A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis

Dissertation
zur Erlangung des mathematisch-naturwissenschaftlichen
Doktorgrades
„Doctor rerum naturalium “
der Georg-August-Universität Göttingen

vorgelegt von
Michael Scheuerer
aus Amberg

Göttingen 2009

Referent: Prof. Dr. Martin Schlather
Koreferent: Prof. Dr. Robert Schaback
Tag der mündlichen Prüfung: 28.10.2009

Acknowledgement

At this place I would like to thank several people who supported my writing of this PhD thesis.

Above all I would like to thank my supervisor Prof. Martin Schlather. He was always available for my questions, gave me advice whenever I requested it, but also let me a lot of freedom to pursue my own ideas and to develop my own research interests within the scope of my PhD project.
I owe the same thanks to my co-supervisor Prof. Robert Schaback. Without his support and readiness to answer my numerous questions in the field approximation theory a thesis like this, across the borderlines of different fields of mathematics, would not have been possible.
This work was initiated by some individual and joint ideas of my two supervisors that came up in the course of their cooperation. I am grateful to have been trusted with the continuation of this project assured of total freedom but still full support by both of them.

Among several of my colleagues who helped me with useful ideas and suggestions I want to highlight Stefan Müller who was another important reference person for me for questions concerning approximation theory.
I am very grateful to Dr. Zakhar Kabluchko for drawing my attention to a proof given in [2] for a theorem on sample path regularity. The part of my research that finally led to the results in Section 5.5 was triggered by the idea of this proof.
I am also indebted to Dr. Emilio Porcu for his careful proofreading of my thesis. His numerous annotations clearly helped to improve this work.

My research was made possible through the financial support I received from the Deutsche Forschungsgemeinschaft in the framework of the Graduiertenkolleg 1023 "Identification in Mathematical Models: Synergy of Stochastic and Numerical Methods". I am grateful for this opportunity and for the various benefits entailed by my membership in this programme.

Finally I want to thank my fellow PhD students, above all Anja, Karola and Andree for helpful discussions and frequent cheering up.

Contents

1 Introduction **1**

2 Basic Notions of Probability Theory **4**
- 2.1 Measure and Probability . . . 4
- 2.2 Integration . . . 9
  - 2.2.1 The Lebesgue integral . . . 9
  - 2.2.2 L^p-Spaces . . . 13
- 2.3 Fourier Transforms of Measures . . . 14
- 2.4 Random Variables and Vectors . . . 15
- 2.5 Conditional Expectation . . . 21
- 2.6 Stochastic Processes . . . 24

3 Hilbert Spaces in Approximation Theory and Stochastics **26**
- 3.1 Reproducing-Kernel Hilbert Spaces . . . 26
- 3.2 Sobolev Spaces . . . 27
- 3.3 Canonical Isomorphism . . . 33

4 Expansions **35**
- 4.1 Mercer Eigenfunction Expansions . . . 35
- 4.2 Karhunen-Loève Expansion . . . 36

5 Sample Path Regularity of Random Fields **40**
- 5.1 Existence of Stochastic Processes . . . 40
- 5.2 Separable Random Fields . . . 42
- 5.3 Sample Path Regularity in the Gaussian Case . . . 46
  - 5.3.1 Continuity . . . 46
  - 5.3.2 Differentiability . . . 49
- 5.4 Measurable Random Fields . . . 57
- 5.5 Sample Path Regularity in the General Case . . . 60

6 Kernel Interpolation / Kriging **77**
- 6.1 Kernel Interpolation . . . 77
- 6.2 Generalized Kernel Interpolation . . . 79

6.3 Simple Kriging 82
6.4 Ordinary and Universal Kriging 83
6.5 Error Analysis 85
6.6 Best Prediction of Random Fields revisited 89
6.7 Kernel Interpolation / Kriging with Wrong Kernels 91

7 Parameter Identification **99**
7.1 Cross Validation 99
7.2 Maximum Likelihood 102
7.3 Comparing CV and ML in the Statistical Context 104
7.3.1 Estimating functions and information criteria 105
7.3.2 Accuracy of parameter estimates 112
7.3.3 Prediction accuracy with estimated parameters 116
7.3.4 Kriging variance prediction with estimated parameters 118
7.4 Maximum Likelihood revisited 123
7.5 Comparing CV and ML in a Numerical Analysis Framework 130
7.5.1 Approximation accuracy 131
7.5.2 Prediction of the L^2-approximation error 135
7.5.3 Choosing the smoothness of the interpolation kernel 137

Notation

$\mathbb{B}^d$	Borel σ-algebra on $\mathbb{R}^d$, page 4
$\mathbb{B}^d_T$	$\mathbb{B}^d \cap T$, page 13
$\mathcal{A}_1 \otimes \mathcal{A}_2$	product σ-algebra of $\mathcal{A}_1$ and $\mathcal{A}_2$, page 5
$\mathcal{C}(T)$	space of continuous functions over T, page 27
$\mathcal{C}^k(T)$	space of continuously differentiable functions over T, page 27
$\mathcal{H}_R$	reproducing kernel Hilbert space associated with a kernel R, page 27
$\mathrm{Cov}(X,Y)$	covariance of two RVs X and Y, page 18
$\mathbb{E}(X)$	expectation of a RV X, page 17
$\mathrm{Exp}(\lambda)$	exponential distribution, page 17
$\mathcal{N}(\mu,\sigma^2)$	Gaussian distribution, page 17
λ^d	Lebesgue measure on $(\mathbb{R}^d, \mathbb{B}^d)$, page 8
$\lfloor a \rfloor$	the biggest integer $\leq a$, page 30
$\mathrm{AC}(T)$	space of 'absolutely continuous on the line' functions over T, page 30
$\overline{\pi}_i$	projection on the subspace perpendicular to the i^{th} coordinate, page 29
π_i	projection on the i^{th} coordinate, page 29
$\mathcal{U}_{[a,b]}$	Uniform distribution, page 17
$\mathrm{Var}(X)$	variance of a RV X, page 17
$vol(T)$	volume (Lebesgue measure) of $T \subseteq \mathbb{R}^d$, page 58
$A \subset\subset B$	A is compactly contained in B, page 13
A_{ω_1}	ω_1-cross-section of an event $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$, page 7

$B_1 \times_i B_2$	Cartesian product of B_1 and B_2 taken in the i^{th} component, page 29
$B_\epsilon(a)$	open ball of radius ϵ centred at a, page 44
$D^\alpha f$	(weak) partial derivative in the direction $\left(e_1^{\alpha_1}, \dots, {e_d}^{\alpha_d}\right)'$, page 27
e_i	i^{th} unit vector in $\mathbb{R}^d$, page 27
$L^p(\Omega, \mathcal{A}, \mu)$	L^p-space over the measure space $(\Omega, \mathcal{A}, \mu)$, page 13
$L^p(T)$	L^p-space over the measure space $(T, \mathbb{B}_T^d, \lambda^d)$, page 13
$L^p_{\text{loc}}(T)$	local L^p-space over the measure space $(T, \mathbb{B}_T^d, \lambda^d)$, page 13
v', M'	transpose of some vector v or some matrix M, page 14
$W^{\mu,p}(T)$	Sobolev space of (fractional) order μ over T, page 30
$W^{k,p}(T)$	Sobolev space of (integer) order k over T, page 28
$X \sim \mathcal{D}$	the RV X is distributed according to the distribution $\mathcal{D}$, page 16
a.e.	("almost everywhere") everywhere outside a set of measure zero, page 7
a.s.	("almost surely") with probability one, page 7
i.i.d.	independently and identically distributed, page 16
RV	random variable, page 15
RVct	random vector, page 15

Chapter 1

Introduction

The present PhD thesis deals with the following generic mathematical problem:

Reconstruct a function $f : T \to \mathbb{R}$, where T is a domain in $\mathbb{R}^d$, based on its values at a finite set of data points ("sampling locations") $\{t_1, \ldots, t_n\} \subset T$.

Such kind of problem arises (directly or indirectly) in applications such as

- surface reconstruction
- numerical solution of partial differential equations
- fluid-structure interaction
- learning theory, neural networks and data mining
- modelling and prediction of environmental variables

Specific instances from different fields of application can be found in [8] and [41]. In order to derive "optimal" procedures for reconstruction and to provide a priori estimates of their precision it is necessary to make assumptions about f. There are basically two different fields of mathematics that deal with the above problem in different ways: approximation theory and spatial statistics.

In approximation theory f is assumed to belong to some Hilbert space $\mathcal{H}$ of functions of certain smoothness. This allows to use Taylor approximation techniques to derive bounds for the approximation error in terms of the density of the data points. Smoothness is a comparatively weak and flexible assumption, and the error bounds allow to control the precision whenever it is possible to control the sampling. In this work the focus will be on kernel interpolation. This procedure allows to adapt very flexibly the degree of smoothness of f and it turns out to be optimal in the sense that it leads to minimal approximation errors with respect to the norm $\|\cdot\|_{\mathcal{H}}$.

In some applications there is only limited or no control over the sampling and one has to get by with the (sometimes very sparse) data that are available. Typical examples are environmental modelling or mining where sampling involves high costs or is limited by lacking accessibility of the variable of interest. Moreover, in these applications the variable of interest is often a very rough function, and together with the sparsity of data this implies that error bounds obtained on the basis of Taylor approximation are only of limited use. A way out is possible if the stronger model assumption that comes with a statistical modelling approach is adequate: the assumption that f is a sample path of a (second-order) random field. Then, again optimal approximation procedures can be derived, and a satisfactory stochastic description of the approximation error is available.

It is quite remarkable that both approaches finally come up with the same type of approximant, despite the different model assumptions and motivations of its construction. Moreover, even the function that characterizes the magnitude of the approximation error appears - with different interpretations - in both frameworks. This motivates a synopsis of the two approaches that have so far been developed completely independent of each other (except for their common interest in classes of positive definite functions).

In this thesis we review and compare the approaches taken in approximation theory and spatial statistics to solve the reconstruction problem sketched above, and we contrast the different model assumptions that come with these approaches. Our main focus is to answer the following questions

1. To what extent do the probabilistic assumptions made in spatial statistics already imply assumptions about the smoothness of f?

2. How sensitive are approximation accuracy and the accuracy of approximation error prediction with respect to changes of the model / kernel parameters?

3. Which procedures can be used for parameter identification and how does the efficiency of those procedures depend on the adequacy of the model assumptions?

Substantial new contributions that considerably exceed the results in the stochastic literature are made in connection with the first question by proving a number of theorems providing an extensive characterization of the smoothness of the sample paths of second-order random fields. Another major contribution of this thesis consists in deriving an alternative interpretation of the maximum likelihood estimator for model parameters in spatial statistics which motivates its use in a non-statistical framework and helps to identify its scope of application.

In order to make this thesis completely self-contained and readable for mathematicians from both fields - statistics and numerical analysis - we give a summary of all relevant notions of probability theory (Chapter 2) and of reproducing kernel Hilbert spaces (RKHSs) and show their connection to the Hilbert spaces associated with stochastic processes (Chapter 3). This connection reappears in Chapter 4 where particular representations of RKHSs and stochastic processes are given that allow to draw first conclusions on the regularity of sample paths. Results of more immediate applicability are then derived - from a completely different starting point - in Chapter 5. After explaining the general principles behind the construction of stochastic processes we state and generalize some results from the literature on continuity and differentiability in the mean square sense. Continuity and differentiability of the sample paths is first discussed for the Gaussian case only. We then propose to focus on criteria for *weak differentiability* as it will turn out that this type of regularity is entirely determined by the second-order structure. Necessary and sufficient conditions on the second-order structure of the process are proved that ensure weak differentiability of any degree, and examples are presented to illustrate these statements.

In Chapter 6 we finally turn to the actual approximation problem, outline and contrast the different approaches to solve it and the different ways to quantify the approximation errors coming with these approaches. We also study the sensitivity of approximation accuracy and accuracy of the prediction of the approximation error to changes of the model parameters. Two standard methods (cross validation and maximum likelihood) for selecting such parameters are introduced in Chapter 7. An alternative derivation of the maximum likelihood procedure is given, allowing to widen its scope of application to the non-statistical framework and to better understand the limits of its applicability. Last but not least we compare the ability of both methods to select parameters that lead to a good reconstruction of f and to an adequate prediction of the approximation error in both a statistical and an approximation theory framework.

In this and the following chapters we are often sloppy with the nomenclature of the mathematical fields “stochastics”, “statistics”, “spatial statistics”, and “geostatistics”. These terms are used as synonyms whenever contrasting the stochastic approach with the deterministic approach. Likewise, when talking about the latter, we use the terms “numerical analysis” or “approximation theory”. The same is done with the nomenclature for the people working in these fields.

We often use a “/” between two expressions corresponding to terminology from spatial statistics and approximation theory when making statements that apply to both frameworks but describe objects with different nomenclature.

Chapter 2

Basic Notions of Probability Theory

In this section we will give some basic definitions and theorems from measure and probability theory, and from the theory of stochastic processes, which we will frequently need in subsequent sections. We mainly follow [3] and [5], and these are also our main references for proofs and further details in this chapter.

2.1 Measure and Probability

Definition 2.1.1. Let Ω be a set. Then $\mathcal{A} \subset 2^{\Omega}$ is called a σ-algebra on Ω if

1. $\Omega \in \mathcal{A}$
2. $A \in \mathcal{A} \;\Rightarrow\; A^c := \Omega \setminus A \in \mathcal{A}$
3. $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A} \;\Rightarrow\; \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{A}$

If $\mathcal{A}$ is a σ-algebra on Ω, then $(\Omega, \mathcal{A})$ is called measurable space and each $A \in \mathcal{A}$ is called a measurable set.

A σ-algebra can be interpreted as an information system on Ω. We will only be allowed to make (probabilistic) statements about subsets of Ω (so-called "events") that are contained in $\mathcal{A}$.

Every intersection of (finitely or infinitely many) σ-algebras in a set Ω is itself a σ-algebras in Ω. It follows that for every system Ξ of subsets of Ω there exists a smallest σ-algebra $\sigma(\Xi)$ containing Ξ. If $\mathcal{A} = \sigma(\Xi)$, then Ξ is called a generator of $\mathcal{A}$.

Example 2.1.2. An important example for a measurable space is $(\mathbb{R}^d, \mathbb{B}^d)$, the real space of dimension d, endowed with the Borel σ-algebra, which is by definition the smallest σ-algebra generated by the open subsets of $\mathbb{R}^d$.

The set of open subsets is not the only generator of $\mathbb{B}^d$. Other generators are

1. The set of all open cuboids (a,b) in $\mathbb{R}^d$ where
$$(a,b) \;:=\; \big\{x \in \mathbb{R}^d \,:\, a_i < x_i < b_i, \;\; \text{for all} \;\; 1 \leq i \leq d\big\}$$

2. The set of all closed cuboids $[a,b]$ in $\mathbb{R}^d$ where
$$[a,b] \;:=\; \big\{x \in \mathbb{R}^d \,:\, a_i \leq x_i \leq b_i, \;\; \text{for all} \;\; 1 \leq i \leq d\big\}$$

3. The set of all right half-open cuboids $[a,b)$ in $\mathbb{R}^d$ where
$$[a,b) \;:=\; \big\{x \in \mathbb{R}^d \,:\, a_i \leq x_i < b_i, \;\; \text{for all} \;\; 1 \leq i \leq d\big\}\,.$$

In many cases, the space Ω of interest is naturally represented as the Cartesian product of spaces Ω_i, $i \in I$, where I is an arbitrary index set. This motivates

Definition 2.1.3. Let $\{(\Omega_i, \mathcal{A}_i)\}_{i\in I}$ a set of measurable spaces, let $\Omega := \times_{i\in I}\, \Omega_i$ and $\pi_j : \Omega \longrightarrow \Omega_j$ the j-th canonical projection. Let
$$\mathcal{G} := \big\{\pi_i^{-1}(A_i) : A_i \in \mathcal{A}_i,\; i \in I\big\}$$
Then the product σ-algebra $\otimes_{i\in I}\mathcal{A}_i$ on Ω is defined as $\sigma(\mathcal{G})$.

By interpreting $\mathbb{R}^d$ as the n-fold Cartesian product of $\mathbb{R}^1$, Definition 2.1.3 yields a product σ-algebra $\otimes_{i=1}^d \mathbb{B}$ on $\mathbb{R}^d$, generated by all sets of the form
$$\begin{aligned}
&\big\{x \in \mathbb{R}^d \,:\, a_i < x_i < b_i, \;\; \text{for } \underline{\text{one}} \;\; 1 \leq i \leq d\big\},\\
&\big\{x \in \mathbb{R}^d \,:\, a_i \leq x_i \leq b_i, \;\; \text{for } \underline{\text{one}} \;\; 1 \leq i \leq d\big\}, \;\text{or}\\
&\big\{x \in \mathbb{R}^d \,:\, a_i \leq x_i < b_i, \;\; \text{for } \underline{\text{one}} \;\; 1 \leq i \leq d\big\}.
\end{aligned}$$
It is well-known that $\mathbb{B}^d = \otimes_{i=1}^d \mathbb{B}$, so we have yet another generator for $\mathbb{B}^d$.

When working with real-valued functions f, it is sometimes necessary that f takes values in the compact extension $\overline{\mathbb{R}} := \mathbb{R} \cup \{-\infty, \infty\}$ of $\mathbb{R}$. The corresponding Borel σ-algebra $\overline{\mathbb{B}}$ then consist of the sets
$$B_0,\; B_0 \cup \{-\infty\},\; B_0 \cup \{\infty\}, \;\; \text{and} \;\; B_0 \cup \{-\infty, \infty\} \qquad \text{with } B_0 \in \mathbb{B}.$$

Definition 2.1.4. Let $(\Omega, \mathcal{A})$ and $(E, \mathcal{B})$ be measurable spaces. A mapping $f : \Omega \to E$ is called $\mathcal{A}/\mathcal{B}$ measurable or simply measurable if
$$f^{-1}(B) \in \mathcal{A} \qquad \text{for all } B \in \mathcal{B}.$$

Example 2.1.5. Continuous mappings $f : \mathbb{R}^d \to \mathbb{R}^n$ are measurable. This follows directly from the definition of continuity ("preimages of open subsets are open") and the next theorem.

Theorem 2.1.6. *(cf. [3, Thm. 1.7.2]) Let $(\Omega, \mathcal{A})$ and $(E, \mathcal{B})$ be measurable spaces with $\mathcal{B} = \sigma(\Xi)$. A mapping $f : \Omega \to E$ is measurable if and only if*

$$f^{-1}(B) \in \mathcal{A} \quad \text{for all } B \in \Xi.$$

There is also a reverse point of view on the measurability of mappings:
Consider a set $((\Omega_i, \mathcal{A}_i))_{i \in I}$ of measurable spaces and a set $(f_i)_{i \in I}$ of measurable mappings $f_i : \Omega \to \Omega^i$, $i \in I$. Define $\sigma(f_i, i \in I)$ as the smallest σ-algebra with respect to which every f_i is still $\mathcal{A}/\mathcal{A}_i$ measurable. This sub-σ-algebra of $\mathcal{A}$ on Ω induced by $(f_i)_{i \in I}$ reflects their information content, and we will come back to this interpretation in subsection 2.5.

We give some results concerning the measurability of functions $f : \Omega \to \overline{\mathbb{R}}$:

Theorem 2.1.7. *([3, Thm. 2.1.2]) A function $f : \Omega \to \overline{\mathbb{R}}$ on $(\Omega, \mathcal{A})$ is $\mathcal{A}/\overline{\mathbb{B}}$ measurable if and only if it satisfies one of the following conditions*

1. $\big\{\omega \,:\, f(\omega) \leq a\big\} \in \mathcal{A} \quad$ *for all* $a \in \mathbb{R}$,
2. $\big\{\omega \,:\, f(\omega) < a\big\} \in \mathcal{A} \quad$ *for all* $a \in \mathbb{R}$,
3. $\big\{\omega \,:\, f(\omega) \geq a\big\} \in \mathcal{A} \quad$ *for all* $a \in \mathbb{R}$,
4. $\big\{\omega \,:\, f(\omega) > a\big\} \in \mathcal{A} \quad$ *for all* $a \in \mathbb{R}$.

Theorem 2.1.8. *([3, Thm. 2.1.3, 2.1.4]) For any two $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions $f, g : \Omega \to \overline{\mathbb{R}}$ on $(\Omega, \mathcal{A})$, the sets*

$$\big\{\omega \,:\, f(\omega) < g(\omega)\big\}, \qquad \text{and } \big\{\omega \,:\, f(\omega) = g(\omega)\big\},$$

(and of course their union and their complements) are all in $\mathcal{A}$. Moreover, the functions $f + g$, $f - g$ and $f \cdot g$ are also $\mathcal{A}/\overline{\mathbb{B}}$ measurable, provided they are defined everywhere on Ω.

Theorem 2.1.9. *([3, Thm. 2.1.5, Cor. 2.1.6, 2.1.7])*
Let $(f_n)_{n \in \mathbb{N}}$ be a sequence of $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions on $(\Omega, \mathcal{A})$, with values in $\overline{\mathbb{R}}$. Then each of the following functions is also $\mathcal{A}/\overline{\mathbb{B}}$ measurable:

$$|f_1|, \quad \sup(f_1, 0), \quad \inf(f_1, 0), \quad \sup_{n \in \mathbb{N}} f_n, \quad \inf_{n \in \mathbb{N}} f_n, \quad \limsup_{n \to \infty} f_n, \quad \liminf_{n \to \infty} f_n.$$

If $(f_n)_{n \in \mathbb{N}}$ is pointwise convergent, i.e. if $\lim\limits_{n \to \infty} f_n(\omega)$ exists in $\overline{\mathbb{R}}$ for each ω, then this limit function is also $\mathcal{A}/\overline{\mathbb{B}}$ measurable.

The following Lemma ([3, Lem. 3.2.1, 3.2.5]) links measurability of sets and mappings w.r.t. a product space to measurability of their cross-sections.

Lemma 2.1.10. *Let* $(\Omega_1, \mathcal{A}_1), (\Omega_2, \mathcal{A}_2)$ *and* $(E, \mathcal{B})$ *be measurable spaces.*
If $A \in \mathcal{A}_1 \otimes \mathcal{A}_2$*, then we have for the cross-sections:*

$$\begin{aligned} A_{\omega_1} &:= \{\omega_2 : (\omega_1, \omega_2) \in A\} \ \in \ \mathcal{A}_2 \qquad \textit{for all } \ \omega_1 \in \Omega_1 \qquad \textit{and} \\ A_{\omega_2} &:= \{\omega_1 : (\omega_1, \omega_2) \in A\} \ \in \ \mathcal{A}_1 \qquad \textit{for all } \ \omega_2 \in \Omega_2 \, . \end{aligned}$$

If $f : \Omega_1 \times \Omega_2 \to E$ *is* $\mathcal{A}_1 \otimes \mathcal{A}_2 \,/\, \mathcal{B}$ *measurable, then*

$f(\omega_1, \cdot)$ is $\mathcal{A}_2 \,/\, \mathcal{B}$ measurable for each fixed ω_1, and
$f(\cdot, \omega_2)$ is $\mathcal{A}_1 \,/\, \mathcal{B}$ measurable for each fixed ω_2 .

We are now ready to introduce the notion of a (probability) measure:

Definition 2.1.11. A set function $\mu : \mathcal{A} \to [0, \infty]$ on a measurable space $(\Omega, \mathcal{A})$ is called a measure on $\mathcal{A}$, and the triple $(\Omega, \mathcal{A}, \mu)$ a measure space, if

1. $\mu(\emptyset) = 0$
2. $(A_n)_{n\in\mathbb{N}} \subset \mathcal{A}, \ A_n \cap A_m = \emptyset \ (n \neq m) \implies \mu\Big(\bigcup\limits_{n\in\mathbb{N}} A_n\Big) = \sum\limits_{n\in\mathbb{N}} \mu(A_n)$

If, in addition, $\mu(\Omega) = 1$ then μ is called a probability measure (and usually denoted by P) and $(\Omega, \mathcal{A}, \mu)$ is called a probability space.

Definition 2.1.12. A measure μ on $(\Omega, \mathcal{A})$ is called σ-finite if there exists some countable or finite sequence of $\mathcal{A}$-sets $(A_n)_{n\in\mathbb{N}}$ so that

$$A_n \nearrow \Omega \quad \text{as } n \to \infty \qquad \text{and} \quad \mu(A_n) < \infty \quad \text{for all } n \in \mathbb{N}.$$

In many situations, the subsets of Ω with measure 0 (called null sets) are of particular interest having the interpretation of “exceptional sets” which are somehow negligible. From this point of view, it is often desirable that subsets of null sets are again null sets, although they might not even be measurable a priori. This motivates the following

Definition 2.1.13. A measure space $(\Omega, \mathcal{A}, \mu)$ is called complete if $A' \subset A$, $A \in \mathcal{A}$ and $P(A) = 0$ imply that $A' \in \mathcal{A}$ (and that $P(A') = 0$).

In any probability space it is possible to enlarge the σ-algebra and extend the measure in such way as to get a complete space [3, Sec. 1.5].

Notation: If some property holds for all $\omega \in \Omega \backslash N$, where $N \subset \Omega$ is a set of μ-measure 0, we say that the property holds (μ-)almost everywhere (a.e.).
In the same way, in a probabilistic context, we say that some statement is true (P-)almost surely (a.s.) if it holds for all ω outside a P-null set.

Example 2.1.14. (Lebesgue Measure)
As noted above, the Borel σ-algebra $\mathbb{B}^d$ on $\mathbb{R}^d$ is generated by the set

$$\Xi \;=\; \big\{[a,b) \subset \mathbb{R}^d \;:\; a,b \in \mathbb{R}^d,\; a_i < b_i \;\text{ for all }\; 1 \le i \le d\big\}$$

of right half-open cuboids in $\mathbb{R}^d$. Now define a measure λ^d on Ξ by

$$\lambda^d\big([a,b)\big) \;:=\; \prod_{i=1}^{d} (b_i - a_i).$$

It can be shown ([3, Sec. 1.4,1.5]) that this measure has a unique extension to $\mathbb{B}^d$. Moreover,

$$B_n \;:=\; [-n,n)^d, \quad n \in \mathbb{N}$$

defines a sequence $(B_n)_{n\in\mathbb{N}}$ in Ξ with $B_n \nearrow \mathbb{R}^d$ and $\lambda^d(B_n) = (2n)^d < \infty$, so λ^d is σ-finite. This measure λ^d is called Lebesgue-Borel measure, its completion is called Lebesgue measure.

Having defined products of measurable spaces, we need to introduce the notion of a product measure.

Definition and Theorem 2.1.15. *Let* $(\Omega,\mathcal{A}) := (\times_{i=1}^n \Omega_i, \otimes_{i=1}^n \mathcal{A}_i)$ *be the product space of measurable spaces* $(\Omega_i,\mathcal{A}_i,\mu_i)$, $i = 1,\dots,n$. *The product measure* $\mu = \otimes_{i=1}^n \mu_i$ *is defined by*

$$\mu\big(\times_{i=1}^n A_i\big) \;:=\; \prod_{i=1}^{n} \mu_i(A_i) \qquad \textit{for all } \; A_i \in \mathcal{A}_i,\; i = 1,\dots,n$$

Such a measure μ *exists and is uniquely determined on* $(\Omega,\mathcal{A})$ *by the preceding requirement (cf.[3, Thm. 3.3.1]).*

For later use we state the first Borel-Cantelli lemma:

Lemma 2.1.16. *([3, Lem. 6.2.1]) Let* $(\Omega,\mathcal{A},P)$ *be a probability space and* $(A_n)_{n\in\mathbb{N}}$ *be a sequence of* $\mathcal{A}$ *measurable events. Then*

$$\sum_{n\in\mathbb{N}} P(A_n) \;<\; \infty \quad \Longrightarrow \quad P\left(\bigcap_{n\in\mathbb{N}} \bigcup_{m\ge n} A_m\right) \;=\; 0.$$

2.2 Integration

2.2.1 The Lebesgue integral

Following [3, Ch. 2] we give the main ideas of integration of a real-valued function f w.r.t. some measure μ. The Lebesgue integral is a special case.

In this and all subsequent sections, $\mathbf{1}_A(x)$ denotes the indicator function of the set A, i.e.

$$\mathbf{1}_A(x) \;=\; \begin{cases} 1, & x \in A \\ 0, & x \notin A \end{cases}$$

Definition 2.2.1. Let $(\Omega, \mathcal{A}, \mu)$ be a measure space. A function $f : \Omega \to \mathbb{R}_+$ is called an <u>elementary</u> or <u>simple</u> function, if it allows the representation

$$f(\cdot) \;=\; \sum_{i=1}^{n} a_i \, \mathbf{1}_{A_i}(\cdot), \quad a_i \geq 0,\; A_i \in \mathcal{A}, \quad i = 1, \dots, n, \quad n \in \mathbb{N}. \tag{2.1}$$

If in addition the sets $A_1, \dots, A_n$ are pairwise disjoint with $\Omega = \bigcup_{i=1}^{n} A_i$, then (2.1) is called <u>normal representation</u> of f.

Clearly, a normal representation of an elementary function f always exists, but it is not unique. However, this is of no concern for integration.

Definition and Lemma 2.2.2. *Let $f : \Omega \to \mathbb{R}_+$ be an elementary function on $(\Omega, \mathcal{A}, \mu)$. Then the number*

$$\int_\Omega f(\omega)\, \mu(d\omega) \;:=\; \sum_{i=1}^{n} a_i \, \mu(A_i) \quad \in \overline{\mathbb{R}}_+$$

is called the <u>(μ-)integral</u> of f (over Ω).

It is independent of the chosen normal representation.

This definition of integrals can be extended to nonnegative $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions f. Such a function can always be represented as the limit of an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of elementary functions. Indeed, by defining

$$f_n(\omega) \;:=\; \sum_{j=1}^{n\cdot 2^n} (j-1)\, 2^{-n} \cdot \mathbf{1}_{\left\{\frac{j-1}{2^n} \leq f(\omega) < \frac{j}{2^n}\right\}}(\omega) \;+\; n \cdot \mathbf{1}_{\{n \leq f(\omega)\}}(\omega), \quad \omega \in \Omega,$$

we obtain such a sequence with $f = \sup_{n\in\mathbb{N}} f_n$ but again, the f_n are not unique.

Definition and Lemma 2.2.3. *Let $f:\Omega\to\overline{\mathbb{R}}_+$ be a $\mathcal{A}/\overline{\mathbb{B}}$ measurable function on $(\Omega,\mathcal{A},\mu)$, and $(f_n)_{n\in\mathbb{N}}$ an increasing sequence of elementary functions with $f=\sup\limits_{n\in\mathbb{N}} f_n$. Then the number*

$$\int_\Omega f(\omega)\,\mu(d\omega) \;:=\; \sup_{n\in\mathbb{N}}\int_\Omega f_n(\omega)\,\mu(d\omega) \quad\in\overline{\mathbb{R}}_+$$

is called the <u>(μ-)integral</u> of f (over Ω).

It is independent of the particular sequence $(f_n)_{n\in\mathbb{N}}$.

Finally the definition of the integral is extended to certain measurable functions of arbitrary sign. To this end, for every function $f:\Omega\to\overline{\mathbb{R}}$, we set

$$f^+ \;:=\; \sup(f,0) \qquad\text{and}\qquad f^- \;:=\; -\inf(f,0).$$

Clearly, $f^+\geq 0$, $f^-\geq 0$ and we have $f=f^+-f^-$ and $|f|=f^++f^-$. Hence, by Theorem 2.1.8 and 2.1.9, if f is $\mathcal{A}/\overline{\mathbb{B}}$ measurable so is f^+ and f^-.

Definition 2.2.4. Let $f:\Omega\to\overline{\mathbb{R}}$ be a $\mathcal{A}/\overline{\mathbb{B}}$ measurable function on $(\Omega,\mathcal{A},\mu)$ so that at least one of the (μ-)integrals

$$\int_\Omega f^+(\omega)\,\mu(d\omega) \qquad\text{and}\qquad \int_\Omega f^-(\omega)\,\mu(d\omega) \tag{2.2}$$

is finite. Then the number

$$\int_\Omega f(\omega)\,\mu(d\omega) \;:=\; \int_\Omega f^+(\omega)\,\mu(d\omega) \;-\; \int_\Omega f^-(\omega)\,\mu(d\omega) \quad\in\overline{\mathbb{R}}$$

is called the <u>(μ-)integral</u> of f (over Ω).

If both (μ-)integrals in (2.2) are finite, then f is said to be <u>(μ-)integrable</u>.

Remark 2.2.5. So far integration was always over the whole of Ω. Now, for any $A\in\mathcal{A}$ we know that if $f:\Omega\to\overline{\mathbb{R}}$ is an $\mathcal{A}/\overline{\mathbb{B}}$ measurable function so is $\mathbf{1}_A f$, and we define

$$\int_A f(\omega)\,\mu(d\omega) \;:=\; \int_\Omega \mathbf{1}_A(\omega)\,f(\omega)\,\mu(d\omega).$$

We note some basic properties of the μ-integral:

Theorem 2.2.6. *Let f,g be (μ-)integrable functions on $(\Omega,\mathcal{A},\mu)$. Then*

1. $f\leq g \quad\Longrightarrow\quad \displaystyle\int_\Omega f(\omega)\,\mu(d\omega) \;\leq\; \int_\Omega g(\omega)\,\mu(d\omega).$

2. *for any $\alpha, \beta \in \mathbb{R}$ the function $\alpha\, f + \beta\, g$ is (μ-)integrable and*

$$\int_\Omega \alpha\, f(\omega) + \beta\, g(\omega)\; \mu(d\omega) \;=\; \alpha \int_\Omega f(\omega)\, \mu(d\omega) + \beta \int_\Omega g(\omega)\, \mu(d\omega).$$

3. $\left| \int_\Omega f(\omega)\, \mu(d\omega) \right| \;\le\; \int_\Omega \big|f(\omega)\big|\, \mu(d\omega).$

As an immediate consequence of part *2.* in Thm. 2.2.6 we note that both integrals in (2.2) are finite (i.e. f is integrable) if and only if $|f|$ is integrable.

One of the big strengths of the μ-integral (which is the Lebesgue integral if μ is the Lebesgue measure) compared to the Riemann integral lies in the validity of the following theorems, which provide sufficient conditions under which the passage to the limit of a sequence of functions and integration can be interchanged.

Theorem 2.2.7. *(Monotone Convergence Theorem, [3, Thm. 2.3.4])*
For an increasing sequence $(f_n)_{n\in\mathbb{N}}$ of nonnegative $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions on $(\Omega, \mathcal{A}, \mu)$ it holds that

$$\int_\Omega \Big(\sup_{n\in\mathbb{N}} f_n\Big)(\omega)\, \mu(d\omega) \;=\; \sup_{n\in\mathbb{N}} \int_\Omega f_n(\omega)\, \mu(d\omega).$$

Lemma 2.2.8. *(Fatou's Lemma, [3, Lem. 2.7.1])*
For every sequence $(f_n)_{n\in\mathbb{N}}$ of nonnegative $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions on $(\Omega, \mathcal{A}, \mu)$ it holds that

$$\int_\Omega \Big(\liminf_{n\to\infty} f_n\Big)(\omega)\, \mu(d\omega) \;\le\; \liminf_{n\to\infty} \int_\Omega f_n(\omega)\, \mu(d\omega).$$

Lemma 2.2.9. *(Dominated Convergence Theorem, [5, Thm. 16.4])*
Let $(f_n)_{n\in\mathbb{N}}$ and f all be $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions on $(\Omega, \mathcal{A}, \mu)$, and let g be a nonnegative μ-integrable function on $(\Omega, \mathcal{A}, \mu)$. If

$$|f_n| \le g \quad \textit{a.e.} \quad \textit{for all } n \in \mathbb{N}, \qquad \textit{and} \quad f_n \to f \quad \textit{a.e.} \quad \textit{as } n \to \infty,$$

then

$$\int_\Omega f(\omega)\, \mu(d\omega) \;=\; \lim_{n\to\infty} \int_\Omega f_n(\omega)\, \mu(d\omega).$$

In the following sections we will consider measure spaces $(\Omega', \mathcal{A}', \mu')$ whose measure μ' is defined indirectly by a $\mathcal{A}/\mathcal{A}'$ measurable mapping T from a measure spaces $(\Omega, \mathcal{A}, \mu)$ to $(\Omega', \mathcal{A}')$ by

$$\mu'(A') \;:=\; \mu\big(T^{-1}(A')\big), \qquad A' \in \mathcal{A}'.$$

The following theorem shows the connection between μ- and μ'-integrals:

Theorem 2.2.10. *(Transformation theorem, [3, Cor. 2.10.2])*
Let $(\Omega, \mathcal{A}, \mu)$ and $(\Omega', \mathcal{A}', \mu')$ be as above, and let $f : \Omega' \to \overline{\mathbb{R}}$ be an $\mathcal{A}'/\overline{\mathbb{B}}$ measurable function. Then the μ'-integrability of f' implies the μ-integrability of $f' \circ T$ and conversely. In this case we have

$$\int_{\Omega'} f'(\omega')\, \mu'(d\omega') \;=\; \int_{\Omega} (f' \circ T)(\omega)\, \mu(d\omega).$$

The next theorem ([3, Thm. 3.2.6, Cor. 3.2.7]) shows the connection between the full integral and the marginal integrals of functions on product spaces.

Theorem 2.2.11. *(Fubini's Theorem)*
Let $(\Omega_i, \mathcal{A}_i, \mu_i)$, $i = 1, 2$ be σ-finite measure spaces and let $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ a $\mathcal{A}_1 \otimes \mathcal{A}_2 \,/\, \overline{\mathbb{B}}$ measurable function. Define F_1, F_2 by

$$F_1(\omega_1) := \int_{\Omega_2} f(\omega_1, \omega_2)\, \mu_2(d\omega_2), \qquad F_2(\omega_2) := \int_{\Omega_1} f(\omega_1, \omega_2)\, \mu_1(d\omega_1).$$

If f is nonnegative, then F_1 and F_2 are $\mathcal{A}_1/\mathcal{B}$ and $\mathcal{A}_2/\mathcal{B}$ measurable, respectively,

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, (\mu_1 \otimes \mu_2)\big(d(\omega_1, \omega_2)\big) \;=\; \int_{\Omega_1} F_1(\omega_1)\, \mu_1(d\omega_1) \tag{2.3}$$

$$\text{and} \quad \int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\, (\mu_1 \otimes \mu_2)\big(d(\omega_1, \omega_2)\big) \;=\; \int_{\Omega_2} F_2(\omega_2)\, \mu_2(d\omega_2) \tag{2.4}$$

(if one side of (2.3) or (2.4) is infinite, so is the other).

If f is $\mu_1 \otimes \mu_2$-integrable, then $f(\omega_1, \cdot)$ is μ_2-integrable for μ_1-almost all ω_1 and $f(\cdot, \omega_2)$ is μ_1-integrable for μ_2-almost all ω_2. Further, F_1 is defined μ_1-a.e., F_2 is defined μ_2-a.e., and again (2.3) and (2.4) hold.

We have already emphasized the importance of null sets and introduced the notion of almost everywhere properties. The following theorem (see [3, Sec. 2.5]) shows the significance of these concepts in integration theory.

Theorem 2.2.12. *Let $f, g : \Omega \to \overline{\mathbb{R}}$ be two $\mathcal{A}/\overline{\mathbb{B}}$ measurable functions on $(\Omega, \mathcal{A}, \mu)$ that are μ-a.e. equal. Then*

1. $\displaystyle\int_{\Omega} f(\omega)\, \mu(d\omega) = 0 \quad\Longleftrightarrow\quad f = 0 \quad a.e.$
2. *if f and g are nonnegative, then* $\displaystyle\int_{\Omega} f(\omega)\, \mu(d\omega) \;=\; \int_{\Omega} g(\omega)\, \mu(d\omega).$
3. *if f is μ-integrable, then so is g and* $\displaystyle\int_{\Omega} f(\omega)\, \mu(d\omega) \;=\; \int_{\Omega} g(\omega)\, \mu(d\omega).$

4. if f is μ-integrable, then it is μ-a.e. finite on Ω.

Note that this allows us to define the integral for a function f defined only almost everywhere on Ω, provided that f can be extended to an integrable function f^* on Ω.

Following [5, Sec. 19], we can now introduce the L^{p}-Spaces.

2.2.2 L^{p}-Spaces

Fix a measure space $(\Omega, \mathcal{A}, \mu)$. For a $\mathcal{A}/\mathbb{B}$ measurable function $f : \Omega \to \mathbb{R}$ and $1 \leq p \leq \infty$ define

$$\|f\|_{L^p(\Omega)} := \left(\int_\Omega |f|^p \, \mu(d\omega) \right)^{\frac{1}{p}}, \quad 1 \leq p < \infty, \qquad \text{and} \tag{2.5}$$

$$\|f\|_{L^\infty(\Omega)} := \operatorname{ess\,sup}_\Omega |f|. \tag{2.6}$$

where $\operatorname{ess\,sup}_\Omega |f| = \inf \big\{ a \in \mathbb{R} : \mu(\{\omega : |f(\omega)| > a\}) = 0 \big\}$.

Then for any $1 \leq p \leq \infty$ we define the function space

$$L^p(\Omega, \mathcal{A}, \mu) := \big\{ f : \Omega \to \mathbb{R} : f \text{ is } \mathcal{A}/\mathbb{B} \text{ measurable and } \|f\|_{L^p(\Omega)} < \infty \big\}.$$

If $\Omega = T \subset \mathbb{R}^d$, $\mathcal{A} = \mathbb{B}^d_T := \mathbb{B}^d \cap T$ and $\mu = \lambda^d$ (restricted to $\mathbb{B}^d_T$), then $\mathbb{B}^d_T$ and λ^d are usually dropped from the notation and one writes $L^p(T)$ instead of $L^p(T, \mathbb{B}^d_T, \lambda^d)$ and

$$\int_T f(x)\, dx \quad \text{instead of} \quad \int_T f(x)\, \lambda^d(dx).$$

In this context, spaces of locally integrable functions are also of interest. Writing $I \subset\subset T$ for a subset I that is compactly contained in T , i.e. $I \subset \overline{I} \subset T$ and $\overline{I}$ is compact, we further define

$$L^p_{\text{loc}}(T) := \big\{ f : T \to \mathbb{R} : f \in L^p(I) \text{ for each } I \subset\subset T \big\}.$$

The great utility of L^p-spaces is due to their good mathematical structure:

Theorem 2.2.13. *Let $(\Omega, \mathcal{A}, \mu)$ be a measure space and $1 \leq p \leq \infty$. If we identify functions that are equal μ-a.e. the space $L^p(\Omega, \mathcal{A}, \mu)$ defined above becomes a normed vector space with the norm defined in (2.5) and (2.6) respectively. Moreover, it is complete under the corresponding metric.*

Theorem 2.2.13 says that, for any $1 \leq p \leq \infty$, $L^p(\Omega, \mathcal{A}, \mu)$ is a Banach space. For $p = 2$ we can even make it a Hilbert space by defining a scalar product

$$(f,g)_\mu \;:=\; \int_\Omega fg\,\mu(d\omega) \qquad f,g \in L^2(\Omega, \mathcal{A}, \mu).$$

In the following section we will frequently encounter this kind of Hilbert space either w.r.t. the Lebesgue measure or w.r.t. some probability measure.

The following Lemma allows to draw conclusion about the integrability of products of functions:

Lemma 2.2.14. *(Hölder's inequality)*
Let $1 \leq p, q \leq \infty$ so that $p^{-1} + q^{-1} = 1$ (with the convention $\infty^{-1} = 0$). For $f \in L^p(\Omega, \mathcal{A}, \mu)$ and $g \in L^q(\Omega, \mathcal{A}, \mu)$ it holds that $f \cdot g$ is μ-integrable and

$$\int_\Omega |f\,g|\,\mu(d\omega) \;\leq\; \|f\|_{L^p(\Omega)} \cdot \|g\|_{L^q(\Omega)}.$$

We conclude this subsection by introducing the concept of weak convergence of sequences of finite measures on $(\mathbb{R}^d, \mathbb{B}^d)$:

Definition 2.2.15. Denote by $\mathcal{C}_b(\mathbb{R}^d)$ the set of all continuous and bounded functions $f : \mathbb{R}^d \to \mathbb{R}$. A series of finite measures $(\mu_n)_{n\in\mathbb{N}}$ on $(\mathbb{R}^d, \mathbb{B}^d)$ is called weakly convergent towards a finite measure μ on $(\mathbb{R}^d, \mathbb{B}^d)$, if

$$\lim_{n\to\infty} \int_\Omega f(\omega)\,\mu_n(d\omega) \;=\; \int_\Omega f(\omega)\,\mu(d\omega) \qquad \text{for all } \; f \in \mathcal{C}_b(\mathbb{R}^d).$$

In this case we write $\mu_n \xrightarrow{w} \mu$.

2.3 Fourier Transforms of Measures

We shall briefly introduce the concept of Fourier transforms of measures, which are a useful tool for working with probability measures. For proofs and further details we refer to [3, Sec. 8.1, 8.2].

Definition 2.3.1. Let μ be a finite measure on the measure space $(\mathbb{R}^n, \mathbb{B}^n)$. Then the function $\widehat{\mu} : \mathbb{R}^n \to \mathbb{R}$ defined by

$$\widehat{\mu}(\tau) \;:=\; \int_{\mathbb{R}^n} e^{i\tau' y}\,\mu(dy) \;=\; \int_{\mathbb{R}^n} \cos(\tau' y)\,\mu(dy) + i \int_{\mathbb{R}^n} \sin(\tau' y)\,\mu(dy) \tag{2.7}$$

is called the Fourier transform of μ (by τ' we denote the transpose of τ).

We note some basic properties of Fourier transforms:

Lemma 2.3.2. *Let μ, ν be a finite measures on the measure space $(\mathbb{R}^n, \mathbb{B}^n)$ and $\widehat{\mu}, \widehat{\nu}$ their Fourier transforms according to (2.7). Then*

1. *$\widehat{\mu}(\tau)$ is defined for every $\tau \in \mathbb{R}^n$;*
2. *$\widehat{\mu}(0) = \mu(\mathbb{R}^n)$;*
3. *$\widehat{\mu}$ is uniformly continuous on $\mathbb{R}^n$;*
4. *μ is a symmetric measure $\iff$ $\widehat{\mu}$ is real-valued and symmetric;*
5. *$\widehat{\mu}(\tau) = \widehat{\nu}(\tau)$ for all $\tau \in \mathbb{R}^n \iff \mu = \nu$.*

Because of the last uniqueness property, Fourier transforms are usually called characteristic functions in the stochastic literature. We will stick to the term "Fourier transform" to avoid confusion with characteristic functions of sets.

The next theorem shows, that weak convergence of measures is equivalent to pointwise convergence of their Fourier transforms:

Theorem 2.3.3. *([3, Thm. 8.2.7]) Let μ be finite measure on $(\mathbb{R}^n, \mathbb{B}^n)$, and $(\mu_n)_{n\in\mathbb{N}}$ a sequence of finite measures on $(\mathbb{R}^n, \mathbb{B}^n)$. Then $\mu_n \xrightarrow{w} \mu$ implies*

$$\widehat{\mu}_n(\tau) \to \widehat{\mu}(\tau) \quad \text{as } n \to \infty, \qquad \text{for all } \tau \in \mathbb{R}^n,$$

and the convergence is uniform on every compact subset of $\mathbb{R}^n$. If in turn there exists a function $f : \mathbb{R}^n \to \mathbb{C}$ that is continuous at 0, so that

$$\widehat{\mu}_n(\tau) \to f(\tau) \quad \text{as } n \to \infty, \qquad \text{for all } \tau \in \mathbb{R}^n,$$

then there exists a finite measure μ on $(\mathbb{R}^n, \mathbb{B}^n)$ with $\widehat{\mu} = f$ and $\mu_n \xrightarrow{w} \mu$.

2.4 Random Variables and Vectors

Definition 2.4.1. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A measurable mapping $X : (\Omega, \mathcal{A}) \to (\mathbb{R}, \mathbb{B})$ is called random variable (RV).

It induces a push-forward measure P_X on $\mathbb{B}$ via

$$P_X(B) := P\big(X^{-1}(B)\big) \qquad \text{for all } B \in \mathbb{B}. \tag{2.8}$$

Instead of push-forward measure we also say distribution of X.

Definition 2.4.2. Let $(\Omega, \mathcal{A}, P)$ be a probability space. A measurable mapping $X : (\Omega, \mathcal{A}) \to (\mathbb{R}^n, \mathbb{B}^n)$ is called random vector (RVct).

Notation: To express that a RV or a random vector X is distributed according to some distribution $\mathcal{D}$, it is common to write $X \sim \mathcal{D}$.

Definition 2.4.3. (see also [3, Thm. 5.4.4]) Let $(X_i)_{i\in I}$ be a set of RVs on $(\Omega, \mathcal{A}, P)$, where I is an arbitrary index set. For any finite subset $J \subset I$ denote by X_J the random vector whose components are the RVs $X_j,\ j \in J$. The RVs $X_i,\ i \in I$ are called (mutually) independent if

$$P_{X_J} = \otimes_{j\in J} P_{X_j} \qquad \text{for any} \quad J \overset{\text{finite}}{\subset} I. \tag{2.9}$$

According to Definition 2.1.15, condition (2.9) is equivalent to

$$P\big(X_j \in B_j,\ j \in J\big) \ = \prod_{j\in J} P(X_j \in B_j), \qquad B_j \in \mathbb{B} \ \text{ for all } \ j \in J,$$

which illustrates the idea behind Definition 2.4.3: changing a marginal (i.e. determined by only one of the X_j) event $B_k,\ k \in J$, affects the joint probability on the left only through the change of the respective marginal probability.
The generalization of Definition 2.4.3 for random vectors is obvious.

Notation: If the RVs $X_i,\ i \in I$, are independent and have identical marginal distributions, we write

$$(X_i)_{i\in I} \overset{i.i.d.}{\sim} \mu.$$

to specify their (common) marginal distribution μ.

The definition of the push-forward measure reduces everything to the probability space $(\Omega, \mathcal{A}, P)$. In practice however, it is often more natural to specify the distribution on the image space $(\mathbb{R}, \mathbb{B})$, without any reference to the original probability space. This can be conveniently done using the following notions:

Definition 2.4.4. Let $X : (\Omega, \mathcal{A}, P) \longrightarrow (\mathbb{R}^n, \mathbb{B}^n)$ be a RV (n=1) or a RVct (n>1). The distribution function of X is given by

$$F(t) := P\big(X_i \le t_i,\ 1 \le i \le n\big) \qquad t = (t_1, \ldots, t_n)' \in \mathbb{R}^n.$$

The distribution function F uniquely determines the push-forward measure P_X. If P_X is absolutely continuous w.r.t. the Lebesgue measure, it can also be characterized by its probability density:

Definition and Theorem 2.4.5. *([3, Thm. 2.9.10]) Let X be a RV or a RVct. If P_X is absolutely continuous w.r.t. the Lebesgue measure λ^d, i.e.*

$$\lambda^d(B) = 0 \ \Rightarrow\ P_X(B) = 0 \quad \textit{for all } B \in \mathbb{B}^n,$$

then there exists a non-negative, integrable function $f : \mathbb{R}^n \to \mathbb{R}$ so that

$$P_X(B) = \int_B f(x)\,dx \qquad \text{for all } B \in \mathbb{B}^n.$$

f is called the probability density function.

We give some examples of important univariate distributions (i.e. $n = 1$):

Example 2.4.6. The uniform distribution $\mathcal{U}_{[a,b]}$ with parameters $a, b \in \mathbb{R}$, $a < b$, is defined by its probability density function

$$f(x) = \frac{1}{b-a}\,\mathbf{1}_{[a,b]}(x), \qquad x \in \mathbb{R}.$$

Example 2.4.7. The exponential distribution $\operatorname{Exp}(\lambda)$ with parameter $\lambda \in \mathbb{R}_+$, is defined by its probability density function

$$f(x) = \lambda\, e^{-\lambda x}\,\mathbf{1}_{[0,\infty)}(x), \qquad x \in \mathbb{R}.$$

Example 2.4.8. The Gaussian or normal distribution $\mathcal{N}(\mu, \sigma^2)$ with parameters $\mu, \sigma \in \mathbb{R}$, $\sigma > 0$, is defined by its probability density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}}\, e^{-\frac{(x-\mu)^2}{2\,\sigma^2}}, \qquad x \in \mathbb{R}.$$

In the case where $\sigma = 0$, the Gaussian distribution $\mathcal{N}(\mu, 0)$ is no longer absolutely continuous w.r.t. the Lebesgue measure. It is then defined by its distribution function

$$F(x) = \mathbf{1}_{[\mu,\infty)}(x), \qquad x \in \mathbb{R}.$$

The special case $\mathcal{N}(0, 1)$ is called standard Gaussian or standard normal distribution.

The parameters μ and σ of the univariate Gaussian distribution will turn out to be its expectation and variance. The latter are the two most important quantities that can be used to characterize random variables.

Definition 2.4.9. For a RV $X \in L^p(\Omega, \mathcal{A}, P))$, the k-th moment is given by

$$\mathbb{E}\big(X^k\big) := \int_\Omega \big(X(\omega)\big)^k\, P(d\omega), \qquad k \in \mathbb{N},\ k \le p.$$

$\mathbb{E}\big(|X|^k\big)$ is called the k-th absolute moment and, for $k \ge 2$, $\mathbb{E}\big(\big(X - \mathbb{E}(X)\big)^k\big)$ is called the k-th centered moment.

The first moment, $\mathbb{E}(X)$, is called the expectation or the mean of X, the second centered moment is called the variance $\operatorname{Var}(X)$ of X (provided that $p \ge 1$ and $p \ge 2$, respectively).

Remark 2.4.10. The existence of the integrals in Definition 2.4.9 follows from Lemma 2.2.14 (Hölder's inequality), which yields for $k < p$

$$\mathbb{E}\big(|X|^k\big) \;\leq\; \big(\mathbb{E}(|X|^p)\big)^{\frac{k}{p}} \underbrace{\big(\mathbb{E}(1)\big)^{\frac{p-k}{p}}}_{=1} \;<\; \infty.$$

We also briefly note the relation $\operatorname{Var}(X) = \mathbb{E}\big(X^2\big) - \big(\mathbb{E}(X)\big)^2$.

For many distributions the mean, the variance and higher moments can explicitly be calculated. We shall only state those for the normal distribution:

Lemma 2.4.11. *Let $X \sim \mathcal{N}(\mu, \sigma^2)$. Then for any $n \in \mathbb{N}$ we have*

$$\mathbb{E}(X) = \mu, \quad \mathbb{E}\big(|X - \mathbb{E}(X)|^{2n-1}\big) = 0, \quad \text{and} \quad \mathbb{E}\big(|X - \mathbb{E}(X)|^{2n}\big) = \tfrac{(2n)!}{2^n n!}\,\sigma^{2n}.$$

In particular, all centered moments exist and are determined by σ^2.

We note the following inequality (see [5, (21.12)]) that bounds the probability of a deviation from 0 in terms of the absolute moments:

Lemma 2.4.12. *(Markov's inequality) For a RV $X \in L^p(\Omega, \mathcal{A}, P))$ it holds that*

$$P\big(|X| > \epsilon\big) \;\leq\; \frac{1}{\epsilon^p} \int_{\{\omega\,:\,|X(\omega)|>\epsilon\}} \big|X(\omega)\big|^p \, P(d\omega) \;\leq\; \frac{1}{\epsilon^p}\,\mathbb{E}\big(|X|^p\big).$$

The special case $P\big(|X - \mathbb{E}(X)| > \epsilon\big) \;\leq\; \frac{1}{\epsilon^2}\operatorname{Var}(X)$ is usually referred to as Chebyshevs's inequality.

A certain subset of random variables, namely those with existing second moment, are of particular interest:

Definition 2.4.13. Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X, Y \in L^2(\Omega, \mathcal{A}, P)$ second-order RVs. The (centered) covariance of X and Y is

$$\operatorname{Cov}(X, Y) := \mathbb{E}\Big(\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big).$$

The RVs X and Y are called uncorrelated if $\operatorname{Cov}(X, Y) = 0$.

Lemma 2.4.14. *(cf. [5, Sec. 21]) Let $X, Y \in L^1(\Omega, \mathcal{A}, P)$ be independent RVs. Then $\mathbb{E}(XY)$ exists and*

$$\mathbb{E}(XY) = \mathbb{E}(X)\,\mathbb{E}(Y).$$

In particular, if $X, Y \in L^2(\Omega, \mathcal{A}, P)$ are independent, then they are uncorrelated.

Using the relation $\operatorname{Var}(X + Y) = \operatorname{Var}(X) + 2\operatorname{Cov}(X, Y) + \operatorname{Var}(Y)$ we obtain

Corollary 2.4.15. *For two independent RVs $X, Y \in L^2(\Omega, \mathcal{A}, P)$ we have*

$$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y).$$

Moments of random vectors are defined by applying the above notions to their components. We are particularly interested in the first two moments:

Definition 2.4.16. Let X be a RVct whose components $X_i,\ i = 1, \ldots, n$, are second-order RVs. Then the vector

$$\mathbb{E}(X) := \big(\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n)\big)'$$

is called expectation or the mean of X and the matrix

$$\mathrm{Cov}(X) := \big(\mathrm{Cov}(X_i, X_j)\big)_{i,j=1,\ldots,n}$$

is called (variance-)covariance matrix of X.

We briefly note that for any second-order random vector X, any vector $b \in \mathbb{R}^n$ and any matrix $A \in \mathbb{R}^{m \times n}$ we have

1. $\mathbb{E}\big(AX + b\big) = A\,\mathbb{E}(X) + b$
2. $\mathrm{Cov}(AX + b) = A\,\mathrm{Cov}(X)\,A'$

Following [5, Sec. 29] we can now generalize the Gaussian distribution to the multivariate case:

Definition 2.4.17. Let X be a random vector where $(X_i)_{i=1,\ldots,n} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let $\mu \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. Then the distribution of

$$Y := A\,X + \mu, \tag{2.10}$$

denoted by $\mathcal{N}(\mu, \Sigma)$, where $\Sigma = AA'$, is called n-variate Gaussian or n-variate normal distribution with mean μ and covariance Σ.

Lemma 2.4.18. *Let $Y \sim \mathcal{N}(\mu, \Sigma)$ be a random vector in $\mathbb{R}^n$.*

1. *Y has mean $\mathbb{E}(Y) = \mu$ and covariance $\mathrm{Cov}(Y) = \Sigma$.*
2. *For $Z := T\,Y + b$ with $b \in \mathbb{R}^m$ and $T \in \mathbb{R}^{m \times n}$, we have*

$$Z \sim \mathcal{N}(T\mu + b, T\,\Sigma\,T')\,.$$

3. *If Σ is regular (i.e. Σ has full rank), then P_Y is absolutely continuous w.r.t. λ^n and its probability density function equals*

$$f(x) = \frac{1}{(2\pi)^{\frac{n}{2}}\,|\Sigma|^{\frac{1}{2}}}\;e^{-\frac{1}{2}\,(x-\mu)'\,\Sigma^{-1}(x-\mu)}.$$

4. *The Fourier transform $\widehat{P_Y}$ of P_Y is given by $\widehat{P_Y}(\tau) = e^{i\,\mu'\tau-\frac{1}{2}\,\tau'\Sigma\,\tau}$, in particular (see part 5. in Lemma 2.3.2) $\mathcal{N}(\mu,\Sigma)$ is well defined by (2.10).*

5. *Then components $Y_1,\dots,Y_n$ of Y are stochastically independent if and only if they are pairwise uncorrelated, i.e. if $\Sigma = I_n$.*

Note that the necessity of $\Sigma = I_n$ in part *5.* follows from Lemma 2.4.14. It is one of the remarkable properties of the multivariate Gaussian distribution that this is also sufficient.

Next, we introduce different notions of convergence of a sequence $(X_n)_{n\in\mathbb{N}}$ of random variables or random vectors on a probability space $(\Omega,\mathcal{A},P)$. In the latter case, convergence is with respect to some suitable norm on $\mathbb{R}^d$:

Definition 2.4.19. The sequence $(X_n)_{n\in\mathbb{N}}$ is called almost surely convergent towards X, if

$$P\Big(\limsup_{n\to\infty}|X_n - X| > \epsilon\Big) \;=\; 0 \qquad \text{for all } \epsilon > 0.$$

In this case we write $X_n \xrightarrow{a.s.} X$.

