ulm university universität uulm

Spatial Statistics

Lecture Notes

Prof. Dr. Evgeny Spodarev

Ulm
2024

Contents

Introduction

Generally speaking, the field of *spatial statistics* deals with the statistical inference of random objects that are embedded into the Euclidean space $\mathbb{R}^d$ such as random closed sets, mosaics, geometric point processes and graphs, shapes, fields etc. Here, the space $\mathbb{R}^d$ can as well be understood as a *space-time continuum* if we separate some dimension of $\mathbb{R}^d$ to be a *time line*.

Modern spatial statistics is a huge area which can not be covered in one lecture course. The field's practical applications range from engineering to medicine and social sciences, and are too numerous to mention them all. To give a feeling of how powerful its methods can be, consider just two specific subareas of spatial statistics:

First, the subarea concerned with random sections of random closed sets is called *stereology* with main applications in medicine and biology, in particular pathology analysis. The second subarea, which deals by applications in geosciences, is named *geostatistics*. Here, the main objects of study are *random surfaces*, which might be evolving in time. We refer to these surfaces as *random fields*. These fields may model gas, ore, crude oil or coal deposits in geology, rough surfaces of metals or composites in materials science, regression forecast surfaces in climate research and environmental contexts as well as in georeferenced economics and insurance.

The latter area, dealing with statistical inference of random fields, is the subject of these lecture notes. It will include the estimation of the mean, covariance, spectral density, variogram as well as various prediction methods for (space-time-)random surfaces. In the context of prediction, we will review different regression, Kriging and metric projection-type methods.

1 Basics of random fields

We start with a short overview of the basic notions of the theory of random fields. Let $\mathcal{B}_{\mathbb{R}^d}$ denote the Borel σ-algebra of subsets of $\mathbb{R}^d$, $d \geq 1$. We equip the Euclidean space $\mathbb{R}^d$ with a norm $\|\cdot\|$, e.g. the Euclidean norm $\|x\|_2 = \sqrt{\langle x, x\rangle}$, $x \in \mathbb{R}^2$, where $\langle \cdot, \cdot \rangle$ is the scalar product in $\mathbb{R}^2$. Denote by $\mathbb{R}_+$ the set of non-negative real numbers, i.e. $\mathbb{R}_+ = [0, \infty)$. Furthermore, let $(\Omega, \mathcal{F}, \mathbb{P})$ be an arbitrary probability space.

Definition 1.1 A *random field* $X = \{X(t), t \in T\}$ is a random function on $(\Omega, \mathcal{F}, \mathbb{P})$ indexed by the elements of some subset $T \subset \mathbb{R}^d$, where $d \geq 1$ is an arbitrary integer, i.e. X is a measurable mapping $X : \Omega \times \mathbb{R}^d \to \mathbb{R}$. That is, for all Borel sets $B \in \mathcal{B}_{\mathbb{R}}$ it holds that $X^{-1}(t)(B) = \{\omega \in \Omega : X(t) \in B\} \in \mathcal{F}$ for all $t \in T$.

For an introduction into the theory of random fields see lecture notes “Random Fields“ [43].

1.1 Random Fields with invariance properties

Definition 1.2 A random field $X = \{X(t), t \in T\}$ whose finite-dimensional distributions are invariant with respect to the action of a group G of transformations of T is called *G-invariant in the strict sense.* That is, for all $t_1, \dots, t_n \in T, n \in \mathbb{N}$, and $g \in G$ it holds that

$$(X(gt_1), \dots, X(gt_n)) \overset{d}{=} (X(t_1), \dots, X(t_n)),$$

where $\overset{d}{=}$ denotes equality in distribution and $gt = g(t)$ for all $t \in T$.

In case the invariance is given only for the first two moments of the field, which are assumed to be finite, we speak of *G-invariance in wide sense.*

Definition 1.3 A random field $X = \{X(t), t \in T\}$ is *G-invariant in wide sense* if it is square-integrable, i.e. $\mathbb{E}[X^2(t)] < \infty$ for all $t \in T$, and the *mean value function* $\mu(t) := \mathbb{E}[X(t)]$, $t \in T$, as well as the *covariance function* $C(s,t) := \mathbf{cov}(X(s), X(t))$ $s, t \in T$, satisfy

$$\mu(gt) = \mu(t) \quad \text{and} \quad C(gs, gt) = C(s,t)$$

for all $s, t \in T$ and $g \in G$.

It is easy to see that any random field X which is G-invariant in strict sense, is also G-invariant in wide sense provided that $\mathbb{E}[X^2(t)] < \infty$ for all $t \in T$.

Remark 1.4 Let G be

(a) the *group of translation of* T, i.e., $g(t) = t + h_g$ for some translation vector $h_g \in \mathbb{R}^d$. Then, the G-invariant random field X is called *stationary* (in the respective sense).

(b) the *group of rotations of* T, SO_d. Then, the G-invariant random field X is called *isotropic* (in the respective sense).

(c) the *group of all rigid motions of* T. Then, the G-invariant random field X is called *motion invariant* (in the respective sense). This is equivalent to X being stationary and isotropic.

If X is square integrable, then properties (a)-(c) imply

(a) $\mu(t) \equiv \text{const},\ C(s,t) = C_0(s-t),\ s,t \in T$, where $C_0 : \mathbb{R}^d \to \mathbb{R}$ is a covariance function,

(b) $\mu(t) = \mu(\|t\|),\ C(s,t) = C(\|s\|,\|t\|), s,t \in T$, where $\|\cdot\|$ is a norm in $\mathbb{R}^d$,

(c) $\mu(t) \equiv \text{const},\ C(s,t) = C_1(\|s-t\|),\ s,t \in T$, where $C_1 : \mathbb{R}_+ \to \mathbb{R}$ is a covariance function.

The same notions of G-invariance can be applied to the *increments* $X_h = \{X_h(t) := X(t+h) - X(t), t \in T\},\ h \in T$, of a random field X. In this case, the stationary property is called *intrinsic*. The intrinsic stationarity in the wide sense is called *intrinsic stationarity of order two*. It holds that

$$\mathbb{E}\left[X_h(t)\right] \equiv f(h) \quad \text{and} \quad \mathbb{E}\left[X_h^2(t)\right] = 2\gamma(h)$$

do not depend on $t \in T$. The function γ is called the *variogram*. It is defined by

$$\gamma(h) := \frac{1}{2}\mathbb{E}\left[(X(t+h) - X(t))^2\right]$$

for all $h \in T$.

Exercise 1.5 Show that the mean value function (if it exists) of any process ($d = 1$) with stationary increments is a linear function, i.e., $\mathbb{E}X(t) = at + c$ for all $t \in \mathbb{R}$, where $a \in \mathbb{R}$ and $c \in \mathbb{R}$ are some constants.

Lastly, we say that a random field $X = \{X(t), t \in T\}$ is *centered* if its mean value function $\mu(t)$ exists and $\mu(t) \equiv 0$ for all $t \in T$.

1.2 Elements of correlation theory

Let $X = \{X(t), t \in T\},\ T \subset \mathbb{R}^d$ be a square-integrable random field which is wide-sense stationary with covariance function $C(s-t) = \mathbf{cov}(X(s), X(t)),\ s,t \in T$. Then, the covariance function C is positive semi-definite, which is a consequence of the following result.

Proposition 1.6 A function $G : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a covariance function of some square-integrable random field if and only if it is positive semi-definite.

Exercise 1.7 Prove Proposition 1.6.
Hint: Calculate the variance of the linear combination $\sum_{i=1}^n x_i X(t_i)$ *for arbitrary* $n \in \mathbb{N},\ t_i \in \mathbb{R}^d,\ x_i \in \mathbb{R}$.

An important aspect of correlation theory is the so-called *spectral representation* of X. By the Bochner-Kchinchin theorem, see lecture notes “Random Fields“ [43, Theorem 2.1.1], any positive semi-definite function $f : \mathbb{R}^d \to \mathbb{R}$, which is continuous at the origin, is a Fourier transform of some symmetric finite measure μ_f on $\mathbb{R}^d$. Thus, for a wide-sense stationary and mean-square continuous field X we have

$$\mathbf{cov}(X(s)), X(t)) = C(s-t) = \int_{\mathbb{R}^d} e^{i\langle x, s-t\rangle} \mu_C(dx).$$

Here, $\langle \cdot, \cdot \rangle$ denotes the Euclidean scalar product in $\mathbb{R}^d$. The finite measure μ_C on $\mathcal{B}_{\mathbb{R}^d}$ is called a *spectral measure* of X. If μ_C is absolutely continuous with respect to the Lebesque measure, then its density f_C is called a *spectral density*, i.e. $\mu_C(dx) = f_C(x)dx$.

It holds that the spectral density $f_C : \mathbb{R}^d \to \mathbb{R}_+$ is integrable on $\mathbb{R}^d$, since

$$\int_{\mathbb{R}^d} f_C(x)dx = \int_{\mathbb{R}^d} \mu_C(dx) = C(0) = \mathbf{var}(X(t)) < \infty.$$

Together with the covariance function C, the spectral density measures the stochastic dependence between $X(s)$ and $X(t)$ for $s,t \in T$. In particular, the behaviour of f_C at the origin $0 \in \mathbb{R}^d$ (e.g. whether $f_C(0) < \infty$ or $f_C(x) \uparrow \infty,\ x \to 0$) can tell whether the field X has *short memory*, i.e.

$$\int_{\mathbb{R}^d} |C(x)|dx < \infty$$

or *long memory*, i.e.

$$\int_{\mathbb{R}^d} |C(x)|dx = \infty.$$

Sometimes, the terms *short* and *long range dependence* are used, instead.

1.3 Examples of random fields

1.3.1 Boolean random fields

Let $\{X_l(t), t \in \mathbb{R}^d\}_{l\in\mathbb{R}}$ be a family of stochastically independent, a.s. continuous random functions with subgraphs having a.s. compact sections. Furthermore, let $\Pi = \{(Y_i, T_i)\}_{i=1}^{\infty}$ be a Poisson point process in $\mathbb{R}^d \times \mathbb{R}$ with intensity measure $v_d \otimes \theta$, where v_d denotes the Lebesgue measure on $\mathbb{R}^d$ and θ is a σ-finite measure on $\mathbb{R}$. The random function $X = \{X(t), t \in \mathbb{R}^d\}$ with

$$X(t) = \sup_{(Y_k,T_k)\in\Pi} X_{T_k}(t - Y_k), \quad t \in \mathbb{R}^d,$$

is called a *Boolean random field*. The functions X_l are referred to as *primary functions*. Boolean random fields are often used for modeling purposes in geo- and materials science.

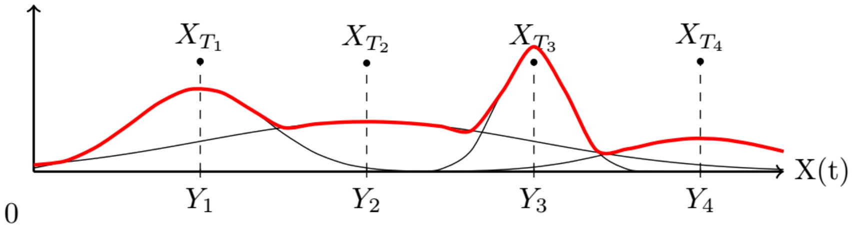


Fig. 1.1: $X(t)$ at “simulation time“ step T_4.

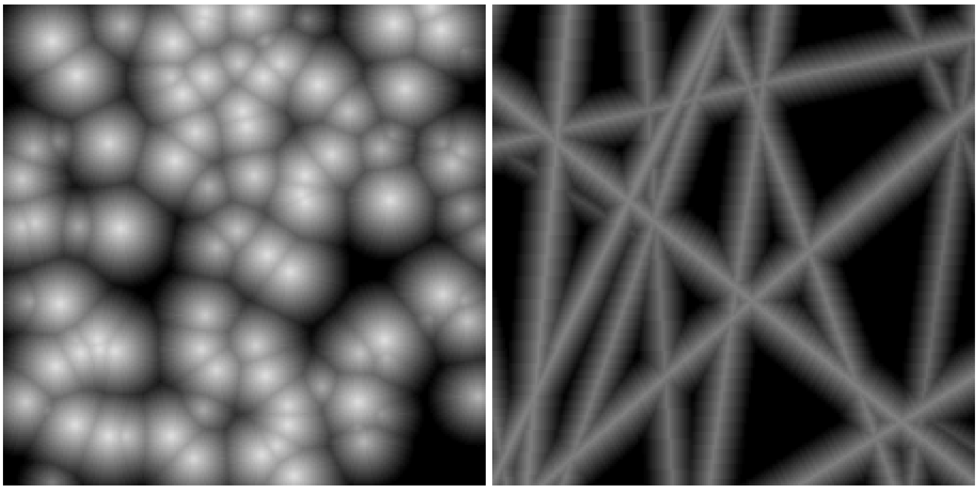


Fig. 1.2: Examples of realizations of a Boolean random function with cone primary functions (left) and built from Poisson lines (right) [38, p. 149].

1.3.2 Gaussian random fields

A widely known and one of the most important class of random fields is the class of Gaussian fields.

Definition 1.8 A random field $X = \{X(t), t \in \mathbb{R}^d\}$ is called *Gaussian* if its finite-dimensional distributions are Gaussian.

The popularity of Gaussian fields for modeling purposes in applications can be explained mainly by the simplicity of their construction and analytic tractability combined with the normal distributions of marginals, which describe many real phenomena due to the central limit theorem.

By Kolmogorov's theorem, the probability law (or distribution) of a Gaussian random field is uniquely defined by its mean value and covariance function, see lecture "Random fields" [43, Theorem 1.1.2.].

Exercise 1.9 Show that for Gaussian random fields stationarity (isotropy, motion invariance) in the strict sense and stationarity (isotropy, motion invariance) in the wide sense are equivalent. In this case we call a Gaussian field just *stationary* (*isotropic*, *motion invariant*)

In the following, consider two particular cases of Gaussian random fields:

(a) An *Ornstein-Uhlenbeck random field* is a centered, i.e. $\mathbb{E}[X(t)] = 0$ for all $t \in \mathbb{R}$, Gaussian random field $X = \{X(t), t \in \mathbb{R}\}$ with covariance function

$$\mathbb{E}\left[X(s)X(t)\right] = \exp\left\{-|s-t|/2\right\}, \quad s, t \in \mathbb{R}^d.$$

It is clearly stationary and isotropic, hence motion invariant.

(b) A *fractional Brownian field* $X = \{X(t), t \in \mathbb{R}^d\}$ is a centered Gaussian field with covariance function

$$\mathbb{E}\left[X(s)X(t)\right] = \frac{1}{2}\|s\|^{2H} + \|t\|^{2H} - \|s-t\|^{2H}, \quad s, t \in \mathbb{R}^d,$$

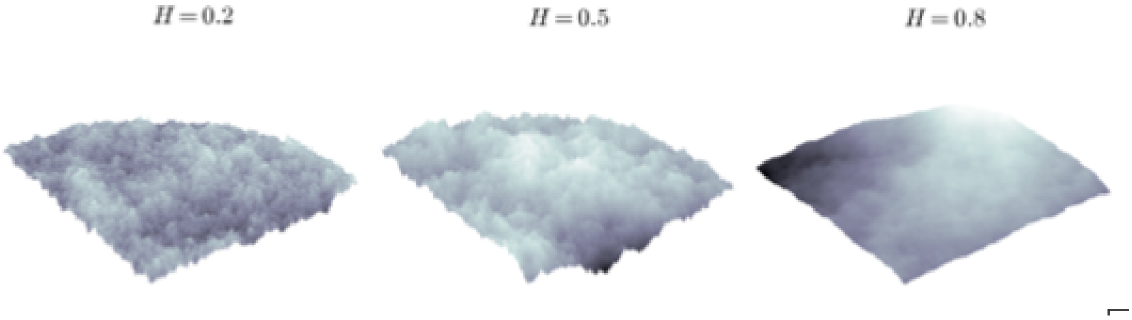


Fig. 1.3: Fractional Brownian fields with different roughness parameter H [38, p.398].

for some $H \in (0,1]$, where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^d$. The process is often denoted by X^H to emphasize its dependence on H.

The parameter H, often referred to as *Hurst index*, is responsible for the regularity of the paths of X, i.e. the greater H the smoother the paths of X. In the one-dimensional case the process X is called the *fractional Brownian motion*, including the *two-sided Wiener process* (defined on the whole real line $\mathbb{R}$) where $H = \frac{1}{2}$. In the case $d > 1$, X is called the *Brownian Lévy field* if $H = \frac{1}{2}$.

It is easy to check that X is intrinsically stationary of order two and isotropic. Its variogram $\gamma(h) = \frac{1}{2} \cdot \|h\|^{2H}$ is clearly motion invariant. However, this field is not wide-sense stationary as its variance is not constant.

Exercise 1.10 Show that a fractional Brownian field X

(a) has stationary increments, which are positively correlated for $H \in (\frac{1}{2}, 1)$ and negatively correlated for $H \in (0, \frac{1}{2})$,

(b) is H-self-similar, i.e. $X(\lambda t) \overset{d}{=} |\lambda|^H X(t)$ for all $\lambda \in \mathbb{R}$ and $t \in \mathbb{R}^d$,

(c) has a version with a.s. Hölder-continuous paths of any order $\beta \in (0, H)$,

(d) has nowhere differentiable paths for any $H \in (0,1)$,

(e) is a linear process for $d = H = 1$, i.e. $X(t) \overset{d}{=} tZ$, for all $t \in \mathbb{R}$ and some random variable $Z \sim N(0,1)$.

1.3.3 Max-stable random fields

Max-stable random fields are used to model extremal events such as highest floods, maximal temperature and precipitation, etc.

Definition 1.11 A random field $X = \{X(t), t \in T\}$ is called *max-stable* if for any $n \in \mathbb{N}$ there exist constants $\alpha_n > 0$, $\beta_n \in \mathbb{R}$ such that

$$X \overset{d}{=} \left\{ \frac{\max\limits_{j=1,\dots,n} X_j(t) - \beta_n}{\alpha_n}, t \in T \right\},$$

where $\{X_j(t), t \in T\}$ are i.i.d. copies of X. In particular, if $T = \{t_0\}$, $t_0 \in \mathbb{R}^d$, then X is called a *max-stable random variable.* On the other hand, if $T = \{t_1, \dots, t_n\} \subset \mathbb{R}^d$, then X is a *max-stable random vector.*

By the theorem of *Fisher-Tippett-Gnedenko* [7] any max-stable random variable (and hence any marginal distribution of a max-stable random field X) has one of the three possible *extreme value distributions*: *Weibull*, *Gumbel* or *Fréchet*.

Among these distributions, only the Fréchet distribution does not have a finite variance. We say that a random variable Y is *Fréchet-distributed* is its cumulative distribution function is given by

$$\mathbb{P}(Y \leq x) = \exp\left\{ -\left(\frac{x-\mu}{\sigma}\right)^{-\alpha} \right\}, \quad x > \mu.$$

We use the notation $Y \sim$ *Fréchet*(α, μ, σ), and for *Fréchet*$(\alpha, 0, 1) =$ *Fréchet*(α). Fréchet(1) is called *standard Fréchet distribution.*

Exercise 1.12 Show that for $Y_1 \sim$ Fréchet(α) it holds that $\mathbb{E}|[Y_1|] < \infty$ if and only if $\alpha > 1$. Moreover, $\mathbb{E}[|Y_1|^k] < \infty$ holds if and only if $\alpha > k$, $k \in \mathbb{N}$. For $\alpha > 1$, check that $\mathbb{E}[Y_1] = \Gamma(1 - \frac{1}{\alpha})$.

Every max-stable random field X can be suitably transformed via a transform Ψ such that the marginals of $\Psi(X)$ have either Weibull, Fréchet or Gumbel distribution. A suitable transformation of $X(t)$ can be chosen as follows. For the cumulative distribution function $F_{X(t)} = \mathbb{P}(X(t) \leq x)$, $x \in \mathbb{R}$, define

$$\Psi(x) = -\frac{1}{\log F_{X(t)}(x)}, \quad x \in \mathbb{R}.$$

Then, it is not difficult to see that $Y(t) = \Psi(X(t)) \sim$ Fréchet(1). Consequently, it is sufficient to consider random fields with marginals standardized to Fréchet(1).

Exercise 1.13 Prove the above.

The stochastic dependence in max-stable random vectors is described by the so-called *tail dependence function.* Consider the $(n-1)$-*dimensional unit simplex*

$$S_n = \{(x_1, ..., x_n) = x \in \mathbb{R}^n_+ : \sum_{j=1}^n x_j = 1\}$$

for $n \geq 2$. The tail dependence function l_n is introduced in the following result.

Theorem 1.14 Let $Y = (Y_1, ..., Y_n)$ be a max-stable random vector. The following are equivalent.

(a) $Y_j \sim$ Fréchet(α) for some $\alpha > 0$, $j = 1, \dots, n$.

(b) There exists a function $l_n : \mathbb{R}_+ \to \mathbb{R}_+$ such that

$$\mathbb{P}\left(Y_1 \leq x_1, \dots, Y_n \leq x_n\right) = \exp\left\{ -l_n(x_1^{-\alpha}, \dots, x_n^{-\alpha}) \right\}, \quad x_1, \dots, x_n > 0.$$

where

$$l_n(x_1, \dots, x_n) = \int_{S_n} \max_{j=1,\dots,n} \{x_j q_j\} \; d\mu(q_1, \dots, q_n).$$

The finite measure μ on S_n satisfies the constraint

$$\int_{S_n} q_j \, d\mu(q_1, \ldots, q_n) = 1, \quad j = 1, \ldots, n,$$

see [34] for a proof. The function l_n is called *tail dependence function* of the vector Y.

Theorem 1.15 The tail dependence function l_n

(a) is convex,

(b) is *homogeneous of order 1*, i.e. $l_n(\lambda x_1, \ldots, \lambda x_n) = \lambda \cdot l_n(x_1, \ldots, x_n)$ for all $\lambda > 0$ and $x_1, \ldots, x_n > 0$,

(c) satisfies

$$\|x\|_{\max} \le l_n(x_1, \ldots, x_n) \le \sum_{j=1}^{n} x_j = \|x\|_1$$

for all vectors $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$, where $\|x\|_{\max} = \max_{j=1\ldots n} |x_j|$ is the maximum norm of x and $\|x\|_1 = \sum_{j=1}^{n} |x_j|$ the l_1-norm.

Exercise 1.16 Prove Theorem 1.15.

Exercise 1.17 Let l_n be the tail-dependence function of a max-stable random vector $Y = (Y_1, \ldots, Y_n)$. Show:

(a) $l_n(x) = \|x\|_{max}$ if and only if $Y_1 = \cdots = Y_n$ a.s., i.e. *complete dependence* holds.

(b) $l_n(x) = \|x\|_1$ if and only if $Y_1, \ldots, Y_n$ are *stochastically independent.*

Recall that the Fréchet distribution has infinite variance. For $n = 2$, the quantity $\theta = l_2(1, 1)$ is called *(pairwise) extremal coefficient*. It serves as an analogue to the covariance for heavy-tailed random variables $Y_1, Y_2 \sim$ Fréchet.

Example 1.18 Let us now give some examples of max-stable random fields.

(i) The *Brown-Resnick random field* is defined as follows. Let $Y = \{Y(t), t \in T\}$ be a centered Gaussian random field with stationary increments and $\sigma^2(t) := \mathbf{var}(Y(t))$, $t \in T$. Let $\Pi = \{\zeta_j\}$ be a Poisson process on $\mathbb{R}$, independent of Y, with intensity measure $\Delta(dx) = e^{-x} dx$. Then, the *Brown-Resnick random field* is given by

$$R(t) := \max_{j \in \mathbb{N}} \left\{ \zeta_j + Y_j(t) - \frac{\sigma^2(t)}{2} \right\}, \quad t \in T,$$

where $\rho_j = \{Y_j(t),\ t \in T\}$ are independent copies of Y. It is stationary and has standard *Gumbel* margins $\mathbb{P}(R(t) \le x) = \exp(-e^{-x})$, $x \in \mathbb{R}$.

Exercise 1.19 Show that the transformation $B(t) := e^{R(t)}$ has Fréchet(1) margins.

The finite-dimensional distributions of $\{R(t),\ t \in T\}$ as well as the tail dependence function l_n are given in the following.

Theorem 1.20 For any $t_1, \ldots, t_n \in T$ it holds that

(a) $\mathbb{P}(R(T_1) \leq y_1, \dots R(t_n) \leq y_n) = \exp\left\{-\mathbb{E}\left[\exp\left\{\max_{j=1,\dots,n}(Y(t_j) - \frac{\sigma^2(t_j}{2} - Y_j)\right\}\right]\right\}.$

(b) the tail dependence function l_n of $(B(t_1), \dots, B(t_n))$ is given by

$$l_n(x_1, \dots, x_n) = \int_0^{+\infty} \left[1 - F_{\Sigma_n}\left(\log\left(\frac{y}{x_1}\right) + \frac{\sigma^2(t_1)}{2}, \dots, \log\left(\frac{y}{x_n}\right) + \frac{\sigma^2(t_n)}{2}\right)\right] dy,$$

where Σ_n is the covariance matrix of $(Y(t_1), \dots, Y(t_n))$, $x_1, \dots, x_n > 0$, and F_{Σ_n} is the multivariate cumulative distribution function of $N(0, \Sigma_n)$.

(ii) The *Smith random field* $S = \{S(t), t \in T\}$ is defined by

$$S(t) := \max_{j \in \mathbb{N}} \{\zeta_j f_\Sigma(t - \epsilon_j)\}, \quad t \in T \subset \mathbb{R}^d,$$

where Σ is a positive definite matrix with dimensions $d \times d$, f_Σ is the probability density function of $N(0, \Sigma)$ and $\tilde{\Pi} = \{(\zeta_j, \varepsilon_j)\}_{j \in \mathbb{N}}$ is a Poisson point process on $(0, \infty) \times \mathbb{R}^d$ with intensity measure $\tilde{\Lambda}(dx, dy) = x^{-2} dx dy$.

Similarly to Theorem 1.20, we may formulate:

Theorem 1.21 For any $t_1, \dots, t_n \in T$ it holds that

(a) $S(t_j) \sim$ Fréchet(1),

(b) $\mathbb{P}(S(t_1) \leq y_1, \dots, S(t_n) \leq y_n) = \exp\left\{-\int_{\mathbb{R}^d} \max_{j=1,\dots,n} \frac{f_\Sigma(t_j - s)}{y_j} ds\right\}$, $y_1, \dots, y_n > 0$,

(c) the tail dependence function l_n of $(S(t_1), \dots, S(t_n))$ is given by

$$l_n(x_1, \dots, x_n) = \int_{\mathbb{R}^d} \max_{j=1,\dots,n} \{x_j f_\Sigma(t_j - s)\} \, ds.$$

(iii) The *extremal Gaussian random field* $G = \{G(t),\ t \in T\}$ is given by

$$G(t) := \max_{j \in \mathbb{N}} \zeta_j(\max\{Y_j(t), 0\}), \quad t \in T \subset \mathbb{R}^d,$$

where $\{Y_j(t),\ t \in T\}$ are independent copies of a stationary centered Gaussian random field $Y = \{Y(t),\ t \in T\}$, and $\tilde{\tilde{\Pi}} = \{\zeta_j\}_{j \in \mathbb{N}}$ is an independent Poisson process on $\mathbb{R}_+$ with intensity measure $\tilde{\tilde{\Lambda}}(dx) = \sqrt{2\pi} x^{-2} dx$.

