Contributions aux processus gaussiens, à la quantification d'incertitudes et à l'inférence post-sélection de modèle

François Bachoc

UNIVERSITÉ PAUL SABATIER (Toulouse 3)

&

INSTITUT DE MATHÈMATIQUES DE TOULOUSE

mémoire présenté pour l'obtention de

L'HABILITATION

à diriger les recherches

présenté par

François BACHOC

Contributions aux processus gaussiens, à la quantification d'incertitudes et à l'inférence post-sélection de modèle

---

Soutenue le 29 novembre 2018 devant le jury composé de :

M. Bernard BERCU	Université Bordeaux 1
Mme Béatrice LAURENT	INSA Toulouse
M. Jean-Michel LOUBES	Université Paul Sabatier
M. Emilio PORCU	Newcastle University
Mme Clémentine PRIEUR	Université Grenoble Alpes
M. Vincent RIVOIRARD	Université Paris Dauphine

D'après les rapports de :

Mme Gerda CLAESKENS	KU Leuven
M. Emilio PORCU	Newcastle University
Mme Clémentine PRIEUR	Université Grenoble Alpes

Remerciements

Je remercie chaleureusement Clémentine Prieur pour m'avoir fait l'honneur de rapporter ce manuscrit d'habilitation. Je suis reconnaissant pour son rapport encourageant et constructif sur mon travail de recherche. *I warmly thank Gerda Claeskens and Emilio Porcu for their reports as well. I am honored by their positive and insightful feedback.*

Je suis très reconnaissant à Bernard Bercu, Béatrice Laurent et Vincent Rivoirard de compléter mon jury d'habilitation. Je suis ravi de recevoir leurs retours sur mon travail. Je remercie chaleureusement Jean-Michel Loubes pour avoir parrainé cette habilitation.

Quelques années après la rédaction de mon manuscrit de thèse en 2013, je souhaite remercier à nouveau Josselin Garnier et Jean-Marc Martinez pour m'avoir donnée ma première initiation à la recherche.

My next research experience was at the university of Vienna from 2013 to 2015. I am grateful to Benedikt M. Pötscher and Hannes Leeb, for their trust when they offered me a post-doctoral position. I also thank them for our research collaboration on post-model-selection inference and for their support. I am also happy to thank David Preinerstorfer and Lukas Steinberger for our work together. I am grateful to Reinhard Furrer for our joint work at this time. At last, I thank my other former colleagues from the university of Vienna, for the good time I spent there.

Mon accueil à l'institut de mathématiques de Toulouse et à l'université Paul Sabatier en 2015 s'est déroulé dans de très bonnes conditions. Je pense que Fabrice Gamboa et Jean-Michel Loubes ont fortement contribué à cela. Je les remercie donc de m'avoir proposé dès le début des collaborations de recherche, de m'avoir ouvert à de nouvelles rencontres scientifiques et pour leurs conseils bienveillants.

Je remercie les membres du support administratif et informatique de l'institut de mathématiques de Toulouse pour leur travail et leur aide. J'ai particulièrement interagi avec Nicole Lhermitte, Françoise Michel, et Céline Rozier, que je remercie à nouveau. Je salue également Sébastien Déjean et Jonas Kahn, avec qui j'ai eu le plaisir de partager un bureau.

Merci à mes collègues de l'institut de mathématiques de Toulouse pour nos collaborations et discussions. Ils font de cet institut un lieu remarquablement propice à l'animation et l'ouverture scientifique.

Je remercie les membres de la chaire de recherche OQUAIDO et des projets ANR PEPITO, RISCOPE et SansSoucis. Leur contact est toujours source de motivation et d'ouverture scientifique pour moi.

J'exprime ma gratitude aux doctorants José Daniel Bétancourt, Baptiste Broto et Andrès Felipe López-Lopera, pour m'avoir fait confiance en tant que co-directeur de thèse. Mon travail avec eux est enrichissant et fécond, et je leur souhaite la meilleure réussite possible. Je remercie également les autres directeurs et co-directeurs de ces trois thèses : Marine Depecker, Nicolas Durrande, Thierry Klein, Jean-Marc Martinez et Olivier Roustant.

Je suis reconnaissant envers mes co-auteurs, pour les différents articles et pré-publications que nous avons écrits ensemble. Le travail de recherche est agréable et enrichissant avec eux.

Finalement, je remercie affectueusement ma famille et mes amis pour leur soutien et les bons moments passés ensemble.

Résumé

L'auteur donne un panorama des contributions de recherche qu'il a obtenues en tant que post-doctorant à l'université de Vienne et que Maître de Conférences à l'université Paul Sabatier. Tout d'abord, l'auteur présente les résultats théoriques qu'il a obtenu pour l'estimation de paramétres de covariance de processus gaussiens, dans les cadres asymptotiques par expansion et par remplissage. Ensuite, il décrit ses autres contributions aux processus gaussiens : grands volumes de données, distributions en entrées, approches séquentielles, contraintes d'inégalités et applications à la quantification d'incertitudes. Enfin, il présente ses contributions à la construction d'intervalles de confiance valides dans un cadre post-sélection de modèle.

Mots clès

Processus gaussiens; expériences numériques; quantification d'incertitudes; estimation de paramètres de covariance; cadres asymptotiques par expansion et par remplissage; contraintes d'inégalités; approches séquentielles; aggrégation de prédicteurs; inférence post-sélection de modèle; intervalles de confiance.

Abstract

The author provides an overview of the research contributions that he made as a post-doctoral fellow at the University of Vienna, and as a 'Maître de conférences' at the University Paul Sabatier. First, the author presents the theoretical results he obtained for covariance parameter estimation of Gaussian processes, in the fixed and increasing-domain asymptotic settings. Then, he describes his other contributions to Gaussian processes: large data sets, distribution inputs, sequential designs, inequality constraints and applications to uncertainty quantification. Finally, he presents his contributions to the construction of valid confidence intervals post-model-selection.

Key words

Gaussian processes; computer experiments; uncertainty quantification, covariance parameter estimation, fixed and increasing-domain asymptotics; inequality constraints; sequential design; predictor aggregation; post-model-selection inference; confidence intervals.

Contents

Chapter 1

Introduction

1.1 List of publications

My publications are available on my web page

`www.math.univ-toulouse.fr/~fbachoc/index.html`

The technical reports of the commissariat à l'énergie atomique et aux énergies alternatives (CEA) can be made available on request.

In the list below, the reference [T1] corresponds to my master thesis. The references [J1], [J2], [J3], [C1], [T2] and [T3] correspond to my PhD thesis. All the other references correspond to research that was performed after my PhD thesis.

I am currently co-supervising three PhD students: Andrès Félipe López-Lopera, Baptiste Broto and José Daniel Bétancourt. The references [J14], [S4], [S5] and [S8] correspond to joint work with them.

Refereed journal articles

[J16] F. Bachoc, G. Blanchard and P. Neuvial . On the post selection inference constant under restricted isometry properties. **Electronic Journal of Statistics**, forthcoming, 2018.

[J15] J. Bect, F. Bachoc and D. Ginsbourger. A supermartingale approach to Gaussian process based sequential design of experiments. **Bernoulli**, forthcoming, 2018.

[J14] A. F. López-Lopera, F. Bachoc, N. Durrande and O. Roustant. Finite-dimensional Gaussian approximation with linear inequality constraints. **SIAM/ASA Journal on Uncertainty Quantification**, forthcoming 2018.

[J13] F. Bachoc, H. Leeb, and B. M. Pötscher. Valid confidence intervals for post-model-selection predictors. **Annals of Statistics**, forthcoming, 2018.

[J12] F. Bachoc, F. Gamboa, J.M. Loubes and N. Venet. Gaussian process regression model for distribution inputs. **IEEE Transactions on Information Theory**, 2017.

[J11] F. Bachoc. Asymptotic analysis of covariance parameter estimation for Gaussian processes in the misspecified case. **Bernoulli**, 24(2):1531-1575, 2018.

[J10] D. Rullière, N. Durrande, F. Bachoc and C. Chevalier. Nested Kriging estimations for datasets with large number of observations. **Statistics and Computing**, 2017.

[J9] F. Bachoc, A. Lagnoux and T.M.N. Nguyen. Cross-validation estimation of covariance parameters under fixed-domain asymptotics. **Journal of Multivariate Analysis**, 160:42-67, 2017.

[J8] D. Velandia, F. Bachoc, M. Bevilacqua, X. Gendre, J.M. Loubes. Maximum likelihood estimation for a bivariate Gaussian process under fixed domain asymptotics. **Electronic Journal of Statistics**, 11(2):2978 - 3007, 2017.

[J7] F. Bachoc, M. Ehler and M.Gräf. Optimal configurations of lines and a statistical application. **Advances in Computational Mathematics**, 43 (1):113–126, 2017.

[J6] R. Furrer, F. Bachoc and J. Du. Asymptotic properties of multivariate tapering for estimation and prediction. **Journal of Multivariate Analysis**, 149:177-191, 2016.

[J5] F. Bachoc and R. Furrer. On the smallest eigenvalues of covariance matrices of multivariate spatial processes. **Stat**, 5:102–107, 2016.

[J4] F. Bachoc, K. Ammar and J.M. Martinez. Improvement of code behavior in a design of experiments by metamodeling. **Nuclear Science and Engineering**, 183(3):387–406, 2016.

[J3] F. Bachoc. Asymptotic analysis of the role of spatial sampling for covariance parameter estimation of Gaussian processes. **Journal of Multivariate Analysis**, 125:1–35, 2014.

[J2] 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.

[J1] F. Bachoc. Cross validation and maximum likelihood estimations of hyper-parameters of Gaussian processes with model mispecification. **Computational Statistics and Data Analysis**, 66:55–69, 2013.

Articles submitted to journals

[S8] F. Bachoc, B. Broto, F. Gamboa, J-M Loubes. Gaussian processes indexed on the symmetric group: prediction and learning. `arxiv.org/abs/1803.06118`, **submitted**, 2018.

[S7] F. Bachoc, M. Bevilacqua and D. Velandia. Composite likelihood estimation for a Gaussian process under fixed domain asymptotics. `arxiv.org/abs/1807.08988`, **submitted**, 2018.

[S6] J.M. Azais, F. Bachoc, A. Lagnoux, T.M.N. Nguyen and T. Klein. Semi-parametric estimation of the variogram of a Gaussian process with stationary increments. `arxiv.org/abs/1806.03135`, **submitted**, 2018.

[S5] F. Bachoc, A. Lagnoux and A. F. López-Lopera. Maximum likelihood estimation for Gaussian processes under inequality constraints. `arxiv.org/abs/1804.03378`, **submitted**, 2018.

[S4] B. Broto, F. Bachoc, M. Depecker and J.M. Martinez. Sensitivity indices for independent groups of variables. `arxiv.org/abs/1801.04095`, **submitted**, 2017.

[S3] M. Ben Salem, F. Bachoc, O. Roustant, F. Gamboa and L. Tomaso. Sequential dimension reduction for learning features of expensive black-box functions. `hal.archives-ouvertes.fr/hal-01688329/document`, resubmitted after decision **major revision**, 2018.

[S2] F. Bachoc, N. Durrande, D. Rullière and C. Chevalier. Some properties of nested Kriging predictors. `arxiv.org/abs/1707.05708`, **submitted**, 2017.

[S1] F. Bachoc, D. Preinerstorfer and L. Steinberger. Uniformly valid confidence intervals post-model-selection. `arxiv.org/abs/1611.01043`, resubmitted after decision **major revision**, 2016.

Preprints (not submitted yet)

[P1] F. Bachoc, A. Suvorikova, J-M Loubes and V. Spokoiny. Gaussian process forecast with multidimensional distributional entries. `arxiv.org/abs/1805.00753`

Refereed conference proceedings

[C6] F. Bachoc, E. Contal, H. Maatouk and D. Rullière. Gaussian processes for computer experiments. **ESAIM: Proceedings and Surveys**, 2017. 17 pages.

[C5] F. Bachoc, D. Preinerstorfer and L. Steinberger. Intervalles de confiance uniformément valides en présence de sélection de modèle. **49èmes Journées de Statistique,** Avignon, May 29- June 2 2017. 6 pages.

[C4] F. Bachoc, F. Gamboa, J.M. Loubes and N. Venet. Modèles de régression gaussienne pour des distributions en entrée. **49èmes Journées de Statistique,** Avignon, May 29 - June 2 2017. 6 pages.

[C3] F. Bachoc, H. Leeb and B. Pötscher. Intervalles de confiance valides en présence de sélection de modèle. **47èmes Journées de Statistique,** Lille, June 1-5 2015. 6 pages.

[C2] F. Bachoc, A. Bachouch, and L. Lenôtre. Hastings-Metropolis algorithm on Markov chains for small-probability estimation. **ESAIM: Proceedings and Surveys**, 48, p.276, 2015. 32 pages.

[C1] F. Bachoc, J. Garnier and J.M. Martinez. Maximum de vraisemblance et validation croisée pour l'estimation des hyper-paramètres de covariance pour le Krigeage. **45èmes Journées de Statistique**, Toulouse, May 27-31 2013. 6 pages.

Technical reports

[T3] J.M. Martinez, A. Marrel, N. Gilardi and F. Bachoc. Krigeage par processus gaussiens Librairie gpLib. **Rapport technique CEA**, 2012. 50 pages.

[T2] F. Bachoc, G. Bois and J.M. Martinez. Contribution à la validation des codes de calcul par processus gaussiens. Application à la calibration du modèle de frottement pariétal de Flica 4. **Rapport technique CEA**, 2012. 42 pages.

[T1] F. Bachoc. Calibration de modèles physiques par méthodes probabilistes. **Rapport technique CEA**, 2010. 78 pages.

1.2 Overview of research activities

Beforehand, I would like to stress out that the works discussed below have been made possible thanks to a number of co-authors, from which I have learned a lot.

The topic of my PhD thesis (defended in October 2013) is Gaussian process models. The first part of the thesis is focused on obtaining methodological and increasing-domain asymptotic results on covariance parameter estimation [J1,J3]. The second part of the thesis is constituted by applications of Gaussian process models to computer experiments and uncertainty quantification [J2].

My first experience of post-PhD research coincides with my participation to the CEMRACS 2013 summer school, where I worked on Monte Carlo algorithms and Markov chains, for a neutron transport application. The resulting publication [C2] turns out to be somehow independent of my other publications and is mentioned in Chapter 5.

At the university of Vienna

When I was a post-doctoral fellow at the university of Vienna, I strengthened, on the one hand, my contributions to covariance parameter estimation (Chapter 2) and to applications to uncertainty quantification (Chapter 3). On the other hand, I started working on the topic of post-selection inference, together with my colleagues at the university of Vienna (Chapter 4).

Consider first covariance parameter estimation. In my PhD article [J1], it was shown that the cross validation estimator could be preferable to the maximum likelihood estimator for prediction purposes. In [J11], I obtained a rigorous result, under increasing-domain asymptotics, supporting this finding. Also, it turns out that some of the techniques introduced in my PhD article [J3] (in the univariate case) could be extended to the multivariate case. This enabled us to provide consistency results for multivariate covariance tapering in [J6]. Similarly, [J3] contains an asymptotic lower bound for the eigenvalues of covariance matrices of stationary univariate processes. In [J5], we extend this bound to multivariate processes, using complex Hermitian matrix tools.

Related to uncertainty quantification, the reference [J4] is an application of Gaussian processes to surrogate modeling of computer models in nuclear engineering. In particular we show how Gaussian processes provide an useful interpretation of the behavior of the computer model, together with anomaly detection tools.

Consider then post-selection inference. For this theme of research, I focus on an approach based on simultaneous coverage and post-selection inference (PoSI) constants, that was suggested in Berk et al. [2013], for covering individual coefficients in linear regression. The article [J13], which first version was written when I was working in Vienna, extends the PoSI framework to the coverage of linear predictors. Another article, that focuses on the link between the computation of the PoSI constants and optimal ('space filling') configurations of lines, is [J7].

At the Institut de Mathématiques de Toulouse

As I started working as a 'Maître de conférences' at the Institut de Mathématiques de Toulouse, I aimed at broadening my areas of research on Gaussian processes. I hence started working on fixed-domain asymptotics for covariance function estimation. With co-authors, we focused in particular on the exponential covariance function, for which the theoretical analysis is facilitated (see Ying [1991]). Chapter 2 presents the two articles [J8,J9] dealing with the exponential covariance function. In [J8], we prove the asymptotic normality of maximum likelihood estimators of microergodic parameters for a bivariate Gaussian process. In [J9], we prove the asymptotic normality of a cross validation estimator of the microergodic parameter (for a univariate process).

With various co-authors, I also started working on more diverse aspects of Gaussian processes. In particular, I started putting an emphasis on the connections between Gaussian processes and contemporary machine learning and data science problems, for instance large data sets, sequential algorithms, specific input data structures, regression under constraints and Monte Carlo procedures. This is the object of Chapter 3.

We address the issue of large data sets in [J10], for which the usual matrix-based formulas of Gaussian processes can not be implemented in practice. In [J10], we suggest an optimal aggregation of Gaussian process models based on smaller data subsets, together with a dedicated cross validation procedure, and we show the good numerical performances of this aggregation, in comparison with other existing procedures. In [S2], we analyze the setting of [J10] theoretically, providing a consistency result and other theoretical guarantees for our aggregation, and giving examples of inconsistency for other usual aggregation procedures.

In [J15], we address iterative designs for Gaussian processes, and focus on a class of strategies called stepwise uncertainty reduction (SUR). The SUR setting provides a fairly general framework that can be applied, for instance, to optimization or failure domain estimation. In [J15], we provide a general consistency result for a large class of SUR strategies, based, in particular, on supermartingale arguments. We also apply this general result to four standard algorithms.

In [J12], we consider Gaussian processes for which the inputs are one-dimensional distributions (instead of real numbers or vectors as is considered above). We consider covariance kernels based on the Monge-Kantorovich, or Wasserstein, transport-based distance. We demonstrate the good empirical performances of the resulting Gaussian process model and provide asymptotic results (in a situation that would correspond to increasing-domain asymptotics for vector inputs) for covariance function estimation.

In [J14,S5], we consider Gaussian processes with specific inequality constraints (boundedness, monotonicity and convexity). These types of constraints indeed regularly occur in applied

situations. In [J14], we extend in several directions the approach of Maatouk and Bay [2017]. This approach is based on a finite dimensional representation of the process, that guarantees that the constraints are satisfied everywhere on the input space. We enable a general aggregation of constraints (for instance, monotonicity and boundedness) and we study the performances of various Markov chain Monte Carlo (MCMC) algorithms to sample from the conditional distribution of the constrained Gaussian process. In addition, we introduce a constrained maximum likelihood estimator and study its consistency. In [S5], we provide a deeper theoretical analysis of maximum likelihood and constrained maximum likelihood. We show that these two estimators have the same asymptotic distribution conditionally to the constraints, and that this asymptotic distribution is the same as for maximum likelihood without constraints. Hence, informally speaking, the constraints have no asymptotic impact on estimation. Nevertheless, we show in simulations that they have an impact in finite sample, and that constrained maximum likelihood is then more accurate.

At the Institut de Mathématiques de Toulouse, I also continued working on post-selection inference (Chapter 4). Recall that in Berk et al. [2013] and [J13], the setting consists in (Gaussian) linear regression. In [S1], we show that the general idea of Berk et al. [2013] can actually be extended to significantly more general frameworks, based on asymptotic approximations. We provide a general construction of confidence intervals, together with uniform asymptotic guarantees. We show in simulations that these intervals offer stronger guarantees than those based on the 'conditional' post-selection inference framework (see for instance Lee et al. [2016]), while, overall, providing similar or shorter intervals. Also, in [J16], we consider the asymptotic order of magnitude of the PoSI constants. We provide an asymptotically optimal upper bound for the case of design matrices satisfying restricted isometry properties (RIP).

Finally, the references [S3,S4,S6,S7,S8,P1] are more recent. They also correspond to research that was carried out at the Institut de Mathématiques de Toulouse. They are briefly presented in Chapter 5.

Chapter 2

Covariance parameter estimation for Gaussian processes

2.1 Introduction

Gaussian processes A Gaussian process ξ on $\mathbb{R}^d$ is a stochastic process from $\mathbb{R}^d$ to $\mathbb{R}$ so that, for any $a \in \mathbb{N}$ and $x_1, \ldots, x_a \in \mathbb{R}^d$, the random vector $(\xi(x_1), \ldots, \xi(x_a))$ is Gaussian Rasmussen and Williams [2006]. In this manuscript, we shall consider that the Gaussian process ξ has mean function zero, so that it is characterized by its covariance function $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. The function k is symmetric non-negative definite, that is, for any $a \in \mathbb{N}$ and $x_1, \ldots, x_a \in \mathbb{R}^d$, the matrix $[k(x_i, x_j)]_{1 \leq i,j \leq a}$ is symmetric non-negative definite. Examples of standard covariance functions are given, for instance, in Stein [1999], Rasmussen and Williams [2006], Roustant et al. [2012], Abrahamsen [1997]. A covariance function k is stationary when $k(x_1, x_2) = k(x_3, x_4)$ whenever $x_2 - x_1 = x_4 - x_3$. For a stationary covariance function k, we let $k(x_1, x_2) = k(x_1 - x_2)$ and identify k as a function from $\mathbb{R}^d$ to $\mathbb{R}$.

Gaussian processes have become popular models for many applications, probably because of their combination of flexibility and simplicity. Indeed, as we shall see below, the conditional distribution of Gaussian processes, given observed values, is particularly tractable in practice. Furthermore, many properties of Gaussian process realizations (e.g. symmetry or smoothness) can be enforced by appropriately selecting the covariance function Stein [1999], Rasmussen and Williams [2006].

Scientific fields where Gaussian processes are applied include machine learning Rasmussen and Williams [2006], geosciences Matheron [1970] and computer experiments Santner et al. [2003]. In this manuscript, we shall mainly discuss computer experiments when considering applications (the term uncertainty quantification is also employed in this context). A computer experiment corresponds to an evaluation of a code function (or computer model) $f_{\text{code}} : \mathbb{R}^d \to \mathbb{R}$. [The domain of f_{code} would typically be a bounded subset of $\mathbb{R}^d$ in practice, but we omit this fact for simplicity of exposition.] Typically, the input $x \in \mathbb{R}^d$ corresponds to geometric or physical parameters of the experiment and the output $f_{\text{code}}(x)$ corresponds to a scalar quantity of interest (usually extracted from complex output data of the computer simulation). For

specific examples, we refer to Santner et al. [2003], Bachoc [2013] and [J4]. The paradigm of Gaussian process models for computer experiments is to consider the (fixed and unknown) code function f_{code} as a realization of a Gaussian process ξ. This provides a Bayesian framework that enables to achieve various tasks of interest, for instance, metamodeling (predicting $f_{\text{code}}(x)$ [J4]), calibration (tuning the computer model for reproduction of field experiments Paulo et al. [2012], [J2]), global optimization Jones et al. [1998] or failure domain estimation Chevalier et al. [2014].

Let us now discuss Gaussian conditioning. Throughout the manuscript, for a function $g : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and for two finite sequences $V = (v_1, \ldots, v_a)$ and $W = (w_1, \ldots, w_b)$ of points in $\mathbb{R}^d$, we let $g(V, W)$ be the $a \times b$ matrix defined by $[g(V, W)]_{i,j} = g(v_i, w_j)$ for $i \in \{1, \ldots, a\}$ and $j \in \{1, \ldots, b\}$. Consider then a Gaussian process ξ observed at $X = (x_1, \ldots, x_n)$ with $x_i \in \mathbb{R}^d$ for $i \in \{1, \ldots, n\}$. Let $y_i = \xi(x_i)$. We call $x_1, \ldots, x_n$ the observation points and $y_1, \ldots, y_n$ the observed values. Then, the Gaussian conditioning theorem (see e.g. Rasmussen and Williams [2006]) implies that, conditionally to $y = (y_1, \ldots, y_n)^\top$, ξ is a Gaussian process with mean and covariance functions m_n and k_n defined by

$$m_n(u) = k(u, X)k(X, X)^{-1}y \tag{2.1}$$

and

$$k_n(u, v) = k(u, v) - k(u, X)k(X, X)^{-1}k(X, v). \tag{2.2}$$

This conditional Gaussianity of ξ, together with the above explicit expressions of the conditional moments, is one of the main reasons why Gaussian processes are popular. The conditional mean function provides an approximation of ξ based on the observed values y (sometimes called a metamodel or surrogate model in the case of computer experiments [J4]). The conditional variance $k_n(u, u)$ is an uncertainty indicator on the value of $\xi(u)$. Finally, the full conditional Gaussianity can be used to define, for instance, probabilistic sequential design strategies Chevalier et al. [2014], [J15].

Covariance function estimation In practice, the covariance function k of the Gaussian process is unknown and is assumed to belong to a parametric set of covariance functions $\{k_\theta, \theta \in \Theta\}$. Here Θ is a subset of $\mathbb{R}^p$ for $p \in \mathbb{N}$ and k_θ is a covariance function on $\mathbb{R}^d$ for $\theta \in \Theta$. A classical example if given by $d = 1$, $\Theta = (0, \infty)^2$, $\theta = (\sigma^2, \ell)$ and $k_\theta(t_1, t_2) = \sigma^2 \exp(-|t_1 - t_2|/\ell)$, where covariance functions of this form are called exponential covariance functions. Many more examples can be found, for instance, in Stein [1999], Rasmussen and Williams [2006], Roustant et al. [2012], Abrahamsen [1997]. In this chapter, we will put a special emphasis on the distinction between the cases where the true covariance function k does belong to $\{k_\theta, \theta \in \Theta\}$ and the case where it does not. We shall refer to the first case as the well-specified case and to the second case as the misspecified case. In the well-specified case, we write $k = k_{\theta_0}$ with $\theta_0 \in \Theta$.

The most standard method for selecting a covariance parameter θ is called (Gaussian) maximum likelihood (ML) Stein [1999], Rasmussen and Williams [2006]. It consists in maximizing, over $\theta \in \Theta$, the Gaussian probability density function at y, when considering that ξ has zero mean function and covariance function k_θ. The ML estimator $\hat{\theta}_{ML}$ can thus be written as

$$\hat{\theta}_{ML} \in \operatorname{argmin}_{\theta \in \Theta} \log(|k_\theta(X, X)|) + y^\top k_\theta(X, X)^{-1}y, \tag{2.3}$$

with the notation of (2.1) and (2.2) and where $|.|$ is the determinant. In practice, the optimization problem (2.3) needs to be tackled numerically and popular procedures are gradient descent-type algorithms or global derivative-free techniques (e.g. genetic algorithms) Roustant et al. [2012]. Furthermore, the cost of evaluating the criterion in (2.3) for a given θ is a $O(n^3)$, which can become problematic when n becomes larger than about $10,000$ (see [J6,J10]).

Another class of procedures for selecting θ is called cross validation (CV). CV is less popular than ML but is nevertheless considered by some authors Sundararajan and Keerthi [2001], Zhang and Wang [2010]. CV was also one of the main topics of my PhD thesis Bachoc [2013]. With the notation of (2.3), let X_{-i} be the finite sequence $(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$ for $i \in \{1, \ldots, n\}$. Let also $y_{-i} = (y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_n)^\top$. Then, let for $u \in \mathbb{R}^d$

$$m_{\theta,n,-i}(u) = k_\theta(u, X_{-i}) k_\theta(X_{-i}, X_{-i})^{-1} y_{-i} \tag{2.4}$$

and

$$k_{\theta,n,-i}(u, v) = k_\theta(u, v) - k_\theta(u, X_{-i}) k_\theta(X_{-i}, X_{-i})^{-1} k_\theta(X_{-i}, v). \tag{2.5}$$

The two above quantities are conditional means and covariances given the observation vector y_{-i} where the observation $y_i = \xi(x_i)$ is left out. Then, CV consists in optimizing, over $\theta \in \Theta$, an empirical criterion evaluating the quality of the conditional distributions for y_i given by $m_{\theta,n,-i}(x_i)$ and $k_{\theta,n,-i}(x_i, x_i)$ for $i = 1, \ldots, n$. To be specific, a common CV estimator is defined by

$$\hat{\theta}_{CV} \in \text{argmin}_{\theta \in \Theta} \sum_{i=1}^{n} \left(y_i - m_{\theta,n,-i}(x_i)\right)^2 \tag{2.6}$$

and consists in minimizing the sum of leave one out square errors. We remark that this estimation technique can be extended to take into account the leave one out conditional variances $k_{\theta,n,-i}(x_i, x_i)$ and we refer to my PhD article [J1] for details.

