\boldsymbol{\beta}d\sigma^2\\&=\left(\frac{2\pi^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-p}{2}\right)}\right)^{-1}\left|\boldsymbol{\Sigma}_\theta^{-1}\right|^{\frac{1}{2}}\left|\left|\boldsymbol{H}^\top\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{H}\right|^{-\frac{1}{2}}\left(\boldsymbol{y}^\top\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{Q}_\theta\boldsymbol{y}\right)^{\frac{n-p}{2}}.\end{aligned} \tag{3.2}$$

Alternatively, the integrated likelihood with $p\geqslant 1$ can also be written

$$L(\boldsymbol{y}|\theta)=\left(\frac{2\pi^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-p}{2}\right)}\right)^{-1}\left|\boldsymbol{H}^\top\boldsymbol{H}\right|^{\frac{1}{2}}\left|\boldsymbol{W}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{W}\right|^{-\frac{1}{2}}\left(\boldsymbol{y}^\top\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}\right)^{\frac{n-p}{2}}. \tag{3.3}$$

If $p = 0$, the integrated likelihood is simply

$$L(\boldsymbol{y}|\theta) = \int L(\boldsymbol{y}|\sigma^2, \theta)/\sigma^2 d\sigma^2 = \left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right)^{-1} \left|\boldsymbol{\Sigma}_\theta^{-1}\right|^{\frac{1}{2}} \left|\left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{y}\right)^{\frac{n}{2}}\right|. \tag{3.4}$$

Proof. The result for $p = 0$ and the first result for $p \geqslant 1$ are from Berger et al. [2001]. Lemma 3.10 yields that

$$\boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top = \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{Q}_\theta. \tag{3.5}$$

So all that remains to be proved is that $|\boldsymbol{\Sigma}_\theta| = \left|\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right| \left|\boldsymbol{H}^\top \boldsymbol{H}\right|^{-1} \left|\boldsymbol{H}^\top \boldsymbol{\Sigma}_\theta^{-1} \boldsymbol{H}\right|^{-1}$. Choose an $n \times p$ matrix $\boldsymbol{P}$ with columns forming an orthonormal basis of the subspace of $\mathbb{R}^n$ spanned by the columns of $\boldsymbol{H}$. $(\boldsymbol{W}\boldsymbol{P})$ is therefore an $n \times n$ orthogonal matrix, so $|\boldsymbol{\Sigma}_\theta| = \left|(\boldsymbol{W}\boldsymbol{P})^\top \boldsymbol{\Sigma}_\theta (\boldsymbol{W}\boldsymbol{P})\right|$. Using Schur's complement, we have

$$|\boldsymbol{\Sigma}_\theta| = \left|\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right| \left|\boldsymbol{P}^\top \boldsymbol{\Sigma}_\theta \left(\boldsymbol{I}_n - \boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta\right) \boldsymbol{P}\right|. \tag{3.6}$$

Lemma 3.10 again yields the result. □

In the following proposition, the first assertion is from Ren et al. [2012].

Proposition 3.2. *The reference prior with ordering $\boldsymbol{\beta} \prec \sigma^2 \prec \theta$ is $\pi(\boldsymbol{\beta}, \sigma^2, \theta) \propto \left(\sigma^2\right)^{-1} \pi(\theta)$, where*

$$\pi(\theta) \propto \sqrt{\mathrm{Tr}\left[\left\{\left(\frac{d}{d\theta}\boldsymbol{\Sigma}_\theta\right)\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{Q}_\theta\right\}^2\right] - \frac{1}{n-p}\left[\mathrm{Tr}\left\{\left(\frac{d}{d\theta}\boldsymbol{\Sigma}_\theta\right)\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{Q}_\theta\right\}\right]^2}. \tag{3.7}$$

Denoting $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$ by $\boldsymbol{\Sigma}_\theta^{\boldsymbol{W}}$, $\pi(\theta)$ can also be written as:

$$\pi(\theta) \propto \sqrt{\mathrm{Tr}\left[\left\{\left(\frac{d}{d\theta}\boldsymbol{\Sigma}_\theta^{\boldsymbol{W}}\right)\left(\boldsymbol{\Sigma}_\theta^{\boldsymbol{W}}\right)^{-1}\right\}^2\right] - \frac{1}{n-p}\left[\mathrm{Tr}\left\{\left(\frac{d}{d\theta}\boldsymbol{\Sigma}_\theta^{\boldsymbol{W}}\right)\left(\boldsymbol{\Sigma}_\theta^{\boldsymbol{W}}\right)^{-1}\right\}\right]^2}. \tag{3.8}$$

Proof. To obtain the second assertion, all we need to do is recognize Example 3 from Chapter 2. This can be done after noticing that $\boldsymbol{W}^\top \boldsymbol{y}$ plays the role of $\boldsymbol{y}$ and $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$ the role of $\boldsymbol{\Sigma}_\theta$. Because $\boldsymbol{W}^\top \boldsymbol{y} \in \mathbb{R}^{n-p}$, $n - p$ plays the role of n.

The first assertion is then a consequence of Lemma 3.10.

□

3.3 Smoothness of the correlation kernel

Lemma 2 of Berger et al. [2001] requires that correlation kernel and design set should be such that $\boldsymbol{\Sigma}_\theta = \mathbf{1}\mathbf{1}^\top + g_0(\theta)\boldsymbol{D} + \boldsymbol{R}_0(\theta)$, where $\mathbf{1}$ is the vector with n entries all equal to 1, $g_0(\theta)$ is a real-valued function such that $\lim_{\theta \to +\infty} g_0(\theta) = 0$, $\boldsymbol{D}$ is a fixed nonsingular matrix and $\boldsymbol{R}_0$ is a mapping from $(0, +\infty)$ to the set of $n \times n$ real matrices $\mathcal{M}_n$ such that $\lim_{\theta \to +\infty} \left\|\frac{1}{g_0(\theta)}\boldsymbol{R}_0(\theta)\right\| = 0$.

What makes this assumption restrictive is the condition that $\boldsymbol{D}$ should be nonsingular, because it holds for rough correlation kernels only. For instance, as was noted by Paulo [2005], it does not hold for the Squared Exponential correlation kernel.

For a given correlation kernel K, $\boldsymbol{D}$ is typically a matrix proportional to the matrix with entries $\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^q$, where q depends on the smoothness of the correlation kernel but should in any case belong to the interval $(0, 2]$. This is because $K(s)-K(0)$ is equivalent to a constant times s^q when $s \to 0+$.

Schoenberg [1937] gives the following result (Theorem 4 in the original paper):

Theorem 3.3. *If $q \in (0, 2)$, the quadratic form $\boldsymbol{\xi} \in \mathbb{R}^n \mapsto \sum_{i,j=0}^n \left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^q \ \xi_i\xi_j$ is nonsingular and its canonical representation contains one positive and n negative squares.*

This means that if the correlation kernel is rough enough to have $q \in (0, 2)$, the assumption that $\boldsymbol{D}$ is nonsingular is reasonable.

Corollary 3.4. *The $n \times n$ matrix with entries $\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^q$ with $q \in (0, 2)$ is nonsingular and has one positive eigenvalue and n negative eigenvalues.*

The picture is dramatically different when the correlation kernel K is smooth enough to have $q = 2$. This happens as soon as K is twice continuously differentiable. Gower [1985]'s Theorem 6 implies the following results:

Theorem 3.5. *If d is the dimension of E_d, the smallest Euclidean subspace containing all points in the design set, then the $n \times n$ matrix with entries $\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^2$ has rank:*

(a) $d + 1$ (one positive eigenvalue, d negative eigenvalues, any other eigenvalue null) if all points in the design set lie on the surface of a hypersphere of E_d ;

(b) $d + 2$ (one positive eigenvalue, $d + 1$ negative eigenvalues, any other eigenvalue null) otherwise.

Corollary 3.6. *The $n \times n$ matrix with entries $\left\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(j)}\right\|^2$ has rank lower or equal to $r+2$.*

For all practical purposes, n is much greater than r, so the matrix $\boldsymbol{D}$ is singular when $q = 2$.

Let us review the values of q for correlation kernels listed in Table 3.1. Matérn correlation kernels [Matérn, 1986] [Handcock and Stein, 1993] with smoothness parameter ν have $q = 2\min(1, \nu)$, thus for $0 < \nu < 1$, $0 < q < 2$ but for $\nu \geqslant 1$, $q = 2$. Spherical correlation kernels [Wackernagel, 1995] have $q = 1$. Power Exponential kernels [De Oliveira et al., 1997] have q equal to their power. This means that all Power Exponential kernels except the Squared Exponential correlation kernel have $0 < q < 2$. In particular, the Exponential kernel (which is also the Matérn kernel with smoothness $\nu = 1/2$) has $q = 1$, but the Squared Exponential kernel has $q = 2$. Rational Quadratic kernels [Yaglom, 1987] have $q = 2$. For easy reference, the review is summarized in Table 3.2.

This review justifies the claim in the abstract that the Squared Exponential kernel, Matérn kernels with smoothness $\nu \geqslant 1$ and Rational Quadratic kernels require a proof of the reference posterior's propriety.

Kernel	$g_0(\theta)$	$\|\boldsymbol{R}_0(\theta)\|$	q	$\boldsymbol{D}$ nonsingular*
Spherical ($r = 1, 2, 3$)	$-3/2\theta^{-1}$	$O(\theta^{-3})$	1	yes
Power Expon. ($q < 2$)	$-\theta^{-q}$	$O(\theta^{-2q})$	q	yes
Squared Exponential	$-\theta^{-2}$	$O(\theta^{-4})$	2	no
Rational Quadratic	$-\nu\theta^{-2}$	$O(\theta^{-4})$	2	no
Matérn ($\nu < 1$)	$\Gamma(-\nu)\nu^{\nu}\Gamma(\nu)^{-1}\theta^{-2\nu}$	$O(\theta^{-2})$	2ν	yes
Matérn ($\nu = 1$)	$-2\theta^{-2}\log(\theta)$	$O(\theta^{-2})$	2	no
Matérn ($1 < \nu < 2$)	$-\Gamma(\nu-1)\nu\Gamma(\nu)^{-1}\theta^{2}$	$O(\theta^{-2\nu})$	2	no
Matérn ($\nu = 2$)	$-2\theta^{2}$	$O(\theta^{-4\log(\theta)})$	2	no
Matérn ($\nu > 2$)	$-\Gamma(\nu-1)\nu\Gamma(\nu)^{-1}\theta^{2}$	$O(\theta^{-4})$	2	no

Table 3.2 – *Summary of the results of Section 3.3. *Answer given assuming $n > r + 2$.*

3.4 Propriety of the reference posterior distribution

Berger et al. [2001] show that the reference posterior distribution on $\boldsymbol{\beta}$ and σ^2 conditionally to θ is proper.
In this section, we prove that the joint reference posterior distribution is proper for Matérn kernels with smoothness $\nu \geqslant 1$, Rational Quadratic kernels and the Squared Exponential kernel.

Proposition 3.7. *For Matérn kernels with smoothness $\nu \geqslant 1$, for Rational Quadratic kernels with parameter $\nu > 0$ and for the Squared Exponential kernel, the "marginal" reference prior distribution $\pi(\theta)$ defined by Proposition 3.2 has the following behavior.*

1. *When $\theta \to 0$,*

$$\pi(\theta) = \begin{cases} o(1) & \text{for Matérn kernels and the Squared Exponential kernel;} \\ O(\theta^{2\nu-1}) & \text{for Rational Quadratic kernels .} \end{cases} \tag{3.9}$$

2. *When $\theta \to +\infty$,*

$$\pi(\theta) = \begin{cases} O(\theta^{-1}) & \text{for Matérn kernels;} \\ o(1) & \text{for Rational Quadratic kernels ;} \\ O(\theta) & \text{for the Squared Exponential kernel .} \end{cases} \tag{3.10}$$

Proof. Denoting any of these kernels by K, K is continuously differentiable.

If K is Squared Exponential, $\lim_{\theta\to 0} \frac{d}{d\theta}K(1/\theta) = 0$. This also holds if K is Matérn with smoothness $\nu \geqslant 1$ (see Abramowitz and Stegun [1964] 9.6.28. and 9.7.2.). If K is Rational Quadratic with parameter $\nu > 0$, $\frac{d}{d\theta}K(1/\theta) \underset{\theta\to 0}{\sim} 2\nu\theta^{2\nu-1}$. Moreover, $\boldsymbol{\Sigma}_\theta$ converges to $\boldsymbol{I}_n$ when $\theta \to 0$, so its inverse does too. The first assertion follows from these facts.

The second assertion is proved by combining Lemma 3.12 with Lemma 3.20/3.21/3.22 for Matérn, Rational Quadratic and Squared Exponential kernels respectively. □

Let $v_1(\theta) \geqslant ... \geqslant v_{n-p}(\theta) > 0$ be the ordered eigenvalues of $\boldsymbol{W}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{W}$.

Lemma 3.8. *For Rational Quadratic and Squared Exponential kernels and for Matérn kernels with smoothness $\nu \geqslant 1$, there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that for every $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, when $\theta \to +\infty$:*

$$\left(\boldsymbol{y}^\top\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}\right)^{-1} = O(v_{n-p}(\theta)). \tag{3.11}$$

The proof of this lemma can be found in Appendix 3.D. Combined with Equation (3.3), it implies that if the observation vector $\boldsymbol{y}$ belongs to $\mathbb{R}^n \setminus \mathcal{H}$, then

$$L(\boldsymbol{y}|\theta)^2 = \prod_{i=1}^{n-p} \frac{O(v_{n-p}(\theta))}{v_i(\theta)} = O(1) \quad \text{when } \theta \to +\infty. \tag{3.12}$$

In the following, when $\boldsymbol{y}$ belongs to $\mathbb{R}^n \setminus \mathcal{H}$, we write that "$\boldsymbol{y}$ looks nondegenerate". This terminology relies on the intuition that if the observation were to take some values within $\mathcal{H}$, it would be better explained by a degenerate Gaussian model. The most compelling example is that of a constant observation vector, for which the Kriging model would be grossly inappropriate.

Theorem 3.9. *For Matérn kernels with noninteger smoothness $\nu > 1$, for Rational Quadratic kernels and for the Squared Exponential kernel, regardless of the design set and of the mean function space, if $\boldsymbol{y}$ looks nondegenerate, then the reference posterior distribution $\pi(\theta|\boldsymbol{y})$ is proper.*

Proof. The first assertion of Proposition 3.7 implies the reference prior $\pi(\theta)$ is integrable in the neighborhood of 0. Furthermore, when $\theta \to 0$, $\boldsymbol{\Sigma}_\theta \to \boldsymbol{I}_n$ so the reference posterior $\pi(\theta|\boldsymbol{y})$ is integrable in the neighborhood of 0 as well.

All that remains to be proved is therefore that the reference posterior is integrable in the neighborhood of $+\infty$. In the following $\theta \to +\infty$, so we rely on the asymptotic expansion of $\boldsymbol{\Sigma}_\theta$, which is detailed in Appendix 3.D.

The proof is somewhat trickier for Matérn kernels with integer smoothness, so we tackle this case at the end. Until further notice, assume the kernel is Rational Quadratic, Squared Exponential or Matérn with noninteger smoothness $\nu > 1$.

For Rational Quadratic and Squared Exponential (resp. Matérn with noninteger smoothness parameter $\nu > 1$) kernels, Appendix 3.D (resp. Appendix 3.D) shows how $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$ can be decomposed as

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} = g(\theta) \left(\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W} + g^\star(\theta) \boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} + \boldsymbol{R}_g(\theta) \right), \tag{3.13}$$

where

- g is a differentiable function;
- $g^\star(\theta) = \theta^{-2l}$ with $l \in (0, +\infty)$ (actually, if the kernel is Rational Quadratic or Squared Exponential, $l \in \mathbb{Z}_+$);
- $\boldsymbol{R}_g$ is a differentiable mapping from $(0, +\infty)$ to $\mathcal{M}_n$ such that $\|\boldsymbol{R}_g(\theta)\| = o(\theta^{-2l})$;
- $\boldsymbol{D}$ and $\boldsymbol{D}^\star$ are both fixed symmetric matrices;
- $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is non-null;
- either $\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W}$ is non-null or $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is nonsingular.

Lemma 3.15 implies that one of the following is true:

1. When θ is large enough, $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W} + g^\star(\theta) \boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W}$ is nonsingular. This case can be further decomposed in the following subcases:
   a) $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is nonsingular;
   b) $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is singular, but $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W} + g^\star(\theta) \boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W}$ is nonsingular when θ is large enough.

2. The vector space $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}\right)$ is non-trivial.

Let us differentiate $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$:

$$\frac{d}{d\theta}\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} = \frac{g'(\theta)}{g(\theta)}\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} + g(\theta)\left(g^{\star\prime}(\theta)\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} + \frac{d}{d\theta}\boldsymbol{R}_g(\theta)\right). \tag{3.14}$$

We can show that $\left\|\frac{d}{d\theta}\boldsymbol{R}_g(\theta)\right\| = o(g^{\star\prime}(\theta))$. This is due to Equation (3.72) for Rational Quadratic and Squared Exponential kernels, and to Equation (3.75) for Matérn kernels with noninteger smoothness.

Lemma 3.11 shows that $\frac{d}{d\theta}\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$ can be replaced by $g(\theta)\left(g^{\star\prime}(\theta)\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} + \frac{d}{d\theta}\boldsymbol{R}_g(\theta)\right)$ in Equation (3.8): $\pi(\theta) \propto w(\theta)$, where

$$\begin{aligned} w(\theta)^2 := \mathrm{Tr}&\left[\left\{g(\theta)\left(g^{\star\prime}(\theta)\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} + \frac{d}{d\theta}\boldsymbol{R}_g(\theta)\right)\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1}\right\}^2\right] \\ &- \frac{1}{n-p}\left[\mathrm{Tr}\left\{g(\theta)\left(g^{\star\prime}(\theta)\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} + \frac{d}{d\theta}\boldsymbol{R}_g(\theta)\right)\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1}\right\}\right]^2. \end{aligned} \tag{3.15}$$

We have $w(\theta) \leqslant \tilde{w}(\theta)$, where

$$\tilde{w}(\theta) := \sqrt{\mathrm{Tr}\left[\left\{g(\theta)\left(g^{\star\prime}(\theta)\boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} + \frac{d}{d\theta}\boldsymbol{R}_g(\theta)\right)\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1}\right\}^2\right]}. \tag{3.16}$$

A specific asymptotic analysis is required in each case. This study is conducted in Appendix 3.E. We summarize the results in Table 3.3.

Case	**Kernels**	$\pi(\theta)$	$L(\boldsymbol{y}\|\theta)$
1.(a)	Matérn ($\nu \in [1,+\infty) \setminus \mathbb{Z}_+$), RQ, SE	$O\left(\theta^{-2l-1}\right)$	$O(1)$
1.(b)	Matérn ($\nu \in [1,+\infty) \setminus \mathbb{Z}_+$), RQ, SE	$O\left(\theta^{-1}\right)$	$O\left(\theta^{-l}\right)$
2.	Matérn ($\nu \in [1,+\infty) \setminus \mathbb{Z}_+$)	$O\left(\theta^{-1}\right)$	$O\left(\theta^{-l}\right)$
2.	RQ, SE (usual case)	$O(\theta)$	$O\left(\theta^{-3}\right)$
2.	RQ, SE (special case)	$O\left(\theta^{-1}\right)$	$O\left(\theta^{-1}\right)$

Table 3.3 – *Asymptotic upper bounds for reference prior $\pi(\theta)$ and likelihood $L(\boldsymbol{y}|\theta)$ for Rational Quadratic (RQ) and Squared Exponential (SE) kernels and Matérn kernels with noninteger smoothness $\nu > 1$ in all three cases. The proof in Appendix 3.E shows that for Rational Quadratic and Squared Exponential kernels, case 2. can be split in two subcases ("usual" and "special").*

The posterior distribution resulting from the reference prior is proper in all cases.

Matérn kernels with integer smoothness are dealt with in Appendix 3.E.

□

3.5 Conclusion

The main result of this chapter is Theorem 3.9, which ensures that the reference prior leads to a proper posterior distribution for a large class of smooth kernels. This class contains the Squared Exponential correlation kernel as well as the important Matérn family [Stein, 1999] with smoothness parameter $\nu \geqslant 1$. Rational Quadratic kernels, whose usage is less widespread are also included within this class.

Berger et al. [2001] proved this result for a class of rough correlation kernels. This class includes the complementary set of the Matérn family – kernels with smoothness parameter $\nu < 1$ – as well as all other Power Exponential kernels. Spherical kernels, which are mostly used in the field of geostatistics also belong to this class.

Combining Theorem 3.9 with the results from Berger et al. [2001], one can appreciate how polyvalent the reference prior is, insofar as it is able to adapt to very different correlation kernels and always leads to a proper posterior. No *ad-hoc* technique is required to derive useable inference, so this approach seems to be flawless from a Bayesian point of view when no explicit prior information is available. Even when explicit prior information is available, following Druilhet and Marin [2007], it can be used to derive Maximum A Posteriori (MAP) estimates or High Probability Density (HPD) sets that are invariant under reparametrization.

Appendix 3.A Algebraic facts

Lemma 3.10. *Let a and b be positive integers and let $\boldsymbol{\Sigma}$ be a nonsingular symmetric $(a+b)\times(a+b)$ matrix. Then, for any $(a+b)\times a$ matrix $\boldsymbol{A}$ with rank a and any $(a+b)\times b$ matrix $\boldsymbol{B}$ with rank b such that $\boldsymbol{A}^\top\boldsymbol{B}$ is the null $a\times b$ matrix,*

$$\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top = \boldsymbol{\Sigma}^{-1}\left(\boldsymbol{I}_{a+b} - \boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\right). \tag{3.17}$$

Proof. Notice that both matrices have the same kernel, namely the subspace of $\mathbb{R}^{a+b}$ spanned by $\boldsymbol{A}$. Indeed, because $\boldsymbol{B}$ has full column rank and $\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}\right)^{-1}$ is nonsingular, the left matrix has the same kernel as $\boldsymbol{B}^\top$. Besides, the a-dimensional subspace of $\mathbb{R}^{a+b}$ spanned by $\boldsymbol{A}$ in included in this kernel. So because the rank of $\boldsymbol{B}^\top$ is b, its kernel has dimension a and the inclusion is an equality.
Similarly, because $\boldsymbol{\Sigma}^{-1}$ is nonsingular, the right matrix has the same kernel as $\boldsymbol{I}_{a+b} - \boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}$. Moreover, because the image of $\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}$ is included within the image of $\boldsymbol{A}$, its dimension is lower or equal to a. The image of $\boldsymbol{I}_{a+b}$ on the other hand has dimension $a+b$, so the image of $\boldsymbol{I}_{a+b} - \boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}$ has dimension greater or equal to b and therefore its kernel has dimension lower or equal to a. Now, a simple computation shows that the a-dimensional subspace of $\mathbb{R}^{a+b}$ spanned by $\boldsymbol{A}$ in included in the kernel, so it is in fact equal to the kernel.
Besides, for any $\boldsymbol{z}\in\mathbb{R}^b$,

$$\boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top(\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{z}) = \boldsymbol{B}\boldsymbol{z}; \tag{3.18}$$

$$\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{I}_{a+b} - \boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}^{-1}\right)(\boldsymbol{\Sigma}\boldsymbol{B}\boldsymbol{z}) = \boldsymbol{B}\boldsymbol{z}. \tag{3.19}$$

So both matrices act the same way on the subspace spanned by $\boldsymbol{\Sigma}\boldsymbol{B}$, which is supplementary to their common kernel, hence the equality. □

Lemma 3.11. *Let m be a positive integer, $\boldsymbol{\Sigma}$ be a nonsingular $m\times m$ matrix, and $\boldsymbol{A}$ and $\boldsymbol{B}$ be $m\times m$ matrices. If there exists a real number t such that*

$$\boldsymbol{A} = t\boldsymbol{\Sigma} + \boldsymbol{B}, \tag{3.20}$$

then

$$\operatorname{Tr}\left[\left\{\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\right\}^2\right] - \frac{1}{m}\left[\operatorname{Tr}\left\{\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\right\}\right]^2 = \operatorname{Tr}\left[\left\{\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right\}^2\right] - \frac{1}{m}\left[\operatorname{Tr}\left\{\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right\}\right]^2. \tag{3.21}$$

Proof. The lemma follows from a direct calculation:

$$\operatorname{Tr}\left[\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\right] = \operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right] + tm \tag{3.22}$$

$$\operatorname{Tr}\left[\left\{\boldsymbol{A}\boldsymbol{\Sigma}^{-1}\right\}^2\right] = \operatorname{Tr}\left[\left\{\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right\}^2\right] + 2t\operatorname{Tr}\left[\boldsymbol{B}\boldsymbol{\Sigma}^{-1}\right] + t^2m \tag{3.23}$$

□

Lemma 3.12. *Let $m > a$ be positive integers, $\boldsymbol{\Sigma}$ be an $m \times m$ symmetric positive definite matrix, $\boldsymbol{\Sigma}'$ be an $m \times m$ symmetric matrix and $\boldsymbol{A}$ be an $m \times a$ matrix with rank a. Denote by $\boldsymbol{Q}$ the matrix $\boldsymbol{I}_m - \boldsymbol{A}\left(\boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{A}\right)^{-1} \boldsymbol{A}^\top \boldsymbol{\Sigma}^{-1}$. Then, if there exist $t_1 \in \mathbb{R}$ and $t_2 \in [0, +\infty)$ such that the matrix $\boldsymbol{F} := t_1 \boldsymbol{\Sigma} - \boldsymbol{\Sigma}'$ is positive semi-definite and verifies $\forall \boldsymbol{\xi} \in \mathbb{R}^m \ \boldsymbol{\xi}^\top \boldsymbol{F} \boldsymbol{\xi} \leqslant t_2 \boldsymbol{\xi}^\top \boldsymbol{\Sigma} \boldsymbol{\xi}$, then*

$$\sqrt{\operatorname{Tr}\left[(\boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Q})^2\right] - \frac{1}{m-a} \operatorname{Tr}\left[\boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}\right]^2} \leqslant (m-a) t_2 \tag{3.24}$$

Proof. Let $\boldsymbol{B}$ be an $m \times (m-a)$ matrix with rank $m-a$ such that $\boldsymbol{A}^\top \boldsymbol{B}$ is the null $a \times (m-a)$ matrix. Such a matrix $\boldsymbol{B}$ can for instance be constructed by computing a Singular Value Decomposition (SVD) of $\boldsymbol{A}$: $\boldsymbol{A} = \boldsymbol{U} \boldsymbol{S} \boldsymbol{V}^\top$. In this decomposition, $\boldsymbol{U}$ and $\boldsymbol{V}$ are orthogonal matrices of size $m \times m$ and $a \times a$ respectively, and $\boldsymbol{S}$ is an $m \times a$ matrix whose only non-null entries are on the main diagonal. Therefore the last $m - a$ rows of $\boldsymbol{S}$ are filled with zeros. So define $\boldsymbol{B}$ as the $m \times (m-a)$ matrix formed by the last $m - a$ columns of $\boldsymbol{U}$.

By applying Lemma 3.10, we obtain that $\boldsymbol{\Sigma}^{-1} \boldsymbol{Q} = \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1} \boldsymbol{B}^\top$.

Because of the properties of the trace, this implies

$$\operatorname{Tr}\left[\boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}\right] = \operatorname{Tr}\left[\boldsymbol{B}^\top \boldsymbol{\Sigma}' \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right] \tag{3.25}$$

$$\operatorname{Tr}\left[(\boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Q})^2\right] = \operatorname{Tr}\left[\left\{\boldsymbol{B}^\top \boldsymbol{\Sigma}' \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right\}^2\right]. \tag{3.26}$$

Similarly, we have

$$\operatorname{Tr}\left[\boldsymbol{F} \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}\right] = \operatorname{Tr}\left[\boldsymbol{B}^\top \boldsymbol{F} \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right] \tag{3.27}$$

$$\operatorname{Tr}\left[(\boldsymbol{F} \boldsymbol{\Sigma}^{-1} \boldsymbol{Q})^2\right] = \operatorname{Tr}\left[\left\{\boldsymbol{B}^\top \boldsymbol{F} \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right\}^2\right]. \tag{3.28}$$

Because $\boldsymbol{B}^\top \boldsymbol{F} \boldsymbol{B} = t_1 \boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B} - \boldsymbol{B}^\top \boldsymbol{\Sigma}' \boldsymbol{B}$, Lemma 3.11 implies

$$\begin{aligned} &\operatorname{Tr}\left[\left\{\boldsymbol{B}^\top \boldsymbol{\Sigma}' \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right\}^2\right] - \frac{1}{m-a} \operatorname{Tr}\left[\boldsymbol{B}^\top \boldsymbol{\Sigma}' \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right]^2 \\ =& \operatorname{Tr}\left[\left\{\boldsymbol{B}^\top \boldsymbol{F} \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right\}^2\right] - \frac{1}{m-a} \operatorname{Tr}\left[\boldsymbol{B}^\top \boldsymbol{F} \boldsymbol{B}\left(\boldsymbol{B}^\top \boldsymbol{\Sigma} \boldsymbol{B}\right)^{-1}\right]^2. \end{aligned} \tag{3.29}$$

Combining the 5 equations above yields

$$\operatorname{Tr}\left[(\boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Q})^2\right] - \frac{1}{m-a} \operatorname{Tr}\left[\boldsymbol{\Sigma}' \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}\right]^2 = \operatorname{Tr}\left[\left(\boldsymbol{F} \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}\right)^2\right] - \frac{1}{m-a} \operatorname{Tr}\left[\boldsymbol{F} \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}\right]^2. \tag{3.30}$$

An elementary computation shows that $\boldsymbol{\Sigma}^{-1} \boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Q}$.
Consider the Cholesky decomposition $\boldsymbol{\Sigma} =: \boldsymbol{L} \boldsymbol{L}^\top$.
Then $\boldsymbol{\Sigma}^{-1} \boldsymbol{Q} = \boldsymbol{Q}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{Q} = \boldsymbol{Q}^\top \left(\boldsymbol{L}^{-1}\right)^\top \boldsymbol{L}^{-1} \boldsymbol{Q}$.

$$\begin{aligned}\operatorname{Tr}\left[\left(\boldsymbol{F}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\right)^2\right] &= \operatorname{Tr}\left[\left(\boldsymbol{F}\boldsymbol{Q}^\top\left(\boldsymbol{L}^{-1}\right)^\top\boldsymbol{L}^{-1}\boldsymbol{Q}\right)^2\right] = \operatorname{Tr}\left[\left(\boldsymbol{L}^{-1}\boldsymbol{Q}\boldsymbol{F}\boldsymbol{Q}^\top\left(\boldsymbol{L}^{-1}\right)^\top\right)^2\right] \\ &\leqslant \operatorname{Tr}\left[\boldsymbol{L}^{-1}\boldsymbol{Q}\boldsymbol{F}\boldsymbol{Q}^\top\left(\boldsymbol{L}^{-1}\right)^\top\right]^2 = \operatorname{Tr}\left[\boldsymbol{F}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\right]^2. \end{aligned} \tag{3.31}$$

The inequality holds because $\boldsymbol{L}^{-1}\boldsymbol{Q}\boldsymbol{F}\boldsymbol{Q}^\top\left(\boldsymbol{L}^{-1}\right)^\top$ is a symmetric positive semi-definite matrix.

Let $(\boldsymbol{\xi}^i)_{1\leqslant i\leqslant m}$ be a basis of unit eigenvectors of $\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}$ such that for every integer $i \in [\![1, m]\!] \setminus [\![1, m-a]\!]$, $\boldsymbol{\xi}^i$ belongs to the kernel of $\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}$. Indeed, because $\boldsymbol{\Sigma}^{-1}\boldsymbol{Q} = \boldsymbol{B}\left(\boldsymbol{B}^\top\boldsymbol{\Sigma}\boldsymbol{B}\right)^{-1}\boldsymbol{B}^\top$, this kernel has the same dimension as the kernel of $\boldsymbol{B}^\top$: a.
Denoting by $(s_i)_{1\leqslant i\leqslant m}$ the family of the eigenvalues corresponding to the family of eigenvectors $(\boldsymbol{\xi}^i)_{1\leqslant i\leqslant m}$, we have for every integer $i \in [\![1, m-a]\!]$ $s_i \neq 0$ and

$$\begin{aligned}(\boldsymbol{\xi}^i)^\top\boldsymbol{\Sigma}\boldsymbol{\xi}^i &= s_i^{-2}\left\{(\boldsymbol{\xi}^i)^\top\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1}\right\}\boldsymbol{\Sigma}\left\{\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\boldsymbol{\xi}^i\right\} \\ &= s_i^{-2}(\boldsymbol{\xi}^i)^\top\boldsymbol{Q}^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\boldsymbol{\xi}^i \\ &= s_i^{-2}(\boldsymbol{\xi}^i)^\top\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\boldsymbol{\xi}^i \\ &= s_i^{-1}. \end{aligned} \tag{3.32}$$

This implies the third equality below:

$$\operatorname{Tr}\left[\boldsymbol{F}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\right] = \sum_{i=1}^{m}\left(\boldsymbol{\xi}^i\right)^\top\boldsymbol{F}\boldsymbol{\Sigma}^{-1}\boldsymbol{Q}\boldsymbol{\xi}^i = \sum_{i=1}^{m-a}s_i\left(\boldsymbol{\xi}^i\right)^\top\boldsymbol{F}\boldsymbol{\xi}^i = \sum_{i=1}^{m-a}\frac{\left(\boldsymbol{\xi}^i\right)^\top\boldsymbol{F}\boldsymbol{\xi}^i}{\left(\boldsymbol{\xi}^i\right)^\top\boldsymbol{\Sigma}\boldsymbol{\xi}^i} \leqslant (m-a)t_2. \tag{3.33}$$

Equations (3.30) and (3.31) yield the result.

□

Entire series

Lemma 3.13. *Let $(\boldsymbol{D}_k)_{k\in\mathbb{N}}$ be a sequence of matrices of the same size. If $\sum_{k\in\mathbb{N}}\boldsymbol{D}_k$ exists and its kernel is the trivial vector space, then there exists a nonnegative integer N such that $\cap_{k=0}^{N}\operatorname{Ker}\boldsymbol{D}_k$ is the trivial vector space.*

Proof. Assume the sum $\sum_{k\in\mathbb{N}}\boldsymbol{D}_k$ exists and its kernel is the trivial vector space. Consider the sequence $(d(n))_{n\in\mathbb{N}}$ where for every nonnegative integer n, $d(n)$ is the dimension of $\cap_{k=0}^{n}\operatorname{Ker}\boldsymbol{D}^{(k)}$. $(d(n))_{n\in\mathbb{N}}$ is a nonincreasing sequence of nonegative integers, so it is convergent. If its limit is strictly greater than 0, then for every nonnegative integer n, there exists a unit vector $\boldsymbol{v}_n$ that belongs to $\cap_{k=0}^{n}\operatorname{Ker}\boldsymbol{D}^{(k)}$. Because the unit sphere is compact, there exists an increasing mapping $\phi : \mathbb{N} \to \mathbb{N}$ such that the subsequence $(\boldsymbol{v}_{\phi(n)})_{n\in\mathbb{N}}$ converges to a limit $\boldsymbol{v}$ such that $\|\boldsymbol{v}\| = 1$. Besides, for every pair of nonnegative integers $n \leqslant n'$, $\boldsymbol{v}_{\phi(n')} \in \cap_{k=0}^{\phi(n)}\operatorname{Ker}\boldsymbol{D}^{(k)}$. Given this set is closed, the limit $\boldsymbol{v}$ also belongs to $\cap_{k=0}^{\phi(n)}\operatorname{Ker}\boldsymbol{D}^{(k)}$. So for every nonnegative integer k, $\boldsymbol{v} \in \operatorname{Ker}\boldsymbol{D}^{(k)}$ and therefore $\boldsymbol{v} \in \cap_{k=0}^{\infty}\operatorname{Ker}\boldsymbol{D}^{(k)}$. So $\boldsymbol{v}$ can only be the null vector, which is absurd since $\|\boldsymbol{v}\| = 1$. We deduce from this contradiction that the limit of the sequence of integers $(d(n))_{n\in\mathbb{N}}$ is 0. Therefore there exists a nonnegative integer N such that $d(N) = 0$. □

Appendix 3.B Maclaurin series

The lemmas in this subsection deal with the following setting.

Let m be a positive integer and let $\boldsymbol{M}$ be a continuous mapping from $\mathbb{R}$ to $\mathcal{M}_m$, the set of $m \times m$ matrices. Assume $\boldsymbol{M}$ admits the following Maclaurin series:

$$\boldsymbol{M}(t) = \sum_{k=0}^{N} a_k(t)\boldsymbol{A}_k + \boldsymbol{B}(t). \tag{3.34}$$

In the above expression, N is a nonnegative integer and for every $k \in [\![0, N]\!]$:

(a) a_k is a continuous mapping $\mathbb{R} \to \mathbb{R}$ such that for all $t \neq 0$, $a_k(t) \neq 0$;

(b) for every nonnegative integer $l < k$, $a_k(t) = o(|a_l(t)|)$ when $t \to 0$;

(c) $\boldsymbol{A}_k$ is a non-null symmetric $m \times m$ matrix.

$\boldsymbol{B}$ is a continuous mapping $\mathbb{R} \to \mathcal{M}_m$ such that for every $t \in \mathbb{R}$, $\boldsymbol{B}(t)$ is a symmetric matrix and when $t \to 0$, $\|\boldsymbol{B}(t)\| = o(|a_N(t)|)$.

Lemma 3.14. *Consider* (3.34). *If* $\cap_{k=0}^{N} \operatorname{Ker} \boldsymbol{A}_k$ *is the trivial vector space and if there exists* $T > 0$ *such that for all* $t \in (-T, T)$ $\boldsymbol{M}(t)$ *is nonsingular, then when* $t \to 0$, $\left\|\boldsymbol{M}(t)^{-1}\right\| = O\left(|a_N(t)|^{-1}\right)$.

Proof. Assume that $\cap_{k=0}^{N} \operatorname{Ker} \boldsymbol{A}_k$ is the trivial vector space and that there exists $T > 0$ such that for all $t \in (-T, T)$, $\boldsymbol{M}(t)$ is a nonsingular matrix.

If $N = 0$, then $\boldsymbol{A}_0$ is nonsingular and the conclusion is trivial.

If $N \geqslant 1$, we may assume without loss of generality that $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$ is a nontrivial vector space, otherwise we could replace N by $N-1$ and $\boldsymbol{B}(t)$ by $\{a_N(t)\boldsymbol{A}_N + \boldsymbol{B}(t)\}$ for all $t \in \mathbb{R}$.

Let d_N be the codimension of $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$. Let $\boldsymbol{W}_N$ be an $m \times (m - d_N)$ matrix whose columns form an orthonormal basis of $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$, and let $\boldsymbol{P}_N$ be an $m \times d_N$ matrix whose columns form an orthonormal basis of its orthogonal complement. Then $(\boldsymbol{P}_N \boldsymbol{W}_N)$ is an orthogonal matrix. For all $t \in \mathbb{R}$, let us replace $\boldsymbol{M}(t)$ by $(\boldsymbol{P}_N \boldsymbol{W}_N)^\top \boldsymbol{M}(t)(\boldsymbol{P}_N \boldsymbol{W}_N)$. Because $(\boldsymbol{P}_N \boldsymbol{W}_N)$ is an orthogonal matrix, the Frobenius norm of $\boldsymbol{M}(t)^{-1}$ is unchanged. Naturally, for all $k \in [\![0, N]\!]$, $\boldsymbol{A}_k$ is replaced by $(\boldsymbol{P}_N \boldsymbol{W}_N)^\top \boldsymbol{A}_k (\boldsymbol{P}_N \boldsymbol{W}_N)$ and for every $t \in \mathbb{R}$, $\boldsymbol{B}(t)$ is replaced by $(\boldsymbol{P}_N \boldsymbol{W}_N)^\top \boldsymbol{B}(t)(\boldsymbol{P}_N \boldsymbol{W}_N)$.

Now, for every $k \in [\![1, N]\!]$, $\boldsymbol{A}_k$ can be decomposed into blocks – a $d_N \times d_N$ block $\boldsymbol{A}'_k$, an $(m - d_N) \times (m - d_N)$ block $\boldsymbol{A}''_k$ and a $d_N \times (m - d_N)$ block $\boldsymbol{A}'''_k$:

$$\boldsymbol{A}_k = \begin{pmatrix} \boldsymbol{A}'_k & \boldsymbol{A}'''_k \\ (\boldsymbol{A}'''_k)^\top & \boldsymbol{A}''_k \end{pmatrix} \tag{3.35}$$

For all $t \in \mathbb{R}$, $\boldsymbol{B}(t)$ can be decomposed in a similar manner (here the $'$ notation is used to distinguish the blocks, not to express some derivative with respect to t):

$$\boldsymbol{B}(t) = \begin{pmatrix} \boldsymbol{B}(t)' & \boldsymbol{B}(t)''' \\ (\boldsymbol{B}(t)''')^\top & \boldsymbol{B}(t)'' \end{pmatrix} \tag{3.36}$$

Now, for any symmetric nonsingular matrix

$$\boldsymbol{C} = \begin{pmatrix} \boldsymbol{C}' & \boldsymbol{C}''' \\ (\boldsymbol{C}''')^\top & \boldsymbol{C}'' \end{pmatrix}, \tag{3.37}$$

denoting by $\boldsymbol{S} := \left\{\boldsymbol{C}' - \boldsymbol{C}''' \left(\boldsymbol{C}''\right)^{-1} \left(\boldsymbol{C}'''\right)^\top\right\}$ the Schur complement of $\boldsymbol{C}''$, the inverse of $\boldsymbol{C}$ is

$$\boldsymbol{C}^{-1} = \begin{pmatrix} \boldsymbol{I} & \boldsymbol{0} \\ -\left(\boldsymbol{C}''\right)^{-1}\left(\boldsymbol{C}'''\right)^\top & \boldsymbol{I} \end{pmatrix} \begin{pmatrix} \boldsymbol{S}^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \left(\boldsymbol{C}''\right)^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I} & -\boldsymbol{C}''' \left(\boldsymbol{C}''\right)^{-1} \\ \boldsymbol{0} & \boldsymbol{I} \end{pmatrix}. \tag{3.38}$$

For every $k \in [\![0, N-1]\!]$, $\boldsymbol{A}_k''$ and $\boldsymbol{A}_k'''$ are null. For all $t \in (-T, T)$, $\boldsymbol{M}(t)$ is nonsingular. Its lower $(m - d_N) \times (m - d_N)$ block is $\left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}$ and its Schur complement $\boldsymbol{S}_N(t)$ is

$$\begin{aligned} \boldsymbol{S}_N(t) := &- \left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\} \left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}^{-1} \left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\}^\top \\ &+ \left\{\sum_{k=0}^{N} a_k(t)\boldsymbol{A}_k' + \boldsymbol{B}(t)'\right\}. \end{aligned} \tag{3.39}$$

Because we are dealing with the finite dimensional vector space of matrices of size $m \times m$, all norms are equivalent. In particular, the Frobenius norm is equivalent to the algebra norm

$$\boldsymbol{A} \mapsto \sup\left\{\sqrt{\boldsymbol{\xi}^\top \boldsymbol{A}^\top \boldsymbol{A}\boldsymbol{\xi} / \boldsymbol{\xi}^\top \boldsymbol{\xi}} \;:\; \boldsymbol{\xi} \in \mathbb{R}^m \setminus \{\boldsymbol{0}_m\}\right\}.$$

So there exists a constant $C_m \in (0, +\infty)$ such that for every $t \in (-T, T)$,

$$\begin{aligned} \left\|\boldsymbol{M}(t)^{-1}\right\| \leqslant C_m \Big(\|\boldsymbol{I}_m\| + \left\|\left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\}\left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}^{-1}\right\|\Big)^2 \\ \Big(\left\|\boldsymbol{S}_N(t)^{-1}\right\| + \left\|\left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}^{-1}\right\|\Big). \end{aligned} \tag{3.40}$$

$\boldsymbol{A}_N''$ is nonsingular, otherwise $\cap_{k=0}^N \operatorname{Ker} \boldsymbol{A}_k$ would be nontrivial. This means that the norm of the matrix $\left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\}\left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}^{-1}$ is bounded when $t \to 0$. Because of Equation (3.40), this implies that there exists $T_N > 0$ and $\lambda_N > 0$ such that for all $t \in (-T_N, T_N)$,

$$\lambda_N \left\|\boldsymbol{M}(t)^{-1}\right\| \leqslant |a_N(t)|^{-1} + \left\|\boldsymbol{S}_N(t)^{-1}\right\|. \tag{3.41}$$

Our goal is to use Equation (3.41) recursively, by having $\boldsymbol{S}_N(t)$ take the place of $\boldsymbol{M}(t)$. To achieve this, a new expression of $\boldsymbol{S}_N(t)$ is required.

$$\boldsymbol{S}_N(t) = \sum_{k=0}^{N-1} a_k(t)\boldsymbol{A}_k' + \boldsymbol{B}_N(t), \tag{3.42}$$

where

$$\begin{aligned} \boldsymbol{B}_N(t) := &- \left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\}\left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}^{-1}\left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\}^\top \\ &+ a_N(t)\boldsymbol{A}_N' + \boldsymbol{B}(t)'. \end{aligned} \tag{3.43}$$

It turns out that when $t \to 0$, the norm of $\boldsymbol{B}_N(t)$ is $O(|a_N(t)|)$. This is due to the fact mentioned above that $\left\{a_N(t)\boldsymbol{A}_N''' + \boldsymbol{B}(t)'''\right\}\left\{a_N(t)\boldsymbol{A}_N'' + \boldsymbol{B}(t)''\right\}^{-1}$ is bounded when $t \to 0$.

