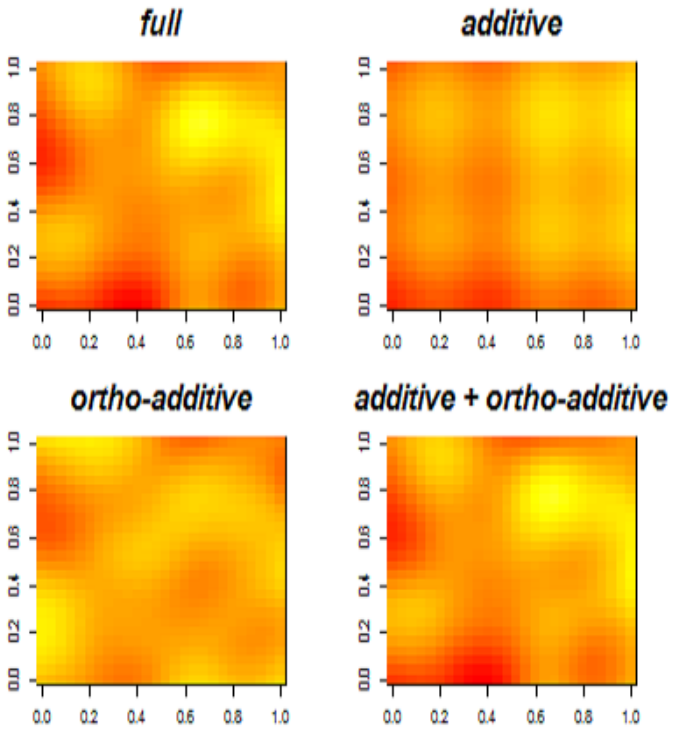
Additivity and Ortho-Additivity in Gaussian Random Fields

Nicolas Lenz

Additivity and Ortho-Additivity in Gaussian Random Fields

Master Thesis by Nicolas Lenz
2013
University of Bern, Mathematical Institute

Supervised by
Dr. David Ginsbourger and Prof. Dr. Dominic Schuhmacher



Contents

1 Introduction **4**

2 Random fields **7**
- 2.1 General definition . . . 7
- 2.2 Gaussian random fields . . . 9
- 2.3 Path regularity . . . 9
- 2.4 Simulation of Gaussian random fields . . . 12
  - 2.4.1 Positive definite kernels . . . 13

3 Decompositions in L^2 **15**
- 3.1 Fundamentals . . . 15
- 3.2 Retrieving additive and ortho-additive functions . . . 17

4 Projecting Gaussian random fields **24**
- 4.1 Projecting Gaussian random field paths . . . 24
- 4.2 Projecting covariance kernels . . . 26
  - 4.2.1 Product kernels . . . 27
  - 4.2.2 The Gaussian kernel . . . 30
- 4.3 Schematic representation of a covariance kernel . . . 31

5 Application and numerical experiments **33**
- 5.1 Kriging . . . 33
- 5.2 Illustrations . . . 34
- 5.3 Effect of a misspecified kernel in two dimensions . . . 36
  - 5.3.1 Setup of the experiment . . . 37
  - 5.3.2 Results . . . 38
- 5.4 Quantifying additivity in higher dimensions . . . 41
  - 5.4.1 Applying maximum likelihood estimation . . . 41
  - 5.4.2 Setup of the experiment . . . 43
  - 5.4.3 Results . . . 44

6 Conclusion and Perspectives **49**

7 Appendix **51**

References **56**

Acknowledgments

I address my special thanks to my advisors David and Dominic for their enduring efforts to advance the thesis and for their carefully balanced motivation and pressure. I very much appreciate the helpful assistance when I became acquainted with the topic, the intense support when I dove deeper and the reviews and feedbacks towards the end of the thesis. In particular I am grateful for facilitating the two journeys to Göttingen and Nice which were at the same time an impulse for the research progress and some of the personal highlights of the last year.

Furthermore I thank Lutz Dümbgen for creating the right conditions to make this thesis possible.

Last but not least I want to express my appreciation for the effort of Nicolas Durrande who provided me with an extra release of the GPy package which facilitated some of the here presented experiments.

1 Introduction

The thesis deals with Gaussian random field (GRF) models and their use in functional approximation. More particularly we focus here on a novel class of kernels leading to Gaussian random fields with paths that are orthogonal to the space of additive functions. One of the main motivations of this work, where a number of concepts from functional analysis and spatial statistics meet, is to modestly contribute to recent extensions of Gaussian process regression techniques for high-dimensional problems.

With growing dimension, kriging suffers from the curse of dimensionality [BC56]. Typical designs require a number of observations that grows exponentially with respect to the dimension. A promising approach to escape the curse is described in [DGR12], where the used kriging models are based on generalized additive models [HT90]. The results confirm the suggestion about additive models that was already expressed in [Cre93, p.284]: *"It would be worth investigating whether this regression technology could be adapted to spatial prediction"*.

[DGR12] shows that Gaussian random fields with an additive covariance structure have additive paths, i.e. paths of the form $f(x_1, \ldots, x_d) = c + f_1(x_1) + \ldots + f_d(x_d)$. Kriging models based on such kernels turned out to be particularly useful in high dimensional. Experiments were carried out with a linear budget of observations in which additive models worked well compared to standard kriging models.

In other cases the assumption of additivity is too restrictive. Therefore it seems reasonable to enrich an additive model in order to allow for non-additivity. Finding a suitable complement for additive kernels is one of the main objectives of this thesis.

In order to reach this goal we start by identifying the additive part of the underlying Gaussian random field. We assume L^2 paths and, inspired by [DGRC13], we offer a decomposition of L^2 into the subspace $\mathcal{A}$ of additive functions and its orthogonal complement which we call the space $\mathcal{O}$ of ortho-additive functions. We derive the according orthogonal projections $\pi_{\mathcal{A}}$ and $\pi_{\mathcal{O}}$.

In the sequel of the thesis we extent $\pi_{\mathcal{A}}$ and $\pi_{\mathcal{O}}$ to projections which can be applied to covariance kernels. Then we consider centered Gaussian random fields whose covariance structures are expressed by projected kernels and state that their paths are additive, respectively ortho-additive.

The thesis shows the full calculation of the additive and the ortho-additive projection of kernels which have a product structure

$$k(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} k_i(x_i, y_i) \ .$$

In particular explicit formulae are presented in the case of a squared exponential (or Gaussian) kernel.

It is also an objective of the thesis that the newly projected kernels are assessed. For this purpose we offer an original cross-check technique which illustrates the quality of kriging predictions generated with the true kernel and with a set of misspecified kernels.

We hence propose to use a GRF model which is based on a kernel of the form $\alpha \ k_{\mathcal{A}} + (1 - \alpha) \ k_{\mathcal{O}}$. $k_{\mathcal{A}}$ and $k_{\mathcal{O}}$ denote respectively the additive and ortho-additive parts of a Gaussian kernel, α is some coefficient of additivity in $[0, 1]$. A numerical experiment is conducted for investigating the properties of the Maximum Likelihood Estimator of α. The obtained results suggest that it is unrealistic to hope recovering α based on a single realization in the proposed experimental setup.

The outline of the thesis is as follows. In Chapter 2 we recall some basic definitions and properties about random fields with a special emphasis on GRFs and the role of the covariance kernel. Then we discuss sufficient conditions for paths of random fields to be measurable, continuous or L^2.

Chapter 3 is about a decomposition of square-integrable functions into a subspace of additive functions and its orthogonal complement, here called the space of ortho-additive functions. First, it presents concisely some results from functional analysis and then, in more detail, how the results can be used to derive a decomposition of L^2 into subspaces containing additive and ortho-additive functions, respectively. We derive the according orthogonal projections.

The interplay between the two preceding chapters is shown in Chapter 4. We extend the additive and ortho-additive projections such that they can be applied to covariance kernels. A GRF with such a projected kernel has additive or ortho-additive paths, respectively. We present the general calculations for projecting a product kernel and the explicit formulae in the case of a Gaussian kernel.

Chapter 5 is dedicated to numerical experiments. First, Section 5.1 recalls the necessary formulae for applying kriging. In Section 5.2 we illustrate and

discuss kriging predictions of an additive and an ortho-additive kernel. This offers some intuition for the experiment in Section 5.3 where we compare a set of kernels in a two-dimensional setting. The idea is to choose two kernels from the set, perform kriging predictions with the first one, based on measurements that were generated with the second one, then repeating it for all possible pairs of kernels. We present the results for an additive kernel, an ortho-additive one and possible variations. Finally, in Section 5.4, we consider kernels which differ in their degree of additivity. In an experiment we generate data with such kernels and try to recover the degree of additivity with maximum likelihood estimation. We present the results and discuss their development with growing dimension.

Chapter 6 concludes the thesis.

2 Random fields

Some central objects that we will consider in this thesis are real-valued random fields. In this chapter we discuss random fields in general and the special case which is known as Gaussian random field. Then we will see some particular properties, characterizations and prospects.

2.1 General definition

A random field is a stochastic process whose index set is some multi-dimensional space.

Definition 1 (Random field)**.** *A (real-valued) random field (RF) is a collection of (real-valued) random variables defined over a common probability space* $(\Omega, \mathcal{F}, P)$ *and indexed by an element of a set* D*. We write* $(Z_x)_{x \in D}$*.*

Note that no particular assumption is made about D. However, for most considerations it is necessary that D is at least equipped with a topology, as will be the case in Section 2.3.

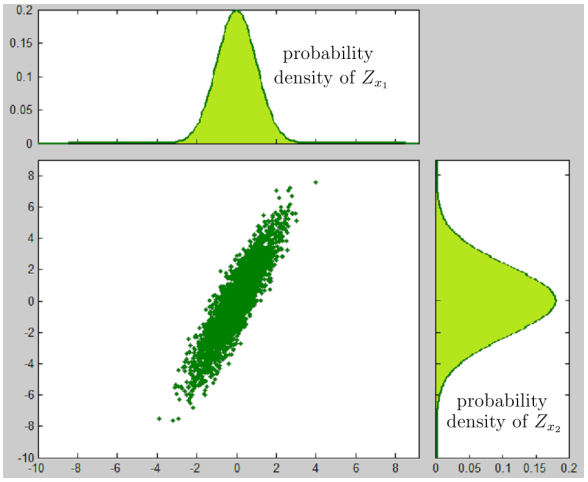
Since a random field Z is indexed by $x \in D$ and defined over a probability space $(\Omega, \mathcal{F}, P)$ we can consider it as a function of two arguments.

$$\begin{aligned} Z : D \times \Omega &\to \mathbb{R} \\ (x, \omega) &\mapsto Z_x(\omega) \end{aligned}$$

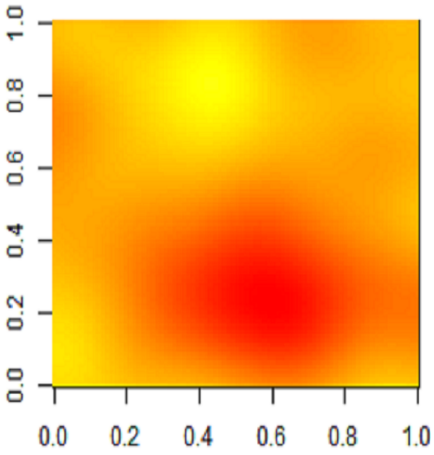
If we fix either x or ω we see two aspects of the random field. Both of them will be important in future investigations. Fixing $x \in D$ gives us a real-valued random variable Z_x. More generally we can fix a set $\{x_1, \ldots, x_n\}$ of n elements of D and we get a random vector $(Z_{x_1}, \ldots, Z_{x_n})$ (like, e.g., in Figure 1a. It describes the random field in the chosen set of locations. Fixing ω we get a realization, i.e. an elementary event of the random field as in Figure 1b. We speak of a trajectory or path. It can be seen as function $Z_\cdot(\omega) : D \to \mathbb{R}$ mapping x to $Z_x(\omega)$ for a fixed $\omega \in \Omega$.

Note that a priori we do not know much about the properties of such paths. In Section 2.3 we will see some sufficient conditions under which the paths are measurable, almost surely square-integrable or continuous.

As a random field is a collection of random variables, for any $x \in D$ we can speak of moments of the random variable Z_x. We assume here and in the following that all Z_x have a finite second moment (i.e. $\mathbb{E}[Z_x^2] < \infty$) such



(a) Samples of a random vector (Z_{x_1}, Z_{x_2}) with joint normal distribution; The respective probability density functions for Z_{x_1} and Z_{x_2}.



(b) Realization of a Gaussian random field having a squared exponential covariance function.

Figure 1: Aspects of a random field

that the expectation function m and the covariance kernel k of the random field can be defined:

$$m(x) := \mathbb{E}(Z_x) \ ,$$
$$k(x_1, x_2) := \mathrm{Cov}(Z_{x_1}, Z_{x_2}) \ .$$

m and k are important descriptors of a random field. In 2.2 we will see a class of random fields that are even entirely characterized by these quantities.

Definition 2. *A random field $(Z_x)_{x \in D}$ with finite second moment is called*

- *first order stationary if m is constant.*
- *second order stationary if it is first order stationary and $k(\boldsymbol{x} + \boldsymbol{h}, \boldsymbol{x})$ does not depend on $\boldsymbol{x}$. In this case we simplify the kernel function $k(\boldsymbol{h}) := k(\boldsymbol{x} + \boldsymbol{h}, \boldsymbol{x})$.*
- *isotropic if it is second order stationary and $k(\boldsymbol{h}) = k(\|\boldsymbol{h}\|)$.*

A special case of a first order stationary random field is the centered random field whose expectation function is trivial.

2.2 Gaussian random fields

We now get to some special kind of random fields.

Definition 3 (Gaussian random field, GRF)**.** *A random field Z over the domain D is called Gaussian if $(Z_{x_1}, \dots, Z_{x_n})$ is a Gaussian vector for any set $\{x_1, \dots, x_n\} \subseteq D$ and any $n \in \mathbb{N}$.*

Gaussian random fields play a special role because they have some nice properties. For one they are entirely characterized by their expectation function and covariance kernel (see, e.g., [Sch09, Lemma 2.4.18]). Moreover the class of Gaussian random fields is closed under certain transformations. For instance the sum of two Gaussian random fields is again Gaussian. And likewise if we apply a bounded linear transformation to a GRF then the result is Gaussian.

Example 1. *If D is in $\mathbb{R}^2$ we can display trajectories as images. Figure 2 shows some trajectories of centered Gaussian random fields with a covariance structure expressed by the exponential kernel based on the l_1-distance in $\mathbb{R}^2$ (upper row)*

$$k_1(x, y) = \sigma^2 \cdot e^{-\frac{|x_1 - y_1|}{\theta_1} - \frac{|x_2 - y_2|}{\theta_2}}$$

and the isotropic squared exponential (or "Gaussian") kernel (lower row)

$$k_2(\boldsymbol{x}, \boldsymbol{y}) = \sigma^2 \cdot e^{-\left(\frac{\|\boldsymbol{x} - \boldsymbol{y}\|}{\theta}\right)^2}.$$

In Example 1 above, notice that k_1 is not isotropic. k_2, however, is isotropic and hence invariant under rotation. Another observation is that the trajectories of the GRF with covariance kernel k_2 seem to be smother. This is due to the fact that $k_2(x_1, x_2)$ is big (close to σ^2) for two points $x_1, x_2 \in D$ that lie close together. So the according random variables Z_{x_1} and Z_{x_2} have a high covariance. However, this is just an intuitive explanation. There has been a lot of research about path regularity as we will see in the next subsection.