Theorem 1.22 For any $t_1, \dots, t_n \in T$ it holds that

a) $G(t_j) \sim$ Fréchet(1),

b) $\mathbb{P}(G(t_1) \leq y_1, \dots, G(t_n) \leq y_n) = \exp\left\{-\mathbb{E}\left[\max_{j=1,\dots,n} \frac{f_\Sigma(t_j - s)}{y_j}\right]\right\}$, $y_1, \dots, y_n > 0$,

c) the tail dependence function l_n of $(G(t_1), \dots, G(t_n))$ is given by

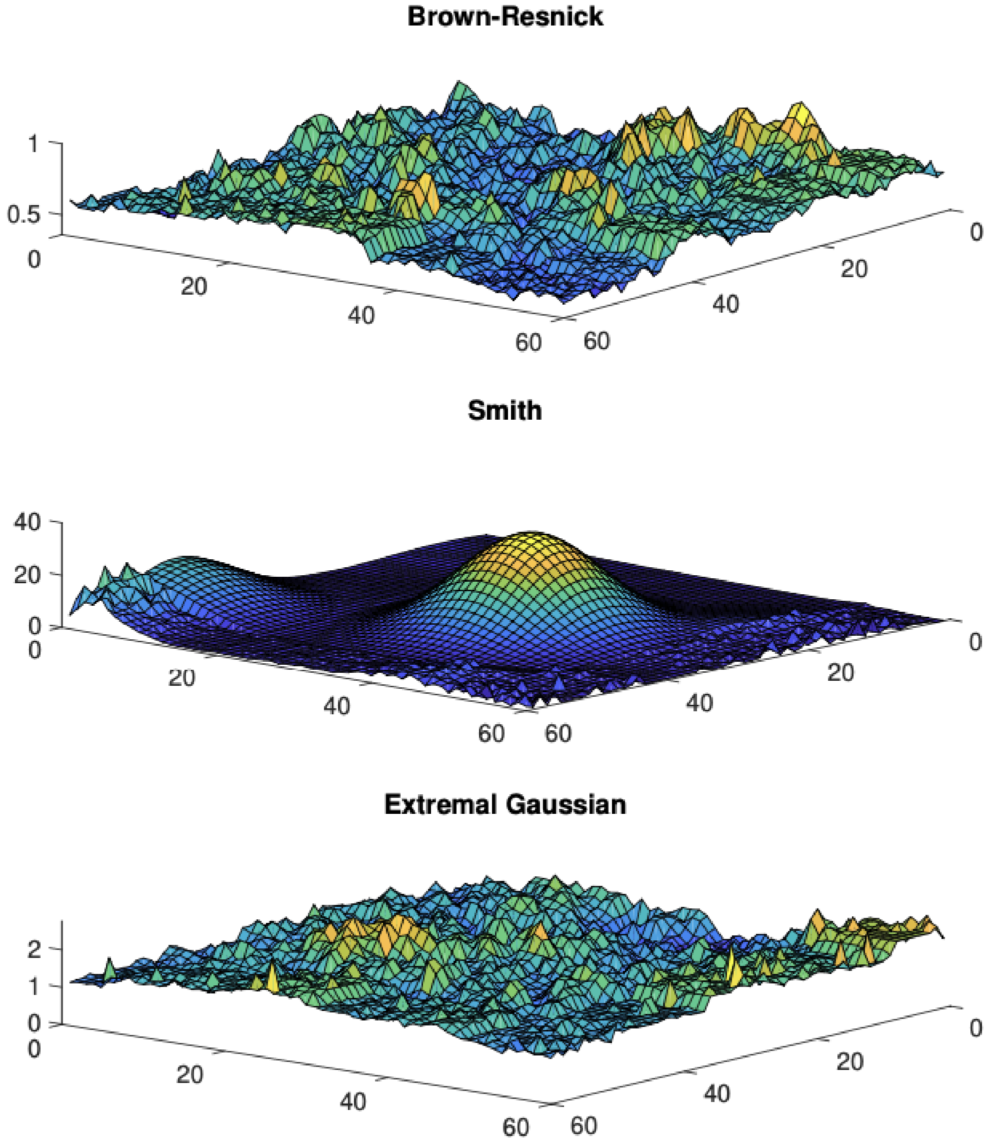
$$l_n(x_1, \dots, x_n) = \int_0^\infty \left[1 - F_{\sigma_n}\left(\frac{y}{x_1}, \dots, \frac{y}{x_n}\right)\right] dy, \quad x_1, \dots, x_n > 0,$$

where F_{σ_n} is chosen as in Theorem 1.20.

Exercise 1.23 Show that for $n = 2$, the tail dependence function in Theorem 1.22 (c) simplifies to

$$l_2(x_1, x_2) = \frac{1}{2}\left(x_1 + x_2 + \sqrt{x_1^2 - \rho x_1 x_2 + x_2^2}\right),$$

where $\rho = \mathbf{corr}(Y(t_1), Y(t_2))$.



Brown-Resnick

Smith

Extremal Gaussian

Fig. 1.4: Simulated realizations of Brown-Resnick, Smith and extremal Gaussian random fields in $d = 2$ [9].

1.3.4 α-stable random fields

The class of α-stable random fields models heavy-tailed phenomena, where the variance (and possibly even the mean) is infinite. Such events are common in insurance and finance, where high volatility and dangerous risks in extreme scenarios are common. First, we introduce α-stable random vectors.

Definition 1.24 A random vector $Y = (Y_1, \dots, Y_n) : \Omega \mapsto \mathbb{R}^n$ is called *stable* if for all $m \geq 2$ there exist $c = c(m) > 0$, $k = k(m) \in \mathbb{R}^n$ such that

$$Y^{(1)} + Y^{(2)} + \dots + Y^{(m)} \overset{d}{=} cY + k,$$

where $\{Y^{(j)}\}_{j=1}^m$ are independent copies of Y.

It can be shown that $c = m^{1/\alpha}$, where $\alpha \in (0, 2]$ is called *stability index* or *index of stability* [37, Theorem 2.1.2].

Exercise 1.25 Show that for $k = k(m)$ from Definition 1.24 it holds that $k(m) = \mu(m - m^{1/\alpha})$, where μ is the shift parameter of $S_\alpha(\mu, \Gamma)$.
Hint: First, show that $\sum_{j=1}^m Y^{(j)} \sim S_\alpha(m\mu, m\Gamma)$ *and* $m^{1/\alpha}Y + k(m) \sim S_\alpha(m^{1/\alpha}\mu + k(m), m\Gamma)$.

An equivalent definition of stable random vectors is given in terms of their characteristic function.

Definition 1.26 A random vector $Y = (Y_1, \dots, Y_n) : \Omega \mapsto \mathbb{R}^n$ is called *stable* if its characteristic function $\varphi_Y(s) = \mathbb{E}\left[e^{i\langle Y, s\rangle}\right]$, $s \in \mathbb{R}^n$, is of the form

$$\varphi_Y(s) = \begin{cases} \exp\left\{-\int_{S^{n-1}} |\langle s, x\rangle|^\alpha \left(1 - i \cdot \mathrm{sgn}(\langle s, x\rangle) \tan\left(\frac{\pi\alpha}{2}\right)\right) \Gamma(dx) + i\langle s, \mu\rangle\right\}, & \alpha \neq 1, \\ \exp\left\{-\int_{S^{n-1}} |\langle s, x\rangle| \left(1 + i\frac{2}{\pi}\mathrm{sgn}(\langle s, x\rangle) \log|\langle s, x\rangle|\right) \Gamma(dx) + i\langle s, \mu\rangle\right\}, & \alpha = 1, \end{cases}$$

where $\Gamma(\cdot)$ is a finite measure on the unit sphere $S^{n-1} \subset \mathbb{R}^n$ and $\mu \in \mathbb{R}^n$.

For $\alpha \in (0, 2)$ the pair (μ, Γ) yields a unique parametrization of the distribution of a stable random vector Y. We say Y is *α-stable* with *shift parameter* μ and *spectral measure* Γ and write $Y \sim S_\alpha(\mu, \Gamma)$. The spectral measure Γ contains all information about the interdependence of the coordinates Y_j, $j = 1, \dots, n$.

In the case $\alpha = 2$, the characteristic function in Definition 1.26 defines a Gaussian random vector Y with

$$\varphi_Y(s) = \mathbb{E}\left[e^{i\langle s, Y\rangle}\right] = \exp\left\{i\langle s, \mu\rangle - \frac{1}{2}s^T\Sigma s\right\}, \quad s \in \mathbb{R}^n,$$

with a positive semi-definite $(n \times n)$- covariance matrix $\Sigma = (\sigma_{jk})_{j,k=1}^n$, $\sigma_{jk} = \mathbf{cov}(Y_j, Y_k)$ and $\mu = \mathbb{E}[Y] \in \mathbb{R}^n$. In the Gaussian case, the spectral measure $\Gamma(\cdot)$ is not unique, see Exercise 1.27 below.

Exercise 1.27 Consider the measures Γ_1, Γ_2 on S^0 with

$$\begin{aligned} \Gamma_1(dx) &= \delta_{\{1/\sqrt{2}, 1/\sqrt{2}\}}(dx) + \delta_{\{-1/\sqrt{2}, -1/\sqrt{2}\}}(dx), \\ \Gamma_2(dx) &= 2\delta_{\{1/\sqrt{2}, 1/\sqrt{2}\}}(dx). \end{aligned}$$

Show that Γ_1, Γ_2 yield the same characteristic function of a Gaussian random vector $Y = (Y_1, Y_2)$ with mean $\mu \in \mathbb{R}^2$ and covariance matrix $\Sigma = \begin{pmatrix} 2 & 2 \\ 2 & 2 \end{pmatrix}$.

A random vector Y is called *symmetric* if $Y \overset{d}{=} -Y$. For symmetric α-stable random vectors, we will use the notation $Y \sim S\alpha S$.

Lemma 1.28 A random vector $Y \sim S_\alpha(\mu, \Gamma)$ is $S\alpha S$ if and only if $\mu = 0$ and Γ is symmetric on S^{n-1}.

The proof of the above lemma is given in [37, Theorem 2.4.3]. For $n = 1$, Definition 1.24 yields an α-stable random variable Y. We will reparametrize its distribution using 4 parameters α, σ, β and μ, as seen in the following representation of the characteristic function of Y. For all $s \in \mathbb{R}$ we have

$$\varphi_Y(s) = \mathbb{E}\left[e^{isY}\right] = \begin{cases} \exp\left\{-\sigma^\alpha |s|^\alpha (1 - i\beta \cdot \operatorname{sgn}(s)\tan(\frac{\pi\alpha}{2})) + i\mu s\right\}, & \alpha \neq 1, \\ \exp\left\{-\sigma\ |s|\ (1 + i\beta \cdot \frac{2}{\pi}\operatorname{sgn}(s)\log(|s|) + i\mu s\right\}, & \alpha = 1. \end{cases}$$

In the univariate case, we use the notation $Y \sim S_\alpha(\sigma, \beta, \mu)$.

Exercise 1.29 Show that the spectral measure Γ of $Y \sim S_\alpha(\sigma, \beta, \mu)$ is given by

$$\Gamma(dx) = \frac{\sigma^\alpha}{2}(1 + \beta)\delta_{\{1\}}(dx) + \frac{\sigma^\alpha}{2}(1 - \beta)\delta_{\{-1\}}(ds)$$

such that

$$\sigma^\alpha = \Gamma(\{1\}) + \Gamma(\{-1\}) \quad \text{and} \quad \beta = \frac{\Gamma(\{1\}) - \Gamma(\{-1\})}{\Gamma(\{1\}) + \Gamma(\{-1\})}$$

Stable distributions are absolutely continuous. However, their densities are not known explicitly except for the cases $\alpha = \frac{1}{2}$ (*Lévy distribution*), $\alpha = 1$ (*Cauchy distribution*), and $\alpha = 2$ (*normal distribution*).

Exercise 1.30 Show:

(a) For $Y \sim S\alpha S$ it holds that $Y \sim S_\alpha(\sigma, 0, 0)$ with characteristic function

$$\varphi_Y(s) = e^{\sigma^\alpha |s|^\alpha}, \quad s \in \mathbb{R}.$$

(b) For $Y \sim S_2(\sigma, 0, \mu)$ it holds that $Y \sim N(\mu, 2\sigma^2)$.

Theorem 1.31 S_α is heavy-tailed, i.e. for $Y \sim S_\alpha(\sigma, \beta, \mu)$ with $\alpha \in (0, 2)$

$$\mathbb{P}(|Y| > x) \sim \frac{c}{x^\alpha}$$

as $x \to \infty$ holds for some $c > 0$.

Consequently, we have

$$\mathbb{E}|Y|^p = \int_0^\infty \mathbb{P}(|Y| > x^{1/p})\, dx \leq c_1 \cdot \int_0^\infty x^{-\alpha/p}\, dx \begin{cases} < \infty, & p \in (0, \alpha) \\ = \infty, & p \geq \alpha. \end{cases}$$

In contrast to this, the normal distribution shows the following *short tail behavior*. For $X \sim N(0, 1)$ it holds that

$$\mathbb{P}(X < -x) = \mathbb{P}(X > x) \sim \frac{1}{\sqrt{2\pi}x} e^{-x^2/s}$$

as $x \to \infty$ and all moments are therefore finite.

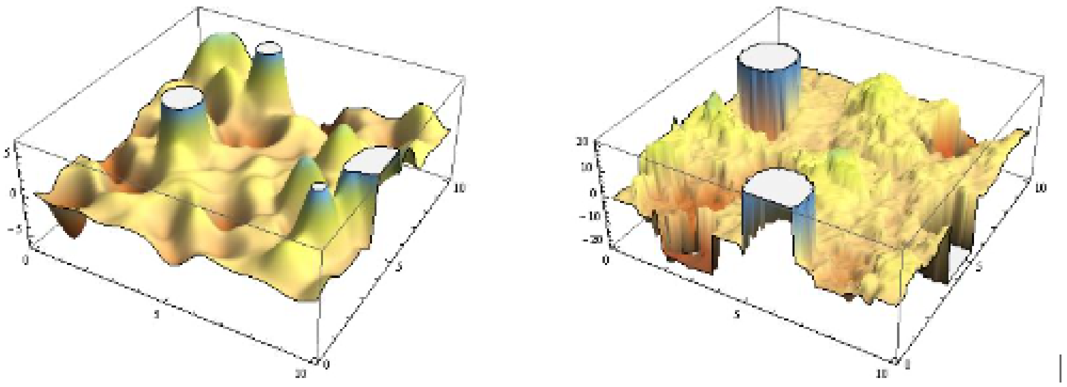


Fig. 1.5: Realizations of $S_{0.8}S$- moving averages with different kernels [38, p.344].

Definition 1.32 A non-Gaussian random field $X = \{X(t), t \in T\}$ is called *α-stable* with index of stability $\alpha \in (0,2)$ if all its finite-dimensional distributions are α-stable (in the sense of Definition 1.24).

An important example for an α-stable random field is given by $X = \{X(t) : t \in T\}$ with $X(t) \overset{d}{:=} \sqrt{A} \cdot Y(t)$, where A and Y independent with

$$A \sim S_{\alpha/2}((\cos(\pi\alpha/4))^{2/\alpha}, 1, 0)$$

and $Y = \{Y(t), t \in T\}$ being a centered Gaussian random field with a positive definite covariance function. The random field X is called *subgaussian*. Its α-stable distribution follows from Exercise 1.33.

Exercise 1.33 Show that for a stable random variable A as above and $Y_0 \sim N(2\sigma^2)$, with A and Y_0 independent, it holds that $X_0 := \sqrt{A} \cdot Y_0 \sim S_\alpha(\sigma, 0, 0)$.
Hint: Compute the conditional characteristic function $\mathbb{E}\left[e^{isX_0} \mid A\right]$.

If the underlying Gaussian random field Y of a subgaussian random field X is stationary, then X is *stationary in the strict sense*.

2 Elementary statistical inference for square-integrable random fields

Let $X = \{X(t), t \in T\}$, $T \subset \mathbb{R}^d$ be a wide-sense stationary, measurable random field with mean $\mu = \mathbb{E}[X(0)]$, covariance function $C(t) = \mathbf{cov}(X(0), X(t))$ and spectral density $f_C(t),\ t \in T$. In this chapter, we describe some non-parametric statistical inference methods for the estimation of μ, C and f_C. We also analyze their asymptotic properties as the number of observations grows to infinity.

We assume that *one single realisation* of X is available and can be observed on *an observation window* $W \subset T$, which is assumed to be non-empty bounded Borel subset of $\mathbb{R}^d$. The values $\{X(t), t \in W\}$ will be called *observations* of X. Sometimes, we will assume $W = \{t_1, \dots, t_n\}$ to be a finite set. For asymptotic inference, we will consider a sequence of observation windows $\{W_n\} \subset T$ growing in the *Van Hove-sense*, i.e.

$$\lim_{n\to\infty} |W_n| = \infty \tag{2.1}$$

and

$$\lim_{n\to\infty} \frac{|\partial W_n \oplus B_r(0)|}{|W_n|} = 0 \text{ for any } r > 0, \tag{2.2}$$

where $|\cdot|$ is the *volume* (d-dimensional Lebesgue measure) in $\mathbb{R}^d$, ∂W is the boundary of W, $B_r(0) = \{x \in \mathbb{R}^d : \|x\|_2 \leq r\}$ is the spherical neighbourhood of the origin with radius r. Additionally, the $\oplus$-operation is the so-called *Minkowski addition* of two sets being their pointwise sum, i.e. $A \oplus B = \{x + y : \ x \in A, y \in B\}$ for $A, B \subset \mathbb{R}^d$.

Requirement (2.1) is understood in the sense that the growth of W_n is unbounded as $n \to \infty$, whereas (2.2) indicates that the boundary of W_n can be neglected in the subsequent asymptotic analysis.

2.1 Estimation of the mean

We consider the estimator

$$\hat{\mu}_n := \int_{W_n} X(t) G(W_n, t) dt, \quad n \in \mathbb{N},$$

of μ, where $G : B_{\mathbb{R}^d} \times T \mapsto \mathbb{R}_+$ is a *weight functional* satisfying $G(W, t) = 0,\ t \in T \backslash W$, and

$$\int_T G(W, t) dt = 1 \tag{2.3}$$

for any bounded Borel window $W \in \mathcal{B}_{\mathbb{R}^d}$.

Example 2.1 (Uniform weights): The simplest weight functional is uniform on W, i.e. G is given by

$$G(W,t) = \frac{I(t \in W)}{|W|}, \quad t \in T.$$

Then, the estimator $\hat{\mu}_n$ becomes

$$\hat{\mu}_n = \frac{1}{|W_n|} \int_{W_n} X(t)dt,$$

which is the well-known sample mean in statistics. Its practical interpretation is based on the discretization of the integral. If the only observed sample of X at the points $t_1, \dots, t_N \in W_n$ is $(X(t_1,) \dots, X(t_N))$, then $\hat{\mu}_n \approx \frac{1}{N} \sum_{j=1}^{N} X(t_j)$.

Lemma 2.2 (Unbiasedness): The estimator $\hat{\mu}_n$ is unbiased, i.e. $\mathbb{E}[\hat{\mu}_n] = \mu$.

Proof Applying Fubini's theorem as well as the stationarity of X and Equation (2.3) yields

$$E\left[\hat{\mu}_n\right] = \mathbb{E}\left[\int\limits_{W_n} X(t)G(W_n,t)dt\right] \overset{\text{Fubini}}{=} \int\limits_{W_n} \underbrace{\mathbb{E}\left[X(t)\right]}_{=\mu} G(W_n,t)dt = \mu \cdot \underbrace{\int\limits_{W_n} G(W_n,t)dt}_{=1} = \mu.$$

□

Consider the function $\Gamma_n(t) := \int\limits_T G(W_n,t)G(W_n,y+t)dy$ for $t \in T$. Note that $\Gamma_n(t) = 0$ if $t \notin W_n \oplus \check{W}_n$, where $\check{K} := -K$ for any set K.

Lemma 2.3 For any $n \in \mathbb{N}$, it holds that

$$\mathbf{var}(\hat{\mu}_n) = \int_T C(t) \cdot \Gamma_n(t)dt.$$

Exercise 2.4 Proof Lemma 2.3.

Next, we examine the asymptotic behaviour of $\hat{\mu}_n$. We take a look at its *mean-square consistency* and *asymptotic normality.*

Theorem 2.5 (L^2-consistency): For a Van Hove sequence $\{W_n\}_{n\in\mathbb{N}}$ of observation windows, assume that there exist constants $c_0,\ \theta > 0$ such that

$$\sup_{t\in T} G(W_n,t) \leq \frac{c_0}{|W_n|}, \ n \in \mathbb{N}, \quad \text{and} \quad \lim_{n\to\infty} |W_n| \cdot \Gamma_n(t) = \theta.$$

If the covariance function C is integrable over T, i.e. $\int_T |C(t)|dt < \infty$, then

$$\lim_{n\to\infty} |W_n| \mathbf{var}(\hat{\mu}_n) = \theta \int_T C(t)dt,$$

and consequently $\mathbb{E}\left[|\hat{\mu}_n - \mu|^2\right] \to 0$ as $n \to \infty$.

Proof The unbiasedness of $\hat{\mu}$ in Lemma 2.2 and Lebesgue's dominated convergence theorem yield

$$\lim_{n\to\infty} |W_n| \mathbf{var}(\hat{\mu}) = \lim_{n\to\infty} |W_n| \int_T C(t)\Gamma_n(t)dt = \int_T C(t) \underbrace{\lim_{n\to\infty} |W_n|\Gamma_n(t)}_{=\theta} dt = \theta \int_T C(t)dt := c_1.$$

Thus, $\mathbf{var}(\hat{\mu}_n) \sim \frac{c_1}{|W_n|} \to 0$ as $n \to \infty$. The unbiasedness of $\hat{\mu}_n$ then implies $\mathbb{E}\left[|\hat{\mu}_n - \mu|^2\right] = \mathbf{var}(\hat{\mu}_n) \to 0$ as $n \to \infty$. □

Under certain mixing and integrability assumptions, the asymptotic normality of $\hat{\mu}_n$ can be proven, i.e.

$$\sqrt{|W_n|} \cdot (\hat{\mu}_n - \mu) \xrightarrow{d} N(0, \sigma^2),$$

where $\sigma^2 := \theta \cdot \int\limits_T C(t)dt > 0$. Let us give an example of such assumptions [20, Theorem 1.7.4], see also [3, Chapter 3].

(I) Assumptions of Theorem 2.5 hold true.

(II) There exists a constant $\delta > 0$ such that $\mathbb{E}\left[|X(0)|^{2+\delta}\right] < \infty$.

(III) It holds that $\sigma^2 > 0$.

(IV) There exist $\beta, \varepsilon > 0$ such that

$$\alpha(r) \leq \beta \cdot r^{-d \cdot \varepsilon}$$

for $\varepsilon \cdot \delta > 2d$ and arbitrary $r \to \infty$, where $\alpha(r)$ is the so-called *α-mixing rate of X*.

To give a formal definition of the α-mixing rate we introduce the following quantity.

Definition 2.6 The *Rosenblatt dependence rate* of two σ-algebras $\mathcal{F}_1, \mathcal{F}_2 \subset \mathcal{F}$ is defined by

$$\alpha(\mathcal{F}_1, \mathcal{F}_2) = \sup_{A \in \mathcal{F}_1, B \in \mathcal{F}_2} |\mathbb{P}(A \cap B) - \mathbb{P}(A) \cdot \mathbb{P}(B)|.$$

It is a measure of stochastic dependence between $\mathcal{F}_1$ and $\mathcal{F}_2$. This can be easily seen from the following lemma.

Lemma 2.7 Let Y_j be $\mathcal{F}_j$-measurable random variables and $p_j > 1$, $j = 1, 2$, such that $\mathbb{E}\left[|Y_j|^{p_j}\right] < \infty$. Then,

$$|\mathbf{cov}(Y_1, Y_2)| \leq 10 \cdot \|Y_1\|_{p_1} \cdot \|Y_1\|_{p_2} (\alpha(\mathcal{F}_1, \mathcal{F}_2))^{\frac{1}{q}},$$

where $q > 1$ satisfies $\frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{q} = 1$ and $\|Y_j\|_{p_j} = (\mathbb{E}[|Y_j|^{p_j}])^{1/p_j}$ is the L^{p_j}-norm of Y_j.

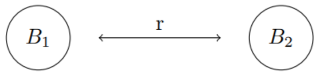
Proof A proof can be found in [20, Lemma 1.0.2] □

Now, for any Borel window $W \in \mathcal{B}_{\mathbb{R}^d}$ introduce $\mathcal{F}_{X(W)} = \sigma(\{X(t), t \in W\})$, i.e. the σ-algebra generated by the random field X within the observation window W. Finally, the *α-mixing rate* $\alpha(r)$ is defined as

$$\alpha(r) = \sup_{\substack{\Delta(B_1, B_2) \geq r \\ B_1, B_2 \in \mathcal{B}_{\mathbb{R}^d}}} \alpha(\mathcal{F}_{X(B_1)}, \mathcal{F}_{X(B_2)}), \quad r > 0,$$

where the supremum above is taken over all σ-algebras generated by X on observation windows B_1, B_2 that have a minimum distance of r to another, see the following for an illustration.

$$\Delta(B_1, B_2) := \inf_{\substack{x \in B_1 \\ y \in B_2}} \|x - y\|_2$$



The mixing conditions (I)-(IV) can be varied in many ways, see e.g. [10].

2.2 Estimation of the covariance function

The covariance function $C(h)$ of a wide-sense stationary, measurable random field $X = \{X(t)), t \in T\}$ observed within a Van-Hove sequence of observation windows $\{W_n\}$ can be estimated by

$$\hat{C}_n(h) := \frac{1}{|W_n \cap (W_n - h)|} \int_{W_n \cap (W_n - h)} X(t)X(t+h)dt - \hat{\mu}_n^2, \quad h \in W,$$

for any Borel observation window W. Alternatively, one might apply

$$\tilde{C}_n(h) := \frac{1}{|W_n|} \int\limits_{W_n} X(t)X(t+h)dt - \hat{\mu}_n^2, \quad h \in W_n \cup (W_n \oplus W),$$

for which the random field X needs to be observed in a larger area $W_n \cup (W_n \oplus W)$. For practical calculations based on a finite sample $\{X(t_j), j = 1, \dots, N\}$, a discretization of $\hat{C}_n$ and $\tilde{C}_n$ is given by

$$\bar{C}_N(h) := \frac{1}{N_n} \sum_{\substack{j,k=1: \\ t_j - t_k \approx h}}^{N} X(t_j)X(t_k) - \bar{\mu}_N^2,$$

where $N_h := \#\{(t_j, t_k), j, k = 1, \dots, N : t_j - t_k \approx h\} > 0,\ h \in W$ and $\bar{\mu}_N := \frac{1}{N} \sum\limits_{j=1}^{N} X(t_j)$.

Lemma 2.8 (Asymptotic unbiasedness): Under the assumption of Theorem 2.5 both $\hat{C}_n(h)$ and $\tilde{C}_n(h)$ are asymptotically unbiased, i.e. it holds that

$$\mathbb{E}\left[\hat{C}_n(h)\right] \to C(h) \quad \text{and} \quad \mathbb{E}\left[\tilde{C}_n(h)\right] \to C(h), \quad n \to \infty.$$

Proof By the stationarity of X and Fubini's theorem, we have

$$\begin{aligned} \mathbb{E}\left[\hat{C}_n(h)\right] &:= \frac{1}{|W_n \cap (W_n - h)|} \int_{W_n \cap (W_n - h)} \underbrace{\left(\mathbb{E}\left[X(t) \cdot X(t+h)\right] - \mu^2\right)}_{=C(h)} dt + \mu^2 - \mathbb{E}\left[\hat{\mu}_n^2\right] \\ &= C(h) - \underbrace{\mathbf{var}(\hat{\mu}_n)}_{\to 0 \text{ by Thm. } 2.5} \to C(h) \end{aligned}$$

as $n \to \infty$. The proof for $\tilde{C}_n(h)$ is analogous. □

Under some additional mixing and integrability conditions similar to (I)-(IV), one can show that the estimators $\hat{C}_n$ and $\tilde{C}_n$ are strongly consistent (uniformly in $h \in W$) and asymptotically normally distributed [20, chapter 4].

Returning to their computational version $\bar{C}_N$, note that the number N_h may be often very small or even zero when the sample points $\{t_1, \dots, t_n\}$ are not lying on a regular grid within $W \subset T$. This is for example the case for large lags h, where $\|h\|_2 \approx \text{diam(W)} := \sup_{x,y \in W} \|x - y\|_2$. The estimator $\bar{C}_N$ becomes unreliable at such lags h. In addition, $\bar{C}_N(\cdot)$ is not positively semi-definite (contrary to $C(\cdot)$).

Similarly to $\hat{C}_n$ and $\tilde{C}_n$, estimates of the *variogram* $\gamma(\cdot)$ for an intrinsically stationary measurable random field X of order two are commonly used in geostatistics. An estimator for γ is given by

$$\hat{\gamma}(h) := \frac{1}{2|W_n \cap (W_n - h)|} \int_{W_n \cap (W_n - h)} \left(X(t) - X(t+h)\right)^2 dt, \quad h \in W,$$

and a discretization is given by

$$\bar{\gamma}(h) := \frac{1}{2N_h} \sum_{\substack{j,k=1 \\ t_j - t_k \sim h}}^{N} (X(t_j) - X(t_k))^2, \quad h \in W.$$

Their asymptotic properties are very similar to those of $\hat{C}_n$ and $\bar{C}_n$, respectively, due to the well-known relation $\gamma(h) = C(0) - C(h)$ when X is mean-square integrable.

2.3 Spectral density estimation

Let f_C be the spectral density of the wide-sense stationary centered random field $X = \{X(t), t \in \mathbb{R}^d\}$ with covariance function C, i.e.

$$f_C(x) = \frac{1}{(2\pi)^d} \int\limits_{\mathbb{R}_d} e^{-i\langle x,t\rangle} C(t) dt, \quad x \in \mathbb{R}^d$$

by means of the Fourier inversion formula, if C is continuous at the origin and $\int_{\mathbb{R}^d} |C(t)| dt < \infty$. The so-called *periodogram* allows us to estimate f_C.

Definition 2.9 (a) The *periodogram* $\hat{f}_C$ of X observed within a window W_n is given by

$$\hat{f}_C(h) := \frac{1}{(2\pi)^d |W_n|} \left| \int_{W_n} \exp\{-i\langle t,h\rangle\} X(t) dt \right|^2, \quad h \in W,$$

where W and $\{W_n\}_{n=1}^{\infty}$ are observation windows.

(b) For a random field $X = \{X(t), t \in T\}$ observed only by a finite sample $(X(t_1), \ldots, X(t_N))$, $N \in \mathbb{N}$, the *empirical periodogram* $\bar{f}_N$ is defined by

$$\bar{f}_N(h) := \frac{1}{(2\pi)^d N} \left| \sum_{j=1}^{N} \exp\{-i\langle t_j,h\rangle\} X(t_j) \right|^2, \ h \in W.$$

Assume that $\{t_1, \ldots, t_N\} = \{0, 1, \ldots, M-1\}^d \cdot \delta$, $\delta > 0$, lie on a regular grid with $N = M$.