Furthermore, $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k'$ is the trivial vector space. Indeed, let $\boldsymbol{v}_1 \in \cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k'$. Then for any vector $\boldsymbol{v}_2 \in \mathbb{R}^{m-d_N}$, $(\boldsymbol{v}_1, \boldsymbol{v}_2)^\top \cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$. Independently from this, for any vector $\boldsymbol{v}_3 \in \mathbb{R}^{d_N}$, $(\boldsymbol{v}_3, \boldsymbol{0}_{m-d_N})^\top$ belongs to the orthogonal complement of $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$. So

$(\boldsymbol{v}_1, \mathbf{0}_{m-d_N})^\top$ belongs both to $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$ and its orthogonal complement: it is the null vector. Therefore $\boldsymbol{v}_1 = \mathbf{0}_{d_N}$.

The two paragraphs above show that Equation (3.42) is formally similar to Equation (3.34): the role of $\boldsymbol{M}(t)$ is held by $\boldsymbol{S}_N(t)$, the role of N by $N-1$, the role of the $\boldsymbol{A}_k$s by the $\boldsymbol{A}'_k$s and the role of $\boldsymbol{B}(t)$ by $\boldsymbol{B}_N(t)$.

Therefore an equation similar to (3.41) can be derived: there exist $T_{N-1} > 0$ and $\lambda_{N-1} > 0$ such that for all $t \in (-T_{N-1}, T_{N-1})$,

$$\lambda_{N-1} \left\| \boldsymbol{S}_N(t)^{-1} \right\| \leqslant |a_{N-1}(t)|^{-1} + \left\| \boldsymbol{S}_{N-1}(t)^{-1} \right\|. \tag{3.44}$$

Here, $\boldsymbol{S}_{N-1}(t)$ is defined with respect to $\boldsymbol{S}_N(t)$ the same way $\boldsymbol{S}_N(t)$ was defined with respect to $\boldsymbol{M}(t)$.

Recursive application of this reasoning until 0 is reached yields the result. □

Lemma 3.15. *Consider* (3.34) *with* $N = 1$. *If* $\operatorname{Ker} \boldsymbol{A}_0 \cap \operatorname{Ker} \boldsymbol{A}_1$ *is the trivial vector space, then there exists* $T > 0$ *such that for all* $t \in (-T, T)$, $\boldsymbol{M}(t)$ *is nonsingular.*

Proof. We use the same notations as in the proof of Lemma 3.14 and redefine matrices the same way: $\boldsymbol{M}(t) := (\boldsymbol{P}_1\boldsymbol{W}_1)^\top \boldsymbol{M}(t)(\boldsymbol{P}_1\boldsymbol{W}_1)$, $\boldsymbol{B}(t) := (\boldsymbol{P}_1\boldsymbol{W}_1)^\top \boldsymbol{B}(t)(\boldsymbol{P}_1\boldsymbol{W}_1)$, $\boldsymbol{A}_0 := (\boldsymbol{P}_1\boldsymbol{W}_1)^\top \boldsymbol{A}_0(\boldsymbol{P}_1\boldsymbol{W}_1)$, $\boldsymbol{A}_1 := (\boldsymbol{P}_1\boldsymbol{W}_1)^\top \boldsymbol{A}_1(\boldsymbol{P}_1\boldsymbol{W}_1)$. For all $t \in \mathbb{R}$, the determinant of $\boldsymbol{M}(t)$ is the product of the determinants of $\boldsymbol{S}_1(t)$ and of $\left\{a_1(t)\boldsymbol{A}''_1 + \boldsymbol{B}(t)''\right\}$. Because $\operatorname{Ker} \boldsymbol{A}_0 \cap \operatorname{Ker} \boldsymbol{A}_1$ is the trivial vector space, $\boldsymbol{A}''_1$ is nonsingular. When $|t|$ is small enough therefore, $\left\{a_1(t)\boldsymbol{A}''_1 + \boldsymbol{B}(t)''\right\}$ is nonsingular too. Moreover, $\lim_{t\to 0} a_0^{-1}(t)\boldsymbol{S}_1(t) = \boldsymbol{A}'_0$ is also nonsingular. These two facts imply that when $|t|$ is small enough, the determinant of $\boldsymbol{M}(t)$ is non-null and $\boldsymbol{M}(t)$ is nonsingular. □

Lemma 3.16. *Consider* (3.34). *If* $\cap_{k=0}^{N} \operatorname{Ker} \boldsymbol{A}_k$ *is the trivial vector space, if the vector space* $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$ *is non-trivial, and if there exists* $T > 0$ *such that for all* $t \in (-T, T)$, $\boldsymbol{M}(t)$ *is positive definite, then there exists a hyperplane* $\mathcal{H}$ *of* $\mathbb{R}^m$ *such that for all* $\boldsymbol{v} \in \mathbb{R}^m \setminus \mathcal{H}$,

$$\liminf_{t\to 0} \boldsymbol{v}\boldsymbol{M}(t)^{-1}\boldsymbol{v} / \left\| \boldsymbol{M}(t)^{-1} \right\| > 0. \tag{3.45}$$

Proof. This result is trivial if $N = 0$ ($\cap_{k=0}^{-1} \operatorname{Ker} \boldsymbol{A}_k$ is an intersection over an empty set, so we take it by convention to be $\mathbb{R}^m$). If $N \geqslant 1$, it follows from the proof of Lemma 3.14. Indeed, the requirements of this lemma are stronger than those of Lemma 3.14, so all intermediate results of its proof are valid. Consider Equation (3.38) while assuming $\boldsymbol{C}$ is positive definite. In the right member, the matrices on the left and on the right are the transpose of one another, so the middle matrix is necessarily positive definite. In particular, both $\boldsymbol{S}^{-1}$ and $\left(\boldsymbol{C}''\right)^{-1}$ are positive definite. Any vector $\boldsymbol{v} \in \mathbb{R}^m$ can be decomposed as $\boldsymbol{v} = (\boldsymbol{v}', \boldsymbol{v}'')^\top$ with $\boldsymbol{v}' \in \mathbb{R}^{d_N}$ and $\boldsymbol{v}'' \in \mathbb{R}^{m-d_N}$. This decomposition yields a lower bound: $\boldsymbol{v}^\top \boldsymbol{C}^{-1} \boldsymbol{v} \geqslant (\boldsymbol{v}'')^\top \left(\boldsymbol{C}''\right)^{-1} \boldsymbol{v}''$. Here, $\boldsymbol{C}$ is $\boldsymbol{M}(t)$, $\boldsymbol{S}$ is $\boldsymbol{S}_N(t)$ and $\boldsymbol{C}''$ is $a_N(t)\boldsymbol{A}''_N + \boldsymbol{B}''(t)$. Let us recall that $\boldsymbol{A}''_N$ is nonsingular and $\|\boldsymbol{B}''(t)\| = o(|a_N(t)|)$ when $t \to 0$. So as long as $\boldsymbol{v}$ is not orthogonal to $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$, $\boldsymbol{v}''$ is non-zero and there exists $\tilde{\lambda}_N(\boldsymbol{v}) > 0$ such that when $|t|$ is small enough, $\boldsymbol{v}^\top \boldsymbol{M}(t)^{-1} \boldsymbol{v} \geqslant \tilde{\lambda}_N(\boldsymbol{v}) |a_N(t)|^{-1}$. Then Lemma 3.14 yields the result for any hyperplane $\mathcal{H}$ of $\mathbb{R}^m$ that contains the orthogonal complement of $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k$. □

Lemma 3.17. *If* $\cap_{k=0}^{N} \operatorname{Ker} \boldsymbol{A}_k \neq \operatorname{Ker} \boldsymbol{A}_0$*, if* $\cap_{k=0}^{N-1} \operatorname{Ker} \boldsymbol{A}_k = \operatorname{Ker} \boldsymbol{A}_0$*, and if there exists* $T > 0$ *such that for all* $t \in (-T, T)$*,* $\boldsymbol{M}(t)$ *is positive definite, then the largest eigenvalue* $v_1(t)$ *and the second largest eigenvalue* $v_2(t)$ *of* $\boldsymbol{M}(t)$ *have the following behavior when* $t \to 0$*:*

(a) $v_1(t)^{-1} = O(|a_0(t)|^{-1})$*;*

(b) $v_2(t)^{-1} = O(|a_N(t)|^{-1})$*.*

Proof. It is equivalent to prove that in this situation, there exists $\lambda > 0$ such that when $|t|$ is sufficiently small $v_1(t) \geqslant \lambda |a_0(t)|$[and $v_2(t) \geqslant \lambda |a_N(t)|$.

When $t \to 0$, we have $a_0(t)^{-1}\boldsymbol{M}(t) \to \boldsymbol{A}_0$, so $\boldsymbol{A}_0$ is either positive or negative semi-definite. Since a_0 is continuous and non-null everywhere except possibly at 0, its sign is therefore constant: nonnegative if $\boldsymbol{A}_0$ is positive semi-definite and nonpositive if $\boldsymbol{A}_0$ is negative semi-definite. Without loss of generality, let us assume that $\boldsymbol{A}_0$ is positive semi-definite and that a_0 is nonnegative. $a_0(t)^{-1}\boldsymbol{M}(t) \to \boldsymbol{A}_0$ implies that $a_0(t)^{-1}v_1(t)$ converges to $\boldsymbol{A}_0$'s greatest eigenvalue, which is strictly greater than 0 because $\boldsymbol{A}_0$ is non-null. This implies the first result.

Now, since $\boldsymbol{A}_0$ is non-null, its rank is greater or equal to 1. If it is greater or equal to 2, then $a_0(t)^{-1}v_2(t)$ converges to the second greatest eigenvalue of $\boldsymbol{A}_0$, so $v_2^{-1}(t) = O(a_0(t)^{-1})$ and the second result holds a fortiori.

Assume from now on that $\boldsymbol{A}_0$ has rank 1. For every nonnegative integer $k < N$, $\boldsymbol{A}_k$ shares $\boldsymbol{A}_0$'s kernel, so $\boldsymbol{A}_k$ is proportional to $\boldsymbol{A}_0$. We may therefore assume without loss of generality that $N = 1$. Since $\boldsymbol{A}_0$ is a symmetric positive semi-definite matrix with rank 1, there exists a vector $\boldsymbol{a}_0$ such that $\boldsymbol{A}_0 = \boldsymbol{a}_0\boldsymbol{a}_0^\top$.

Choose for all $t \in (-T, T)$ a unit eigenvector $\boldsymbol{V}_1(t)$ corresponding to the eigenvalue $v_1(t)$ of $\boldsymbol{M}(t)$. Then choose a unit eigenvector $\boldsymbol{V}_2(t)$ corresponding to the eigenvalue $v_2(t)$ such that $\boldsymbol{V}_1(t)^\top \boldsymbol{V}_2(t) = 0$ (it is always possible to choose $\boldsymbol{V}_2(t)$ that way because $\boldsymbol{M}(t)$ is symmetric). When $t \to 0$, $\boldsymbol{V}_1(t) \to \boldsymbol{a}_0/\|\boldsymbol{a}_0\|$. Since $\boldsymbol{V}_1(t)^\top \boldsymbol{V}_2(t) = 0$ for all $t \in (-T, T)$, we have $\lim_{t\to 0} \boldsymbol{a}_0^\top \boldsymbol{V}_2(t) = 0$.

$$\begin{aligned} v_2(t) &= a_0(t) \left(\boldsymbol{a}_0^\top \boldsymbol{V}_2(t)\right)^2 + a_1(t)\boldsymbol{V}_2(t)^\top \boldsymbol{A}_1 \boldsymbol{V}_2(t) + \boldsymbol{V}_2(t)^\top \boldsymbol{B}(t) \boldsymbol{V}_2(t) \\ &\geqslant a_1(t) \left\{ \boldsymbol{V}_2(t)^\top \boldsymbol{A}_1 \boldsymbol{V}_2(t) + o(1) \right\}. \end{aligned} \tag{3.46}$$

Because $\boldsymbol{M}(t)$ is positive definite for all $t \in (-T, T)$, the restriction of $\boldsymbol{A}_1$ to $\operatorname{Ker} \boldsymbol{A}_0$ is either positive semi-definite (making a_1 nonnegative) or negative semi-definite (making a_1 nonpositive). Moreover, it is non-null.

Since $v_2(t) = \max\{\boldsymbol{\xi}\boldsymbol{M}(t)\boldsymbol{\xi} \mid \boldsymbol{\xi} \in \mathbb{R}^m \text{ and } \|\boldsymbol{\xi}\| = 1 \text{ and } \boldsymbol{\xi}^\top \boldsymbol{v}_1(t) = 0\}$, the above implies the following: $\liminf_{t\to 0} |a_1(t)|^{-1} v_2(t) > 0$. So the second result also holds when the rank of $\boldsymbol{A}_0$ is 1.

□

Appendix 3.C Spectral decomposition

For the following lemmas, we need to set up a few notations. First, denote by $\widehat{K}_r$ the r-dimensional Fourier transform of the isotropic correlation kernel K:

$$\widehat{K}_r(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} K(\|\boldsymbol{x}\|) e^{-i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{x} \quad \text{and} \quad K(\|\boldsymbol{x}\|) = \int_{\mathbb{R}^r} \widehat{K}_r(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{\omega}. \tag{3.47}$$

For all $\theta \in (0, +\infty)$, using the correlation kernel $K_\theta(\cdot) = K(\cdot/\theta)$, the correlation matrix $\boldsymbol{\Sigma}_\theta$ is such that $\forall \boldsymbol{\xi} \in \mathbb{R}^n$:

$$\boldsymbol{\xi}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\xi} = \sum_{j,k=1}^n \xi_j \xi_k K\left(\frac{\left\| \boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)} \right\|}{\theta} \right) = \int_{\mathbb{R}^r} \widehat{K_r}(\boldsymbol{\omega}) \left| \sum_{j=1}^n \xi_j e^{i \left\langle \boldsymbol{\omega} \middle| \frac{\boldsymbol{x}^{(j)}}{\theta} \right\rangle} \right|^2 d\boldsymbol{\omega} = M_r \theta^r I_\theta(\boldsymbol{\xi}). \tag{3.48}$$

The factors in the last equality depend on the kernel and are given in Table 3.4.

Kernel	M_r	$I_\theta(\boldsymbol{\xi})$
Matérn	$\frac{\Gamma(\nu+\frac{r}{2})(2\sqrt{\nu})^{2\nu}}{\pi^{\frac{r}{2}}\Gamma(\nu)}$	$\int_{\mathbb{R}^r} \left(4\nu + \theta^2 \|\boldsymbol{s}\|^2\right)^{-\frac{r}{2}-\nu} \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}$
Rational Quadratic	$\frac{2^{1-\nu}}{(2\pi)^{\frac{r}{2}}\Gamma(\nu)}$	$\int_{\mathbb{R}^r} (\theta\|\boldsymbol{s}\|)^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}}(\theta\|\boldsymbol{s}\|) \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}$
Squared Exponential	$(2\sqrt{\pi})^{-r}$	$\int_{\mathbb{R}^r} \exp\left(-\frac{\theta^2\|\boldsymbol{s}\|^2}{4}\right) \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}$

Table 3.4 – *M_r and $I_\theta(\boldsymbol{\xi})$ for the three considered correlation kernel families. $\mathcal{K}_\nu$ is the modified Bessel function of second kind with parameter ν [Abramowitz and Stegun, 1964] (9.6.)*

The spectral decomposition of correlation kernels is a powerful tool.
To use it, recall Proposition 1.31:

Lemma 3.18. *Let μ be a positive measure on $\mathbb{R}^r$ with finite non-null total mass that is absolutely continuous with respect to the Lebesgue measure. Then the mapping $K : \mathbb{R}^r \to \mathbb{R}$ defined by*

$$K(\boldsymbol{x}) = \int_{\mathbb{R}^r} e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\mu(\boldsymbol{\omega}) \tag{3.49}$$

is positive definite. Moreover, for any $\boldsymbol{\xi} \in \mathbb{R}^n \setminus \{\mathbf{0}_n\}$,

$$\sum_{k,l \in [\![1,n]\!]} \xi_k \xi_l K(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)}) > 0. \tag{3.50}$$

Let us use spectral decomposition to show this useful fact about Matérn kernels:

Lemma 3.19. *For Matérn kernels, when $\theta \to +\infty$ $\left\| \boldsymbol{\Sigma}_\theta^{-1} \right\| = O(\theta^{2\nu})$.*

Proof. When $\theta \geqslant 2\sqrt{\nu}$, for any $\boldsymbol{\xi} \in \mathbb{R}^n$,

$$I_\theta(\boldsymbol{\xi}) \geqslant 2^{-\frac{r}{2}-\nu} \theta^{-r-2\nu} \int_{\|\boldsymbol{s}\| \geqslant 1} \|\boldsymbol{s}\|^{-r-2\nu} \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}. \tag{3.51}$$

Define the mapping $K^{aux} : \mathbb{R}^r \to \mathbb{R}$ by

$$K^{aux}(\boldsymbol{x}) = \int_{\mathbb{R}^r} e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} \|\boldsymbol{s}\|^{-r-2\nu} \mathbf{1}_{\|\boldsymbol{s}\| \geqslant 1} d\boldsymbol{s}. \tag{3.52}$$

By Lemma 3.18, the $n \times n$ matrix $\boldsymbol{M}$ with (i, i')-th element $K^{aux}\left((\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')})\right)$ is positive definite. For any $\boldsymbol{\xi} \in \mathbb{R}^n$, $\int_{\|\boldsymbol{s}\| \geqslant 1} \|\boldsymbol{s}\|^{-r-2\nu} \left|\sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle}\right|^2 d\boldsymbol{s} = \boldsymbol{\xi}^\top \boldsymbol{M} \boldsymbol{\xi}$. Denote by M the smallest eigenvalue of $\boldsymbol{M}$. For any $\boldsymbol{\xi} \in \mathbb{R}^n$, when $\theta \geqslant 2\sqrt{\nu}$, $I_\theta(\boldsymbol{\xi}) \geqslant 2^{-\frac{r}{2}-\nu} M \|\boldsymbol{\xi}\|^2 \theta^{-r-2\nu}$. Equation (3.48) implies the result. □

More generally, it can be used to study the behavior of the reference prior. From Equation (3.48), we obtain that $\forall \theta \in (0, +\infty)$, $\forall \boldsymbol{\xi} \in \mathbb{R}^n$:

$$\boldsymbol{\xi}^\top \left(\frac{d}{d\theta} \boldsymbol{\Sigma}_\theta\right) \boldsymbol{\xi} = M_r r \theta^{r-1} I_\theta(\boldsymbol{\xi}) + M_r \theta^r \frac{d}{d\theta} I_\theta(\boldsymbol{\xi}). \tag{3.53}$$

The next three lemmas are used to prove the second assertion of Proposition 3.7. Since the proof varies for each of the three different kernel families considered, I_θ is written I_θ^M for Matérn kernels, I_θ^{RQ} for Rational Quadratic kernels and I_θ^{SE} for the Squared Exponential kernel.

Lemma 3.20. *For Matérn kernels, $\boldsymbol{F}_\theta := r\theta^{-1}\boldsymbol{\Sigma}_\theta - \frac{d}{d\theta}\boldsymbol{\Sigma}_\theta$ is a symmetric positive definite matrix. Furthermore, for any $\boldsymbol{\xi} \in \mathbb{R}^n$, $\boldsymbol{\xi}^\top \boldsymbol{F}_\theta \boldsymbol{\xi} \leqslant (2\nu + r)\theta^{-1} \boldsymbol{\xi}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\xi}$.*

Proof. For any $\theta \in (0, +\infty)$ and any $\boldsymbol{\xi} = (\xi_1, ..., \xi_n)^\top \in \mathbb{R}^n$,

$$\begin{aligned} \frac{d}{d\theta} I_\theta^M(\boldsymbol{\xi}) &= (-2)\left(\frac{r}{2} + \nu\right)\theta \int_{\mathbb{R}^r} \|\boldsymbol{s}\|^2 \left(4\nu + \theta^2 \|\boldsymbol{s}\|^2\right)^{-\frac{r}{2}-\nu-1} \left|\sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle}\right|^2 d\boldsymbol{s} \\ &= -(2\nu + r)\theta^{-1} \int_{\mathbb{R}^r} \frac{\theta^2 \|\boldsymbol{s}\|^2}{4\nu + \theta^2 \|\boldsymbol{s}\|^2} \left(4\nu + \theta^2 \|\boldsymbol{s}\|^2\right)^{-\frac{r}{2}-\nu} \left|\sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle}\right|^2 d\boldsymbol{s} \end{aligned} \tag{3.54}$$

Since the ratio in the integrand is smaller than 1, for any $\theta \in (0, +\infty)$ and any non-null vector $\boldsymbol{\xi} \in \mathbb{R}^n$,

$$0 < -\frac{d}{d\theta} I_\theta^M(\boldsymbol{\xi}) \leqslant (2\nu + r)\theta^{-1} I_\theta^M(\boldsymbol{\xi}) \tag{3.55}$$

Combining Equations (3.48), (3.53) and (3.55) yields the result.

□

Lemma 3.21. *For Rational Quadratic isotropic correlation kernels, $\boldsymbol{F}_\theta := r\theta^{-1}\boldsymbol{\Sigma}_\theta - \frac{d}{d\theta}\boldsymbol{\Sigma}_\theta$ is a symmetric positive definite matrix. If θ is large enough, it verifies $\forall \boldsymbol{\xi} \in \mathbb{R}^n$, $\boldsymbol{\xi}^\top \boldsymbol{F}_\theta \boldsymbol{\xi} \leqslant (r+2)\boldsymbol{\xi}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{\xi}$.*

Proof. In the following, denote by $\mathcal{K}_\nu$ the modified Bessel function of second kind with parameter ν. [Abramowitz and Stegun, 1964](9.6.)

For any $\theta \in (0, +\infty)$ and any $\boldsymbol{\xi} \in \mathbb{R}^n$,

$$\frac{d}{d\theta} I_\theta^{RQ}(\boldsymbol{\xi}) = \int_{\mathbb{R}^r} \|\boldsymbol{s}\| \frac{d}{dz} \left\{ z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}}(z) \right\}_{z=\theta\|\boldsymbol{s}\|} \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}. \tag{3.56}$$

We now compute $\frac{d}{dz}\left\{ z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}} z \right\}$. Following Abramowitz and Stegun [1964] (9.6.28.),

$$\frac{d}{dz}\left\{ z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}} z \right\} = \begin{cases} -z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}-1}(z) & \text{if } \nu - \frac{r}{2} \geqslant 0. \\ -z^{\nu-\frac{r}{2}} \mathcal{K}_{\frac{r}{2}-\nu-1}(z) + (2\nu - r) z^{\nu-\frac{r}{2}-1} \mathcal{K}_{\frac{r}{2}-\nu}(z) & \text{if } \nu - \frac{r}{2} < 0. \end{cases} \tag{3.57}$$

Combining this with Equations (3.48) and (3.53) proves that $\boldsymbol{F}_\theta$ is positive definite.

We now have to deal with the behavior of $\mathcal{K}_{\left|\nu-\frac{r}{2}\right|-1}(z)$ when $z \to 0$ and when $z \to +\infty$. Let us start with $z \to 0$. Using Abramowitz and Stegun [1964] (9.6.9.), we obtain:

$$\mathcal{K}_{\left|\nu-\frac{r}{2}\right|-1}(z) \sim \begin{cases} \frac{z}{2(\left|\nu-\frac{r}{2}\right|-1)} \mathcal{K}_{\nu-\frac{r}{2}}(z) & \text{if } \left|\nu - \frac{r}{2}\right| > 1. \\ -z\log(z) \mathcal{K}_{\nu-\frac{r}{2}}(z) & \text{if } \left|\nu - \frac{r}{2}\right| = 1. \\ \frac{\Gamma\left(1-\left|\nu-\frac{r}{2}\right|\right)}{2^{|2\nu-r|-1}\Gamma\left(\left|\nu-\frac{r}{2}\right|\right)} z^{|2\nu-r|-1} \mathcal{K}_{\nu-\frac{r}{2}}(z) & \text{if } 0 < \left|\nu - \frac{r}{2}\right| < 1. \\ -\frac{1}{z\log(z)} \mathcal{K}_{\nu-\frac{r}{2}}(z) & \text{if } \left|\nu - \frac{r}{2}\right| = 0. \end{cases} \tag{3.58}$$

So, for any $\nu > 0$, there exists $a_{r,\nu} > 0$ such that, as long as z is small enough, $\mathcal{K}_{\left|\nu-\frac{r}{2}\right|-1}(z) \leqslant a_{r,\nu} z^{-1} \mathcal{K}_{\nu-\frac{r}{2}}(z)$.
Moreover, Abramowitz and Stegun [1964] (9.7.2.) states that when $z \to +\infty$, $\mathcal{K}_{\left|\nu-\frac{r}{2}\right|-1}(z) \sim \mathcal{K}_{\nu-\frac{r}{2}}(z) \sim \exp(-z)\sqrt{\pi}/\sqrt{2z}$.
Because $\mathcal{K}_{\left|\nu-\frac{r}{2}\right|-1}$ is a continuous function on $(0, +\infty)$, the two results above imply that

$$\forall \lambda > 1 \quad \exists a'_{r,\nu} > 0 \quad \forall z > 0 \quad z\mathcal{K}_{\left|\nu-\frac{r}{2}\right|-1} \leqslant \max(a'_{r,\nu}, \lambda z) \mathcal{K}_{\nu-\frac{r}{2}}(z). \tag{3.59}$$

Now, $\frac{d}{d\theta} I_\theta^{RQ}(\boldsymbol{\xi}) = J_\theta^1(\boldsymbol{\xi}) + J_\theta^2(\boldsymbol{\xi})$, with

$$\begin{aligned} J_\theta^1(\boldsymbol{\xi}) &:= \int_{\|s\| \leqslant 1} \|\boldsymbol{s}\| \frac{d}{dz} \left\{ z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}}(z) \right\}_{z=\theta\|\boldsymbol{s}\|} \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}; \\ J_\theta^2(\boldsymbol{\xi}) &:= \int_{\|s\| > 1} \|\boldsymbol{s}\| \frac{d}{dz} \left\{ z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}}(z) \right\}_{z=\theta\|\boldsymbol{s}\|} \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s}. \end{aligned} \tag{3.60}$$

Set $0 < \epsilon < 1$. When $z \in [\epsilon, 1]$, $\frac{d}{dz}\left\{ z^{\nu-\frac{r}{2}} \mathcal{K}_{\nu-\frac{r}{2}}(z) \right\}$ is bounded away from 0, so there exists m_ϵ such that

$$\theta |J_\theta^1(\boldsymbol{\xi})| \geqslant m_\epsilon \int_{\frac{\epsilon}{\theta} < \|\boldsymbol{s}\| < \frac{1}{\theta}} \left| \sum_{j=1}^n \xi_j e^{i\langle \boldsymbol{s} | \boldsymbol{x}^{(j)} \rangle} \right|^2 d\boldsymbol{s} = m_\epsilon \theta^{-r} \int \mathbf{1}_{\epsilon < \|\boldsymbol{\omega}\| < 1} \left| \sum_{j=1}^n \xi_j e^{i\left\langle \boldsymbol{\omega} \middle| \frac{\boldsymbol{x}^{(j)}}{\theta} \right\rangle} \right|^2 d\boldsymbol{\omega}. \tag{3.61}$$

The Lebesgue measure on $\{\boldsymbol{\omega} \in \mathbb{R}^r : \epsilon < \|\boldsymbol{\omega}\| < 1\}$ is a finite positive measure. Lemma 3.18 asserts that the mapping $K^\epsilon : \mathbb{R}^r \to \mathbb{R}$ defined by

$$K^\epsilon(\boldsymbol{x}) = \int_{\epsilon < \|\boldsymbol{\omega}\| < 1} e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{\omega} \tag{3.62}$$

is positive definite. It is an isotropic covariance kernel: K^ϵ only depends on $\boldsymbol{x}$ through its norm $\|\boldsymbol{x}\|$. Let $\boldsymbol{\Sigma}_\theta^\epsilon$ be the correlation matrix corresponding with K^ϵ: its (i,i')-th element is $K^\epsilon\left((\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(i')})/\theta\right)$. The Lebesgue measure on $\{\boldsymbol{\omega}\in\mathbb{R}^r : \epsilon<\|\boldsymbol{\omega}\|<1\}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^r$, so Lemma 3.18 also asserts that $\boldsymbol{\Sigma}_\theta^\epsilon$ is positive definite.

Moreover, for every nonnegative integer k, $\int\|\boldsymbol{\omega}\|^k\mathbf{1}_{\epsilon<\|\boldsymbol{\omega}\|<1}d\boldsymbol{\omega}$ is smaller than the mass of the Lebesgue measure on $\{\boldsymbol{\omega}\in\mathbb{R}^r : \epsilon<\|\boldsymbol{\omega}\|<1\}$. The Maclaurin series of K^ϵ has therefore infinite radius of convergence. For any nonnegative integer k, denote by $\boldsymbol{D}^{(k)}$ the $n\times n$ matrix with (i,i')-th element $\left\|\boldsymbol{x}^{(i)}-\boldsymbol{x}^{(i')}\right\|^{2k}$. Because the Maclaurin series has infinite radius of convergence, $\boldsymbol{\Sigma}_\theta^\epsilon$ is equal to its asymptotic expansion regardless of the value of θ. There exist real numbers a_k $(k\in\mathbb{N})$ such that for all $\theta\in(0,+\infty)$, $\boldsymbol{\Sigma}_\theta^\epsilon=\sum_{k=0}^\infty a_k\theta^{-2k}\boldsymbol{D}^{(k)}$. Because $\boldsymbol{\Sigma}_\theta^\epsilon$ is positive definite, Lemma 3.13 ensures there exists a nonnegative integer N such that the vector space $\cap_{k=0}^N\operatorname{Ker}a_k\boldsymbol{D}^{(k)}$ is trivial.

Applying Lemma 3.14 then yields that when $\theta\to+\infty$ $\left\|(\boldsymbol{\Sigma}_\theta^\epsilon)^{-1}\right\|=O(\theta^{2N})$. Because the greatest eigenvalue of a positive definite matrix is the smallest eigenvalue of its inverse, this implies the existence of a constant $c_\epsilon>0$ such that when θ is large enough

$$\min_{\boldsymbol{\xi}\in\mathbb{R}^n,\|\boldsymbol{\xi}\|=1}\boldsymbol{\xi}^\top\boldsymbol{\Sigma}_\theta^\epsilon\boldsymbol{\xi}\geqslant c_\epsilon\theta^{-2N}. \tag{3.63}$$

So when θ is large enough, for every $\boldsymbol{\xi}\in\mathbb{R}^n$,

$$\int\mathbf{1}_{\epsilon<\|\boldsymbol{\omega}\|<1}\left|\sum_{j=1}^n\xi_j e^{i\left\langle\boldsymbol{\omega}\middle|\frac{\boldsymbol{x}^{(j)}}{\theta}\right\rangle}\right|^2 d\boldsymbol{\omega}\geqslant c_\epsilon\|\boldsymbol{\xi}\|^2\theta^{-2N}. \tag{3.64}$$

This provides a lower bound for $|J_\theta^1(\boldsymbol{\xi})|$:

$$\theta|J_\theta^1(\boldsymbol{\xi})|\geqslant\tilde{m}_\epsilon c_\epsilon\|\boldsymbol{\xi}\|^2\theta^{-r-2N}. \tag{3.65}$$

Besides, we have when θ is large enough, for every $\boldsymbol{\xi}\in\mathbb{R}^n$,

$$\begin{aligned}\theta|J_\theta^2(\boldsymbol{\xi})|&\leqslant n^2\|\boldsymbol{\xi}\|^2\int_{\|\boldsymbol{s}\|>1}(\theta\|\boldsymbol{s}\|)^{\nu-\frac{r}{2}-\frac{1}{2}}\exp(-\theta\|\boldsymbol{s}\|)d\boldsymbol{s}\\&=n^2\|\boldsymbol{\xi}\|^2\frac{2\pi^{\frac{r-1}{2}}}{\Gamma\left(\frac{r-1}{2}\right)}\int_1^{+\infty}(\theta t)^{\nu-\frac{r}{2}-\frac{1}{2}}\exp(-\theta t)t^{r-1}dt\\&\leqslant n^2\|\boldsymbol{\xi}\|^2\frac{2\pi^{\frac{r-1}{2}}}{\Gamma\left(\frac{r-1}{2}\right)}\Gamma\left(\nu+\frac{r-1}{2}\right)\theta^{\nu-\frac{r}{2}-\frac{1}{2}}\exp(-(\theta-1)).\end{aligned} \tag{3.66}$$

From Equations (3.65) and (3.66), we gather that when $\theta\to+\infty$, $\sup_{\|\boldsymbol{\xi}\|=1}|J_\theta^2(\boldsymbol{\xi})|=o\left(\inf_{\|\boldsymbol{\xi}\|=1}|J_\theta^1(\boldsymbol{\xi})|\right)$, so for any $\lambda>1$, when θ is large enough, $-\frac{d}{d\theta}I_\theta^{RQ}(\boldsymbol{\xi})\leqslant-\lambda J_\theta^1(\boldsymbol{\xi})$.

Denote by $(r-2\nu)_+$ the quantity $\max(0, r-2\nu)$. Then, combining Equations (3.57) and (3.59), there exists $a'_{r,\nu} > 0$ such that

$$\begin{aligned}
&-\theta\frac{d}{d\theta}I_\theta^{RQ}(\boldsymbol{\xi}) \\
\leqslant&\lambda\int_{\|\boldsymbol{s}\|\leqslant 1}\max\left(a'_{r,\nu},(\lambda+(r-2\nu)_+)\theta\|\boldsymbol{s}\|\right)(\theta\|\boldsymbol{s}\|)^{\nu-\frac{r}{2}}\mathcal{K}_{\nu-\frac{r}{2}}(\theta\|\boldsymbol{s}\|)\left|\sum_{j=1}^n\xi_j e^{i\langle\boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s} \\
\leqslant&\lambda\max\left(a'_{r,\nu},(\lambda+(r-2\nu)_+)\theta\right)\int_{\|\boldsymbol{s}\|\leqslant 1}(\theta\|\boldsymbol{s}\|)^{\nu-\frac{r}{2}}\mathcal{K}_{\nu-\frac{r}{2}}(\theta\|\boldsymbol{s}\|)\left|\sum_{j=1}^n\xi_j e^{i\langle\boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s}.
\end{aligned} \tag{3.67}$$

When θ is large enough, $a'_{r,\nu} \leqslant \lambda\theta$ so

$$-\theta\frac{d}{d\theta}I_\theta^{RQ}(\boldsymbol{\xi}) \leqslant \lambda(\lambda+(r-2\nu)_+)\theta I_\theta^{RQ}(\boldsymbol{\xi}). \tag{3.68}$$

From this, we obtain that for any non-null vector $\boldsymbol{\xi}\in\mathbb{R}^n$, for any $\lambda > 1$, provided θ is large enough,

$$0 < -\frac{d}{d\theta}I_\theta^{RQ}(\boldsymbol{\xi}) \leqslant \lambda(\lambda+(r-2\nu)_+)I_\theta^{RQ}(\boldsymbol{\xi}). \tag{3.69}$$

Combining Equations (3.48), (3.53) and (3.69) yields the result.

□

Lemma 3.22. *For the Squared Exponential kernel, $\boldsymbol{F}_\theta := r\theta^{-1}\boldsymbol{\Sigma}_\theta - \frac{d}{d\theta}\boldsymbol{\Sigma}_\theta$ is a symmetric positive definite matrix. If θ is large enough, it verifies $\forall\boldsymbol{\xi}\in\mathbb{R}^n$, $\boldsymbol{\xi}^\top\boldsymbol{F}_\theta\boldsymbol{\xi} \leqslant \theta\boldsymbol{\xi}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{\xi}$.*

Proof. For any $\theta\in(0,+\infty)$ and any $\boldsymbol{\xi}\in\mathbb{R}^n$,

$$\frac{d}{d\theta}I_\theta^{SE}(\boldsymbol{\xi}) = \int_{\mathbb{R}^r} -\frac{\theta\|\boldsymbol{s}\|^2}{2}\exp\left(-\frac{\theta^2\|\boldsymbol{s}\|^2}{4}\right)\left|\sum_{j=1}^n\xi_j e^{i\langle\boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s} \tag{3.70}$$

So $\boldsymbol{F}_\theta$ is positive definite.

Similarly to the Rational Quadratic case (cf. proof of Lemma 3.21), one can show that for any $\lambda > 1$, for large enough θ and for any non-null vector $\boldsymbol{\xi}\in\mathbb{R}^n$,

$$0 < -\frac{d}{d\theta}I_\theta^{SE}(\boldsymbol{\xi}) \leqslant \lambda\int_{\|\boldsymbol{s}\|\leqslant 1}\frac{\theta\|\boldsymbol{s}\|^2}{2}\exp\left(-\frac{\theta^2\|\boldsymbol{s}\|^2}{4}\right)\left|\sum_{j=1}^n\xi_j e^{i\langle\boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s} \leqslant \frac{\lambda}{2}\theta I_\theta^{SE}(\boldsymbol{\xi}). \tag{3.71}$$

Combining Equations (3.48), (3.53) and (3.71) yields the result.

□

Appendix 3.D Asymptotic study of the correlation matrix $\boldsymbol{\Sigma}_\theta$

Rational Quadratic and Squared Exponential kernels

For all $\nu > 0$, the series expansion of the mapping $x \mapsto (1+x)^{-\nu}$ at $x = 0$ has radius of convergence 1. Moreover, the series expansion of the exponential function has infinite radius

of convergence. From these facts follows that when θ is large enough, if a Rational Quadratic kernel or a Squared Exponential kernel is used,

$$\boldsymbol{\Sigma}_\theta = \sum_{k=0}^{\infty} \frac{a_k}{\theta^{2k}} \boldsymbol{D}^{(k)}. \tag{3.72}$$

In the above expression, for every k, $\boldsymbol{D}^{(k)}$ is the $n \times n$ matrix with (i, i')-th element $\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|^{2k}$ and a_k is a non-null real number. To be precise, $a_k = (-1)^k \left(\prod_{l=0}^{k}(\nu + l)\right)/k!$ for Rational Quadratic kernels and $a_k = (-1)^k/k!$ for the Squared Exponential kernel.

Equation (3.72) implies

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} = \sum_{k=0}^{\infty} \frac{a_k}{\theta^{2k}} \boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W}. \tag{3.73}$$

$\boldsymbol{\Sigma}_\theta$ is positive definite and the kernel of $\boldsymbol{W}$ is trivial so $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$ is positive definite. Let k_1 be the smallest nonnegative integer such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is non-null. Define $\boldsymbol{D} := a_{k_1} \boldsymbol{D}^{(k_1)}$. If $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is nonsingular, then define $k_2 := k_1 + 1$ and $\boldsymbol{D}^\star := a_{k_2} \boldsymbol{D}^{(k_2)}$. If $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is singular, then because $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}$ is nonsingular, there must exist an integer $k > k_1$ such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W}$ is non-null. Then let k_2 be the smallest of them and define $\boldsymbol{D}^\star := a_{k_2} \boldsymbol{D}^{(k_2)}$. Now, define the mappings $g(\theta) = \theta^{-2k_1}$ and $g^\star(\theta) = \theta^{-2(k_2 - k_1)}$.
Finally, define

$$\boldsymbol{R}_g(\theta) = g(\theta)^{-1} \sum_{k=k_2+1}^{\infty} \frac{a_k}{\theta^{2k}} \boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W}. \tag{3.74}$$

Notice that $\|\boldsymbol{R}_g(\theta)\| = o(g^\star(\theta))$ and that $\|\frac{d}{d\theta} \boldsymbol{R}_g(\theta)\| = o(g^{\star\prime}(\theta))$.

Matérn kernels with noninteger smoothness ν

If a Matérn kernel with noninteger smoothness $\nu > 0$ (whether greater or smaller than 1) is used, we can write $\boldsymbol{\Sigma}_\theta$ as [Abramowitz and Stegun, 1964] (9.6.2. and 9.6.10.):

$$\boldsymbol{\Sigma}_\theta = \sum_{k=0}^{\lfloor \nu \rfloor} \frac{a_k}{\theta^{2k}} \boldsymbol{D}^{(k)} + \frac{a_\nu}{\theta^{2\nu}} \boldsymbol{D}^{(\nu)} + \boldsymbol{R}(\theta). \tag{3.75}$$

Like in the case of Rational Quadratic and Squared Exponential kernels, for every k, $\boldsymbol{D}^{(k)}$ is the $n \times n$ matrix with (i, i')-th element $\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|^{2k}$. The a_k's, of course, are different: $a_k = (-1)^k \Gamma(\nu - k) \nu^k / (k! \Gamma(\nu))$. Moreover, $\boldsymbol{D}^{(\nu)}$ is the $n \times n$ matrix with (i, i')-th element $\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|^{2\nu}$, $a_\nu = \Gamma(-\nu)\nu^\nu / \Gamma(\nu)$ and $\boldsymbol{R}$ is a differentiable mapping from $(0, +\infty)$ to the set of real $n \times n$ matrices $\mathcal{M}_n$ such that $\|\boldsymbol{R}(\theta)\| = O(\theta^{-2(\lfloor \nu \rfloor + 1)})$ and $\|\frac{d}{d\theta} \boldsymbol{R}(\theta)\| = O(\theta^{-2(\lfloor \nu \rfloor + 1) - 1})$. Lemma (3.19) implies that when θ is large enough $\boldsymbol{\Sigma}_\theta - \boldsymbol{R}(\theta)$ is positive definite.

Equation (3.75) implies

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} = \sum_{k=0}^{\lfloor \nu \rfloor} \frac{a_k}{\theta^{2k}} \boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W} + \frac{a_\nu}{\theta^{2\nu}} \boldsymbol{W}^\top \boldsymbol{D}^{(\nu)} \boldsymbol{W} + \boldsymbol{W}^\top \boldsymbol{R}(\theta) \boldsymbol{W}. \tag{3.76}$$

When θ is large enough, $\boldsymbol{\Sigma}_\theta - \boldsymbol{R}(\theta)$ is positive definite. Since the kernel of $\boldsymbol{W}$ is trivial, when θ is large enough, $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} - \boldsymbol{W}^\top \boldsymbol{R}(\theta) \boldsymbol{W}$ is positive definite. If it exists, let k_1 be

the smallest nonnegative integer smaller than ν such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is non-null and define $\boldsymbol{D} := a_{k_1} \boldsymbol{D}^{(k_1)}$ and $g(\theta) := \theta^{-2k_1}$. If not, then define $\boldsymbol{D} := a_\nu \boldsymbol{D}^{(\nu)}$ and $g(\theta) := \theta^{-2\nu}$. Either way $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is non-null.

If k_1 exists and $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is nonsingular, then define $k_2 := k_1 + 1$ if $k_1 < \lfloor \nu \rfloor$ and $k_2 = \nu$ if $k_1 = \lfloor \nu \rfloor$. Then define $\boldsymbol{D}^\star := a_{k_2} \boldsymbol{D}^{(k_2)}$ and $g^\star(\theta) := g(\theta)^{-1} \theta^{-2k_2}$.

If k_1 exists and $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is singular, then there must exist $k \in [\![k_1 + 1, \lfloor \nu \rfloor]\!] \cup \{\nu\}$ such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W}$ is non-null. Let k_2 be the smallest number among all such k. Define $\boldsymbol{D}^\star := a_{k_2} \boldsymbol{D}^{(k_2)}$ and $g^\star(\theta) := g(\theta)^{-1} \theta^{-2k_2}$.