Definition 2.4.20. The sequence $(X_n)_{n\in\mathbb{N}}$ is called stochastically convergent towards X, if

$$\lim_{n\to\infty}\; P\big(|X_n - X| > \epsilon\big) \;=\; 0 \qquad \text{for all } \epsilon > 0.$$

In this case we write $X_n \xrightarrow{sto.} X$.

Definition 2.4.21. Assuming that $X, X_n \in L^2(\Omega,\mathcal{A},P)$ for all $n\in\mathbb{N}$, the sequence $(X_n)_{n\in\mathbb{N}}$ is called convergent in the mean square towards X, if

$$\lim_{n\to\infty}\;\mathbb{E}\big(|X_n - X|^2\big) \;=\; 0.$$

In this case we write $X_n \xrightarrow{m.s.} X$.

If $X_n \xrightarrow{m.s.} X$, then the first and second moments must also converge, since

$$\mathbb{E}\big(|X_n - X|^2\big) \;=\; \mathbb{E}(X_n^2) - 2\,\underbrace{\mathbb{E}(X_n X)}_{\le\sqrt{\mathbb{E}(X_n^2)\,\mathbb{E}(X^2)}} + \mathbb{E}(X^2) \;\ge\; \Big(\sqrt{\mathbb{E}(X_n^2)} - \sqrt{\mathbb{E}(X^2)}\Big)^2$$

and

$$\mathbb{E}\big(|X_n - X|^2\big) \;\ge\; \big(\mathbb{E}(|X_n - X|)\big)^{\frac{1}{2}} \;\ge\; \big|\mathbb{E}(X_n) - \mathbb{E}(X)\big|^{\frac{1}{2}}.$$

The following theorems (collecting results from [3, Sec. 2.11, 7.7]) clarify the relations between the different types of convergence:

Theorem 2.4.22. *For $(X_n)_{n\in\mathbb{N}}$ and X as above, we have the implications*

1. $X_n \xrightarrow{m.s.} X \quad \Longrightarrow \quad X_n \xrightarrow{sto.} X.$
2. $X_n \xrightarrow{a.s.} X \quad \Longrightarrow \quad X_n \xrightarrow{sto.} X.$
3. $X_n \xrightarrow{sto.} X \quad \Longrightarrow \quad P_{X_n} \xrightarrow{w} P_X.$

The converse statements are not true in general, and there is no implication between a.s. and m.s. convergence. For part *2.* of Theorem 2.4.22 however there exists at least some kind of converse statement:

Theorem 2.4.23. *The sequence $(X_n)_{n\in\mathbb{N}}$ converges stochastically towards X if and only if from every subsequence of $(X_n)_{n\in\mathbb{N}}$ we can extract a further subsequence which converges to X a.s.*

2.5 Conditional Expectation

We introduce the notion of conditional expectation of RVs. It can be generalized to RVcts by applying it componentwise.

Theorem 2.5.1. *([3, Thm. 10.1.1]) Let $X \in L_1(\Omega, \mathcal{A}, P)$, $\mathcal{A}' \subset \mathcal{A}$ a sub-σ-algebra on Ω and $P|_{\mathcal{A}'}$ the restriction of P on $\mathcal{A}'$. Then there exists a random variable $\tilde{X} \in L_1(\Omega, \mathcal{A}', P|_{\mathcal{A}'})$ satisfying the condition*

$$\int_{A'} X(\omega)\, P(d\omega) \;=\; \int_{A'} \tilde{X}(\omega)\, P(d\omega) \qquad \text{for all } \; A' \in \mathcal{A}'.$$

$\tilde{X}$ is unique up to $P|_{\mathcal{A}'}$-null sets, is usually denoted by $\mathbb{E}[X|\mathcal{A}']$ and is called conditional expectation of X given $\mathcal{A}'$.

The conditional expectation $\mathbb{E}[X|\mathcal{A}']$ reflects the information about X contained in $\mathcal{A}'$. In practice we are interested in the information about X contained in another RV Y or, more generally, in a set $(Y_i)_{i\in I}$ of RVs on the same probability space $(\Omega, \mathcal{A}, P)$. As noted in subsection 2.1, the sub-σ-algebra $\sigma(Y_i,\ i \in I)$ on Ω generated by the set $(Y_i)_{i\in I}$ reflects its information content, and so we call

$$\mathbb{E}\big[X \,\big|\, Y_i,\ i \in I\big] \;:=\; \mathbb{E}\big[X \,\big|\, \sigma(Y_i,\ i \in I)\big]$$

the conditional expectation of X given $(Y_i)_{i\in I}$.

For $X \in L_2(\Omega, \mathcal{A}, P)$ we can give an equivalent definition of the conditional expectation as an orthogonal projection:

Proposition 2.5.2. *Let $(\Omega, \mathcal{A}, P)$ be a probability space and $X \in L_2(\Omega, \mathcal{A}, P)$, further let $\mathcal{A}' \subset \mathcal{A}$ be a sub-σ-algebra on Ω. Denote by $\Pi_{\mathcal{A}'}$ the orthogonal projection of $L_2(\Omega, \mathcal{A}, P)$ on $L_2(\Omega, \mathcal{A}', P|_{\mathcal{A}'})$. Then*

$$\Pi_{\mathcal{A}'} X \;=\; \mathbb{E}[X|\mathcal{A}'] \quad a.s.$$

The following two properties of conditional expectations emphasize its meaning as projection on some "less informative" σ-algebra [3, Sec. 10.1].

Lemma 2.5.3. *Let $(Y_i)_{i \in I}$ a set of RVs on $(\Omega, \mathcal{A}, P)$, and $X \in L_1(\Omega, \mathcal{A}, P)$.*

1. *If $\sigma(X) \subset \sigma(Y_i,\ i \in I)$, then $\mathbb{E}\big[X \,\big|\, Y_i, i \in I\big] = X \quad P$-a.s.*
2. *If X is independent of $(Y_i)_{i \in I}$, then $\mathbb{E}\big[X \,\big|\, Y_i, i \in I\big] = \mathbb{E}(X) \quad P$-a.s.*

In the first case, $(Y_i)_{i \in I}$ contains exhaustive information about X and so X is projected onto itself, while in the second case of independent RVs, no information about X is contained in $(Y_i)_{i \in I}$ and $\Pi_{\sigma(Y_i,\, i \in I)}$ is simply the projection on the constant RVs.

We note some more properties (see [3, Sec. 10.1] and [5, Sec. 34]), which are more technical, but will be needed in later chapters.

Lemma 2.5.4. *Let $X, X_1, X_2 \in L_1(\Omega, \mathcal{A}, P)$, Y an $\mathcal{A}'/\mathbb{B}$ measurable RV on $(\Omega, \mathcal{A}, P)$ where $\mathcal{A}' \subset \mathcal{A}$ is a sub-σ-algebra, and $a_1, a_2 \in \mathbb{R}$.*

1. $\mathbb{E}\big(\mathbb{E}[X|\mathcal{A}']\big) \;=\; \mathbb{E}(X)$
2. $\mathbb{E}[a_1 X_1 + a_2 X_2|\mathcal{A}'] \;=\; a_1\, \mathbb{E}[X_1|\mathcal{A}'] + a_2\, \mathbb{E}[X_2|\mathcal{A}'] \quad P$*-a.s.*
3. $\mathbb{E}[YX|\mathcal{A}'] \;=\; Y\, \mathbb{E}[X|\mathcal{A}'] \quad P$*-a.s.*

Note that the integrals are with respect to different measures: the outer expectation in part *1.* in Lemma 2.5.4 for instance, is with respect to $P|_{\mathcal{A}'}$ and not with respect to P as usual. Here and in the future, we will suppress this subtle difference in the notation to keep notation simple.

Factorization

Lemma 2.5.5. *([3, Lem. 10.2.1]) Let X be a RVs and Y be an n-dimensional random vector on $(\Omega, \mathcal{A}, P)$. X is $\sigma(Y)/\mathbb{B}$ measurable if and only if there exists a $\mathbb{B}^n/\mathbb{B}$ measurable function*

$$g : \mathbb{R}^n \to \mathbb{R} \quad \textit{so that} \quad X \;= g \circ Y$$

Lemma 2.5.5 allows us to define a $\mathbb{B}/\mathbb{B}$ measurable mapping $y \mapsto \mathbb{E}[X|Y=y]$ with the property

$$\int_{Y^{-1}(B)} X(\omega)\, P(d\omega) \;=\; \int_B \mathbb{E}[X|Y=y]\, P_Y(dy) \qquad \text{for all } \; B \in \mathbb{B}.$$

$\mathbb{E}[X|Y=y]$ is called factorized conditional expectation and assigns to every observed value y the expected value of X given that $Y=y$. It is P_Y-a.s. and inherits all of the properties of the conditional expectation.

Apart from the restriction that $\mathbb{E}[X|Y=y]$ must be $\mathbb{B}^n/\mathbb{B}$ measurable it can be of arbitrary form. It is another remarkable property of the multivariate Gaussian distribution that the conditioning of some of its components on the remaining ones leads to a very simple form:

Proposition 2.5.6. *Let $(X_1, X_2)'$ be a random vector of size n_1+n_2 that is distributed according to a multivariate Gaussian distribution, i.e.*

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix} \right).$$

Then the factorized conditional expectation of X_1 given $X_2 = x_2$ equals

$$\mathbb{E}[X_1|X_2=x_2] \;=\; \mu_1 + \Sigma_{12}\,\Sigma_{22}^{-1}(x_2-\mu_2).$$

Recalling the projection property of $\mathbb{E}[X|Y]$ from Proposition 2.5.2 and that $\mathbb{E}[X|Y=y]$ is the function g such that $\mathbb{E}[X|Y]=g(Y)$, we can interpret the factorized conditional expectation as the best predictor of X given $Y=y$. Proposition 2.5.6 states that in the case of a multivariate Gaussian distribution the best such predictor function g is linear in y, and only depends on the means and the covariances.

2.6 Stochastic Processes

Definition 2.6.1. A (real valued) stochastic process X on a set T is a set $(X_t)_{t\in T}$ of random variables $X_t : \Omega \to \mathbb{R}$ sharing the same probability space $(\Omega, \mathcal{A}, P)$.
When $T \subseteq \mathbb{R}^d$ X is also called a random field (RF).

There are two ways of looking at a stochastic process:

1. For fixed $t \in T$, $X_t : \omega \mapsto X_t(\omega)$ is simply a random variable,
2. For fixed $\omega \in \Omega$, $X_{\bullet}(\omega) : t \mapsto X_t(\omega)$ is a sample path.

Taking the path point of view amounts to interpreting a stochastic process as a map $X : (\Omega, \mathcal{A}, P) \to \left(\mathbb{R}^T, \mathbb{B}^T\right)$ into the space $\mathbb{R}^T$ of real-valued functions on T with product σ-algebra $\mathbb{B}^T$.
If T is a metric or at least a topological space (as e.g. in the case of a RF), one can ask whether the sample paths of $(X)_{t\in T}$ are continuous functions. We postpone this problem to section 5.

Definition 2.6.2. A stochastic process $(X)_{t\in T}$ over the probability space $(\Omega, \mathcal{A}, P)$ is of second order if $X_t \in L^2(\Omega, \mathcal{A}, P)$ for all $t \in T$. Then

$$K(s,t) \;:=\; \operatorname{Cov}(X_s, X_t) \quad \text{for all } s,t \in T$$

is called the covariance function or covariance kernel on $T \times T$ and

$$m(t) \;:=\; \mathbb{E}(X_t) \quad \text{for all } t \in T.$$

is called the mean function. Sometimes one uses the second moment function

$$R(s,t) \;:=\; \mathbb{E}(X_s X_t) \quad \text{for all } s,t \in T$$

and it holds that

$$R(s,t) \;=\; K(s,t) + m(s)\,m(t) \quad \text{for all } s,t \in T. \tag{2.11}$$

The following theorem is a consequence of the properties of the covariance (see Definition 2.4.16 and the subsequent remark):

Theorem 2.6.3. *The covariance and the second moment function of a second order stochastic process $(X)_{t\in T}$ are symmetric (i.e. $R(s,t) = R(t,s)$ and $K(s,t) = K(t,s)$ for all $s,t \in T$) and positive semidefinite (see Definition 3.1.1) functions on $T \times T$.*

The converse, i.e. the existence of a second order stochastic process on T for any positive semidefinite function on $T \times T$, is also true as is shown later in Corollary 5.1.2.

Definition 2.6.4. A stochastic process $(X)_{t \in T}$ is called a Gaussian process, if for any finite subset $\{t_1, \dots, t_n\} \subset T$ the random vector $(X_{t_1}, \dots, X_{t_n})'$ is n-variate normally distributed.

Let us now turn to the special case of RFs. This is the case we will deal with in all subsequent chapters. We shall introduce some important subclasses:

Definition 2.6.5. A random field $(X)_{t \in T}$ is called (strictly) stationary if for any finite subset $\{t_1, \dots, t_n\} \subset T$ and any $h \in \mathbb{R}^d$ with $\{t_1 + h, \dots, t_n + h\} \subset T$ it holds that

$$P_{(X_{t_1}, \dots, X_{t_n})'} \; = \; P_{(X_{t_1+h}, \dots, X_{t_n+h})'} \, ,$$

i.e. if the finite dimensional marginal distributions are shift-invariant.

Definition 2.6.6. A second-order random field $(X)_{t \in T}$ is called (weakly) stationary if the mean function $m(\cdot)$ is constant and the covariance function $K(s,t)$ depends on s and t only via $t - s$, i.e. if

$$K(s,t) \; = \; \Phi(t-s) \qquad \text{for all } s, t \in T.$$

for some function $\Phi : T \to \mathbb{R}$ (which we will call covariance function as well).

From Lemma 2.4.18 we can see that the multivariate Gaussian distribution is completely determined by its mean and its covariance. Hence, if a random field $(X)_{t \in T}$ is Gaussian and weakly stationary, it is also strictly stationary, and we will no longer distinguish between the two notions in this case.

Chapter 3

Hilbert Spaces in Approximation Theory and Stochastics

3.1 Reproducing-Kernel Hilbert Spaces

In this section, we shall introduce the notion of a reproducing-kernel Hilbert space (RKHS), one of the basic notions to describe the classes of functions that are dealt with in approximation theory. To this end, we study continuous functions $R : T \times T \to \mathbb{R}$ (called kernels in the following), where $T \subseteq \mathbb{R}^d$ is an arbitrary region which contains at least one point.
Requiring R to be continuous is not always necessary, but allowing for discontinuous kernels would complicate many of our considerations, so we shall stick to continuity as one of our working assumptions in this and all subsequent chapters.

Another important property we need to impose on kernels is the following:

Definition 3.1.1. A continuous kernel $R : T \times T \to \mathbb{R}$ is called positive semidefinite on $T \subset \mathbb{R}^d$ if for all $n \in \mathbb{N}$, all pairwise distinct $\{t_1, \dots, t_n\} \subset T$, and all $a \in \mathbb{R}^n \setminus \{0\}$ we have

$$\sum_{j=1}^{n} \sum_{k=1}^{n} a_j \, a_k \, R(t_j, t_k) \;\geq\; 0. \tag{3.1}$$

If the sum in (3.1) is strictly > 0, then R is called positive definite on $T \subset \mathbb{R}^d$.

Now for a positive definite kernel R on $T \subset \mathbb{R}$ with $R(s,t) = R(t,s)$ define

$$H_R := \left\{ \sum_{i=1}^{m} a_i \, R(t_i, \cdot) \;:\; a_i \in \mathbb{R},\; t_i \in T,\; m \in \mathbb{N} \right\}. \tag{3.2}$$

with inner product

$$\left(\sum_{i=1}^{m} a_i \, R(s_i, \cdot), \sum_{j=1}^{n} b_j \, R(t_j, \cdot) \right)_{\mathcal{H}_R} \;:=\; \sum_{i=1}^{m} \sum_{j=1}^{n} a_i \, b_j \, R(s_i, t_j) \tag{3.3}$$

By the positive definiteness of R we have $(f,f)_{\mathcal{H}_R} \geq 0$ for all $f \in H_R$, and $(f,f)_{\mathcal{H}_R} = 0$ if and only if $f \equiv 0$, so the inner product (3.3) defines a norm $\|f\|_{\mathcal{H}_R} = (f,f)_{\mathcal{H}_R}^{1/2}$ on H_R. Furthermore, for any $f \in H_R$, we have

$$\big(f, R(t,\cdot)\big)_{\mathcal{H}_R} = \left(\sum_{i=1}^{m} a_i\, R(s_i,\cdot), R(t,\cdot)\right)_{\mathcal{H}_R} = \sum_{i=1}^{m} a_i\, R(s_i,t) = f(t) \tag{3.4}$$

This is the reproducing kernel property.

The closure of H_R under $\|\cdot\|_{\mathcal{H}_R}$ is a space of real-valued functions, denoted by $\mathcal{H}_R$, and called the reproducing kernel Hilbert space (RKHS) of R. By the continuity of the inner product, the reproducing equation (3.4) carries over to $\mathcal{H}_R$.

From (3.4) and from the continuity of R it follows that any $f \in \mathcal{H}_R$ is continuous since we have

$$|f(t) - f(s)| = |(f, R(t,\cdot) - R(s,\cdot))_{\mathcal{H}_R}| \leq \|f\|_{\mathcal{H}_R} \cdot \|R(t,\cdot) - R(s,\cdot)\|_{\mathcal{H}_R}$$

and

$$\|R(t,\cdot) - R(s,\cdot)\|_{\mathcal{H}_R}^2 = R(t,t) + R(s,s) - 2\,R(s,t).$$

3.2 Sobolev Spaces

Following [12, Sec. 5.2] we introduce an important class of RKHSs, the Sobolev spaces. Each one of these spaces guarantees a certain smoothness of the functions it contains. They will turn out to be the natural function spaces for the sample paths of second order random fields (see Section 5.5).

Notation: Here and in all following chapters an open subset $T \subseteq \mathbb{R}^d$ is called a domain. Its boundary is denoted by ∂T, its closure by $\overline{T}$.

For an (arbitrary) domain $T \subseteq \mathbb{R}^d$ we denote by $\mathcal{C}(T)$ the space of continuous (real valued) functions, by $\mathcal{C}^k(T)$ the space of k times continuously differentiable functions and by $\mathcal{C}^\infty(T)$ the space of infinitely differentiable functions $f : T \to \mathbb{R}$.

Further we denote by $\mathcal{C}_c(T)$, $\mathcal{C}_c^k(T)$ and $\mathcal{C}_c^\infty(T)$ respectively the corresponding classes of function which in addition have compact support in T, and by $\mathcal{C}(\overline{T})$, $\mathcal{C}^k(\overline{T})$ and $\mathcal{C}^\infty(\overline{T})$ respectively the classes of functions whose partial derivatives up to order $0, k$ or ∞ respectively can be extended continuously to $\overline{T}$.

Finally, for $f \in \mathcal{C}^k(T)$ and a multi-index $\alpha \in \mathbb{N}_0^d$ of order $|\alpha| \leq k$, $|\alpha| := \sum_{i=1}^d \alpha_i$, let

$$D^\alpha f = \frac{\partial^{|\alpha|} f}{\partial e_1^{\alpha_1} \cdots \partial e_d{}^{\alpha_d}},$$

where e_i is the unit vector in $\mathbb{R}^d$ in the direction of the i^{th} coordinate axis.

The notion of smoothness that comes with Sobolev spaces is a weakening of the notion of partial derivatives (cf. [12, Sec. 5.2.1]).

Definition and Lemma 3.2.1. *Suppose that $f, g \in L^1_{\text{loc}}(T)$, and $\alpha \in \mathbb{N}^d$ is a multi-index. We say that g is the α^{th} weak partial derivative of f, written*

$$D^\alpha f \;=\; g,$$

provided

$$\int_T f(x)\, D^\alpha \varphi(x)\, dx \;=\; (-1)^{|\alpha|} \int_T g(x)\, \varphi(x)\, dx$$

for all functions $\varphi \in \mathcal{C}^\infty_c(T)$ (so called test functions). A weak α^{th} partial derivative of f, if it exists, is uniquely defined up to a set of measure 0.

If $f \in \mathcal{C}^k(T)$, then for any α with $|\alpha| \le k$ an α^{th} weak partial derivative of f exists and coincides (up to a set of measure 0) with the ordinary α^{th} partial derivative. A simple example of a function that is differentiable in the weak, but not in the ordinary sense, is the function

$$f : \mathbb{R} \to \mathbb{R}, \quad x \mapsto |x|$$

with weak derivative

$$D^1 f(x) \;=\; \begin{cases} -1, & x \le 0 \\ 1, & x > 0 \end{cases}$$

Note that the notion of weak differentiability is always a global one, there is no weak counterpart for differentiability of f at a certain point $t \in T$.

We can now introduce a new class of function spaces, whose members have weak derivatives of order $k \in \mathbb{N}_0$ lying in some L^p space:

Definition 3.2.2. Let T be a domain in $\mathbb{R}^d$ and $1 \le p \le \infty$. The Sobolev space $W^{k,p}(T)$ consists of all locally integrable functions $f : T \to \mathbb{R}$ such that for each multi-index α with $|\alpha| \le k$, $D^\alpha f$ exists in the weak sense and belongs to $L^p(T)$. If it only belongs to $L^p_{\text{loc}}(T)$, we obtain the local Sobolev space $W^{k,p}_{\text{loc}}(T)$.

For $f \in W^{k,p}(T)$ we define its norm to be

$$\|f\|_{W^{k,p}(T)} \;:=\; \begin{cases} \left(\sum_{|\alpha| \le k} \|D^\alpha f\|^p_{L^p(T)} \right)^{1/p}, & 1 \le p < \infty \\ \sum_{|\alpha| \le k} \operatorname{ess\,sup}_T \, |D^\alpha f| \,, & p = \infty. \end{cases}$$

We note some elementary properties of weak derivatives ([12, Sec. 5.2.3]):

Theorem 3.2.3. *Assume $f, g \in W^{k,p}(T)$, $|\alpha| \le k$. Then*

1. *$D^\alpha f \in W^{k-|\alpha|,p}(T)$ and $D^\alpha(D^\beta f) = D^\beta(D^\alpha f) = D^{\alpha+\beta} f$ for all multi-indices α, β with $|\alpha| + |\beta| \le k$.*

2. *For each $a, b \in \mathbb{R}$, $a\, f + b\, g \in W^{k,p}(T)$ and $D^\alpha(a\, f + b\, g) = a\, D^\alpha f + b\, D^\alpha g$, $|\alpha| \le k$.*

3. *If I is an open subset of T, then $f \in W^{k,p}(I)$.*

Like the L^p-spaces, Sobolev spaces have a good mathematical structure:

Theorem 3.2.4. *For each $k \in \mathbb{N}$ and $1 \le p \le \infty$, the Sobolev space $W^{k,p}(T)$ is a Banach space. The special case $W^{k,2}(T)$ is a Hilbert space.*

Next, give a characterization of the class of weakly differentiable functions f as functions which are absolutely continuous on a.e. line parallel to the coordinate axes (taken from [24, Sec. 5.6]). This will be useful in the study of regularity properties of stochastic processes in Chapter 5. First we recall the definition of absolute continuity in the one-dimensional case:

Definition 3.2.5. A function $f : I \to \mathbb{R}$ where either $I = \mathbb{R}$ or $I = [a, b]$, $a, b \in \mathbb{R}$, is said to be absolutely continuous (on I) if for every $\epsilon > 0$, there exists a $\delta > 0$ such that

$$\sum_{i=1}^{n} (\beta_i - \alpha_i) \;<\; \delta \quad\Longrightarrow\quad \sum_{i=1}^{n} \big|f(\beta_i) - f(\alpha_i)\big| \;<\; \epsilon,$$

whenever $(\alpha_1, \beta_1), \ldots, (\alpha_n, \beta_n)$ are disjoint subintervals of I.

Theorem 3.2.6. *([32, Thm. 8.17, 8.18]) Let $g \in L^1(I)$ with I as above. For some $x_0 \in I$ define the function f on I by*

$$f(x) \;:=\; \int_{x_0}^{x} g(t)\, dt.$$

Then f is absolutely continuous and $f' = g$ a.e. on I ($f' = \frac{df}{dx}$ in the ordinary sense).

Conversely, if $f : I \to \mathbb{R}$ is absolutely continuous, then f is differentiable (in the ordinary sense) a.e. on I, $f' \in L^1(I)$, and it holds that

$$f(x) \;:=\; f(x_0) \;+\; \int_{x_0}^{x} f'(t)\, dt \qquad \text{for all } x \in I.$$

The notion of absolute continuity of functions f on $[a, b] \subset \mathbb{R}$ is generalized to open and connected subsets T of $\mathbb{R}^d$ (domains) by considering the restrictions of f to all straight lines parallel to the coordinate axes that intersect T. More precisely, let

$$\pi_i : \mathbb{R}^d \to \mathbb{R} \qquad \text{and} \qquad \overline{\pi}_i : \mathbb{R}^d \to \mathbb{R}^{d-1}$$

denote the projections of some point in $\mathbb{R}^d$ on the i^{th} coordinate and on all other coordinates respectively. For some set $B_1 \in \mathbb{B}^{d-1}$ and some set $B_2 \in \mathbb{B}$ denote by

$$B_1 \times_i B_2 \;:=\; \big\{t \in \mathbb{R}^d \,:\, \overline{\pi}_i(t) \in B_1,\;\; \pi_i(t) \in B_2\big\}$$

the Cartesian product of B_1 and B_2 that is taken in the i^{th} component.

Now fix $i \in \{1, \ldots, d\}$. For some $\bar{t} \in \overline{\pi}_i(T)$ consider the line $l_i(\bar{t}) := \bar{t} \times_i \mathbb{R}$. For every such line there exists a (finite or infinite) sequence $(J_n)_{n \in \mathcal{I}(\bar{t})}$, $\mathcal{I}(\bar{t}) \subseteq \mathbb{N}$ of disjoint open intervals $J_n = J_n(\bar{t}) \subset \mathbb{R}$ such that

$$l_i(\bar{t}) \cap T \;=\; \bigcup_{n \in \mathcal{I}(\bar{t})} \bar{t} \times_i J_n(\bar{t}).$$

Definition 3.2.7. Let T be a domain in $\mathbb{R}^d$. A real valued function f defined on T is said to be absolutely continuous on the line $l_i(\bar{t})$ if the function

$$f_{\bar{t}}(\xi) \;:=\; f\big((\bar{t}_1, \ldots, \bar{t}_{i-1}, \xi, \bar{t}_i, \ldots, \bar{t}_{d-1})\big), \qquad \xi \in \bigcup_{n \in \mathcal{I}(\bar{t})} J_n(\bar{t})$$

is absolutely continuous on every compact subinterval of $J_n(\bar{t})$ for any $n \in \mathcal{I}(\bar{t})$.

Definition 3.2.8. For $1 \le i \le d$ and T as above the space $\mathrm{AC}_i(T)$ consists of all functions f that are absolutely continuous on $l_i(\bar{t})$ for almost every $\bar{t} \in \overline{\pi}_i(T)$. Further we set

$$\mathrm{AC}(T) \;=\; \bigcap_{i=1}^{d} \mathrm{AC}_i(T)$$

We note the following relation between $\mathrm{AC}(T)$ and $W^{1,1}_{\mathrm{loc}}(T)$ (see [24, Lem. 5.6.2 and Thm. 5.6.3] or [28, Sec. 1.1.3]):

Lemma 3.2.9. *Let $f \in \mathrm{AC}_i(T) \cap L^1_{\mathrm{loc}}(T)$, and denote by $\frac{\partial f}{\partial e_i}$ the ordinary partial derivative of f in the direction e_i (it exists a.e. in T since f is absolutely continuous on a.e. line l_i). If $\frac{\partial f}{\partial e_i} \in L^1_{\mathrm{loc}}(T)$, then it is a weak partial derivative of f.*

Theorem 3.2.10. *Let $f \in W^{1,1}_{\mathrm{loc}}(T)$ and $D^{e_i} f$ a weak derivative of f in the direction e_i. Then there exists a function $g \in \mathrm{AC}_i(T)$ which is equal to f a.e. on T and whose ordinary partial derivative $\frac{\partial g}{\partial e_i}$ is equal to $D^{e_i} f$ a.e. on T.*

The notion of Sobolev spaces can be extended to non-integer orders ([24, Sec. 6.8]) which altogether yields a class of function spaces with continuously parametrized degree of smoothness.

For $1 \le p < \infty$, $\mu \in \mathbb{R}_+ \backslash \mathbb{N}$ and $k := \lfloor \mu \rfloor$ (the biggest integer $\le \mu$) define

$$|f|_{W^{\mu,p}(T)} \;:=\; \left(\sum_{|\alpha|=k} \int_T \int_T \frac{|D^\alpha f(x) - D^\alpha f(y)|^p}{\|x-y\|^{d+p(\mu-k)}} \, dx\, dy \right)^{1/p}$$

$$\|f\|_{W^{\mu,p}(T)} \;:=\; \left(\|f\|^p_{W^{k,p}(T)} + |f|^p_{W^{\mu,p}(T)} \right)^{1/p}$$

Definition and Theorem 3.2.11. *Let $T \subseteq \mathbb{R}^d$ and $1 \le p < \infty$. For $\mu \in \mathbb{R}_+ \backslash \mathbb{N}$ the Sobolev space $W^{\mu,p}(T)$, defined by*

$$W^{\mu,p}(T) \;:=\; \left\{ f \in W^{k,p}(T) \;:\; \|f\|_{W^{\mu,p}(T)} < \infty \right\}$$

is a Banach space. The corresponding local Sobolev space $W^{\mu,p}_{\mathrm{loc}}(T)$ is defined by

$$W^{\mu,p}_{\mathrm{loc}}(T) \;:=\; \left\{ f \in W^{k,p}_{\mathrm{loc}}(T) \;:\; \|f\|_{W^{\mu,p}(V)} < \infty \;\; \textit{for all} \;\; V \subset\subset T \right\}.$$

Now that we have defined the full scale of Sobolev spaces we shall outline their connection to the general idea of RKHS introduced in Section 3.1. To this end we introduce a particular class of radially symmetric kernels (i.e. $R(s,t) = \Phi(\|t-s\|)$) that will turn out to be the reproducing kernels for the Sobolev spaces.

Definition 3.2.12. For $\tau > \frac{d}{2}$ the Whittle-Matérn kernel is given by

$$\Phi_\tau(h) \;:=\; \frac{(2\pi)^{\frac{d}{2}}\,\|h\|^{\tau-\frac{d}{2}}}{2^{\tau-1}\,\Gamma(\tau)}\;\mathcal{K}_{\tau-\frac{d}{2}}(\|h\|)\,, \tag{3.5}$$

where $\mathcal{K}_\tau$ is the modified Bessel function of the third kind. It is the Fourier transform (see Definition 2.3.1) of a measure on $(\mathbb{R}^d, \mathbb{B}^d)$ that is absolutely continuous w.r.t. λ^d with density

$$\varphi_\tau(\omega) \;=\; \left(1+\|\omega\|^2\right)^{-\tau}. \tag{3.6}$$

In the Numerical Analysis literature this kernel is also called Sobolev kernel. Its great use in both Numerical Analysis and Spatial statistics (as a covariance function, see Chapter 5) is due to its property to quantify the smoothness of the associated RKHS (see next Theorem) and the associated random field (see Sections 5.3 and 5.5) respectively.

Theorem 3.2.13. *([41, Cor. 10.13]) Suppose that $\Phi \in L_1(\mathbb{R}^d) \cap \mathcal{C}(\mathbb{R}^d)$ satisfies*

$$c_1\left(1+\|\omega\|^2\right)^{-\tau} \;\le\; \widehat{\Phi}(\omega) \;\le\; c_2\left(1+\|\omega\|^2\right)^{-\tau}, \qquad \omega\in\mathbb{R}^d$$

with $\tau > \frac{d}{2}$ and two positive constants $c_1 \le c_2$. Then the RKHS $\mathcal{H}_\Phi$ coincides with the Sobolev space $W^{\tau,2}(\mathbb{R}^d)$, and the norms $\|\cdot\|_{\mathcal{H}_\Phi}$ and $\|\cdot\|_{W^{\tau,2}(\mathbb{R}^d)}$ are equivalent.

Note that the kernel Φ_τ itself (and hence any finite linear combinations of kernel translates) is contained in the Sobolev space $W^{\mu,2}(\mathbb{R}^d)$ if and only if $\mu < 2\tau - \frac{d}{2}$ which directly follows from

$$\left(1+\|x\|^2\right)^{-s} \in L^1(\mathbb{R}^d) \quad\Longleftrightarrow\quad s > \tfrac{d}{2},$$

and from the alternative characterization of $W^{\mu,2}(\mathbb{R}^d)$ as (cf. [41, p. 141]

$$W^{\mu,2}(\mathbb{R}^d) \;=\; \left\{f\in L^2(\mathbb{R}^d)\;:\;\widehat{f}(\cdot)\,(1+\|\cdot\|^2)^{\mu/2}\in L_2(\mathbb{R}^d)\right\}.$$

We have already noted that Sobolev spaces are characterized by the degree of smoothness of the functions they contain, where smoothness (differentiability) was always in the weak sense. The next two theorems (see [10, Ch. 3.3, 3.4 and 4.2]) are just two of many imbedding theorems for Sobolev spaces and provide a link to the "classical" notion of smoothness.

Definition 3.2.14. A domain $T \subset \mathbb{R}^d$ is called a bounded $\mathcal{C}^k$ domain in $\mathbb{R}^d$ if it is bounded, connected, and if the boundary ∂T can be covered by finitely many open balls $B_j \subset \mathbb{R}^d$, $j = 1,\ldots,m$, centred at ∂T such that - upon relabeling and reorienting the coordinate axes if necessary - we have

$$T\cap B_j \;=\; \left\{x\in B_j\;:\;x_d > \gamma_j(x_1,\ldots,x_{d-1})\right\}$$

with functions $\gamma_j \in \mathcal{C}^k(\mathbb{R}^{d-1})$. For $d = 1$ it simply means open bounded interval.

If T is a bounded $\mathcal{C}^k$ domain for every $k \in \mathbb{N}$, it is called a bounded $\mathcal{C}^\infty$ domain.

Theorem 3.2.15. *Let T be a bounded $\mathcal{C}^\infty$ domain in R^d. Then, for $\mu > k + \frac{d}{2}$ we have the implication*

$$f \in W^{\mu,2}(T) \quad \Longrightarrow \quad \exists\, \tilde{f} \in \mathcal{C}^k(\overline{T}) \;\; \textit{so that} \;\; \tilde{f} = f \quad \textit{a.e. on } T.$$

This result can be generalized by introducing a class of functions that continuously parametrizes the degree of smoothness in the classical sense.
For a bounded and continuous function $f : T \to \mathbb{R}$ on a domain $T \subseteq \mathbb{R}^d$ we write

$$\|f\|_{\mathcal{C}(\overline{T})} \;=\; \sup_{t \in T} |f(t)|.$$

Moreover, for $0 < \beta \le 1$ we define the β^{th} Hölder seminorm of $f : T \to \mathbb{R}$ by

$$|f|_{\mathcal{C}^{0,\beta}(\overline{T})} \;=\; \sup_{\substack{s,t \in T \\ s \neq t}} \frac{|f(t) - f(s)|}{\|t - s\|^\beta}\,.$$

Definition 3.2.16. The Hölder space $\mathcal{C}^{k,\beta}(\overline{T})$, $k \in \mathbb{N}_0$, $0 < \beta \le 1$ consists of all functions $f \in \mathcal{C}^k(\overline{T})$ for which the norm

$$\|f\|_{\mathcal{C}^{k,\beta}(\overline{T})} \;:=\; \sum_{|\alpha| \le k} \|D^\alpha f\|_{\mathcal{C}(\overline{T})} \;+\; \sum_{|\alpha| = k} |D^\alpha f|_{\mathcal{C}^{0,\beta}(\overline{T})}$$

is finite. We define the local Hölder space $\mathcal{C}^{k,\beta}_{\text{loc}}(T)$ by

$$\mathcal{C}^{k,\beta}_{\text{loc}}(T) \;:=\; \Big\{ f \in \mathcal{C}^k(T) \;:\; \|f\|_{\mathcal{C}^{k,\beta}(\overline{V})} < \infty \;\; \text{for all} \;\; V \subset\subset T \Big\}.$$

Theorem 3.2.17. *Let T be a bounded $\mathcal{C}^\infty$ domain in R^d, further let $k \in \mathbb{N}_0$ and $0 < \beta < 1$. Then, for $\mu > k + \beta + \frac{d}{2}$ we have the implication*

$$f \in W^{\mu,2}(T) \quad \Longrightarrow \quad \exists\, \tilde{f} \in \mathcal{C}^{k,\beta}(\overline{T}) \;\; \textit{so that} \;\; \tilde{f} = f \quad \textit{a.e. on } T.$$

For later use we finally prove the following

Lemma 3.2.18. *For the Whittle-Matérn kernel Φ_τ from Definition 3.2.12 it holds that*

$$\Phi_\tau \in \mathcal{C}^{k,\beta}_{\text{loc}}(\mathbb{R}^d) \quad \Longrightarrow \quad 2\tau \ge k + \beta + d, \qquad k = 0, 1, \quad 0 < \beta \le 1.$$

For $k = 1$ and $\beta = 1$ we even have the strict inequality

$$\Phi_\tau \in \mathcal{C}^{1,1}_{\text{loc}}(\mathbb{R}^d) \quad \Longrightarrow \quad 2\tau > 2 + d.$$

Proof: Using the expansion

$$\begin{aligned} \Phi_\tau(h) &= a_{0,\tau} + O\big(\|h\|^{2\tau-d}\big), && \tfrac{d}{2} < \tau < 1 + \tfrac{d}{2} \\ \Phi_\tau(h) &= a_{0,\tau} + O\big(\|h\|^{2\tau-d}\big|\log\|h\|\big|\big), && \tau = 1 + \tfrac{d}{2} \end{aligned} \tag{3.7}$$

(stated in [39, p. 31] with a different parametrization) around the origin we see that

$$\lim_{h\to 0} \frac{|\Phi_\tau(0) - \Phi_\tau(h)|}{\|h\|^\beta} < \infty \quad \Longleftrightarrow \quad 2\tau \geq \beta + d$$

which shows the first implication for the case $k = 0$. For $k = 1$, we calculate the Lagrange form of the Taylor expansion (order 0) of Φ_τ at the origin. Due to the radial symmetry of Φ_τ, we may w.l.o.g. assume that $h = a \cdot e_1$, $a > 0$. Then there exists some $0 \leq \xi \leq a$ so that

$$\big|\Phi_\tau(0) - \Phi_\tau(ae_1)\big| = \left|\frac{\partial\Phi_\tau}{\partial e_1}(\xi e_i) \cdot \|ae_1\|\right| = \underbrace{\left|\frac{\partial\Phi_\tau}{\partial e_1}(\xi e_1) - \frac{\partial\Phi_\tau}{\partial e_1}(0)\right|}_{\leq C\,\xi^\beta} \cdot a \leq C\,a^{1+\beta}$$

and so we must have

$$\Phi_\tau(h) = \Phi_\tau(0) + O\big(\|h\|^{1+\beta}\big) \quad \text{or} \quad \Phi_\tau(h) = \Phi_\tau(0) + o\big(\|h\|^{1+\beta}\big). \tag{3.8}$$

For $0 < \beta < 1$ we are still in the first case of (3.7) and we conclude $2\tau \geq 1 + \beta + d$. If $\beta = 1$, the second expansion is relevant and for (3.8) to hold we must have $2\tau > 2 + d$ as asserted.

□

3.3 Canonical Isomorphism

Following [4, p. 61-65], we now go back to Definition 2.4.13 in Section 2.4 and consider the space $L^2(\Omega, \mathcal{A}, P)$ of all second-order random variables on the probability space $(\Omega, \mathcal{A}, P)$. Furthermore, we assume that $(X_t)_{t\in T}$ is a random field over that probability space, and we have the Hilbert space $\mathcal{H}_R$ in which the second moment kernel $R(s,t) = \mathbb{E}(X_sX_t)$ of $(X_t)_{t\in T}$ is reproducing. Throughout this section, R is considered to be fixed.

Consider the space S_X of all linear combinations of random variables from our stochastic process $(X_t)_{t\in T}$, i.e.

$$S_X := \left\{ \sum_{j=1}^n a_j X_{t_j} \;:\; a_j \in \mathbb{R},\ t_j \in T,\ n \in \mathbb{N} \right\}.$$

This clearly is a subspace of $L^2(\Omega, \mathcal{A}, P)$, and we know the L^2 inner product

$$\langle X_t, X_s \rangle = \mathbb{E}(X_tX_s) = R(t,s) \quad \text{for all } s,t \in T$$

on its generators. We can map the space S_X to $\mathcal{H}_R$ by the map

$$\Psi_X\left(\sum_{j=1}^n a_j X_{t_j}\right) := \sum_{j=1}^n a_j\, R(t_j, \cdot), \tag{3.9}$$

in particular we have

$$\Psi_X(X_t) \;=\; R(t,\cdot) \quad \text{for all } t \in T.$$

and we will always write Ψ_X to stress the dependence of Ψ on X.

We still need to prove that the map is well-defined. Assume that the zero random variable $Z \equiv 0 \in S_X$ has a nontrivial representation

$$Z \;=\; \sum_{j=1}^{n} a_j X_{t_j}\,.$$

Then we have to prove that the function

$$\Psi_X(Z) \;=\; \Psi_X\left(\sum_{j=1}^{n} a_j X_{t_j}\right) \;=\; \sum_{j=1}^{n} a_j\, R(t_j,\cdot)$$

vanishes everywhere. We check this via

$$\begin{aligned}\Psi_X(Z)(t) \;&=\; \sum_{j=1}^{n} a_j R(t_j,t) \;=\; \sum_{j=1}^{n} a_j\, \mathbb{E}(X_{t_j}X_t)\\ &=\; \mathbb{E}\left(\sum_{j=1}^{n} a_j X_{t_j} X_t\right) \;=\; \mathbb{E}(ZX_t) \;=\; 0 \quad \text{for all } t \in T.\end{aligned}$$

Thus Ψ_X is an isometry between S_X and $\Psi_X(S_X) \subset \mathcal{H}_R$, and it extends continuously to the Hilbert space closures. This is why we call Ψ_X in (3.9) the canonical isomorphism. We define $\mathcal{S}_X$ to be the Hilbert space closure of S_X under the L^2 inner product $\langle .,. \rangle$, and we know that the $\mathcal{H}_R$ closure of $\Psi_X(S_X)$ is all of $\mathcal{H}_R$ (cf. [41, Ch. 10]). Thus the closure $\mathcal{S}_X$ of S_X under $\langle .,. \rangle$ is isometrically isomorphic to $\mathcal{H}$, and it still is a closed subspace of $L^2(\Omega, \mathcal{A}, P)$.

Summarizing, for each second-order random field $(X_t)_{t\in T}$, the canonical isomorphism induces two isometric Hilbert spaces: a space $\mathcal{H}_R$ of functions on T and a subspace $\mathcal{S}_X$ of random variables in $L^2(\Omega, \mathcal{A}, P)$. The function space $\mathcal{H}_R$ is only dependent on the kernel R, while the space $\mathcal{S}_X$ of random variables still depends on the particular process $(X_t)_{t\in T}$, its Hilbert space structure, however, being only dependent on R.

Finally we note that if $(X_t)_{t\in T}$ is Gaussian, then any random vector Z with components $Z_1, \ldots, Z_m \in \mathcal{S}_X$ follows an m-variate Gaussian distribution. For $Z_j \in S_X$, $j = 1, \ldots, m$, this follows from part *2.* in Lemma 2.4.18, because in this case Z is a linear transformation of some random vector $(X_{t_1}, \ldots, X_{t_k})'$ which is multivariate Gaussian by definition. Now if Z is the L^2-limit of random vectors $Z^{(n)} \sim \mathcal{N}(\mu_n, \Sigma_n)$, we have (Theorem 2.4.22)

$$\mu_n \to \mu,\;\; \Sigma_n \to \Sigma \quad \text{and} \;\; P_{Z^{(n)}} \xrightarrow{w} P_Z \qquad \text{as } n \to \infty.$$

Using Lemma 2.3.3 and part *4.* in Lemma 2.4.18 we obtain $Z \sim \mathcal{N}(\mu, \Sigma)$, in particular the limit distribution is m-variate Gaussian as well.

Chapter 4

Expansions

4.1 Mercer Eigenfunction Expansions

In this section we give another characterization of a RKHS in terms of the eigenfunctions of a linear operator associated with the reproducing kernel. This operator, $T_R : L^2(T) \to L^2(T)$, is given by

$$T_R(f)(t) \;=\; \int_T R(s,t)\, f(s)\, ds, \quad f \in L^2(T), \quad t \in T.$$

For the eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ and the eigenfunctions $(\varphi_n)_{n\in\mathbb{N}}$ of T_R we have the following theorem (see [13, Thm. 13.5], [23, Thm. 3.a.1])

Theorem 4.1.1. *(Mercer) Let $R : T \times T \to \mathbb{R}$ be a continuous symmetric positive definite kernel that satisfies*

$$\int_T R(s,t)\, f(s)\, f(t)\, ds\, dt \;>\; 0, \quad \text{for all } \; f \in L^2(T). \tag{4.1}$$

Then there is an orthonormal basis $(\varphi_n)_{n\in\mathbb{N}}$ in $L^2(T)$ consisting of eigenfunctions of T_R such that the corresponding sequence of eigenvalues $(\lambda_n)_{n\in\mathbb{N}}$ is nonnegative. The eigenfunctions corresponding to non-zero eigenvalues are continuous on T, and R has the representation

$$R(s,t) \;=\; \sum_{j=1}^{\infty} \lambda_j\, \varphi_j(s)\varphi_j(t) \tag{4.2}$$

where the convergence is absolute and uniform on $T \times T$.

This representation can now be used to give an alternative characterization of $\mathcal{H}_R$:

Theorem 4.1.2. *([1, Lem. 3.2.2]) $(\sqrt{\lambda_n}\,\varphi_n)_{n\in\mathbb{N}}$ is an orthonormal basis for $\mathcal{H}_R$, and we have*

$$\mathcal{H}_R \;=\; \left\{ f \;:\; f(t) = \sum_{j=1}^{\infty} c_{f,j}\, \varphi_j(t), \quad t \in T, \quad \sum_{j=1}^{\infty} \frac{c_{f,j}^2}{\lambda_j} \;<\; \infty \right\}. \tag{4.3}$$

The inner product on $\mathcal{H}_R$ can be rewritten as

$$(f, g)_{\mathcal{H}_R} = \sum_{j=1}^{\infty} \frac{c_{f,j}\, c_{g,j}}{\lambda_j}\,. \tag{4.4}$$

Note that the above series expansion is just a matter of the kernel R and the domain T. It is completely independent of whether the kernel has a stochastic background or not. However, (following [1, Sec. 3.1]) we can use it for an alternative representation of a random field:

4.2 Karhunen-Loève Expansion

Let $(X_t)_{t\in T}$ be a random field with second moment function R and assume that T and the kernel R are such that a Mercer expansion exists.

From Theorem 4.1.2, using the canonical isomorphism (cf. Section 3.3), we obtain an orthonormal basis $(\xi_n)_{n\in\mathbb{N}}$ for $\mathcal{S}_X \subset L^2(\Omega, \mathcal{A}, P)$ by setting $\xi_n := \Psi_X^{-1}\left(\sqrt{\lambda_n}\,\varphi_n\right)$. Thus we have the representation

$$X_t = \sum_{j=1}^{\infty} \xi_j\, \mathbb{E}(X_t\, \xi_j), \qquad \text{for all } t \in T \tag{4.5}$$

where the series converges in $L^2(\Omega, \mathcal{A}, P)$. By using that Ψ is an isometry, we have

$$\mathbb{E}(X_t\, \xi_j) = \left(R(t,\cdot), \sqrt{\lambda_j}\,\varphi_j\right)_{\mathcal{H}_R} = \sqrt{\lambda_j}\,\varphi_j(t),$$

where the last equality follows from the reproducing kernel property of $\mathcal{H}_R$. Putting both together yields the Karhunen-Loève expansion

$$X_t = \sum_{j=1}^{\infty} \xi_j\, \sqrt{\lambda_j}\,\varphi_j(t), \qquad \text{for all } t \in T, \tag{4.6}$$

with an orthonormal sequence $(\xi_n)_{n\in\mathbb{N}}$ of random variables.

If $(X_t)_{t\in T}$ has zero mean, then $K = R$, and all RVs in $\mathcal{S}_X$ (in particular $(\xi_n)_{n\in\mathbb{N}}$) also have zero mean. If, in addition, $(X_t)_{t\in T}$ is Gaussian, then it follows from the last paragraph in Section 3.3 that any finite subset of $(\xi_n)_{n\in\mathbb{N}}$ is multivariate normally distributed. But then, by Lemma 2.4.18 and the orthonormality of $(\xi_n)_{n\in\mathbb{N}}$ we even have

$$(\xi_n)_{n\in\mathbb{N}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)\,.$$

The equivalence in (4.6) is only in $L^2(\Omega, \mathcal{A}, P)$, i.e. the sum is, in general, convergent, in each t, only in the mean square sense. The following result ([1, Thm. 3.1.2]) shows that much more is true if we know that $(X_t)_{t\in T}$ has continuous sample paths a.s.

Theorem 4.2.1. *If $(X_t)_{t\in T}$ has continuous sample paths a.s., then the sum in (4.6) converges uniformly on T with a.s.*

Based on the representation in (4.3), we can now directly compare the different model assumptions of geostatisticians and numerical analysts:
The common assumption of the latter that $f \in \mathcal{H}_R$ implies that the squared coefficients $c_{f,j}^2$ divided by the eigenvalues λ_j of the Mercer expansion are summable. The typical assumption of the former that f is a sample path of a zero-mean Gaussian RF implies that the $c_{f,j}$'s are realizations of independent RVs C_j with $C_j \sim \mathcal{N}(0, \lambda_j)$, $j \in \mathbb{N}$.
The next proposition shows (provided that R and T are such that (4.1) is satisfied), that these two assumptions can never be true at the same time (although they lead to the same interpolation scheme, see Chapter 6).

Proposition 4.2.2. *Let $(X_t)_{t\in T}$ be a zero-mean Gaussian RF with covariance kernel K $(= R)$ that has continuous sample paths a.s. Assume that (4.1) is satisfied. For a positive sequence $(w_n)_{n\in\mathbb{N}}$ and λ_j, φ_j from (4.2) let*

$$\mathcal{H}_R^{(\mathbf{w})} := \left\{ \sum_{j=1}^{\infty} c_j \, \varphi_j(\cdot) \; : \; \sum_{j=1}^{\infty} \frac{c_j^2 \, w_j}{\lambda_j} < \infty \right\}.$$

Then for the sample paths $X_{\bullet}(\omega)$ of $(X_t)_{t\in T}$ it holds that

$$\begin{aligned} \sum_{j=1}^{\infty} w_j \; < \; \infty \qquad &\Longrightarrow \qquad X_{\bullet}(\omega) \; \in \; \mathcal{H}_R^{(\mathbf{w})} \quad a.s. \\ \sum_{j=1}^{\infty} w_j \; = \; \infty \qquad &\Longrightarrow \qquad X_{\bullet}(\omega) \; \notin \; \mathcal{H}_R^{(\mathbf{w})} \quad a.s. \end{aligned}$$

Proof: The preceding arguments show that under the assumptions of the proposition, the sample paths can a.s. be represented as

$$X_{\bullet}(\omega) \; = \; \sum_{j=1}^{\infty} \xi_j(\omega) \, \sqrt{\lambda_j} \, \varphi_j(\cdot), \qquad (\xi_n)_{n\in\mathbb{N}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1) \, .$$

so it follows that $X_{\bullet}(\omega) \; \in \; \mathcal{H}_R^{(\mathbf{w})} \quad \Longleftrightarrow \quad \sum_{j=1}^{\infty} \xi_j^2(\omega) \, w_j \; < \; \infty.$

The first implication then follows via monotone convergence

$$\mathbb{E}\left(\sum_{j=1}^{\infty} \xi_j^2 \, w_j \right) \; = \; \sum_{j=1}^{\infty} \mathbb{E}\left(\xi_j^2\right) \, w_j \; = \; \sum_{j=1}^{\infty} w_j$$

by noting that a RV that assumes the value ∞ with positive probability cannot have finite expectation.

Now assume that $\sum_{j=1}^n w_j \overset{n\to\infty}{\longrightarrow} \infty$. Then, for any $\epsilon > 0$ we have

$$\begin{aligned} P\left(\sum_{j=1}^n \xi_j^2\, w_j \le (1-\epsilon)\sum_{j=1}^n w_j\right) &= P\left(\sum_{j=1}^n (\xi_j^2\, w_j - w_j) \le -\epsilon \sum_{j=1}^n w_j\right) \\ &\le P\left(\left|\sum_{j=1}^n (\xi_j^2\, w_j - w_j)\right| \ge \epsilon \sum_{j=1}^n w_j\right), \end{aligned}$$

and hence, by the Chebyshev-inequality (Lemma 2.4.12)

$$P\left(\sum_{j=1}^n \xi_j^2\, w_j \le (1-\epsilon)\sum_{j=1}^n w_j\right) \le \frac{\mathrm{Var}\left(\sum_{j=1}^n \xi_j^2\, w_j\right)}{\epsilon^2 \left(\sum_{j=1}^n w_j\right)^2} = \frac{2\, \sum_{j=1}^n w_j}{\epsilon^2 \left(\sum_{j=1}^n w_j\right)^2},$$

where the second equality holds due to the independence of $(\xi_n)_{n\in\mathbb{N}}$ and

$$\mathrm{Var}(\xi_j^2) = \mathbb{E}(\xi_j^4) - \big(\mathbb{E}(\xi_j^2)\big)^2 \overset{\text{Lemma 2.4.11}}{=} 3 - 1 = 2.$$

Now for arbitrary $M > 0$ we can find $n_0 \in \mathbb{N}$ so that $(1-\epsilon)\sum_{j=1}^n w_j \ge M$ for all $n \ge n_0$ and for these n it holds that

$$P\left(\sum_{j=1}^n \xi_j^2\, w_j \le M\right) \le P\left(\sum_{j=1}^n \xi_j^2\, w_j \le (1-\epsilon)\sum_{j=1}^n w_j\right) \le \frac{2}{\epsilon^2\, \sum_{j=1}^n w_j}.$$

Using the dominated convergence theorem, we obtain

$$P\left(\sum_{j=1}^\infty \xi_j^2\, w_j \le M\right) \le \lim_{n\to\infty} \frac{2}{\epsilon^2\, \sum_{j=1}^n w_j} = 0$$

which shows, that the assumption $\sum_{j=1}^n w_j \overset{n\to\infty}{\longrightarrow} \infty$ implies that $\sum_{j=1}^\infty \xi_j^2\, w_j$ exceeds any bound a.s. which proves the second implication.

□

For the special sequence of weights $w_j = 1$, $j \in \mathbb{N}$, we have $\mathcal{H}_R^{(\mathbf{w})} = \mathcal{H}_R$, so under the assumptions of Proposition 4.2.2 the sample paths of $(X_t)_{t\in T}$ are outside $\mathcal{H}_R$ a.s. Without the assumption that $(X_t)_{t\in T}$ is Gaussian, however, it is possible to construct a zero-mean random field with covariance function R and sample paths in $\mathcal{H}_R$ a.s.

Example 4.2.3. Let U and V be independent RVs on $(\Omega, \mathcal{A}, P)$ with

$$U \sim \mathcal{U}_{[0,1]} \quad \text{and} \quad P(V=-1) = P(V=1) = 0.5,$$

and set

$$\xi_j := V \cdot 2^{j/2} \cdot \mathbf{1}_{\left(1-(\frac{1}{2})^{j-1},\, 1-(\frac{1}{2})^j\right]}(U).$$

Using Theorem 2.2.11 (Fubini) we have

$$\begin{aligned} \mathbb{E}(\xi_j) &= \mathbb{E}(V) \cdot 2^{j/2} \cdot P\big(1 - (\tfrac{1}{2})^{j-1} < U \le 1 - (\tfrac{1}{2})^j\big) \\ &= 0 \cdot 2^{j/2} \cdot (\tfrac{1}{2})^j \;=\; 0, \qquad \text{for all } j \in \mathbb{N}, \end{aligned}$$

and

$$\begin{aligned} \mathbb{E}(\xi_j^2) &= \mathbb{E}(V^2) \cdot 2^{j} \cdot P\big(1 - (\tfrac{1}{2})^{j-1} < U \le 1 - (\tfrac{1}{2})^j\big) \\ &= 1 \cdot 2^{j} \cdot (\tfrac{1}{2})^j \;=\; 1, \qquad \text{for all } j \in \mathbb{N}. \end{aligned}$$

For $i \neq j$ the two conditions

$$U \in \big(1 - (\tfrac{1}{2})^{i-1}, 1 - (\tfrac{1}{2})^i\big] \quad \text{and} \quad U \in \big(1 - (\tfrac{1}{2})^{j-1}, 1 - (\tfrac{1}{2})^j\big]$$

can never hold at the same time, so it follows that $\mathbb{E}(\xi_j \xi_k) = 0$.

Hence, $(\xi_j)_{j\in\mathbb{N}}$ is an orthonormal sequence of centred RVs, and we can use it to define a random field $(X_t)_{t\in T}$ by its Karhunen-Loève representation (4.6). By the construction of $(\xi_j)_{j\in\mathbb{N}}$ for each $\omega \in \Omega$ there is only one nonzero coefficient $\sqrt{\lambda_j}\,\xi_j(\omega)$, so it follows from (4.3) that $X_{\bullet}(\omega) \in \mathcal{H}_R$.

The sample paths of the random field from Example 4.2.3 would be considered very untypical by a geostatistician, since they do not reflect what he has in mind when modelling some spatial variable by a (stationary) random field. The reason for this is that this random field is not ergodic, a property that is tacitly assumed in geostatistical modeling and that basically means that the behaviour of any sample path of $(X_t)_{t\in T}$ reflects the probabilistic properties of this random field (see [8, Sec. 1.1.6] for a proper definition and discussion of ergodicity in the geostatistical context).
The non-ergodicity in Example 4.2.3 results from the fact that for every ω, only one component φ_j is seen, so these paths do clearly not reflect the probabilistic structure of $(X_t)_{t\in T}$. The reason for this is that the members of the sequence $(\xi_j)_{j\in\mathbb{N}}$ are highly dependent (although uncorrelated). Conversely, the independence of $(\xi_j)_{j\in\mathbb{N}}$ was (besides the existence of fourth moments) was the crucial point in the proof of Proposition 4.2.2 for which the Gaussian assumption was needed. It is therefore plausible to conjecture that the result still holds for a more general class of stationary RFs but so far we did not pursue this issue any further (see however Proposition 5.5.5 where this issue reappears in a similar context).

Chapter 5

Sample Path Regularity of Random Fields

In this chapter we describe how random fields that have prescribed covariance or second moment functions can be constructed. We will also look at the sample paths generated by these random fields and study their regularity properties. Throughout the whole chapter we will always consider random fields on an index set $T \subseteq \mathbb{R}^d$.

5.1 Existence of Stochastic Processes

The most straight-forward image space (from a stochastic point of view) for a random field $(X_t)_{t\in T}$ is the whole of $\mathbb{R}^T$ with the product σ-algebra $\mathbb{B}^T$. If we define the probability space $(\Omega, \mathcal{A}, P)$ by $\Omega := \mathbb{R}^T$, $\mathcal{A} := \mathbb{B}^T$, then the process $(X_t)_{t\in T}$ is simply the identity on $(\Omega, \mathcal{A})$. The probability measure P can now directly be interpreted as a probability structure on the measurable subsets of paths. Potentially any function $f : T \to \mathbb{R}$ can occur as a path of $(X_t)_{t\in T}$, but only special subsets of functions will occur with positive probability.
In other words: while the image space of $(X_t)_{t\in T}$ is completely unspecific, it is the probability measure P that determines the properties of the paths, that are actually observed (in Section 5.2 we will however see, that further assumptions on $(X_t)_{t\in T}$ are needed in order that $\mathbb{B}^T$ is an appropriate σ-algebra to study path properties, and in order that these properties are uniquely determined by P).