Lemma 2.10 If $f_C \in C(T)$, i.e. if f_C is a continuous function on T, then the estimators $\hat{f}_C$ and $\bar{f}_N$ are asymptotically unbiased almost everywhere on W, i.e.

$$\mathbb{E}\left[\hat{f}_C(h)\right] \to f_C(h), \quad n \to \infty, \quad \text{and} \quad \mathbb{E}\left[\bar{f}_N(h)\right] \to f_C(h), \quad N \to \infty,$$

for almost all $h \in W$.

Proof (a) It holds that $|z|^2 = z \cdot \bar{z}$ for all $z \in \mathbb{C}$, where $\bar{z}$ is the complex conjugate of z.

Fubini's theorem yields

$$
\begin{aligned}
\mathbb{E}\left[\hat{f}_C(h)\right] &= \frac{1}{(2\pi)^d |W_n|} \int_{W_n}\int_{W_n} \exp\left\{-i\langle t,h\rangle + i\langle s,h\rangle\right\} \underbrace{\mathbb{E}[X(s)X(t)]}_{=C(t-s)}\, dsdt \\
&= \frac{1}{(2\pi)^d |W_n|} \cdot \int_{\mathbb{R}^d}\int_{\mathbb{R}^d} e^{-i\langle h,t-s\rangle} C(t-s)\cdot \mathbb{1}(s\in W_n, \underbrace{t-s}_{=y} \in W_n \oplus \tilde{W}_n) dsdt \\
&\overset{t\mapsto y}{=} \frac{1}{(2\pi)^d} \cdot \underbrace{\int_{W_n} \frac{ds}{|W_n|}}_{=1} \cdot \int_{W_n\oplus\tilde{W}_n} e^{-i\langle h,y\rangle} C(y)dy \\
&\to \frac{1}{(2\pi)^d} \cdot \int_{\mathbb{R}^d} e^{-i\langle h,y\rangle} C(y)dy = f_C(h), \quad h\in T,
\end{aligned}
$$

as $n\to\infty$, since $W_n \oplus \check{W}_n \to \mathbb{R}^d$ for any Van-Hove-sequence $\{W_n\}$ and the Fourier inversion formula holds.

(b) To prove the statement for $\mathbb{E}[\bar{f}_N(h)]$, consider for simplicity $d=1$ and $t_j = j$, $j = 0,\dots,N-1$. Then, similar to (a), it holds that

$$
\begin{aligned}
\mathbb{E}\left[\bar{f}_N(h)\right] &= \frac{1}{2\pi N} \sum_{j,k=0}^{N-1} e^{-i(t_j-t_k)h} \underbrace{\mathbb{E}[X(t_j)X(t_k)]}_{C(t_j-t_k} \\
&= \frac{1}{2\pi N} \sum_{j,k=0}^{N-1} e^{-i(t_j-t_k)h} \cdot \int_{\mathbb{R}} e^{i(t_j-t_k)t} f_C(t)dt \\
&= \int_{\mathbb{R}} \frac{1}{2\pi N} \left| \sum_{j=0}^{N-1} e^{ij(t-h)} \right|^2 f_C(t)dt \\
&= \int_{\mathbb{R}} \varphi_N(t-h) f_C(t)dt \to f_C(h)
\end{aligned}
$$

as $N\to\infty$, where

$$
\varphi_N(\lambda) = \frac{1}{2\pi N} \left| \sum_{j=0}^{N-1} e^{i\lambda j} \right|^2 = \frac{1}{2\pi N} \left| \frac{\sin(\frac{\lambda}{2}N)}{\sin(\frac{\lambda}{2})} \right|^2
$$

is the so-called *Fejér kernel*. The convergence above holds for almost all $h\in W$ (with respect to the Lebesgue measure on $\mathbb{R}$) by the properties of the Fejér kernel.

□

However, the variance of $\hat{f}_C$ or $\bar{f}_N$, respectively, does not vanish with increasing n or N, which makes it a bad estimator of f_C. Indeed, it holds that $\mathbf{var}(\hat{f}_C(h)) \to f_C^2(h)$ as $n\to\infty$ [36, p. 129], since consistency is not given. In order to correct this estimate, we consider *smoothed versions* of $\hat{f}_C$ and $\bar{f}_N$, which are defined by

$$
\hat{f}_C^*(h) := \int_{\mathbb{R}^d} G_m(h-t)\hat{f}_C(t)dt, \quad h\in W,
$$

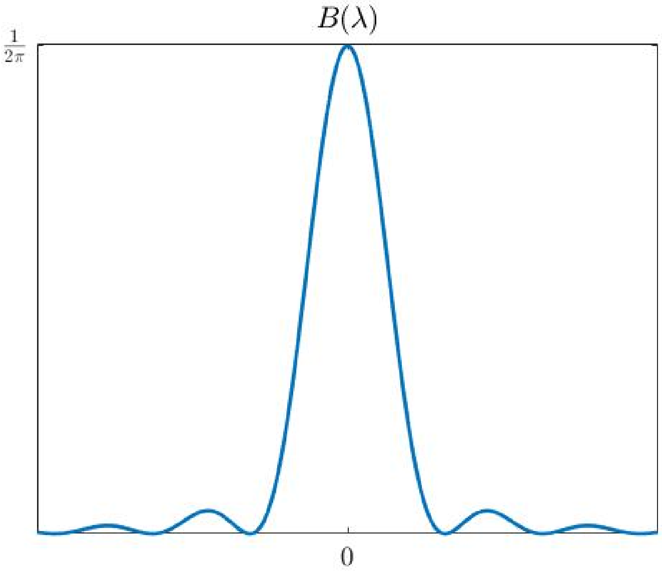


Fig. 2.1: Bartlett's kernel

and

$$\bar{f}^*_N(h) := \int_{\mathbb{R}^d} G_m(h-t)\bar{f}_N(t)dt, \quad t \in W,$$

where $G_m : \mathbb{R}^d \to \mathbb{R}_+$ is a square-integrable *smoothing kernel*, which approximates the Dirac delta function as $m \to \infty$ and $\int_{\mathbb{R}^d} G_m(t)dt = 1$ for all $m \in \mathbb{N}$, i.e. $G_m(t) \to \delta_0(t)$, $t \in T$, as $m \to \infty$ and $\int_{\mathbb{R}^d} G^2_m(t)dt < \infty$ for all $m \in \mathbb{N}$.

Remark 2.11 The asymptotic unbiasedness does not hold for $\bar{f}_N(h)$ if the sampling locations $t_1, \dots, t_n$ are irregularly spaced [8, Section 3.2].

Example 2.12 Let $\{a_m\}_{m\in\mathbb{N}}$ be a sequence with $a_m \to \infty$ and $\frac{a_m}{m} \to 0$ as $m \to \infty$. For $d = 1$ consider the following examples for smoothing kernel functions.

(a) *Bartlett's kernel*: $G_m(t) = a_m \cdot B(a_m \cdot t)$,

$$B(\lambda) = \frac{1}{2\pi} \cdot \left(\frac{\sin(\frac{\lambda}{2})}{\frac{\lambda}{2}} \right)^2 .$$

(b) *Parzen's kernel*: $G_m(t) = a^d_m \cdot P(a_m \cdot t)$, $d \in \{1, 2, 3\}$,

$$P(\lambda) = \frac{3}{8\pi} \cdot \left(\frac{\sin(\frac{\lambda}{4})}{\frac{\lambda}{4}} \right)^4 .$$

(c) *Zhurbenko's kernel*: $G_m(t) = a_m \cdot Z(a_m \cdot t)$,

$$Z(\lambda) = \frac{\alpha+1}{2\alpha} \cdot (1 - |\lambda|^\alpha) \cdot \mathbf{1}(|\lambda| \le 1), \quad \alpha(0, 2].$$

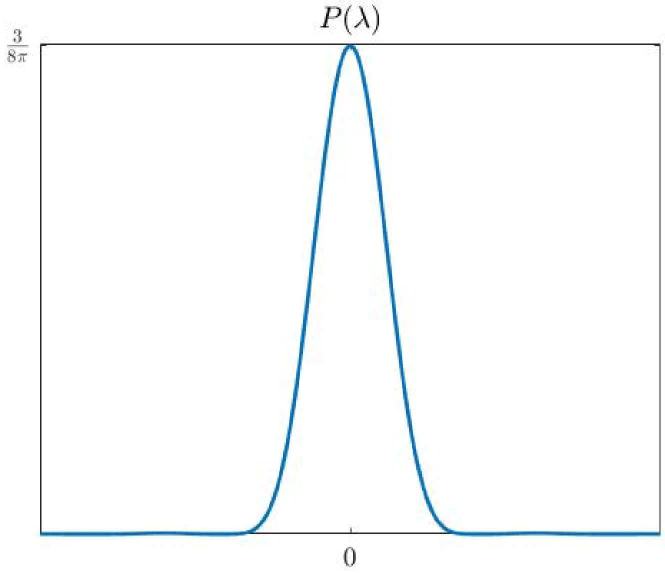


Fig. 2.2: Parzen's kernel

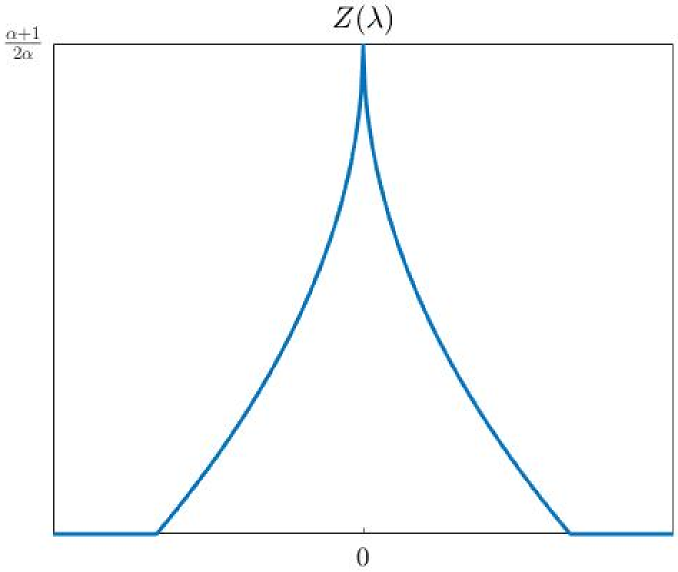


Fig. 2.3: Zhurbenko's kernel

All of these functions are even and have a pronounced peak (sharp maximum) at $\lambda = 0$, see Figures 2.1 - 2.3. The kernel $Z(\lambda)$ has compact support $[-1,1]$, whereas the support of $B(\lambda)$ and $P(\lambda)$ is the whole real line.

As seen above, we may put $G_m(t) = a_m^d \cdot H(a_m \cdot t)$ for any $d \geq 1$, where H is an even function H which integrates to 1, i.e. $\int_{\mathbb{R}^d} H(t)dt = 1$. For instance, one might choose

$$H(t) = p_d \cdot \left(\frac{\sin(\|t\|_d/4)}{\|t\|_d/4} \right)^4$$

with

$$p_d = \left(w_d \cdot 4^d \int_0^\infty \frac{\sin^4(\lambda)}{\lambda^{s-d}} \right)^{-1}, \quad d = 2,3,$$

where $w_d \in \{2\pi, 4\pi\}$ is the surface area of the unit sphere S^{d-1} in $\mathbb{R}^d$ and $\|\cdot\|_d$ is the Euclidean norm in $\mathbb{R}^d$ $d = 2,3$.

Exercise 2.13 Compute the normalizing constant z_d of

$$H(t) = z_d(1 - \|t\|_2^\alpha)\mathbf{1}(\|t\|_2 \leq 1), \quad \alpha \in (0,2],$$

for $d > 1$.

The smoothed periodograms $\hat{f}_C^*$, $\bar{f}_N^*$ are asymptotically unbiased as well:

Lemma 2.14 Let $G_m(t) = a_m^d \cdot H(a_m \cdot t),\ t \in T,\ m \in \mathbb{N}$, where $H : \mathbb{R}^d \to \mathbb{R}_+$ is an even bounded function with $\int_{\mathbb{R}^d} H(t)dt = 1$ and $a_m \to \infty$ as $m \to \infty$. Under the assumptions of Lemma 2.10, it holds that

$$\lim_{m\to\infty} \lim_{n\to\infty} \mathbb{E}\left[\hat{f}_C^*(h)\right] = f_C(h),$$
$$\lim_{m\to\infty} \lim_{N\to\infty} \mathbb{E}\left[\bar{f}_N^*(h)\right] = f_C(h)$$

for $h \in W$.

Proof We proof the assertion only for $\hat{f}_C^*$. Fubini's theorem yields

$$\mathbb{E}\left[\hat{f}_C^*(h)\right] = \int_{\mathbb{R}^d} G_m(h-t) \cdot \mathbb{E}\left[\hat{f}_C(h)\right] dt \to \int_{\mathbb{R}^d} G_m(h-t) f_C(t)dt,\ h \in W,$$

as $n \to \infty$, where the convergence is a result of Lemma 2.10 and Lebesgue's dominated convergence theorem. Then,

$$\int_{\mathbb{R}^d} G_m(h-t)f_C(t)dt = \int_{\mathbb{R}^d} a_m^d \cdot H(a_m(h-t)) \cdot f_C(t)dt \overset{y=a_m(t-h)}{=} \int_{\mathbb{R}^d} \underbrace{H(-y)}_{=H(y)} f_C(h + \frac{y}{a_m})dy$$
$$\to f_C(h) \cdot \underbrace{\int_{\mathbb{R}^d} H(y)dy}_{=1} = f_C(h), \quad h \in W,$$

as $m \to \infty$. □

In contrast to $\hat{f}_C(\cdot)$ and $\bar{f}_N(\cdot)$, the variance of the smoothed estimators $\hat{f}^*_C(\cdot)$ and $\bar{f}^*_N(\cdot)$ tends to zero as $N \to \infty$ for any fixed $m \in \mathbb{N}$. Indeed, we have

$$\mathbf{var}\left(\bar{f}^*_N(h)\right) = o\left(\frac{f^2_C(h)}{N}\int_{\mathbb{R}^d} G^2_m(t)dt\right),$$

see [36, p. 134], and thus $\hat{f}^*_C$ *and* $\bar{f}^*_N$ *are weakly consistent estimators* for the spectral density f_C by means of Chebyshev's inequality, since for all $h \in W$ it holds that

$$\begin{aligned}\mathbb{P}\left(\left|\bar{f}^*_N(h) - f_C(h)\right| > \varepsilon\right) &\leq \mathbb{P}\left(\left|\bar{f}^*_N(h) - \mathbb{E}\left[\bar{f}^*_N(h)\right]\right| > \frac{\varepsilon}{2}\right) + \mathbb{P}\left(\left|\mathbb{E}\left[\bar{f}^*_N(h)\right] - f_C(h)\right| > \frac{\varepsilon}{2}\right)\\ &\leq \frac{\mathbf{var}\left(\bar{f}^*_N(h)\right)}{\varepsilon^2/4} \to 0,\end{aligned}$$

as $a_m \to \infty$ and the function $H(\cdot)$ is chosen in a way such that $\frac{1}{N}\int_{\mathbb{R}^d} G^2_m(t)dt \to 0$ as $m, N \to \infty$. In general, it holds that $\int_{\mathbb{R}^d} G^2_m(t)dt \to \infty$ as $m \to \infty$, hence its convergence to infinity must be slower than N.

Remark 2.15 (a) The selection ot the bandwidth a_m is studied in [23, 31, 24].

(b) Asymptotic normality of $\hat{f}^*_C(\cdot)$ and $\bar{f}^*_N(\cdot)$ can be shown as in [36, Theorem 7, p.118]; see also [19] and [44].

In the literature, one can find further (parametric and non-parametric) spectral density estimates such as the Whittle likelihood [53, 12, 32, 48] and Kernel density estimators [8]. See also [15] and references therein. For a Bayesian approach we refer to [54, 46] and for other methods see [25, 1, 2, 5, 6, 18, 28, 47, 52].

3 Prediction of stationary random fields

Let $X = \{X(t), t \in T\}$ be a square-integrable and stationary (in a sense to be specified later) random field, $T \subset \mathbb{R}^d$. Assume that the sample $\{X(t_j),\ j = 1, \ldots, N\}$ is observable.

Problem: How can we predict the value of $X(t)$ for $t \notin \{t_1, \ldots, t_N\}$ based on this sample?

Denote by $\mathcal{F}_{t_1,\ldots,t_N} = \sigma(\{X(t_j), j = 1, \ldots, n\})$ the σ-algebra generated by $\{X(t_j), j = 1, \ldots, n\}$. Evidently, the predictor $\hat{X}(t)$ of $X(t)$ has to be $\mathcal{F}_{t_1,\ldots,t_N}$-measurable, and also *optimal* in some particular sense. A list of *desirable properties of* $\hat{X}(t)$ is given in the following.

(i) *Exactness:* $\hat{X}(t) = X(t)$ a.s. if $t = t_j,\ j = 1, \ldots, N$.

(ii) *Unbiasedness:* $\mathbb{E}[\hat{X}(t)] = \mathbb{E}[X(t)],\ t \in T$.

(iii) *Continuity:* Almost every path realization of $\hat{X}(t)$ is a continuous function in $t \in T$.

(iv) *Consistency:* $\hat{X}(t) \to X(t)$ as $N \to \infty$, where this convergence may be understood in the a.s., weakly or quadratic-mean sense.

(v) *Exactness in distribution:* $\hat{X}(t) \overset{d}{=} X(t)$, $t \in W$.

Later on, several criteria of optimality will be considered. One of the most common ones is

$$\mathbb{E}\left[|\hat{X}(t) - X(t)|^p\right] = \min_{Y \in L^p(\Omega, \mathcal{F}_{t_1,\ldots,t_N}, P)} \mathbb{E}\left[|Y - X(t)|^p\right]$$

for some $p \in \mathbb{N}$, e.g. $p = 1, 2$, where $L^p(\Omega, \mathcal{F}_{t_1,\ldots,t_N}, P)$ is the space of all $\mathcal{F}_{t_1,\ldots,t_N}$-measurable random variables Y with $\mathbb{E}[|Y|^p] < \infty$. Let us consider the case $p = 2$ in more detail.

3.1 L^2-optimal prediction as a conditional expectancy

It is well known that

$$\hat{X}(t) := \mathbb{E}\left[X(t) \mid \mathcal{F}_{t_1,\ldots,t_N}\right] = \underset{Y \in L^p(\Omega, \mathcal{F}_{t_1,\ldots,t_N}, p)}{\operatorname{argmin}} \mathbb{E}\left[(X(t) - Y)^2\right], \tag{3.1}$$

compare [41, Theorem 1.4.7]. It holds that there exists a Borel-measurable function φ such that $\hat{X}(t) = \varphi(X(t_1), \ldots, X(t_N))$ [41, Lemma 1.4.11]. However, the function φ is usually not explicitly known. In very few cases, it is known to be linear, as for instance in the case of Gaussian or some α-stable random fields

$$\varphi(X(t_1), \ldots, X(t_N)) = \lambda_1 \cdot X(t_1) + \cdots + \lambda_N \cdot X(t_N)$$

Here, the procedure of finding weights $\lambda_j = \lambda_j(t_1, \ldots, t_N, t),\ j = 1, \ldots, N$, satisfying (3.1) is called *linear regression*.

3.1.1 Linear regression for Gaussian random fields

We consider linear regression in the Gaussian case and begin with $N = 1$. The goal is to find $\mathbb{E}[X(t) \mid X(t_1)]$, if the stationary random field is Gaussian with the mean $\mathbb{E}[X(t)] = \mu$ and covariance function $C(t) = \mathbf{cov}(X(0), X(t))$, which is positive definite. In this case, the random vector $(X(t), X(t_1))$ has a bivariate normal distribution with probability density function

$$f_{X(t),X(t_1)}(x, y) = \frac{1}{2\pi\sigma^2\sqrt{1-\rho^2}} \exp\left\{-\frac{1}{2(1-\rho^2)\sigma^2}\left[(x-\mu)^2 - 2\rho(x-\mu)(y-\mu) + (y-\mu)^2\right]\right\}$$

for $x, y \in \mathbb{R}$, where $\rho = \mathbf{corr}(X(t), X(t_1)) \in (-1, 1)$ and $\sigma^2 = C(0) = \mathbf{var}(0) > 0$. The conditional density $f_{X(t)|X(t_1)}(x \mid y)$ of $X(t)$ given $X(t_1) = y$ is equal to

$$\begin{aligned} f_{X(t)|X(t_1)}(x \mid y) &= \frac{f_{X(t),X(t_1)}(x, y)}{f_{X(t_1)}(y)} \\ &= \frac{1}{\sqrt{2\pi\sigma^2(1-\rho^2)}} \exp\left\{\frac{-1}{2(1-\rho^2)\sigma^2}\left((x-\mu)^2 - 2\rho(x-\mu)(y-\mu) + (y-\mu)^2\right) + \frac{1-\rho^2(y-\mu)^2}{2\sigma(1-\rho^2)}\right\} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma} \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}\left((x-\mu)^2 - 2\rho(x-\mu)(y-\mu) + \rho^2(y-\mu)^2\right)\right\} \\ &= \frac{1}{\sqrt{2\pi(1-\rho^2)}\sigma} \exp\left\{-\frac{1}{2\sigma^2(1-\rho^2)}\left(x - \underbrace{(\mu + \rho(y-\mu))}_{=:\mu(y)}\right)^2\right\}, \end{aligned}$$

which is equal to the probability density function of a $N(\mu(y), (1-\rho^2)\sigma^2)$ distributed random variable. The following equations are well-known. It holds that

$$\mathbb{E}[X(t) \mid X(t_1) = y] = \int_{\mathbb{R}} x \cdot f_{X(t)|X(t_1)}(x \mid y)dx = \mu(y) = \mu + \rho(y-\mu) \tag{3.2}$$

and

$$\mathbf{var}(X(t) \mid X(t_1) = y) = \int_{\mathbb{R}} (x - \mu(y))^2 \cdot f_{X(t)|X(t_1)}(x \mid y)dx = \sigma^2(1-\rho^2). \tag{3.3}$$

The conditional variance above does not depend on y and represents the minimal variance in (3.2). Hence, the following lemma holds.

Lemma 3.1 Let $X = \{X(t), t \in T\}$ be a stationary Gaussian random field with mean $\mathbb{E}[X(t)] = \mu$, positive definite covariance function $C(t) = \mathbf{cov}(X(0), X(t))$ and $\sigma^2 = C(0) > 0$. Then,

$$\hat{X}(t) = \mathbb{E}[X(t) \mid X(t_1)] = \mu + \frac{C(t-t_1)}{\sigma^2}(X(t_1) - \mu)$$

with

$$\mathbb{E}\left[(\hat{X}(t) - X(t))^2\right] = \sigma^2 - \frac{C^2(t-t_1)}{\sigma^2}.$$

Proof Follows from (3.2) and (3.3) with $\rho = \frac{C(t-t_1)}{\sigma^2}$. □

Let I_N be the $(N \times N)$-dimensional identity matrix. We can now formulate and prove the following more general result.

Theorem 3.2 Let X be as in Lemma 3.1. Then,

$$\hat{X}(t) = \mathbb{E}[X(t) \mid X(t_1), \dots, X(t_N)] = \mu + \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \tag{3.4}$$

and

$$\mathbb{E}\left[(\hat{X}(t) - X(t))^2\right] = \sigma^2 - \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \Sigma^{\intercal}_{t,t_1,\dots,t_N}, \tag{3.5}$$

where $\Sigma_{t,t_1,\dots,t_N} = (C(t-t_1), \dots, C(t-t_N))$ and $\Sigma_{t_1,\dots,t_N} = (C(t_j - t_k))^N_{j,k=1}$.

Proof Consider a random variable

$$Y = X(t) - \mu - \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \tilde{X},$$

where $\tilde{X} = (X(t_1) - \mu, \dots, X(t_N) - \mu)^{\intercal}$. It holds that

$$\mathbb{E}[Y \cdot \tilde{X}^{\intercal}] = \underbrace{\mathbb{E}[(X(t) - \mu) \cdot \tilde{X}^{\intercal}]}_{=\Sigma_{t,t_1,\dots,t_N}} - \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \underbrace{\mathbb{E}[\tilde{X} \cdot \tilde{X}^{\intercal}]}_{=\Sigma_{t_1,\dots,t_N}} = 0.$$

Hence, Y and $\tilde{X}$ are uncorrelated, and therefore also stochastically independent, since they are jointly Gaussian. It follows that

$$\mathbb{E}[Y \mid \underbrace{(X(t_1), \dots, X(t_N)]}_{=\tilde{X}+\mu(1,\dots,1)}) = \mathbb{E}[Y] = 0,$$

and consequently

$$\mathbb{E}[Y \mid (X(t_1), \dots, X(t_N))] = \mathbb{E}[X(t) - \mu \mid X(t_1), \dots, X(t_N)] - \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \tilde{X}.$$

We can conclude that

$$\mathbb{E}[X(t) \mid X(t_1), \dots, X(t_N)] = \mu + \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \tilde{X},$$

which proves (3.4).

To show that (3.5) holds, note that $Y = X(t) - \mathbb{E}[X(t) \mid X(t_1), \dots, X(t_N)]$, and Y and $(X(t_1), \dots, X(t_N))^{\intercal}$ are stochastically independent. We can compute

$$\begin{aligned}
&\mathbf{var}(X(t) \mid X(t_1), \dots, X(t_N)) \\
&:= \mathbb{E}\Bigg[\underbrace{(X(t) - \overbrace{\mathbb{E}[X(t) \mid X(t_1), \dots, X(t_N)]}^{=\hat{X}(t)})^2}_{=Y^2} \mid X(t_1), \dots, X(t_N)\Bigg] \\
&= \mathbb{E}[Y^2 \mid X(t_1), \dots, X(t_N)] \\
&= \mathbb{E}[Y^2] \\
&= \mathbb{E}\left[\left(X(t) - \mu - \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \tilde{X}\right)^2\right] \\
&= \underbrace{\mathbb{E}\left[(X(t) - \mu)^2\right]}_{=\sigma^2} - 2\Sigma_{t,t_1,\dots,t_N} \Sigma^{-1}_{t_1,\dots,t_N} \Sigma^{\intercal}_{t,t_1,\dots,t_N} + \Sigma_{t,t_1,\dots,t_N} \underbrace{\Sigma^{-1}_{t_1,\dots,t_N} \underbrace{\mathbb{E}[\tilde{X}\tilde{X}^{\intercal}]}_{=\Sigma_{t_1,\dots,t_N}}}_{=I_N} \Sigma^{-1}_{t_1,\dots,t_N} \Sigma^{\intercal}_{t,t_1,\dots,t_N} \\
&= \sigma^2 - \Sigma_{t,t_1,\dots,t_N} \Sigma^{-1}_{t_1,\dots,t_N} \Sigma^{\intercal}_{t,t_1,\dots,t_N},
\end{aligned}$$

which is deterministic. Thus,

$$\mathbb{E}\left[(\hat{X}(t)-X(t))^2\right] = \mathbb{E}\left[\mathbb{E}[Y^2 \mid X(t_1),\dots,X(t_N)]\right] = \sigma^2 - \Sigma_{t,t_1,\dots,t_N} \cdot \Sigma^{-1}_{t_1,\dots,t_N} \cdot \Sigma^{\mathsf{T}}_{t,t_1,\dots,t_N},$$

which coincides with (3.5). □

Remark 3.3 Lemma 3.1 and Theorem 3.2 hold (with obvious modifications $\mu \mapsto \mu(t)$, $C(t-s) \mapsto C(s,t)$, $\sigma^2 \mapsto C(t,t)$) also for non-stationary random fields.

Corollary 3.4 Let $X(t_1),\dots,X(t_N)$ be stochastically independent, and X be as in Lemma 3.1. Then,

$$\hat{X}(t) = \mu + \frac{1}{\sigma^2} \cdot \sum_{j=1}^{N} C(t-t_j) \cdot (X(t_j)-\mu), \tag{3.6}$$

$$\mathbb{E}\left[(\hat{X}(t)-X(t))^2\right] = \sigma^2 - \frac{1}{\sigma^2} \cdot \sum_{j=1}^{N} C^2(t-t_j). \tag{3.7}$$

Proof Use Equations (3.4) and (3.5) with $\Sigma_{t_1,\dots,t_N} = \sigma^2 \cdot I_N$. □

Notice that Equations (3.6) - (3.7) are a direct generalization of Lemma 3.1.

3.1.2 Linear regression for α-stable random fields

Let $X = \{X(t), t \in T\}$ be a strictly stationary α-stable random field with index of stability $\alpha \in (1,2)$, so that $\mathbb{E}[|X(t)|] < \infty$ for all $t \in T$, which is necessary for conditional expectations to exist. Assume for simplicity that $X(t)$ is symmetric α-stable (write $X(t) \sim S\alpha S$), i.e.

$$\varphi_{X(t)}(s) = \mathbb{E}[e^{isX(t)}] = e^{-\sigma^\alpha |s|^\alpha}, \quad s \in \mathbb{R}.$$

Note that X is centered, i.e. $\mathbb{E}[X(t)] \equiv 0$ for all $t \in T$.

Problem: When does

$$\mathbb{E}[X(t) \mid X(t_1),\dots,X(t_N)] = \lambda_1 \cdot X(t_1) + \dots + \lambda_N \cdot X(t_N) \quad a.s. \tag{3.8}$$

hold?

First, we mention a very general result on characteristic functions.