If k_1 does not exist, then $\boldsymbol{W}^\top \boldsymbol{D}^{(\nu)} \boldsymbol{W}$ is necessarily nonsingular. Define $\boldsymbol{D}^\star$ as the null $n \times n$ matrix and $g^\star(\theta) = g(\theta)^{-1} \theta^{-2\nu - (\lfloor \nu \rfloor + 1 - \nu)}$.

Finally, define

$$\boldsymbol{R}_g(\theta) := g(\theta)^{-1} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} - g(\theta) \boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W} - g(\theta) g^\star(\theta) \boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} \right). \tag{3.77}$$

In all situations, $\|\boldsymbol{R}_g(\theta)\| = o(g^\star(\theta))$ and $\|\frac{d}{d\theta} \boldsymbol{R}_g(\theta)\| = o(g^{\star\prime}(\theta))$.

Matérn kernels with integer smoothness ν

Finally, if a Matérn kernel with integer smoothness ν is used, we can write $\boldsymbol{\Sigma}_\theta$ as [Abramowitz and Stegun, 1964] (9.6.11.):

$$\boldsymbol{\Sigma}_\theta = \sum_{k=0}^{\nu-1} \frac{a_k}{\theta^{2k}} \boldsymbol{D}^{(k)} + \tilde{a}_\nu \left(\frac{\log(\theta)}{\theta^{2\nu}} \boldsymbol{D}^{(\nu)} + \frac{1}{\theta^{2\nu}} \tilde{\boldsymbol{D}}^{(\nu)} \right) + \boldsymbol{R}(\theta). \tag{3.78}$$

a_k and $\boldsymbol{D}^{(k)}$ ($k \in [\![0, \nu - 1]\!]$) and $\boldsymbol{D}^{(\nu)}$ have the same definitions as for Matérn kernels with noninteger smoothness $\nu > 1$.
$\tilde{a}_\nu = (-1)^\nu 2\nu^\nu / (\nu - 1)!$ and $\tilde{\boldsymbol{D}}^{(\nu)}$ is the $n \times n$ matrix with null diagonal and (i, i')-th element ($i \neq i'$) given by $\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|^{2\nu} \left\{ -0.5 \log\left(\|\boldsymbol{x}^{(i)} - \boldsymbol{x}^{(i')}\|^2 \right) - 0.5 \log(\nu) - 2\gamma + \sum_{l=1}^{\nu} l^{-1} \right\}$, where γ is Euler's constant.
Finally, $\boldsymbol{R}$ is a differentiable mapping from $(0, +\infty)$ to the set of real $n \times n$ matrices $\mathcal{M}_n$ such that $\|\boldsymbol{R}(\theta)\| = O(\log(\theta) \theta^{-2(\nu+1)})$ and $\|\frac{d}{d\theta} \boldsymbol{R}(\theta)\| = O(\log(\theta) \theta^{-2(\nu+1)-1})$.

Equation (3.78) implies

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} = \sum_{k=0}^{\nu-1} \frac{a_k}{\theta^{2k}} \boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W} + \frac{\log(\theta)}{\theta^{2\nu}} \tilde{a}_\nu \boldsymbol{W}^\top \boldsymbol{D}^{(\nu)} \boldsymbol{W} + \frac{\tilde{a}_\nu}{\theta^{2\nu}} \boldsymbol{W}^\top \tilde{\boldsymbol{D}}^{(\nu)} \boldsymbol{W} + \boldsymbol{W}^\top \boldsymbol{R}(\theta) \boldsymbol{W}. \tag{3.79}$$

When θ is large enough, $\boldsymbol{\Sigma}_\theta - \boldsymbol{R}(\theta)$ is positive definite. Since the kernel of $\boldsymbol{W}$ is trivial, when θ is large enough, $\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} - \boldsymbol{W}^\top \boldsymbol{R}(\theta) \boldsymbol{W}$ is positive definite. If it exists, let k_1 be the smallest nonnegative integer smaller or equal to ν such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is non-null and define $\boldsymbol{D} := a_{k_1} \boldsymbol{D}^{(k_1)}$ and $g(\theta) = \theta^{-2k_1}$ ($k_1 < \nu$) or $\boldsymbol{D} := \tilde{a}_\nu \boldsymbol{D}^{(\nu)}$ and $g(\theta) := \log(\theta) \theta^{-2\nu}$ ($k_1 = \nu$). If not, then define $\boldsymbol{D} := \tilde{a}_\nu \tilde{\boldsymbol{D}}^{(\nu)}$ and $g(\theta) := \theta^{-2\nu}$. Either way $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is non-null.

If $\boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W}$ is nonsingular, then

— either k_1 exists and $k_1 < \nu - 1$, in which case define $\boldsymbol{D}^\star := a_{k_1+1} \boldsymbol{D}^{(k_1+1)}$ and $g^\star(\theta) := \theta^{-2}$;

— or k_1 exists and $k_1 = \nu - 1$, in which case define $\boldsymbol{D}^\star := \tilde{a}_\nu \boldsymbol{D}^{(\nu)}$ and $g^\star(\theta) := \log(\theta)\theta^{-2}$;
— or k_1 exists and $k_1 = \nu$, in which case define $\boldsymbol{D}^\star := \tilde{a}_\nu \tilde{\boldsymbol{D}}^{\nu}$ and $g^\star(\theta) := \log(\theta)^{-1}$;
— or k_1 does not exist, in which case define $\boldsymbol{D}^\star$ as the null $n \times n$ matrix and $g^\star(\theta) := \theta^{-1}$.

If $\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{W}$ is singular, then k_1 necessarily exists:
— either $k_1 < \nu$. Then there are two possibilities. The first is that there exists a smallest integer $k_2 \in [\![k_1 + 1, \nu]\!]$ such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_2)} \boldsymbol{W}$ is non-null, in which case define $\boldsymbol{D}^\star := a_{k_2} \boldsymbol{D}^{(k_2)}$ and $g^\star(\theta) := \theta^{-2(k_2 - k_1)}$ ($k_2 < \nu$) or $\boldsymbol{D}^\star := \tilde{a}_\nu \boldsymbol{D}^{(\nu)}$ and $g^\star(\theta) := \log(\theta)\theta^{-2(\nu - k_1)}$ ($k_2 = \nu$). The second is that no such k_2 exists, but then $\boldsymbol{W}^\top \tilde{\boldsymbol{D}}^{(\nu)} \boldsymbol{W}$ is necessarily non-null, so define $\boldsymbol{D}^\star := \tilde{a}_\nu \tilde{\boldsymbol{D}}^{(\nu)}$ and $g^\star(\theta) := \theta^{-2(\nu - k_1)}$.
— or $k_1 = \nu$. Then $\boldsymbol{W}^\top \tilde{\boldsymbol{D}}^{(\nu)} \boldsymbol{W}$ is necessarily non-null, so define $\boldsymbol{D}^\star := \tilde{a}_\nu \tilde{\boldsymbol{D}}^{(\nu)}$ and $g^\star(\theta) := \log(\theta)^{-1}$.

Finally, define

$$\boldsymbol{R}_g(\theta) := g(\theta)^{-1} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} - g(\theta) \boldsymbol{W}^\top \boldsymbol{D} \boldsymbol{W} - g(\theta) g^\star(\theta) \boldsymbol{W}^\top \boldsymbol{D}^\star \boldsymbol{W} \right). \tag{3.80}$$

In all situations, $\|\boldsymbol{R}_g(\theta)\| = o(g^\star(\theta))$ and $\|\frac{d}{d\theta} \boldsymbol{R}_g(\theta)\| = o(g^{\star\prime}(\theta))$.

Proof of Lemma 3.8

For Rational Quadratic and Squared Exponential kernels, Equation (3.72) implies thanks to Lemma 3.13 that there exists $k' \in \mathbb{N}$ such that $\cap_{k=0}^{k'} \operatorname{Ker} \left(\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W} \right)$ is the trivial vector space and $\cap_{0 \leqslant k < k'} \operatorname{Ker} \left(\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W} \right)$ is a non-trivial vector space (if $k' = 0$, the intersection is done over an empty index set, so we take it to be $\mathbb{R}^{n-p}$ by convention).

Lemma 3.14 implies that there exists a constant $c_{k'} > 0$ such that for large enough θ, $v_{n-p}(\theta) \geqslant c_{k'} \theta^{-2k'}$. Thanks to Lemma 3.16, there exists a hyperplane $\mathcal{H}_{n-p}$ of $\mathbb{R}^{n-p}$ such that for every $\boldsymbol{y}' \in \mathbb{R}^{n-p} \setminus \mathcal{H}_{n-p}$, there exists $c_{\boldsymbol{y}'} > 0$ such that for large enough θ,

$$(\boldsymbol{y}')^\top \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} \right)^{-1} \boldsymbol{y}' \geqslant c_{\boldsymbol{y}'} \left\| \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} \right)^{-1} \right\|. \tag{3.81}$$

So for every $\boldsymbol{y} \in \mathbb{R}^n$ such that $\boldsymbol{W}^\top \boldsymbol{y} \in \mathbb{R}^{n-p} \setminus \mathcal{H}_{n-p}$, there exists $c_{\boldsymbol{y}} > 0$ such that for large enough θ

$$\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{y} \geqslant c_{\boldsymbol{y}} \left\| \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} \right)^{-1} \right\|. \tag{3.82}$$

Because the matrix $\boldsymbol{W}^\top$ has full row rank, the vector space of all $\boldsymbol{v} \in \mathbb{R}^n$ such that $\boldsymbol{W}^\top \boldsymbol{v} \in \mathcal{H}_{n-p}$ is included within a hyperplane $\mathcal{H}_n$ of $\mathbb{R}^n$, so for every $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}_n$, there exists $c_{\boldsymbol{y}} > 0$ such that for large θ the above equation holds.

For Matérn kernels with noninteger smoothness $\nu > 0$ (resp. with integer smoothness $\nu > 0$), Equation (3.75) (resp. Equation (3.78)) allows a similar argument. Indeed, Lemma 3.19 asserts that $\left\| \boldsymbol{\Sigma}_\theta^{-1} \right\| = O(\theta^{2\nu})$, so $\cap_{k=0}^{\lfloor \nu \rfloor} \operatorname{Ker} \left(\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W} \right) \cap \operatorname{Ker} \left(\boldsymbol{W}^\top \boldsymbol{D}^{(\nu)} \boldsymbol{W} \right)$ (resp. $\cap_{k=0}^{\nu} \operatorname{Ker} \left(\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W} \right) \cap \operatorname{Ker} \left(\boldsymbol{W}^\top \tilde{\boldsymbol{D}}^{(\nu)} \boldsymbol{W} \right)$) is necessarily the trivial vector space.

Appendix 3.E Details of the proof of Theorem 3.9

Here the last part of the proof of Theorem 3.9 is given in detail.

Rational Quadratic, Squared Exponential and Matérn ($\nu \in [1, +\infty) \setminus \mathbb{Z}_+$) kernels

Let us first tackle the case of Rational Quadratic and Squared Exponential kernels and of Matérn kernels with noninteger smoothness $\nu > 1$.

In case 1. (a), Lemma 3.14 yields $\tilde{w}(\theta) = O(g^{\star\prime}(\theta))$.
This implies $\pi(\theta) = O(g^{\star\prime}(\theta)) = O(\theta^{-2l-1})$, so the reference prior is proper. Given the likelihood function is bounded (cf. Equation 3.12), the reference posterior is proper as well.

In case 1. (b), Lemma 3.14 yields $\tilde{w}(\theta) = O(g^{\star\prime}(\theta)g^{\star}(\theta)^{-1})$.
This implies $\pi(\theta) = O(g^{\star\prime}(\theta)g^{\star}(\theta)^{-1}) = O(\theta^{-1})$. Moreover, $v_{n-p}(\theta) = O(g(\theta)g^{\star}(\theta))$. As the rank of $\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{W}$ is at least one, $v_1(\theta)^{-1} = O(g(\theta)^{-1})$. Gathering all this, $v_{n-p}(\theta)/v_1(\theta) = O(g^{\star}(\theta))$, so Equation (3.12) implies $L(\boldsymbol{y}|\theta) = O(g^{\star}(\theta))^{1/2}) = O(\theta^{-l})$. The reference posterior is then proportional to $L(\boldsymbol{y}|\theta)\pi(\theta) = O(\theta^{-l-1})$ and is proper.

In case 2., we must distinguish between Matérn kernels and the others. For Matérn kernels with noninteger smoothness $\nu > 1$, Propostion 3.7 asserts that the reference prior is $O(\theta^{-1})$ so the argument used in case 1.(b) still holds. For Rational Quadratic and Squared Exponential kernels, Equation (3.72) implies

$$\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W} = \sum_{k=0}^{\infty} \frac{a_k}{\theta^{2k}} \boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W}. \tag{3.83}$$

Let k_1 be the smallest nonnegative integer such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k)} \boldsymbol{W}$ is not the null matrix. Then

$$g(\theta)\boldsymbol{W}^\top \boldsymbol{D}\boldsymbol{W} = a_{k_1}\theta^{-2k_1} \boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}. \tag{3.84}$$

and for some integer $k_2 > k_1$,

$$g(\theta)g^{\star}(\theta)\boldsymbol{W}^\top \boldsymbol{D}^{\star}\boldsymbol{W} = a_{k_2}\theta^{-2k_2} \boldsymbol{W}^\top \boldsymbol{D}^{(k_2)} \boldsymbol{W}. \tag{3.85}$$

Things are easiest if $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)} \boldsymbol{W}$ is null, because then $k_2 > k_1+1$. Since we are dealing with case 2., $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{\star} \boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}\right)$ is not the trivial vector space so Equation (3.83) yields $v_{n-p}(\theta) = O(\theta^{-2(k_2+1)}) = O(\theta^{-2(k_1+3)})$. Besides, the smallest eigenvalue of $\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1}$ verifies $v_1(\theta)^{-1} = O(\theta^{-2k_1})$ so $v_{n-p}(\theta)/v_1(\theta) = O(\theta^{-6})$. Recall Proposition 3.7 asserts that $\pi(\theta) = O(\theta)$. The reference posterior is proportional to $L(\boldsymbol{y}|\theta)\pi(\theta) = O(\theta^{-2})$ and thus proper.

In the following, assume $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)} \boldsymbol{W}$ is not null. Then $k_2 = k_1 + 1$.

If we assume that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ has rank greater or equal to 2, then a similar reasoning can be applied. The two smallest eigenvalues of $\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_\theta \boldsymbol{W}\right)^{-1}$ verify $v_1(\theta)^{-1} = O(\theta^{-2k_1})$ and $v_2(\theta)^{-1} = O(\theta^{-2k_1})$. $v_{n-p}(\theta) = O(\theta^{-2(k_2+1)}) = O(\theta^{-2(k_1+2)})$. From this we obtain $v_{n-p}(\theta)/v_1(\theta) = O(\theta^{-4})$ and $v_{n-p}(\theta)/v_2(\theta) = O(\theta^{-4})$. Equation (3.12) then implies $L(\boldsymbol{y}|\theta) = O(\theta^{-2-2})$. The reference posterior is proportional to $L(\boldsymbol{y}|\theta)\pi(\theta) = O(\theta^{-3})$ and thus proper.

Now, assume that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ has rank 1. We need to distinguish between two possibilites: either $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+2)} \boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)} \boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}\right)$ is the trivial vector space, or it is not. If it is not, Equation (3.83) yields $v_{n-p}(\theta) = O(\theta^{-2(k_1+3)})$ and the conclusion is the same as in the case where $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)} \boldsymbol{W}$ is null.

Let us now deal with the situation where $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+2)}\boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$ is the trivial vector space. Two further subcases must be distinguished here: either $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$ is equal to $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$, or it is strictly included within $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$.

If it is strictly included, then Lemma 3.17 is applicable and $v_1(\theta)^{-1} = O(\theta^{-2k_1})$ and $v_2(\theta)^{-1} = O(\theta^{-2(k_1+1)})$. Because this is case 2., $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$ is not the trivial vector space, so Equation (3.83) yields $v_{n-p}(\theta) = O(\theta^{-2(k_1+2)})$. From there we obtain $v_{n-p}(\theta)/v_1(\theta) = O(\theta^{-4})$ and $v_{n-p}(\theta)/v_2(\theta) = O(\theta^{-2})$. Equation (3.12) then implies $L(\boldsymbol{y}|\theta) = O(\theta^{-3})$, so the reference posterior is proper.

In the second subcase, $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W}\right) \cap \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right) = \mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$. Since $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}\right)$ is a hyperplane of $\mathbb{R}^{n-p}$, this implies that $\mathrm{Ker}\left(\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W}\right)$ is the same hyperplane: $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W}$ and $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}$ are symmetric matrices of rank 1 with the same kernel. So there exists $b \in \mathbb{R} \setminus \{0\}$ such that $\boldsymbol{W}^\top \boldsymbol{D}^{(k_1+1)}\boldsymbol{W} = b\boldsymbol{W}^\top \boldsymbol{D}^{(k_1)}\boldsymbol{W}$. If we redefine $g(\theta) := a_{k_1}\theta^{-2k_1} + a_{k_1+1}b\theta^{-2(k_1+1)}$, $g^\star(\theta) := a_{k_1+2}\theta^{-2(k_1+2)}g(\theta)^{-1}$, $\boldsymbol{D}^\star := \boldsymbol{D}^{(k_1+2)}$ and $\boldsymbol{R}_g(\theta) := g(\theta)^{-1}\sum_{k=k_1+3}^{\infty} a_k\theta^{-2k}$, the situation is similar to case 1.(b), except that $g^\star(\theta)$ is not necessarily of the form θ^{-2l}.
However, $g^{\star\prime}(\theta) = g^\star(\theta)(-2(k_1+2)\theta^{-1} - g'(\theta)g(\theta)^{-1})$ and $g'(\theta)g(\theta)^{-1} = O(\theta^{-1})$, which implies $g^{\star\prime}(\theta)g^\star(\theta)^{-1} = O(\theta^{-1})$. In addition, $g^\star(\theta) = O(\theta^{-2})$. Therefore the arguments of the study of case 1.(b) apply: $\pi(\theta) = O(\theta^{-1})$ and $L(\boldsymbol{y}|\theta) = O(g^\star(\theta)^{1/2}) = O(\theta^{-1})$. The reference posterior is proportional to $L(\boldsymbol{y}|\theta)\pi(\theta) = O(\theta^{-2})$: it is proper. This particular subcase, because it is analoguous to case 1.(b) is called "special". All other subcases of case 2. collectively form the "usual" case.

Matérn ($\nu \in \mathbb{Z}_+$) kernels

We now address the case where the correlation kernel is Matérn with integer smoothness ν. The proof strategy remains the same as for the other kernels, but the execution is a little trickier.

It still relies on the asymptotic expansion of $\boldsymbol{\Sigma}_\theta$. For Matérn kernels with integer smoothness ν, the decomposition is detailed in Appendix 3.D.

First, assume either $\boldsymbol{D}$ is *not* proportional to $\boldsymbol{D}^{(\nu)}$ or $\boldsymbol{D}^\star$ is *not* proportional to $\tilde{\boldsymbol{D}}^{(\nu)}$. In Equation (3.13), $g^\star(\theta)$ may be $\theta^{-2l}\log(\theta)$ instead of θ^{-2l}. Then its derivative is $g^{\star\prime}(\theta) = \theta^{-2l-1}(1-2l\log(\theta))$. In case 1.(a), the reference prior (and posterior) is $O(g^{\star\prime}(\theta)) = O(\theta^{-2l-1}\log(\theta))$ and thus proper. It is useless to distinguish cases 1.(b) and 2. because thanks to Proposition 3.7, the reference prior is $O(\theta^{-1})$. In either case, the rank of $\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}$ is at least one, so $v_{n-p}(\theta)/v_1(\theta) = O(g^\star(\theta))$. Equation (3.12) implies $L(\boldsymbol{y}|\theta) = O(g^\star(\theta)^{1/2}) = O(\theta^{-l}\log(\theta)^{1/2})$, so the reference posterior is proportional to $L(\boldsymbol{y}|\theta)\pi(\theta) = O(\theta^{-l-1}\log(\theta)^{1/2})$ and thus proper.

Now, assume $\boldsymbol{D}$ is proportional to $\boldsymbol{D}^{(\nu)}$ *and* $\boldsymbol{D}^\star$ is proportional to $\tilde{\boldsymbol{D}}^{(\nu)}$. In Equation (3.13), $g^\star(\theta) = \log(\theta)^{-1}$. Its derivative is $g^{\star\prime}(\theta) = -\theta^{-1}\log(\theta)^{-2}$. In case 1.(a), the reference prior (resp. posterior) is $O(g^{\star\prime}(\theta)) = O(\theta^{-1}\log(\theta)^{-2})$ and is thus proper. In case 1.(b), the reference prior is $O(g^{\star\prime}(\theta)g^\star(\theta)^{-1}) = O(\theta^{-1}\log(\theta)^{-1})$. Besides, as the rank of $\boldsymbol{W}^\top\boldsymbol{D}\boldsymbol{W}$ is at

least one, Lemma 3.17 yields $v_{n-p}(\theta)/v_1(\theta) = O(g^\star(\theta))$, so Equation (3.12) implies $L(\boldsymbol{y}|\theta) = O(g^\star(\theta)^{1/2}) = O(\log(\theta)^{-1/2})$. The reference posterior is then proportional to $L(\boldsymbol{y}|\theta)\pi(\theta) = O(\theta^{-1}\log(\theta)^{-3/2})$: it is proper. Case 2. cannot occur because Lemma 3.19 asserts that $\left\|\boldsymbol{\Sigma}_\theta^{-1}\right\| = O(\theta^{2\nu})$.

Part II

Compromise

Chapter 4

Optimal compromise between incompatible conditional probability distributions

This chapter covers Section 2 of the article Muré [2017].

Abstract

Models are often defined through conditional rather than joint distributions, but it can be difficult to check whether the conditional distributions are compatible. When they are not, we give meaning to the intuition that the stationary probability distribution of the associated Pseudo-Gibbs sampler is the optimal compromise between the conditional distributions.

Résumé

Les modèles statistiques sont souvent définis par lois conditionnelles plutôt que jointes. Il peut cependant être difficile de contrôler la compatibilité des lois conditionnelles. Dans le cas où elles ne seraient pas compatibles, nous donnons sens à l'intuition selon laquelle la distribution stationnaire du pseudo-échantillonneur de Gibbs correspondant est le compromis optimal entre les lois conditionnelles.

4.1 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. 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 [2001] 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) [Roverato, 2004]. Moreover, an abstraction (simplification) of a DAG can be compatible or not [Yet and Marsh, 2014]. In this dissertation 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.

Accordingly, the problem of efficiently determining whether a given system of conditionals is compatible has received considerable attention over the years. Kuo et al. [2017], 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. Obviously, this solution, like its predecessors deals only with discrete conditional probability distributions.

However, even if a system contains incompatible conditional probability distributions, it does not follow that it is useless. Since Heckerman et al. [2000], practitioners have been using systems of conditional probability distributions without reference to compatibility. Gelman and Raghunathan [2001] state that "in general, reasonable-seeming conditional models will not be compatible with any single joint distribution". It is, indeed, 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 [2017] provide a detailed analysis and geometrical interpretation of the behavior of Pseudo-Gibbs Samplers for discrete conditional distributions. In particular, they show how the scanning order determines its stationary distribution. In Section 4.2 of the present chapter, we provide some theoretical foundation for the intuition that the stationary distribution of a PGS with random scanning order is, in case of existence and uniqueness, the best "compromise" between incompatible conditionals.

4.2 Optimal compromise: a general theory

Definitions and notations

In this section we introduce the concepts necessary to define the optimal compromise between potentially incompatible conditional distributions.

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

Definition 4.1. *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 $\Omega = \bigtimes_{i=1}^{r} \Omega_i = \Omega_1 \times ... \times \Omega_r$ 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 $\left(\bigtimes_{j \neq i} \Omega_j\right) \times \mathcal{A}_i \to [0, 1]$.
Intuitively (we formalize this below), every π_i should be assembled with a distribution m^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^i ($i \in [\![1, r]\!]$) as an $(r-1)$-dimensional distribution. If the m^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 $(m^i)_{i \in [\![1,r]\!]}$ can make all joint distributions equal, we have to look for a "compromise" between the Markov kernels.

Remark (Producing incompatibility is easy). Take $r = 2$ and $\Omega_1 = \Omega_2 = \mathbb{R}$ and endow $\mathbb{R}$ with the Borel sigma-algebra. Assume that for every $t \in \mathbb{R}$, $\pi_1(\cdot|t)$ and $\pi_2(\cdot|t)$ are absolutely continuous with respect to the Lebesgue measure and denote by $f_1(\cdot|t)$ and $f_2(\cdot|t)$ their respective density functions. A necessary condition [Arnold et al., 2001] for the compatibility of π_1 and π_2 is the existence of two mappings u and v defined on $\mathbb{R}$ such that for almost every real numbers x and t (in the sense of the Lebesgue measure),

$$\frac{f_1(x|t)}{f_2(t|x)} = u(t)v(x). \tag{4.1}$$

Definition 4.2. *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* $S_{-i} = \bigtimes_{j \neq i} S_j$ *(with* $S_j \in \mathcal{A}_j$ *for every* $j \neq i$*),*

$$\phi_{-i}(S_{-i}) = \phi(\bigtimes_{j<i} S_j \times \Omega_i \times \bigtimes_{k>i} S_k). \tag{4.2}$$

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

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

For every $i \in [\![1, r]\!]$ and every probability distribution m^i on $\bigotimes_{j \neq i} \mathcal{A}_j$, denote by $\pi_i m^i$ 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 $S_i \in \mathcal{A}_i$,

$$\pi_i m^i(S_{<i} \times S_i \times S_{>i}) = \int_{S_{<i} \times S_{>i}} \pi_i(S_i|\omega_{-i}) dm^i(\omega_{-i}). \tag{4.3}$$

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 for every $i \in [\![1, r]\!]$, m^i is the i-th $(r-1)$-marginal distribution of $\pi_i m^i$:

$$\left(\pi_i m^i\right)_{-i} = m^i. \tag{4.4}$$

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

Definition 4.3. *A sequence of $(r-1)$-dimensional distributions $(m^i)_{i\in[\![1,r]\!]}$ (each m^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 $i \in [\![1,r]\!]$*

$$m^i = \frac{1}{r}\sum_{j=1}^{r}(\pi_j m^j)_{-i}. \tag{4.5}$$

So the "common ground" we require for the sequence of distributions $(\pi_i m^i)_{i\in[\![1,r]\!]}$ 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 4.3 below.

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

Definition 4.4. *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 the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$.*

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.

Definition 4.5. *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:*

$$E_\lambda(P) = \sum_{i=1}^{r}\int_{\mathcal{A}}\left[\frac{d(\pi_i P_{-i})}{d\lambda}(\omega) - \frac{dP}{d\lambda}(\omega)\right]^2 d\lambda(\omega). \tag{4.6}$$

Remark. 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 4.5 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.

Deriving the optimal compromise

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

Definition 4.6. *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:*

$$P_G = \frac{1}{r}\sum_{i=1}^{r} \pi_i (P_G)_{-i}. \tag{4.7}$$

Proposition 4.7. *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 4.4.*

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

$$P_G = \frac{1}{r}\sum_{i=1}^{r} \pi_i (P_G)_{-i}(A) = 0. \tag{4.8}$$

So for every integer $i \in [\![1, r]\!]$, $\pi_i(P_G)_{-i}(A) = 0$. This fulfills the first condition of Definition 4.4.

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

$$(P_G)_{-i} = \left(\frac{1}{r}\sum_{j=1}^{r} \pi_j (P_G)_{-j}\right)_{-i} = \frac{1}{r}\sum_{j=1}^{r} \left(\pi_j (P_G)_{-j}\right)_{-i}. \tag{4.9}$$

□

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 4.8. *If a sequence of $(r-1)$-dimensional probability distributions $(m^i)_{i\in[\![1,r]\!]}$ (each m^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 $(\pi_i)_{i\in[\![1,r]\!]}$.*

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

$$P = \frac{1}{r}\sum_{i=1}^{r} \pi_i m^i. \tag{4.10}$$

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

$$P_{-i} = \frac{1}{r}\sum_{j=1}^{r} \left(\pi_j m^j\right)_{-i} = m^i, \tag{4.11}$$

where the last equality is due to $(m^i)_{i\in[\![1,r]\!]}$ being compatible with $(\pi_i)_{i\in[\![1,r]\!]}$. Plugging this into Equation (4.10), we obtain that P is a Gibbs compromise. Equation(4.11) then yields the result. □

Remark (Equivalence relationship 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 relationship 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.

The main result follows from Proposition 4.8:

Theorem 4.9. *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.*

Proof. If the distribution P_G is the unique Gibbs compromise between the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$, then Proposition 4.8 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 $\omega \in \Omega$,

$$\frac{d(\pi_i P_{-i})}{d\lambda}(\omega) = \frac{d(\pi_i (P_G)_{-i})}{d\lambda}(\omega). \tag{4.12}$$

Moreover, Equation (4.10) implies that for λ-almost any $\omega \in \Omega$, $\frac{dP_G}{d\lambda}(\omega)$ is the arithmetic average between the $\frac{d(\pi_i(P_G)_{-i})}{d\lambda}(\omega)$ $(i \in [\![1,r]\!])$, so it minimizes the mean squared error. Together with Equation (4.12), this implies that for λ-almost any $\omega \in \Omega$,

$$\sum_{i=1}^{r} \left[\frac{d(\pi_i(P_G)_{-i})}{d\lambda}(\omega) - \frac{dP_G}{d\lambda}(\omega)\right]^2 \leqslant \sum_{i=1}^{r} \left[\frac{d(\pi_i P_{-i})}{d\lambda}(\omega) - \frac{dP}{d\lambda}(\omega)\right]^2. \tag{4.13}$$

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 $\omega \in S$,

$$\forall i \in [\![1,r]\!] \quad \frac{d(\pi_i P_{-i})}{d\lambda}(\omega) = \frac{d(\pi_i (P_G)_{-i})}{d\lambda}(\omega) \quad \text{and} \quad \frac{dP}{d\lambda}(\omega) \neq \frac{d(P_G)}{d\lambda}(\omega). \tag{4.14}$$

So for every $\omega \in S$, Equation (4.13) 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. □

Remark (Equivalence class and convexity – continued). A similar proof can be used to show the following results. In cases where there are several different Gibbs compromises, 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 4.9 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 4.9, 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 $f = (f_1, ..., f_r) : \bigtimes_{i\in[\![1,r]\!]} \Omega_i \to \bigtimes_{i\in[\![1,r]\!]} \tilde{\Omega}_i$ and for every $i \in [\![1,r]\!]$ $f_{-i} = (f_1, .., f_{i-1}, f_{i+1}, ..., f_r) : \bigtimes_{j\neq i} \Omega_i \to \bigtimes_{j\neq i} \tilde{\Omega}_i$.

Also let $\tilde{\pi}_i$ be the Markov kernel $\left(\bigtimes_{j\neq i} \tilde{\Omega}_j\right) \times \tilde{\mathcal{A}}_i \to [0,1]$ such that for every $\omega_{-i} \in \bigtimes_{j\neq i} \Omega_j$ 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 4.10. *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 $(\tilde{\pi}_i)_{i\in[\![1,r]\!]}$.*

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

$$\begin{aligned}
\tilde{P}_G\left(\bigtimes_{i\in[\![1,r]\!]} \tilde{S}_i\right) &= P_G\left(f^{-1}\left(\bigtimes_{i\in[\![1,r]\!]} \tilde{S}_i\right)\right) \\
&= P_G\left(\bigtimes_{i\in[\![1,r]\!]} f_i^{-1}(\tilde{S}_i)\right) \\
&= \frac{1}{r}\sum_{i=1}^{r} \int_{\bigtimes_{j\neq i} f_j^{-1}(\tilde{S}_j)} \pi_i(f_i^{-1}(\tilde{S}_i)|\omega_{-i}) d\left\{(P_G)_{-i}\right\}(\omega_{-i}) \\
&= \frac{1}{r}\sum_{i=1}^{r} \int_{\bigtimes_{j\neq i} f_j^{-1}(\tilde{S}_j)} \tilde{\pi}_i(\tilde{S}_i|f_{-i}(\omega_{-i})) d\left\{(P_G)_{-i}\right\}(\omega_{-i}) \\
&= \frac{1}{r}\sum_{i=1}^{r} \int_{\bigtimes_{j\neq i} \tilde{S}_j} \tilde{\pi}_i(\tilde{S}_i|\tilde{\omega}_{-i}) d\left\{(P_G)_{-i} * f_{-i}\right\}(\tilde{\omega}_{-i}). \qquad (4.15)
\end{aligned}$$

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

$$\begin{aligned}
(P_G)_{-i} * f_{-i}\left(\bigtimes_{j\neq i} \tilde{T}_{-i}\right) &= (P_G)_{-i}\left(\bigtimes_{j\neq i} f_j^{-1}(\tilde{T}_j)\right) \\
&= P_G\left(\bigtimes_{j<i} f_j^{-1}(\tilde{T}_j) \times f_i^{-1}(\tilde{\Omega}_i) \times \bigtimes_{k>i} f_k^{-1}(\tilde{T}_k)\right) \\
&= P_G * f\left(\bigtimes_{j<i} \tilde{T}_j \times \tilde{\Omega}_i \bigtimes_{k>i} \tilde{T}_k\right) \\
&= (P_G * f)_{-i}\left(\bigtimes_{j\neq i} \tilde{T}_j\right). \qquad (4.16)
\end{aligned}$$

So $(P_G)_{-i} * f_{-i} = (P_G * f)_{-i} = (\tilde{P}_G)_{-i}$. Then, returning to Equation 4.15,

$$\tilde{P}_G\left(\bigtimes_{i\in[\![1,r]\!]} \tilde{S}_i\right) = \frac{1}{r}\sum_{i=1}^{r}\int_{\times_{j\neq i}\tilde{S}_j} \tilde{\pi}_i(\tilde{S}_i|\tilde{\omega}_{-i})d\left\{(\tilde{P}_G)_{-i}\right\}(\tilde{\omega}_{-i}). \tag{4.17}$$

Finally, we obtain

$$\tilde{P}_G = \frac{1}{r}\sum_{i=1}^{r}\tilde{\pi}_i(\tilde{P}_G)_{-i}. \tag{4.18}$$

$\tilde{P}_G$ 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 $\tilde{Q}_G = \tilde{Q}_G * f^{-1} * f = P_G * f = \tilde{P}_G$.

□

4.3 Discussion of the notion of compromise

This section is devoted to discussing the merits, in our opinion, of Definitions 4.3 and 4.4. The first subsection shows that strengthening their requirements often makes the set of all compromises empty. Conversely, the second subsection provides examples showing that relaxing their requirements leads to undesirable behavior in compromises.

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 4.11. *A pair of probability distributions m^1 (resp. m^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 $\pi_1 m^1$ and $\pi_2 m^2$ verify*

$$\left(\pi_1 m^1\right)_{-2} = m^2 \text{ and } \left(\pi_2 m^2\right)_{-1} = m^1. \tag{4.19}$$

While this definition may seem more restrictive at first glance than Definition 4.3, both definitions are in fact equivalent when applied to a pair of Markov kernels, because $(r-1)$-dimensional distributions are simply 1-dimensional distributions in this case. Indeed, following directly from Definition 4.3, we have this result which holds for any r:

Proposition 4.12. *If a sequence of $(r-1)$-dimensional probability distributions $(m^i)_{i\in[\![1,r]\!]}$ (each m^i being a probability distribution on $\bigotimes_{j\neq i}\mathcal{A}_j$) is compatible (in the sense of Definition*

4.3) with the sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$*, then all joint distributions in the sequence* $(\pi_i m^i)_{i\in[\![1,r]\!]}$ *share the same marginals, that is* $\forall i,j,k \in [\![1,r]\!]$,

$$\forall S_k \in \mathcal{A}_k, \quad \pi_i m^i \left(\bigtimes_{k'<k} \Omega_{k'} \times S_k \times \bigtimes_{k''>k} \Omega_{k''} \right) = \pi_j m^j \left(\bigtimes_{k'<k} \Omega_{k'} \times S_k \times \bigtimes_{k''>k} \Omega_{k''} \right). \tag{4.20}$$

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 $(\omega_1,\omega_2) \in \Omega_1 \times \Omega_2$,

$$P(\{\omega_1\} \times \{\omega_2\}|\omega_3) = \frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3)P(\Omega_1 \times \{\omega_2\}|\omega_3) + \frac{1}{2}\pi_1(\omega_2|\omega_1,\omega_3)P(\{\omega_1\} \times \Omega_2|\omega_3). \tag{4.21}$$

Thanks to Theorem 4.9, $P(\cdot|\omega_3)$ is the optimal compromise. Moreover, notice that Equation (4.21) 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 (4.21) 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 (4.21) 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 $(\omega_1,\omega_2,\omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$,

$$\begin{aligned}
\phi(\{\omega_1\} \times \{\omega_2\} \times \{\omega_3\}) &= P(\{\omega_1\} \times \{\omega_2\}|\omega_3)\phi(\Omega_1 \times \Omega_2 \times \{\omega_3\}) && (4.22)\\
&= Q(\{\omega_2\} \times \{\omega_3\}|\omega_1)\phi(\{\omega_1\} \times \Omega_2 \times \Omega_3) && (4.23)\\
&= R(\{\omega_1\} \times \{\omega_3\}|\omega_2)\phi(\Omega_2 \times \{\omega_2\} \times \Omega_3). && (4.24)
\end{aligned}$$

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 (4.22) and (4.24) yields

$$\begin{aligned}
&\frac{\phi(\{\omega_1\}\times\{\omega_2\}\times\{\omega_3\})}{\phi(\Omega_1\times\{\omega_2\}\times\{\omega_3\})}\\
=&\frac{P(\{\omega_1\}\times\{\omega_2\}|\omega_3)}{P(\Omega_1\times\{\omega_2\}|\omega_3)} = \frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3) + \frac{1}{2}\pi_2(\{\omega_2\}|\omega_1,\omega_3)\frac{P(\{\omega_1\}\times\Omega_2|\omega_3)}{P(\Omega_1\times\{\omega_2\}|\omega_3)} \qquad (4.25)\\
=&\frac{R(\{\omega_1\}\times\{\omega_3\}|\omega_2)}{R(\Omega_1\times\{\omega_2\}|\omega_3)} = \frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3) + \frac{1}{2}\pi_3(\{\omega_3\}|\omega_1,\omega_2)\frac{R(\{\omega_1\}\times\Omega_3|\omega_2)}{R(\Omega_1\times\{\omega_3\}|\omega_2)}. \qquad (4.26)
\end{aligned}$$

Now combine Equations (4.25) and (4.26):

$$\frac{1}{2}\pi_2(\{\omega_2\}|\omega_1,\omega_3)\frac{\phi(\{\omega_1\}\times\Omega_2\times\{\omega_3\})}{\phi(\Omega_1\times\{\omega_2\}\times\{\omega_3\})} = \frac{1}{2}\pi_3(\{\omega_3\}|\omega_1,\omega_2)\frac{\phi(\{\omega_1\}\times\{\omega_2\}\times\Omega_3)}{\phi(\Omega_1\times\{\omega_2\}\times\{\omega_3\})}. \qquad (4.27)$$

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 Equation (4.21)), but only *on average* over $\omega_3\in\Omega_3$. So a compromise ϕ should still verify Equation (4.22), but it would only need to verify this weakened version of Equation (4.21) for every $(\omega_1,\omega_2)\in\Omega_1\times\Omega_2$:

$$\begin{aligned}
&\sum_{\omega_3\in\Omega_3} P(\{\omega_1\}\times\{\omega_2\}|\omega_3)\phi(\Omega_1\times\Omega_2\times\{\omega_3\})\\
=&\sum_{\omega_3\in\Omega_3}\left[\frac{1}{2}\pi_1(\{\omega_1\}|\omega_2,\omega_3)P(\Omega_1\times\{\omega_2\}|\omega_3) + \frac{1}{2}\pi_2(\{\omega_2\}|\omega_1,\omega_3)P(\{\omega_1\}\times\Omega_2|\omega_3)\right]\times\\
&\qquad\phi(\Omega_1\times\Omega_2\times\{\omega_3\}). \qquad (4.28)
\end{aligned}$$

Because Equation (4.22) is still expected to hold, Equation (4.28) is equivalent to

$$\begin{aligned}
&\phi_{-3}(\{\omega_1\}\times\{\omega_2\})\\
=&\frac{1}{2}\sum_{\omega_3\in\Omega_3}\pi_1(\{\omega_1\}|\omega_2,\omega_3)\phi_{-1}(\{\omega_2\}\times\{\omega_3\}) + \frac{1}{2}\sum_{\omega_3\in\Omega_3}\pi_2(\{\omega_2\}|\omega_1,\omega_3)\phi_{-2}(\{\omega_1\}\times\{\omega_3\}).
\end{aligned} \qquad (4.29)$$

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" 3-dimensional distribution ϕ to requirements about its $(3-1)$-marginal distributions ϕ_{-1}, ϕ_{-2} and ϕ_{-3}. Definition 4.3 is just another formulation of these requirements, which are taken to define a compatible sequence of $(r-1)$-dimensional distributions.

Proposition 4.8 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!

Weaker definitions of compromises are inconvenient

As was shown in the previous subsection, the requirements given by Definition 4.3 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.

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 4.5 over all distributions absolutely continuous with respect to λ? The following 2-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:

$$\pi_1(X_1 = 1|\omega_2) = \mathbf{1}_{\{\omega_2=0\}} + 1/2\,\mathbf{1}_{\{\omega_2=1\}}; \tag{4.30}$$

$$\pi_2(X_2 = 1|\omega_1) = 1/2\,\mathbf{1}_{\{\omega_1=0\}} + \mathbf{1}_{\{\omega_1=1\}}. \tag{4.31}$$

In this situation, let us define the reference measure λ as the counting measure.

Denote by π_C the optimal compromise (maximizing 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 $E(\pi_C) = 2/25 > E(\pi_E) = 1/15$ and

$$\begin{aligned}\pi_C(\{\omega_1\} \times \{\omega_2\}) = 1/10\,\big(&\mathbf{1}_{\{(\omega_1,\omega_2)=(0,0)\}} + \mathbf{1}_{\{(\omega_1,\omega_2)=(1,0)\}} \\ &+3\,\mathbf{1}_{\{(\omega_1,\omega_2)=(0,1)\}} + 5\,\mathbf{1}_{\{(\omega_1,\omega_2)=(1,1)\}}\big)\,;\end{aligned} \tag{4.32}$$

$$\begin{aligned}\pi_E(\{\omega_1\} \times \{\omega_2\}) = 1/30\,\big(&3\,\mathbf{1}_{\{(\omega_1,\omega_2)=(0,0)\}} + \mathbf{1}_{\{(\omega_1,\omega_2)=(1,0)\}} \\ &+11\,\mathbf{1}_{\{(\omega_1,\omega_2)=(0,1)\}} + 15\,\mathbf{1}_{\{(\omega_1,\omega_2)=(1,1)\}}\big)\,.\end{aligned} \tag{4.33}$$

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 Definition 4.3), one may "cheat" by having the marginals disadvantage inconvenient values for the parameters.

Why the compatibility requirements cannot be weakened: example in 3 dimensions.

In the two-dimensional case, because of Proposition 4.12, Definitions 4.11 and 4.3 give the same meaning to the concept of compatibility of marginals, so Definition 4.3 may be thought of as a generalization of Definition 4.11 to cases with more than two dimensions. However, another generalization of 4.11 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 4.2.