The leave one out conditional means and variances $m_{\theta,n,-i}(x_i)$ and $k_{\theta,n,-i}(x_i, x_i)$, for $i = 1, \ldots, n$, can be computed all at once, at the cost of a single matrix inversion. Indeed, we have for $i = 1, \ldots, n$

$$m_{\theta,n,-i}(x_i) = y_i - \frac{\left[k(X, X)^{-1} y\right]_i}{\left[k(X, X)^{-1}\right]_{i,i}} \tag{2.7}$$

and

$$k_{\theta,n,-i}(x_i, x_i) = \frac{1}{\left[k(X, X)^{-1}\right]_{i,i}}. \tag{2.8}$$

The equations (2.7) and (2.8) are shown in Dubrule [1983], in the more general case of a non-zero linearly parametrized mean function. They are called virtual leave one out formulas in Bachoc [2013] and [J1]. Thanks to them, the cost of evaluating (2.6) (or other criteria based on leave one out) for a given θ is a $O(n^3)$, similarly as for ML. The CV estimation procedure (2.6), based on the virtual leave one out formulas, is for instance implemented in the `R` package `DiceKriging` Roustant et al. [2012].

Existing asymptotic results Let us now consider the asymptotic framework where $n \to \infty$ for covariance parameter estimation. In the literature, two main settings are considered: increasing and fixed-domain asymptotics Stein [1999], Zhang and Zimmerman [2005]. Traditionally, increasing-domain asymptotics corresponds to the case where, as $n \to \infty$, there exists a fixed

minimum distance between any two distinct x_i and x_j. As a consequence, the observation points can not be restricted to a bounded subset of $\mathbb{R}^d$. Conversely, fixed-domain asymptotics refers to the case where $(x_1, \ldots, x_n)$ become dense in a fixed bounded subset of $\mathbb{R}^d$ as $n \to \infty$.

The nature of the asymptotic results that can be obtained differs significantly between these two frameworks. Considered first increasing-domain asymptotics. Then, it has been shown in Mardia and Marshall [1984], Cressie and Lahiri [1993] that, in the well-specified case and for fairly general sets of stationary covariance functions $\{k_\theta, \theta \in \Theta\}$, the ML estimator $\hat{\theta}_{ML}$ converges to the true parameter θ_0. Furthermore, a central limit theorem holds with the usual parametric rate of convergence $n^{1/2}$. During my PhD thesis, I provided a result of this kind for ML in [J3]. In [J3], it is also shown that the CV estimator $\hat{\theta}_{CV}$ in (2.6) is consistent and asymptotically Gaussian distributed. An interpretation of these increasing-domain asymptotic results is that, as $n \to \infty$, there are more and more observation points very distant from each other, yielding approximate independence between the corresponding observations, and thus, so to speak, an increasing amount of information.

The situation is different under fixed-domain asymptotics. Indeed, two types of covariance parameters can be distinguished: microergodic and non-microergodic parameters Ibragimov and Rozanov [1978], Stein [1999]. A covariance parameter is microergodic if, for two different values of it, the two corresponding Gaussian measures are orthogonal, see Ibragimov and Rozanov [1978], Stein [1999]. It is non-microergodic if, even for two different values of it, the two corresponding Gaussian measures are equivalent. Non-microergodic parameters cannot be estimated consistently, but have an asymptotically negligible impact on prediction (at least in the fixed parameter case) Stein [1988, 1990a,b], Zhang [2004]. On the other hand, it is at least possible to consistently estimate microergodic covariance parameters, and misspecifying them can have a strong negative impact on prediction.

Under fixed-domain asymptotics, references indicating which covariance parameters are microergodic are for instance Stein [1999], Zhang [2004], Anderes [2010]. References providing asymptotic properties of ML estimators of microergodic parameters are Ying [1991, 1993], Loh and Lam [2000], Loh [2005], Kaufman and Shaby [2013], Bevilacqua et al. [2018]. We refer to, for instance, the introduction section in [J9] for a more detailed discussion of this topic and for further references. We would just like to point out that, generally speaking, the existing results for ML estimation under fixed-domain asymptotics are relatively scarce, and restricted to specific parametric models of covariance functions (e.g. to exponential covariance functions in Ying [1991]).

2.2 Increasing-domain asymptotic analysis of the misspecified setting [J11]

In my PhD article [J1], it is shown numerically that, in the misspecified case, the CV estimator $\hat{\theta}_{CV}$ in (2.6) can yield smaller mean square prediction errors than the ML estimator $\hat{\theta}_{ML}$ in (2.3). Furthermore, it is also pointed out in [J1] that CV is not recommended for regularly spaced observation points $x_1, \ldots, x_n$ (e.g. regular grids).

In [J11], we provide asymptotic results that confirm these findings. For each $n \in \mathbb{N}$, we con-

sider independent random observation points $(X_1, \ldots, X_n)$ such that X_i is uniformly distributed on $[0, n^{1/d}]^d$ for $i = 1, \ldots, n$. Hence, we address the case of irregularly spaced (random and independents) observation points, for which the CV principle has the most ground. Furthermore, we address an asymptotic setting that corresponds to the increasing-domain framework since the observation domain is unbounded with volume n equal to the number of observation points. It should be mentioned that, in this setting, there does not exist a fixed minimal non-zero distance between any two distinct observation points, contrary to most of the increasing-domain asymptotic literature for ML and CV for Gaussian processes. This fact significantly complicates the proofs, see [J11] for more details and discussion.

We consider noisy observations of the Gaussian process and let $y_i = \xi(X_i) + \epsilon_i$, where $(\epsilon_1, \ldots, \epsilon_n)$, $(X_1, \ldots, X_n)$ and ξ are independent and where $(\epsilon_1, \ldots, \epsilon_n)$ are independent with $\mathcal{N}(0, \delta_0)$ distribution. The observation variance $\delta_0 \geq 0$ is fixed and unknown. We consider a parametric model $\{(k_\theta, \delta_\theta), \theta \in \Theta\}$, with k_θ a stationary covariance function and $\delta_\theta > 0$, for the covariance function k and the noise variance δ_0. We set ourselves in the misspecified case and do not assume that $(k, \delta_0) \in \{(k_\theta, \delta_\theta), \theta \in \Theta\}$. Here, Θ is a compact subset of $\mathbb{R}^p$ for a fixed $p \in \mathbb{N}$.

We make the following two technical assumptions, for which $||v||$ is the Euclidean norm of a vector v.

Condition 2.1. *The covariance function k is stationary and continuous on $\mathbb{R}^d$. There exists $C_0 < +\infty$ so that for $t \in \mathbb{R}^d$,*

$$|k(t)| \leq \frac{C_0}{1 + ||t||^{d+1}}.$$

In addition, for any $l \in \mathbb{N}$, for any two-by-two distinct points $x_1, ..., x_l$, the matrix $(k(x_i - x_j))_{1 \leq i,j \leq l}$ is invertible. Finally we have $\delta_0 \geq 0$.

Condition 2.2. *For all $\theta \in \Theta$, the covariance function k_θ is stationary. For all fixed $t \in \mathbb{R}^d$, $k_\theta(t)$ is $p+1$ times continuously differentiable with respect to θ. For all $i_1, ..., i_p \in \mathbb{N}$ so that $i_1 + ... + i_p \leq p+1$, there exists $A_{i_1,...,i_p} < +\infty$ so that for all $t \in \mathbb{R}^d$, $\theta \in \Theta$,*

$$\left| \frac{\partial^{i_1}}{\partial \theta_1^{i_1}} ... \frac{\partial^{i_p}}{\partial \theta_p^{i_p}} k_\theta(t) \right| \leq \frac{A_{i_1,...,i_p}}{1 + ||t||^{d+1}}.$$

There exists a constant $C_{inf} > 0$ so that, for any $\theta \in \Theta$, $\delta_\theta \geq C_{inf}$. Furthermore, δ_θ is $p+1$ times continuously differentiable with respect to θ. For all $i_1, ..., i_p \in \mathbb{N}$ so that $i_1 + ... + i_p \leq p+1$, there exists $B_{i_1,...,i_p} < +\infty$ so that for all $\theta \in \Theta$,

$$\left| \frac{\partial^{i_1}}{\partial \theta_1^{i_1}} ... \frac{\partial^{i_p}}{\partial \theta_p^{i_p}} \delta_\theta \right| \leq B_{i_1,...,i_p}.$$

In Conditions 2.1 and 2.2, we can mention that the covariance functions are assumed to be stationary and to vanish with the rate $1/||t||^{d+1}$ as $||t|| \to \infty$. This type of assumption is standard in the increasing-domain asymptotic litterature. We require the existence of $p+1$ partial derivatives with respect to θ in the parametric model. This is a stronger requirement than in some other works under increasing-domain asymptotics (e.g. [J3]), which was here necessary because of the random and independent observation points. Nevertheless, many covariance

models yield an infinite differentiability of the covariance functions with respect to θ, for instance the Matérn model Stein [1999]. Finally, the condition $\delta_\theta \geq C_{inf}$ is crucial and enables us to bound the spectral norms of $n \times n$ inverse covariance matrices (see [J11] for more details). It is mainly in order to have the technical condition $\delta_\theta \geq C_{inf}$ that we consider observation errors, although they also correspond to a significant number of application cases, such as measure errors [J2], Monte Carlo computer experiments Le Gratiet and Garnier [2014] or the presence of a nugget effect Andrianakis and Challenor [2012].

Then, we let $\hat{\theta}_{CV}$ be the CV estimator defined in (2.6) (adapted to the presence of measure errors, see [J11] for details). We also let, for $\theta \in \Theta$,

$$E_{n,\theta} = \frac{1}{n} \int_{[0,n^{1/d}]^d} \left(m_{n,\theta}(t) - \xi(t)\right)^2 dt \tag{2.9}$$

be the integrated square prediction error, where $m_{n,\theta}(t)$ is defined as in (2.1) (with k replaced by k_θ and with an adaptation to the presence of measure errors, see [J11] for details).

In [J11] we explain that, in (2.9), the variable t is formally equivalent to a new observation point X_{n+1} so that the mean value of $E_{n,\theta}$ is equal to the mean value of the CV criterion in (2.6), with $n+1$ points instead of n. This holds notably because $(X_1, \ldots, X_n)$ are independent. From this observation, and with the rather technical proof given in [J11], whose main steps are summarized in Section A.4 there, we obtain the following theorem.

Theorem 2.3. *Under Conditions 2.1 and 2.2, we have, as $n \to \infty$,*

$$E_{n,\hat{\theta}_{CV}} = \inf_{\theta \in \Theta} E_{n,\theta} + o_p(1),$$

where the $o_p(1)$ in the above display is a function of $(X_1, \ldots, X_n)$, $(\epsilon_1, \ldots, \epsilon_n)$ and ξ only that goes to 0 in probability as $n \to \infty$.

Theorem 2.3 means that, here, CV is asymptotically optimal for the integrated square prediction error. This confirms the empirical findings in [J1]. In [J11] we also show that ML is asymptotically optimal for the Kullback Leibler divergence criterion $D_{n,\theta}$. The quantity $D_{n,\theta}$ is the normalized Kullback Leibler divergence between the true distribution of the observation vector and the one corresponding to the covariance parameter θ (see [J11]). We show that the ML estimator $\hat{\theta}_{ML}$ (adapted from (2.3) to take the observation errors into account) asymptotically minimizes $D_{n,\theta}$. The message of [J11] is thus that, in the misspecified case where there is not a single true parameter θ_0, different quality criteria are minimized by different parameters. The integrated square prediction error is then asymptotically minimized by CV and the Kullback Leibler divergence is asymptotically minimized by ML.

We illustrate this in Figure 2.1, which corresponds to Figure 1 in the online supplement to [J11] to which we refer for more details. In Figure 2.1, we consider $n = 100$ and a model where k_θ is the Matérn covariance function and where $\theta = (\sigma^2, \ell)$ where σ^2 is the variance and ℓ is the correlation length. We consider a well-specified case where $(k, \delta_0) \in \{(k_\theta, \delta_\theta), \theta \in \Theta\}$ and a misspecified case where the observation error variance is enforced to an incorrect value. We show the histograms of the ML and CV estimates of ℓ (in a Monte Carlo simulation) and the histograms of $E_{n,\hat{\theta}}$ and $D_{n,\hat{\theta}}$ where $\hat{\theta}$ is the ML or CV estimator. We observe that, in the well-specified case, ML is preferable to CV in all aspects: ML provides a better estimation of

the true value of ℓ and smaller values of $E_{n,\hat{\theta}}$ and $D_{n,\hat{\theta}}$. However, in the misspecified case, CV provides smaller values of $E_{n,\hat{\theta}}$ while ML provides smaller values of $D_{n,\hat{\theta}}$. Furthermore, in the misspecified case, the ML and CV histograms of ℓ are centered at different values.

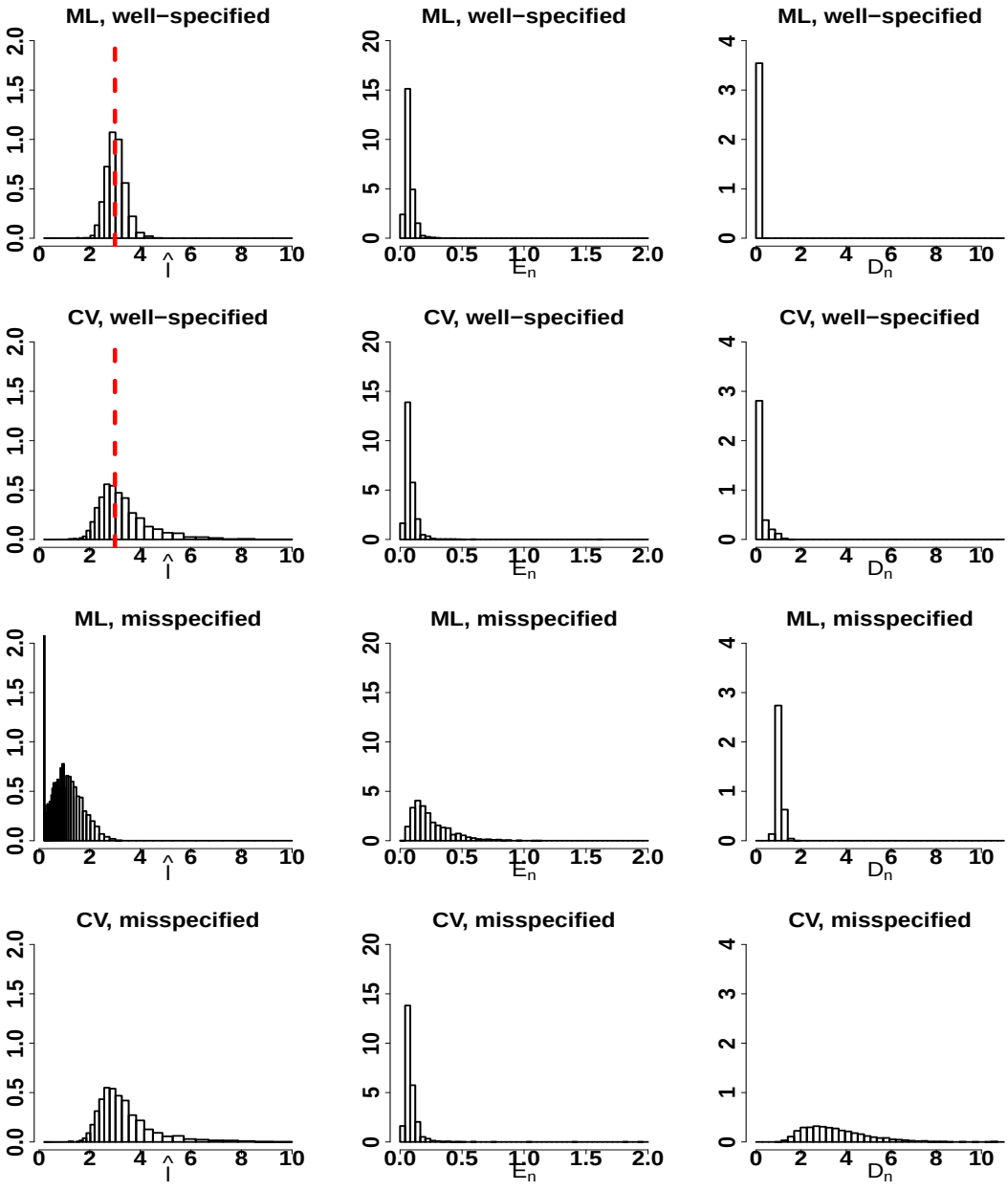


Figure 2.1: Illustration of [J11] in Section 2.2 for $n = 100$. The histograms of $\hat{\ell}$, $D_{n,\hat{\theta}}$ and $E_{n,\hat{\theta}}$ are reported for ML and CV and in the well-specified and misspecified cases. In the well-specified case, the dashed red vertical lines denote the true value of ℓ.

2.3 Increasing-domain asymptotic analysis of multivariate tapering [J6]

Evaluating the likelihood function in (2.3) has a computational cost of $O(n^3)$ in time and $O(n^2)$ in storage. Hence, it typically becomes too costly to evaluate this function (let alone optimizing it) when n is much larger than, say, $10,000$. On the other hand there is a need to exploit large data sets, for instance with n above 10^6 with satellite data. Among the variety of techniques to tackle this issue (see for instance the references in the introduction section in [J10]), we shall focus on covariance tapering here Kaufman et al. [2008], Furrer et al. [2006].

We consider a taper function $T : \mathbb{R}^d \to \mathbb{R}$ with compact support, for which the function $x_1, x_2 \to T(x_1 - x_2)$ is a symmetric non-negative definite function on $\mathbb{R}^d \times \mathbb{R}^d$. We also consider a taper range $\gamma_n > 0$ and let $\bar{k}_\theta : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be the tapered covariance function, defined by $\bar{k}_\theta(x_1, x_2) = k_\theta(x_1, x_2)T([x_1 - x_2]/\gamma_n)$. The tapered likelihood estimator is then defined by

$$\hat{\theta}_{tML} \in \operatorname{argmin}_{\theta\in\Theta} \log(\left|\bar{k}_\theta(X, X)\right|) + y^\top \bar{k}_\theta(X, X)^{-1} y. \tag{2.10}$$

The benefit of tapering is that the covariance matrix $\bar{k}_\theta(X, X)$ is sparse, because the function $\bar{k}_\theta$ has a compact support. Thus, sparse linear algebra procedures can be used, that allow for much larger values of n than when handling full matrices (see for instance the `R` package `SPAM` Furrer and Sain [2010]). The taper range γ_n is chosen by the user. Small values of γ_n yield sparser matrices, which is computationally beneficial, but yields a larger difference between $\bar{k}_\theta$ and k_θ, which degrades the statistical accuracy. The opposite occurs when γ_n is large: the tapered likelihood function is closer to the (untapered) likelihood function but is more costly to evaluate. Finally, let us mention that (2.10) corresponds to the one-taper equation and that a two-taper equation also exists Kaufman et al. [2008], Furrer et al. [2006]. The two-taper equation yields an estimator with a better bias but more difficult to compute, and for which the asymptotic analysis is different, see [J6] for more details.

In [J6], we provide an asymptotic analysis of the tapered maximum likelihood estimator, in the case of a stationary multivariate Gaussian process. A multivariate Gaussian process is a collection of Gaussian processes $\xi^{(r)} = (\xi_1, \ldots, \xi_r)$, with $r \in \mathbb{N}$ fixed, so that any linear combination of $(\xi_1, \ldots, \xi_r)$ is also a Gaussian process. We consider that $\xi_1, \ldots, \xi_r$ have mean function zero. Hence, $\xi^{(r)}$ is characterized by its matrix covariance function $k^{(r)} = (k_{a,b})_{a,b=1,\ldots,r}$, with $k_{a,b} : \mathbb{R}^d \to \mathbb{R}$ ($k_{a,b}$ is stationary) for $a, b = 1, \ldots, r$. We have $\operatorname{cov}(\xi_a(x_1), \xi_b(x_2)) = k_{a,b}(x_1 - x_2)$ for $a, b \in \{1, \ldots, r\}$ and $x_1, x_2 \in \mathbb{R}^d$. For $a \in \{1, \ldots, r\}$, $k_{a,a}$ is a stationary covariance function and for $a, b \in \{1, \ldots, r\}$, $a \neq b$, $k_{a,b}$ is called a (stationary) cross covariance function. The matrix covariance function is symmetric non-negative definite, that is, for any $x_1, \ldots, x_m \in \mathbb{R}^d$, the $rm \times rm$ matrix

$$(k_{a,b}(x_i - x_j))_{(a,i)\in\{1,\ldots,r\}\times\{1,\ldots,m\},(b,j)\in\{1,\ldots,r\}\times\{1,\ldots,m\}}$$

is symmetric non-negative definite. Examples of existing matrix covariance functions can be found in Genton et al. [2015], Gneiting et al. [2010]. The multivariate process $\xi^{(r)}$ is observed at the observation points $X = (x_1, \ldots, x_n)$.

Based on a matrix covariance function, conditioning is carried out by a natural extension of (2.1) and (2.2). We refer to [J6] for details. When considering a parametric set

$\{k_\theta^{(r)} = (k_{a,b,\theta})_{a,b=1,\ldots,r}, \theta \in \Theta\}$ of stationary matrix covariance functions, the tapered likelihood estimator is defined by a natural extension of (2.10). In particular, we consider a matrix taper function $T^{(r)} = (T_{a,b})_{a,b=1,\ldots,r}$, with $T_{a,b} : \mathbb{R}^d \to \mathbb{R}$ with compact support for $a, b = 1, \ldots, r$. We also consider a taper rang sequence $(\gamma_n)_{n\in\mathbb{N}}$ with $\gamma_n > 0$. Finally, we consider the well-specified case here and let $k_{\theta_0}^{(r)}$ be the true matrix covariance function of $\xi^{(r)}$ with $\theta_0 \in \Theta$.

Then ML estimator is of the form

$$\hat{\theta}_{ML} \in \text{argmin}_{\theta\in\Theta} L_\theta \tag{2.11}$$

and the tapered ML estimator is of the form

$$\hat{\theta}_{tML} \in \text{argmin}_{\theta\in\Theta} \bar{L}_\theta \tag{2.12}$$

where the expressions of L_θ and $\bar{L}_\theta$ are similar to (2.3) and (2.10) and are given in [J6].

We define the Fourier transform of a function $f : \mathbb{R}^d \to \mathbb{R}$ by $\tilde{f}(\omega) = (2\pi)^{-d} \int_{\mathbb{R}^d} e^{-\imath w^\top x} f(x)dx$ with $\imath^2 = -1$. We make the following assumptions for our asymptotic results below.

Condition 2.4. *For all fixed $x \in \mathbb{R}^d$, $a, b = 1, \ldots, r$, $k_{a,b,\theta}(x)$ is continuously differentiable with respect to θ. There exist constants $A < +\infty$ and $\alpha > 0$ so that for all $i = 1, \ldots, p$, for all $x \in \mathbb{R}^d$ and for all $\theta \in \Theta$,*

$$|k_{a,b,\theta}(x)| \leq \frac{A}{1 + ||x||^{d+\alpha}} \quad \text{and} \quad \left|\frac{\partial}{\partial\theta_i} k_{a,b,\theta}(x)\right| \leq \frac{A}{1 + ||x||^{d+\alpha}}.$$

The Fourier transforms $\tilde{k}_{a,b,\theta}(\omega)$ are jointly continuous in ω and θ and the inverse Fourier transform thereof exist. The smallest eigenvalue of the matrix $(\tilde{k}_{a,b,\theta}(\omega))_{a,b=1\ldots,r}$ is strictly positive for all w and θ.

Condition 2.5. *For all $a, b = 1, \ldots, r$, the taper function $T_{a,b}$ is continuous at 0 and satisfies $T_{a,b}(0) = 1$ and $|T_{a,b}(x)| \leq 1$ for all $x \in \mathbb{R}^d$. The taper range $\gamma = \gamma_n$ satisfies $\gamma_n \to_{n\to\infty} +\infty$.*

Condition 2.6. *There exists a constant $\Delta > 0$ so that for all $n \in \mathbb{N}$ and for all $a \neq b$, $||x_a - x_b|| \geq \Delta$.*

Condition 2.4 implies a smoothness with respect ot θ and a decrease with respect to $||x||$ of the covariance and cross covariance functions, similarly to Condition 2.2 in Section 2.2. We remark that, for a model of stationary cross covariance functions, the (complex) matrix $(\tilde{k}_{a,b,\theta}(\omega))_{a,b=1\ldots,r}$ is Hermitian non-negative definite Wackernagel [2003]. Hence, we require a little more in Condition 2.4 when imposing that the eigenvalues of $(\tilde{k}_{a,b,\theta}(\omega))_{a,b=1\ldots,r}$ are non-zero. This enables us to have an asymptotic lower bound on the eigenvalues of the $nr \times nr$ covariance matrices obtained from $k_\theta^{(r)}$, see Section 2.4 and [J5].

Condition 2.5 is very mild and means that the functions $k_{a,b,\theta}$ and $k_{a,b,\theta}T_{a,b}(\cdot/\gamma)$ will be close to each other when γ is large. Finally Condition 2.6 is a standard increasing-domain asymptotics condition, as is discussed above.

Next, we provide the consistency results in [J6] for the tapered maximum likelihood estimator. First we show that the likelihood and tapered likelihood criteria are asymptotically equivalent.

Theorem 2.7. *Assume that Conditions 2.4, 2.5 and 2.6 hold. Then, as* $n \to \infty$,

$$\sup_{\theta \in \Theta} |L_\theta - \bar{L}_\theta| = o_p(1).$$

As a consequence, under mild conditions on the likelihood function, implying the consistency of the ML estimator, the tapered ML estimator is also consistent.

Corollary 2.8. *Consider the same setting as in Theorem 2.7. Assume that for all* $\kappa > 0$ *there exists* $\epsilon > 0$ *so that*

$$\inf_{||\theta - \theta_0|| \geq \kappa} L_\theta - L_{\theta_0} \geq \epsilon + o_p(1),$$

where the $o_p(1)$ *may depend on* ϵ *and* κ *and goes to* 0 *in probability as* $n \to \infty$. *Then, as* $n \to \infty$,

$$\widehat{\theta}_{ML} \to_p \theta_0 \qquad \text{and} \qquad \widehat{\theta}_{tML} \to_p \theta_0.$$

We now consider the prediction of ξ_1 and let $m_{n,\theta}(u)$ be the conditional mean of $\xi_1(u)$ given $(\xi_a(x_i))_{a=1,\ldots,r,i=1,\ldots,n}$, under the matrix covariance function $k_\theta^{(r)}$. We also let $\bar{m}_{n,\theta}(u)$ be the conditional mean of $\xi_1(u)$ given $(\xi_a(x_i))_{a=1,\ldots,r,i=1,\ldots,n}$, under the tapered matrix covariance function $(k_{a,b,\theta} T_{a,b}(\cdot/\gamma_n))_{a,b=1\ldots,r}$. In both cases we refer to [J6] for the explicit expressions. Then, we show in [J6] that the predictions obtained with and without tapering are asymptotically equivalent.

Theorem 2.9. *Assume that Conditions 2.4, 2.5 and 2.6 hold. Let* $(x_{\text{new},n})_{n \in \mathbb{N}}$ *be a fixed sequence in* $\mathbb{R}^d$. *Then, as* $n \to \infty$,

$$\sup_{\theta \in \Theta} \left| [m_{n,\theta}(x_{\text{new},n}) - \xi_1(x_{\text{new},n})]^2 - [\bar{m}_{n,\theta}(x_{\text{new},n}) - \xi_1(x_{\text{new},n})]^2 \right| = o_p(1). \tag{2.13}$$

Assume furthermore that for any fixed θ, a *and* b, *the functions* $k_{a,b,\theta}$ *and* $T_{a,b}$ *are continuous. Let* $\mathcal{D}_n$ *be a sequence of measurable subsets of* $\mathbb{R}^d$ *with positive Lebesgue measures and let* f_n *be a sequence of continuous probability density functions on* $\mathcal{D}_n$. *Then, as* $n \to \infty$,

$$\sup_{\theta \in \Theta} \left| \int_{\mathcal{D}_n} [m_{n,\theta}(x) - \xi_1(x)]^2 f_n(x) dx - \int_{\mathcal{D}_n} [\bar{m}_{n,\theta}(x) - \xi_1(x)]^2 f_n(x) dx \right| = o_p(1). \tag{2.14}$$

We illustrate Theorem 2.7 and Corollary 2.8 in Figure 2.2. For two settings (see [J6] for their specifications), we estimate a standard deviation $\sigma_{1,2}$ and a correlation length $\rho_{1,1}$. For $r = 2$ (bivariate case) we consider different values of n and of the taper range γ. We observe that, for a fixed γ, as n increases, the variance decreases significantly but the bias remains non-zero and approximately constant. On the contrary, for fixed n, the bias decreases as γ increases. We also observe that the tapered ML estimator becomes more and more accurate as both n and γ increase, which illustrates the consistency result.