Theorem 3.5 Let $Z = (Z_0, Z_1, \dots, Z_N)$ be a random vector with $\mathbb{E}[|Z_j|] < \infty$, $j = 0,\dots,N$, and joint characteristic function $\varphi_Z(s) = \mathbb{E}[e^{i\langle s,Z\rangle}]$, $s = (s_0, s_1, \dots, s_N) \in \mathbb{R}^{N+1}$. Then,

$$\begin{aligned}
&\mathbb{E}[Z_0 \mid Z_1,\dots,Z_N] = \sum_{j=0}^{N} \lambda_j Z_j \quad a.s.\\
\iff &\frac{\partial}{\partial s_0}\varphi_Z(s_0,s_1,\dots,s_N)\Big|_{s_0=0} = \lambda_1 \frac{\partial}{\partial s_1}\varphi_Z(s_0,s_1,\dots,s_N) + \dots + \lambda_N \frac{\partial}{\partial s_N}\varphi_Z(s_0,s_1,\dots,s_N).
\end{aligned} \tag{3.9}$$

Proof See [29, Theorem 3.1] □

A random field $X = \{X(t), t \in T\}$ is a $S\alpha S$ *stationary Subgaussian random field*, $\alpha \in (1,2)$ if

$$X(t) \overset{d}{=} \sqrt{A} \cdot Y(t),$$

where $Y = \{Y(t), t \in T\}$ is a centered Gaussian random field with positive-definite covariance function $C(\cdot)$, independent of $A \sim S_{\alpha/2}((\cos(\frac{\pi\alpha}{4}))^{2/\alpha}, 1, 0)$.

Corollary 3.6 Let $X = \{X(t), t \in T\}$ be a $S\alpha S$ stationary Subgaussian random field. Then, the regression (3.8) is always linear and coincides with (3.3) for $\mu = 0$, i.e.

$$\mathbb{E}\left[X(t_0) \mid X(t_1), \ldots, X(t_N)\right] = \Sigma_{t_0,t_1,\ldots,t_N} \cdot \Sigma^{-1}_{t_1,\ldots,t_N} (X(t_1), \ldots, X(t_N))^\intercal.$$

Proof Check the validity of condition (3.9) for $Z = (X(t_0), \ldots, X(t_N))$ with characteristic function $\varphi_Z(s) = \exp\{-(s^\intercal \Sigma s)^{\alpha/2}\}$, where $\Sigma = (C(t_j - t_k))^N_{j,k=0}$. □

Exercise 3.7 Proof Corollary 3.6.

Corollary 3.8 Let $X = \{X(t), t \in T\}$ be a $S\alpha S$ random field, $\alpha \in (1,2)$, and let Γ be the spectral measure on S^N of the $S\alpha S$ random vector $(X(t_0), \ldots, X(t_N))^\intercal$. Then,

$$\begin{aligned} &\mathbb{E}\left[X(t_0) \mid X(t_1), \ldots, X(t_N)\right] = \lambda_1 X(t_1) + \cdots + \lambda_N X(t_N) \quad a.s. \\ \Longleftrightarrow &\int\limits_{S^N} (x - \lambda_1 x_1 - \cdots - \lambda_N x_N)(s_1 x_1 + \cdots + s_N x_N)^{\langle \alpha - 1 \rangle} \Gamma(dx) = 0 \end{aligned} \tag{3.10}$$

for all $s_1, \ldots, s_N \in \mathbb{R}$, where $a^{\langle p \rangle} := |a|^p \cdot \mathrm{sgn}(a)$ for $a, p \in \mathbb{R}$ and $dx = dx_0 dx_1 \ldots dx_N$.

Note that, condition (3.10) is always satisfied for $N = 1$.

Proposition 3.9 Under the conditions of Corollary 3.7, it holds that $\mathbb{E}[X(t_0) \mid X(t_1)] = \lambda_1 \cdot X(t_1)$ a.s., where

$$\lambda_1 = \frac{[X(t_0), X(t_1)]_\alpha}{\sigma^\alpha}$$

with scale parameter σ of $X(t_1) \sim S\alpha S(\sigma)$ and *covariation*

$$[X(t_0), X(t_1)]_\alpha = \int_{S^1} x_1 \cdot x_2^{\langle \alpha - 1 \rangle} \Gamma(dx_1, dx_2)$$

of $X(t_0)$, $X(t_1)$ and Γ being the spectral measure of the vector $(X(t_0), X(t_1))^\intercal$.

Proof Write (3.10) for $N = 1$ in the form

$$s_1^{\langle \alpha - 1 \rangle} \underbrace{\int\limits_{S^2} x_o \cdot x_1^{\langle \alpha - 1 \rangle} \Gamma(dx)}_{=[X(t_0),X(t_1)]_\alpha} = \lambda_1 \cdot s_1^{\langle \alpha - 1 \rangle} \underbrace{\int\limits_{S^2} \overbrace{x_1 \cdot x_1^{\langle \alpha - 1 \rangle}}^{=|x_1|^\alpha} \Gamma(dx)}_{=\sigma^2},$$

where $dx = dx_1 dx_2$. After canceling $s_1^{\langle \alpha - 1 \rangle}$ out, we get $\lambda_1 = \frac{[X(t_0),X(t_1)]_\alpha}{\sigma^\alpha}$. □

Remark 3.10 (Properties of the covariation): The covariation is a dependence measure between jointly $S\alpha S$ random variables, $\alpha \in (1,2]$, which may be considered a generalization of covariance for $\alpha = 2$. Indeed, for $(Y_1, Y_2) \sim S2S = N(0, \Sigma)$, with covariance matrix

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

and spectral measure Γ it holds that$[Y_1, Y_2]_2 = \frac{1}{2}\mathbf{cov}(Y_1, Y_2)$, as the following comparison between $\varphi_{(Y_1,Y_2)}(s)$ of a Gaussian and a $S\alpha S$ random vector shows:

$$\begin{aligned} \varphi_{(Y_1,Y_2)}(s) &= \mathbb{E}\left[e^{i(s_1Y_1+s_2Y_2)}\right] \\ &= \exp\left\{-\int_{S^2}(s_1x_1+s_2x_2)^2\Gamma(dx)\right\} \\ &= \exp\left\{-\left(s_1^2\int_{S^2}x_1^2\Gamma(dx) + 2s_1s_2[Y_1,Y_2]_2 + s_2^2\int_{S^2}x_2^2\Gamma(dx)\right)\right\} \\ &= \exp\left\{-\frac{1}{2}\left(s_1^2\cdot\mathbf{var}Y_1 + 2s_1s_2\cdot\mathbf{cov}(Y_1,Y_2) + s_2^2\cdot\mathbf{var}(Y_2)\right)\right\} \end{aligned}$$

Setting $s_1 = 0$ or $s_2 = 0$ yields

$$\mathbf{var}(Y_j) = 2\int_{S^2} x_j^2\Gamma(dx), \quad j = 1, 2,$$

hence

$$[Y_1, Y_2]_2 = \frac{1}{2}\mathbf{cov}(Y_1, Y_2).$$

However, in the case $\alpha \in (1,2)$ the covariation $[X(t_0), X(t_1)]_\alpha$ of a $S\alpha S$ random vector (Y_1, Y_2) is *not a symmetric function* of Y_1, Y_2, and it is linear only with respect to its first argument.

The following proposition shows *necessary conditions* for Equation (3.8) to hold for any $N \geq 1$ that are much simpler than condition (3.10):

Proposition 3.11 Under the assumptions of Corollary 3.7, if

$$\mathbb{E}[X(t_0) \mid X(t_1), \ldots, X(t_N)] = \sum_{j=1}^{N} \lambda_j X(t_j) \quad a.s. \tag{3.11}$$

then the coefficients $\lambda_1, \ldots, \lambda_N$ must satisfy the following system of linear equations:

$$\sum_{j=1}^{N} \lambda_j [X(t_j), X(t_k)]_\alpha = [X(t_0), X(t_k)]_\alpha, \quad k = 1, \ldots, N. \tag{3.12}$$

Proof Set $c_{jk} = [X(t_j), X(t_k)]_\alpha$, $j, k = 0, \ldots, N$. It holds that $c_{jj} = \sigma_j^\alpha$, see the proof of Proposition 3.8. Using its result, we infer for $j = 1, \ldots, N$

$$\begin{aligned} \frac{c_{0j}}{\sigma_j^\alpha}\cdot X(t_j) &= \mathbb{E}\left[X(t_0) \mid X(t_j)\right] = \mathbb{E}\left[\mathbb{E}\left[X(t_0) \mid \sum_{k=1}^{N} X(t_k)\right] \mid X(t_j)\right] \\ &\overset{\text{Eq. (3.11)}}{=} \mathbb{E}\left[\sum_{k=1}^{N}\lambda_k X(t_k) \mid X(t_j)\right] = \lambda_j X(t_j) + \sum_{k\neq j}\lambda_k \mathbb{E}[X(t_k) \mid X(t_j)] \\ &\overset{\text{Prop. 3.9}}{=} \lambda_j X(t_j) + \sum_{k\neq j}\lambda_k \frac{c_{kj}}{\sigma_j^\alpha} X(t_j) = \left(\frac{1}{\sigma_j^\alpha}\sum_{k=1}^{N}\lambda_k c_{kj}\right) X(t_j). \end{aligned}$$

This leads to $c_{0j} = \sum_{k=1}^{N} \lambda_k c_{kj},\ j = 1, \dots, N$, which ultimately yields (3.12). □

Since $[Y_1, Y_2]_\alpha$ is not linear in its second argument, we can easily construct an example of a non-linear regression, where the necessary condition (3.11) does not hold.

Example 3.12 Let Y_1, Y_2, Y_3 be stochastically independent $S\alpha S$ random variables, $\alpha \in (1, 2)$. Consider the $S\alpha S$ random vector $X = (X_0, X_1, X_2)^\intercal$ with $X_0 = Y_1,\ X_1 = Y_1 + Y_2,\ X_2 = Y_1 + Y_3$. Then,

$$\mathbb{E}[X_0 \mid X_1, X_2] \neq \lambda_1 X_1 + \lambda_2 X_2 \quad a.s.$$

Exercise 3.13 Check that condition (3.11) in Example 3.12 is not satisfied.
Hint: Consider $[X_0, X_1 + \theta X_2]_\alpha$ *and* $\mathbb{E}[X_0 \mid X_1 + \theta X_2]$ *as a function of* $\theta > 0$.

Finally, we state the following positive result about regression:

Theorem 3.14 Let $X = \{X(t), t \in T\}$ be a $S\alpha S$ random field, $\alpha \in (1, 2)$, and let locations $t_0, t_1, \dots, t_N \in T$, $N \geq 1$, be chosen such that $X(t_1), \dots, X(t_N)$ are stochastically independent. Then,

$$\mathbb{E}\left[X(t_0) | X(t_1), \dots, X(t_N)\right] = \sum_{j=1}^{N} \lambda_j X(t_j) \quad a.s.$$

where

$$\lambda_j = \frac{[X(t_0), X(t_j)]_\alpha}{\sigma_j^\alpha}$$

and σ_j^α is the scale parameter of $X(t_j),\ j = 1, \dots, N$.

Proof See [37, Corollary 4.1.5]. □

3.2 Kriging methods

The previous section, in particular Section 3.1.2, illustrated how prediction as a conditional mean does not always lead to feasible computable forecasts, since the regression $\mathbb{E}[X(t) \mid X(t_1), \dots, X(t_n)]$ may not be linear. From an application's point of view however, linear forecast methods are very easy to use and thus desirable to have. Hence, the need for linear L^2-theory of prediction for square-integrable random functions arose. After pioneering publications [21, 39], where the linear predictor

$$\hat{X}(t) = \sum_{j=1}^{N} \lambda_j X(t_j) + \lambda_0, \quad t \in W, \tag{3.13}$$

was used with weights $\lambda_1, \dots, \lambda_N$, which solve the minimization problem

$$E\left[\left(\hat{X}(t) - X(t)\right)^2\right] \to \min_{\lambda_1, \dots, \lambda_N}, \tag{3.14}$$

the German geologist D.G. Krige was first to apply these predictors for gold ore mining prediction in South Africa (1951). These linear prediction methods, further developed by the French school of mathematical geology(1960s, G. Mathéron), were subsequently called linear inter- or extrapolation, or *Kriging*, some of which we explore in the sequel. More detailed accounts on Kriging methods can be found in the books [45, 51, 4, 22].

3.2.1 Simple Kriging

Let $X = \{X(t), t \in T\}$ be a (possibly non-stationary) square-integrable random field, i.e. $\mathbb{E}[X^2(t)] < \infty, t \in T$, with known covariance function $C(s,t) = \mathbf{cov}(X(s), X(t)),\ s,t \in T$, and mean function $\mu(t) = \mathbb{E}[X(t)],\ t \in T$.

Goal: Find a linear forecast (3.13) such that the prediction error $\mathbb{E}\left[\left(\hat{X}(t) - X(t)\right)^2\right]$ is minimal.

It is easily seen that

$$\mathbb{E}\left[\left(\hat{X}(t) - X(t)\right)^2\right] = \mathbf{var}\left(\hat{X}(t) - X(t)\right) + \left(\mathbb{E}\left[\hat{X}(t) - X(t)\right]\right)^2 \to \min_{\lambda_1,\dots,\lambda_N}$$

if $\mathbb{E}[\hat{X}(t) - X(t)] \equiv 0$, i.e. the predictor $\hat{X}(t)$ is unbiased and $\mathbb{E}[\hat{X}(t)] = \mathbb{E}[X(t)] = \mu(t),\ t \in W$.

Plugging in the linear form of $\hat{X}(t)$ from Equation (3.13) into this relation yields $\mu(t) = \sum_{j=1}^{N} \lambda_j \mu(t_j) + \lambda_0$, hence, equivalently $\lambda_0 = \mu(t) - \sum_{j=1}^{N} \lambda_j \mu(t_j)$, which allows for

$$\hat{X}(t) = \mu(t) + \sum_{j=1}^{N} \lambda_j (X(t_j) - \mu(t_j)), \quad t \in W.$$

The latter expression implies that if $\mu(t)$ is explicitly known, then $X(t)$ can be centered by subtracting its mean in the forecast $\hat{X}(t)$.

Let $t = t_0$ and consider

$$\Psi(\lambda) = \mathbf{var}(\hat{X}(t_0) - X(t_0)) = \mathbb{E}\left[\left(\sum_{j=0}^{N} \lambda_j (X(t_j) - \mu(t_j))\right)^2\right]$$

with $\lambda_0 = -1$ as the target function to be minimized with respect to $\lambda = (\lambda_1, \dots, \lambda_N) \in \mathbb{R}^N$. Note that we used the fact that $\mathbb{E}\left[\sum_{j=0}^{N} \lambda_j (X(t_j) - \mu(t_j))\right] = 0$ for all $\lambda_1, \dots, \lambda_N$ in the above.

The necessary conditions of an extremum are $\frac{\partial \Psi(\lambda)}{\partial \lambda_j} = 0$, $j = 1, \dots, N$. Since

$$\Psi(\lambda) = \sum_{j,k=0}^{N} \lambda_j \lambda_k \underbrace{\mathbf{cov}(X(t_j), X(t_k))}_{:=C(t_j,t_k)} \tag{3.15}$$

by linearity of the expectation, we get

$$\frac{\partial \Psi(\lambda)}{\partial \lambda_j} = 2 \sum_{\substack{k=0 \\ k \neq i}}^{N} \lambda_k C(t_j, t_k) + 2\lambda_j C(t_j, t_j) = 0, \quad j = 1, \dots, N,$$

which, together with $\lambda_0 = -1$, yields the system of linear equations

$$\sum_{k=1}^{N} \lambda_j \cdot C(t_j, t_k) = C(t_0, t_j), \quad j = 1, \dots, N,$$

or, in matrix form,

$$\Sigma \lambda = c_0, \tag{3.16}$$

where $\Sigma = (C(t_j, t_k))_{k,j=1}^{N}$, $\lambda = (\lambda_1, \dots, \lambda_N)^\intercal \in \mathbb{R}^N$ and $c_0 = (C(t_o, t_1), \dots, C(t_0, t_k))^\intercal$.

Theorem 3.15 Let $X = \{X(t), t \in T\}$ be a square-integrable random field with known mean function $\mu(t)$, $t \in T$, and positive definite covariance function $C(s,t)$, $s,t \in T$. Then, the simple Kriging method yields the unique predictor

$$\hat{X}(t) = \mu(t) + \sum_{j=1}^{N} \lambda_j (X(t_j) - \mu(t_j))$$

with $\lambda = (\lambda_1, \dots, \lambda_N)^\intercal = \Sigma^{-1} \cdot c_0$.

Proof The quadratic function $\Psi(\lambda)$ has a unique minimum if Σ is invertible, since it is a paraboloid function with $\Psi(\lambda) \geq 0$ for all λ. Then, the vector λ, which satisfies $\frac{\partial \Psi}{\partial \lambda_j} = 0$, $j = 1, \dots, N$, coincides with the unique solution of Equation (3.16). □

Let us investigate the properties of simple Kriging forecast.

Theorem 3.16 (Properties of simple Kriging): Under the assumptions of Theorem 3.15, the following holds.

(i) *Exactness:* $\hat{X}(t_j) = X(t_j)$, $j = 1, \dots, N$.

(ii) *Smoothness:* If μ, C are C^k-smooth, $k \in \mathbb{N}_0$, then so is $\hat{X}(\cdot)$.

(iii) *Shrinkage:* $\mathbf{var}(\hat{X}(t_0)) \leq \mathbf{var}(X(t_0))$, $t \in W$.

(iv) *Orthogonality:* $\mathbb{E}\left[(\hat{X}(t_0) - X(t_0))Y\right] = 0$ for all $Y \in L_N$, where L_N is the linear span of $X(t_j)$, $j = 1, \dots, N$, i.e. $L_N = \mathbf{span}\{X(t_1), \dots, X(t_N)\}$

(v) If X is Gaussian, then the simple Kriging predictor coincides with *Gaussian linear regression*, i.e.
$$\hat{X}(t) = \mathbb{E}\left[X(t_0) \mid X(t_1), \dots, X(t_N)\right] \quad a.s., \quad t_0 \in W.$$

Proof (i) It is easy to see that for $t_0 = t_j$, the vector $\lambda = (0, \dots, 0, \underset{j}{1}, \dots, 0)$ is the unique solution of the system of linear equations in Equation (3.16).

(ii) The smoothness of

$$\hat{X}(t_0) = \mu(t_0) + \sum_{j=1}^{N} \lambda_j \cdot (X(t_j) - \mu(t_j)) = \mu(t_0) + \bar{X} \cdot \Sigma^{-1} c_0$$

with $\bar{X} = (X(t_1) - \mu(t_1), \dots, X(t_N) - \mu(t_N))$ with respect to $t_0 \in W$ is evidently controlled by the corresponding smoothness of functions $\mu(\cdot)$, $C(\cdot, t_j)$.

(iii) One can easily see from Equation (3.16) that

$$\mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] = \mathbf{var}(X(t_0)) - \mathbf{var}(\hat{X}(t_0)) \geq 0,$$

hence

$$\mathbf{var}(X(t_0)) \geq \mathbf{var}(\hat{X}(t_0)), \ t_0 \in W.$$

Indeed, Equations (3.16) and (3.15) together with $\lambda_0 = -1$ imply that

$$\Psi(\lambda) = \lambda^\intercal \Sigma \lambda - 2\lambda^\intercal \cdot \underbrace{\Sigma\lambda}_{=c_0} + \underbrace{C(t_0, t_0)}_{=\mathbf{var}(X(t_0))} = \mathbf{var}(X(t_0)) - \underbrace{\lambda^\intercal \Sigma \lambda}_{=\mathbf{var}(\hat{X}(t_0))} .$$

(iv) For any $Y \in L_N$, there exist $\gamma_1, \dots \gamma_N \in \mathbb{R}$ such that $Y = \sum_{j=1}^N \gamma_j X(t_j)$. Then,

$$\begin{aligned}
&\mathbb{E}\left[Y \cdot \left(\hat{X}(t_0) - X(t_0)\right)\right] \\
&= \mathbb{E}\left[\sum_{j=1}^N \gamma_j \cdot X(t_j) \cdot \sum_{j=0}^N \lambda_j (X(t_j) - \mu(t_j))\right] \\
&= \sum_{j=1}^N \sum_{k=0}^N \gamma_j \lambda_k \mathbb{E}\left[X(t_j) \cdot (X(t_k) - \mu(t_k))\right] \\
&= \sum_{j=1}^N \sum_{k=0}^N \gamma_j \lambda_k \Big(\underbrace{\mathbb{E}\left[X((t_j) - \mu(t_j)) \cdot (X(t_k) - \mu(t_k))\right]}_{=C(t_j,t_k)} + \mu(t_j) \cdot \underbrace{\mathbb{E}\left[X(t_k) - \mu(t_k)\right]}_{=0} \Big) \\
&= \sum_{j=1}^N \gamma_j \left(\sum_{k=1}^N \lambda_k C(t_j, t_k) - C(t_0, t_k) \right) = \gamma^\intercal \underbrace{\Sigma\lambda}_{=c_0} - \gamma^\intercal \cdot c_0 = 0,
\end{aligned}$$

where $\gamma = (\gamma_1, \dots, \gamma_N)^\intercal$.

(v) The assertion follows from Equation (3.4), which evidently holds also for non-stationary Gaussian random fields, see Remark 3.3.

□

Remark 3.17 (a) The shrinkage property (iii) in Theorem 3.16 means that the simple Kriging estimate is less dispersed than the original random field. The simple Kriging predictor $\hat{X}$ thus provides a linear smoothing procedure which does not perfectly imitate the path properties of the original field X. In particular, we have $\hat{X}(t) \overset{d}{\neq} X(t),\ t \in W$.

Other prediction methods which yield forecasts that are equal in marginal distribution to X are e.g. conditional simulation and excursion metric prediction , see later sections of Chapter 3.

(b) Property (iv) has the following geometric interpretation. The simple Kriging predictor $\hat{X}(t_0) = \mathrm{Proj}_{L_N} X(t_0)$ is the orthogonal projection of the unobserved random variable $X(t_0)$ onto the linear subspace L_N formed by available observations $X(t_j)$, $j = 1, \dots, N$, i.e.

$$\mathrm{Proj}_{L_N} X(t_0) = \underset{Y \in L_N}{\mathrm{argmin}} \langle X(t_0) - Y, X(t_0) \rangle,$$

where $\langle \xi, \eta \rangle = \mathbb{E}\left[\xi \cdot \eta\right]$ for square-integrable random variables ξ, η.

(c) For Gaussian random fields X, the property

$$\mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2 \mid X(t_1), \dots, X(t_N)\right] = \mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] \quad a.s., \quad t_0 \in W,$$

shown in the proof of Theorem 3.2, Equation (3.7) is called *homoscedasticity*.

(d) Another property of Gaussian simple Kriging is the *conditional unbiasedness*, i.e.

$$\mathbb{E}\left[X(t_0) \mid \hat{X}(t_0)\right] = \hat{X}(t_0) \quad a.s., \quad t_0 \in W.$$

Indeed,

$$\begin{aligned}\mathbb{E}\left[X(t_0) \mid \hat{X}(t_0)\right] &= \mathbb{E}\left[\mathbb{E}[X(t_0) \mid \hat{X}(t_0)] \mid X(t_1, \ldots, X(t_N)\right] \\ &= \mathbb{E}[\underbrace{\mathbb{E}[X(t_0) \mid X(t_1), \ldots, X(t_N)]}_{=\hat{X}(t_0)} \mid \hat{X}(t_0)] = \hat{X}(t_0)\end{aligned}$$

by properties of the conditional expectation, since $\sigma(\hat{X}(t_0)) \subset \sigma(X(t_1), \ldots, X(t_N))$, where $\sigma(L)$ denotes the σ-algebra generated by a family of random variables L.

Remark 3.18 The practical application of the simple Kriging estimates to spatial data is tampered by the prerequisite to have an explicit knowledge of the mean value function $\mu(t)$ (also called *drift*) and the covariance function $C(s,t)$. Since both are in general unknown, they have to be inferred statistically from the available spatial data.

Estimation of the drift

It is assumed that the unobserved mean function $\mu(\cdot)$ can be decomposed into a series

$$\mu(t) = \sum_{j \in \mathbb{N}} \mu_j \Psi_j(t), \quad t \in T,$$

with respect to some orthonormal basis $\{\Psi_j\}_{j\in\mathbb{N}}$ in $L^2(T)$. Then, this series can be truncated at some detail level $M \in \mathbb{N}$ and the coefficients

$$\mu_j = \langle \mu, \Psi_j \rangle_2 = \int_{\mathbb{R}^d} \mu(t)\Psi_j(t)dt, \quad j = 1, \ldots, M$$

are estimated from the data. Ideally, many paths of X have to be observed for drift estimation, since the estimation of a non-constant mean value function based on only one single path is highly unreliable.

Denoting the estimates of μ_j by $\hat{\mu}_j$, we get

$$\hat{\mu}(t) = \sum_{j=1}^{M} \hat{\mu}_j \Psi_j(t).$$

The squared estimation error

$$\begin{aligned}\|\mu(t) - \hat{\mu}(t)\|_2^2 &= \| \sum_{j=1}^{M} (\mu_j - \hat{\mu}_j)\Psi(t) + \sum_{j=N+1}^{\infty} \mu_j \Psi_j(t)\|_2^2 \\ &\overset{(*)}{=} \sum_{j=1}^{M} (\mu_j - \hat{\mu}_j)^2 + \sum_{j=M+1}^{\infty} \mu_j^2,\end{aligned}$$

where $(*)$ follows from Parceval's identity, has to be kept minimal, thus a trade-off between the number of basis functions M and the quality of estimates $\hat{\mu}_j$, $j = 1, \ldots, M$, has to be accepted.

Common examples of bases in use include the Fourier basis, Wavelets, B-splines, etc. However, a simple Kriging based on an esimated drift $\hat{\mu}$ is prone to large errors. A way out would be the use of other non-stationary methods of geostatistics such as e.g. universal Kriging [51, Chapter 38].

Kriging with estimated covariance function

Since also the covariance function $C(s,t)$, $s,t \in T$, is unknown, an estimate from spatial data using inference methods from Section 2.2 is needed. However, the estimation result $\hat{C}$ is not positively semi-definite, so that its immediate use in the linear system of simple Kriging equations, see (3.16), is not recommended. The matrix $\hat{\Sigma} := (\hat{C}(t_j,t_k))_{j,k=1}^N$ is often singular or ill-conditioned. To avoid the issue of numerical instability, a parametric covariance model C_θ, $\theta \in \Theta \in \mathbb{R}$ is fitted to the estimator $\hat{C}$ such that the mean square error between $\hat{\Sigma}$ and $\Sigma_\theta = (C_\theta(t_j,t_k))_{j,k=1}^N$ is kept minimal, i.e.

$$\hat{\theta} = \underset{\theta\in\Theta}{\operatorname{argmin}} \sum_{j,k=1}^{N} (C_\theta(t_j,t_k) - \hat{C}(t_j,t_k))^2 \cdot w_{jk}, \tag{3.17}$$

where weights $w_{jk} \geq 0$ with $\sum_{j,k=1}^N w_{jk} = 1$ are often taken to be uniform, i.e. $w_{jk} = N^{-2}$. Then, the matrix $\Sigma_{\hat{\theta}} = (C_{\hat{\theta}}(t_j,t_k))_{j,k=1}^N$ and the vector $\hat{c}_0 = (C_{\hat{\theta}}(t_0,t_1),\dots,C_{\hat{\theta}}(t_0,t_N))^\intercal$ are used in (3.16) to compute the vector of simple Kriging weights

$$\lambda = \Sigma_{\hat{\theta}}^{-1} \cdot \hat{c}_0.$$

The choice of the parametric model C_θ is usually made after a visual inspection of the estimate $\hat{C}$ based on statistical experience [42, Section 2.1.4]. Since the estimate $\hat{C}$ often exhibits discontinuities at the origin, one might suggest to use a family of models displaying the so-called *nugget effect* σ_t^2 as well as *geometric anisotropy.*

Remark 3.19 (Nugget effect and geometric anisotropy): Let

$$C_\theta(s,t) = \sigma_t^2 \cdot I(s=t) + C^0(\sqrt{(s-t)^\intercal Q(s-t)}), \quad s,t \in T,$$

where $\sigma^2 > 0$, $t \in T$, $C^0(\cdot)$ is a covariance function of a motion invariant random field on $\mathbb{R}^d$, and Q is a positive definite $(d \times d)$-matrix responsible for anisotropy. This matrix can be chosen as

$$Q = R\Lambda R^\intercal,$$

where $\Lambda = \operatorname{diag}(\lambda_1,\dots,\lambda_d)$ is a diagonal matrix with diagonal entries $\lambda_1,\dots,\lambda_d > 0$, and R is a rotation matrix in $\mathbb{R}^d$ parametrized by Euler angles $\theta_1,\dots,\theta_{d-1}$.

If we assume for simplicity $\sigma_t^2 \equiv \sigma_0^2 > 0$, then the parameter vector θ with include σ^2, $\theta_1,\dots,\theta_{d-1},\lambda_1,\dots,\lambda_d$ as well as parameters of C^0.