Usually it is most convenient to define P by specifying the distribution of $X_\mathbf{t} := (X_{t_1}, \dots, X_{t_m})'$ for every vector of indices $\mathbf{t} \in T^m$, $m \in \mathbb{N}$. If this is done in a consistent way, then such a probability measure P on $(\Omega, \mathcal{A})$ exists.

Theorem 5.1.1 (Kolmogorov). *([15, Ch. I, §4, Thm. 2])*
Let $T \subseteq \mathbb{R}^d$ *and*

$$\{\mu_\mathbf{t} = \mu_{t_1,\dots,t_m} \ : \ \mathbf{t} \in T^m, \ m \in \mathbb{N}\}$$

be a system of probability distributions that respect the two subsequent consistency conditions

$$\mu_{t_1,\dots,t_m}(\times_{i=1}^m B_i) \ = \ \mu_{t_{\pi(1)},\dots,t_{\pi(m)}}(\times_{i=1}^k B_{\pi(i)}), \qquad B_i \in \mathbb{B} \quad \forall i \in J \tag{5.1}$$

for any permutation π of $\mathbf{t} := (t_1, \ldots, t_m)'$, and

$$\mu_{t_1,\ldots,t_{m-1}}(\times_{i=1}^{m-1} B_i) \;=\; \mu_{t_1,\ldots,t_m}(\times_{i=1}^{m-1} B_i \times \mathbb{R}), \qquad B_i \in \mathbb{B} \quad \forall i \in J. \tag{5.2}$$

Then there exists a unique probability measure P on $(\mathbb{R}^T, \mathbb{B}^T)$ with

$$P_{X_\mathbf{t}} = \mu_\mathbf{t} \qquad \text{for all} \quad \mathbf{t} \in T^m,\ m \in \mathbb{N}.$$

Corollary 5.1.2. *For any function $m(\cdot)$ on T and any positive definite function $K(\cdot,\cdot)$ on $T \times T$ with $K(s,t) = K(t,s)$ there exists a Gaussian random field $(X_t)_{t\in T}$ with mean function m and covariance kernel K.*

Proof: It suffices to prove the corollary for $m(\cdot) \equiv 0$, the general case follows by simply adding the desired mean function which doesn't change the covariance structure.

Define $(\Omega, \mathcal{A})$ and $(X_t)_{t\in T}$ as above, and for any $\mathbf{t} \in T^m$ let

$$C(\mathbf{t}) \;:=\; \begin{pmatrix} K(t_1,t_1) & \cdots & K(t_1,t_m) \\ \vdots & \ddots & \vdots \\ K(t_m,t_1) & \cdots & K(t_1,t_m) \end{pmatrix}.$$

From theorem 5.1.1 we then get the existence of a probability measure P on $(\Omega, \mathcal{A})$ that makes $(X_t)_{t\in T}$ a random field with the prescribed properties by verifying the consistency conditions (5.1) and (5.2) for the system of probability distributions defined by $\mu_\mathbf{t} := \mathcal{N}(\mathbf{0}, \mathbf{C}(\mathbf{t}))$.

Let $\mathbf{t}_\pi := (t_{\pi(1)}, \ldots, t_{\pi(m)})'$, $\mathbf{t}_\downarrow := (t_1, \ldots, t_{m-1})'$ and let $\psi_\pi : X_\mathbf{t} \mapsto X_{\mathbf{t}_\pi}$ and $\psi_\downarrow : X_\mathbf{t} \mapsto X_{\mathbf{t}_\downarrow}$ denote the corresponding permutation and projection maps. We have to show that

$$\psi_\pi(\mu_\mathbf{t}) = \mu_{\mathbf{t}_\pi} \qquad \text{and} \quad \psi_\downarrow(\mu_\mathbf{t}) = \mu_{\mathbf{t}_\downarrow}$$

For the present case, this can be verified by calculating the respective characteristic functions. For a multivariate Gaussian distribution, these have a nice and simple form with respect to their dependence on their covariance matrices (see lemma 2.4.18) and using Theorem 2.2.10 we obtain:

$$\begin{aligned} \widehat{\psi_\downarrow(\mu_\mathbf{t})}(\tau) \;&=\; \int_{\mathbb{R}^{m-1}} e^{i\tau'\mathbf{t}_\downarrow}\, \big(\psi_\downarrow(\mu_\mathbf{t})\big)(d\mathbf{t}_\downarrow) \;=\; \int_{\mathbb{R}^m} e^{i\tau'\psi_\downarrow(\mathbf{t})}\, \mu_\mathbf{t}(d\mathbf{t}) \\ &=\; \int_{\mathbb{R}^m} e^{i\binom{\tau}{0}'\mathbf{t}}\, \mu_\mathbf{t}(d\mathbf{t}) \;=\; e^{-\frac{1}{2}\binom{\tau}{0}'C(\mathbf{t})\binom{\tau}{0}} \;=\; e^{-\frac{1}{2}\tau'C(\mathbf{t}_\downarrow)\,\tau} \;=\; \widehat{\mu_{\mathbf{t}_\downarrow}}(\tau) \end{aligned}$$

and

$$\begin{aligned} \widehat{\psi_\pi(\mu_\mathbf{t})}(\tau) \;&=\; \int_{\mathbb{R}^m} e^{i\tau'\mathbf{t}_\pi}\, \big(\psi_\pi(\mu_\mathbf{t})\big)(d\mathbf{t}_\pi) \;=\; \int_{\mathbb{R}^m} e^{i\tau'\psi_\pi(\mathbf{t})}\, \mu_\mathbf{t}(d\mathbf{t}) \\ &=\; e^{-\frac{1}{2}\,\psi_\pi^{-1}(\tau)'\,C(\mathbf{t})\,\psi_\pi^{-1}(\tau)} \;=\; e^{-\frac{1}{2}\tau'C(\mathbf{t}_\pi)\,\tau} \;=\; \widehat{\mu_{\mathbf{t}_\pi}}(\tau) \end{aligned}$$

The validity of the consistency conditions then follows from the uniqueness of Fourier transforms (cf. Lemma 2.3.2).

5.2 Separable Random Fields

The above construction of random fields is straightforward and appealing from a point of view, that is focused on the finite dimensional distributions (as is taken e.g. in kriging). However, we will see that in order to study sample path properties such as continuity or differentiability, the information about the finite dimensional distributions alone is insufficient, and must be supplemented by the additional assumption of separability.

To motivate this assumption, we state the following

Proposition 5.2.1. *For an open subset $T \subset \mathbb{R}^d$ let $\mathcal{C}(T) \subset \mathbb{R}^T$ denote the subset of all continuous functions $f : T \to \mathbb{R}$. Then*

$$\mathcal{C}(T) \notin \mathbb{B}^T,$$

i.e. the set of continuous functions on T is not measurable.

Hence, if we stick to our construction of a random field $(X_t)_{t\in T}$ according to Theorem 5.1.1, we may not even ask for the probability of $(X_t)_{t\in T}$ having continuous paths.

Proposition 5.2.1 is just a special example for the more general fact that a subset $A \subset \mathbb{R}^T$ cannot lie in $\mathbb{B}^T$ unless there exists a countable subset D of T with the property that, if $x \in A$ and $x(t) = y(t)$ for all $t \in D$, then $y \in A$ (see [5, Thm. 36.3]). This is a consequence of the definition of a product σ-algebra which starts from events defined by finite projections only (cf. Definition 2.1.3) and generalizes to countable projections by intersection. Properties of $(X_t)_{t\in T}$ such as continuity, that effectively involve all the points in $t \in T$, in general are not in the product σ-algebra.

Moreover, defining the probability measure P on $(\Omega, \mathcal{A}) = (\mathbb{R}^T, \mathbb{B}^T)$ by specifying the finite-dimensional distributions $P_{X_\mathbf{t}}$ does not determine the process $(X_t)_{t\in T}$ uniquely:

Example 5.2.2. (from [5, Sec. 38]) Let $(\Omega, \mathcal{A}, P) = \big([0,1], \mathbb{B} \cap [0,1], \mathcal{U}_{[0,1]}\big)$ and define the two processes

$$\begin{aligned} X_t(\omega) &= 0 \qquad \text{for all } t, \omega \\ Y_t(\omega) &= \begin{cases} 0 & t \neq \omega \\ 1 & t = \omega \end{cases} \end{aligned}$$

Then $P\big(X_t = Y_t\big) = 1$ for all $t \in T$, which implies that they have the same finite-dimensional distributions, but

$$\begin{aligned} &P\big(X_{\boldsymbol{\cdot}} \text{ is continuous on } [0,1]\big) = 1, \\ &P\big(Y_{\boldsymbol{\cdot}} \text{ is continuous on } [0,1]\big) = 0. \end{aligned}$$

It is because the position of the discontinuity has a continuous distribution that the two processes have the same finite-dimensional distributions.

Definition 5.2.3. Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two random fields on some index set T over the same probability space $(\Omega, \mathcal{A}, P)$. If

$$P\big(X_t = Y_t\big) = 1 \quad \text{for all} \;\; t \in T$$

then $(Y_t)_{t\in T}$ is called a version of $(X_t)_{t\in T}$.

A way out of this dilemma, that important path properties lead to unmeasurable events, and are not fully determined by finite-dimensional distributions of $(X_t)_{t\in T}$, is to restrict attention to separable processes:

Definition 5.2.4. [15, Ch. III, §2, Def. 2]
A random field $(X_t)_{t\in T}$ over the probability space $(\Omega, \mathcal{A}, P)$ is called separable, if there exists a countable dense subset $D \subset T$ and a set $N \in \mathcal{A}$ of probability 0, so that for any open set $I \subset T$ and any closed set $B \subset \mathbb{R}$ the two sets

$$\begin{array}{lclll} A_{B,I} & := & \{\omega \,:\, X_t(\omega) \in B & \text{for all} & t \in I\} \\ A_{B,I\cap D} & := & \{\omega \,:\, X_t(\omega) \in B & \text{for all} & t \in I \cap D\} \end{array}$$

differ from each other only on a subset of N.

Note that $A_{B,I\cap D}$ is measurable while $A_{B,I}$ in general is not, but for a separable process we can switch from $A_{B,I}$ to $A_{B,I\cap D}$ by adding/subtracting a subset $\tilde{N}$ of N. It is convenient to require all of these subsets to be measurable with $P(\tilde{N}) = 0$, therefore from now we will tacitly assume that $(\Omega, \mathcal{A}, P)$ is complete (see Def. 2.1.13 and subsequent remark).

A priori, separability is defined with respect to some (countable) set of separability D. Under a mild additional condition, which will always be fulfilled in our case of continuous covariance kernels (see Section 5.3), the choice of D is arbitrary.

Definition 5.2.5. A random field $(X_t)_{t\in T}$ is called stochastically continuous at point t if for any $\epsilon > 0$

$$P\big(\big|X_{t+h} - X_t\big| > \epsilon\big) \to 0 \quad \text{as} \quad h \to 0$$

The random field is called stochastically continuous if it is stochastically continuous at every $t \in T$.

Theorem 5.2.6. *([15, Ch. III, §2, Thm. 5]) Let $(X_t)_{t\in T}$ be a separable random field. If $(X_t)_{t\in T}$ is stochastically continuous, then any countable dense set $D \subset T$ may serve as its set of separability.*

We prove the next statement from [5, Sec. 38] based on Definition 5.2.4:

Lemma 5.2.7. *Let $(X_t)_{t\in T}$ be a separable random field. Then*

$$\sup_{t\in I} X_t \;=\; \sup_{t\in I\cap D} X_t \quad \textit{a.s.} \quad \textit{and} \quad \inf_{t\in I} X_t \;=\; \inf_{t\in I\cap D} X_t \quad \textit{a.s.}$$

for any open set $I \in T$.

Proof: Using the same notation as in Definition 5.2.4 and letting $B = [a, b]$ with $a, b \in \mathbb{R}, a < b$, we note that

$$A_{[a,b],I} \; := \; \Big\{\omega \, : \, a \leq \inf_{t\in I} X_t(\omega) \; \leq \; \sup_{t \in I} X_t(\omega) \leq b\Big\}$$

By Definition there exists a set $N = \bigcup_{a,b\in\mathbb{Q},\, a<b} N_{a,b} \subset \Omega$ of probability 0 so that

$$A_{[a,b],I\cap D} \setminus A_{[a,b],I} \subset N \qquad \text{for all } a, b \in \mathbb{Q}, \; a < b,$$

Now we have the implication

$$\sup_{t\in I} X_t(\omega) \; \neq \; \sup_{t\in I\cap D} X_t(\omega) \implies \exists\, b \in \mathbb{Q} \; : \; \sup_{t\in I\cap D} X_t(\omega) \; \leq \; b \; < \; \sup_{t\in I} X_t(\omega)$$

so if the suprema on the left differ, then necessarily $\omega \in A_{[a,b],I\cap D} \setminus A_{[a,b],I}$ for some $a \in \mathbb{Q}$, $a < b$, so this can only happen with probability 0. The argument for the infimum is the same. □

If in addition it is known that the finite-dimensional distributions of a random field $(X_t)_{t\in T}$ allow for continuous sample paths, then the assumption of separability entails a certain uniqueness:

Lemma 5.2.8. *Let $(X_t)_{t\in T}$ be a separable random field and let $(Y_t)_{t\in T}$ be a version of $(X_t)_{t\in T}$ having continuous sample paths a.s. Then*

$$P\big(X_t = Y_t \quad \textit{for all} \;\; t \in T\big) = 1.$$

In particular $(X_t)_{t\in T}$ has continuous sample paths a.s.

Proof: (generalizes [2, Ch. 1, Sec. 4, Prop. 1.9] to the case $d > 1$)

Let D and N be as in the definition of separability. Let

$$A' = \big\{\omega \, : \, X_t(\omega) = Y_t(\omega) \;\; \text{for all } t \in D\big\}.$$

Since $(Y_t)_{t\in T}$ is a version of $(X_t)_{t\in T}$ we have $P(A') = 1$. Further let

$$\begin{aligned} \overline{A} \quad &= \bigcap_{J=B_\epsilon(a)\cap T \,:\, \epsilon\in\mathbb{Q}_+, a\in\mathbb{Q}^d} \left\{\omega \, : \, \sup_{t\in J} X_t(\omega) = \sup_{t\in J\cap D} X_t(\omega)\right\} \quad \text{and} \\ \underline{A} \quad &= \bigcap_{J=B_\epsilon(a)\cap T \,:\, \epsilon\in\mathbb{Q}_+, a\in\mathbb{Q}^d} \left\{\omega \, : \, \inf_{t\in J} X_t(\omega) = \inf_{t\in J\cap D} X_t(\omega)\right\}. \end{aligned}$$

where $B_\epsilon(a)$ denotes the open ball of radius ϵ centred at a. It follows from Lemma 5.2.7 that $P\big(\overline{A} \cap \underline{A}\big) = 1$ and w.l.o.g. we can also assume that $(Y_t)_{t\in T}$ has continuous sample paths for all $\omega \in A' \cap \overline{A} \cap \underline{A}$.

Now let $\omega \in A' \cap \overline{A} \cap \underline{A}$ and $t \in T$. For any $\epsilon \in \mathbb{Q}_+$ we can choose $a \in \mathbb{Q}^d$ so that $t \in B_\epsilon(a)$. Defining $J(\epsilon) := B_\epsilon(a) \cap T$ we have

$$X_t(\omega) \; \leq \; \sup_{s\in J(\epsilon)} X_s(\omega) \; = \; \sup_{s\in J(\epsilon)\cap D} X_s(\omega) \; = \; \sup_{s\in J(\epsilon)\cap D} Y_s(\omega) \; \leq \; \sup_{s\in J(\epsilon)} Y_s(\omega)$$

Letting $\epsilon \to 0$ it follows from the continuity of $Y_{.}(\omega)$ that

$$X_t(\omega) \;\leq\; \limsup_{\epsilon\to 0}\ \sup_{s\in J(\epsilon)} Y_s(\omega) = Y_t(\omega).$$

In a similar way one proves $X_t(\omega) \geq Y_t(\omega)$, so the two versions are identical for all $\omega \in A' \cap \overline{A} \cap \underline{A}$, which is a set of probability 1.

□

Note from the proof that we could replace the requirement that $(Y_t)_{t\in T}$ is a version of $(X_t)_{t\in T}$ by the weaker requirement that

$$P\big(X_t = Y_t\big) = 1 \;\text{ for all }\; t \in D,$$

where D is the set of separability. In particular, if $(X_t)_{t\in T}$ is stochastically continuous, then the statement of Lemma 5.2.8 still holds as long as

$$P\big(X_t = Y_t\big) = 1 \;\text{ for a.e. }\; t \in T,$$

since by Theorem 5.2.6 we may choose D such that it does not contain any of the exceptional points.

The next theorem (a special case of [15, Ch. III, §2, Thm. 2]) shows, that among all versions of $(X_t)_{t\in T}$ we can always find a separable one:

Theorem 5.2.9. *For any random field $(X_t)_{t\in T}$ there exists on the same probability space a separable version $(Y_t)_{t\in T}$ taking on values in the compact extension $(\bar{\mathbb{R}}, \bar{\mathbb{B}})$ of $(\mathbb{R}, \mathbb{B})$.*

When constructing the paths of a separable version $(Y_t)_{t\in T}$, it may be necessary to assign this path the additional value ∞, but for every fixed $t \in T$ the probability of this is zero. Combining Theorem 5.2.9 with Kolmogorov's existence theorem shows that for any consistent system of probability distributions $\{\mu_{\mathbf{t}} : \mathbf{t} \in T^m,\ m \in \mathbb{N}\}$ there exists a separable process with these finite-dimensional distributions.

We conclude this section with a lemma (actually a corollary of [5, Thm. 38.2]) that generalizes Lemma 5.2.8 in that it allows to compare processes over different probability spaces.

Lemma 5.2.10. *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two random fields over the probability spaces $(\Omega, \mathcal{A}, P)$ and $(\Omega', \mathcal{A}', P')$ respectively having the same finite-dimensional distributions. If*

$$A'_{\mathcal{C}} \;:=\; \{\omega' \in \Omega' \,:\, Y_{\bullet}(\omega') \textit{ is continuous on } T\}$$

lies in $\mathcal{A}'$ with $P(A'_{\mathcal{C}}) = 1$, and if $(X_t)_{t\in T}$ is separable, then for

$$A_{\mathcal{C}} \;:=\; \{\omega \in \Omega \,:\, X_{\bullet}(\omega) \textit{ is continuous on } T\}$$

we also have $P(A_{\mathcal{C}}) = 1$.

Lemma 5.2.10 presents a useful link between the point of view of a numerical analyst and a statistician. The former would probably not think of constructing a random field as in Section 5.1 but rather start from a space like $\Omega = \mathcal{C}(T)$ with Borel σ-algebra $\mathcal{A}$ on it, define $(X_t)_{t\in T}$ as the embedding of $\mathcal{C}(T)$ in $\mathbb{R}^T$ and use P to assign a probability structure on $(\Omega, \mathcal{A})$. This way the continuity of the paths of $(X_t)_{t\in T}$ is trivial. Lemma 5.2.10 now tells us, that any separable process $(Y_t)_{t\in T}$ with the same finite-dimensional distributions as $(X_t)_{t\in T}$ will also necessarily have continuous paths a.s.

5.3 Sample Path Regularity in the Gaussian Case

We are mainly interested in regularity properties such as continuity and differentiability of the sample paths of a random field. Nevertheless we shall also introduce the (weaker) concepts of mean square continuity and mean square differentiability as these are directly linked to the second-order structure of a random field.

In this and all subsequent sections we always assume that $(X_t)_{t\in T}$ is a second-order random field with second-moment function R, covariance function K and mean function m. As in the preceding Chapters, we assume that R, K and m are continuous.

5.3.1 Continuity

Definition 5.3.1. A random field $(X_t)_{t\in T}$ is called continuous in the mean square sense (briefly m.s. continuous) at point t if

$$\mathbb{E}\big(X_{t+h} - X_t\big)^2 \to 0 \quad \text{as} \quad h \to 0.$$

The random field is called continuous in the mean square sense (briefly m.s. continuous) if it is m.s. continuous at every $t \in T$.

Remark 5.3.2. A related notion is that of stochastic continuity at t (see Definition 5.2.5). By Lemma 2.4.12 (Markov's inequality) m.s. continuity at t implies stochastic continuity at t.

Continuity of a covariance kernel and m.s. continuity of the corresponding random field are linked in a natural way (cf. [27, P-2-1]):

Theorem 5.3.3. *Let $(X_t)_{t\in T}$ be a random field with second-moment kernel $R : T \times T \to \mathbb{R}$. Then the following statements are equivalent:*

1. *$(X_t)_{t\in T}$ is m.s. continuous,*
2. *R is continuous on $T \times T$,*
3. *R is continuous on the diagonal of $T \times T$.*

The natural link is between m.s. continuity and the second-moment function R. From the stochastic modelling point of view it is however preferable to work with the covariance function K rather than R due to the interpretation of a random field as some random fluctuation, controlled by K, around some deterministic mean function m. Recalling that

$$R(s,t) \;=\; K(s,t) + m(s)\, m(t) \qquad \text{for all } s,t \in T,$$

we have that in the case of a continuous mean function m continuity of R is equivalent to continuity of K. Throughout the section, we will therefore assume that $(X_t)_{t\in T}$ is centred (i.e. $m(t) \equiv 0$), bearing in mind that the general case can be obtained by adding some continuous mean function m.

In the same way, we will also study all other regularity properties only for centred random fields. Adding a mean function that has itself the required regularity will always give the general case.

In the special case of a stationary random field we note the following

Corollary 5.3.4. *Let $(X_t)_{t\in T}$ be a weakly stationary centred random field with covariance kernel $K(\cdot,\cdot) = \Phi(\cdot - \cdot)$. Then the following statements are equivalent:*

1. *$(X_t)_{t\in T}$ is m.s. continuous,*
2. *Φ is continuous on T,*
3. *Φ is continuous at the origin.*

In the case where Φ is defined (and positive definite) on the whole of $\mathbb{R}^d$ we can obtain a further equivalent statement (cf. e.g. [41, Thm. 6.6]):

Theorem 5.3.5. *(Bochner)*
A function $\Phi : \mathbb{R}^d \to \mathbb{R}$ is continuous and positive definite if and only if a finite symmetric non-negative measure ν on $\mathbb{R}^d$ exists so that

$$\Phi(h) = \int_{\mathbb{R}^d} \cos(h'\omega)\, \nu(d\omega) \qquad \textit{for all} \;\; h \in \mathbb{R}^d.$$

In stochastics, ν is called spectral measure and we will use it frequently in this chapter to formulate alternative sufficient conditions for the regularity of sample paths. If ν is absolutely continuous w.r.t. the Lebesgue measure λ^d (see also Definition 2.4.5), then there exists a non-negative, integrable function $\varphi : \mathbb{R}^n \to \mathbb{R}$ so that

$$\nu(B) = \int_B \varphi(\omega)\, d\omega \quad \text{for all } B \in \mathbb{B}^n.$$

φ is called the spectral density. In Definition 3.2.12 we have already introduced the Whittle-Matérn class of covariance functions which was characterized by its spectral density

$$\varphi(\omega) \;=\; \left(1 + \|\omega\|^2\right)^{-\tau}.$$

Note that all statements so far only depend on R respectively K and do not require any further assumption on the finite dimensional distributions of $(X_t)_{t\in T}$. This is no longer true (see Example 5.5.1) for the following Theorem ([1, Thm. 1.4.1]) on a.s. sample path continuity, the proof of which explicitly uses the assumption that $(X_t)_{t\in T}$ is Gaussian.

Define the (pseudo-)metric d on T by

$$d^2(s,t) \;:=\; \mathbb{E}\big((X_t - X_s)^2\big) \;=\; K(s,s) + K(t,t) - 2\,K(s,t).$$

Theorem 5.3.6. *Let $(X_t)_{t \in T}$ be a separable centred Gaussian process on a compact index set $T \subset \mathbb{R}^d$. If for some $0 < C < \infty$ and $\delta, \eta > 0$,*

$$d^2(s,t) \;\leq\; \frac{C}{\big|\log \|t-s\|\big|^{1+\delta}} \tag{5.3}$$

for all $s, t \in T$ with $\|t-s\| < \eta$, then the paths of $(X_t)_{t \in T}$ are a.s. continuous and bounded on T.

The restriction to compact $T \subset \mathbb{R}^d$ is not a serious problem. As far as continuity is concerned, if T is σ-compact (i.e. if it can be represented as a countable union of compact sets) then a.s. continuity on its compact subsets immediately implies a.s. continuity over T itself (the same is not true for boundedness).
Condition (5.3) is quite sharp, but not necessary. There are simple examples of processes (see [1, Sec. 1.4.1]) with a high level of nonstationarity that are a.s. continuous but do not satisfy (5.3). However, for the class of stationary processes, which are most important in practice, the following theorem ([1, Cor. 1.5.5]) shows, that this condition is reasonably definite.

Theorem 5.3.7. *Let $(X_t)_{t \in T}$ be a separable centred stationary Gaussian process on an open index set $T \subseteq \mathbb{R}^d$ with covariance function Φ. If for some $0 < C_1, C_2 < \infty$,*

$$\frac{C_1}{\big(-\log \|h\|\big)^{1+\delta_1}} \;\leq\; \Phi(0) - \Phi(h) \;\leq\; \frac{C_2}{\big(-\log \|h\|\big)^{1+\delta_2}},$$

for all $\|h\|$ small enough, then the paths of $(X_t)_{t \in T}$ are a.s. continuous if $\delta_2 > 0$ and a.s. discontinuous if $\delta_1 < 0$.

Via Tauberian theory (cf. e.g. [6]), which translates the behaviour of Φ at the origin to that of ν at infinity, one can also obtain a sufficient condition for path continuity on the spectral measure ([1, Sec. 1.4.1]):

Theorem 5.3.8. *Let $(X_t)_{t \in T}$ be a separable, centred and stationary Gaussian process on an open $T \subset \mathbb{R}^d$ with spectral measure ν. If the integral*

$$\int_{\mathbb{R}^d} \big(\log(1 + \|\omega\|)\big)^{1+\delta} \; \nu(d\omega)$$

converges for some $\delta > 0$, then $(X_t)_{t \in T}$ has continuous sample paths a.s. If it diverges for some $\delta < 0$, then $(X_t)_{t \in T}$ has discontinuous sample paths a.s.

5.3.2 Differentiability

As before, we denote by e_i the unit vector in $\mathbb{R}^d$ in direction of the i-th coordinate axis. We study the behaviour of the difference quotients

$$X_t^{(i,h)} := \frac{X_{t+he_i} - X_t}{h}, \qquad t \in T, \quad h \in \mathbb{R} \text{ s.t. } t + he_i \in T.$$

Definition 5.3.9. A random field $(X_t)_{t\in T}$ has a m.s. partial derivative at t in the direction e_i if there exists a RV $X_t^{(i)} \in L^2(\Omega, \mathcal{A}, P)$ so that

$$\mathbb{E}\left(X_t^{(i,h)} - X_t^{(i)}\right)^2 \to 0 \qquad \text{as } h \to 0.$$

If $(X_t)_{t\in T}$ has a m.s. partial derivative in the direction e_i at all $t \in T$, we say that it has a m.s. partial derivative in the direction e_i, and we denote by $(X_t^{(i)})_{t\in T}$ the corresponding m.s. partial derivative process.

If $(X_t)_{t\in T}$ has a m.s. partial derivative in all directions e_i, $i = 1, \ldots, d$, we say that it has mean square partial derivatives.

M.s. differentiability hence corresponds to m.s. convergence (see Definition 2.4.21) of $X_t^{(i,h)}$ to $X_t^{(i)}$. When asking for a.s. differentiability of the sample paths, we actually want to establish a.s. convergence (see Definition 2.4.19). As was pointed out in the remark subsequent to Theorem 2.4.22, neither of these notions implies the other, but in this subsection we shall derive additional conditions under which the respective implications hold.

First, we shall however explore the connection between m.s. differentiability and differentiability of R (or K and m). To this end we need to study the limit behaviour of the covariances

$$\begin{aligned} R_{hh'}^{(i)}(s,t) &:= \mathbb{E}\left(X_s^{(i,h')} X_t^{(i,h)}\right) \\ &= \frac{R(s+h'e_i, t+he_i) - R(s+h'e_i, t) - R(s, t+he_i) + R(s,t)}{hh'}. \end{aligned}$$

Theorem 5.3.10. *Let $(X_t)_{t\in T}$ be a random field with second-moment function R. Then the following statements are equivalent:*

1. *$(X_t)_{t\in T}$ has a m.s. partial derivative in the direction e_i at t,*
2. *The generalized mixed partial derivative $D^{i,i}R(t,t) := \lim_{h,h'\to 0} R_{hh'}^{(i)}(t,t)$ exists, i.e. for every $\epsilon > 0$ exist a δ so that*

$$|h|, |h|' < \delta \quad \Longrightarrow \quad \left|R_{hh'}^{(i)}(t,t) - D^{i,i}R(t,t)\right| < \epsilon.$$

If $D^{i,i}R(t,t)$ exists for all $t \in T$, then $D^{i,i}R(s,t)$ exists for all $s,t \in T$.

Proof: (generalizes [15, Ch. IV, §3, Thm. 4] to $d > 1$)

(1) ⇒ (2): If $X_t^{(i)}$ exists as limit in $L^2(\Omega, \mathcal{A}, P)$, we have

$$\begin{aligned}&\left|\lim_{h,h'\to 0} R^{(i)}_{hh'}(t,t) - \mathbb{E}\left(X_t^{(i)}\, X_t^{(i)}\right)\right| \\ &\quad = \lim_{h,h'\to 0}\left|\mathbb{E}\left(\big(X_t^{(i,h)} - X_t^{(i)}\big)\, X_t^{(i,h')}\right) + \mathbb{E}\left(\big(X_t^{(i,h')} - X_t^{(i)}\big)\, X_t^{(i)}\right)\right|\end{aligned}$$

Using the continuity of the scalar product in $L^2(\Omega, \mathcal{A}, P)$ we get

$$\begin{aligned}\left|\mathbb{E}\left(\big(X_t^{(i,h')} - X_t^{(i)}\big)\, X_t^{(i)}\right)\right| \quad &\to \quad 0, \qquad \text{as } h' \to 0 \qquad \text{and}\\ \left|\mathbb{E}\left(\big(X_t^{(i,h)} - X_t^{(i)}\big)\, X_t^{(i,h')}\right)\right| \quad &\leq \quad \underbrace{\left|\mathbb{E}\left(\big(X_t^{(i,h)} - X_t^{(i)}\big)^2\right)\right|^{\frac{1}{2}}}_{\to 0 \text{ as } h\to 0}\underbrace{\left|\mathbb{E}\left(\big(X_t^{(i,h')}\big)^2\right)\right|^{\frac{1}{2}}}_{\leq M\ \forall h'},\end{aligned}$$

so the limit exists as h and h' tend to 0 independently.

In the same way, if both $X_s^{(i)}$ and $X_t^{(i)}$ exist as limits in $L^2(\Omega, \mathcal{A}, P)$, we obtain the existence of $\lim_{h,h'\to 0} R^{(i)}_{hh'}(s,t)$.

(2) ⇒ (1): If the limit $\lim_{h,h'\to 0} R^{(i)}_{hh'}(t,t)$ exists, then it follows that

$$\mathbb{E}\left(\left(X_t^{(i,h)} - X_t^{(i,h')}\right)^2\right) \;=\; R^{(i)}_{hh}(t,t) - 2\,R^{(i)}_{hh'}(t,t) + R^{(i)}_{h'h'}(t,t)$$

tends to 0 as h and h' tend to 0. Hence, for any null sequence $(h_n)_{n\in\mathbb{N}}$, $X_t^{(i,h_n)}$ is a Cauchy sequence in $L^2(\Omega, \mathcal{A}, P)$ and has a limit $X_t^{(i)}$.

□

It follows directly from Definition 5.3.9 that $(X_t^{(i)})_{t\in T}$ is itself a second-order random field and we have

$$\begin{aligned}\mathbb{E}\big(X_s^{(i)}\, X_t\big) \;&=\; \lim_{h'\to 0}\frac{1}{h'}\big(R(s+h'e_i, t) - R(s,t)\big) \;=\; \frac{\partial R}{\partial_1 e_i}(s,t),\\ \mathbb{E}\big(X_s\, X_t^{(i)}\big) \;&=\; \lim_{h\to 0}\frac{1}{h}\big(R(s, t+h\,e_i) - R(s,t)\big) \;=\; \frac{\partial R}{\partial_2 e_i}(s,t),\\ \mathbb{E}\big(X_s^{(i)}\, X_t^{(i)}\big) \;&=\; \lim_{h\to 0}\lim_{h'\to 0} R^{(i)}_{hh'}(s,t) \;=\; \frac{\partial^2 R}{\partial_1 e_i\,\partial_2 e_i}(s,t),\end{aligned}$$

where $\frac{\partial R}{\partial_1 e_i}$ and $\frac{\partial R}{\partial_2 e_i}$ denote the partial derivatives of R in the direction e_i with respect to the first and the second argument respectively.

Moreover, by the Hölder inequality we have

$$\mathbb{E}\left(\left|X_t^{(i,h)} - X_t^{(i)}\right|\right) \;\leq\; \left|\mathbb{E}\left(\big(X_t^{(i,h)} - X_t^{(i)}\big)^2\right)\right|^{\frac{1}{2}},$$

and hence the mean function of $(X_t^{(i)})_{t\in T}$ exist and is given by

$$\mathbb{E}\left(X_t^{(i)}\right) \;=\; \lim_{h\to 0}\mathbb{E}\left(X_t^{(i,h)}\right) \;=\; \frac{\partial m}{\partial e_i}(t).$$

M.s. differentiability of $(X_t)_{t\in T}$ allows to bound $d(s,t)$ (as defined before Thm. 5.3.6) in terms of the euclidean distance of s and t:

Lemma 5.3.11. *If the random field $(X_t)_{t\in T}$ on a compact $T \subset \mathbb{R}^d$ has m.s. partial derivatives and if these are m.s. continuous, then for some $0 < C < \infty$ and some $\eta > 0$ we have*

$$d(s,t) \;\le\; C\,\|t-s\|$$

for all $s,t \in T$ with $\|t-s\| < \eta$.

Proof: First fix $i \in \{1,\ldots,d\}$ and define the function $g : T \times [1,1] \to \mathbb{R}$ by

$$g(t,h) \;:=\; \begin{cases} R^{(i)}_{hh}\,(t,t), & h \neq 0 \\ D^{i,i}R\,(t,t), & h = 0 \end{cases}$$

We show that g is continuous with respect to the metric $\max(\|t\|,|h|)$. At any point (t,h) with $h \neq 0$ this is an obvious consequence of the continuity of R. For any point $(t,0)$ it follows from the continuity of $D^{i,i}R$ and

$$\begin{aligned} |g(s,h')-g(t,0)| &\le |g(s,h')-g(t,h')| + |g(t,h')-g(t,0)| \\ &= \underbrace{\left|R^{(i)}_{h'h'}(t,t) - R^{(i)}_{h'h'}(s,s)\right|}_{\to 0 \text{ as } s\to t} + \underbrace{\left|R^{(i)}_{h'h'}(t,t) - D^{i,i}R\,(t,t)\right|}_{\to 0 \text{ as } h'\to 0}. \end{aligned}$$

Since T was assumed compact, g is even uniformly continuous on $T \times [1,1]$ and it follows that for every $\epsilon > 0$ there exists an $\eta_i > 0$ so that for all $|h| \le \eta_i$

$$\left|R^{(i)}_{hh}(t,t) - D^{i,i}R\,(t,t)\right| \;\le\; \epsilon \qquad \text{for all } t \in T.$$

Defining $C_i^2 := \epsilon + \sup_{t\in T} \left|D^{i,i}R\,(t,t)\right|$ this implies

$$d^2(t,t+he_i) \;=\; R^{(i)}_{hh}(t,t)\cdot h^2 \;\le\; C_i^2\,h^2 \qquad \text{for all } |h| \le \eta_i.$$

Finally, setting $C := d \cdot \max_{1\le i\le d} C_i$ and $\eta := \min_{1\le i\le d} \eta_i$ we have for $\|t-s\| < \eta$

$$\begin{aligned} d(s,t) &= d\left(s, s + \sum_{i=1}^d (t_i - s_i)e_i\right) \\ &\le \sum_{i=1}^d \underbrace{d\left(s + \sum_{j=1}^{i-1}(t_i - s_i)e_i,\; s + \sum_{j=1}^{i}(t_i-s_i)e_i\right)}_{\le C_i\,|t_i - s_i| \,\le\, C\,\|t-s\|\,/\,d} \;\le\; C\,\|t-s\| \end{aligned}$$

which completes the proof.

□

By applying Theorem 5.3.6 we obtain

Corollary 5.3.12. *If the random field $(X_t)_{t\in T}$ is separable, Gaussian, and has m.s. partial derivatives, and if these are m.s. continuous, then $(X_t)_{t\in T}$ has continuous sample paths a.s.*

From now on, we will again restrict ourselves to centred random fields and formulate all subsequent conditions in terms of the covariance kernel K. In the special case of a stationary random field where $K(\cdot,\cdot) = \Phi(\cdot - \cdot)$ we have the following corollary to Theorem 5.3.10:

Corollary 5.3.13. *Let $(X_t)_{t\in T}$ be a weakly stationary centred random field with covariance function Φ and spectral measure ν. Then the following statements are equivalent:*

1. *$(X_t)_{t\in T}$ has a m.s. partial derivative in the direction e_i,*
2. *$(X_t)_{t\in T}$ has a m.s. partial derivative at some $t_0 \in T$ in the direction e_i,*
3. *The second partial derivative $\frac{\partial^2\Phi}{(\partial e_i)^2}$ exists on T,*
4. *The second partial derivative $\frac{\partial^2\Phi}{(\partial e_i)^2}$ exists at the origin,*
5. *The i-th spectral moment $M_i := \int_{\mathbb{R}^d} \omega_i^2\, \nu(d\omega)$ exists and is finite.*

Proof:

(1) ⇒ (3), (2) ⇒ (4): follows from Theorem 5.3.10 by noting that

$$\lim_{h,h'\to 0} K^{(i)}_{hh'}(s,t) = -\frac{\partial^2\Phi}{(\partial e_i)^2}(t-s)\,, \qquad \lim_{h,h'\to 0} K^{(i)}_{hh'}(t,t) = -\frac{\partial^2\Phi}{(\partial e_i)^2}(0).$$

(1) ⇒ (2), (3) ⇒ (4): trivial

(4) ⇒ (5): (from [9, Thm. 6.4.1]) If $\frac{\partial^2\Phi}{(\partial e_i)^2}(0)$ exists and is finite, we have

$$\frac{\partial^2\Phi}{(\partial e_i)^2}(0) = \lim_{h\to 0} \frac{\Phi(he_i) - 2\,\Phi(0) + \Phi(-he_i)}{h^2} = -2 \lim_{h\to 0} \int_{\mathbb{R}^d} \frac{1-\cos(\omega_i h)}{h^2}\, \nu(d\omega)$$

by the symmetry of Φ and Bochner's Theorem. Using Fatou's lemma we get

$$\begin{aligned}\int_{\mathbb{R}^d} \omega_i^2\, \nu(d\omega) &= 2\int_{\mathbb{R}^d} \lim_{h\to 0} \frac{1-\cos(\omega_i h)}{h^2}\, \nu(d\omega) \\ &\le 2 \lim_{h\to 0} \int_{\mathbb{R}^d} \frac{1-\cos(\omega_i h)}{h^2}\, \nu(d\omega) = -\frac{\partial^2\Phi}{(\partial e_i)^2}(0).\end{aligned}$$

(5) ⇒ (1): (oral communication with Prof. Schlather, see also [27, P-3-5])
We show that the existence of M_i implies the existence of

$$\lim_{h,h'\to 0} \Phi^{(i)}_{hh'}(0) = \lim_{h,h'\to 0} \frac{\Phi((h-h')e_i) - \Phi(he_i) - \Phi(h'e_i) + \Phi(0)}{hh'}\,.$$

From Theorem 5.3.5 (Bochner) we have

$$\Phi^{(i)}_{hh'}(0) \;=\; \int_{\mathbb{R}^d} I_{hh'}(\omega_i)\,\nu(d\omega),$$

where

$$I_{hh'}(\omega_i) \;:=\; \frac{\cos(\omega_i(h-h')) - \cos(\omega_i h) - \cos(\omega_i h') + 1}{hh'}\,.$$

By expanding the cosine terms it is easy to see that

$$I_{hh'}(\omega_i) \;=\; \omega_i^2\left(1 + o(\omega_i h) + o(\omega_i h')\right) \qquad \text{as } \omega_i h, \omega_i h' \to 0, .$$

Next, using $|\sin(x)| \le |x|$, $1-\cos(x) \le \min\left(\frac{x^2}{2}, 2\right)$ and the trigonometric identity $\cos(x-y) = \cos(x)\cos(y) + \sin(x)\sin(y)$ one can derive the inequality

$$\left|\cos(x-y) - \cos(x) - cos(y) + 1\right| \;\le\; 2\,|x|\,|y|\,.$$

Applying this to the enumerator of $I_{hh'}(\omega_i)$ yields a dominating, integrable function

$$\left|I_{hh'}(\omega_i)\right| \;\le\; \frac{2\,|\omega_i h|\,|\omega_i h'|}{|h||h'|} \;=\; 2\,\omega_i^2$$

and via dominated convergence we get

$$\lim_{h,h'\to 0} C^{(i)}_{hh'}(0) \;=\; \int_{\mathbb{R}^d} \lim_{h,h'\to 0} I_{hh'}(\omega_i)\,\nu(d\omega) \;=\; \int_{\mathbb{R}^d} \omega_i^2\,\nu(d\omega).$$

But this implies, according to Theorem 5.3.10, that $(X_t)_{t\in T}$ has a m.s. partial derivative in the direction e_i. □

For later use we state another characterization of m.s. differentiability in the stationary case. Clearly, the existence of $\frac{\partial^2\Phi}{(\partial e_i)^2}$ at the origin implies

$$\left|\,\Phi^{(i)}_{hh}(0)\,\right| \;=\; \left|\,\frac{2\,\Phi(0) - 2\,\Phi(he_i)}{h^2}\,\right| \;\le\; b, \qquad \text{for all } \; h\in\mathbb{R} \tag{5.4}$$

for some $b\in\mathbb{R}$. It is quite remarkable that the converse is also true:

Lemma 5.3.14. *([27, P-3-5]) Let $(X_t)_{t\in T}$ be a weakly stationary centred random field with covariance function Φ. If (5.4) holds for some $b\in\mathbb{R}$, then $(X_t)_{t\in T}$ has a m.s. partial derivative in the direction e_i.*

Another remarkable property of weakly stationary random fields is that m.s. differentiability automatically implies that random fields $(X^{(i)}_t)_{t\in T}$ of the m.s. partial derivatives are m.s. continuous.

Proposition 5.3.15. *The m.s. partial derivative $(X_t^{(i)})_{t\in T}$ of a weakly stationary centred random field $(X_t)_{t\in T}$ with spectral measure ν, has itself the spectral measure $\nu_{(i)}$ defined by*

$$\nu_{(i)}(B) \;:=\; \int_B \omega_i^2\, \nu(d\omega) \qquad \textit{for all}\;\; B \in \mathbb{B}^d.$$

In particular the m.s. partial derivatives of a weakly stationary random field are always m.s. continuous.

Proof: In the same way as above, we write

$$\begin{aligned}\frac{\partial^2\Phi}{(\partial e_i)^2}\,(t) \;&=\; \lim_{h\to 0}\, \frac{\Phi(t+he_i) - 2\,\Phi(t) + \Phi(t-he_i)}{h^2} \\ &=\; \lim_{h\to 0}\, \int_{\mathbb{R}^d} \frac{\cos(t'\omega + \omega_i h) - 2\cos(t'\omega) + \cos(t'\omega - \omega_i h)}{h^2}\; \nu(d\omega).\end{aligned}$$

Using the identity $\cos(x+y) = \cos(x)\cos(y) - \sin(x)\sin(y)$ it is easily checked that the last expression simplifies to

$$\frac{\partial^2\Phi}{(\partial e_i)^2}\,(t) \;=\; \lim_{h\to 0}\, \int_{\mathbb{R}^d} \cos(t'\omega)\; \underbrace{\frac{\cos(\omega_i h) - 2 + \cos(-\omega_i h)}{h^2}}_{= -I_{hh}(\omega_i)}\; \nu(d\omega)\,,$$

and by the dominated convergence theorem we see that

$$\begin{aligned}\frac{\partial^2\Phi}{(\partial e_i)^2}\,(t) \;&=\; -\int_{\mathbb{R}^d} \cos(t'\omega)\, \left(\lim_{h\to 0}\; I_{hh}(\omega_i)\right)\; \nu(d\omega) \\ &=\; -\int_{\mathbb{R}^d} \cos(t'\omega)\; \omega_i^2\; \nu(d\omega) \;=\; -\int_{\mathbb{R}^d} \cos(t'\omega)\; \nu_{(i)}(d\omega)\,.\end{aligned}$$

By Theorem 5.3.5 (Bochner) $\frac{\partial^2\Phi}{(\partial e_i)^2}$ must be continuous and hence m.s. continuity of $(X_t^{(i)})_{t\in T}$ follows from Corollary 5.3.4.

□

Since $(X_t^{(i)})_{t\in T}$ is itself a second-order random field, it is straightforward to define higher-order m.s. partial derivatives of $(X_t)_{t\in T}$ and to formulate the corresponding counterparts of the above theorems.

The next theorem shows, that if a separable Gaussian RF $(X_t)_{t\in T}$ has k-th order m.s. partial derivatives, and if these have continuous sample paths, then $(X_t)_{t\in T}$ has paths in $\mathcal{C}^k(T)$.

If all $|\alpha|$-th order generalized mixed partial derivatives $D^{\alpha,\alpha}K$ (defined as in Theorem 5.3.10 by repeated application of $D^{i,i}$) exist, we define

$$d_\alpha^2(s,t) \;:=\; D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t) - 2\,D^{\alpha,\alpha}K(s,t).$$

Theorem 5.3.16. *Let $(X_t)_{t\in T}$ be a separable centred Gaussian process on an open subset $T \subset \mathbb{R}^d$ with covariance kernel K. If $D^{\alpha,\alpha}K$ exists for all α with $|\alpha| \leq k$, and if for some $0 < C < \infty$ and $\delta, \eta > 0$ it holds that*

$$d_\alpha^2(s,t) \;\leq\; \frac{C}{\big|\log\|t-s\|\big|^{1+\delta}} \tag{5.5}$$

for all α with $|\alpha| = k$ and for all $s,t \in T$ with $\|t-s\| < \eta$, then the sample paths of $(X_t)_{t\in T}$ are in $\mathcal{C}^k(T)$ a.s.

Corollary 5.3.17. *Let $(X_t)_{t\in T}$ be a separable, stationary centred Gaussian process on an open $T \subset \mathbb{R}^d$ with spectral measure ν. If*

$$\int_{\mathbb{R}^d} \big(\log(1+\|\omega\|)\big)^{1+\delta}\; \|\omega\|^{2k}\; \nu(d\omega) \;<\; \infty \tag{5.6}$$

for some $\delta > 0$, then the sample paths of $(X_t)_{t\in T}$ are in $\mathcal{C}^k(T)$ a.s.

Proof of Theorem 5.3.16: (The proof given here generalizes the idea given in [2, Ch. 1, Sec. 4.3] for $d = 1$ to higher space dimensions)

We state the proof for $k = 1$, the case $k > 1$ is proved by applying the following arguments to separable versions of the k^{th}-order m.s. partial derivatives and repeating the steps of the proof for $k-1,\ldots,1$. Fix $i \in \{1,\ldots,d\}$.

First note that the m.s. partial derivatives $(X_t^{(i)})_{t\in T}$ are themselves Gaussian processes as a result of the stability of Gaussian random variables under passage to the limit. According to Theorem 5.2.9 there exists a separable version $(Y_t^{(i)})_{t\in T}$ of $(X_t^{(i)})_{t\in T}$, and due to (5.5) this version a.s. has continuous paths on T that are bounded on every compact $Q \subset T$.
We show that there exists a set $N_i \in \Omega$ with $P(N_i) = 0$, so that $Y_{\boldsymbol{\cdot}}^{(i)}(\omega)$ is the partial derivative of $X_{\boldsymbol{\cdot}}(\omega)$ for all $\omega \in \Omega \setminus N_i$.

Let $[a,b] := \big\{x \in \mathbb{R}^d \,:\, a_i \leq x_i \leq b_i, \;\text{ for all }\; 1 \leq i \leq d\big\}$ be a closed cuboid in $\mathbb{R}^d$, and define the random field $(Y_t^{[i]})_{t\in[a,b]}$ by

$$Y_t^{[i]}(\omega) := X_{\underline{t}}(\omega) + \int_{a_i}^{t_i} Y^{(i)}_{\underline{t}+(h-a_i)e_i}(\omega)\,dh \qquad t \in [a,b] \tag{5.7}$$

where $\underline{t} := (t_1,\ldots,t_{i-1},a_i,t_{i+1},\ldots,t_d)'$.
Note that the existence of all $D^{i,i}K$, $i = 1,\ldots,d$ implies that $(X_t)_{t\in T}$ has m.s. partial derivatives. Now either (5.5) or the existence of higher order derivatives of K guarantee that these m.s. partial derivatives are m.s. continuous and hence, by Corollary 5.3.12, $(X_t)_{t\in T}$ has continuous sample paths a.s. Since the (marginal) integral function of a continuous function is continuous, it follows that $(Y_t^{[i]})_{t\in[a,b]}$ has continuous paths a.s.

Now, for all $t \in [a, b]$ we have

$$\begin{aligned}\mathbb{E}\left(Y_t^{[i]}\,X_t\right) &= K(\underline{t}, t) + \int_{a_i}^{t_i} \tfrac{\partial K}{\partial_1 e_i}\left(\underline{t} + (h - a_i)e_i, t\right) dh \\ &= K(\underline{t}, t) \;+\; K(t, t) \;-\; K(\underline{t}, t) \quad = \quad K(t, t).\end{aligned}$$

In the same way we can verify that $\mathbb{E}\left((Y_t^{[i]})^2\right) = K(t, t)$, and hence

$$\mathbb{E}\left(\left(Y_t^{[i]} - X_t\right)^2\right) \;=\; \mathbb{E}\left((X_t)^2\right) - 2\,\mathbb{E}\left(Y_t^{[i]}\,X_t\right) + \mathbb{E}\left((Y_t^{[i]})^2\right) \;=\; 0,$$

which implies that $(Y_t^{[i]})_{t\in[a,b]}$ is a version of $(X_t)_{t\in T}$ restricted to $[a, b]$. Since it has continuous sample paths a.s. and $(X_t)_{t\in T}$ was assumed separable, we have, by Lemma 5.2.8, a set $N_{a,b}^i$ of probability 0 so that for all $\omega \in \Omega \setminus N_{a,b}^i$ with

$$X_{\bullet}(\omega) = Y_{\bullet}^{[i]}(\omega) \qquad \text{on } [a, b].$$

Now T can be represented as the countable union of (overlapping) closed cuboids $[a, b]$, so we obtain that the sample paths of $(X_t)_{t\in T}$ have continuous partial derivatives in direction e_i for all $\omega \in \Omega \setminus N_i$ with $P(N_i) = 0$.

Repeating the same argument for all other partial derivatives yields a P-null-set $N = \bigcup_{i=1}^d N_i$ outside of which $X_{\bullet}(\omega)$ is continuously differentiable.

□

In both, Theorem 5.3.16 and Corollary 5.3.17, the existence of the k-th order m.s. partial derivatives is part of the sufficient condition for a.s. sample path differentiability. In Proposition 5.5.5 we show that for stationary Gaussian random fields it is also necessary. The following counterexample however shows, that it is no longer necessary if the assumption of $(X_t)_{t\in T}$ being Gaussian is dropped without substitution.

Example 5.3.18. Let ν be a symmetric probability measure on $(\mathbb{R}^d, \mathbb{B}^d)$.
Define $(X_t)_{t\in T}$ by

$$X_t(\omega) \;:=\; \sqrt{2}\,\cos\left(t'\,\Theta(\omega) + \Psi(\omega)\right), \tag{5.8}$$

where Θ and Ψ are independent RVs with $\Theta \sim \nu$ and $\Psi \sim \mathcal{U}_{[0,2\pi]}$. For the mean function of $(X_t)_{t\in T}$ we have by Fubini's theorem

$$\mathbb{E}(X_t) \;=\; \frac{\sqrt{2}}{2\pi} \int_{\mathbb{R}^d} \underbrace{\int_0^{2\pi} \cos\left(t'\theta + \psi\right) d\psi}_{=0} \;\nu(d\theta) \;=\; 0,$$

for the covariance function we obtain ([43, p. 92])

$$\mathbb{E}(X_s X_t) \;=\; \int_{\mathbb{R}^d} \cos\left((t - s)'\theta\right)\,\nu(d\theta)\,.$$

so $(X_t)_{t\in T}$ is a weakly stationary centred process with spectral measure ν. Hence, by choosing ν accordingly (see Corollary 5.3.13), we can realize any order of m.s. differentiability. However, it follows immediately from (5.8) that $(X_t)_{t\in T}$ has always sample paths in $\mathcal{C}^\infty(\mathbb{R}^d)$.

We illustrate the preceding theorems by applying them to the Whittle-Matérn model for covariance functions introduced in Section 3.2:

Example 5.3.19. (Whittle-Matérn model, part 1)
Consider a separable, stationary centred Gaussian random field $(X_t)_{t\in T}$ on an open set $T \subseteq \mathbb{R}^d$ with spectral density

$$\varphi_\tau(\omega) \;=\; \left(1+\|\omega\|^2\right)^{-\tau}, \qquad \tau > \tfrac{d}{2}.$$

First note that the finiteness of the integral in (5.6) is only an issue of its finiteness on $B_r^c(0) := \mathbb{R}^d \setminus B_r(0)$ for an arbitrary $r > 0$. Now, for every $\epsilon > 0$ there exists an $\eta > 0$ so that

$$\left(\log(1+r)\right)^{1+\delta} \;\leq\; r^\epsilon \qquad \text{for all } r \geq \eta.$$

Hence, we have the inequality

$$\int_{B_\eta^c(0)} \left(\log(1+\|\omega\|)\right)^{1+\delta} \|\omega\|^{2k}\, \nu(d\omega) \;\leq\; \int_{B_\eta^c(0)} \underbrace{\|\omega\|^{2k+\epsilon}\left(1+\|\omega\|^2\right)^{-\tau}}_{\leq \|\omega\|^{2(k-\tau)+\epsilon}} d\omega$$

and it follows that (5.6) holds if $k < \tau - \frac{d+\epsilon}{2}$.

Since ϵ can be chosen arbitrarily small, we can conclude

$$\tau \;>\; k + \tfrac{d}{2} \qquad \Longrightarrow \qquad X_\cdot(\omega) \;\in\; \mathcal{C}^k(T) \quad \text{a.s.}$$

In Section 5.5 we prove a theorem that will finally yield

$$\tau \;\leq\; k + \tfrac{d}{2} \qquad \Longrightarrow \qquad X_\cdot(\omega) \;\notin\; \mathcal{C}^k(T) \quad \text{a.s.}$$

(we will obtain a statement even stronger than that, see Example 5.5.6).

5.4 Measurable Random Fields

In Section 5.2 we have introduced the notion of separability, which turned out to be a suitable means to overcome the problem of non-measurability of events related to path properties. Moreover we have seen that it ensures a certain uniqueness of a random field $(X_t)_{t\in T}$, provided that finite dimensional distributions allow for continuous sample paths. The notion of a measurable random field, which will be introduced in this section, will play a similar role in our discussion of sample paths properties of general second-order processes in Section 5.5. Throughout this (and the subsequent) section, we will always tacitly assume that $T \subseteq \mathbb{R}^d$ is Lebesgue measurable.

Definition 5.4.1. Let $(X_t)_{t\in T}$ be a random field over the probability space $(\Omega, \mathcal{A}, P)$. Let $\mathcal{A} \otimes \mathbb{B}_T^d$ the product σ-algebra of $\mathcal{A}$ and $\mathbb{B}_T^d$, and $\overline{\mathcal{A} \otimes \mathbb{B}_T^d}$ its completion with respect to the measure $P \otimes \lambda^d$. Then $(X_t)_{t\in T}$ is called measurable if it is $\overline{\mathcal{A} \otimes \mathbb{B}_T^d} \,/\, \mathbb{B}$ measurable as a map

$$X : (\Omega \times T) \to \mathbb{R}.$$

It follows from Theorem 2.1.10 that the paths of a measurable random field $(X_t)_{t\in T}$ are $\mathbb{B}^d_T\,/\,\mathbb{B}$ measurable. The following Theorem (stated in [15, Ch. III, §3, Thm. 1] in more generality) gives a condition for the existence of a measurable separable version of $(X_t)_{t\in T}$:

Theorem 5.4.2. *If the random field $(X_t)_{t\in T}$ is stochastically continuous, then there exists a measurable (and separable) version $(Y_t)_{t\in T}$ of $(X_t)_{t\in T}$.*

In the case where $(Y_t)_{t\in T}$ is also required to be separable, it may again assume values in the compact extension $(\bar{\mathbb{R}}, \bar{\mathbb{B}})$ of $(\mathbb{R}, \mathbb{B})$.

Note that the condition that $(X_t)_{t\in T}$ is stochastically continuous is always fulfilled in our case. Indeed, it follows from Remark 5.3.2 and Theorem 5.3.3, that our working assumption that $(X_t)_{t\in T}$ is second-order with continuous mean and covariance function automatically implies stochastic continuity. The following Proposition is a another consequence of this working assumption:

Proposition 5.4.3. *The sample paths of a measurable random field $(X_t)_{t\in T}$ are in $L^2_{\text{loc}}(T)$ a.s. If in addition*

$$\int_T R(t,t)\,dt \;<\; \infty, \tag{5.9}$$

then the sample paths of $(X_t)_{t\in T}$ are in $L^2(T)$ a.s.

Proof: For any compact subset $I \subset T$ Theorem 2.2.11 (Fubini) yields

$$\mathbb{E}\left(\int_I X_t^2\,dt\right) \;=\; \int_I \mathbb{E}\left(X_t^2\right)dt \;=\; \int_I R(t,t)\,dt \;<\; \infty. \tag{5.10}$$

But then necessarily it must hold that

$$P\left(\int_I X_t^2\,dt < \infty\right) = 1, \tag{5.11}$$

which implies that $X_\cdot(\omega) \in L^2_{\text{loc}}(T)$ a.s. If condition (5.9) holds, we obtain the same conclusions in (5.10) and (5.11) for T instead of I.

□

Remark 5.4.4. In the weakly stationary centred case we have $R(t,t) \equiv \Phi(0)$, so condition (5.9) holds if and only if T is bounded.

If $(X_t)_{t\in T}$ is stationary Gaussian then, by Lemma 2.4.11, we have

$$\int_T \mathbb{E}\big(X_t^{2p}\big)\,dt \;=\; vol(T)\,\frac{(2n)!}{2^n n!}\,\big(\Phi(0)\big)^{2p} \;<\; \infty \qquad \text{for any } \; p \in \mathbb{N}.$$

With the same arguments as in (5.10) and (5.11) we conclude that for bounded T and for any $p \in \mathbb{N}$ the paths of $(X_t)_{t\in T}$ are in $L^p(T)$ a.s.

The restriction to measurable random fields entails a certain uniqueness of processes with given finite dimensional distributions:

Lemma 5.4.5. *Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be measurable RFs over the probability space $(\Omega, \mathcal{A}, P)$. If*

$$P\big(\{\omega \,:\, X_t(\omega) = Y_t(\omega)\}\big) \;=\; 1 \qquad \text{for } \lambda^d\text{-almost all } t \in T, \tag{5.12}$$

then there exists a P-null set $N \subset \Omega$ so that

$$\lambda^d\big(\{t \in T \,:\, X_t(\omega) \neq Y_t(\omega)\}\big) = 0 \qquad \text{for all } \;\omega \in \Omega \setminus N, \tag{5.13}$$

i.e. the sample paths of $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ are a.s. identical as functions in $L^2(T)$.

Conversely, if almost all sample paths of $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ differ only on a subset of T with Lebesgue measure 0, then (5.12) must hold.