2.4 Smallest eigenvalues of covariance matrices of multivariate processes [J5]

For the proofs of Theorems 2.7 and 2.9, and for proofs in other references Shaby and Ruppert [2012], Bevilacqua et al. [2015], a key element is the existence of a lower bound on the smallest

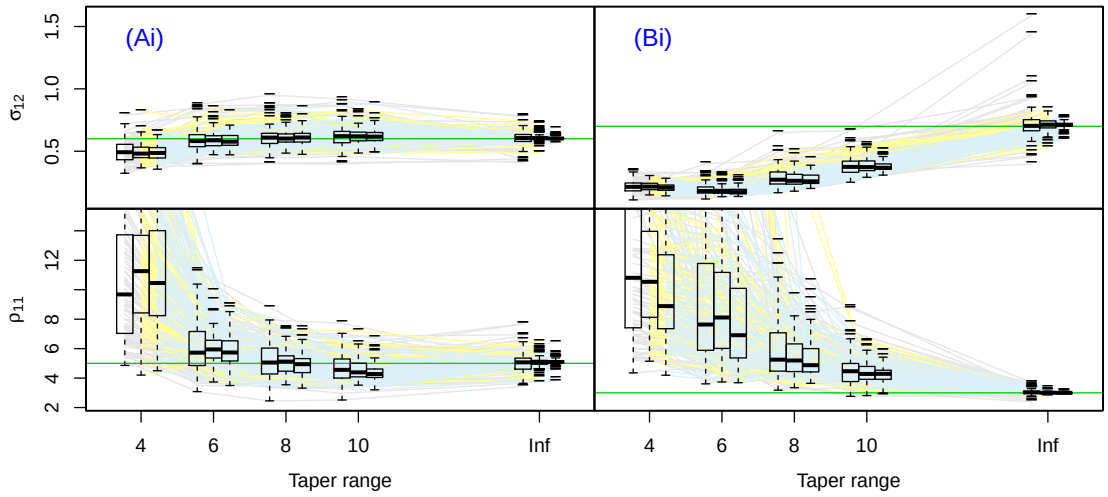


Figure 2.2: Illustration of [J6] in Section 2.3. Effect of increasing n and the taper range γ. The boxplots correspond to $n = 400$ (gray), 1024 (yellow), 2500 (light blue), left to right for each taper range. The horizontal lines denote the true covariance parameter values. The case $\gamma = \infty$ corresponds to the untapered ML estimator. The two simulation settings are (Ai) and (Bi).

eigenvalues of the covariance matrices corresponding to various parameters θ. For instance, in the context of Section 2.3, the $nr \times nr$ covariance matrix $k_\theta^{(r)}(X, X)$ defined by

$$(k_{a,b,\theta}(x_i, x_j)_{(a,i)\in\{1,\dots,r\}\times\{1,\dots,n\},(b,j)\in\{1,\dots,r\}\times\{1,\dots,n\}},$$

is the covariance matrix of the $nr \times 1$ observation vector, under covariance parameter θ.

In Shaby and Ruppert [2012], Bevilacqua et al. [2015], the existence of this lower bound is assumed. The contribution of [J5], that we now present, is to show that this lower bound holds under conditions that are relatively mild and simple to check, for given models of matrix covariance functions. These conditions are implied by Conditions 2.4, 2.5 and 2.6, so that the lower bound need not be assumed for Theorems 2.7 and 2.9. In fact, working on the article [J6] motivated us to study the eigenvalue lower bound in [J5].

Let us consider again a multivariate Gaussian process $\xi^{(r)} = (\xi_1, \dots, \xi_r)$, with $r \in \mathbb{N}$, with stationary matrix covariance function $k^{(r)}$. We consider r sequences of points $(x_i^{(1)})_{i\in\mathbb{N}}, \dots, (x_i^{(r)})_{i\in\mathbb{N}}$. For $a = 1, \dots, r$, $(x_i^{(a)})_{i\in\mathbb{N}}$ is the sequence of observation points for ξ_a. We thus remark that we allow here for the different processes to be observed at different observation points (non-collocated observations) while the processes are observed at the same points in Section 2.3 (collocated). We have the following conditions.

Condition 2.10. *There exists a finite fixed constant $A > 0$ and a fixed constant $\tau > 0$ so that the functions $k_{a,b}$, $a, b = 1, \dots, r$, satisfy, for all $x \in \mathbb{R}^d$,*

$$|k_{a,b}(x)| \leq \frac{A}{1 + ||x||^{d+\tau}}. \tag{2.15}$$

We define the Fourier transform of a function $g : \mathbb{R}^d \to \mathbb{R}$ as in Section 2.3. Then, from (2.15), the covariance functions $k_{a,b}$ have Fourier transforms $\tilde{k}_{a,b}$ that are continuous and bounded. Also, note that, for any $\omega \in \mathbb{R}^d$, $\tilde{k}^{(r)}(\omega) = \{\tilde{k}_{a,b}(\omega)\}_{a,b=1,\dots,r}$ is a Hermitian complex matrix, that has real non-negative eigenvalues $0 \leq \lambda_1\{\tilde{k}^{(r)}(\omega)\} \leq \cdots \leq \lambda_r\{\tilde{k}^{(r)}(\omega)\}$.

Condition 2.11. *We have, for $a, b = 1, \ldots, r$,*

$$k_{a,b}(x) = \int_{\mathbb{R}^d} \tilde{k}_{a,b}(\omega) e^{\imath \omega^\top x} \mathrm{d}\omega. \tag{2.16}$$

Also, we have $0 < \lambda_1\{\tilde{k}^{(r)}(\omega)\}$ for all $\omega \in \mathbb{R}^d$.

The condition (2.16) is very weak and satisfied by most standard covariance and cross covariance functions. On the other hand, the condition $0 < \lambda_1\{\tilde{k}^{(r)}(\omega)\}$ for all $\omega \in \mathbb{R}^d$ is less innocuous, and is further discussed at the end of the section.

We assume, similarly as in Section 2.3, the existence of a minimal distance between two different observation points of the same process.

Condition 2.12. *There exists a fixed $\Delta > 0$ so that for all $a = 1, \ldots, r$, $\inf_{i,j \in \mathbb{N}; i \neq j} |x_i^{(a)} - x_j^{(a)}| \geq \Delta$.*

For all $n_1, \ldots, n_p \in \mathbb{N}$, let, for $0 \leq a \leq r$, $N_a = n_1 + \cdots + n_a$, with the convention that $N_0 = 0$. Let also $N = N_r$. Then, let Σ be the $N \times N$ covariance matrix, filled as follows: For $u = N_{a-1} + i$ and $v = N_{b-1} + j$, with $1 \leq a, b \leq r$, $1 \leq i \leq n_a$ and $1 \leq j \leq n_b$, $\Sigma_{u,v} = k_{a,b}(x_i^{(a)} - x_j^{(b)})$.

The eigenvalue lower bound of [J5] is then the following, with $\lambda_1(A)$ the smallest eigenvalue of a symmetric real matrix A.

Theorem 2.13. *Assume that Conditions 2.10, 2.11 and 2.12 are satisfied. Then, we have*

$$\inf_{n_1, \ldots, n_p \in \mathbb{N}} \lambda_1(\Sigma) > 0.$$

We remark that the version of Theorem 2.13 corresponding to $r = 1$ (univariate case) is shown in my PhD article [J3]. The proof of Theorem 2.13 extends that in [J3], and necessitates to introduce complex Hermitian matrices and matrix Fourier functions.

In [J5] we also extend Theorem 2.13 to the case of a parametric family of stationary covariance and cross covariance functions $\{k_\theta^{(r)} = (k_{a,b,\theta})_{a,b=1,\ldots,r}, \theta \in \Theta\}$ with Θ compact in $\mathbb{R}^p$. We have the following, with the same notation as in Section 2.3.

Theorem 2.14. *Assume that, for all $\theta \in \Theta$, the functions $\{k_{a,b,\theta}\}_{1 \leq a,b \leq r}$ satisfy Condition 2.10, where A and τ can be chosen independently of θ. Assume that (2.16) holds with $\{k_{a,b}\}_{1 \leq a,b \leq r}$ replaced by $\{k_{a,b,\theta}\}_{1 \leq a,b \leq r}$, for all $\theta \in \Theta$. Assume that $\tilde{k}_\theta^{(r)}(\omega) = (\tilde{k}_{a,b,\theta}(\omega))_{a,b=1,\ldots,r}$ is jointly continuous in ω and θ with strictly positive smallest eigenvalue for all ω and θ. Assume finally that Condition 2.12 is satisfied.*

Then, with Σ_θ being as Σ with $\{k_{a,b}\}_{1 \leq a,b \leq r}$ replaced by $\{k_{a,b,\theta}\}_{1 \leq a,b \leq r}$, we have

$$\inf_{\theta \in \Theta, n_1, \ldots, n_p \in \mathbb{N}} \lambda_1(\Sigma_\theta) > 0.$$

It is known that if $\lambda_1\{\tilde{k}^{(r)}(\omega)\} > 0$ for almost all $\omega \in \mathbb{R}^d$, then $\lambda_1(\Sigma) > 0$ whenever the points $(x_i^{(a)})_{1 \leq i \leq n_a}$ are two-by-two distinct for all $a = 1 \ldots, r$. We show nevertheless in [J5] that this condition can be insufficient for Theorem 2.13 to hold. More precisely, we exhibit a counterexample in the univariate case and in dimension one, with the triangular covariance function.

2.5 Fixed-domain asymptotic results for the exponential covariance function [J8,J9]

In this section, we consider fixed-domain asymptotics. As we have discussed, increasing domain asymptotic results for ML and CV can typically be obtained for general parametric families of covariance functions (for instance as in Sections 2.2, 2.3 and 2.4). On the other hand, fixed-domain asymptotic results for ML are typically obtained for more restrictive classes of covariance functions, which specific structures can be exploited. Important examples are exponential covariance functions Ying [1991, 1993], Matérn covariance functions in dimension $d \leq 3$ Kaufman and Shaby [2013], squared exponential covariance functions Loh and Lam [2000], Matérn 3/2 covariance functions Loh [2005] and generalized Wendland covariance functions Bevilacqua et al. [2018]. In addition, there are no fixed-domain asymptotic results for CV, to our knowledge, at the exception of [J9] that we present here.

In this section, we focus on (stationary) exponential covariance functions k_{ρ,σ^2} on $\mathbb{R}$, defined by, for $(\rho, \sigma^2) \in (0,\infty)^2$ and for $t \in \mathbb{R}$, $k_{\rho,\sigma^2}(t) = \sigma^2 e^{-\rho|t|}$. A Gaussian process ξ with covariance function k_{ρ,σ^2} is called a stationary Ornstein-Uhlenbeck process and is Markovian. Considering these covariance functions makes the analysis of ML much more tractable under fixed-domain asymptotics Ying [1991]. Similarly, inverse covariance matrices obtained from k_{ρ,σ^2} are tridiagonal with explicit expressions Antognini and Zagoraiou [2010]. This also makes the analysis of CV (see (2.7) and (2.8)) more tractable.

Cross validation [J9] In [J9], we consider a Gaussian process ξ on $[0,1]$ with covariance function k_{θ_0} with $\theta_0 = (\rho_0, \sigma_0^2) \in (0,\infty)^2$ and $k_{\theta_0}(t) = \sigma_0^2 e^{-\rho_0|t|}$. For each $n \in \mathbb{N}$ the process ξ is observed at $x_1^{(n)}, \ldots, x_n^{(n)}$ where $(x_i^{(n)})_{n\in\mathbb{N}, i=1\ldots,n}$ is a triangular array of points in $[0,1]$. We let $(x_1, \ldots, x_n) = (x_1^{(n)}, \ldots, x_n^{(n)})$ for concision.

We consider the parametric model $\{k_\theta; \theta \in \Theta = [a,A] \times [b,B]\}$ with fixed $0 < a \leq A < \infty$ and $0 < b \leq B < \infty$, and with $\theta = (\rho, \sigma^2)$ and k_{ρ,σ^2} defined as above. We study the CV estimator $\hat{\theta}_{lCV}$ based on the logarithmic score, defined by

$$\hat{\theta}_{lCV} \in \operatorname{argmin}_{\theta\in\Theta} \sum_{i=1}^{n} \left(\log\left(k_{\theta,n,-i}(x_i,x_i)\right) + \frac{\left(\xi(x_i) - m_{\theta,n,-i}(x_i)\right)^2}{k_{\theta,n,-i}(x_i,x_i)} \right) \tag{2.17}$$

with the notation of Section 2.1. The rationale for this CV estimator is that $\log(2\pi)+\log(k_{\theta,n,-i}(x_i,x_i))+[\xi(x_i)-m_{\theta,n,-i}(x_i))^2]/k_{\theta,n,-i}(x_i,x_i)$ is equal to -2 times the conditional log-likelihood of $\xi(x_i)$, given $(\xi(x_1),...,\xi(x_{i-1}),\xi(x_{i+1}),...,\xi(x_n))$, under the covariance parameters ρ, σ^2. This estimator is used in Rasmussen and Williams [2006], Zhang and Wang [2010].

We remark that we are in the well-specified setting for covariance parameter estimation. As already known Ibragimov and Rozanov [1978], Ying [1991], Zhang [2004], the parameters ρ_0 and σ_0^2 are non-microergodic and can not be estimated consistently. The product $\rho_0\sigma_0^2$ is microergodic and Ying [1991] shows the consistency and asymptotic normality of the ML estimator $\hat{\rho}_{ML}\hat{\sigma}^2_{ML}$ obtained from (2.3).

In [J9] we show the strong consistency of $\hat{\rho}_{lCV}\hat{\sigma}^2_{lCV}$. In the sequel, we let $\Delta_i = x_i - x_{i-1}$ for $n \in \mathbb{N}$, $i = 2, \ldots, n$.

Theorem 2.15. *Assume that*

$$\limsup_{n\to+\infty} \max_{i=2,\dots,n} \Delta_i = 0. \tag{2.18}$$

Assume that there exists $(\tilde{\rho}, \tilde{\sigma}^2)$ *in* Θ *so that* $\tilde{\rho}\tilde{\sigma}^2 = \theta_0\sigma_0^2$. *Then we have*

$$\hat{\rho}_{lCV}\hat{\sigma}^2_{lCV} \overset{a.s.}{\to} \rho_0\sigma_0^2. \tag{2.19}$$

Then, we show a central limit theorem in [J9].

Theorem 2.16. *Consider the same assumptions as in Theorem 2.15. Assume further that either* $aB < \rho_0\sigma_0^2$; $Ab > \rho_0\sigma_0^2$ *or* $aB > \rho_0\sigma_0^2$; $Ab < \rho_0\sigma_0^2$ *hold. Then we have*

$$\frac{\sqrt{n}}{\rho_0\sigma_0^2\tau_n}(\hat{\rho}_{lCV}\hat{\sigma}^2_{lCV} - \rho_0\sigma_0^2) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,1). \tag{2.20}$$

The quantity τ_n^2 *depends on how the underlying design points* $\{x_1,\dots,x_n\}$ *have been chosen. More precisely,*

$$\tau_n^2 = \frac{2}{n}\sum_{i=3}^{n-1}\left[\left(\frac{\Delta_{i+1}}{\Delta_i+\Delta_{i+1}} + \frac{\Delta_{i-1}}{\Delta_i+\Delta_{i-1}}\right)^2 + 2\frac{\Delta_i\Delta_{i+1}}{(\Delta_i+\Delta_{i+1})^2}\right]. \tag{2.21}$$

In Theorem 2.16, the condition $aB < \rho_0\sigma_0^2$; $Ab > \rho_0\sigma_0^2$ or $aB > \rho_0\sigma_0^2$; $Ab < \rho_0\sigma_0^2$ ensures that the derivative with respect to ρ or σ^2 of the sum in (2.17) will be equal to zero for n large enough almost surely, by applying Theorem 2.15. This is used in the proof of Theorem 2.16. A similar assumption is made in Ying [1991], where the parameter domain for (ρ,σ^2) is $(0,\infty)\times[b,B]$ or $[a,A]\times(0,\infty)$.

In the following proposition, we show that the quantity τ_n^2 in Theorem 2.16 is lower and upper bounded, so that the rate of convergence is always $\sqrt{n}$ in this theorem.

Proposition 2.17. *We have, for any choice of the triangular array of design points* $\{x_1,...,x_n\}$ *satisfying* (2.18),

$$2 \le \liminf_{n\to\infty}\tau_n^2 \le \limsup_{n\to\infty}\tau_n^2 \le 4. \tag{2.22}$$

By the previous proposition, the asymptotic variance of the limiting distribution of $\hat{\rho}_{lCV}\hat{\sigma}^2_{lCV} - \rho_0\sigma_0^2$ is always larger than that of the ML estimator. Indeed, we have $(\sqrt{n}/[\rho_0\sigma_0^2])(\hat{\rho}_{ML}\hat{\sigma}^2_{ML} - \rho_0\sigma_0^2) \xrightarrow[n\to\infty]{\mathcal{D}} \mathcal{N}(0,2)$, see Ying [1991]. This fact is quite expected as ML estimates usually perform best when the covariance model is well-specified, as is the case here. It is also interesting to remark that the asymptotic variance is always the same for ML while it depends on the triangular array of observation points for CV.

As one can check easily, the regular design, defined by $\Delta_i = \frac{1}{n-1}$ for all $i = 2,\dots,n$, does not yield the limiting variance of the ML estimator. Instead, we have $\tau_n^2 \to_{n\to\infty} 3$ for this design. However, in Proposition 2.18, we exhibit a particular design realizing the limiting variance of the ML estimator: $\lim_{n\to\infty}\tau_n^2 = 2$. In fact, the bounds in (2.22) are sharp as shown in the following proposition.

Proposition 2.18. *(i) Let* $\{x_1,\dots,x_n\}$ *be such that* $x_1 = 0$, *for* $i = 2,...,n-1$,

$$\Delta_i = \begin{cases} (1-\gamma_n)\frac{2}{n} & \text{if } i \text{ is even,} \\ \frac{2\gamma_n}{n} & \text{if } i \text{ is odd,} \end{cases}$$

where $\gamma_n \in (0,1)$*, and* $\Delta_n = 1 - \sum_{i=2}^{n-1} \Delta_i$*. Then, taking* $\gamma_n = 1/n$*, we get* $\tau_n^2 \underset{n\to\infty}{\to} 4$*.*

(ii) Let $\{x_1, \dots, x_n\}$ *and* $0 < \alpha < 1$ *be such that* $x_1 = 0$*,* $\Delta_i = 1/(i\,!)$ *for* $i = \lfloor n^\alpha \rfloor + 1, \dots, n$ *and* $\Delta_2 = \cdots = \Delta_{\lfloor n^\alpha \rfloor} = (1 - r_n)/(\lfloor n^\alpha \rfloor - 1)$ *with* $r_n = \sum_{i=\lfloor n^\alpha \rfloor + 1}^{n} \Delta_i$*. Then* $\sum_{i=2}^{n} \Delta_i = 1$ *and* $\tau_n^2 \underset{n\to\infty}{\to} 2$*.*

Finally, for proving Theorems 2.15 and 2.16, we follow the same general proof architecture as in Ying [1991] for ML, but our proofs contain several new elements. In particular, the computations are globally more involved, Taylor expansions with two variables (each variable being an interpoint distance Δ_i) are needed and central limit theorems for dependent random variables are used. We refer to [J9] for more discussion and for the proofs.

Maximum likelihood in the bivariate case [J8] In [J8] we consider a bivariate Gaussian process $\xi^{(2)} = (\xi_1, \xi_2)$ with matrix covariance function $k_{\theta_0}^{(2)} = (k_{i,j,\theta_0})_{i,j=1,2}$. We let $\theta_0 = (\rho_0, \sigma_{1,0}^2, \sigma_{2,0}^2, c_0) \in (0,\infty)^3 \times (-1,1)$ and we let for $i, j \in \{1,2\}$ and $x_1, x_2 \in \mathbb{R}$,

$$\operatorname{cov}(\xi_i(x_1), \xi_j(x_2)) = k_{i,j,\theta_0}(x_1 - x_2) = \sigma_{i,0}\sigma_{j,0}\left[\mathbf{1}_{i=j} + c_0 \mathbf{1}_{i\neq j}\right] e^{-\rho_0 |x_1 - x_2|}.$$

The parameters $\sigma_{1,0}^2$ and $\sigma_{2,0}^2$ are the variances of ξ_1 and ξ_2. The parameter ρ_0 is an inverse correlation length and c_0 is the correlation between $\xi_1(x)$ and $\xi_2(x)$ for all $x \in \mathbb{R}$. We remark that here the correlation functions of ξ_1 and ξ_2 are the same and that the correlation between $\xi_1(x)$ and $\xi_2(x)$ does not depend on x. While this can be a restriction in practice, it has two main benefits. First, it is guaranteed that $k_{\theta_0}^{(2)}$ is indeed a matrix covariance function for any $\theta_0 \in (0,\infty)^3 \times (-1,1)$. Second, when ξ_1 and ξ_2 are observed at the same n observation points $x_1, \dots, x_n$, then the resulting $2n \times 2n$ covariance matrix can be written as a Kronecker product involving the $n \times n$ correlation matrix $(e^{-\rho_0 |x_i - x_j|})_{i,j=1,\dots,n}$ (which inverse is tridiagonal and has an explicit expression as discussed above). We refer to [J8] for more details. This last fact is crucial for the analysis in [J8] (furthermore, this Kronecker product also entails a computational benefit for evaluating the likelihood criterion).

We consider the parametric model of bivariate covariance functions $\{k_\theta^{(2)}; \theta \in \Theta\}$, where $\theta = (\rho, \sigma_1^2, \sigma_2^2, c)$, where $k_\theta^{(2)}$ is defined as $k_{\theta_0}^{(2)}$ with θ_0 replaced by θ and where Θ is a compact subset of $(0,\infty)^3 \times (-1,1)$.

In Zhang and Cai [2015], it is shown that the parameters ρ, σ_1^2 and σ_2^2 are non-microergodic and that the parameters $\rho\sigma_1^2$, $\rho\sigma_2^2$ and c are microergodic.

Let us consider that for each $n \in \mathbb{N}$, ξ_1 and ξ_2 are observed at $(x_1, \dots, x_n) = (x_1^{(n)}, \dots, x_n^{(n)})$ where $(x_i^{(n)})_{n\in\mathbb{N}, i=1,\dots,n}$ is a triangular array of points in $[0,1]$, so that $(x_1, \dots, x_n)$ are two by two distinct and become dense in $[0,1]$ as $n \to \infty$. Let us consider the ML estimator $\hat{\theta}_{ML} = (\hat{\rho}_{ML}, \hat{\sigma}_{1,ML}^2, \hat{\sigma}_{2,ML}^2, \hat{c}_{ML})$ of θ_0, defined as in (2.11) (see also [J8]). Then, in [J8], we show the consistency and asymptotic normality of the ML estimators of the microergodic parameters. We state the asymptotic normality result here.

Theorem 2.19. *If* θ_0 *belongs to the interior of* Θ *then, as* $n \to \infty$*,*

$$\sqrt{n}\begin{pmatrix} \hat{\rho}_{ML}\hat{\sigma}_{1,ML}^2 - \rho_0\sigma_{1,0}^2 \\ \hat{\rho}_{ML}\hat{\sigma}_{2,ML}^2 - \rho_0\sigma_{2,0}^2 \\ \hat{c} - c_0 \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \Sigma), \tag{2.23}$$

where $\Sigma = \begin{pmatrix} 2(\rho_0\sigma_{1,0}^2)^2 & 2(\rho_0 c_0\sigma_{01}\sigma_{02})^2 & \rho_0 c_0\sigma_{1,0}^2(1-c_0^2) \\ 2(\rho_0 c_0\sigma_{01}\sigma_{02})^2 & 2(\rho_0\sigma_{2,0}^2)^2 & \rho_0 c_0\sigma_{2,0}^2(1-c_0^2) \\ \rho_0 c_0\sigma_{1,0}^2(1-c_0^2) & \rho_0 c_0\sigma_{2,0}^2(1-c_0^2) & (c_0^2-1)^2 \end{pmatrix}$.

We remark that the asymptotic covariance matrix does not depend on the triangular array of observation points, similarly as ML in the univariate case Ying [1991] and contrary to CV [J9].

Chapter 3

Other contributions to Gaussian processes

3.1 Application to metamodeling in nuclear engineering [J4]

In the application article [J4], we investigate the use of Gaussian processes for metamodeling of computer codes in nuclear engineering. We consider the Germinal code Roche and Pelletier [2000] for which Gaussian processes globally compare favorably to kernel methods and neural networks. In addition, we show how they can help interpreting the behavior and instabilities of the code.

The context The Germinal code enables to simulate the thermalmechanical behavior of a fuel pin. A fuel pin is a nuclear component which is a part of a fuel assembly in fast breeding reactors. A schematic illustration is given in Figure 3.1. In our setting, the Germinal code works schematically as follows. A number of scalar parameters of interest is chosen. A preprocessor is built, which constructs, for each vector of these parameters, a (more complex) input file which can be interpreted by the Germinal code. The Germinal code then produces a (also potentially complex) output file, from which a scalar variable of interest is extracted, by a postprocessor. This process is illustrated in Figure 3.2 and is standard in large scale studies involving computer models.

Hence, a computer experiment here consists in selecting an input point x in $[0,1]^d$ with $d = 11$ (after renormalization of the physical inputs) and in observing $f_{\text{code}}(x) \in \mathbb{R}$. The components in x are related to the use cycle of the fuel pin, to its geometry, to the input power map and to the volume of expansion for fission gas. The output $f_{\text{code}}(x)$ is the fusion margin, which is important because it is an indicator of melting phenomena, which are hazardous and need to be avoided. An observation of $f_{\text{code}}(x)$ corresponds to the process in Figure 3.2 and takes around one minute. We refer to [J4] for more details on the context of this study.

Prediction results We model f_{code} as a realization of a Gaussian process with mean function zero and which covariance function k is assumed to belong to a parametric set $\{k_\theta; \theta \in \Theta\}$ where

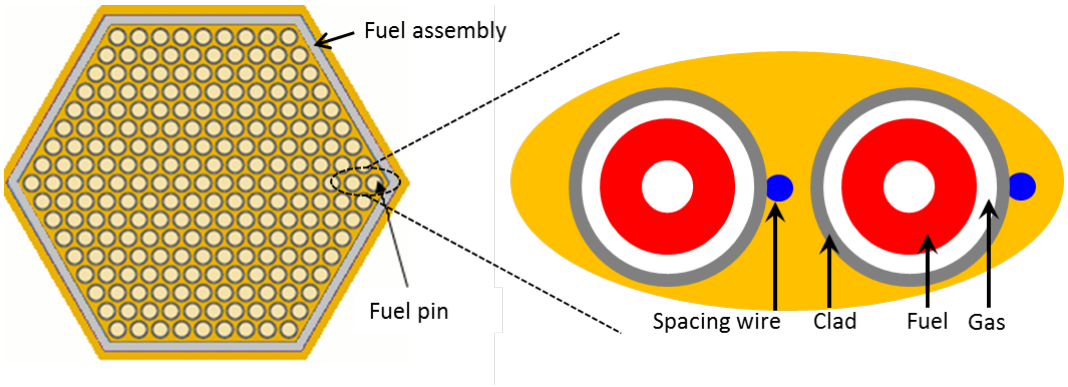


Figure 3.1: Context of [J4] in Section 3.1. A schematic representation of a fuel pin and a fuel assembly in nuclear fast-neutron reactors.

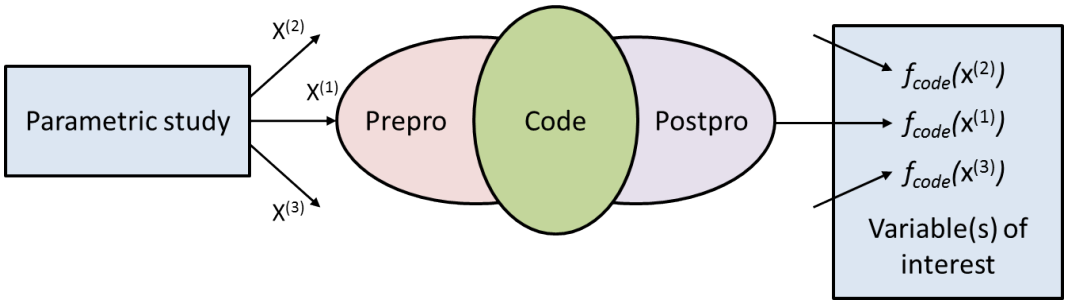


Figure 3.2: Context of [J4] in Section 3.1. Process for using the Germinal code.