Definition 4.13. *A sequence of $(r-1)$-dimensional distributions $(m^i)_{i\in[\![1,r]\!]}$ is* weakly compatible *with a sequence of Markov kernels $(\pi_i)_{i\in[\![1,r]\!]}$ if Equation (4.20) holds.*

Proposition 4.12 means compatibility in the sense of Definition 4.3 implies weak compatibility in the sense of Definition 4.13, 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 4.14. *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 4.15. *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 (4.6).*

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 $(\omega_1,\omega_2,\omega_3)\in\Omega_1\times\Omega_2\times\Omega_3$

$$\pi_1(X_1=1|\omega_2,\omega_3)=1/2; \tag{4.34}$$

$$\pi_2(X_2=1|\omega_1,\omega_3)=1/2; \tag{4.35}$$

$$\pi_3(X_3=1|\omega_1,\omega_2)=\mathbf{1}_{\{\omega_1=\omega_2\}}+1/2\,\mathbf{1}_{\{\omega_1\neq\omega_2\}}. \tag{4.36}$$

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 $(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2$

$$P(\{\omega_1\} \times \{\omega_2\}|\omega_3) = 1/4. \tag{4.37}$$

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 $(\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$

$$\begin{aligned}\pi_C(\{\omega_1\} \times \{\omega_2\} \times \{\omega_3\}) = \frac{1}{24}\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=0\}} + \frac{1}{12}\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=0\}} \\ + \frac{5}{24}\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=1\}} + \frac{1}{6}\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=1\}}.\end{aligned} \tag{4.38}$$

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

$$(\pi_C)_{-3}(\{\omega_1\} \times \{\omega_2\}) = 1/4. \tag{4.39}$$

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 4.8 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 $(\omega_1, \omega_2, \omega_3) \in \Omega_1 \times \Omega_2 \times \Omega_3$,

$$\begin{aligned}\pi_W(\{\omega_1\} \times \{\omega_2\} \times \{\omega_3\}) \approx 0.04\,\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=0\}} + 0.10\,\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=0\}} \\ + 0.19\,\mathbf{1}_{\{\omega_1=\omega_2,\omega_3=1\}} + 0.17\,\mathbf{1}_{\{\omega_1\neq\omega_2,\omega_3=1\}}.\end{aligned} \tag{4.40}$$

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

$$(\pi_W)_{-3}(\{\omega_1\} \times \{\omega_2\}) \approx 0.23\,\mathbf{1}_{\{\omega_1=\omega_2\}} + 0.27\,\mathbf{1}_{\{\omega_1\neq\omega_2\}}. \tag{4.41}$$

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

$$(\pi_W)_{-3}(\{\omega_1\} \times \Omega_2) = 1/2, \tag{4.42}$$

$$(\pi_W)_{-3}(\Omega_1 \times \{\omega_2\}) = 1/2, \tag{4.43}$$

$(\pi_W)_{-3}$ 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.

4.4 Conclusion

In this chapter, we proposed a theory defining the notions of compromise and espectially optimal compromise between potentially incompatible conditional distributions, i.e. Markov kernels. 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.

In the rest of this thesis, we focus on one particular application of this theory: deriving Objective Bayesian inference for correlation lenghts in Kriging models with anisotropic correlation kernels. However, the field of potential applications is much larger. It encompasses all models that rely on conditional rather than joint probability distributions.

Part III

Application

Chapter 5

Application of the Optimal Compromise to Simple Kriging models with Matérn correlation kernels

This chapter covers the article Muré [2017], with the exception of its second section.

Abstract

The notion of optimal compromise between possibly incompatible conditional distributions 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.

Résumé

La notion de compromis optimal entre lois conditionnelles potentiellement incompatibles permet une analyse bayésienne objective des paramètres de corrélation de modèles de krigeage : nous utilisons des lois *a posteriori* de Jeffreys univariées conditionnelles plutôt que la très usitée loi *a posteriori* de Jeffreys multivariée. Cette stratégie rend la procédure pleinement bayésienne abordable. Des exemples numériques montrent que ses performances fréquentistes en termes de taux de couverture des intervalles prédictifs sont quasi-optimales.

5.1 Introduction

The present chapter is the point where all previous chapters come together.

Recalling Chapter 1, 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 [Santner et al., 2003]. 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 [Journel and Huijbregts, 1978], but is also frequently used in the context of computer experiments and machine learning under the label "Gaussian Process regression" [Rasmussen and Williams, 2006]. In this chapter, 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 and will be tackled in Chapter 6.

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 [2001], because the likelihood function may often be quite flat [Li and Sudjianto, 2005]. In a Bayesian framework, this uncertainty is represented by a prior distribution on the parameters.

Recalling Chapter 2, the Objective Bayesian paradigm as explained by Berger [2006] 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 [Bernardo, 2005] 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 [Bernardo, 2005]. 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 – 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. 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 [Berger and Bernardo, 1992].

Berger et al. [2001] were the first to derive a reference prior for the parameters of a Gaussian Process regression model –cf. Chapter 3. 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 [Paulo, 2005, Ren et al., 2012, Kazianka and Pilz, 2012, Ren et al., 2013, Gu et al., 2018] have therefore resolved to treat all correlation parameters as a single multidimensional parameter. It is in order to avoid having to do this that we make use of Potentially Incompatible Gibbs Samplers as defined in Chapter 4.

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 (Potentially Incompatible Gibbs Sampler, cf. previous chapter).

Theorem 5.2 is the main result with respect to the application. First, under reasonable assumptions, the PIGS admits one single stationary probability distribution, which is the optimal compromise in the sense of the theory of Chapter 4. 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 5.3 and 5.4 illustrate the inferential and predictive performance of the stationary distribution respectively.

5.2 Optimal compromise between Objective Posterior conditional distributions in Gaussian Process regression

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 $\mathcal{N}(\mathbf{0}_n, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}})$.

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

$$L(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\theta}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} \boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}\right\} . \tag{5.1}$$

Conditional on $\boldsymbol{\theta}$, this is a scale model. The reference prior with parameter ordering $\sigma^2 \prec \boldsymbol{\theta}$ is therefore given by:

$$\pi(\sigma^2, \boldsymbol{\theta}) = \pi(\sigma^2|\boldsymbol{\theta})\pi(\boldsymbol{\theta}) \quad \text{with} \quad \pi(\sigma^2|\boldsymbol{\theta}) \propto 1/\sigma^2. \tag{5.2}$$

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

Let us integrate σ^2 out of the likelihood (5.1):

$$\begin{aligned} L^1(\boldsymbol{y} \mid \boldsymbol{\theta}) &\propto \int_0^\infty L(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\theta}) \pi(\sigma^2 \mid \boldsymbol{\theta})\, d(\sigma^2) \\ &\propto \int_0^\infty L(\boldsymbol{y} \mid \sigma^2, \boldsymbol{\theta}) (\sigma^2)^{-1}\, d(\sigma^2) = \left(\frac{2\pi^{\frac{n}{2}}}{\Gamma\left(\frac{n}{2}\right)}\right)^{-1} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} \boldsymbol{y}\right)^{-\frac{n}{2}} . \end{aligned} \tag{5.3}$$

This model is the one from the remark under Example 3 in Chapter 2. The "marginal" reference prior on $\boldsymbol{\theta}$ is proportional to the square root of the determinant of the $r \times r$ matrix with (i, j)-th entry

$$\mathrm{Tr}\left[\left(\frac{\partial}{\partial\theta_i}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\left(\frac{\partial}{\partial\theta_j}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)\right] - \frac{1}{n}\mathrm{Tr}\left[\frac{\partial}{\partial\theta_i}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]\mathrm{Tr}\left[\frac{\partial}{\partial\theta_j}\left(\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right] . \tag{5.4}$$

See also Ren et al. [2012] for an other way of reaching this conclusion. Their solution is based on astute algebraic considerations.

Viewing $\boldsymbol{\theta}$ as a single parameter has the disadvantage of requiring the use of a multidimensional Jeffreys-rule prior distribution, which may show the sort of undesirable behavior mentioned in the introduction, of which Example 4 from Chapter 2 is an illustration.

Alternatively, we could draw inspiration from the one-dimensional case in the following way. Suppose that we knew every entry of $\boldsymbol{\theta}$ except one, θ_i. Then, according to Equation (5.4), the prior density on θ_i knowing all entries θ_j $(j \neq i)$ would be

$$\pi_i(\theta_i \mid \theta_j \ \forall j \neq i) \propto \sqrt{\operatorname{Tr}\left[\left(\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\frac{\partial}{\partial\theta_i}(\boldsymbol{\Sigma}_{\boldsymbol{\theta}})\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\right]^2}. \tag{5.5}$$

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 (4.1).

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 4.2 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 4.9 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,

$$\frac{\pi_G(\boldsymbol{\theta}|\boldsymbol{y})}{L^1(\boldsymbol{y}|\boldsymbol{\theta})} = \frac{\pi_G(\boldsymbol{\theta})}{\tilde{L}(\boldsymbol{y})}. \tag{5.6}$$

As we have no explicit expression of $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$, we have no way to check whether Equation (5.6) 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 4.9 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 5.A.

Definitions

In this chapter, 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 $g(\boldsymbol{x}) = \int_{\mathbb{R}^r} \widehat{g}(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{\omega}$ and $\widehat{g}(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} g(\boldsymbol{x}) e^{-i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{x}$.
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) $\forall \boldsymbol{x} \in \mathbb{R}^r$,

$$K_{r,\nu}(\boldsymbol{x}) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu}\|\boldsymbol{x}\|\right)^\nu \mathcal{K}_\nu \left(2\sqrt{\nu}\|\boldsymbol{x}\|\right) \ ; \tag{5.7}$$

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

$$\widehat{K}_{r,\nu}(\boldsymbol{\omega}) = \frac{M_r(\nu)}{(\|\boldsymbol{\omega}\|^2 + 4\nu)^{\nu + \frac{r}{2}}} \text{ with } M_r(\nu) = \frac{\Gamma(\nu + \frac{r}{2})(2\sqrt{\nu})^{2\nu}}{\pi^{\frac{r}{2}} \Gamma(\nu)}. \tag{5.8}$$

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

(i) $\forall \boldsymbol{x} \in \mathbb{R}^r$,

$$K^{tens}_{r,\nu}(\boldsymbol{x}) = \prod_{j=1}^{r} K_{1,\nu}(\boldsymbol{x}_j) \ ; \tag{5.9}$$

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

$$\widehat{K}^{tens}_{r,\nu}(\boldsymbol{\omega}) = \prod_{j=1}^{r} \widehat{K_{1,\nu}}(\boldsymbol{\omega}_j). \tag{5.10}$$

(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 $\boldsymbol{x} \mapsto \sigma^2 K^{tens}_{r,\nu}\left(\frac{\boldsymbol{x}}{\boldsymbol{\theta}}\right)$ (resp. $\boldsymbol{x} \mapsto \sigma^2 K^{tens}_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})$).

Thanks to Proposition 4.10, 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 5.A) involve a technical assumption on the design set:

Definition 5.1. *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.*

Randomly sampled design sets almost surely have coordinate-distinct points – for instance Latin Hypercube Sampling. Cartesian product design sets, however, do not.

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 anisostropic 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 5.2. *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*

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

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

The Markov kernel $P_{\boldsymbol{y}}$ is uniformly ergodic. This means that, denoting by $\| \cdot \|_{TV}$ the total variation norm,

$$\lim_{n \to \infty} \sup_{\boldsymbol{\theta}^{(0)} \in (0, +\infty)^r} \| P_{\boldsymbol{y}}^n(\boldsymbol{\theta}^{(0)}, \cdot) - \pi_G(\cdot|\boldsymbol{y}) \|_{TV} = 0$$

$$(resp. \lim_{n \to \infty} \sup_{\boldsymbol{\mu}^{(0)} \in (0, +\infty)^r} \| P_{\boldsymbol{y}}^n(\boldsymbol{\mu}^{(0)}, \cdot) - \pi_G(\cdot|\boldsymbol{y}) \|_{TV} = 0). \tag{5.11}$$

Remark. 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 Propostion 4.10, the Gibbs compromise does not either. The parametrization using inverse correlation lenghts $\boldsymbol{\mu}$ is more convenient for proving this theorem, however, so we use it exclusively in the rest of this section.

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 5.2 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".

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. [2003] provides this useful result for prediction:

Proposition 5.3. *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. 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.

In practice, the distribution of $Y(\boldsymbol{x}_0)$ conditionally to $\boldsymbol{Y} = \boldsymbol{y}$ when $\boldsymbol{\theta}$ is unknown can be obtained once from the Gibbs reference posterior distribution on $\boldsymbol{\theta}$ has been sampled. 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.

5.3 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 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 Likeilhood Estimator (MLE) which does the same with the likelihood function.

Methodology

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

Our test cases are 3-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 5.4. *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 [Hotelling, 1953] (that is, the inverse hyperbolic tangent function) to every (non-1) 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.
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.

Results

This subsection provides results obtained on 3-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 5.1 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.

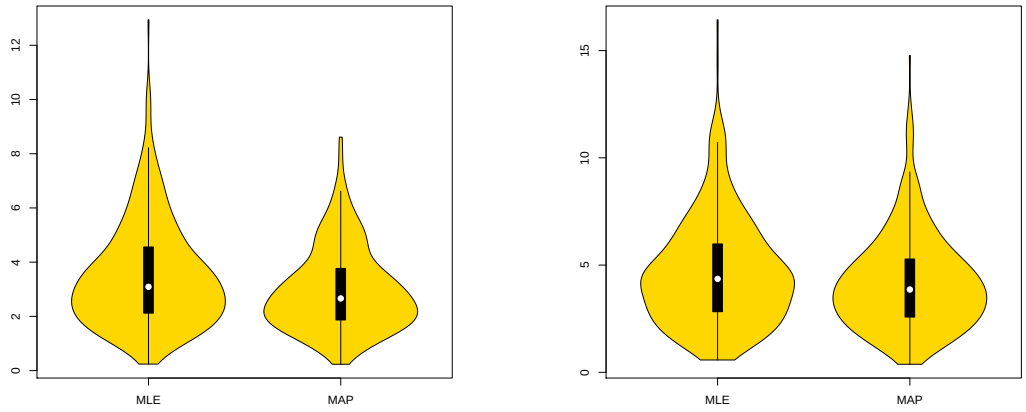
Most of the "true" vectors of correlation lengths featured in Table 5.1 were selected in a way to showcase the behavior of both estimators in strongly anisotropic cases, but one (0.5 - 0.5 - 0.5) also showcases their behavior if the true kernel is actually isotropic. And the final one (0.8 - 1 - 0.9) 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 (Root Mean Square Error) was between 9 and 15% lower, as showcased in Table 5.1.

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

Table 5.1 – *RMSE (where the error is measured in terms of the distance in Definition 5.4) 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.*

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 5.1 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)



(a) *0.4 – 0.8 – 0.2* **(b)** *0.8 – 1.0 – 0.9*

Figure 5.1 – *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).*

5.4 Comparison of the predictive distributions associated with the estimators (MLE and MAP) and the full posterior distribution

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. We consider 95% intervals: the lower bound is the 2.5% quantile and the upper bound the 97.5% quantile of predictive distribution $\hat{P}_{MLE}(\boldsymbol{y}_0 \mid \boldsymbol{y})$, $\hat{P}_{MAP}(\boldsymbol{y}_0 \mid \boldsymbol{y})$ and $P(\boldsymbol{y}_0 \mid \boldsymbol{y})$. For the sake of comprehensiveness, we also consider predictive intervals associated with the "true" predictive 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), $\hat{P}_{MLE}(\boldsymbol{y}_0 \mid \boldsymbol{y})$, $\hat{P}_{MAP}(\boldsymbol{y}_0 \mid \boldsymbol{y})$ 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 of each of the four predictive 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, $m = 500$).

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 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 predictive distributions, whether they are based on the MLE or MAP estimator, or 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 5.2 show that using the full posterior distribution (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.

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 5.2 – *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.*

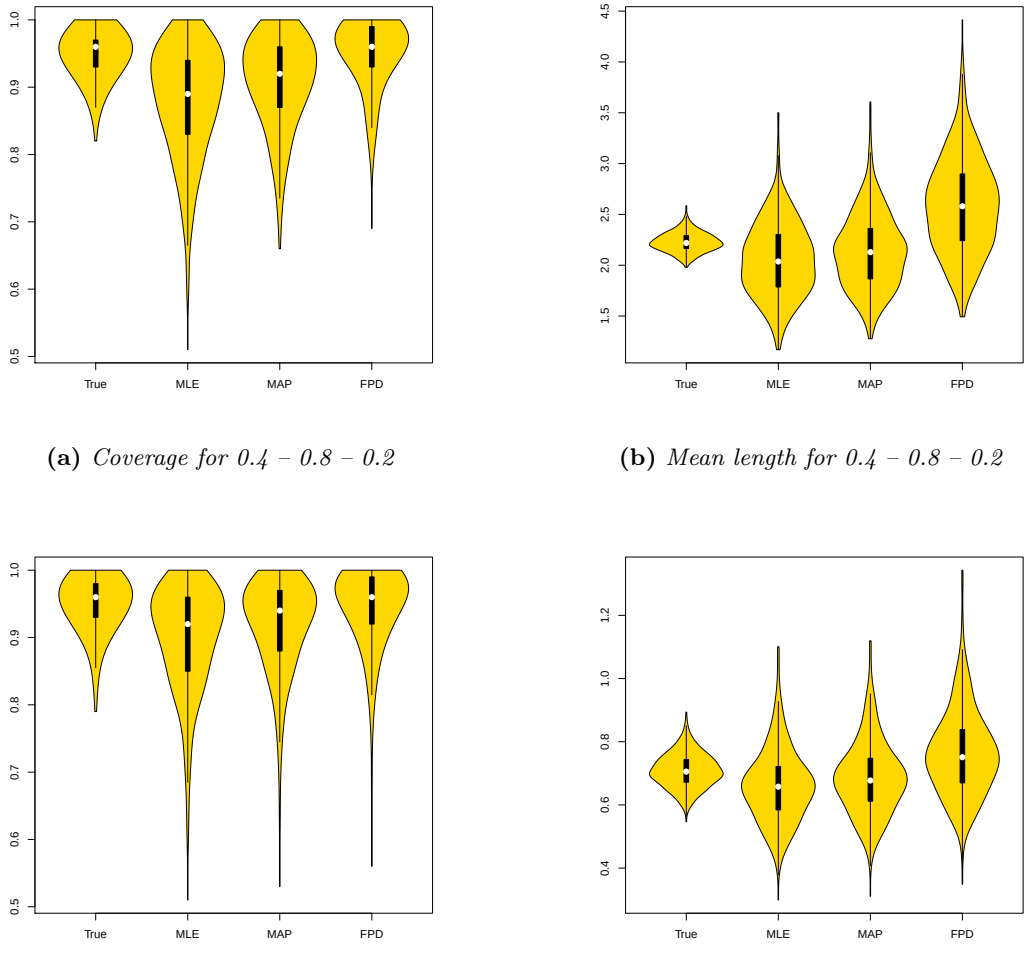
Let us now focus on the average (with respect to the uniform design sets and realizations of the Gaussian Process) of the mean (over the test set for a given realization of the Gaussian Process and a given uniform design set) length of prediction intervals. The results are given in Table 5.3, where the 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.

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)

Table 5.3 – *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.*

Predictive intervals derived from the full posterior distribution (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 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 5.2.

In Figure 5.2, 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.



(a) *Coverage for 0.4 – 0.8 – 0.2*

(b) *Mean length for 0.4 – 0.8 – 0.2*

(c) *Coverage for 0.8 – 1.0 – 0.9*

(d) *Mean length for 0.8 – 1.0 – 0.9*

Figure 5.2 – *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).*

A higher-dimensional case

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

$$A(\boldsymbol{x}) = 20 + \exp(1) - 20\exp\left(-0.2\sqrt{\frac{1}{10}\sum_{i=1}^{10} x_i^2}\right) - \exp\left(\frac{1}{10}\sum_{i=1}^{10}\cos(2\pi x_i)\right). \tag{5.12}$$

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 Latin Hypercube Sampling (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 5.4 and 5.5.

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 5.4 – *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.*

As is shown in Table 5.4, 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.

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

Table 5.5 – *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.*

As expected, prediction intervals derived from the Full Posterior Distribution 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.

5.5 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.
Although further investigation is needed to fully understand the properties of the optimal compromise, its invariance by reparametrization and its respect of pairwise independence show that it preserves important features of the conditional distributions.

This construction 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.

The next step, which is taken in Chapter 6, is to extend this 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 be done in the same way Berger et al. [2001] 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 relationship between correlation length $\boldsymbol{\theta}$ and smoothness ν. Unfortunately, asymptotic theory is not of much help in this regard, as Anderes [2010] 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 [2004] 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 [1999] (section 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 5.A Proofs of Section 5.2

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.

Differentiating the Matérn correlation kernel

Lemma 5.5. *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:*

$$\frac{\partial}{\partial \mu_i}\left(\sigma^2 K_{r,\nu}^{tens}\left(\boldsymbol{x}\boldsymbol{\mu}\right)\right) = -\frac{\sigma^2(2\sqrt{\nu})^2}{\Gamma(\nu)2^{\nu-1}}|x_i|^2\mu_i\left(2\sqrt{\nu}|x_i|\mu_i\right)^{\nu-1}\mathcal{K}_{\nu-1}\left(2\sqrt{\nu}|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right). \tag{5.13}$$

This can be rewritten as:

$$\frac{\partial}{\partial \mu_i}\left(\sigma^2 K_{r,\nu}^{tens}\left(\boldsymbol{x}\boldsymbol{\mu}\right)\right) = \begin{cases} \sigma^2\frac{2\nu}{\nu-1}|x_i|^2\mu_i K_{1,\nu-1}\left(|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) & \text{if } \nu>1 \\ \sigma^2 4|x_i|^2\mu_i\mathcal{K}_0\left(2|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) & \text{if } \nu=1 \\ \sigma^2\frac{2\nu^\nu\Gamma(1-\nu)}{\Gamma(\nu)}|x_i|^{2\nu}\mu_i^{2\nu-1}K_{1,1-\nu}\left(|x_i|\mu_i\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) & \text{if } \nu<1. \end{cases} \tag{5.14}$$

Proof. The first assertion is a simple matter of differentiating Equation (5.9). In the following calculation, the fourth line is given by formula 9.6.28 (page 376) in Abramowitz and Stegun [1964].

$$\begin{aligned} \frac{\partial}{\partial \mu_i}\left(\sigma^2 K_{r,\nu}^{tens}\left(\boldsymbol{x}\boldsymbol{\mu}\right)\right) &= \sigma^2\frac{\partial}{\partial \mu_i}\left(K_{1,\nu}\left(x_i\mu_i\right)\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) \\ &= \sigma^2 x_i\left(K'_{1,\nu}\left(x_i\mu_i\right)\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) \\ &= \sigma^2 x_i\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}\left.\frac{d}{dy}\right|_{y=2\sqrt{\nu}x_i\mu_i}\left[y^\nu\mathcal{K}_\nu(y)\right]\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right) \\ &= \sigma^2 x_i\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}\left[-y\cdot y^{\nu-1}\mathcal{K}_{\nu-1}(y)\right]_{y=2\sqrt{\nu}x_i\mu_i}\right)\prod_{j\neq i}K_{1,\nu}\left(|x_j|\mu_j\right). \end{aligned} \tag{5.15}$$

From there, Equation (5.13) follows immediately. Rewriting it in the form given in (5.14) 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 5.6. *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:*

$$\frac{\partial}{\partial \mu_i}\left(\sigma^2 K_{r,\nu}\left(\boldsymbol{x}\boldsymbol{\mu}\right)\right) = \frac{\sigma^2(2\sqrt{\nu})^2}{\Gamma(\nu)2^{\nu-1}}|x_i|^2\mu_i\left(2\sqrt{\nu}\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|\right)^{\nu-1}\mathcal{K}_{\nu-1}\left(2\sqrt{\nu}\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|\right). \tag{5.16}$$

This can be rewritten as:

$$\frac{\partial}{\partial \mu_i}\left(\sigma^2 K_{r,\nu}\left(\boldsymbol{x}\boldsymbol{\mu}\right)\right) = \begin{cases} \sigma^2\frac{2\nu}{\nu-1}|x_i|^2\mu_i K_{1,\nu-1}\left(\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|\right) & \text{if } \nu>1 \\ \sigma^2 4|x_i|^2\mu_i\mathcal{K}_0\left(2\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|\right) & \text{if } \nu=1 \\ \sigma^2 2\nu^\nu\frac{\Gamma(1-\nu)}{\Gamma(\nu)}\frac{1}{\mu_i}\left(\frac{|x_i|\mu_i}{\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|^{1-\nu}}\right)^2 K_{1,1-\nu}\left(\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|\right) & \text{if } \nu<1. \end{cases} \tag{5.17}$$

Proof. The first assertion is a simple matter of differentiating Equation (5.7). In the following calculation, the fourth line is given by formula 9.6.28 (page 376) in Abramowitz and Stegun [1964].

$$\begin{aligned}\frac{\partial}{\partial\mu_i}\left(\sigma^2 K_{r,\nu}\left(\boldsymbol{x}\boldsymbol{\mu}\right)\right) &= \sigma^2\frac{\partial}{\partial\mu_i}\left(K_{1,\nu}\left(\|\boldsymbol{x}\boldsymbol{\mu}\|\right)\right)\\ &= \sigma^2 x_i^2\mu_i\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|^{-1}K'_{1,\nu}\left(\|\boldsymbol{x}\boldsymbol{\mu}\|\right)\\ &= \sigma^2 x_i^2\mu_i\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|^{-1}\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}\left.\frac{d}{dy}\right|_{y=2\sqrt{\nu}\|\boldsymbol{x}\boldsymbol{\mu}\|}[y^\nu\mathcal{K}_\nu(y)]\right)\\ &= \sigma^2 x_i^2\mu_i\left\|\boldsymbol{x}\boldsymbol{\mu}\right\|^{-1}\left(\frac{2\sqrt{\nu}}{\Gamma(\nu)2^{\nu-1}}[-y\cdot y^{\nu-1}\mathcal{K}_{\nu-1}(y)]_{y=2\sqrt{\nu}\|\boldsymbol{x}\boldsymbol{\mu}\|}\right).\end{aligned} \tag{5.18}$$

From there, Equation (5.16) follows immediately. Rewriting it in the form given in (5.17) 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 $\nu < 1$). □

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 5.7. *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 $\nu \neq 1$, 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 5.5 or 5.6 after recalling that 1) a Matérn kernel is a bounded function, 2) $\forall\nu \geqslant 0$, as $z \to +\infty$, $\mathcal{K}_\nu(z) \sim \sqrt{\pi}\exp(-z)/\sqrt{2z}$ (Abramowitz and Stegun [1964] 9.7.2) and 3) as $z \to 0$, $\mathcal{K}_0(z) \sim -\log(z)$ (Abramowitz and Stegun [1964] 9.6.8). □

Let us define

$$f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := \sqrt{[\boldsymbol{\mathcal{I}}(\boldsymbol{\mu})]_{ii}}\,; \tag{5.19}$$

$$\pi_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) / \int_0^\infty f_i(\mu_i = t \mid \boldsymbol{\mu}_{-i})dt. \tag{5.20}$$

Proposition 5.8. *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 5.7 implies that $\int_0^\infty f_i(\mu_i = t \mid \boldsymbol{\mu}_{-i})dt$ is finite and, thanks to the dominated convergence theorem, that it is continuous at $\boldsymbol{\mu}_{-i} = \tilde{\boldsymbol{\mu}}_{-i}$. □

Definition 5.9. *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 5.10. *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 5.11. *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

$$h_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := \sqrt{\mathrm{Tr}\left[\left(\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)^2\right]} = \left\|\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\|. \tag{5.21}$$

Lemma 5.12. *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 5.11. □

Corollary 5.13. *For any well-behaved correlation kernel, there exist* $S > 0$, *and* $0 < a < b$ *such that, whenever* $\|\boldsymbol{\mu}\| \geqslant S$,

$$a\, h_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \leqslant f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \leqslant b\, h_i(\mu_i \mid \boldsymbol{\mu}_{-i}). \tag{5.22}$$

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 k-th row and l-th column.

Lemma 5.14. *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* $\|\boldsymbol{\mu}_{-i}\| \geqslant S$,

$$\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \geqslant c\frac{\|\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\|}{\sum_{k\neq l}\boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}}. \tag{5.23}$$

Proof. If a well-behaved correlation kernel is used, then for any for any $\epsilon > 0$, Corollary 5.13 implies that

$$\int_0^{+\infty} f_i(\mu_i = t|\boldsymbol{\mu}_{-i})dt \leqslant b\int_0^{+\infty} h_i(\mu_i = t|\boldsymbol{\mu}_{-i})dt \leqslant -b\sum_{k\neq l}\int_0^{+\infty}\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(kl)}dt. \tag{5.24}$$

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 $\mu_i \to +\infty$, $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(kl)} \to 0$ and when $\mu_i = 0$, $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{(kl)} = \boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}$. From this, we gather that

$$\int_0^{+\infty} f_i(\mu_i = t|\boldsymbol{\mu}_{-i})dt \leqslant b\sum_{k\neq l}\boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}. \tag{5.25}$$

From this, we deduce that

$$\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) = \frac{f_i(\mu_i|\boldsymbol{\mu}_{-i})}{\int_0^{+\infty} f_i(\mu_i = t|\boldsymbol{\mu}_{-i})dt} \geqslant \frac{a}{b}\frac{\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\|}{\sum_{k\neq l}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}}. \tag{5.26}$$

□

This Lemma has the following immediate consequence:

Proposition 5.15. *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 $M_i^{(kl)} : (0, +\infty) \to (0, +\infty)$; $t \mapsto \left|\frac{d}{dt}\boldsymbol{\Sigma}^{(kl)}_{\mu_i=t,\boldsymbol{\mu}_{-i}=\mathbf{0}_{r-1}}\right|$.

$$\left\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\| \geqslant \frac{1}{n}\sum_{k\neq l}\left|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}}\right| = \frac{1}{n}\sum_{k\neq l} M_i^{(kl)}(\mu_i)\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}} \geqslant \frac{1}{n}\min_{k\neq l} M_i^{(kl)}(\mu_i)\sum_{k\neq l}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}_{-i}}. \tag{5.27}$$

This fact, joined with Lemma 5.14, yields the result. □

Proposition 5.16. *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 for all $t \geqslant S_1$, $|K'(\sqrt{t^2 + a^2})| \geqslant M_1\sqrt{t^2 + a^2}K(\sqrt{t^2 + a^2})$. 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

$$|K'(\sqrt{t^2 + a^2})| \geqslant M_1M_2(a)\sqrt{t^2 + a^2}K(t). \tag{5.28}$$

Independently from this, we have the following algebraic fact:

$$\left\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\| \geqslant \frac{1}{n}\sum_{k\neq l}\left|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}}\right|. \tag{5.29}$$

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:

$$\left|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}}\right| = -\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}^{(kl)}_{\boldsymbol{\mu}} = \left(x_i^{(k)} - x_i^{(l)}\right)^2\mu_i\frac{K'\left(\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|\right)}{\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|}. \tag{5.30}$$

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 (5.28) yields the existence of some number $m_i^{(kl)}(\mu_i) > 0$ such that

$$\frac{K'\left(\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|\right)}{\|(\boldsymbol{x}^{(k)} - \boldsymbol{x}^{(l)})\boldsymbol{\mu}\|} \geqslant m_i^{(kl)}(\mu_i) K(\|(\boldsymbol{x}_{-i}^{(k)} - \boldsymbol{x}_{-i}^{(l)})\boldsymbol{\mu}_{-i}\|) = m_i^{(kl)}(\mu_i)\boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}. \tag{5.31}$$

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

$$\left\|\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\| \geqslant \frac{m_i(\mu_i)}{n}\sum_{k\neq l}\boldsymbol{\Sigma}_{\boldsymbol{\mu}_{-i}}^{(kl)}. \tag{5.32}$$

Then, applying Lemma 5.14 yields the result. □

Proposition 5.17. *Matérn one-dimensional kernels with smoothness parameter $\nu > 1$ have the properties (P1) and (P2) of Proposition 5.16.*

Proof. (P1) is given by Lemma 5.6, 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 $t \to +\infty$, $\mathcal{K}_\nu(t) \sim \sqrt{\pi}\exp(-t)/\sqrt{2t}$ (Abramowitz and Stegun [1964] 9.7.2), hence $K_\nu(t) \sim 2/\Gamma(\nu)(\sqrt{\nu}t)^\nu\sqrt{\pi/(4\sqrt{\nu}t)}\exp(-2\sqrt{\nu}t)$. Moreover, this last equivalence relation also implies (P2). □

Proposition 5.18. *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 5.15 or 5.16 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 5.8 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 5.19. *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 5.18, yields the result. □

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 5.20. *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 $\boldsymbol{\xi}=(\xi_1,...,\xi_n)\in\mathbb{R}^n$,*

$$\begin{aligned}&\sum_{j,k=1}^{n}\xi_j\xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)}-\boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right)\\ &\geqslant 2^{-\frac{r}{2}-\nu}M_r(\nu)f_{r,\nu}(\|\boldsymbol{\mu}\|_\infty)\sum_{j,k=1}^{n}\xi_j\xi_k\widetilde{K}_{r,\nu}\left(\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty}\left(\boldsymbol{x}^{(j)}-\boldsymbol{x}^{(k)}\right)\right),\end{aligned}\tag{5.33}$$

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

$$\begin{aligned}&\sum_{j,k=1}^{n}\xi_j\xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)}-\boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right)\\ =&\int_{\mathbb{R}^r}\widehat{K_{r,\nu}}(\boldsymbol{\omega})\left|\sum_{j=1}^{n}\xi_j e^{i\langle\boldsymbol{\omega}|\boldsymbol{x}^{(j)}\boldsymbol{\mu}\rangle}\right|^2 d\boldsymbol{\omega}\\ =&M_r(\nu)\|\boldsymbol{\mu}\|_\infty^{-r}\int_{\mathbb{R}^r}\left(4\nu+\|\boldsymbol{\mu}\|_\infty^{-2}\|\boldsymbol{s}\|^2\right)^{-\frac{r}{2}-\nu}\left|\sum_{j=1}^{n}\xi_j e^{i\langle\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty}\boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s}\\ \geqslant&2^{-\frac{r}{2}-\nu}M_r(\nu)f_{r,\nu}(\|\boldsymbol{\mu}\|_\infty)\int_{\mathbb{R}^r\setminus B(0,1)}\|\boldsymbol{s}\|^{-r-2\nu}\left|\sum_{j=1}^{n}\xi_j e^{i\langle\boldsymbol{s}|\frac{\boldsymbol{\mu}}{\|\boldsymbol{\mu}\|_\infty}\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s}.\end{aligned}\tag{5.34}$$

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 5.21. *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 $\boldsymbol{\mu}\in(\mathbb{R}_+)^r$,*

$$\forall\boldsymbol{\xi}=(\xi_1,...,\xi_n)\in\mathbb{R}^n,\ \sum_{j,k=1}^{n}\xi_j\xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)}-\boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right)\geqslant c_{\boldsymbol{x}}\|\boldsymbol{\xi}\|^2 2^{-\frac{r}{2}-\nu}M_r(\nu)f_{r,\nu}(\|\boldsymbol{\mu}\|_\infty)\tag{5.35}$$

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 5.20. □

Proposition 5.22. *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 5.21. In the case of tensorized Matérn kernels, analoguous results to Lemma 5.20 and then Lemma 5.21 may be used. □

Abramowitz and Stegun [1964] 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 $\psi : \mathbb{N} \setminus \{0\} \to \mathbb{R}\ ;\ k \mapsto -\gamma + \sum_{i=1}^{k-1} i^{-1}$:

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

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

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

$$\begin{aligned}
&K_{1,\nu}(z)\\
=&\frac{\pi}{\Gamma(\nu)\sin(\nu\pi)}\left(\sum_{0\leqslant k<\nu} \frac{\nu^k z^{2k}}{k!\Gamma(-\nu+k+1)} - \frac{\nu^\nu z^{2\nu}}{\Gamma(\nu+1)} + o\left(z^{2\nu}\right)\right)\\
=&\frac{\pi}{\Gamma(\nu)\sin(\nu\pi)\Gamma(-\nu+1)}\left(\sum_{0\leqslant k<\nu} \frac{\Gamma(-\nu+1)}{k!\Gamma(-\nu+k+1)}\nu^k z^{2k} - \frac{\Gamma(-\nu+1)}{\Gamma(\nu+1)}\nu^\nu z^{2\nu} + o\left(z^{2\nu}\right)\right)\\
=&\sum_{0\leqslant k<\nu} \frac{\Gamma(-\nu+1)}{k!\Gamma(-\nu+k+1)}\nu^k z^{2k} + \frac{\Gamma(-\nu)}{\Gamma(\nu)}\nu^\nu z^{2\nu} + o\left(z^{2\nu}\right)\\
=&\sum_{0\leqslant k<\nu} (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)}\nu^k z^{2k} + \frac{\Gamma(-\nu)}{\Gamma(\nu)}\nu^\nu z^{2\nu} + o\left(z^{2\nu}\right).
\end{aligned} \tag{5.36}$$

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

$$\boldsymbol{D}^k(i,j) := (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)}\nu^k \left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right\|^{2k}. \tag{5.37}$$

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

$$\boldsymbol{D}^\nu(i,j) := \frac{\Gamma(-\nu)}{\Gamma(\nu)}\nu^\nu \left\|\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right\|^{2\nu} \qquad \text{if } \nu \in [0,+\infty) \setminus \mathbb{N}. \tag{5.38}$$

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

$$\boldsymbol{\Sigma}_\mu = \sum_{0\leqslant k<\nu} \mu^{2k}\boldsymbol{D}^k + \mu^{2\nu}\boldsymbol{D}^\nu + \boldsymbol{R}_\mu. \tag{5.39}$$

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

$$\boldsymbol{D}_i^k(m,p) := (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)}\nu^k \left|x_i^{(m)} - x_i^{(p)}\right|^{2k} \tag{5.40}$$

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

$$\boldsymbol{D}_i^\nu(m,p) := \frac{\Gamma(-\nu)}{\Gamma(\nu)}\nu^\nu \left|x_i^{(m)} - x_i^{(p)}\right|^{2\nu}. \tag{5.41}$$

For every $i \in [\![1, r]\!]$, if the points in the design set differed only through their i-th 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

$$\boldsymbol{\Sigma}_{\mu_i} = \sum_{0 \leqslant k < \nu} \mu_i^{2k} \boldsymbol{D}_i^k + \mu_i^{2\nu} \boldsymbol{D}_i^\nu + \boldsymbol{R}_{\mu_i} \tag{5.42}$$

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

Note the following identities:

$$\boldsymbol{D}_i^0 = \mathbf{1}\mathbf{1}^\top; \tag{5.43}$$

$$\boldsymbol{D}_i^1 = -\frac{\Gamma(\nu-1)}{\Gamma(\nu)} \nu \left\{ \mathbf{1} \left(\boldsymbol{X}_i^{\circ 2}\right)^\top + \left(\boldsymbol{X}_i^{\circ 2}\right) \mathbf{1}^\top - 2\boldsymbol{X}_i \boldsymbol{X}_i^\top \right\}; \tag{5.44}$$

$$\begin{aligned} \boldsymbol{D}_i^2 = \frac{\Gamma(\nu-2)}{\Gamma(\nu)} \nu^2 \Big\{ & \mathbf{1} \left(\boldsymbol{X}_i^{\circ 4}\right)^\top + \left(\boldsymbol{X}_i^{\circ 4}\right) \mathbf{1}^\top - 4\boldsymbol{X}_i \left(\boldsymbol{X}_i^{\circ 3}\right)^\top \\ & -4 \left(\boldsymbol{X}_i^{\circ 3}\right) \boldsymbol{X}_i^\top + 6 \left(\boldsymbol{X}_i^{\circ 2}\right) \left(\boldsymbol{X}_i^{\circ 2}\right)^\top \Big\}. \end{aligned} \tag{5.45}$$

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

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{tens} = \overset{\circ}{\prod_{i \in [\![1,r]\!]}} \boldsymbol{\Sigma}_{\mu_i} \tag{5.46}$$

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

$$\boldsymbol{D}^k(\boldsymbol{\mu})(m,p) := (-1)^k \frac{\Gamma(\nu-k)}{k!\Gamma(\nu)} \nu^k d_{m,p}(\boldsymbol{\mu})^{2k} \tag{5.47}$$

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

$$\boldsymbol{D}^\nu(\boldsymbol{\mu})(m,p) := \frac{\Gamma(-\nu)}{\Gamma(\nu)} \nu^\nu d_{m,p}(\boldsymbol{\mu})^{2\nu} \qquad \text{if } \nu \in [0, +\infty) \setminus \mathbb{N}. \tag{5.48}$$

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

$$\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{geom} = \sum_{0 \leqslant k < \nu} \boldsymbol{D}^k(\boldsymbol{\mu}) + \boldsymbol{D}^\nu(\boldsymbol{\mu}) + \boldsymbol{R}_{\boldsymbol{\mu}}^{geom} \tag{5.49}$$

where $\|\boldsymbol{\mu}\|^{-2\nu} \, \|\boldsymbol{R}_{\boldsymbol{\mu}}^{geom}\| \to 0$ as $\|\boldsymbol{\mu}\| \to 0$.
Similar identities to those of Equation (5.43) can be derived to make Equation (5.49) more explicit for small values of ν.

$$\boldsymbol{D}^0(\boldsymbol{\mu}) = \mathbf{1}\mathbf{1}^\top; \tag{5.50}$$

$$\boldsymbol{D}^1(\boldsymbol{\mu}) = -\frac{\nu}{\nu-1}\left\{\sum_{i=1}^r \mu_i^2\left(\mathbf{1}\left(\boldsymbol{X}_i^{\circ 2}\right)^\top + \left(\boldsymbol{X}_i^{\circ 2}\right)\mathbf{1}^\top - 2\boldsymbol{X}_i\boldsymbol{X}_i^\top\right)\right\}; \tag{5.51}$$

$$\begin{aligned}\boldsymbol{D}^2(\boldsymbol{\mu}) = \frac{\nu^2}{(\nu-1)(\nu-2)}\Bigg\{&\sum_{i,j\in[\![1,r]\!]} \mu_i^2\mu_j^2\Big(\mathbf{1}\left(\boldsymbol{X}_i^{\circ 2}\circ\boldsymbol{X}_j^{\circ 2}\right)^\top + \left(\boldsymbol{X}_i^{\circ 2}\circ\boldsymbol{X}_j^{\circ 2}\right)\mathbf{1}^\top \\ &- 2\boldsymbol{X}_i\left(\boldsymbol{X}_i\circ\boldsymbol{X}_j^{\circ 2}\right)^\top - 2\left(\boldsymbol{X}_i\circ\boldsymbol{X}_j^{\circ 2}\right)\boldsymbol{X}_i^\top - 2\boldsymbol{X}_j\left(\boldsymbol{X}_j\circ\boldsymbol{X}_i^{\circ 2}\right)^\top - 2\left(\boldsymbol{X}_j\circ\boldsymbol{X}_i^{\circ 2}\right)\boldsymbol{X}_j^\top \\ &+ \left(\boldsymbol{X}_i^{\circ 2}\right)\left(\boldsymbol{X}_j^{\circ 2}\right)^\top + \left(\boldsymbol{X}_j^{\circ 2}\right)\left(\boldsymbol{X}_i^{\circ 2}\right)^\top + 4\left(\boldsymbol{X}_i\circ\boldsymbol{X}_j\right)\left(\boldsymbol{X}_i\circ\boldsymbol{X}_j\right)^\top\Big)\Bigg\}.\end{aligned} \tag{5.52}$$

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

$$\text{For } \nu\in(0,1): \quad \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{tens} = \mathbf{1}\mathbf{1}^\top + \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{D}_i^\nu + \boldsymbol{R}_{\boldsymbol{\mu}}^{tens}. \tag{5.53}$$

$$\text{For } \nu\in(1,2): \quad \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{tens} = \mathbf{1}\mathbf{1}^\top + \boldsymbol{D}^1(\boldsymbol{\mu}) + \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{D}_i^\nu + \boldsymbol{R}_{\boldsymbol{\mu}}^{tens}. \tag{5.54}$$

$$\text{For } \nu\in(2,3): \quad \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{tens} = \mathbf{1}\mathbf{1}^\top + \boldsymbol{D}^1(\boldsymbol{\mu}) + \frac{\nu-2}{\nu-1}\boldsymbol{D}^2(\boldsymbol{\mu}) + \sum_{i=1}^r \mu_i^4\boldsymbol{D}_i^2 + \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{D}_i^\nu + \boldsymbol{R}_{\boldsymbol{\mu}}^{tens}. \tag{5.55}$$

In the three expressions above, $\|\boldsymbol{\mu}\|^{-2\nu}\,\|\boldsymbol{R}_{\boldsymbol{\mu}}^{tens}\| \to 0$ 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 $\boldsymbol{v}\in k_\nu$,

$$\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{geom}\boldsymbol{v} = \boldsymbol{v}^\top\boldsymbol{D}^\nu(\boldsymbol{\mu})\boldsymbol{v} + \boldsymbol{v}^\top\boldsymbol{R}_{\boldsymbol{\mu}}^{geom}\boldsymbol{v}, \tag{5.56}$$

$$\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{tens}\boldsymbol{v} = \sum_{i=1}^r \mu_i^{2\nu}\boldsymbol{v}\boldsymbol{D}_i^\nu\boldsymbol{v} + \boldsymbol{v}^\top\boldsymbol{R}_{\boldsymbol{\mu}}^{tens}\boldsymbol{v}. \tag{5.57}$$

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

$$\max(\boldsymbol{v}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{geom}\boldsymbol{v}, \boldsymbol{v}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{tens}\boldsymbol{v}) \leqslant c\|\boldsymbol{\mu}\|^{2\nu}\boldsymbol{v}^\top\boldsymbol{v}. \tag{5.58}$$

Proposition 5.23. *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 $\boldsymbol{\Sigma_\mu} = \boldsymbol{O}_\nu \boldsymbol{O}_\nu^\top \boldsymbol{\Sigma_\mu} \boldsymbol{O}_\nu \boldsymbol{O}_\nu^\top$. Consider the following decomposition of $\boldsymbol{O}_\nu^\top \boldsymbol{\Sigma_\mu} \boldsymbol{O}_\nu$:

$$\boldsymbol{O}_\nu^\top \boldsymbol{\Sigma_\mu} \boldsymbol{O}_\nu = \begin{pmatrix} \boldsymbol{A_\mu} & \boldsymbol{B_\mu} \\ \boldsymbol{B_\mu}^\top & \boldsymbol{C_\mu} \end{pmatrix} \tag{5.59}$$

where the blocks $\boldsymbol{A_\mu}$, $\boldsymbol{B_\mu}$ and $\boldsymbol{C_\mu}$ are respectively $(n - d_\nu) \times (n - d_\mu)$, $(n - d_\mu) \times d_\mu$ and $d_\mu \times d_\mu$ matrices. Note that $\boldsymbol{A_\mu}$ and $\boldsymbol{C_\mu}$ represent the restriction of the scalar product defined by $\boldsymbol{\Sigma_\mu}$ to $k_\nu^\perp$ and k_ν respectively. When $\|\boldsymbol{\mu}\|$ is small enough, defining $c > 0$ as in Equation (5.58), $\|\boldsymbol{C_\mu}\| \leqslant c\|\boldsymbol{\mu}\|^{2\nu}$.

$$\boldsymbol{O}_\nu^\top \boldsymbol{\Sigma_\mu}^{-1} \boldsymbol{O}_\nu = \begin{pmatrix} \boldsymbol{I}_{n-d_\nu} & \boldsymbol{0} \\ -\boldsymbol{B_\mu} \boldsymbol{C_\mu}^{-1} & \boldsymbol{I}_{d_\nu} \end{pmatrix} \begin{pmatrix} \left(\boldsymbol{A} - \boldsymbol{B_\mu} \boldsymbol{C_\mu}^{-1} \boldsymbol{B_\mu}^\top\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C_\mu}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{I}_{n-d_\nu} & -\boldsymbol{C_\mu}^{-1} \boldsymbol{B_\mu}^\top \\ \boldsymbol{0} & \boldsymbol{I}_{d_\nu} \end{pmatrix}. \tag{5.60}$$

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

$$\boldsymbol{O}_\nu^\top \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \end{pmatrix}, \tag{5.61}$$

$$\boldsymbol{y}^\top \boldsymbol{\Sigma_\mu}^{-1} \boldsymbol{y} = \begin{pmatrix} \boldsymbol{y}_1 - \boldsymbol{C_\mu}^{-1} \boldsymbol{B_\mu} \boldsymbol{y}_2 \\ \boldsymbol{y}_2 \end{pmatrix}^\top \begin{pmatrix} \left(\boldsymbol{A} - \boldsymbol{B_\mu} \boldsymbol{C_\mu}^{-1} \boldsymbol{B_\mu}^\top\right)^{-1} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{C_\mu}^{-1} \end{pmatrix} \begin{pmatrix} \boldsymbol{y}_1 - \boldsymbol{C_\mu}^{-1} \boldsymbol{B_\mu} \boldsymbol{y}_2 \\ \boldsymbol{y}_2 \end{pmatrix}. \tag{5.62}$$

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

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

□

Proposition 5.24. *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 $\boldsymbol{\Sigma_\mu}$. We can now rewrite $L(\boldsymbol{y}|\boldsymbol{\mu})$ as

$$L(\boldsymbol{y}|\boldsymbol{\mu})^2 \propto \prod_{k=1}^{n} \left[v_k(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma_\mu}^{-1} \boldsymbol{y}\right)^{-1} \right]. \tag{5.63}$$

Proposition 5.23 asserts that for any $\boldsymbol{y} \in \mathbb{R}^n$ that is not orthogonal to k_ν, $\left(\boldsymbol{y}^\top \boldsymbol{\Sigma_\mu}^{-1} \boldsymbol{y}\right)^{-1} = O(\|\boldsymbol{\mu}\|^{2\nu})$ for $\|\boldsymbol{\mu}\| \to 0$. Besides, Proposition 5.22 asserts that $\left\|\boldsymbol{\Sigma_\mu}^{-1}\right\| = O(\|\boldsymbol{\mu}\|^{-2\nu})$, so $\left(\boldsymbol{y}^\top \boldsymbol{\Sigma_\mu}^{-1} \boldsymbol{y}\right)^{-1} = O\left(\left\|\boldsymbol{\Sigma_\mu}^{-1}\right\|^{-1}\right)$.
This implies that for every integer $i \in [\![1, r]\!]$, $v_k(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{\Sigma_\mu}^{-1} \boldsymbol{y}\right)^{-1} = O(1)$.