Example 3.20 (Exponential model): Choose $d=2$, $C^0 : \mathbb{R}_+ \to \mathbb{R}_+$ with $C^0(x) = a \cdot e^{-|x|}$ and

$$Q = \begin{pmatrix} \cos(\theta_1) & -\sin(\theta_1) \\ \sin(\theta_1) & \cos(\theta_1) \end{pmatrix} \cdot \begin{pmatrix} \eta_1 & 0 \\ 0 & \eta_2 \end{pmatrix} \cdot \begin{pmatrix} \cos(\theta_1) & \sin(\theta_1) \\ -\sin(\theta_1) & \cos(\theta_1) \end{pmatrix}.$$

We get $C_\theta(s,t) = C_\theta^1(s-t)$, where $C_\theta(y) = \sigma_0^2 \cdot I(y=0) + a \exp\left\{-\sqrt{y^\intercal Q y}\right\}$, $y \in \mathbb{R}^2$, with $\theta = (\sigma_0^2, a, \theta_1, \lambda_1, \lambda_2)$.

Remark 3.21 Since the covariance estimates are non-reliable for high values of $\|s-t\|_2$, see Figure 3.1 and exhibit large oscillation artefacts (due to the simple fact that the number of pairs (t_j,t_k), $j,k=1,\dots,N$ with $\|t_j-t_k\|_2 \approx \operatorname{diag}(W) := \max\limits_{s,t\in W} \|s-t\|$ is rather small), it may be reasonable to punish these artefacts in (3.17) by taking non-uniform weigths w_{jk} that are nearly zero for such lags (j,k) of $C_\theta^1(y)$, $y = (y_1,y_2)^\intercal \in \mathbb{R}^2$ from Example 3.20.

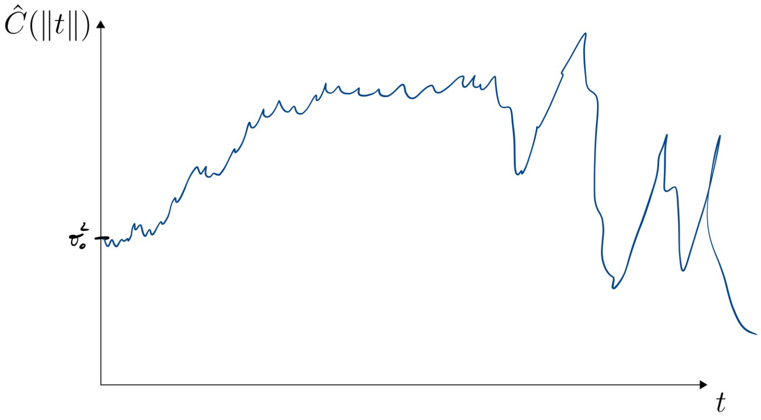


Fig. 3.1: An empirical isotropic covariance function $\hat{C}(\|t\|)$.

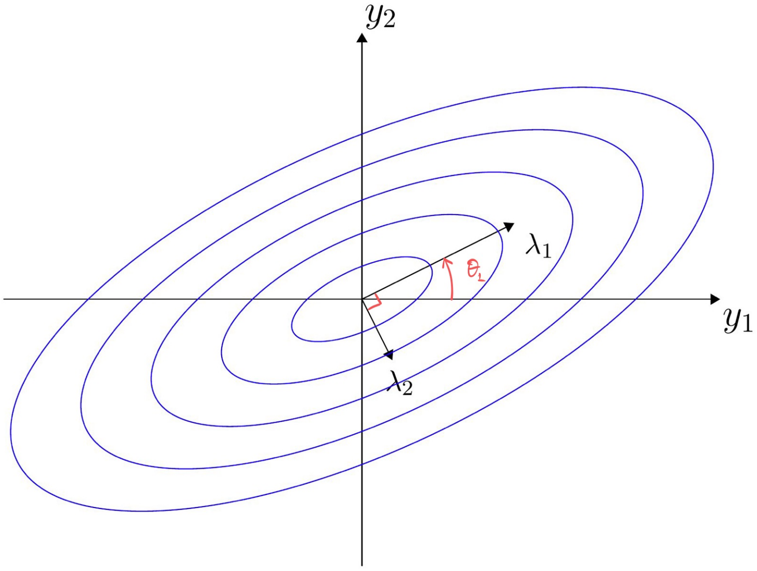


Fig. 3.2: Contour lines $\{y \in \mathbb{R}^2 : C^1_\theta(y) = \text{const}\}$ of $C^1_\theta(y)$, $y = (y_1, y_2)^\intercal \in \mathbb{R}^2$, from Example 3.20.

3.2.2 Ordinary Kriging

Assume that $X = \{X(t), t \in T\}$, $T \subset \mathbb{R}^d$, is a wide-sense stationary random field with mean $\mu = \mathbb{E}[X(0)]$ and covariance function $C(t) = \mathbf{cov}(X(0), X(t))$, $t \in T$, where the value of μ is unknown. Assuming the linear form of the forecast $\hat{X}(t)$ as in (3.13) and the minimization of the mean-square error (3.14), one sees that the unbiasedness of $\hat{X}(t)$ leads to

$$\mu = \mathbb{E}[X(t)] = \lambda_0 + \sum_{j=1}^{N} \lambda_j \underbrace{\mathbb{E}[X(t_j)]}_{=\mu} = \lambda_0 + \mu \sum_{j=1}^{N} \lambda_j,$$

which yields

$$\mu\left(1-\sum_{j=0}^{N}\lambda_j\right)=\lambda_0.$$

Since the explicit form of $\hat{X}(t)$ must not depend on μ, we conclude that $\lambda_0=0,\ \sum_{j=1}^{N}\lambda_j=1$ is the only possible choice.

Consequently, the *ordinary Kriging estimate* $\hat{X}(t_0)=\sum_{j=1}^{N}\lambda_j X(t_j),\ t_0\in W$ should satisfy

$$\mathcal{E}(\lambda)=\mathbf{var}\left(\hat{X}(t_0)-X(t_0)\right)=\mathbb{E}\left[(\hat{X}(t_0)-X(t_0))^2\right]\to\min_{\lambda=(\lambda_1,\ldots,\lambda_N)^\intercal}$$

such that $\sum_{j=1}^{N}\lambda_j=1$, where the function $\mathcal{E}(\lambda)$ is given by Equation (3.15), i.e.

$$\mathcal{E}(\lambda)=\sum_{j,k=1}^{N}\lambda_j\lambda_k C(t_j-t_k)-2\sum_{j=1}^{N}\lambda_j\cdot C(t_0-t_j)+\underbrace{C(0)}_{=\sigma^2}=\lambda^\intercal\Sigma\lambda-2\lambda^\intercal C_0+\sigma^2.$$

The quantities $\Sigma,\ c_0$ are defined in Section 3.2.1.

Consider the Lagrange function of the constraint minimization problem

$$\begin{cases}\lambda^\intercal\Sigma\lambda-2\lambda^\intercal c_0+\sigma^2\to\min\limits_{\lambda\in\mathbb{R}^N},\\ \lambda^\intercal\cdot e=1,\end{cases}\tag{3.18}$$

where $e=(1,\ldots,1)\in\mathbb{R}^N$. Then,

$$L(\lambda,\gamma)=\lambda^\intercal\Sigma\lambda-2\lambda^\intercal\cdot c_0+\sigma^2+2\beta(\lambda^\intercal\cdot e-1),\quad \beta\in\mathbb{R},$$

is the so-called *Lagrange multiplier*.

Taking partial derivates of $L(\lambda,\gamma)$ with respect to λ_j,β and setting them equal to zero, we obtain the following system of linear equations in λ,β,

$$\begin{cases}\frac{\partial L}{\partial\lambda_j}=\sum_{k=1}^{N}C(t_j-t_k)+\beta-C(t_0-t_j)=0,\quad j=1,\ldots,N,\\ \frac{\partial L}{\partial\gamma}=\sum_{j=1}^{N}\lambda-1=0,\end{cases}\tag{3.19}$$

or in matrix form

$$\begin{cases}\Sigma\lambda=c_0-\beta\cdot e,\\ \lambda^\intercal\cdot e=1.\end{cases}\tag{3.20}$$

The relation $\gamma(t)=C(0)-C(t)$ connects the covariance C to the variogram γ of X. One can easily rewrite (3.19) and(3.20) in a form, which is common in geostatistical literature, i.e.

$$\begin{cases}\sum_{k=1}^{N}\lambda_k\gamma(t_j-t_k)-\beta=\gamma(t_0-t_j),\quad j=1,\ldots,N,\\ \sum_{j=1}^{N}\lambda_j=1,\end{cases}\tag{3.21}$$

or in matrix form

$$\begin{cases}\Gamma\cdot\lambda=\beta\cdot e+\Gamma_0,\\ \lambda^\intercal\cdot e=1,\end{cases}\tag{3.22}$$

where $\Gamma=(\gamma(t_j-t_k))_{j,k=1}^{N}$ and $\Gamma_0=(\gamma(t_0-t_1),\ldots,\gamma(t_0-t_N))^\intercal$.

Theorem 3.22 If the matrix Σ (or Γ) is non-singular, then the constrained optimization problem (3.18) has the unique solution

$$\lambda = \Gamma^{-1}(\beta \cdot e + \Gamma_0), \tag{3.23}$$

where

$$\beta = \frac{1 - e^\intercal \Gamma^{-1}\Gamma_0}{e^\intercal \Gamma^{-1} e}. \tag{3.24}$$

The corresponding mean square error of the ordinary Kriging is given by

$$\sigma_{OK}^2 := \mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] = \lambda^\intercal \Gamma_0 - \beta \tag{3.25}$$

with λ, β as above in (3.23) and (3.24).

Proof Relation (3.24) is evident. To show (3.23) multiply λ in (3.24) with e and set it equal to 1, i.e.

$$1 = e^\intercal \lambda = \beta e^\intercal \Gamma^{-1} e + e^\intercal \Gamma^{-1}\Gamma_0,$$

hence

$$\beta = \frac{1 - e^\intercal \Gamma^{-1}\Gamma_0}{e^\intercal \Gamma^{-1} e},$$

which proves (3.24). To show (3.25) rewrite $\mathcal{E}(\lambda)$ in terms of Γ. It follows

$$\sigma_{OK}^2 = \mathcal{E}(\lambda) = -\lambda^\intercal \Gamma \lambda + 2C(0) - 2\lambda^\intercal c_0 = -\lambda^\intercal \Gamma \lambda + 2\lambda^\intercal \Gamma_0.$$

Here, we used the relations $\lambda^\intercal e = 1$ and $\gamma(t) = C(0) - C(t)$. Plugging (3.22) into the above yields

$$\sigma_{OK}^2 = -\lambda^\intercal \underbrace{(\beta e + \Gamma_0)}_{=\Gamma\lambda} + 2\lambda^\intercal \Gamma_0 = -\beta \underbrace{\lambda^\intercal e}_{=1} - \lambda^\intercal \Gamma_0 + 2\lambda^\intercal \cdot \Gamma_0 = \lambda^\intercal \Gamma_0 - \beta.$$

□

We may also give σ_{OK}^2 explicitly

$$\begin{aligned}
\sigma_{OK}^2 &= \Gamma_0^\intercal \cdot \lambda - \beta \\
&= \Gamma_0^\intercal \cdot \Gamma^{-1} \cdot \Gamma_0 + \beta \left(\Gamma_0^\intercal \cdot \Gamma^{-1} \cdot e - 1\right) \\
&= \Gamma_0^\intercal \cdot \Gamma^{-1} \cdot \Gamma_0 - \beta \left(1 - e^\intercal \cdot \Gamma^{-1} \cdot \Gamma_0\right) \\
&= \Gamma_0^\intercal \cdot \Gamma^{-1} \cdot \Gamma_0 - \frac{(1 - e^\intercal \cdot \Gamma^{-1} \cdot \Gamma_0)^2}{e^\intercal \Gamma^{-1} e}.
\end{aligned}$$

Remark 3.23 An advantage of expressing the ordinary Kriging system as in (3.22) is that it is also applicable to intrinsically stationary random fields X of order 2, i.e., fields that may not possess a finite variance, but have wide-sense stationary increments.

Theorem 3.24 (Properties of ordinary Kriging): For an intrinsically stationary random field $X = \{X(t), t \in T\}$ with observations $X(t_j)$, $j = 1, \dots, N$ and variogram $\gamma(\cdot)$ such that the matrix $\Gamma = (\gamma(t_k - t_j))_{k,j=1}^N$ is invertible, the oridinary Kriging predictor $\hat{X}(t) = \sum_{j=1}^N \lambda_j \cdot X(t_j)$ with $\lambda = (\lambda_1, \dots, \lambda_N)^\intercal$ satisfying (3.23) possesses the following properties.

(i) *Exactness:* $\hat{X}(t_j) = X(t_j)$ a.s., $j = 1, \dots, N$.

(ii) *Orthogonality:* For any $Y \in \left\{\sum_{j=1}^N a_j X(t_j) : \sum_{j=1}^N a_j = 0\right\} = \tilde{L}_N$ it holds $\langle \hat{X}(t_0) - X(t_0), Y\rangle_2 = 0$ for all $t_0 \in W$.

(iii) *Conditional bias reduction:* For all $t_0 \in W$ it holds that

$$\mathbb{E}\left[\left(\mathbb{E}\left[X(t_0) \mid \hat{X}(t_0)\right] - \hat{X}(t_0)\right)^2\right] = \mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] - \mathbb{E}\left[\mathbf{var}\left(X(t_0) \mid \hat{X}(t_0)\right)\right]. \tag{3.26}$$

Proof (1) Set $t_0 = t_j$ for some $j \in \{1, \dots, N\}$ and show that $\lambda = (0, \dots, 0, \overset{j}{1}, 0, \dots, 0)$ is a solution of (3.22), where $\beta = 0$ and

$$\Gamma_0 = (\gamma(t_j - t_1), \dots, \underbrace{\gamma(t_j - t_j)}_{=\gamma(0)=0}, \dots, \gamma(t_j - t_N))^\intercal.$$

Indeed, we have $\beta = 0$ by (3.24) since $e^\intercal \Gamma^{-1}\Gamma_0 = e^\intercal(0, \dots, 0, 1, 0, \dots, 0) = 1$ by definition of the inverse matrix Γ^{-1}, while Γ_0 is the j-th column of Γ. Then, the system (3.22) reduces to $\Gamma \cdot \lambda = \Gamma_0$, which holds evidently by the explicit form of Γ, λ and Γ_0.

(2) Let $Y \in \tilde{L}_N$, i.e. $Y = \sum_{j=1}^N a_j \cdot X(t_j)$ with $\sum_{j=1}^N a_j = 0$. Then, with $\lambda_0 = -1$ we have

$$\begin{aligned}
&\mathbb{E}\left[Y \cdot \left(\hat{X}(t_0) - X(t_0)\right)\right] \\
&= \mathbb{E}\left[\sum_{j=1}^N a_j X(t_j) \cdot \sum_{k=0}^N \lambda_k X(t_k)\right] = \sum_{j=1}^N \sum_{k=0}^N a_j \lambda_k \cdot \mathbb{E}[X(t_j)X(t_k)] \\
&= \sum_{j,k=1}^N a_j \lambda_k \cdot \underbrace{\left(\mathbb{E}[X(t_j)X(t_k)] - \mu^2\right)}_{=C(t_j - t_k)} + \mu^2 \cdot 0 - \sum_{j=1}^N a_j \underbrace{\mathbb{E}[X(t_0)X(t_j)] - \mu^2}_{=C(t_0 - t_j)} - \mu^2 \cdot 0 \\
&= \sum_{j=1}^N a_j \Bigg(\underbrace{\sum_{k=1}^N \lambda_k \cdot C(t_j - t_k) - C(t_0 - t_j)}_{=-\beta \text{ by } (3.19)} \Bigg) = -\beta \underbrace{\sum_{j=1}^N a_j}_{=0} = 0.
\end{aligned}$$

due to $\sum\limits_{j=1}^N a_j = 0$ and $\sum\limits_{j=1}^N \lambda_j = 1$

(3) To prove relation (3.26), we apply the *law of total variance*

$$\mathbf{var}(Y) = \mathbf{var}(\mathbb{E}[Y \mid Z]) + \mathbb{E}[\mathbf{var}(Y \mid Z)] \tag{3.27}$$

for any square-integrable random variables Y, Z. We may write for $Z = \hat{X}(t_0), Y =$

$\hat{X}(t_0) - X(t_0)$ that

$$
\begin{aligned}
&\mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] \overset{(*)}{=} \mathbf{var}(\hat{X}(t_0) - X(t_0)) \\
&= \mathbb{E}\left[\mathbf{var}(\hat{X}(t_0) - X(t_0) \mid \hat{X}(t_0)\right] + \underbrace{\mathbf{var}\left(\mathbb{E}\left[\hat{X}(t_0) - X(t_0) \mid \hat{X}(t_0)\right]\right)}_{=\mathbb{E}\bigl[(\hat{X}(t_0)-\mathbb{E}[X(t_0)|\hat{X}(t_0))^2\bigr]-\Bigl(\underbrace{\mathbb{E}[\hat{X}(t_0)]}_{=\mu}-\underbrace{\mathbb{E}[\mathbb{E}[X(t_0)\mid\hat{X}(t_0)]]}_{=\mathbb{E}X(t_0)=\mu}\Bigr)^2} \\
&= \mathbb{E}\Bigl[\mathbb{E}\Bigl[X(t_0) \mid \hat{X}(t_0)\Bigr] - \hat{X}(t_0)\Bigr]^2 + \mathbb{E}\Biggl[\underbrace{\mathbb{E}\Bigl[[\hat{X}(t_0) - X(t_0) - \mathbb{E}(\hat{X}(t_0) - X(t_0) \mid \hat{X}(t_0)]^2 \mid \hat{X}(t_0)\Bigr]}_{=\mathbf{var}(X(t_0)|\hat{X}(t_0)}\Biggr] \\
&= \mathbb{E}\left[\left(\mathbb{E}[X(t_0) \mid \hat{X}(t_0)] - \hat{X}(t_0)\right)^2\right] + \mathbb{E}\left[\mathbf{var}(X(t_0) \mid \hat{X}(t_0)\right],
\end{aligned}
$$

where $(*)$ follows from $\mathbb{E}[\hat{X}(t_0)] = \mathbb{E}[X(t_0)] = \mu$.

□

Remark 3.25 (a) The shrinkage property $\mathbf{var}(X(t_0)) \geq \mathbf{var}(\hat{X}(t_0))$ of simple Kriging does in general not hold for ordinary Kriging anymore. Indeed, we have

$$
\begin{aligned}
0 \leq \mathcal{E}(\lambda) &= \mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] \\
&\overset{(3.17)}{=} \underbrace{\lambda^\intercal \Sigma \lambda}_{=\mathbf{var}(\hat{X}(t_0))} - 2\lambda^\intercal c_0 + \underbrace{\sigma^2}_{=\mathbf{var}(X(t_0))} \\
&\overset{(3.20)}{=} \lambda^\intercal \Sigma \lambda(t_0) - 2\lambda^\intercal \Sigma \lambda - 2\beta \underbrace{\lambda^\intercal e}_{=1} + \sigma^2 \\
&= \sigma^2 - \lambda^\intercal \Sigma \lambda - 2\beta
\end{aligned}
$$

hence

$$\mathbf{var}(X(t_0)) \geq \mathbf{var}(\hat{X}(t_0)) + 2\beta,$$

whereas the β given in (3.24) belongs to $\mathbb{R}$, i.e. it can be ≥ 0 as well as < 0. It follows that $\mathbf{var}(X(t_0)) \geq \mathbf{var}(\hat{X}(t_0))$ if $\beta \geq 0$.

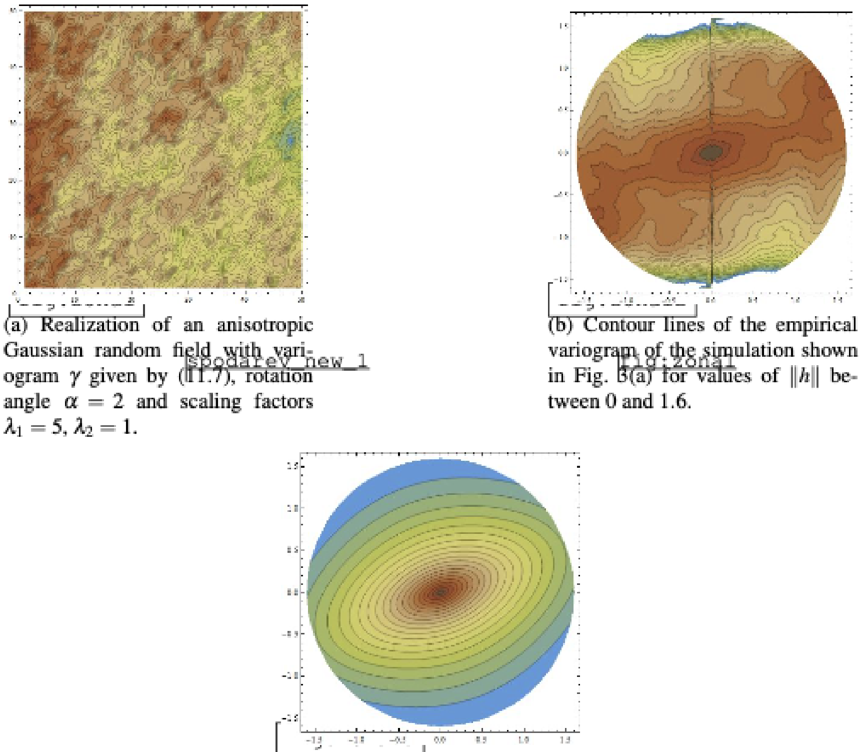
(b) The smoothness property holds for ordinary Kriging as follows. If $C \in C^k(T)$ or $\gamma \in C^k(T)$, $k \in \mathbb{N}_0$, then so is $\hat{X} \in C^k(T)$ a.s. This can be seen from Equations (3.23)-(3.26). It holds that

$$\lambda = \lambda(t_0) = \beta \cdot \Gamma^{-1} e + \Gamma^{-1}\Gamma_0 = \frac{(1 - e^\intercal \cdot \Gamma^{-1} \cdot \Gamma_0)}{e^\intercal \Gamma^{-1} e} \Gamma^{-1} e + \Gamma^{-1}\Gamma_0,$$

where only the term Γ_0 depends on t_0, and so λ inherits the smoothness of γ.

(c) The remarks about Kriging with estimated covariance function hold for ordinary Kriging as well. However, it is more common to estimate the variogram γ of X, fit a valid parametric model γ_θ to $\hat{\gamma}$ via weighted least squares and solve the system of linear equations (3.22) as

$$
\begin{cases}
\Gamma_{\hat{\theta}} \cdot \lambda = \beta e + \Gamma_{\hat{\theta},0}, \\
\lambda^\intercal e = 1,
\end{cases}
$$



(a) Realization of an anisotropic Gaussian random field with variogram γ given by (11.7), rotation angle $\alpha = 2$ and scaling factors $\lambda_1 = 5, \lambda_2 = 1$.

(b) Contour lines of the empirical variogram of the simulation shown in Fig. 3(a) for values of $\|h\|$ between 0 and 1.6.

(c) Theoretical variogram function $\gamma = \gamma_1 + \gamma_2$.

Fig. 3.3: Realization of an anisotropic Gaussian random field with the corresponding empirical and theoretical variogram from Example 3.26.

where $\Gamma_{\hat{\theta}} = (\gamma_{\hat{\theta}}(t_j - t_k))_{j,k=1}^N$, $\Gamma_{\hat{\theta},0} = (\gamma_{\hat{\theta}}(t_0 - t_1), \dots, \gamma_{\hat{\theta}}(t_0 - t_N))^\intercal$, and $\{\gamma_\theta, \theta \in \Theta\}$ is a parametric family of conditionally negative-definite functions [42, Proposition 2.2.1] with

$$\hat{\theta} = \operatorname*{argmin}_{\theta \in \Theta} \sum_{j,k=1}^N w_{jk}(\hat{\gamma}(t_j - t_k) - \gamma_\theta(t_j - t_k))^2.$$

In the construction of γ_θ, a nugget effect and geometric anisotropy can be transferred from the covariance functions via the relation $\gamma(h) = C(0) - C(h)$, accordingly.

Example 3.26 Let $X = \{X(t), t \in \mathbb{R}^2\}$ be a stationary anisotropic Gaussian random field ($d = 2$) with variogram

$$\gamma(t) = \gamma_1(t) + \gamma_2(t), \quad t \in \mathbb{R}^2,$$

where

$$\gamma_1(t) = 1 - e^{-\|t\|_2} \quad \text{and} \quad \gamma_2(t) = 1 - \exp -\frac{\sqrt{t^\intercal Q t}}{5},$$

and the matrix Q is given as in Example 3.20 with $\theta_1 = 114,59°$, $\eta_1 = 5, \eta_2 = 1$. A simulated realization of X, an estimate $\hat{\gamma}$ of γ as well as γ itself are given in Figure 3.3.

Example 3.27 Let $X = \{X(t), t \in \mathbb{R}^2\}$ be a centered motion invariant Gaussian random field observed in a window $W = [0, 10]^2$ on the grid $\{(3j, 2k),\ j, k \in [0, 3] \cap \mathbb{N}_0\}$. Let X have the

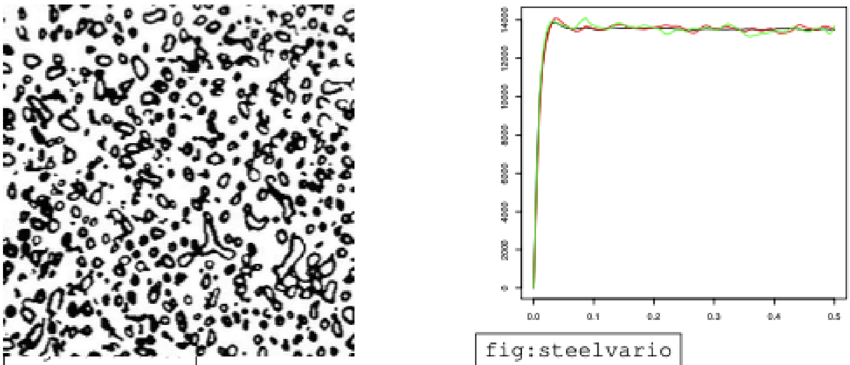


fig:steelvario

(a) Microscopic image of a steel surface.

(b) Estimates for the x-direction (red), y-direction (green), all directions (black) for values $0 \leq h \leq 0.5$.

Fig. 3.4: Microscopic steel image (left) and its empirical variogram estimated in different directions (right).

Whittle-Matern-type covariance function with nugget effect, i.e. we have

$$\mathbf{cov}(X(0), X(t)) = C(t) = 2 \cdot \mathbf{1}(t = 0) + 2\kappa_1(2\|t\|_2) \cdot \|t\|_2 \cdot \mathbf{1}(t \neq 0), \quad t \in \mathbb{R}^2,$$

where

$$\kappa_n(x) = \lim_{\nu \to n} \frac{\pi}{2\sin(\pi\nu)} \left(e^{i\frac{\pi}{2}\nu} \cdot J_{-\nu}(xe^{i\frac{\pi}{2}}) - e^{-i\frac{\pi}{2}\nu} \cdot J_{\nu}(xe^{-i\frac{\pi}{2}}) \right), \quad x \in \mathbb{R}, \quad n \in \mathbb{N},$$

is the *modified Bessel function of the third kind*, and

$$J_\nu(x) = \sum_{r=0}^{\infty} \frac{(-1)^r (\frac{x}{2})^{2r+\nu}}{\Gamma(\nu + r + 1) r!}, \quad x \in \mathbb{C}, \quad \nu \in \mathbb{R},$$

is the *Bessel function of the first kind of order ν*.

After estimating the variogram of X, see Section 2.2, by $\hat{\gamma}$ from a realisation of X given in Figure 3.5(a), a *Whittle-Matern-type family of variogram models*

$$\gamma_\theta(t) = \mathbf{1}(t \neq 0) \left[\sigma^2 + b - b \cdot 2^{1-\nu} (a\|t\|_2)^\nu \kappa_\nu(a\|t\|_2) \right], \quad t \in \mathbb{R}^2,$$

with $\theta = (b, \nu, a, b)$ is fitted to $\hat{\gamma}$ by ordinary least squares, see Figure 3.5(c) (true variogram γ is in red, $\hat{\gamma}$ in green, and $\gamma_{\hat{\theta}}$ in black with the estimate $\hat{\theta} = (0.933, 1, 1.967, 1.067)$). Then, the ordinary Kriging is performed with $\gamma_{\hat{\theta}}$, its result being shown in Figure 3.5(b). As it is seen, the Kriging result $\hat{X}$ smooths out the rough surface of X.

3.2.3 Universal Kriging

Assume that $X = \{X(t), t \in T\}$, $T \subset \mathbb{R}^d$ is a non-stationary random field with drift $\mu(t) = \mathbb{E}[X(t)], t \in T$, where

$$\mu(t) = \sum_{j=0}^{M} \mu_j \Psi_j(t)$$

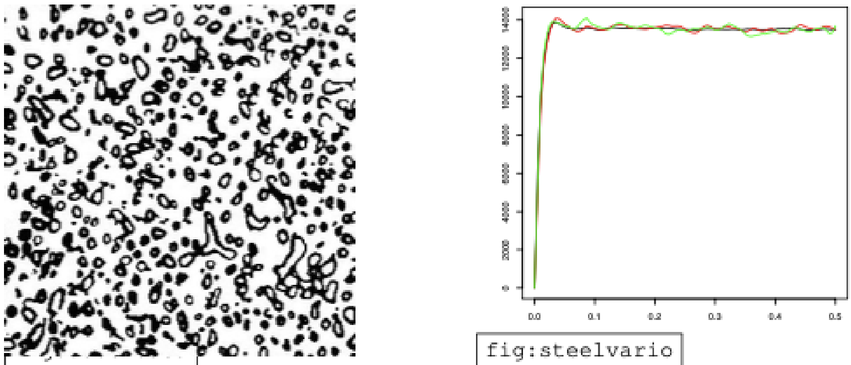


fig:steelvario

(a) Microscopic image of a steel surface.