$\theta = (\sigma^2, \ell_1, \ldots, \ell_d, \delta)$ and where $k_\theta(x_1, x_2) = \sigma^2 \check{k}_{\ell_1,\ldots,\ell_d}(x_1 - x_2) + \delta^2 \mathbf{1}_{x_1=x_2}$ with $\check{k}_{\ell_1,\ldots,\ell_d}$ the (stationary) Matérn 3/2 correlation function with correlation lengths $\ell_1, \ldots, \ell_d$ (see [J4]). Since $\check{k}_{\ell_1,\ldots,\ell_d}(x)$ is a continuous function of x, the term $\delta^2 \mathbf{1}_{x_1=x_2}$ is called a nugget effect Andrianakis and Challenor [2012] and is here meant to model the numerical instabilities of the code. More precisely, it can happen in practice that $f_{\text{code}}(x_1)$ is significantly different from $f_{\text{code}}(x_2)$ even though x_1 and x_2 are close. The value of δ is interpreted as the average order of magnitude of these instabilities. The covariance parameters can be estimated by ML and we let the resulting Gaussian process conditional mean and variance functions be denoted by $m_{n,\hat{\theta}}$ and $k_{n,\hat{\theta}}$ (see [J4]).

We compare the predictions obtained from the Gaussian process model with those obtained by kernel methods and neural networks. The kernel method predictions are computed as described in Wahba [1990]. The neural network predictions are obtained from the uncertainty quantification platform Uranie Gaudier [2010], developed at the French alternative energies and atomic energy commission (CEA).

The three methods are based on a learning basis $(x_1, f_{\text{code}}(x_1)), \ldots, (x_n, f_{\text{code}}(x_n))$ with $n = 3,807$. It takes a few hours to fit the Gaussian process model (covariance parameter estimation) and to fit the neural network (architecture and coefficient optimization by gradient descent and cross validation). On the other hand, there is no need to optimize parameters for the kernel method.

Then, the time required to compute a prediction $\hat{f}(x)$ for a new point x ($\hat{f} = m_{n,\hat{\theta}}$ for the Gaussian process model) is about 0.004 seconds for Gaussian processes and kernel methods and about 0.00015 seconds for the neural network. These prediction times are much shorter than the computation time for Germinal, and allow for a massive number of predictions. In the context of [J4], the final aim is to use these predictions to perform a multiobjective optimization of the fuel pin design. Neural networks are beneficial for prediction time here.

We consider a test base $(x_{t,1}, f_{\text{code}}(x_{t,1})), \ldots, (x_{t,N}, f_{\text{code}}(x_{t,N}))$ with $N = 1,613$, from which we compute the root mean square error (RMSE) (that should be minimal) defined by

$$\text{RMSE}^2 = \frac{1}{N} \sum_{i=1}^{N} \left(\hat{f}_{\text{code}}(x_{t,i}) - f_{\text{code}}(x_{t,i}) \right)^2 . \tag{3.1}$$

Furthermore, the criterion RMSE can be estimated from the learning base for the three predictors. We let $\widehat{\text{RMSE}}$ be its estimate. For Gaussian processes it is computed using the fast CV formula (2.7). For kernel methods, it is computed using an analog of (2.7) and for neural networks it is computed by directly predicting the observed values of the learning base. We refer to [J4] for more details.

The values of RMSE and $\widehat{\text{RMSE}}$ for the three predictors are given in Table 3.1. We observe that Gaussian process models provide the smallest value of RMSE. Furthermore, Gaussian processes and kernel methods enable to better estimate the value of RMSE than neural networks, that are slightly over-optimistic.

Numerical instability analysis and detection For the results of Table 3.1, the estimate of the nugget variance is $\hat{\delta} = 28.5°$. This value is comparable with the value of RMSE for the Gaussian process model. This is an indication that numerical instabilities exist in the code

	$\widehat{\text{RMSE}}$	RMSE
Neural networks	34.5°	38.5°
Gaussian processes	35.6°	36.1°
Kernel methods	44.3°	44.5°

Table 3.1: Context of [J4] in Section 3.1. Values of RMSE and $\widehat{\text{RMSE}}$ for the three predictors of the Germinal code function.

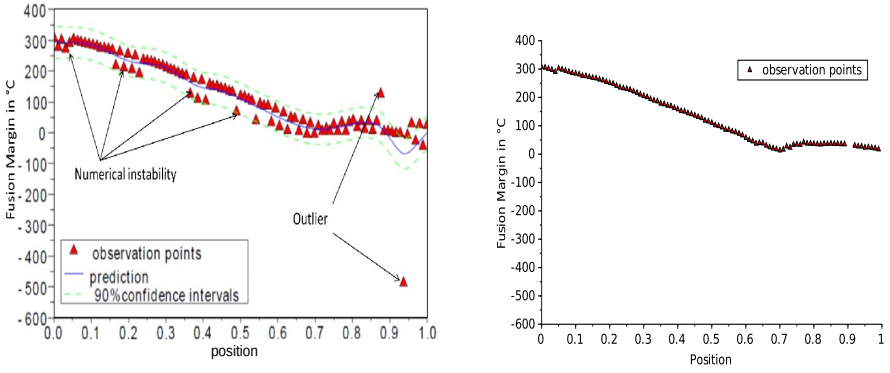


Figure 3.3: Context of [J4] in Section 3.1. The values of the Germinal code are plotted for input points along a segment of $[0,1]^{11}$. The x-axis indicates the position on the segment. The y-axis indicates the code output (fusion margin). The left hand-side corresponds to the first version of the preprocessor. We observe numerical instabilities. We also show the Gaussian process prediction and confidence intervals (see [J4] for details) that show that the nugget estimate $\hat{\delta}$ is appropriate. We also observe two outlier values that can be automatically detected by Gaussian processes, kernel methods and neural networks [J4]. The right hand-side corresponds to the improved version of the preprocessor. The improvement results in the absence of numerical instabilities on this segment.

function f_{code}. In order to investigate them, we plot in Figure 3.3 the values of the code function, for input points along a segment of $[0,1]^{11}$. Figure 3.3 indeed clearly highlights the presence of numerical instabilities. In [J4] we explain how they can be analyzed and how the preprocessor can be improved. This results in a new version of the code function f_{code}, for which, with the same segment of input points, no instabilities are visible anymore, as we show in Figure 3.3. We show in [J4] that, with the new version of f_{code}, the new estimate of the nugget variance is $\hat{\delta} = 19.8°$ and the new error criterion is RMSE $= 27.2°$. Hence, the conclusion of [J4] is that the estimator of the nugget variance in Gaussian process models is a good quantifier of the presence of numerical instabilities.

3.2 Aggregation of submodels for large data sets [J10,S2]

As discussed in Section 2.3, the ML or CV estimation of covariance parameters can become computationally too costly when n is too large (typically above $10,000$ in practice). Furthermore, for a Gaussian process with covariance function k, computing the conditional mean and covariance functions in (2.1) and (2.2) entails a similar issue (cost of $O(n^3)$ in time and $O(n^2)$ in storage).

Classical methods of the literature are dedicated to this problem, in particular inducing points Hensman et al. [2013], Nickson et al. [2015], low rank approximations Stein [2014], Gaussian Markov Random Fields Rue and Held [2005] and compactly supported covariance functions and covariance tapering Stein [2013], Kaufman et al. [2008], Furrer et al. [2006] (see also Section 2.3). Also, some methods aggregate submodels or 'experts' based on subsets of the data Hinton [2002], Tresp [2000], Deisenroth and Ng [2015].

Aggregation of submodels In [J10,S2], we address this last type of methods, based on aggregating several Gaussian process submodels. We let $X = (x_1, \ldots, x_n)$ be a sequence of observation points in $\mathbb{R}^d$. We let $X_1, \ldots, X_q$ be sequences of observation points so that $X_1, \ldots, X_q$ constitute a partition of X. We let n_i be the number of elements of X_i so that we have $n_1 + \ldots + n_q = n$. For $i = 1, \ldots, q$, we let $y^{(i)}$ be the $n_i \times 1$ vector $(\xi([X_i]_1), \ldots, \xi([X_i]_{n_i}))^\top$ where $X_i = ([X_i]_1, \ldots, [X_i]_{n_i})$. Then, from (2.1), for $x \in \mathbb{R}^d$, $m_{i,n_i}(x) = k(x, X_i)k(X_i, X_i)^{-1}y^{(i)}$ and $k_{i,n_i}(x, x) = k(x, x) - k(x, X_i)k(X_i, X_i)^{-1}k(X_i, x)$ are the conditional mean and variance of $\xi(x)$ given $y^{(i)}$.

In the literature, several aggregated predictors of the form

$$m_{\text{agg}}(x) = \sum_{i=1}^{q} \alpha_i(k_{1,n_1}(u,u), ..., k_{q,n_q}(u,u), k(u,u))m_{i,n_i}(u) \tag{3.2}$$

are suggested, with $\alpha_i : \mathbb{R}^{q+1} \to \mathbb{R}$. They are meant to approximate the full conditional mean given the n observations of ξ. These aggregation techniques include product of expert (POE) Hinton [2002], generalized product or expert (GPOE) Cao and Fleet [2014], Bayesian committee machines (BCM) Tresp [2000] and robust Bayesian committee machines (RBCM) Deisenroth and Ng [2015]. We refer to [J10,S2] for the expressions of α_i for them. These aggregation methods also provide an aggregated variance $k_{\text{agg}}(x, x)$, that is meant to approximate the full conditional variance. The computational benefit of (3.2) is clear, since the time cost is $O(n_1^3) + \ldots + O(n_q^3) + O(q)$, where the $O(n_i^3)$ is the cost of computing $m_{i,n_i}(x)$ for $i = 1, \ldots, q$ and where the $O(q)$ is the aggregation cost. This is much smaller than n^3 when q is large.

However, we show in [S2] that aggregations given by (3.2) can lead to mean square prediction errors that do not go to zero as $n \to \infty$, when considering triangular arrays of observation points that become dense in a compact set D. In other words these aggregations are asymptotically inconsistent.

Proposition 3.1. *Let D be a compact nonempty subset of $\mathbb{R}^d$. Let ξ be a Gaussian process on D with mean zero and stationary covariance function k. Assume that k is defined on $\mathbb{R}^d$, continuous and satisfies $k(x, y) > 0$ for two distinct points $x, y \in D$ such that D contains two open balls with strictly positive radii and centers x and y. Assume also that k has a positive*

spectral density (defined by $\tilde{k}(\omega) = (2\pi)^{-d} \int_{\mathbb{R}^d} k(x) \exp(-\imath x^\top \omega) dx$ *with* $\imath^2 = -1$ *and for* $\omega \in \mathbb{R}^d$*). Assume that there exists* $0 \leq A < \infty$ *and* $0 \leq T < \infty$ *such that* $1/\tilde{k}(\omega) \leq A(1 + ||\omega||^T)$*.*

For any triangular array of observation points $(x_i^{(n)})_{n\in\mathbb{N},i=1,\ldots,n}$ *and for* $n \in \mathbb{N}$*, we let* X *be the sequence* $(x_1^{(n)}, \ldots, x_n^{(n)})$ *and we let* $X_1, ..., X_{q_n}$ *be sequences constituting a partition of* X*. Finally, for* $n \in \mathbb{N}$ *we let* $m_{\text{agg}}(x)$ *be obtained from* (3.2) *with* q *replaced by* q_n*. We also assume that*

$$\alpha_i(v_1(x), ..., v_{q_n}(x), v) \leq \frac{a(v_i(x), v)}{\sum_{l=1}^{q_n} b(v_l(x), v)},$$

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

Then, there exists a triangular array of observation points $(x_i^{(n)})_{n\in\mathbb{N},i=1,\ldots,n}$ *such that* $\lim_{n\to\infty} \sup_{x\in D} \min_{i=1,\ldots,n} ||x_i^{(n)} - x|| = 0$*, a triangular array of sequences* $X_1, ..., X_{q_n}$ *forming a partition of* X*, with* $q_n \to_{n\to\infty} \infty$ *and* $q_n/n \to_{n\to\infty} 0$*, and such that there exists* $x_0 \in D$ *such that*

$$\liminf_{n\to\infty} \mathbb{E}(\xi(x_0) - m_{\text{agg}}(x_0))^2 > 0. \tag{3.3}$$

The intuitive explanation of the above proposition is that the aggregation methods for which it applies ignore the correlations between the different predictors $m_{i,n_i}(x)$ for $i = 1, \ldots, q$. Hence, for prediction points around which the density of observation points is smaller than on average, too much weight can be given to predictors based on distant observation points.

In Proposition 3.1, the assumptions made on k are satisfied by many stationary covariance functions, including those of the Matérn model, with the notable exception of the Gaussian covariance function (Proposition 1 in Vazquez and Bect [2010b]). We show in [S2] that the proposition applies to the POE, GPOE, BCM and RBCM methods introduced above. Hence, Proposition 3.1 constitutes a significant theoretical drawback, and warning for practical use, for an important class of aggregation techniques in the literature.

Our suggested aggregation procedure We let $R(x)$ be the covariance matrix of $m_{n_1,\ldots,n_q}(x) = (m_{1,n_1}(x), \ldots, m_{q,n_q}(x))^\top$ and we let $r(x)$ be the $q\times 1$ covariance vector between $(m_{1,n_1}(x), \ldots, m_{q,n_q}(x))$ and $\xi(x)$ for $x \in \mathbb{R}^d$. Then, in [J10], we suggest to approximate the full conditional mean $m_n(x)$ by the optimal linear combination of $m_{1,n_1}(x), \ldots, m_{q,n_q}(x)$. That it we suggest the aggregated predictor

$$m_{\text{agg}}(x) = r(x)^\top R(x)^{-1} m_{n_1,\ldots,n_q}(x) \tag{3.4}$$

and the corresponding aggregated predictive variance (approximating the full conditional variance)

$$k_{\text{agg}}(x, x) = k(x, x) - r(x)^\top R(x)^{-1} r(x).$$

The most computationally costly step for computing $m_{\text{agg}}(x)$ is to compute the $q\times q$ matrix $R(x)$, see [J10]. Unfortunately, this computation depends on the point x where one aims at predicting $\xi(x)$. In [J10] we show that, for appropriate choices of q and $n_1, \ldots, n_q$, the computational cost for performing r predictions is $O(n)$ in storage and $O(rn^2)$ in time. Hence there is a significant improvement in storage compared to the computation of the full conditional mean,

and a potentially significant improvement in time if r is smaller than n. Finally, the computation of $m_{\text{agg}}(x)$ is easy to parallelize, which is contrary to that of the full conditional mean.

In [J10,S2], we show that the aggregated prediction and variance $m_{\text{agg}}(x)$ and $k_{\text{agg}}(x,x)$ have several good theoretical properties. They have similar interpolation properties as the full conditional mean and variance and they provide a Gaussian conditional distribution given $(m_{1,n_1}(x),\ldots,m_{q,n_q}(x))$. Furthermore, they provide exact conditional means and variances (given the n observation points) for a Gaussian process with a slightly different covariance function (where this difference can be bounded). We refer to [J10,S2] for details on these theoretical properties.

Consider now covariance parameter estimation. In [J10] we suggest a stochastic gradient descent algorithm for optimizing the leave one out mean square error in (2.6), with $m_{\theta,n,-i}(x_i)$ replaced by a leave one out version of (3.4). The motivation for stochastic gradient is that the cost of computing r leave one out errors, with the aggregation procedure, is proportional to r. Hence, at each step of the algorithm, the gradient is evaluated on a random sample of indices in (2.6). This estimation procedure is implemented in [J10] for $n = 10,000$, in which case the computation time is around a few hours with a mono-threaded implementation.

Finally, a tree version of the aggregation procedure is also suggested in [J10] and implementations of the procedure are publicly available on the web-site `http://www.clementchevalier.com/index.php/r-packages`.

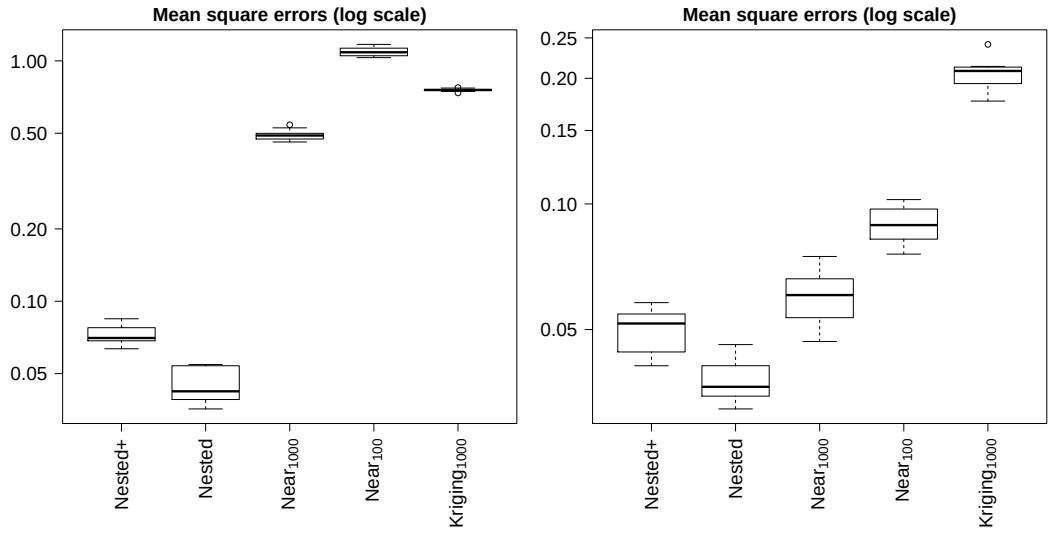
Numerical results We present some practical comparisons of our suggested aggregation procedures with other methods. For a given test base defined similarly as in Section 3.1, we consider the $\text{MSE} = \text{RMSE}^2$ criterion (that should be minimized) and the mean negative log probability (MNLP) criterion defined by

$$\text{MNLP} = \frac{1}{n_t}\sum_{i=1}^{n_t}\frac{1}{2}\log(2\pi k_{\text{agg}}(x_{t,i},x_{t,i})) + \frac{(m_{\text{agg}}(x_{t,i},x_{t,i}) - f(x_{t,i}))^2}{2k_{\text{agg}}(x_{t,i},x_{t,i})},$$

where f is any function that is modeled as a Gaussian process realization and where m_{agg} and k_{agg} are the mean and variance provided by any aggregation procedure. The MNLP takes into account both the prediction errors and the predictive variances and should be minimized.

In Figure 3.4, we show boxplots of mean square errors (over random design and prediction points) for predicting the analytic function Hartman 18 in dimension $d = 18$, where the total number of observation points is 10^6 (see [J10] for details). We observe that the mean square errors obtained from our aggregation procedures are significantly smaller than those obtained from nearest neighbor based methods.

Finally, in Table 3.2, we show prediction results in the context of experimental data on the behavior of a steel test piece subject to cycles of tension-compression. For $n = 10,000$ and $d = 6$, we compare our proposed aggregation procedure to the product of experts and Bayesian committee machines techniques presented above and to the predictor that uses the subset of observation points X_i yielding the smallest conditional variance, with $i \in \{1,\ldots,q\}$. Our suggested procedure provides the best prediction criteria.



(a) Hartman 18, concatenated parameters

(b) Hartman 18, estimated parameters

Figure 3.4: Context of [J10] in Section 3.2. Boxplots of mean square errors (in log scale), for our procedure (Nested), a variant (Nested+), predictions based on nearest neighbors (Near_{1000} and Near_{100}) and predictions based on $1,000$ random points (Kriging_{1000}). We use either concatenated covariance parameters (left panel) or estimated covariance parameters (right panel). We refer to [J10] for the full specification of the setting of this figure.

	SPV	PoE	GPoE1	GPoE2	BCM	RBCM	Nested
MSE	0.00416	0.0662	0.0033	0.0662	0.604	0.0625	**0.00321**
MNLP	−1.86	7.25	−0.949	−0.765	107	27.2	**−1.97**

Table 3.2: Context of [J10] in Section 3.2. Bold figures indicate each line's best performing aggregation method. The aggregation methods under study are our proposal (Nested), product of experts (PoE, GPoE1 and GPoE2), Bayesian committee machines (BCM and RBCM) and the predictor using the subset of observation point X_i, $i \in \{1, \ldots, q\}$, yielding the smallest conditional variance (SPV). We refer to [J10] for the full details.

3.3 Consistency of stepwise uncertainty reduction strategies [J15]

Sequential designs In [J15], we address iterative (or sequential) designs (of experiments) for Gaussian processes. We consider here a continuous Gaussian process ξ with values in $\mathbb{R}$, defined on a compact set $\mathbb{X}$ of $\mathbb{R}^d$ (the setting in [J15] is slightly more general). Then, in the case of a sequential design, for $i \in \mathbb{N}$, the observation point $i+1$ is selected as a function of $\xi(x_1), \dots, \xi(x_i)$. Hence, in this section, the observation points 1 to i become random points $X_1, \dots, X_i \in \mathbb{X}$.

Iterative designs are now routinely used for estimating quantities of interest related to the Gaussian process realization ξ. For instance, iterative designs exist for mono-objective optimization Mockus et al. [1978], Jones et al. [1998], multi-objective optimization Williams et al. [2000], Emmerich et al. [2006], Picheny [2014], Binois [2015], Gramacy et al. [2016], Feliot et al. [2016] and for estimating contour lines, probabilities of failures, profile optima and excursion sets Ranjan et al. [2008], Vazquez and Bect [2009], Picheny et al. [2010], Bect et al. [2012], Zuluaga et al. [2013], Chevalier et al. [2014], Ginsbourger et al. [2014], Wang et al. [2016]. Common applications include for instance computer experiments Jones et al. [1998], Forrester et al. [2008] and machine learning Shahriari et al. [2016].

Stepwise uncertainty reduction Here, we focus on a class of sequential designs called stepwise uncertainty reduction. We let $\mathbb{M}$ be the set of Gaussian measures, that is the set of distributions of continuous Gaussian processes on $\mathbb{X}$ (an element ν in $\mathbb{M}$ is characterized by a mean and covariance function). We refer to [J15] for more technical details.

Then, we consider a function $\mathcal{H} : \mathbb{M} \to [0, +\infty)$ that we call an uncertainty functional. The interpretation is that, for each $\nu \in \mathbb{M}$, $\mathcal{H}(\nu)$ quantifies the uncertainty we have on the realization of a Gaussian process with distribution ν. Smaller values of $\mathcal{H}(\nu)$ indicate less uncertainty.

Let P_n^ξ be the conditional distribution of ξ given $\xi(X_1), \dots, \xi(X_n)$. Let m_n, k_n, $\mathbb{E}_n$ and $\mathbb{P}_n$ be the corresponding mean function, covariance function, expectation and probability. For $x \in \mathbb{X}$, let $\mathrm{Cond}_{x,\xi(x)}(P_n^\xi)$ be the conditional distribution of ξ given $\xi(X_1), \dots, \xi(X_n), \xi(x)$. Conditionally to $\xi(X_1), \dots, \xi(X_n)$, the quantity $\mathcal{H}\left(\mathrm{Cond}_{x,\xi(x)}(P_n^\xi)\right)$ is the uncertainty after having observed $\xi(x)$. Hence, a good choice of observation point x results in a small value of $\mathcal{H}\left(\mathrm{Cond}_{x,\xi(x)}(P_n^\xi)\right)$. The quantity $\mathcal{H}\left(\mathrm{Cond}_{x,\xi(x)}(P_n^\xi)\right)$ is random because $\xi(x)$ is random given $\xi(X_1), \dots, \xi(X_n)$. Hence, we consider its expectation

$$J_n(x) = \mathbb{E}_n\left(\mathcal{H}\left(\mathrm{Cond}_{x,\xi(x)}(P_n^\xi)\right)\right).$$

A stepwise uncertainty reduction (SUR) strategy then consists in letting the observation point X_{n+1} be defined by

$$X_{n+1} \in \operatorname*{argmin}_{x \in \mathbb{X}} J_n(x). \tag{3.5}$$

Thus, X_{n+1} is a function of $\xi(X_1), \dots, \xi(X_n)$. Furthermore, there is no need to observe $\xi(x)$ to compute $J_n(x)$. In practice, J_n often needs to be minimized numerically. Hence, SUR strategies are applied to cases where observing $\xi(x)$ is costly (for instance when $\xi(x)$ represents the output of a computer model), and in particular more costly than minimizing J_n.

Let us consider two examples of uncertainty functionals. First, let

$$\mathcal{H}(\nu) = \mathbb{E}(\sup_{x\in\mathbb{X}} \xi_\nu(x)) - \max_{x\in\mathbb{X}:k_\nu(x,x)=0} m_\nu(x) \tag{3.6}$$

where m_ν and k_ν are the mean and covariance functions of the distribution $\nu \in \mathbb{M}$ and where ξ_ν is a Gaussian process with distribution ν. For this uncertainty functional, the quantities of interest are the global maximum and maximizer of ξ. The functional can be interpreted as the average difference between the global maximum of ξ_ν and the observed maximum (over the input points at which the variance is zero). Minimizing the corresponding J_n yields

$$X_{n+1} \in \operatorname*{argmax}_{x\in\mathbb{X}} \mathbb{E}_n \left(\left[\max_{u\in\mathbb{X};k_n(u,u|x)=0} \xi(u) - \max_{u\in\mathbb{X};k_n(u,u)=0} \xi(u) \right]^+ \right),$$

where $k_n(u,u|x)$ is the conditional variance of $\xi(u)$ given $\xi(X_1),\dots,\xi(X_n),\xi(x)$. The above display corresponds to the celebrated expected improvement criterion for global optimization Mockus et al. [1978], Jones et al. [1998]. The second example is defined by

$$\mathcal{H}(\nu) = \int_{\mathbb{X}} p_\nu(u)\,(1-p_\nu(u))\,du \tag{3.7}$$

where $p_\nu(u) = \mathbb{E}(\mathbf{1}_{\xi_\nu(u)\geq T})$, where ξ_ν is a Gaussian process with distribution ν and $T \in \mathbb{R}$ is a fixed threshold. In this case, the quantity of interest is the excursion set $\{u \in \mathbb{X}; \xi(u) \geq T\}$.

Consistency results In [J15], we prove the consistency of SUR strategies, under regularity conditions, and where the uncertainty functional possesses the following supermartingale property.

Definition 3.2. *An uncertainty functional $\mathcal{H}$ possesses the supermartingale property if, for any $\nu \in \mathbb{M}$ and $x \in \mathbb{X}$, we have $\mathbb{E}(\mathcal{H}(\mathrm{Cond}_{x,\xi_\nu(x)}(\nu))) \leq \mathcal{H}(\nu)$, where ξ_ν is a Gaussian process with distribution ν and where $\mathrm{Cond}_{x,\xi_\nu(x)}(\nu)$ is the conditional distribution of ξ_ν given $\xi_\nu(x)$.*

The supermartingale property is rather natural, since it means that adding a new observation point decreases the uncertainty on average. In [J15], we prove the following general consistency result.

Theorem 3.3. *Let $\mathcal{H}$ denote a non-negative, measurable functional on $\mathbb{M}$ with the supermartingale property. Let (X_n) be a sequence of random points in $\mathbb{X}$ so that (3.5) holds. Then $\mathcal{H}(P_n^\xi) - \inf_{x\in\mathbb{X}} J_n(x)$ goes to zero almost surely. If, moreover, continuity conditions given in Theorem 3.10 in [J15] hold, then $\mathcal{H}(P_n^\xi) \to 0$ almost surely.*

In Theorem 3.3, consistency means that the uncertainty measure converges to zero. In [J15], we also provide in Corollary 3.17 a similar general consistency result for uncertainty functionals related to a loss function for estimating a quantity of interest. In this Corollary 3.17, there are fewer continuity conditions that need to be satisfied. Finally, in [J15], we apply the general results to four examples of SUR strategies, including those defined by (3.6) and (3.7). We give the details in the next theorem.