Clearly, $\lim_{\|\boldsymbol{\mu}\| \to 0} |\boldsymbol{1}^\top \boldsymbol{v}_1(\boldsymbol{\mu})| = \|\boldsymbol{1}\|$ 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 $\lim_{\|\boldsymbol{\mu}\|\to 0} \mathbf{1}^\top \boldsymbol{v}_2(\boldsymbol{\mu}) = 0$.

$$
\begin{aligned}
& v_2(\boldsymbol{\mu}) \\
=& (\mathbf{1}^\top \boldsymbol{v}_2(\boldsymbol{\mu}))^2 + 2\nu(\nu-1)^{-1}\sum_{i=1}^r \mu_i^2 \left(\boldsymbol{X}_i^\top \boldsymbol{v}_2(\boldsymbol{\mu})\right)^2 - 2\mu_i^2 \left(\mathbf{1}^\top \boldsymbol{v}_2(\boldsymbol{\mu})\right)\left(\boldsymbol{X}_i^{\circ 2\top} \boldsymbol{v}_2(\boldsymbol{\mu})\right) + O\left(\|\boldsymbol{\mu}\|^4\right) \\
\geqslant & 2\nu(\nu-1)^{-1}\sum_{i=1}^r \mu_i^2 (\boldsymbol{X}_i^\top \boldsymbol{v}_2(\boldsymbol{\mu}))^2 + o\left(\|\boldsymbol{\mu}\|^2\right). \qquad (5.64)
\end{aligned}
$$

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 $\boldsymbol{X}_i^\top \boldsymbol{w}_i(\boldsymbol{\mu}) > 0$.

$$
\boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \geqslant 2\nu(\nu-1)^{-1} r^{-1} \|\boldsymbol{\mu}\|^2 (\boldsymbol{X}_{i(\boldsymbol{\mu})}^\top \boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}))^2 + o\left(\|\boldsymbol{\mu}\|^2\right). \qquad (5.65)
$$

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

$$
\boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{w}_{i(\boldsymbol{\mu})}(\boldsymbol{\mu}) \geqslant c_2 \|\boldsymbol{\mu}\|^2. \qquad (5.66)
$$

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

Proposition 5.25. *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 5.7, $\left\|\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\| = O(\mu_i^{-1+2\nu})$. Putting all this together, $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i}) = O(\mu_i^{-1+\nu})$. □

Proposition 5.26. *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,

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

Proof. Given Proposition 5.8 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

$$\begin{aligned} &\int_0^\infty L(\boldsymbol{y}|\mu_i = t, \boldsymbol{\mu}_{-i}) f_i(\mu_i = t|\boldsymbol{\mu}_{-i}) dt \underset{\|\boldsymbol{\mu}_{-i}\| \to 0}{\longrightarrow} \\ &\qquad \int_0^\infty L(\boldsymbol{y}|\mu_i = t, \boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1}) f_i(\mu_i = t|\boldsymbol{\mu}_{-i} = \mathbf{0}_{r-1}) dt < +\infty. \end{aligned} \tag{5.67}$$

Lemma 5.7 implies that there exists $M_i > 0$ such that

$$\begin{aligned} L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i}) &= L(\boldsymbol{y}|\boldsymbol{\mu}) \left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| \left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\|^{-1} f_i(\mu_i|\boldsymbol{\mu}_{-i}) \\ &\leqslant L(\boldsymbol{y}|\boldsymbol{\mu}) \left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| M_i \mu_i^{-2}. \end{aligned} \tag{5.68}$$

Lemma 5.10 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 5.24 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 5.25 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 the product $L(\boldsymbol{y}|\boldsymbol{\mu}) f_i(\mu_i|\boldsymbol{\mu}_{-i})$ and integrable over $\mu_i \in (0, +\infty)$, so the dominated convergence theorem is applicable.

□

Lower bound for conditional reference posterior densities

The following Lemma provides the key to proving Theorem 5.2:

Lemma 5.27. *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:*

$$\pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) \geqslant m_{i,\boldsymbol{y}}(\mu_i) > 0. \tag{5.69}$$

Proof. This proof consists in combining Proposition 5.19 and Proposition 5.26, which respectively deal with large and small values of $\|\boldsymbol{\mu}_{-i}\|$.

Proposition 5.19 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

$$\inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in [0, +\infty)^{r-1} \setminus N_i(\mu_i)\} > 0. \tag{5.70}$$

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 5.26 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 $[0,+\infty)^r$,

$$\inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in N_i(\mu_i)\} > 0. \tag{5.71}$$

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

$$\inf\{\pi(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i}) : \boldsymbol{\mu}_{-i} \in [0,+\infty)^{r-1}\} > 0. \tag{5.72}$$

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 5.2. Lemma 5.27 implies that $\forall \boldsymbol{\mu}^{(0)} \in (0,+\infty)^r$,

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

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

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

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:

$$\forall \boldsymbol{\mu}^{(0)} \in (0,+\infty)^r \qquad P_{\boldsymbol{y}}^n(\boldsymbol{\mu}^{(0)}, d\boldsymbol{\mu}) \geqslant \epsilon_{\boldsymbol{y}} \left(\frac{1}{\epsilon_{\boldsymbol{y}}} f_{\boldsymbol{y}}(\boldsymbol{\mu}) \right) d\boldsymbol{\mu}. \tag{5.73}$$

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 $\pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i})$.

□

Chapter 6

A Comprehensive Bayesian Treatment of the Universal Kriging model with Matérn correlation kernels

This chapter mostly adheres to the article Muré [2018].

Abstract

The Gibbs reference posterior distribution provides an objective full-Bayesian solution to the problem of prediction of a stationary Gaussian process with Matérn anisotropic kernel. A full-Bayesian approach is possible, because the posterior distribution is expressed as the invariant distribution of a uniformly ergodic Markovian kernel for which we give an explicit expression. In this chapter, we show that it is appropriate for the Universal Kriging framework, that is when an unknown function is added to the stationary Gaussian process. We give sufficient conditions for the existence and propriety of the Gibbs reference posterior that apply to a wide variety of practical cases and illustrate the method with several examples. Finally, simulations of Gaussian processes suggest that the Gibbs reference posterior has good frequentist properties in terms of coverage of prediction intervals.

Résumé

Le posterior de référence de Gibbs fournit une solution pleinement bayésienne au problème de la prédiction des valeurs prises par un processus gaussien stationnaire avec noyau de Matérn anisotrope. Une approche pleinement bayésienne est possible parce que la loi *a posteriori* est exprimée comme loi invariante d'un noyau markovien uniformément ergodique dont nous donnons une expression explicite. Dans ce chapitre, nous montrons que cette approche est indiquée dans le cadre du krigeage universel, c'est-à-dire quand une fonction inconnue est ajoutée à un processus gaussien stationnaire. Nous donnons des conditions suffisantes à l'existence et la propreté du posterior de référence de Gibbs qui s'appliquent à une large collection de cas pratiques et illustrons la méthode par plusieurs exemples. Enfin, des simulation des processus gaussien suggèrent que le posterior de référence de Gibbs a de bonnes propriétés fréquentistes en termes de couverture et d'intervalles prédictifs.

6.1 Introduction

In Simple Kriging models, the Gaussian Process is assumed to have zero mean and be stationary on the domain $\mathcal{D}$, so its distribution can be characterized by a positive variance parameter σ^2 and by an autocorrelation function K. The Universal Kriging framework adds another parameter : a mean function f. If f is known, then subtracting it from the process returns us to the Simple Kriging framework. Allowing for an unknown mean function provides greater flexibility in the modeling by enabling some degree of non-stationarity [Santner et al., 2003, section 2.3.2].

In practice, the mean function f is assumed to belong to a p-dimensional ($p \in \mathbb{N}$) vector space $\mathcal{F}_p$, which is specified by means of a basis $(f_1, ..., f_p)$. Being a linear combination of $f_1, ..., f_p$, the mean function f is then encoded by the vector of linear coefficients $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^\top$: $f = \beta_1 f_1 + ... + \beta_p f_p$.
Therefore, what separates the Universal Kriging framework from its Simple counterpart is the addition of the p-dimensional parameter $\boldsymbol{\beta}$.

Assuming the autocorrelation function to be characterized by a vector of correlation lengths $\boldsymbol{\theta}$, we are faced with the inference problem of estimating $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$. In the previous chapter, an objective posterior distribution on $(\sigma^2, \boldsymbol{\theta})$ was proposed in the context of Simple Kriging. In this chapter we address the more general framework of Universal Kriging in order to obtain a distribution on $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$. The developments in both chapters are based on Bernardo's reference prior theory. The idea to use this theory in the context of Kriging first appeared in Berger et al. [2001], and then was successively extended by Paulo [2005], Kazianka and Pilz [2012], Ren et al. [2012], Ren et al. [2013] and Gu [2016].

To use it we first need to order the parameters [Bernardo, 2005]. Because our main goal is to maximize the predictive capacity of the model, we are unable to outright say which parameter we care about most. However, a few common sense observations help: first, in order to profit from the work done in the Simple Kriging case, we separate $\boldsymbol{\beta}$, which refers to the mean function, from $(\sigma^2, \boldsymbol{\theta})$, which yields the covariance structure. Within the latter, $\boldsymbol{\theta}$ should have the priority over σ^2, because while σ^2 can very easily be accurately estimated once $\boldsymbol{\theta}$ is known, the reverse is not true. The same consideration will make us prioritize $(\sigma^2, \boldsymbol{\theta})$ over $\boldsymbol{\beta}$, because while knowing $\boldsymbol{\beta}$ reduces the problem to the Simple Kriging case, knowing $(\sigma^2, \boldsymbol{\theta})$ reduces it to a much simpler regression problem.

In Section 6.2 we derive the reference posterior distribution on $(\boldsymbol{\beta}, \sigma^2)$ and the corresponding predictive distribution at unobserved points, both conditional to the observed data and the correlation parameter $\boldsymbol{\theta}$.
In Section 6.3, we derive analytical formulas for the reference prior on $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ in the case where $\boldsymbol{\theta}$ is a one-dimensional parameter.
In Section 6.4, we prove the main result of the chapter: in the context of a Matérn anisotropic correlation kernel [Matérn, 1986, Handcock and Stein, 1993] – see Appendix 6.A for precise definitions – under a few conditions, the Gibbs reference posterior on a multidimensional $\boldsymbol{\theta}$ exists. Combined with the "partial" reference posterior distribution on $(\boldsymbol{\beta}, \sigma^2)$ conditional to $\boldsymbol{\theta}$, it provides a proper objective posterior distribution on all parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ given the observed data. It is significant that this proper objective posterior distribution is well

defined for Matérn anisotropic correlation kernels, because this class of correlation kernels has remarkable properties (see Stein [1999] or Bachoc [2013a, chapter 2]). Notably, it allows the user to specify the smoothness of the realizations of the Gaussian Process.
In Section 6.5, we evaluate the predictive performance of the Universal Kriging model with the Gibbs reference posterior distribution both in the context of a well-specified model and when emulating deterministic functions. We compare the full-Bayesian approach relying on the Gibbs reference posterior with plug-in approaches, where the parameters are assumed to be equal to either the Maximum Likelihood Estimator (MLE) or the Maximum A Posteriori (MAP) estimator.

6.2 Analytical treatment of the location-scale parameters

Suppose our design set contains n observation points. n must be greater than p, otherwise the model is not identifiable. Let $\boldsymbol{H}$ be the $n \times p$ matrix whose columns contain the values of the p basis functions at the n observation points. Let us assume that the rank of $\boldsymbol{H}$ is p, because if it were not, the model would also not be identifiable.
Let $\boldsymbol{y}$ be the vector of the n observations. Then $\boldsymbol{y}$ is a Gaussian vector and its distribution is

$$\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{H}\boldsymbol{\beta}, \sigma^2 \boldsymbol{\Sigma}_{\boldsymbol{\theta}}), \tag{6.1}$$

where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}}$ is a correlation matrix that only depends on the design set and on the vector of correlation lengths $\boldsymbol{\theta}$.

In terms of likelihood, we have

$$L(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2} (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\beta})^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1} (\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\beta})\right\} . \tag{6.2}$$

The aim of this section is to get the parameters $\boldsymbol{\beta}$ and σ^2 out of the way in order to focus on the more interesting parameter $\boldsymbol{\theta}$. For now, assume that $\boldsymbol{\theta}$ is known, which is to say that the correlation function is completely known.

Reference prior and integrated likelihood when $\boldsymbol{\theta}$ is known.

Clearly, $\boldsymbol{\beta}$ is a location parameter and $\sigma := \sqrt{\sigma^2}$ is a scale parameter for this model. Therefore, the joint reference prior is $\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\theta}) \propto 1/\sigma^2$ regardless of the order of the parameters $(\boldsymbol{\beta}, \sigma^2)$.

We now derive the posterior distributions $\pi(\boldsymbol{\beta}|\boldsymbol{y}, \sigma^2, \boldsymbol{\theta})$ and $\pi(\sigma^2|\boldsymbol{y}, \boldsymbol{\theta})$ as well as the integrated likelihoods $L^0(\boldsymbol{y}|\sigma^2, \boldsymbol{\theta}) := \int L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) d\boldsymbol{\beta}$ and $L^1(\boldsymbol{y}|\boldsymbol{\theta}) := \iint L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})/\sigma^2 d\boldsymbol{\beta} d\sigma^2$.

Gaussian theory makes it convenient to split $\boldsymbol{y}$ into two components: one that belongs to the subspace of $\mathbb{R}^n$ spanned by $\boldsymbol{H}$, and one that is orthogonal to the subspace spanned by $\boldsymbol{H}$. In order not to have to deal with degenerate Gaussian vectors, we define an $n \times p$ matrix $\boldsymbol{P}$ with full rank which spans the same subspace as $\boldsymbol{H}$ (Actually, for the time being, we may as well set $\boldsymbol{P} = \boldsymbol{H}$.) and an $n \times (n-p)$ matrix $\boldsymbol{W}$ with full rank which spans its orthogonal space. Thus $\boldsymbol{W}^\top \boldsymbol{H} = \boldsymbol{W}^\top \boldsymbol{P} = \boldsymbol{0}_{n-p,p}$ and

$$\boldsymbol{W}^\top \boldsymbol{y} | \sigma^2, \boldsymbol{\theta} \sim \mathcal{N}(\boldsymbol{0}_{n-p}, \sigma^2 \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}) \; ; \tag{6.3}$$

$$\begin{aligned} \boldsymbol{P}^\top \boldsymbol{y} | \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \boldsymbol{W}^\top \boldsymbol{y} \sim \mathcal{N}(&\boldsymbol{P}^\top \boldsymbol{H} \boldsymbol{\beta} + \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}, \\ &\sigma^2 \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P} - \sigma^2 \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P}). \end{aligned} \tag{6.4}$$

$\boldsymbol{\beta}$ having flat prior density, $\boldsymbol{P}^\top \boldsymbol{y} - \boldsymbol{P}^\top \boldsymbol{H} \boldsymbol{\beta}$ has the same distribution whether $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ or whether $\boldsymbol{P}^\top \boldsymbol{y}$, σ^2 and $\boldsymbol{\theta}$ are known. Therefore, the posterior distribution of $\boldsymbol{P}^\top \boldsymbol{H} \boldsymbol{\beta}$ if σ^2 and $\boldsymbol{\theta}$ are known is:

$$\begin{aligned} \boldsymbol{P}^\top \boldsymbol{H} \boldsymbol{\beta} | \sigma^2, \boldsymbol{\theta}, \boldsymbol{W}^\top \boldsymbol{y}, \boldsymbol{P}^\top \boldsymbol{y} \sim \mathcal{N}(&\boldsymbol{P}^\top \boldsymbol{y} - \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}, \\ &\sigma^2 \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P} - \sigma^2 \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P}). \end{aligned} \tag{6.5}$$

From there, we get the posterior distribution of $\boldsymbol{\beta}$ if σ^2 and $\boldsymbol{\theta}$ are known:

$$\begin{aligned} \boldsymbol{\beta} | \sigma^2, \boldsymbol{\theta}, \boldsymbol{y} \sim &\mathcal{N}((\boldsymbol{P}^\top \boldsymbol{H})^{-1} \boldsymbol{P}^\top \boldsymbol{y} - (\boldsymbol{P}^\top \boldsymbol{H})^{-1} \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}, \\ &\sigma^2 \Big((\boldsymbol{P}^\top \boldsymbol{H})^{-1} \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P} (\boldsymbol{H}^\top \boldsymbol{P})^{-1} \\ &\quad - (\boldsymbol{P}^\top \boldsymbol{H})^{-1} \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P} (\boldsymbol{H}^\top \boldsymbol{P})^{-1}) \Big) \end{aligned} \tag{6.6}$$

Moreover, (6.4) implies that the integrated likelihood of $\boldsymbol{P}^\top \boldsymbol{y}$, i.e. its likelihood averaged over the Lebesgue measure (the prior distribution on $\boldsymbol{\beta}$), is $|\boldsymbol{P}^\top \boldsymbol{H}|^{-1}$, where $|\cdot|$ denotes the absolute value of the determinant.

$$\boldsymbol{P}^\top \boldsymbol{y} | \sigma^2, \boldsymbol{\theta}, \boldsymbol{W}^\top \boldsymbol{y} \sim \text{Improper “uniform” distribution on } \mathbb{R}^p. \tag{6.7}$$

This means that if $\boldsymbol{\beta}$ is unknown, then $\boldsymbol{P}^\top \boldsymbol{y}$ can yield no information about σ^2 and $\boldsymbol{\theta}$. When $\boldsymbol{\beta}$ is unknown, all information about σ^2 and $\boldsymbol{\theta}$ is carried by $\boldsymbol{W}^\top \boldsymbol{y}$, because as is shown by (6.3), the predictive distribution on $\boldsymbol{W}^\top \boldsymbol{y}$ knowing σ^2 and $\boldsymbol{\theta}$ does not depend on $\boldsymbol{\beta}$.

A straightforward calculation yields that the posterior distribution of σ^2 is Inverse-Gamma:

$$\sigma^2 | \boldsymbol{\theta}, \boldsymbol{y} \sim \mathcal{IG}(\text{shape} = (n-p)/2, \text{rate} = \boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W} \right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}/2). \tag{6.8}$$

The posterior distribution on σ^2 (knowing $\boldsymbol{\theta}$) does not take into account $\boldsymbol{P}^\top \boldsymbol{y}$, because all information contained in $\boldsymbol{P}^\top \boldsymbol{y}$ is given in the posterior distribution of $\boldsymbol{\beta}$ conditional to σ^2 and $\boldsymbol{\theta}$.

We conclude this subsection with the formulas for the likelihoods with the parameters $\boldsymbol{\beta}$ and σ^2 successively integrated out.

$$L^0(\boldsymbol{y}|\sigma^2,\boldsymbol{\theta}) = \int L(\boldsymbol{y}|\boldsymbol{\beta},\sigma^2,\boldsymbol{\theta})d\boldsymbol{\beta}$$
$$= \left|\boldsymbol{P}^\top\boldsymbol{H}\right|^{-1}\left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n-p}{2}}|\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}|^{-\frac{1}{2}}\exp\left\{-\frac{\boldsymbol{y}^\top\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}}{2\sigma^2}\right\}; \tag{6.9}$$

$$L^1(\boldsymbol{y}|\boldsymbol{\theta}) = \int L^0(\boldsymbol{y}|\sigma^2,\boldsymbol{\theta})/\sigma^2 d\sigma^2$$
$$= \left|\boldsymbol{P}^\top\boldsymbol{H}\right|^{-1}\left(\frac{2\pi^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-p}{2}\right)}\right)^{-1}|\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}|^{-\frac{1}{2}}\left(\boldsymbol{y}^\top\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}\right)^{-\frac{n-p}{2}}. \tag{6.10}$$

Posterior predictive distribution when $\boldsymbol{\theta}$ is known.

Following Santner et al. [2003] (Theorem 4.1.2., case (4)), we derive conditionally to $\boldsymbol{\theta}$ the posterior predictive distribution of the values taken by the process at unobserved points.
In order to simplify notations in this subsection, all the distributions we consider are, until further notice, conditional to σ^2 and $\boldsymbol{\theta}$ even with no explicit mention. Equation (6.1) can be usefully restated in the following way:

$$\begin{pmatrix}\boldsymbol{P}^\top\boldsymbol{y}-\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}\\ \boldsymbol{W}^\top\boldsymbol{y}\end{pmatrix}\Bigg|\,\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}\sim\mathcal{N}\left(\boldsymbol{0}_n,\sigma^2\begin{pmatrix}\boldsymbol{P}^\top\\ \boldsymbol{W}^\top\end{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\begin{pmatrix}\boldsymbol{P} & \boldsymbol{W}\end{pmatrix}\right). \tag{6.11}$$

Because the prior distribution on $\boldsymbol{\beta}$ is flat, $\begin{pmatrix}\boldsymbol{P}^\top\boldsymbol{y}-\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}\\ \boldsymbol{W}^\top\boldsymbol{y}\end{pmatrix}$ and its opposite have the same distribution when conditional respectively to $\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}$ and $\boldsymbol{P}^\top\boldsymbol{y}$.

$$\begin{pmatrix}\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{P}^\top\boldsymbol{y}\\ -\boldsymbol{W}^\top\boldsymbol{y}\end{pmatrix}\Bigg|\,\boldsymbol{P}^\top\boldsymbol{y}\sim\mathcal{N}\left(\boldsymbol{0}_n,\sigma^2\begin{pmatrix}\boldsymbol{P}^\top\\ \boldsymbol{W}^\top\end{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\begin{pmatrix}\boldsymbol{P} & \boldsymbol{W}\end{pmatrix}\right) \tag{6.12}$$

Let $\boldsymbol{y}_0$ be the values of the Gaussian Process at the n_0 unobserved points. We denote by $\boldsymbol{H}_{0,0}$ the $n_0\times p$ matrix whose columns contain the values of the p basis functions at the unobserved points, by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,0}$ the $n_0\times n_0$ correlation matrix of $\boldsymbol{y}_0$, by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}$ the $n_0\times n$ correlation matrix between $\boldsymbol{y}_0$ and $\boldsymbol{y}$ and by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0}$ its transpose. It is also convenient to define the $n_0\times n$ matrix $\boldsymbol{H}_{0,\cdot}=\boldsymbol{H}_{0,0}\left(\boldsymbol{P}^\top\boldsymbol{H}\right)^{-1}$ and its transpose $\boldsymbol{H}_{\cdot,0}$. With these notations, the distribution of $\boldsymbol{y}_0$ when $\boldsymbol{y}$ and $\boldsymbol{\beta}$ are known is

$$\boldsymbol{y}_0|\boldsymbol{\beta},\boldsymbol{y}\sim\mathcal{N}\left(\boldsymbol{H}_{0,0}\boldsymbol{\beta}+\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y}-\boldsymbol{H}\boldsymbol{\beta}),\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,0}-\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0}\right) \tag{6.13}$$

Now, the distribution of $\boldsymbol{y}_0$ when both $\boldsymbol{y}$ and $\boldsymbol{\beta}$ are known, together with the distribution of $\begin{pmatrix}\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{P}^\top\boldsymbol{y}\\ -\boldsymbol{W}^\top\boldsymbol{y}\end{pmatrix}$ when $\boldsymbol{P}^\top\boldsymbol{y}$ is known, jointly define some probability distribution on the vector $\begin{pmatrix}\boldsymbol{y}_0\\ \boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}-\boldsymbol{P}^\top\boldsymbol{y}\\ -\boldsymbol{W}^\top\boldsymbol{y}\end{pmatrix}$.
This distribution is given in the following proposition. In order to give it a concise expression, it is convenient to require that $\boldsymbol{P}\boldsymbol{P}^\top+\boldsymbol{W}\boldsymbol{W}^\top=\boldsymbol{I}_n$, which simply means that the columns of $\boldsymbol{P}$ and $\boldsymbol{W}$ form an orthonormal basis of $<\boldsymbol{H}>$ and its orthogonal space respectively.

Proposition 6.1. *Assume that* $\boldsymbol{P}\boldsymbol{P}^\top + \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}_n$. *Then the probability distribution on the vector of* $\mathbb{R}^{n_0+n}$ $(\boldsymbol{y}_0, \boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{P}^\top\boldsymbol{y}, -\boldsymbol{W}^\top\boldsymbol{y})^\top$ *conditional to* $\boldsymbol{P}^\top\boldsymbol{y}$ *is the following multivariate normal distribution:*

$$\mathcal{N}\left(\begin{pmatrix}\boldsymbol{E}_0\\ \boldsymbol{0}_n\end{pmatrix}, \sigma^2\begin{pmatrix}\boldsymbol{S}_{\boldsymbol{\theta},0,0} & \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}\\ \boldsymbol{S}_{\boldsymbol{\theta},\cdot,0} & \begin{pmatrix}\boldsymbol{P}^\top\\ \boldsymbol{W}^\top\end{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\begin{pmatrix}\boldsymbol{P} & \boldsymbol{W}\end{pmatrix}\end{pmatrix}\right). \tag{6.14}$$

We use the following notations:

$$\begin{aligned}
\boldsymbol{E}_0 &:= \boldsymbol{H}_{0,\cdot}\boldsymbol{P}^\top\boldsymbol{y}\\
\boldsymbol{S}_{\boldsymbol{\theta},0,0} &:= \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,0} + \boldsymbol{H}_{0,\cdot}\boldsymbol{P}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{P}\boldsymbol{H}_{\cdot,0} - \boldsymbol{H}_{0,\cdot}\boldsymbol{P}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{P}\boldsymbol{H}_{\cdot,0}\\
\boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}\begin{pmatrix}\boldsymbol{P}^\top\\ \boldsymbol{W}^\top\end{pmatrix} &:= \boldsymbol{H}_{0,\cdot}\boldsymbol{P}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\\
\boldsymbol{S}_{\boldsymbol{\theta},\cdot,0} &:= \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}^\top
\end{aligned}$$

Proof. First, notice that the mean vector of the Normal distribution given by Equation 6.13 can be rewritten as

$$(\boldsymbol{H}_{0,\cdot} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{P})\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{W}\boldsymbol{W}^\top\boldsymbol{y} + \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{P}\boldsymbol{P}^\top\boldsymbol{y}, \tag{6.15}$$

which is a linear mapping of the vector $(\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}, \boldsymbol{W}^\top\boldsymbol{y}, \boldsymbol{P}^\top\boldsymbol{y})^\top$. Now, Equation 6.12 tells us that conditional to $\boldsymbol{P}^\top\boldsymbol{y}$, $(\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}, \boldsymbol{W}^\top\boldsymbol{y}, \boldsymbol{P}^\top\boldsymbol{y})^\top$ is a (degenerate) Gaussian vector, so Gaussian theory implies that conditional to $\boldsymbol{P}^\top\boldsymbol{y}$, $(\boldsymbol{y}_0, \boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}, \boldsymbol{W}^\top\boldsymbol{y}, \boldsymbol{P}^\top\boldsymbol{y})^\top$ is a Gaussian vector and therefore $(\boldsymbol{y}_0, \boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{P}^\top\boldsymbol{y}, -\boldsymbol{W}^\top\boldsymbol{y})^\top$ is one as well. So all that remains to be shown is that its mean and covariance are those given by Proposition 6.1.
To do this, we compute $\boldsymbol{E}_{\boldsymbol{\theta}}^{\boldsymbol{y},\boldsymbol{\beta}}$ and $\sigma^2\boldsymbol{S}_{\boldsymbol{\theta}}^{\boldsymbol{y},\boldsymbol{\beta}}$, the conditional mean and variance of $\boldsymbol{y}_0$ given $\boldsymbol{W}^\top\boldsymbol{y}$, $\boldsymbol{P}^\top\boldsymbol{y}$ and $\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta}$ and check that they fit the parameters of (6.13).

$$\begin{aligned}
\boldsymbol{E}_{\boldsymbol{\theta}}^{\boldsymbol{y},\boldsymbol{\beta}} &= \boldsymbol{E}_0 + \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}\begin{pmatrix}\boldsymbol{P}^\top\\ \boldsymbol{W}^\top\end{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\begin{pmatrix}\boldsymbol{P} & \boldsymbol{W}\end{pmatrix}\begin{pmatrix}\boldsymbol{P}^\top\boldsymbol{H}\boldsymbol{\beta} - \boldsymbol{P}^\top\boldsymbol{y}\\ -\boldsymbol{W}^\top\boldsymbol{y}\end{pmatrix}\\
&= \boldsymbol{H}_{0,0}\boldsymbol{\beta} + \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{\beta});
\end{aligned} \tag{6.16}$$

$$\begin{aligned}
\boldsymbol{S}_{\boldsymbol{\theta}}^{\boldsymbol{y},\boldsymbol{\beta}} &= \boldsymbol{S}_{\boldsymbol{\theta},0,0} - \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}\begin{pmatrix}\boldsymbol{P}^\top\\ \boldsymbol{W}^\top\end{pmatrix}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\begin{pmatrix}\boldsymbol{P} & \boldsymbol{W}\end{pmatrix}\boldsymbol{S}_{\boldsymbol{\theta},\cdot,0}\\
&= \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,0} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0}.
\end{aligned} \tag{6.17}$$

□

From this point onwards, distributions are no longer implicitly conditional to σ^2 and $\boldsymbol{\theta}$.

Corollary 6.2. *Assume that* $\boldsymbol{P}\boldsymbol{P}^\top + \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}_n$. *The predictive distribution when* $\boldsymbol{\beta}$ *is unknown – i.e. the distribution of* $\boldsymbol{y}_0$ *conditional to* $\boldsymbol{y}$, σ^2 *and* $\boldsymbol{\theta}$ *– is Normal. With the notations of Proposition 6.1, it has mean vector* $\boldsymbol{E}_0 - \boldsymbol{S}_{\boldsymbol{\theta},\cdot,0}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}$ *and covariance matrix*

$$\sigma^2\left\{\boldsymbol{S}_{\boldsymbol{\theta},0,0} - \boldsymbol{S}_{\boldsymbol{\theta},\cdot,0}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}\right\}.$$

Corollary 6.3. *Assume that* $\boldsymbol{P}\boldsymbol{P}^\top + \boldsymbol{W}\boldsymbol{W}^\top = \boldsymbol{I}_n$. *The predictive distribution when both* $\boldsymbol{\beta}$ *and* σ^2 *are unknown – i.e. the distribution of* $\boldsymbol{y}_0$ *conditional to* $\boldsymbol{y}$ *and* $\boldsymbol{\theta}$ *– is multivariate Student with* $n-p$ *degrees of freedom. With the notations of Proposition 6.1, it has location vector* $\boldsymbol{E}_0 - \boldsymbol{S}_{\boldsymbol{\theta},\cdot,0}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}$ *and scale matrix*

$$\frac{\boldsymbol{y}^\top\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{y}}{n-p}\left\{\boldsymbol{S}_{\boldsymbol{\theta},0,0} - \boldsymbol{S}_{\boldsymbol{\theta},\cdot,0}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\boldsymbol{W}^\top\boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}\right\}.$$

6.3 Reference prior on a one-dimensional $\boldsymbol{\theta}$

In this section, $\boldsymbol{\theta}$ is assumed to be a scalar parameter, which we emphasize by writing it θ. This is the easy case: Propositions 6.4 and 6.5 recall Proposition 3.2.

Proposition 6.4. *The reference prior on* θ *is* $\pi(\theta) \propto$

$$\sqrt{\mathrm{Tr}\left[\left\{\boldsymbol{W}^\top\frac{\partial}{\partial\theta}(\boldsymbol{\Sigma}_\theta)\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{W}\right)^{-1}\right\}^2\right] - \frac{1}{n-p}\mathrm{Tr}\left\{\boldsymbol{W}^\top\frac{\partial}{\partial\theta}(\boldsymbol{\Sigma}_\theta)\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_\theta\boldsymbol{W}\right)^{-1}\right\}^2}. \tag{6.18}$$

This result is in keeping with the previous work of [Berger et al., 2001]:

Proposition 6.5. *The reference prior on* θ *can also be written as:*

$$\pi(\theta) \propto \sqrt{\mathrm{Tr}\left[\left\{\frac{\partial}{\partial\theta}(\boldsymbol{\Sigma}_\theta)\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{Q}_\theta\right\}^2\right] - \frac{1}{n-p}\left[\mathrm{Tr}\left\{\frac{\partial}{\partial\theta}(\boldsymbol{\Sigma}_\theta)\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{Q}_\theta\right\}\right]^2}, \tag{6.19}$$

where $\boldsymbol{Q}_\theta := \boldsymbol{I}_n - \boldsymbol{H}\left(\boldsymbol{H}^\top\boldsymbol{\Sigma}_\theta^{-1}\boldsymbol{H}\right)^{-1}\boldsymbol{H}^\top\boldsymbol{\Sigma}_\theta^{-1}$.

6.4 The Gibbs reference posterior on a multi-dimensional $\boldsymbol{\theta}$

Definition

In the case of a multidimensional $\boldsymbol{\theta}$, reference prior theory gives a choice between 1) considering $\boldsymbol{\theta}$ as a single parameter or 2) defining an ordering on the scalar parameters $\theta_1, ..., \theta_r$. Both possibilities are unsatisfactory, albeit in different ways. Concerning 1), Jeffreys' prior is unsuited to dealing with multidimensional parameters [Robert et al., 2009] and besides, the posterior may be improper. Concerning 2), further integration of the likelihood (6.10) would be analytically intractable, even if it were possible to define a non-arbitrary ordering of the coordinates of $\boldsymbol{\theta}$.

We propose a quasi-posterior distribution based on the reference posterior of models where only one coordinate of $\boldsymbol{\theta}$ is unknown. For any integer $i \in [\![1, r]\!]$, we collectively denote by $\boldsymbol{\theta}_{-i}$ all coordinates of $\boldsymbol{\theta}$ except the i-th: $\boldsymbol{\theta}_{-i} = (\theta_j)_{j\in[\![1,r]\!]\setminus\{i\}}$.
Consider now $\pi_i(\theta_i|\boldsymbol{\theta}_{-i})$, the reference prior distribution on θ_i conditional to $\boldsymbol{\theta}_{-i}$ and the associated reference posterior distribution $\pi_i(\theta_i|\boldsymbol{y}, \boldsymbol{\theta}_{-i}) \propto L^1(\boldsymbol{y}|\boldsymbol{\theta})\pi_i(\theta_i|\boldsymbol{\theta}_{-i})$,.
The conditional reference prior $\pi_i(\theta_i|\boldsymbol{\theta}_{-i})$ is given by $\pi_i(\theta_i|\boldsymbol{\theta}_{-i}) \propto$

$$\sqrt{\mathrm{Tr}\left[\left\{\boldsymbol{W}^\top\partial_{\theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\right\}^2\right] - \frac{1}{n-p}\mathrm{Tr}\left\{\boldsymbol{W}^\top\partial_{\theta_i}\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\theta}}\boldsymbol{W}\right)^{-1}\right\}^2}. \tag{6.20}$$

Now consider the sequence of conditional posterior distributions $(\pi_i(\theta_i|\boldsymbol{y},\boldsymbol{\theta}_{-i}))_{i\in[\![1,r]\!]}$. These conditional distributions are incompatible in the sense that there exists no joint probability distribution $\pi(\boldsymbol{\theta}|\boldsymbol{y})$ which agrees with all of them. We may however define the Gibbs reference posterior as a compromise between the conditionals in this sequence. In Chapter 4 we provided theoretical foundation for what such a compromise could be. In the end, we showed it to be the stationary probability distribution of a Markovian kernel $P_{\boldsymbol{y}} : (0,+\infty)^r \times \mathcal{B}((0,+\infty)^r)$, where $\mathcal{B}((0,+\infty)^r)$ denotes the Borel algebra on $((0,+\infty)^r)$. $P_{\boldsymbol{y}}$ is defined by the following expression, where $\boldsymbol{\theta}^{(0)} \in (0,1)^r$ and δ_t denotes the shifted Dirac measure $\delta(\cdot - t)$:

$$P_{\boldsymbol{y}}(\boldsymbol{\theta}^{(0)}, d\boldsymbol{\theta}) = \frac{1}{r}\sum_{i=1}^{r} \pi_i(\theta_i|\boldsymbol{y},\boldsymbol{\theta}_{-i}^{(0)}) d\theta_i \; \delta_{\boldsymbol{\theta}_{-i}^{(0)}}(d\boldsymbol{\theta}_{-i}). \tag{6.21}$$

The goal of this section is to provide sufficient conditions for the existence (and thus, propriety) of this stationary probability distribution $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$ and to show that the Markov Chain Monte-Carlo (MCMC) algorithm based on the Markovian kernel $P_{\boldsymbol{y}}$ converges to it, that is, $P_{\boldsymbol{y}}$ is uniformly ergodic. This means that denoting by $P_{\boldsymbol{y}}^n$ the Markov kernel produced by n successive applications of $P_{\boldsymbol{y}}$ and by $\|\cdot\|_{TV}$ the total variation norm,

$$\lim_{n\to\infty} \sup_{\boldsymbol{\theta}^{(0)}\in(0,+\infty)^r} \|P_{\boldsymbol{y}}^n(\boldsymbol{\theta}^{(0)},\cdot) - \pi_G(\cdot|\boldsymbol{y})\|_{TV} = 0. \tag{6.22}$$

In the following results, when we write that "$P_{\boldsymbol{y}}$ is uniformly ergodic", we mean that Equation (6.22) holds.

Existence

The results in this subsection deal with the following setting:

- The spatial domain is the unit cube $(0,1)^r$ $(r > 0)$.
- The mean function space $\mathcal{F}_p$ has dimension $p \geqslant 0$.
- The Universal Kriging model uses a Matérn anisotropic geometric or tensorized correlation kernel with smoothness parameter $\nu > 0$.
- Design sets contain $n > 0$ points, so we identify $(0,1)^{rn}$ with the set of all design sets in the spatial domain $(0,1)^r$. Let $Q(r,n)$ be the Lebesgue measure on $(0,1)^{rn}$.

In the following, we change parametrization for the sake of convenience: define $\boldsymbol{\mu}$ such that $\forall i \in [\![1,r]\!]$, $\mu_i = 1/\theta_i$. The conditionals are invariant to such a change, and therefore both the Markovian kernel $P_{\boldsymbol{y}}$ and, if it exists, its stationary probability remain the same. Abusing notations, the likelihood $L^1(\boldsymbol{y}|\boldsymbol{\theta})$ is denoted by $L(\boldsymbol{y}|\boldsymbol{\mu})$ when expressed in the $\boldsymbol{\mu}$-parametrization. Define the functions f_i by $f_i(\mu_i|\boldsymbol{\mu}_{-i}) :=$

$$\sqrt{\operatorname{Tr}\left[\left(\boldsymbol{W}^\top \frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{W}\right)^{-1}\right)^2\right] - \frac{1}{n-p}\operatorname{Tr}\left[\boldsymbol{W}^\top \frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{W}\right)^{-1}\right]^2}. \tag{6.23}$$

Then, following Equation (6.20), the conditional density π_i is in the $\boldsymbol{\mu}$-parametrization given by:

$$\pi_i(\mu_i|\boldsymbol{\mu}_{-i}) \propto f_i(\mu_i|\boldsymbol{\mu}_{-i}). \tag{6.24}$$

We need to make some assumptions which are detailed below.

Assumption 3. *Any vector in the subspace of $\mathbb{R}^n$ spanned by $\boldsymbol{H}$ is either null or has strictly more than $2r$ non-null elements when expressed in the canonical base.*

Remark. It is not apparent, but the purpose of Assumption 3 is to control the behavior of the $f_i(\mu_i|\boldsymbol{\mu}_{-i})$ ($i \in [\![1, r]\!]$) when $\|\boldsymbol{\mu}\| \to \infty$. See the proofs in Appendix 6.B for details.

This assumption is not very restrictive, as the two following results show.