2.3 Path regularity

Without further assumptions on the random field, little can be said about the regularity properties of its paths: it is not even guaranteed that they

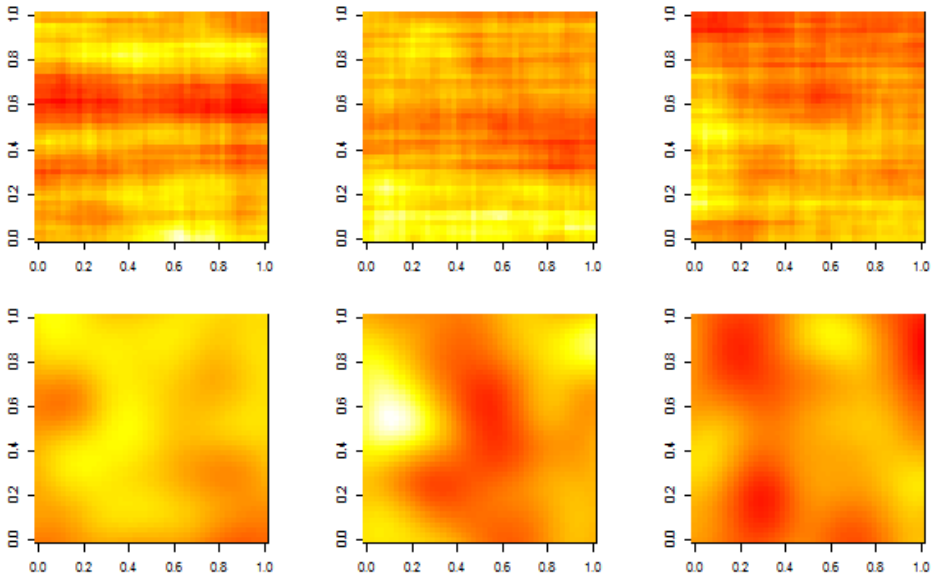


Figure 2: Paths of centered Gaussian random fields over $D = [0, 1]^2$.
Upper row: using an exponential kernel with $\sigma^2 = 1$, $\theta_1 = 1$ and $\theta_2 = 0.25$.
Lower row: using an isotropic Gaussian kernel with $\sigma^2 = 1$ and $\theta = 0.25$.

are (Borel-)measurable functions! In this section, we present some selected state-of-the-art results on the functional properties of paths depending on assumptions on the distribution of Z.

A general overview of the regularity of random field paths is given in [Sch09, Chapter 5] or [AT10, Chapter 1.3]. There we can find criteria that guarantee measurable, L^2, continuous or differentiable paths. We follow [Sch09] and equip the domain D with Lebesgue measure λ.

For a start we introduce the conditions under which a random field is said to be separable (going back to [Doo53]).

Definition 4 (Separable random field)**.** *A random field $(Z_x)_{x \in D}$ over the probability space $(\Omega, \mathcal{F}, P)$ is called separable if there exists a countable and dense subset $S \subseteq D$ and a set $N \subseteq \mathcal{F}$ with $P(N) = 0$ such that for any $G \subseteq D$ open and any $F \subseteq \mathbb{R}$ closed the following sets*

$$A_{F,G} := \{\omega : Z_x(\omega) \in F \; \forall x \in G\}$$
$$A_{F,G \cap S} := \{\omega : Z_x(\omega) \in F \; \forall x \in G \cap S\}$$

differ only on a subset of N.

In the context of this thesis we are mainly interested in random fields with almost surely L^2 paths. The following theorem (see [Sch09, Theorem 5.3.6]) states some sufficient conditions under which the paths are a.s. continuous.

Theorem 1. *Let $(Z_x)_{x\in D}$ be a separable centered Gaussian random field on a compact domain $D \subseteq \mathbb{R}^d$. If for some constants $0 < C < \infty$ and $\delta, \eta > 0$ we have*

$$\mathbb{E}[(Z_x - Z_y)^2] \leq \frac{C}{\left| ln \|x - y\| \right|^{1+\delta}} \tag{1}$$

for all $x, y \in D$ with $\|x - y\| < \eta$, then the paths of $(Z_x)_{x\in D}$ are almost surely continuous and bounded on D.

Remark 1. *Since D is here assumed compact, this will in particular ensure that the paths are almost surely in $L^2(D, \lambda)$.*

Assuming that the random field is centered is no serious restriction. If we have a random field $(Z_x)_{x\in D}$ with non-trivial expectation m then we consider the centered random field $(Z_x - m(x))_{x\in D}$ and add the expectation to its paths.

Condition (1) is met (by far) by the continuous kernels that are most frequently used in practice. We see it illustrated in Figure 3 for the Gaussian kernel (as encountered in Example 1). We choose $C = \delta = 1$ and some appropriate value for η, e.g. $\eta = 0.03$. Assuming that k is the covariance kernel of $(Z_x)_{x\in D}$, we get $\mathbb{E}[(Z_x - Z_y)^2] = k(x, x) + k(y, y) - 2k(x, y)$. We see that (1) is satisfied.

There is another way that leads to almost surely L^2 paths [Sch09, Section 5.4]. For this purpose we introduce the notion of a measurable random field:

Definition 5 (Measurable random field)**.** *Let $(Z_x)_{x\in D}$ be a random field over the probability space $(\Omega, \mathcal{F}, P)$ and a measurable domain $(D, \mathcal{M})$. Let $\mathcal{F} \otimes \mathcal{M}$ be the product σ-algebra of $\mathcal{F}$ and $\mathcal{M}$, and $\overline{\mathcal{F} \otimes \mathcal{M}}$ its completion w.r.t. the measure $P \otimes \lambda$. Then $(Z_x)_{x\in D}$ is called measurable if it is measurable as a map*

$$Z : (\Omega \times D, \overline{\mathcal{F} \otimes \mathcal{M}}) \to (\mathbb{R}, \mathcal{B}) \ . \tag{2}$$

Remark 2. *From Definition 5 we get directly that the paths of a measurable random field are measurable, for a path is the restriction of the map (2) to a fixed $\omega \in \Omega$.*

Then, by the following proposition, we again get a criterion for paths that are almost surely (locally) L^2. Notice, though, that this time we do not know whether the paths are continuous.

Proposition 1. *The sample paths of a measurable random field $(Z_x)_{x\in D}$ are in $L^2_{loc}(D)$ almost surely. If in addition we have*

$$\int_D k(x,x)\, d\lambda(x) < \infty$$

then the paths of $(Z_x)_{x\in D}$ are in $L^2(D)$ a.s.

In [Sch09, Proposition 5.4.3] λ is still the Lebesgue measure. The proof can be extended, though, to any σ-finite measure.

2.4 Simulation of Gaussian random fields

In Chapter 5 we will see some experiments concerning Gaussian random fields. Amongst other things we will need a method to simulate their outcome, i.e. to generate random field paths. This section is dedicated to such simulations and how they are carried out in practice.

Let $(Z_x)_{x\in D}$ be a Gaussian random field indexed by D. Let m and k be respectively its expectation function and covariance kernel. In order to

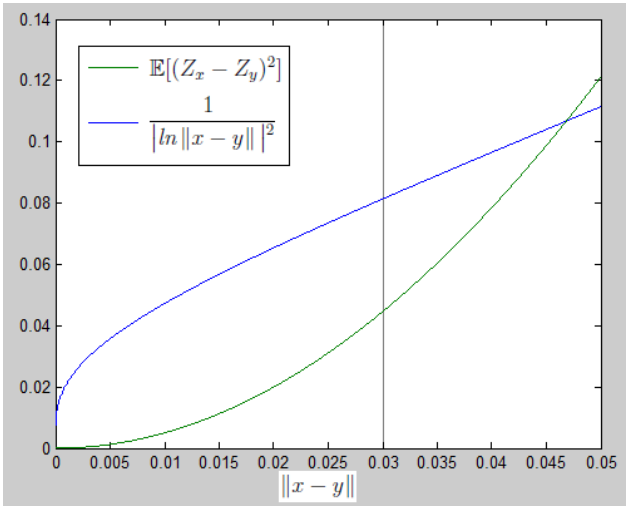


Figure 3: Visualization of condition (1) for a Gaussian kernel

generate a random field path we start by a discretization of the domain. That means we choose a set $X = \{x_1, \dots x_n\} \subseteq D$ of points, e.g. lying on a grid. Then we calculate the expectation vector $M := (m(x_1), \dots, m(x_n))^T$ and the covariance matrix K, an $n \times n$ matrix with entries $K_{ij} = k(x_i, x_j)$ $(1 \leq i, j \leq n)$. From Definition 3 we know that $Z_X := (Z_{x_1}, \dots, Z_{x_n})$ is a Gaussian vector.

$$Z_X \sim \mathcal{N}(M, K)$$

There are numerous procedures to generate random vectors having a multivariate normal distribution. Often used are the Cholesky decomposition or the Mahalanobis transform [GS11]. K has to be a positive definite matrix. We will see in Section 2.4.1 below that this is guaranteed in our case.

If the set of points X has a special structure then there may even exist some advanced techniques that allow simulations with better performance. This is the case if X lies on a regular grid. Then we can do a circular embedding [DN97]. Another method is called the turning band method. It can be applied when the covariance kernel is isotropic (see [Ste99, Chapter 2.2]).

For our purposes we have chosen to apply a Cholesky decomposition[1] which is the standard procedure in this case (see [GG09, Chapter 4.7]).

As symmetric positive definite matrix, K can be decomposed as

$$K = LL^T \ ,$$

where L is a lower triangular matrix with nonnegative diagonal entries. If $N \sim \mathcal{N}(\mathbf{0}, I_n)$ (i.e. N has a standard multivariate normal distribution) then $(LN + M) \sim \mathcal{N}(M, K)$.

2.4.1 Positive definite kernels

An important object, when talking about a random field $(Z_x)_{x \in D}$, is the corresponding covariance kernel k. It is a symmetric real-valued function taking two arguments from D. A significant property of a kernel is positive definiteness. It is closely related to the positive definiteness of a matrix.

Definition 6 (Positive definite kernel)**.** *A kernel $k : D \times D \to \mathbb{R}$ is said to be positive definite if for any $n \in \mathbb{N}$, for any subset $\{x_1, \dots, x_n\}$ of D and*

[1] After the French mathematician André-Louis Cholesky (1875-1918)

for all choices of real values $\alpha_1, \ldots, \alpha_n$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j k(x_i, x_j) \geq 0 \ .$$

The corresponding definition for matrices requires strict positiveness. The loose definition is more convenient and well-established [Cre93, SS02], although the literature is not consistent in this case. We can also find the term positive semi-definite (e.g. [GG09]), or non-negative definite [AT10].

In the case of a covariance kernel we have the following useful result:

Theorem 2 (Loève[2])**.** *$k : D \times D \to \mathbb{R}$ is a covariance kernel of some real-valued random field $(Z_x)_{x \in D}$ if and only if it is symmetric positive definite.*

Proof.
"$\Rightarrow$": Consider the random field $(Z_x)_{x \in D}$ with covariance kernel k. For arbitrary $n \in \mathbb{N}$, $x_1, \ldots, x_n \subseteq D$ and $\alpha \in \mathbb{R}^n$ we have

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \ k(x_i, x_j) = \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \ \mathrm{Cov}(Z_{x_i}, Z_{x_j})$$
$$= \mathrm{Cov}\left(\sum_{i=1}^{n} \alpha_i Z_{x_i}, \sum_{j=1}^{n} \alpha_j Z_{x_j} \right) = \mathrm{Var}\left(\sum_{j=1}^{n} \alpha_j Z_{x_j} \right) \geq 0 \ .$$

"$\Leftarrow$": If we have any positive definite kernel we can construct a Gaussian random field with the according covariance function. The details for this direction of the proof are omitted here. Instead we recommend, e.g., [Sch09, Corollary 5.1.2]. □

[2]Michel Loève 1907-1979, French American mathematician, known amongst other things for the Karhunen–Loève expansion

3 Decompositions in L^2

We come now to a rather different topic. This chapter is dedicated to some classical Hilbert space theory. In particular we discuss orthogonal projections. For now it will not be related to what we did in Chapter 2. We will see the connection only in Chapter 4 where we apply orthogonal projections to the trajectories of a random field.

3.1 Fundamentals

In the sequel of the chapter we are going to define an orthogonal decomposition of a Hilbert space. In order to do this we use some standard results from functional analysis. This section is based on [Tre13]. The proofs are omitted.

Lemma 1. *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space, $M \subseteq H$. Then $M^\perp$ is a closed linear subspace of H.*

We recall that $M^\perp$ is the orthogonal complement of M in H, i.e. the set of all elements of H that are orthogonal to M (which means orthogonal to all elements of M). Note that in Lemma 1 above, M needs to be just a set, not a subspace, and does not have to be closed.

Definition 7 (Direct orthogonal sum)**.** *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and $M, N \subseteq H$ two linear subspaces. Then H is called the direct orthogonal sum of M and N, written as $H = M \oplus N$ if $M \perp N$ and*

$$\forall x \in H \; \exists!\, x_1 \in M, x_2 \in N : x = x_1 + x_2 \,.$$

There exists literature that makes the distinction between a direct sum and a direct orthogonal sum and the notation is not consistent. Notice that we write $\oplus$ and require orthogonality. If H can be written as a direct orthogonal sum of two subspaces M and N then we have directly $M = N^\perp$ and $N = M^\perp$ and hence - using Lemma 1 - both subspaces are closed (in H).

The definition is equivalent if we exchange $\exists!$ by $\exists$. The uniqueness of the decomposition is guaranteed by the orthogonality of M and N.

Definition 8 (Orthogonal projection)**.** *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space and $M \subseteq H$ a closed linear subspace. Then the linear operator $\pi : H \to H$, $\pi x = x_1$ where $x = x_1 + x_2 \in M \oplus M^\perp$ is called orthogonal projection of H onto M.*

Notice that $M^\perp$ is closed by Lemma 1. This and the fact that M is closed itself ensure that H can always be written as a direct orthogonal sum $M \oplus M^\perp$.

The co-domain of the projection π is often called range and denoted by $\mathcal{R}(\pi)$.

Lemma 2. *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space, M_0, M_1 closed linear subspaces of H and π_0, π_1 the orthogonal projections onto M_0, M_1 respectively. Then the following are equivalent:*

- $\pi_0\pi_1 = 0$ *(i.e. $\forall x \in H : \pi_0\pi_1 x = 0$)*
- $M_0 \perp M_1$
- $\pi_0 + \pi_1$ *is an orthogonal projection*

In this case we have $\mathcal{R}(\pi_0 + \pi_1) = M_0 \oplus M_1$.

Lemma 3. *Let $(H, \langle\cdot,\cdot\rangle)$ be a Hilbert space, M_0, M_1 closed linear subspaces of H and π_0, π_1 the orthogonal projections onto M_0, M_1 respectively. Then the following are equivalent:*

- $\pi_0\pi_1 = \pi_0$ *(i.e. $\forall x \in H : \pi_0\pi_1 x = \pi_0 x$)*
- $M_0 \subseteq M_1$
- $\forall x \in H : \|\pi_0 x\| \leq \|\pi_1 x\|$ *(we write $\pi_0 \leq \pi_1$)*
- $\forall x \in H : \langle \pi_0 x, x\rangle \leq \langle \pi_1 x, x\rangle$
- $\pi_1 - \pi_0$ *is an orthogonal projection*

In this case we have $\mathcal{R}(\pi_1) = \mathcal{R}(\pi_0) \oplus \mathcal{R}(\pi_1 - \pi_0)$.

3.2 Retrieving additive and ortho-additive functions

We consider a set $D \subseteq \mathbb{R}^d$ which has the form of a d-dimensional hypercuboid:

$$D = D_1 \times \ldots \times D_d \,,$$

D_i being bounded closed real intervals ($1 \leq i \leq d$) of positive length. We equip D with a product measure

$$\mu = \mu_1 \otimes \ldots \otimes \mu_d \,.$$

For the sake of simplicity we restrict μ_i, and hence μ, to be probability measures. In this section we consider the Hilbert space $L^2(D, \mu)$ (which in the following is simply called L^2) and subspaces thereof.