Theorem 3.4. *Let $(X_n)_{n\in\mathbb{N}}$ be defined by (3.5) and (3.6). Let $M_n = \max_{u\in\mathbb{X};k_n(u,u)=0} \xi(u)$. Then almost surely and in L^1 as $n \to \infty$, $\mathcal{H}(P_n^\xi) \to 0$ and $M_n \to \sup_{u\in\mathbb{X}} \xi(u)$.*

Let also $(X_n)_{n\in\mathbb{N}}$ *be defined by* (3.5) *and* (3.7). *Then almost surely and in* L^1 *as* $n \to \infty$, $\mathcal{H}(P_n^\xi) \to 0$ *and, with* $p_n(u) = \mathbb{P}_n(\xi(u) \geq T)$,

$$\int_{\mathbb{X}} \left(\mathbf{1}_{\xi(u)\geq T} - p_n(u)\right)^2 du \to 0$$

and

$$\int_{\mathbb{X}} \left(\mathbf{1}_{\xi(u)\geq T} - \mathbf{1}_{p_n(u)\geq 1/2}\right)^2 du \to 0.$$

Finally, let us mention that there exist several references providing consistency or rate of convergence results for specific sequential designs with Gaussian processes, for instance expected improvement Vazquez and Bect [2010a], Bull [2011] (Theorem 3.4 provides complementary or additional results compared to these references) or Gaussian process upper confidence bounds Srinivas et al. [2012]. In [J15], we restrict ourselves to proving consistency, but aim at a more general framework than for existing references. Furthermore, to our knowledge, no results exist, except in [J15], for proving the consistency of sampling criteria for excursion domain estimation as in (3.7). Finally, note that it is shown that, for instance, the expected improvement criterion can yield inconsistent estimates of the global maximum for some fixed continuous functions Yarotsky [2013]. Our results can be interpreted as showing that these functions yielding inconsistency have probability zero, under the probability measure corresponding to the Gaussian process model used.

3.4 Distribution inputs [J12]

In [J12], we aim at considering a Gaussian process ξ which input set is the set of probability measures on $\mathbb{R}$, with a finite moment of order 2, that we write $W_2(\mathbb{R})$. In applications, there are indeed cases where output values can be observed for distributions inputs, for instance in social sciences Flaxman et al. [2015] or engineering Radulescu et al. [2009].

Covariance functions based on the Monge-Kantorovich transport distance We consider Gaussian processes with mean function zero and aim at constructing covariance functions based on the Monge-Kantorovich (or Wasserstein) distance, with quadratic cost. For two distributions μ and ν on $\mathbb{R}$ with finite variances, this distance is defined as

$$W_2(\mu,\nu) = \sqrt{\inf_{\pi\in\Pi(\mu,\nu)} \int_{\mathbb{R}^2} |x-y|^2 d\pi(x,y)},$$

where $\Pi(\mu,\nu)$ is the set of distributions on $\mathbb{R}^2$ with first and second marginal distributions equal to μ and ν. A feature of the Monge-Kantorovich distance for distributions on $\mathbb{R}$ is that this minimization problem can be solved explicitly. More precisely, for a distribution μ on $\mathbb{R}$, let F_μ be its cumulative distribution function and let $F_\mu^{-1}(t) = \inf\{u \in \mathbb{R}; F_\mu(u) \geq t\}$ be its quantile function. Then we have

$$W_2(\mu,\nu) = \sqrt{\int_{\mathbb{R}} (F_\mu^{-1}(t) - F_\nu^{-1}(t))^2}.$$

From the above expression, and from Schoenberg theorem (which is proved in Berg et al. [1984]), we show in [J12] that the following functions $k_{\sigma^2,\lambda,H} : W_2(\mathbb{R}) \times W_2(\mathbb{R}) \to \mathbb{R}$, defined by

$$k_{\sigma^2,\lambda,H}(\mu,\nu) = \sigma^2 \exp\left(-\lambda W_2(\mu,\nu)^{2H}\right) \tag{3.8}$$

are non-negative definite for $(\sigma^2, \lambda, H) \in (0,\infty)^2 \times (0,1]$. In [J12], we also suggest an other family of covariance functions that extends the Fractional Brownian motion on $\mathbb{R}$, for which we show that the corresponding covariance matrices are invertible, for sets of two-by-two distinct input distributions.

Increasing-domain asymptotic results In [J12], we show that the proof techniques of my PhD article [J3] can be extended to the case of distribution inputs. We consider a general parametric family of covariance functions $\{k_\theta; \theta \in \Theta\}$ on $W_2(\mathbb{R})$ and a triangular array of input distributions satisfying the following conditions.

Condition 3.5. *We consider a triangular array of input distributions* $\{\mu_1, ..., \mu_n\} = \{\mu_1^{(n)}, ..., \mu_n^{(n)}\}$ *so that for all* $n \in \mathbb{N}$ *and* $1 \leq i \leq n$, μ_i *has support in* $[i, i+K]$ *with a fixed* $K < \infty$.

Condition 3.5 is inspired by increasing-domain asymptotic settings for Gaussian processes on the real line. In [J12], we show the consistency and asymptotic normality of the ML estimator of θ_0 (in the well-specified case). We also show that predicting with the ML estimator is asymptotically as good as predicting with the true covariance parameter. These results hold under technical conditions that are listed in [J12]. We also show that, in a specific representative example, with random input distributions which densities are renormalized exponential of Gaussian processes, all these technical conditions are satisfied. The proofs of the asymptotic results in [J12] can be divided into two groups. For the proofs in the first group, we show that the arguments of [J3] can be extended. The proofs in the second group are, on the other hand, specific to distribution inputs and to the Monge-Kantorovich distance.

Numerical results We first compare the covariance functions in (3.8) with covariance functions based on projecting the densities of the input distributions on finite dimensional spaces. Projecting functional inputs on finite-dimensional spaces is indeed classical with computer experiments Nanty et al. [2016], Muehlenstaedt et al. [2016]. We address random distribution inputs, and let the output values be obtained from an analytical function. As quality criteria for the Gaussian process models, we consider the RMSE criterion as in (3.1) and the proportion of test values that are covered by the 0.9 confidence intervals. We write $\text{CIR}_{0.9}$ for this second criterion and we refer to [J12] for the full experimental setting. The results are given in Table 3.3, where our suggested kernel provides the best values of the quality criteria.

In another experimental setting, we compare our suggested Gaussian process model with the kernel regression procedure of Póczos et al. [2013]. The results are given in Table 3.4 and, again, show that our suggested procedure performs better.

3.5 Inequality constraints [J14,S5]

In [J14,S5], we consider a zero-mean Gaussian process ξ with covariance function k indexed on $[0,1]^d$. We consider conditioning ξ by the event $\xi \in \mathcal{E}$, where $\mathcal{E}$ is, for instance, a set of bounded, increasing or convex functions. This event corresponds to inequality constraints for ξ, as opposed to the more common case of equality constraints (for instance $\xi(x_i) = y_i$ for $x_i \in [0,1]^d$ and $y_i \in \mathbb{R}$).

model	RMSE	$CIR_{0.9}$
'distribution'	0.094	0.92
'Legendre' order 5	0.49	0.92
'Legendre' order 10	0.34	0.89
'Legendre' order 15	0.29	0.91
'PCA' order 5	0.63	0.82
'PCA' order 10	0.52	0.87
'PCA' order 15	0.47	0.93

Table 3.3: Context of [J12] in Section 3.4. Values of different quality criteria for the 'distribution', 'Legendre' and 'PCA' Gaussian process models. The 'distribution' Gaussian process model is based on (3.8), while 'Legendre' and 'PCA' are based on linear projections of the input distributions on finite-dimensional spaces. These finite dimensional spaces are based on the Legendre polynomials and on principal component analysis and have dimensions indicated by 'order'.

model	RMSE	$CIR_{0.9}$
'distribution'	0.21	0.91
'kernel regression'	0.93	

Table 3.4: Context of [J12] in Section 3.4. Comparison of our suggested procedure 'distribution' with the kernel regression procedure of Póczos et al. [2013] (this method does not provide confidence intervals so that $CIR_{0.9}$ is not calculated for it).

Gaussian processes with inequality constraints are studied in Da Veiga and Marrel [2012], Golchi et al. [2015], Riihimäki and Vehtari [2010]. They provide valuable models in applications such as computer networking (monotonicity) Golchi et al. [2015], social system analysis (monotonicity) Riihimäki and Vehtari [2010] and econometrics (monotonicity or positivity) Cousin et al. [2016].

Finite dimensional representation Let $d = 1$ and consider equally-spaced knots $\{t_j = (j-1)\Delta_m\}_{j=1,\dots,m}$ with $\Delta_m = 1/(m-1)$ for $m \in \mathbb{N}$. Then, Maatouk and Bay [2017] suggests to consider a finite-dimensional Gaussian process, denoted by ξ_m, as the piecewise linear interpolation of ξ at knots $t_1, \cdots, t_m$:

$$\xi_m(x) = \sum_{j=1}^{m} \xi(t_j)\phi_j(x), \tag{3.9}$$

where $\phi_1 \cdots, \phi_m$ are hat basis functions given by

$$\phi_j(x) := \begin{cases} 1 - \left|\frac{x-t_j}{\Delta_m}\right| & \text{if } \left|\frac{x-t_j}{\Delta_m}\right| \leq 1, \\ 0 & \text{otherwise.} \end{cases} \tag{3.10}$$

We use ξ_m as an approximation of ξ, because inequality constraints on the function ξ_m can be shown to be equivalent to finitely many inequality constraints constraints on $\epsilon = (\xi(t_1), \dots, \xi(t_m))^\top$. More precisely, let

$$\mathcal{E}_\kappa := \begin{cases} \{f \in C([0,1], \mathbb{R}) \text{ s.t. } \ell \leq f(x) \leq u, \ \forall x \in [0,1]\} & \text{if } \kappa = 0, \\ \{f \in C([0,1], \mathbb{R}) \text{ s.t. } f \text{ is non-decreasing}\} & \text{if } \kappa = 1, \\ \{f \in C([0,1], \mathbb{R}) \text{ s.t. } f \text{ is convex}\} & \text{if } \kappa = 2, \end{cases} \tag{3.11}$$

which correspond to boundedness, monotonicity, and convexity constraints and where $-\infty \leq \ell < u \leq +\infty$ are fixed. Let

$$\mathcal{C}_\kappa := \begin{cases} \{c \in \mathbb{R}^m; \ \forall\, j = 1, \cdots, m \ : \ \ell \leq c_j \leq u\} & \text{if } \kappa = 0, \\ \{c \in \mathbb{R}^m; \ \forall\, j = 2, \cdots, m \ : \ c_j \geq c_{j-1}\} & \text{if } \kappa = 1, \\ \{c \in \mathbb{R}^m; \ \forall\, j = 3, \cdots, m \ : \ c_j - c_{j-1} \geq c_{j-1} - c_{j-2}\} & \text{if } \kappa = 2. \end{cases} \tag{3.12}$$

Then, it is observed in Maatouk and Bay [2017] that we have, for $\kappa = 0, 1, 2$, $\xi_m \in \mathcal{E}_\kappa$ if and only if $\epsilon \in \mathcal{C}_\kappa$. Thus, for instance, simulations of trajectories of ξ_m conditionally to $\xi_m \in \mathcal{E}_\kappa$ can be obtained, by simulating realizations of the vector ϵ, conditionally to $\epsilon \in \mathcal{C}_\kappa$.

The definitions of the hat functions ϕ_j and of the sets $\mathcal{C}_\kappa$ can be extended to $d \geq 2$, by tensorization Maatouk and Bay [2017], and as we also explain in [J14]. Nevertheless, this framework is currently typically limited to dimensions $d = 1, 2, 3$ in practice. We are currently working on other approaches for larger dimensions.

One of the contributions of [J14] is to suggest a generalization of the sets $\mathcal{E}_\kappa$ and $\mathcal{C}_\kappa$ in (3.11) and (3.12). Indeed, we consider inequality constraints of the form $\epsilon \in \mathcal{C}$, where

$$\mathcal{C} = \left\{ c \in \mathbb{R}^m; \ \forall\, i = 1, \dots, q \ : \ \ell_i \leq \sum_{j=1}^{m} \lambda_{i,j} c_j \leq u_i \right\},$$

where the $\lambda_{i,j}$'s encode the linear operations and where the l_i's and u_i's represent the lower and upper bounds. In [J14], we explain how to specify $\mathcal{C}$ to impose combinations of constraints, such as boundedness and monotonicity.

Simulating from the posterior distribution Let $d \in \mathbb{N}$ here. Let $\Lambda = [\lambda_{i,j}]_{i=1,\dots,q,j=1,\dots,m}$ and let $\ell = (\ell_1, \dots, \ell_q)^\top$ and $u = (u_1, \dots, u_q)^\top$. For $n \in \mathbb{N}$ and $x_1, \dots, x_n \in [0,1]^d$ let $\Phi = [\phi_j(x_i)]_{i=1,\dots,n,j=1,\dots,m}$. Let $\mathcal{E}$ be a set of functions so that $\xi_m \in \mathcal{E}$ if and only if $\epsilon \in \mathcal{C}$. Then, as explained in [J14], simulating conditional trajectories of ξ_m given $\xi_m(x_1) = y_1, \dots, \xi_m(x_n) = y_n$ and given $\xi_m \in \mathcal{E}$, boils down to simulating conditional realizations of ϵ given $\Phi\epsilon = y = (y_1, \dots, y_n)^\top$ and given $\ell \leq \Lambda\epsilon \leq u$. This in turn boils down to simulating a Gaussian vector ϵ_c (which mean vector and covariance matrix are expressed in [J14]), conditionally to $\ell \leq \epsilon_c \leq u$.

This type of simulation is performed in Maatouk and Bay [2017] by a method called rejection sampling from the mode. In [J14], we study several other advanced Monte Carlo methods for this simulation. We consider the Hastings-Metropolis algorithm Murphy [2012], the Gibbs sampler Taylor and Benjamini [2017], exponential tilting Botev [2017] and Hamiltonian Monte Carlo Pakman and Paninski [2014]. We refer to [J14] for detailed simulation results. Generally speaking, we find that exponential tilting and Hamiltonian Monte Carlo provide the most efficient results, and we recommend to use the Hamiltonian Monte Carlo method.

Constrained maximum likelihood We let $d \in \mathbb{N}$ here. For simplicity of exposition, in the rest of Section 3.5, we do not consider the finite-dimensional Gaussian process ξ_m anymore and we consider inequality constraints of the type $\xi \in \mathcal{E}_\kappa$ with

$$\mathcal{E}_\kappa = \begin{cases} f : [0,1]^d \to \mathbb{R}, f \text{ is } C^0 \text{ and } \forall x \in \mathbb{X},\ \ell \leq f(x) \leq u & \text{if } \kappa = 0, \\ f : [0,1]^d \to \mathbb{R}, f \text{ is } C^1 \text{ and } \forall x \in \mathbb{X},\ \forall i = 1, \cdots, d,\ \frac{\partial}{\partial x_i} f(x) \geq 0 & \text{if } \kappa = 1, \\ f : [0,1]^d \to \mathbb{R}, f \text{ is } C^2 \text{ and } \forall x \in \mathbb{X},\ \frac{\partial^2}{\partial x^2} f(x) \text{ is a non-negative definite matrix} & \text{if } \kappa = 2. \end{cases} \tag{3.13}$$

We refer to [J14] for the various algorithmic and theoretical extensions when ξ_m is considered instead of ξ. Consider a parametric family $\{k_\theta; \theta \in \Theta\}$ of covariance functions on $[0,1]^d$ and assume that ξ has covariance function k_{θ_0} with $\theta_0 \in \Theta$. Consider also an observation vector $y = (\xi(x_1), \dots, \xi(x_n))^\top$. For estimating the parameter θ_0, it is of course possible to use the ML estimator presented in Section 2.1. Since this estimator does not explicitly use the information $\xi \in \mathcal{E}_\kappa$, we call it the unconstrained ML estimator. In [J14], we remark that the logarithm of the conditional density function of y given $\xi \in \mathcal{E}_\kappa$, under covariance function k_θ, is

$$\mathcal{L}_{\mathcal{C},n}(\theta) = \log p_\theta(y) + \log P_\theta(\xi \in \mathcal{E}_\kappa | y) - \log P_\theta(\xi \in \mathcal{E}_\kappa), \tag{3.14}$$

where the first term is the logarithm of the probability density function of y, under covariance function k_θ (unconstrained log-likelihood), and the last two terms depend on the inequality constraints. In (3.14), we let P_θ be the probability when ξ has covariance function k_θ. Then, we suggest in [J14] the constrained ML estimator given by

$$\widehat{\theta}_{\text{cML}} \in \arg\max_{\theta \in \Theta} \mathcal{L}_{\mathcal{C},n}(\theta). \tag{3.15}$$

The second and third term in (3.14) need to be approximated numerically, which makes constrained ML computationally more costly than unconstrained ML. In [J14], we consider

the numerical integration procedure of Genz [1992] and the exponential tilting Monte Carlo procedure Botev [2017].

Consistency results We consider an infinite sequence $(x_i)_{i\in\mathbb{N}}$ of observation points that is dense in $[0,1]^d$. In [J14], we show that, roughly speaking, consistency results for unconstrained ML can be transferred into consistency results for constrained ML. The next result corresponds to the propositions given in Section 5.3 of [J14].

Proposition 3.6. *Let ξ be a zero-mean Gaussian process on $[0,1]^d$ which covariance function k satisfies technical conditions given in [J14]. Let Θ be a compact subset of $(0,\infty)^{d+1}$. Let k_θ be the covariance function of $x \to \sigma\xi(\alpha_1 x_1, \cdots, \alpha_d x_d)$ for $\theta = (\sigma^2, \alpha_1, \cdots, \alpha_d) \in \Theta$. Let $\theta_0 = (1, \cdots, 1)$. Remark that $k = k_{\theta_0}$ and assume that $\theta_0 \in \Theta$.*

Let $\kappa \in \{0,1,2\}$ be fixed. Then, we have $\inf_{\theta\in\Theta} P_\theta(\xi \in \mathcal{E}_\kappa) > 0$. Assume that $\forall \varepsilon > 0$ and $\forall M < \infty$,

$$P\left(\sup_{\|\theta-\theta_0\|\geq\varepsilon} (\log p_\theta(y) - \log p_{\theta_0}(y)) \geq -M \right) \xrightarrow[n\to\infty]{} 0.$$

Then,

$$P\left(\sup_{\|\theta-\theta_0\|\geq\varepsilon} (\mathcal{L}_{\mathcal{C},n}(\theta) - \mathcal{L}_{\mathcal{C},n}(\theta_0)) \geq -M \;\middle|\; \xi \in \mathcal{E}_\kappa \right) \xrightarrow[n\to\infty]{} 0.$$

Consequently

$$\text{argmax}_{\theta\in\Theta}\ \log p_\theta(y) \xrightarrow[n\to\infty]{P} \theta_0, \qquad \text{and} \qquad \text{argmax}_{\theta\in\Theta}\ \mathcal{L}_{\mathcal{C},n}(\theta) \xrightarrow[n\to\infty]{P|\xi\in\mathcal{E}_\kappa} \theta_0,$$

where $\xrightarrow[n\to\infty]{P}$ denotes the convergence in probability under the distribution of ξ, and $\xrightarrow[n\to\infty]{P|\xi\in\mathcal{E}_\kappa}$ denotes the convergence in probability under the distribution of ξ given $\xi \in \mathcal{E}_\kappa$.

The technical conditions on k in Proposition 3.6 are not really restrictive and allow for many stationary covariance functions, in particular the Matérn covariance functions. In [J14], extensions of Proposition 3.6, to the case where the finite-dimensional approximation ξ_m is used, are given.

Asymptotic normality results In [S5], the aim is to provide more quantitative results than in Proposition 3.6. This can be done, at the cost of considering less general families of covariance functions.

First, we consider a family of covariance functions $\{\sigma^2 k, \sigma^2 \in [\sigma_l^2, \sigma_u^2]\}$ with fixed $0 < \sigma_l^2 < \sigma_u^2 < +\infty$. We let ξ have mean function $\sigma_0^2 k$ with $\sigma_l^2 < \sigma_0^2 < \sigma_u^2$. In this case, it is well-known that the ML estimator $\hat{\sigma}^2_{ML}$ satisfies

$$\sqrt{n}\left(\hat{\sigma}^2_{ML} - \sigma_0^2\right) \to^{\mathcal{L}}_{n\to\infty} \mathcal{N}(0, 2\sigma_0^4).$$

The first result in [S5] is that this asymptotic distribution is unaffected by conditioning by the event $\xi \in \mathcal{E}_\kappa$ for $\kappa = 0,1,2$. For a sequence of random vectors or variables $(X_n)_{n\in\mathbb{N}}$ on $\mathbb{R}^l$, that are functions of ξ, and for a probability distribution μ on $\mathbb{R}^l$, we write

$$X_n \xrightarrow[n\to\infty]{\mathcal{L}|\xi\in\mathcal{E}_\kappa} \mu$$

when, for any bounded continuous function $g : \mathbb{R}^l \to \mathbb{R}$, we have

$$\mathbb{E}\left[\, g(X_n) |\, \xi \in \mathcal{E}_\kappa \right] \xrightarrow[n\to\infty]{} \int_{\mathbb{R}^l} g(x)\mu(dx).$$

Theorem 3.7. *Assume that k and the sequence of observation points satisfy technical conditions given in [S5]. Then, we have, for $\kappa = 0, 1, 2$,*

$$\sqrt{n}\left(\hat{\sigma}^2_{ML} - \sigma_0^2\right) \xrightarrow[n\to\infty]{\mathcal{L}|\xi\in\mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4).$$

The conditions on k in Theorem 3.8 are not really restrictive. Then, in [S5] we show that the constrained ML estimator $\hat{\sigma}^2_{cML}$, defined in (3.15), has the same asymptotic distribution as the ML estimator, conditionally to $\xi \in \mathcal{E}_\kappa$.

Theorem 3.8. *Assume that k and the sequence of observation points satisfy technical conditions given in [S5]. Then, we have, for $\kappa = 0, 1, 2$,*

$$\sqrt{n}\left(\hat{\sigma}^2_{cML} - \sigma_0^2\right) \xrightarrow[n\to\infty]{\mathcal{L}|\xi\in\mathcal{E}_\kappa} \mathcal{N}(0, 2\sigma_0^4).$$

Next, in [S5], we consider the Matérn family of covariance functions $\{k_{\theta,\nu}; \theta \in \Theta\}$, with Θ compact in $(0, +\infty)^2$, with $\theta = (\sigma^2, \rho)$ and with

$$k_{\theta,\nu}(x, x') = \sigma^2 K_\nu\left(\frac{||x - x'||}{\rho}\right) = \frac{\sigma^2}{\Gamma(\nu)2^{\nu-1}}\left(\frac{||x - x'||}{\rho}\right)^\nu \kappa_\nu\left(\frac{||x - x'||}{\rho}\right).$$

Here, $\sigma^2 > 0$ is the variance, $\rho > 0$ is the correlation length and $\nu > 0$ (assumed to be known) is the smoothness parameter of the process. The function κ_ν is the modified Bessel function of the second kind of order ν (see Abramowitz and Stegun [1964]) and Γ is the Gamma function. We let ξ have covariance function $k_{\theta_0,\nu}$ with $\theta_0 = (\sigma_0^2, \rho_0)$ in the interior of Θ.

We let $d \leq 3$. Then, as shown in Zhang [2004], only the parameter $\sigma^2/\rho^{2\nu}$ is microergodic. Consequently, there are no consistent estimators of σ_0^2 and ρ_0 but $\sigma_0^2/\rho_0^{2\nu}$ can be consistently estimated. It is shown in Kaufman and Shaby [2013], using results in Du et al. [2009], that the ML estimator $(\hat{\sigma}^2_{ML}, \hat{\rho}^{2\nu}_{ML})$ satisfies

$$\sqrt{n}\left(\frac{\hat{\sigma}^2_{ML}}{\hat{\rho}^{2\nu}_{ML}} - \frac{\sigma_0^2}{\rho_0^{2\nu}}\right) \to^{\mathcal{L}}_{n\to\infty} \mathcal{N}\left(0, 2\left(\frac{\sigma_0^2}{\rho_0^{2\nu}}\right)^2\right).$$

In [S5], we show the same types of results as for the case of the estimation of a single variance parameter. Conditionally to $\xi \in \mathcal{E}_\kappa$, ML and constrained ML have the same asymptotic distribution, which coincides with that of ML without inequality constraints.

Theorem 3.9. *Assume that ν and the sequence of observation points satisfy technical conditions given in [S5]. Then, we have, for $\kappa = 0, 1, 2$,*

$$\sqrt{n}\left(\frac{\hat{\sigma}^2_{ML}}{\hat{\rho}^{2\nu}_{ML}} - \frac{\sigma_0^2}{\rho_0^{2\nu}}\right) \to^{\mathcal{L}|\xi\in\mathcal{E}_\kappa}_{n\to\infty} \mathcal{N}\left(0, 2\left(\frac{\sigma_0^2}{\rho_0^{2\nu}}\right)^2\right)$$

and

$$\sqrt{n}\left(\frac{\hat{\sigma}^2_{cML}}{\hat{\rho}^{2\nu}_{cML}} - \frac{\sigma_0^2}{\rho_0^{2\nu}}\right) \to^{\mathcal{L}|\xi\in\mathcal{E}_\kappa}_{n\to\infty} \mathcal{N}\left(0, 2\left(\frac{\sigma_0^2}{\rho_0^{2\nu}}\right)^2\right).$$

The proof of Theorem 3.9 involves the results of Zhang [2004], Du et al. [2009], Kaufman and Shaby [2013] on the Matérn model, and tools from extrema of Gaussian processes and from reproducing kernel Hilbert spaces. In [S5], it is explained where the technical assumptions of the above theorem are needed, and how these assumptions are interpreted.

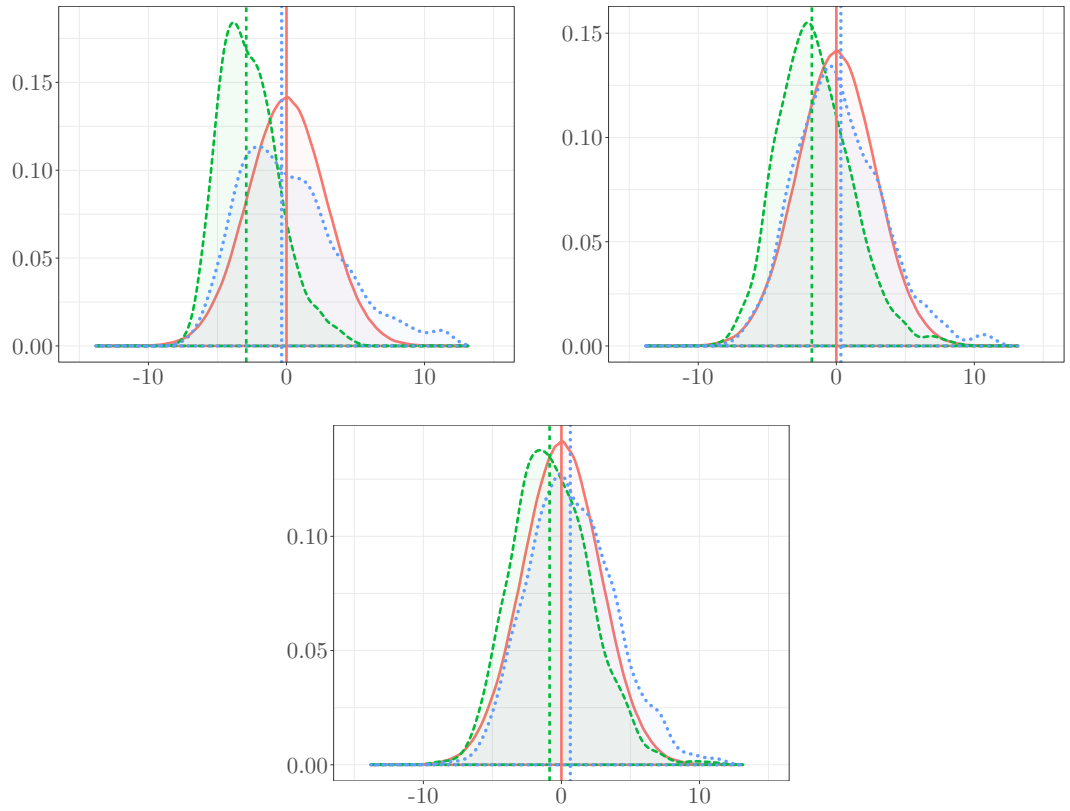


Figure 3.5: Context of [S5] in Section 3.5. Estimates of the distributions of $n^{1/2}(\hat{\sigma}^2_{ML}/\hat{\rho}^{2\nu}_{ML} - \sigma_0^2\rho_0^{2\nu})$ (green, dashed lines) and of $n^{1/2}(\hat{\sigma}^2_{cML}/\hat{\rho}^{2\nu}_{cML} - \sigma_0^2\rho_0^{2\nu})$ (blue, dotted lines). We also show the limit Gaussian distribution (red, solid lines). The vertical lines represent the median values of the distributions. The values of n are 20 (top left), 50 (top right) and 80 (bottom).