Proposition 6.6. *In Ordinary Kriging – that is with $p = 1$ and $\mathcal{F}_p$ being the space of constant functions – if $n > 2r$, Assumption 3 is automatically verified.*

Proof. In this setting, $\boldsymbol{H}$ is a non-null constant $n \times 1$ matrix, so Assumption 3 is trivially verified. □

Proposition 6.7. *Assume that the design set is such that any subset with cardinal $r+1$ forms a simplex. Then in Universal Kriging, if the mean function space $\mathcal{F}_p$ is included within the vector space of polynomials of degree 0 and 1, and if $n > 3r$, Assumption 3 is automatically verified.*

Proof. Let $\boldsymbol{y}^*$ belong to the subspace of $\mathbb{R}^n$ spanned by $\boldsymbol{H}$. Assume that it has $2r$ or fewer non-null elements when expressed in the canonical base. Conversely, it has at least $n-2r$ null elements. If $n > 3r$, then this means that there exists a function $f^* \in \mathcal{F}_p$ (the one represented by $\boldsymbol{y}^*$) which admits at least $r + 1$ zeros on the design set. However, given the premise of Proposition 6.7, these $r + 1$ points form a simplex, so they span an affine space of dimension r. As f^* is a polynomial with r unknowns of degree 0 or 1, this implies that $f^* = 0$. □

Remark. $Q(r, n)$-almost all design sets fit the premise of Proposition 6.7.

In some cases, Assumption 3 is sufficient for our purposes. Define $\mathbf{1}$ as the vector of $\mathbb{R}^n$ with all components in the canonical basis equal to 1.

Proposition 6.8. *In the setting described above, if $0 < \nu < 1$ and $n > p + 1$, then for $Q(r, n)$-almost all design sets, if $\mathbf{1}$ does not belong to the vector space spanned by $\boldsymbol{H}$, then Assumption 3 implies that there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that $\forall \boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, $P_{\boldsymbol{y}}$ is uniformly ergodic.*

The proof of this Proposition can be found in Appendix 6.B.
Naturally, the above result is somewhat unsatisfactory since most users will want to include non-null constant functions in $\mathcal{F}_p$.
Consider now the following assumption.

Assumption 4. *There exists $\epsilon_{\boldsymbol{y}} > 0$ such that $L(\boldsymbol{y}|\boldsymbol{\mu}) = O(\|\boldsymbol{\mu}\|^{\epsilon_{\boldsymbol{y}}})$ when $\|\boldsymbol{\mu}\| \to 0$.*

Remark. Assumption 4 essentially means that the model should find perfect correlation unlikely.

The following theorem, which is proved in Appendix 6.B, is our essential tool for dealing with the case where non-null constant functions are included in $\mathcal{F}_p$.

Theorem 6.9. *In the setting described above, if $\nu > 1$ and $n > p + 2r + 2$, then for $Q(r, n)$-almost all design sets, Assumptions 3 and 4 imply that $P_{\boldsymbol{y}}$ is uniformly ergodic.*

The next two results, which are proved in Appendix 6.B, concern particular settings where Assumptions 3 and 4 are both verified and therefore Theorem 6.9 yields the uniform ergodicity of $P_{\boldsymbol{y}}$.

Proposition 6.10. *Consider the particular case of the above described setting where $p = 1$ and $\mathcal{F}_p$ is the space of all constant functions (Ordinary Kriging), and assume that one of the following conditions is satisfied:*

1. *$1 < \nu < 2$ and $n > 2r + 3$;*
2. *$2 < \nu < 3$ and $n > (r + 1)(r/2 + 2)$.*

Then, for $Q(r, n)$-almost all design sets, there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that $\forall \boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, $P_{\boldsymbol{y}}$ is uniformly ergodic.

Proposition 6.11. *Consider the particular case of the above described setting where $\mathcal{F}_p$ is included within the space of all polynomials of degree 0 and 1 (so $p \leqslant r + 1$) and assume that the following condition is satisfied:*

— *$2 < \nu < 3$ and $n > r(r + 1)/2 + 2r + 3$.*

Then, for $Q(r, n)$-almost all design sets, there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that $\forall \boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, $P_{\boldsymbol{y}}$ is uniformly ergodic.

Remark. In Propositions 6.8, 6.10 and 6.11, the condition that the observation $\boldsymbol{y}$ should not belong to a given negligible (for the Lebesgue measure) subset of $\mathbb{R}^n$ is fairly natural: for the Kriging model to be adequate, $\boldsymbol{y}$ must not look like a realization of a degenerate Gaussian vector. Theorem 6.9 does not really dispense with it, as it is implied by Assumption 4.

To sum up the results of this section, to ensure that the Gibbs reference posterior exists and can be accessed through Gibbs sampling, one should check that one of the following assertions is true:

— $\nu > 1$, $n > r + p + 2$ and both Assumptions 3 and 4 are verified;
— $\mathcal{F}_p$ contains only constant functions and $1 < \nu < 2$ and $n > r + 3$;
— $\mathcal{F}_p$ contains only polynomials of degree 0 and 1, $2 < \nu < 3$ and $n > r(r+1)/2+2r+3$;
— $0 < \nu < 1$, $n > p + 1$, no non-null constant function belongs to $\mathcal{F}_p$ and Assumption 3 is verified.

6.5 Comparison of the predictive performance of the full-Bayesian approach versus MLE and MAP plug-in approaches

In this section, we evaluate the predictive performance resulting from the Gibbs reference posterior distribution $\pi_G(\boldsymbol{\theta}|\boldsymbol{y})$ in the context of a well-specified model, and then when emulating some deterministic real functions. We contrast the full-Bayesian approach, in which the Full Gibbs reference Posterior Distribution (FPD) is used, with two plug-in approaches: one where the Maximum Likelihood Estimator (MLE) and the other where the Maximum A Posteriori (MAP) estimator is assumed to be the true value of $\boldsymbol{\theta}$. All approaches make use of the reference posterior $\pi(\boldsymbol{\beta}, \sigma^2|\boldsymbol{y}, \boldsymbol{\theta})$.

We use the following terminology. We call Simple Kriging the Kriging model where the mean function is assumed to be known, whether this assumption is correct or known. We call Ordinary Kriging any Universal Kriging model where the mean function space is the space

of constant functions. We call Affine Kriging any Universal Kriging model where the mean function space is the space of affine functions.

Well-specified model

We first consider well-specified models, specifically Kriging models with unknown parameters $(\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})$ emulating actual Gaussian processes with variance $\sigma^2 = 1$ and Matérn anisotropic geometric autocorrelation kernel with smoothness $\nu = 5/2$. Moreover, the true mean function of the Gaussian process belongs to the assumed mean function space $\mathcal{F}_p$.

The spatial domain is the unit cube $(0,1)^r$ and the considered design sets all contain n points independently chosen according to the Lebesgue measure on the domain $(0,1)^r$.

The following tables give the average coverage and average mean length of prediction intervals. To define these notions, we introduce the following notations:

- $\boldsymbol{Y}$ is the Gaussian process, and $\boldsymbol{Y}(\boldsymbol{x})$ is the vector of the values taken by said process at the points in the design set $\boldsymbol{x}$;
- T is a random variable which follows the Uniform distribution on the unit cube $(0,1)^r$. It represents the "test" point;
- $\boldsymbol{X}$ is the random design set following the Uniform distribution on $((0,1)^r)^n$.
- $\boldsymbol{Y}$, T and $\boldsymbol{X}$ are mutually independent;
- f is a function defined on $((0,1)^r)^n \times \mathbb{R}^n \times (0,1)^r$ which associates to $(\boldsymbol{x}, \boldsymbol{y}, t)$ the prediction interval at t of the Gaussian process, based on the knowledge of its value $\boldsymbol{y}$ on the design set $\boldsymbol{x}$.

Definition 6.12. *The average coverage is the probability (with respect to the distributions of $\boldsymbol{X}$, $\boldsymbol{Y}$ and T) that $\boldsymbol{Y}(T) \in f(\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X}), T)$.*

Definition 6.13. *The average mean length is the expectation (with respect to the distributions of $\boldsymbol{X}$, $\boldsymbol{Y}$ and T) of the length of $f(\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X}), T)$.*

The average coverage $\mathbb{P}[\boldsymbol{Y}(T) \in f(\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X}), T)]$ is numerically computed as

$$\mathbb{P}[\boldsymbol{Y}(T) \in f(\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X}), T)] = \mathbb{E}[\mathbb{P}[\boldsymbol{Y}(T) \in f(\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X}), T) | \boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X})]]$$

over 500 random design sets and for each design set 1000 random test points. The average mean length is computed in a similar fashion.

In this subsection and the following one, we take $n = 30$ and $r = 3$.

In the first set of simulations, we use a well-specified Ordinary Kriging model, with the unknown mean 5. As $r = 3$ and $n = 30$, Proposition 6.10 is applicable.

The results given in Table 6.1 show that using the full posterior distribution (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.

The results given in Table 6.2 show that Predictive Intervals arising from the full Gibbs reference posterior distribution (FPD) are on average somewhat larger than those resulting

Average Coverage				
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	0.88	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	0.94
0.8 – 1.0 – 0.9	0.95	0.90	0.92	0.94

Table 6.1 – *For a Gaussian Process with constant mean function equal to 5, variance parameter 1 and smoothness parameter 5/2, average coverage of 95% Prediction Intervals produced by an Ordinary Kriging model. "True" stands for the Simple Kriging prediction based on the knowledge of the true mean parameter, variance parameter and vector of correlation lengths.*

from knowledge of the true parameters, while intervals arising from both types of parameter estimation (MLE and MAP) are too short.

Average Mean Length				
Corr. lengths	**True**	**MLE**	**MAP**	**FPD**
0.4 – 0.8 – 0.2	2.23	2.06	2.14	2.58
0.5 – 0.5 – 0.5	1.69	1.55	1.59	1.83
0.7 – 1.3 – 0.4	1.09	1.02	1.07	1.20
0.8 – 0.3 – 0.6	1.63	1.51	1.57	1.81
0.8 – 1.0 – 0.9	0.71	0.66	0.69	0.76

Table 6.2 – *For a Gaussian Process with constant mean function equal to 5, variance parameter 1 and smoothness parameter 5/2, average mean length of 95% Prediction Intervals produced by an Ordinary Kriging model. "True" stands for the Simple Kriging prediction based on the knowledge of the true mean parameter, variance parameter and vector of correlation lengths.*

Consider now Universal Kriging models where the true mean function is the polynomial $(x_1, x_2, x_3) \mapsto 5 + 4x_1 + 3x_2 + 2x_3$, and the model (correctly) assumes that it belongs to the 4-dimensional space ($p = 4$) spanned by the functions mapping (x_1, x_2, x_3) to 1, x_1, x_2 and x_3 respectively. For such Affine Kriging models, Proposition 6.11 is applicable.

As shown in Table 6.3, Predictive Intervals resulting from both plug-in approaches (MLE, MAP) and from the full posterior distribution perform a little worse than in the Ordinary Kriging setting, but their relative performances stay the same.

Table 6.4 shows that the average mean lengths of Predictive Intervals are not very different in Affine Kriging than in Ordinary Kriging when it comes to the FPD. However, they are

Average Coverage				
Corr. lengths	**True**	**MLE**	**MAP**	**FPD**
0.4 – 0.8 – 0.2	0.95	0.87	0.90	0.94
0.5 – 0.5 – 0.5	0.95	0.87	0.89	0.92
0.7 – 1.3 – 0.4	0.95	0.89	0.92	0.94
0.8 – 0.3 – 0.6	0.95	0.87	0.90	0.93
0.8 – 1.0 – 0.9	0.95	0.89	0.92	0.93

Table 6.3 – *For a Gaussian Process with mean function $(x_1, x_2, x_3) \mapsto 5 + 4x_1 + 3x_2 + 2x_3$, variance parameter 1 and smoothness parameter 5/2, average coverage of 95% Prediction Intervals produced by an Affine Kriging model. "True" stands for the Simple Kriging prediction based on the knowledge of the true mean function, variance parameter and vector of correlation lengths.*

larger in Affine Kriging than in Ordinary Kriging when it comes to the MLE and the MAP. Interestingly, Predictive Intervals resulting from the MAP have about the same size as Predictive Intervals derived when all parameters are known. Those derived using the MLE are shorter, and those derived from the FPD are larger.

Average Mean Length				
Corr. lengths	**True**	**MLE**	**MAP**	**FPD**
0.4 – 0.8 – 0.2	2.23	2.14	2.23	2.59
0.5 – 0.5 – 0.5	1.69	1.57	1.66	1.83
0.7 – 1.3 – 0.4	1.09	1.04	1.10	1.20
0.8 – 0.3 – 0.6	1.63	1.54	1.61	1.80
0.8 – 1.0 – 0.9	0.71	0.67	0.71	0.75

Table 6.4 – *For a Gaussian Process with mean function* $(x_1, x_2, x_3) \mapsto 5 + 4x_1 + 3x_2 + 2x_3$*, variance parameter 1 and smoothness parameter 5/2, average mean length of 95% Prediction Intervals produced by an Affine Kriging model. "True" stands for the Simple Kriging prediction based on the knowledge of the true mean function, variance parameter and vector of correlation lengths.*

For reference, we give the tables obtained in the Simple Kriging case, that is the case where the Gaussian Process is known to have null mean function. Table 6.5 gives the average coverages and Table 6.6 the average mean lengths.

Average Coverage				
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 6.5 – *For a Gaussian Process with null mean function, variance parameter 1 and smoothness parameter 5/2, average coverage of 95% Prediction Intervals produced by a Simple Kriging model. "True" stands for the prediction based on the knowledge of the true variance parameter and the true vector of correlation lengths.*

Average Mean Length				
Corr. lengths	**True**	**MLE**	**MAP**	**FPD**
0.4 – 0.8 – 0.2	2.23	2.05	2.13	2.59
0.5 – 0.5 – 0.5	1.69	1.55	1.58	1.84
0.7 – 1.3 – 0.4	1.09	1.02	1.07	1.21
0.8 – 0.3 – 0.6	1.63	1.51	1.56	1.82
0.8 – 1.0 – 0.9	0.71	0.66	0.69	0.76

Table 6.6 – *For a Gaussian Process with null mean function, variance parameter 1 and smoothness parameter 5/2, average mean length of 95% Prediction Intervals produced by a Simple Kriging model. "True" stands for the prediction based on the knowledge of the true variance parameter and the true vector of correlation lengths.*

The performance of Ordinary Kriging when the mean function is constant is nearly the same as that of Simple Kriging when the mean function is known.

The performance of Affine Kriging when the mean function is affine, however, is noticeably poorer than the performance of Simple Kriging when the mean function is known: its average coverage is lower. This is not too surprising, since the prediction problem is more difficult.

Average Coverage				
Corr. lengths	**True**	**MLE**	**MAP**	**FPD**
0.4 – 0.8 – 0.2	0.95	0.77	0.81	0.88
0.5 – 0.5 – 0.5	0.95	0.80	0.82	0.89
0.7 – 1.3 – 0.4	0.95	0.82	0.86	0.91
0.8 – 0.3 – 0.6	0.95	0.79	0.83	0.89
0.8 – 1.0 – 0.9	0.95	0.82	0.86	0.91

Table 6.7 – *For a Gaussian Process with mean function $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$, variance parameter 1 and smoothness parameter 5/2, average coverage of 95% Prediction Intervals resulting from Simple Kriging (assuming the mean function is null for MLE/MAP/FPD and knowing it is $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$ for "True"). "True" stands for the prediction based on the knowledge of the true mean function, variance parameter and vector of correlation lengths.*

Misspecified models

In this subsection, we deal with the performance of Kriging in cases where the Gaussian Process does not fit all assumptions.

First, we evaluate the performance of Universal Kriging in a context where the true mean function does not belong to the assumed mean function space $\mathcal{F}_p$. Precisely, we consider a Gaussian process with mean function $(x_1, x_2, x_3) \mapsto 5 + 4x_1 + 3x_2 + 2x_3$ and evaluate the performance of Simple Kriging (assuming the mean function to be null) with respect to that of Affine Kriging, which is the correct model in this situation.
Tables 6.7 and 6.8 show that Simple Kriging performs significantly worse than Affine Kriging when the mean function is $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$, both in terms of average coverage and average mean length of Predictive Intervals. Relative performances of MLE, MAP and FPD once again stay the same, though.

Average Mean Length				
Corr. lengths	**True**	**MLE**	**MAP**	**FPD**
0.4 – 0.8 – 0.2	2.23	2.23	2.36	2.78
0.5 – 0.5 – 0.5	1.69	1.61	1.66	1.92
0.7 – 1.3 – 0.4	1.09	1.03	1.13	1.28
0.8 – 0.3 – 0.6	1.63	1.54	1.63	1.87
0.8 – 1.0 – 0.9	0.71	0.64	0.68	0.77

Table 6.8 – *For a Gaussian Process with mean function $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$, variance parameter 1 and smoothness parameter 5/2, average mean length of 95% Prediction Intervals resulting from Simple Kriging (assuming the mean function is null for MLE/MAP/FPD and knowing it is $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$ for "True"). "True" stands for the prediction based on the knowledge of the true mean function, variance parameter and vector of correlation lengths.*

This observation may lead us to investigate how Simple Kriging behaves with respect to Affine Kriging when the Gaussian Process is smoother than expected. Table 6.9 gives the average coverage and average mean length of Prediction Intervals resulting from the same procedure as before – that is, the correlation kernel is assumed to be Matérn with smoothness 5/2 – but the Gaussian Process actually has a Squared Exponential correlation kernel (with correlation lengths 0.4, 0.8 and 0.2). These results can be compared with those from Table 6.10, which gives the results obtained when both the actual and the assumed correlation kernel are Matérn with smoothness 5/2 (and the true correlation lengths are also 0.4, 0.8 and

0.2). It is apparent that performance is better when the actual kernel is Squared Exponential, both in terms of average coverage and average mean length. Recalling that this kernel can be seen as the limit of the Matérn kernel when the smoothness parameter goes to infinity, we conclude that a smoother process leads to an increase in performance for Simple, Ordinary and Affine Kriging. For Affine Kriging, the smoother process makes Prediction Intervals on average shorter, while the average coverage remains about the same. For Simple Kriging and to a lesser degree Ordinary Kriging, the smoother process makes Prediction Intervals on average shorter, while also increasing average coverage.

Squared Exponential Kernel	**Average coverage**			**Average mean length**		
Kriging model	**MLE**	**MAP**	**FPD**	**MLE**	**MAP**	**FPD**
Simple Kriging*	0.83	0.86	0.92	1.63	1.76	2.02
Ordinary Kriging	0.88	0.90	0.93	1.70	1.79	2.01
Affine Kriging	0.89	0.91	0.93	1.63	1.70	1.88

Table 6.9 – *For a Gaussian Process with mean function* $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$*, variance parameter 1, and squared exponential correlation kernel with correlation lengths 0.4 - 0.8 - 0.2, average coverage and average mean length of 95% Prediction Intervals resulting from different types of Kriging (assuming the smoothness parameter to be 5/2). *With Simple Kriging, the mean function is (wrongly here) assumed to be constant with null value.*

Matérn kernel $\nu = 5/2$	**Average coverage**			**Average mean length**		
Kriging model	**MLE**	**MAP**	**FPD**	**MLE**	**MAP**	**FPD**
Simple Kriging*	0.77	0.81	0.88	2.23	2.36	2.77
Ordinary Kriging	0.84	0.86	0.91	2.30	2.37	2.71
Affine Kriging	0.87	0.90	0.94	2.14	2.23	2.59

Table 6.10 – *For a Gaussian Process with mean function* $(x_1, x_2, x_3) \mapsto 5+4x_1+3x_2+2x_3$*, variance parameter 1 and Matérn kernel with correlation lengths 0.4 - 0.8 - 0.2 and smoothness parameter 5/2, average coverage and average mean length of 95% Prediction Intervals resulting from different types of Kriging. *With Simple Kriging, the mean function is (wrongly here) assumed to be constant with null value.*

All else being equal, smoother processes result in a better quality of prediction for Simple, Ordinary and Affine Kriging, because the observed values of the process yield more information about the value of the process in the neighborhoods of the observation points. This even makes up to some degree for the misspecification of the mean function, so the improvement is greater in the case of Simple Kriging.

Emulating deterministic functions

In this subsection, we test the ability of the model to predict deterministic functions, namely the 7-dimensional Ackley and Rastrigin functions. The Ackley and the Rastrigin functions have the following expressions:

$$A(\boldsymbol{x}) = 20 + \exp(1) - 20\exp\left(-0.2\sqrt{\frac{1}{7}\sum_{i=1}^{7} x_i^2}\right) - \exp\left(\frac{1}{7}\sum_{i=1}^{7}\cos(2\pi x_i)\right); \tag{6.25}$$

$$R(\boldsymbol{x}) = 70 + \sum_{i=1}^{7}\left(x_i^2 - 10\cos(2\pi x_i)\right). \tag{6.26}$$

Naturally, the notions of average coverage and average mean length for Prediction intervals make no sense in this setting, since we can no longer average our results over the distribution of a Gaussian process. Denoting d the deterministic function, and using previous notations, we may define:

Definition 6.14. *The coverage is the probability (with respect to the distribution of $\boldsymbol{X}$ and T) that $d(T) \in f(\boldsymbol{X}, d(\boldsymbol{X}), T)$.*

Definition 6.15. *The mean length is the expectation (with respect to the distribution of $\boldsymbol{X}$ and T) of the length of $f(\boldsymbol{X}, d(\boldsymbol{X}), T)$.*

The coverage $\mathbb{P}[\boldsymbol{Y}(T) \in f(\boldsymbol{X}, d(\boldsymbol{X}), T)]$ is numerically computed as

$$\mathbb{P}[\boldsymbol{Y}(T) \in f(\boldsymbol{X}, \boldsymbol{Y}(\boldsymbol{X}), T)] = \mathbb{E}[\mathbb{P}[\boldsymbol{Y}(T) \in f(\boldsymbol{X}, d(\boldsymbol{X}), T)|\boldsymbol{X}]]$$

over 500 design sets and for each design set 1000 test points. The mean length is computed in a similar fashion.

When emulating the Ackley or the Rastrigin function, we take $r = 7$ and $n = 100$.

We must stress that there is no reason that the coverage of 95% Prediction Intervals, whether produced by MLE or MAP plug-in methods or by the full Gibbs reference posterior distribution should be 95%, but depending on whether or not Kriging can be considered a good surrogate model for the Ackley or Rastrigin function, the coverage of 95% Prediction Intervals may be more or less close to the 95% target figure.

First, we consider an Ordinary Kriging model with anisotropic geometric Matérn kernel of smoothness $\nu = 5/2$. Because $r = 7$ and $n = 100$, Proposition 6.10 is applicable.

When emulating the Ackley function (cf. Table 6.11), regardless of the Kriging method used, the full posterior distribution significantly improves the average coverage of Prediction Intervals when compared to the MLE or the MAP, with a comparatively small trade-off regarding the mean length of these intervals. This result is consistent with results obtained with actual realizations of Gaussian processes.

When we emulate the Rastrigin function (cf. Table 6.12), coverages come closer to the average coverages given in Tables 6.1, 6.3 and 5.2. But the more significant fact of the improvement of the coverage by the full posterior distribution is as true here as in the Ackley case. We may simply infer from this that the Rastrigin function can more plausibly be seen as a realization of a Gaussian Process than the Ackley function.

Let us now compare the performance of different Kriging models: Simple (mean function assumed null), Ordinary and Affine. When emulating the Ackley function, Ordinary and Affine Kriging models yield slightly higher Prediction Interval coverages than Simple Kriging,

Emulated function: Ackley	Coverage			Mean length		
Kriging model	**MLE**	**MAP**	**FPD**	**MLE**	**MAP**	**FPD**
Simple Kriging	0.84	0.87	0.90	0.35	0.36	0.39
Ordinary Kriging	0.87	0.88	0.91	0.37	0.38	0.41
Affine Kriging	0.87	0.90	0.91	0.37	0.39	0.41

Table 6.11 – *Coverage and mean length of 95% Prediction Intervals when emulating the 7-dimensional Ackley function (Matérn anisotropic geometric correlation kernel with smoothness $\nu = 5/2$).*

Emulated function: Rastrigin	Coverage			Mean length		
Kriging model	**MLE**	**MAP**	**FPD**	**MLE**	**MAP**	**FPD**
Simple Kriging	0.94	0.94	0.96	28.3	28.3	30.2
Ordinary Kriging	0.91	0.92	0.94	26.2	26.7	28.3
Affine Kriging	0.90	0.91	0.92	25.9	26.5	27.2

Table 6.12 – *Coverage and mean length of 95% Prediction Intervals when emulating the 7-dimensional Rastrigin function (Matérn anisotropic geometric correlation kernel with smoothness $\nu = 5/2$).*

at the cost of slightly higher mean lengths. When emulating the Rastrigin function, we actually observe the reverse phenomenon.

From this study, we cannot conclusively ascertain whether Universal Kriging, at least in the form of Ordinary or Affine Kriging, yields better results than Simple Kriging. All that can be said is that these Kriging methods are more or less conservative, but even this depends on the emulated function.

In the following example (cf. Table 6.13), we add the linear function $(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \mapsto 100 \sum_{i=1}^{7} x_i$ to the 7-dimensional Rastrigin function. We may expect this modification of the Rastrigin function to be more accurately emulated by Affine Kriging than by Simple Kriging.

Rastrigin + $100 \sum_{i=1}^{7} x_i$	Coverage			Mean length		
Kriging model	**MLE**	**MAP**	**FPD**	**MLE**	**MAP**	**FPD**
Simple Kriging	0.88	0.92	0.94	25.9	29.3	31.1
Ordinary Kriging	0.87	0.91	0.93	25.7	28.5	30.4
Affine Kriging	0.90	0.91	0.92	26.0	26.6	27.3

Table 6.13 – *Coverage and mean length of 95% Prediction Intervals when emulating the 7-dimensional Rastrigin function augmented by a linear function (Matérn anisotropic geometric correlation kernel with smoothness $\nu = 5/2$).*

The addition of the linear function causes a decrease in performance for Prediction Intervals of both Simple and Ordinary Kriging, in the sense that coverage decreases while mean length increases for MAP and FPD. And the coverage of MLE sinks so much – from 94% to 88% for Simple Kriging and from 91% to 87% for Ordinary Kriging – that its performance may also be said to decrease, even though its mean length is slightly lower.

The performance of Affine Kriging is unchanged, however, whether one considers the MLE or MAP plug-in methods or the method using the full posterior distribution. This suggests

that with a stronger linear component, Affine Kriging would be clearly preferable to Simple or Ordinary Kriging.

To test this, we emulate the 7-dimensional Rastrigin function, to which we add a stronger linear term: $(x_1, x_2, x_3, x_4, x_5, x_6, x_7) \mapsto 120 \sum_{i=1}^{7} x_i$.

Rastrigin + $120 \sum_{i=1}^{7} x_i$	Coverage			Mean length		
Kriging model	**MLE**	**MAP**	**FPD**	**MLE**	**MAP**	**FPD**
Simple Kriging	0.90	0.94	0.96	27.6	33.0	37.6
Ordinary Kriging	0.88	0.92	0.94	26.9	30.5	32.3
Affine Kriging	0.90	0.92	0.92	26.0	26.9	27.3

Table 6.14 – *Coverage and mean length of 95% Prediction Intervals when emulating the 7-dimensional Rastrigin function augmented by a linear function (Matérn anisotropic geometric correlation kernel with smoothness $\nu = 5/2$).*

For Simple Kriging, Prediction Intervals coverage and mean length are both higher when $120 \sum_{i=1}^{7} x_i$ is added to the Rastrigin function (Table 6.14) rather than $100 \sum_{i=1}^{7} x_i$ (Table 6.13). This is also true, though to a lesser extent, of Ordinary Kriging. The performance of Affine Kriging, on the other hand, still remains the same because it can account for any linear term by seeing it as part of the mean function. Simple Kriging (assuming the mean function to be null) and Ordinary Kriging do not have this luxury and must assume a greater variance for the Gaussian process, which results in more conservative Predictive Intervals.

Gathering the results obtained above, we conclude that Universal Kriging only significantly improves performance if the trend belongs to the assumed mean function space $\mathcal{F}_p$ and if it stands out. In other words, the signal/noise ratio must be high, where the signal is here the "true" mean function and the noise is the stationary Gaussian Process added to it. When no trend of the expected form can be discerned, like when emulating the Ackley or Rastrigin function through Affine Kriging, then there is no significant benefit to using Universal instead of Simple Kriging. When the ratio is high, as in the case of the Rastrigin function with the addition of the greater linear term $120 \sum_{i=1}^{7} x_i$, Universal Kriging (if the mean function space $\mathcal{F}_p$ is adequately defined) improves upon Simple Kriging, which becomes overly conservative. Further, when the emulated function is particularly smooth, Simple Kriging becomes capable of capturing the trend to some extent even if the mean function is misspecified, thanks to the mechanics of Gaussian conditioning.

6.6 Conclusion

In this chapter, we provided an Objective Bayesian solution to the problem of taking into account parameter uncertainty when performing prediction based on a Universal Kriging model with anisotropic Matérn autocorrelation kernel. The reference posterior on the location parameter $\boldsymbol{\beta}$ and the variance parameter σ^2 is coupled with the Gibbs reference posterior on the vector of correlation lengths $\boldsymbol{\theta}$. By using the Gibbs reference posterior, which is the optimal compromise between the conditional reference posteriors on one correlation length θ_i based on the knowledge of all other correlation length θ_j $(j \neq i)$, we bypass the problem of determining an ordering on the correlation lengths. Moreover, this solution allows for

Gibbs sampling of the posterior distribution, which makes full-Bayesian inference or prediction tractable.

We proved that the Gibbs reference posterior exists and is proper in several Universal Kriging settings, depending on the number of available observation points and on the smoothness parameter of the Matérn kernel.

Numerical simulations show that Prediction Intervals produced by the full-Bayesian procedure based on the Gibbs reference posterior have better coverage than those produced by the Maximum Likelihood Estimator or even the Maximum A Posteriori estimator, and that their mean length is only moderately greater.

In addition, these simulations showed that when emulating deterministic functions, there is no obvious advantage to using Universal Kriging over Simple Kriging, unless the trend strongly stands out and belongs to the assumed mean function space.

From a theoretical standpoint, the Universal Kriging setting poses specific problems when compared to the Simple Kriging setting. As was shown (to our knowledge for the first time) by Berger et al. [2001], the behavior of the integrated likelihood changes significantly depending on whether functions that take a non-null constant value on the design set are included in the mean function space $\mathcal{F}_p$. The integrated likelihood often fails to vanish in the neighborhood of perfect correlation in Ordinary Kriging models and *a fortiori* in more complex Universal Kriging models where the constant term of the mean function is unknown. Berger et al. [2001] show in the isotropic framework that the reference prior adapts to this situation by being proper (at least for sufficiently rough correlation kernels – we provided a proof for smoother kernels in Chapter 3). In the anisotropic framework however, we were not able to prove the existence of the Gibbs reference posterior in such a situation, which is why we require Assumption 4. Although it is possible that closer analysis may allow us to relax this requirement, we find it more likely that Assumption 4 is the price we pay for defining the Gibbs reference posterior as a compromise between incompatible conditional reference posterior distributions. Indeed, each conditional maximizes the expected information of the model when all but one correlation length are fixed at finite values, i.e. in a context where *perfect correlation is impossible*, whatever may be the value of the unfixed correlation length. Therefore, it is conceivable that in the absence of penalization by the integrated likelihood of the kind given by Assumption 4, the conditionals may place too much weight on high values of the unfixed correlation length for the Gibbs reference posterior to be well defined.

Taking this restriction into account in Theorem 6.9, we proved that the Gibbs reference posterior exists and is the limit of a uniformly converging Markov Chain Monte-Carlo (MCMC) algorithm for commonly used Matérn anisotropic geometric and tensorized correlation kernels when the design set has enough points (cf. Propositions 6.10 and 6.11). More generally, we would conjecture that for any noninteger smoothness $\nu \in (1, +\infty)$, and if the mean function space $\mathcal{F}_p$ does not contain polynomials of degree higher than $[\nu] - 1$, there exists some lower bound on the cardinal of the design set over which the Gibbs reference posterior exists and the MCMC algorithm uniformly converges to it. However, this lower bound may be too high for practical purposes.

Future work may involve gaining a better understanding of the significance of the Gibbs reference posterior as a compromise between the incompatible reference conditionals on correlation

lengths. This method was primarily intended as a practical means of solving the problem of giving an objective posterior distribution on correlation lengths in the case of anisotropic correlation kernels, where the reference posterior is intractable and may not be proper. But its theoretical properties beyond its propriety, its invariance under reparametrizations of the type $f((\theta_1, ..., \theta_r)^\top) = (f_1(\theta_1), ..., f_r(\theta_r))^\top$ and its apparent good frequentist performances remain unknown.

Appendix 6.A Matérn kernels

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 $g(\boldsymbol{x}) = \int_{\mathbb{R}^r} \widehat{g}(\boldsymbol{\omega}) e^{i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{\omega}$ and $\widehat{g}(\boldsymbol{\omega}) = (2\pi)^{-r} \int_{\mathbb{R}^r} g(\boldsymbol{x}) e^{-i\langle \boldsymbol{\omega} | \boldsymbol{x} \rangle} d\boldsymbol{x}$. 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) $\forall \boldsymbol{x} \in \mathbb{R}^r$,

$$K_{r,\nu}(\boldsymbol{x}) = \frac{1}{\Gamma(\nu) 2^{\nu-1}} \left(2\sqrt{\nu}\|\boldsymbol{x}\|\right)^\nu \mathcal{K}_\nu \left(2\sqrt{\nu}\|\boldsymbol{x}\|\right) \ ; \tag{6.27}$$

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

$$\widehat{K}_{r,\nu}(\boldsymbol{\omega}) = \frac{M_r(\nu)}{(\|\boldsymbol{\omega}\|^2 + 4\nu)^{\nu + \frac{r}{2}}} \text{ with } M_r(\nu) = \frac{\Gamma(\nu + \frac{r}{2})(2\sqrt{\nu})^{2\nu}}{\pi^{\frac{r}{2}} \Gamma(\nu)}. \tag{6.28}$$

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

(i) $\forall \boldsymbol{x} \in \mathbb{R}^r$,

$$K^{tens}_{r,\nu}(\boldsymbol{x}) = \prod_{j=1}^r K_{1,\nu}(\boldsymbol{x}_j) \ ; \tag{6.29}$$

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

$$\widehat{K}^{tens}_{r,\nu}(\boldsymbol{\omega}) = \prod_{j=1}^r \widehat{K_{1,\nu}}(\boldsymbol{\omega}_j). \tag{6.30}$$

(d) let us adopt the following convention: 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 $\boldsymbol{x} \mapsto \sigma^2 K^{tens}_{r,\nu}\left(\frac{\boldsymbol{x}}{\boldsymbol{\theta}}\right)$ (resp. $\boldsymbol{x} \mapsto \sigma^2 K^{tens}_{r,\nu}(\boldsymbol{x}\boldsymbol{\mu})$).

Appendix 6.B Proofs of the existence of the Gibbs reference posterior

The proof of the existence and uniqueness of the Gibbs reference posterior that was used in Chapter 5 to deal with the Simple Kriging setting is inadequate in the Universal Kriging setting because the projection $\boldsymbol{W}^\top$ may make key facts used in that chapter untrue. In the following, we provide replacements for the parts of the proof in Appendix 6.A that are invalid in the Universal Kriging setting.
The proof contained two parts, one dealing with "low correlations", that is $\|\boldsymbol{\mu}\| \to +\infty$ and one with "high correlation", that is $\|\boldsymbol{\mu}\| \to 0$.

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

Concerning the part about $\|\boldsymbol{\mu}\| \to +\infty$, we need to make sure that Corollary 5.13 remains true.

Define the functions h_i by

$$h_i(\mu_i \mid \boldsymbol{\mu}_{-i}) := \sqrt{\operatorname{Tr}\left[\left(\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)^2\right]} = \left\|\frac{\partial}{\partial \mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\|. \tag{6.31}$$

The conclusion of Corollary 5.13 is that that there exist $S > 0$ and $0 < a < b$ such that, whenever $\|\boldsymbol{\mu}\| \geqslant S$,

$$a\, h_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \leqslant f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \leqslant b\, h_i(\mu_i \mid \boldsymbol{\mu}_{-i}). \tag{6.32}$$

We need to find conditions under which this is true. While the right inequality is obvious, the left inequality is harder to show.

Fix $\boldsymbol{\alpha} = \boldsymbol{\mu}/\|\boldsymbol{\mu}\|_\infty$. Then define $\boldsymbol{L}_{i,\boldsymbol{\alpha}} = \lim\limits_{\|\boldsymbol{\mu}\|\to\infty} \frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}} / \left\|\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right\|_\infty$.

We now give an explicit form for $\boldsymbol{L}_{i,\boldsymbol{\alpha}}$. Let $\boldsymbol{X}$ be the $n \times r$ matrix representing the design set, and let $\boldsymbol{X}_{\boldsymbol{\alpha}}$ be the matrix $\boldsymbol{X}\mathit{Diag}(\boldsymbol{\alpha})$, where $\mathit{Diag}(\boldsymbol{\alpha})$ is the $r \times r$ diagonal matrix whose diagonal is the vector $\boldsymbol{\alpha}$.

Proposition 6.16. *If the Matérn kernel is anisotropic geometric, then $L_{i,\boldsymbol{\alpha}}$ is the symmetric $n \times n$ matrix with null diagonal whose nondiagonal coefficients are given by the following rule: its (a, b) coefficient ($a, b \in [\![1, n]\!]$ and $a \neq b$) is -1 if the a-th and b-th point in the design set $\boldsymbol{X}_{\boldsymbol{\alpha}}$ achieve minimal Euclidean distance within this design set, and 0 otherwise.*

Proof. We only prove the result when $\nu > 1$, but the proof is very similar in the case where $0 < \nu \leqslant 1$.

Abramowitz and Stegun [1964] (formula 9.7.2.) yields an equivalent for the one-dimensional Matérn kernel when $t \to +\infty$:

$$K_{1,\nu}(t) \sim \frac{\sqrt{\pi/2}}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}t)^{\nu-1/2}\exp(-2\sqrt{\nu}t) \tag{6.33}$$

From Abramowitz and Stegun [1964] (formula 9.6.28.), we obtain that:

$$\begin{aligned} K'_{1,\nu}(t) &= -\frac{2\nu t}{\nu-1}K_{1,\nu-1}\left(\sqrt{\frac{\nu}{\nu-1}}t\right) \\ &\sim -2\sqrt{\nu}\frac{\sqrt{\pi/2}}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}t)^{\nu-1/2}\exp(-2\sqrt{\nu}t) \\ &\sim -2\sqrt{\nu}K_{1,\nu}(t) \end{aligned} \tag{6.34}$$

The result follows after recalling that $\frac{\partial}{\partial\mu_i}K_{r,\nu}(\boldsymbol{\mu}\boldsymbol{x}) = \mu_i x_i^2 \|\boldsymbol{\mu}\boldsymbol{x}\|^{-1}K'_{1,\nu}(\|\boldsymbol{\mu}\boldsymbol{x}\|)$. When $\|\boldsymbol{\mu}\| \to \infty$,

$$\frac{\partial}{\partial\mu_i}K_{r,\nu}(\boldsymbol{\mu}\boldsymbol{x}) \sim -2\sqrt{\nu}\mu_i x_i^2\|\boldsymbol{\mu}\boldsymbol{x}\|^{-1}\frac{\sqrt{\pi/2}}{\Gamma(\nu)2^{\nu-1}}(2\sqrt{\nu}\|\boldsymbol{\mu}\boldsymbol{x}\|)^{\nu-1/2}\exp(-2\sqrt{\nu}\|\boldsymbol{\mu}\boldsymbol{x}\|). \tag{6.35}$$

In the case where $0 < \nu \leqslant 1$, $\frac{\partial}{\partial\mu_i}K_{r,\nu}(\boldsymbol{\mu}\boldsymbol{x})$ also has an equivalent when $\|\boldsymbol{\mu}\| \to \infty$ whose prominent factor is $\exp(-2\sqrt{\nu}\|\boldsymbol{\mu}\boldsymbol{x}\|)$, so the end result is the same. □

Proposition 6.17. *If the Matérn kernel is tensorized with smoothness $\nu > 1$, then $L_{i,\boldsymbol{\alpha}}$ is the matrix with nonpositive coefficients such that $\|L_{i,\boldsymbol{\alpha}}\|_\infty = 1$ which is proportional to the symmetric matrix described hereafter: it has null diagonal and its nondiagonal coefficients are given by the following rule: its (a,b) coefficient ($a, b \in [\![1,n]\!]$ and $a \neq b$) is 0 if the a-th and b-th point in the design set $\boldsymbol{X}_{\boldsymbol{\alpha}}$ do not achieve minimal 1-distance within this design set, and $\alpha_i^{\nu-1/2} \left|x_i^{(a)} - x_i^{(b)}\right|^{\nu+1/2} \prod_{j\neq i} \alpha_j^{\nu-1/2} \left|x_j^{(a)} - x_j^{(b)}\right|^{\nu-1/2}$ if they do.*

Remark. If the Matérn kernel is tensorized with smoothness $0 < \nu \leqslant 1$, then the same rule applies but with different formula when minimal 1-distance is achieved.

Proof. The proof is similar to that of Proposition 6.16. □

Corollary 6.18. *For Matérn anisotropic geometric and tensorized kernels, if the design set $\boldsymbol{X}$ is randomly chosen according to the Uniform probability distribution on $(0,1)^{rn}$, then almost surely, whatever $i \in [\![1,n]\!]$ and $\boldsymbol{\alpha}$ in $\mathbb{R}^r$ such that $\|\boldsymbol{\alpha}\|_\infty = 1$, $L_{i,\boldsymbol{\alpha}}$ has rank lower or equal to $2r$.*

Proof. Almost surely, whatever $\boldsymbol{\alpha}$ in $\mathbb{R}^r$ such that $\|\boldsymbol{\alpha}\|_\infty = 1$, the design set $\boldsymbol{X}_{\boldsymbol{\alpha}}$ has at most r couples of distinct points achieving equal distance (whether that distance be the 1- or 2-distance). *A fortiori*, it has at most r couples of distinct points achieving minimal distance. □

With fixed $\boldsymbol{\alpha}$, as $\|\boldsymbol{\mu}\| \to \infty$, we have

$$f_i(\mu_i \mid \boldsymbol{\mu}_{-i}) \sim \left\| \frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \right\|_\infty \sqrt{\operatorname{Tr}\left[\left(\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}\right)^2\right] - \frac{1}{n-p} \operatorname{Tr}\left[\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}\right]^2} \tag{6.36}$$

We may recognize the factor under the square root as the variance (multiplied by $n-p$) of the eigenvalues (accounting for multiplicity) of the matrix $\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}$. If the premise of Corollary 6.18 holds, and if $2r < n-p$, then it is null if and only if $\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}$ is the null matrix. Assumption 3 is designed to prevent this from happening.

Proposition 6.19. *Assume $2r < n-p$. For Matérn anisotropic geometric or tensorized correlation kernels, if the design set $\boldsymbol{X}$ is randomly chosen according to the Uniform probability distribution on $(0,1)^{rn}$, then almost surely, Assumption 3 implies that*

$$\min_{i\in[\![1,n]\!], \|\boldsymbol{\alpha}\|_\infty = 1} \sqrt{\operatorname{Tr}\left[\left(\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}\right)^2\right] - \frac{1}{n-p} \operatorname{Tr}\left[\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}\right]^2} > 0. \tag{6.37}$$

Proof. First, set $i \in [\![1,n]\!]$ and $\boldsymbol{\alpha}$ in $\mathbb{R}^r$ such that $\|\boldsymbol{\alpha}\|_\infty = 1$. We prove that $\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}$ is not the null matrix.

Assume that it is and that Assumption 3 holds. Assumption 3 implies that the intersection of the vector space spanned by $\boldsymbol{P}$ and the image of $\boldsymbol{L}_{i,\boldsymbol{\alpha}}$ is $\{\boldsymbol{0}_n\}$. Therefore, for any $\boldsymbol{z} \in \mathbb{R}^{n-p}$, if $\boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W} z \neq \boldsymbol{0}_n$, then $\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W} z \neq \boldsymbol{0}_{n-p}$, which contradicts the assumption that $\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}$ is the null matrix. So $\boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}$ is the null $n \times (n-p)$ matrix, and thus the vector space spanned by $\boldsymbol{W}$ is included in the kernel of $\boldsymbol{L}_{i,\boldsymbol{\alpha}}$. This implies that $\boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{P} \boldsymbol{P}^\top = \boldsymbol{L}_{i,\boldsymbol{\alpha}}$, and then that $\boldsymbol{P} \boldsymbol{P}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} = \boldsymbol{L}_{i,\boldsymbol{\alpha}}$. However, per Propositions 6.16 and 6.17, all vectors in the image of $\boldsymbol{L}_{i,\boldsymbol{\alpha}}$ have at most $2r$ non-null elements when expressed in the canonical base of $\mathbb{R}^n$, so Assumption

3 implies that $\boldsymbol{P}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}}$ is the null $p \times n$ matrix, and thus that $\boldsymbol{L}_{i,\boldsymbol{\alpha}}$ is the null $n \times n$ matrix, which is untrue.