Proof: Since $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ are measurable processes, so is their difference. Hence the indicator process $\big(\mathbf{1}_{\{X_t\neq Y_t\}}\big)_{t\in T}$ is measurable and by assumption we have

$$\mathbb{E}\big(\mathbf{1}_{\{X_t\neq Y_t\}}\big) \;=\; P\big(X_t \neq Y_t\big) \;=\; 0 \qquad \text{for } \lambda^d\text{-almost all } t \in T.$$

Applying Theorem 2.2.11 (Fubini) yields

$$\mathbb{E}\left(\int_T \mathbf{1}_{\{X_t\neq Y_t\}}\, dt\right) \;=\; \int_T \mathbb{E}\big(\mathbf{1}_{\{X_t\neq Y_t\}}\big)\, dt \;=\; 0,$$

so for all ω outside a set of probability 0 we have

$$\lambda^d\big(\{t \in T \,:\, X_t(\omega) \neq Y_t(\omega)\}\big) \;=\; \int_T \mathbf{1}_{\{X_t(\omega)\neq Y_t(\omega)\}}\, dt \;=\; 0.$$

Reversing the steps of the proof shows the converse implication.

□

We already note a special result on measurability that we will need in a proof in the next section.

Lemma 5.4.6. *Let f be a real-valued function on the product space of the measure spaces $(\Omega, \mathcal{A}, \mu)$ and $([a,b], \mathbb{B}_{[a,b]}, \lambda^1)$, where $a, b \in \mathbb{Q}$, $a < b$. If $f(\cdot, t)$ is $\mathcal{A}/\mathbb{B}$ measurable for every fixed $t \in [a,b]$ and if $f(\omega, \cdot)$ is continuous on $[a,b]$ for every fixed $\omega \in \Omega$, then f is $\mathcal{A}\otimes\mathbb{B}_{[a,b]} \,/\, \mathbb{B}$ measurable.*

Proof: According to Theorem 2.1.7 we must show that for all $M \in \mathbb{R}$

$$G_M \;:=\; \big\{(\omega, t) \in \Omega \times [a,b] \,:\, f(\omega, t) < M\big\} \;\in\; \mathcal{A} \otimes \mathbb{B}_{[a,b]}.$$

First note that

$$G_M \;=\; \bigcup_{t_l, t_u \in \mathbb{Q},\, a \le t_l < t_u \le b} \underbrace{\big\{\omega \,:\, f(\omega, t) < M \quad \forall t \in [t_l, t_u]\big\}}_{=:\, A_{M, t_l, t_u}} \times [t_l, t_u]\,. \tag{5.14}$$

Indeed, since f is continuous in t, $f(\omega, t) < M$ implies $f(\omega, s) < M$ for all s in some interval $[t_l, t_u] \ni t$, where $t_l, t_u \in \mathbb{Q}$, $a \leq t_l < t_u \leq b$. Hence, if $(\omega, t) \in G_M$ it is also contained in the set on the rhs of (5.14). The converse inclusion is obvious.

The rhs of (5.14) is a countable union of sets of the form $A_{M,t_l,t_u} \times B$, $B \in \mathbb{B}_{[a,b]}$ and thus is in $\mathcal{A} \otimes \mathbb{B}_{[a,b]}$ provided that $A_{M,t_l,t_u} \in \mathcal{A}$.

Now for any $t_l, t_u \in \mathbb{Q}$, $a \leq t_l < t_u \leq b$ and any $M \in \mathbb{R}$ by using again that f is continuous in t we obtain

$$\begin{aligned} A_{M,t_l,t_u} &= \bigcup_{m \in \mathbb{Q},\, m<M} \big\{\omega \,:\, f(\omega,t) \leq m \quad \forall t \in [t_l, t_u]\big\} \\ &= \bigcup_{m \in \mathbb{Q},\, m<M} \big\{\omega \,:\, f(\omega,t) \leq m \quad \forall t \in [t_l, t_u] \cap \mathbb{Q}\big\} \\ &= \bigcup_{m \in \mathbb{Q},\, m<M} \; \bigcap_{t \in [t_l,t_u] \cap \mathbb{Q}} \big\{\omega \,:\, f(\omega,t) \leq m\big\}\,. \end{aligned}$$

By assumption, $\big\{\omega \,:\, f(\omega,t) \leq m\big\} \in \mathcal{A}$ for any fixed $t \in [a,b]$ (and any $m \in \mathbb{R}$) and so $A_{M,t_l,t_u} \in \mathcal{A}$ as a countable union and intersection of $\mathcal{A}$ measurable sets and this completes the proof.

□

5.5 Sample Path Regularity in the General Case

We have seen in Section 5.3 that m.s. continuity and m.s. differentiability are linked to the probabilistic structure of a random field $(X_t)_{t \in T}$ only through the covariance function. In contrast to that, the above results on a.s. sample path continuity and a.s. sample path differentiability were formulated for the special case of a Gaussian random field, and we shall give an example that shows, that the above theorems indeed do not hold in the general case.

It will however turn out, that m.s. differentiability implies a.s. weak differentiability (as defined in Section 3.2) of the sample paths, whatever the particular distribution of $(X_t)_{t \in T}$.

Example 5.5.1. Let $K(s,t) = e^{-\|t-s\|}$ be the so-called exponential covariance function on $T = \mathbb{R}^d$. It easily verified that condition (5.3) holds e.g. for $\delta = 1$ and so a separable centred Gaussian random field with exponential covariance function has continuous sample paths a.s.

Now define a process $(X_t)_{t \in T}$ with the same covariance function as follows:

The starting point is a homogeneous Poisson point process on $\mathbb{R}^d$ with intensity 1. This is a stochastic process that assigns to each $\omega \in \Omega$ a countable set of points $\{\zeta_1(\omega), \zeta_2(\omega), \ldots\} \subset \mathbb{R}^d$ with the following properties (see e.g. [25, Sec. 11.1])

1. The number $N(B)$ of points inside a set $B \in \mathbb{B}^d$ is a Poisson RV with parameter $vol(B)$, i.e.
$$P\big(N(B) = k\big) \;=\; e^{-vol(B)}\,\frac{vol(B)^k}{k!}, \qquad \text{for all} \;\; k \in \mathbb{N}_0. \qquad (5.15)$$
2. If $B_1, \ldots, B_m \in \mathbb{B}^d$ are pairwise disjoint, then the RVs $N(B_1), \ldots, N(B_m)$ are mutually independent.

For simplicity we take $d = 1$ (an $\mathbb{R}^d$-counterpart of this example can be constructed via Poisson tessellation, see [25, Sec. 12.3]). Then we can relabel the random point sets $(\zeta_n)_{n\in\mathbb{N}}$ to $(\zeta_n)_{n\in\mathbb{Z}}$ such that $\zeta_i(\omega) \leq \zeta_{i+1}(\omega)$ for all $i \in \mathbb{Z}$ and all $\omega \in \Omega$.

Now let $(U_i)_{i\in\mathbb{Z}} \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ and independent of $(\zeta_i)_{i\in\mathbb{Z}}$ and define

$$X_t \;=\; U_n \quad \text{for} \;\; t \in [\zeta_n, \zeta_{n+1})$$

Since the $(U_i)_{i\in\mathbb{Z}}$ are centred so is $(X_t)_{t\in T}$. By (5.15) and and the independence of $(U_i)_{i\in\mathbb{Z}}$ and $(\zeta_n)_{n\in\mathbb{Z}}$ we get

$$\mathbb{E}\big(X_s X_t\big) \;=\; \mathbb{E}\big(U_1^2\big) \cdot P\big(\exists\, n \in \mathbb{Z} \,:\, s,t \in [\zeta_n, \zeta_{n+1})\big) \;=\; 1 \cdot P\big(N([s,t]) = 0\big) \;=\; e^{-|t-s|}.$$

Hence, $(X_t)_{t\in T}$ has indeed the prescribed covariance function, but obviously does not have continuous sample paths. To complete this counter-example it remains to show that $(X_t)_{t\in T}$ is separable.

Let I be a closed, B an open, and D an arbitrary dense subset of $\mathbb{R}$. With the notation from Definition 5.2.4) we have

$$A_{B,I} \;=\; \big\{\omega \,:\, U_n(\omega) \in B \quad \forall\, n \in \mathbb{Z} \;\; \text{with} \;\; \big[\zeta_n(\omega), \zeta_{n+1}(\omega)\big) \cap I \neq \emptyset\big\}.$$

Since I is open, we have the implication

$$\big[\zeta_n(\omega), \zeta_{n+1}(\omega)\big) \cap I \neq \emptyset \quad \Longrightarrow \quad \big(\zeta_n(\omega), \zeta_{n+1}(\omega)\big) \cap I \neq \emptyset.$$

Consequently, both intersections must also contain points from D, so it follows that $A_{B,I}$ and $A_{B,I\cap D}$ coincide, which proves separability and shows that it is really the fact that $(X_t)_{t\in T}$ is not Gaussian that causes Theorem 5.3.6 to fail. (Note that the univariate marginal distributions are Gaussian, but the multivariate ones are not multivariate Gaussian).

Having shown that Theorem 5.3.6 does no longer hold if we just drop the assumption that $(X_t)_{t\in T}$ is Gaussian, we might ask for other criteria that ensure a.s. sample path continuity in the general case. A condition for weakly stationary random fields on $\mathbb{R}^d$ is derived in [22]:

Denote by $p_d(h)$ the polynomial of degree d given by the Taylor series expansion of $\Phi(h)$ around $h = 0$. If $\Phi \in \mathcal{C}^d(\mathbb{R}^d)$ and if for some $0 < C < \infty$ and $\delta > 0$

$$\big|\Phi(h) - p_d(h)\big| \;\leq\; \frac{C\,\|h\|^d}{(-\log\|h\|)^{3+\delta}} \tag{5.16}$$

for all $\|h\|$ small enough, then the sample paths of $(X_t)_{t\in T}$ are continuous a.s.

As can be expected, the loss of information about the distribution of the random field $(X_t)_{t\in T}$ must be compensated for by requiring much more smoothness from the covariance kernel than in Theorem 5.3.7.

It is therefore quite remarkable that if one settles for weak differentiability, nothing is lost compared to the Gaussian case, as it turns out that this kind of regularity is completely determined by the second-order structure:

Theorem 5.5.2. *Let $(X_t)_{t\in T}$ be a measurable centred random field on an open subset $T \subseteq \mathbb{R}^d$ with covariance function K. If $D^{\alpha,\alpha}K$ exists and is continuous on the diagonal of $T \times T$ for all $|\alpha| \leq k$, then the sample paths of $(X_t)_{t\in T}$ are in $W^{k,2}_{\text{loc}}(T)$ a.s. If in addition*

$$\int_T D^{\alpha,\alpha}K(t,t)\,dt \;<\; \infty \qquad \text{for all } \alpha \text{ with } \quad 0 \leq |\alpha| \leq k, \tag{5.17}$$

then the sample paths of $(X_t)_{t\in T}$ are in $W^{k,2}(T)$ a.s.

In the stationary case we can again give a criterion in terms of the spectral measure:

Corollary 5.5.3. *Let $(X_t)_{t\in T}$ be a measurable, stationary centred random field on an open subset $T \subseteq \mathbb{R}^d$ with spectral measure ν. If*

$$\int_{\mathbb{R}^d} \|\omega\|^{2k}\,\nu(d\omega) < \infty, \tag{5.18}$$

then the sample paths of $(X_t)_{t\in T}$ are in $W^{k,2}_{\text{loc}}(T)$ a.s.

If in addition T is bounded, then the sample paths of $(X_t)_{t\in T}$ are in $W^{k,2}(T)$ a.s.

The idea of the proof is similar to the one in the proof of Theorem 5.3.16. For the definition of the marginal integrals in (5.7) we only need integrability of $(Y_t^{(i)})_{t\in Q}$. The problem is then, however, that the integrated random field $(Y_t^{[i]})_{t\in Q}$ is not continuous in general, and we must use another argument to show that its sample paths coincide a.e. with those of $(X_t)_{t\in Q}$. Another problem arising from the lack of continuity of $(Y_t^{[i]})_{t\in Q}$ is that more caution is needed when patching together these RFs to a RF defined on the whole of T. In order to obtain the statement for any open subset $T \subseteq \mathbb{R}^d$, we therefore need the following technical lemma (with the notation of Section 3.2):

Lemma 5.5.4. *For any open subset $T \subseteq \mathbb{R}^d$ and $i \in \{1, \dots, d\}$ there exists a sequence $(Q_n)_{n\in\mathbb{N}}$ of bounded measurable subsets of T and a sequence $(s_n)_{n\in\mathbb{N}}$ of real numbers $s_n \in \pi_i(Q_n)$ with the following properties:*

(a) $T = \bigcup_{n\in\mathbb{N}} Q_n$.

(b) $Q_n = Q_n^0 \times_i [a_n, b_n], \qquad Q_n^0 \in \mathbb{B}^{d-1}, \quad a_n, b_n \in \mathbb{Q}, \quad n \in \mathbb{N}.$
i.e. each Q_n is cylindrical in the direction of the i-th coordinate axis

(c) If the points $t_j \in Q_j$ and $t_k \in Q_k$, $j < k$, are the endpoints of a line segment $l \subseteq T$ that is parallel to the i-th coordinate axis, then $l \subseteq Q_k$ and $s_k = s_j$.

Proof: First we construct a sequence $(\tilde{Q}_n)_{n\in\mathbb{N}}$ of cubes with $T = \bigcup_{n\in\mathbb{N}} \tilde{Q}_n$ such that

$$j < k \quad \Longrightarrow \quad \pi_i(\tilde{Q}_k) \subset \pi_i(\tilde{Q}_j) \quad \text{or} \quad \pi_i(\tilde{Q}_j) \cap \pi_i(\tilde{Q}_k) = \emptyset.$$

Such a sequence is obtained as follows:

We consider cubes of the form $\tilde{Q} = \tilde{Q}^0 \times_i [\tilde{a}, \tilde{b}]$, $\tilde{a}, \tilde{b} \in \mathbb{Q}$, $\tilde{a} < \tilde{b}$, where $\tilde{Q}^0$ is a right half-open (see Example 2.1.2) cube in $\mathbb{R}^{d-1}$ with edge length $\tilde{b} - \tilde{a}$. Denote by $\mathcal{Q}_\rho$ the set of all cubes of that type with edge length ρ and corners on the grid $\rho\,\mathbb{Z}^d$. For $m \in \mathbb{N}$ define

$$\mathcal{T}_m := \Big\{ Q \in \mathcal{Q}_{2^{-m}} \,:\, Q \subset T \cap [-m, m]^d,\ \lambda^d(Q \cap \bar{Q}) = 0 \ \text{ for all } \bar{Q} \in \textstyle\bigcup_{j<m} \mathcal{T}_j \Big\}.$$

Every $\mathcal{T}_m$ consists of finitely many cubes and by enumerating the set $\bigcup_{m\in\mathbb{N}} \mathcal{T}_m$ starting with all cubes in $\mathcal{T}_1$ and continuing with $\mathcal{T}_2, \mathcal{T}_3, \ldots$, we obtain a sequence $(\tilde{Q}_n)_{n\in\mathbb{N}}$ with the prescribed properties from above. We write $\tilde{Q}_n = \tilde{Q}_n^0 \times_i [\tilde{a}_n, \tilde{b}_n]$.

We will now modify the sequence $(\tilde{Q}_n)_{n\in N}$ to obtain a sequence $(Q_n)_{n\in N}$ and a sequence $(s_n)_{n\in N}$ of real numbers with the prescribed properties (a) - (c).

Start with $Q_1 := \tilde{Q}_1$ and $s_1 := \tilde{a}_1$, then (b) and (c) trivially hold so far. Assume now that these properties hold for the sets $Q_1, \ldots, Q_n$ and points $s_1, \ldots, s_n$ that have been constructed from $\tilde{Q}_1, \ldots, \tilde{Q}_{\tilde{n}}$, and assume that in addition $\bigcup_{j=1}^{\tilde{n}} \tilde{Q}_j \subset \bigcup_{j=1}^{n} Q_j$. For all $t \in \tilde{Q}_{\tilde{n}+1}$ define the open line segment

$$l_{\tilde{n}+1,t} \ := \ \big\{ s \in T \,:\, s = t + \gamma\, e_i,\ \gamma \in \mathbb{R} \ \text{ and } \ \eta s + (1-\eta)t \in T \ \ \forall \eta \in [0,1] \big\}$$

which is parallel to the i^{th} coordinate axis and entirely contained in T. Further let

$$\alpha_{\tilde{n}+1,t} \ := \ \pi_i \Big(\min \Big\{ s \in l_{\tilde{n}+1,t} \,\cap\, \textstyle\bigcup_{j=1}^{\tilde{n}+1} \tilde{Q}_j \Big\} \Big)$$

and

$$\beta_{\tilde{n}+1,t} \ := \ \pi_i \Big(\max \Big\{ s \in l_{\tilde{n}+1,t} \,\cap\, \textstyle\bigcup_{j=1}^{\tilde{n}+1} \tilde{Q}_j \Big\} \Big)$$

the minimal (maximal) value of the i^{th} coordinate of all points from $l_{\tilde{n}+1,t}$ that are contained in the union of $\tilde{Q}_1, \ldots, \tilde{Q}_{\tilde{n}+1}$. For any $t \in \tilde{Q}_{\tilde{n}+1}$ we have one of the following alternatives:

(i) $\alpha_{\tilde{n}+1,t} = \tilde{a}_j$ and $\beta_{\tilde{n}+1,t} = \tilde{b}_k$ for some $1 \le j, k \le \tilde{n}$, $j \ne k$,

(ii) $\alpha_{\tilde{n}+1,t} = \tilde{a}_j$ and $\beta_{\tilde{n}+1,t} = \tilde{b}_{\tilde{n}+1}$ for some $1 \le j \le \tilde{n}$,

(iii) $\alpha_{\tilde{n}+1,t} = \tilde{a}_{\tilde{n}+1}$ and $\beta_{\tilde{n}+1,t} = \tilde{b}_k$ for some $1 \le k \le \tilde{n}$, or

(iv) $\alpha_{\tilde{n}+1,t} = \tilde{a}_{\tilde{n}+1}$ and $\beta_{\tilde{n}+1,t} = \tilde{b}_{\tilde{n}+1}$.

This allows to construct disjoint sets $Q_{n+1}, \ldots, Q_{n+p}$ by setting

$$Q_{n+r} \ := \ \overline{\pi}_i\big(\big\{ t \in \tilde{Q}_{\tilde{n}+1} \,:\, \alpha_{\tilde{n}+1,t} = \tilde{a}_{j_r},\ \beta_{\tilde{n}+1,t} = \tilde{b}_{k_r} \big\} \big) \times_i [\tilde{a}_{j_r}, \tilde{b}_{k_r}] \tag{5.19}$$

for $r = 1, \ldots, p$, where $(j_1, k_1), \ldots, (j_r, k_r)$ is an enumeration of all pairs of indices $1 \le j, k \le \tilde{n} + 1$, for which one of the above alternatives applies.

Clearly $\tilde{a}_{j_r} \le \tilde{a}_{\tilde{n}+1} < \tilde{b}_{\tilde{n}+1} \le \tilde{b}_{k_r}$ for all $r = 1, \ldots, p$, and hence $\tilde{Q}_{\tilde{n}+1} \subset \bigcup_{r=1}^{p} Q_{n+r}$. Consequently, property (a) follows from $T = \bigcup_{n\in\mathbb{N}} \tilde{Q}_n$ by induction.

Note that this conclusion still holds if we omit all sets Q_{n+r} that correspond to alternative (i) in the construction of $Q_{n+1}, \ldots, Q_{n+p}$. Indeed, for any $t \in \tilde{Q}_{\tilde{n}+1}$ so that $l_{\tilde{n}+1,t}$ intersects both

$\tilde{Q}_j$ and $\tilde{Q}_k$ with j,k as for (i), there must already be a set of the form (5.19), constructed in an earlier step by alternative (ii) or (iii), that contains the points $\overline{\pi}_i(t) \times_i [\tilde{a}_j, \tilde{b}_k)$.

We assume therefore that each $Q_{n+1}, \ldots, Q_{n+p}$ corresponds to (ii), (iii) or (iv). Then we can relabel these sets in such a way that for some $0 \leq u \leq v \leq p$ we have

$$\begin{array}{lll} \tilde{a}_{j_1} < \ldots < \tilde{a}_{j_u} < \tilde{a}_{\tilde{n}+1}, & \tilde{a}_{j_{u+1}} = \ldots = \tilde{a}_{j_v} = \tilde{a}_{\tilde{n}+1}, & \text{and} \\ \tilde{b}_{k_1} = \ldots = \tilde{b}_{k_u} = \tilde{b}_{\tilde{n}+1}, & \tilde{b}_{k_{u+1}} > \ldots > \tilde{b}_{k_v} > \tilde{b}_{\tilde{n}+1}. & \end{array}$$

We use this to prove that $\overline{\pi}_i\big(Q_{n+1}\big), \ldots, \overline{\pi}_i\big(Q_{n+p}\big)$ are measurable. First, for $1 \leq r \leq u$, we have $y \in \overline{\pi}_i(Q_{n+r})$ if and only if

- $y \in \overline{\pi}_i\big(\tilde{Q}_{j_r}\big) \cap \overline{\pi}_i\big(\tilde{Q}_{\tilde{n}+1}\big)$,
- $y \notin \overline{\pi}_i\big(\tilde{Q}_{n+s}\big)$ for all $1 \leq s < r$, and
- $y \in \bigcap_{q \in [\tilde{a}_{j_r}, \tilde{b}_{\tilde{n}+1}]} T_i(q) \ := \ J$,

where $T_i(\cdot)$ is a the cross-section of T orthogonal to the i-th coordinate axis, i.e.

$$T_i(q) \ := \ \{\tau \in \mathbb{R}^{d-1} \ : \ (\tau_1, \ldots, \tau_{i-1}, q, \tau_{i+1}, \ldots, \tau_d) \in T\}.$$

Hence, measurability of $\overline{\pi}_i(Q_{n+r})$ follows if we can show measurability of J.

If $y \in J$, then the line segment $l \ := \ y \times_i [\tilde{a}_{j_r}, \tilde{b}_{\tilde{n}+1}]$ is completely contained in T. Now l is compact, and so its distance to the boundary ∂T of T assumes its minimum at some point $s \in l$. Since T is open we must have $\delta := \mathrm{dist}(s, \partial T) > 0$ which implies $B_\delta(y) \subset J$. Hence, J is open and in particular $J \in \mathbb{B}^{d-1}$.

The same argument can be used to prove $\overline{\pi}_i(Q_{n+r}) \in \mathbb{B}^{d-1}$ for $u+1 \leq r \leq v$. Finally, if $v < p$, Q_{n+p} corresponds to alternative (iv) and we have

$$\overline{\pi}_i(Q_{n+p}) \ = \ \overline{\pi}_i\big(\tilde{Q}_{\tilde{n}+1}\big) \setminus \bigcup_{s=1}^{p-1} \overline{\pi}_i\big(\tilde{Q}_{n+s}\big) \quad \in \mathbb{B}^{d-1},$$

which concludes the verification of property (b).

Property (c) of $(Q_n)_{n \in N}$ is an obvious consequence of its construction, so it only remains to provide a suitable choice of s_{n+r}, $r = 1, \ldots, p$. If Q_{n+r} corresponds to alternative (iv) we can simply set $s_{n+r} := \tilde{a}_{\tilde{n}+1}$. Otherwise, it necessarily holds that

$$\overline{\pi}_i(Q_{n+r}) \ \subset \ \overline{\pi}_i(Q_j) \qquad \text{for some } j \leq n,$$

so we find that $s_{n+r} := s_j$ is a suitable choice that respects property (c) and this completes the proof.

□

Proof of Theorem 5.5.2: We state the proof for $k = 1$, the case $k > 1$ is obtained by applying the steps of the proof recursively.

Fix $i \in \{1, \ldots, d\}$. Our assumptions imply the existence of a m.s. partial derivative $(X_t^{(i)})_{t \in T}$

of $(X_t)_{t\in T}$, and a measurable version $(Y_t^{(i)})_{t\in T}$ of it. We show that $Y_{\bullet}^{(i)}(\omega)$ is a weak partial derivative of $X_{\bullet}(\omega)$ for almost every ω.

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

$$Q_n \;=\; Q_n^0 \times_i [a_n, b_n]$$

with a_n, b_n, and Q_n^0 from Lemma 5.5.4. Since $Q_n \subset\subset T$ and K is continuous on $T \times T$ we have

$$\int_{Q_n} \mathbb{E}\left(\big(Y_t^{(i)}\big)^2 \right) dt \;=\; \int_{Q_n} D^{i,i}K(t,t)\, dt \;<\; \infty.$$

Hence, by Fubini's theorem there exists a $P\otimes\lambda^{d-1}$-null set $N_{i,n}^0 \subset \Omega \times Q_n^0$ so that the marginal integral

$$\int_{a_n}^{b_n} Y^{(i)}_{\underline{t}+(h-s_n)\,e_i}(\omega)\; dh, \qquad \underline{t} \;:=\; t^0 \times_i \{s_n\},$$

exists for all $(\omega, t^0) \in \big(\Omega \times Q_n^0\big) \setminus N_{i,n}^0$. Then $N_{i,n} := N_{i,n}^0 \times_i [a_n, b_n]$ is a $P \otimes \lambda^d$-null set and, denoting by $\underline{t}$ the orthogonal projection of $t \in Q_n$ on $Q_n^0 \times_i \{s_n\}$, we can define for all $(\omega, t) \in \big(\Omega \times Q_n\big) \setminus N_{i,n}$

$$Y_{n,t}^{[i]}(\omega) \;:=\; X_{\underline{t}}(\omega) \;+\; \int_{s_n}^{t_i} Y^{(i)}_{\underline{t}+(h-s_n)\,e_i}(\omega)\; dh. \tag{5.20}$$

For $(\omega, t) \in N_{i,n}$ we set $Y_{n,t}^{[i]}(\omega) := X_{\underline{t}}(\omega)$. By Lemma 2.1.10 $X_{\underline{\bullet}}(\cdot)$ is $P \otimes \lambda^{d-1}$ measurable and for every fixed $t_i \in [a_n, b_n]$ Fubini's theorem implies that the second term of the rhs of (5.20), set to 0 on $N_{i,n}$, is $P \otimes \lambda^{d-1}$ measurable as well. For every fixed $(\omega, t^0) \in \big(\Omega \times Q_n^0\big)$ $Y_{n,t}^{[i]}(\omega)$ is continuous as a function of t_i and hence, by Lemma 5.4.6, $(Y_{n,t}^{[i]})_{t\in Q_n}$ is measurable.

Now repeat this construction for all Q_n, $n \in \mathbb{N}$, and set

$$N_i^0 \;:=\; \textstyle\bigcup_{n\in\mathbb{N}} N_{i,n}^0, \qquad N_i \;:=\; \big(N_i^0 \times_i \mathbb{R}\big) \cap T.$$

Let $(\omega, t) \in (\Omega \times T) \setminus N_i$ and assume that $t \in Q_j \cap Q_k$, $j < k$. Then it follows from (c) in Lemma 5.5.4 that $s_j \in Q_k$ and $s_k = s_j$, so the rhs of (5.20) for j and k coincide and we have

$$Y_{j,t}^{[i]}(\omega) \;=\; Y_{k,t}^{[i]}(\omega).$$

Moreover we have $T = \bigcup_{n\in\mathbb{N}} Q_n$ and so the random field $(Y_t^{[i]})_{t\in T}$ is well-defined by

$$Y_t^{[i]}(\omega) \;:=\; \begin{cases} Y_{n,t}^{[i]}(\omega) & \text{if } \;(\omega, t) \in (\Omega \times T) \setminus N_i \;\text{ and }\; t \in Q_n \\ 0 & \text{if } \;(\omega, t) \in N_i \end{cases}$$

An alternative representation of $(Y_t^{[i]})_{t\in T}$ is given by

$$Y_t^{[i]}(\omega) \;=\; \mathbf{1}_{(\Omega\times T)\setminus N_i}(\omega, t) \;\cdot\; \sup_{n\in\mathbb{N}} Y_{n,t}^{[i]}(\omega).$$

and from this and Theorem 2.1.8 and 2.1.9, it follows that $(Y_t^{[i]})_{t\in T}$ is measurable.

Next, for some $(\omega, t^0) \in \big(\Omega \times \overline{\pi}_i(T)\big) \setminus N_i^0$ let l be an arbitrary closed line segment on the line $t^0 \times_i \mathbb{R}$ that is completely contained in T. By (c) in Lemma 5.5.4 there exists a set Q_k so that $l \subset Q_k$, and so it immediately follows from (5.20) that $Y_{\boldsymbol{\cdot}}^{[i]}(\omega)$ is absolutely continuous on l. Since l was arbitrary, $Y_{\boldsymbol{\cdot}}^{[i]}(\omega)$ is absolutely continuous on the line $t^0 \times_i \mathbb{R}$ by Definition 3.2.7. Now $\big(P \otimes \lambda^{d-1}\big)(N_i^0) = 0$ implies

$$P\big(\big\{\omega \,:\, \lambda^{d-1}(N_i^0(\omega)) > 0\big\}\big) \;=\; 0,$$

where $N_i^0(\omega) := \{t^0 \in \overline{\pi}_i(T) \,:\, (\omega, t^0) \in N_i^0\}$ denotes the N_i^0 cross section for fixed ω, and altogether we conclude that

$$Y_{\boldsymbol{\cdot}}^{[i]}(\omega) \;\in\; \mathrm{AC}_i(T) \qquad \text{for almost every } \omega \in \Omega, \tag{5.21}$$

with "classical" partial derivative $Y_{\boldsymbol{\cdot}}^{(i)}(\omega)$ (defined a.e. on T).

Finally we show that the sample paths of $(Y_t^{[i]})_{t \in T}$ and $(X_t)_{t \in T}$ are a.s. identical in $L^2(T)$. Note that $\big(P \otimes \lambda^d\big)(N_i) = 0$ and consequently

$$\lambda^d\big(\big\{t \in T \,:\, P(N_i(t)) > 0\big\}\big) \;=\; 0$$

where $N_i(t) := \{\omega : (\omega, t) \in N_i\}$ denotes the N_i cross section for fixed t. This means that for λ^d-almost every $t \in T$, $Y_t^{[i]}$ is a.s. defined according to (5.20).

But then, using the same notation as above, we have for almost every $t \in T$

$$\begin{aligned}
\mathbb{E}\left(Y_t^{[i]}\, X_t\right) \;&=\; K(\underline{t}, t) + \int_{s_n}^{t_i} \frac{\partial K}{\partial_1 e_i}\big(\underline{t} + (h - s_n)e_i, t\big)\, dh \\
&=\; K(\underline{t}, t) \;+\; K(t,t) \;-\; K(\underline{t}, t) \;\;=\;\; K(t,t) \qquad \text{and}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}\left(Y_t^{[i]}\, Y_t^{[i]}\right) \;&=\; K(\underline{t}, \underline{t}) + 2\int_{s_n}^{t_i} \frac{\partial K}{\partial_1 e_i}\big(\underline{t} + (h - s_n)e_i, \underline{t}\big)\, dh \\
&\quad + \int_{s_n}^{t_i}\int_{s_n}^{t_i} D^{i,i}K\big(\underline{t} + (h - s_n)e_i, \underline{t} + (h' - s_n)e_i\big)\, dh'\, dh \\
&=\; K(\underline{t}, \underline{t}) + \int_{s_n}^{t_i} \frac{\partial K}{\partial_1 e_i}\big(\underline{t} + (h - s_n)e_i, t\big)\, dh \\
&\quad + \int_{s_n}^{t_i} \frac{\partial K}{\partial_1 e_i}\big(\underline{t} + (h - s_n)e_i, \underline{t}\big)\, dh \\
&=\; K(\underline{t}, \underline{t}) + K(t,t) - K(\underline{t}, t) + K(t, \underline{t}) - K(\underline{t}, \underline{t}) \;\;=\;\; K(t,t).
\end{aligned}$$

Putting both together we have

$$\mathbb{E}\left(\big(Y_t^{[i]} - X_t\big)^2\right) \;=\; \mathbb{E}\left((X_t)^2\right) - 2\,\mathbb{E}\left(Y_t^{[i]}\, X_t\right) + \mathbb{E}\left((Y_t^{[i]})^2\right) \;=\; 0,$$

and therefore

$$P\left(\left\{\omega : X_t(\omega) = Y_t^{[i]}(\omega)\right\}\right) = 1 \quad \text{for } \lambda^d\text{-almost all } t \in T.$$

But then it follows from Lemma 5.4.5 that the sample paths of $(Y_t^{[i]})_{t\in T}$ and $(X_t)_{t\in T}$ are indeed a.s. identical in $L^2(T)$. By Proposition 5.4.3 this implies in particular

$$Y_{\bullet}^{[i]}(\omega) \in L^2_{\text{loc}}(T) \quad \text{for almost every } \omega \in \Omega.$$

and from this and (5.21) we obtain from Lemma 3.2.9 that $Y_{\bullet}^{(i)}(\omega)$ is a.s. an i^{th} weak derivative of $Y_{\bullet}^{[i]}(\omega)$ and hence also of $(X_t)_{t\in T}$.

Repeating these arguments for all m.s. partial derivatives $(X_t^{(1)})_{t\in T}, \ldots, (X_t^{(d)})_{t\in T}$, completes the proof.

□

Theorem 5.5.2 shows that in order for a random field $(X_t)_{t\in T}$ to have weakly differentiable sample paths it is sufficient that it is m.s. differentiable, and that the m.s. partial derivatives are m.s. continuous. If $(X_t)_{t\in T}$ is stationary and Gaussian, then we can prove that this is also necessary:

Proposition 5.5.5. *Let $(X_t)_{t\in T}$ be a measurable centred stationary Gaussian process on an open subset $T \subseteq \mathbb{R}^d$. If $(X_t)_{t\in T}$ does not have m.s. partial derivatives of order k, then its sample paths a.s. do not have weak derivatives of order k.*

Proof: We give the proof for $k = 1$, the case $k > 1$ follows in the same way.

For some sequence $(h_n)_{n\in\mathbb{N}}$ of real numbers with $\lim\limits_{n\to\infty} h_n = 0$ define the set

$$A \;:=\; \left\{(\omega, t) \in \Omega \times T \;:\; \limsup_{n\to\infty} \left|X_t^{(i,h_n)}(\omega)\right| < \infty\right\}.$$

and denote its the cross sections by

$$A_\omega \;:=\; \left\{t \in T \,:\, (\omega, t) \in A\right\} \quad \text{and} \quad A_t \;:=\; \left\{\omega \in \Omega \,:\, (\omega, t) \in A\right\}.$$

Note that the fact that $(X_t)_{t\in T}$ is a $\overline{\mathcal{A} \otimes \mathbb{B}_T} \,/\, \mathbb{B}$ measurable random field implies that A is an $\overline{\mathcal{A} \otimes \mathbb{B}_T}$ measurable set (see Theorem 2.1.9).

If $(X_t)_{t\in T}$ does not have a m.s. partial derivative at t in the direction e_i (which can hold for either no or all $t \in T$, see Corollary 5.3.13), then it follows from Lemma 5.3.14 that $(h_n)_{n\in\mathbb{N}}$ can be chosen such that

$$\lim_{n\to\infty} \Phi^{(i)}_{h_n h_n}(0) \;=\; \infty$$

Now for the Gaussian density function φ_{μ,σ^2} (see Example 2.4.8) we have

$$\varphi_{\mu,\sigma^2}(x) \;\leq\; \varphi_{\mu,\sigma^2}(\mu) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}},$$

and from that we obtain for any fixed $t \in T$ and any $M < \infty$

$$P\left(\left|X_t^{(i,h_n)}\right| \leq M\right) \;\leq\; \frac{2\,M}{\sqrt{2\pi\,\Phi_{h_n h_n}^{(i)}(0)}} \;\longrightarrow\; 0, \qquad \text{as } \; n \to \infty.$$

We can therefore choose a subsequence $(n_k)_{k\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ so that

$$P\left(\left|X_t^{(i,h_n)}\right| \leq k\right) \;\leq\; 2^{-k} \qquad \text{for all } \; n \geq n_k.$$

Then

$$\sum_{k=1}^{\infty} P\left(\left|X_t^{(i,h_{n_k})}\right| \leq k\right) \;<\; \infty,$$

and so Lemma 2.1.16 (Borel-Cantelli) yields

$$P\left(\bigcap_{m\in N}\bigcup_{k>m}\left\{\omega \,:\, \left|X_t^{(i,h_{n_k})}(\omega)\right| \leq k\right\}\right) \;=\; 0.$$

This means that for a.e. ω and every $m \in \mathbb{N}$ there exists a $k = k(\omega) > m$ so that $\left|X_t^{(i,h_{n_k})}(\omega)\right| > k$, so we have $P(A_t) = 0$ for any $t \in T$ and hence, by Fubini's theorem

$$\left(P \otimes \lambda^d\right)(A) \;=\; 0. \tag{5.22}$$

Now let $W \subset \Omega$ the set of all ω for which $X_{\bullet}(\omega)$ has an i^{th} weak derivative. For every such ω, according to Theorem 3.2.10, there exists a function $Y_{\bullet}(\omega) \in \mathrm{AC}_i(T)$ that coincides with $X_{\bullet}(\omega)$ λ^d-a.e. on T. Hence, the following sets

$$\begin{aligned} I_\infty(\omega) \;&:=\; \left\{t \in T \,:\, \limsup_{n\to\infty}\left|Y_t^{(i,h_n)}(\omega)\right| = \infty\right\}, \\ I_{\neq,0}(\omega) \;&:=\; \left\{t \in T \,:\, Y_t(\omega) \neq X_t(\omega)\right\}, \qquad \text{and} \\ I_{\neq,n}(\omega) \;&:=\; \left\{t \in T \,:\, t + h_n e_i \in T \;\text{ and }\; Y_{t+h_n e_i}(\omega) \neq X_{t+h_n e_i}(\omega)\right\} \end{aligned}$$

are all λ^d-null sets, and so is the set

$$I(\omega) \;:=\; I_\infty(\omega) \,\cup\, \bigcup_{n=0}^{\infty} I_{\neq,n}(\omega)$$

For all $t \in T \setminus I(\omega)$ it holds that

$$\limsup_{n\to\infty}\left|X_t^{(i,h_n)}(\omega)\right| \;=\; \limsup_{n\to\infty}\left|Y_t^{(i,h_n)}(\omega)\right| \;<\; \infty.$$

so we have $T \setminus A_\omega \subset I(\omega)$ and therefore $\lambda^d(A_\omega) = vol(T)$ for all $\omega \in W$. But then, using (5.22) we get

$$0 \;=\; \left(P \otimes \lambda^d\right)(A) \;\geq\; \left(P \otimes \lambda^d\right)\left((W \times T) \cap A\right) \;\overset{\text{Fubini}}{=}\; P(W) \cdot vol(T)$$

and it follows that $P(W) = 0$ which completes the proof.

□

The assumption that $(X_t)_{t\in T}$ is Gaussian was only needed for the implication

$$\lim_{n\to\infty} \Phi^{(i)}_{h_n h_n}(0) \;=\; \infty \quad\Longrightarrow\quad \lim_{n\to\infty} P\left(\left|X_t^{(i,h_n)}\right| \le M\right) \;=\; 0 \tag{5.23}$$

for any $M < \infty$. The statement of Proposition 5.5.5 should therefore hold under much weaker assumptions (we could e.g. replace the assumption of Gaussianity by the requirement that (5.23) must hold). However, as we can again see from Example 5.3.18, it is not possible to drop this assumption without any substitution.

We continue Example 5.3.19 and illustrate the preceding Theorems with the Whittle-Matérn model:

Example 5.5.6. (Whittle-Matérn model, part 2)
Consider a measurable, (weakly) stationary centred random field $(X_t)_{t\in T}$ on an open and bounded domain $T \subset \mathbb{R}^d$ with spectral density

$$\varphi_\tau(\omega) \;=\; \left(1 + \|\omega\|^2\right)^{-\tau}, \qquad \tau > \tfrac{d}{2}.$$

As in Example 5.3.19 we may compute the integral over $B_r^c(0) := \mathbb{R}^d \setminus B_r(0)$ only rather than $\mathbb{R}^d$ in condition (5.18). For $k < \tau - \frac{d}{2}$ we have

$$\int_{B_r^c(0)} \|\omega\|^{2k}\, \nu(d\omega) \;=\; \int_{B_r^c(0)} \underbrace{\|\omega\|^{2k}\left(1 + \|\omega\|^2\right)^{-\tau}}_{\le \|\omega\|^{2(k-\tau)}} d\omega \;<\; \infty,$$

and so Corollary 5.5.3 yields

$$\tau \;>\; k + \tfrac{d}{2} \qquad\Longrightarrow\qquad X_{\bullet}(\omega) \;\in\; W^{k,2}(T) \quad \text{a.s.}$$

On the other hand, if $k \ge \tau - \frac{d}{2}$, we find that (w.l.o.g. $r > 1$)

$$\int_{B_r^c(0)} \|\omega\|^{2k}\, \nu(d\omega) \;=\; \int_{B_r^c(0)} \|\omega\|^{2k} \underbrace{\left(1 + \|\omega\|^2\right)^{-\tau}}_{\ge 2^{-\tau}\,\|\omega\|^{-2\tau}} d\omega \;=\; \infty,$$

and by Proposition 5.5.5 and Corollary 5.3.13, if $(X_t)_{t\in T}$ is Gaussian, we conclude

$$\tau \;\le\; k + \tfrac{d}{2} \qquad\Longrightarrow\qquad X_{\bullet}(\omega) \;\notin\; W^{k,2}(T) \quad \text{a.s.}$$

If T is not bounded, both statements hold for $W^{k,2}_{\text{loc}}(T)$ instead of $W^{k,2}(T)$.

Now, we go one step further and give conditions for $(X_t)_{t\in T}$ to have sample paths in some fractional order Sobolev space $W^{\mu,2}(T)$ a.s.:

Theorem 5.5.7. *Let $(X_t)_{t\in T}$ be as in Theorem 5.5.2. If it holds for some $k < \mu < k+1$, $0 < C < \infty$ and $\delta, \eta > 0$ that*

$$d_\alpha^2(s,t) \;\le\; \frac{C\left(D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t)\right)\|t-s\|^{2(\mu-k)}}{\big|\log\|t-s\|\big|^{1+\delta}} \tag{5.24}$$

for all α with $|\alpha| = k$ and all $s,t \in T$ with $\|s-t\| < \eta$, then the sample paths of $(X_t)_{t\in T}$ are in $W^{\mu,2}_{\mathrm{loc}}(T)$ a.s.

If in addition condition (5.17) holds, then the sample paths are in $W^{\mu,2}(T)$ a.s.

Proof: According to Definition 3.2.11 we must show that for a.e. $\omega \in \Omega$

$$\sum_{|\alpha|=k} \int_T\int_T \frac{|D^\alpha X_t(\omega) - D^\alpha X_s(\omega)|^2}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt \;<\; \infty.$$

This follows again from Theorem 2.2.11 (Fubini) if we can ensure that

$$\int_T\int_T \frac{\mathbb{E}\left(|D^\alpha X_t - D^\alpha X_s|^2\right)}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt \;<\; \infty \qquad \text{for all } \alpha \text{ with } |\alpha| = k.$$

To this end we split the integral in two parts and verify that

$$\int_T\int_T \mathbf{1}_{\{\|t-s\|\ge\eta\}}(s,t)\, \frac{\mathbb{E}\left(|D^\alpha X_t - D^\alpha X_s|^2\right)}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt \;<\; \infty \tag{5.25}$$

and

$$\int_T\int_T \mathbf{1}_{\{\|t-s\|<\eta\}}(s,t)\, \frac{\mathbb{E}\left(|D^\alpha X_t - D^\alpha X_s|^2\right)}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt \;<\; \infty. \tag{5.26}$$

Taking $D^\alpha X_t = Y_t^{(\alpha)}$ (from the proof of Theorem 5.5.2) we prove (5.25), noting that

$$\begin{aligned}
\mathbb{E}\left(|D^\alpha X_t - D^\alpha X_s|^2\right) &= D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t) - 2\, D^{\alpha,\alpha}K(s,t)\\
&\le D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t) \;+\; 2\sqrt{D^{\alpha,\alpha}K(s,s)\, D^{\alpha,\alpha}K(t,t)}\\
&\le \tfrac{3}{2}\left(D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t)\right).
\end{aligned}$$

Then, denoting by $\mathcal{S}_d$ the surface of the d-dimensional unit sphere, and using condition (5.17), we have for any $0 < \eta < 1$

$$\begin{aligned}
&\int_T\int_T \mathbf{1}_{\{\|t-s\|\ge\eta\}}(s,t)\, \frac{\mathbb{E}\left(|D^\alpha X_t - D^\alpha X_s|^2\right)}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt\\
&\le \int_T\int_T \mathbf{1}_{\{\|t-s\|\ge\eta\}}(s,t)\, \frac{\frac{3}{2}\left(D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t)\right)}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt\\
&\le 3\underbrace{\int_T D^{\alpha,\alpha}K(t,t)\, dt}_{<\infty} \cdot \underbrace{\int_{\mathbb{R}^d\setminus B_\eta(0)} \frac{1}{\|h\|^{d+2(\mu-k)}}\, dh}_{=\mathcal{S}_d \int_\eta^\infty r^{-1-2(\mu-k)}\, dr < \infty} \;<\; \infty\,.
\end{aligned}$$

In the same way, now using condition (5.24), we obtain

$$\begin{aligned}
&\int_T \int_T \mathbf{1}_{\{\|t-s\|<\eta\}}(s,t)\, \frac{\mathbb{E}\big(|D^\alpha X_t - D^\alpha X_s|^2\big)}{\|t-s\|^{d+2(\mu-k)}}\, ds\, dt \\
&\leq\ C \int_T \int_T \mathbf{1}_{\{\|t-s\|<\eta\}}(s,t)\, \frac{D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t)}{\|t-s\|^d \,\big|\log\|t-s\|\big|^{1+\delta}}\, ds\, dt \\
&\leq\ 2\,C \underbrace{\int_T D^{\alpha,\alpha}K(t,t)\, dt}_{<\infty} \cdot \underbrace{\int_{B_\eta(0)} \frac{1}{\|h\|^d \,\big|\log\|h\|\big|^{1+\delta}}\, dh}_{<\infty} \quad < \quad \infty,
\end{aligned}$$

where we have used that

$$\int_{B_\eta(0)} \frac{1}{\|h\|^d \,\big|\log\|h\|\big|^{1+\delta}}\, dh \ =\ \int_0^\eta \frac{\mathcal{S}_d}{r\ |\log(r)|^{1+\delta}}\, dr \ =\ \int_{-\infty}^{\log\eta} \frac{\mathcal{S}_d}{|r|^{1+\delta}}\, dr.$$

This shows (5.26) and completes the proof.

□

Corollary 5.5.8. *Let $(X_t)_{t\in T}$ be a measurable (weakly) stationary centred random field on an open subset $T \subset \mathbb{R}^d$ with spectral measure ν. If*

$$\int_{\mathbb{R}^d} \big(\log(1+\|\omega\|)\big)^{1+\delta}\, \|\omega\|^{2\mu}\, \nu(d\omega) < \infty \tag{5.27}$$

for some $\mu > 0$, $\mu \notin \mathbb{N}$, and some $\delta > 0$, then the sample paths of $(X_t)_{t\in T}$ are in $W^{\mu,2}_{\mathrm{loc}}(T)$ a.s. If in addition T is bounded, then the paths of $(X_t)_{t\in T}$ are in $W^{\mu,2}(T)$ a.s.

Proof: We start as in the proof of Theorem 5.5.7 with $k := \lfloor\mu\rfloor$. Then (5.25) directly follows from the boundedness of T, and we show that (5.26) is implied by condition (5.27).

Note by repeated application of Corollary 5.3.13 and Proposition 5.3.15 that $(X_t)_{t\in T}$ has k^{th} order m.s. derivatives with spectral measures ν_α and spectral moments

$$M_\alpha \ :=\ \int_{\mathbb{R}^d} \nu_\alpha(d\omega) \ =\ \int_{\mathbb{R}^d} \underbrace{\omega_1^{2\alpha_1}\cdots\omega_d^{2\alpha_d}}_{=:\,\omega^{2\alpha}}\ \nu(d\omega).$$

Using the simple inequality

$$1-\cos(x) \ \leq\ \begin{cases} \frac{x^2}{2}, & |x| \leq 2 \\ 2, & |x| > 2 \end{cases}$$

we obtain

$$\begin{aligned}
&\int_T \int_T \mathbf{1}_{\{\|t-s\|<\eta\}}(s,t)\ \frac{2\cdot(-1)^{|\alpha|}\big(D^{2\alpha}\,\Phi(0) - D^{2\alpha}\,\Phi(t-s)\big)}{\|t-s\|^{d+2(\mu-k)}}\ ds\,dt \\
\leq\ & vol(T)\int_{B_\eta(0)} \frac{2\cdot(-1)^{|\alpha|}\big(D^{2\alpha}\,\Phi(0) - D^{2\alpha}\,\Phi(h)\big)}{\|h\|^{d+2(\mu-k)}}\ dh \\
=\ & vol(T)\int_{B_\eta(0)}\int_{\mathbb{R}^d} \frac{2\big(1-\cos(h'\omega)\big)}{\|h\|^{d+2(\mu-k)}}\ \nu_\alpha(d\omega)\,dh \\
=\ & vol(T)\int_{B_\eta(0)}\int_{\mathbb{R}^d} \underbrace{\frac{(h'\omega)^2\cdot\mathbf{1}_{\{(h'\omega)^2\leq 2\}}(\omega,h)}{\|h\|^{d+2(\mu-k)}}}_{I_1} + \underbrace{\frac{4\cdot\mathbf{1}_{\{(h'\omega)^2> 2\}}(\omega,h)}{\|h\|^{d+2(\mu-k)}}}_{I_2}\ \nu_\alpha(d\omega)\,dh
\end{aligned}$$

Now, if $(h'\omega)^2 > 2$, we also have $\|h\|^2\|\omega\|^2 > 2$ by Cauchy-Schwarz inequality and hence

$$\frac{1}{\|h\|^{2(\mu-k)}} < \frac{\|\omega\|^{2(\mu-k)}}{2^{\mu-k}} < \|\omega\|^{2(\mu-k)}.$$

W.l.o.g. we can assume $\eta < 1$, and we get for $\|h\| < \eta$

$$\big|\log\|h\|\big|^{1+\delta} = \left(\log\left(\frac{1}{\|h\|}\right)\right)^{1+\delta} \leq \big(\log\|\omega\|\big)^{1+\delta} < \big(\log(1+\|\omega\|)\big)^{1+\delta},$$

and hence I_2 can be bounded by

$$\frac{4\cdot\|\omega\|^{2(\mu-k)}\cdot\big(\log(1+\|\omega\|)\big)^{1+\delta}}{\|h\|^d\,\big|\log\|h\|\big|^{1+\delta}}$$

Next, note the Cauchy-Schwarz inequality further implies

$$-\log\|h\| \leq -\log|h'\omega| + \log\|\omega\|$$

and hence, assuming again that $\eta < 1$, it follows that

$$\big|\log\|h\|\big|^{1+\delta} \leq \big(\big|\log|h'\omega|\big| + \log(1+\|\omega\|)\big)^{1+\delta}.$$

For any $\epsilon > 0$ it holds that $a^\epsilon\cdot|\log(a)| \to 0$ as $a \to 0$, and this implies that there exist constants $C_1, C_2 > 0$, so that for $|h'\omega| \leq \sqrt{2}$

$$(h'\omega)^{2-2(\mu-k)}\cdot\big(\big|\log|h'\omega|\big| + \log(1+\|\omega\|)\big)^{1+\delta} \leq \big(C_1 + C_2\,\log(1+\|\omega\|)\big)^{1+\delta}.$$

Now if $\|\omega\| > 1$, then $\log(1+\|\omega\|)$ is bounded away from 0, and C_1 can be absorbed into C_2, i.e.

$$\mathbf{1}_{\{\|\omega\|>1\}}\cdot\big(C_1 + C_2\,\log(1+\|\omega\|)\big)^{1+\delta} \leq \tilde{C_2}\,\big(\log(1+\|\omega\|)\big)^{1+\delta}.$$

Conversely, if $\|\omega\| \leq 1$, then $\log(1+\|\omega\|)$ is bounded as well, and we have

$$\mathbf{1}_{\{\|\omega\|\leq 1\}} \cdot \big(C_1 + C_2 \log(1+\|\omega\|)\big)^{1+\delta} \;\leq\; \mathbf{1}_{\{\|\omega\|\leq 1\}} \cdot \tilde{C}_1,$$

so it finally follows that I_1 can be bounded by

$$\frac{\tilde{C}_1 + \tilde{C}_2\,\|\omega\|^{2(\mu-k)} \cdot \big(\log(1+\|\omega\|)\big)^{1+\delta}}{\|h\|^d\,\big|\log\|h\|\big|^{1+\delta}}.$$

By condition (5.27) we have

$$\begin{aligned} C^* \;&:=\; \int_{\mathbb{R}^d} \big(\log(1+\|\omega\|)\big)^{1+\delta}\,\|\omega\|^{2(\mu-k)}\,\nu_\alpha(d\omega) \\ &=\; \int_{\mathbb{R}^d} \big(\log(1+\|\omega\|)\big)^{1+\delta}\,\|\omega\|^{2(\mu-k)}\,\underbrace{\omega^{2\alpha}}_{\leq\,\|\omega\|^{2k}}\,\nu(d\omega) \;<\; \infty, \end{aligned}$$

and hence, in the same way as in Theorem 5.5.7, it follows that

$$\begin{aligned} &\int_T\int_T \mathbf{1}_{\{\|t-s\|<\eta\}}(s,t)\,\frac{2\cdot(-1)^{|\alpha|}\,\big(D^{2\alpha}\,\Phi(0) - D^{2\alpha}\,\Phi(t-s)\big)}{\|t-s\|^{d+2(\mu-k)}}\,ds\,dt \\ &\qquad\qquad \leq\; vol(T)\int_{B_\eta(0)} \frac{\tilde{C}_1\,M_\alpha + \tilde{C}_2\,C^* + 4\,C^*}{\|h\|^d\,\big|\log\|h\|\big|^{1+\delta}}\,dh \;<\; \infty, \end{aligned}$$

which completes the proof.

□

By application of the imbedding theorem for Sobolev spaces from Section 3.2 we can now state sufficient conditions for a (not necessarily Gaussian) second-order random field to have continuously differentiable sample paths a.s.:

Theorem 5.5.9. *Let $(X_t)_{t\in T}$ be a separable centred RF on a domain $T \subseteq \mathbb{R}^d$ with covariance function K. If for some $\mu > k + \frac{d}{2}$ $D^{\alpha,\alpha}K$ exists for all $|\alpha| \leq \lfloor\mu\rfloor$ and*

$$d_\alpha^2(s,t) \;\leq\; \frac{C\,\big(D^{\alpha,\alpha}K(s,s) + D^{\alpha,\alpha}K(t,t)\big)\,\|t-s\|^{2(\mu-\lfloor\mu\rfloor)}}{\big|\log\|t-s\|\big|^{1+\delta}}$$

for all $|\alpha| = \lfloor\mu\rfloor$, some $0 < C < \infty$ some $\delta > 0$, and all $s,t \in T$ where $\|s-t\|$ is small, then the sample paths of $(X_t)_{t\in T}$ are in $\mathcal{C}^k(T)$ a.s.

Corollary 5.5.10. *Let $(X_t)_{t\in T}$ be a separable (weakly) stationary centred RF on a domain $T \subseteq \mathbb{R}^d$ with spectral measure ν. If*

$$\int_{\mathbb{R}^d} \big(\log(1+\|\omega\|)\big)^{1+\delta}\,\|\omega\|^{2\mu}\,\nu(d\omega) < \infty$$

for some $\mu > k + \frac{d}{2}$ and $\delta > 0$, then the sample paths of $(X_t)_{t\in T}$ are in $\mathcal{C}^k(T)$ a.s.

Proof of Theorem 5.5.9: First, instead of T consider the restriction $(X_t)_{t\in V}$ to some bounded $\mathcal{C}^\infty$ domain $V \subset\subset T$. By Theorem 5.4.2, we can pass on to a separable measurable version $(Y_t)_{t\in V}$ of $(X_t)_{t\in V}$. Then the assumptions of either Theorem 5.5.9 or Corollary 5.5.10 imply

$$Y_{\bullet}(\omega) \;\in\; W^{\mu,2}(V) \qquad \text{for all } \;\omega \in \Omega \setminus N,$$

where N is a set of probability zero. By Theorem 3.2.15 and our choice of μ we can modify $Y_{\bullet}(\omega)$, if necessary, on a λ^d-null set $I(\omega) \subset V$ and we obtain a random field $(\tilde{Y}_t)_{t\in V}$ with sample paths in $\mathcal{C}^k(V)$ a.s. so that

$$\lambda^d\big(\{t \in V \,:\, \tilde{Y}_t(\omega) \neq Y_t(\omega)\}\big) = 0 \qquad \text{for all } \;\omega \in \Omega \setminus N.$$

The a.s. continuity of its sample paths implies that $(\tilde{Y}_t)_{t\in V}$ is measurable and hence, by Lemma 5.4.5, it follows that

$$P\big(\{\omega \,:\, \tilde{Y}_t(\omega) = Y_t(\omega)\}\big) \;=\; 1 \qquad \text{for } \lambda^d\text{-almost all } t \in V.$$

Using Lemma 5.2.8 (and the subsequent remark) we conclude

$$X_{\bullet}(\omega) \;\equiv\; Y_{\bullet}(\omega) \;\equiv\; \tilde{Y}_{\bullet}(\omega) \qquad \text{for almost all } \;\omega \in \Omega,$$

and hence then the sample paths of $(X_t)_{t\in V}$ are in $\mathcal{C}^k(V)$ a.s.

Now the domain T can be covered by a countable union of open balls that are contained in T. Applying the above arguments to each one of these balls yields the desired result.

□

Remark 5.5.11. Note that the same proof, if Theorem 3.2.17 is applied instead of Theorem 3.2.15, yields the following result:

If the conditions of Theorem 5.5.7 and Corollary 5.5.8 hold for $k+\beta$ instead of k, then the sample paths of $(X_t)_{t\in T}$ are in $\mathcal{C}^{k,\beta}_{\mathrm{loc}}(T)$ a.s.

Example 5.5.12. (Whittle-Matérn model, part 3)
Consider a measurable, (weakly) stationary centred random field $(X_t)_{t\in T}$ on an open and bounded domain $T \subset \mathbb{R}^d$ with spectral density

$$\varphi_\tau(\omega) \;=\; \big(1 + \|\omega\|^2\big)^{-\tau}, \qquad \tau > \tfrac{d}{2}.$$

By Corollary 5.5.8, the same calculations as in Example 5.3.19 finally yield

$$\tau \;>\; \mu + \tfrac{d}{2} \qquad \Longrightarrow \qquad X_{\bullet}(\omega) \;\in\; W^{\mu,2}(T) \quad \text{a.s.}$$

If T is not bounded the statement holds for $W^{\mu,2}_{\mathrm{loc}}(T)$ instead of $W^{\mu,2}(T)$.

Recall from Section 3.2 that the RKHS associated with the Whittle-Matérn kernel K corresponding to the above spectral measure is $W^{\tau,2}(T)$. Hence, the Sobolev space containing the sample paths of a random field with covariance function K is “rougher“ by a bit more than $\frac{d}{2}$. In the light of kernel interpolation / kriging this means that a statistician and a numerical analyst who interpolate a sample path of a (weakly stationary) second order random field would use kernels the smoothnesses of which differ by about $\frac{d}{2}$.