Simulation results The conclusion of the above asymptotic results is that there is no asymptotic difference between unconstrained and constrained ML, and that conditioning by $\xi \in \mathcal{E}_\kappa$ does not impact the asymptotic distribution of ML. Hence, the inequality constraints do not play a role asymptotically.

In [S5], in simulations, we consider small or moderate values of n, in the aim of assessing how fast the finite sample distributions of the estimators converge to the asymptotic ones, and if we can observe differences between the estimators.

In Figure 3.5, we show estimates of the distributions of $n^{1/2}(\hat{\sigma}^2_{ML}/\hat{\rho}^{2\nu}_{ML} - \sigma_0^2\rho_0^{2\nu})$ and of $n^{1/2}(\hat{\sigma}^2_{cML}/\hat{\rho}^{2\nu}_{cML} - \sigma_0^2\rho_0^{2\nu})$ in the case of the estimation of σ_0^2 for the Matérn model with known ν and ρ_0 and for boundedness constraints (see [S5] for the full setting). The conclusion of Figure 3.5 is that the distributions are close to the asymptotic one for $n = 80$ and that the constrained ML is more accurate for smaller values of n. Hence, in practice, we recommend to use constrained ML (which is computationally more costly) for smaller values of n and unconstrained ML for larger values of n. A practical way of anticipating whether constrained ML and ML would give significantly different results or not is suggested in the conclusion of [S5].

Chapter 4

Valid confidence intervals post-model-selection

4.1 Introduction

Post-model-selection inference The general topic of this chapter is post-model-selection inference. Post-model-selection inference differs quite significantly in nature from Gaussian processes, that are the general topic of Chapters 2 and 3. Arguably, post-model-selection inference is currently quite an active topic. Various recent contributions are discussed below, and are also presented in [J13,S1].

In this chapter, we will mainly consider Gaussian linear regression models (except in Section 4.5 where the setting is more general). Hence, we consider a $n \times 1$ Gaussian observation vector $y = \mu + \epsilon$, with $n \in \mathbb{N}$, where μ is a fixed vector and where $\epsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ distribution. We also consider a $n \times p$ design matrix X where $p \in \mathbb{N}$ is the number of explanatory variables (or regressors).

A model is then a subset M of $\{1, \ldots, p\}$. We use the following notation: For $M \subseteq \{1, ..., p\}$, we write M^c for the complement of M in $\{1, ..., p\}$. We write $|M|$ for the cardinality of M. With $m = |M|$, let us write $M = \{j_1, ..., j_m\}$ in case $m \geq 1$. For an $l \times p$ matrix T, we let $T[M]$ be the matrix of dimension $l \times m$ obtained from T by retaining only the columns of T with indices $j \in M$ and deleting all others. For a $p \times 1$ vector v, we let $v[M]$ be the vector obtained by retaining only the components of v with indices in M.

A model M yields the restricted design matrix $X[M]$ of size $n \times |M|$. A model selection procedure, or model selector, is then a function $\hat{M} : \mathbb{R}^n \to \mathcal{M}$ where $\mathcal{M} \subset \{M; M \subset \{1, \ldots, p\}\}$ is called the universe of possible models. Examples of model selection procedures are AIC, BIC and the lasso Tibshirani [1996].

A general informal definition of post-model-selection inference would be to characterize it as statistical inference (e.g. constructing tests or confidence intervals) based on the selected model $\hat{M}$. Since $\hat{M}$ is random, this is in contrasts to 'classical' inference which would be based on a fixed M, not depending on data. The construction of valid statistical procedures in post-model-selection situations is quite challenging (cf. Leeb and Pötscher [2005, 2006, 2008], Kabaila

and Leeb [2006] and Pötscher [2009], and the references given in that literature). Furthermore, it is known that, in these situations, pointwise asymptotics (for instance over parameters of the true observation distribution) can be misleading Leeb and Pötscher [2005]. Hence, uniform asymptotic results are focused on in several references, for instance Charkhi and Claeskens [2018]. In the articles [J13,S1] presented in this chapter, the asymptotic results given hold uniformly over classes of distributions for the observations.

A short literature review Among various recent references in the post-model-selection context, we can first mention those addressing sparse high dimensional settings with a focus on lasso-type model selection procedures Belloni et al. [2011, 2014], van de Geer et al. [2014], Zhang and Zhang [2014]. In these references, it is considered that the mean vector of y can be written as $X\beta_0$, with β_0 a sparse vector of $\mathbb{R}^p$ and $p > n$. The components of β_0 are the targets of inference, and the lasso model selection procedure $\hat{M}$ is a tool for constructing tests and confidence intervals.

In a different paradigm, in the references Tibshirani et al. [2016], Lee and Taylor [2014], Fithian et al. [2015], Lee et al. [2016], Tibshirani et al. [2018], Berk et al. [2013], each model $M \in \mathcal{M}$ for which $X[M]^\top X[M]$ is invertible, defines a target

$$\begin{aligned} \beta_M^{(n)} &= (X[M]^\top X[M])^{-1} X[M]^\top \mu \\ &\in \underset{\beta_M \in \mathbb{R}^{|M|}}{\operatorname{argmin}} ||\mu - X[M]\beta_M||. \end{aligned} \tag{4.1}$$

One of the main motivations for this paradigm is that targets of inference can be defined also in the misspecified case where μ does not belong to the column space of X (when $p \leq n$). When $p > n$, the existence of a true sparse β_0, as above, is not needed. In Berk et al. [2013] and [J13], further discussion is provided on the interpretation of the model dependent target $\beta_M^{(n)}$, and of its interest.

Then, in Tibshirani et al. [2016], Lee and Taylor [2014], Fithian et al. [2015], Lee et al. [2016], Tibshirani et al. [2018], Berk et al. [2013], the targets of inference are the components of the random vector $\beta_{\hat{M}}^{(n)}$ (that has random dimension). In Tibshirani et al. [2016], Lee and Taylor [2014], Fithian et al. [2015], Lee et al. [2016], Tibshirani et al. [2018], $\hat{M}$ is assumed to be a specific model selector, defined by polyhedral constraints, in particular the lasso or a procedure based on sequential testing. Then, in these references, confidence intervals that are valid (i.e. have coverage probability larger than or equal to the nominal level), conditionally to $\hat{M}$, are obtained.

On the other hand, in Berk et al. [2013], confidence intervals that are valid for any model selection procedure $\hat{M}$ are suggested (the term uniformly valid may be used). In this chapter, we extend the general ideas of Berk et al. [2013] in different directions.

Validity regardless of the model selection procedure is of fundamental importance, because many procedures used in practice are almost impossible to formalize: researchers typically use combinations of visual inspection and numerical algorithms, and sometimes they simply select models that let them reject many hypotheses, that is, they are hunting for significance. These often unreported and informal practices of model selection prior to conducting the actual analysis may also play a key role in the current crisis of reproductibility. Thus, to establish and popularize

statistical methods that are in some sense robust to 'bad practice' is highly desirable.

The construction of uniformly valid confidence intervals of Berk et al. [2013] Assume that with $\mathcal{M} \subset \{M; M \subset \{1, \ldots, p\}\}$, each $M \in \mathcal{M}$ is non-empty and is so that X_M has full column rank. In Berk et al. [2013], a family of confidence intervals $(\mathrm{CI}_{i,M}; i \in M \in \mathcal{M})$ for the components of the vectors $\beta_M^{(n)}$ is introduced, taking the form

$$\mathrm{CI}_{i,M} = (\hat{\beta}_M)_{i.M} \pm \hat{\sigma} \|s_{i,M}\| K_1(X, \mathcal{M}, \alpha, r); \tag{4.2}$$

the different quantities involved, which we now define, are standard ingredients for univariate confidence intervals for regression coefficients, except for the last factor (the 'PoSI constant') which will account for multiplicity of covariates and models. The confidence interval is centered at $\hat{\beta}_M := (X[M]^\top X[M])^{-1} X[M]^\top y$, the ordinary least squares estimator of $\beta_M^{(n)}$; also, if $M = \{j_1, \ldots, j_{|M|}\}$ with $j_1 < \ldots < j_{|M|}$, for $i \in M$ we denote by $i.M$ the number $k \in \mathbb{N}$ for which $j_k = i$, that is, the rank of the i-th element in the subset M. The quantity $\hat{\sigma}$ is assumed to be an observable random variable, such that $\hat{\sigma}^2$ is independent of $P_X Y$ and is distributed as σ^2/r times a chi-square distributed random variable with r degrees of freedom (P_X denoting the orthogonal projection onto the column space of X). We allow for $r = \infty$ corresponding to $\hat{\sigma} = \sigma$, i.e., the case of known variance (also called Gaussian limiting case). In Berk et al. [2013], it is assumed that $\hat{\sigma}$ exists and it is shown that this indeed holds in some specific situations. Nevertheless, the assumptions on $\hat{\sigma}$ are not innocuous, in our opinion, and further discussion is provided in [J13,S1]. The existence of $\hat{\sigma}$ will need to be assumed in [J13] but not in [S1]. The next quantity to define is

$$s_{i,M}^\top := (e_{i.M}^{|M|})^\top (X[M]^\top X[M])^{-1} X[M]^\top \in \mathbb{R}^n, \tag{4.3}$$

where e_a^b is the a-th base column vector of $\mathbb{R}^b$. Finally, $K_1 = K_1(X, \mathcal{M}, \alpha, r) \geq 0$ is called a PoSI constant. Let also

$$\bar{s}_{i,M} = \begin{cases} s_{i,M}/\|s_{i,M}\|, & \text{if } \|s_{i,M}\| \neq 0; \\ 0 \in \mathbb{R}^n & \text{else.} \end{cases} \tag{4.4}$$

Let U be a Gaussian vector with zero mean vector and identity covariance matrix on $\mathbb{R}^n$. Let N be a random variable, independent of U, and so that rN^2 follows a chi-square distribution with r degrees of freedom. If $r = \infty$, then we let $N = 1$. For $\alpha \in (0, 1)$, $K_1(X, \mathcal{M}, \alpha, r)$ is defined as the $1 - \alpha$ quantile of

$$\frac{1}{N} \max_{M \in \mathcal{M}, i \in M} \left| \bar{s}_{i,M}^\top U \right|. \tag{4.5}$$

It is shown in Berk et al. [2013] that we have,

$$\inf_{\mu \in \mathbb{R}^n, \sigma > 0} \mathbb{P}_{\mu,\sigma} \left(\forall M \in \mathcal{M}, \quad \forall i \in M, \quad (\beta_M^{(n)})_{i.M} \in \mathrm{CI}_{i,M} \right) \geq 1 - \alpha, \tag{4.6}$$

where $\mathbb{P}_{\mu,\sigma}$ indicates the dependence on the mean vector μ of y and on the variance σ^2. Hence, the PoSI confidence intervals guarantee a simultaneous coverage of all the projection-based regression coefficients, over all models M in the set $\mathcal{M}$. Hence, for any model selection procedure $\hat{M}$ we have $(\beta_{\hat{M}}^{(n)})_{i.\hat{M}} \in \mathrm{CI}_{i,\hat{M}}$ for all $i \in \hat{M}$ with probability larger than $1 - \alpha$.

Order of magnitude of the PoSI constant The confidence intervals in (4.2) are similar in form to the standard (or 'naive') confidence intervals that one would use for a single fixed model M and a fixed $i \in M$. For a 'naive' interval, K_1 would be replaced by a standard Gaussian or Student quantile. Of course, the 'naive' intervals do not account for multiplicity and do not have uniform coverage over $i \in M \in \mathcal{M}$ (see [J13,S1]). Hence K_1 is the inflation factor or correction over standard intervals to get uniform coverage; it must go to infinity as $p \to \infty$ Berk et al. [2013]. Studying the asymptotic order of magnitude of K_1 is thus an important problem, as this order of magnitude corresponds to the price one has to pay in order to obtain universally valid post model selection inference.

When $n \geq p$, it is shown in Berk et al. [2013] that K_1 is asymptotically no smaller than $\sqrt{\log(p)}$ and asymptotically no larger than $\sqrt{p}$. These two lower and upper bound are reached by respectively orthogonal design matrices and equicorrelated design matrices (see Berk et al. [2013]). It is currently a difficult problem to characterize the order of magnitude of K_1 for very general design matrices X. The aforementioned results in Berk et al. [2013], and those given in [J13] thus tackle specific design matrices, in particular orthogonal and equicorrelated.

4.2 Confidence intervals for predictors in linear regression [J13]

Adaptation of the PoSI confidence intervals to prediction In [J13], we consider a fixed $p \times 1$ vector x_0 of new values of the explanatory variables. We consider as a model dependent scalar target of inference the quantity $x_0[M]^\top \beta_M^{(n)}$. This quantity is an (infeasible) predictor of a new variable y_0, that would typically be studied if the model M were to be selected. We refer to [J13] for more discussion of this predictor and of its optimality properties.

We construct confidence intervals $(\mathrm{CI}_{x_0,M}; M \in \mathcal{M})$ where $\mathrm{CI}_{x_0,M}$ is as $\mathrm{CI}_{i,M}$ in Section 4.1, with $(\hat{\beta}_M)_{i.M}$ replaced by $x_0[M]^\top \hat{\beta}_M$, with $s_{i,M}$ replaced by $s_{x_0,M}$ and with $K_1 = K_1(X, \mathcal{M}, \alpha, r)$ replaced by $K_1(x_0) = K_1(x_0, X, \mathcal{M}, \alpha, r)$. More precisely, we let

$$s_{x_0,M}^\top := x_0[M]^\top (X[M]^\top X[M])^{-1} X[M]^\top \in \mathbb{R}^n, \tag{4.7}$$

we let $\bar{s}_{x_0,M}$ be defined as in (4.4) with $s_{i,M}$ replaced by $s_{x_0,M}$, and we let $K_1(x_0, X, \mathcal{M}, \alpha, r)$ be defined as in (4.5), with $\bar{s}_{i,M}$ replaced by $\bar{s}_{x_0,M}$. These confidence intervals are a straightforward extension of those of Berk et al. [2013]. We naturally obtain

$$\inf_{\mu \in \mathbb{R}^n, \sigma > 0} \mathbb{P}_{\mu,\sigma} \left(\forall M \in \mathcal{M}, \quad x_0[M]^\top \beta_M^{(n)} \in \mathrm{CI}_{x_0,M} \right) \geq 1 - \alpha. \tag{4.8}$$

Other PoSI constants In applications, it can happen that $x_0[M^c]$ is not observed, specifically if model selection is performed in the aim of observing fewer variables in the future, for cost reduction. For example, in a medical application one may want to avoid measuring prognostic variables that require invasive procedures or that incur high monetary costs, see, e.g., Castera et al. [2015]. Cost considerations in the context of model selection or prediction are also common in fields such as industrial process control or engineering (Jaupi [2014], Souders and Stenbakken

[1991]). We remark that, indeed, the knowledge of $x_0[M^c]$ is not needed to define $x_0[M]^\top \beta_M^{(n)}$ and to compute $x_0[M]^\top \hat{\beta}_M$. However, $K_1(x_0)$ is not computable if x_0 is not entirely observed.

Hence, in [J13], we suggest other constants $K_2(x_0[M], M)$, $K_3(x_0[M], M)$ and K_4, where K_2 and K_3 do not depend on $x_0[M^c]$ and K_4 does not depend on x_0 at all. These constants satisfy $K_1(x_0) \leq K_2 \leq K_3 \leq K_4$, so that the confidence intervals obtained as $(\mathrm{CI}_{x_0,M}; M \in \mathcal{M})$, with $K_1(x_0)$ replaced by K_2, K_3 or K_4 satisfy (4.8).

We refer to [J13] for the full construction of K_2, K_3 and K_4. In short words, the constant K_2 is obtained by maximizing the value of $K_1(x_0)$ over $x_0[M^c]$, the constant K_3 is obtained by a 'partial union bound' and the constant K_4 is obtained from a union bound. We point out that a version of K_4, in the context of Berk et al. [2013], is already suggested in an unpublished version of Berk et al. [2013] (see [J13] for a link).

Algorithms for approximating $K_1(x_0)$ to K_4 are given in [J13]. The computation of $K_1(x_0)$ has computational cost proportional to $|\mathcal{M}|$, that is to 2^p if $n \geq p$ and if all submodels are allowed for by the model selection procedure. In this case, the computation of $K_1(x_0)$ is feasible for $p \leq 30$, say. Computing K_2 requires to maximize the value of $K_1(x_0)$ numerically, so that it entails a larger cost than for $K_1(x_0)$. Computing K_2 is thus often prohibitive. Computing K_3 has a similar complexity as for $K_1(x_0)$. Finally, K_4 has a small computational cost, and we find that its computation (with the `R` software) is accurate for $p = 1,000$ or smaller. For larger values of p, one may run into numerical issues for the computation of extreme quantiles of the Beta distribution (see [J13]).

Large p results for the PoSI constants In [J13], we provide asymptotic results for the order of magnitude of $K_1(x_0)$ to K_4 as $p \to \infty$. In the next proposition, we consider the case where X is orthogonal (for $n \geq p$). We show that there exist vectors x_0 so that $K_1(x_0)$ has order of magnitude $\sqrt{p}$. This is in stark contrast with the setting of Berk et al. [2013], where in this case K_1 has order of magnitude $\sqrt{\log(p)}$. Furthermore, we show that, for models M not too close to the full model $\{1, \ldots, p\}$, $K_2(x_0[M], M)$ has order of magnitude $\sqrt{p}$ for all $x_0[M]$. This illustrates the price one has to pay for not observing $x_0[M^c]$, since there also exist vectors x_0 so that $K_1(x_0) = O(1)$.

Proposition 4.1. *Consider the known-variance case (i.e., $r = \infty$ and $\hat{\sigma}^2 = \sigma^2$) and assume that for every $p \geq 1$ the model universe $\mathcal{M}$ used is the power set of $\{1, ..., p\}$. Let $0 < \alpha < 1$, be given, not depending on p. Set $\gamma = \sup_{b>0} \phi(b)/\sqrt{1 - \Phi(b)} \approx 0.6363$, where ϕ and Φ are the probability density function and the cumulative distribution function of the standard Gaussian distribution.*

(a) For any $p \geq 1$ let $X = X(p)$ be an $n(p) \times p$ matrix with (non-zero) orthogonal columns. For any such sequence X one can find a corresponding sequence of $p \times 1$ vectors x_0 such that $K_1(x_0)$ satisfies

$$\liminf_{p \to \infty} K_1(x_0)/\sqrt{p} \geq \gamma.$$

Furthermore, for any sequence X as above one can find another sequence of (non-zero) $p \times 1$ vectors x_0 such that $K_1(x_0) = O(1)$.

(b) Let $\delta \in [0,1)$ be given. Then $K_2(x_0[M], M)$ satisfies

$$\liminf_{p\to\infty} \inf_{x_0\in\mathbb{R}^p} \inf_{X\in\mathsf{X}(p)} \inf_{M\in\mathcal{M},|M|\leq\delta p} K_2(x_0[M], M)/\sqrt{p} \geq \gamma\sqrt{1-\delta},$$

where $\mathsf{X}(p) = \bigcup_{n\geq p}\{X : X \text{ is } n\times p \text{ with non-zero orthogonal columns}\}$.

In the next proposition, we show that K_3 and K_4 are asymptotically equivalent as $p\to\infty$. We also give a simple asymptotic approximation to K_4. This next proposition exploits results in Zhang [2017].

Proposition 4.2. *Consider the known-variance case (i.e., $r=\infty$ and $\hat{\sigma}^2=\sigma^2$). Let the universe of models $\mathcal{M}$ satisfy technical conditions given in [J13]. For $n\in\mathbb{N}$, let $\mathsf{X}_{n,p}(\mathcal{M})$ denote the set of all $n\times p$ matrices of rank $\min(n,p)$ with the property that $X[M]$ has full column-rank for every $\varnothing \neq M\in\mathcal{M}$. Furthermore, let α, $0<\alpha<1$, be given (neither depending on p nor n). Let $n(p)\in\mathbb{N}$ be a sequence such that $n(p)\to\infty$ for $p\to\infty$ and such that $\mathsf{X}_{n(p),p}(\mathcal{M})\neq\varnothing$ for every $p\geq 1$. Then we have*

$$\lim_{p\to\infty} \sup_{M\in\mathcal{M},M\neq\{1,\ldots,p\}} \sup_{x_0\in\mathbb{R}^p} \sup_{X\in\mathsf{X}_{n(p),p}(\mathcal{M})} |1-(K_3(x_0[M],M)/K_4)| = 0. \tag{4.9}$$

Furthermore,

$$K_4/\sqrt{\min(n(p),p)\left(1-|\mathcal{M}|^{-2/(\min(n(p),p)-1)}\right)} \to 1$$

as $p\to\infty$.

The above proposition enables to express simply the order of magnitude of K_4 where $\mathcal{M}=\mathcal{M}_s=\{M\subset\{1,\ldots,p\};|M|\leq s\}$ as a function of the sparsity parameter s. We refer to Section 4.4, corresponding to [J16].

The design-independent target A potential criticism of the target $x_0[M]^\top\beta_M^{(n)}$ is that it is a predictor of x_0 but it depends on X. For this reason, this target is called design-dependent in [J13]. Hence, the target $x_0[M]^\top\beta_M^{(n)}$ is especially relevant when there is a link between x_0 and the lines of X.

In [J13], we thus also consider the case where x_0 and the lines of X are realizations from a common distribution $\mathcal{L}$ (x_0 and X are independent from ϵ). When considering this case, we also let p be fixed and let the full linear model be well-specified. That is we let $\mu=X\beta$ for a $p\times 1$ vector β. Under this framework, letting Σ be the (uncentered) second moment matrix of X, we call $x_0[M]^\top\beta_M^{(\star)}$ the design-independent target, with

$$\beta_M^{(\star)} = \beta[M] + (\Sigma[M,M])^{-1}\Sigma[M,M^c]\beta[M^c]. \tag{4.10}$$

Here, for a $p\times p$ matrix R and for $M_1,M_2\subset\{1,\ldots,p\}$, we let $R[M_1,M_2]$ be the $|M_1|\times|M_2|$ matrix obtained by retaining only the lines of R with indices in M_1 and the columns of R with indices in M_2. The target $x_0[M]^\top\beta_M^{(\star)}$ does not depend on the realization of X and has several optimality properties that make it preferable to $x_0[M]^\top\beta_M^{(n)}$, see [J13].

In [J13], we show that the above confidence intervals, meant to cover $x_0[M]^\top\beta_M^{(n)}$, are also asymptotically valid to cover $x_0[M]^\top\beta_M^{(\star)}$. In the next theorem $P_{\beta,\sigma}$ indicates the dependence in β,σ when evaluating probabilities

Theorem 4.3. *Suppose that* $n^{1/2}((1/n)X^\top X - \Sigma) = O_p(1)$. *Suppose that the model selection procedure* $\hat{M}$ *satisfies a technical assumption given in [J13]. Let* $\mathrm{CI}_{x_0,M}$ *be as in* (4.8)*, with* $K_1(x_0)$ *potentially replaced by* K_2*,* K_3 *or* K_4*. Then we have*

$$\inf_{x_0\in\mathbb{R}^p,\beta\in\mathbb{R}^p,\sigma>0} P_{\beta,\sigma}\left(x_0^\top[\hat{M}]\beta^{(\star)}_{\hat{M}} \in \mathrm{CI}_{x_0,M}\middle| X\right) \geq (1-\alpha) + o_p(1), \tag{4.11}$$

where the $o_p(1)$ *term above depends only on* X *and converges to zero in probability as* $n \to \infty$.

Simulation results In Table 4.1, we show the results of an extensive Monte Carlo study. We show estimates of the minimal coverage probabilities (over β and σ) of the above confidence intervals based on $K_1(x_0)$, K_3 and K_4, as well as of the naive confidence interval that ignores the randomness of the selected model. We let $p = 10$ and $n = 20$ or $n = 100$ and generate the lines of X randomly with equicorrelated covariance matrix. As model selectors we consider here AIC, BIC, the lasso, SCAD Fan and Li [2001], and MCP Zhang [2010]. We refer to [J13] for the full setting of the simulation study.

We observe that, as holds from the theory, the confidence intervals based on $K_1(x_0)$, K_3 and K_4 always cover the design-dependent target with minimal probability above the nominal level. However, for $n = 20$ and for the lasso, SCAD and MCP, they do not cover the design-independent target with minimal probability above the nominal level. Hence, for this small sample size, Theorem 4.3 does not provide an accurate description of the finite sample situation. For $n = 100$, the confidence intervals based on $K_1(x_0)$, K_3 and K_4 cover both targets with minimal probability above the nominal level. Furthermore, the coverage probabilities are almost equal between the two targets. This illustrates Theorem 4.3 (see also Lemma C.1 in the online supplement to [J13]). Finally, we see that the naive confidence intervals have minimal coverage probabilities significantly below the nominal level. This illustrates the need for studying specific inference procedures that take the post-model-selection context into account.

In [J13], we also compare the PoSI confidence intervals, based on $K_1(x_0)$, K_3 and K_4, with those in Lee et al. [2016], that are specific to the lasso and conditionally valid. We observe that the PoSI intervals have lengths that are similar, in the median sense, to those from Lee et al. [2016], and are shorter, when considering more extreme quantiles. In addition, we show that the intervals from Lee et al. [2016] have minimal coverage probabilities that can break down when the regularization parameter of the lasso is data-dependent instead of fixed. In contrasts, the confidence intervals based on $K_1(x_0)$, K_3 and K_4 have minimal coverage probabilities above the nominal level for any model selection procedure.

4.3 Links with optimal configurations of lines [J7]

In [J7], we investigate the links between the computation of the PoSI constant $K_1(x_0)$, defined in the previous section, and the problem of finding evenly spaced lines in the Euclidean space.

Indeed, one can show (see Berk et al. [2013], [J13] and [J7]) that $K_1(x_0)$ can be written as $f_{p,r,\alpha}(L(X, x_0))$, where $L(X, x_0)$ is a set of at most N lines of $\mathbb{R}^p$, where we let $N = 2^p - 1$. The lines of $L(X, x_0)$ depend on the design matrix X and on the vector x_0. We have

$$f_{p,r,\alpha} : \mathcal{D}_{\leq N} \to \mathbb{R}_+,$$

Data set	n	Model selector	Target							
			design-dependent $x_0[\hat{M}]'\beta^{(n)}_{\hat{M}}$				design-independent $x_0[\hat{M}]'\beta^{(\star)}_{\hat{M}}$			
			K_{naive}	K_1	K_3	K_4	K_{naive}	K_1	K_3	K_4
Equicorrelated	20	AIC	0.83	0.99	1.00	1.00	0.79	0.98	0.99	0.99
	20	BIC	0.81	0.99	1.00	1.00	0.74	0.98	0.99	0.99
	20	LASSO	0.88	1.00	1.00	1.00	0.39	0.71	0.79	0.79
	20	SCAD	0.88	0.99	1.00	1.00	0.67	0.92	0.95	0.96
	20	MCP	0.86	0.99	1.00	1.00	0.66	0.93	0.96	0.96
	100	AIC	0.84	0.99	1.00	1.00	0.84	0.99	1.00	1.00
	100	BIC	0.86	0.99	1.00	1.00	0.86	0.99	1.00	1.00
	100	LASSO	0.88	1.00	1.00	1.00	0.88	1.00	1.00	1.00
	100	SCAD	0.88	0.99	1.00	1.00	0.89	1.00	1.00	1.00
	100	MCP	0.88	0.99	1.00	1.00	0.89	0.99	1.00	1.00

Table 4.1: Context of [J13] in Section 4.2. Monte Carlo estimates of the minimal coverage probabilities (w.r.t. β and σ) of various confidence intervals. The naive confidence interval corresponds to K_{naive}. The nominal coverage probability is $1-\alpha = 0.95$ and $p = 10$.

where $\mathcal{D}_{\leq N}$ is the set of all sets of at most N lines of $\mathbb{R}^p$. A line is defined as the set $u\mathbb{R}$ with $u \in \mathbb{S}^{p-1}$, with $\mathbb{S}^{p-1}$ the unit sphere of $\mathbb{R}^p$.

The value of $f_{p,r,\alpha}(L)$, for $L \in D_{\leq N}$, is defined as the unique $K > 0$ such that

$$\mathbb{E}_V\Big(F_{p,r}\big(\frac{K^2}{p \cdot \max\limits_{u\mathbb{R}=\ell\in L}\langle u, V\rangle^2}\big)\Big) = 1-\alpha, \tag{4.12}$$

where $F_{p,r}$ is the cumulative distribution function of the F-distribution with parameters p and r, and V is a uniformly distributed random vector on $\mathbb{S}^{p-1}$.