So, under Assumption 3, whatever $i \in [\![1, n]\!]$ and $\boldsymbol{\alpha}$ in $\mathbb{R}^r$ such that $\|\boldsymbol{\alpha}\|_\infty = 1$, $\boldsymbol{W}^\top \boldsymbol{L}_{i,\boldsymbol{\alpha}} \boldsymbol{W}$ is not the null matrix and thus has a non-null eigenvalue. Moreover, $2r < n - p$ implies, according to Corollary 6.18, that it also almost surely has a null eigenvalue, so the standard deviation of its eigenvalues is positive. As the number of possible matrices $\boldsymbol{L}_{i,\boldsymbol{\alpha}}$ (with $i \in [\![1, n]\!]$ and $\boldsymbol{\alpha}$ in $\mathbb{R}^r$ such that $\|\boldsymbol{\alpha}\|_\infty = 1$) is almost surely finite, this yields the result. □

Corollary 6.20. *Assume $2r < n - p$. For Matérn anisotropic geometric or tensorized correlation kernels, if the design set $\boldsymbol{X}$ is randomly chosen according to the Uniform probability distribution on $(0, 1)^{rn}$, then almost surely, Assumption 3 implies that there exist $S > 0$ and $0 < a < b$ such that whenever $\|\boldsymbol{\mu}\| \geqslant S$, Equation (6.32) holds.*

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

In the part of the proof in Appendix 5.A concerning $\|\boldsymbol{\mu}\| \to 0$, we used a the series expansion of $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}$. This expansion may be heavily modified by premultiplication by $\boldsymbol{W}^\top$ and postmultiplication by $\boldsymbol{W}$.

In the case where $\nu < 1$, there is no material change unless the vector $\mathbf{1}$ belongs to the vector space spanned by $\boldsymbol{H}$.

Proof of Proposition 6.8 . Because $\mathbf{1}$ does not belong to the vector space spanned by $\boldsymbol{H}$, $\boldsymbol{W}^\top \mathbf{1}\mathbf{1}^\top \boldsymbol{W}$ has rank 1 and so the proof of this result is the same as in the Simple Kriging case. □

If $\mathbf{1}$ *does* belong to the vector space spanned by $\boldsymbol{H}$, further study would be needed to assess whether or not the above theorem still applies, essentially because we cannot count on $L(\boldsymbol{y}|\boldsymbol{\mu})$ vanishing as $\|\boldsymbol{\mu}\| \to 0$.

Let us now focus on the case where $\nu > 1$. We reproduce in Lemma 6.21 key facts given by Lemma 5.7 and Proposition 5.22:

Lemma 6.21. *For any Matérn anisotropic geometric or tensorized correlation kernel with smoothness parameter $\nu > 1$, if a coordinate-distinct design set is used, there exists $a > 0$ such that when $\|\boldsymbol{\mu}\| \to 0$:*

1. $\left\| \frac{\partial}{\partial \mu_i} \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \right\| = O(\mu_i)$;
2. $\left\| \boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} \right\| = O(\|\boldsymbol{\mu}\|^{-a})$.

The newt result follows immediately.

Corollary 6.22. *For any Matérn anisotropic geometric or tensorized correlation kernel with smoothness parameter $\nu > 1$, if a coordinate-distinct design set is used, there exist $a > 0$, $m > 0$ and $S > 0$ such that, for any $\boldsymbol{\mu} \in (0, +\infty)^r$ such that $\|\boldsymbol{\mu}\| \leqslant S$ and $\mu_i \leqslant \|\boldsymbol{\mu}\|^a$, $f_i(\mu_i|\boldsymbol{\mu}_{-i}) \leqslant m$.*

We combine the previous fact with a useful universal majoration of $f_i(\mu_i|\boldsymbol{\mu}_{-i})$.

Proposition 6.23. *For an r-dimensional anisotropic geometric or tensorized Matérn correlation kernel with smoothness parameter ν pertaining to a design set containing n coordinate-distinct points, $\forall \boldsymbol{\mu} \in [0,+\infty)^r$ such that $\mu_i > 0$,*

$$f_i(\mu_i|\boldsymbol{\mu}_{-i}) \leqslant (n-p)(2\nu+r)\mu_i^{-1} \tag{6.38}$$

Proof. Whatever $x, y \in \mathbb{R}$, $K_{1,\nu}(x-y) = \int_{\mathbb{R}} \widehat{K}_{1,\nu}(\omega) e^{i\omega(x-y)} d\omega$.
For the sake of concision, we only consider the case where the Matérn kernel is anisotropic geometric, as the changes in the case of a tensorized kernel are straightforward.

Moreover, we start by proving the result in the case where $\boldsymbol{W}$ is the identity matrix $\boldsymbol{I}_n$ (Simple Kriging case).

$$\begin{aligned}\sum_{j,k=1}^{n} \xi_j \xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right) &= \int_{\mathbb{R}^r} \widehat{K_{r,\nu}}(\boldsymbol{\omega}) \left|\sum_{j=1}^{n} \xi_j e^{i\omega_i \mu_i x_i^{(j)} + i\left\langle \boldsymbol{\omega}_{-i} \middle| \boldsymbol{\mu}_{-i} \boldsymbol{x}_{-i}^{(j)} \right\rangle}\right|^2 d\boldsymbol{\omega} \\ &= M_r(\nu)\mu_i^{-1} I_{\boldsymbol{\mu}}(\boldsymbol{\xi}),\end{aligned} \tag{6.39}$$

where

$$M_r(\nu) = \frac{\Gamma(\nu+\frac{r}{2})(2\sqrt{\nu})^{2\nu}}{\pi^{\frac{r}{2}}\Gamma(\nu)}; \tag{6.40}$$

$$I_{\boldsymbol{\mu}}(\boldsymbol{\xi}) = \int_{\mathbb{R}^r} \left(4\nu + \mu_i^{-2} s_i^2 + \left\|\frac{\boldsymbol{s}_{-i}}{\boldsymbol{\mu}_{-i}}\right\|^2\right)^{-\frac{r}{2}-\nu} \left|\sum_{j=1}^{n} \xi_j e^{i\langle \boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s}. \tag{6.41}$$

We also have

$$\frac{d}{d\mu_i} \sum_{j,k=1}^{n} \xi_j \xi_k K_{r,\nu}\left(\left(\boldsymbol{x}^{(j)} - \boldsymbol{x}^{(k)}\right)\boldsymbol{\mu}\right) = -M_r(\nu)\mu_i^{-1} I_{\boldsymbol{\mu}}(\boldsymbol{\xi}) + M_r(\nu)\mu_i^{-1} \frac{d}{d\mu_i} I_{\boldsymbol{\mu}}(\boldsymbol{\xi}) \tag{6.42}$$

$$\begin{aligned}&\frac{d}{d\mu_i} I_{\boldsymbol{\mu}}(\boldsymbol{\xi}) \\ =&2\left(\frac{r}{2}+\nu\right)\mu_i^{-3} \int_{\mathbb{R}^r} s_i^2 \left(4\nu + \mu_i^{-2} s_i^2 + \left\|\frac{\boldsymbol{s}_{-i}}{\boldsymbol{\mu}_{-i}}\right\|^2\right)^{-\frac{r}{2}-\nu-1} \left|\sum_{j=1}^{n} \xi_j e^{i\langle \boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s} \\ =&(2\nu+r)\mu_i^{-3} \int_{\mathbb{R}^r} \frac{s_i^2}{4\nu + \mu_i^{-2} s_i^2 + \left\|\frac{\boldsymbol{s}_{-i}}{\boldsymbol{\mu}_{-i}}\right\|^2} \left(4\nu + \mu_i^{-2} s_i^2 + \left\|\frac{\boldsymbol{s}_{-i}}{\boldsymbol{\mu}_{-i}}\right\|^2\right)^{-\frac{r}{2}-\nu} \left|\sum_{j=1}^{n} \xi_j e^{i\langle \boldsymbol{s}|\boldsymbol{x}^{(j)}\rangle}\right|^2 d\boldsymbol{s}\end{aligned} \tag{6.43}$$

From this, we obtain that for any non-null vector $\boldsymbol{\xi} \in \mathbb{R}^n$,

$$0 < \frac{d}{d\mu_i} I_{\boldsymbol{\mu}}(\boldsymbol{\xi}) \leqslant (2\nu+r)\mu_i^{-1} I_{\boldsymbol{\mu}}(\boldsymbol{\xi}) \tag{6.44}$$

Now let us define the matrix $\boldsymbol{F}_{\boldsymbol{\mu}}$ as the matrix representing in the canonical base of $\mathbb{R}^n$ the positive definite quadratic form $\boldsymbol{\xi} \mapsto M_r(\nu)\mu_i^{-1} \frac{d}{d\mu_i} I_{\boldsymbol{\mu}}(\boldsymbol{\xi})$. From the previous calculations, we

gather that $\frac{d}{d\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}} = -\mu_i^{-1}\boldsymbol{\Sigma}_{\boldsymbol{\mu}} + \boldsymbol{F}_{\boldsymbol{\mu}}$. This in turn yields $\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1} = -\mu_i^{-1}\boldsymbol{I}_n + \boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}$ and $\left(\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2 = \mu_i^{-2}\boldsymbol{I}_n + \left(\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2 - 2\mu_i^{-1}\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}$.

$$\operatorname{Tr}\left[\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right] = -n\mu_i^{-1} + \operatorname{Tr}\left[\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]. \tag{6.45}$$

$$\operatorname{Tr}\left[\left(\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] = n\mu_i^{-2} + \operatorname{Tr}\left[\left(\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] - 2\mu_i^{-1}\operatorname{Tr}\left[\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]. \tag{6.46}$$

$$\operatorname{Tr}\left[\left(\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]^2 = \operatorname{Tr}\left[\left(\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]^2. \tag{6.47}$$

$\boldsymbol{F}_{\boldsymbol{\mu}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}$ being two symmetric positive definite matrices, their product $\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}$ is diagonalizable and all its eigenvalues are positive. Thus $\operatorname{Tr}\left[\left(\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] \leqslant \operatorname{Tr}\left[\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]^2$.

Let $(\boldsymbol{\xi}_{\boldsymbol{\mu}}^j)_{1\leqslant j\leqslant n}$ be a basis of unit eigenvectors of $\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}$. Then

$$\operatorname{Tr}\left[\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right] = \sum_{j=1}^{n}\left(\boldsymbol{\xi}_{\boldsymbol{\mu}}^j\right)^\top \boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\boldsymbol{\xi}_{\boldsymbol{\mu}}^j = \sum_{j=1}^{n}\frac{\left(\boldsymbol{\xi}_{\boldsymbol{\mu}}^j\right)^\top \boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{\xi}_{\boldsymbol{\mu}}^j}{\left(\boldsymbol{\xi}_{\boldsymbol{\mu}}^j\right)^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{\xi}_{\boldsymbol{\mu}}^j} \leqslant n(2\nu + r)\mu_i^{-1}. \tag{6.48}$$

This implies that

$$\operatorname{Tr}\left[\left(\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]^2 \leqslant n(n-1)(2\nu + r)^2\mu_i^{-2} \tag{6.49}$$

$$\sqrt{\operatorname{Tr}\left[\left(\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right)^2\right] - \frac{1}{n}\operatorname{Tr}\left[\left(\frac{\partial}{\partial\mu_i}\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\right)\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right]^2} \leqslant n(2\nu + r)\mu_i^{-1} \tag{6.50}$$

Now, if $\boldsymbol{W}$ is not the identity matrix, then the previous proof still holds, albeit with some alterations. Instead of considering all non-null vectors $\boldsymbol{\xi} \in \mathbb{R}^n$, we consider only those which can be expressed as $\boldsymbol{W}\boldsymbol{\xi}_{\boldsymbol{W}}$, with $\boldsymbol{\xi}_{\boldsymbol{W}}$ belonging to $\mathbb{R}^{n-p}$. In the same vein, once it comes to computing $\operatorname{Tr}\left[\boldsymbol{W}^\top\boldsymbol{F}_{\boldsymbol{\mu}}\boldsymbol{W}\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{W}\right)^{-1}\right]$, we use a basis $(\boldsymbol{\xi}_{\boldsymbol{W},\boldsymbol{\mu}}^j)_{1\leqslant j\leqslant n-p}$ of unit eigenvectors of $\left(\boldsymbol{W}^\top\boldsymbol{\Sigma}_{\boldsymbol{\mu}}\boldsymbol{W}\right)^{-1}$.

□

Proposition 6.24. *With a Matérn anisotropic geometric or tensorized correlation kernel with smoothness $\nu > 1$, if a design set with coordinate-distinct points is used, then Assumption 4 implies that there exists $\epsilon' > 0$ such that $L(\boldsymbol{y}|\boldsymbol{\mu})f_i(\mu_i|\boldsymbol{\mu}_{-i}) = O(\mu_i^{-1+\epsilon'})$ when $\|\boldsymbol{\mu}\| \to 0$.*

Proof. Assumption 4 ensures that $L(\boldsymbol{y}|\boldsymbol{\mu})$ is bounded as a function of $\boldsymbol{\mu}$. Because of Corollary 6.22, and using said Corollary's notations, we know that there exists $M > 0$ such that, for any $\boldsymbol{\mu} \in (0, +\infty)^r$ such that $\mu_i \leqslant \|\boldsymbol{\mu}\|^a$, $L(\boldsymbol{y}|\boldsymbol{\mu})f_i(\mu_i|\boldsymbol{\mu}_{-i}) \leqslant M$.

Let us now focus on the $\boldsymbol{\mu} \in (0, +\infty)^r$ such that $\mu_i \geqslant \|\boldsymbol{\mu}\|^a$. Then $\|\boldsymbol{\mu}\|^\epsilon \leqslant \mu_i^{\epsilon/a}$. Choosing $\epsilon' = \epsilon/a$, combining Assumption 4 and Proposition 6.23 yields the result. □

With the help of Proposition 6.24, using essentially the proof of Proposition 5.26, we obtain the following result.

Proposition 6.25. *In a Universal Kriging model with a Matérn anisotropic geometric or tensorized correlation kernel with smoothness $\nu > 1$, if a design set with coordinate-distinct points is used, then Assumption 4 implies that the conditional posterior distribution $\pi_i(\mu_i|\boldsymbol{y}, \boldsymbol{\mu}_{-i})$, seen as a function of $\boldsymbol{\mu}$, is continuous over $\{\boldsymbol{\mu} \in [0, +\infty)^r : \mu_i \neq 0\}$.*

Proofs of the main results

Proof of Theorem 6.9. With the help of Proposition 6.25 and Corollary 6.20, a similar result to Lemma 5.27 can be proved. Then the proof of Theorem 6.9 works like the proof of Theorem 5.2. □

In order to be able to state the remaining results in a concise manner, we implicitely assume from this point onwards that the correlation kernel is Matérn anisotropic geometric or tensorized. Consider the following set of conditions:

1. $1 < \nu < 2$ and $n > r + 1$;
2. $2 < \nu < 3$ and $n > (r + 1)(r/2 + 2)$.

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

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

Proposition 6.26. *In the case of Universal Kriging where the mean function space is included within the space of polynomials of degree 0 or 1, if one of the previous conditions is satisfied, for any vector $\boldsymbol{y} \in \mathbb{R}^n$ not orthogonal to k_ν, when $\|\boldsymbol{\mu}\| \to 0$,*

$$\|\boldsymbol{\mu}\|^{-2\nu} = O\left(\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right). \tag{6.51}$$

Proof. The proof is similar to that of Proposition 5.23 after noticing that $\boldsymbol{W}\boldsymbol{W}^\top$ is a projection on a vector space that contains k_ν. □

Proposition 6.27. *In the case of Ordinary Kriging, under the conditions of Theorem 6.9, if one of the previous conditions is satisfied, then there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that, provided $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, Assumption 4 is true.*

Proposition 6.28. *In the case of Universal Kriging where the mean function space is included within the space of polynomials of degree 0 or 1, if $2 < \nu < 3$ and $n > r(r + 1)/2 + 2r + 3$, then there exists a hyperplane $\mathcal{H}$ of $\mathbb{R}^n$ such that, provided $\boldsymbol{y} \in \mathbb{R}^n \setminus \mathcal{H}$, Assumption 4 is true.*

Proof of Propositions 6.27 and 6.28. Let $v_1(\boldsymbol{\mu}) \geqslant v_2(\boldsymbol{\mu}) \geqslant ... \geqslant v_{n-p}(\boldsymbol{\mu})$ be the ordered eigenvalues of $\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}$. We can now rewrite $L(\boldsymbol{y}|\boldsymbol{\mu})$ as

$$L(\boldsymbol{y}|\boldsymbol{\mu})^2 \propto \prod_{k=1}^{n-p} \left[v_k(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right)^{-1} \right]. \tag{6.52}$$

Proposition 6.26 asserts that for any $\boldsymbol{y} \in \mathbb{R}^n$ that is not orthogonal to k_ν, the following holds: $\left(\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right)^{-1} = O(\|\boldsymbol{\mu}\|^{2\nu})$ for $\|\boldsymbol{\mu}\| \to 0$.

Besides, Proposition 5.22 asserts that $\left\|\boldsymbol{\Sigma}_{\boldsymbol{\mu}}^{-1}\right\| = O(\|\boldsymbol{\mu}\|^{-2\nu})$, so a fortiori $\left\|\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1}\right\| = O(\|\boldsymbol{\mu}\|^{-2\nu})$ and therefore $\left(\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right)^{-1} = O\left(\left\|\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1}\right\|^{-1}\right)$. This implies that for every integer $i \in [\![1, r]\!]$, $v_k(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right)^{-1} = O(1)$.

Moreover, $\left\|\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right\| = O(v_1(\boldsymbol{\mu}))$ so $v_1(\boldsymbol{\mu})^{-1} = O\left(\left\|\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right\|^{-1}\right)$. In all cases considered, $\|\boldsymbol{\mu}\|^{2\lfloor\nu\rfloor} = O\left(\left\|\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right\|\right)$, which implies that the following asymptotic upper bound holds: $v_1(\boldsymbol{\mu})^{-1} \left(\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\mu}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right)^{-1} = O\left(\|\boldsymbol{\mu}\|^{2\nu-2\lfloor\nu\rfloor}\right)$.

In the end,

$$L(\boldsymbol{y}|\boldsymbol{\mu})^2 = O\left(\|\boldsymbol{\mu}\|^{2\nu-2\lfloor\nu\rfloor}\right). \tag{6.53}$$

□

Proof of Propositions 6.10 and 6.11. Propositions 6.10 and 6.11 are obtained by combining Theorem 6.9 with Propositions 6.27 and 6.28 respectively. □

Chapter 7

Trans-Gaussian Kriging in a Bayesian framework: a case study

This chapter corresponds to the article Muré [2018b].

Abstract

In the context of Gaussian Process Regression or Kriging, we propose a full-Bayesian solution to deal with hyperparameters of the covariance function. This solution can be extended to the Trans-Gaussian Kriging framework, which makes it possible to deal with spatial data sets that violate assumptions required for Kriging. It is shown to be both elegant and efficient. We propose an application to computer experiments in the field of nuclear safety, where it is necessary to model non-destructive testing procedures based on eddy currents to detect possible wear in steam generator tubes.

Résumé

Dans le but de faciliter la régression par processus gaussiens ou krigeage, nous proposons une solution pleinement bayésienne pour traiter les hyperparamètres de la fonction de covariance. Cette solution peut être étendue au krigeage trans-gaussien, ce qui permet d'utiliser des jeux de données qui violent les hypothèses du krigeage. La méthode se révèle à la fois élégante et efficace. Nous proposons une application aux simulations numériques dans le domaine de la sûreté nucléaire, domaine dans lequel il est nécessaire de modéliser les tests non destructifs fondés sur les courants de Foucault visant à détecter de possibles usures dans les tubes de générateurs de vapeur.

7.1 Introduction

Non-destructive testing (NDT) is a group of techniques used in industry to evaluate the properties of components or systems without destroying them. Numerical simulations can be used to calibrate NDT techniques and determine adequate threshold levels for defect detection. These simulations can be computationally expensive. As a result, we may want to replace them by a cheaper surrogate model.

The NDT problem that motivates this chapter is the following one. In order to improve nuclear safety, Électricité de France (EDF) has developed a Non-Destructive Evaluation for testing the presence of a defect in Steam Generator Tubes. In nuclear power plants, steam generators

are the interface between primary and secondary water circuits. High-pressured water from the primary circuit flows into the steam generator tubes. To prevent the tubes from vibrating under this solicitation, they are held in place by anti-vibration bars (AVB). Rubbing may in time leave defects in the tubes, however. To detect them, a SAX (axial) probe is moved down the tube. This Non-Destructive Exam was modeled in C3D, a computer code which can accurately represent any possible defect geometry [Maurice et al., 2013]. As this code uses the finite element method, a wide range of parameters may be taken into account. Furthermore, any degree of accuracy can be reached as long as the mesh is fine enough [Thomas et al., 2015]. Non-destructive inspections based on eddy currents exploit the way the induction flux changes as the probe approaches a defect. If the tube were a perfect cylinder, the coils of the probe would get the same flux by cylindrical symmetry. A large enough difference between both fluxes therefore signals a defect. This differential flux is a complex quantity, but in the presented application only its imaginary part is used for defect detection, or rather the difference between maximum and minimum imaginary part as the probe moves through the tube. It is then converted to a tension by Lenz's law.

Following both expert reports and data simulations, eight geometrical parameters (see Figure 7.1) and one non-geometrical parameter were identified as having the greatest influence on the output of C3D. In order to be able to define POD curves, they were given probability distributions reflecting expert opinion.

1. $a \sim U[a_a, b_a]$: defect depth (mm);
2. $E \sim N(a_e, b_e)$: pipe thickness (mm) based on measurements of 5000 pipes;
3. $h_1 \sim U[a_{h_1}, b_{h_1}]$: distance between the AVB and the top of the defect (mm);
4. $h_2 \sim U[a_{h_2}, b_{h_2}]$: distance between the AVB and the bottom of the defect (mm);
5. $e_{BAV1} \sim U[-a + a_{e_{BAV1}}, b_{e_{BAV1}}]$: length of the gap between the AVB on the side of the defect and the tube (mm);
6. $e_{BAV2} \sim U[a_{e_{BAV2}}, b_{e_{BAV2}}]$: length of the gap between the AVB on the other side and the tube (mm);
7. $h_{BAV} \sim U[a_{h_{BAV}}, b_{H_{BAV}}]$: shift between both AVBs (mm);
8. $inc \sim \mathcal{N}_{trunc,0}(inc_a, inc_b)$: inclination of the AVB on the side of the defect.
   $\mathcal{N}_{trunc,0}(inc_a, inc_b)$ denotes a Normal distribution with mean inc_a and variance inc_b truncated at 0: its support is $[0, +\infty)$ (mm);
9. $cond \sim \mathcal{N}(cond_a, cond_b)$: conductivity of the tube (MS/m).

The parameter of interest being a, all other parameters are collectively denoted by the eight-dimensional vector $\boldsymbol{x} = (E, h_1, h_2, e_{BAV1}, e_{BAV2}, h_{BAV}, inc, cond)^\top$. $\boldsymbol{X}$ is the corresponding random variable: its joint distribution is the product of the distributions described above. For ease of manipulation, $\boldsymbol{X}$ is reparametrized in such a way as to follow the Uniform distribution on the 8-dimensional cube $(0,1)^8$. We also reparametrize a in the same way. The whole input space becomes the unit cube $(0,1)^r$ with $r = 9$.

Our ultimate goal is to compute the Probability Of Detection (POD) of a defect as a function of its depth a. We formalize this notion in Section 7.2. Practical computation of the POD requires the use of surrogate models. Section 7.3 provides the mathematical framework for dealing with surrogate model uncertainty. Finally, Section 7.4 presents the resulting POD curves in the case of steam generator defect detection.

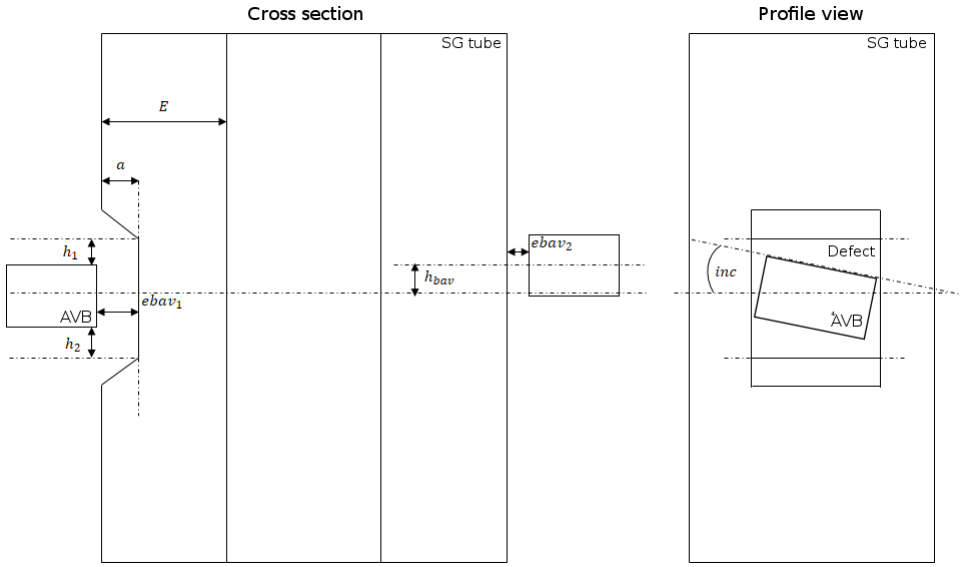


Figure 7.1 – *Geometrical parameters of the computer model C3D. Left: Profile view of a steam generator tube where the left AVB caused wear. Right: View of the side with the defect.*

7.2 Probability Of Detection (POD)

Industrial practice often uses Functional Risk Curves (FRC) as a means of expressing the "probability" of some (un)desirable event based on the value of some critical parameter. Probability Of Detection (POD) curves are a particular kind of FRC. They represent the probability that a given testing procedure has of detecting a defect as a function of some parameter of interest. Two factors can justify the probabilistic framework:

1. the testing procedure may incorporate some randomness;
2. the result may depend not only on the parameter of interest but also on unknown nuisance parameters.

We denote by a the parameter of interest and by $\boldsymbol{x}$ the nuisance parameters. The set of all values possibly taken by $\boldsymbol{x}$ is endowed with a probability distribution which should reflect their frequency in real life. Let $\boldsymbol{X}$ be a random variable following this probability distribution. For given a and $\boldsymbol{x}$, $z(a, \boldsymbol{x})$ denotes the output of the testing procedure. Depending on the specific testing procedure used, it may or may not be random. In any case, it falls to the modeler to specify the probability distribution of $\boldsymbol{X}$ and/or the probability distributions of $z(a, \boldsymbol{x})$ for all a and $\boldsymbol{x}$. In this chapter, we only deal with deterministic testing procedures so POD curves are entirely determined by the distribution of $\boldsymbol{X}$.

Typically, the output z is a scalar quantity. When this quantity lies beyond a predetermined threshold s, it signals the presence of a defect, so the POD curve is the function defined by

$$POD(a) = \mathbb{P}(z(a, \boldsymbol{X}) > s). \tag{7.1}$$

Being a deterministic mapping does not make z easy to determine, however. Computing the POD curve would in theory require conducting physical tests on a wide sample of objects with a wide variety of defects – i.e. a wide variety of parameters a and $\boldsymbol{x}$. Because of the

prohibitive costs involved, phenomenological numerical simulations are used to simulate the testing procedure. Throughout this chapter, such simulations – collectively called a "computer experiment" – are assumed to be perfectly accurate.

Unfortunately, even these numerical simulations are too costly to allow a Monte-Carlo approach. Therefore, $\boldsymbol{z}$ needs to be approximated by some less costly model. In order to control the ensuing error, statistical surrogate models are used – linear regression or polynomial chaos for example. In this chapter, we focus on Kriging, otherwise known as Gaussian Process regression, and on Trans-Gaussian Kriging, where a transformation step is performed on the output space before Kriging is performed. When a surrogate model is used, z must be replaced by Z, a random mapping whose probability distribution represents the uncertainty about z.

Surrogate models introduce some ambiguity in the definition of the POD curve. Should it be approximated by $POD(a) \approx \mathbb{P}(Z(a, \boldsymbol{X}) > s)$? Although natural, this approximation conflates the uncertainty about $\boldsymbol{X}$ and about Z, although the two are different in nature. The distribution of $\boldsymbol{X}$ is understood in *frequentist* terms: over the year, an experimenter conducting tests on multiple pieces of equipment will encounter varied values of $\boldsymbol{X}$. The distribution of Z represents *epistemic* uncertainty, which is irrelevant to the actual tests conducted in real life and could be mitigated if more computational resources were available. This point is illustrated in the description of a concrete application below.

Kriging and Trans-Gaussian Kriging surrogate models depend on parameters, which will henceforth be named "hyperparameters" to differentiate them from the "physical" parameters a and $\boldsymbol{X}$. Hyperparameters are tricky, because they have tremendous influence so careful surrogate model calibration is normally required. The Bayesian paradigm, being a complete inferential approach (Robert et al. [2011] provides a review of its range from a practical standpoint) can be used to avoid this step. In this chapter, we propose eliminating hyperparameters from the model by means of Bayesian averaging out. Prior elicitation is done with the help of Bernardo's reference prior theory [Bernardo, 2005].

7.3 An Objective Bayesian outlook to Trans-Gaussian Kriging

Kriging Likelihood

For the sake of simplicity, we assume that the parameter of interest a belongs to $[0, 1]$ and that $\boldsymbol{x}$, which represents all nuisance parameters belongs to $[0, 1]^{r-1}$ for some positive integer r. The scalar output of interest is $y(a, \boldsymbol{x})$.
Kriging is a very flexible surrogate model for computer experiments [Santner et al., 2003]. To use it, the computer experiment must first be conducted for $n \in \mathbb{N}$ values $(a^{(i)}, \boldsymbol{x}^{(i)})$ (called observations points) of the parameters. The set of observation points $\left((a^{(i)}, \boldsymbol{x}^{(i)})\right)_{i \in [\![1,n]\!]}$ is called the design set. The vector of outputs $\boldsymbol{y} = (y(a^{(1)}, \boldsymbol{x}^{(1)}), ..., y(a^{(n)}, \boldsymbol{x}^{(n)}))^\top \in \mathbb{R}^n$ is called the observation vector.

Kriging models the uncertainty about y by defining Y as a Gaussian process subjected to the condition that for every integer $i \in [\![1, n]\!]$, $Y(a, \boldsymbol{x}_i) = y(a, \boldsymbol{x}_i)$. Because this conditioning is critical, we need a way to differentiate the conditioned from the unconditioned process, so let $\mathcal{Y}$ be the unconditioned Gaussian process.

In the literature, $\mathcal{Y}$ is often assumed to be stationary. That is, the distribution of the unconditioned Gaussian Process (before knowing the observed values) should be invariant by translation. This simplifying assumption is often reasonable, because should the emulated function be non-stationary, the distribution of the Gaussian Process conditional to the observations would reflect this non-stationarity, provided the number of observations is sufficient. Moreover, assuming non-stationarity would require some sort of prior knowledge of the kind of non-stationarity that is expected. For example, warped Gaussian processes [Snelson et al., 2004] use the knowledge of the presence of a discontinuity.

In some contexts, an additional assumption can be made: that the distribution of the Gaussian Process before observing the data is isotropic. When it is possible to make this assumption, the correlation kernel of the Gaussian Process is usually parametrized by two positive hyperparameters: variance σ^2 and correlation length θ. Seeing the Gaussian Process as a response surface, one may think of $\sigma := \sqrt{\sigma^2}$ as the scale of variation of the output and of θ as the scale of variation of the input.

While isotropy is a natural assumption in geostatistics – the original application of Kriging [Matheron, 1960] – it is rarely adequate when dealing with computer experiments. Each of the r dimensions in the input space $[0,1]^r$ corresponds to one parameter, and the parameters can be heterogeneous. In such contexts, a correlation length θ_i is necessary for every parameter ($i \in [\![1, r]\!]$). The covariance kernel is then anisotropic and we denote by $\boldsymbol{\theta}$ the vector $(\theta_i)_{i=1}^r$. The best kind of anisotropic kernel for interpretation is anisotropic geometric, but tensorized kernels are often used for simplicity [Stein, 1999, page 54].

The easiest way to introduce non-stationarity is through the addition of some non-constant deterministic function. In the Universal Kriging framework, this function is assumed to belong to a given small-dimensional vectorial space $\mathcal{F}_p$. One assumes therefore that there exists in $\mathcal{F}_p$ a function f that adequately approximates the deterministic function one seeks to emulate, and the stationary Gaussian process then merely models our perception of the error made when using such an approximation.

Let $(f_j)_{j=1}^p$ be a basis of $\mathcal{F}_p$ and let $\boldsymbol{\beta} = (\beta_j)_{j=1}^p \in \mathbb{R}^p$ be the vector such that $f = \sum_{j=1}^p \beta_j f_j$. Naturally, p should be smaller than the number of observation points n or the model would not be identifiable. Denote by $\boldsymbol{H}$ the $n \times p$ matrix whose (i,j)-th element is $f_j\left(\boldsymbol{x}^{(i)}\right)$. Again for the sake of identifiability, assume that $\boldsymbol{H}$ is of full rank. Let then $\boldsymbol{y}$ be the vector of length n whose i-th element is the observation of the Gaussian process made at $\boldsymbol{x}^{(i)}$. It is a Gaussian vector with mean vector $\boldsymbol{H\beta}$. Let $\boldsymbol{\Sigma_\theta}$ be its correlation matrix.
The likelihood of the parameters $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ when observing $\boldsymbol{y}$ is then:

$$L(\boldsymbol{y} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma_\theta}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{y} - \boldsymbol{H\beta}\right)^\top \boldsymbol{\Sigma_\theta}^{-1}\left(\boldsymbol{y} - \boldsymbol{H\beta}\right)\right\} . \tag{7.2}$$

Transformation

If even more flexibility is needed, like in some cases arising naturally from examples in the natural sciences – see the example below – one can relax the stationarity assumption and even the assumption that the random process is Gaussian through Trans-Gaussian Kriging [De Oliveira et al., 1997]. The idea is to assume that a random field $\mathcal{Z}(a, \boldsymbol{x})$ *would be* Gaussian

and stationary if the output space O (which is assumed to be an interval of $\mathbb{R}$) were transformed in an adequate fashion. Practically speaking, one must choose a parametric family of nondecreasing differentiable transformations $g_\alpha : O \to \mathbb{R}$. For some α, $\mathcal{Y} := g_\alpha(\mathcal{Z})$ is a stationary Gaussian random field.

The detection of defects on Steam Generator tubes is most importantly impacted by the depth of the defect. As a first approximation, one may say that the greater the defect depth, the greater the chances of detecting it.

Figure 7.2 (left) illustrates the importance of the length of the defect a. It represents the measured tensions for 100 defective tubes with various defect depths. The depths are normalized so that 1 represents the thickness of a tube coating. Using the parametrization presented at the end of the introduction, the 100 points form a space-filling design set for the 9-dimensional cube $(0,1)^9$.

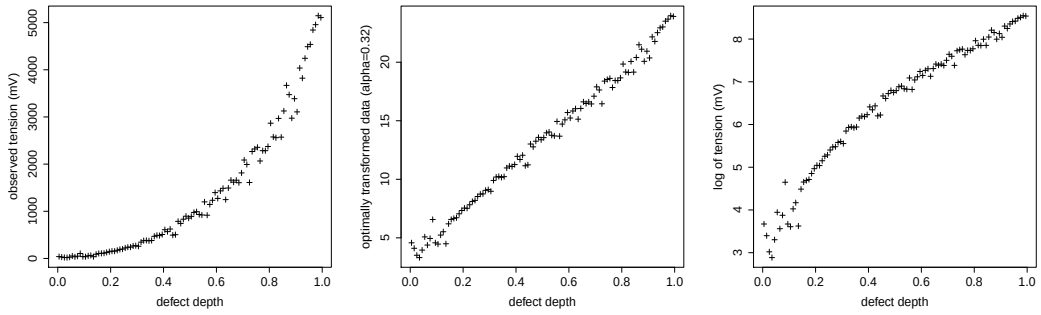


Figure 7.2 – *Left: Measured tension as a mapping of defect depth for 100 defective tubes. The other "nuisance" geometrical parameters are representative of possible geometries. Right: Same data after applying the logarithm function (B_α or C_α with $\alpha = 0$) to all measured tensions. Center: Same data after applying the optimal transformation (C_α with $\alpha = 0.32$) to all measured tensions.*

Because the measured tension is necessarily nonnegative, the Gaussian assumption of the Kriging surrogate model is inadequate. Moreover, the spread becomes greater and greater as the defect depth increases, which contradicts the assumption of stationarity. Because Figure 7.2 (left) gives the impression of being based on the graph of some exponential mapping, our first instinct was to apply the logarithm to the observations. The result is given by Figure 7.2 (right).

Although this transformation seems fruitful, it is clearly too strong for our purpose. While the original data yielded what looked like a strongly convex mapping of defect depth, the log-transformed data yield a somewhat concave mapping of defect depth. Moreover, the spread which was originally increasing with depth seems now to be decreasing with depth. Some intermediate transformation between the identity mapping and the logarithm mapping was needed. A possible choice could have been the Box-Cox power transform family:

$$B_\alpha(t) = \begin{cases} \frac{t^\alpha - 1}{\alpha} & \text{if } \alpha > 0 \\ \log(t) & \text{if } \alpha = 0 \end{cases} \tag{7.3}$$

The Box-Cox power transformation fits our requirements since all mappings B_α for $\alpha \in (0,1)$ are intermediate transformations between the logarithm ($\alpha = 0$) and the identity mapping ($\alpha = 1$, although the data are uniformly decremented by 1, which is of no consequence).

However, for any $\alpha > 0$, B_α is a bijection from $(0, +\infty)$ to $(-1/\alpha, +\infty)$, whereas we would like a bijection from $(0, +\infty)$ to $(-\infty, +\infty)$. How can the Gaussian assumption be credible otherwise? In the Box-Cox family, only the logarithm mapping ($\alpha = 0$) fits this requirement.

The following alteration to the Box-Cox family is suitable:

$$C_\alpha(t) = \begin{cases} \frac{1}{\alpha}\sinh(\alpha \log(t)) & \text{if } \alpha > 0 \\ \log(t) & \text{if } \alpha = 0 \end{cases} \tag{7.4}$$

Every mapping C_α is a bijection from $(0, +\infty)$ to $(-\infty, +\infty)$. Moreover, C_1 is equivalent to the linear mapping $t \mapsto 0.5t$ when $t \to \infty$, so this family too can be considered to contain intermediate mappings between the logarithm ($\alpha = 0$) and the identity mapping ($\alpha = 1$). Figure 7.3 illustrates both Box-Cox and alternative transformation families.

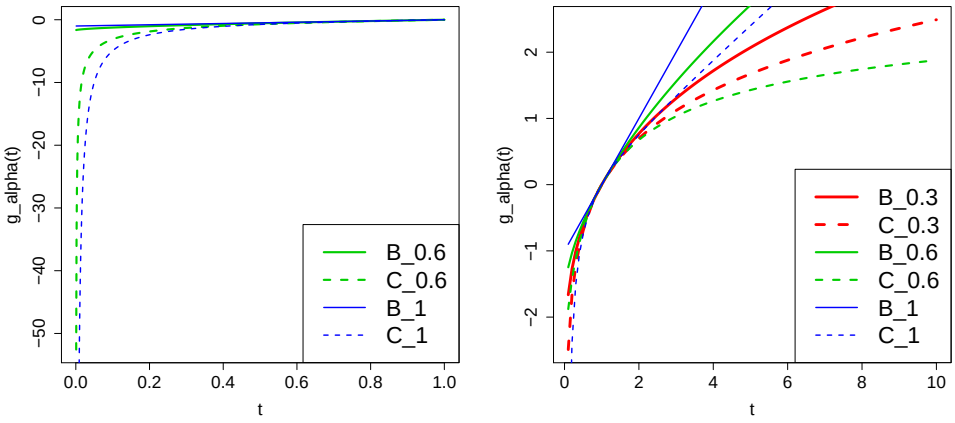


Figure 7.3 – *Box-Cox and alternative transformation families for small values of α (left) and greater values of α (right).*

We use this transformation family (i.e. $g_\alpha = C_\alpha$ for all $\alpha \in [0, +\infty)$) to apply the Bayesian framework for Trans-Gaussian Kriging described in the previous section. Our mean function space $\mathcal{F}_2$ is the space of affine functions of the parameter of interest a. This choice reflects the fact that a is the main influence on the output of the computer code. Our basis functions are the constant function of value 1 and the coordinate function $(a, \boldsymbol{x}) \mapsto a$. Our correlation kernel is the Matérn anisotropic geometric kernel of smoothness $\nu = 5/2$. The transformation family is C_α, $\alpha \in [0, +\infty)$.

Trans-Gaussian Kriging Likelihood

In Trans-Gaussian Kriging, Z is a random field whose distribution is that of $\mathcal{Z}$ conditioned by $\mathcal{Z}(a^{(i)}, \boldsymbol{x}^{(i)}) = z(a^{(i)}, \boldsymbol{x}^{(i)})$ for every $i \in [\![1, n]\!]$, and $\boldsymbol{z}$ is the vector $\left(z(a^{(1)}, \boldsymbol{x}^{(1)}), ..., z(a^{(n)}, \boldsymbol{x}^{(n)})\right)^\top$.

Denoting by $g_\alpha(\boldsymbol{z})$ the vector $(g_\alpha(z_1), ..., g_\alpha(z_n))^\top$, the likelihood of hyperparameters $\boldsymbol{\beta}$, σ^2, $\boldsymbol{\theta}$ and α is:

$$\begin{aligned} & L(\boldsymbol{z} \mid \boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta}, \alpha) \\ &= \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} |\boldsymbol{\Sigma}_{\boldsymbol{\theta}}|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2\sigma^2}\left(g_\alpha(\boldsymbol{z}) - \boldsymbol{H}\boldsymbol{\beta}\right)^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}}^{-1}\left(g_\alpha(\boldsymbol{z}) - \boldsymbol{H}\boldsymbol{\beta}\right)\right\} \prod_{i=1}^{n} g'_\alpha(z_i). \end{aligned} \tag{7.5}$$

With Trans-Gaussian Kriging, the "Gaussian" parameters $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ can no longer be interpreted as pertaining to mean function, variance and correlation respectively. They can only be interpreted in relation to the transformation parameter α.

[Box and Cox, 1964] propose estimating the "Gaussian" parameters conditionally to the transformation parameter, and to treat the value of the transformation parameter with maximum likelihood as the true value. This point of view was criticized by Bickel and Doksum [1981], who showed that when using the Box-Cox transformation family, the estimators for the transformation parameter and for $\boldsymbol{\beta}$ are highly correlated, and they argued for taking this effect into account when performing inference on the parameters. Unfortunately, when the computational budget is too low to be able to make many observations, the Kriging likelihood can be rather flat even without adding a transformation parameter into the mix [Li and Sudjianto, 2005].

These difficulties are the reason why we propose in this chapter using an Objective Bayesian framework to integrate $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ out of the model – conditionally to α, in keeping with Box and Cox [1964]'s insight. This way, we only need to estimate the parameter α, with the likely consequence that the "integrated" one-dimensional likelihood is not as flat as the regular multi-dimensional likelihood.

Objective Bayesian treatment of the Gaussian parameters

In this subsection, we operate again under the framework of Subsection 7.3. This framework is the standard Kriging setting, which is parametrized by $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$. The likelihood is given by Equation (7.2).