The next in our series of examples with the Whittle-Matérn model is the Non-Gaussian counterpart of Example 5.3.19:

Example 5.5.13. (Whittle-Matérn model, part 4)
Consider a separable, (weakly) stationary centred random field $(X_t)_{t\in T}$ on a domain $T \subseteq \mathbb{R}^d$ with spectral density

$$\varphi_\tau(\omega) \;=\; \left(1+\|\omega\|^2\right)^{-\tau}, \qquad \tau > \tfrac{d}{2}.$$

By Corollary 5.5.10, the same calculations as in Example 5.3.19 yield

$$\tau \;>\; k+d \qquad \Longrightarrow \qquad X_{\bullet}(\omega) \;\in\; \mathcal{C}^k(T) \quad \text{a.s.}$$

The price for not requiring $(X_t)_{t\in T}$ to be Gaussian is hence an increase of the required smoothness of the covariance function by $\frac{d}{2}$.
By Remark 5.5.11 we can finally obtain the more precise characterisation

$$\tau \;>\; k+\beta+d \qquad \Longrightarrow \qquad X_{\bullet}(\omega) \;\in\; \mathcal{C}^{k,\beta}_{\text{loc}}(T) \quad \text{a.s.}$$

The sufficient conditions given in Theorem 5.5.2 and Corollary 5.5.3 for $(X_t)_{t\in T}$ to have sample paths a.s. in some integer order Sobolev space $W^{k,2}(T)$ were proved to be even necessary in the stationary Gaussian case. The following example shows that the sufficient conditions for the sample paths to be in some fractional order Sobolev space $W^{\mu,2}(T)$ a.s. (Theorem 5.5.7 and Corollary 5.5.8) are also at least very sharp:

Example 5.5.14. (Whittle-Matérn model, part 5)
Let $(\zeta_n)_{n\in\mathbb{N}}$ be a labeling of the (random) point sets of a homogeneous Poisson point process on $\mathbb{R}^d$ with intensity 1 (see Example 5.5.1) and define

$$X_t \;:=\; \sum_{j=1}^{\infty} \phi_{\tau/2}(t-\zeta_j) \quad \text{where} \quad \phi_s(h) \;=\; \frac{(2\pi)^{\frac{d}{2}}\,\|h\|^{s-\frac{d}{2}}}{2^{s-1}\,\Gamma(s)}\;\mathcal{K}_{s-\frac{d}{2}}(\|h\|)\,. \tag{5.28}$$

This defines (cf. [22, Example 5]) a stationary random field $(X_t)_{t\in\mathbb{R}^d}$ with mean

$$\int_{\mathbb{R}^d} \phi_{\tau/2}(h)\,dh, \qquad t \in \mathbb{R}^d$$

and covariance function $\Phi_\tau = (\phi_{\tau/2} * \phi_{\tau/2}) = (2\pi)^d\,\phi_\tau\,.$

This is, up to the constant factor $(2\pi)^d$, the same Whittle-Matérn model with parameter τ that was also used in several preceding examples. We have already seen that

$$\tau \;>\; k+d \;\;\Longrightarrow\;\; X_{\bullet}(\omega) \;\in\; \mathcal{C}^k(\mathbb{R}^d) \quad \text{a.s.} \quad \text{and}$$

$$\tau \;>\; k+\beta+d \;\;\Longrightarrow\;\; X_{\bullet}(\omega) \;\in\; \mathcal{C}^{k,\beta}_{\text{loc}}(\mathbb{R}^d) \quad \text{a.s.}$$

However, we noted in Lemma 3.2.18 that $\tau < k+\beta+d$ implies $\phi_{\tau/2} \notin \mathcal{C}^{k,\beta}_{\text{loc}}(\mathbb{R}^d)$ for $k=0,1$ and $0<\beta\leq 1$, and that the same implication even holds for $\tau = k+\beta+d$ if $k=\beta=1$. From the properties of a homogeneous Poisson point process it follows that a.e. realization contains at least one point, and since $\phi_{\tau/2}>0$ we can conclude for $k=0,1$ and $0<\beta\leq 1$:

$$\tau \;<\; k+\beta+d \;\;\Longrightarrow\;\; X_{\bullet}(\omega) \;\notin\; \mathcal{C}^{k,\beta}_{\text{loc}}(\mathbb{R}^d) \quad \text{a.s.}$$

Further, since $\mathcal{C}^{k+1}(\mathbb{R}^d) \subset \mathcal{C}^{k,1}_{\text{loc}}(\mathbb{R}^d)$, we have

$$\begin{aligned} \tau &< 1+d &\implies X_{\bullet}(\omega) &\notin \mathcal{C}^1(\mathbb{R}^d) \quad \text{a.s.} \quad \text{and} \\ \tau &\leq 2+d &\implies X_{\bullet}(\omega) &\notin \mathcal{C}^2(\mathbb{R}^d) \quad \text{a.s.} \end{aligned}$$

In order for this example not to be overloaded by technicalities we do not attempt to prove that the RF defined by (5.28) is separable. We rather appeal to the reader's intuition that any other version of it would be even more irregular and would still not have sample paths in $\mathcal{C}^{k,\beta}_{\text{loc}}(\mathbb{R}^d)$ or $\mathcal{C}^k(\mathbb{R}^d)$ respectively.

Chapter 6

Kernel Interpolation / Kriging

In this section we present the approach taken in spatial statistics and approximation theory, respectively, to reconstruct a function $f : T \to \mathbb{R}$ based on its values at a finite set of sampling locations $\mathcal{T} := \{t_1, \ldots, t_n\} \subset T$ which we will always assume to be distinct.
Despite of starting off with different model assumptions, both approaches result in the same type of interpolants. We will compare the different concepts of evaluating the approximation errors that correspond to the two modelling approaches and elaborate the different notions of optimality. We finally comment on the influence of the kernel that is chosen for interpolation.

6.1 Kernel Interpolation

In approximation theory, the model assumption is that $f \in \mathcal{H}_R$ where $\mathcal{H}_R$ is the RKHS that corresponds to a kernel R as defined in Section 3.1. The idea is now to consider the linear subspace

$$V_{R,\mathcal{T}} \;:=\; \operatorname{span}\big\{R(\cdot, t_1), \ldots, R(\cdot, t_n)\big\} \;\subset\; \mathcal{H}_R,$$

and to build an interpolant s_R of f from this subspace, i.e. we set

$$s_R(t) \;:=\; \sum_{j=1}^{n} \alpha_j \, R(t, t_j) \tag{6.1}$$

with coefficients $\alpha_1, \ldots, \alpha_n$ yet to be determined. Now, for s_R to interpolate f at the locations where values are known, it must satisfy

$$\Big(s_R(t_k) \;=\; \Big) \sum_{j=1}^{n} \alpha_j \, R(t_k, t_j) \;=\; f(t_k), \qquad k = 1, \ldots, n. \tag{6.2}$$

If this system is solvable (which is ensured if R is strictly positive definite), then the coefficients $\alpha_1, \ldots, \alpha_n$ are uniquely determined. Apart from being an interpolant of f at $\mathcal{T}$, s_R is also the best approximation to f (cf. [41, Thm. 13.1]):

Theorem 6.1.1. *Suppose that R is a symmetric, continuous and strictly positive definite kernel on $T \subseteq \mathbb{R}^d$. Suppose further that $f \in \mathcal{H}_R$ is known only at $\mathcal{T}$. Then s_R is the best approximation to f from the subspace $V_{R,\mathcal{T}}$ in the sense that*

$$\|f - s_R\|_{\mathcal{H}_R} \;\leq\; \|f - s\|_{\mathcal{H}_R} \qquad \textit{for all} \;\; s \in V_{R,\mathcal{T}}.$$

Hence, s_R is the orthogonal projection of f onto $V_{R,\mathcal{T}}$.

The interpolant s_R can be rewritten in an alternative form (cf. [41, Sec. 11.1]). To this end, we define the cardinal basis functions $u_1^*, \dots, u_n^*$ that are of the same form as s_R, i.e.

$$u_i^*(t) \;=\; \sum_{j=1}^{n} \alpha_j^{(i)}\, R(t, t_j), \qquad i = 1, \dots, n, \tag{6.3}$$

and satisfy the Lagrange conditions

$$u_i^*(t_k) \;=\; \delta_{ik}, \qquad i, k = 1, \dots, n. \tag{6.4}$$

If R is strictly positive definite, then this system is solvable and all cardinal basis functions are uniquely determined. But then, we obviously have

$$s_R(t) \;=\; \sum_{i=1}^{n} u_i^*(t)\, f(t_i) \tag{6.5}$$

since the rhs of (6.5) has the form (6.1) and satisfies the interpolation condition (6.2). This is the so-called Lagrange form of s_R, and another optimality result ([41, Thm. 13.3]) can be given for the cardinal basis functions:

Theorem 6.1.2. *Suppose that R is a symmetric, continuous and strictly positive definite kernel on $T \subseteq \mathbb{R}^d$. Then for $f \in \mathcal{H}_R$ and fixed $t \in T$ it holds that*

$$\sup_{f \in \mathcal{H}_R \,:\, \|f\|_{\mathcal{H}_R} = 1} \big| f(t) - s_R(t) \big| \quad \leq \quad \sup_{f \in \mathcal{H}_R \,:\, \|f\|_{\mathcal{H}_R} = 1} \left| f(t) - \sum_{i=1}^{n} u_i\, f(t_i) \right|$$

for all choices of $u_1, \dots, u_n \in \mathbb{R}$.

In other words: the interpolant s_R is pointwise more accurate in a worst case sense than any other linear combination of the given data.

Another characterization of s_R is given through the following minimal property ([41, Thm. 13.2]):

Theorem 6.1.3. *Let R be as above. Then s_R according has minimal norm $\|\cdot\|_{\mathcal{H}_R}$ among all functions $s \in \mathcal{H}_R$ that interpolate the data $f(t_1), \dots, f(t_n)$ at $\mathcal{T}$.*

6.2 Generalized Kernel Interpolation

Sometimes it is desirable that certain types of functions (e.g. constant or linear functions) are reproduced exactly. Some motivation for that in the application of interpolation methods to the numerical solution of partial differential equations is given in [13, Sec. 6.1].

Let $\mathcal{P} := \operatorname{span}\{p_1, \dots, p_q\}$ be a subspace of $\mathcal{C}(T)$. One will usually have in mind the space $\pi_m(T)$ of polynomials of degree m restricted to T, but we will formulate all results for a general subspace of finite dimension. A first challenge that must be dealt with when functions from $\mathcal{P}$ are to be used for interpolation is to guarantee the solvability of a system of interpolation conditions involving $p_1, \dots, p_q$. This motivates the following notion:

Definition 6.2.1. A finite subset $\mathcal{T} \subset T$ containing at least q points is called $\underline{\mathcal{P}\text{-unisolvent}}$ if the zero function is the only function from $\mathcal{P}$ that vanishes on $\mathcal{T}$.

Note that the points $\{t_1, \dots, t_n\}$ are $\mathcal{P}$-unisolvent if and only if the matrix

$$P = \begin{pmatrix} p_1(t_1) & \cdots & p_q(t_1) \\ \vdots & \ddots & \vdots \\ p_1(t_n,) & \cdots & p_q(t_n) \end{pmatrix} \tag{6.6}$$

has full column rank. Criteria for $\pi_m(\mathbb{R}^d)$-unisolvency are discussed in [41, Sec. 2.2] and [13, Sec. 6.1].

Now, the interpolant $s_{R,\mathcal{P}}$ is formed by both the basis functions $R(\cdot, t_1), \dots, R(\cdot, t_n)$ and the basis functions $p_1, \dots, p_q$:

$$s_{R,\mathcal{P}}(t) \; := \; \sum_{j=1}^{n} \alpha_j \, R(t, t_j) \; + \; \sum_{k=1}^{q} \beta_k \, p_k(t). \tag{6.7}$$

In order to determine the coefficients $\alpha_1, \dots, \alpha_n$ and $\beta_1, \dots, \beta_q$ we require s_R again to satisfy the interpolation conditions

$$\sum_{j=1}^{n} \alpha_j \, R(t_i, t_j) \; + \; \sum_{k=1}^{q} \beta_k \, p_k(t_i) \; = \; f(t_i), \qquad i = 1, \dots, n. \tag{6.8}$$

However, this is no longer enough. To guarantee a unique decomposition of $s_{R,\mathcal{P}}$ into the part made up of the kernel translates $R(\cdot, t_1), \dots, R(\cdot, t_n)$ and the part made up of the basis functions from $\mathcal{P}$ we additionally require

$$\sum_{j=1}^{n} \alpha_j \, p_k(t_j) = 0, \qquad k = 1, \dots, q. \tag{6.9}$$

The system of equations (6.8), (6.9) can be written in compact form

$$\begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} = \begin{pmatrix} \mathrm{f} \\ \mathbf{0} \end{pmatrix} \tag{6.10}$$

with f $= \big(f(t_1), \dots, f(t_n)\big)'$, coefficient vectors $\alpha = (\alpha_1, \dots, \alpha_n)'$, $\beta = (\beta_1, \dots, \beta_q)'$, the matrix P corresponding to the basis functions from $\mathcal{P}$ (see (6.6)), and the system matrix

$$A \;=\; \begin{pmatrix} R(t_1,t_1) & \cdots & R(t_1,t_n) \\ \vdots & \ddots & \vdots \\ R(t_n,t_1) & \cdots & R(t_n,t_n) \end{pmatrix}. \tag{6.11}$$

If R is strictly positive definite, then A has full rank, and together with the remark after Definition 6.2.1 this implies that (6.10) has a unique solution. This implies in particular that any function from $\mathcal{P}$ is reproduced exactly by the expansion (6.7).

Remark 6.2.2. Within the generalized kernel interpolation framework discussed in this subsection, it is possible to weaken the requirement of R being positive definite to requiring only conditional positive definiteness with respect to $\mathcal{P}$. In this work we do however not want to go beyond positive definiteness and refer to [41, Ch. 8] for a detailed discussion of this concept.

As above, $s_{R,\mathcal{P}}$ can be rewritten in the Lagrange form (6.5), now with cardinal basis functions $u_1^*, \dots, u_n^*$ of the form

$$u_i^*(t) \;=\; \sum_{j=1}^{n} \alpha_j^{(i)}\, R(t,t_j) \;+\; \sum_{k=1}^{q} \beta_k^{(i)}\, p_k(t), \qquad i = 1, \dots, n, \tag{6.12}$$

with coefficients $\alpha_1^{(i)}, \dots, \alpha_n^{(i)}$ and $\beta_1^{(i)}, \dots, \beta_q^{(i)}$, $i = 1, \dots, n$ satisfying (6.9) and the Lagrange conditions (6.4). The fact that functions from $\mathcal{P}$ are reproduced exactly entails the following additional property of the cardinal basis functions

$$p(t) \;=\; \sum_{i=1}^{n} u_i^*(t)\, p(t_i) \qquad \text{for every } t \in T \text{ and all } p \in \mathcal{P}. \tag{6.13}$$

This is the form in which the $\mathcal{P}$-reproduction condition will appear in the framework of spatial statistics in Section 6.4.

We shall at least briefly comment on the consequences of working with these generalized kernel interpolants $s_{R,\mathcal{P}}$ on the link with RKHS theory. We follow [41, Ch. 10.3], a much more detailed outline of the theory can be found there.

At the beginning of Section 6.1 kernel interpolation was introduced as a projection of $f \in \mathcal{H}_R$ onto the subspace spanned by the kernel translates $R(\cdot,t_1), \dots, R(\cdot,t_n)$. Now that the basis functions from $\mathcal{P}$ are involved in the projection as well the space of functions to be interpolated must be redefined. First, we choose a $\mathcal{P}$-unisolvent set $\{\xi_1, \dots, \xi_q\} \subset T$ and define the projection operator

$$\Pi_{\mathcal{P}} : \mathcal{C}(T) \to \mathcal{P}, \qquad \Pi_{\mathcal{P}}(f) \;=\; \sum_{k=1}^{q} f(\xi_k)\, p_k.$$

In order to account for the kernel interpolation part we define

$$H_R/\mathcal{P} \;\;:=\;\; \left\{ \sum_{i=1}^{m} a_i\, R(\cdot,t_i) \;:\; a_i \in \mathbb{R},\; t_i \in T,\; m \in \mathbb{N}, \right. \\ \left. \text{with } \sum_{i=1}^{m} a_i\, p(t_i) = 0 \;\text{ for all } p \in \mathcal{P} \right\}.$$

which is the counterpart of H_R in (3.2) but with an additional restriction imposed on the coefficients a_i. We use the same inner product $\|\cdot\|_{\mathcal{H}_R}$ as for H_R (see (3.3)), define $\mathcal{H}_R/\mathcal{P}$ to be the closure of H_R under this inner product, and further define the mapping

$$\mathcal{R} : \mathcal{H}_R/\mathcal{P} \to \mathcal{C}(T), \qquad \mathcal{R}(f) \;=\; f - \Pi_{\mathcal{P}} f.$$

Due to the restrictions on the coefficients a_i this mapping is injective ([41, Lem. 10.15]), and this allows us to define the space

$$\mathcal{H}_{R,\mathcal{P}} \;:=\; \mathcal{R}\big(\mathcal{H}_R/\mathcal{P}\big) \;+\; \mathcal{P} \tag{6.14}$$

equipped with the semi-inner product

$$(f,g)_{\mathcal{H}_{R,\mathcal{P}}} \;:=\; \big(\mathcal{R}^{-1}(f - \Pi_{\mathcal{P}} f),\, \mathcal{R}^{-1}(g - \Pi_{\mathcal{P}}\, g)\big)_{\mathcal{H}_R}.$$

as an appropriate function space for our extended interpolation framework.

It can be shown ([41, Cor. 10.23]) that neither the space $\mathcal{H}_{R,\mathcal{P}}$ nor the inner product $(\cdot,\cdot)_{\mathcal{H}_{R,\mathcal{P}}}$ depend on the choice of the set $\{\xi_1,\dots,\xi_q\}$ used to define $\Pi_{\mathcal{P}}$.

The space $\mathcal{H}_{R,\mathcal{P}}$ need no longer have a reproducing kernel but it is the natural generalization of $\mathcal{H}_R$ for the extended interpolation framework of this subsection. In particular we note the following generalizations of Theorem 6.1.1 - 6.1.3 (see [41, Sec. 13.1]):

Theorem 6.2.3. *Suppose that R is a symmetric, continuous and strictly positive definite kernel on $T \subseteq \mathbb{R}^d$. Suppose further that $\mathcal{T} := \{t_1,\dots,t_n\}$ is $\mathcal{P}$-unisolvent and that $f \in \mathcal{H}_{R,\mathcal{P}}$ is known only at $\mathcal{T}$. Then $s_{R,\mathcal{P}}$ is the best approximation to f from the subspace $V_{R,\mathcal{P},\mathcal{T}}$ in the sense that*

$$\|f - s_{R,\mathcal{P}}\|_{\mathcal{H}_{R,\mathcal{P}}} \;\le\; \|f - s\|_{\mathcal{H}_{R,\mathcal{P}}} \qquad \text{for all}\;\; s \in V_{R,\mathcal{P},\mathcal{T}}.$$

where

$$V_{R,\mathcal{P},\mathcal{T}} := \left\{ s = \sum_{j=1}^{n} \alpha_j\, R(\cdot,t_j) \;:\; \sum_{j=1}^{n} \alpha_j\, p(t_j) = 0 \;\; \text{for all}\;\; p \in \mathcal{P} \right\} \;+\; \mathcal{P}.$$

Hence, $s_{R,\mathcal{P}}$ is the orthogonal projection of f onto $V_{R,\mathcal{P},\mathcal{T}}$.

Theorem 6.2.4. *Suppose that R is a symmetric, continuous and strictly positive definite kernel on $T \subseteq \mathbb{R}^d$. Suppose further that $\mathcal{T}$ is $\mathcal{P}$-unisolvent and let $t \in T$ be fixed. Then we have*

$$\sup_{f \in \mathcal{H}_{R,\mathcal{P}} \,:\, \|f\|_{\mathcal{H}_{R,\mathcal{P}}}=1} \big|f(t) - s_{R,\mathcal{P}}(t)\big| \quad\le\quad \sup_{f \in \mathcal{H}_{R,\mathcal{P}} \,:\, \|f\|_{\mathcal{H}_{R,\mathcal{P}}}=1} \left| f(t) - \sum_{i=1}^{n} u_i\, f(t_i) \right|$$

for all choices of $u_1,\dots,u_n \in \mathbb{R}$ with $\sum_{i=1}^{n} u_i\, p(t_i) \;=\; p(t)$ for all $p \in \mathcal{P}$.

Theorem 6.2.5. *Suppose that R is a symmetric, continuous and strictly positive definite kernel on $T \subseteq \mathbb{R}^d$. Suppose further that $\mathcal{T}$ is $\mathcal{P}$-unisolvent and that values $f(t_1), \dots, f(t_n)$ are given. Then $s_{R,\mathcal{P}}$ has minimal norm $\| \cdot \|_{\mathcal{H}_{R,\mathcal{P}}}$ under all functions $s \in \mathcal{H}_{R,\mathcal{P}}$ that interpolate the given data at $\mathcal{T}$.*

Finally we note the orthogonality property (cf. [41, Lem. 10.24]):

Lemma 6.2.6. *Let R and f be as above and suppose that $\mathcal{T}$ is $\mathcal{P}$-unisolvent. Then we have for the interpolation errors*

$$(f - s_{R,\mathcal{P}}, s)_{\mathcal{H}_{R,\mathcal{P}}} \;=\; 0 \qquad \textit{for all} \;\; s \in V_{R,\mathcal{P},\mathcal{T}}.$$

In particular, it holds that

$$\|s_{R,\mathcal{P}}\|^2_{\mathcal{H}_{R,\mathcal{P}}} \;+\; \|f - s_{R,\mathcal{P}}\|^2_{\mathcal{H}_{R,\mathcal{P}}} \;=\; \|f\|^2_{\mathcal{H}_{R,\mathcal{P}}}.$$

Both the definitions of s_R and $s_{R,\mathcal{P}}$ as well as the preceding theorems that motivate this definition were linked to the assumption that f belongs to $\mathcal{H}_R$ or $\mathcal{H}_{R,\mathcal{P}}$ respectively. It is therefore quite remarkable that the completely different model assumptions made in spatial statistics lead to the same interpolant.

6.3 Simple Kriging

In spatial statistics f is assumed to be a sample path of a second-order RF $(X_t)_{t\in T}$, i.e. $f = X_\bullet(\omega)$ for some $\omega \in \Omega$. The observations $f(t_1), \dots, f(t_n)$ are then realizations of the RVs $X_{t_1}, \dots, X_{t_n}$. To predict the value of $(X_t)_{t\in T}$ at some (unobserved) location $t \in T$, we consider all linear predictors of the form

$$Y_t \;=\; \sum_{i=1}^{n} \lambda_i(t)\, X_{t_i} \tag{6.15}$$

which are themselves random variables. The prediction of f at t given the observations $f(t_1), \dots, f(t_n)$ is then

$$y(t) \;:=\; \sum_{i=1}^{n} \lambda_i(t)\, f(t_i)\,. \tag{6.16}$$

We are now looking for the “best” linear predictor Y_t^*, which we define to be the RV with minimal distance to X_t in $L^2(\Omega, \mathcal{A}, P)$, i.e.

$$\mathbb{E}\big((X_t - Y_t^*)^2\big) \;\le\; \mathbb{E}\big((X_t - Y_t)^2\big) \qquad \text{for all } Y_t \text{ of the form (6.15).}$$

If R is the second moment function of $(X_t)_{t\in T}$ this amounts to minimizing

$$\begin{aligned} &\mathbb{E}(X_t X_t) - 2\sum_{i=1}^{n} \lambda_i(t)\, \mathbb{E}\big(X_{t_i} X_t\big) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i(t)\, \lambda_j(t)\, \mathbb{E}(X_{t_i} X_{t_j}) \\ &\quad = \;\; R(t,t) - 2\sum_{i=1}^{n} \lambda_i(t)\, R(t_i, t) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i(t)\, \lambda_j(t)\, R(t_i, t_j), \end{aligned} \tag{6.17}$$

and by standard arguments concerning quadratic forms, it is seen that the minimum is attained for kriging weights $\lambda_1^*(t), \dots, \lambda_n^*(t)$ satisfying the linear system

$$\sum_{i=1}^{n} \lambda_i^*(t)\, R(t_i, t_j) \;=\; R(t, t_j), \qquad j = 1, \dots n. \tag{6.18}$$

If R is strictly positive definite then this system is uniquely solvable, and this implies

1. the kriging weights $\lambda_1^*(t), \dots, \lambda_n^*(t)$ have the form (6.3)
2. "prediction" at the sampling points $t_1, \dots, t_n$ yields

   $$\lambda_i^*(t_k) \;=\; \delta_{ik}, \qquad i, k = 1, \dots, n,$$

   i.e. the kriging weights satisfy the Lagrange conditions (6.4).

It follows that the kriging predictor (6.16) coincides with the Lagrange form of s_R if the second moment function R of $(X_t)_{t\in T}$ is used as interpolation kernel.

Note that the kriging procedure does not require any knowledge of the underlying random field apart from the second moment function. This is because we defined optimal prediction in terms of the $L^2(\Omega, \mathcal{A}, P)$ distance, which depends on $(X_t)_{t\in T}$ only through R.
Now, from a statistician's point of view working with the second moment function is quite unnatural. It is much more intuitive to work with the covariance function K which describes the structure of the random fluctuations of $(X_t)_{t\in T}$ around the (deterministic) mean function m, while R is a combination of K and m (see (2.11)).
The simplest way to pass from R to K is to assume $m(t) \equiv 0$. In this case the kernels R and K coincide, so we can simply replace R with K in all of the above equations. This approach, which models f as a sample path of a zero mean random field, is called simple kriging.

6.4 Ordinary and Universal Kriging

In most cases the assumption of a zero mean random field does not seem realistic. Another possibility to pass from R to K is to further restrict the class of potential predictors by requiring them to be unbiased, i.e.

$$\mathbb{E}(Y_t) \;=\; \mathbb{E}(X_t) \quad \left(\Longleftrightarrow \quad \sum_{i=1}^{n} \lambda_i(t)\, m(t_i) \;=\; m(t) \right) \quad \text{for all } \; t \in T \tag{6.19}$$

This additional condition ensures that the mean function is reproduced exactly. From a statistical point of view, such a requirement is plausible as it prevents systematic over- or underestimation of X_t. However, as we are interested in reconstructing a sample path of $(X_t)_{t\in T}$ rather than its mean function, it is by no means necessary. We will return to this question at the end of this subsection and give another motivation for condition (6.19) based on practical considerations. Before, we shall study its implications on the form of the optimal predictor.

Introducing the auxiliary function

$$G_R(s,t) \;:=\; R(s,t) - \sum_{j=1}^{n} \lambda_j(s)\, R(t_j,t)$$

we can now write (6.17) (the $L^2(\Omega,\mathcal{A},P)$ distance of Y_t to X_t) as

$$\mathbb{E}\big((X_t - Y_t)^2\big) \;=\; G_R(t,t) - \sum_{i=1}^{n} \lambda_i(t)\, G_R(t,t_i)$$

and using (2.11) we see that

$$\begin{aligned} G_R(s,t) \;&=\; K(s,t) - \sum_{j=1}^{n} \lambda_j(s)\, K(t_j,t) + m(s)\, m(t) - \sum_{j=1}^{n} \lambda_j(s)\, m(t_j)\, m(t) \\ &=\; G_K(s,t) + m(t) \underbrace{\left(m(s) - \sum_{j=1}^{n} \lambda_i(s)\, m(t_j) \right)}_{=0} \;=\; G_K(s,t). \end{aligned}$$

Hence, if we restrict to the class of unbiased predictors we can again replace R with K in the target function (6.17) that we want to minimize.

The corresponding equation system, however, is no longer of the form (6.18). While the target function is still the same, we now have to take into account additional constraints that ensure the unbiasedness of our predictor. In spatial statistics one usually considers models where means behave like

$$m(t) \;:=\; \sum_{k=1}^{q} \beta_k\, p_k(t) \tag{6.20}$$

with known functions $p_1,\ldots,p_q$, and unknown coefficients $\beta_1,\ldots,\beta_q$. Such a mean function is also called a trend.

A very simple but common assumption is $m(t) \equiv \beta_1$, i.e. the mean function is constant (but unknown), and the corresponding procedure is called ordinary kriging.

If the trend has the more general form (6.20), the corresponding interpolation procedure is called universal kriging.

Now, if $m(t)$ is of the form (6.20), the unbiasedness condition (6.19) becomes

$$\sum_{k=1}^{q} \beta_k\, p_k(t) \;=\; \sum_{k=1}^{q} \beta_k \sum_{i=1}^{n} \lambda_i(t)\, p_k(t_i) \qquad \text{for all } \; t \in T.$$

This condition must hold for any set of coefficients $\beta_1,\ldots,\beta_q$, and so we have q conditions

$$p_k(t) \;=\; \sum_{i=1}^{n} \lambda_i(t)\, p_k(t_i), \qquad k = 1,\ldots,q, \tag{6.21}$$

restricting the n kriging weights $\lambda_1^*(t), \dots, \lambda_n^*(t)$. For minimizing the $L^2(\Omega, \mathcal{A}, P)$ distance of Y_t to X_t subject to (6.21) we need Lagrange multipliers $\zeta_1(t), \dots, \zeta_q(t)$, and we find that the optimal kriging weights must now satisfy

$$\sum_{i=1}^{n} \lambda_i^*(t)\, K(t_i, t_j) \;+\; \sum_{k=1}^{q} \zeta_k^*(t)\, p_k(t_j) \;=\; K(t, t_j), \qquad j = 1, \dots, n. \tag{6.22}$$

If the set $\mathcal{T}$ of sampling points is $\mathcal{P}$-unisolvent with $\mathcal{P} := \operatorname{span}\{p_1, \dots, p_q\}$, then the linear system given by (6.21) and (6.22) uniquely solvable.

Solving this system for the kriging weights $\lambda_1^*(t), \dots, \lambda_n^*(t)$ shows that they now must have the form (6.12), and calculating their values at $t_1, \dots, t_n$ shows that the Lagrange conditions are satisfied. Hence, the universal kriging predictor coincides with the Lagrange form of the generalized kernel interpolants $s_{R,\mathcal{P}}$, the special case of ordinary kriging corresponds to interpolants that reproduce constant functions.

Beside the idea that the prediction should be unbiased, another motivation for universal kriging can now be given based on the representation (6.7):

The second term of the predictor is a linear combination of basis functions $p_1, \dots, p_q$, which account for the global trend of f. Such a component can be very useful to reduce the prediction error at locations in sparsely sampled subdomains of T, especially when preliminary analyses suggest that such global trends are present. The first term is a linear combination of kernel translates $K(\cdot, t_1), \dots, K(\cdot, t_n)$ which accounts for (local) deviations from this global trend and provides additional accuracy in more densely sampled subdomains. Universal kriging therefore compromises between modelling global and local features, and such a compromise is often adequate in practical situations.

6.5 Error Analysis

No matter how we proceed, in all cases we obtain the same expression for the expected squared prediction error (also called kriging variance)

$$\mathbb{E}\big((X_t - Y_t^*)^2\big) \;=\; R(t,t) - 2\sum_{i=1}^{n} \lambda_i^*(t)\, R(t_i, t) + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i^*(t)\, \lambda_j^*(t)\, R(t_i, t_j), \tag{6.23}$$

where we note once again that for all of the different kriging approaches discussed above we can replace R with K, provided that our respective assumption on the form of the mean function is correct.

The rhs of equation (6.23) is also well-known to numerical analysts as the square of the (optimal) power function $P_{R,\mathcal{P}}(t)$ for the kernel R (see [41, Sec. 11.1]). Its interpretation in Numerical Analysis is as follows:

Let $\Pi_{R,\mathcal{P}} : \mathcal{H}_{R,\mathcal{P}} \to \mathcal{H}_{R,\mathcal{P}}$ be the map that projects any function $g \in \mathcal{H}_{R,\mathcal{P}}$ on its interpolant $s_{R,\mathcal{P}}$ at $\mathcal{T}$ with the convention that $\mathcal{P} = \{0\}$ means standard kernel interpolation. Then

$$P_{R,\mathcal{P}}(t) \;=\; \sup_{g \in \mathcal{H}_{R,\mathcal{P}} : \|g\|_{\mathcal{H}_{R,\mathcal{P}}} \neq 0} \frac{\big|g(t) - \Pi_{R,\mathcal{P}}(g)(t)\big|}{\|g\|_{\mathcal{H}_{R,\mathcal{P}}}}, \tag{6.24}$$

i.e. $P_{R,\mathcal{P}}$ is the norm of the pointwise error functional of $\Pi_{R,\mathcal{P}}$ (see [34, Sec. 4.1]).

If the function f to be interpolate is in $\mathcal{H}_{R,\mathcal{P}}$, then equation (6.24) yields a bound

$$|f(t) - s_{R,\mathcal{P}}(t)| \;\le\; P_{R,\mathcal{P}}(t) \cdot \|f\|_{\mathcal{H}_{R,\mathcal{P}}} \qquad \text{for all } t \in T \tag{6.25}$$

for the maximal interpolation error, independent of any stochastic assumption about f. Now, for $P_{R,\mathcal{P}}(t)$ itself, a variety of asymptotic bounds in terms of the fill distance

$$h_{\mathcal{T},T} \;:=\; \sup_{t\in T}\, \min_{1\le j\le n} \|t - t_j\|$$

is available (cf. e.g. [41, Sec. 11.2-11.6]), and we shall state one of those bounds that is applicable in the case where $\mathcal{P} = \{0\}$ and $\mathcal{H}_R$ is norm-equivalent to some Sobolev Space $W^{\tau,2}(T)$. In particular we assume that $R(\cdot,\cdot) = \Phi(\cdot - \cdot)$ and that the Fourier transform of Φ satisfies

$$c_1\big(1 + \|\omega\|^2\big)^{-\tau} \;\le\; \widehat{\Phi}(\omega) \;\le\; c_2\big(1 + \|\omega\|^2\big)^{-\tau}, \qquad \omega \in \mathbb{R}^d, \tag{6.26}$$

for some positive constants $c_1 \le c_2$ and $\tau > \frac{d}{2}$.

Definition 6.5.1. ([41, Def. 3.6]) A domain $T \subseteq \mathbb{R}^d$ is said to satisfy an interior cone condition if there exists an angle $\theta \in (0, \frac{\pi}{2})$ and a radius $r > 0$ such that for every $t \in T$ a unit vector $\xi(t)$ exists such that the cone

$$C\big(t, \xi(t), \theta, r\big) \;:=\; \big\{t + \lambda u \,:\, h \in \mathbb{R}^d,\ \|u\| = 1, u'\xi(t) \ge \cos\theta,\ \lambda \in [0, r]\big\}$$

is contained in T.

Theorem 6.5.2. *([41, Cor. 11.33], cf. also [30, Sec. 4]) Suppose that $T \subset \mathbb{R}^d$ is a bounded domain with Lipschitz boundary that satisfies an interior cone condition with radius r and angle θ. Let $f \in W^{\tau,2}(T)$ and $\Pi_R(f)$ its kernel interpolant based on its values at $\mathcal{T} := \{t_1, \dots, t_n\} \subset T$. Further assume that $R(\cdot,\cdot) = \Phi(\cdot - \cdot)$ so that (6.26) holds for $\tau = k + s$, where k is a positive integer and $0 \le s < 1$. If $l \in \mathbb{N}_0$ satisfies $k > l + \frac{d}{2}$, then there exist positive constants h_0 and C, so that the interpolation error can be bounded by*

$$\big\|f - \Pi_R(f)\big\|_{W^{l,\gamma}(T)} \;\le\; C\, h_{\mathcal{T},T}^{\tau - l - d(\frac{1}{2} - \frac{1}{\gamma})_+}\, \big\|f\big\|_{W^{\tau,2}(T)}\,, \qquad 1 \le \gamma \le \infty,$$

provided that $\mathcal{T}$ has fill distance $h_{\mathcal{T},T} < h_0$.

Corollary 6.5.3. *Under the assumptions of Theorem 6.5.2, the power function is bounded by*

$$P_R(t) \;\le\; C\, h_{\mathcal{T},T}^{\tau - \frac{d}{2}}, \qquad \textit{for all } t \in T,$$

provided that $\mathcal{T}$ has fill distance $h_{\mathcal{T},T} < h_0$.

It is obvious from equation (6.24) that $P_R(t)$ only depends on the domain T, the set $\mathcal{T}$ of sampling locations, and the kernel R used for interpolation. It does not depend on the particular function f that is interpolated, and thus Corollary 6.5.3 is valid also in the simple kriging framework:

If simple kriging of a weakly stationary second-order zero-mean random field with covariance function K $(=R)$ is carried out on a domain T satisfying the assumptions of Theorem 6.5.2, then the kriging variance is bounded by

$$\mathbb{E}\big((X_t - Y_t^*)^2\big) \;\leq\; C\, h_{\mathcal{T},T}^{2\tau-d} \qquad \text{for all } t \in T,$$

provided that $\mathcal{T}$ has fill distance $h_{\mathcal{T},T} < h_0$.

For later use we proof the following

Lemma 6.5.4. *If $f \in \mathcal{H}_{R,\mathcal{P}}$ has the particular form*

$$f \;=\; a_0\, R(\cdot, t_0) \;+\; \sum_{j=1}^{n} a_j\, R(\cdot, t_j) \;+\; \sum_{k=1}^{q} b_k\, p_k \tag{6.27}$$

with coefficients that satisfy $a_0 \neq 0$, $\;a_0\, p(t_0) \,+\, \sum_{i=1}^{n} a_i\, p(t_i) = 0$ for all $p \in \mathcal{P}$, then

$$|f(t_0) - s_{R,\mathcal{P}}(t_0)| \;=\; P_{R,\mathcal{P}}(t_0) \cdot \|f - s_{R,\mathcal{P}}\|_{\mathcal{H}_{R,\mathcal{P}}}.$$

Proof: Define the error function $e_f := f - s_{R,\mathcal{P}}$. Since e_f has the form (6.27) as well, it has minimal norm $\|\cdot\|_{\mathcal{H}_{R,\mathcal{P}}}$ among all functions that have the same values on $\mathcal{T} \cup \{t_0\}$ (see Theorem 6.2.5). This form moreover implies $e_f(t_0) \neq 0$ because otherwise we would have $e_f \equiv 0$ (by the uniqueness of kernel interpolants) which is impossible since $a_0 \neq 0$. Using Lemma 6.2.6 and the linearity of Π_R we get

$$\begin{aligned}
P_{R,\mathcal{P}}(t_0) \;&\leq\; \sup_{g \,\in\, \mathcal{H}_{R,\mathcal{P}} \,:\, \|g\|_{\mathcal{H}_{R,\mathcal{P}}} \neq 0} \frac{\big|g(t_0) - \Pi_{R,\mathcal{P}}(g)(t_0)\big|}{\|g - \Pi_{R,\mathcal{P}}(g)\|_{\mathcal{H}_{R,\mathcal{P}}}} \\
&=\; \sup_{g \,\in\, \mathcal{H}_{R,\mathcal{P}} \,:\, g(t_0) - \Pi_{R,\mathcal{P}}(g)(t) \,=\, e_f(t_0)} \frac{\big|g(t) - \Pi_{R,\mathcal{P}}(g)(t)\big|}{\|g - \Pi_{R,\mathcal{P}}(g)\|_{\mathcal{H}_{R,\mathcal{P}}}} \;=\; \frac{|e_f(t_0)|}{\|e_f\|_{\mathcal{H}_{R,\mathcal{P}}}}\,.
\end{aligned}$$

The other inequality follows from (6.25) since $e_f \in \mathcal{H}_{R,\mathcal{P}}$.

□

Corollary 6.5.5. *Let u_k^* be the k^{th} cardinal basis function according to (6.12), (6.4) and $P_{R,\mathcal{P},[-k]}$ the power function for kernel interpolation based on the function values at $\mathcal{T} \setminus \{t_k\}$ only. Then we have*

$$P_{R,\mathcal{P},[-k]}(t_k) \;=\; \|u_k^*\|_{\mathcal{H}_{R,\mathcal{P}}}^{-1}.$$

We conclude this subsection by showing how the different notions of optimality are linked to the way prediction errors are measured. A comparison of both modelling approaches is best possible by looking at the Lagrange forms (6.5) and (6.15) respectively. As usual let

$\mathcal{T} := \{t_1, \dots, t_n\}$ be the set of sampling locations. Denote by $\mathcal{F}(T)$ the space of all functions $u : T \to \mathbb{R}^n$ and define the set

$$\mathcal{F}_{\mathcal{P}}(T) \;:=\; \left\{ u \in \mathcal{F} \;:\; p(t) \;=\; \sum_{i=1}^{n} u_i(t)\, p(t_i) \;\; \text{for all} \;\; t \in T \;\; \text{and all } p \in \mathcal{P} \right\}$$

of permissible weight functions. For $u \in \mathcal{F}_{\mathcal{P}}(T)$ define the projection

$$\Pi_u : \mathbb{R}^T \to \operatorname{span}\{u_1, \dots, n_n\}, \quad f \mapsto \sum_{i=1}^{n} u_i\, f(t_i),$$

which generalizes the projection Π_R defined above. Indeed, for $f \in \mathcal{H}_{R,\mathcal{P}}$ and u^* according to (6.12), (6.4) and (6.13) we have $\Pi_{R,\mathcal{P}}(f) = \Pi_{u^*}(f)$. Finally, we generalize the power function and define

$$P_u(t) \;=\; \sup_{g \in \mathcal{H}_{R,\mathcal{P}} : \|g\|_{\mathcal{H}_{R,\mathcal{P}}} \neq 0} \frac{\left| g(t) - \Pi_u(g)(t) \right|}{\|g\|_{\mathcal{H}_{R,\mathcal{P}}}}, \qquad u \in \mathcal{F}_{\mathcal{P}}(T),$$

which is the norm of the pointwise error functional for approximation according to (6.5) with weight functions $u_1, \dots, u_n$. Now, by Theorem 6.2.4, we have

$$P_{u^*}(t) \;\le\; P_u(t) \qquad \text{for all} \;\; t \in T \;\; \text{and all } u \in \mathcal{F}_{\mathcal{P}}(T),$$

and hence the cardinal basis functions $u_1^*, \dots, u_n^*$ can be defined pointwise as the minimizers of $P_u(t)$ in u. On the other hand, the proof of [41, Thm. 13.3] shows

$$P_u(t) \;=\; R(t,t) - 2\sum_{i=1}^{n} u_i(t)\, R(t_i, t) + \sum_{i=1}^{n}\sum_{j=1}^{n} u_i(t)\, u_j(t)\, R(t_i, t_j), \tag{6.28}$$

which is the same expression as (6.17), the expected squared prediction error at t, the minimizers of which were defined to be the optimal kriging weights.

As a summary we note:

1. In both approximation theory and spatial statistics the optimal interpolant of a function f at sampling points $t_1, \dots, t_n$ can be represented in the form

   $$s_{u^*}(t) \;=\; \Pi_{u^*}(f)(t) \;=\; \sum_{i=1}^{n} u_i^*(t)\, f(t_i),$$

   with weight functions $u_1^*, \dots, u_n^*$ defined pointwise as the minimizers of (6.28) in u, where minimization ranges over all $u \in \mathcal{F}_{\mathcal{P}}(T)$ (this ensures that s_{u^*} reproduces functions in $\mathcal{P}$).

2. For numerical analysts the target function $P_u(t)$ is interpreted as

   $$P_u(t) \;=\; \sup_{g \in \mathcal{H}_{R,\mathcal{P}} : \|g\|_{\mathcal{H}_{R,\mathcal{P}}} = 1} \left| g(t) - \Pi_u(g)(t) \right|.$$

   They assume f to belong to some space $\mathcal{H}_{R,\mathcal{P}}$ and consider an approximant as optimal if it minimizes for every fixed $t \in T$ the worst possible approximation error over all choices of $f \in \mathcal{H}_{R,\mathcal{P}}$ with $\|f\|_{\mathcal{H}_{R,\mathcal{P}}} = 1$.

3. For spatial statisticians the target function $P_u(t)$ is interpreted as

$$P_u(t) \;=\; \mathbb{E}\Big(\big(X_t - \Pi_u(X_\bullet)(t)\,\big)^2\Big)\,.$$

They leave the function space of f unspecified (i.e. $\mathbb{R}^T$) but assume a probability structure on this space. Specifically they assume f to be a sample path of a second-order RF $(X_t)_{t\in T}$ with mean function $m \in \mathcal{P}$ and covariance function K, and consider a predictor of f as optimal if it minimizes the expected squared prediction error at every $t \in T$.

6.6 Best Prediction of Random Fields revisited

Simple kriging can also be viewed as a projection of the unknown RV X_t on the linear subspace generated by the RVs at the sampling locations $t_1, \ldots, t_n$ where $(X_t)_{t\in T}$ is observed.

Recall the definition of $\mathcal{S}_X$ in section 3.3 as the Hilbert space closure of the space S_X of all linear combinations of RVs from $(X_t)_{t\in T}$ under the inner product

$$\langle X_t, X_s\rangle \;=\; \mathbb{E}(X_t X_s) \;=\; R(t,s), \quad s,t \in T.$$

The RVs $X_{t_1}, \ldots, X_{t_n}$ define a linear subspace S_Y of $\mathcal{S}_X$

$$S_Y \;:=\; \left\{ \sum_{j=1}^{n} a_j X_{t_j} \;:\; a_j \in \mathbb{R} \right\},$$

which inherits the inner product defined above. Now simple arguments from Hilbert space theory show that the construction of the interpolation process $(Y_t^*)_{t\in T}$ by pointwise minimization of (6.28) is equivalent to calculating the orthogonal projection of $X_t \in \mathcal{S}_X$ on S_Y for each $t \in T$. This projection property yields a decomposition into two orthogonal RFs

$$(X_t)_{t\in T} \;=\; (Y_t^*)_{t\in T} \;+\; (\varepsilon_t)_{t\in T}. \tag{6.29}$$

Indeed, defining $\rho(t) := \big(R(t,t_1), \ldots, R(t,t_n)\big)'$, $X_{\mathcal{T}} := \big(X_{t_1}, \ldots, X_{t_n}\big)'$, and A as in (6.11), we can combine the kriging system (6.18) and the representation (6.15) and rewrite it in the compact form

$$Y_t^* \;=\; \rho(t)' A^{-1} X_{\mathcal{T}}.$$

The second moment function of $(Y_t^*)_{t\in T}$ is then given by

$$\mathbb{E}\big(Y_t^*\, Y_s^*\big) \;=\; \rho(t)' A^{-1} \rho(s)\,. \tag{6.30}$$

Defining the kriging error process $(\varepsilon_t)_{t\in T}$ by $\varepsilon_t := X_t - Y_t^*$, we note as immediate consequence of $(Y_t^*)_{t\in T}$ being a pointwise orthogonal projection on S_Y:

$$\langle \varepsilon_t, Z\rangle \;=\; 0 \qquad \text{for all } t \in T, \;\; \text{for all } \; Z \in S_Y. \tag{6.31}$$

Hence, the random fields $(Y_t^*)_{t\in T}$ and $(\varepsilon_t)_{t\in T}$ are indeed orthogonal and $(\varepsilon_t)_{t\in T}$ has second moment function

$$\mathbb{E}\big(\varepsilon_t\, \varepsilon_s\big) \;=\; R(t,s) - \rho(t)' A^{-1} \rho(s)\,. \tag{6.32}$$

Under the simple kriging assumption ($m(t) \equiv 0$) the second moment functions in (6.30) and (6.32) coincide with the respective covariance functions and (6.31) implies that any ε_t, $t \in T$ is uncorrelated with any RV $Z \in S_Y$.

Remark 6.6.1. If $(X_t)_{t\in T}$ is a Gaussian random field then so is $(\varepsilon_t)_{t\in T}$. More precisely, it follows from the arguments in the last paragraph in Section 3.3 that any RVct

$$(\varepsilon_{s_1}, \dots, \varepsilon_{s_m}, X_{t_1}, \dots, X_{t_n})', \quad m \in \mathbb{N},$$

is multivariate Gaussian and hence, by Lemma 2.4.18, $(\varepsilon_t)_{t\in T}$ is independent of S_Y.

The interpretation of simple kriging as the pointwise orthogonal projection on a linear subspace illustrates why only the covariance information of the underlying random field is needed to construct interpolants and to provide a probabilistic error analysis. However, the restriction to *linear* subspaces can limit the potential precision of the resulting predictors (see [39, Sec. 1.4] for an example where linear prediction is suboptimal) compared to the conditional expectation

$$\mathbb{E}\big[X_t \,\big|\, X_{t_1}, \dots, X_{t_n}\big],$$

which was found to be the best $\sigma\big(X_{t_1}, \dots, X_{t_n}\big) \,/\, \mathbb{B}$ *measurable* predictor (see Proposition 2.5.2). However, in the special case where $(X_t)_{t\in T}$ is a Gaussian random field, it follows from Proposition 2.5.6 that $\mathbb{E}\big[X_t \,\big|\, X_{t_1}, \dots, X_{t_n}\big]$ is a linear function of $X_{t_1}, \dots, X_{t_n}$ and coincides with the kriging predictor Y_t^* from (6.15), so simple kriging already yields the best prediction in the statistical sense that can be obtained.

In the light of the preceding remarks, we can once again contrast the different points of view from spatial statistics and numerical analysis (see end of Section 6.5) in the special case of simple kriging ($\mathcal{P} = \{0\}$):

1. With respect to prediction, spatial statisticians take a somewhat Bayesian point of view. They impose a "prior distribution" on $\mathbb{R}^T$ assuming a zero-mean Gaussian RF with covariance function K. The posterior distribution, i.e. the distribution given the observations $\big(f(t_1), \dots, f(t_n)\big)' =: \mathrm{f}$ is then a Gaussian RF with mean function

$$y^*(t) \;=\; \kappa(t)'\, A^{-1}\, \mathrm{f}\,,$$

where $\kappa(t) := \big(K(t, t_1), \dots, K(t, t_n)\big)'$, and covariance function

$$K_\varepsilon(s,t) \;=\; K(t,s) - \kappa(t)'\, A^{-1}\, \kappa(s)\,.$$

2. Numerical analysts on the contrary take a minimax point of view. They limit the space of considered functions to $\mathcal{H}_R$ and seek to minimize at each location $t \in T$ the maximal approximation error over all choices of $f \in \mathcal{H}_R$.

In the case of ordinary and universal kriging X_t is projected on the space

$$S_{Y,\mathcal{P},t} \;:=\; \left\{ \sum_{j=1}^{n} a_j X_{t_j} \;:\; a_j \in \mathbb{R}, \quad \sum_{j=1}^{n} a_j\, p(t_j) = p(t) \quad \forall\, p \in \mathcal{P} \right\},$$

which is no longer a linear space. For ordinary kriging it is an affine space, for universal kriging it even depends on t, and the uncorrelatedness of the random fields $(Y_t^*)_{t\in T}$ and $(\varepsilon_t)_{t\in T}$ can no longer be guaranteed. However we note

Lemma 6.6.2. *Let the RV Z be a contrast of $X_{t_1}, \ldots, X_{t_n}$ with respect to $\mathcal{P}$, i.e.*

$$Z \;=\; \sum_{j=1}^{n} a_j X_{t_j}, \quad a_j \in \mathbb{R}, \qquad \sum_{j=1}^{n} a_j\, p(t_j) = 0 \qquad \textit{for all } \; p \in \mathcal{P}$$

Then ε_t is uncorrelated with Z for any $t \in T$.

Proof: Using the equation (6.22) for the kriging predictor Y_t^* we get

$$\begin{aligned}
\operatorname{Cov}(\varepsilon_t, Z) \;&=\; \sum_{j=1}^{n} a_j \left(K(t,t_j) - \sum_{i=1}^{n} \lambda_i^*(t)\, K(t_i,t_j) \right) \\
&=\; \sum_{j=1}^{n} a_j \sum_{k=1}^{n} \zeta_k^*(t)\, p_k(t_j) \;=\; \sum_{k=1}^{n} \zeta_k^*(t) \underbrace{\sum_{j=1}^{n} a_j\, p_k(t_j)}_{=0} \;=\; 0
\end{aligned}$$

□

6.7 Kernel Interpolation / Kriging with Wrong Kernels

So far we have proceeded as if we knew the correct covariance function K that should be used for constructing the interpolant. However, in practice it is usually unknown and an appropriate choice for it must be “guessed” based on the available data. In this case we can no longer assume to be using the covariance function that exactly corresponds to the actual second-order structure of the random field $(X_t)_{t\in T}$, so it is reasonable to ask how much our prediction deviates from the prediction based on the true second-order structure.
This question also arises from the perspective of a numerical analyst, who works under the assumption $f \in \mathcal{H}_R$ or $f \in \mathcal{H}_{R,\mathcal{P}}$, which is linked to the smoothness of the interpolation kernel R.

We start with the numerical analysts' point of view but restrict our discussion to the case $\mathcal{P} = \{0\}$. We assume that $f \in \mathcal{H}_R = W^{\mu,2}(T)$ (in the sense that $\mathcal{H}_R$ and $W^{\mu,2}(T)$ coincide as vector spaces and are norm equivalent) but we use a kernel $\tilde{R}$ with $\mathcal{H}_{\tilde{R}} = W^{\tau,2}(T)$.

For $\tau < \mu$, i.e. in the case were the kernel used for interpolation is too rough, we always have $f \in W^{\mu,2}(T) \subset W^{\tau,2}(T)$ and hence Theorem 6.5.2 is still applicable and yields for any $l \in \mathbb{N}_0$ that satisfies $\tau > l + \frac{d}{2}$:

$$\left\| f - \Pi_{\tilde{R}}(f) \right\|_{W^{l,\gamma}(T)} \;\leq\; C h_{T,T}^{\tau - l - d\left(\frac{1}{2} - \frac{1}{\gamma}\right)} \; \|f\|_{W^{\tau,2}(T)}.$$

This means that choosing the interpolation kernel $\tilde{R}$ too rough will in general imply losing the advantages of R, so that everything works as if we were in the $\tilde{R}$ setting.

For $\tau > \mu$, i.e. in the case were the kernel used for interpolation is too smooth, f is no longer in the RKHS of $\tilde{R}$, and the traditional RKHS techniques do not apply. However, using different arguments, [30] proved the following "escape" theorem, that extends Theorem 6.5.2 to the case where $f \notin W^{\tau,2}(T)$. Beside the fill distance $h_{\mathcal{T},T}$ it involves another characteristic of $\mathcal{T}$, the so-called separation radius

$$q_{\mathcal{T}} \;:=\; \tfrac{1}{2}\, \min_{j \neq k} \|t_j - t_k\|$$

which is half of the smallest distance between any two distinct points in $\mathcal{T}$. The mesh ratio $\rho_{\mathcal{T},T} := h_{\mathcal{T},T} / \, q_{\mathcal{T}}$ then characterizes the uniformity of the set $\mathcal{T}$ of sampling points in T.

Theorem 6.7.1. *([30, Thm. 4.2]) Let $T \subset \mathbb{R}^d$ be a bounded domain with Lipschitz boundary that satisfies an interior cone condition with radius r and angle θ. Suppose that $f \in W^{\mu,2}(T)$ and let $\Pi_{\tilde{R}}(f)$ its kernel interpolant at $\mathcal{T}$ with kernel $\tilde{R}(\cdot,\cdot) = \Phi(\cdot - \cdot)$ so that (6.26) holds. If $\mu \leq \tau$, $\mu = k + s$ for some positive integer $k > \frac{d}{2}$ and $0 \leq s < 1$, then there exist positive constants h_0 and C, so that the interpolation error can be bounded by*

$$\big\|f - \Pi_R(f)\big\|_{W^{\kappa,2}(T)} \;\leq\; C\, h_{\mathcal{T},T}^{\mu-\kappa}\, \rho_{\mathcal{T},T}^{\tau-\kappa}\, \|f\|_{W^{\mu,2}(T)}, \qquad \text{for any } \; 0 \leq \kappa \leq \mu,$$

provided that $\mathcal{T}$ has fill distance $h_{\mathcal{T},T} < h_0$.

Hence, if $\rho_{\mathcal{T},T}$ is bounded by some constant $\rho < \infty$, the interpolation error w.r.t. $\|\cdot\|_{W^{\kappa,2}(T)}$ tends to zero at the rate $\mu - \kappa$ as long as $\mathcal{H}_{\tilde{R}} = W^{\tau,2}(T) \supset W^{\mu,2}(T) \ni f$.

Remark 6.7.2. The final conclusion on the issue of misspecifying the smoothness of the interpolation kernel R seems to be that the approximation order is maintained even if $\tilde{R}$ is too smooth, but is reduced to the order that one could expect in the $\mathcal{H}_R$ setting if $\tilde{R}$ is too rough. This is not yet the whole truth, however. [40] shows that for kernels of the type considered above the approximation order doubles if f is in a certain (smooth) subspace $\mathcal{H}^*_{\tilde{R}}$ of $\mathcal{H}_{\tilde{R}}$. Thus, a kernel $\tilde{R}$ with about half of the smoothness of the "right" kernel R may still yield the same approximation order. In general, $\mathcal{H}^*_{\tilde{R}}$ must not only guarantee double smoothness, but also certain boundary conditions of its functions. We shall refer to [40] and [35] for details and further results (the second author discusses the general case $\mathcal{H}_{R,\mathcal{P}}$ with $\mathcal{P} = \pi_m(\mathbb{R}^d)$ and expresses the smoothness and boundary conditions of $\mathcal{H}^*_{\tilde{R}}$ in a general RKHS framework).

Now, we take the spatial statistician's point of view. Denote by $\tilde{\lambda}^*(t)$ the vector of optimal kriging weights at t according to either (6.18) or (6.21) and (6.22), respectively, but now with respect to the covariance function $\tilde{K}$, and by $\tilde{Y}^*_t$ the corresponding ($\tilde{K}$-optimal) kriging prediction at t. The L^2 interpolation error of $\tilde{Y}^*_t$ is then

$$\begin{aligned}\mathbb{E}_K\big((X_t - \tilde{Y}^*_t)^2\big) \;&=\; \mathbb{E}_K\left(X_t - \sum_{i=1}^n \tilde{\lambda}^*_i(t)\, X_{t_i}\right)^2 \\ &=\; K(t,t) - 2\sum_{i=1}^n \tilde{\lambda}^*_i(t)\, K(t,t_i) + \sum_{i=1}^n\sum_{j=1}^n \tilde{\lambda}^*_i(t)\, \tilde{\lambda}^*_j(t)\, K(t_i,t_j),\end{aligned}$$

which is a non-optimal “mixed” power function for K (respectively R).

We have seen (Remark 6.6.1) that for simple kriging, when $(X_t)_{t\in T}$ is in addition assumed to be Gaussian, the kriging error process $(\varepsilon_t)_{t\in T}$ is independent of $X_{t_1},\ldots,X_{t_n}$. In this case, we can make the effect of misspecification of the covariance function used for kriging explicit by calculating the kriging error variance conditionally on the available data. By splitting the interpolation error and using the properties of conditional expectation we obtain

$$\begin{aligned}
&\mathbb{E}_K\big[(X_t-\tilde{Y}_t^*)^2 \,\big|\, X_{t_1},\ldots,X_{t_n}\big]\\
&\quad= \mathbb{E}_K\big[(X_t-Y_t^*+Y_t^*-\tilde{Y}_t^*)^2 \,\big|\, X_{t_1},\ldots,X_{t_n}\big]\\
&\quad= \mathbb{E}_K\big[(X_t-Y_t^*)^2 \,\big|\, X_{t_1},\ldots,X_{t_n}\big] \;+\; \mathbb{E}_K\big[(Y_t^*-\tilde{Y}_t^*)^2 \mid X_{t_1},\ldots,X_{t_n}\big]\\
&\qquad + 2\,\mathbb{E}_K\big[(X_t-Y_t^*)(Y_t^*-\tilde{Y}_t^*) \,\big|\, X_{t_1},\ldots,X_{t_n}\big]\\
&\quad= \mathbb{E}_K\big((X_t-Y_t^*)^2\big) \;+\; (Y_t^*-\tilde{Y}_t^*)^2 \;+\; 2\,(Y_t^*-\tilde{Y}_t^*)\cdot\underbrace{\mathbb{E}_K\big(X_t-Y_t^*\big)}_{=0}\\
&\quad= P_K^2(t) \;+\; (Y_t^*-\tilde{Y}_t^*)^2
\end{aligned}$$

The conditional expectation of the squared prediction error given the data is then

$$\mathbb{E}_K\big[(X_t-\tilde{Y}_t^*)^2 \,\big|\, X_{t_1}=f(t_1),\ldots,X_{t_n}=f(t_n)\big] \;=\; P_K^2(t) \;+\; (y_t^*-\tilde{y}_t^*)^2.$$

Hence, subject to our assumptions, the kriging variance increases by a deterministic term $(y_t^*-\tilde{y}_t^*)^2$ that depends on $K,\tilde{K}$, the given data $f(t_1),\ldots,f(t_n)$, and the geometry of $\mathcal{T}\cup\{t\}$. If the misspecification of $\tilde{K}$ is small enough to at least guarantee that $\tilde{y}_t^*\in\mathcal{H}_K$, we can use (6.25) to get

$$\mathbb{E}\big[(X_t-\tilde{Y}_t^*)^2 \,\big|\, X_{t_1}=f(t_1),\ldots,X_{t_n}=f(t_n)\big] \;\leq\; P_K^2(t)\cdot\big(1+\|y_t^*-\tilde{y}_t^*\|_{\mathcal{H}_R}^2\big)$$

The term $\|y_t^*-\tilde{y}_t^*\|_{\mathcal{H}_K}^2$ is an upper bound for the relative increase of the kriging variance when the covariance function $\tilde{K}$ is used instead of the (correct) covariance function K.