As we have seen in the previous sections, the function $f_{p,r,\alpha}(L)$ can not be computed exactly since it is a quantile, and can be costly to approximate for large values of $N = 2^p - 1$. The quantity $K_4 = K_4(p, r, \alpha)$ of Section 4.2 is an upper bound satisfying

$$\sup_{\{\ell_1,\ldots,\ell_N\}\in\mathcal{D}_N} f_{p,r,\alpha}(\{\ell_1,\ldots,\ell_N\}) \leq K_4(p,r,\alpha), \tag{4.13}$$

with $\mathcal{D}_N$ the set of all sets of N lines of $\mathbb{R}^p$. In [J7], we aim at evaluating the above supremum in order to assess how tight the upper bound K_4 is for small values of p. We indeed remark that, for large values of p, the asymptotic results in Berk et al. [2013] imply that K_4 is tight in (4.13). In order to evaluate the supremum, we suggest a procedure based on two steps. In a first step, we compute N evenly-spaced lines by minimizing potential energies (see e.g. Cohn and Kumar [2007], Hoggar [1982]). Then we evaluate $f_{p,r,\alpha}$ only for these evenly space lines. This method is computationally beneficial, since potential energies are much less expensive to compute than $f_{p,r,\alpha}$, and since their minimization can be carried out independently of r and α.

When, computing sets of lines that are optimal for potential energies, we obtained interesting results with respect to the notion of universal optimality Cohn and Kumar [2007]. For $p = 3$, the

optimal set of lines we find coincides with a set of lines that is known to be universally optimal Cohn et al. [2016]. For $p = 4, 5$, our results suggest that there is no universally optimal sets of $2^p - 1$ lines. For $p = 6$, there seems to exist a universally optimal configuration of $2^p - 1 = 63$ lines, which we identified as the best packing configuration provided in Conway et al. [1996].

In Figure 4.1, we show the ratio S_K/K_4, where S_K is an evaluation of the supremum above obtained by four different methods. For the 'evenly-spaced lines' method, S_K is obtained as described above, from sets of lines minimizing potential energies. For the 'local evenly-spaced lines' and 'very local evenly-spaced lines' methods, we aim at maximizing $f_{p,r,\alpha}$, by a stochastic algorithm sampling sets of lines locally (or very locally) around the sets of lines minimizing potential energies. These two methods are meant to assess whether these latter sets of lines are local maxima of the function $f_{p,r,\alpha}$. Finally, the 'naive Monte Carlo' method consists in maximizing $f_{p,r,\alpha}$ directly, by a stochastic algorithm. We refer to [J7] for more details.

In Figure 4.1, we thus observe that the 'evenly spaced lines' method performs significantly better than the 'naive Monte Carlo' one. We also see that the upper bound K_4 is fairly tight in (4.13). We also see, from the results of the 'local evenly-spaced lines' and 'very local evenly-spaced lines' methods, that the sets of lines minimizing potential energies show evidences to be (at least) local maxima of $f_{p,r,\alpha}$. Finally, we see that for the dimensions $p = 3, 6$, for which there exists or seems to exist a universally optimal set of lines, the upper bound K_4 is tighter in (4.13). This indicates that the concept or property of universal optimality is indeed beneficial.

4.4 An upper bound on the PoSI constant under restricted isometry properties [J16]

In [J16], we consider the PoSI constant $K_1 = K_1(X, \mathcal{M})$ defined in Section 4.1. As shown in Berk et al. [2013], when $n \geq p$, and when X is an orthogonal matrix, $K_1(X, \mathcal{M})$ has order of magnitude $\sqrt{\log(p)}$ as $n, p \to \infty$.

Consider now the case where $\mathcal{M} = \mathcal{M}_s = \{M \subset \{1, \ldots, p\}; |M| \leq s\}$ is the set of s-sparse models, with $s \leq \min(n, p)$. Then, we show in [J16] that $K_1(X, \mathcal{M}_s)$ has order of magnitude no larger than $\sqrt{s \log(p/s)}$, independently of X (this bound actually aleady appears in an intermediary version of Zhang [2017]). We call this upper bound the general sparsity based upper bound.

For a square symmetric non-negative matrix A, we let $\text{corr}(A) = (\text{diag}(A)^\dagger)^{1/2} A (\text{diag}(A)^\dagger)^{1/2}$, where $\text{diag}(A)$ is obtained by setting all the non-diagonal elements of A to zero and where $B^\dagger$ is the Moore-Penrose pseudo-inverse of B. We define the restricted isometry property (RIP) constant of a $n \times p$ matrix X with sparsity level s as

$$\delta(X, s) = \sup_{|M| \leq s} ||\text{corr}(X[M]^\top X[M]) - I_{|M|}||_{op}, \tag{4.14}$$

where $||.||_{op}$ is the largest absolute eigenvalue. This definition is similar to the common definition of the RIP constant [Foucart and Rauhut, 2013, Chap.6], except that in this common definition $\text{corr}(X[M]^\top X[M])$ is replaced by $X[M]^\top X[M]$. Here we consider $\text{corr}(X[M]^\top X[M])$ because $K_1(X, \mathcal{M}_s)$ depends on X only through $\text{corr}(X[M]^\top X[M])$ [J16].

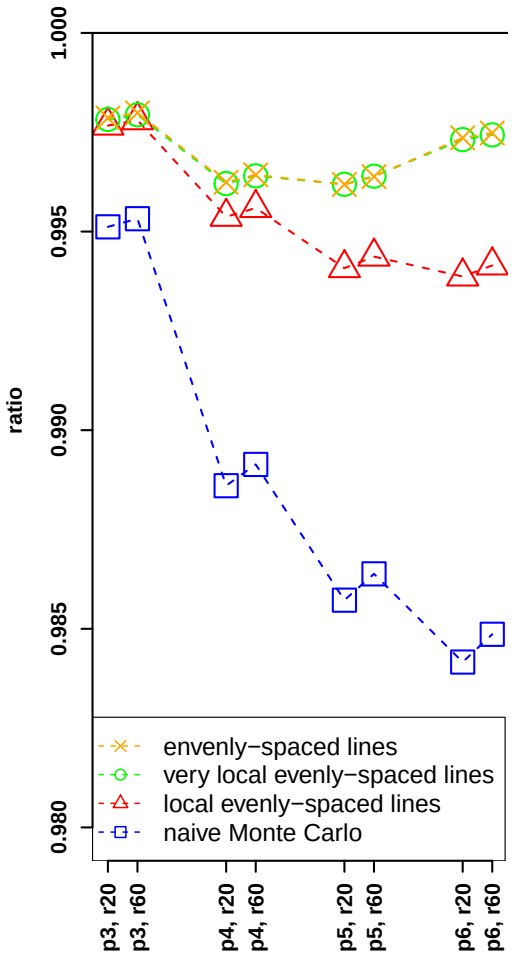


Figure 4.1: Context of [J7] in Section 4.3. Plot of the ratio S_K/K_4, by the four different methods described in the discussion of the figure.

Then, in [J16], we consider design matrices X with small RIP constants, and provide a non-asymptotic upper bound on $K_1(X, \mathcal{M}_s)$ in this case. The upper bound is as follows

Theorem 4.4. *Let X be a $n \times p$ matrix with $n, p \in \mathbb{N}$. Let $\delta = \delta(X, s)$. We have*

$$K_1(X, \mathcal{M}_s) \leq g(r, \alpha) \left[\sqrt{2\log(2p)} + 2\delta \left(\frac{\sqrt{1+\delta}}{1-\delta} \right) \sqrt{2s\log(6p/s)} \right] + h(r, \alpha),$$

where $g(r, \alpha)$ and $h(r, \alpha)$ are continuous functions of α on $(0, 1)$ for fixed $r \in \mathbb{N} \cup \{+\infty\}$ and, for any fixed $\alpha \in (0, 1)$, $g(r, \alpha) \to 1$ and $h(r, \alpha) \to \sqrt{2\log(4/\alpha)}$ as $r \to +\infty$.

When $\delta \to 0$, this upper bound can be summarized as $K_1(X, \mathcal{M}_s) = O(\sqrt{\log(p)} + \delta\sqrt{s\log(p/s)})$. Hence, it provides an interpolation between the order of magnitude in the orthogonal case and the general sparsity based upper bound. Furthermore, we show in [J16] that the upper bound of Theorem 4.4 is tight in many asymptotic regimes of p, s and δ. More precisely we have the following.

Proposition 4.5. *Let $(s_p, \delta_p)_{p \geq 0}$ be sequences of values such that $s_p < p$, $\delta_p > 0$, $\delta_p \to 0$ and $\delta_p \sqrt{s_p}\sqrt{1 - \log s_p / \log p + \log 6 / \log p} \to \infty$ as $p \to \infty$. Then Theorem 4.4 implies*

$$\sup_{\substack{n \in \mathbb{N} \\ s \leq s_p, X \in \mathbb{R}^{n \times p} \\ s.t. \ \ \delta(X,s) \leq \delta_p}} K_1(X, \mathcal{M}_{s_p}) \leq B\delta_p \sqrt{s_p}\sqrt{\log(6p/s_p)}, \tag{4.15}$$

where B is a constant. Moreover, there exists a sequence of design matrices X_p such that $\delta(X_p, s_p) \leq \delta_p$ and

$$K_1(X_p, \mathcal{M}_{s_p}) \geq A\delta_p \sqrt{s_p}\sqrt{\log\left(\min(1/\delta_p^2, \lfloor (p-1)/s_p \rfloor)\right)}, \tag{4.16}$$

where A is a constant and where $\lfloor \cdot \rfloor$ is the floor function.

In particular, if $\delta_p = O(p^{-\lambda})$ for some $\lambda > 0$ and if $\lfloor (p-1)/s_p \rfloor \geq 2$, then the above upper and lower bounds have the same rate.

In Proposition 4.5, the lower bound is obtained by generalizing the equicorrelated design matrix construction of Berk et al. [2013].

In [J16], we also study the case of large $n \times p$ random matrices and give the order of magnitude of the upper bound as a function of n and p. We discuss how large n should be in this case for this upper bound to be of smaller order than the general sparsity based upper bound. We then discuss the orders of magnitude of the lengths of the subsequent PoSI confidence intervals, and compare these orders of magnitude with upper bounds based on Euclidean norms provided in Kuchibhotla et al. [2018a]. We refer to [J16] for more details.

4.5 Extension to more general settings [S1]

The construction of universally valid confidence intervals in Berk et al. [2013] and [J13] is specific to the Gaussian linear case. A key ingredient for this construction, for instance in Section 4.1 and with $r = \infty$ (known variance case) for simplicity, is to define K_1 as the $1 - \alpha$ quantile of

$$\max_{M \in \mathcal{M}, i \in M} \left| \bar{s}_{i,M}^\top U \right|,$$

where U is a Gaussian vector. A general interpretation of this is that the vector of the $p2^{p-1}$ (normalized) differences between true and estimated projection based targets is a Gaussian vector, from which we consider a quantile of its supremum norm.

In [S1], we extend the construction of uniformly valid confidence intervals to general non Gaussian non linear settings, by using the same principle. That is, we show that the (normalized) vector of all the differences between estimators and targets converges to a Gaussian vector. We then construct confidence intervals, based on a PoSI constant similar to K_1 and show their asymptotic validity in the post-model-selection context. Furthermore, we show how to estimate, consistently or in a conservative way, the asymptotic covariance matrix of the aforementionned Gaussian vector. This is a significant improvement of Berk et al. [2013] and [J13], where the existence of an estimator $\hat{\sigma}^2$ of the variance, with strong distribution properties, has to be assumed, and can not be guaranteed to hold in fully general settings (see for instance Proposition 3.5 in [J13]).

General setting and joint asymptotic normality result We consider a general situation where we observe a data set $y \in \mathbb{R}^{n\times\ell}$ with distribution $\mathbb{P}_n$. We denote the i-th row of the data vector (matrix) y by $y_i \in \mathbb{R}^{1\times\ell}$, so that $y = (y_1^\top, \dots, y_n^\top)^\top$, and write $\mathbb{P}_{i,n}$ for the marginal distribution corresponding to that row. Throughout, we assume that the data generating distribution is of product form, that is $\mathbb{P}_n = \bigotimes_{i=1}^n \mathbb{P}_{i,n}$. Suppose further that one wants to conduct inference on $\mathbb{P}_n$, and intends to use as a working model an element of M_n, a set consisting of d nonempty sets of distributions $\mathbb{M}_{1,n}, \dots, \mathbb{M}_{d,n}$ on the Borel sets of $\mathbb{R}^{n\times\ell}$. Throughout d is fixed, i.e., does not depend on n. We emphasize that it is *not* assumed that $\mathbb{P}_n$ is contained in one of the sets $\mathbb{M}_{j,n}$ for $j = 1, \dots, d$. That is, the candidate set M_n might be *misspecified*.

For each model $\mathbb{M} \in \mathsf{M}_n$ we consider a corresponding target of inference $\theta^*_{\mathbb{M},n} = \theta^*_{\mathbb{M},n}(\mathbb{P}_n)$, say, which we take as given throughout the present section. Furthermore we assume that for every $\mathbb{M}_{j,n} \in \mathsf{M}_n$ the target is an element of a Euclidean space of finite dimension $m(\mathbb{M}_{j,n})$ which does not depend on n. We also assume that for every model $\mathbb{M} \in \mathsf{M}_n$ an estimator $\hat{\theta}_{\mathbb{M},n} : \mathbb{R}^{n\times\ell} \to \mathbb{R}^{m(\mathbb{M})}$ of the corresponding target $\theta^*_{\mathbb{M},n}$ is available. For a specific example, we refer to the model-dependent target $\beta_M^{(n)}$ and its estimator $\hat{\beta}_M$ in Section 4.1. We finally consider a model selection procedure $\hat{\mathbb{M}}_n : \mathbb{R}^{n\times\ell} \to \mathsf{M}_n$.

In this general framework, we assume an asymptotic linearity of the differences between estimators and targets, for each fixed model. We let $\mathbb{V}_n$ denote the covariance matrix under $\mathbb{P}_n$ and we let $A^{1/2}$ be the unique symmetric square root of a symmetric non-negative definite matrix A. If A is also invertible we let $A^{-1/2} = (A^{-1})^{1/2}$. For a vector v, we may let $v^{(j)}$ be its j-th component. We let $\hat{\theta}_n(y) = \hat{\theta}_n = (\hat{\theta}^\top_{\mathbb{M}_1,n}, \dots, \hat{\theta}^\top_{\mathbb{M}_d,n})^\top$ and $\theta^*_n = (\theta^{*\top}_{\mathbb{M}_1,n}, \dots, \theta^{*\top}_{\mathbb{M}_d,n})^\top$. We denote the dimension of $\hat{\theta}_n$ by $k = \sum_{j=1}^d m(\mathbb{M}_{j,n})$, which does not depend on n.

Condition 4.6. *There exist functions $g_{i,n} : \mathbb{R}^{1\times\ell} \to \mathbb{R}^k$ for $i = 1, \dots, n$, and $\Delta_n : \mathbb{R}^{n\times\ell} \to \mathbb{R}^k$, possibly depending on θ^*_n, so that for $y \in \mathbb{R}^{n\times\ell}$*

$$\hat{\theta}_n(y) - \theta^*_n \quad = \quad \sum_{i=1}^n g_{i,n}(y_i) + \Delta_n(y), \tag{4.17}$$

where, writing $r_n(y) := \sum_{i=1}^n g_{i,n}(y_i)$, it holds for every $i \in \{1, \dots, n\}$ and every $j \in \{1, \dots, k\}$

that

$$\mathbb{E}_{i,n}\left(g_{i,n}^{(j)}\right)=0 \quad \textit{and} \quad 0<\mathbb{V}_n\left(r_n^{(j)}\right)<\infty. \tag{4.18}$$

Furthermore, for every coordinate $j\in\{1,\ldots,k\}$ we have, with $\{\cdot\}$ the indicator function,

$$\begin{aligned}&\mathbb{V}_n^{-1}\left(r_n^{(j)}\right)\sum_{i=1}^{n}\int_{\mathbb{R}^{1\times\ell}}\left[g_{i,n}^{(j)}\right]^2\left\{|g_{i,n}^{(j)}|\geq\varepsilon\mathbb{V}_n^{\frac{1}{2}}(r_n^{(j)})\right\}d\mathbb{P}_{i,n}\to 0\\ &\textit{for every } \epsilon>0,\end{aligned} \tag{4.19}$$

and

$$\mathbb{P}_n\left(\left|\mathbb{V}_n^{-1/2}\left(r_n^{(j)}\right)\Delta_n^{(j)}\right|\geq\epsilon\right)\to 0 \textit{ for every } \epsilon>0. \tag{4.20}$$

This condition is satisfied in many applications, and is a relatively standard tool of asymptotic statistics. A benefit of this condition is that it is formulated in terms of rescaled summands, which enables a larger range of applications (see the discussion after Condition 1 in [S1]).

Next, we show a joint asymptotic normality result for the vector of differences between targets and estimators. We let d_w denote a distance metrizing weak convergence of probability measures on the Borel sets of the respective Euclidean space (cf. the discussion in Dudley [2002] pp. 393 for specific examples). We let $A^\dagger$ be Moore-Penrose pseudo inverse of a square matrix A and we let $A^{\dagger/2}=(A^\dagger)^{1/2}$. We let $\mathrm{diag}(A)$ be obtained by setting all the offdiagonal elements of A to zero. Finally, we abbreviate $A_i=A_{i,i}$ for $i\in\{1,\ldots,b\}$ when A is $b\times b$.

Lemma 4.7. *Under Condition 4.6,*

$$d_w\left(\mathbb{P}_n\circ\left[\mathrm{diag}(\mathbb{V}_n(r_n))^{\dagger/2}\left(\hat{\theta}_n-\theta_n^*\right)\right],\mathcal{N}\left(0,\mathrm{corr}(\mathbb{V}_n(r_n))\right)\right)\to 0. \tag{4.21}$$

In the above lemma, we state the asymptotic normality result in terms of difference between measures, because we do not need to assume that the asymptotic correlation matrix $\mathrm{corr}(\mathbb{V}_n(r_n))$ stabilizes as $n\to\infty$.

Confidence intervals based on consistent estimators of $\mathbb{V}_n(r_n)$ For $\alpha\in(0,1)$ and a covariance matrix Γ, we denote by $K_1(\Gamma,\alpha)$ the $1-\alpha$-quantile of the distribution of the supremum-norm $\|Z\|_\infty$ of $Z\sim N(0,\Gamma)$. We remark that the PoSI constant K_1 in Section 4.1 can be defined by the function $K_1(\Gamma,\alpha)$ applied to a certain $p2^{p-1}\times p2^{p-1}$ matrix Γ.

Then, the next theorem shows that, when a consistent estimator of $\mathbb{V}_n(r_n)$ (in terms of the correlation matrix and of the diagonal elements) is available, one can construct PoSI confidence intervals that are asymptotically valid. Given $\mathbb{M}=\mathbb{M}_{j,n}\in\mathsf{M}_n$ we abbreviate $\rho(\mathbb{M}):=\sum_{l=1}^{j-1}m(\mathbb{M}_{l,n})$ where sums over an empty index set are to be interpreted as 0.

Theorem 4.8. *Let $\alpha\in(0,1)$, suppose Condition 4.6 holds, and let $\hat{S}_n:\mathbb{R}^{n\times\ell}\to\mathbb{R}^{k\times k}$ be a sequence of functions so that for every $\varepsilon>0$*

$$\mathbb{P}_n\left(\|\mathrm{corr}(\hat{S}_n)-\mathrm{corr}\left(\mathbb{V}_n(r_n)\right)\|+\|\mathrm{diag}(\mathbb{V}_n(r_n))^{-1}\mathrm{diag}(\hat{S}_n)-I_k\|\geq\epsilon\right) \tag{4.22}$$

converges to 0. Define for every $\mathbb{M}\in\mathsf{M}_n$ and every $j=1,\ldots,m(\mathbb{M})$ the confidence interval

$$\mathrm{CI}_{1-\alpha,\mathbb{M}}^{(j),\mathrm{est}} \quad = \quad \hat{\theta}_{\mathbb{M},n}^{(j)}\pm\sqrt{[\hat{S}_n]_{\rho(\mathbb{M})+j}}\;K_1\left(\mathrm{corr}(\hat{S}_n),\alpha\right). \tag{4.23}$$

Then, $\mathbb{P}_n\left(\theta^{*(j)}_{\mathbb{M},n} \in \mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\mathbb{M}} \text{ for all } \mathbb{M} \in \mathsf{M}_n \text{ and all } j = 1,\dots,m(\mathbb{M})\right)$ *converges to* $1-\alpha$ *as* $n \to \infty$*. In particular, for every model selection procedure* $\hat{\mathbb{M}}_n$*, we have*

$$\liminf_{n\to\infty} \mathbb{P}_n\left(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{est}}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1,\dots,m(\hat{\mathbb{M}}_n)\right) \geq 1-\alpha. \tag{4.24}$$

The confidence intervals in the above theorem can be interpreted similarly as the PoSI confidence intervals in Sections 4.1 and 4.2. Indeed, they are centered at an asymptotically unbiased estimator and have half-length equal to an estimate of the standard deviation of the estimator multiplied by a PoSI constant that adjusts for multiplicity.

Confidence intervals based on estimators that consistently overestimate the diagonal entries of $\mathbb{V}_n(r_n)$ In case of strong model misspecification, it is typically difficult to obtain an estimator $\hat{S}_n$ satisfying the condition in Theorem 4.8 (see [S1]).

Hence, we show in [S1] that asymptotically valid confidence intervals can be obtained, in a larger range of situations than in Theorem 4.8. The two ingredients required for this are an upper bound on the PoSI constant $K_1\left(\mathrm{corr}\left(\mathbb{V}_n(r_n)\right),\alpha\right)$ and observable asymptotic upper bounds of the diagonal elements of $\mathbb{V}_n(r_n)$.

Theorem 4.9. *Let* $\alpha \in (0,1)$*, and suppose Condition 4.6 is satisfied. For every* n *and every* $j = 1,\dots,k$ *let* $\hat{\nu}^2_{j,n} \geq 0$ *be an estimator of* $\mathbb{V}_n(r_n^{(j)})$ *so that*

$$\mathbb{P}_n\left(\sqrt{\frac{[\mathbb{V}_n(r_n)]_j}{\hat{\nu}^2_{j,n}}} \geq 1+\epsilon\right) \to 0 \text{ for every } \epsilon > 0. \tag{4.25}$$

Let $\hat{K}_n \geq 0$ *be so that* $\hat{K}_n \geq K_1(\mathrm{corr}(\mathbb{V}_n(r_n)),\alpha)$ *holds eventually. For every* $\mathbb{M} \in \mathsf{M}_n$ *and every* $j = 1,\dots,m(\mathbb{M})$*, define the confidence interval*

$$\mathrm{CI}^{(j),\mathrm{oest}}_{1-\alpha,\mathbb{M}} \quad = \quad \hat{\theta}^{(j)}_{\mathbb{M},n} \pm \sqrt{\hat{\nu}^2_{\rho(\mathbb{M})+j,n}}\ \hat{K}_n. \tag{4.26}$$

Then, for every (measurable) model selection procedure $\hat{\mathbb{M}}_n$*, we have*

$$\liminf_{n\to\infty} \mathbb{P}_n\left(\theta^{*(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{oest}}_{1-\alpha,\hat{\mathbb{M}}_n} \text{ for all } j = 1,\dots,m(\hat{\mathbb{M}}_n)\right) \geq 1-\alpha. \tag{4.27}$$

In the above theorem, $\hat{K}_n$ can be computed similarly to K_4 in Section 4.2. We also remark that the rank of $\mathrm{corr}(\mathbb{V}_n(r_n))$ can be bounded in specific cases, resulting in a smaller upper bound $\hat{K}_n$, see [S1]. In [S1] we also provide a general method for computing the upper bounds $\hat{\nu}^2_{j,n}$.

Application to binary regression Now, we let the observation distribution $\mathbb{P}_n$ belong to the set $\mathbf{P}^{(\mathrm{bin})}_n(\tau)$, for a fixed $\tau > 0$. The set $\mathbf{P}^{(\mathrm{bin})}_n(\tau)$ is defined as follows: the distribution of a random vector $Y_n = (Y_{1,n},\dots,Y_{n,n})^\top$ is an element of $\mathbf{P}^{(\mathrm{bin})}_n(\tau)$ if and only if the n coordinates of Y_n are independent, each coordinate $Y_{i,n}$ takes on either 0 or 1, and $\mathrm{var}(Y_{i,n}) \geq \tau$. That is, we observe binary variables for which the probabilities of 0 or 1 are not too close to zero.

For the models, we consider binary regression based on a $n \times p$ design matrix X (with p fixed). Wet let M_n be the finite set of models, with fixed cardinality. Each model $\mathbb{M} \in \mathsf{M}_n$ is characterized by a link function $h : \mathbb{R} \to (0,1)$ and a subset M of $\{1,\dots,p\}$. Then, $\mathbb{M}$ is parametrized by

$\beta \in \mathbb{R}^{|M|}$ and, under a parameter β, the distribution in $\mathbb{M}$ has independent components, with values in $\{0,1\}$, where the component $i \in \{1,\dots,n\}$ has mean value $h(X_{i,n}[M]\beta)$ where $X_{i,n}$ is the line i of X.

For each model $\mathbb{M} \in \mathsf{M}_n$, we define a target vector $\beta^*_{\mathbb{M},n} \in \mathbb{R}^{|M|}$ as a Kullback Leibler divergence minimizer. We also define the maximum likelihood estimator $\hat{\beta}_{\mathbb{M},n}$. We refer to [S1] for more details. We show several uniform asymptotic results related to the existence of $\hat{\beta}_{\mathbb{M},n}$ and to its consistency and asymptotic normality. We remark that, in the binary regression context, for some points in the sample space $\{0,1\}^n$, the maximum likelihood estimator in the binary regression model does not exist Wedderburn [1976]. Our results show that this issue occurs with probability going to zero as $n \to \infty$, uniformly.

We construct PoSI confidence intervals based on the conservative principle of Theorem 4.9. For each $\mathbb{M} \in \mathsf{M}_n$ and for every $j = 1,\dots,m(\mathbb{M})$, the confidence interval for $\beta^{*,(j)}_{\mathbb{M},n}$ is written $\mathrm{CI}^{(j),\mathrm{bin}}_{1-\alpha,\mathbb{M}}$. These intervals are uniformly valid, as shown in the following theorem.

Theorem 4.10. *Let $\alpha \in (0,1)$ and $\tau > 0$ and suppose that technical conditions given in [S1] hold. Let $\hat{\mathbb{M}}_n$ be a model selection procedure, i.e., a map from the sample space $\{0,1\}^n$ to M_n. Then*

$$\liminf_{n\to\infty} \inf_{\mathbb{P}_n \in \mathbf{P}_n^{(\mathrm{bin})}(\tau)} \mathbb{P}_n\left(\beta^{*,(j)}_{\hat{\mathbb{M}}_n,n} \in \mathrm{CI}^{(j),\mathrm{bin}}_{1-\alpha,\hat{\mathbb{M}}_n} \ \forall j = 1,\dots,m(\hat{\mathbb{M}}_n)\right) \geq 1-\alpha. \tag{4.28}$$

In [S1], we also apply the general results in Theorems 4.8 and 4.9 to homoscedastic and heteroscedastic linear regression. We use estimators of variances that are always observable (without external assumptions as in Berk et al. [2013] and [J13]). In the homoscedastic case, we show quantitatively how the estimator of the variance goes from consistent to conservative, as the degree of model misspecification increases.

Numerical results In [S1], we provide simulation results for homoscedastic linear regression and binary regression. For linear regression, we compare our suggested confidence intervals with confidence intervals that are dedicated to the least angle regression model selector Tibshirani et al. [2018]. For binary regression, we compare our suggested confidence intervals with confidence intervals that are dedicated to the lasso model selector Taylor and Tibshirani [2017]. In both cases, we also study the naive intervals, that ignore the model selection step. Overall, we find that our suggested confidence intervals compare favorably to the others. Indeed, they are usually of similar length or shorter than those in Tibshirani et al. [2018] and Taylor and Tibshirani [2017]. Furhermore, our confidence intervals have coverage proportions above the nominal level in all the cases under consideration, while the coverage proportions of the intervals in Tibshirani et al. [2018] and Taylor and Tibshirani [2017] can become very small in some cases. The naive confidence intervals are of course shorter than the ones we suggest, but at the price of also having coverage proportions below the nominal level.