Following Berger et al. [2001], we make use of the reference prior on the parameters $(\boldsymbol{\beta}, \sigma^2)$ conditional to $\boldsymbol{\theta}$: $\pi_R(\boldsymbol{\beta}, \sigma^2|\boldsymbol{\theta}) \propto 1/\sigma^2$. The integrated likelihood $L^1(\boldsymbol{y}|\boldsymbol{\theta})$ is given by successively integrating $\boldsymbol{\beta}$ and σ^2 out. To do this, we introduce $\boldsymbol{W}$ as an $n \times (n-p)$ matrix such that $\boldsymbol{W}^\top \boldsymbol{H}$ is the null $(n-p) \times p$ matrix and $\boldsymbol{W}^\top \boldsymbol{W}$ is the $(n-p) \times (n-p)$ identity matrix. $\boldsymbol{W}$ can be computed by performing Standard Value Decomposition (SVD) on $\boldsymbol{H}$. The singular vectors corresponding to the null singular values form the columns of $\boldsymbol{W}$. We find

$$\begin{aligned} L^1(\boldsymbol{y}|\boldsymbol{\theta}) &= \int L(\boldsymbol{y}|\boldsymbol{\beta}, \sigma^2, \boldsymbol{\theta})/\sigma^2 d\boldsymbol{\beta} d\sigma^2 \\ &= \left(\frac{2\pi^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-p}{2}\right)}\right)^{-1} |\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}|^{-\frac{1}{2}} \left(\boldsymbol{y}^\top \boldsymbol{W}\left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}\right)^{-\frac{n-p}{2}}. \end{aligned} \tag{7.6}$$

We briefly explain how prediction can be performed in our Bayesian Kriging model.

First, assume that all hyperparameters – $\boldsymbol{\beta}$, σ^2 and $\boldsymbol{\theta}$ – are known. We wish to predict the value taken by the Gaussian process Y at an unobserved point $(a_0, \boldsymbol{x}_0)$.

Let $\boldsymbol{P}$ be the $n \times p$ matrix whose columns are obtained by applying the Gram-Schmidt process to the columns of $\boldsymbol{H}$.

Denote by $\boldsymbol{H}_{0,0}$ the transpose of the vector of length p containing the values of the p basis functions at$(a_0, \boldsymbol{x}_0)$, by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}$ the $1 \times n$ correlation matrix between the Gaussian process at $(a_0, \boldsymbol{x}_0)$ and the Gaussian process at the design set, and finally by $\boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0}$ its transpose.

Define the $1 \times p$ matrix $\boldsymbol{H}_{0,\cdot} = \boldsymbol{H}_{0,0} \left(\boldsymbol{P}^\top \boldsymbol{H}\right)^{-1}$ and its transpose $\boldsymbol{H}_{\cdot,0}$. The following matrix definitions are necessary to express the predictive distribution:

$$
\begin{aligned}
E_0 &:= \boldsymbol{H}_{0,\cdot} \boldsymbol{P}^\top \boldsymbol{y} \\
S_{\boldsymbol{\theta},0,0} &:= 1 + \boldsymbol{H}_{0,\cdot} \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{P} \boldsymbol{H}_{\cdot,0} - \boldsymbol{H}_{0,\cdot} \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta},\cdot,0} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot} \boldsymbol{P} \boldsymbol{H}_{\cdot,0} \\
\boldsymbol{S}_{\boldsymbol{\theta},0,\cdot} &:= \left(\boldsymbol{H}_{0,\cdot} \boldsymbol{P}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} - \boldsymbol{\Sigma}_{\boldsymbol{\theta},0,\cdot}\right) (\boldsymbol{P}\boldsymbol{W}) \\
\boldsymbol{S}_{\boldsymbol{\theta},\cdot,0} &:= \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot}^\top
\end{aligned}
$$

We recall here Corollary 6.3:

Proposition 7.1. *The predictive distribution of $Y(a_0, \boldsymbol{x}_0)$ averaged over both $\boldsymbol{\beta}$ and σ^2 is the non-standardized Student's t-distribution with $n-p$ degrees of freedom, location parameter $E_0 - \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}$ and scale parameter*

$$
\sqrt{\frac{\boldsymbol{y}^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{y}}{n-p} \left\{S_{\boldsymbol{\theta},0,0} - \boldsymbol{S}_{\boldsymbol{\theta},0,\cdot} \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top \boldsymbol{S}_{\boldsymbol{\theta},\cdot,0}\right\}}.
$$

Practically speaking, if $n - p$ is large the Student t-distribution can be approximated by a Normal distribution with mean equal to the location parameter and standard deviation equal to the scale parameter.

Further integrating the predictive distribution requires averaging over the Gibbs reference posterior distribution on $\boldsymbol{\theta}$. The Gibbs reference posterior distribution is derived through Objective Bayesian techniques and has nice theoretical as well as practical properties, like invariance by reparametrization and good frequentist performance for prediction intervals, as was seen in Sections 5.4 and 6.5. This integration can be done numerically, by using a sample from the Gibbs reference posterior that can be obtained easily from a Gibbs sampler.

The Likelihood problem

Let us go back to Trans-Gaussian Kriging. In this framework, it is desirable to integrate $\boldsymbol{\theta}$ out of the model in order to derive an integrated likelihood of the transformation parameter α only.

The problem with the Gibbs reference posterior approach is that, given there is no actual prior distribution yielding the Gibbs reference posterior, it is not possible to integrate $\boldsymbol{\theta}$ out

of the model with likelihood $L^1(\boldsymbol{y}|\boldsymbol{\theta})$. This is a major hurdle since the actual model (taking into account that $\boldsymbol{y}$ is the result of a transformation) is:

$$\begin{aligned} & L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha) \\ &= \left(\frac{2\pi^{\frac{n-p}{2}}}{\Gamma\left(\frac{n-p}{2}\right)}\right)^{-1} |\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}|^{-\frac{1}{2}} \left(g_\alpha(\boldsymbol{z})^\top \boldsymbol{W} \left(\boldsymbol{W}^\top \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{W}\right)^{-1} \boldsymbol{W}^\top g_\alpha(\boldsymbol{z})\right)^{-\frac{n-p}{2}} \prod_{i=1}^n g'_\alpha(z_i). \end{aligned} \tag{7.7}$$

Because we cannot integrate $\boldsymbol{\theta}$ out, we need to compute some other aggregate of all possible $L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)$ when $\boldsymbol{\theta}$ varies. The most obvious solution is to average $L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)$ over the Gibbs reference posterior distribution $\pi_G(\boldsymbol{\theta}|\boldsymbol{z},\alpha) := \pi_G(\boldsymbol{\theta}|g_\alpha(\boldsymbol{z}))$, but this solution gives too much weight to the likelihood: if a prior $\pi_G(\boldsymbol{\theta}|\alpha)$ existed, then this would be equivalent to computing $\int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)^2 \pi_G(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta} / \int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi_G(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta}$.

A second possibility is to compute the Maximum A Posteriori estimator $\widehat{\boldsymbol{\theta}}_{MAP}$ for $\boldsymbol{\theta}$ and then proceed with $L^{MAP}(\boldsymbol{z}|\alpha) := L^1(\boldsymbol{z}|\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}_{MAP},\alpha)$. The idea is that $\widehat{\boldsymbol{\theta}}_{MAP}$ should be a "typical" value of $\boldsymbol{\theta}$, but this approach unfortunately makes the aggregate dependent on the parametrization chosen for the correlation lengths, which is a bad property for a quantity treated as a likelihood.

A third possibility is

$$L^{LOG}(\boldsymbol{z}|\alpha) := \exp\left[\int \log\{L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\}\pi_G(\boldsymbol{\theta}|\boldsymbol{z},\alpha) d\boldsymbol{\theta}\right] \tag{7.8}$$

Compared to L^{MAP}, L^{LOG} has the advantage that it does not depend on the parametrization, insofar as $\boldsymbol{\theta}$ could be replaced by any vector $(h_1(\theta_1), ..., h_r(\theta_r))^\top$ as long as all h_i ($i \in [\![1, r]\!]$) are bijections. Moreover, it does not rely on a particular estimate of the "true" $\boldsymbol{\theta}$, which makes L^{LOG} more robust that L^{MAP}.

Fundamentally though, L^{MAP} and L^{LOG} are justified by a simple heuristic. This heuristic is based on two approximations.

1. Asymptotic: the Gibbs reference posterior distribution is approximated by the probability distribution $\mathcal{N}(\widehat{\boldsymbol{\theta}}_{MAP}; \mathcal{I}(\widehat{\boldsymbol{\theta}}_{MAP})^{-1})$ where $\mathcal{I}(\boldsymbol{\theta})$ is the Fisher information matrix. This means that we assume that the conclusion of ,the Bernstein - von Misès theorem applies as though we were in an asymptotic framework.
2. Jeffreys: the Gibbs reference posterior is the posterior distribution of $\boldsymbol{\theta}$ corresponding to the Jeffreys-rule prior on $\boldsymbol{\theta}$ denoted by $\pi(\boldsymbol{\theta}|\alpha)$. In order to have simple notations, we do not normalize it – it is possibly not proper anyway – so, denoting by $|\cdot|$ the matrix determinant, $\pi(\boldsymbol{\theta}|\alpha) = |\mathcal{I}(\boldsymbol{\theta})|^{1/2}$.

Define $K_{\boldsymbol{z}}(\alpha) \in \mathbb{R}$ such that $K_{\boldsymbol{z}}(\alpha) \int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha) d\boldsymbol{\theta} = \prod_{i=1}^n g'_\alpha(z_i)$. This is equivalent to asserting that $\pi_G(\boldsymbol{\theta}|\boldsymbol{z},\alpha) \prod_{i=1}^n g'_\alpha(z_i) = K_{\boldsymbol{z}}(\alpha) L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha)$. In the following, we show that the two approximations imply that $\prod_{i=1}^n g'_\alpha(z_i)/K_{\boldsymbol{z}}(\alpha)$ is related to both $L^{MAP}(\boldsymbol{z}|\alpha)$ and $L^{LOG}(\boldsymbol{z}|\alpha)$.

First, we have

$$\begin{aligned} & K_{\boldsymbol{z}}(\alpha)L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha) \\ =&(2\pi)^{-r/2}|\mathcal{I}(\boldsymbol{\theta})|^{1/2}\exp\left\{-\frac{(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{MAP})^\top\mathcal{I}(\widehat{\boldsymbol{\theta}}_{MAP})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{MAP})}{2}\right\}\prod_{i=1}^n g'_\alpha(z_i). \end{aligned} \tag{7.9}$$

In particular,

$$K_{\boldsymbol{z}}(\alpha)L^1(\boldsymbol{z}|\widehat{\boldsymbol{\theta}}_{MAP},\alpha) = (2\pi)^{-r/2}\prod_{i=1}^n g'_\alpha(z_i). \tag{7.10}$$

So the integrated likelihood $\int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}$ is proportional to $L^{MAP}(\boldsymbol{z}|\alpha)$:

$$\int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta} = (2\pi)^{r/2}L^{MAP}(\boldsymbol{z}|\alpha). \tag{7.11}$$

Furthermore,

$$\int \log L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi_G(\boldsymbol{\theta}|\boldsymbol{z},\alpha)d\boldsymbol{\theta} = \int \log\{K_{\boldsymbol{z}}(\alpha)L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\}\pi_G(\boldsymbol{\theta}|\boldsymbol{z},\alpha)d\boldsymbol{\theta} - \log K_{\boldsymbol{z}}(\alpha) \tag{7.12}$$

Notice that

$$\log\{K_{\boldsymbol{z}}(\alpha)L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\} = -\frac{r}{2}\log(2\pi) - \frac{(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{MAP})^\top\mathcal{I}(\widehat{\boldsymbol{\theta}}_{MAP})(\boldsymbol{\theta}-\widehat{\boldsymbol{\theta}}_{MAP})}{2} + \sum_{i=1}^n \log g'_\alpha(z_i), \tag{7.13}$$

so

$$\int \log L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi_G(\boldsymbol{\theta}|\boldsymbol{z},\alpha)d\boldsymbol{\theta} = -\frac{r}{2}\log(2\pi) - \frac{r}{2} + \sum_{i=1}^n \log g'_\alpha(z_i) - \log K_{\boldsymbol{z}}(\alpha). \tag{7.14}$$

Finally, we see that the integrated likelihood $\int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}$ is also proportional to $L^{LOG}(\boldsymbol{z}|\alpha)$:

$$\log\left\{\int L^1(\boldsymbol{z}|\boldsymbol{\theta},\alpha)\pi(\boldsymbol{\theta}|\alpha)d\boldsymbol{\theta}\right\} = \log L^{LOG}(\boldsymbol{z}|\alpha) + \frac{r}{2}\log(2\pi) + \frac{r}{2}. \tag{7.15}$$

Interestingly, this heuristic also predicts that

$$\log\{L^{MAP}(\boldsymbol{z}|\alpha)\} - \log\{L^{LOG}(\boldsymbol{z}|\alpha)\} = r/2. \tag{7.16}$$

This prediction provides a sanity check for the heuristic, which we use in the application below.

7.4 Industrial Application

Integrating out Kriging hyperparameters

Figure 7.2 shows that α can only reasonably belong to $[0,1]$, so we endow $[0,1]$ with the fine grid $0.01 * [\![0,100]\!]$. For every element α in this grid, we sample 100 points $\boldsymbol{\theta}_\alpha^{(j)} \in (0,+\infty)^r$ from the Gibbs reference posterior distribution. This is done by performing 9000 iterations of the Gibbs algorithm and keeping only one out of 90.

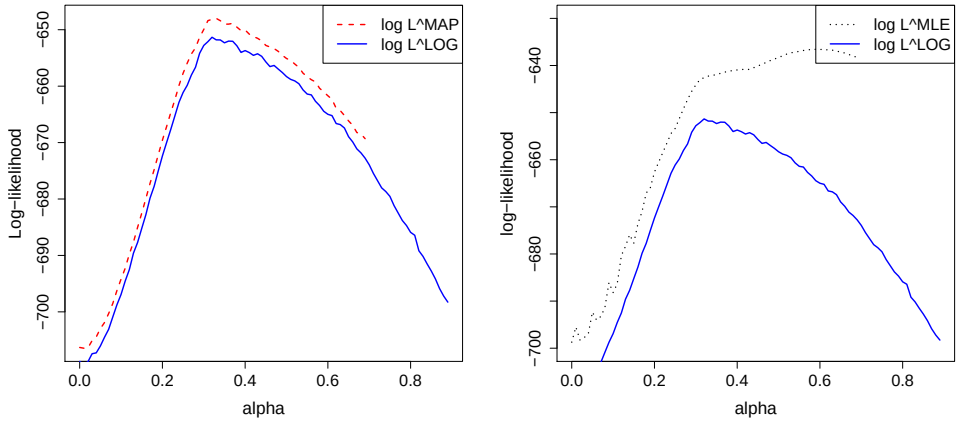


Figure 7.4 – *Left: Logarithm of $L^{MAP}(\boldsymbol{z}|\alpha)$ (red dotted line) and $L^{LOG}(\boldsymbol{z}|\alpha)$ (blue solid line). L^{MAP} is not represented for greater values of α because the MAP cannot be reliably computed. Right: Logarithm of $L^{LOG}(\boldsymbol{z}|\alpha)$ (blue solid line) and log-likelihood $L^{MLE}(\boldsymbol{z}|\alpha)$ if the MLE estimator for $\boldsymbol{\theta}$ is used (black dotted line). L^{MAP} and L^{MLE} are not represented for greater values of α, because the MAP cannot be reliably computed and the MLE favors correlation lengths so high that correlation matrices are ill-conditioned.*

Figure 7.4 (left) gives the logarithm of the pseudo-likelihoods L^{MAP} and L^{LOG}.

L^{MAP} and L^{LOG} reach their maxima at $\alpha = 0.34$ and $\alpha = 0.32$ respectively. Interestingly, the average value of $L^{MAP}(\boldsymbol{z}|\alpha) - L^{LOG}(\boldsymbol{z}|\alpha)$ for $\alpha \in [0.25, 0.55]$ is 3.4 and its standard deviation 0.3. This is quite close to the prediction (7.16) that this difference should be constant, especially considering that the values of both functions spread over an interval of length greater than 50. However, the value 3.4 is substantially lower than the predicted $r/2 = 4.5$, which suggests that not all 9 parameters are influent.

On account of L^{LOG} not depending on the estimate of the MAP, we accept the value of α maximizing L^{LOG}, $\alpha = 0.32$ as the "true" alpha. In this respect, we follow the suggestion of Box and Cox [1964]. Naturally, a completely Bayesian treatment would require defining a prior distribution on α in order to integrate it out. In this case however, both versions of the likelihood are so pronounced that any reasonable prior would have little effect on the posterior distribution. In the interest of completeness though, we do offer in the next section a comparison between the prediction of the Maximum Likelihood approach setting $\alpha = 0.32$ and a "full-Bayesian" approach using the Uniform prior distribution on $[0, 1]$.

The question of whether or not to use a Bayesian approach to deal with α instead of the MLE may be debatable, but it is not when dealing with $\boldsymbol{\theta}$: the Bayesian approach is clearly superior. Figure 7.4 (right) shows the log-likelihood of α when taking $\boldsymbol{\theta}$ to be equal to its MLE (dotted curve). It favors high values of α, which means the data are weakly transformed so the MLE on $\boldsymbol{\theta}$ favors very high correlation lengths and correlation matrices become ill-conditioned.

Let us consider again the observation data presented in Figure 7.2 (left). If we apply the transformation C_α with $\alpha = 0.32$, we obtain Figure 7.2 (center). Notice that the observations seem to be placed along a straight line. This is no coincidence: it reflects our choice of mean function space $\mathcal{F}_2$. The optimal transformation parameter is in first approximation the one that makes the data match our assumption about the mean function.

To confirm this, consider a surrogate model of transformed linear regression. It can be seen as a Trans-Gaussian Kriging model with null correlation lengths. We provide the log-likelihood of the transformation parameter for such a model in Figure 7.4 (right). It reaches its maximum at $\alpha = 0.30$, very near $\alpha = 0.32$ where L^{LOG} reaches its maximum. This shows that the correlation structure parametrized by $\boldsymbol{\theta}$ has little impact on the maximum likelihood of α. $\boldsymbol{\beta}$ and σ^2 having been integrated out of the model, the choice of mean function space $\mathcal{F}_p$ is necessarily the primary explanation for the the likelihood function of α reaching its maximum where it does.

Probability Of defect Detection

With this machinery, we can now return to the question of POD curves. For any point in the $r = 9$-dimensional input space of C3D, the Trans-Gaussian Kriging surrogate model can provide a probability that the detection threshold $s = 200mV$ is crossed. We need a few definitions.

Definition 7.2. *Let $a \in [0, 1]$. The* actual POD*, denoted by $POD(a)$, is the probability for a defect of depth a to be detected: $POD(a) = \mathbb{P}(z(a, \boldsymbol{X}) > s)$*

This probability refers to actual randomness, in the sense that the geometrical characteristics of a defect are considered random. It is a probability in the frequentist sense but cannot be accessed without prohibitive computational costs because it would involve running C3D over a large set of geometries $\boldsymbol{x}$.

Definition 7.3. *The* surrogate model safety *denoted by $SAFE(a, \boldsymbol{x})$ is the probability, according to the surrogate model, of a particular defect characterized by $(a, \boldsymbol{x})$ being detected: $SAFE(a, \boldsymbol{x}) = \mathbb{P}(Z(a, \boldsymbol{x}) > s)$.*

Contrarily to actual POD, surrogate model safety does not refer to any randomness in the frequentist meaning but instead expresses the "opinion" of the surrogate model. It represents epistemic uncertainty and is a probability in the Bayesian sense.

Definition 7.4. *Let $a \in [0, 1]$. The* mean POD*, denoted by $POD_{mean}(a)$, is the average of surrogate model safety over all defects of depth a: $POD_{mean}(a) = \mathbb{E}(SAFE(a, \boldsymbol{X})) = \mathbb{P}(Z(a, \boldsymbol{X}) > s)$.*

The mean POD is perhaps the best approximation of the actual POD available to us. However it is difficult to interpret since it aggregates two very different kinds of uncertainty: uncertainty about the defect geometry, which is random in kind, and epistemic uncertainty, which refers to imprecision on the part of the surrogate model.

Definition 7.5. *Let $a \in [0, 1]$ and $\gamma \in [0, 1]$. The* POD *at safety level γ, denoted by $POD_\gamma(a)$, is the probability for surrogate model safety to be greater or equal to γ: $POD_\gamma(a) = \mathbb{P}(SAFE(a, \boldsymbol{X}) \geqslant \gamma)$.*

The POD at safety level γ for a given depth length a is the probability of the surrogate model being confident about the defect being detected, with γ denoting the required confidence level. Its aim is to constrain epistemic uncertainty in order to provide a figure that reflects actual

randomness. Its interpretation is therefore clearer than that of the mean POD. However, it is of interest only if γ is high. In the following, we compute it with $\gamma = 95\%$ and $\gamma = 99\%$.

Computationally speaking, because the defect depth a belongs to the interval $[0, 1]$, we endow this interval with the fine grid $0.01 * [\![0, 100]\!]$. For every value of a in this grid, we generate a 1000-points sample from the probability distribution of $\boldsymbol{X}$. For every a therefore, we gather 1000 probabilities of defect detection. With this, we may:

1. compute the mean;
2. count how many are greater or equal to safety levels (95% and 99%).

The first quantity is a Monte-Carlo approximation of the mean POD, the second of the POD at safety level 95% or 99%.

The mean POD and POD at safety level 95% and 99% are drawn in Figure 7.5. On the left, α is taken to be 0.32, the value for which L^{LOG} reaches its maximum. On the right, they are also depicted and accompanied by the curves obtained with integrated α.

Integrated α here means that the Uniform prior distribution on $[0, 1]$ was used for transformation parameter α. Because the likelihood L^{LOG} has a very sharp peak, integration was performed only over the interval $[0.30, 0.37]$. The complement set was deemed to have too low posterior probability to be worth taking into account. Integration was performed by using the rectangle method over $[0.30, 0.37]$ with gap 0.01. This is admittedly a rough approximation, and is intended to serve as a glimpse into a Bayesian treatment of every single parameter of the problem. Figure 7.5 (right) shows that the full-Bayesian method differs only slightly from the Maximum Likelihood method. Interestingly, the full-Bayesian method seems slightly less conservative than the Maximum Likelihood method. We therefore adopt in this case the Maximum Likelihood method for simplicity and conservativeness both.

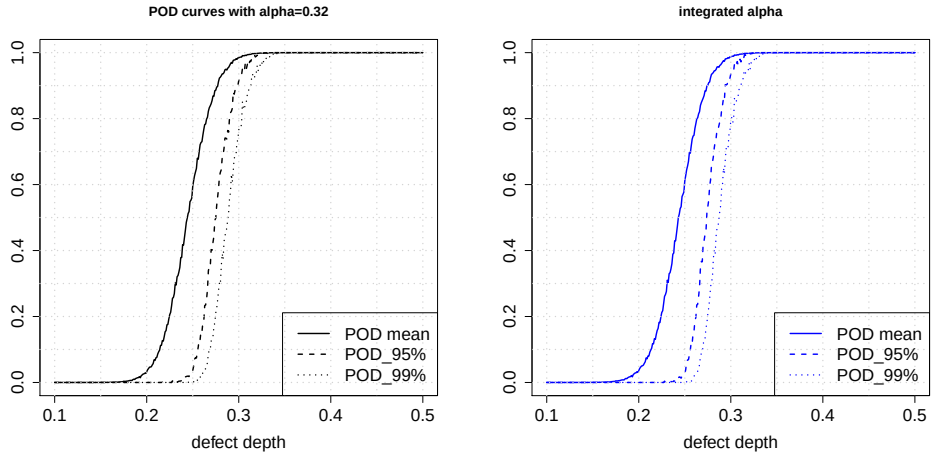


Figure 7.5 – *Left: POD curves with $\alpha = 0.32$. Right: POD curves with $\alpha = 0.32$ and for integrated α (points). Since curves with $\alpha = 0.32$ and integrated α are almost confused, it was not necessary to use different graphical styles for the different POD curves derived with integrated α.*

7.5 Conclusion

In this chapter, we demonstrated how an Objective Bayesian framework could be applied to a Trans-Gaussian Kriging surrogate model. Contrary to Maximum Likelihood approaches, it makes the likelihood function of the transformation parameter clearly discriminate between all possible candidates. Moreover, it provides a way to naturally incorporate hyperparameter uncertainty into the prediction (at least as far as "Gaussian parameters" are concerned, because one of our findings was precisely that there is not much uncertainty about the transformation parameter).

This is especially useful in the context of calibration of NDT techniques by numerical simulation. It makes a clear distinction between randomness and epistemic uncertainty possible. This distinction can then be incorporated into Probability Of defect Detection (POD) curves.

Concerning the particular problem of detecting defects in Steam Generator tubes, our framework provides a theoretically sound solution for estimating surrogate model uncertainty. This makes interpretation of results easier and thereby increases the reliability of the results of the study.

Conclusion

What was done

This thesis is the result of the combination of methods originating from geostatistics, Objective Bayesian statistics and from Markov chain theory. All three domains are part of the solution to an industrial problem put forward by EDF.

This problem was how to compute Probability Of defect Detection (POD) curves for non-destructive testing of defects in steam generator tubes. Because a computer simulation of the procedure was available, such curves could theoretically have been obtained by repeated calls to the code under a Monte-Carlo procedure. Budgetary constraints made such an approach untractable, however.

It is precisely to deal with such cases that surrogate models like Kriging exist. However, the scarcity of available data made reliable estimation of Kriging hyperparameters impossible. Hence the use of Objective Bayesian techniques to circumvent the need to estimate these parameters in the first place.

The first part of the present dissertation describes the tools that were the basis of the solution: Kriging and reference analysis. It recalls how Berger et al. [2001] combined these tools to provide a complete Objective Bayesian analysis of the Kriging model with isotropic covariance kernel and fixes problems in the original proof.

The third part of the dissertation builds on Berger et al. [2001]'s success to extend the analysis to Kriging models with anisotropic kernels. It shows how such a Bayesian analysis makes the derivation of POD curves natural even in the absence of reliable estimate for the Kriging hyperparameters. It also demonstrates how this approach is compatible with the elementary non-Gaussian technique called "trans-Gaussian Kriging" which is basically regular Kriging with an added parametric transformation. Even though complexity increases with each new parameter, it sketches a full-Bayesian analysis of the full model, including the transformation parameter.

This would not have been possible without the notion of optimal compromise between potentially incompatible conditional distributions, which is introduced in the second part. This notion is of great practical interest because it allows Gibbs sampling and thereby makes the full-Bayesian solution tractable. An alternative was developed recently [Gu et al., 2018]: it uses the reference prior obtained by grouping all covariance parameters in the reference prior algorithm. Practical implementations however rely on the Maximum A Posteriori estimator instead of conducting a full-Bayesian analysis.

What remains to be done

There is much that remains to be understood about the optimal compromise between incompatible conditional distributions. By definition, it fits an optimality criterion, but its properties remain largely unknown. What is the precise link between each of the incompatible conditionals and the corresponding conditional in the compromise? My intuition is that the optimal compromise tends to blur dependence relationships, but this idea would need to be quantified.

Is it possible to quantify how much each conditional is taken into account in the compromise? Could changes to the definition allow for the regulation of the importance of each conditional in the definition of the compromise?

For example, an intuitive idea would be to modify the scanning probabilities. If we wanted one of the conditionals to be better taken into account, we could alter the Gibbs algortithm and make the probability of choosing it greater.

But whether the impact of a conditional predominantly depends on this probability or on other factors is unclear. A rarely chosen conditional could still have a tremendous impact on one of the marginals of the resulting joint distribution and thereby impact the joint distribution itself.

And above all, a compromise being optimal does not actually mean that it is *good*. Criteria should be developed to quantify how satisfying a compromise is. In the application presented in the dissertation, results in terms of frequentist behavior were surprisingly good, but they remain empirical rather than formal.

Bibliography

M. Abramowitz and I. A. Stegun, editors. *Handbook of mathematical functions with formulas, graphs, and mathematical tables*, volume 55 of *Applied Mathematics Series*. National Bureau of Standards, 1964.

R. J. Adler. *An introduction to continuity, extrema, and related topics for general Gaussian processes*. Hayward, CA: Institute of Mathematical Statistics, 1990.

E. Anderes. On the consistent separation of scale and variance for Gaussian random fields. *Annals of Statistics*, 38(2):870–893, 2010.

B. C. Arnold, E. Castillo, and J. M. Sarabia. Conditionally specified distributions: an introduction. *Statistical Science*, 16(3):268–269, 2001.

F. Bachoc. *Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer models*. PhD thesis, Université Paris Diderot, 2013a.

F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of gaussian processes with model misspecification. *Computational Statistics & Data Analysis*, 66:55–69, 2013b.

F. Bachoc, G. Bois, J. Garnier, and J.-M. Martinez. Calibration and improved prediction of computer models by Universal Kriging. *Nuclear Science and Engineering*, 176(1):81–97, 2014.

F. Bachoc, E. Contal, H. Maatouk, and D. Rullière. Gaussian processes for computer experiments. *ESAIM: Proceedings and Surveys*, 60:163–179, 2017a.

F. Bachoc, F. Gamboa, J.-M. Loubes, and N. Venet. A Gaussian Process Regression Model for Distribution Inputs. *IEEE Transactions on Information Theory*, 2017b.

Richard F Bass. *Probabilistic techniques in analysis*. Springer Science & Business Media, 1995.

J. Bect, F. Bachoc, and D. Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. *arXiv:1608.01118*, 2016.

J. Berger. The case for objective bayesian analysis. *Bayesian analysis*, 1(3):385–402, 2006.

J. O. Berger and J. M. Bernardo. On the Development of Reference Priors. *Bayesian statistics*, 4(4):35–60, 1992.

J. O. Berger, V. De Oliveira, and B. Sansó. Objective Bayesian analysis of spatially correlated data. *Journal of the American Statistical Association*, 96(456):1361–1374, 2001.

J. O. Berger, J. M. Bernardo, and D. Sun. The formal definition of reference priors. *Annals of Statistics*, 37(2):905–938, 2009.

James O Berger, Jose M Bernardo, and Dongchu Sun. Overall objective priors. *Bayesian Analysis*, 10(1):189–221, 2015.

J. M. Bernardo. Reference Posterior Distributions for Bayesian Inference. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41(2):113–147, 1979a.

J. M. Bernardo. Expected information as expected utility. *The Annals of Statistics*, 7(3): 686–690, 1979b.

J. M. Bernardo. Reference analysis. In D. Dey and C. Rao, editors, *Handbook of statistics*, volume 25, pages 17–90. Elsevier, 2005.

J. M. Bernardo. Integrated objective Bayesian estimation and hypothesis testing. In J. M. Bernardo, M. J. Bayarri, and J. O. Berger, editors, *Bayesian statistics 9*, pages 1–68. Oxford University Press, 2011.

P. J. Bickel and K. A. Doksum. An analysis of transformations revisited. *Journal of the American Statistical Association*, 76(374):296–311, 1981.

P. Billingsley. *Probability and measure* (Third Edition). Wiley, New York, 1995.

M. Binois, D. Ginsbourger, and O. Roustant. Quantifying uncertainty on Pareto fronts with Gaussian process conditional simulations. *European journal of operational research*, 243(2): 386–394, 2015.

S. Bochner. *Vorlesungen über Fouriersche Integrale*, volume 12. Akad. Verl.-Ges., 1932.

G. E. P. Box and D. R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society. Series B (Methodological)*, 26:211–252, 1964.

R. L. Bras and I. Rodriguez-Iturbe. *Random functions and hydrology*. Courier corporation, 1985.

C. Chevalier, J. Bect, D. Ginsbourger, E. Vazquez, V. Picheny, and Y. Richet. Fast parallel kriging-based stepwise uncertainty reduction with application to the identification of an excursion set. *Technometrics*, 56(4):455–465, 2014.

G. Christakos. *Random Field Models in Earth Sciences*. Academic press, New York, 1992.

B. S. Clarke and A. R. Barron. Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical planning and Inference*, 41(1):37–60, 1994.

G. Damblin, M. Couplet, and B. Iooss. Numerical studies of space-filling designs: optimization of Latin Hypercube Samples and subprojection properties. *Journal of Simulation*, 7(4):276–289, 2013.

A. P. Dawid. Invariant prior distributions. In S. Kotz and N. L. Johnson, editors, *Encyclopedia of statistical sciences*, pages 228–236. John Wiley, New York, 1983.

A. P. Dawid and S. L. Lauritzen. Compatible prior distributions. *Bayesian methods with applications to science, policy and official statistics*, pages 109–118, 2001.

V. De Oliveira, B. Kedem, and D. A. Short. Bayesian Prediction of Transformed Gaussian Random Fields. *Journal of the American Statistical Association*, 92:1422–1433, 1997.

P. Druilhet and J.-M. Marin. Invariant HPD and MAP based on Jeffreys measure. *Bayesian Analysis*, 2(4):681–692, 2007.

R. M. Dudley. *Real Analysis and Probability*. Chapman and Hall/CRC, 1989.

S. E. Fienberg. Does it make sense to be an" objective bayesian"?(comment on articles by berger and by goldstein). *Bayesian Analysis*, 1(3):429–432, 2006.

Andrew Gelman and TE Raghunathan. Comment on "Conditionally specified distributions: an introduction" by b.c. arnold, e. castillo and j.m. sarabia. *Statistical Science*, 16(3): 268–269, 2001.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian Data Analysis, Third Edition*. CRC Press, Chapman & Hall, 2013.

M. Ghosh. Objective Priors: An Introduction for Frequentists. *Statistical Science*, 26(2): 187–202, 2011.

I. I. Gihman and A. V. Skorohod. *The theory of stochastic processes*, volume 1. Springer-Verlag, Berlin, 1974.

I. J. Good. A Derivation of the Probabilistic Explication of Information. *Journal of the Royal Statistical Society. Series B (Methodological)*, 28(3):578–581, 1966.

P. Goovaerts. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New York, 1997.

J.C. Gower. Properties of Euclidean and non-Euclidean distance matrices. *Linear Algebra and its Applications*, 67:81–97, 1985.

M. Gu. *Robust Uncertainty Quantification and Scalable Computation for Computer Models with Massive Output*. PhD thesis, Duke University, 2016.

M. Gu, X. Wang, and J. O. Berger. Robust gaussian stochastic process emulation. *Annals of Statistics, In Press*, 2018.

M. S. Handcock and M. L. Stein. A Bayesian Analysis of Kriging. *Technometrics*, 35:403–410, 1993.

M. S. Handcock and J. R. Wallis. An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). *Journal of the American Statistical Association*, 89 (426):368–390, 1994.

J. Hartigan. Invariant prior distributions. *The Annals of Mathematical Statistics*, 35(2): 836–845, 1964.

D. Heckerman, D. M. Chickering, C. Meek, R. Rounthwaite, and C. Kadie. Dependency networks for inference, collaborative filtering, and data visualization. *Journal of Machine Learning Research*, 1:49–75, 2000.

H. Hotelling. New light on the correlation coefficient and its transforms. *Journal of the Royal Statistical Society. Series B (Methodological)*, 15(2):193–232, 1953.

B. Iooss and L. Le Gratiet. Uncertainty and sensitivity analysis of functional risk curves based on Gaussian processes. *Reliability Engineering & System Safety*, 2017.

E. H. Isaaks and M. R. Srivastava. *Applied geostatistics.* Oxford University Press, New York, 1989.

E. T. Jaynes. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4 (3):227–241, 1968.

E. T. Jaynes. On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, 70 (9):939–952, 1982.

H. Jeffreys. *Theory of Probability.* London: Oxford University Press, 3 edition, 1961.

A. G. Journel and Ch. J. Huijbregts. *Mining geostatistics.* Academic press, New York, 1978.

K. Kamary and C. P. Robert. Reflecting about selecting noninformative priors. *Journal of Applied and Computational Mathematics*, 3(5):1–7, 2014.

K. Kamary, K. Mengersen, C. P. Robert, and J. Rousseau. Testing hypotheses via a mixture estimation model. *arXiv:1412.2044*, 2014.

R. E. Kass and L. Wasserman. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, 91(435):1343–1370, 1996.

H. Kazianka and J. Pilz. Objective bayesian analysis of spatial data with uncertain nugget and range parameters. *Canadian Journal of Statistics*, 40(2):304–327, 2012.

M. C. Kennedy and A. O'Hagan. Bayesian calibration of computer models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(3):425–464, 2001.

D. Khoshnevisan. *Multiparameter Processes: an Introduction to Random Fields.* Springer Science & Business Media, 2002.

P. K. Kitadinis. *Introduction to Geostatistics: Applications in Hydrogeology.* Cambridge University Press, New York, 1997.

K. L. Kuo and Y. J. Wang. Pseudo-Gibbs sampler for discrete conditional distributions. *Annals of the Institute of Statistical Mathematics*, pages 1–13, 2017.

K.-L. Kuo, C.-C. Song, and T. J. Jiang. Exactly and almost compatible joint distributions for high-dimensional discrete conditional distributions. *Journal of Multivariate Analysis*, 157:115–123, 2017.

P. S. Laplace. *Essai philosophique sur les probabilités*. Vve Courcier, Paris, 1814.

L. Le Gratiet and J. Garnier. Recursive co-Kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 4(5), 2014.

L. Le Gratiet, B. Iooss, G. Blatman, T. Browne, S. Cordeiro, and B. Goursaud. Model Assisted Probability of Detection curves: New statistical tools and progressive methodology. *Journal of Nondestructive Evaluation*, 36(1):8, 2017.

Loic Le Gratiet. Bayesian analysis of hierarchical multifidelity codes. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1):244–269, 2013.

P. M. Lee. On the axioms of information theory. *The Annals of Mathematical Statistics*, 35 (1):415–418, 1964.

R. Li and A. Sudjianto. Analysis of computer experiments using penalized likelihood in Gaussian Kriging models. *Technometrics*, 47(2):111–120, 2005.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

D. V. Lindley. Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 102–107, 1958.

B. Liseo. The elimination of nuisance parameters. In D. Dey and C. Rao, editors, *Handbook of Statistics*, volume 25, chapter 7, pages 193–219. Elsevier-Sciences, 2006.

J.-M. Marin, K. Mengersen, and C. P. Robert. Bayesian modelling and inference on mixtures of distributions. In D. Dey and C. Rao, editors, *Handbook of statistics*, volume 25, pages 459–507. Elsevier-Sciences, 2005.

J.-M. Marin, P. Pudlo, C. P. Robert, and R. J. Ryder. Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180, 2012.

J.-M. Marin, N. S. Pillai, C. P. Robert, and J. Rousseau. Relevant statistics for Bayesian model choice. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76(5):833–859, 2014.

G. Matheron. Krigeage d'un panneau rectangulaire par sa périphérie. *Note géostatistique*, 28, 1960.

B. Matérn. *Spatial Variation*. Springer-Verlag, Berlin, 2nd edition, 1986.

L. Maurice, V. Costan, and P. Thomas. Axial probe eddy current inspection of steam generator tubes near anti-vibration bars: performance evaluation using finite element modeling. In *Proceedings of JRC-NDE, Cannes*, pages 638–644, 2013.

K. Mengersen, C. P. Robert, and M. Titterington. *Mixtures: estimation and applications*, volume 896. John Wiley & Sons, 2011.

J. Muré. Optimal compromise between incompatible conditional probability distributions and Objective Bayesian Kriging. `https://arxiv.org/pdf/1703.07233`, 2017.

J. Muré. A Comprehensive Bayesian Treatment of the Universal Kriging parameters with Matérn correlation kernels. `https://arxiv.org/pdf/1801.01007`, 2018.

J. Muré. Propriety of the reference posterior distribution in Gaussian Process regression. `https://arxiv.org/pdf/1805.08992`, 2018a.

J. Muré. Trans-Gaussian Kriging in a Bayesian framework : a case study. `https://arxiv.org/pdf/1805.09038`, 2018b.

L. Nachbin. *The Haar Integral*. New York: van Nostrand, 1965.

E. Padonou and O. Roustant. Polar Gaussian Processes and Experimental Designs in Circular Domains. *SIAM/ASA Journal on Uncertainty Quantification*, 4(1):1014–1033, 2016.

R. Paulo. Default priors for Gaussian processes. *Annals of Statistics*, 33(2):556–582, 2005.

C. Perrin, G.and Soize, S. Marque-Pucheu, and J. Garnier. Nested polynomial trends for the improvement of Gaussian process-based predictors. *Journal of Computational Physics*, 346:389–402, 2017.

L. Pronzato. Minimax and maximin space-filling designs: some properties and methods for construction. *Journal de la Société Française de Statistique*, 158(1):7–36, 2017.

L. Pronzato and M.-J. Rendas. Bayesian local Kriging. *Technometrics*, 59(3):293–304, 2017.

C. E. Rasmussen and C. K. I. Williams. *Gaussian processes for machine learning*. MIT Press, 2006.

C. Ren, D. Sun, and C. He. Objective bayesian analysis for a spatial model with nugget effects. *Journal of Statistical Planning and Inference*, 142(7):1933–1946, 2012.

C. Ren, D. Sun, and S. K. Sahu. Objective bayesian analysis of spatial models with separable correlation functions. *Canadian Journal of Statistics*, 41(3):488–507, 2013.

P. Ressel. De Finetti-type theorems: an analytical approach. *The Annals of Probability*, 13 (3):898–922, 1985.

C. P. Robert. *The Bayesian Choice : From Decision-Theoretic Foundations to Computational Implementation.* Springer-Verlag, New York, 2007.

C. P. Robert and J. Rousseau. How Principled and Practical Are Penalised Complexity Priors? *Statistical Science*, 32(1):36–40, 2017.

C. P. Robert, N. Chopin, and J. Rousseau. Harold Jeffreys's Theory of Probability Revisited. *Statistical Science*, 24(2):141–172, 2009.

C. P. Robert, J.-M. Marin, and J. Rousseau. Bayesian inference and computation. *Handbook of statistical systems biology. West Sussex: John Wiley & Sons, Ltd*, pages 39–65, 2011.

O. Roustant, E. Padonou, Y. Deville, A. Clément, G. Perrin, J. Giorla, and H. Wynn. Group kernels for Gaussian process metamodels with categorical inputs. *arXiv:1802.02368*, 2018.

G. Roverato, A.and Consonni. Compatible prior distributions for directed acyclic graph models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(1): 47–61, 2004.

D. Rullière, N. Durrande, F. Bachoc, and C. Chevalier. Nested Kriging predictions for datasets with a large number of observations. *Statistics and Computing*, 28(4):849–867, 2018.

T. J. Santner, B. J. Williams, and W. I. Notz. *The Design and Analysis of Computer Experiments*. Springer-Verlag, New York, 2003.

I.T. Schoenberg. On certain Metric Spaces arising from Euclidean Spaces by a change of metric and their imbedding in Hilbert Space. *Annals of Mathematics*, 38(4):787–793, 1937.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support with Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, 2012.

J. W. Seaman III, J. W. Seaman Jr, and J. D. Stamey. Hidden dangers of specifying noninformative priors. *The American Statistician*, 66(2):77–84, 2012.

Teddy Seidenfeld. Entropy and uncertainty. *Philosophy of Science*, 53(4):467–491, 1986.

C. E. Shannon. A Mathematical Theory of Communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

D. Simpson, H. Rue, A. Riebler, T. G. Martins, and S. H. Sørbye. Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors. *Statistical science*, 32(1):1–28, 2017.

E. Snelson, C. E. Rasmussen, and Z. Ghahramani. Warped gaussian processes. In *Advances in neural information processing systems*, volume 16, pages 337–344, 2004.

M. L. Stein. *Interpolation of Spatial Data. Some Theory for Kriging*. Springer Series in Statistics. Springer-Verlag, New York, 1999.

P. Thomas, B. Goursaud, L. Maurice, and S. Cordeiro. Eddy-current non destructive testing with the finite element tool Code_Carmel3d. In *11th International Conference on Nondestructive Evaluation, Jeju*, 2015.

F. Tuyl, R. Gerlach, and K. Mengersen. Consensus priors for multinomial and binomial ratios. *Journal of Statistical Theory and Practice*, 10(4):736–754, 2016.

H. Wackernagel. *Multivariate Geostatistics*. Springer-Verlag, Berlin, 1995.

J. Watson and C. Holmes. Approximate models and robust decisions. *Statistical Science*, 31 (4):465–489, 2016.

B. L. Welch and H. W. Peers. On formulae for confidence points based on integrals of weighted likelihoods. *Journal of the Royal Statistical Society. Series B (Methodological)*, 25(2):318–329, 1963.

W. Xueou, D. J. Nott, C. C. Drovandi, K. Mengersen, and M. Evans. Using history matching for prior choice. *Technometrics*, 2018.

A. M. Yaglom. *Correlation Theory of Stationary and Related Random Functions 1. Basic Results*. Springer-Verlag, New York, 1987.

R. Yang and J. O. Berger. *A catalog of noninformative priors*. Institute of Statistics and Decision Sciences, Duke University, 1996.

B. Yet and W. Marsh. Compatible and incompatible abstractions in Bayesian networks. *Knowledge-Based Systems*, 62:84–97, 2014.

H. Zhang. Inconsistent Estimation and Asymptotically Equal Interpolations in Model-based Geostatistics. *Journal of the American Statistical Association*, 99(465):250–261, 2004.