Let us start with some remarks on notation. We recall that the elements of L^2 are equivalence classes of (real-valued) functions in $\mathcal{L}^2$ with respect to μ-almost everywhere equality. If in the further text the term function is used then it may refer to an element of $\mathcal{L}^2$ (which is indeed a function according to the common definition) or likewise to the corresponding equivalence class in L^2.

For such functions capital letters are used, whereas lowercase letters refer to constants or functions in one variable. For instance the notation

$$C(\boldsymbol{x}) = c \quad \text{or} \quad C = c \ \mu\text{-almost everywhere}$$

refers to a function $C : D \to \mathbb{R}$ with the constant value c on the whole domain (or to the equivalence class of functions that are equal to c μ-almost everywhere). We will sometimes use the symbol $\mathbf{1}$ for the constant one-function, i.e. we could also write $C(\boldsymbol{x}) = c \cdot \mathbf{1}(\boldsymbol{x})$. We will use it quite liberally, though, e.g. also for functions which are defined only on a subspace of D.

We will now start with a simple decomposition of L^2 into two subspaces. It is at the same time an introduction for the more complicated decomposition that follows.

The subspace of L^2 that contains all elements which are constant on the whole domain is important enough to us to have its own name. Let us call it $L^2_{\mathcal{C}}$.

$$L^2_{\mathcal{C}} := \left\{ F \in L^2 \text{ s.t. } \exists c \in \mathbb{R} \text{ with } F = c \text{ almost everywhere} \right\}$$

Lemma 4. *$L^2_{\mathcal{C}}$ is a closed linear subspace of L^2.*

Proof. When adding two elements of $L^2_{\mathcal{C}}$ or multiplying one with a scalar we get again an element of $L^2_{\mathcal{C}}$. Thus $L^2_{\mathcal{C}}$ is clearly a linear subspace of L^2.

There exists a natural isometric isomorphism between $L^2_{\mathcal{C}}$ and $\mathbb{R}$. Since such an isomorphism preserves completeness we conclude that $L^2_{\mathcal{C}}$ is complete. A subspace of the complete space L^2 is closed if and only if it is complete. □

Let us define the space $L^2_{\mathcal{Z}}$ of zero-mean functions[3]:

$$L^2_{\mathcal{Z}} := \left\{ Z \in L^2 \text{ s.t. } \int_D Z \, d\mu = 0 \right\} .$$

And let us state the following result:

Lemma 5. *L^2 is the direct orthogonal sum of $L^2_{\mathcal{Z}}$ and $L^2_{\mathcal{C}}$.*

Proof. First we get that $L^2 = L^2_{\mathcal{C}} \oplus (L^2_{\mathcal{C}})^\perp$, for $L^2_{\mathcal{C}}$ is a closed linear subspace of L^2. Then we can show that $L^2_{\mathcal{Z}} = (L^2_{\mathcal{C}})^\perp$.

$$\begin{aligned}
Z \in (L^2_{\mathcal{C}})^\perp &\iff \langle C, Z \rangle = 0 \; \forall C \in L^2_{\mathcal{C}} \\
&\iff \int_D CZ \, d\mu = 0 \; \forall C \in L^2_{\mathcal{C}} \\
&\iff c \cdot \int_D Z \, d\mu = 0 \; \forall c \in \mathbb{R} \\
&\iff \int_D Z \, d\mu = 0 \\
&\iff Z \in L^2_{\mathcal{Z}}
\end{aligned}$$

□

We can define orthogonal projections onto $L^2_{\mathcal{C}}$ and $L^2_{\mathcal{Z}}$, which we call respectively $\pi_{\mathcal{C}}$ and $\pi_{\mathcal{Z}}$.

$$F \overset{\pi_{\mathcal{C}}}{\longmapsto} F_{\mathcal{C}} := \int_D F \, d\mu \cdot \mathbf{1} \quad \text{and} \quad F \overset{\pi_{\mathcal{Z}}}{\longmapsto} F_{\mathcal{Z}} := F - F_{\mathcal{C}} \; .$$

What we have achieved is a first orthogonal decomposition of L^2 with its according orthogonal projections.

[3]The letter Z (or $\mathcal{Z}$) that we will use in the context of zero-mean functions has nothing to do with a random field $(Z_x)_{x \in D}$

We now want to go a step further and decompose L^2 into subspaces L_i^2 containing elements that depend only on the ith input variable ($1 \leq i \leq d$). By

$$F_i(\boldsymbol{x}) = f_i(x_i) \text{ with } \boldsymbol{x} = (x_1, \ldots, x_i, \ldots, x_d)$$

we refer to functions $F_i : D \to \mathbb{R}$ which depend only on the ith variable, i.e. the functions f_i are univariate on D_i. In its strict sense the subscript i for the functions is not mandatory. But we write it for the sake of clarity. For every $i = 1, \ldots, d$ we subsume all equivalence classes of functions which possess such a univariate nature in the i-th variable and are moreover centered under the term $L_i^2(D, \mu)$ or L_i^2, i.e.

$$L_i^2 := \left\{ \begin{array}{r} F_i \in L^2(D, \mu) \text{ s.t. } \exists f_i \in L^2(D_i, \mu_i) \text{ with } F_i(\boldsymbol{x}) = f_i(x_i) \\ \text{and } \int_{D_i} f_i \, d\mu_i = 0 \end{array} \right\}. \tag{3}$$

Lemma 6. *For $i \in \{1, \ldots, d\}$ the spaces L_i^2 are pairwise orthogonal closed linear subspaces of L^2.*

Proof. First we introduce some new notation. For the following calculations we will use the subscript $-i$ to consider the space D without the ith dimension, i.e. $D_{-i} = D_1 \times \ldots \times D_{i-1} \times D_{i+1} \times \ldots \times D_d$. In the analogous way $\mu_{-i} = \mu_1 \otimes \ldots \otimes \mu_{i-1} \otimes \mu_{i+1} \otimes \ldots \otimes \mu_d$ and $\mathbf{1}_{-i}$ is the constant one-function defined on D_{-i}.

The spaces L_i^2 are clearly linear subspaces of L^2. We show that they are closed and pairwise orthogonal.

Straight from the definition (3) we can assign to any $F_i \in L_i^2(D, \mu)$ the element $f_i \in L^2(D_i, \mu_i)$ (or the other way round we can construct F_i from any f_i). The obtained bijection is an isometric[4] isomorphism preserving completeness. Hence L_i^2 inherits the completeness from $L^2(D_i, \mu_i)$ and is therefore closed in L^2.

There remains to be shown that the subspaces are orthogonal. We prove it by calculation. Consider arbitrary $F_i \in L_i^2$ and $F_j \in L_j^2$ ($i \neq j$).

[4] $\|F_i\| = \langle F_i, F_i \rangle = \int_D F_i F_i \, d\mu = \int_{D_i} f_i f_i \, d\mu_i \cdot \int_{D_{-i}} d\mu_{-i} = \langle f_i, f_i \rangle = \|f_i\|$.

$$\begin{aligned}
\langle F_i, F_j \rangle &= \int_D F_i \cdot F_j \, d\mu \\
&\overset{i \neq j}{=} \int_{D_i} \int_{D_j} \int_{D_{-i-j}} f_i(x_i) \cdot \mathbf{1}_{-i} \cdot f_j(x_j) \cdot \mathbf{1}_{-j} \, d\mu_{-i-j} \, d\mu_j \, d\mu_i \\
&= \int_{D_i} f_i(x_i) \cdot \int_{D_j} f_j(x_j) \cdot \int_{D_{-i-j}} \mathbf{1}_{-i-j} \, d\mu_{-i-j} \, d\mu_j \, d\mu_i \\
&= \int_{D_i} f_i(x_i) \, d\mu_i \cdot \int_{D_j} f_j(x_j) \, d\mu_j = 0
\end{aligned}$$

We conclude that L_i^2 $(1 \leq i \leq d)$ are pairwise orthogonal. □

A consequence of Lemma 6 is that there exist orthogonal projections π_i onto L_i^2. We will soon see how they look like. Since $L_i^2 \subseteq L_{\mathcal{Z}}^2$ we will first consider orthogonal projections $\tilde{\pi}_i$ from $L_{\mathcal{Z}}^2$ onto L_i^2.

Lemma 7. *For $i \in \{1, \ldots, d\}$ the map $\tilde{\pi}_i$, defined as follows, is the orthogonal projection onto L_i^2*

$$\begin{aligned}
\tilde{\pi}_i : L_{\mathcal{Z}}^2 &\to L_{\mathcal{Z}}^2 \\
F_{\mathcal{Z}} &\mapsto \int_{D_{-i}} F_{\mathcal{Z}} \, d\mu_{-i} \cdot \mathbf{1}_{-i} \, .
\end{aligned}$$

Proof. We will show (according to Definition 8) that any $F_{\mathcal{Z}} \in L_{\mathcal{Z}}^2$ can be decomposed into two elements $\tilde{\pi}_i F_{\mathcal{Z}} \in L_i^2$ and $(F_{\mathcal{Z}} - \tilde{\pi}_i F_{\mathcal{Z}}) \in (L_i^2)^{\perp}$.

By construction $\tilde{\pi}_i F_{\mathcal{Z}}$ is the product of a univariate function in the ith variable and the constant one-function on D_{-i}. We can also check that the result has still zero-mean.

$$\int_D \tilde{\pi}_i F_{\mathcal{Z}} \, d\mu = \int_D \int_{D_{-i}} F_{\mathcal{Z}} \, d\mu_{-i} \, \mathbf{1}_{-i} \, d\mu = \int_{D_{-i}} \underbrace{\int_D F_{\mathcal{Z}} \, d\mu}_{0} \, \mathbf{1}_{-i} \, d\mu_{-i} = 0$$

So for an arbitrary $F_{\mathcal{Z}}$ from $L_{\mathcal{Z}}^2$ we have $\tilde{\pi}_i F_{\mathcal{Z}} \in L_i^2$. Moreover we can assure ourselves that $(F_{\mathcal{Z}} - \tilde{\pi}_i F_{\mathcal{Z}}) \perp L_i^2$, i.e. $\langle F_{\mathcal{Z}} - \tilde{\pi}_i F_{\mathcal{Z}}, F \rangle = 0 \; \forall F \in L_i^2$. In order to convince ourselves we recall that F is constant on D_{-i}.

$$
\begin{aligned}
\langle F_{\mathcal{Z}} - \tilde{\pi}_i F_{\mathcal{Z}}, F\rangle &= \int_D \left(F_{\mathcal{Z}} - \int_{D_{-i}} F_{\mathcal{Z}}\, d\mu_{-i}\, \mathbf{1}_{-i} \right) \cdot F\, d\mu \\
&= \int_{D_i} \int_{D_{-i}} \left(F_{\mathcal{Z}} - \int_{D_{-i}} F_{\mathcal{Z}}\, d\mu_{-i}\, \mathbf{1}_{-i} \right) \cdot F\, d\mu_{-i}\, d\mu_i \\
&= \int_{D_i} \int_{D_{-i}} \left(F_{\mathcal{Z}} - \int_{D_{-i}} F_{\mathcal{Z}}\, d\mu_{-i}\, \mathbf{1}_{-i} \right) d\mu_{-i} \cdot F\, d\mu_i \\
&= \int_{D_i} \left(\int_{D_{-i}} F_{\mathcal{Z}}\, d\mu_{-i} - \int_{D_{-i}} \underbrace{\int_{D_{-i}} F_{\mathcal{Z}}\, d\mu_{-i}}_{\text{constant on } D_{-i}}\, \mathbf{1}_{-i}\, d\mu_{-i} \right) \cdot F\, d\mu_i \\
&= \int_{D_i} \left(\int_{D_{-i}} F_{\mathcal{Z}}\, d\mu_{-i} \cdot \underbrace{\left(1 - \int_{D_{-i}} \mathbf{1}_{-i}\, d\mu_{-i} \right)}_{0} \right) \cdot F\, d\mu_i = 0
\end{aligned}
$$

□

But as said before there exists also a projection π_i which is defined on the whole space L^2. From Lemma 3 (and since $L_i^2 \subseteq L_{\mathcal{Z}}^2$) we know that such a projection satisfies $\pi_i = \pi_i \pi_{\mathcal{Z}}$. But in this case this has to be equal to $\tilde{\pi}_i \pi_{\mathcal{Z}}$.

Corollary 1. *For $i \in \{1, \ldots, d\}$ the map π_i, defined as follows, is the orthogonal projection onto L_i^2*

$$
\begin{aligned}
\pi_i : L^2 &\to L^2 \\
F &\mapsto \int_{D_{-i}} \pi_{\mathcal{Z}} F\, d\mu_{-i} \cdot \mathbf{1}_{-i}\ .
\end{aligned}
$$

Proof. π_i is nothing but $\tilde{\pi}_i \pi_{\mathcal{Z}}$. □

Note that the spaces L_i^2 are also orthogonal to $L_{\mathcal{C}}^2$ since $L_i^2 \subseteq L_{\mathcal{Z}}^2 = L_{\mathcal{C}}^{2\perp}$.

So we have defined a lot of closed linear subspaces of L^2. But what we are really interested in is the space which is composed of all of them.

Definition 9 (Space $\mathcal{A}$ of additive functions)**.** *We call*

$$
\mathcal{A} = L_{\mathcal{C}}^2 \oplus \left(\bigoplus_{i=1}^{d} L_i^2 \right)
$$

the space of additive functions.

In order to be correct we recall that the elements of $\mathcal{A}$ are equivalence classes with respect to μ-almost everywhere equality. Anyway we will use the term additive function for these equivalence classes as well as for the according functions of $\mathcal{L}^2$.

In the literature (e.g. [HT90]) we can find that additive functions on $D \subseteq \mathbb{R}^d$ are simply defined as a sum of a constant function and d univariate functions in x_i $(1 \leq i \leq d)$. In other publications (e.g. [DGRC13]) the univariate functions are assumed to be centered. Apart from the difference between functions and equivalence classes, the definitions are equivalent. In our definition we also impose that the elements in the spaces L_i^2 have to be centered. This guarantees orthogonality between the different subspaces of $\mathcal{A}$. It does not affect the whole space, though.

From Lemma 2 we get the according orthogonal projection onto $\mathcal{A}$:

$$\pi_{\mathcal{A}} = \pi_{\mathcal{C}} + \sum_{i=1}^{d} \pi_i \ .$$

In the following text not just $\mathcal{A}$ will be important to us but also its orthogonal complement $\mathcal{A}^{\perp}$.

Definition 10 (Space $\mathcal{O}$ of ortho-additive functions)**.** *Let us define by*

$$\mathcal{O} := \left\{ F \in L^2 : \forall i \in \{1, \ldots, d\} \int_{D_{-i}} F \, d\mu_{-i} = 0 \right\}$$
$$\text{and} \int_D F \, d\mu = 0 \Bigg\}$$

the space of "ortho-additive" functions.