We study the effect of misspecifying the covariance function on prediction for a particular class of stationary covariance functions, i.e. $K(\cdot,\cdot)=\Phi(\cdot-\cdot)$, the Whittle-Matérn class

$$\Phi_{r,\nu}(h) \;=\; \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{2\nu^{1/2}\|h\|}{r}\right)^{\nu}\mathcal{K}_\nu\left(\frac{2\nu^{1/2}\|h\|}{r}\right), \qquad r,\nu>0, \tag{6.33}$$

where $\mathcal{K}_\nu$ is the modified Bessel function of the third kind. This class was already introduced in Definition 3.2.12, but here we use an alternative parametrization, that was recommended by [18], and has the following attractive features:

1. $\Phi_{r,\nu}$ is independent of the state space dimension d and $\Phi_{r,\nu}(0)=1$.

2. the parameter r rescales the argument h and therefore determines how quickly the correlations of the RF decay with distance.

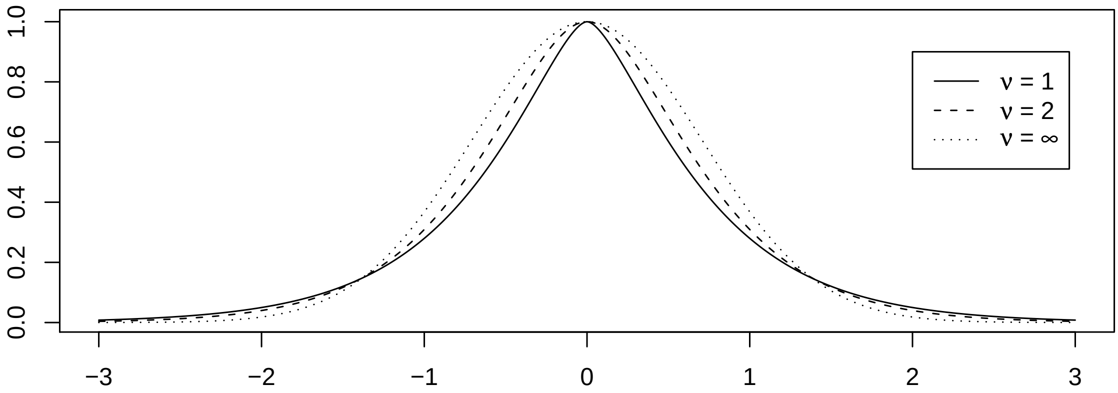

Figure 6.1: Plots of the covariance function $\Phi_{1,\nu}$ for different values of ν.

3. the parameter ν parametrizes the smoothness of $\Phi_{r,\nu}$ at $h = 0$ and hence of the associated RF. Unlike the original smoothness parameter τ it has only moderate influence on the covariances at longer distances (see Figure 6.1).

Due to the last point the interpretation of r is largely independent of ν, which is also illustrated by the fact that

$$\lim_{\nu\to\infty} \Phi_{r,\nu}(h) \;=\; e^{-\|h\|^2/r^2}$$

(see [39, p. 50]). This limit model is the Gaussian model with scale parameter r that has the same interpretation as above.

In later calculations we will need the partial derivatives of $\Phi_{r,\nu}$ with respect to the parameters r and ν. These are given by

$$\begin{aligned}
\tfrac{\partial}{\partial r}\,\Phi_{r,\nu}(h) &= \frac{\varrho^{\nu+1}\,\mathcal{K}_{\nu-1}(\varrho)}{2^{\nu-1}\Gamma(\nu)\,r}\\
\tfrac{\partial}{\partial \nu}\,\Phi_{r,\nu}(h) &= \frac{\varrho^{\nu}\left(\log\left(\frac{\varrho}{2}\right)-\psi(\nu)\right)\,\mathcal{K}_{\nu}(\varrho)}{2^{\nu-1}\Gamma(\nu)} + \frac{\varrho^{\nu}\,\left(\frac{\partial}{\partial\nu}\,\mathcal{K}_{\nu}\right)(\varrho)}{2^{\nu-1}\Gamma(\nu)} - \frac{\varrho^{\nu+1}\,\mathcal{K}_{\nu-1}(\varrho)}{2^{\nu}\,\nu\,\Gamma(\nu)}
\end{aligned}$$

where ψ denotes the Digamma function, and $\varrho := \frac{2\nu^{1/2}\|h\|}{r}$.

For the derivatives with respect to ν at $\nu = 0.5, 1$ and 2, we can use the formulae

$$\begin{aligned}
\tfrac{\partial}{\partial \nu}\,\mathcal{K}_{\nu}(z)\big|_{\nu=0.5} &= -\left(\frac{\pi}{2z}\right)^{1/2} e^{z}\,\mathrm{Ei}(-2z) \qquad \text{where} \qquad \mathrm{Ei}(z) = \int_{-\infty}^{z} \frac{e^t}{t}dt\\
\tfrac{\partial}{\partial \nu}\,\mathcal{K}_{\nu}(z)\big|_{\nu=1} &= z^{-1}\mathcal{K}_0(z)\\
\tfrac{\partial}{\partial \nu}\,\mathcal{K}_{\nu}(z)\big|_{\nu=2} &= 2z^{-2}\mathcal{K}_0(z) + 2z^{-1}\mathcal{K}_1(z)
\end{aligned}$$

([16, 8.486(1), 9., 21.]) and the functional relations

$$\mathcal{K}_{\pm 1/2}(z) \;=\; \sqrt{\frac{\pi}{2z}}\, e^{-z} \quad \text{and} \quad \mathcal{K}_2(z) \;=\; \frac{2}{z}\,\mathcal{K}_1(z) \;+\; \mathcal{K}_0(z)$$

([16, 8.469, 3., and 8.486, 17]) to obtain the explicit representations

$$\begin{aligned}
\tfrac{\partial}{\partial\nu}\,\Phi_{r,\nu}(h)\big|_{\nu=0.5} \;&=\; \Big(\log(\varrho) - \psi(\tfrac{1}{2}) - \varrho\Big)\, e^{-\varrho} \;-\; \mathrm{Ei}(-2\varrho)\, e^{\varrho} \\
\tfrac{\partial}{\partial\nu}\,\Phi_{r,\nu}(h)\big|_{\nu=1} \;&=\; \Big(\log\big(\tfrac{\varrho}{2}\big) - \psi(1)\Big)\varrho\,\mathcal{K}_1(\varrho) \;+\; \Big(1 - \tfrac{\varrho^2}{2}\Big)\,\mathcal{K}_0(\varrho) \\
\tfrac{\partial}{\partial\nu}\,\Phi_{r,\nu}(h)\big|_{\nu=2} \;&=\; \Big(\log\big(\tfrac{\varrho}{2}\big) - \psi(1) - \tfrac{\varrho^2}{8}\Big)\,\varrho\,\mathcal{K}_1(\varrho) \;+\; \Big(1 + \big(\log\big(\tfrac{\varrho}{2}\big) - \psi(2)\big)\tfrac{\varrho^2}{2}\Big)\,\mathcal{K}_0(\varrho)
\end{aligned}$$

We consider prediction of $f : [-l_{\max}, l_{\max}] \to \mathbb{R}$ at $t = 0$ based on its values at

$$\mathcal{T} \;:=\; \{\pm\,\delta, \pm\,2\delta, \ldots\} \;\cap\; [-l_{\max}, l_{\max}].$$

f is assumed to be a sample path of a zero mean weakly stationary second-order RF with covariance function Φ_θ, $\theta = (r, \nu)$ according to the model (6.33).

Now suppose that prediction is done by simple kriging with the (incorrect!) covariance function $\Phi_{\tilde\theta}$, $\tilde\theta = (\tilde r, \tilde\nu)$. We can write the simple kriging system (6.18) in compact form as

$$A_{\tilde\theta}\,\tilde\lambda^* \;=\; b_{\tilde\theta}$$

where $b_{\tilde\theta} = \big(\Phi_{\tilde\theta}(t - t_1), \ldots, \Phi_{\tilde\theta}(t - t_n)\big)'$, and $A_{\tilde\theta}$ is the system matrix (6.11) with $\Phi_{\tilde\theta}$ instead of R. The expected squared prediction error (at $t = 0$) is then

$$\mathcal{V}_\theta(\tilde\theta) \;:=\; \mathbb{E}_{\Phi_\theta}\big((X_t - \tilde Y_t^*)^2\big) \;=\; 1 \;-\; 2\, b_\theta'\, A_{\tilde\theta}^{-1}\, b_{\tilde\theta} \;+\; b_{\tilde\theta}'\, A_{\tilde\theta}^{-1}\, A_\theta\, A_{\tilde\theta}^{-1}\, b_{\tilde\theta}\,,$$

and we will study the second derivatives of $\mathcal{V}_\theta(\tilde\theta)$ with respect to $\tilde\theta_1 = \tilde r$ and $\tilde\theta_2 = \tilde\nu$ at $\tilde\theta = \theta$. The motivation for this comes from the Taylor expansion

$$\mathcal{V}_\theta(\tilde\theta) \;=\; \mathcal{V}_\theta(\theta) \;+\; (\tilde\theta - \theta)'\,(\nabla\mathcal{V}_\theta)(\theta) \;+\; \tfrac{1}{2}\,(\tilde\theta - \theta)'\,\mathrm{Hess}(\mathcal{V}_\theta)(\theta)\,(\tilde\theta - \theta) \;+\; O\big(\|\tilde\theta - \theta\|^3\big),$$

where $(\nabla\mathcal{V}_\theta)(\cdot)$ is the gradient and $\mathrm{Hess}(\mathcal{V}_\theta)(\cdot)$ is the Hessian of $\mathcal{V}_\theta(\tilde\theta)$ w.r.t. $\tilde\theta$.

Note that

$$\mathcal{V}_\theta(\theta) \;=\; P^2_{\Phi_\theta}(0) \quad \text{and} \quad (\nabla\mathcal{V}_\theta)(\theta) \;=\; 0$$

since $\tilde\theta = \theta$ is the optimal choice and hence the minimizer of $\mathcal{V}_\theta(\tilde\theta)$. Consequently, we have

$$\mathcal{V}_\theta(\tilde\theta) \;\approx\; P^2_{\Phi_\theta}(0) \;+\; \tfrac{1}{2}\,(\tilde\theta - \theta)'\,\mathrm{Hess}(\mathcal{V}_\theta)(\theta)\,(\tilde\theta - \theta) \tag{6.34}$$

when the deviation of $\tilde\theta$ from θ is small. The second derivatives of $\mathcal{V}_\theta(\tilde\theta)$ at $\tilde\theta = \theta$ can give us an indication about how strongly a misspecification of r and ν, respectively, increases the expected squared prediction error.

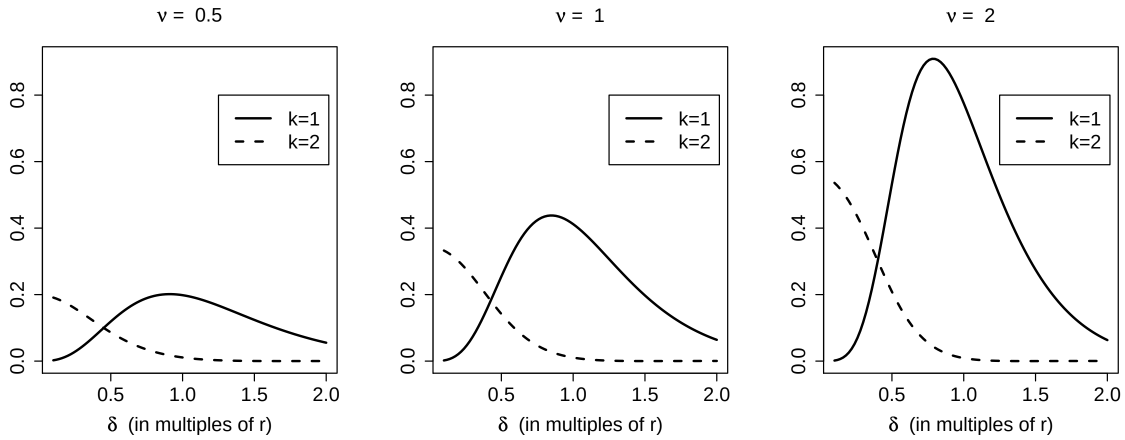

Figure 6.2: The variable $\Delta^k_{rel}\,\mathcal{V}_\theta$ as a function of δ for different values of ν. $k = 1$ corresponds to prediction sensitivity to deviations from r, $k = 2$ corresponds to prediction sensitivity to deviations from ν.

These derivatives are given by

$$\begin{aligned}\frac{\partial^2 \mathcal{V}_\theta}{\partial\tilde\theta_k \partial\tilde\theta_l}(\theta) \quad &= \quad 2\, b'_\theta\, A_\theta^{-1}\, \frac{\partial}{\partial\theta_l} A_\theta\, A_\theta^{-1}\, \frac{\partial}{\partial\theta_k} A_\theta\, A_\theta^{-1}\, b_\theta + 2\, \frac{\partial}{\partial\theta_l} b'_\theta\, A_\theta^{-1}\, \frac{\partial}{\partial\theta_k} b_\theta \\ &\qquad -2\, \frac{\partial}{\partial\theta_l} b'_\theta\, A_\theta^{-1}\, \frac{\partial}{\partial\theta_k} A_\theta\, A_\theta^{-1}\, b_\theta - 2\, b'_\theta\, A_\theta^{-1}\, \frac{\partial}{\partial\theta_l} A_\theta\, A_\theta^{-1}\, \frac{\partial}{\partial\theta_k} b_\theta\end{aligned}$$

In our calculations we use $l_{\max} = 10$, $r = 1$ and plot the variables

$$\Delta^k_{rel}\,\mathcal{V}_\theta \;:=\; \frac{1}{2}\left(\frac{\partial^2 \mathcal{V}_\theta}{(\partial\tilde\theta_k)^2}(\theta)\cdot\theta_k^2\right) \Big/ P^2_{\Phi_\theta}(0), \qquad k = 1, 2. \tag{6.35}$$

for different values of ν as a function of δ (the distance between $t = 0$ and the nearest sampling point). The definition of $\Delta^k_{rel}\,\mathcal{V}_\theta$ is motivated as follows:

- Multiplying the second partial derivatives by θ_k^2 passes from absolute deviations $(\tilde\theta_k - \theta_k)$ to relative deviations $(\tilde\theta_k/\theta_k - 1)$ in (6.34)
- Dividing by $P^2_{\Phi_\theta}(0)$ passes from absolute to relative increase of $P^2_{\Phi_\theta}(0)$

This facilitates the interpretation and makes it independent of the magnitude (and physical dimension) of r, ν and $P^2_{\Phi_\theta}(0)$. For instance, $10^{-2}\cdot\Delta^k_{rel}\,\mathcal{V}_\theta$ gives the relative approximate increase of the expected squared prediction error for a relative deviation of 10^{-1} from the correct parameter value.

Figure 6.2 shows plots of $\Delta^k_{rel}\,\mathcal{V}_\theta$ for $\nu = 0.5, 1$ and 2. δ is given in multiples of the scaling parameter r (the curves are then virtually independent of the choice of r as long as $r \ll l_{\max}$). The following conclusions can be drawn from these plots

1. If the sampling points are very dense (i.e. $\delta \ll r$) then prediction accuracy depends mainly on the correct choice of the smoothness parameter ν and hardly on the scale parameter r.

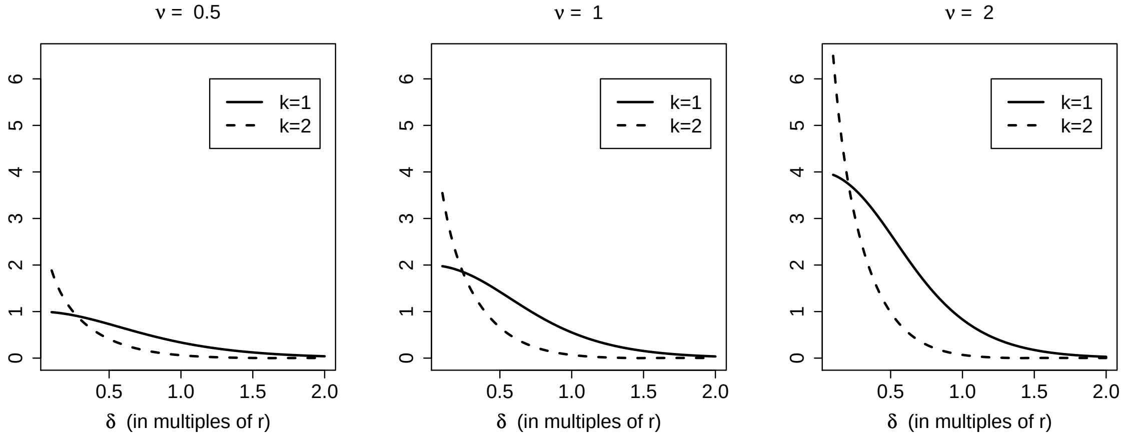

Figure 6.3: The variable $-\Delta_{rel}^k P_{\Phi_\theta}^2(0)$ as a function of δ for different values of ν. $k=1$ and $k=2$ correspond to the changes of the predicted kriging variance due to deviations from r and ν respectively.

2. The influence of ν decreases rapidly as the sampling points get thinner. The influence of r increases first, becomes maximal when $\delta \approx r$ and then goes down again.

3. The magnitude of $\Delta_{rel}^k \mathcal{V}_\theta$ increases with increasing smoothness. For the values of ν used here (these are realistic in statistical applications) it is quite moderate: a 10% misspecification of either r or ν increases the expected squared prediction error by at most 0.2% (for $\nu = 0.5$), 0.4% (for $\nu = 1$) and 0.9% (for $\nu = 2$) respectively.

In spatial statistics one is not only interested in the best possible prediction of f at an unknown location $t \in T$, but also in obtaining reliable information about the magnitude of the prediction error. Such information is given through the kriging variance

$$P_{\Phi_\theta}^2(t) \;=\; \mathbb{E}_{\Phi_\theta}\big((X_t - Y_t^*)^2\big).$$

Now, if $\Phi_{\tilde\theta}$ is falsely assumed to be the covariance function, one would take $P_{\Phi_{\tilde\theta}}^2(t)$ as kriging variance. Therefore we shall also study the effect of misspecifying θ on the value of $P_{\Phi_\theta}^2(t)$. In the above setting, we now consider the first order approximation

$$P_{\Phi_{\tilde\theta}}^2(0) \;\approx\; P_{\Phi_\theta}^2(0) \,+\, (\tilde\theta - \theta)' \left(\nabla P_{\Phi_\theta}^2(0)\right)$$

where $\nabla P_{\Phi_\theta}^2(0)$ is the gradient of $P_{\Phi_\theta}^2(0)$ w.r.t. θ, and we study the variables

$$\Delta_{rel}^k P_{\Phi_\theta}^2(0) \;=\; \left(\frac{\partial P_{\Phi_\theta}^2(0)}{\partial\theta_k}\cdot\theta_k\right) \Big/\, P_{\Phi_\theta}^2(0), \qquad k = 1, 2,$$

which give the relative increase of $P_{\Phi_\theta}^2(0)$. For instance, if the relative deviation of $\tilde\theta$ from θ is 10^{-1}, then the relative increase of $P_{\Phi_\theta}^2(0)$ is $10^{-1}\cdot\Delta_{rel}^k P_{\Phi_\theta}^2(0)$.

Figure 6.3 shows plots of $-\Delta_{rel}^k P_{\Phi_\theta}^2(0)$ for $\nu = 0.5, 1$ and 2. We note the following conclusions from these plots

1. Here, too, the influence of the smoothness parameter ν is big if the sampling points are very dense and decreases rapidly as δ increases.
2. The influence of r is now also decreasing as δ increases. For δ very small its influence is biggest, but smaller than that of ν, however the decline of its influence is slower.
3. The magnitude of $\Delta^k_{rel}\, P^2_{\Phi_\theta}(0)$ increases with increasing smoothness.
4. $P^2_{\Phi_\theta}(0)$ is much more sensitive to parameter misspecification than $\mathcal{V}_\theta(\tilde{\theta})$:
   increasing r by 10% decreases the kriging variance by up to 10% (for $\nu = 0.5$), 20% (for $\nu = 1$) and 40% (for $\nu = 2$), respectively.

Summing up, based on these studies we can expect that the prediction accuracy is relatively robust against parameter misspecification, but our ability to predict the magnitude of the prediction error may suffer substantially if we fail to identify the true underlying covariance function.

In the preceding sensitivity study we took a statistical point of view, but many of our conclusions also apply to the numerical analysis framework. For example, we observed that in finite settings the scaling of the covariance function (interpolation kernel) plays a role for prediction (approximation) accuracy as well although it did not appear in the statements on approximation orders at the beginning of this subsection.
We come back to this issue in the next section, where we discuss methods to identify the true covariance function or, in the language of numerical analysis, an optimal interpolation kernel.

Chapter 7

Parameter Identification

In Section 6 interpolation methods for functions $f \in \mathcal{H}_R$ and for sample paths $X_{\bullet}(\omega)$ of a second order RF $(X_t)_{t\in T}$ with covariance function K were derived. These methods were based on the assumption that the appropriate reproducing kernel R or covariance function K, respectively, are known. In practice, this is not the case in general, and hence there is a need for algorithms that select R or K based on the available data $\big(f(t_1),\dots,f(t_n)\big)' =: \mathrm{f}$. The ideas of what a "good" choice of a kernel or covariance function is are not the same for statisticians and numerical analysts. For the latter the main interest is in selecting a kernel that yields the best possible approximation of f. The former are trying to identify "true" covariance function K of $(X_t)_{t\in T}$ which, by construction, leads to the best (linear) prediction in the stochastic sense.

In this section we describe the procedure of leave-one-out cross validation (LOOCV) which was proposed by Rippa ([31]) in the context of kernel interpolation, and the maximum likelihood estimator (MLE) which is one of the standard methods in the context of spatial statistics. Both methods assume that R or K is from a parametric family of kernels $\{R_\theta : \theta \in \Theta \subset \mathbb{R}^p\}$ and they try to realize a near-optimal choice of the parameter θ. We show that the MLE, despite its probabilistic background, has a reasonable interpretation also in the kernel interpolation framework and we assess the performance of both methods from both a statistical and a numerical analysis point of view.

7.1 Cross Validation

Cross validation is a very general idea that has long been used in the statistic literature. The algorithm proposed by Rippa corresponds to LOOCV, one of its variants. The idea is to split off one single location t_k at a time, calculate the partial interpolant $s_{R_\theta,\mathcal{P},[-k]}$ of all data pairs except $\big(t_k, f(t_k)\big)$ and, from it, the approximation errors at t_k

$$\varepsilon_k \;:=\; f(t_k) - s_{R_\theta,\mathcal{P},[-k]}(t_k), \qquad k = 1,\dots,n.$$

If we let $\varepsilon_\theta = (\varepsilon_1,\dots,\varepsilon_n)'$ be the vector of cross validation errors, the parameter θ is then chosen as the minimizer of some norm $\|\cdot\|_p$ of ε_θ, i.e.,

$$\theta_{\mathrm{CV_p}} \;=\; \operatorname*{argmin}_{\theta\in\Theta} \;\|\varepsilon_\theta\|_p\,,$$

assuming that $\|\varepsilon_\theta\|_p$ should depend on θ in the same way as $\|f - s_{R_\theta,\mathcal{P}}\|_{L_p(T)}$. The plots in [31] show, that for $p = 1, 2$ such an assumption is plausible.

Another way of looking at the LOOCV errors would be to consider the error function $f_{\varepsilon_k} := s_{R_\theta,\mathcal{P}} - s_{R_\theta,\mathcal{P},[-k]}$ of the surrogate problem of interpolating $s_{R_\theta,\mathcal{P}}$ based on the data at the locations $\{t_1, \dots, t_n\} \setminus \{t_k\}$. Then we have $\varepsilon_k = f_{\varepsilon_k}(t_k)$.

The LOOCV procedure does not make any explicit assumption on the function f that is to be reconstructed, so in principle it can be used in both the statistical and the numerical analysis framework. Its performance however depends on the implicit assumption that the general behaviour of f with respect to interpolation is reflected well by its behaviour on the discrete subset $\mathcal{T} := \{t_1, \dots, t_n\} \subset T$. Its performance relative to other methods will therefore strongly depend on f and $\mathcal{T}$. This issue will be studied later in this section.

An inconvenient feature about LOOCV from the computational aspect seems to be that n interpolants $s_{R_\theta,\mathcal{P},[-1]}, \dots, s_{R_\theta,\mathcal{P},[-n]}$ have to be calculated for each choice of θ. The following proposition generalizes a similar statement in [31] to the general interpolation framework with interpolants of the form (6.7) that reproduce functions from a finite dimensional space $\mathcal{P}$. It shows how the LOOCV error vector can be obtained with the same computational effort that is needed to calculate a single interpolant.

Consider the interpolation system (6.10). It follows from the calculation rules for block matrices that the inverse of a block matrix has a block structure as well and so we can write

$$\begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix}^{-1} = \begin{pmatrix} \Psi & \Xi \\ \Xi' & * \end{pmatrix}.$$

We shall use the convention that for interpolants of the simpler form (6.1) we understand $\mathcal{P} = \{0\}$ and hence (6.10) simply becomes $A\alpha = \mathrm{f}$. Then, this case is also covered by the following Proposition.

Proposition 7.1.1. *The LOOCV errors $\varepsilon_1, \dots, \varepsilon_n$ defined above are given by*

$$\varepsilon_k \;=\; \frac{\alpha_k}{\Psi_{kk}}, \qquad k = 1, \dots, n. \tag{7.1}$$

Proof: (generalizes the proof in [31])
Denote by

$$\begin{pmatrix} A_{[-k]} & P_{[-k]} \\ (P_{[-k]})' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha_{[-k]} \\ \beta_{-[k]} \end{pmatrix} = \begin{pmatrix} \mathrm{f}_{[-k]} \\ \mathbf{0} \end{pmatrix}$$

the equation system corresponding to $s_{R_\theta,\mathcal{P},[-k]}$. $A_{[-k]}$ is obtained by removing the k^{th} row and the k^{th} column of A, $P_{[-k]}$ is obtained by removing the k^{th} row from P, and $\mathrm{f}_{[-k]}$ is obtained by removing the k^{th} element of f.

Using the same notation for some matrix M and compatible vectors y, z we note

$$My = z, \quad y_k = 0 \quad\Longrightarrow\quad M_{[-k]}\, y_{[-k]} = z_{[-k]} \tag{7.2}$$

Denoting by $\Psi_{\bullet k}$ and $\Xi_{\bullet k}$ the k^{th} column of Ψ and Ξ respectively, and by $e_k \in \mathbb{R}^{n+q}$ the k^{th} canonical unit vector we can write

$$\begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Psi_{\bullet k} \\ \Xi_{\bullet k} \end{pmatrix} = e_k \,, \qquad k = 1, \ldots, n. \tag{7.3}$$

Necessarily we must have $\Psi_{kk} \neq 0$, otherwise (7.2) and (7.3) would imply

$$\begin{pmatrix} A_{[-k]} & P_{[-k]} \\ (P_{[-k]})' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Psi^{(k)}_{\bullet k} \\ \Xi_{\bullet k} \end{pmatrix} = \mathbf{0}.$$

But since $\begin{pmatrix} A_{[-k]} & P_{[-k]} \\ (P_{[-k]})' & \mathbf{0} \end{pmatrix}$ is invertible, this would mean $\begin{pmatrix} \Psi_{\bullet k} \\ \Xi_{\bullet k} \end{pmatrix} = \mathbf{0}$ which is impossible since (7.3) has a unique (nonzero) solution.

Now consider the vector

$$\begin{pmatrix} \vartheta^{(k)} \\ \eta^{(k)} \end{pmatrix} := \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \frac{\alpha_k}{\Psi_{kk}} \cdot \begin{pmatrix} \Psi_{\bullet k} \\ \Xi_{\bullet k} \end{pmatrix}$$

We have

$$\begin{aligned} \begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vartheta^{(k)} \\ \eta^{(k)} \end{pmatrix} &= \begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \alpha \\ \beta \end{pmatrix} - \frac{\alpha_k}{\Psi_{kk}} \cdot \begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \Psi_{\bullet k} \\ \Xi_{\bullet k} \end{pmatrix} \\ &= \Big(\mathrm{f}_1, \ldots, \mathrm{f}_{k-1}, \mathrm{f}_k - \tfrac{\alpha_k}{\Psi_{kk}}, \mathrm{f}_{k+1}, \cdots, \mathrm{f}_n, 0, \ldots, 0 \Big)' \end{aligned}$$

and since $\vartheta_k^{(k)} = 0$, the uniqueness of the coefficient vectors and (7.2) imply

$$\begin{pmatrix} \alpha_{[-k]} \\ \beta_{[-k]} \end{pmatrix} = \Big(\vartheta_1^{(k)}, \ldots, \vartheta_{k-1}^{(k)},\, \vartheta_{k+1}^{(k)}, \ldots, \vartheta_n^{(k)},\, \eta_1^{(k)}, \ldots, \eta_q^{(k)} \Big)'$$

so we obtain for the interpolant $s_{R_\theta,[-k]}$ at t_k

$$\begin{aligned} s_{R_\theta,[-k]}(t_k) &= \sum_{\substack{i=1 \\ i \neq k}}^{n} \alpha_{[-k],i}\, R_\theta(t_k, t_i) + \sum_{j=1}^{q} \beta_{[-k],j}\, p_j(t_k) \\ &= \sum_{i=1}^{n} \vartheta_i^{(k)}\, R_\theta(t_k, t_i) + \sum_{j=1}^{q} \eta_j^{(k)}\, p_j(t_k)\,. \end{aligned}$$

The last term however is simply the k^{th} row of $\begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \begin{pmatrix} \vartheta^{(k)} \\ \eta^{(k)} \end{pmatrix}$ which we found to be equal to $\mathrm{f}_k - \frac{\alpha_k}{\Psi_{kk}}$ and this completes the proof.

□

An interpretation of the variables $\Psi_{11}, \ldots, \Psi_{nn}$ can be given by the following

Lemma 7.1.2. *For the power function $P_{R_\theta,\mathcal{P},[-k]}$ corresponding to $s_{R_\theta,\mathcal{P},[-k]}$ we have the relation*

$$P^2_{R_\theta,\mathcal{P},[-k]}(t_k) \;=\; \Psi^{-1}_{kk}.$$

Proof: Let u^*_k be the k^{th} cardinal basis function according to (6.12), (6.4) with coefficient vectors $\alpha^{(k)}$, $\beta^{(k)}$. For functions of this form, the norm $\|\cdot\|_{\mathcal{H}_{R,\mathcal{P}}}$ can be calculated explicitly (see Section 6.2) and we obtain

$$\|u^*_k\|^2_{\mathcal{H}_{R,\mathcal{P}}} \;=\; \big(\alpha^{(k)}\big)' A\,\alpha^{(k)} \;=\; \big(\alpha^{(k)}\big)' e_k - \big(\alpha^{(k)}\big)' P\,\beta^{(k)} \;=\; \big(\alpha^{(k)}\big)' e_k,$$

since, by condition (6.9), we have $P'\alpha^{(k)} = \mathbf{0}$. Using (6.10) we get

$$\|u^*_k\|^2_{\mathcal{H}_{R,\mathcal{P}}} \;=\; \begin{pmatrix} \alpha^{(k)} \\ \beta^{(k)} \end{pmatrix}' \begin{pmatrix} e_k \\ \mathbf{0} \end{pmatrix} \;=\; \begin{pmatrix} e_k \\ \mathbf{0} \end{pmatrix}' \begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} e_k \\ \mathbf{0} \end{pmatrix} \;=\; \Psi_{kk}\,,$$

and the assertion of the Lemma follows from Corollary 6.5.5.

□

7.2 Maximum Likelihood

The idea of maximum likelihood is also a very general idea that is used in many different fields in statistics for parameter identification. Yet it is not as general as cross validation because it is always based on very specific model assumptions under which the maximum likelihood estimator (MLE) and its counterpart corresponding to universal kriging, the restricted maximum likelihood estimator (REML), are derived. In our case these assumptions are:

- f is a sample path of a second order random field $(X_t)_{t\in T}$
- $(X_t)_{t\in T}$ has mean function m and covariance function K_θ
- $(X_t)_{t\in T}$ is Gaussian

While the first two are standard working assumptions in the kriging framework the last one has been argued against even by spatial statisticians for being too specific (recall that this assumption is not needed for kriging). In the framework of kernel interpolation none of these assumptions is necessary and so it is not clear whether the MLE also makes sense in this context. In Section (7.4) we will show that (RE)ML estimation still can be given a meaningful interpretation and often yields good results even if the above model assumptions are not met.

For now, we shall however take the classical approach to derive the MLE and we assume the model above, initially with $m(t) \equiv 0$. There usually is a special parameter in each covariance model which in the stationary case can be interpreted as the variance

$$\mathrm{Var}(X_t) \;=\; K(t,t) \;\overset{\text{stationary case}}{=}\; \Phi(0)$$

of the random field $(X_t)_{t\in T}$. We shall treat it separately from the other parameters and consider covariance models of the form

$$\upsilon\, K_\theta, \quad \upsilon \in \mathbb{R}_+,\ \theta \in \Theta,$$

where K_θ is normalized in an appropriate way. In the stationary case we require $\Phi_\theta(0) = 1$, then υ indeed parametrizes the variance. In the nonstationary case we could simply set $K_\theta(t_0, t_0) = 1$ for a fixed $t_0 \in T$.
Treating υ separately is convenient since it will also turn out to allow a special treatment in estimation. Moreover we note that none of the interpolants derived in Chapter 6 depend on υ, in particular this parameter is completely meaningless in the framework of numerical analysis. In spatial statistics it is important, but only for the assessment of the interpolation error.

For this model the joint probability density of $X_{\mathcal{T}} := (X_{t_1}, \ldots, X_{t_n})'$ is given by

$$\varphi_{\upsilon,\theta}(x) \;=\; \frac{1}{(2\pi)^{\frac{n}{2}}\, \upsilon^{\frac{n}{2}}\, |A_\theta|^{\frac{1}{2}}}\; e^{-\frac{1}{2\upsilon}\, x'\, A_\theta^{-1} x}.$$

(see Lemma 2.4.18) where we define $A_\theta \in \mathbb{R}^{n\times n}$ as in (6.11) with K_θ instead of R. The MLE then chooses $\vartheta := (\upsilon, \theta)$ so that the corresponding probability distribution for the RVs $X_{t_1}, \ldots, X_{t_n}$ has maximal density ("likelihood") at f:

$$\vartheta_{\text{MLE}} \;=\; \underset{\vartheta\, \in \mathbb{R}_+ \times \Theta}{\arg\max}\;\; \varphi_\vartheta(\mathrm{f}).$$

The idea is that this covariance model is the most likely one to have produced the observations $f(t_1), \ldots, f(t_n)$. Since $\log(\cdot)$ is a monotone function, it is equivalent to maximize the log likelihood

$$l(\vartheta; \mathrm{f}) \;=\; -\tfrac{n}{2}\,\log(2\pi) - \tfrac{n}{2}\,\log(\upsilon) - \tfrac{1}{2}\,\log(|A_\theta|) - \tfrac{1}{2\upsilon}\,\mathrm{f}'\,A_\theta^{-1}\mathrm{f} \tag{7.4}$$

which is more convenient to work with. Now for any $\theta \in \Theta$ the maximum in υ is attained by

$$\upsilon_{\text{MLE}}(\theta) \;=\; \tfrac{1}{n}\,\mathrm{f}'\,A_\theta^{-1}\mathrm{f},$$

and by plugging this back into (7.4) we obtain the profile log likelihood in θ

$$l(\theta; \mathrm{f}) \;=\; -\tfrac{n}{2}\,\log(2\pi) - \tfrac{n}{2}\,\big(1 - \log(n)\big) - \tfrac{1}{2}\,\log(|A_\theta|) - \tfrac{n}{2}\,\log\left(\mathrm{f}'\,A_\theta^{-1}\mathrm{f}\right), \tag{7.5}$$

which we can use to define the MLE for θ:

$$\theta_{\text{MLE}} \;=\; \underset{\theta\in\Theta}{\operatorname{argmin}}\;\; l(\theta; \mathrm{f})\,.$$

So far, for the derivation of the MLE, we have made the additional assumption that $m(t) \equiv 0$ which corresponds to the assumption in simple kriging. We now want to relax this assumption to the one made in universal kriging that the mean function is given by

$$m(t) \;:=\; \sum_{k=1}^{q} \beta_k\, p_k(t)$$

with known functions p_k, and unknown coefficients β_k, $k = 1, \dots, q$ (see (6.20)). This leaves us with the problem that in addition to υ and θ we also have to deal with the unknown parameters $\beta = (\beta_1, \dots, \beta_q)'$, which can be considered as nuisance parameters since our interest is in estimating θ.

One way to handle this is to pass from f to contrasts $\tilde{\mathrm{f}} := Q\mathrm{f}$, $Q \in \mathbb{R}^{n\times n}$, so that the distribution of the corresponding RVs

$$\tilde{X}_{\mathcal{T}} \;=\; QX_{\mathcal{T}}$$

is independent of β. This approach leads to restricted maximum likelihood (REML) estimation and the idea is similar to what is done in universal kriging: the linear combinations of $X_{t_1}, \dots, X_{t_n}$ that may be used for estimation are limited to those that filter out the mean. Defining P as in (6.6) we see that e.g. $Q = I - P(P'P)^{-1}P'$ is a suitable choice which yields (see Lemma 2.4.18):

$$X_{\mathcal{T}} \;\sim\; \mathcal{N}(P\beta, \upsilon A_\theta) \quad\Longrightarrow\quad \tilde{X}_{\mathcal{T}} \;\sim\; \mathcal{N}\big(\mathbf{0}, \upsilon\, QA_\theta Q'\big)\;.$$

Now, $QA_\theta Q'$ has rank $n - q$ and is therefore not invertible. So, the likelihood for $\tilde{X}_{\mathcal{T}}$ cannot be defined in the same way as above. Instead, a subset of $n - q$ linearly independent components must be chosen and the likelihood is then defined for the joint distribution of those components only. It can be shown (see [29, Exc. 7.10-7.13]) that the likelihood functions for different choices of Q and of the $n - q$ linearly independent components are proportional. A particular representation of the (restricted) log likelihood is given by (see [19])

$$\begin{aligned} l(\vartheta; \mathrm{f}) \;&=\; -\tfrac{n-q}{2}\log(2\pi) - \tfrac{n-q}{2}\log(\upsilon) - \tfrac{1}{2}\log(|A_\theta|) - \tfrac{1}{2}\log\big(|P'A_\theta^{-1}P|\big) \\ &\quad +\tfrac{1}{2}\log(|P'P|) - \tfrac{1}{2\upsilon}\,\mathrm{f}'\,\big(A_\theta^{-1} - A_\theta^{-1}P(P'A_\theta^{-1}P)^{-1}P'A_\theta^{-1}\big)\,\mathrm{f}\;. \end{aligned} \tag{7.6}$$

For given $\theta \in \Theta$ the maximum in υ is attained by

$$\upsilon_{\mathrm{MLE}}(\theta) \;=\; \tfrac{1}{n-q}\,\mathrm{f}'\,\big(A_\theta^{-1} - A_\theta^{-1}P(P'A_\theta^{-1}P)^{-1}P'A_\theta^{-1}\big)\,\mathrm{f}\,,$$

and plugging this back into (7.6) we obtain the restricted profile log likelihood in θ

$$\begin{aligned} l(\theta; \mathrm{f}) \;&=\; -\tfrac{n-q}{2}\log(2\pi) - \tfrac{n-q}{2}\big(1 - \log(n-q)\big) - \tfrac{1}{2}\log(|A_\theta|) - \tfrac{1}{2}\log\big(|P'A_\theta^{-1}P|\big) \\ &\quad +\tfrac{1}{2}\log(|P'P|) - \tfrac{n-q}{2}\log\big(\mathrm{f}'\,\big(A_\theta^{-1} - A_\theta^{-1}P(P'A_\theta^{-1}P)^{-1}P'A_\theta^{-1}\big)\,\mathrm{f}\,\big)\;. \end{aligned} \tag{7.7}$$

7.3 Comparing CV and ML in the Statistical Context

In this section we shall assume that the function f to be interpolated is a sample path of a second order Gaussian RF $(X_t)_{t\in T}$. Then both of the above methods can be used to identify the (vector of) covariance parameter(s) θ and one may ask which of them is more efficient. We show that the MLE and the LOOCV procedure with l_2-norm for the error vector (from now denoted by CV2) both fit into the framework of unbiased estimating functions and we give a criterion that allows for a theoretical comparison of estimation procedures belonging to this very general class. We carry out a simulation study to assess the quality of these theoretically motivated precision measures for MLE and CV2, and we compare the performance of MLE, CV1 and CV2 with respect to their ability to select a parameter value that yields a good interpolate.

7.3.1 Estimating functions and information criteria

Consider the MLE with the log likelihood (7.4) (to keep calculations simple we only consider the zero mean case here). Throughout this subsection we shall assume the following regularity conditions

(a) K_θ is twice differentiable w.r.t. θ

(b) differentiation and integration can be interchanged in

$$\mathbb{E}\big(\,\Lambda(\vartheta; X_{\mathcal{T}})\,\Lambda(\vartheta; X_{\mathcal{T}})'\,\big)$$

(the score function Λ is defined below).

(c) for all $G \in \mathcal{G}$ differentiation and integration can be interchanged in

$$\mathbb{E}\big(\,G(\vartheta; X_{\mathcal{T}})\,\Lambda(\vartheta; X_{\mathcal{T}})'\,\big) \quad \text{and} \quad \mathbb{E}\big(\,\Lambda(\vartheta; X_{\mathcal{T}})\,G(\vartheta; X_{\mathcal{T}})'\,\big)$$

($\mathcal{G}$ will be a class of unbiased estimating functions, details are given later).

Assume in addition that $-l(\vartheta; \mathrm{f})$ is convex at least in a neighbourhood of the true parameter ϑ_0. Then maximizing $l(\vartheta; \mathrm{f})$ is equivalent to finding the root of the so called score function $\Lambda(\vartheta; \mathrm{f}) := \nabla l(\vartheta; \mathrm{f})$ with components

$$\begin{aligned}
\Lambda_v(\vartheta; \mathrm{f}) &= -\tfrac{n}{2v} + \tfrac{1}{2v^2}\,\mathrm{f}' A_\theta^{-1}\mathrm{f} \\
\Lambda_{\theta_1}(\vartheta; \mathrm{f}) &= -\tfrac{1}{2}\,\mathrm{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_1}A_\theta\right) + \tfrac{1}{2v}\,\mathrm{f}' A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_1}A_\theta\,A_\theta^{-1}\mathrm{f} \\
\vdots \qquad & \qquad\qquad \vdots \qquad\qquad\qquad \vdots \\
\Lambda_{\theta_p}(\vartheta; \mathrm{f}) &= -\tfrac{1}{2}\,\mathrm{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_p}A_\theta\right) + \tfrac{1}{2v}\,\mathrm{f}' A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_p}A_\theta\,A_\theta^{-1}\mathrm{f}\,.
\end{aligned}$$

The CV2, if the (squared) l_2-norm of the vector ε_θ of LOOCV errors is minimized, can also be represented in this form. As with the MLE we only consider the zero mean case. Then, by Proposition 7.1.1, we have $\varepsilon_\theta = D_\theta A_\theta^{-1}\mathrm{f}$ where D_θ is a diagonal matrix with diagonal elements

$$(D_\theta)_{ii} = \left(e_i' A_\theta^{-1} e_i\right)^{-1}, \qquad i = 1, \ldots, n.$$

Now, the squared l_2-norm of ε_θ can be written as

$$\|\varepsilon_\theta\|^2 \;=\; \mathrm{f}' A_\theta^{-1} D_\theta^2 A_\theta^{-1}\mathrm{f},$$

and by equating the gradient of $\|\varepsilon_\theta\|^2$ to 0 we obtain θ_{CV2} as the root of the p-variate function $G(\theta;\mathrm{f})$ with components

$$\begin{array}{rcl} G_{\theta_1}(\theta;\mathrm{f}) & = & -2\,\mathrm{f}' A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_1} A_\theta\, A_\theta^{-1} D_\theta^2\, A_\theta^{-1}\mathrm{f} \;+\; 2\,\mathrm{f}' A_\theta^{-1}\, D_\theta\,\tfrac{\partial}{\partial\theta_1}\, D_\theta A_\theta^{-1}\mathrm{f}\,, \\ \vdots & & \vdots \qquad\qquad\qquad\qquad\qquad \vdots \\ G_{\theta_p}(\theta;\mathrm{f}) & = & -2\,\mathrm{f}' A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_p} A_\theta\, A_\theta^{-1} D_\theta^2\, A_\theta^{-1}\mathrm{f} \;+\; 2\,\mathrm{f}' A_\theta^{-1}\, D_\theta\,\tfrac{\partial}{\partial\theta_p}\, D_\theta A_\theta^{-1}\mathrm{f}\,. \end{array} \tag{7.8}$$

Such a function G of both the data and the parameter θ defining an estimate of θ as its root is called estimating function. The score function encountered above in the maximum likelihood context is a special case. An additional property of estimating function that is often required is unbiasedness, i.e.

$$\mathbb{E}_{\vartheta_0}\big(G(\theta_0;X_\mathcal{T})\big) \;=\; \mathbf{0},$$

where $\mathbb{E}_{\vartheta_0}$ denotes the expectation under the probability measure corresponding to ϑ_0. Such a property is reasonable as it ensures that θ_0 is the root of G at least in expectation. Using standard results on the expectation of quadratic forms (cf. [37, p. 55]) it is easy to verify that both score function and the above estimating function corresponding to CV2 are unbiased (independently of the assumption of a multivariate Gaussian distribution).

The LOOCV error we defined in section 7.1 is completely independent of the parameter υ. In spatial statistics however, we are interested in identifying υ as well in order to calculate the kriging error variance. In the LOOCV framework an estimator for υ can be derived as follows:

Using representation (7.1) of the LOOCV errors we find that

$$\mathbb{E}_\vartheta(\varepsilon_k) \;=\; 0, \qquad \mathbb{E}_\vartheta\big(\varepsilon_k^2\big) \;=\; (\Psi_{kk})^{-2}\cdot e_k' A_\theta^{-1}\, \underbrace{\mathbb{E}_\vartheta\big(X_\mathcal{T} X_\mathcal{T}'\big)}_{=\,\upsilon A_\theta}\, A_\theta^{-1} e_k \;=\; \upsilon\,(\Psi_{kk})^{-1}\,,$$

and hence, for given θ, a reasonable estimator of υ is given by

$$\upsilon_{\text{CV2}}(\theta) \;:=\; \frac{1}{n}\sum_{i=1}^{n} \Psi_{ii}\cdot\varepsilon_i^2 \;=\; \tfrac{1}{n}\,\mathrm{f}' A_\theta^{-1} D_\theta A_\theta^{-1}\mathrm{f}\,,$$

where we have used that in the simple kriging case $\Psi_{kk} = (A_\theta^{-1})_{kk} = (D_\theta)_{kk}^{-1}$.

The estimator $\upsilon_{\text{CV2}}(\theta)$ of υ is clearly not optimal in the statistical sense since it ignores the correlations between the different components of ε. Taking these into account would lead to the estimator $\upsilon_{\text{MLE}}(\theta)$ obtained by the maximum likelihood principle. However, $\upsilon_{\text{CV2}}(\theta)$ is more in the spirit of LOOCV since θ_{CV2} does not account for these correlations either (see discussion in Section 7.4). Moreover it will be easier to give an interpretation to $\upsilon_{\text{CV2}}(\theta)$ also in the framework of Numerical Analysis (Section 7.5). Now, by adding a further component

$$G_\upsilon(\vartheta;\mathrm{f}) \;=\; \tfrac{1}{n}\,\mathrm{f}' A_\theta^{-1} D_\theta A_\theta^{-1}\mathrm{f} - \upsilon \tag{7.9}$$

to (7.8) we obtain an unbiased, $(p{+}1)$-variate estimating function $G(\vartheta;\mathrm{f})$ for the parameters υ and θ.

Within the framework of unbiased estimating functions, measures of information can be defined that allow to compare the performance of estimators. Following [39, p. 174] we start by defining the Fisher information

$$\mathcal{I}(\vartheta_0) \;:=\; \mathbb{E}_{\vartheta_0}\big(\Lambda(\vartheta_0; X_{\mathcal{T}})\,\Lambda(\vartheta_0; X_{\mathcal{T}})'\,\big)\,. \tag{7.10}$$

Under the regularity conditions (a) and (b) stated above it holds that

$$\mathcal{I}(\vartheta_0) \;=\; -\,\mathbb{E}_{\vartheta_0}\big(J_\Lambda(\vartheta_0; X_{\mathcal{T}})\,\big)\,, \tag{7.11}$$

where J_Λ denotes the Jacobi matrix of Λ. Now, if $\mathcal{I}(\vartheta_0)$ is "large" (in the sense that the smallest eigenvalue is large) and $\mathcal{I}(\vartheta_0)^{-1}\,J_\Lambda(\vartheta_0; X_{\mathcal{T}}) \approx I_n$ with high probability, then standard asymptotic theory (cf. [14], [21]) suggests that

$$\vartheta_{\text{MLE}} \;\overset{\text{approx.}}{\sim}\; \mathcal{N}\big(\vartheta_0, \mathcal{I}(\vartheta_0)^{-1}\big)\,. \tag{7.12}$$

In spatial statistics it is often difficult to prove rigorously that MLEs do indeed have this behaviour. In fact, there are even different asymptotic frameworks to which one can appeal (cf. e.g. [45]): increasing domain asymptotics, in which the minimum distance between sampling points is bounded away from zero and thus the spatial domain of observation is unbounded, and fixed domain asymptotics in which observations are taken ever more densely in a fixed and bounded domain. While under increasing domain asymptotics conditions are known (cf. [26]) that ensure that ϑ_{MLE} converges to ϑ_0 a.s. (such estimators are called consistent) with asymptotic distribution as indicated above, it was shown by several authors that under fixed domain asymptotics even consistency is not ensured for all parameters (cf. [44] for such a result on the Whittle-Matérn class). For a fixed set of observations it is not clear a priori which of these two frameworks is is adequate. Some answers to that can be found in [45]. Despite all these difficulties, we note that the Fisher information *may* give a good indication about the accuracy of the MLE and in our simulation study presented later in this section we shall calculate it and compare it with the empirical results from the simulated estimates.

The Fisher information was motivated by the asymptotic theory for MLEs. If instead of the score function Λ we consider some unbiased estimating function G, then we can define the information criterion

$$\mathcal{E}_G(\vartheta_0) \;:=\; \big(W(\vartheta_0)\big)'\,\big(H(\vartheta_0)\big)^{-1}\,\big(W(\vartheta_0)\big)\,, \tag{7.13}$$

where

$$\begin{aligned} H(\vartheta_0) &= \mathbb{E}_{\vartheta_0}\big(G(\vartheta_0; X_{\mathcal{T}})\,G(\vartheta_0; X_{\mathcal{T}})'\,\big) \quad \text{and} \\ W(\vartheta_0) &= -\,\mathbb{E}_{\vartheta_0}\big(J_G(\vartheta_0; X_{\mathcal{T}})\,\big)\,, \end{aligned}$$

which is a natural generalization of the Fisher information and will allow us to compare the MLE with the CV2. Note that multiplying any component of G with an arbitrary non-zero constant does not change $\mathcal{E}_G(\vartheta_0)$, which is an important property since these operations do not change the root of G either. The role of $\mathcal{E}_G(\vartheta_0)$ as an information measure can be motivated as follows ([20, Sec. 1.3]):

1. Since the parameter estimate is given by the root of G, $G(\vartheta_0)$ should be as close to zero as possible. Thus, the “smaller“ $H(\vartheta_0)$ the more accurate the corresponding estimator can be expected to be.

2. On the other hand, the gradients $\nabla G_\upsilon, \nabla G_{\theta_1}, \ldots, \nabla G_{\theta_p}$ should be as steep as possible because then the root of G will be somewhere near ϑ_0 if $G(\vartheta_0)$ does not differ from zero too much. Thus the “bigger” $W(\vartheta_0)$ the more accurate the corresponding estimator can be expected to be.

If G depends on more then one parameter, “small” $H(\vartheta_0)$ and “big” $W(\vartheta_0)$ are understood in the sense that all eigenvalues of H and W should be small or big respectively.

In this general framework of unbiased estimating functions there are also a number of theorems (see [20, Ch. 12]) providing conditions subject to which the corresponding estimators are consistent and asymptotically normal, so that for n large enough we have

$$\vartheta_{\mathrm{G}} \overset{\text{approx.}}{\sim} \mathcal{N}\big(\vartheta_0, \mathcal{E}_G(\vartheta_0)^{-1}\big)\,. \tag{7.14}$$

Although in our situation these are even harder to verify than those for the MLE framework, $\mathcal{E}_G(\vartheta_0)$ appears to be a reasonable information measure. It is used by several authors (e.g. [38], [7]) to compare the accuracy of estimators, and so shall we in the following discussion.

In our situation of estimating the covariance parameters of a zero mean Gaussian RF the entries of the $(1+p)\times(1+p)$ Fisher information matrix $\mathcal{I}(\vartheta)$ are

$$\begin{aligned}
\mathcal{I}_{\upsilon\upsilon}(\vartheta) &= \tfrac{1}{2}\,\tfrac{n}{\upsilon^2} \qquad\qquad (7.15)\\
\mathcal{I}_{\upsilon\theta_j}(\vartheta) &= \mathcal{I}_{\theta_j\upsilon}(\vartheta) = \tfrac{1}{2\upsilon}\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\right),\\
\mathcal{I}_{\theta_k\theta_j}(\vartheta) &= \tfrac{1}{2}\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\right).
\end{aligned}$$

These formulae follow from the general formula given in [39, p. 179]. The entries of the $(1+p)\times(1+p)$ matrices H and W that define the information criterion for the CVE with estimating functions according to (7.8), (7.9) are

$$\begin{aligned}
H_{\upsilon\upsilon}(\vartheta) &= \tfrac{2\upsilon^2}{n^2}\operatorname{tr}\left(A_\theta^{-1}D_\theta\,A_\theta^{-1}D_\theta\right), \qquad\qquad (7.16)\\
H_{\upsilon\theta_j}(\vartheta) &= -\tfrac{4\upsilon^2}{n}\operatorname{tr}\left(A_\theta^{-1}D_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right) + \tfrac{4\upsilon^2}{n}\operatorname{tr}\left(A_\theta^{-1}D_\theta\,A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right)\\
H_{\theta_k\theta_j}(\vartheta) &= 4\upsilon^2\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right)\\
&\quad + 4\upsilon^2\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\right)\\
&\quad - 8\upsilon^2\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right)\\
&\quad - 8\upsilon^2\operatorname{tr}\left(A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_k}D_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right)\\
&\quad + 8\upsilon^2\operatorname{tr}\left(A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_k}D_\theta\,A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right),
\end{aligned}$$

($H_{\theta_k \upsilon}(\vartheta)$ is determined by symmetry of H) and

$$\begin{aligned} W_{\upsilon\upsilon}(\vartheta) &= -1, \quad W_{\theta_k \upsilon}(\vartheta) \;=\; 0, \\ W_{\upsilon\theta_j}(\vartheta) &= -\tfrac{2\upsilon}{n}\,\mathrm{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta\right) \,+\, \tfrac{\upsilon}{n}\,\mathrm{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}D_\theta\right), \\ W_{\theta_k\theta_j}(\vartheta) &= 2\upsilon\,\mathrm{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right) \,-\, 2\upsilon\,\mathrm{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right). \end{aligned} \tag{7.17}$$

The formulae for $H(\vartheta)$ are obtained by symmetrizing the first terms of $G_{\theta_1}, \ldots, G_{\theta_1}$ and using the formula

$$\mathrm{Cov}\big(Z'AZ, Z'BZ\big) \;=\; 2\,\mathrm{tr}\,(AVBV)\,, \qquad \text{if } \; Z \sim \mathcal{N}(\mathbf{0}, V)\,,$$

for quadratic forms $Z'AZ$ and $Z'BZ$ where Z is a n-variate RVct and $A, B \in \mathbb{R}^{n\times n}$ are symmetric matrices (see [37, p. 66]).

To obtain formulae for $W(\vartheta)$ we first calculate the derivatives of $G(\vartheta; \mathrm{f})$, starting with $\frac{\partial}{\partial \upsilon}$:

$$\tfrac{\partial}{\partial\upsilon}G_\upsilon(\vartheta;\mathrm{f}) \;=\; -1, \qquad \tfrac{\partial}{\partial\upsilon}G_{\theta_k}(\vartheta;\mathrm{f}) \;=\; 0\,, \quad j = 1, \ldots, p.$$

Next, we have for $j = 1, \ldots, p$:

$$\tfrac{\partial}{\partial\theta_j}G_\upsilon(\vartheta;\mathrm{f}) \;=\; -\tfrac{2}{n}\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta\,A_\theta^{-1}\mathrm{f} \,+\, \tfrac{1}{n}\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}D_\theta\,A_\theta^{-1}\mathrm{f}\,,$$

and finally, for $j, k = 1, \ldots, p$:

$$\begin{aligned} \tfrac{\partial}{\partial\theta_j}G_{\theta_k}(\vartheta;\mathrm{f}) \;=\;& 2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\mathrm{f} \,-\, 2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\mathrm{f} \\ & + 2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\mathrm{f} \,-\, 4\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\,A_\theta^{-1}\mathrm{f} \\ & + 2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}\mathrm{f} \,-\, 2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_k}D_\theta\,A_\theta^{-1}\mathrm{f} \\ & + 2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}D_\theta\,\tfrac{\partial}{\partial\theta_k}D_\theta\,A_\theta^{-1}\mathrm{f} \,+\, 2\,\mathrm{f}'A_\theta^{-1}D_\theta\,\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}D_\theta\,A_\theta^{-1}\mathrm{f} \\ & - 2\,\mathrm{f}'A_\theta^{-1}D_\theta\,\tfrac{\partial}{\partial\theta_k}D_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}\mathrm{f} \end{aligned}$$

With respect to the derivatives of D_θ we note

$$\begin{aligned} \big(\tfrac{\partial}{\partial\theta_k}D_\theta\big)_{ii} \;&=\; \frac{e_i'\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}e_i}{(e_i'\,A_\theta^{-1}e_i)^2} \\ \big(\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}D_\theta\big)_{ii} \;&=\; -2\,\frac{e_i'\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}e_i}{(e_i'\,A_\theta^{-1}e_i)^2} \;+\; \frac{e_i'\,A_\theta^{-1}\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}A_\theta\,A_\theta^{-1}e_i}{(e_i'\,A_\theta^{-1}e_i)^2} \\ & \qquad + 2\,\frac{e_i'\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}e_i \,\cdot\, e_i'\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}e_i}{(e_i'\,A_\theta^{-1}e_i)^3}\,. \end{aligned}$$

Now, by application of the formula

$$\mathbb{E}\big(Z'AZ\big) \;=\; \mathrm{tr}\,(AV)\,, \qquad \text{if } \; Z \sim \mathcal{N}(\mathbf{0}, V)\,, \;\; \text{and} \;\; A \in \mathbb{R}^{n\times n}$$

(see [37, p. 55]) and by using the identity

$$\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\,B\right) \;=\; \sum_{i=1}^{n} e_i'\,A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}e_i\;\cdot\;B_{ii} \;=\; \operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,B\right),$$

which hold for any diagonal matrix $B \in \mathbb{R}^{n\times n}$, we obtain

$$\begin{aligned}
\mathbb{E}_\vartheta\big(\tfrac{\partial}{\partial\theta_j}G_{\theta_k}(\vartheta;X_{\mathcal{T}})\big) \;=\;& 2\upsilon\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right) \;-\; 2\upsilon\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\right)\\
&+2\upsilon\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right) \;-\; 4\upsilon\,\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right)\\
&+2\upsilon\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\,A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_j}A_\theta\right) \;-\; 2\upsilon\,\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right)\\
&+2\upsilon\,\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right) \;+\; 2\upsilon\,\operatorname{tr}\left(A_\theta^{-1}D_\theta\,\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}D_\theta\right)\\
&-2\upsilon\,\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right)
\end{aligned}$$

Noting that

$$\begin{aligned}
\operatorname{tr}\left(A_\theta^{-1}D_\theta\,\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}D_\theta\right) \;=\;& \operatorname{tr}\left(\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}D_\theta\right)\\
\;=\;& -\,2\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\right) \;+\; \operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial^2}{\partial\theta_j\partial\theta_k}A_\theta\,A_\theta^{-1}D_\theta^2\right)\\
&+2\,\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right)
\end{aligned}$$

and combining all the terms finally yields $W_{\theta_k\theta_j}(\vartheta)$ as stated above:

$$W_{\theta_k\theta_j}(\vartheta) \;=\; 2\upsilon\,\operatorname{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_k}A_\theta\,A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_j}A_\theta\,A_\theta^{-1}D_\theta^2\right) \;-\; 2\upsilon\,\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_k}D_\theta\,\tfrac{\partial}{\partial\theta_j}D_\theta\right).$$

Note that if we work with the inverse of $\mathcal{E}_G(\vartheta_0)$, then the block with the entries associated with the estimation of θ is independent of the information on the estimation of υ, and is therefore not influenced by our particular choice of $\upsilon_{\mathrm{CV2}}(\theta)$. Indeed, since $H(\vartheta)$ and $W(\vartheta)$ are of the form

$$W(\vartheta) \;=\; \begin{pmatrix} W_{\upsilon\upsilon} & W_{\upsilon\theta}\\ \mathbf{0} & W_{\theta\theta}\end{pmatrix}, \qquad H(\upsilon,\theta) \;=\; \begin{pmatrix} H_{\upsilon\upsilon} & H_{\upsilon\theta}\\ H_{\upsilon\theta}' & H_{\theta\theta}\end{pmatrix},$$

we obtain by applying standard rules for the inversion of block matrices

$$\big(W(\vartheta)\big)^{-1} \;=\; \begin{pmatrix} W_{\upsilon\upsilon}^{-1} & -W_{\upsilon\upsilon}^{-1}\,W_{\upsilon\theta}\,W_{\theta\theta}^{-1}\\ \mathbf{0} & W_{\theta\theta}^{-1}\end{pmatrix}$$

(note that $W_{\theta\theta}$ is symmetric). Using this and writing "$*$" for convenience for all terms that are not relevant for our claim we find

$$\begin{aligned}
\big(\mathcal{E}_G(\vartheta_0)\big)^{-1} \;=\;& \begin{pmatrix} * & *\\ \mathbf{0} & W_{\theta\theta}^{-1}\end{pmatrix}\begin{pmatrix} H_{\upsilon\upsilon} & H_{\upsilon\theta}\\ H_{\upsilon\theta}' & H_{\theta\theta}\end{pmatrix}\begin{pmatrix} * & \mathbf{0}\\ * & W_{\theta\theta}^{-1}\end{pmatrix}\\
\;=\;& \begin{pmatrix} * & *\\ * & W_{\theta\theta}^{-1}\,H_{\theta\theta}\,W_{\theta\theta}^{-1}\end{pmatrix},
\end{aligned}$$

and so the block with the information on θ is the same as if we had only considered θ_{CV2} from the beginning.