In Table 4.2 below (which is extracted from Tables 3 and 4 of the supplementary material to [S1], to which we refer for the full details), we illustrate these conclusions in a Monte Carlo study in the case of binary regression. We show the coverage proportions, median lengths and 0.9-quantile lengths of our suggested confidence intervals, of those in Taylor and Tibshirani [2017] and of the naive ones. We consider three settings with the lasso model model selector and one

model	cov. 0.9			med.			qua.		
selector	P	L	N	P	L	N	P	L	N
lasso (1)	0.99	0.89	0.84	4.26	7.44	2.09	6.97	43.33	3.42
lasso (2)	1.00	0.85	0.68	1.63	2.31	0.74	1.90	13.52	0.84
lasso (3)	1.00	0.25	0.98	2.22	1.23	1.01	2.83	3.50	1.24
sig. hun.	0.95		0.39	4.40		2.63	6.22		3.63

Table 4.2: Context of [S1] in Section 4.5. Coverage proportion with nominal level 0.9 (cov. 0.9), median length (med.) and 0.9-quantile length (qua.) of our suggested confidence intervals (P), of those in Taylor and Tibshirani [2017] (L) and of the naive ones (N) in a Monte Carlo study. We consider three settings with the lasso model model selector and one setting with a 'significance hunting' procedure.

setting with a 'significance hunting' procedure, that we find representative of some application practices. The conclusions drawn from the table are similar to those given above.

Chapter 5

Conclusion

5.1 Other research contributions [C2,S3,S4,S6,S7,S8,P1]

Here, we briefly summarize various other research contributions.

In [C2], we consider a neutron transport application, where the objective is to evaluate the probability that a particle reaches a sensitive target. The particle evolves as a discrete Markov chain, and can be subject to absorption, collision and medium change. The probability of interest is very small. We suggest an adaptation of the particle Monte Carlo algorithm of Guyader et al. [2011], together with the Hastings-Metropolis algorithm on Markov chain trajectories. We show that the resulting algorithm gives good results in practice.

In [S3], in the context of Gaussian process models, we suggest an iterative procedure for joint variable selection and optimization (or more general purposes). We also provide theoretical insight on the link between correlation lengths and variable importance. In particular, we show that, under relatively general conditions, the likelihood function of a Gaussian process model goes to infinity when the correlation length of an inactive variable goes to infinity. This is a theoretical justification for flagging as inactive the variables corresponding to large estimated correlation lengths, in practice.

In [S4], we address the computation of Shapley indices Owen and Prieur [2017] in sensitivity analysis. We show that, in the special case where the function is linear, and the inputs are jointly Gaussian with partitioned covariance matrix, significant savings in terms of computation cost can be obtained. This linear Gaussian setting occurs, for instance, in nuclear engineering, when considering cross section uncertainties Kawano et al. [2006].

In [S6] we consider a Gaussian process on $[0, 1]$ with known smoothness. We estimate the (microergodic) parameter C driving the behavior at zero of the variogram (that we assume to be stationary). The estimator we consider is based on quadratic variations. We show its consistency and asymptotic normality, for a large range of Gaussian processes and of sequences defining the variations. In simulations, we study the impact of the choice of the sequence on the asymptotic variance, and show the benefit of combining estimators obtained from several sequences.

In [S7], we study weighted pairwise likelihood estimators of the microergodic parameter for the exponential covariance function (see Section 2.5). We show that using the unconditional

likelihood of pairs of observations can yield inconsistent estimators if the range of admissible values for the variance is too large. In contrasts, we show that using the conditional likelihood based on pairs always yields a consistent estimator. We show asymptotic normality results and compare the asymptotic variances of pairwise likelihood estimators with that of the maximum likelihood estimator.

In [S8] we address Gaussian processes which inputs are permutations, or partial rankings. We provide functions based on permutation distances that are symmetric non-negative definite, so that they can be used to build parametric models of covariance functions. In the case of partial rankings, we provide computational simplifications of the computation of covariance functions, also based on these distances. Finally, we show that the increasing-domain asymptotic methods of [J3,J12] can be adapted to the case of permutation inputs. This enables us to show the consistency and asymptotic normality of the maximum likelihood estimator of the covariance parameters.

In [P1], we aim at extending the covariance functions of [J12], that are based on the Wasserstein distance for one-dimensional distributions, to multi-dimensional distributions. More precisely, we aim at constructing covariance functions on the space of distributions of $\mathbb{R}^d$, based on the notion of optimal transport. In this case, the functions of [J12] (see also (3.8)), that are indeed covariance functions for $d = 1$, would not be guaranteed to be covariance functions for $d > 1$. We suggest a construction of covariance functions, based on computing distances to Wasserstein barycenters Le Gouic and Loubes [2017]. We also provide some theoretical guarantees for these covariance functions.

5.2 Ongoing work

Here, I briefly describe my ongoing research collaborations.

With Anne Ruiz Gazen, Klaus Nordhausen and Joni Virta, we are preparing a manuscript, on a procedure for recovering independent components, for multivariate spatial Gaussian processes. We provide consistency and asymptotic normality results, under increasing-domain asymptotics.

On a related topic, with Klaus Nordhausen and Joni Virta, we are aiming at obtaining asymptotic results for procedures similar to the one mentioned above. We study a different asymptotic settings, for which other proof techniques are needed.

With Edouard Pauwels, we aim at obtaining theoretical results on the use of the Christoffel function Lasserre and Pauwels [2017] for distribution learning.

With Céline Helbert and Victor Picheny, we are working on a sequential design strategy based on Gaussian processes, for optimization in the case of computation failures. This topic entails a combination of Gaussian process based classification and optimization. An application to fan design in the car industry motivates this work.

With Yann Richet and Thomas Santner, we are working on a nuclear engineering application, where the objective is to find functional inputs of a computer model, that both are physically realistic and yield large output values for the computer model.

With Fabrice Gamboa and Jean-Michel Loubes, we are aiming at obtaining quantitative and graphical indicators for the interpretation of black box machine learning models. This work falls

within the scope of interpretability in machine learning.

Finally, with Thierry Klein and José Bétancourt, in the context of the French national research agency (ANR) grant 'Riscope', we are working on an application of Gaussian process models to early warning in he context of coastal flood hazards.

5.3 Open problems and prospects

Let us discuss some open problems and prospects, related to the topics of this habilitation manuscript.

On a theoretical standpoint, when considering the fixed-domain asymptotic framework for Gaussian processes, it seems that there exist several questions that are relatively fundamental and open. First, when considering maximum likelihood estimation of covariance parameters, even consistency can only be proved in specific settings. The settings where this can be done that are the most general are arguably these of Kaufman and Shaby [2013] (Matérn covariance functions in dimension $d \leq 3$) and Bevilacqua et al. [2018] (generalized Wendland covariance functions in dimension $d \leq 3$). Yet, the results obtained under increasing-domain asymptotics are significantly more general.

It is hence an important open problem to show the consistency of the maximum likelihood estimator in other classical settings. For instance, no results are known for the estimation of the smoothness parameter of the Matérn covariance function, or for the estimation of the correlation lengths when $d \geq 4$, although maximum likelihood is very commonly used in practice in these settings.

Consider next the case where a Gaussian process model, with a fixed given continuous covariance function, is fitted to a fixed continuous function. Then, it is not known, in full generality, whether the Gaussian process predictions will converge to the values of the fixed continuous function. Some results exist in Vazquez and Bect [2010b], Hangelbroek et al. [2010]. This question is important, in my opinion, because in practice Gaussian process models are indeed usually fitted to deterministic continuous functions (e.g. computer models). Covariance functions for which this convergence (consistency) would hold, for all continuous functions, could then be considered as 'robust choices'.

It would also be beneficial to obtain rates of convergence, in the same context as the consistency results of [J15], for sequential strategies based on Gaussian processes. Indeed, to my knowledge, most available rates of convergence concern optimization problems. On the other hand, Gaussian processes are useful in practice for other problems, such as failure domain estimation.

In the same vein as [J12,S8,P1], it seems that there is nowadays a need for statistical models for diverse and non-standard data. Modern machine learning techniques are indeed applied, for instance, to texts, graphs or points on manifolds. Gaussian process models that would be applicable to these contexts would be useful for many practical problems, in my opinion.

Consider then post-model-selection inference. It seems there that the PoSI approach tackles an important problem, by providing valid inference uniformly over the model selection procedure. Nevertheless, a concern that practitioners can have is that the confidence intervals may be larger

than what they are used to. The constant K_1 provides confidence intervals that cannot be made smaller if one wants to retain the full theoretical guarantees of the PoSI approach. Nevertheless, the constant K_1 can currently not be computed in practice for a large number p of explanatory variables. It is hence replaced by K_4 (see Sections 4.2 and 4.5) in these cases, but K_4 may or may not be a tight upper bound for K_1, depending on the design matrix X. Hence, finding more efficient algorithms for computing K_1, or obtaining other upper bounds, is key to the development of the PoSI approach, in my opinion.

Finally, in [S1] the PoSI approach is extended to more general settings than Gaussian linear regression. In terms of asymptotic theory, the results in [S1] are for a fixed number of models (equivalently a fixed number of variables in linear and binary regression). It would be interesting to obtain asymptotic guarantees also for a number of models going to infinity (the high dimensional case). Technical tools for this could be given in Chernozhukov et al. [2013], where Gaussian approximation results for maximums of sums of high dimensional random vectors are provided. We remark that, recently, Kuchibhotla et al. [2018b] seem to have followed a similar path, in the special case of linear models.

Appendix A

References

Bibliography

P. Abrahamsen. A review of Gaussian random fields and correlation functions. Technical report, Norwegian computing center, 1997.

M. Abramowitz and I. A. Stegun. Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables. Dover, New York, ninth Dover printing, tenth GPO printing edition, 1964.

E. Anderes. On the consistent separation of scale and variance for Gaussian random fields. The Annals of Statistics, 38:870–893, 2010.

I. Andrianakis and P. G. Challenor. The effect of the nugget on Gaussian process emulators of computer models. Computational Statistics and Data Analysis, 2012.

A. B. Antognini and M. Zagoraiou. Exact optimal designs for computer experiments via Kriging metamodelling. Journal of Statistical Planning and Inference, 140:2607–2617, 2010.

F. Bachoc. Parametric estimation of covariance function in Gaussian-process based Kriging models. Application to uncertainty quantification for computer experiments. PhD thesis, Université Paris-Diderot - Paris VII, 2013. Available at `https://tel.archives-ouvertes.fr/tel-00881002/document`.

J. Bect, D. Ginsbourger, L. Li, V. Picheny, and E. Vazquez. Sequential design of computer experiments for the estimation of a probability of failure. Statistics and Computing, 22 (3): 773–793, 2012.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference for high-dimensional sparse econometric models. Advances in Economics and Econometrics. 10th World Congress of the Econometric Society, Volume III,, pages 245–295, 2011.

A. Belloni, V. Chernozhukov, and C. Hansen. Inference on treatment effects after selection among high-dimensional controls. The Review of Economic Studies, **81**:608–650, 2014.

C. Berg, J. P. R. Christensen, and P. Ressel. Harmonic analysis on semigroups. Springer-Verlag, 1984.

R. Berk, L. Brown, A. Buja, K. Zhang, , and L. Zhao. Valid post-selection inference. The Annals of Statistics, 41(2):802–837, 2013.

M. Bevilacqua, A. Fassò, C. Gaetan, E. Porcu, and D. Velandia. Covariance tapering for multivariate Gaussian random fields estimation. Statistical Methods & Applications, pages 1–17, 2015.

M. Bevilacqua, T. Faouzi, R. Furrer, and E. Porcu. Estimation and prediction using generalized wendland covariance functions under fixed domain asymptotics. Annals of Statistics (in press), 2018.

M. Binois. Uncertainty quantification on pareto fronts and high-dimensional strategies in bayesian optimization, with applications in multi-objective automotive design. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne, 2015.

Z. I. Botev. The normal law under linear restrictions: simulation and estimation via minimax tilting. Journal of the Royal Statistical Society: Series B (Statistical Methodology), 79(1): 125–148, 2017. ISSN 1467-9868.

A. D. Bull. Convergence rates of efficient global optimization algorithms. Journal of Machine Learning Research, 12:2879–2904, 2011.

Y. Cao and D. J. Fleet. Generalized product of experts for automatic and principled fusion of Gaussian process predictions. Modern Nonparametrics 3: Automating the Learning Pipeline workshop at NIPS, Montreal, 2014.

L. Castera, H. Chan, M. Arrese, N. Afdhal, P. Bedossa, M. Friedrich-Rust, K.-H. Han, and M. Pinzani. EASL-ALEH clinical practice guidelines: non-invasive tests for evaluation of liver disease severity and prognosis. Journal of Hepatology, 63:237–264, 2015.

A. Charkhi and G. Claeskens. Asymptotic post-selection inference for the Akaike information criterion. Biometrika, 2018.

V. Chernozhukov, D. Chetverikov, and K. Kato. Gaussian approximations and multiplier bootstrap for maxima of sums of high-dimensional random vectors. The Annals of Statistics, 41 (6):2786–2819, 2013.

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.

H. Cohn and A. Kumar. Universally optimal distribution of points on spheres. Journal of the American Mathematical Society, 20(1):99–148, 2007.

H. Cohn, A. Kumar, and G. Minton. Optimal simplices and codes in projective spaces. Geometry & Topology, 20(3):1289–1357, 2016.

J. H. Conway, R. H. Hardin, and N. J. Sloane. Packing lines, planes, etc.: Packings in grassmannian spaces. Experimental mathematics, 5(2):139–159, 1996.

A. Cousin, H. Maatouk, and D. Rullière. Kriging of financial term-structures. European Journal of Operational Research, 255(2):631–648, 2016.

N. Cressie and S. Lahiri. The asymptotic distribution of REML estimators. Journal of Multivariate Analysis, 45:217–233, 1993.

S. Da Veiga and A. Marrel. Gaussian process modeling with inequality constraints. Annales de la faculté des sciences de Toulouse Mathématiques, 21(3):529–555, 4 2012.

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

J. Du, H. Zhang, and V. S. Mandrekar. Fixed-domain asymptotic properties of tapered maximum likelihood estimators. The Annals of Statistics, 37(6A):3330–3361, 12 2009.

O. Dubrule. Cross validation of Kriging in a unique neighborhood. Mathematical Geology, 15: 687–699, 1983.

R. M. Dudley. Real analysis and probability. Cambridge University Press, 2002.

M. Emmerich, K. Giannakoglou, and B. Naujoks. Single-and multiobjective optimization assisted by Gaussian random field metamodels. IEEE Transactions on Evolutionary Computation, 10 (4), 2006.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association, 96:1348–1360, 2001.

P. Feliot, J. Bect, and E. Vazquez. A Bayesian approach to constrained single- and multi-objective optimization. Journal of Global Optimization, 67(1):97–133, 2016.

W. Fithian, D. Sun, and J. Taylor. Optimal inference after model selection. arXiv:1410.2597, 2015.

S. R. Flaxman, Y.-X. Wang, and A. J. Smola. Who supported obama in 2012?: Ecological inference through distribution regression. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 289–298. ACM, 2015.

A. Forrester, A. Sobester, and A. Keane. Engineering design via surrogate modelling: a practical guide. Wiley, 2008.

S. Foucart and H. Rauhut. A mathematical introduction to compressive sensing. Basel: Birkhäuser, 2013.

R. Furrer and S. R. Sain. SPAM: A sparse matrix R package with emphasis on MCMC methods for Gaussian Markov random fields. Journal of Statistical Software, 36(10):1–25, 2010.

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

F. Gaudier. URANIE: The CEA DEN uncertainty and sensitivity platform. In Procedia - Social and Behavioral Sciences, volume 2, pages 7660–7661, 2010.

M. G. Genton, W. Kleiber, et al. Cross-covariance functions for multivariate geostatistics. Statistical Science, 30(2):147–163, 2015.

A. Genz. Numerical computation of multivariate normal probabilities. Journal of Computational and Graphical Statistics, 1:141–150, 1992.

D. Ginsbourger, J. Baccou, C. Chevalier, N. Garland, F. Perales, and Y. Monerie. Bayesian adaptive reconstruction of profile optima and optimizers. SIAM/ASA Journal on Uncertainty Quantification, 2(1):490–510, 2014.

T. Gneiting, W. Kleiber, and M. Schlather. Matérn cross-covariance functions for multivariate random fields. Journal of the American Statistical Association, 105(491):1167–1177, 2010.

S. Golchi, D. R. Bingham, H. Chipman, and D. A. Campbell. Monotone emulation of computer experiments. SIAM/ASA Journal on Uncertainty Quantification, 3(1):370–392, 2015.

R. Gramacy, G. Gray, S. Le Digabel, H. Lee, P. Ranjan, G. Wells, and S. Wild. Modeling an augmented Lagrangian for blackbox constrained optimization. Technometrics (with discussion), 58(1):1–11, 2016.

A. Guyader, N. Hengartner, and E. Matzner-Løber. Simulation and estimation of extreme quantiles and extreme probabilities. Applied Mathematics & Optimization, 64(2):171–196, 2011.

T. Hangelbroek, F. J. Narcowich, and J. D. Ward. Kernel approximation on manifolds i: bounding the Lebesgue constant. SIAM Journal on Mathematical Analysis, 42(4):1732–1760, 2010.

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

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

S. G. Hoggar. t-designs in projective spaces. European Journal of Combinatorics, 3(3):233–254, 1982.

I. Ibragimov and Y. Rozanov. Gaussian Random Processes. Springer-Verlag, New York, 1978.

L. Jaupi. Variable selection methods for multivariate process monitoring. In S. Ao, L. Gelman, D. Hukins, A. Hunter, and A. Korsunsky, editors, Proceedings of the World Congress of Engineering 2014, volume II, pages 1116–1120, 2014.

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

P. Kabaila and H. Leeb. On the large-sample minimal coverage probability of confidence intervals after model selection. Journal of the American Statistical Association, 101:619–629, 2006.

C. Kaufman and B. Shaby. The role of the range parameter for estimation and prediction in geostatistics. Biometrika, 100:473–484, 2013.

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

T. Kawano, K. Hanson, S. Frankle, P. Talou, M. Chadwick, and R. Little. Evaluation and propagation of the 239pu fission cross-section uncertainties using a monte carlo technique. Nuclear Science and Engineering, 153(1):1–7, 2006.

A. K. Kuchibhotla, L. D. Brown, A. Buja, E. I. George, and L. Zhao. A model free perspective for linear regression: Uniform-in-model bounds for post selection inference. arXiv preprint arXiv:1802.05801, 2018a.

A. K. Kuchibhotla, L. D. Brown, A. Buja, E. I. George, and L. Zhao. Valid post-selection inference in assumption-lean linear regression. arXiv preprint arXiv:1806.04119, 2018b.

J. B. Lasserre and E. Pauwels. The empirical christoffel function in statistics and machine learning. arXiv preprint arXiv:1701.02886, 2017.

T. Le Gouic and J.-M. Loubes. Existence and consistency of wasserstein barycenters. Probability Theory and Related Fields, 168(3-4):901–917, 2017.

L. Le Gratiet and J. Garnier. Asymptotic analysis of the learning curve for gaussian process regression. Machine Learning, pages 1–27, 2014.

J. D. Lee and J. E. Taylor. Exact post model selection inference for marginal screening. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, Advances in Neural Information Processing Systems 27, pages 136–144. Curran Associates, Inc., 2014.

J. D. Lee, D. L. Sun, Y. Sun, and J. E. Taylor. Exact post-selection inference, with application to the lasso. The Annals of Statistics, 44(3):907–927, 2016.

H. Leeb and B. M. Pötscher. Model selection and inference: Facts and fiction. Econometric Theory, 21:21–59, 2005.

H. Leeb and B. M. Pötscher. Performance limits for estimators of the risk or distribution of shrinkage-type estimators, and some general lower risk-bound results. Econometric Theory, 22:69–97, 2 2006.

H. Leeb and B. M. Pötscher. Model selection. In T. G. Andersen, R. A. Davis, J.-P. Kreiß, and T. Mikosch, editors, Handbook of Financial Time Series, pages 785–821, New York, NY, 2008. Springer.

W. Loh. Fixed domain asymptotics for a subclass of Matérn type Gaussian random fields. The Annals of Statistics, 33:2344–2394, 2005.

W. Loh and T. Lam. Estimating structured correlation matrices in smooth Gaussian random field models. The Annals of Statistics, 28:880–904, 2000.

H. Maatouk and X. Bay. Gaussian process emulators for computer experiments with inequality constraints. Mathematical Geosciences, 49(5):557–582, 2017.

K. Mardia and R. Marshall. Maximum likelihood estimation of models for residual covariance in spatial regression. Biometrika, 71:135–146, 1984.

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

J. B. Mockus, V. Tiesis, and A. Žilinskas. The application of Bayesian methods for seeking the extremum. In L. C. W. Dixon and G. P. Szegö, editors, Towards Global Optimization, volume 2, pages 117–129, North Holland, New York, 1978.

T. Muehlenstaedt, J. Fruth, and O. Roustant. Computer experiments with functional inputs and scalar outputs by a norm-based approach. Statistics and Computing, pages 1–15, 2016.

K. P. Murphy. Machine Learning: A Probabilistic Perspective (Adaptive Computation And Machine Learning Series). The MIT Press, 2012.

S. Nanty, C. Helbert, A. Marrel, N. Pérot, and C. Prieur. Sampling, metamodeling, and sensitivity analysis of numerical simulators with functional stochastic inputs. SIAM/ASA Journal on Uncertainty Quantification, 4(1):636–659, 2016.

T. Nickson, T. Gunter, C. Lloyd, M. A. Osborne, and S. Roberts. Blitzkriging: Kronecker-structured stochastic Gaussian processes. arXiv preprint arXiv:1510.07965, 2015.

A. B. Owen and C. Prieur. On Shapley value for measuring importance of dependent inputs. SIAM/ASA Journal on Uncertainty Quantification, 5(1):986–1002, 2017.

A. Pakman and L. Paninski. Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. Journal of Computational and Graphical Statistics, 23(2):518–542, 2014.

R. Paulo, G. Garcia-Donato, and J. Palomo. Calibration of computer models with multivariate output. Computational Statistics and Data Analysis, 56:3959–3974, 2012.

V. Picheny. A stepwise uncertainty reduction approach to constrained global optimization. In Proceedings of the 17^{th} International Conference on Artificial Intelligence and Statistics (AISTATS), 2014.

V. Picheny, D. Ginsbourger, O. Roustant, R. Haftka, and N.-H. Kim. Adaptive designs of experiments for accurate approximation of target regions. Journal of Mechanical Design, 132 (7), 2010.

B. Póczos, A. Singh, A. Rinaldo, and L. Wasserman. Distribution-free distribution regression. In Proceedings of the 16th International Conference on Artificial Intelligence and Statistics, volume 31 of JMLR Proceedings, pages 507–515, 2013.

B. M. Pötscher. Confidence sets based on sparse estimators are necessarily large. Sankhya, 71: 1–18, 2009.

G. Radulescu, D. E. Mueller, and J. C. Wagner. Sensitivity and uncertainty analysis of commercial reactor criticals for burnup credit. Nuclear Technology, 167(2):268–287, 2009.

P. Ranjan, D. Bingham, and G. Michailidis. Sequential experiment design for contour estimation from complex computer codes. Technometrics, 50(4):527–541, November 2008.

C. Rasmussen and C. Williams. Gaussian Processes for Machine Learning. The MIT Press, Cambridge, 2006.

J. Riihimäki and A. Vehtari. Gaussian processes with monotonicity information. In Journal of Machine Learning Research: Workshop and Conference Proceedings, volume 9, pages 645–652, 2010.

L. Roche and M. Pelletier. Modelling of the thermomechanical and physical processes in FR fuel pins using the GERMINAL code. In MOX Fuel Cycle Technologies for Medium and Long Term Deployment, page 322, 2000.

O. Roustant, D. Ginsbourger, and Y. Deville. DiceKriging, DiceOptim: Two R packages for the analysis of computer experiments by Kriging-based metamodeling and optimization. Journal of Statistical Software, 51(1), 2012.

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

T. Santner, B. Williams, and W. Notz. The Design and Analysis of Computer Experiments. Springer, New York, 2003.

B. A. Shaby and D. Ruppert. Tapered covariance: Bayesian estimation and asymptotics. Journal of Computational and Graphical Statistics, 21(2):433–452, 2012.

B. Shahriari, K. Swersky, Z. Wang, R. Adams, and N. de Freitas. Taking the human out of the loop: A review of bayesian optimization. Proceedings of the IEEE, 104(1):148–175, 2016.

T. Souders and G. Stenbakken. Cutting the high cost of testing. IEEE Spectrum, 28:48–51, 1991.

N. Srinivas, K. A., S. Kakade, and M. Seeger. Information-theoretic regret bounds for Gaussian process optimization in the bandit setting. IEEE Transactions on Information Theory, 58: 3250–3265, 2012.

M. Stein. Asymptotically efficient prediction of a random field with a misspecified covariance function. The Annals of Statistics, 16:55–63, 1988.

M. Stein. Bounds on the efficiency of linear predictions using an incorrect covariance function. The Annals of Statistics, 18:1116–1138, 1990a.

M. Stein. Uniform asymptotic optimality of linear predictions of a random field using an incorrect second-order structure. The Annals of Statistics, 18:850–872, 1990b.

M. Stein. Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York, 1999.

M. L. Stein. Statistical properties of covariance tapers. Journal of Computational and Graphical Statistics, 22(4):866–885, 2013.

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

S. Sundararajan and S. Keerthi. Predictive approaches for choosing hyperparameters in Gaussian processes. Neural Computation, 13:1103–1118, 2001.

J. Taylor and Y. Benjamini. RestrictedMVN: multivariate normal restricted by affine constraints. `https://cran.r-project.org/web/packages/restrictedMVN/index.html`, 2017. [Online; 02-Feb-2017].

J. Taylor and R. Tibshirani. Post-selection inference for l1-penalized likelihood models. Canadian Journal of Statistics, pages 1–21, 2017.

R. Tibshirani. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), pages 267–288, 1996.

R. J. Tibshirani, J. Taylor, R. Lockhart, and R. Tibshirani. Exact post-selection inference for sequential regression procedures. Journal of the American Statistical Association, 111(514): 600–620, 2016.

R. J. Tibshirani, A. Rinaldo, R. Tibshirani, L. Wasserman, et al. Uniform asymptotic inference and the bootstrap after model selection. The Annals of Statistics, 46(3):1255–1287, 2018.

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

S. van de Geer, P. Bühlmann, Y. Ritov, and R. Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. The Annals of Statistics, **42**:1166–1202, 2014.

E. Vazquez and J. Bect. A sequential Bayesian algorithm to estimate a probability of failure. In 15th IFAC Symposium on System Identification (SYSID 2009), 2009.

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

E. Vazquez and J. Bect. Pointwise consistency of the kriging predictor with known mean and covariance functions. In mODa 9–Advances in Model-Oriented Design and Analysis, pages 221–228. Springer, 2010b.

H. Wackernagel. Multivariate Geostatistics: An Introduction with Applications. Springer Berlin Heidelberg, 2003.

G. Wahba. Spline Models for Observational Data. Society for Industrial and Applied Mathematics, 1990.

H. Wang, G. Lin, and J. Li. Gaussian process surrogates for failure detection: a Bayesian experimental design approach. Journal of Computational Physics, 313:247–259, 2016.

R. W. M. Wedderburn. On the existence and uniqueness of the maximum likelihood estimates for certain generalized linear models. Biometrika, 63(1):27–32, 1976.

B. J. Williams, T. J. Santner, and W. I. Notz. Sequential design of computer experiments to minimize integrated response functions. Statistica Sinica, 10:1133–1152, 2000.

D. Yarotsky. Examples of inconsistency in optimization by expected improvement. Journal of Global Optimization, 56(4):1773–1790, 2013.

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

Z. Ying. Maximum likelihood estimation of parameters under a spatial sampling scheme. The Annals of Statistics, 21:1567–1590, 1993.

C. Zhang. Nearly unbiased variable selection under minimax concave penalty. The Annals of Statistics, 38(2):894–942, 2010.

C.-H. Zhang and S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. Journal of the Royal Statistical Society. Series B (Methodological), 76: 217–242, 2014.

H. Zhang. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. Journal of the American Statistical Association, 99:250–261, 2004.

H. Zhang and W. Cai. When doesn't cokriging outperform kriging? Statistical Science, 30(2): 176–180, 2015.

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

H. Zhang and D. Zimmerman. Toward reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92:921–936, 2005.

K. Zhang. Spherical cap packing asymptotics and rank-extreme detection. IEEE Transactions on Information Theory, 63(7):4572–4584, 2017.

M. Zuluaga, A. Krause, G. Sergent, and M. Püschel. Active learning for level set estimation. In International Joint Conference on Artificial Intelligence (IJCAI), 2013.