Proposition 2. *$\mathcal{O}$ is the orthogonal complement of $\mathcal{A}$ in L^2.*

Proof. We have to show $\mathcal{O} = \mathcal{A}^{\perp}$. "$\subseteq$" and "$\supseteq$" can be shown simultaneously.

$$
\begin{aligned}
& F \in \mathcal{A}^{\perp} \\
\iff & F \in L_{\mathcal{C}}^{2\perp} \cap \left(\bigcap_{i=1}^{d} L_i^{2\perp} \right) \\
\iff & \pi_{\mathcal{C}} F = 0 \text{ and } \pi_i F = 0 \; \forall i \in \{1, \dots, d\} \\
\iff & \int_D F \, d\mu = 0 \text{ and } \int_{D_{-i}} F - \pi_{\mathcal{C}} F \, d\mu_{-i} = 0 \; \forall i \in \{1, \dots, d\} \\
\iff & \int_D F \, d\mu = 0 \text{ and } \int_{D_{-i}} F \, d\mu_{-i} = 0 \; \forall i \in \{1, \dots, d\} \\
\iff & F \in \mathcal{O}
\end{aligned}
$$

□

Any orthogonal complement in a Hilbert space is a closed linear subspace. The according orthogonal projection is

$$\pi_{\mathcal{O}} = \mathrm{id} - \pi_{\mathcal{A}} \, .$$

Example 2. *We consider the function $F(x_1, x_2) = x_1 x_2 + x_1$ defined on the Domain $[0, 1]^2$. We can calculate the projections of this function.*

$$
\begin{aligned}
(\pi_{\mathcal{C}} F)(x_1, x_2) &= \frac{3}{4} \\
(\pi_1 F)(x_1, x_2) &= \frac{3}{2} x_1 - \frac{3}{4} \\
(\pi_2 F)(x_1, x_2) &= \frac{1}{2} x_2 - \frac{1}{4} \\
(\pi_{\mathcal{O}} F)(x_1, x_2) &= x_1 x_2 - \frac{1}{2} x_1 - \frac{1}{2} x_2 + \frac{1}{4}
\end{aligned}
$$

4 Projecting Gaussian random fields

We are ready now to make the link between the previous two chapters. We will apply the projections defined in Chapter 3 to random field paths.

For the whole chapter let $D = D_1 \times \ldots \times D_d \subseteq \mathbb{R}^d$ be a d-dimensional hyper-cuboid with measure μ as in Section 3.2. Consider a measurable Gaussian random field $(Z_x)_{x \in D}$ indexed by D. Again we use the abbreviation L^2 for the Hilbert space $L^2(D, \mu)$.

4.1 Projecting Gaussian random field paths

From Proposition 1 we know that $(Z_x)_{x \in D}$ has a.s. L^2 paths (for D is compact and μ finite). This means that we can project these paths by applying an orthogonal projection $\pi : L^2 \to L^2$.

In Figure 4 in the first row we see some examples of trajectories of a centered Gaussian random field. They are defined over the unit square and are generated using the Gaussian covariance kernel

$$k(x, y) = \sigma^2 \cdot e^{-\left(\frac{\|x-y\|}{\theta}\right)^2} . \tag{4}$$

The values are represented by colors. Positive values are yellow, values around zero are orange and negative values are red. The values follow a Gaussian distribution for any coordinate of the domain. We can see that two points of the field that are close to each other tend to have similar values. This behavior is driven by the covariance kernel.

What does it mean if we now project the trajectories? The result can be seen in the second to fourth row. In the second row are projections into the space $\mathcal{A}$ of additive functions. We can clearly see the additive structure of $C + Z_1(x_1) + Z_2(x_2)$ in terms of horizontal and vertical stripes.

The third row is even more interesting. At first glance one might think that these trajectories have been generated by the same kernel as in the first row. But in fact these are the projections of such paths into $\mathcal{O}$. This means that they have certain properties that we can check. First of all the projected trajectories have to be centered, i.e. the images have to be orange on average. This is not clearly visible. It can be checked numerically, though. The values in the images are also column- and row-wise centered. That means that in any row and in any column there is as much yellow as there is red. Knowing

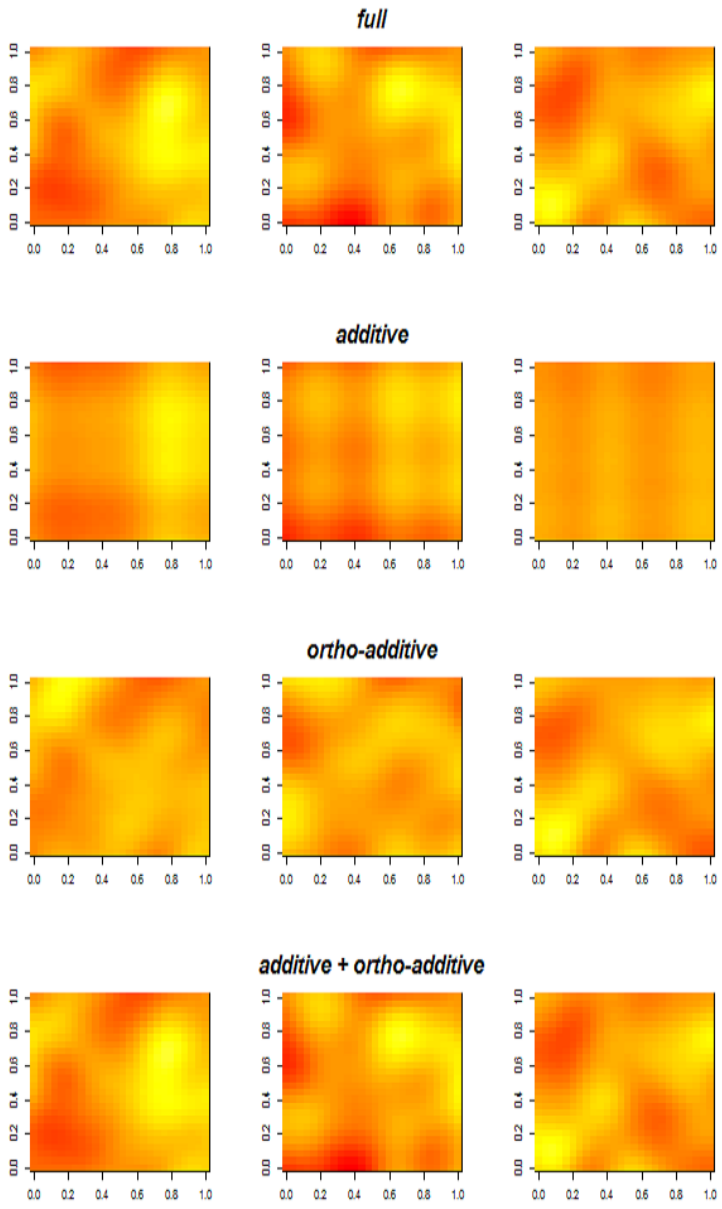


Figure 4: Three realizations of a GRF (first row), their projections to $\mathcal{A}$ (second row), to $\mathcal{O}$ (third row) and the sum of the two latter (last row)

this we can clearly see a difference between the corresponding images in the first and in the third row.

The last row shows trajectories which are the sum of the additive and the ortho-additive projection. These images are indeed equal to the images in the first row. This is clear by construction. Anyway, we emphasize it because in Section 4.2 we will see some other kind of additive and ortho-additive projections which do not sum up to the identity.

4.2 Projecting covariance kernels

We present here a general result which has an interesting application in our context [TV07, Corollary 3.7].

Property 1. *Let X and Y be real separable Banach spaces, μ be a Gaussian measure on $\mathcal{B}(X)$, $\eta : X \to Y$ be a bounded linear operator and $\nu = \mu \circ \eta^{-1}$. Then ν is a Gaussian measure on $\mathcal{B}(Y)$ with mean $m_\nu = \eta m_\mu \in Y$ and with covariance operator $C_\nu = \eta C_\mu \eta^* : Y^* \to Y$, where η^* is the adjoint operator of η and Y^* is the dual space of Y.*

Let X be the space of continuous functions on D equipped with the supremum norm (and the hereby induced topology). Consider an orthogonal projection $\pi : X \to X$ (hence we consider $Y = X$). It is in particular a bounded linear operator. X is a separable[5] Banach space. We already know that the random field πZ is again Gaussian.

From Property 1 follows that for the projection π there exists an operator which can be applied to covariance kernels. We call it $\pi \otimes \pi$, where $\otimes$ denotes the tensor product between two operators.

The relation between π and $\pi \otimes \pi$ can be express as follows:

$$\mathrm{Cov}(\pi Z, \pi Z)(x, y) = \big((\pi \otimes \pi)k\big)(x, y) \ .$$

So the covariance of a projected random field is expressed by a projection of the covariance kernel of the original random field. Notice that $(\pi \otimes \pi)\ k$ is positive definite by Theorem 2.

For two (different) projections π_1, π_2 respectively, we can derive a more general projection $\pi_1 \otimes \pi_2$.

$$\mathrm{Cov}(\pi_1 Z, \pi_2 Z)(x, y) = \big((\pi_1 \otimes \pi_2)k\big)(x, y)$$

[5]We point out the distinction from a separable random field (Definition 4)

However the resulting kernel $(\pi_1 \otimes \pi_2)k$ defines a cross-covariance between two projections of a GRF and is no more positive definite, in general.

Let us go a step further and consider a family Π of projections such that

$$\mathrm{id}_{L^2(D,\mu)} = \sum_{\pi_i \in \Pi} \pi_i$$

then we have

$$\mathrm{id}_{L^2(D\times D,\mu\otimes\mu)} = \sum_{\pi_i \in \Pi} \pi_i \otimes \sum_{\pi_j \in \Pi} \pi_j = \sum_{\pi_i \in \Pi} \sum_{\pi_j \in \Pi} \pi_i \otimes \pi_j \ .$$

Notice that in order to calculate the projected kernels we define the following operators where the superscripts l and r are used to indicate that we apply the operator to the left (i.e. to the first variable) or to the right (second variable), respectively.

$$\begin{aligned} \pi^l : \mathbb{R}^{D\times D} &\to \mathbb{R}^{D\times D} \\ k &\mapsto (\pi^l\ k)(x,y) = \big(\pi\ k(.,y)\big)(x) \end{aligned} \tag{5}$$

$$\begin{aligned} \pi^r : \mathbb{R}^{D\times D} &\to \mathbb{R}^{D\times D} \\ k &\mapsto (\pi^r\ k)(x,y) = \big(\pi\ k(x,.)\big)(y) \end{aligned} \tag{6}$$

The cross-covariance of $(\pi_1 Z)_x$ and $(\pi_2 Z)_y$ is $\big(\pi_1^l \pi_2^r k\big)(x,y)$. For the sake of legibility we shorten the terms for the projections conveniently: We will use π_{ij} for $\pi_i^l \pi_j^r$ and π_i for π_{ii}.

4.2.1 Product kernels

Consider the projections $\pi_{\mathcal{A}}$ and $\pi_{\mathcal{O}}$ seen in Section 3.2. Using notation (5) and (6), we know how to decompose a kernel into an additive and an ortho-additive part[6]. However, it is not always practicable to find an explicit expression in a closed form. But there exist kernels for which the calculations are feasible. Notice that in the following calculations we will use a shortened notation concerning integrals by writing, e.g., $\int dx$ instead of $\int d\mu(x)$.

One of the main research findings of this master thesis is subsumed under the following proposition.

[6]Following our shorthand notation introduced above, we write $\pi_{\mathcal{A}} k$ for $\pi_{\mathcal{A}}^l \pi_{\mathcal{A}}^r k$ and accordingly for $\pi_{\mathcal{O}} k$

Proposition 3 (Decomposition of a product kernel)**.** *Let k be a symmetric product kernel on the non-empty domain $D = [a_1, b_1] \times \ldots \times [a_d, b_d]$, i.e. a symmetric kernel which fulfills*

$$k(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} k_i(x_i, y_i) \tag{7}$$

for some set of symmetric positive definite kernels k_i defined on $[a_i, b_i]^2$

We can calculate $\pi_{\mathcal{A}} k$ and $\pi_{\mathcal{O}} k$ explicitly in terms of

$$\begin{aligned} E_i(x_i, a_i, b_i) &:= \int_{a_i}^{b_i} k_i(x_i, y_i)\, dy_i \\ \text{and} \quad \mathcal{E}_i(a_i, b_i) &:= \int_{a_i}^{b_i} E_i(x_i, a_i, b_i)\, dx_i \ . \end{aligned} \tag{8}$$

The integrals in (8) can always be calculated at least numerically.

Proof. We can calculate $\pi_{\mathcal{A}} k$ and $\pi_{\mathcal{O}} k$ straight forward by applying the projections from Section 3.2. However we have to do almost the whole decomposition of k into $(d+2)^2$ terms. The main clue to the result is that the projected kernels depend solely on the following auxiliary quantities which we get from (8):

- $E_i(x_i) := E_i(x_i, a_i, b_i) = \int_{a_i}^{b_i} k_i(x_i, y_i)\, dy_i$
- $E(\boldsymbol{x}) := E(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{d} E_i(x_i, a_i, b_i)$
- $\mathcal{E}_i := \mathcal{E}_i(a_i, b_i) = \int_{a_i}^{b_i} E_i(x_i, a_i, b_i)\, dx_i$
- $\mathcal{E} := \mathcal{E}(\boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{d} \mathcal{E}_i(a_i, b_i)$

Notice that we get simultaneously $E_i(y_i)$ and $E(\boldsymbol{y})$ due to the symmetry of k. A helpful simplification to get neater equations is the assumption that our kernel is strictly positive in the following way:

$$k_i(x_i, y_i) > 0 \ \forall\ x_i, y_i \in [a_i, b_i] \quad (1 \leq i \leq d) \ . \tag{9}$$

This assumption along with the fact that the domain is non-empty guarantees us that all our auxiliary quantities are likewise strictly positive. We

can hence write them into the denominator of a fraction. The proof could be done without this assumption, though.

The computation is a matter of endurance. We give here the strategy. The full calculations can be found in the appendix.

We will first apply the right-hand-side projections $\pi^r_{\mathcal{C}}k$ and $\pi^r_j k$ for all $j \in \{1, \ldots, d\}$. This will lead to

$$\pi^r_{\mathcal{A}}k = \pi^r_{\mathcal{C}}k + \sum_{i=1}^{d} \pi^r_i k \ .$$

Then with the left-hand-side projections we get $\pi_{\mathcal{C}\mathcal{A}}k$, $\pi_{i\mathcal{A}}k$ $(1 \leq i \leq d)$ and finally

$$\pi_{\mathcal{A}}k = \pi_{\mathcal{C}\mathcal{A}}k + \sum_{i=1}^{d} \pi_{i\mathcal{A}}k \ .$$

In a second step we calculate the ortho-additive part of a kernel. From the previous results we can derive

$$\pi^r_{\mathcal{O}}k = k - \pi^r_{\mathcal{A}}k$$

and with the intermediate results $\pi_{\mathcal{C}\mathcal{O}}k$ and $\pi_{i\mathcal{O}}k$ $(1 \leq i \leq d)$ we get

$$\pi_{\mathcal{O}}k = \pi^r_{\mathcal{O}}k - \pi_{\mathcal{C}\mathcal{O}}k - \sum_{i=1}^{d} \pi_{i\mathcal{O}}k \ .$$

When we carry out all these calculations we obtain the two equations which complete the proof:

$$\pi_{\mathcal{A}}k = \frac{A(\boldsymbol{x})A(\boldsymbol{y})}{\mathcal{E}} + \mathcal{E} \cdot \sum_{i=1}^{d} \left(\frac{k_i(x_i, y_i)}{\mathcal{E}_i} - \frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2} \right)$$

$$\pi_{\mathcal{O}}k = k(\boldsymbol{x}, \boldsymbol{y}) - E(\boldsymbol{x}) \cdot \left(1 - d + \sum_{j=1}^{d} \frac{k_j(x_j, y_j)}{E_j(x_j)} \right) - E(\boldsymbol{y}) \cdot \left(1 - d + \sum_{i=1}^{d} \frac{k_i(x_i, y_i)}{E_i(y_i)} \right)$$
$$+ \frac{A(\boldsymbol{x})A(\boldsymbol{y})}{\mathcal{E}} + \mathcal{E} \cdot \sum_{i=1}^{d} \left(\frac{k_i(x_i, y_i)}{\mathcal{E}_i} - \frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2} \right)$$

$$\text{with} \quad A(\boldsymbol{x}) = \mathcal{E} \left(1 - d + \sum_{i=1}^{d} \frac{E_i(x_i)}{\mathcal{E}_i} \right) .$$

□

We finish this section with an important remark. We refer to the very last statement of Section 4.1. There we noticed that the sum of the additive projection of a random field path and its ortho-additive projection is equal to the original (unprojected) path.