Denote by $\mathcal{G}$ the class of all unbiased estimating functions that we consider as potential candidates for estimating υ and θ. An estimating function $G^* \in \mathcal{G}$ is called O_F-optimal (fixed sample optimal) if

$$\mathcal{E}_{G^*}(\vartheta) - \mathcal{E}_G(\vartheta)$$

is nonnegative definite for all $G \in \mathcal{G}$ and all $\vartheta \in \mathbb{R}_+ \times \Theta$. If $\Lambda(\vartheta; \mathrm{f})$ belongs to $\mathcal{G}$ then it is (subject to the regularity conditions stated at the beginning of this subsection) the O_F-optimal estimating function in $\mathcal{G}$ (see [20, Ch. 2]).

Consequently, under the assumption that $(X_t)_{t\in T}$ is a zero mean Gaussian RF with covariance function υK_θ it follows that MLE is superior to CV2 in the sense of O_F-optimality. The following result ([20, Thm. 8.1]) claims O_F-optimality within a very general class of estimating functions also for the REML estimator.

As above, let f be a vector of observations of $(X_t)_{t\in T}$ at locations $\{t_1, \ldots, t_n\} \subset T$, and let $X_{\mathcal{T}} := (X_{t_1}, \ldots, X_{t_n})'$ the vector of the corresponding RVs. Assuming a mean function

$$m(t) \;=\; \sum_{k=1}^{q} \beta_k \, p_k(t)$$

we have $\mathbb{E}(X_{\mathcal{T}}) = P\beta$ with P as in (6.6), assumed to have full rank. For an arbitrary matrix $Q \in \mathbb{R}^{n\times n}$ with rank $n-q$ and $QP = \mathbf{0}$ we consider the contrasts

$$\tilde{\mathrm{f}} \;:=\; Q\mathrm{f} \quad \text{and} \quad \tilde{X}_{\mathcal{T}} \;:=\; QX_{\mathcal{T}}.$$

For the covariance of $\tilde{X}_{\mathcal{T}}$ we set $\mathrm{Cov}(\tilde{X}_{\mathcal{T}}) = \upsilon A_\theta =: V_\vartheta$. In the expectation that it is quadratic forms of the data that should be used to estimate covariance parameters we then consider the class of (unbiased) estimating functions

$$\mathcal{G}_0 \;=\; \Big\{ G = (G_1, \ldots, G_{p+1})' \;:\; G_k(\vartheta; \tilde{\mathrm{f}}) \;=\; \tilde{\mathrm{f}}'\, S_k\, \tilde{\mathrm{f}} \;-\; \mu_{S_k}\,,\quad 1 \le k \le p+1 \Big\},$$

where $\mu_{S_k} = \mathbb{E}\big(\tilde{X}'_{\mathcal{T}} S_k \tilde{X}_{\mathcal{T}}\big)$.

Theorem 7.3.1. *Assume that $X_{\mathcal{T}} \sim \mathcal{N}(P\beta, V_\vartheta)$. Then G^* is an O_F-optimal estimating function in $\mathcal{G}_0$ if*

$$S^*_k \;=\; (QV_\vartheta Q')^-\, \big(Q\, \tfrac{\partial}{\partial\vartheta_j} V_\vartheta\, Q'\big)\, (QV_\vartheta Q')^-, \qquad 1 \le k \le p+1\,,$$

for any g-inverse $(QV_\vartheta Q')^-$. Furthermore, the S^*_k do not depend on Q.

In order to see the connection to REML, consider the derivatives of the restricted log likelihood function (7.6) :

$$\begin{aligned}
\tfrac{\partial}{\partial\upsilon}\, l(\vartheta; \mathrm{f}) \;&=\; -\tfrac{n-q}{2}\,\tfrac{1}{\upsilon} \;+\; \tfrac{1}{2\upsilon^2}\, \mathrm{f}'\, A_\theta^{-1}\, \Pi_P\, \mathrm{f},\\
\tfrac{\partial}{\partial\theta_k}\, l(\vartheta; \mathrm{f}) \;&=\; -\tfrac{1}{2}\,\mathrm{tr}\left(\Pi_{\bar{P}}\, \tfrac{\partial}{\partial\theta_k} A_\theta\, A_\theta^{-1}\right) \;+\; \tfrac{1}{2\upsilon}\, \mathrm{f}'\, A_\theta^{-1}\, \Pi_{\bar{P}}\, \tfrac{\partial}{\partial\theta_k} A_\theta\, A_\theta^{-1}\, \Pi_{\bar{P}}\, \mathrm{f}
\end{aligned}$$

where $\Pi_{\bar{P}} := I_n - P\big(P' A_\theta^{-1} P\big)^{-1} P' A_\theta^{-1}$ is a projector on the orthogonal complement of the range space of the matrix P. Using the relation

$$\tfrac{1}{\upsilon}\, A_\theta^{-1}\, \Pi_{\bar{P}} \;=\; Q'(QV_\vartheta Q')^{-}Q$$

(see [20, eq. (8.2)]) and using that $\Pi_{\bar{P}}$ is idempotent we obtain

$$\begin{aligned} \tilde{\mathrm{f}}'\, S_k^*\, \tilde{\mathrm{f}} \;&=\; \tfrac{1}{\upsilon^2}\, \mathrm{f}'\, A_\theta^{-1}\, \Pi_{\bar{P}}\, \tfrac{\partial}{\partial\vartheta_k} V_\vartheta\, A_\theta^{-1}\, \Pi_{\bar{P}}\, \mathrm{f}\,, \\ \mu_{S_k^*} \;&=\; \tfrac{1}{\upsilon^2}\, \mathbb{E}\big(\tilde{X}_{\mathcal{T}}'\, A_\theta^{-1}\, \Pi_{\bar{P}}\, \tfrac{\partial}{\partial\vartheta_k} V_\vartheta\, A_\theta^{-1}\, \Pi_{\bar{P}}\, \tilde{X}_{\mathcal{T}}\big) \;=\; \tfrac{1}{\upsilon}\, \mathrm{tr}\left(\Pi_{\bar{P}}\, \tfrac{\partial}{\partial\vartheta_k} V_\vartheta\, A_\theta^{-1}\right). \end{aligned}$$

Resubstituting $V_\vartheta = \upsilon A_\theta$, noting that $\mathrm{tr}\,(\Pi_{\bar{P}}) = n - q$, and comparing the resulting estimating function with $\nabla\, l(\vartheta; \mathrm{f})$ finally yields

Corollary 7.3.2. *Under the assumptions of Theorem 7.3.1 the REML estimator is O_F-optimal among all estimators corresponding to an estimating function in $\mathcal{G}_0$.*

Remark 7.3.3. The assumption in Theorem 7.3.1 of $X_{\mathcal{T}}$ that follows a multivariate Gaussian distribution can be weakened to certain assumptions on some third and fourth moments of

$$V_\vartheta^{-1/2}\big(X_{\mathcal{T}} - P\beta\big).$$

Due to this transformation with $V_\vartheta^{-1/2}$ these conditions are not very transparent in our context of RFs. Nevertheless we note that (RE)ML can make sense even if the distribution assumption under which they were derived do not hold.

7.3.2 Accuracy of parameter estimates

We present the results of a simulation study in which we compare the performance of MLE and CV2. We simulated 300 centred stationary Gaussian RFs with covariance function $\upsilon\Phi_{r,\nu}$ as defined in (6.33).

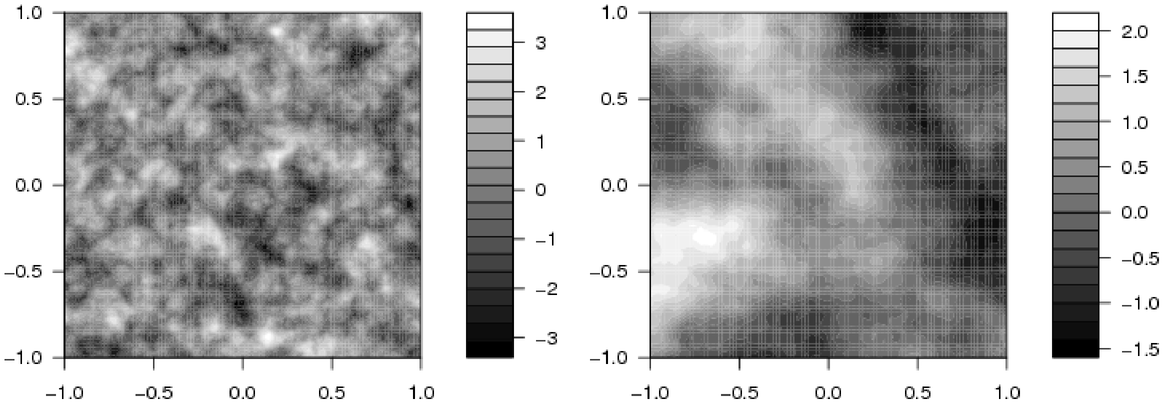

Figure 7.1: Examples of simulated sample paths of a Gaussian RF with covariance function $\Phi_{r_0,1}$ with scale parameter $r_0 = 0.1$ (left) and $r_0 = 1$ (right).

The simulation was carried out with parameter values $\upsilon_0 = 1$, $\nu_0 = 1$ for two different scale parameters $r_0 = 0.1$ and $r_0 = 1$ on an equidistant 100×100 grid $\mathcal{Q} \subset [-1,1]^2$ using the R-package “RandomFields” (cf. [36]). According to the results of Section 5.3 the corresponding sample paths are just barely not differentiable. One of the respective 300 realizations is visualized in Figure 7.1 to illustrate the different structures that are observed due to the different scaling.

First, we compare the ability of MLE and CV2 to estimate the covariance parameters υ, r and ν based on $n = 100, 200, 300, 400, 500$ observations of the respective sample path. The n sampling locations are chosen randomly from the simulation grid $\mathcal{Q}$ (identical probabilities, no replacement) such that

$$\{t_1, \dots, t_{100}\} \subset \ldots \subset \{t_1, \dots, t_{500}\} \subset \mathcal{Q}.$$

This way it is ensured that the information strictly increases. These sets of locations are then used for both choices of r_0 and all 300 respective realizations.

For the moment, we focus on the estimation of r and ν, a discussion of the estimation and use of υ will follow later in this subsection. Figures 7.2 and 7.3 show plots of the estimated values of r against ν obtained by MLE and CV2. Especially for small n both estimators sometimes yield very large estimates of ν (typically when $r_0 = 0.1$) or very large estimates of r (typically when $r_0 = 1$). Parameter estimation was carried out subject to the constraints $r \leq 40$ and $\nu \leq 16$ and in particular the CV2 estimates for $r_0 = 1$ attain these bounds quite often, even for bigger values of n.

With respect to the asymptotic approximations (7.12) and (7.14) to the distribution of the estimates it can be suspected that they may describe the behaviour of the MLE quite well for large n. The same may be true, to a lesser extent, for the CV2 in the case where $r_0 = 0.1$. For $r_0 = 1$, however, the distribution of the CV2 estimates seems far from being multivariate Gaussian, and we will see that also the dispersion of $(r, \nu)_{\mathrm{CV2}}$ around (r_0, ν_0) is only poorly described by the information criterion derived above.

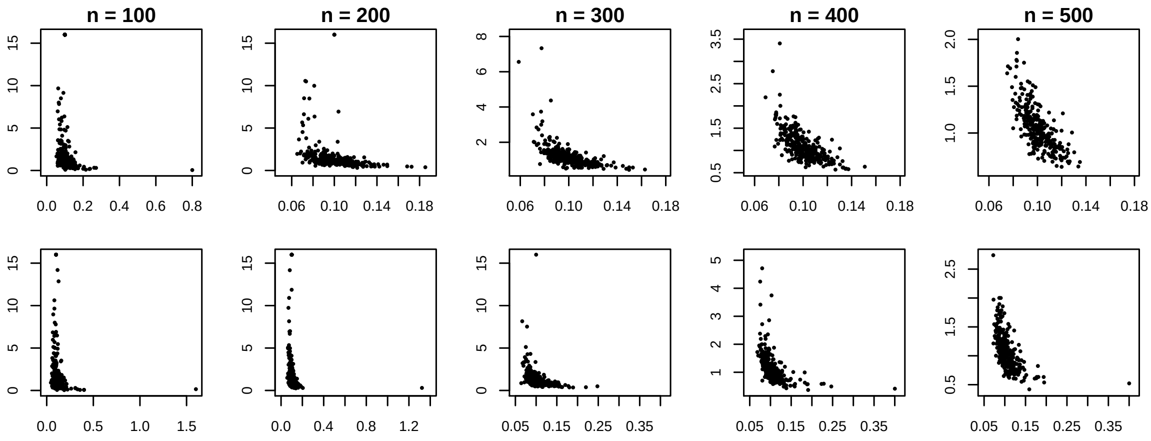

Figure 7.2: Plots of the estimates of r (x-axis) against ν (y-axis) obtained by MLE (top row) and CV2 (bottom row) for the 300 sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$.

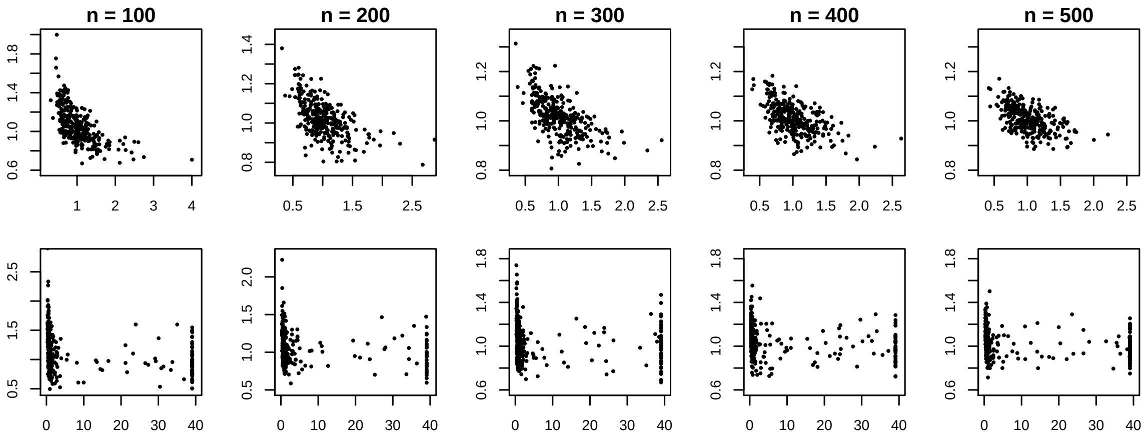

Figure 7.3: Plots of the estimates of r (x-axis) against ν (y-axis) obtained by MLE (top row) and CV2 (bottom row) for the 300 sample paths simulated with $\vartheta_0 = (1, 1, 1)$.

In order to measure the precision of the estimates quantitatively, we calculate

1. the empirical means ($r_{\star,i}$ denotes the i^{th} estimate of r with procedure $\star$)

$$\overline{r}_\star \;:=\; \tfrac{1}{300} \sum_{i=1}^{300} r_{\star,i}, \quad \overline{\nu}_\star \;:=\; \tfrac{1}{300} \sum_{i=1}^{300} \nu_{\star,i}, \qquad \star = \text{MLE, CV2},$$

2. the empirical mean squared errors for $\star = \text{MLE, CV2}$ defined by

$$\text{MSE}(r_\star) \;:=\; \tfrac{1}{300} \sum_{i=1}^{300} (r_{\star,i} - r_0)^2, \qquad \text{MSE}(\nu_\star) \;:=\; \tfrac{1}{300} \sum_{i=1}^{300} (\nu_{\star,i} - \nu_0)^2,$$

3. the diagonal elements of the inverse Fisher information

$$(\mathcal{I}(\vartheta_0)^{-1})_{rr}, \quad \text{and} \quad (\mathcal{I}(\vartheta_0)^{-1})_{\nu\nu},$$

4. the diagonal elements of the inverse of the information criterion for CV2

$$(\mathcal{E}_G(\vartheta_0)^{-1})_{rr}, \quad \text{and} \quad (\mathcal{E}_G(\vartheta_0)^{-1})_{\nu\nu}.$$

$\mathcal{I}(\vartheta_0)$ and $\mathcal{E}_G(\vartheta_0)$ can be calculated numerically using the formulae (7.15)-(7.17).

The empirical means of the 300 parameter estimates are given in Table 7.1. As could be expected from the plots, in the simulations with $r_0 = 0.1$ both $\overline{r}_{\text{MLE}}$ and $\overline{r}_{\text{CV2}}$ are quite close to r_0, while $\overline{\nu}_{\text{MLE}}$ and $\overline{\nu}_{\text{CV2}}$ are considerably larger than ν_0 for small n but approach the true value as n increases. In the simulations with $r_0 = 1$ both $\overline{\nu}_{\text{MLE}}$ and $\overline{\nu}_{\text{CV2}}$ are reasonably close to ν_0 even for small n and get even closer as n increases. The same is true for the ML estimates of r_0 but not for the CV estimates, which are substantially bigger than r_0 for all n and do not show any tendency of convergence. Without the constraint $r_{\text{CV2}} \leq 40$ (imposed for computational reasons) this overestimation of r_0 would presumably be even more dramatic.

n	$\overline{r}_{\text{MLE}}$	$\overline{r}_{\text{CV2}}$	$\overline{\nu}_{\text{MLE}}$	$\overline{\nu}_{\text{CV2}}$
100	0.1100	0.1205	3.693	1.885
200	0.0999	0.1077	1.609	2.155
300	0.1011	0.1051	1.192	1.338
400	0.1008	0.1065	1.007	1.152
500	0.1009	0.1064	1.058	1.067

n	$\overline{r}_{\text{MLE}}$	$\overline{r}_{\text{CV2}}$	$\overline{\nu}_{\text{MLE}}$	$\overline{\nu}_{\text{CV2}}$
100	1.041	10.25	1.054	1.119
200	1.019	7.23	1.023	1.066
300	1.020	8.80	1.016	1.046
400	1.017	9.28	1.011	1.033
500	1.017	11.96	1.006	1.025

Table 7.1: Empirical means of the respective 300 parameter estimates for r and ν for the sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$ (left) and $\vartheta_0 = (1, 1, 1)$ (right) respectively.

Tables 7.2 and 7.3 show the empirical mean squared errors of the respective 300 parameter estimates by ML and CV2 and the corresponding entries of the inverse information matrices.

In both cases $r_0 = 0.1$ and $r_0 = 1$, the dispersion of $(r, \nu)_{\text{MLE}}$ around (r_0, ν_0) is predicted reasonably well by the inverse Fisher information if n is large. For small n, the MSE is dominated by the large deviation of $\overline{r}_{\text{MLE}}$ from r_0, and $\overline{\nu}_{\text{MLE}}$ from ν_0 respectively (see above) but as n increases this bias disappears and (7.12) seems to be an acceptable approximation whatever the intricacy of establishing a rigorous asymptotic theory. Analysing the decrease of $(\mathcal{I}^{-1})_{rr}$ and $(\mathcal{I}^{-1})_{\nu\nu}$ with increasing n we also note that the accuracy of ν_{MLE} improves faster than that of r_{MLE}, particularly so in the case $r_0 = 1$. This could be expected since ν is mainly linked to the local behaviour of the sample paths and so identifying ν_0 should benefit substantially from an increase of the density of sampling locations.

For the CV2 estimates we find that approximation (7.14) of the dispersion of $(r, \nu)_{\text{CV2}}$ around (r_0, ν_0) is inadequate even for $n = 500$. In the case $r_0 = 1$ this could be expected since we already saw that r_{CV2} does not even seem to converge to r_0. For $r_0 = 0.1$ simulation results not presented here show that for a bigger number of sampling locations (at least 800) the estimates behave indeed like realizations from a multivariate normally distributed RV with mean θ_0 and covariance $\mathcal{E}_G^{-1}$.

Nevertheless the information criterion for the CV2 estimating functions can help us to identify the shortcomings of LOOCV from a statistical point of view. Note for instance in the case $r_0 = 1$ that $(\mathcal{E}_G^{-1})_{rr}$ is increasing rather than decreasing as the number of observations grows. The explanation for this very peculiar effect is that more LOOCV-components (due to more data) do not necessarily imply more information:

- The LOOCV-components are correlated, quite strongly so for components belonging to close-by sampling locations. These correlations can lead to a distortion of the estimate and additional observations may for instance enhance an already existing tendency due to the first LOOCV components to select a very large value of r.

- Perhaps more important is that LOOCV only assesses the accuracy of the prediction (at the left-out locations). As noted in Section 6.7, if the sampling locations are very close compared to r_0, prediction is insensitive to deviations from the true scale parameter. The sampling locations are indeed very dense relative to r_0 in the case $r_0 = 1$ and become even denser as n increases and so at the same time the data become less informative about r.

n	$(\mathcal{I}^{-1})_{rr}$	$\mathrm{MSE}(r_{\mathrm{MLE}})$	$(\mathcal{E}_G^{-1})_{rr}$	$\mathrm{MSE}(r_{\mathrm{CV2}})$
100	$9.99 \cdot 10^{-4}$	$26.7 \cdot 10^{-4}$	$12.1 \cdot 10^{-4}$	$99.9 \cdot 10^{-4}$
200	$3.47 \cdot 10^{-4}$	$3.33 \cdot 10^{-4}$	$4.70 \cdot 10^{-4}$	$54.9 \cdot 10^{-4}$
300	$2.15 \cdot 10^{-4}$	$2.4 \cdot 10^{-4}$	$3.41 \cdot 10^{-4}$	$5.74 \cdot 10^{-4}$
400	$1.54 \cdot 10^{-4}$	$1.72 \cdot 10^{-4}$	$2.68 \cdot 10^{-4}$	$8.99 \cdot 10^{-4}$
500	$1.24 \cdot 10^{-4}$	$1.24 \cdot 10^{-4}$	$2.33 \cdot 10^{-4}$	$7.33 \cdot 10^{-4}$

n	$(\mathcal{I}^{-1})_{\nu\nu}$	$\mathrm{MSE}(\nu_{\mathrm{MLE}})$	$(\mathcal{E}_G^{-1})_{\nu\nu}$	$\mathrm{MSE}(\nu_{\mathrm{CV2}})$
100	$8.53 \cdot 10^{-1}$	$361 \cdot 10^{-1}$	$14.3 \cdot 10^{-1}$	$91.7 \cdot 10^{-1}$
200	$2.14 \cdot 10^{-1}$	$44.2 \cdot 10^{-1}$	$4.42 \cdot 10^{-1}$	$113 \cdot 10^{-1}$
300	$1.07 \cdot 10^{-1}$	$5.19 \cdot 10^{-1}$	$2.04 \cdot 10^{-1}$	$16.1 \cdot 10^{-1}$
400	$0.60 \cdot 10^{-1}$	$1.21 \cdot 10^{-1}$	$1.13 \cdot 10^{-1}$	$3.0 \cdot 10^{-1}$
500	$0.41 \cdot 10^{-1}$	$0.57 \cdot 10^{-1}$	$0.76 \cdot 10^{-1}$	$1.08 \cdot 10^{-1}$

Table 7.2: Inverse Fisher information and CV2 information criterion for r and ν when the true parameters are $\vartheta_0 = (1, 0.1, 1)$, and empirical mean squared errors of the corresponding 300 estimates.

n	$(\mathcal{I}^{-1})_{rr}$	$\mathrm{MSE}(r_{\mathrm{MLE}})$	$(\mathcal{E}_G^{-1})_{rr}$	$\mathrm{MSE}(r_{\mathrm{CV2}})$
100	$1.307 \cdot 10^{-1}$	$1.928 \cdot 10^{-1}$	$11.55 \cdot 10^{-1}$	$3286 \cdot 10^{-1}$
200	$1.001 \cdot 10^{-1}$	$1.131 \cdot 10^{-1}$	$13.40 \cdot 10^{-1}$	$2133 \cdot 10^{-1}$
300	$0.895 \cdot 10^{-1}$	$0.961 \cdot 10^{-1}$	$17.85 \cdot 10^{-1}$	$2737 \cdot 10^{-1}$
400	$0.821 \cdot 10^{-1}$	$0.875 \cdot 10^{-1}$	$20.86 \cdot 10^{-1}$	$2797 \cdot 10^{-1}$
500	$0.769 \cdot 10^{-1}$	$0.739 \cdot 10^{-1}$	$23.39 \cdot 10^{-1}$	$3920 \cdot 10^{-1}$

n	$(\mathcal{I}^{-1})_{\nu\nu}$	$\mathrm{MSE}(\nu_{\mathrm{MLE}})$	$(\mathcal{E}_G^{-1})_{\nu\nu}$	$\mathrm{MSE}(\nu_{\mathrm{CV2}})$
100	$2.217 \cdot 10^{-2}$	$3.635 \cdot 10^{-2}$	$8.69 \cdot 10^{-2}$	$13.2 \cdot 10^{-2}$
200	$0.906 \cdot 10^{-2}$	$0.914 \cdot 10^{-2}$	$3.87 \cdot 10^{-2}$	$4.79 \cdot 10^{-2}$
300	$0.551 \cdot 10^{-2}$	$0.615 \cdot 10^{-2}$	$2.86 \cdot 10^{-2}$	$3.35 \cdot 10^{-2}$
400	$0.385 \cdot 10^{-2}$	$0.381 \cdot 10^{-2}$	$1.91 \cdot 10^{-2}$	$2.15 \cdot 10^{-2}$
500	$0.297 \cdot 10^{-2}$	$0.259 \cdot 10^{-2}$	$1.55 \cdot 10^{-2}$	$1.60 \cdot 10^{-2}$

Table 7.3: Inverse Fisher information and CV2 information criterion for r and ν when the true parameters are $\vartheta_0 = (1, 1, 1)$, and empirical mean squared errors of the corresponding 300 estimates.

The same effect is not observed with the MLE which also implicitly involves a prediction of the kriging variance (this will become more obvious in the next section). We have seen that this kriging variance is sensitive to changes of r *especially* when sampling locations are dense, and this translates into much better estimates compared to CV2 in the case $r_0 = 1$.
For the same reason, we should however expect that the estimates θ_{CV2}, even when they are imprecise, should yield reasonably well predictions because they deviate from θ_0 in a way that hardly affects prediction accuracy. This will be studied next.

7.3.3 Prediction accuracy with estimated parameters

Estimating the parameters of the covariance function of a RF $(X_t)_{t \in T}$ is only an intermediate step to predict a sample path f of $(X_t)_{t \in T}$ at an unobserved location based on observations

$f(t_1), \ldots, f(t_n)$. In Section 6.3 and 6.4, kriging was derived as optimal (in the sense that the expected squared error is minimized) unbiased prediction, but optimality was based on the assumption that the correct covariance function is used.

In this subsection we shall therefore investigate to which degree the prediction accuracy degrades when the parameter used for kriging is estimated. We use the same respective 300 simulated sample paths with covariance function $\upsilon\, \Phi_{r,\nu}$ and parameters $\vartheta_0 = (1, 0.1, 1)$, and $\vartheta_0 = (1, 1, 1)$, respectively. As a measure of the prediction accuracy we use the root of the mean squared prediction errors (RMSE) on our simulation grid $\mathcal{Q}$

$$\text{RMSE}(\theta) \;:=\; \sqrt{\frac{1}{|\mathcal{Q}|} \sum_{t \in \mathcal{Q}} \big(f(t) - s_{\Phi_\theta}(t)\big)^2},$$

where $|\mathcal{Q}|$ denotes the number of points in $\mathcal{Q}$ and s_{Φ_θ} is the (simple) kriging interpolant corresponding to the covariance function Φ_θ. Results are presented for prediction based on the first 200 and all 500 observations, respectively, at the sampling locations from above.

We compare the RMSEs that are obtained for

1. the best possible choice θ_{opt} of θ, i.e. the minimizer of RMSE(θ) on a fine grid on $[0, 40] \times [0, 16]$,
2. the “correct” parameter θ_0, i.e. the parameter of the covariance function according to which the sample paths were simulated,
3. the maximum likelihood estimate θ_{MLE}, and
4. the cross validation estimates θ_{CV1} and θ_{CV2} corresponding to minimization of the l_1- and l_2-norm respectively of the vector of LOOCV errors.

n	RMSE(θ_{opt})	RMSE(θ_0)	RMSE(θ_{MLE})	RMSE(θ_{CV1})	RMSE(θ_{CV2})
200	$8.152 \cdot 10^{-1}$ –	$8.176 \cdot 10^{-1}$ + 0.30 %	$8.215 \cdot 10^{-1}$ + 0.77 %	$8.253 \cdot 10^{-1}$ + 1.24 %	$8.246 \cdot 10^{-1}$ + 1.16 %
500	$6.620 \cdot 10^{-1}$ –	$6.636 \cdot 10^{-1}$ + 0.24 %	$6.647 \cdot 10^{-1}$ + 0.41 %	$6.667 \cdot 10^{-1}$ + 0.71 %	$6.659 \cdot 10^{-1}$ + 0.58 %

Table 7.4: Average RMSE and average relative increase over the optimal RMSE for the kriging predictions of the 300 sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$.

n	RMSE(θ_{opt})	RMSE(θ_0)	RMSE(θ_{MLE})	RMSE(θ_{CV1})	RMSE(θ_{CV2})
200	$1.414 \cdot 10^{-1}$ –	$1.431 \cdot 10^{-1}$ + 1.20 %	$1.434 \cdot 10^{-1}$ + 1.42 %	$1.447 \cdot 10^{-1}$ + 2.33 %	$1.445 \cdot 10^{-1}$ + 2.22 %
500	$8.954 \cdot 10^{-2}$ –	$9.016 \cdot 10^{-2}$ + 0.69 %	$9.021 \cdot 10^{-2}$ + 0.74 %	$9.052 \cdot 10^{-2}$ + 1.10 %	$9.054 \cdot 10^{-2}$ + 1.12 %

Table 7.5: Average RMSE and average relative increase over the optimal RMSE for the kriging predictions of the 300 sample paths simulated with $\vartheta_0 = (1, 1, 1)$.

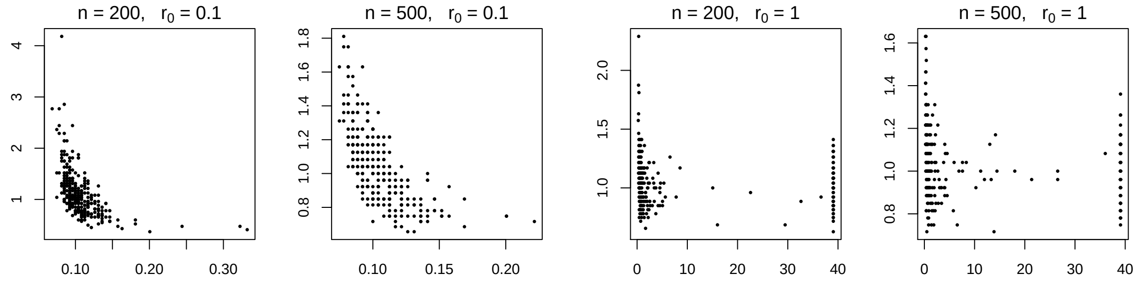

Figure 7.4: Plots of the values of r (x-axis) against ν (y-axis) that yield the best predictions (based on 200 and 500 observations, respectively) for the 300 sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$ (left two plots) and $\vartheta_0 = (1, 1, 1)$ (right two plots).

Averages over the respective 300 RMSEs and the average increase of the RMSE over the optimal RMSE when θ is chosen by one of the estimation procedures are given in Tables 7.4 and 7.5. The following points can be noted:

- The loss of prediction accuracy because of not knowing the optimal parameter θ_{opt} is quite moderate in our simulation setup. This could be expected from the results of Section 6.7.

- MLE estimates yield better predictions than CV1 and CV2. However, even in the case $r_0 = 1$ where CV2 produces very poor parameter estimates, the corresponding predictions are not dramatically worse than those of the MLE.

- In the case $r_0 = 1$ the prediction accuracy using θ_{MLE} is hardly worse than the prediction accuracy that is obtained by using the "correct" parameter θ_0.

The last point is quite surprising at first. However, as one can see in Figure 7.4, the optimal parameter θ_{opt} itself is strongly dispersed around θ_0. This emphasizes that θ_0 is the optimal choice only in expectation whereas the best prediction for an individual sample path may be obtained for a different value.

7.3.4 Kriging variance prediction with estimated parameters

We shall now investigate to which degree the precision of the kriging variance prediction degrades when it is based on the estimated parameter rather than the true one. We study the predicted kriging variance based on 500 observations in the case $r_0 = 0.1$ and based on 200 observations in the case $r_0 = 1$. For these choices the kriging variance under the true parameters takes on a broad range of values between 0 and 1 (see Figure 7.5)

In order to predict the kriging variance we need to estimate the parameter υ in addition to r and ν. An estimator in the CV2 context was derived in Section 7.3.

In Figure 7.6 boxplots of the 300 estimates υ_{MLE} and υ_{CV2} are given for the two cases $r_0 = 0.1$ and $r_0 = 1$. In the latter case a lot of the CV2 estimates are totally off the mark. While $\upsilon_{\text{MLE}} < 2$ at least for 93 % of the sample paths, we have $\upsilon_{\text{CV2}} < 2$ in only 63 % of all 300 cases. The reason for this is that the estimates of υ are strongly (positively) correlated with those of r and ν. As we saw before, especially r_{CV2} is often very big in the case $r_0 = 1$

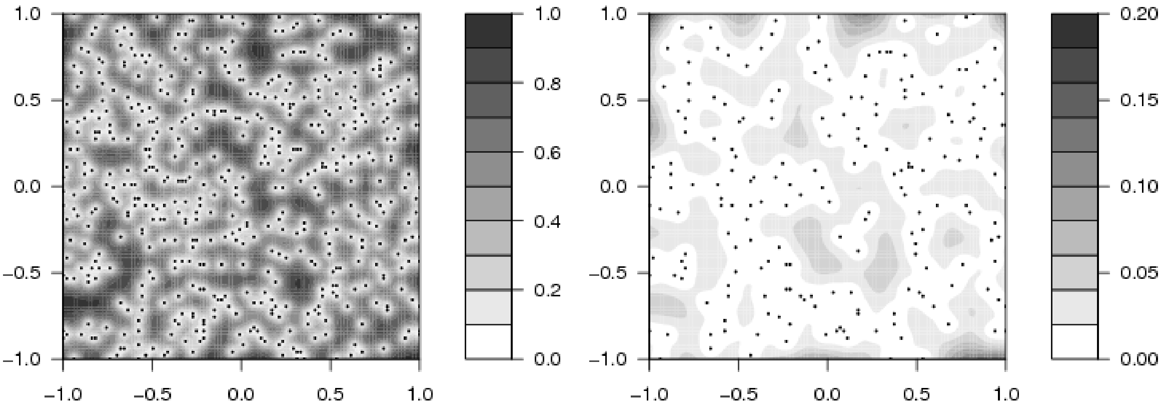

Figure 7.5: Kriging variances under the (true) parameters $\vartheta_0 = (1, 0.1, 1)$ based on 500 observations (left) and $\vartheta_0 = (1, 1, 1)$ based on 200 observations (right). The black dots indicate the sampling locations.

and consequently so is υ_{CV2}. The following results will however show that both of these considerable overestimation partially compensate each other when it comes to predicting the kriging variance which typically gets small for large values of r and big for large values of υ.

In order to assess if reasonable predictions of the kriging variance can be obtained on the basis of the estimates $(\upsilon, r, \nu)_{\mathrm{MLE}}$ and $(\upsilon, r, \nu)_{\mathrm{CV2}}$ we proceed as follows:

(a) For each sample path $f^{(j)}$, $j = 1, \dots, 300$, we compute

- the parameter estimates $\vartheta_{\star,j} = (\upsilon, \theta)_{\star,j} = (\upsilon, r, \nu)_{\star,j}\,, \quad \star = \text{MLE, CV2}$

(b) the corresponding predictions $s^{(j)}_{\Phi_\theta}$ of this sample path

(c) the square root $P^{(j)}_{\upsilon\Phi_\theta}$ of the kriging variance, calculated with $\vartheta = \vartheta_{\star,j}$

(d) Now, for $t \in T$ define the standardized prediction errors

$$E_j(\vartheta, t) \quad = \quad \frac{f^{(j)}(t) - s^{(j)}_{\Phi_\theta}(t)}{P^{(j)}_{\upsilon\Phi_\theta}}, \qquad j = 1, \dots, 300.$$

If the true parameter ϑ_0 is used for the calculation of $s^{(j)}_{\Phi_\theta}$ and $P^{(j)}_{\upsilon\Phi_\theta}$, the empirical distribution of $E_j(\vartheta_0, t)$, $j = 1, \dots, 300$, should be approximately standard Gaussian for any fixed t.

1. We compute the mean of the squared standardized prediction errors

$$\mathrm{MSSPE}(\vartheta, t) \;:=\; \tfrac{1}{300} \sum_{j=1}^{300} \big(E_j(\vartheta, t)\big)^2$$

According to (b) we should expect that $\mathrm{MSSPE}(\vartheta_0, t) \approx 1$ for all $t \in T$.

Hence, by comparing $\mathrm{MSSPE}(\vartheta_0, t)$, $\mathrm{MSSPE}(\vartheta_{\mathrm{MLE}}, t)$ and $\mathrm{MSSPE}(\vartheta_{\mathrm{CV2}}, t)$ we get an impression about how well $P^{(j)}_{\upsilon\Phi_\theta}$ describes the magnitude of the prediction errors when ϑ is estimated from the data.

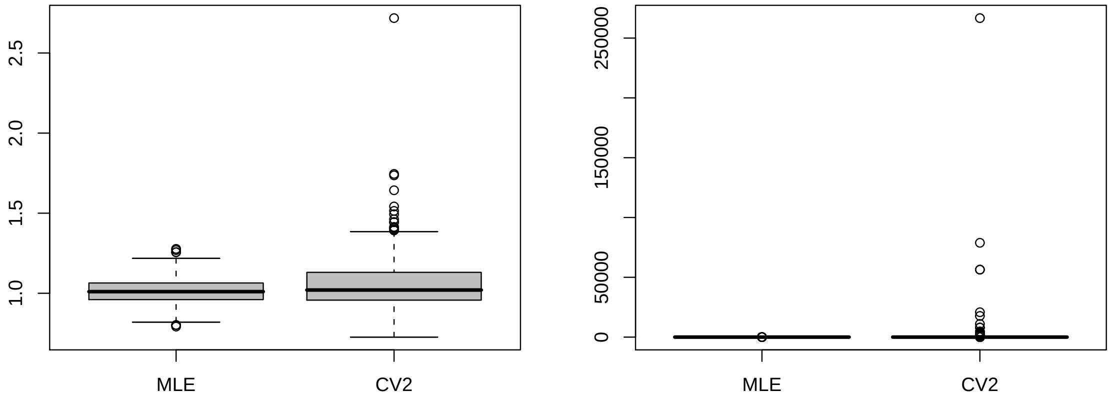

Figure 7.6: Estimates of υ for the sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$, based on 500 observations (left plot), and with parameters $\vartheta_0 = (1, 1, 1)$, now based on 200 observations (right plot).

In Figure 7.7 we visualize $\text{MSSPE}(\vartheta, \cdot)$ via filled contour plots. In the different plots the correct value ϑ_0 and the estimates ϑ_{MLE} or ϑ_{CV2} respectively are used for prediction and for the calculation of the kriging variance.

Even $\text{MSSPE}(\vartheta_0, \cdot)$ deviates from 1 considerably as a consequence of the randomness of the prediction errors. The magnitude of these deviations could be reduced only by increasing the number of simulations. We are now interested in how far $\text{MSSPE}(\vartheta_{\text{MLE}}, \cdot)$ and $\text{MSSPE}(\vartheta_{\text{CV2}}, \cdot)$ differ from $\text{MSSPE}(\vartheta_0, \cdot)$. A more quantitative study of these variables yields the following findings:

- While, in the case $r_0 = 0.1$, $\text{MSSPE}(\vartheta_0, \cdot) > 1.1$ for about 11 % of the points of $\mathcal{Q}$ (the evaluation grid) the same threshold is exceeded at about 15 % and 17 % respectively of these points if ϑ is estimated by MLE and CV2.
  At the same time we have $\text{MSSPE}(\vartheta_0, \cdot) < 0.9$ for about 11 % of the points of $\mathcal{Q}$, the respective percentages for MLE and CV2 are 8 % and 7 %.

  This shows that the magnitude of the prediction errors tends to be underestimated. The difference to the "true" kriging variance prediction is not extremely big though, neither for MLE nor for CV2.

- The results for $r_0 = 1$ are similar but a bit more pronounced. The percentages of points that exceed 1.1 (fall below 0.9) are 11 % (13 %), 22 % (7 %) and 26 % (6 %) respectively for $\vartheta = \vartheta_0, \vartheta_{\text{MLE}}$ and ϑ_{CV2}.

  The deviations from $\text{MSSPE}(\vartheta_0, \cdot)$ are larger in this case, but they are still not dramatic which is quite remarkable especially for the CV2 estimates which were seen to be very poor in the case $r_0 = 1$.

So far we have studied the quality of pointwise kriging variance predictions. We will now study the possibility to predict some global measure of prediction accuracy like the L^2-prediction error

$$\left\| f - s_{\Phi_\theta} \right\|_{L^2(T)} = \left(\int_T \left(f(t) - s_{\Phi_\theta}(t) \right)^2 dt \right)^{1/2}.$$

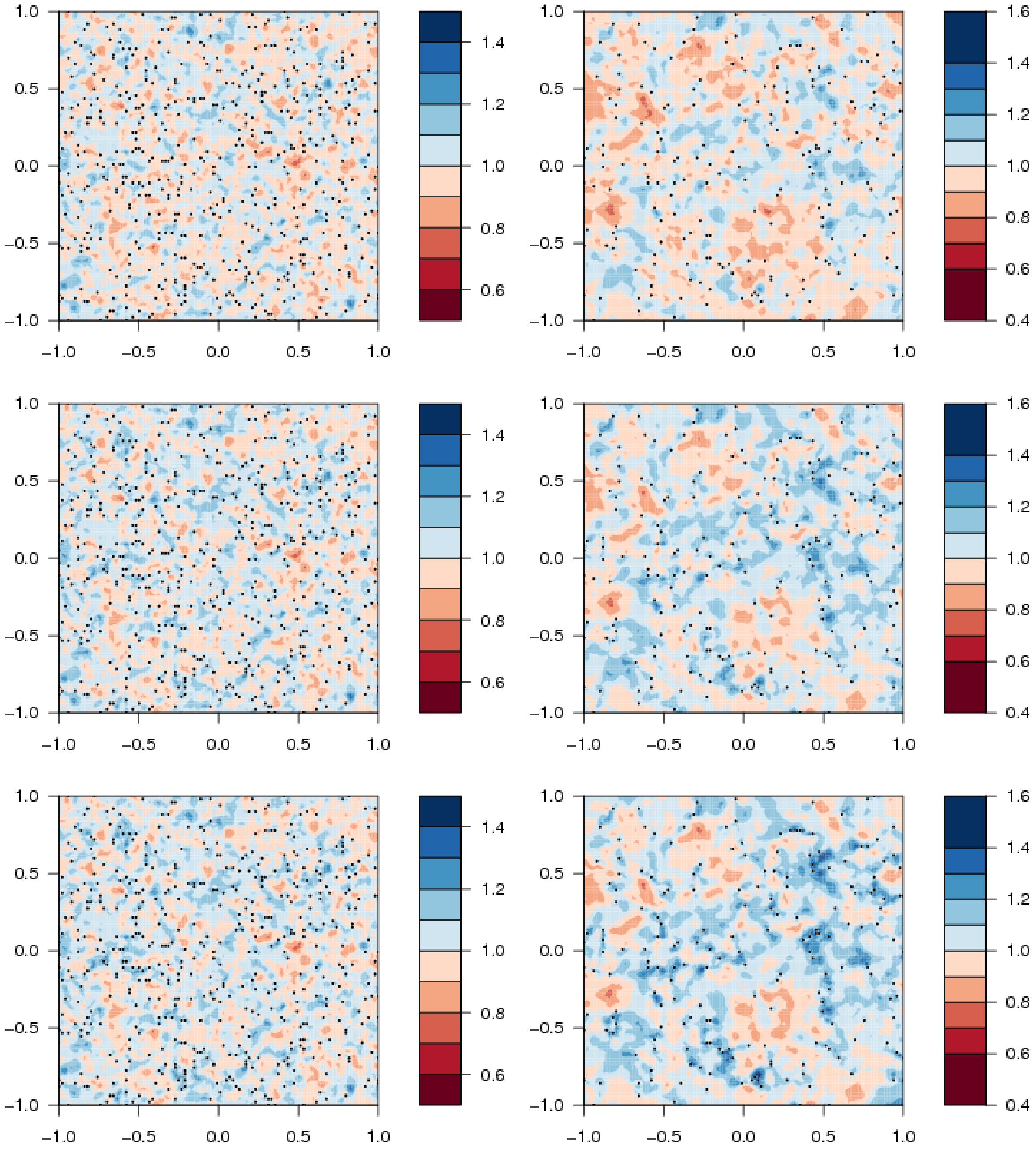

Figure 7.7: Contour plots of MSSPE($\vartheta_0, \cdot$) (top row), MSSPE($\vartheta_{\mathrm{MLE}}, \cdot$) (middle row), and MSSPE($\vartheta_{\mathrm{CV2}}, \cdot$) (bottom row) for the sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$ and $n = 500$ (left) and for the sample paths with $\vartheta_0 = (1, 1, 1)$ and $n = 200$ (right).

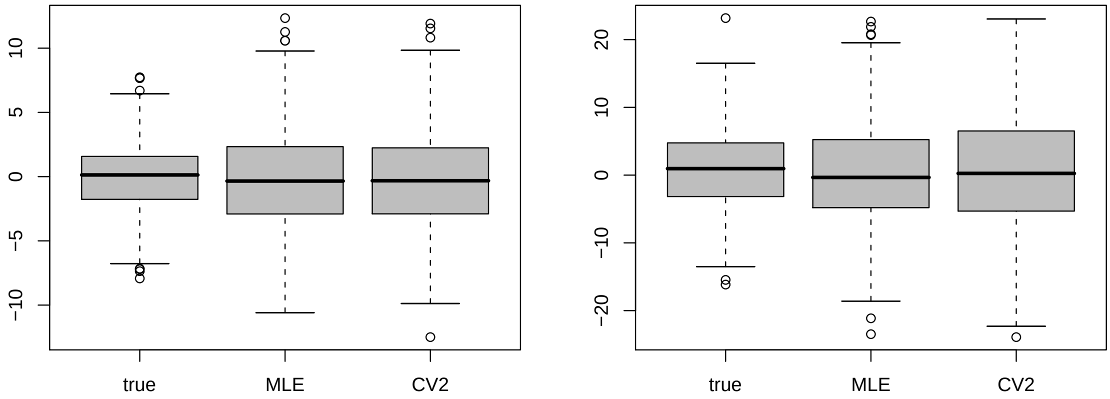

Figure 7.8: Relative deviation (in %) of the predicted RMSE from the actual RMSE for the sample paths simulated with $\vartheta_0 = (1, 0.1, 1)$, based on 500 observations (left) and with $\vartheta_0 = (1, 1, 1)$, based on 200 observations (right).

In our context where f is assumed to be a sample path of a RF $(X_t)_{t\in T}$, s_{Φ_θ} is a sample path of the interpolation process $(Y_t^*)_{t\in T}$, and we have by Fubini's theorem

$$\mathbb{E}\left(\|X_\cdot - Y_\cdot^*\|_{L^2(T)}^2\right) \;=\; \int_T \mathbb{E}\left((X_t - Y_t^*)^2\right)\, dt \;=\; \int_T P_{\upsilon\Phi_\theta}^2(t)\, dt\,.$$

If the $(X_t)_{t\in T}$ is ergodic (see Section 4.2) and r is small compared to the diameter of T we can expect the random fluctuations of the squared prediction errors around their mean to partially average out over T and hence the variance of $\|X_\cdot - Y_\cdot^*\|_{L^2(T)}^2$ to be small. We then have

$$\|f - s_{\Phi_\theta}\|_{L^2(T)} \quad\approx\quad \left(\int_T P_{\upsilon\Phi_\theta}^2(t)\, dt\right)^{1/2} \tag{7.18}$$

so we may use the rhs as a prediction of the L^2-prediction error.

We assess the accuracy of this prediction in our simulation setup. Instead of the L^2-prediction error we consider predicting $\mathrm{RMSE}(\theta_\star)$, $\star =$ MLE, CV2. This is more or less equivalent if our evaluation grid $\mathcal{Q}$ is reasonably fine since

$$\int_T \left(f(t) - s_{\Phi_\theta}(t)\right)^2 dt \quad\approx\quad \frac{vol(T)}{|\mathcal{Q}|} \sum_{t\in\mathcal{Q}} \left(f(t) - s_{\Phi_\theta}(t)\right)^2 \tag{7.19}$$

$$\text{and} \qquad \int_T P_{\upsilon\Phi_\theta}^2(t)\, dt \quad\approx\quad \frac{vol(T)}{|\mathcal{Q}|} \sum_{t\in\mathcal{Q}} P_{\upsilon\Phi_\theta}^2(t)\,. \tag{7.20}$$

In our two experiments on kriging variance prediction from above we calculate the rhs of (7.19) and (7.20) with $\vartheta = \vartheta_0, \vartheta_{\mathrm{MLE}}$ and ϑ_{CV2}. The accuracy of the prediction of $\mathrm{RMSE}(\theta_\star)$ is then illustrated by calculating the relative deviation of the predicted RMSE over the actual RMSE for the different choices of ϑ.

The deviations of the predicted RMSE from the actual RMSE are illustrated in Figure 7.8. The boxplots for $\vartheta = \vartheta_0$ give an idea about how appropriate the approximation (7.18) is apart

from the additional uncertainty due to unknown model parameters. Predictions are more precise in the case where $r_0 = 0.1$ which could be expected since at this scale the fluctuations of the squared prediction errors around their mean are more likely to average out.
A comparison with the boxplots for $\vartheta = \vartheta_{\text{MLE}}$ and $\vartheta = \vartheta_{\text{CV2}}$ shows how much the RMSE predictions deteriorate when estimated parameters are used for both predicting f and predicting the kriging variance. In the case where $r_0 = 0.1$ there is no noticeable advantage of the predictions corresponding to ϑ_{MLE} over those corresponding to ϑ_{CV2}. In the case $r_0 = 1$ there is a slight advantage of MLE over CV2 but again both methods yield good predictions of the RMSE with deviations mostly smaller than 20%. Note in particular that the CV2 is competitive to MLE w.r.t. estimation of the L^2-prediction error although it yields very poor estimates of υ.

The results of the comparison between MLE and CV2 made in this subsection can be summarized as follows

- The MLE yields (sometimes considerably) better parameter estimates than the CV2 when the model assumptions (Gaussian RF) are true.
- The parameter estimates obtained by ML lead to better kriging predictions than those obtained by CV1 and CV2. In our examples the difference was not very big though, and neither was the difference of both methods to the case where prediction is carried out with the optimal parameters.
- The parameter estimates from both methods allow for a satisfactory prediction of the kriging variance. Even bad estimates seem to be at least consistent with themselves in the sense that they still lead to more or less acceptable predictions of the kriging variance. MLE is again slightly ahead of CV2.

All of the simulation results of this subsection were obtained for the case where f is indeed a sample path of a stationary Gaussian RF and do therefore not allow any conclusion about the performance of MLE and LOOCV in the kernel interpolation framework. In particular for the MLE, the derivation of which was explicitly based on these assumptions, it is not clear if its application in the context of approximation theory is meaningful at all.
We will show in the next section that the MLE can indeed be used in a much more general framework and we will conduct a further simulation study with typical examples from approximation theory in Section 7.5.

7.4 Maximum Likelihood revisited

We already noted that the (RE)ML estimator, which was derived under the assumption of a Gaussian RF, may perform well also in Non-Gaussian frameworks. In this section we go even further and motivate its use in the framework of Numerical Analysis which does not assume any probabilistic model behind the creation of the observations $f(t_1), \dots, f(t_n)$ at all. As above we denote by $\mathcal{T}$ the set of sampling locations (in kernel interpolation also called centres).

Our starting point is the LOOCV procedure which has a meaningful interpretation independent of any model assumptions. It can be argued however, that this procedure does not use the available information in an optimal way, in particular the following two points of criticism can be made (see also the discussion of the results for r_{CV2} in the simulation study in Section 7.3):

1. The same observations $f(t_1), \ldots, f(t_n)$ are used for the calculation of every component ε_k, either as value to be predicted or as data the interpolant is fitted to. This can lead to distortions, especially with irregular patterns of sampling locations. To see this assume that the distance between two sampling location t_i and t_j is small compared with the average distance. Then $s_{R_\theta,\mathcal{P},[-i]}(t_i)$ is determined mainly by $f(t_j)$ and vice versa, so that the components ε_i and ε_j basically contain the same information about $f - s_{R_\theta,\mathcal{P}}$. This "redundancy" is not accounted for by LOOCV.

2. The accuracy of $s_{R_\theta,\mathcal{P},[-k]}$ as a predictor for $f(t_k)$ does not only depend on θ, but strongly depends on the geometry of $\mathcal{T}$. Even for a good choice of θ, data points near the margin or isolated data points will in general be predicted worse, i.e. lead to bigger values of ε_k, than data points in densely sampled areas of T. This is also not taken into account by LOOCV.

The latter point of criticism suggests that the LOOCV components should be weighted with weights that reflect the prediction accuracy that can be expected on the basis of the geometry of $\mathcal{T}$. A suitable such measure for the "potential" prediction accuracy at $t \in T$ is the power function $P_{R_\theta,\mathcal{P}}(t)$ introduced in (6.24). $P_{R_\theta,\mathcal{P}}$ is the norm of the pointwise error functional of the interpolation process and thus gives an indication about the magnitude of $f - s_{R_\theta,\mathcal{P}}$ independent of the actual f.

Writing $P_{R_\theta,\mathcal{P},[-k]}$ for the power function corresponding to $s_{R_\theta,\mathcal{P},[-k]}$ we now propose to pass to the weighted LOOCV errors

$$\varepsilon_k^{(w)} := \frac{\varepsilon_k}{P_{R_\theta,\mathcal{P},[-k]}(t_k)}, \qquad k = 1, \ldots, n. \tag{7.21}$$

When $\varepsilon_\theta^{(w)} := \big(\varepsilon_1^{(w)}, \ldots, \varepsilon_n^{(w)}\big)'$ is used instead of ε_θ components corresponding to hard-to-predict locations are no longer dominating the norm of the error vector.

Remark 7.4.1. In Section 7.1 we pointed out that the LOOCV error component ε_k is the value at t_k of the error function $f_{\varepsilon_k} := s_{R_\theta\mathcal{P}} - s_{R_\theta\mathcal{P},[-k]}$. Now, f_{ε_k} has the form (6.27) with respect to the centres $\mathcal{T} \setminus \{t_k\}$, and so Lemma 6.5.4 yields

$$\big|\varepsilon_k^{(w)}\big| = \|f_{\varepsilon_k}\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}. \tag{7.22}$$

Hence, instead of the LOOCV interpolation errors at the left-out centres we are now considering the norms $\|\cdot\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}$ of the LOOCV error functions $f_{\varepsilon_1}, \ldots, f_{\varepsilon_n}$.

The weighting according to (7.21) however raises a new problem: the power functions $P_{R_\theta,\mathcal{P},[-k]}$, $k = 1, \ldots, n$, themselves depend on θ which has the consequence that minimization of the weighted errors favours values of θ that lead to big power functions. This calls for a correction factor that penalizes big values of the power functions. When the l_2 norm of the weighted errors is used we propose the following weighted cross validation (WCV) procedure

$$\theta_{\text{WCV}} \;=\; \arg\min_{\theta\in\Theta} \left\{ \|\varepsilon_\theta^{(w)}\|^2 \cdot \sqrt[n]{\prod_{i=1}^{n} P^2_{R_\theta,\mathcal{P},[-i]}(t_i)} \right\}.$$

The rationale behind using the geometric mean of $P_{R_\theta,[-1]}(t_1),\ldots,P_{R_\theta,[-n]}(t_n)$ as a correction factor rather than e.g. the arithmetic mean is that this corresponds to averaging on a logarithmic scale. In our situation, the most important feature about the correction factor is its behaviour under changes of θ, and the terms

$$\tfrac{\partial}{\partial\theta_l}\log\left(P^2_{R_\theta,\mathcal{P},[-k]}(t_k)\right) \;=\; \frac{\frac{\partial}{\partial\theta_l}P^2_{R_\theta,\mathcal{P},[-k]}(t_k)}{P^2_{R_\theta,\mathcal{P},[-k]}(t_k)}, \qquad k=1,\ldots,n$$

have the big advantage that they can more reasonably be assumed to be of the same magnitude. Hence, the penalty factor, too, depends quite uniformly on all the terms involved.