(b) Estimates for the x-direction (red), y-direction (green), all directions (black) for values $0 \le h \le 0.5$.

Fig. 3.5: Application of ordinary Kriging to simulated data from Example 3.27

is a finite sum of orthonormal basis functions $\{\Psi_j, j \in \mathbb{N}_0\}$, $\Psi_0(t) \equiv 1$ as in Remark 3.18. Let the random field $Y = \{Y(t) = X(t) - \mu(t), t \in T\}$ of residuals be wide-sense stationary with covariance function $C(t) = \mathbf{cov}(Y(0), Y(t)), t \in T$.

As in Sections 3.2.1 and 3.2.2, we are looking for a linear predictor of the form

$$\hat{X}(t_0) = \sum_{j=1}^{N} \lambda_j X(t_j)$$

for $t_0 \notin \{t_1, \ldots, t_N\}$, where $X(t_j), j = 1, \ldots, N$, are a sample of observed values, subject to $\mathbb{E}[\hat{X}(t_0)] = \mathbb{E}[X(t_0)] = \mu(t_0)$. It follows that

$$\sum_{k=0}^{M} \mu_k \Psi_k(t_0) = \mu(t_0) = \sum_{j=1}^{N} \lambda_j \mu(t_j) = \sum_{k=0}^{M} \mu_k \Psi_k(t),$$

which yields

$$\sum_{k=0}^{M} \mu_k \left(\Psi_k(t_0) - \sum_{j=1}^{N} \lambda_j \Psi_k(t_j) \right) = 0.$$

Since μ_k are non-zero, the above equation is a source of the so-called *universality constraints*

$$\sum_{j=1}^{N} \lambda_j \Psi_k(t_j) = \Psi_k(t_0), \quad k = 0, \ldots, M. \tag{3.28}$$

For $k = 0$ the condition $\sum_{j=1}^{N} \lambda_j = 1$ known for ordinary Kriging appears. Minimizing the target function

$$\mathcal{E}(\lambda) = \mathbb{E}\left[\left(X(t_0) - \hat{X}(t_0) \right)^2 \right]$$

subject to the universality constraints with respect to $\lambda = (\lambda_1, \ldots, \lambda_N)$ via the *Lagrange function*

$$L(\lambda, \beta) = \mathcal{E}(\lambda) + \sum_{k=0}^{M} \beta_k \left(\sum_{j=1}^{N} \lambda_j \Psi_k(t_j) - \Psi_k(t_0) \right)$$

with the *Lagrange multipliers* $\beta_0, \dots, \beta_M$ leads to the *system of linear equations for universal Kriging* (compare (3.19) - (3.20))

$$\begin{cases} \sum_{j=1}^N \lambda_j C(t_i - t_j) - \sum_{k=0}^M \beta_k \Psi_k(t_i) = C(t_i - t_0), & i = 1, \dots, N, \\ \sum_{j=1}^N \lambda_j \Psi_k(t_j) = \Psi_k(t_0), & k = 0, \dots, M, \end{cases}$$

or, in matrix form,

$$\begin{pmatrix} \Sigma & \bar{\Psi} \\ \bar{\Psi}^\intercal & \bar{0} \end{pmatrix} \begin{pmatrix} \lambda \\ -\beta \end{pmatrix} = \begin{pmatrix} c_0 \\ \Psi^0 \end{pmatrix}, \tag{3.29}$$

where

$$\begin{aligned} \Sigma &= (C(t_i - t_j))_{i,j=1}^N, \\ \bar{\Psi} &= (\Psi_k(t_i))_{i=1}^N{}_{k=0}^M, \\ c_0 &= (C(t_1 - t_0), \dots, C(t_N - t_0))^\intercal, \\ \Psi^0 &= (\Psi_0(t_0), \dots, \Psi_M(t_0))^\intercal. \end{aligned}$$

Lemma 3.28 If C is positive definite, then there exists a unique solution for the system of linear equations (3.29).

Proof Since is Σ is invertible, the whole matrix $\begin{pmatrix} \Sigma & \bar{\Psi} \\ \bar{\Psi}^\intercal & \bar{0} \end{pmatrix}$ is invertible if the matrix $\bar{\Psi}$ has full rank, i.e. if its columns $(\Psi_k(t_1), \dots, \Psi_k(t_N))^\intercal$, $k = 0, \dots, N$, are linearly independent. This is true since $\{\Psi_k\}_{k=0}^\infty$ is an orthonormal basis. □

Remark 3.29 To solve the system (3.29), we assumed that C and function Ψ_j are explicitly known. However, in practice the function C has to be estimated from the data $X(t_1), \dots, X(t_N)$, which is e.g. possible by inferring the covariance of estimated residuals $Y^*(t_j) = X(t_j) - \hat{\mu}(t_j)$, where $\hat{\mu}(\cdot)$ is the estimated drift. This is a source of additional bias to the universal Kriging [51, p.303-307].

Remark 3.30 The drift estimation previously mentioned in Remark 3.18 can be practically exercised using an approach very similar to (3.29). In the expression $\mu(t) = \sum_{k=0}^M \mu_k \Psi_k(t)$, assume that μ_k themselves are realizations of random variables M_k and estimated via

$$\hat{\mu}_k = \sum_{j=1}^N \alpha_j^{(k)} X(t_j), \quad k = 0, \dots, M.$$

The unbiasedness of the estimator, i.e. $\mathbb{E}[\hat{\mu}_k] = \mathbb{E}[M_k]$, $k = 0, \dots, M$, yields the constraints

$$\sum_{j=1}^N \alpha_j^{(k)} \Psi_l(t_j) = \delta_{kl}, \quad k, l = 0, \dots, M. \tag{3.30}$$

Indeed, similar to (3.28), we may write

$$\mu_k = \mathbb{E}\left[\hat{\mu}_k\right] = \sum_{j=1}^N \alpha_j^{(k)} \mathbb{E}[X(t_j)] = \sum_{j=1}^N \alpha_j^{(k)} \mu(t_j) = \sum_{j=1}^N \alpha_j^{(k)} \sum_{l=0}^M \mu_l \Psi_l(t_j) = \sum_{l=0}^M \sum_{j=1}^N \alpha_j^{(k)} \Psi_l(t_j) \mu_l,$$

where

$$\delta_{kl} = \begin{cases} 1, & k = l, \\ 0, & k \neq l. \end{cases}$$

Solving the minimization problem

$$\mathbb{E}\left[(\hat{\mu}_k - \mu_k)^2\right] \to \min_{\alpha_1^{(k)},\ldots,\alpha_N^{(k)}}, \quad k = 0, \ldots, M,$$

subject to the constraints (3.30) via the Lagrange formalism leads to the system of linear equations

$$\begin{cases} \sum_{j=1}^{N} \alpha_j^{(k)} C(t_i - t_j) - \sum_{l=1}^{M} \beta_j^{(k)} \Psi_l(t_i) = 0, & i = 1, \ldots, N, \\ \sum_{j=1}^{N} \alpha_j^{(k)} \Psi_l(t_j) = \delta_{kl}, & l = 0, \ldots, M, \; k = 0, \ldots, M, \end{cases} \tag{3.31}$$

where $\beta_l^{(k)}$, $k, l = 0, \ldots, M$ are the *Lagrange multipliers*, see Equation (3.29), which can be shown to be

$$\beta_l^{(k)} = \mathbf{cov}(\hat{\mu}_k, \hat{\mu}_l), \quad k, l = 0, \ldots, M.$$

Exercise 3.31 Check the above.

The *estimated drift* is then given by

$$\hat{\mu}(t_0) = \sum_{k=0}^{M} \hat{\mu}_k \Psi_k(t_0) = \sum_{k=0}^{M} \sum_{j=1}^{N} \alpha_j^{(k)} X(t_j) \Psi_k(t_0).$$

However, the uncertainty of estimating C as mentioned in Remark 3.29 is still present. It can be also shown that the *drift estimation variance* equals

$$\mathbb{E}\left[(\hat{\mu}(t_0) - \mu(t_0))^2\right] = \Psi^{0\mathsf{T}} \cdot \left(\hat{\Psi}^{\mathsf{T}} \cdot \Sigma^{-1} \cdot \hat{\Psi}\right)^{-1} \cdot \Psi^0$$

using the same notation as in (3.29). More on drift estimation can be found in [4, Section 3.4.5, 3.4.6].

Remark 3.32 (a) The *variance of universal Kriging* is equal to

$$\sigma_{UK}^2 = \mathbb{E}\left[\left(\hat{X}(t_0) - X(t_0)\right)^2\right] = C(0) - \lambda^{\mathsf{T}} \cdot c_0 + \beta^{\mathsf{T}} \cdot \Psi^0,$$

where $(\lambda, -\beta)$ is the solution of the system (3.29).

(b) Similarly to ordinary Kriging , see Equation (3.22), one can rewrite the system of linear equations in (3.29) in terms of the variogram $\gamma(t) = \frac{1}{2}\mathbb{E}[Y(0) - Y(t)]^2$ of the residual random field Y. Formally, the values $C(t_i - t_j)$, $i, j = 0, \ldots, N$, sould be replaced there by $-\gamma(t_i - t_j)$.

(c) *Additivity property:* The universal Kriging predictor $\hat{X}(t_0) = \sum_{j=1}^{N} \lambda_j X(t_j)$ can be decomposed into

$$\hat{X}(t_0) = \hat{X}_{SK}(t_0) + \hat{X}^*(t_0), \tag{3.32}$$

where

(i) $\hat{X}_{SK}(t_0) = \sum_{j=1}^{N} \lambda_j^{SK} X(t_j)$ is the simple Kriging predictor of $X(t_0)$ as if X were centered, i.e. after the substraction of the "known" mean from the data, i.e. $\lambda^{SK} = (\lambda_1^{SK}, \dots, \lambda_N^{SK})^\intercal = \Sigma^{-1} c_0$, compare (3.16).

(ii) The term

$$\hat{X}^*(t_0) = \hat{\mu}(t_0) - \sum_{j=1}^{N} \lambda_j^{SK} \hat{\mu}(t_j) \tag{3.33}$$

is the *drift correction*, where the estimated drift $\hat{\mu}$ is given in Remark 3.30.

Combining (i) and (ii) yields

$$\hat{X}(t_0) = \hat{\mu}(t_0) + \sum_{j=1}^{N} (X(t_j) - \hat{\mu}(t_j)).$$

Indeed, substracting $\Sigma\lambda^{SK} = c_0$ from the system $\Sigma\lambda - \bar{\Psi}\beta = c_0$ out of (3.29) yields $\Sigma(\lambda - \lambda^{SK}) - \bar{\Psi}\beta = 0$.

Let $\lambda^D := \lambda - \lambda^{SK}$. Then ,the above relation together with the second equation $\bar{\Psi}^\intercal \lambda = \Psi^0$ of (3.29) rewrites as

$$\begin{cases} \Sigma \cdot \lambda^D - \bar{\Psi} \cdot \beta = 0, \\ \bar{\Psi}^\intercal \cdot \lambda = \Psi^0 - \bar{\Psi}^\intercal \cdot \lambda^{SK}, \end{cases}$$

where the first equation coincides with the first one from (3.31) and both give birth to the correction term $\hat{X}^*(t_0)$.

Remark 3.33 (Further properties of universal Kriging): Similar to Theorem 3.24, the properties of *exactness, orthogonality* and *conditional bias reduction* hold for universal kriging as follows.

(a) $\hat{X}(t_j) = X(t_j)$ a.s., since $\lambda = (0, \dots, 0, 1, 0, \dots, 0)$ and $\beta = (0, \dots, 0)$ yield $\sigma_{UK}^2 = 0$, compare Remark 3.32.

(b) For any

$$Y \in \left\{ \sum_{j=1}^{N} a_j X(t_j) : \sum_{j=1}^{N} a_j \Psi_k(t_j) = 0,\ k = 0, 1, \dots, M \right\}$$

it holds that

$$\left\langle \hat{X}(t_0) - X(t_0), Y \right\rangle_2 = 0, \quad t_0 \in W.$$

(c) Equation (3.26) is still valid, which means that $\lambda = (\lambda_1, \dots, \lambda_N)$ minimizing $\mathbb{E}[(\hat{X}(t_0) - X(t_0))^2]$ also reduces the conditional bias $\mathbb{E}[X(t_0) \mid \hat{X}(t_0)] - \hat{X}(t_0)$.

3.3 Geoadditive regression models

Prediction of spatial phenomena can occur, beyond random fields, also in the context of classical (non)-linear regression. For that, consider the following setting. Let Y_i be absolutely continuous *target* or *response* random variables. Assume the regression model

$$Y_i = \sum_{j=0}^{k} \beta_j x_{ij} + \sum_{j=1}^{m} g_j(z_{ij}) + f_{\text{geo}}(t_i) + \varepsilon_i, \quad i = 1, \dots, n, \tag{3.34}$$

where

(i) $\beta^\intercal x_i = \sum_{j=0}^{k} \beta_j x_{ij}$ with parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\intercal$ and covariates $x_i = (1, x_{i1}, \ldots, x_{ik})^\intercal$, where $x_{i0} = 1$ is the *linear part* and $\beta, x_i \in \mathbb{R}^{k+1}$, $i = 1, \ldots, k$.

(ii) $z_{i1}, \ldots, z_{im}$, $i = 1, \ldots, n$ are continuous covariates, and the unknown functions $g_1, \ldots, g_m$ are smooth enough and all together form the *non-linear additive part*.

(iii) The unknown smooth function $f_{\text{geo}} : \mathbb{R}^d \to \mathbb{R}$ contains georeferenced information provided at *spatial locations* $t_i \in \mathbb{R}^d$, $i = 1, \ldots, n$

(iv) The random variable ε_i is the *regression error* with $\mathbb{E}\varepsilon_i = 0$ and $\mathbb{E}\varepsilon_i^2 = \sigma^2 > 0$. It is usually assumed that ε_i are uncorrelated or even stochastically independent. Sometimes, Gaussian errors are common, i.e. $\varepsilon_i \sim N(0, \sigma^2)$.

The regression model in Equation (3.34) is called *geoadditive*. Its purpose is to yield estimates for $\beta, g_j, f_{\text{geo}}$ given the data $Y = (Y_1, \ldots, Y_n), Z = (z_{ij})_{i=1}^{n}{}_{j=1}^{m}, X = (x_{ij})_{i=1}^{n}{}_{j=1}^{k}, t = (t_1, \ldots, t_n)^\intercal \in \mathbb{R}^{n \times d}$.

The spatial locations t_i can either attain a finite number of values, e.g. postal code centers of a region, and are thus naturally modeled as vertices of a finite spatial graph. Alternatively, they may have an uncountable range of values.

Similarly to the drift $\mu(\cdot)$ in universal Kriging, we assume

$$f_{\text{geo}}(t) = \sum_{j=1}^{M} \gamma_j \cdot \Psi_j(t), \quad M \in \mathbb{N}, \tag{3.35}$$

where $\Psi = \{\Psi_j\}_{j=1}^{\infty}$ is an orthonormal basis of functions in a certain functional space. For smoothing procedures, the basis functions Ψ_j may be assumed to have a certain degree of smoothness, such as tensor products of univariate splines or Fourier basis. In higher dimensions $d \gg 2$, the additive structure

$$f_{\text{geo}}(t) = \sum_{l=1}^{d} f_l(t^l), \quad t = (t^1, \ldots, t^d)^\intercal \in \mathbb{R}^d,$$

is often used to diminish the effect of the curse of dimensionality. Here, $f_l : \mathbb{R} \to \mathbb{R}$ are univariate functions which may themselves have a structure as in (3.35). The goal is to estimate the regression coefficients $\gamma_1, \ldots, \gamma_M$ in (3.35).

Similarly, functions g_j are assumed to have the form

$$g_j(z) = \sum_{l=1}^{N} \alpha_{jl} \cdot \varphi_l(z), \quad j = 1, \ldots, M, \tag{3.36}$$

where $\varphi = \{\varphi_l\}_{l=1}^{\infty}$ is another orthonormal basis in a certain functional space. The regression coefficients $\alpha_{1l}, \ldots, \alpha_{Nl}$, $l = 1, \ldots, N$, have to be estimated from the data.

Example 3.34 (Basis functions): (a) *Tensor product bases:* Let $d = 2$, $t = (t^1, t^2)^\intercal \in \mathbb{R}^2$. For an orthonormal basis in $L^2(\mathbb{R})$ consisting of functions $\{\Psi_j^1\}_{j \in \mathbb{N}}$, we may form

$$\Psi_{j_1 j_2}(t) = \Psi_{j_1}^1(t^1) \cdot \Psi_{j_2}^1(t^2), \quad j_1, j_2 \in \mathbb{N},$$

thus yielding

$$f_{\text{geo}}(t) = \sum_{j1=1}^{M1} \sum_{j_2=1}^{M_2} \gamma_{j_1 j_2} \Psi_{j_1 j_2}, \quad t \in \mathbb{R}^2.$$

Similarly, the construction can be easily adapted to any dimension $d > 2$.

(b) *B-Splines:* As an example of univariate bases $\{\Psi_j^1\}_{j\in\mathbb{N}}$ consider *B-Splines* on an interval $(a, b] \subset \mathbb{R}$. Their construction is iterative. Without loss of generality set $a = 0, b = 1$ and let $0 \le c_1 < c_2 < \cdots < c_p = 1$ be a decomposition of $[0, 1]$ into disjoint intervals $[c_j, c_{j+1})$ for $j = 1, \ldots, p - 1$.

For any $z \in [0, 1]$ consider $l = 0$ first and set

$$B_j^0(z) = \mathbf{1}(z \in [c_j, c_{j+1})), \quad j = 1, \ldots, p-1,$$

and for higher orders $l \ge 1$ proceed with

$$B_j^l(z) = \frac{z - c_{j-l}}{c_j - c_{j-l}} B_{j-1}^{l-1}(z) + \frac{c_{j+1} - z}{c_{j+1} - c_{j+1-l}} B_j^{l-1}(z), \quad j = 1, \ldots, p-1.$$

For this calculation, we need $2l$ outer knots $c_{1-l}, \ldots, c_0, \ldots, c_{p+1}, \ldots, c_l$ lying outside of the interval $[0, 1]$. For simplicity, c_j can be often chosen equidistantly over $[0, 1]$ and beyond.

B-Splines have many interesting properties:

(i) *Local basis*: It holds that $B_j^l(z) > 0$ only on (c_{j-l}, c_{j+1-l}) and $B_j^l(z) = 0$, elsewhere on $[0, 1]$. Vice versa, at any $z \in [0, 1]$, only $l + 1$ functions B_j^l are positive. If c_j are chosen equidistantly, then all B_j^l have the same shape and are only shifted along the z-axis.

(ii) *Unity decomposition*: $\sum_{j=1}^p B_j^l(z) = 1$ for all $z \in [0, 1]$, $l \in \mathbb{N}_0$.

(iii) *Uniformly bounded*: $0 \le B_j^l(z) \le 1$ for all $z \in [0, 1]$, $j = 1, \ldots, p - 1$, $l \in \mathbb{N}_0$.

(iv) *Derivates*:

$$\frac{\partial B_j^l(z)}{\partial z} = l \left(\frac{B_{j-1}^{l-1}(z)}{c_j - c_{j-1}} - \frac{B_j^{l-1}(z)}{c_{j+1} - c_{j+1-l}} \right),$$

which yields

$$\frac{\partial}{\partial z} \left[\sum_{j=1}^p \gamma_j B_j^l(z) \right] = l \cdot \sum_{j=1}^p \frac{\gamma_j - \gamma_{j-1}}{c_j - c_{j-l}} B_{j-1}^{l-1}(z), \quad z \in [0, 1],$$

for any fixed $l \ge 1$ and any coefficients $\gamma_0, \gamma_1, \ldots, \gamma_p \in \mathbb{R}$, $\gamma_0 = 0$.

(c) *Splines and the truncated power series:* Another example of $\{\Psi_j^1\}_{j\in\mathbb{N}}$ is given by

$$\Psi_1^1(z) \equiv 1, \Psi_2^1(z) = z, \ldots, \Psi_{l+1}^1(z) = z^l, \Psi_{l+j}(z) = (z - c_j)_+, \quad j = 2, \ldots, p,$$

where l is chosen large enough, and the points $\{c_j\}_{j=1}^p$ are as in Example (b). The sum

$$\sum_{j=1}^{l+1} \gamma_j z^j + \sum_{j=2}^p \gamma_{l+j} (z - c_j)_+^l$$

is called a *polynomial spline with truncated power series*. Here,

$$a_+ = \begin{cases} a, & a \geq 0, \\ 0, & a < 0, \end{cases}$$

denotes the non-negative part of $a \in \mathbb{R}$. The second sum of truncated monomials $(z - c_j)_+^l$ is designed to catch sudden changes of slope in the functional data.

We now return our focus to the geoadditive regression model in (3.34). Using the assumptions on f_{geo} and g_j in (3.35) and (3.36) we can state the regression model as

$$Y_i = \sum_{j=0}^{K} \beta_j x_{ij} + \sum_{j=1}^{m} \sum_{l=1}^{N} \alpha_{jl} \varphi_l(z_{ij}) + \sum_{j=1}^{M} \gamma_j \Psi_j(t_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{3.37}$$

or in matrix form

$$Y = X\beta + \sum_{l=1}^{N} \varphi_l(Z)\alpha_l + \Psi_t \gamma + \varepsilon,$$

where

$$\begin{aligned}
Y &= (Y_1, \ldots, Y_n)^\intercal, \\
X &= (x_{ij})_{i=1,\ldots,n,\ j=0,\ldots,K}, \\
\beta &= (\beta_0, \beta_1, \ldots, \beta_k)^\intercal, \\
\alpha_l &= (\alpha_{jl})_{j=1}^{m}, \quad \varphi_l(Z) = (\varphi_l(z_{ij}))_{i=1,\ldots,n,\ j=1,\ldots,m}, \qquad l = 1, \ldots, N, \\
\Psi_t &= (\Psi_j(t_i))_{j=1,\ldots,M,\ i=1,\ldots,n}, \\
\gamma &= (\gamma_1, \ldots, \gamma_M)^\intercal, \\
\varepsilon &= (\varepsilon_1 \ldots, \varepsilon_n)^\intercal.
\end{aligned}$$

We may combine everything in one single matrix $\tilde{X}$, i.e.

$$\tilde{X} = \underbrace{\left.\left(\begin{array}{c|c|c|c|c} \vdots & \vdots & & \vdots & \vdots \\ X & \varphi_1(Z) & \cdots & \varphi_N(Z) & \Psi_t \\ \vdots & \vdots & & \vdots & \vdots \end{array} \right)\right\} n}_{k+1+m \cdot N+M}, \quad \tilde{\beta} = \begin{pmatrix} \beta \\ \alpha_1 \\ \vdots \\ \alpha_N \\ \gamma \end{pmatrix}$$

Summarizing, the linear model (3.37) is given by

$$Y = \tilde{X} \cdot \tilde{\beta} + \varepsilon. \tag{3.38}$$

The parameter vector $\tilde{\beta}$ can be estimated using the *ordinary least squares* procedure, i.e.

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta} \in \mathbb{R}^{k+1+mN+M}}{\operatorname{argmin}} \|Y - \tilde{X}\tilde{\beta}\|_2^2 \tag{3.39}$$

Theorem 3.35 Let the matrix X have full rank K, and let $n > k + 1 + mN + M$. If $\{\varphi_l\}_l$ and $\{\Psi_j\}_j$ are linearly independent, then there exists a unique solution solution of (3.39), which is given by

$$\hat{\tilde{\beta}} = \left(\tilde{X}^\intercal \tilde{X}\right)^{-1} \tilde{X}^\intercal Y.$$

Proof Since all the matrices X, $\varphi_l(Z)$, $l = 1, \ldots, N$, and Ψ_t have full rank, the matrix $\tilde{X}$ has full rank equal to $k + 1 + mN + M$. It follows that $\tilde{X}^\intercal \tilde{X}$ is invertible. Hence, we can compute the partial derivatives of $\|Y - \tilde{X}\tilde{\beta}\|_2^2$ and set them equal to zero. The solution is then given by (3.39) It is unique since the target function represents a paraboloid. □

Sometimes, it is also desirable to control the smoothness of the solution of the regression equations. For that, a *penalized regression* is usually performed, which minimizes the energy of functional basis approximation given by the integral of its second derivative. For instance, if the j-th non-linear part is given by $g_j(z) = \sum_{l=1}^{N} \alpha_{jl} \cdot \varphi_l(z)$, we define its *energy* by

$$\mathfrak{E}(g_j) := \int_{\mathbb{R}} \left(g_j''(z)\right)^2 dz = \sum_{i,l=1}^{N} \alpha_{jl}\alpha_{ji} \cdot \underbrace{\int_{\mathbb{R}} \varphi_l''(z)\varphi_i''(z)dz}_{K_{il}} = \alpha_j^\intercal K \alpha_j,$$

where $K = (K_{il})_{i,l=1,\ldots,N}$ and $\alpha_j = (\alpha_{j1}, \ldots, \alpha_{jN})^\intercal$. Similarly, doing so for any g_j, $j = 1, \ldots, m$, and the geoadditive part f_{geo} with

$$\mathfrak{E}(f_{geo}) := \int\limits_{\mathbb{R}^d} \triangle f_{geo}(t)dt,$$

where $\triangle = \sum_{j=1}^{d} \frac{\partial^2}{\partial t^{j\,2}}$ is the Laplace operator, we may come to the *penalized regression*:

$$\|Y - \tilde{X}\tilde{\beta}\|_2^2 + \lambda \cdot \tilde{\beta}^\intercal \tilde{K} \tilde{\beta} \to \min_{\tilde{\beta} \in \mathbb{R}^{k+1+mN+M}}, \tag{3.40}$$

where the *penalty factor* $\lambda \geq 0$ is chosen experimentally.

The matrix $\tilde{K}$ is a block matrix (similar to $\tilde{X}$):

$$\tilde{K} = \left(\begin{array}{c|c|c|c|c} \vdots & \vdots & & \vdots & \vdots \\ 0 & K\varphi_1 & \cdots & K\varphi_N & K\Psi_t \\ \vdots & \vdots & & \vdots & \vdots \end{array}\right).$$

Analogously to Theorem 3.35, the solution of (3.40) is given by

$$\hat{\tilde{\beta}} = \left(\tilde{X}^\intercal \tilde{X} + \lambda \tilde{K}\right)^{-1} \tilde{X}^\intercal Y$$

for all $\lambda \geq 0$ such that $\tilde{X}^\intercal \tilde{X} + \lambda \tilde{K}$ is invertible.

Example 3.36 (Penalization): (a) *Penalization with B-Splines*: If for any fixed degree l and $j_0 = 1, \ldots, m$ we have

$$g_{j_0}(z) = \sum_{j=1}^{p} \alpha_{j_0 j} \cdot B_j^l(z),$$

where $B_j^l(\cdot)$ are B-Splines, then using property (iv) in Example 3.34 (b) we may write

$$g_{j_0}'(Z) = l \cdot \sum_{j=1}^{p} \underbrace{\frac{\alpha_{j_0 j} - \alpha_{j_0 j-1}}{c_j - c_{j-l}}}_{=d_{j-1}} B_{j-1}^{l-1}(z),$$

$$g''_{j_0}(Z) = l(l-1) \cdot \sum_{j=1}^{p} \underbrace{\frac{d_{j-1} - d_{j-2}}{c_j - c_{j-l}}}_{=e_{j-2}} B_{j-2}^{l-2}(z),$$

which yields

$$\mathfrak{E}(g_{j_0}) = \alpha_{j_0}^{\intercal} K \alpha_{j_0} = (l(l-1))^2 \sum_{j=1}^{p} e_{i-2} e_{j-2} \cdot \int_{\mathbb{R}} B_{i-2}^{l-2}(z) B_{j-2}^{l-2}(z) dz.$$

An alternative (easier) way to write the penalization with B-Splines is to get rid of the energy and replace it by a quadratic form $\alpha_{j_0}^{\intercal} K_r \alpha_{j_0}$, where $K_r := D_r^{\intercal} D_r$, and $D_r = D_1 D_{r-1}$ is the recursively defined matrix of differences of order r, i.e.

$$D_1 = \begin{pmatrix} -1 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & -1 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & -1 & 1 & & & \vdots \\ \vdots & & & \ddots & & & \vdots \\ \vdots & & & & \ddots & & 0 \\ 0 & \dots & \dots & \dots & 0 & -1 & 1 \end{pmatrix}$$

is a $((d-1) \times d)$-matrix differences of first order with

$$D_1 \alpha_{j_0} = \begin{pmatrix} \alpha_{j_0 2} - \alpha_{j_0 1} \\ \vdots \\ \alpha_{j_0 p} - \alpha_{j_0 p-1} \end{pmatrix},$$

and

$$D_2 = \begin{pmatrix} 1 & -2 & 1 & 0 & \dots & \dots & \dots & 0 \\ 0 & 1 & -2 & 1 & 0 & \dots & \dots & 0 \\ 0 & 0 & -1 & 1 & & & \vdots & \\ \vdots & & & \ddots & & & \vdots & \\ \vdots & & & & \ddots & & 0 & \\ 0 & \dots & \dots & 0 & 1 & -2 & 1 & \end{pmatrix}$$

is a $((d-2) \times d)$-matrix of differences of second order. Ultimately, we get

$$\lambda \alpha_{j_0}^{\intercal} K_r \alpha_{j_0} = \lambda \cdot \alpha_{j_0}^{\intercal} D_r^{\intercal} D_r \alpha_{j_0} = \lambda \cdot \|D_r \alpha_{j_0}\|_2^2 = \lambda \sum_{j=r+1} (\Lambda_r \alpha_{j_0 j})^2,$$

where $\Lambda_r = \Lambda \Lambda_{r-1}$ is the difference of order $r \geq 2$, $\Lambda \alpha_{j_0 j} = \alpha_{j_0 j} - \alpha_{j_0 j-1}$.