However, if we project the covariance kernel k of a random field then we get

$$\pi k = \pi_{\mathcal{A}} k + \pi_{\mathcal{A}\mathcal{O}} k + \pi_{\mathcal{O}\mathcal{A}} k + \pi_{\mathcal{O}} k \ .$$

The sum of the additive and the ortho-additive part is no more equal to k, for there are the cross-covariance terms $\pi_{\mathcal{A}\mathcal{O}} k$ and $\pi_{\mathcal{O}\mathcal{A}} k$.

4.2.2 The Gaussian kernel

For a concrete example we consider now the Gaussian kernel on the domain $D = [a_1, b_1] \times \ldots \times [a_d, b_d]$

$$k(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} \sigma_i^2 \cdot e^{-\left(\frac{x_i - y_i}{\theta_i}\right)^2} \ .$$

It complies with the necessary condition (7) and also (9). We can calculate the first auxiliary integral numerically using the cumulative distribution of the standard normal distribution.

$$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \, e^{-\frac{t^2}{2}} \, dt$$

or in our case (to simplify the results) the error function

$$\operatorname{erf}(x) = 2 \cdot \Phi(\sqrt{2}x) - 1 \ .$$

We get

$$E_i(x_i) = \int_{a_i}^{b_i} \sigma_i^2 \cdot e^{-\left(\frac{x_i - y_i}{\theta_i}\right)^2} \, dy_i = \frac{\sigma_i^2 \theta_i \sqrt{\pi}}{2} \cdot \left(\operatorname{erf}\Big(\frac{b_i - x_i}{\theta_i}\Big) + \operatorname{erf}\Big(\frac{x_i - a_i}{\theta_i}\Big) \right) .$$

In order to calculate $\mathcal{E}_i$ we have to integrate once more. To do so we can benefit of the fact that the integral of the error function is known (see e.g. in [Wei10, p. 934]):

$$\int \operatorname{erf}(x) \, dx = x \cdot \operatorname{erf}(x) + \frac{1}{\sqrt{\pi}} \cdot e^{-x^2} \ .$$

We can calculate the second integral

$$\begin{aligned}\mathcal{E}_i = \frac{\sigma_i^2\theta_i\sqrt{\pi}}{2} \cdot \Bigg(& -(b_i - b_i) \cdot \operatorname{erf}\Big(\frac{b_i - b_i}{\theta_i}\Big) + (a_i - b_i) \cdot \operatorname{erf}\Big(\frac{a_i - b_i}{\theta_i}\Big) \\ & + (b_i - a_i) \cdot \operatorname{erf}\Big(\frac{b_i - a_i}{\theta_i}\Big) - (a_i - a_i) \cdot \operatorname{erf}\Big(\frac{a_i - a_i}{\theta_i}\Big) \Bigg) \\ + \frac{\sigma_i^2\theta_i^2}{2} \cdot \Bigg(& -e^{-\frac{(b_i-b_i)^2}{\theta_i^2}} + e^{-\frac{(a_i-b_i)^2}{\theta_i^2}} + e^{-\frac{(b_i-a_i)^2}{\theta_i^2}} - e^{-\frac{(a_i-a_i)^2}{\theta_i^2}} \Bigg)\end{aligned}$$

and after some simplifications we get

$$\mathcal{E}_i = \sigma_i^2\theta_i\sqrt{\pi} \cdot (b_i - a_i) \cdot \operatorname{erf}\Big(\frac{b_i - a_i}{\theta_i}\Big) + \sigma_i^2\theta_i^2 \cdot \left(e^{-\left(\frac{b_i-a_i}{\theta_i}\right)^2} - 1 \right).$$

4.3 Schematic representation of a covariance kernel

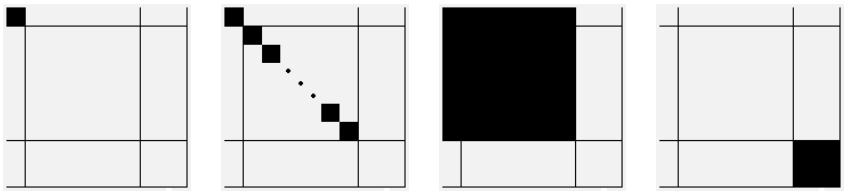
In the following we will use projected kernels as seen in Section 4.2. The family of projections $\Pi = \{\pi_{\mathcal{C}}, \pi_1, \ldots, \pi_d, \pi_{\mathcal{O}}\}$ decomposes a kernel k into $(d+2)^2$ parts:

$$k(\boldsymbol{x}, \boldsymbol{y}) = \sum_{\pi_i \in \Pi} \sum_{\pi_j \in \Pi} \Big(\big(\pi_i \otimes \pi_j\big)\ k \Big)(\boldsymbol{x}, \boldsymbol{y})\ . \tag{10}$$

Every part of the right-hand-side sum in (10) can again be used as kernel and any sum of two or more parts just as much. Even if we consider just the positive definite kernels this gives as a big number of kernels. Instead of giving a name to each and any of them, we identify a projected kernel schematically by a $(d+2) \times (d+2)$ matrix as in Figure 5.

The first row and the first column of the matrix correspond to the projection $\pi_{\mathcal{C}}$. The d rows and d columns in the middle stand for the projections π_i, $1 \leq i \leq d$; and the last row and column for the ortho-additive projection. Hence, Image 5a depicts the constant kernel $(\pi_{\mathcal{C}} \otimes \pi_{\mathcal{C}})k$ and Image 5b the additive kernel as it was proposed in [DGR12][7]. The other images, 5c and

[7] We will sometimes call it the sparse additive kernel. Notice the important distinction between the sparse additive kernel $(\pi_{\mathcal{C}} \otimes \pi_{\mathcal{C}})\ k + \sum_{i=1}^{d} (\pi_i \otimes \pi_i)\ k$ and the full additive kernel $(\pi_{\mathcal{A}} \otimes \pi_{\mathcal{A}})\ k$.



(a) constant (b) sparse add. (c) full additive (d) ortho-add.

Figure 5: Schematic representation of projected kernels

5d, correspond to the full additive kernel $(\pi_{\mathcal{A}} \otimes \pi_{\mathcal{A}})k$, and the ortho-additive kernel $(\pi_{\mathcal{O}} \otimes \pi_{\mathcal{O}})k$, respectively.

There is a nice property of the here proposed schematic representation. Kernels whose corresponding matrix has a block-diagonal shape, e.g. the kernels that are depicted in Figure 5, are positive definite.

5 Application and numerical experiments

5.1 Kriging

Under the notion of kriging[8] we understand an interpolation and prediction technique stemming from geology and mining engineering. The original problem was to find veins of metal ore based on information from a limited number of boreholes. The technique provides a best linear unbiased predictor accounting for the spatial dependency of measurements by means of covariances (or variogram values, in a slightly different setup like, e.g., in [Cre93]).

Most applications of kriging are in two or three dimensions. In the thesis at hand we are interested in the d-dimensional formulation, though. Furthermore, we concentrate on the case where the random field is assumed Gaussian.

Let $(Z_x)_{x\in D}$ be a Gaussian random field over a domain $D \subseteq \mathbb{R}^d$. Assume that we have a vector Z containing the outcome of the random field at some given set $X = \{x_1, \dots, x_n\} \subseteq D$ of locations and want to predict the outcome of a location x_0. For now, let $(Z_x)_{x\in D}$ be centered. We assume furthermore that we know the covariance kernel k of the random field[9].

We consider the linear predictor

$$\hat{Z}_{x_0} = \lambda^T Z$$

with some vector of weights λ.

As shown in any standard book about kriging or Gaussian process modeling, e.g. in [GG09, Chapter 1.9.1], we can derive formulae for the Best Linear Unbiased Predictor (BLUP) of Z_{x_0} as well as for the variance τ^2 of the prediction error by minimizing the mean squared error

$$\mathrm{MSE}(x_0) = \mathbb{E}[(Z_{x_0} - \hat{Z}_{x_0})^2]$$

and solving for λ.

Using the covariance kernel k to get $\sigma^2_{x_0} := k(x_0, x_0)$, the vector $k(x_0) := (k(x_0, x_1), \dots k(x_0, x_n))^T$ and the $n \times n$ matrix K (assumed invertible) with

[8] After its recently deceased founding father Danie Krige 1919-2013, South African mining engineer and professor.

[9] In practice it is in general not known and has to be estimated

entries $K_{ij} = k(x_i, x_j)$ the optimal weights write $\lambda = K^{-1}k(x_0)$ and so the kriging predictor and variance are respectively

$$\hat{Z}_{x_0} = k(x_0)^T K^{-1} Z \quad \text{and} \quad \tau^2(x_0) = \sigma^2_{x_0} - k(x_0)^T K^{-1} k(x_0) \; .$$

If we are interested in the case of a non-centered GRF (i.e. a GRF with non-trivial expectation function m) we can generalize the formula for the predictor $\hat{Z}_{x_0}$. The prediction variance remains the same.

$$\begin{aligned} \hat{Z}_{x_0} &= m(x_0) + k(x_0)^T K^{-1}(Z - m(X)), \\ \text{where} \quad m(X) &= (m(x_1, \ldots, m(x_n))^T \end{aligned}$$

These equations are known under the name Simple Kriging. Possible extensions and generalizations have been discussed in detail in [GS11] and can be found for instance in [Cre93] or [GG09]. In the thesis at hand we will remain for convenience with the Simple Kriging settings.

5.2 Illustrations

Kriging is performed under the assumption that the covariance kernel is known. In the next section we we will see an experiment that analyses a possible kernel misspecification. Before we get to that point we first want to do some illustrative introduction.

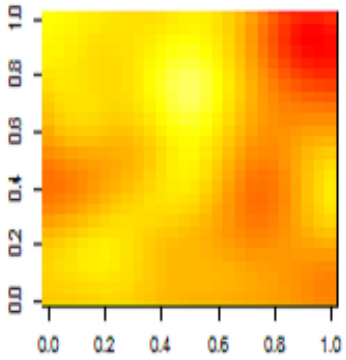


Figure 6: One realization of a centered Gaussian random field with a Gaussian covariance kernel

For this purpose we look at a discretized path that has been generated using a Gaussian kernel (see Figure 6). It consists of 26×26 pixels (i.e. 26×26 values which are arranged on a grid). Now let us assume that we know only some of these values, namely 6×6 values which lie themselves on a grid (a coarser one which covers the same domain, though).

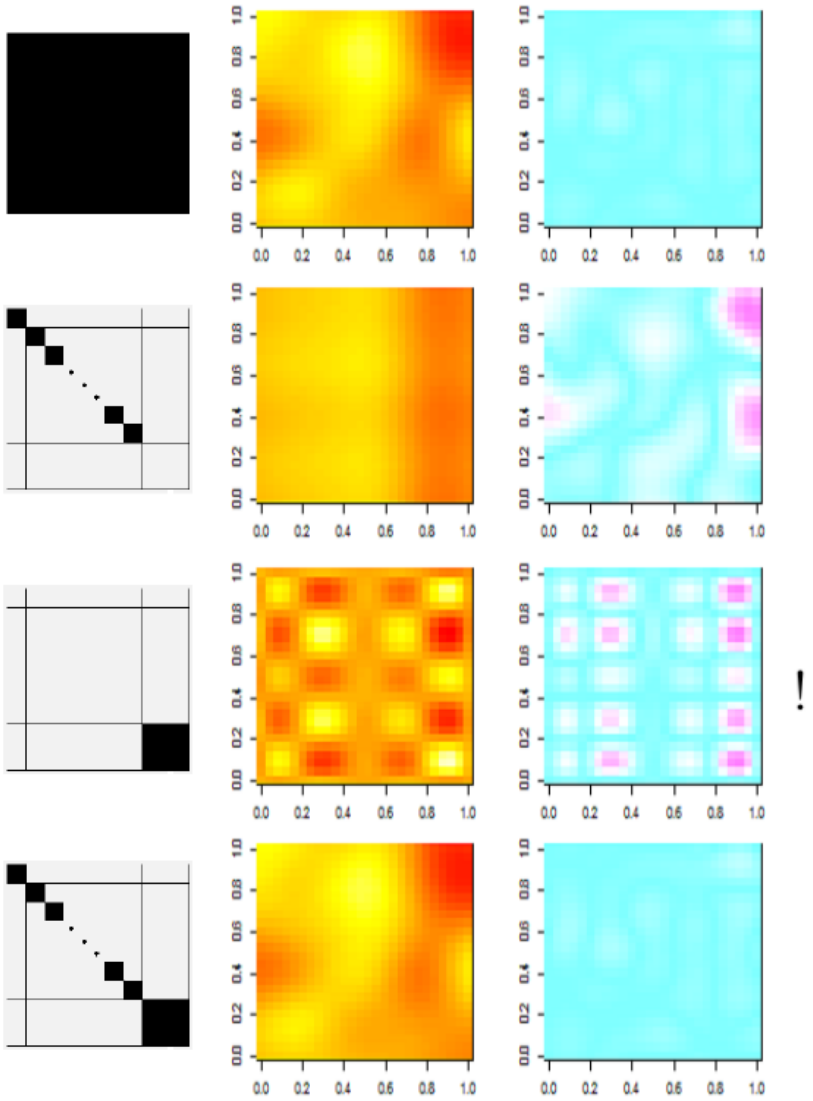


Figure 7: Predictions and prediction errors with various covariance kernel

Under these conditions we can apply kriging and predict the values that are not known. We assume, though, that we do not know the true covariance kernel either. So we perform Kriging with several kernels. In Figure 7 we see the kriging predictions that have been generated with four different kernels: the Gaussian one, its sparse additive and ortho-additive projections and the sum of the two latter. In each case we find the schematic representation of the respective kernel on the left. In the middle there are the predictions whereas on the right-hand-side we see the errors (compared to the true value from the simulated path).

We can see that the first prediction is quite good. The turquoise picture on the right indicates that there is only a minor difference between the simulated path and the prediction.

In the second row we find the results when using a sparse additive kernel. We clearly see a difference to the simulated path and this difference is also reflected in the error on the right.

When we consider the third prediction, though, we discover that it is even worse. Even much worse because the pictures are displayed using a different color scheme (this is what the exclamation mark on the very right stands for). The fourth prediction, finally, is done using a kernel which is not full but still consists of some additive and some ortho-additive part. We see that the prediction quality is comparable to the one that we had in the first row.

With these illustrations we cannot come to any conclusion. It was just one simulated path. But we will see in the subsequent section that the phenomena that we have seen here are typical for predictions with the selected kernels.

5.3 Effect of a misspecified kernel in two dimensions

A crucial step in the application of kriging is the choice of a covariance kernel (or a variogram) and its parameters. The number of possible kernels is unlimited and with the projected kernels that were introduced earlier there are even more. Typically one of the most popular kernels is picked and some parameters (e.g. variance or scale parameters) are chosen in a way such that the kernel fits best to the data. The definition and examples of variograms have been shown in [GS11] or can be found in [Cre93, Chapter 3] or [GG09, Chapters 1.3 and 5.1].

The choice of a suitable kernel is still an open area of research. In the following, though, we change the point of view and assume that a kernel has been chosen. We focus on the consequences if the choice was bad.