Remark 7.4.2. The weighting of the error components in (7.21) by the power function can be motivated in the same way in the framework of spatial statistics. There, moreover, a formal justification of the proposed correction factor can be given:

Consider the log-target function

$$\log\left(\|\varepsilon_\theta^{(w)}\|^2\right) + \log\left(F(\theta)\right)$$

with correction factor $F(\theta)$. Taking partial derivatives and multiplying by $\|\varepsilon_\theta^{(w)}\|^2$ leads to an estimating function $G(\theta;\mathrm{f})$ with components

$$G_{\theta_l}(\theta;\mathrm{f}) \;=\; \tfrac{\partial}{\partial\theta_l}\|\varepsilon_\theta^{(w)}\|^2 \;+\; \|\varepsilon_\theta^{(w)}\|^2\cdot\tfrac{\partial}{\partial\theta_l}\log\left(F(\theta)\right), \qquad l=1,\ldots,p.$$

For simplicity we only discuss the case $\mathcal{P}=\{0\}$ in which, using Proposition 7.1.1 and Lemma 7.1.2, we have

$$\begin{aligned} \|\varepsilon_\theta^{(w)}\|^2 \;&=\; \mathrm{f}'A_\theta^{-1}D_\theta\,A_\theta^{-1}\mathrm{f} \qquad \text{and} \\ \tfrac{\partial}{\partial\theta_l}\|\varepsilon_\theta^{(w)}\|^2 \;&=\; -2\,\mathrm{f}'A_\theta^{-1}\tfrac{\partial}{\partial\theta_l}A_\theta\,A_\theta^{-1}D_\theta A_\theta^{-1}\mathrm{f} \;+\; \mathrm{f}'A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_l}D_\theta\,A_\theta^{-1}\mathrm{f}\,. \end{aligned}$$

Noting that $\mathrm{tr}\left(A_\theta^{-1}\tfrac{\partial}{\partial\theta_l}A_\theta\,A_\theta^{-1}D_\theta\right) = \mathrm{tr}\left(A_\theta^{-1}\,\tfrac{\partial}{\partial\theta_l}D_\theta\right) = \mathrm{tr}\left(D_\theta^{-1}\,\tfrac{\partial}{\partial\theta_l}D_\theta\right)$ we obtain

$$\mathbb{E}\big(G_{\theta_l}(\theta;X_\mathcal{T})\big) \;=\; -\mathrm{tr}\left(D_\theta^{-1}\,\tfrac{\partial}{\partial\theta_l}D_\theta\right) \;+\; n\cdot\tfrac{\partial}{\partial\theta_l}\log\left(F(\theta)\right), \qquad l=1,\ldots,p,$$

and hence the estimating function $G(\theta;\mathrm{f})$ is unbiased if and only if

$$\tfrac{\partial}{\partial\theta_l}\log\left(F(\theta)\right) \;=\; \tfrac{1}{n}\,\mathrm{tr}\left(D_\theta^{-1}\,\tfrac{\partial}{\partial\theta_l}D_\theta\right), \qquad l=1,\ldots,p.$$

Now, if we take

$$F(\theta) \ := \ \sqrt[n]{\prod_{i=1}^{n} P^2_{R_\theta,\mathcal{P},[-i]}(t_i)}$$

as proposed above, we get (again by using Lemma 7.1.2) for any $1 \le l \le p$:

$$\tfrac{\partial}{\partial\theta_l} \log\left(F(\theta)\right) \ = \ \tfrac{1}{n}\tfrac{\partial}{\partial\theta_l}\log\left(\prod_{i=1}^{n}\Psi_{ii}^{-1}\right) \ = \ \tfrac{1}{n}\tfrac{\partial}{\partial\theta_l}\log\left(|D_\theta|\right) \ = \ \tfrac{1}{n}\operatorname{tr}\left(D_\theta^{-1}\tfrac{\partial}{\partial\theta_l}D_\theta\right),$$

and so this choice of $F(\theta)$ is exactly the correction needed to obtain an unbiased estimating function.

The weighted cross validation procedure proposed above is equivalent to a procedure proposed in the geostatistical literature by [33] who assume the LOOCV errors to be multivariate Gaussian.

In our motivation of the weighting (7.21) we did not make any explicit assumption about the LOOCV errors. Implicitly however, we assume that for some $\upsilon > 0$

$$\begin{aligned} \left(f(t_k) - s_{R_\theta,\mathcal{P},[-k]}(t_k)\right)^2 \ &\approx \ \upsilon \cdot P^2_{R_\theta,\mathcal{P},[-k]}(t_k), \qquad k = 1,\dots,n, \\ \left(f(t) - s_{R_\theta,\mathcal{P}}(t)\right)^2 \ &\approx \ \upsilon \cdot P^2_{R_\theta,\mathcal{P}}(t), \qquad t \in T \end{aligned} \tag{7.23}$$

with “ $\approx$ ” in the sense that the deviations are not systematic and “average out”. We will assess the adequacy of this assumption in the numerical experiments in Section 7.5 together with a similar assumption that comes with the following further refinement of WCV.

The weighting of the LOOCV errors was suggested in response to our criticism on LOOCV concerning the ignorance of the possibly different magnitudes of $\varepsilon_1,\dots,\varepsilon_n$. Another point of criticism was the ignorance of the relations (“dependencies”) between different error components that are present due to the multiple use of the data $f(t_1),\dots,f(t_n)$.

One way to deal with this is to pass from the leave-one-out principle to a sequential approach, i.e. instead of using the data at all locations $\mathcal{T}\setminus\{t_k\}$ to predict $f(t_k)$, we only use the data at the locations $\{t_1,\dots,t_{k-1}\}$. Denoting by $s_{R_\theta,\mathcal{P},[<k]}$ the corresponding interpolant with the convention that $s_{R_\theta,\mathcal{P},[\le q]} \equiv 0$, the approximation errors now considered are

$$\tilde\varepsilon_k := f(t_k) - s_{R_\theta,\mathcal{P},[<k]}(t_k), \qquad k = q+1,\dots,n.$$

$\tilde\varepsilon_k$ is the value at t_k of the error function $\tilde f_{\varepsilon_k} := s_{R_\theta,\mathcal{P},[\le k]} - s_{R_\theta,\mathcal{P},[<k]}$ which, unlike the error function f_{ε_k} introduced in Section 7.1, is now only defined for $k > q$.

In this sequence of $n-q$ surrogate interpolation problems each data pair is still used several times, but now we have

$$\left(\tilde f_{\varepsilon_i}, \tilde f_{\varepsilon_j}\right)_{\mathcal{H}_{R_\theta,\mathcal{P}}} \ = \ 0, \qquad \text{for all} \quad q < i \neq j \le n \tag{7.24}$$

which follows from Lemma 6.2.6 by noting that $\tilde f_{\varepsilon_i} \in V_{R_\theta,\mathcal{P},\{t_1,\dots,t_{j-1}\}}$ for all $i < j$.

In other words: the error functions $\tilde f_{\varepsilon_{q+1}},\dots,\tilde f_{\varepsilon_n}$ are pairwise orthogonal functions in $\mathcal{H}_{R_\theta,\mathcal{P}}$ and can therefore be expected to yield essentially different information about the interpolation behaviour of the given data.

Remark 7.4.3. In the statistical context the motivation for the sequential approach instead of the leave-one-out principle is similar. If we regard $\tilde{\varepsilon}_{q+1}, \dots, \tilde{\varepsilon}_n$ as RVs, condition (6.21) on the kriging weights implies that $\tilde{\varepsilon}_k$ is a contrast of $X_{t_1}, \dots, X_{t_k}$. But then, by Lemma 6.6.2, $\tilde{\varepsilon}_i$ and $\tilde{\varepsilon}_j$ are uncorrelated for any $1 \le i < j \le b$.

We can now combine both ideas and work with the weighted sequential approximation errors

$$\tilde{\varepsilon}_k^{(w)} := \frac{\tilde{\varepsilon}_k}{P_{R_\theta,\mathcal{P},[<k]}(t_k)}, \qquad k = q+1, \dots, n. \tag{7.25}$$

If we use the (squared) euclidean norm of the error vector $\tilde{\varepsilon}_\theta^{(w)} = \big(\tilde{\varepsilon}_{q+1}^{(w)}, \dots, \tilde{\varepsilon}_n^{(w)}\big)'$, the same arguments as above suggest that we should introduce a correction factor and minimize

$$\|\tilde{\varepsilon}_\theta^{(w)}\|^2 \;\cdot\; \sqrt[n-q]{\prod_{i=q+1}^{n} P^2_{R_\theta,\mathcal{P},[<i]}(t_i)}\,. \tag{7.26}$$

The following two Propositions provide the basis for better interpretability and computationally effective calculation of this target function.

Proposition 7.4.4. *Let $\tilde{\varepsilon}_\theta^{(w)}$ be the vector of weighted sequential approximation errors as defined above. Then it holds that*

$$\|\tilde{\varepsilon}_\theta^{(w)}\|^2 \;=\; \|s_{R_\theta,\mathcal{P}}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}} \;=\; \mathrm{f}'\left(A_\theta^{-1} \;-\; A_\theta^{-1}P\big(P'A_\theta^{-1}P\big)^{-1}P'A_\theta^{-1}\right)\mathrm{f}.$$

In particular $\|\tilde{\varepsilon}_\theta^{(w)}\|^2$ does not depend on the ordering of $t_1, \dots, t_n$.

Proof: As in Remark 7.4.1 we see that $|\tilde{\varepsilon}_k^{(w)}| = \|\tilde{f}_{\varepsilon_k}\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}, \quad k = q+1, \dots, n.$
Using the orthogonality relation (7.24) we get

$$\begin{aligned}
\|\tilde{\varepsilon}_\theta^{(w)}\|^2 \;&=\; \sum_{i=q+1}^{n} \|s_{R_\theta,\mathcal{P},[\le i]} - s_{R_\theta,\mathcal{P},[<i]}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}} \qquad (7.27)\\
&=\; \sum_{i=q+1}^{n} \left(\|s_{R_\theta,\mathcal{P},[\le i]}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}} - \|s_{R_\theta,\mathcal{P},[<i]}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}}\right) \;=\; \|s_{R_\theta,\mathcal{P}}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}}\,.
\end{aligned}$$

Now, due to the special form (6.7) of $s_{R_\theta,\mathcal{P}}$ its norm $\|\cdot\|_{\mathcal{H}_{R,\mathcal{P}}}$ can be calculated explicitly (see Section 6.2) and we obtain

$$\|s_{R_\theta,\mathcal{P}}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}} \;=\; \alpha' A_\theta\, \alpha \;=\; \alpha'\mathrm{f} - \alpha' P\,\beta \;=\; \alpha'\mathrm{f},$$

since by condition (6.9) we have $P'\alpha = \mathbf{0}$. Using (6.10) we get

$$\|s_{R_\theta,\mathcal{P}}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}} = \begin{pmatrix} \alpha \\ \beta \end{pmatrix}' \begin{pmatrix} \mathrm{f} \\ \mathbf{0} \end{pmatrix} = \begin{pmatrix} \mathrm{f} \\ \mathbf{0} \end{pmatrix}' \begin{pmatrix} A_\theta & P \\ P' & \mathbf{0} \end{pmatrix}^{-1} \begin{pmatrix} \mathrm{f} \\ \mathbf{0} \end{pmatrix},$$

and the asserted representation follows from standard rules for the inversion of block matrices.
□

Proposition 7.4.5. *For the product of sequential power functions it holds that*

$$\prod_{i=q+1}^{n} P^2_{R_\theta,\mathcal{P},[<i]}(t_i) \quad \sim \quad \left|A_\theta\right| \left|P' A_\theta^{-1} P\right|$$

with a proportionality constant that does not depend on θ. In particular the correction factor in (7.26) does not depend on the ordering of $t_1, \ldots, t_n$.

Proof: To simplify notation we drop the subscript θ from A_θ during the proof. We rewrite (6.10) in the form

$$\begin{pmatrix} \mathbf{0} & P' \\ P & A \end{pmatrix} \begin{pmatrix} \beta \\ \alpha \end{pmatrix} = \begin{pmatrix} \mathbf{0} \\ \mathrm{f} \end{pmatrix}$$

and prove the assertion of the proposition by induction.

Let $A_k \in \mathbb{R}^{k\times k}$ and $P_k \in \mathbb{R}^{k\times q}$ denote the submatrices of A and P that correspond to the interpolation system for the first k data points only, further let

$$M_k := \begin{pmatrix} \mathbf{0} & P_k' \\ P_k & A_k \end{pmatrix},$$

$a_k := \big(R_\theta(t_1, t_k), \ldots, R_\theta(t_{k-1}, t_k)\big)'$ and $p_k := \big(p_1(t_k), \ldots, p_q(t_k)\big)'$.

Let's first assume, that $\{p_1, \ldots, p_q\}$ is a Lagrange basis of P. In this case $P_q = \mathrm{Id}_q$ and we have

$$\left|M_q\right| \;=\; \left|\begin{pmatrix} \mathbf{0} & \mathrm{Id}_q \\ \mathrm{Id}_q & A_q \end{pmatrix}\right| \;=\; (-1)^q \left|\begin{pmatrix} \mathrm{Id}_q & A_q \\ \mathbf{0} & \mathrm{Id}_q \end{pmatrix}\right| \;=\; (-1)^q$$

Let $q < k \le n$ and write M_k in block form as M_{k-1} augmented by the k^{th} row and k^{th} column

$$M_k = \begin{pmatrix} \mathbf{0} & P_{k-1}' & p_k \\ P_{k-1} & A_{k-1} & a_k \\ p_k' & a_k' & R_\theta(t_k, t_k) \end{pmatrix}$$

with Schur complement

$$S_k \;=\; R_\theta(t_k, t_k) \;-\; \begin{pmatrix} p_k \\ a_k \end{pmatrix}' \begin{pmatrix} \mathbf{0} & P_{k-1}' \\ P_{k-1} & A_{k-1} \end{pmatrix}^{-1} \begin{pmatrix} p_k \\ a_k \end{pmatrix}.$$

By standard rules for determinants of block matrices we have $\left|M_k\right| = \left|M_{k-1}\right| \cdot S_k$ and hence

$$\left|\begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix}\right| \;=\; \left|\begin{pmatrix} \mathbf{0} & P_n' \\ P_n & A_n \end{pmatrix}\right| \;=\; (-1)^q \prod_{i=q+1}^{n} S_i\,.$$

Now, the inversion rule for block matrices implies

$$S_k^{-1} \;=\; \begin{pmatrix} \mathbf{0} \\ e_k \end{pmatrix}' \begin{pmatrix} \mathbf{0} & P_k' \\ P_k & A_k \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{0} \\ e_k \end{pmatrix}$$

and the same arguments as in the proof of Lemma 7.1.2 show that $S_k = P^2_{R_\theta,\mathcal{P},[<k]}(t_k)$.

Noting that $-P'AP$ is the Schur complement of $\begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix}$ we apply again the rule for determinants of block matrices and obtain

$$\prod_{i=q+1}^{n} P^2_{R_\theta,\mathcal{P},[<i]}(t_i) \;=\; (-1)^q \left| \begin{pmatrix} A & P \\ P' & \mathbf{0} \end{pmatrix} \right| \;=\; |A| \, |P'A^{-1}P|.$$

If $\{p_1, \ldots, p_q\}$ is not a Lagrange basis, there exists a regular matrix B so that

$$\tilde{p}_k \;:=\; \sum_{j=1}^{q} B_{jk}\, p_j, \qquad k = 1, \ldots, n$$

is a Lagrange basis, and hence $\tilde{P}_q = (PB)_q = I_q$ as needed above. But then

$$|A| \; |P'A^{-1}P| \;=\; |A| \; \left|\tilde{P}'A^{-1}\tilde{P}\right| \; / \; |B|^2 \;=\; \prod_{i=q+1}^{n} P^2_{R_\theta,\mathcal{P},[<i]}(t_i) \; / \; |B|^2$$

so the determinant only changes by a factor that is independent of θ.

□

Via Proposition 7.4.4 we can give the following interpretation to the target function (7.26):

1. $\|s_{R_\theta,\mathcal{P}}\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}}$ reflects the behaviour of $\|f\|^2_{\mathcal{H}_{R_\theta,\mathcal{P}}}$

2. $\sqrt[n-q]{\prod_{i=q+1}^{n} P^2_{R_\theta,\mathcal{P},[<i]}(t_i)}$ reflects the behaviour of $P^2_{R_\theta,\mathcal{P}}(t)$

In the light of the standard bound (6.25) for approximation errors it is desirable to make both factors small, and the target function (7.26) realizes a compromise between these two objectives. This seems to be an adequate response to the observation in Section 7.3 that it may be beneficial to involve the power function (as a measure of potential prediction accuracy) in the parameter estimation process. On the other hand, it can be seen as a remedy to the problem of the procedure presented in [13, Sec. 17.2.1] which is based exclusively on the power function which, as remarked by the author, ignores the dependency of $\|f\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}$ on θ.

If we minimize the logarithm of (7.26) and plug in the expressions derived in the two preceding propositions, a comparison with (7.7) shows that the resulting target function coincides, up to some unimportant constants, with the negative (restricted) profile log likelihood. Hence, in this section we have given a motivation for (RE)ML that is independent of any assumption on the mechanism that created f. In the following we shall refer to the procedure that minimizes (7.26) as MLE or REML estimator.

Once again we shall point out that implicitly we did make an additional assumption to justify the weighting of the approximation errors, namely that for some $\upsilon > 0$

$$\begin{aligned} \big(f(t_k) - s_{R_\theta,\mathcal{P},[<k]}(t_k)\big)^2 &\approx \upsilon \cdot P^2_{R_\theta,\mathcal{P},[<k]}(t_k), \qquad k = q+1,\dots,n, \\ \big(f(t) - s_{R_\theta,\mathcal{P}}(t)\big)^2 &\approx \upsilon \cdot P^2_{R_\theta,\mathcal{P}}(t), \qquad t \in T \end{aligned} \tag{7.28}$$

with "$\approx$" in the sense that the deviations are not systematic and do somehow "average out". Whenever this assumption is satisfied we should expect (RE)ML to perform better than LOOCV. When (7.28) is inappropriate however, (RE)ML may produce parameter estimates that are systematically too small or too big. Two typical situations where this is likely to happen are

- If the sequence $\|s_{R_\theta,\mathcal{P},[\le q]}\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}, \|s_{R_\theta,\mathcal{P},[\le q+1]}\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}, \dots$ converges fast.

  In this case one can see from (7.27) that the sequence $\tilde{\varepsilon}^{(w)}_{q+1}, \tilde{\varepsilon}^{(w)}_{q+2}, \dots$ quickly tends to 0, i.e. the magnitude of $\tilde{\varepsilon}_k$ decreases faster than $P_{R_\theta,\mathcal{P},[<k]}(t_k)$ and hence (7.28) cannot hold.

  If the norms of the interpolants approach $\|f\|_{\mathcal{H}_{R_\theta,\mathcal{P}}}$ only slowly (or not at all in case $f \notin \mathcal{H}_{R_\theta,\mathcal{P}}$), then for moderate n the magnitudes of $\tilde{\varepsilon}_{q+1}, \dots, \tilde{\varepsilon}_n$ should be similar and so (7.28) is more plausible.

  In the context of in spatial statistics we usually have $f \notin \mathcal{H}_{R_\theta,\mathcal{P}}$ (see Proposition 4.2.2) and so this caveat does not apply.

- If f behaves substantially different in different subregions of T. This can also entail systematic deviations from (7.28) that may not average out.

  Consider for instance a situation where the data are scattered and f exhibits strong fluctuations in exactly those subregions of T where the sampling locations are sparse. The corresponding components of $\tilde{\varepsilon}_\theta$ will then be assigned small weights, although it is in these subregions where the prediction errors will be largest.

  This is again in contrast to the situation in spatial statistics where the model assumptions usually imply that the behaviour of f does not radically differ from one subdomain of T to another, and so the weighting of the error components should never entail any systematic deviations.

We will see instances of both situations in the numerical examples in Section 7.5.

7.5 Comparing CV and ML in a Numerical Analysis Framework

Once again, we compare the performance of LOOCV and ML but now for functions $f : [0,1]^2 \to \mathbb{R}$ without any probabilistic background.

Specifically, we use three test functions (F1, F5, F9) from [31] and use the procedures MLE, WCV, CV1, CV2 to select the parameter of the interpolation kernel that is then used to

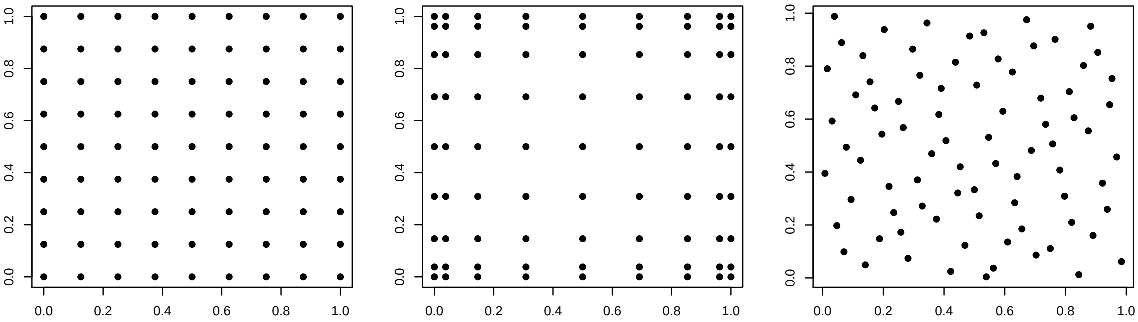

Figure 7.9: Different alignments of sampling locations (here for $n = 81$) used in the experiments of this section: equidistant points (left), tensor-product Chebyshev points (middle) Halton points (right).

approximate f based on its values at $\mathcal{T} := \{t_1, \ldots, t_n\}$. Experiments are carried out with $25, 81,$ and 289 data points. Since in many of the applications of approximation theory the centres are not fixed in advance but can be chosen freely, we also try out different alignments of data points:

- equidistant points
- tensor-product Chebyshev points
- Halton points

The tensor-product Chebyshev points are taken from [13] and have the advantage that they provide more information about f near the boundaries of T where the approximation accuracy is often lower. Halton points are an example of a quasi-random number sequence that can be created e.g. by the R-package “randtoolbox”. We use them to represent the situation of scattered data. For details about their definition we refer to [17] and [42]. The different types of alignments (abbreviated with E-n, C-n and H-n respectively) are illustrated in Figure 7.9.

7.5.1 Approximation accuracy

Interpolation will be carried out in the standard framework where $\mathcal{P} = \{0\}$ with a scaled version (with scaling parameter c) of the inverse multiquadrics (in spatial statistics called cauchy model):

$$\Phi_c(h) \;=\; \left(1 + \left(\frac{\|h\|}{c}\right)^2\right)^{-\frac{1}{2}}$$

as interpolation kernel. As in Section 7.3 we define an equidistant 100×100 evaluation grid $\mathcal{Q} \subset [0,1]^2$ and compare the root of the mean squared approximation errors RMSE(c) on our simulation grid $\mathcal{Q}$ for c_{opt}, c_{MLE}, c_{WCV}, c_{CV1}, and c_{CV2}.

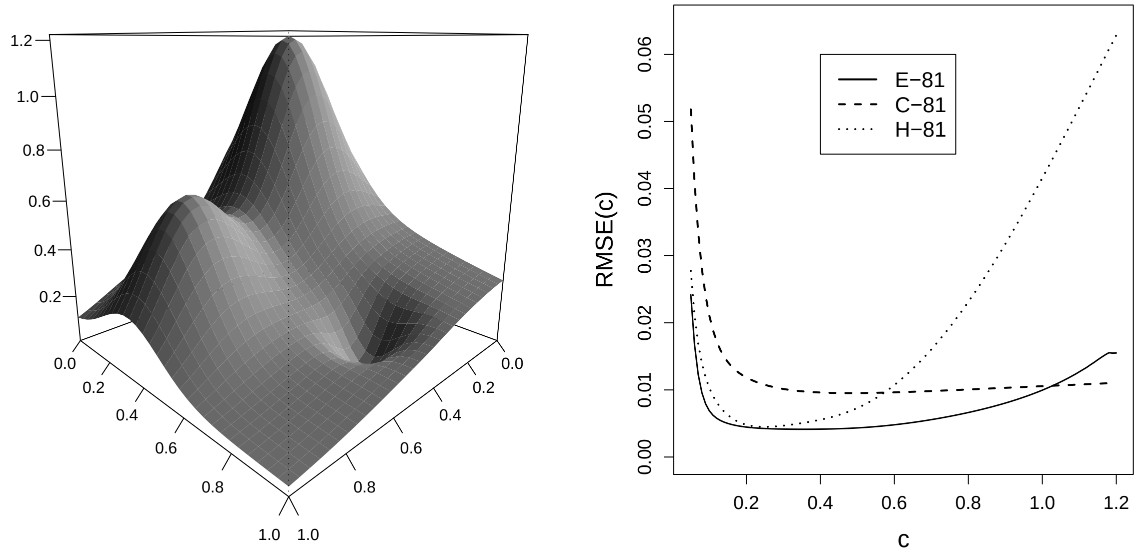

Figure 7.10: Perspective plot of F1 (left) and RMSE-curves for F1 for different point sets (right).

Our first test function is Franke's function (F1):

$$\begin{aligned} f(x,y) \;&=\; 0.75\,\exp\left(-\frac{(9x-2)^2+(9y-2)^2}{4}\right)+0.75\,\exp\left(-\frac{(9x+1)^2}{49}-\frac{9y+1}{10}\right)\\ &\quad+0.5\,\exp\left(-\frac{(9x-7)^2+(9y-3)^2}{4}\right)-0.2\,\exp\left(-(9x-4)^2-(9y-7)^2\right)\end{aligned}$$

A plot of F1 and the RMSE-curves for the experiments with 81 centres are given in Figure 7.10. We shall use it to make some general remarks:

First of all note that there is indeed for each curve a finite value of c where the RMSE is minimal. This is not self-evident and there are indeed examples where the minimum is attained for $c \to \infty$.
A remarkable observation that can be made in Figure 7.10 is that the RMSE-curves differ substantially for the different point alignments. While choosing c too small leads to large approximation errors with all of the point sets, the use of very big c's produces big errors when Halton points are used, but yields near-optimal approximants when tensor-product Chebyshev points are used.

In Figure 7.11 we have depicted the kernel interpolants of F1 based on 81 Halton points for different values of c. For small c our radial interpolation kernels are highly peaked and therefore cannot produce a good interpolant. If c is chosen too big, however, the corresponding interpolant approximates f well in the interior of $[0,1]^2$ but produces undesirable oscillations near the boundaries which blow up the approximation error (this is due to a connection between kernel interpolation and polynomial interpolation, see e.g. [11]). The same does not happen with Chebyshev centres which are comparatively dense near the boundaries and prevent such oscillations. The price for this, however, is a smaller data point density - and hence lower approximation accuracy - in the interior.

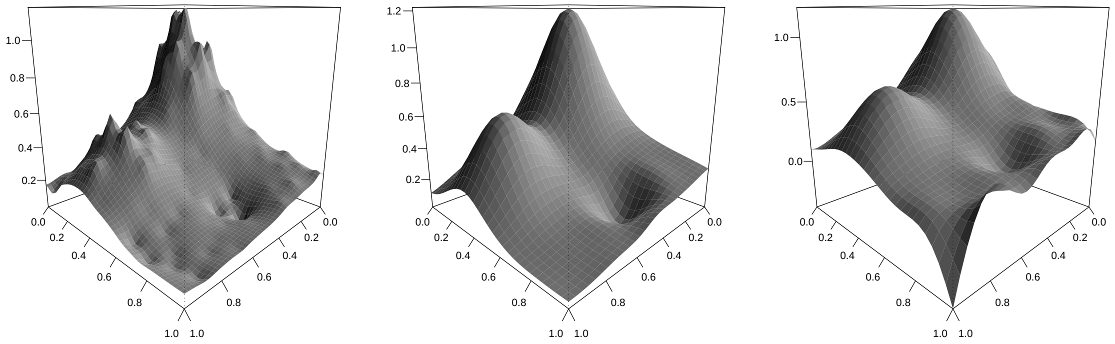

Figure 7.11: Approximation of F1 with 81 Halton points for $c = 0.03$ (left), 0.25 (middle) and 0.8 (right).

Note that the approximations of F1 are much better than those of the simulated sample paths of Gaussian RFs in Section 7.3. This is not surprising since F1 is much smoother than those sample paths. On the other hand, for the same reason, the sensitivity of the approximation accuracy to deviations of c from c_{opt} which was found to be small for the simulated RFs is much bigger for F1.
Table 7.6 shows the optimal parameter and the parameters chosen by the different procedures. The corresponding RMSEs are given in Table 7.7 with the results from the respective best procedure printed bold. Neither a clear over- nor a clear underperformance of any of the methods can be reported based on these results. In particular the MLE is competitive also in a non-statistical context but is no longer superior to LOOCV. The cases where c_{MLE} deviates from c_{opt} quite strongly can often (but not always) be explained by deviations from the assumption (7.28). We illustrate this by calculating the components of $\tilde{\varepsilon}_c$ in the experiment with 81 Halton points, once computed for $c = c_{\text{opt}}$ and once for $c = c_{\text{MLE}}$ (Figure 7.12). For the optimal c the first components tend to be bigger than the later components. The MLE tries to "correct" this by choosing a value of c that results in more uniform magnitudes of the different components. In this particular case however such a "correction" implies moving away from $c = c_{\text{opt}}$. We shall study an example where the magnitudes of the components of $\tilde{\varepsilon}_c$ are even much more different.

Our second test function is a scaled Gaussian kernel (function F5 from [31]):

$$f(x,y) \;=\; \frac{\exp\left(-\frac{81}{4}\left((x-0.5)^2+(y-0.5)^2\right)\right)}{3}$$

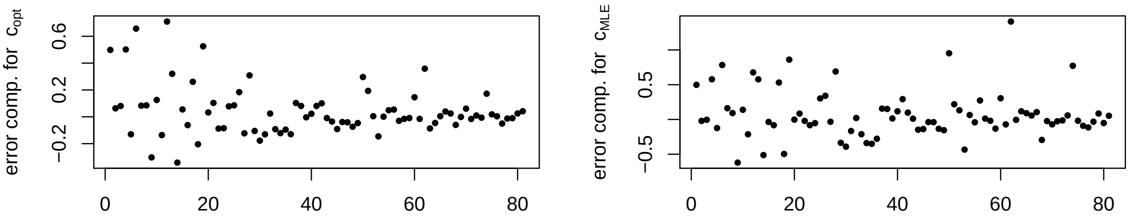

Figure 7.12: Components of $\tilde{\varepsilon}_c$ for $c = c_{\text{opt}}$ (left) and $c = c_{\text{MLE}}$ (right) in the setup with F1 and H-81.

	c_{opt}	c_{MLE}	c_{WCV}	c_{CV1}	c_{CV2}
E-25	0.28	0.31	0.25	0.26	0.28
C-25	0.47	0.32	0.22	0.20	0.25
H-25	0.20	0.40	0.67	0.45	0.45
E-81	0.35	0.38	0.34	0.33	0.39
C-81	0.48	0.44	0.42	0.45	0.43
H-81	0.25	0.36	0.28	0.31	0.28
E-289	0.43	0.39	0.45	0.46	0.47
C-289	0.45	0.39	0.50	0.50	0.50
H-289	0.44	0.39	0.37	0.46	0.46

	c_{opt}	c_{MLE}	c_{WCV}	c_{CV1}	c_{CV2}
E-25	0.27	0.20	0.31	0.20	0.40
C-25	0.18	0.28	0.40	0.16	0.37
H-25	0.31	0.34	0.39	0.42	0.39
E-81	0.58	0.59	0.72	0.62	0.69
C-81	0.39	0.54	0.77	0.77	0.95
H-81	0.45	0.61	0.57	0.46	0.48
E-289	0.66	0.78	0.67	0.71	0.71
C-289	0.78	0.77	0.59	0.72	0.72
H-289	0.71	0.76	0.65	0.67	0.67

Table 7.6: Optimal parameter and parameter estimates for test function F1 (left) and F5 (right).

	RMSE(c_{opt})	RMSE(c_{MLE})	RMSE(c_{WCV})	RMSE(c_{CV1})	RMSE(c_{CV2})
E-25	$2.586 \cdot 10^{-2}$	$2.604 \cdot 10^{-2}$	$2.603 \cdot 10^{-2}$	$2.593 \cdot 10^{-2}$	**$2.586 \cdot 10^{-2}$**
C-25	$4.270 \cdot 10^{-2}$	**$4.835 \cdot 10^{-2}$**	$6.675 \cdot 10^{-2}$	$7.286 \cdot 10^{-2}$	$5.927 \cdot 10^{-2}$
H-25	$3.299 \cdot 10^{-2}$	**$4.765 \cdot 10^{-2}$**	$7.986 \cdot 10^{-2}$	$5.266 \cdot 10^{-2}$	$5.266 \cdot 10^{-2}$
E-81	$4.140 \cdot 10^{-3}$	$4.145 \cdot 10^{-3}$	**$4.142 \cdot 10^{-3}$**	$4.145 \cdot 10^{-3}$	$4.150 \cdot 10^{-3}$
C-81	$9.531 \cdot 10^{-3}$	$9.551 \cdot 10^{-3}$	$9.577 \cdot 10^{-3}$	**$9.542 \cdot 10^{-3}$**	$9.562 \cdot 10^{-3}$
H-81	$4.492 \cdot 10^{-3}$	$5.147 \cdot 10^{-3}$	$4.567 \cdot 10^{-3}$	$4.741 \cdot 10^{-3}$	**$4.567 \cdot 10^{-3}$**
E-289	$3.823 \cdot 10^{-5}$	**$4.053 \cdot 10^{-5}$**	$4.095 \cdot 10^{-5}$	$4.431 \cdot 10^{-5}$	$4.924 \cdot 10^{-5}$
C-289	$3.918 \cdot 10^{-4}$	$3.996 \cdot 10^{-4}$	**$3.949 \cdot 10^{-4}$**	**$3.949 \cdot 10^{-4}$**	**$3.949 \cdot 10^{-4}$**
H-289	$6.267 \cdot 10^{-5}$	$9.399 \cdot 10^{-5}$	$1.252 \cdot 10^{-4}$	**$7.094 \cdot 10^{-5}$**	**$7.094 \cdot 10^{-5}$**

Table 7.7: Optimal RMSE and RMSE corresponding to the estimated parameters for test function F1.

	RMSE(c_{opt})	RMSE(c_{MLE})	RMSE(c_{WCV})	RMSE(c_{CV1})	RMSE(c_{CV2})
E-25	$1.468 \cdot 10^{-3}$	$4.104 \cdot 10^{-3}$	**$2.290 \cdot 10^{-3}$**	$4.104 \cdot 10^{-3}$	$4.747 \cdot 10^{-3}$
C-25	$7.560 \cdot 10^{-3}$	$1.073 \cdot 10^{-2}$	$1.302 \cdot 10^{-2}$	**$8.259 \cdot 10^{-3}$**	$1.258 \cdot 10^{-2}$
H-25	$5.789 \cdot 10^{-3}$	**$5.882 \cdot 10^{-3}$**	$6.330 \cdot 10^{-3}$	$6.719 \cdot 10^{-3}$	$6.330 \cdot 10^{-3}$
E-81	$4.012 \cdot 10^{-6}$	**$5.355 \cdot 10^{-6}$**	$1.027 \cdot 10^{-4}$	$2.002 \cdot 10^{-5}$	$7.297 \cdot 10^{-5}$
C-81	$5.230 \cdot 10^{-5}$	**$1.180 \cdot 10^{-4}$**	$2.357 \cdot 10^{-4}$	$2.357 \cdot 10^{-4}$	$3.644 \cdot 10^{-4}$
H-81	$1.006 \cdot 10^{-5}$	$4.268 \cdot 10^{-5}$	$2.916 \cdot 10^{-5}$	**$1.050 \cdot 10^{-5}$**	$1.255 \cdot 10^{-5}$
E-289	$2.345 \cdot 10^{-10}$	$3.633 \cdot 10^{-10}$	**$3.046 \cdot 10^{-10}$**	$1.422 \cdot 10^{-9}$	$1.422 \cdot 10^{-9}$
C-289	$7.099 \cdot 10^{-11}$	**$1.054 \cdot 10^{-10}$**	$1.446 \cdot 10^{-9}$	$3.009 \cdot 10^{-10}$	$3.009 \cdot 10^{-10}$
H-289	$1.004 \cdot 10^{-9}$	$1.946 \cdot 10^{-9}$	$2.004 \cdot 10^{-9}$	**$1.039 \cdot 10^{-9}$**	**$1.039 \cdot 10^{-9}$**

Table 7.8: Optimal RMSE and RMSE corresponding to the estimated parameters for test function F5.

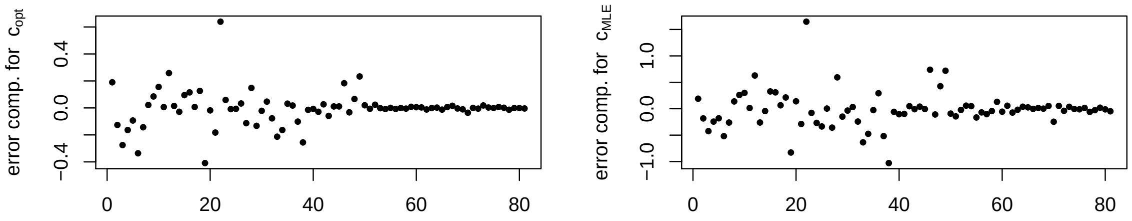

Figure 7.13: Components of $\tilde{\varepsilon}_c$ for $c = c_{\text{opt}}$ (left) and $c = c_{\text{MLE}}$ (right) in the setup with F5 and H-81.

F5 is an example of a function where assumption (7.28) cannot hold since the sequence $\|s_{\Phi_c,[\le 1]}\|_{\mathcal{H}_{\Phi_c}}, \|s_{\Phi_c,[\le 2]}\|_{\mathcal{H}_{\Phi_c}}, \ldots$ of the RKHS-norms of the interpolants converges to $\|f\|_{\mathcal{H}_{\Phi_c}}$ rather quickly. As discussed at the end of Section 7.4, this implies that the magnitudes of the components $\tilde{\varepsilon}_k$ for big k are considerably smaller than those for small k and so it is not clear if the MLE produces reasonable estimates.

The results in Table 7.6 and 7.8 show that the MLE again performs comparable to the LOOCV procedures although we can indeed observe a certain tendency of the MLE to choose c too big for larger $n = 81$ and $n = 289$.

In Figure 7.13 we have again calculated the components of $\tilde{\varepsilon}_c$ in the experiment with 81 Halton points, now for F5 and the respective values of c_{opt} and c_{MLE}. We observe that the error components with indices ≥ 50 are quite close to zero which results from the fact that $\|s_{\Phi_c,[\le 50]}\|_{\mathcal{H}_{\Phi_c}}$ is already very close to $\|f\|_{\mathcal{H}_{\Phi_c}}$. As before, for the choice $c = c_{\text{MLE}}$ the magnitudes of the error components become a bit more similar although here the dissimilarities are only reduced but not eliminated. This “correction” again explains the preference of bigger c's by the MLE, but at least in this example the resulting bias does not render the MLE uncompetitive.

In our comparison of MLE, CV1 and CV2 for the other test functions from [31] we found that in most cases MLE yields comparable or even slightly superior choices of c than CV1 and CV2. Noticeable exceptions are the test functions that favour very big values of c, but in this case the additional problem of ill-conditioned interpolation systems arises, a discussion of which would be beyond the scope of this work.

Instead, we want to study if assumption (7.28) even allows for a prediction of the L^2-approximation error.

7.5.2 Prediction of the L^2-approximation error

If assumption (7.28) is adequate it implies in our situation

$$\text{(a)} \quad \upsilon \quad \approx \quad \frac{1}{n}\sum_{k=1}^{n}\left(\tilde{\varepsilon}_k^{(w)}\right)^2 \;=\; \frac{1}{n}\left\|s_{\Phi_c}\right\|_{\Phi_c} \;=\; \frac{1}{n}\,\mathrm{f}' A_c^{-1}\,\mathrm{f}$$

$$\text{(b)} \quad \|f - s_{\Phi_c}\|^2_{L^2(T)} \;=\; \int_T \left(f(t) - s_{\Phi_c}(t)\right)^2 dt \quad \approx \quad \upsilon \int_T P^2_{\Phi_c}(t)\, dt$$

	Pred. RMSE (MLE)	rel. error	Pred. RMSE (CV2)	rel. error
E-25	$5.411 \cdot 10^{-2}$	+ 108 %	$5.345 \cdot 10^{-2}$	+ 107 %
C-25	$6.376 \cdot 10^{-2}$	+ 32 %	$6.598 \cdot 10^{-2}$	+ 11 %
H-25	$4.830 \cdot 10^{-2}$	+ 1 %	$4.385 \cdot 10^{-2}$	- 17 %
E-81	$3.861 \cdot 10^{-3}$	- 7 %	$3.427 \cdot 10^{-3}$	- 17 %
C-81	$5.082 \cdot 10^{-3}$	- 47 %	$5.204 \cdot 10^{-3}$	- 46 %
H-81	$6.868 \cdot 10^{-3}$	+ 33 %	$5.095 \cdot 10^{-3}$	+ 11 %
E-289	$9.421 \cdot 10^{-5}$	+ 132 %	$2.703 \cdot 10^{-5}$	- 45 %
C-289	$1.999 \cdot 10^{-4}$	- 50 %	$9.867 \cdot 10^{-5}$	- 75 %
H-289	$3.061 \cdot 10^{-4}$	+ 226 %	$2.625 \cdot 10^{-4}$	+ 270 %

Table 7.9: Predicted RMSEs for F1 and relative deviation of this prediction from the actual RMSE.

Approximation (a) suggest a method to determine υ. By comparing with the derivation of the profile log likelihood in (7.5) we see that the υ here coincides with the variance parameter in the statistical context and (a) coincides with υ_{MLE}.

Approximation (b) shows how we can use the estimate of υ suggested by (a) to obtain a prediction of the L^2-approximation error. Instead of $\|f - s_{\Phi_c}\|_{L^2([0,1]^2)}$, however, we use RMSE$(c)$ which can be interpreted as a discrete approximation. In the same way, we approximate the rhs of (b) by

$$\upsilon \int_{[0,1]^2} P^2_{\Phi_c}(t)\, dt \quad \approx \quad \frac{1}{|\mathcal{Q}|} \sum_{t \in \mathcal{Q}} P^2_{\upsilon\Phi_c}(t)\,. \tag{7.29}$$

where $|\mathcal{Q}|$ denotes the number of points in our evaluation grid $\mathcal{Q}$. If assumption (7.28) is adequate, then this yields a prediction of RMSE(c).

In the LOOCV framework, we can make the assumption

$$\begin{aligned} \big(f(t_k) - s_{R_\theta,\mathcal{P},[-k]}(t_k)\big)^2 \quad &\approx \quad \upsilon \cdot P^2_{R_\theta,\mathcal{P},[-k]}(t_k), \qquad k = 1, \ldots, n, \\ \big(f(t) - s_{R_\theta,\mathcal{P}}(t)\big)^2 \quad &\approx \quad \upsilon \cdot P^2_{R_\theta,\mathcal{P}}(t), \qquad t \in T \end{aligned}$$

similar to (7.28) which leads to the estimate υ_{CV2} introduced in Section 7.3. Plugging this estimate into (7.29) we have again a prediction of RMSE(c).

Tables 7.9 and 7.10 show the predicted RMSEs that are obtained as described above by estimating $(\upsilon, c)_\star$, $\star =$ MLE, CV2 and calculating (7.29) with kernel $\upsilon_\star \Phi_{c_\star}$ over the grid $\mathcal{Q}$. In addition, we calculate the relative deviation of this predicted RMSE from RMSE$(c_\star)$, the actual RMSE from Tables 7.7 and 7.8 that is obtained when $c_\star$ is used for interpolation.

Even for test function F1 the predictions are not very good in general, neither for $(\upsilon, c)_{\text{MLE}}$ nor for $(\upsilon, c)_{\text{CV2}}$. They may serve as a guess on the magnitude of the L^2-approximation error but are far less accurate than in the statistical context.

The situation is even worse for test function F5. The bad results for the MLE in the examples with $n = 81$ and $n = 289$ could be anticipated since we already saw that assumption (7.28) is totally inappropriate in this case. Specifically, the value of υ that would be adequate for

	Pred. RMSE (MLE)	rel. error	Pred. RMSE (CV2)	rel. error
E-25	$2.712 \cdot 10^{-2}$	+ 560 %	$1.681 \cdot 10^{-2}$	+ 254 %
C-25	$2.651 \cdot 10^{-2}$	+ 147 %	$2.546 \cdot 10^{-2}$	+ 102 %
H-25	$1.794 \cdot 10^{-2}$	+ 205 %	$1.268 \cdot 10^{-2}$	+ 100 %
E-81	$4.678 \cdot 10^{-4}$	+ 8635 %	$5.857 \cdot 10^{-5}$	- 20 %
C-81	$9.938 \cdot 10^{-4}$	+ 742 %	$9.291 \cdot 10^{-4}$	+ 153 %
H-81	$8.749 \cdot 10^{-4}$	+ 1950 %	$9.095 \cdot 10^{-5}$	+ 624 %
E-289	$5.884 \cdot 10^{-7}$	+ 161900 %	$2.766 \cdot 10^{-10}$	- 81 %
C-289	$3.715 \cdot 10^{-7}$	+ 352400 %	$1.309 \cdot 10^{-9}$	+ 335 %
H-289	$3.784 \cdot 10^{-6}$	+ 109400 %	$1.293 \cdot 10^{-9}$	+ 24 %

Table 7.10: Predicted RMSEs for F5 and relative deviation of this prediction from the actual RMSE.

the first error components is much bigger than the value for the later components and for the predictions at unobserved data sites. As a consequence, v, and hence the predicted L^2-approximation error, is grossly overestimated in these cases. For $n = 25$ and for the CV2 predictions this explanation does not apply but still we note that the predictions are even worse than those for F1.

As a conclusion of the experiments so far in this section we note

1. In the context of kernel interpolation both ML and LOOCV methods can be recommended for selecting unknown kernel parameters. At least in the cases where the issue of ill-conditioned interpolation systems does not come into play the parameters chosen by these procedures in general lead to satisfactory approximants.

2. A reasonable prediction of the L^2-approximation error as obtained under statistical model assumptions is in general not available in the context of kernel interpolation. While the MLE estimates turned out to be not very sensitive to violations of assumption (7.28) when the purpose is approximating f, these assumptions need to be satisfied sufficiently well in order to yield adequate predictions of the L^2-approximation error.

We finally study if (RE)ML or LOOCV can be used to guess the smoothness of f.

7.5.3 Choosing the smoothness of the interpolation kernel

The two test function considered above are both in $\mathcal{C}^\infty\big([0,1]^2\big)$ which justifies the use of the infinitely smooth inverse multiquadric kernel. Now we want to consider a test function f with finite smoothness and use the Whittle-Matérn kernel with the parametrization as in (6.33). By Theorem 6.7.1 we know that we can expect optimal convergence rates also for kernels whose corresponding RKHS is smoother than f, and so from the point of view of prediction accuracy it is not clear how smooth the interpolation kernel should finally be. It is therefore interesting to ask

- Which kernel smoothness yields the best predictions? In particular, are kernels prefered whose corresponding RKHS contains f?

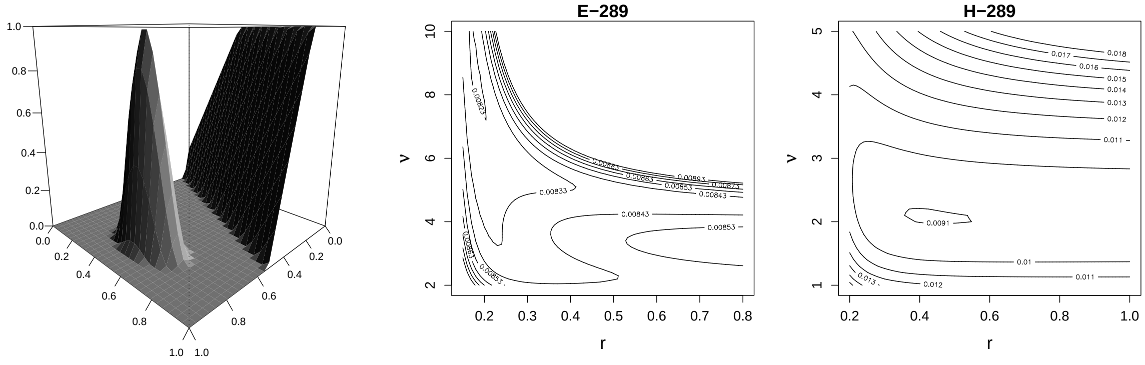

Figure 7.14: Perspective plot of F9 (left) and contour plots of RMSE(r, ν) for E-289 and H-289 (right).

- Which kernel smoothness is suggested by (RE)ML and LOOCV methods based on the available data?

A first answer to this issue is obtained by looking again at the simulation study in Section 7.3. The simulated sample paths studied there were all (according to the theoretical results from Section 5.5) just barely not in $W^{1,2}\big([-1,1]^2\big)$ and we could just ignore their stochastic background, take a numerical analyst's point of view, and use this smoothness information only. In any case a look at Figure 7.4 then tells us that the best prediction is obtained for a smoothness parameter ν somewhere around 1.0, which corresponds to a kernel that is reproducing in $W^{2,2}\big([-1,1]^2\big)$. This suggests that we should use a kernel for interpolation whose corresponding RKHS is by $d/2$ smoother than f. We shall investigate if the same conclusion holds for a test function without stochastic background.

We study the interpolation behaviour of test function F9 from [31]:

$$f(x,y) \;=\; \begin{cases} 1 & \text{if} \quad y-\xi \geq 1/2, \\ 2(y-\xi) & \text{if} \quad 0 \leq y-\xi \leq 1/2, \\ (\cos(4\pi r(\xi,y)\,)+1)/2 & \text{if} \quad r(\xi,y) \leq 1/4, \\ 0, & \text{otherwise} \end{cases}$$

where

$$r(\xi,y) = \sqrt{\left(\xi - \frac{3}{2}\right)^2 + \left(y - \frac{1}{2}\right)^2}, \qquad \xi = 2.1\,x - 0.1$$

This function has jumps in its first derivative, but the jumps occur only on a set of λ^2-measure zero and one can see that f is just barely not in $W^{1.5,2}([0,1]^2)$. It is depicted in Figure 7.14 on the left.

We only give results for interpolation based on the point sets E-289 and H-289. Using less than 289 centres does not give a satisfactory reproduction of f and Chebyshev points also do not seem appropriate for a function with such irregular behaviour in the interior.

Optimal parameters, parameter estimates and corresponding RMSEs for these two setups are given in Tables 7.11 and 7.12.

	r_{opt}	r_{MLE}	r_{WCV}	r_{CV1}	r_{CV2}	ν_{opt}	ν_{MLE}	ν_{WCV}	ν_{CV1}	ν_{CV2}
E-289	0.18	0.222	0.126	0.153	0.147	8.6	3.45	15.3	10.4	11.3
H-289	0.42	0.241	0.235	0.580	0.277	2.1	2.86	2.84	2.24	2.63

Table 7.11: Optimal parameter and parameter estimates for test function F9.

	RMSE(θ_{opt})	RMSE(θ_{MLE})	RMSE(θ_{WCV})	RMSE(θ_{CV1})	RMSE(θ_{CV2})
E-289	$8.220 \cdot 10^{-3}$	$8.326 \cdot 10^{-3}$	$8.570 \cdot 10^{-3}$	$\mathbf{8.263 \cdot 10^{-3}}$	$8.287 \cdot 10^{-3}$
H-289	$9.082 \cdot 10^{-3}$	$9.661 \cdot 10^{-3}$	$9.687 \cdot 10^{-3}$	$\mathbf{9.161 \cdot 10^{-3}}$	$9.385 \cdot 10^{-3}$

Table 7.12: Optimal RMSE and RMSE corresponding to the estimated parameters for test function F9.

First note that the optimal choices of r and ν strongly depend on the geometry of the point sets. With respect to smoothness we observe that the values of ν that yield optimal interpolants are relatively big, especially so for E-289. This is remarkable because $W^{1.5,2}([0,1]^2)$ is very rough and corresponds to a reproducing kernel $\Phi_{r,0.5}$. Hence, we might expect the optimal ν to be close to 0.5 but this is neither true for the optimal nor for the estimated ν's. The precision of the approximation with estimated parameters is again quite satisfactory, now with superior performance of CV1 and CV2.

To sum up we can say that our parameter selection procedures (MLE, CV1, CV2) are suitable to identify parameters that yield good approximants but cannot be used to identify the smoothness of f. In statistical setup this was possible indirectly by identifying the covariance function and applying the theorems from Section 5.5.

The situation there is however different in that the sample paths typically have the same regularity *all over* T whereas our example here was irregular only on a λ^2-null set. This is obviously a point where the stochastic model assumption (which assumes a probability measure on the space of all function) is quite crucial and the methodology that can be used to estimate the smoothness of f in the statistical context cannot be carried over to the context of approximation theory.

Bibliography

[1] R.J. Adler and J.E. Taylor. Random Fields and Geometry. Springer, 2007.

[2] J.M. Azaïs and M. Wschebor. Level Sets and Extrema of Random Processes and Fields. John Wiley & Sons, 2009.

[3] H. Bauer. Probability Theory and Elements of Measure Theory. Academic Press Inc., London, New York, Toronto, Sydney, San Francisco, second english edition, 1981.

[4] A. Berlinet and C. Thomas-Agnan. Reproducing Kernel Hilbert Spaces in Probability and Statistics. Kluwer, 2004.

[5] P. Billingsley. Probability and Measure. John Wiley & Sons, New York, Chichester, Brisbane, third edition, 1995.

[6] Goldie C.M. Bingham, N.H. and J.L. Teugels. Regular Variation. Cambridge University Press, Cambridge, 1987.

[7] P. Caragea and R.L. Smith. Asymptotic properties of computationally efficient alternative estimators for a class of multivariate normal models. Journal of Multivariate Analysis, 98(7):1417–1440, 2007.

[8] J.-P. Chilès and P. Delfiner. Geostatistics. Modeling Spatial Uncertainty. John Wiley & Sons, New York, Chichester, 1999.

[9] K.L. Chung. A course in Probability Theory. Harcourt, Brace & World, Inc., New York, 1968.

[10] Haroske D.D. and Triebel H. Distributions, Sobolev Spaces, Elliptic Equations. European Mathematical Society, 2008.

[11] T.A. Driscoll and B. Fornberg. Interpolation in the limit of increasingly flat radial basis functions. Comp. Math. Appl., 43:413–422, 2002.

[12] L.C. Evans. Partial Differential Equations. American Mathematical Society, 2002.

[13] G.E. Fasshauer. Meshfree Approximation Methods with Matlab. World Scientific Publishing Co. Pte. Ltd., 2007.

[14] T.S. Ferguson. A Course in Large Sample Theory. Chapman & Hall, London, 1996.

[15] I.I. Gihman and A.V. Skorohod. The Theory of Stochastic Processes I. Springer, Berlin, Heidelberg, New York, 1974.

[16] I.S. Gradshteyn and I.M. Ryzhik. Table of Integrals, Series and Products. Academic Press, seventh edition, 2007.

[17] J.H. Halton. On the efficiency of certain quasi-random sequences of points in evaluating muti-dimensional integrals. Numer. Math., 2:84–90, 1960.

[18] M.S. Handcock and Wallis J.R. An approach to statistical spatial-temporal modeling of meteorological fields (with discussion). J. Amer. Statist. Assoc., 89(7):368–390, 1994.

[19] D.A. Harville. Bayesian inference for variance components using only error contrasts. Biometrika, 61:383–385, 1974.

[20] C.C. Heyde. Quasi-Likelihood and its Application. Springer, New York, 1997.

[21] I.A. Ibragimov and R.Z. Has'minskii. Statistical Estimation: Asymptotic Theory. Springer, New York, 1981. trans. S. Kotz.

[22] J.T. Kent. Continuity properties for random fields. Ann. Probab., 17(4):1432–1440, 1989.

[23] H. König. Eigenvalue distribution of compact operators. Birkhäuser Verlag, 1986.

[24] John O. Kufner, A. and S. Fučík. Function Spaces. Noordhoff International Publishing, Leyden, 1977.

[25] C. Lantuéjoul. Geostatistical Simulation. Springer, Berlin, 2002.

[26] K.V. Mardia and R.J. Marshall. Maximum likelihood estimation of models for residual covariance in spatial statistics. Biometrika, 71:135–146, 1984.

[27] G. Matheron. Leçons sur les fonctions aléatoires d'ordre 2. Technical Report C-53, Ecole des Mines de Paris, 1972.

[28] V.G. Maz'ja. Sobolev Spaces. Springer Verlag, Berlin, Heidelberg, New York, Tokyo, 1980.

[29] P. McCullagh and J.A. Nelder. Generalized Linear Models. Chapman & Hall, London, second edition, 1989.

[30] Ward J.D. Narcowich, F.J. and H. Wendland. Sobolev error estimates and a bernstein inequality for scattered data interpolation via radial basis functions. Constr. Approx., 24:175–186, 2006.

[31] S. Rippa. An algorithm for selecting a good value for the parameter c in radial basis function interpolation. Adv. Comput. Math., 11:193–210, 1999.

[32] W. Rudin. Real and Complex Analysis. McGraw-Hill series in higher mathematics, second edition, 1974.

[33] F.J. Samper and Neuman S.P. Estimation of spatial covariance structures by adjoint state maximum likelihood cross validation: 1. theory. Water Resour. Res., 25:351–362, 1989.

[34] R. Schaback. Reconstruction of multivariate functions from scattered data. Technical report, 1997.

[35] R. Schaback. Improved error bounds for scattered data interpolation by radial basis functions. Math. Comp., 68:201–216, 1999.

[36] M. Schlather. Simulation and analysis of random fields. R News, 1(2):18–20, 2001.

[37] S.R. Searle. Linear Models. John Wiley & Sons, Inc., New York, Chichester, Weinheim, Brisbane, Singapore, Toronto, 1997.

[38] Chi Z. Stein, M.L. and L.J Welty. Approximating likelihoods for large spatial data sets. J. R. Statist. Soc B, 66(2):275–296, 2004.

[39] M.L. Stein. Interpolation of Spatial Data. Springer, Heidelberg, New York, 1999.

[40] H. Wendland. Sobolev-type error estimates for interpolation by radial basis functions. pages 337–344, 1997.

[41] H. Wendland. Scattered Data Approximation. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press, Cambridge, 2005.

[42] Luk W.-S. Wong, T.-T. and P.-A. Heng. Sampling with hammersley and halton points. J. Graphics Tools, 2:9–24, 1997.

[43] A.M. Yaglom. Correlation Theory of Stationary and Related Random Functions I, Basic Results. Springer, New York, Berlin, 1987.

[44] H. Zhang. Inconsistent estimation and asymptotically equivalent interpolations in model-based geostatistics. J. Am. Statist. Assoc., 99:250–261, 2004.

[45] H. Zhang and D.L. Zimmerman. Towards reconciling two asymptotic frameworks in spatial statistics. Biometrika, 92(4):921–936, 2005.

Curriculum Vitae

Name:	Michael Scheuerer
Date of birth:	07.04.1981
Place of birth:	Amberg
Nationality:	German
Address:	Zimmermannstr. 13, Ap. 70 37075 Göttingen, Germany
Phone:	+49 (0)551 995 1627
E-mail:	scheuer@math.uni-goettingen.de

School education & Military service

1987-1991	Dreifaltigkeitsschule I (primary school), Amberg
1991-2000	Gregor-Mendel-Gymnasium (secondary school), Amberg
2000-2001	Gebirgstransportbataillon 83 (military service), Kümmersbruck

Study and Scientific career

2001-2006	Study of mathematics at Universität Bayreuth, Germany and Université de Marne-la-Vallée, France (for one semester)
09/2006	Diplom in mathematics at Universität Bayreuth, Germany (with distinction)
since 01/2007	PhD student in the DFG graduate programme “Identification in Mathematical Models” at Georg-August-Universität Göttingen, Germany

Publications

(under revision)	An alternative procedure for selecting a good value for the parameter c in RBF-interpolation. Submitted to Adv. Comput. Math.

Language abilities

German	native
English	fluent
French	advanced level
Spanish	intermediate level