(b) *Truncated power series*: For a regression as in Example 3.34 (c), the most popular form of penalization is to keep the sum of coefficients of truncated powers

$$\lambda \cdot \sum_{j=l+2}^{p} \gamma_j^2 = \lambda \cdot \gamma^{\intercal} \cdot K \gamma$$

with $\gamma = (\gamma_1, \dots, \gamma_{l+p})^{\intercal}$, $K = diag(\underbrace{0, \dots, 0}_{l+1}, \underbrace{1, \dots, 1}_{p-1})$ minimal.

Example 3.37 (Radial basis functions and thin plate splines): Another way to construct the basis $\{\Psi_j\}_{j=1}^{\infty}$ in the geoadditive part of the regression model is via the so-called *radial functions*, which are defined by

$$\Psi_X(t) = B(\|x - t\|_2),$$

where the function $B : \mathbb{R}_+ \mapsto \mathbb{R}$ depends only on the distance between an observation point $t \in \mathbb{R}^d$ and a knot $x \in \mathbb{R}^d$. Hence, instead of counting Ψ_j with $j \in \mathbb{N}$, this functional system is parameterized via a finite system of knots $x_j \in W,\ j = 1, \dots, M$, where W is our spatial obervation window, see Figure 3.6. This kind of basis is advisable for isotropic spatial effects.

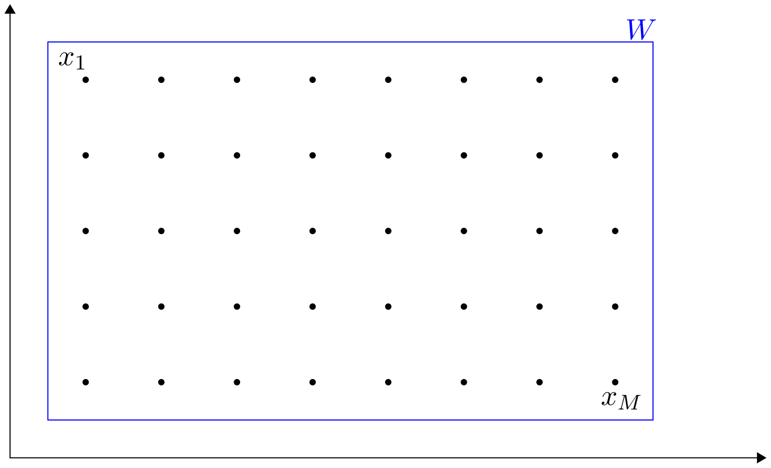


Fig. 3.6: Finite system of knots $x_1, \dots, x_M$ in a window $W \subset \mathbb{R}^2$.

As an example for the function B, consider the so-called *thin plate spline* given by $B(r) = r^2 \cdot \log(r),\ r > 0$. As a georeferenced penalization criterion, the penalization

$$\mathfrak{E}(f_{geo}) = \int_{\mathbb{R}^d} \left(d^2 f_{geo}(t)\right)^2 dt$$

is often used instead of

$$\mathfrak{E}(f_{geo}) := \int_{\mathbb{R}^d} \left(\Lambda^2 f_{geo}(t)\right)^2 dt,$$

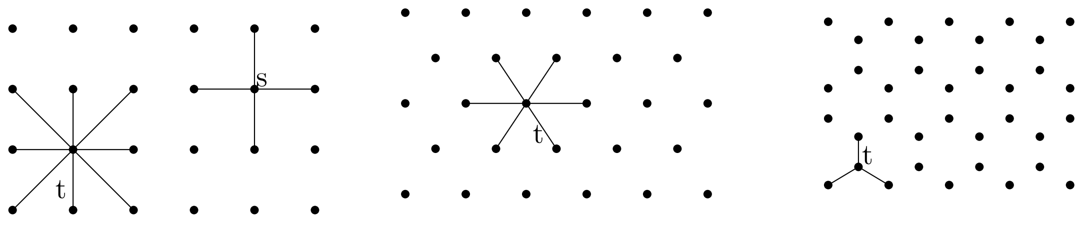
where $d^2 f_{geo}$ is the *second differential* of f_{geo}, i.e.

$$d^2 f_{geo}(t) = \left[\sum_{j=1}^{d} \frac{\partial^2}{\partial t^{j^2}} + 2\sum_{i<j} \frac{\partial}{\partial t^i} \frac{\partial}{\partial t^j}\right] f_{geo(t)}.$$

Other common examples of radial functions B are $B(r) = r^l$ with l odd, $B(r) = \sqrt{r^2 + c^2}$, $c > 0$ constant, or $B(r)$ originating from a covariance function of an isotropic random field, i.e. $C(s, t) = B(\|s - t\|_2),\ s, t \in \mathbb{R}^d$. This random field may represent spatial effects in regression model (3.34) giving rise to the part f_{geo}.

Example 3.38 (Markov random fields): Sometimes, the spatial location variable t may be discrete, attaining a finite number of values, e.g. as in the case of postal codes. We assume that $t \in V$, where $\Gamma = (V, E)$ is a *finite non-oriented geometric graph* with a finite set of *vertices* $V = t_1, \dots, t_M \subset \mathbb{R}^d$ and *edges* E. We say that the vertices t_i, t_j are neighbors (write $t_i \sim t_j$)

if they are connected by an edge $(t_i, t_j) \in E$. For instance, the set of vertices V may be a finite regular grid of locations within an observation window $W \subset \mathbb{R}^d$ with an intuitive neighboring relation, compare Figure 3.7.



(a) Regular square grid with a 4- or 8-neighbourhood relation

(b) Regular triangulation of plane with a 6-neighbourhood relation

(c) Regular hexagonal lattice in $\mathbb{R}^2$ with a 3-neighbourhood relation

Fig. 3.7: Planar regular grids with corresponding neighbourhood relation

In this case, the georeferenced part is *piecewise constant*: $f_{geo}(t) = \gamma_t,\ t \in V$. A reasonable penalization criterion is that values of γ_t for neighboring vertices t do not differ too much, i.e.

$$\mathfrak{E}(f_{geo}) = \sum_{(s,t)\in E} (\gamma_s - \gamma_t)^2.$$

In matrix form this can be expressed as

$$\mathfrak{E}(f_{geo}) = \gamma^\intercal K \gamma,$$

with $\gamma = (\gamma_t)_{t\in V}$, $K = (K_{st})_{s,t\in V}$ is the *adjacency matrix* of the graph Γ, i.e. for $s, t \in V$ we have

$$K_{st} = \begin{cases} \deg(s), & s = t, \\ -1, & s \neq t, s \sim t, \\ 0, & s \neq t, s \not\sim t, \end{cases}$$

where $\deg(s)$ is the *degree* of a vertex $s \in V$, i.e., the number of all neighbors of s, i.e.

$$\deg(s) = \#\{t \in V : t \sim s\}.$$

However, the order of vertices in V is important for the particular structure of K. It is desirable to numerate $t \in V$ so that all non-zero elements of K are located close to the main diagonal to produce a band matrix with a small band width.

A popular model used in econometrics is the so-called *spatial autoregressive process*, which is defined by

$$Y_s = X_s^\intercal \beta + \alpha \sum_{t\sim s} w_{st} Y_t + \varepsilon_s, \quad s \in V,$$

where $X_s = (1, X_{s1}, \ldots, X_{sk})^\intercal$, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\intercal$ represents the linear part (see (3.29), (3.31)), the constant $\alpha \in [0, 1)$ is an *autoregressive parameter*, $w_{st} = w_{ts}$ are symmetric weights such that $w_{ss} = 0$, $s \in V$, and the errors $\varepsilon_d \sim N(0, \sigma^2)$ are i.i.d., $s \in V$. Then, the field $\{Y_s, s \in V\}$ is called a *Markov random field.*

3.4 Quantile regression

A drawback of geoadditive regression is the consideration of means of target variables, i.e. $\mathbb{E}[Y] = \tilde{X}\tilde{\beta}$ in terms of (3.37), since $\mathbb{E}[\varepsilon] = 0$. Here, $\mathbb{E}[Y] = (\mathbb{E}[Y_1], \ldots, \mathbb{E}[Y_n])^\intercal$, $\mathbb{E}[\varepsilon] = (\mathbb{E}[\varepsilon_1], \ldots, \mathbb{E}[\varepsilon_n])^\intercal$. In addition, the method of least squares imposes the assumption $\mathbb{E}[Y_i^2] < \infty$, $i = 1, \ldots, n$, which excludes heavy tailed regression errors ε_i with $\mathbb{E}[\varepsilon_i^2] = \infty$.

The goal of this Section is to construct a regression model which is not based on the (conditional) means but on the (conditional) quantiles of Y, allowing for a more general structure of the error vector ε. For a random variable Z, its quantile of order $\alpha \in (0,1)$ is given by

$$F_Z^-(\alpha) = \inf x \in \mathbb{R} : F_Z(x) \geq \alpha := q_\alpha, \tag{3.41}$$

where $F_Z(x) = \mathbb{P}(Z \leq x)$, $x \in \mathbb{R}$, is the cumulative distribution function of Z. It has the property

$$F_Z^-(\alpha) = \underset{x \in \mathbb{R}}{\operatorname{argmin}}\, \mathbb{E}\left[w_\alpha(Z,x) \cdot |Z - x|\right], \tag{3.42}$$

where

$$w_\alpha(Z,x) = \begin{cases} 1-\alpha, & Z < x, \\ 0, & Z = x, \\ \alpha, & Z > x. \end{cases}$$

For $\alpha = \frac{1}{2}$, the quantile $q_{1/2}(Z)$ is called *median* of Z. Here, $w_{1/2}(Z,x) = \frac{1}{2} \cdot \mathbf{1}(Z \neq x)$.

Exercise 3.39 Show relation (3.42).

If a sample of i.i.d. realizations $(Z_1, \ldots, Z_n)$ of Z is given, the *empirical quantile* $\hat{q}_\alpha(Z)$ of Z of order $\alpha \in (0,1)$ is defined similarly to (3.41) as

$$\frac{1}{n}\sum \mathbf{1}(Z_i \leq \hat{q}_\alpha) \geq \alpha, \quad \frac{1}{n}\sum \mathbf{1}(Z_i \geq \hat{q}_\alpha) \geq 1 - \alpha,$$

or equivalently similar to (3.42) as

$$\hat{q}_\alpha(Z) = \underset{x}{\operatorname{argmin}} \sum_{i=1}^n \left(w_\alpha(Z_i,x) \cdot |Z - x|\right). \tag{3.43}$$

The *linear α-quantile regression* is given by

$$Y = X^\intercal\beta + \varepsilon,$$

where $X^\intercal\beta$ is the linear part with $X = (1, x_1, \ldots, x_k)^\intercal$, $\beta = (\beta_0, \beta_1, \ldots, \beta_k)^\intercal$, and the regression error random variable ε satisfies $F_\varepsilon(0) = \alpha$, where $F_\varepsilon(x) = \mathbb{P}(\varepsilon \leq x),\ x \in \mathbb{R}$. This implies

$$\alpha = F_\varepsilon(0) = \mathbb{P}(\varepsilon \leq 0) = \mathbb{P}(\underbrace{X^\intercal\beta + \varepsilon}_{=Y} \leq X^\intercal\beta) = F_Y(X^\intercal\beta).$$

Hence, the α-quantile of the target random variable Y is given by $X^\intercal\beta$, i.e.

$$q_\alpha(Y) = X^\intercal\beta, \quad \alpha \in (0,1). \tag{3.44}$$

For instance, $\alpha = \frac{1}{2}$ yields the *median regression*.

Now, the *α-quantile linear regression model*, $\alpha \in (0,1)$, given by

$$Y_i = X_i^\intercal \beta + \varepsilon_i, \ i = 1, \dots, n,$$

where $Y = (Y_1, \dots, Y_n)^\intercal$ is the vector of the target random variables Y_i, $i = 1, \dots, n$, $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\intercal$ is the vector of regression error random variables subject to assumptions that $\varepsilon_1, \dots, \varepsilon_n$ are independent and $F_{\varepsilon_i}(0) = \alpha$, $i = 1, \dots, n$. The matrix $X = (x_{ij})_{i=1,\dots,n,\ j=0,\dots,k}$ is the *design matrix* with rows $X_i = (1, x_{i1}, \dots, x_{ik})$, $i = 1 \dots n$ and $\beta = (\beta_0, \dots, \beta_k)^\intercal$ is the vector of *regression coefficients*. The estimation of β is based on relation (3.43), i.e.

$$\hat{\beta} = \operatorname*{argmin}_{\beta \in \mathbb{R}^{k+1}} \sum_{i=1}^{n} w_\alpha \left(Y_i, X_i^\intercal \beta\right) |Y_i - X_i^\intercal \beta| \tag{3.45}$$

The above minimization problem is usually solved numerically via linear programming or functional gradient descent boosting. Let us explain the first of these two methods.

To minimize the target functional in (3.45), rewrite the regression errors as

$$Y_i - X_i^\intercal \beta = \varepsilon_i = (\varepsilon_i)_+ - (-\varepsilon_i)_+, \quad i = 1 \dots n,$$

with $a_+ = a \cdot \mathbf{1}(a \geq 0)$ such that

$$\sum_{i=1}^{n} w_\alpha \left((Y_i, X_i^\intercal \beta)|Y_i - X_i^\intercal \beta|\right) = \alpha \sum_{i=1}^{n} u_i + (1-\alpha) \sum_{i=1}^{n} v_i = \alpha e^\intercal u + (1-\alpha) e^\intercal v,$$

where

$$u_i := (\varepsilon_i)_+ = (Y_i - X_i^\intercal \beta)_+, \quad v_i := (-\varepsilon_i)_+ = (X_i^\intercal \beta - Y_i)_+,$$

with $u = (u_1, \dots, u_n)^\intercal$, $v = (v_1, \dots, v_n)^\intercal$, $e = (1, \dots, 1)^\intercal \in \mathbb{R}^n$. Then, the quantile regression is given in total by the equation

$$Y_i = X_i^\intercal \beta + u_i - v_i, \quad i = 1 \dots n,$$

or in matrix form

$$Y = X\beta + u - v. \tag{3.46}$$

In terms of constraint minimization (3.45), this can be rewritten as

$$\begin{cases} \alpha e^\intercal u + (1-\alpha) e^\intercal v \to \min\limits_{\beta, u, v}, \\ X\beta + u - v = Y, \end{cases}$$

which is a linear programming problem with polyhedral constraints.

Theorem 3.40 (Properties of quantile regression):

(a) *Invariance under monotone transforms:* If $T_i : \mathbb{R} \mapsto \mathbb{R}$ is a monotone transformation of the data, then the α-quantile regression (3.45) based on Y_i, $X_i^\intercal \beta$ and $T_i(Y_i)$, $T_i(X_i^\intercal \beta)$ yield the same results.

(b) *Asymptotic normality:* If ε_i, $i = 1, \dots, n$, are i.i.d with probability density function f_ε, then

$$X^\intercal X \cdot (\hat{\beta} - \beta) \xrightarrow{d} N\left(0, \frac{\alpha(1-\alpha)}{f_\varepsilon^2(\delta)} \cdot I_{k+1}\right), \tag{3.47}$$

where I_{k+1} is the $((k+1) \times (k+1))$-dimensional unity matrix.

Remark 3.41 Although we will not prove the above theorem, let us comment on its assertions.

(a) The invariance property is clear since the quantiles are kept under monotonic transforms T_i, i.e. $q_\alpha(T_i(Y_i)) = T_i(X_i^\intercal \beta)$, $i = 1 \ldots n$, for any $\alpha \in (0,1)$.

(b) Although the asymptotic variance in (3.47) is seemingly minimal if $\alpha \to 0$ or $\alpha \to 1$, the requirement $F_\varepsilon(0) = \alpha$ suggests that $f_\varepsilon(0) \to 0$ as well in such cases, which will dominate the quantity $\frac{\alpha(1-\alpha)}{f_\varepsilon^2(0)}$ letting it diverge to ∞. This observation is in line with the fact that estimating extremal quantiles $q_\alpha(Y_i)$, i.e. for $\alpha \to 0$ or $\alpha \to 1$, is difficult. Hence, the the regression with $\alpha \approx \frac{1}{2}$, e.g. the median regression, will yield a smaller asymptotic variance.

Using Equation (3.45), we can propose a similar framework for a *non-linear α-quantile regression*

$$Y_i = g(Z_i) + \varepsilon_i, \quad i = 1, \ldots, n,$$

where the non-linear part $g : [a,b] \mapsto \mathbb{R}$, $g \in C^1[a,b]$, $a < b$, is approximated by a truncated expansion, i.e.

$$g(x) \approx \sum_{l=1}^{N} \alpha_l \varphi_l(x)$$

with respect to some function basis $\{\varphi_l\}_{l=1}^{\infty} \subset L^2[a,b]$, and $F_{\varepsilon_i}(0) = \alpha \in (0,1)$, $i = 1, \ldots, n$. Under the notation $\vec{\alpha} = (\alpha_1, \ldots, \alpha_N)^\intercal$, one looks for the *$\alpha$-quantile regression estimate* $\hat{\vec{\alpha}}$ of $\vec{\alpha}$ as

$$\hat{\vec{\alpha}} = \operatorname*{argmin}_{\hat{\alpha} \in \mathbb{R}^N} \sum_{i=1}^{n} w_\alpha \left((Y_i, \sum_{l=1}^{N} \alpha_l \varphi_l(Z_i)) \cdot |Y_i - \sum_{l=1}^{N} \alpha_l \varphi_l(Z_i)| + \lambda \cdot V(g') \right), \tag{3.48}$$

where $\lambda \geq 0$ is the *penalization factor*. Here, the*penalty* is the total variation of the first derivate of g, i.e.

$$V(g') = \sup_{\{x_j\} \subset [a,b]} \sum_{j=1}^{n} |f'(x_{x+1}) - f'(x_j)|. \tag{3.49}$$

The supremum is taken over all partitions of $[a,b]$ into disjoint intervals $(x_j, x_{j+1}]$ with $a \leq x_1 < x_2 < \cdots < x_{n-1} < x_n \leq b$.

If we additionally assume $g \in C^2[a,b]$, we may write the penalty as

$$V(g') = \int_a^b |g''(x)| dx. \tag{3.50}$$

This is similar to the penalty of the usual (geo-)additive linear regression, which was given by $\int_a^b (g''(x))^2 dx$. The use of the L^1-norm of g'' instead of the L^2-norm allows for the use of linear programming methods for the minimization of the target functional in (3.48).

Since we approximate $g(x) \approx \sum_{l=1}^{N} \alpha_l \varphi_l(x)$, it follows that $g'(x) \approx \sum_{l=1}^{N} \alpha_l \varphi_l'(x)$ and

$$V(g') = \sup_{\{x_j\} \subset [a,b]} \sum_{j=1}^{n} \left| \sum_{l=1}^{N} \alpha_l (\varphi_l'(x_{j+1}) - \varphi_l'(x_j)) \right|$$

or

$$V(g') = \int_a^b \left| \sum_{l=1}^{N} \alpha_l \varphi_l''(x) \right| dx,$$

respectively. For practical purposes, we may use $x_i = Z_{(i)},\ i = 1, \ldots, n$, which yields

$$V(g') \approx \sum_{i=1}^{n} \left| \sum_{l=1}^{N} \alpha_l (\varphi_l'(Z_{(i+1)}) - \varphi_l'(Z_{(i)})) \right|.$$

Similar to (3.34) and (3.48), the geo-additive α-quantile regression can be computed by

$$q_\alpha(Y_i) = X_i^\intercal \beta + \sum_{j=1}^{m} g_j(z_{ij}) + f_{geo}(t_i), \quad i = 1, \ldots, n.$$

Using the truncated series expansions $g_j(z) = \sum_{l=1}^{N} \alpha_{jl} \varphi_l(z)$ and $f_{geo}(t) = \sum_{j=1}^{M} \gamma_j \cdot \Psi(t)$ leads to $q_\alpha(Y) = \tilde{X}\tilde{\beta}$, compare (3.38).

The α-quantile estimate of $\tilde{\beta}$ is given by

$$\hat{\tilde{\beta}} = \underset{\tilde{\beta} \in \mathbb{R}^{K+1+m \cdot N+M}}{\operatorname{argmin}} \left[\sum_{i=1}^{n} w_\alpha \left((Y_i, (\tilde{X} \cdot \tilde{\beta})_i) \cdot |Y_i - (\tilde{X} \cdot \tilde{\beta})| \right) + \lambda \cdot p_{\tilde{\beta}} \right],$$

where the penalty is

$$p_{\tilde{\beta}} = \sum_{j=1}^{m} V(g_j') + V(\nabla g_{geo})$$

with

$$V(g') = \sum_{i=1}^{n} | \sum_{l=1}^{N} \alpha_{jl} (\varphi_l'(Z_{(i+1)j}) - \varphi_l'(Z_(i)j))|, \quad j = 1, \ldots, m,$$

as well as

$$V(\nabla f_{geo}) = \int_{\mathbb{R}^d} |\Delta f_{geo}(t)| dt \quad \text{or} \quad V(\nabla f_{geo}) = |\partial^2 f_{geo}(t)| dt.$$

Here, the penalty factor is $\lambda \geq 0$ and $(\tilde{X} \cdot \tilde{\beta})_i$ is the i-th coordinate of the vector $\tilde{X} \cdot \tilde{\beta}$.

Remark 3.42 (Formulation via a loss function): The α-quantile regression (3.45) can be reformulated as follows. For any strictly increasing real function G, define the *"tick" function* for any $\alpha \in (0, 1)$ by

$$\rho_\alpha(x) = (\alpha - \mathbf{1}(x \leq 0))x, \quad x \in \mathbb{R}.$$

Let

$$L(x, y) = \rho_\alpha(G(x) - f(y)), \quad x, y \in \mathbb{R},$$

be the *loss function* with the property $L(x, y) \geq 0$ and $L(x, y) = 0$ if and only if $x = y$. Then, the α-quantile regression estimate is

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{k+1}}{\operatorname{argmin}}\ \mathbb{E}\left[L(Y, X^\intercal \beta)\right]$$

for the regression (3.44), or, in a data setting,

$$\hat{\beta} = \underset{\beta \in \mathbb{R}^{k+1}}{\operatorname{argmin}} \sum_{i=1}^{n} L(Y_i, X_i^\intercal \beta).$$

Analogously, the unpenalized non-linear α-quantile regression $Y = g(Z) + \varepsilon$ is given by

$$\hat{g} = \underset{g}{\operatorname{argmin}}\ \mathbb{E}\left[L(Y, g(Z))\right]$$

or

$$\hat{g} = \underset{g}{\operatorname{argmin}} \sum_{i=1}^{n} L(Y_i, g(Z_i)),$$

respectively, where $g(Z)$ can be further linearized as in (3.48).

3.5 Prediction via level sets

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a complete probability space and $X = \{X(t), t \in T\}$, $T \subset \mathbb{R}^d$ be a strictly stationary real-valued measure random field on $(\Omega, \mathcal{F}, \mathbb{P})$ with marginal distribution $F_X(x) = \mathbb{P}(X(s) \leq x)$, $x \in \mathbb{R}$. Assume that X is observed at locations $t_1, \ldots, t_N \in W$, where $W \subset \mathbb{R}^d$ is a compact non-empty observation window. For $t \notin \{t_1, \ldots, t_N\}$, we predict the value $X(t)$ by the linear predictor

$$\hat{X}(t) = \sum_{j=1}^{N} \lambda_j X(t_j)$$

such that weights $\lambda_1, \ldots, \lambda_N$ are chosen to minimize a certain mean error criterion subject to the additional constraint

$$\hat{X}(t) \overset{d}{=} X(t), \quad t \in W,$$

i.e., $\mathbb{P}(\hat{X}(t) \leq x) = \mathbb{P}(X(t) \leq x)$, $x \in \mathbb{R}$. Compared to Kriging, which does not keep the marginal distribution of X, this property is sometimes very desirable and is attained most of the time by the so-called conditional simulation. However, to be able to mimic the law of $X(t)$ by a linear predictor $\hat{X}(t)$, it is necessary that X belongs to the so-called *infinitely divisible class*.

Definition 3.43 (a) The probability law of a random vector $Y : \Omega \to \mathbb{R}^m$ is called *infinitely divisible* if for all $n \in \mathbb{N}$ there exist i.i.d. random vectors $Y_{n1}, \ldots, Y_{nn}$ such that

$$Y \overset{d}{=} \sum_{j=1}^{n} Y_{nj}.$$

(b) A random field $X = \{X(t), t \in T\}$ is called infinitely divisible if all its finite dimensional distributions are infinitely divisible.

Examples of infinitely divisible random functions are Lévy processes and α-stable random fields. Under the assumption that X is infinitely divisible it is guaranteed that the linear combination $\sum_{j=1}^{N} \lambda_j X(t_j)$ may have a distribution of the same type as $X(t)$.

Apart from the mean-square error $\mathbb{E}[(X(t) - \hat{X}(t))^2]$, which is used to compute Kriging predictors, other error criteria, which do not impose the restriction of square-integrability onto the field X, i.e. $\mathbb{E}[X^2(t)] < \infty$, $t \in T$, are possible. One of them is based on the comparison of the so-called *level sets* or *excursion sets* of X and $\hat{X}$.

Definition 3.44 The *excursion set* of a random field X at a level $u \in \mathbb{R}$ observed on a window W is given by

$$A_X(u) := \{t \in W : X(t) > u\}.$$

Since X is measurable, the set $A_X(u)$ is Borel for any $\omega \in \Omega$, and thus its volume $|A_X(u)|$ exists and is a random variable bounded a.s. by $|A_X(u)| \leq |W|$, where

$$|A_X(u)| = \int\limits_W \mathbf{1}(X(t) > u)dt.$$

The error criterion measuring the prediction error of X by $\hat{X}$ is given as an *error-in-measure*

$$\mathbb{E}\left[|A_X(u)\Delta A_{\hat{X}}(u)|\right]$$

at an excursion level $u \in \mathbb{R}$, where $A_{\hat{X}}(u) = \{t \in W : \hat{X}(t) > u\}$ and the symmetric difference is defined by

$$A_X(u)\Delta A_{\hat{X}}(u) := \left(A_X(u)\backslash A_{\hat{X}}(u)\right) \cup \left(A_{\hat{X}}(u)\backslash A_X(u)\right).$$

Choosing an excursion level u according to a finite non-zero measure ν on $(\mathbb{R}, \mathcal{B}_\mathbb{R})$ allows us to give the *overall mean extrapolation error* as

$$\int\limits_{\mathbb{R}} \mathbb{E}\left[|A_X(u)\Delta A_{\hat{X}}(u)|\right] \nu(du).$$

The measure ν can be chosen to be discrete, i.e. $\nu(\cdot) = \sum_{j=1}^k \delta_{u_j}(\cdot)$ with ν being concentrated at atoms u_j. Alternatively, it can be an absolutely continuous probability measure. Later on, a choice $\nu(\cdot) = \mathbb{P}_{X(0)}(\cdot)$ is proven to be quite reasonable, since it provides meaningful levels u such that $A_X(u) \neq \emptyset$ with positive probability.

To compute the weights $\lambda_1, \ldots, \lambda_N$ for the linear predictor $\hat{X}(t)$ we need to solve the minimization problem

$$\begin{cases} \int_\mathbb{R} \mathbb{E}\left[|A_X(u)\Delta A_{\hat{X}}(u)|\right] \nu(du) \to \min\limits_{\lambda_1,\ldots,\lambda_N}, \\ \hat{X}(t) \overset{d}{=} X(t), \; t \in W. \end{cases} \tag{3.51}$$

The target functional above does not depend on $t \in W$, since it provides an average over all $t \in W$. However, it would be desirable to let $\lambda_1, \ldots, \lambda_N$ depend on the point $t \in W$. In order to do so, we modify (3.51) as follows.