5.3.1 Setup of the experiment

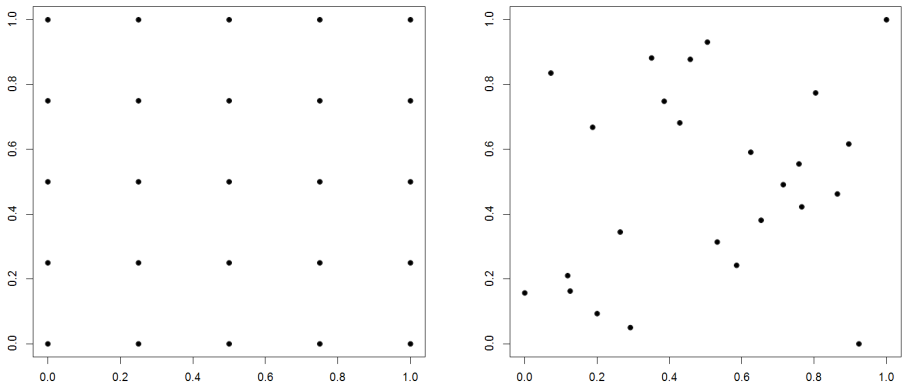
We present here a general experiment that can be run for an arbitrary set of kernels.

1. Determine the domain, the set of kernels and the set of locations for the measurements X_m as well as for the predictions X_p
2. For each pair of kernels (k_i, k_j) repeat the following N times:
   (a) Simulate data simultaneously on X_m (Z_m: the measurements) and on X_p (Z_p: control values) using the kernel k_i
   (b) Based on the measurements Z_m in X_m predict the values $\hat{Z}_p$ at the locations X_p using the second kernel k_j
   (c) Estimate the ISE (Integrated Squared Error) by

$$\frac{1}{\#X_p} \sum_{x \in X_p} [\hat{Z}_p(x) - Z_p(x)]^2 \ , \quad \#X_p \text{ denoting the cardinality of } X_p$$

The concrete numerical experiment for this thesis was performed under the following conditions:

- The domain: $D = [0, 1]^2$. And the used kernels:
  - The Gaussian kernel (with $\sigma^2 = 1$, $\theta_1 = \theta_2 = 0.2$)
  - Its full additive part
  - Its (sparse) additive part
  - The ortho-additive part
  - The sum of the full additive and the ortho-additive part
  - The sum of the sparse additive and the ortho-additive part
- X_p: A grid consisting of 45×45 points
- There were two runs with $N = 500$ in both cases but different X_m:
  (a) A coarser grid with 5×5 points
  (b) An LHS design consisting of 25 points



(a) grid (b) LHS

Figure 8: Design of experiment (25 points in each case)

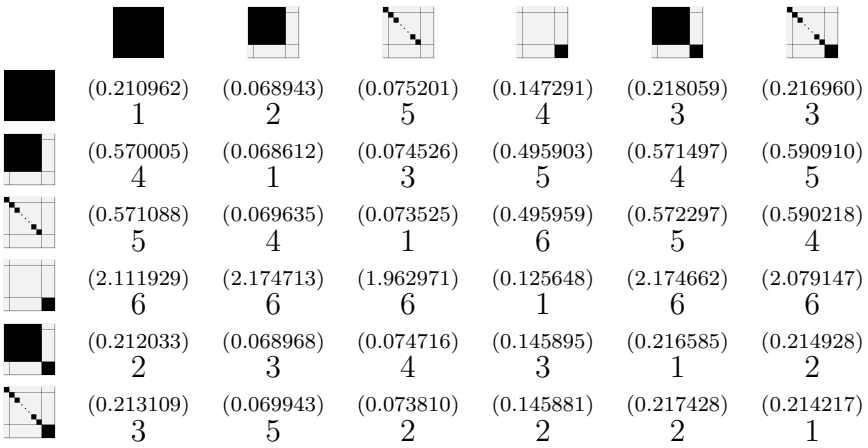
X_m is called the design of experiment (DoE). The two designs that are used in the experiment are shown in Figure 8. LHS stands for latin hypercube sampling [MBC79]. This method generates n points in a d-dimensional domain which are distributed such that their projections onto each of the coordinate axes always contains a point in every interval $[\frac{i-1}{n}, \frac{i}{n}]$ $(1 \leq i \leq n)$.

The experiment was written using the free statistical software R [R D08]. The execution of the code on an average personal computer lasted about 12 hours.

5.3.2 Results

The results are displayed in Table 1 (for measurements on a grid) and Table 2 (for the LHS arrangement of measurements). Any entry of the tables contains (in brackets) the average of the ISE after 500 runs along with the column-wise rank. We can see in the respective column header the kernel that was used for simulating data and in the row header the kernel which was used for the predictions.

The motivation for the actual experiments was the assumption that a dataset in practice contains additive as well as non-additive information. Even with a Gaussian kernel we have different possibilities to treat the additive and the non-additive part of the data. The format of the experiment gives us the possibility to compare several of them.



(0.210962) 1	(0.068943) 2	(0.075201) 5	(0.147291) 4	(0.218059) 3	(0.216960) 3
(0.570005) 4	(0.068612) 1	(0.074526) 3	(0.495903) 5	(0.571497) 4	(0.590910) 5
(0.571088) 5	(0.069635) 4	(0.073525) 1	(0.495959) 6	(0.572297) 5	(0.590218) 4
(2.111929) 6	(2.174713) 6	(1.962971) 6	(0.125648) 1	(2.174662) 6	(2.079147) 6
(0.212033) 2	(0.068968) 3	(0.074716) 4	(0.145895) 3	(0.216585) 1	(0.214928) 2
(0.213109) 3	(0.069943) 5	(0.073810) 2	(0.145881) 2	(0.217428) 2	(0.214217) 1

Table 1: Results of the first experiment using a grid design. Average mean integrated squared error (in brackets) and column-wise rank

Regarding the results in Table 1 we can draw several conclusions. For one they reflect that the predictions are best if we use the kernel which also produced the data, i.e. the column-wise minimal average ISE is always on the diagonal of the table. It is not surprising. But we can also observe more interesting connections.

When we have a look at the first column (where the data was simulated using the full Gaussian kernel) we see that the predictions using the kernels in the last two rows are almost as good as the predictions with the Gaussian kernel itself. Using an only additive or an only ortho-additive kernel we do not get good predictions. Concerning the additive kernel there is a huge difference with respect to the DoE. We will come back to that point later.

When the data is produced by an additive kernel (in the second and third column of the two tables) a different image is conveyed. The predictions by an additive kernel are close to perfect if we have an LHS design. Having a grid design, though, it does not help much to know that the data is additive. All the kernels that we used, except the only ortho-additive one, had almost the same prediction quality.

For only ortho-additive data we get a similar image. With a grid design all but the only additive kernels have similar prediction quality. In case of an LHS design the differences are bigger with outstanding bad results using the only additive kernels.

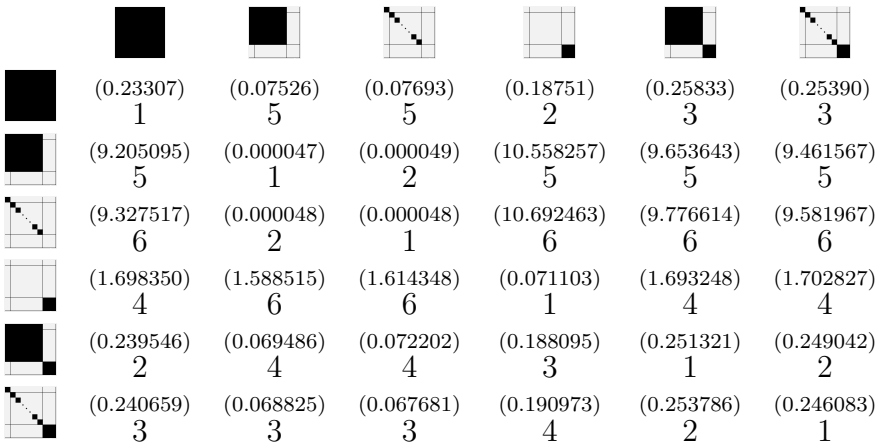


Table 2: Results of the second experiment using LHS. Average mean integrated squared error (in brackets) and column-wise rank

In the last two columns there is no further surprise. They emphasize, though, that the results differ with respect to the DoE.

Combined the results suggest that the (solely) ortho-additive kernel is not suitable for predictions if the data is generated by any other kernel. Of course the experiment treats only a small number of kernels but it is plausible that a kernel which does not make allowances for additive data cannot depict the complexity of some (more general) dataset. The same applies to the sparse additive and the full additive kernel (which cannot gather the non-additive part). But, interestingly, we can see it only in the experiment with the LHS design. This shows us that the DoE can have a critical impact.

The actual phenomenon concerning the solely additive kernels looks (in terms of the average ISE) worse than it is and it can be explained. Other experiments (for instance [DGR12]) have shown earlier that an LHS design is suitable when working with an additive kernel. In the current case, however, there occurred two problems. For one there was an over-fitting towards the measurements which was reflected by strongly oscillating predictions. In addition there were bad predictions at the borders of the domain because the LHS design does not cover the domain as well as the grid does. The latter, though, is a problem with which also the other kernels are confronted.

Finally there is one more observation which is probably the most important. By comparing the average ISE we notice that there is almost no difference between the second and the third row, as well as between the fifth and the last row. And even the differences between the first and the last two rows are marginal! That means that our predictions with a sparse kernel (e.g. the last one) offer almost the same quality as the predictions with the full kernel. Or, in the context of the double decomposition the experiment suggests that neglecting the non-diagonal parts of the Gaussian kernel does not gravely harm the prediction quality.

5.4 Quantifying additivity in higher dimensions

The results in Section 5.3 suggest that an additive kernel can be enriched with an extra ortho-additive term. This makes sense, for random fields in general have some non-additive part. This section presents a special class of kernels which allow to quantify the additivity by some coefficient. Then it shows how the coefficient can be recovered from data that was produced by such a kernel.

For the experiment we propose the following model. We work with a kernel k_ψ which is a weighted sum of the additive and ortho-additive projection of a Gaussian kernel, $k_{\mathcal{A}}$ and $k_{\mathcal{O}}$.

$$k_\psi := \sigma^2(\alpha\, k_{\mathcal{A}} + (1-\alpha)\, k_{\mathcal{O}}), \quad \text{for some} \quad \sigma^2 > 0,\ \alpha \in [0,1]$$

ψ denotes the parameter vector (α, σ^2). We call α the coefficient of additivity.

The experiment is, again, performed under laboratory conditions. We fix the coefficient of additivity α of k_ψ and generate data with this kernel. Then we try to recover α from the data by maximum likelihood estimation.

5.4.1 Applying maximum likelihood estimation

Maximum likelihood estimation (MLE) is a technique to estimate the parameters of a statistical model relying on data. We recapitulate shortly how it can be applied in the context of a GRF model (see also [GS11] or [GG09]). We treat the specific case of a kernel k_ψ as above.

From a centered Gaussian random field $(Z_x)_{x\in D}$ we consider some realization. It is represented by a vector of measurements Z at $\{x_1, \ldots, x_n\} \subseteq D$.

The according random vector follows a multivariate Gaussian distribution, $\boldsymbol{Z} \sim \mathcal{N}(\mathbf{0}, K_\psi)$ where K_ψ is the $n \times n$ covariance matrix with respect to k_ψ. Now we try to find the most likely parameter vector ψ for the present realization.

The likelihood function L is the density function of the according distribution (in our case the multivariate Gaussian), seen as a function of ψ. The vector of measurements Z is fixed.

$$L(\psi) = \frac{1}{(2\pi)^{\frac{n}{2}}\sqrt{|K_\psi|}}\, e^{-\frac{1}{2}Z^T K_\psi^{-1} Z}$$

Instead of L we consider the log-likelihood function $l = \log(L)$. It is equivalent to maximize either l or L, for the logarithm is a monotonically increasing bijection.

$$l(\psi) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\log(|K_\psi|) - \frac{Z^T K_\psi^{-1} Z}{2}$$

Notice that K_ψ is $\sigma^2 K_c(\alpha)$ where $K_c(\alpha) = \alpha\, K_{\mathcal{A}} + (1-\alpha)\, K_{\mathcal{O}}$, with $K_{\mathcal{A}}$ and $K_{\mathcal{O}}$ being $n \times n$ covariance matrices with respect to the kernels $k_{\mathcal{A}}$ and $k_{\mathcal{O}}$, respectively.

$$l(\psi) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2}\log(|K_c(\alpha)|) - \frac{Z^T K_c(\alpha)^{-1} Z}{2\sigma^2}$$

When we consider the partial derivative of l with respect to σ^2 we can see that for any α we can calculate

$$\sigma_*^2(\alpha) = \frac{Z^T K_c(\alpha)^{-1} Z}{n}$$

that minimizes l. In this case it is equivalent to maximize the profile log-likelihood function

$$l_p(\alpha) := l(\alpha, \sigma_*^2(\alpha)) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma_*^2(\alpha)) - \frac{1}{2}\log(|K_c(\alpha)|) - \frac{n}{2}$$

instead of l. Or, as we will do in the experiments, we can find the argument that minimizes $-2\, l_p(\alpha)$.

In the case where we have several independent realizations of $(Z_x)_{x \in D}$ we can look at each log-likelihood function individually or we can derive an

overall log-likelihood. The result is simply the mean of the log-likelihood functions of the different realizations.

Let us finally remark that, even though we considered this option for monitoring purposes, it is not perfectly rigorous to do that directly with profile log-likelihood functions, for different realizations lead to different σ_*^2 for a given α.

5.4.2 Setup of the experiment

In the following we describe the structure of the experiment and the chosen parameters.

1. We consider six kernels in dimension $d = 2, 3, \ldots, 9$
   (a) We start with a Gaussian kernel k_{gau} as in Equation (4), with $\sigma^2 = 1$ and $\theta = 0.2$. We extract its additive and ortho-additive part and construct five instances of k_ψ with $\alpha \in \{0.00, 0.25, 0.50, 0.75, 1.00\}$ and $\sigma^2 = 1$ in each case.
   (b) We use k_{gau} itself with $\theta = 0.2$
2. For each kernel k and each dimension d we
   (a) generate a random design of experiment containing $10 \cdot d$ points. We used two different types of design:
      i. An LHS design as in Section 5.3
      ii. A design containing uniformly distributed points in $[0, 1]^d$
   (b) calculate the covariance matrix using k and the current design
   (c) simulate a realization Z of the random vector $\boldsymbol{Z}$
   (d) evaluate $-2l_p$ at the values $\{0.00, 0.01, \ldots, 0.99\}$
   (e) find the argument $\hat{\alpha}$ for which $-2\, l_p$ is minimal

For every kernel, every dimension and both designs we made, at first, 200 iterations. Then we ran the experiment with 1000 iterations.

Notice that the likelihood function is not evaluated at $\alpha = 1.00$ (full additive). When we consider data which is not full additive then the likelihood function typically becomes very small or is even exactly zero. In the latter case the logarithm is undefined. Therefore we decided to neglect $\alpha = 1.00$.

The source code for the experiment was written in the high-level programming language Python [Fou12] using the well-established extension numpy for high-dimensional calculations [Oli07] and the plotting package matplotlib [Hun07]. For the specific calculations concerning Gaussian random fields we made use of GPy [HFA+13][10]. The code was executed on a Linux instance with average power. The experiment for one DoE and 1000 iterations lasted about two days.

5.4.3 Results

The summarized results of the experiment are presented in Figure 11. But we will first consider one specific case, discussing the results in dimension 5 for $\alpha = 0.75$ when working with a design of uniformly distributed points.

We shortly recall the experiment setting for this case: A kernel k_ψ in dimension 5 is constructed with $\sigma^2 = 1$, $\alpha = 0.75$. Then a design of 50 points, uniformly distributed in $[0, 1]^5$, is created and a path of the Gaussian random field with kernel k_ψ is simulated at the design. Then the log-likelihood func-

[10] A special release containing an ortho-additive kernel was provided by Nicolas Durrande to support this thesis!

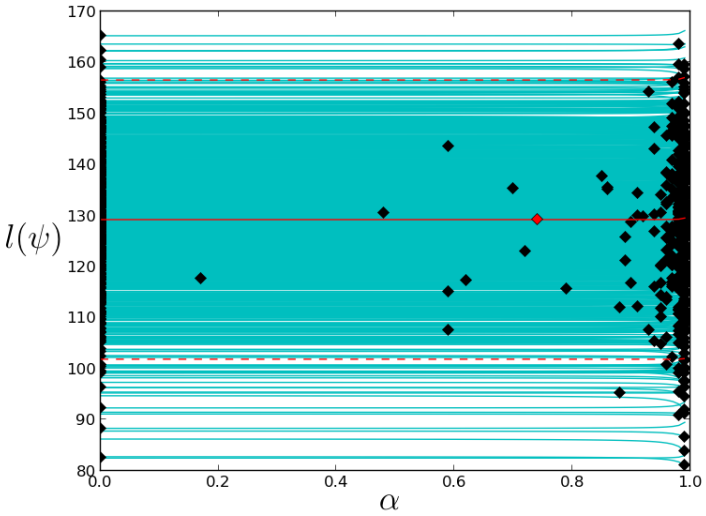


Figure 9: Results in dimension 5 and for $\alpha = 0.75$. 1000 log-likelihood curves (thin cyan lines) with their minima (black diamonds), the mean curve (thick red line) with its minimum (red diamond)

tion l is calculated for 100 equally spaced values in $[0, 1]$. The whole thing is repeated 1000 times.

The likelihood curves are drawn in cyan in the graph in Figure 9. We can see that all the curves have a rather flat shape. Only to the very right, close to $\alpha = 1$, some of them are bent up or down. The arguments $\hat{\alpha}$ which minimize the curves are often close to or exactly 0 or 1. More information about the empirical distribution of $\hat{\alpha}$ is contained in Figure 10 in terms of a histogram (upper left part of the figure). In the histogram we see (even better than in Figure 9) that the values are divided into 2 sets.

The histogram of the variances σ_*^2 (upper right part of Figure 10) shows that most values are smaller than 2. In the plot underneath we see that only for high values of α the variance becomes big. The curves that are presented appear to have a smooth shape.

From the empirical distribution of $\hat{\alpha}$ we cannot recover the original value

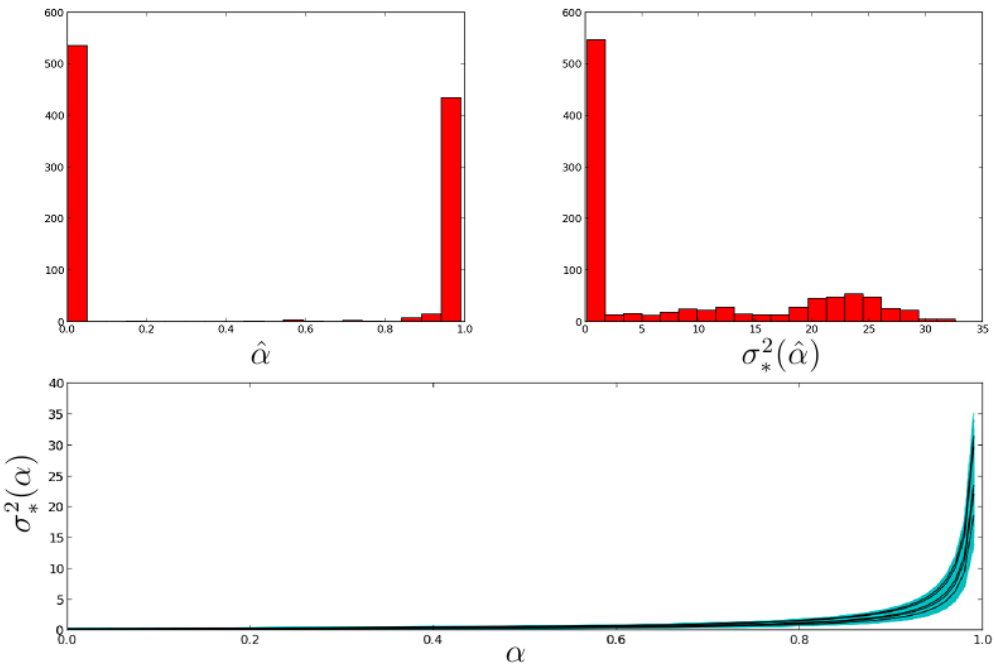


Figure 10: Upper left and upper right, respectively: Histograms with 1000 values of $\hat{\alpha}$ and $\sigma_*^2(\hat{\alpha})$; Bottom: curves showing $\sigma_*^2(\alpha)$ for 1000 realizations of the random vector $\boldsymbol{Z}$ (cyan lines, 5 exemplary curves in black)

of 0.75. This is why we decided to extend the experiment.

Although the procedure is inappropriately simplified (see the last remark of Section 5.4.1) we took the mean over all the profile log-likelihood functions. The correct approach is computationally more expensive but should certainly be considered in the future. Hence, the following results should be interpreted with caution.

In Figure 9 the mean of the $-2\ l_p$ curves is depicted as a red line. The argument which minimizes it is 0.74 (marked by the red diamond). This is reasonably close to the initially chosen original value $\alpha = 0.75$. In the following we use this value as an estimator and call it $\hat{\alpha}$.

We gathered the results for all dimensions and for all kernels of the form k_ψ in Figure 11. There are four sub-figures. The upper two concern the case in which the DoE consists of uniformly distributed points in the domain $[0, 1]^d$. The lower two correspond to LHS designs. To the left and to the right are respectively the results when running 200 and 1000 iterations.

In each graph we see five dashed black lines. Each of them corresponds to one of the values that we have chosen for α, i.e. $0.00, 0.25, 0.50, 0.75$ and 1.00. In the left part of the sub-figures we see clearly to what value each line belongs. For every dimension in $\{2, 3, \ldots, 9\}$ the estimated values for α are plotted. The estimate $\hat{\alpha} = 0.74$ that we discussed before can be found in the upper right graph, marked by a blue circle.

When we have a closer look at the results we notice first that the estimates in low dimensions are much better than in high dimensions. We can see that even with 200 iterations we can find reasonable estimates for α when the dimension does not exceed 4. With 1000 iterations we can go to 5. For $d > 5$ the experiment fails to recover the value of α. It appears, though, that high additivity may be easier to recover than low additivity.

A special case is $\alpha = 1.00$. The estimate is always 0.99 because we decided to skip 1.00 as mentioned earlier. Additional experiments (not reproduced here) with just a few iterations confirmed that if we would allow the true value, the estimate would be correct.

The type of design, in the experiment, does not seem to influence the results gravely. Still it would be interesting to perform the experiments with another design. We recall that in the experiment in Section 5.3 the effect of the DoE was remarkable.

Finally, there are the plain red lines which are related to data that was produced using a Gaussian kernel. Here we have to be cautious, for we perform MLE now with a misspecified kernel k_ψ. The estimated coefficient

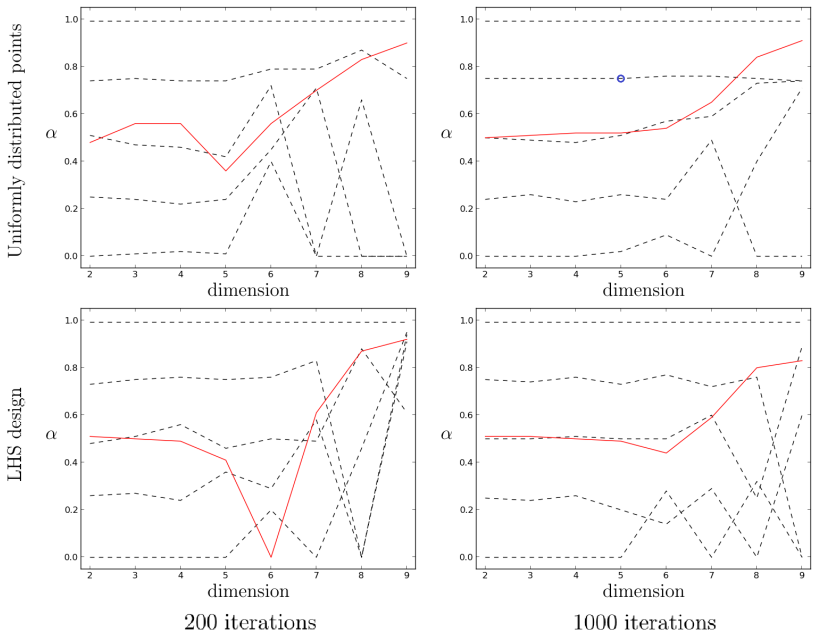


Figure 11: Recovering/Estimating the coefficient of additivity of a dataset. Four settings of the experiment, each containing five curves (dashed, black) corresponding to $\alpha = 0.00, 0.25, 0.50, 0.75, 1.00$, respectively, and one red line related to the Gaussian kernel

of additivity $\hat{\alpha}$ belongs to k_ψ. Its relation to the Gaussian kernel has to be interpreted. The estimates say: if the data (which, in fact, stems from a Gaussian kernel) was generated by a kernel of the form k_ψ, then it is most likely that the coefficient of additivity of the kernel was about 0.5 (at least for $d \leq 6$).

This is reasonable and we can explain it. We could do the same experiment with almost any other kernel. We just have to keep in mind that $k_{\mathcal{A}}$ and $k_{\mathcal{O}}$, in that case, would be the projections of that very kernel. Consequently the experiment would most likely return 0.5. We say almost any kernel because the experiment is pointless when the kernel itself is purely additive or ortho-additive, and hence one of the projections trivial.

Towards higher dimensions ($d > 6$), again, the estimates do not appear reliable. In all the four subgraphs, though, the red curves increase with growing dimension. It is not clear whether this is a coincidence or if there is an explanation.

The experiment indicates that it is not possible to recover the coefficient of additivity of a random field with kernel k_ψ by maximum likelihood estimation, when we know just one realization and have the here chosen design parameters (type of design and number of points).

On the other hand it is possible to combine the information of several realizations in order to calculate some combined likelihood. Based on this we can apply MLE. Although we present here some simplified procedure, which should be improved, the experiment still suggests that it is possible to recover the coefficient of additivity of a kernel from data, in the case when we have many realizations. However, the calculations are computationally intensive. We recall that we had a linear budget concerning the number of points in the design. The number of iterations was constant. With the chosen parameters the values for α are well recovered in low dimensions. The results for $d > 6$, however, do not appear reliable. In this case, we suggest either to adjust the DoE (e.g. by adding more points) or to base the calculations on more realizations.

An open question is how the approach could be extended to estimate the additivity for an arbitrary Gaussian random field. In our calculations the coefficient of additivity α always belongs to a concrete basis kernel. It can only be interpreted relatively to the basis kernel. Further investigations are necessary in this regard.

In addition it would be interesting to decompose another kernel into an additive and an ortho-additive part. With this result we could try to recover the coefficient of additivity of one tunable kernel using another one. This would demonstrate the effect when the kernel was misspecified even before the decomposition.

6 Conclusion and Perspectives

The thesis deals with Gaussian random fields and identifies a close relationship between their covariance kernel and the associated paths. In order to get a deeper understanding of this relationship we examined the space of the random field paths and the space of kernels using some results from functional analysis. We focused on the correspondence between applying orthogonal projections in both spaces.

We introduced orthogonal projections which divide a path of a centered Gaussian random field into an additive part and its orthogonal complement, here called the ortho-additive part. We derived according projections for the covariance kernel, i.e. projections which allow us to create random fields whose paths are additive or ortho-additive, respectively.

We studied these projections in depth. The thesis reveals general formulae for the projection of product kernels and presents the calculation in the case of a squared exponential (or Gaussian) kernel.

Of rather practical importance is the procedure to evaluate newly derived kernels, suggested in Section 5.3. A cross-check technique is introduced which compares a number of random fields having different covariance kernels. The technique proposes to simulate random field paths with each kernel in turn and do kriging predictions with the other kernels. We applied this procedure to the Gaussian kernel, its additive and ortho-additive projection and the sum of the two latter (which turns out to be different from the original kernel!). The technique proves useful to compare kernels and identify those which are particularly robust in the face of misspecification. In the concrete case we identified that neglecting some cross-covariances may not necessarily lead to a loss of prediction power.

In Section 5.4 we introduced GRFs with covariance kernels that have an adjustable additive part (by means of some coefficient). We used the kernels in an experiment to estimate the additivity of such a random field, based on a set of its paths. The first estimates were carried out when the true coefficient was known. Although the here used simplified MLE procedure should be investigated deeper, the estimates were reasonably close to the true values for up to 5-dimensional datasets. The here suggested technique is applicable to recover the coefficient of additivity. However, the calculations are computationally expensive.

Summarized we see two kinds of achievements of the thesis. First, some rather concrete results. Namely the decomposition of random field paths and kernels into additive and ortho-additive parts. And the explicit calculations concerning the Gaussian kernel.

The other achievements of the thesis are more general. They include guidelines and procedures. The thesis can also be used as a tutorial about how to decompose a kernel and work with the resulting partition.

We see certainly potential for future improvements. Amongst other things the detection and quantification of additivity in Gaussian random fields could be investigated much deeper.

In addition to enhancing the here presented results it would be interesting to consider a completely different decomposition of kernels. There are plenty of possible orthogonal function decompositions in L^2 (see for instance [DHRL13]). For any of them we could do a double decomposition for a kernel. This would provide us with new classes of kernels which could be investigated with the variants of the experimental protocol introduced here.

7 Appendix

Here we perform detailed calculations for the proof of Proposition 3 in Section 4.2.1 concerning projections of a kernel k defined on the domain $D = [a_1, b_1] \times \ldots \times [a_d, b_d]$, which fulfills

$$k(\boldsymbol{x}, \boldsymbol{y}) = \prod_{i=1}^{d} k_i(x_i, y_i)$$
$$k_i(x_i, y_i) > 0 \; \forall \; x_i, y_i \in [a_i, b_i] \quad (1 \leq i \leq d)$$

and for which the quantities

$$E_i(x_i, a_i, b_i) \; := \int_{a_i}^{b_i} k_i(x_i, y_i) \; dy_i$$
$$\text{and} \quad \mathcal{E}_i(a_i, b_i) \; := \int_{a_i}^{b_i} E(x_i, a_i, b_i) \; dx_i$$

are known.

Notice that positivity - the latter of the conditions - is not compulsory. It guarantees, though, that the denominators in the formulae below are non-zero. But for a kernel which is not positive everywhere we can write the formula without denominators.

We are using the auxiliary quantities (which can be computed at least numerically):

- $E_i(x_i) := E_i(x_i, a_i, b_i) = \int_{a_i}^{b_i} k_i(x_i, y_i) \; dy_i$
- $E(\boldsymbol{x}) := E(\boldsymbol{x}, \boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{d} E_i(x_i, a_i, b_i)$
- $\mathcal{E}_i := \mathcal{E}_i(a_i, b_i) = \int_{a_i}^{b_i} E(x_i, a_i, b_i) \; dx_i$
- $\mathcal{E} := \mathcal{E}(\boldsymbol{a}, \boldsymbol{b}) = \prod_{i=1}^{d} \mathcal{E}_i(a_i, b_i)$.

For the sake of simplicity and legibility the formulae are shortened and simplified as follows. The dependent variable of a function is often omitted. Instead of

$$\big(\pi_{\mathcal{C}}^{r} \; k\big)(\boldsymbol{x}, \boldsymbol{y}) = \left(\int_D k(\tilde{\boldsymbol{x}}, \tilde{\boldsymbol{y}}) \; d\tilde{\boldsymbol{y}} \right)(\boldsymbol{x}, \boldsymbol{y}) = E(\boldsymbol{x})$$

we write the minimum that is required to cover the essential information, like

$$\pi^r_{\mathcal{C}}\, k = \int_D k\; d\boldsymbol{y} = E(\boldsymbol{x})\ .$$

Moreover some parts of the calculations that occur often (like the application of Fubini) are skipped, e.g.

$$\int_D k\; d\boldsymbol{y} = \int_D \prod_{i=1}^{d} k_i(x_i, y_i)\; d\boldsymbol{y} = \prod_{i=1}^{d} \int_D k_i(x_i, y_i)\; dy_i = \prod_{i=1}^{d} E_i(x_i) = E(\boldsymbol{x})\ ,$$

and we will benefit of the assumption that the auxiliary quantities are strictly positive by writing them into the denominator, like

$$\prod_{\substack{j=1\\ j\neq i}}^{d} E_j(x_j) = \frac{E(\boldsymbol{x})}{E_i(x_i)}\ .$$

In the following calculations, terms that are needed to prove Proposition 3 are printed in bold type. The other terms are written down in the interests of completeness.

Applying the right-hand-side projections we get

$$\begin{aligned}
\boldsymbol{\pi^r_{\mathcal{C}} k} &= \int_D k\; d\boldsymbol{y} = E(\boldsymbol{x})\\
\boldsymbol{\pi^r_j k} &= \int_{D_{-j}} k - \pi^r_{\mathcal{C}} k\; d\boldsymbol{y}_{-j} = \int_{D_{-j}} k\; d\boldsymbol{y}_{-j} - \int_{D_{-j}} \pi^r_{\mathcal{C}} k\; d\boldsymbol{y}_{-j}\\
&= k_j(x_j, y_j) \cdot \frac{E(\boldsymbol{x})}{E_j(x_j)} - E(\boldsymbol{x}) = E(\boldsymbol{x}) \left(\frac{k_j(x_j, y_j)}{E_j(x_j)} - 1 \right)\\
\boldsymbol{\pi^r_{\mathcal{A}} k} &= \pi^r_{\mathcal{C}} k + \sum_{j=1}^{d} \pi^r_j k = E(\boldsymbol{x}) \left(1 - d + \sum_{j=1}^{d} \frac{k_j(x_j, y_j)}{E_j(x_j)} \right)\\
\boldsymbol{\pi^r_{\mathcal{O}} k} &= k - \pi^r_{\mathcal{A}} k = E(\boldsymbol{x}) \left(\frac{k(\boldsymbol{x}, \boldsymbol{y})}{E(\boldsymbol{x})} - 1 + d - \sum_{j=1}^{d} \frac{k_j(x_j, y_j)}{E_j(x_j)} \right)
\end{aligned}$$

and accordingly

$$\begin{aligned}
\boldsymbol{\pi}^{\boldsymbol{l}}_{\boldsymbol{\mathcal{C}}}\boldsymbol{k} &= E(\boldsymbol{y}) \\
\boldsymbol{\pi}^{\boldsymbol{l}}_{\boldsymbol{i}}\boldsymbol{k} &= E(\boldsymbol{y})\left(\frac{k_i(x_i,y_i)}{E_i(y_i)} - 1\right) \\
\pi^l_{\mathcal{A}}k &= E(\boldsymbol{y})\left(1-d+\sum_{i=1}^{d}\frac{k_i(x_i,y_i)}{E_i(y_i)}\right) \\
\pi^l_{\mathcal{O}}k &= E(\boldsymbol{y})\left(\frac{k(\boldsymbol{x},\boldsymbol{y})}{E(\boldsymbol{y})} - 1 + d - \sum_{i=1}^{d}\frac{k_i(x_i,y_i)}{E_i(y_i)}\right).
\end{aligned}$$

When combining the two projections we can compute the rest. We introduce one last auxiliary quantity which will occur several times in the context of additivity:

$$A(\boldsymbol{x}) := \mathcal{E}\left(1-d+\sum_{i=1}^{d}\frac{E_i(x_i)}{\mathcal{E}_i}\right).$$

With all the auxiliary quantities we can express all the projected kernels.

$$\begin{aligned}
\pi_{\mathcal{C}}k &= \mathcal{E} \\
\pi_{\mathcal{C}j}k &= \mathcal{E}\left(\frac{E_j(y_j)}{\mathcal{E}_j} - 1\right) \\
\boldsymbol{\pi}_{\boldsymbol{\mathcal{C}\mathcal{A}}}\boldsymbol{k} &= \int_D E(\boldsymbol{x})\left(1-d+\sum_{j=1}^{d}\frac{k_j(x_j,y_j)}{E_j(x_j)}\right)d\boldsymbol{x} \\
&= (1-d)\int_D E(\boldsymbol{x})\,d\boldsymbol{x} + \sum_{j=1}^{d}\int_D E(\boldsymbol{x})\frac{k_j(x_j,y_j)}{E_j(x_j)}\,d\boldsymbol{x} \\
&= (1-d)\,\mathcal{E} + \sum_{j=1}^{d}\mathcal{E}\frac{E_j(y_j)}{\mathcal{E}_j} = \mathcal{E}\left(1-d+\sum_{j=1}^{d}\frac{E_j(y_j)}{\mathcal{E}_j}\right) = A(\boldsymbol{y}) \\
\boldsymbol{\pi}_{\boldsymbol{\mathcal{C}\mathcal{O}}}\boldsymbol{k} &= \pi^l_{\mathcal{C}}k - \pi_{\mathcal{C}\mathcal{A}}k = E(\boldsymbol{y}) - A(\boldsymbol{y})
\end{aligned}$$

$$\pi_{i\mathcal{C}}k = \mathcal{E}\left(\frac{E_i(x_i)}{\mathcal{E}_i} - 1\right)$$

$$\pi_{ij}k = \mathcal{E}\left(\frac{E_i(x_i)}{\mathcal{E}_i} - 1\right)\left(\frac{E_j(x_j)}{\mathcal{E}_j} - 1\right) \text{ for } i \neq j$$

$$\pi_i k = \pi_{ii}k = \mathcal{E}\left(\frac{k_i(x_i, y_i) - E_i(x_i) - E_i(y_i) + \mathcal{E}_i}{\mathcal{E}_i}\right)$$

$$\begin{aligned}
\boldsymbol{\pi_{i\mathcal{A}}k} &= \int_{D_{-i}} \pi^r_{\mathcal{A}}k - \pi_{\mathcal{C}\mathcal{A}}k \; d\boldsymbol{x}_{-i} \\
&= \int_{D_{-i}} E(\boldsymbol{x})\left(1 - d + \sum_{j=1}^{d} \frac{k_j(x_j, y_j)}{E_j(x_j)}\right) - A(\boldsymbol{y}) \; d\boldsymbol{x}_{-i} \\
&= \mathcal{E}\left[(1-d)\cdot\frac{E_i(x_i)}{\mathcal{E}_i} + \frac{k_i(x_i, y_i)}{\mathcal{E}_i} + \sum_{\substack{j=1 \\ j\neq i}}^{d} \frac{E_i(x_i)\; E_j(y_j)}{\mathcal{E}_i\; \mathcal{E}_j}\right] - A(\boldsymbol{y}) \\
&= A(\boldsymbol{y})\left(\frac{E_i(x_i)}{\mathcal{E}_i} - 1\right) + \mathcal{E}\left(\frac{k_i(x_i, y_i)}{\mathcal{E}_i} - \frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)
\end{aligned}$$

$$\begin{aligned}
\boldsymbol{\pi_{i\mathcal{O}}k} &= \pi^l_i k - \pi_{i\mathcal{A}}k = E(\boldsymbol{y})\left(\frac{k_i(x_i, y_i)}{E_i(y_i)} - 1\right) - A(\boldsymbol{y})\left(\frac{E_i(x_i)}{\mathcal{E}_i} - 1\right) \\
&\quad - \mathcal{E}\left(\frac{k_i(x_i, y_i)}{\mathcal{E}_i} - \frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)
\end{aligned}$$

$$\pi_{\mathcal{A}\mathcal{C}}k = A(\boldsymbol{x})$$

$$\pi_{\mathcal{A}j}k = A(\boldsymbol{x})\left(\frac{E_j(y_j)}{\mathcal{E}_j} - 1\right) + \mathcal{E}\left(\frac{k_j(x_j, y_j)}{\mathcal{E}_j} - \frac{E_j(x_j)E_j(y_j)}{\mathcal{E}_j^2}\right)$$

$$\begin{aligned}
\boldsymbol{\pi_{\mathcal{A}}k} &= \pi_{\mathcal{C}\mathcal{A}}k + \sum_{i=1}^{d} \pi_{i\mathcal{A}}k \\
&= A(\boldsymbol{y}) + \sum_{i=1}^{d}\left[A(\boldsymbol{y})\left(\frac{E_i(x_i)}{\mathcal{E}_i} - 1\right) + \mathcal{E}\left(\frac{k_i(x_i, y_i)}{\mathcal{E}_i} - \frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)\right] \\
&= \frac{A(\boldsymbol{x})A(\boldsymbol{y})}{\mathcal{E}} + \mathcal{E}\cdot\sum_{i=1}^{d}\left(\frac{k_i(x_i, y_i)}{\mathcal{E}_i} - \frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)
\end{aligned}$$

$$
\begin{aligned}
\pi_{\mathcal{A}\mathcal{O}}k &= E(\boldsymbol{y})\left(1-d+\sum_{i=1}^{d}\frac{k_i(x_i,y_i)}{E_i(y_i)}\right)-\frac{A(\boldsymbol{x})A(\boldsymbol{y})}{\mathcal{E}}\\
&\quad-\mathcal{E}\cdot\sum_{i=1}^{d}\left(\frac{k_i(x_i,y_i)}{\mathcal{E}_i}-\frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)\\
\pi_{\mathcal{O}\mathcal{C}}k &= E(\boldsymbol{x})-A(\boldsymbol{x})\\
\pi_{\mathcal{O}j}k &= E(\boldsymbol{x})\left(\frac{k_j(x_j,y_j)}{E_j(x_j)}-1\right)-A(\boldsymbol{x})\left(\frac{E_j(y_j)}{\mathcal{E}_j}-1\right)\\
&\quad-\mathcal{E}\left(\frac{k_j(x_j,y_j)}{\mathcal{E}_j}-\frac{E_j(x_j)E_j(y_j)}{\mathcal{E}_j^2}\right)\\
\pi_{\mathcal{O}\mathcal{A}}k &= E(\boldsymbol{x})\left(1-d+\sum_{i=1}^{d}\frac{k_i(x_i,y_i)}{E_i(y_i)}\right)-\frac{A(\boldsymbol{x})A(\boldsymbol{y})}{\mathcal{E}}\\
&\quad-\mathcal{E}\cdot\sum_{i=1}^{d}\left(\frac{k_i(x_i,y_i)}{\mathcal{E}_i}-\frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)\\
\boldsymbol{\pi_{\mathcal{O}}k} &= \pi^r_{\mathcal{O}}k-\pi_{\mathcal{C}\mathcal{O}}k-\sum_{i=1}^{d}\pi_{i\mathcal{O}}k\\
&= E(\boldsymbol{x})\left(\frac{k(\boldsymbol{x},\boldsymbol{y})}{E(\boldsymbol{x})}-1+d-\sum_{j=1}^{d}\frac{k_j(x_j,y_j)}{E_j(x_j)}\right)-E(\boldsymbol{y})+A(\boldsymbol{y})\\
&\quad-\sum_{i=1}^{d}\left[E(\boldsymbol{y})\left(\frac{k_i(x_i,y_i)}{E_i(y_i)}-1\right)-A(\boldsymbol{y})\left(\frac{E_i(x_i)}{\mathcal{E}_i}-1\right)\right.\\
&\quad\left.-\mathcal{E}\left(\frac{k_i(x_i,y_i)}{\mathcal{E}_i}-\frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\right)\right]\\
&= k(\boldsymbol{x},\boldsymbol{y})+\mathcal{E}\cdot\sum_{i=1}^{d}\Big(\frac{k_i(x_i,y_i)}{\mathcal{E}_i}-\frac{E_i(x_i)E_i(y_i)}{\mathcal{E}_i^2}\Big)+\frac{A(\boldsymbol{x})A(\boldsymbol{y})}{\mathcal{E}}\\
&\quad-E(\boldsymbol{x})\cdot\left(1-d+\sum_{j=1}^{d}\frac{k_j(x_j,y_j)}{E_j(x_j)}\right)-E(\boldsymbol{y})\cdot\left(1-d+\sum_{i=1}^{d}\frac{k_i(x_i,y_i)}{E_i(y_i)}\right)
\end{aligned}
$$

References

[AT10] R.J. Adler and J.E. Taylor. *Random Fields and Geometry*. Springer Monographs in Mathematics. Springer, 2010.

[BC56] R. Bellman and RAND CORP SANTA MONICA CALIF. *Dynamic Programming and Lagrange Multipliers*. Defense Technical Information Center, 1956.

[Cre93] N.A.C. Cressie. *Statistics for spatial data*. Wiley series in probability and mathematical statistics: Applied probability and statistics. J. Wiley, 1993.

[DGR12] N. Durrande, D. Ginsbourger, and O. Roustant. Additive covariance kernels for high-dimensional gaussian process modeling. *Annales de la Faculté de Sciences de Toulouse*, Tome 21(3):p. 481–499, 2012.

[DGRC13] N. Durrande, D. Ginsbourger, O. Roustant, and L. Carraro. Anova kernels and rkhs of zero mean functions for model-based sensitivity analysis. *Journal of Multivariate Analysis*, 115(0):57 – 67, 2013.

[DHRL13] N. Durrande, J. Hensman, M. Rattray, and N. D. Lawrence. Gaussian process models for periodicity detection. hal-00805468, 2013.

[DN97] C. R. Dietrich and G. N. Newsam. Fast and exact simulation of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, 18(4):1088–1107, July 1997.

[Doo53] J.L. Doob. *Stochastic processes*. Wiley publications in statistics. Wiley, 1953.

[Fou12] Python Software Foundation. Python (version 2.7.3) [software]., 2012.

[GG09] C. Gaetan and X. Guyon. *Spatial Statistics and Modeling*. Springer Series in Statistics. Springer Verlag, 2009.

[GS11] D. Ginsbourger and D. Schuhmacher. Spatial statistics. Lecture notes from the Spatial Statistics course in fall 2011 at the University of Bern, 2011.

[HFA+13] J. Hensman, N. Fusi, R. Andrade, N. Durrande, A. Saul, M. Zwiessele, and N.D. Lawrence. Gpy (version 0.3.2) - gaussian processes framework in python [software]., 2013.

[HT90] T. Hastie and R. Tibshirani. *Generalized Additive Models*. Monographs on statistics and applied probability. Chapman and Hall, 1990.

[Hun07] J.D. Hunter. Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, 9:90–95, 2007.

[MBC79] M. D. McKay, R. J. Beckman, and W. J. Conover. A comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, 21(2):pp. 239–245, 1979.

[Oli07] T.E. Oliphant. Python for scientific computing. *Computing in Science & Engineering*, 9:10–20, 2007.

[R D08] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, 2008. ISBN 3-900051-07-0.

[Sch09] M. Scheuerer. *A Comparison of Models and Methods for Spatial Interpolation in Statistics and Numerical Analysis*. PhD thesis, University of Göttingen, 2009.

[SS02] B. Schölkopf and A.J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Adaptive computation and machine learning. MIT Press, 2002.

[Ste99] M.L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Series in Statistics. Springer New York, 1999.

[Tre13] C. Tretter. Functional analysis. Lecture notes from the Functional Analysis course in spring 2013 at the University of Bern, 2013.

[TV07] V. Tarieladze and N. Vakhania. Disintegration of gaussian measures and average-case optimal algorithms. *Journal of Complexity*, 23:851 – 866, 2007.

[Wei10] E.W. Weisstein. *CRC Concise Encyclopedia of Mathematics, Second Edition*. Taylor & Francis, 2010.