Theorem 3.45 The minimization problem in (3.51) is equivalent to the maximization problem

$$\begin{cases} \int_W \int_\mathbb{R} \mathbb{P}\left(X(t) > u, \hat{X}(t) > u\right) \nu(du)dt \to \max\limits_{\lambda_1,\ldots,\lambda_N}, \\ \hat{X}(t) \overset{d}{=} X(t), \; t \in W. \end{cases}$$

Proof First, rewrite

$$\mathbf{1}(A_X(u)\Delta A_{\hat{X}(u)}) = \mathbf{1}(X(t) > u) + \mathbf{1}(\hat{X}(t) > u) - 2\mathbf{1}(X(t) > u)\mathbf{1}(\hat{X}(t) > u).$$

Then, applying Fubini's theorem yields

$$\begin{aligned}\mathbb{E}\left[|A_X(u)\Delta A_{\hat{X}(u)}|\right] &= \mathbb{E}\left[\int\limits_W \mathbf{1}(A_X(u)\Delta A_{\hat{X}(u)})dt\right]\\ &= \int\limits_W \mathbb{P}\left(t \in A_X(u)\Delta A_{\hat{X}(u)}\right)dt\\ &= \int\limits_W \left[\mathbb{P}\left(X(t)>u\right)+\mathbb{P}\left(\hat{X}(t)>u\right)-2\mathbb{P}\left(X(t)>u,\hat{X}(t)>u\right)\right]dt\\ &= 2|W|\mathbb{P}(X(0)>u)-2\int\limits_W \mathbb{P}\left(X(t)>u,\hat{X}(t)>u\right)dt,\end{aligned}$$

where the last equality is due to the constraint in (3.51), i.e. $\hat{X}(t)\overset{d}{=}X(t)\overset{d}{=}X(0)$, and the stationarity of X, i.e.

$$\begin{aligned}\int\limits_W \mathbb{P}(X(t)>u)dt &= \mathbb{P}(X(t)>u)\cdot\int\limits_W dt = |W|\cdot\mathbb{P}(X(0)>u),\\ \int\limits_W \mathbb{P}(\hat{X}(t)>u)dt &= \int\limits_W \mathbb{P}(X(0)>u)dt = |W|\cdot\mathbb{P}(X(0)>u).\end{aligned}$$

Hence, the target functional in (3.51) can be rewritten as

$$\int\limits_{\mathbb{R}} \mathbb{E}\left[|A_X(u)\Delta A_{\hat{X}(u)}|\right]\nu(du) = 2|W|\int\limits_{\mathbb{R}}(1-F_X(u))\nu(du)-2\int\limits_{\mathbb{R}}\int\limits_W \mathbb{P}(X(t)>u,\hat{X}(t)>u)dt\nu(du).$$

Since the first term on the right-hand side does not depend on $\lambda_1,\ldots,\lambda_N$, the above expression is minimal when

$$\int_W\int_{\mathbb{R}}\mathbb{P}\left(X(t)>u,\hat{X}(t)>u\right)\nu(du)dt \tag{3.52}$$

is maximal. □

In view of Theorem 3.45, we omit the integration over W with respect to t, and modify our prediction problem as

$$\left.\begin{cases}\int\limits_{\mathbb{R}}\mathbb{P}(X(t)>u,\hat{X}(t)>u)\nu(du)\to\max\limits_{\lambda_1,\ldots,\lambda_N},\\ \hat{X}(t)\overset{d}{=}X(0),\end{cases}\right\}\text{ for any } t\in W, \tag{3.53}$$

since the integral in (3.52) is maximal if its integrand is maximal for all $t \in W$. Thus, the problem (3.51) yields a geometric motivation to the final formulation (3.52), see Figure 3.8

To solve the maximization problem (3.53) for each $t \in W$, knowledge of the uni- and bivariate probability law of X is required. The advantage of the criterion (3.53) is that no integrability assumptions on X are needed. Thus, extrapolation of heavy-tailed random fields such as α-stable random fields is possible. However, the choice of the measure ν may heavily influence the results, and its optimality is still an open problem. Nonetheless, for Gaussian random fields, the choice of ν is irrelevant, as we will see in the next section.

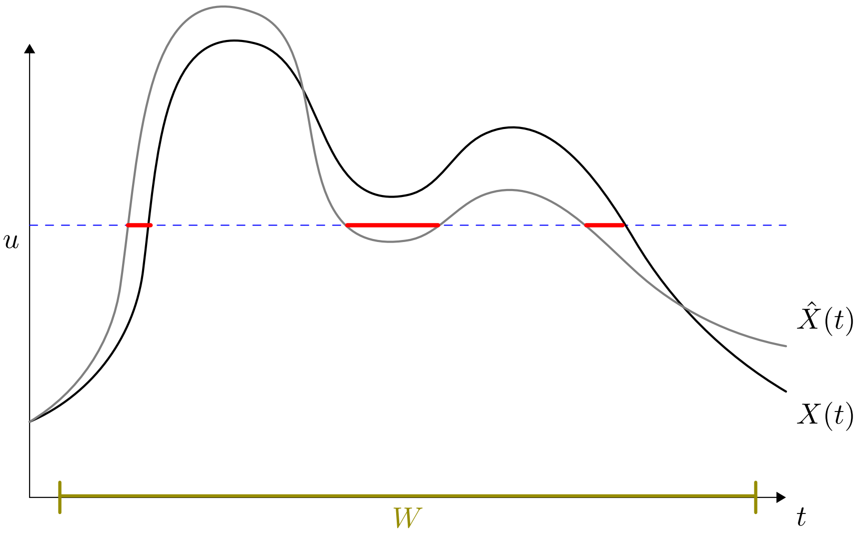


Fig. 3.8: The symmetric difference $A_X(u)\Delta A_{\hat{X}}(u)$ is shown in red for a process $X = \{X(t), t \in \mathbb{R}\}$ and its linear predictor $\hat{X}(t)$, $t \in \mathbb{R}$ $(d = 1)$. The overall length of these intervals is minimized to get a better fit of the path of $\hat{X}$ to the path of X.

3.5.1 Gaussian level set prediction

Let $X = \{X(t), t \in T\}$, $T \subset \mathbb{R}^d$, be a stationary measurable random field with mean $\mathbb{E}[X] = \mu$, covariance function $C(t) = \mathbf{cov}(X(0), X(t))$, $t \in T$, and variance $\sigma^2 = C(0) = \mathbf{var}(X(0)) > 0$. In order to solve the maximization problem (3.53), we first reformulate the constraint $\hat{X}(t) \overset{d}{=} X(0)$ in terms of $\lambda = (\lambda_1, \dots, \lambda_N)$. For ease of notation, we will omit the dependence of $\lambda_j = \lambda_j(t)$ on the location $t \in W$, $j = 1, \dots, N$, in the following.

Note that, since X is Gaussian, the constraint $\hat{X}(t) \overset{d}{=} X(0)$ is equivalent to

$$\begin{cases} \mathbb{E}[\hat{X}(t)] = \mu \\ \mathbf{var}(\hat{X}(t)) = \sigma^2 \end{cases} \Rightarrow \begin{cases} \sum_{j=1}^N \lambda_j \mu = \mu \\ \sum_{i,j=1}^N \lambda_i \lambda_j C(t_i - t_j) = \sigma^2 \end{cases} \Rightarrow \begin{cases} \lambda^\intercal e = 1 \\ \lambda^\intercal \Sigma \lambda = \sigma^2 \end{cases},$$

where $e = (1, \dots, 1)^\intercal$, $\Sigma = (C(t_i - t_j))_{i,j=1}^N$. We refer to $\lambda^\intercal e = 1$ as the *simplex constraint* and to $\lambda^\intercal \Sigma \lambda = \sigma^2$ as an *ellipsoid constraint*. In general, then mean $\mu \neq 0$ is unknown. In the case that $\mu = 0$ is known, the simplex constraint would not be needed and hence could just be omitted.

Recall the usual notation $c_t = (C(t - t_1), \dots, C(t - t_N))^\intercal$.

Lemma 3.46 The maximization problem (3.53) for stationary measurable Gaussian random fields X with unknown mean μ is given by

$$\begin{cases} c_t^\intercal \lambda \to \max\limits_{\lambda \in \mathbb{R}^N}, \\ \lambda^\intercal \Sigma \lambda = \sigma^2, \\ \lambda^\intercal e = 1. \end{cases} \tag{3.54}$$

if $\nu \not\equiv 0$.

Proof We use the following representation of the Gaussian bivariate law [13, p.9]. Since the bivariate vector $(X(t), \hat{X}(t))$ is jointly Gaussian, it holds that

$$\mathbb{P}(X(t) > u, \hat{X}(t) > u) = \bar{\phi}^2_{\mu,\sigma}(u) + \frac{1}{2\pi} \int\limits_0^{\sin^{-1}(\rho_t)} \exp\left\{ -\frac{(u-\mu)^2}{\sigma^2} \cdot \frac{1-\sin\theta}{\cos^2\theta} \right\} d\theta,$$

where $\rho_t = \mathbf{corr}(X(t), \hat{X}(t))$, the function $\phi_{\mu,\sigma}(x) = \frac{1}{\sqrt{2\pi}\sigma} \int\limits_{-\infty}^{x} e^{\frac{(y-\mu)^2}{2\sigma^2}} dy$, $x \in \mathbb{R}$, is the cumulative distribution function of $N(\mu, \sigma^2)$ and $\bar{\phi}_{\mu,\sigma}(x) = 1 - \phi_{\mu,\sigma}(x)$ is the tail probability function. Then, using Fubini's theorem, the target function in (3.53) simplifies to

$$\begin{aligned} F(\lambda, t) &:= \int_{\mathbb{R}} \mathbb{P}(X(t) > u, \hat{X}(t) > u) \nu(du) \\ &= \int_{\mathbb{R}} \bar{\phi}_{\mu,\sigma}(u) \nu(du) + \frac{1}{2\pi} \int_0^{\sin^{-1}(\rho_t)} \underbrace{\int_{\mathbb{R}} \exp\left\{ -\frac{(u-\mu)^2}{\sigma^2} \cdot \frac{1-\sin\theta}{\cos^2\theta} \right\} \nu(du)}_{=g(\theta)} \, d\theta. \end{aligned}$$

It follows that

$$\rho_t = \sigma^{-2} \cdot \sum_{j=1}^{N} \mathbf{cov}(X(t), X(t_j)) = \sigma^{-2} \cdot c_t^\intercal \cdot \lambda.$$

Since $g(\theta) > 0$ for all $\theta \in [0, \pi/2)$, we get

$$\int\limits_0^{\sin^{-1}(\rho_t)} g(\theta) d\theta \to \max_{\lambda \in \mathbb{R}^N} \quad \Longleftrightarrow \quad \sin^{-1}(\rho_t) \to \max_{\lambda \in \mathbb{R}^N},$$

which is equivalent to

$$\rho_t \to \max_{\lambda \in \mathbb{R}^N},$$

since $\sin^{-1}$ is a monotonically increasing function. □

Since $|\lambda^\intercal \cdot c_t| = \|Pr_{c_t}\lambda\|_2 \cdot \|c_t\|_2$, where $Pr_{c_t}\lambda$ is the orthogonal projection of λ onto c_t, it follows $\|Pr_{c_t}\lambda\|_2 \to \max\limits_{\lambda}$. The geometric interpretation of maximization problem (3.54) is given in Figure 3.9.

The problem (3.54) as a linear programming problem with a linear and a quadratic constraint appears to be the special case of *second order cone programming (SOCP)* or *quadratically constrained quadratic problem (QCQP)*. We solve it again via the Lagrangian formalism (as in the universal Kriging case).

Introduce numbers $b_0 := c_t^\intercal \Sigma^{-1} c_+$, $b_1 := e^\intercal \Sigma^{-1} c_t$, $b_2 := e^\intercal \Sigma^{-1} e$ if Σ is non-degenerate. Now we are ready to formulate and prove the following existence and uniqueness result.

Theorem 3.47 For an unknown $\mu \neq 0$, assume that there exists no $\beta \in \mathbb{R}$ such that $c_t = \beta e$, i.e. c_t and e are not parallel to each other, and let Σ be positive definite. Then, there exists a unique solution to (3.54) given by

$$\lambda = \Sigma^{-1} \left(\sqrt{\frac{\sigma^2 b_2 - 1}{b_0 b_2 - b_1^2}} (c_t - \frac{b_1}{b_2} e) + \frac{1}{b_2} e \right). \tag{3.55}$$

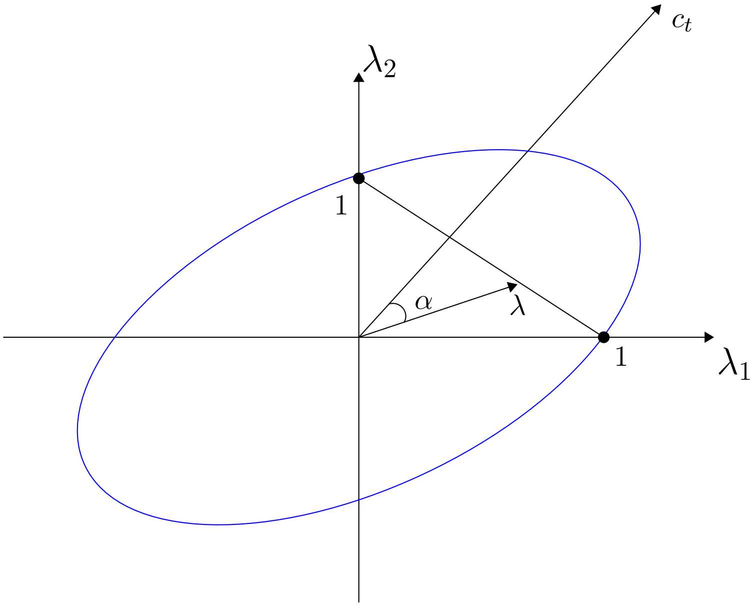


Fig. 3.9: The SOCP problem (3.54) for $N = 2$. Here, $c_t^\intercal \lambda = \|c_t\|_2 \|\lambda\|_2 \cos(\alpha) \to \max_\lambda$.

Proof Let $K := \{\lambda \in \mathbb{R}^n : \lambda^\intercal e = 1,\ \lambda^\intercal \Sigma \lambda = \sigma^2\}$ be the set of weights that satisfy the constraints in (3.54). For the vectors $e_j = (0, \ldots, 0, 1, 0 \ldots, 0)^\intercal,\ j = 1, \ldots, N$, it obviously holds $e_j \in K$. Since the function $\lambda^\intercal c_t$ is linear, it is continuous on the compact set K and thus attains its maximal value on K.

The *Lagrange function* for the problem (3.54) is given by

$$L(\lambda, \gamma, \delta) = c_t^\intercal \lambda + \gamma(\lambda^\intercal \Sigma \lambda - \sigma^2) + \delta(e^\intercal \lambda - 1),$$

where $\gamma, \delta \in \mathbb{R}$ are the *Lagrange multipliers*. Denote $\nabla_\lambda L := \left(\frac{\partial L}{\partial \lambda_1}, \ldots, \frac{\partial L}{\partial \lambda_N}\right)$. Computing $\nabla_\lambda L$ and setting

$$\nabla_\lambda L(\lambda, \gamma, \delta) = c_t + 2\gamma \Sigma \lambda + \delta e = 0$$

in order to find extreme points λ yields

$$2\gamma\lambda = -\Sigma^{-1}(c_t + \delta e). \tag{3.56}$$

Since $c_t \neq \delta e$ by assumption, it follows that $\gamma \neq 0$. Multiply (3.56) from the left by $e^\intercal$ and use $\lambda^\intercal e = e^\intercal \lambda = 1$ yields

$$2\gamma = e^\intercal \Sigma^{-1}(c_t + \delta e) = \underbrace{e^\intercal \Sigma^{-1} c_t}_{=b_1} + \delta \underbrace{e^\intercal \Sigma^{-1} e}_{=b_2} = b_1 + \delta b_2, \tag{3.57}$$

and combining (3.56) and (3.57) we get

$$\lambda = \frac{\Sigma^{-1}(c_t + \delta e)}{e^\intercal \Sigma^{-1}(c_t + \delta e)}. \tag{3.58}$$

We compute δ by plugging the above expression for λ into $\lambda^\intercal \Sigma \lambda = \sigma^2$, i.e.

$$(c_t^\intercal + \delta e^\intercal) \underbrace{\Sigma^{-1} \Sigma \Sigma^{-1}}_{=\Sigma^{-1}} (c_t + \delta e) = \sigma^2 (e^\intercal \Sigma^{-1} c_t + \delta e^\intercal \Sigma^{-1} e)^2,$$

or, in short form,

$$b_2(\sigma^2 b_2 - 1)\delta^2 + 2b_1(\sigma^2 b^2 - 1) + \delta^2 b_1^2 = b_0,$$

which results in

$$(b_2\delta + b_1)^2 = \frac{b_0 b_2 - b1^2}{\sigma^2 b_2 - 1}. \tag{3.59}$$

It can be shown that $\sigma^2 b_2 > 1$ and $b_0 b_2 \geq b_1^2$ due to positive definiteness of Σ. Moreover, $b_0 b_2 = b_1^2$ holds if and only if $c_t = \beta e$ for some $\beta \in \mathbb{R}$, which is prohibited by assumption.

Hence, equation (3.59) has two distinct soultions given by

$$\delta_{1,2} = -\frac{b_1}{b_2} \pm \frac{1}{b_2}\sqrt{\frac{b_0 b_2 - b_1^2}{\sigma^2 b_2 - 1}}.$$

The corresponding values of $\lambda_{1,2}$ can be calculated accordingly such that

$$c_t^\mathsf{T}\lambda_1 = \frac{b_1}{b_2} + \frac{1}{b_2}\sqrt{(b_0 b_2 - b_1^2)(\sigma^2 b_2 - 1)} \geq c_t^\mathsf{T}\lambda_2 = \frac{b_1}{b_2} - \frac{1}{b_2}\sqrt{(b_0 b_2 - b_1^2)(\sigma^2 b_2 - 1)},$$

hence λ_1 maximizes $c_t^\mathsf{T}\lambda$, leading to the unique solution

$$\lambda = \sqrt{\frac{\sigma^2 b_2 - 1}{b_0 b_2 - b_1^2}}\Sigma^{-1}(c_t - \frac{b_1}{b_2}e) + \frac{1}{b_2}\Sigma^{-1}e$$

if and only if $b_0 b_2 - b_1^2 \neq 0$, i.e. if and only if there exists no $\beta \in \mathbb{R}$ such that $c_t = \beta e$. □

Remark 3.48 (a) In Equation (3.55), the vectors $c_t - \frac{b_1}{b_2}e$ and e are orthogonal. Indeed, it holds that

$$\begin{aligned} e^\mathsf{T}(c_t - \frac{b_1}{b_2}e) &= \frac{b_2 e^\mathsf{T} c_t - b_1 \overbrace{e^\mathsf{T} e}^{=N}}{b_2} = b_2^{-1}(e^\mathsf{T}\Sigma^{-1}ee^\mathsf{T}c_t - Ne^\mathsf{T}\Sigma^{-1}c_t) \\ &= b_2^{-1}(\underbrace{e^\mathsf{T}e}_{=N}\, e^\mathsf{T}\Sigma^{-1}c_t - Ne^\mathsf{T}\Sigma^{-1}c_t) = 0, \end{aligned}$$

since Σ^{-1} and ee^T commute because Σ^{-1} is symmetric and $ee^\mathsf{T} = \begin{pmatrix} 1 & \dots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \dots & 1 \end{pmatrix}$.

(b) If c_t and e are parallel, i.e. there exists a $\beta \in \mathbb{R}$ such that $c_t = \beta e$, then $c_t^\mathsf{T} \equiv \mathit{const}$ for all $\lambda \in K$. Indeed, $c_t = \beta e$ implies $c_t^\mathsf{T}\lambda = \beta \underbrace{e^\mathsf{T}\lambda}_{1} = \beta$ for some $\beta \in \mathbb{R}$. Then, all $\lambda \in K$ are solutions to the problem (3.54), e.g. $\lambda = e_j,\ j = 1, \dots, N$, are solutions as well.

(c) The linear predictor $\hat{X}(t) = \sum_{j=1}^{N}\lambda_j X(t_j)$ is exact in the sense that $\hat{X}(t) = X(t_j),\ j = 1, \dots, n$. Indeed, here it follows from $c_t = \Sigma e_j$ that

$$\begin{aligned} b_0 &= e_j^\mathsf{T}\Sigma\underbrace{\Sigma^{-1}\Sigma}_{=I_n}e_j = e_j^\mathsf{T}e_j = C(t_j - t_j) = \sigma^2, \\ b_1 &= e^\mathsf{T}\underbrace{\Sigma^{-1}\Sigma}_{=I_n}e_j = e^\mathsf{T}e_j = 1, \end{aligned}$$

hence by (3.55) we have weights

$$\lambda = \Sigma^{-1}\left(\sqrt{\frac{\sigma^2 b_2 - 1}{\sigma^2 b_2 - 1}}(c_{t_j} - \frac{1}{b_2}) + \frac{1}{b_2}e\right) = \Sigma^{-1}c_{t_j} = e_j.$$

We can conclude that

$$\hat{X}(t_j) = e_j^\intercal (X(t_1), \dots, X(t_N))^\intercal = X(t_j), \quad j = 1, \dots, N.$$

Example 3.49 Consider the case $N = 2$ and compute weights $\lambda = (\lambda_1, \lambda_2)^\intercal$ in (3.55). Here, $\hat{X}(t) = \lambda_1 X(t_1) + \lambda_2 X(t_2)$, $K = \{(1,0),(0,1)\}$ is an intersection of an ellipsoid $\lambda^\intercal \Sigma \lambda = \sigma^2$ with the line $\lambda_1 + \lambda_2 = 1$. Hence, by (3.55) λ is of the form

$$\lambda = \begin{cases} (1,0), & C(t-t_1) > C(t-t_2), \\ (0,1), & C(t-t_1) < C(t-t_2), \\ (1,0) \text{ or } (0,1), & C(t-t_1) = C(t-t_2), \end{cases}$$

which yields

$$\hat{X}(t) = \begin{cases} X(t_1), & C(t-t_1) > C(t-t_2), \\ X(t_2), & C(t-t_1) < C(t-t_2), \\ X(t_1) \text{ or } X(t_2), & C(t-t_1) = C(t-t_2), \end{cases}$$

since the target functional to maximize over K is given by

$$\lambda^\intercal c_t = \lambda_1 C(t-t_1) + \lambda_2 C(t-t_2).$$

Remark 3.50 (a) If μ is known, we may set $\mu = 0$ without loss of generality and solve (3.53) ignoring the constraint $\lambda^\intercal e = 1$. This leads to the solution

$$\lambda = \sigma \frac{\Sigma^{-1} c_t}{\sqrt{c_t^\intercal \Sigma^{-1} c_t}}. \tag{3.60}$$

(b) We may compute the *square extrapolation error* given by

$$\mathbb{E}\left[\left(X(t) - \hat{X}(t)\right)^2\right] = \begin{cases} 2(\sigma^2 - \frac{b_1}{b_2} - \frac{1}{b_2}\sqrt{(b_0 b_2 - b_1^2)(\sigma^2 b_2 - 1)}), & \mu \text{ unknown}, \\ 2\sigma(\sigma - \sqrt{c_t^\intercal \Sigma^{-1} c_t}), & \mu = 0 \text{ known}. \end{cases}$$

This can be done using the explicit form of λ from (3.54) or (3.60) and

$$\mathbb{E}\left[\left(X(t) - \hat{X}(t)\right)^2\right] = \underbrace{\lambda^\intercal \Sigma \lambda}_{=\sigma^2} - 2c_t^\intercal \lambda + \sigma^2 = 2(\sigma^2 - c_t^\intercal \lambda).$$

Another important property of the prediction method (3.53) is its L^2 and a.s. consistency, which states that $\hat{X}(t) \to X(t)$ as $N \to \infty$ in the following sense.

Theorem 3.51 Let the covariance function C be positive definite, and assume there exists no $\beta \in \mathbb{R}$ such that $c_t = \beta e$, $e = (1, \dots, 1)^\intercal$.

(a) Assume that C is continuous, and let $\min\limits_{j=1,\ldots,N} \|t_j - t\|_2 \to 0$ as $N \to \infty$. Then, it holds that

$$\mathbb{E}\left[\left(\hat{X}(t) - X(t)\right)^2\right] \to 0, \quad N \to \infty.$$

(b) Let C be Hölder-continuous at zero with Hölder-index $\alpha > 0$. Assume that $\{t_1, \ldots, t_N\} \subset T_N := (h_N \mathbb{Z}^d) \cap W$ for some compact observation window W such that $\sum_{N=1}^{\infty} h_N^{\alpha} < \infty$. Then, it holds that

$$\hat{X}(t) \overset{a.s.}{\to} X(t), \quad N \to \infty.$$

Proof (a) Since $\lambda = e_j \in K$, we have

$$e_j^{\intercal} c_t = C(t - t_j) \leq \lambda^{\intercal} c_t, \quad j = 1, \ldots, N.$$

Taking $j_N := \operatorname*{argmin}\limits_{j=1,\ldots,N} \|t_j - t\|_2$, we can bound the mean-square error by

$$\mathbb{E}[\hat{X}(t) - X(t)]^2 = 2(\sigma^2 - c_t^{\intercal}\lambda) \leq 2(\sigma^2 - C(t - t_{j_N})) \to 0, \quad N \to \infty, \tag{3.61}$$

since C is continuous by assumption.

(b) Since the points t_j lie in the set T_N for all $j = 1, \ldots, N$, it holds that $\|t_{j_N} - t\|_2 \leq \sqrt{d} h_N$. Furthermore, the Hölder-continuity of C implies that there exist constants $C_1, C_2 > 0$ such that

$$|C(0) - C(t)| \leq C_1 \|t\|_2^{\alpha}$$

for all t with $\|t\|_2 \leq C_2$.

Then,

$$\begin{aligned}\sum_{N=1}^{\infty} \mathbb{E}\left[\left(\hat{X}(t) - X(t)\right)^2\right] &\overset{(3.61)}{\leq} 2 \sum_{N=1}^{\infty} |C(\delta) - C(t - t_j)| \\ &\leq 2C_1 \sum_{N=1}^{\infty} \|(t - t_j\|_2 \leq 2d^{\alpha/2} c_1 \sum_{N=1}^{\infty} h_N^{\alpha} < \infty.\end{aligned}$$

By the Tschebyschew inequality it follows that

$$\sum_{N=1}^{\infty} \mathbb{P}(|\hat{X}(t) - X(t)| > \varepsilon) \leq \frac{1}{\varepsilon^2} \sum_{N=1}^{\infty} \mathbb{E}\left[\left(\hat{X}(t) - X(t)\right)^2\right] < \infty$$

for all $\varepsilon > 0$, and thus $\hat{X}(t) \to X(t)$ a.s. as $N \to \infty$ by the lemma of Borel-Cantelli. □

Remark 3.52 It follows from the proof of Theorem 3.51 (b) that

$$\mathbb{E}\left[\left(\hat{X}(t) - X(t)\right)^2\right] \leq 2C_1 \min_{j=1,\ldots,N} \|t - t_j\|_2^{\alpha}.$$

That is, the speed of convergence of $\hat{X}(t)$ to $X(t)$ as $N \to \infty$ depends on the roughness of the paths of X, which is encoded in a larger constant C_1 or smaller index $\alpha > 0$.

In Figure 3.10 we consider a Gaussian process X with exponential covariance function $C(t) = e^{-|t|}$ and compare prediction results $\hat{X}$ for observation points $t_j = j$, $j = 1, \ldots, 100$, in Figure 3.10a and for $t_j = 0.2j$, $j = 1, \ldots, 500$, see Figure 3.10b. Figure 3.11 compares results of our level set predictor $\hat{X}$ for a Gaussian process X with Gaussian covariance $C(t) = e^{-t^2/2}$ and known mean $\mu = 0$ as well as the case of an unknown mean. The results are additionally compared to simple Kriging and ordinary Kriging.

3.5.2 Excursion metric projections

In the minimization problem (3.51) the target function was rewritten in the proof of Theorem 3.45 as

$$\int_{\mathbb{R}} \mathbb{E}\left[|A_X(u)\Delta A_{\hat{X}}(u)|\right]\nu(du) = \int_W \int_{\mathbb{R}} \underbrace{\mathbb{P}\left(t \in A_X(u)\Delta A_{\hat{X}}(u)\right)}_{=\Delta_{X,\hat{X}}(u)} \nu(du)dt,$$

where $\Delta_{X,\hat{X}}(u) = \mathbb{P}(\{X(t) > u\}\Delta\{\hat{X}(t) > u\})$. The inner integral can be seen as a measure of distance between the random variables $X(t)$ and $\hat{X}(t)$, which leads to the following definition.

Definition 3.53 For a finite measure ν on $\mathbb{R}$ and random variables $Y_1, Y_2 : \Omega \to \mathbb{R}$, we call

$$\mathbf{E}_\nu(Y_1, Y_2) := \int_{\mathbb{R}} \mathbb{P}\left(\{Y_1 > u\}\Delta\{Y_2 > u\}\right)\nu(du)$$

the *excursion pseudo-metric.*

Without loss of generality, assume ν to be a probability measure in the sequel. Let $L^0(\Omega, \mathcal{F}, \mathbb{P})$ be the space of all random variables on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The fact that $\mathbf{E}_\nu$ is a *pseudo-metric* can be seen from

(i) $\mathbf{E}_\nu : L^0(\Omega, \mathcal{F}, \mathbb{P})^2 \to [0, 1]$ is symmetric,

(ii) It satisfies the triangle inequality $\mathbf{E}_\nu(Y_1, Y_2) \leq \mathbf{E}_\nu(Y_1, Y_3) + \mathbf{E}_\nu(Y_3, Y_2)$ for any $Y_1, Y_2, Y_3 \in L^0(\Omega, \mathcal{F}, \mathbb{P})$.

Since $\mathbf{E}_\nu(Y_1, Y_2) = 0$ does not imply that $Y_1 = Y_2$ a.s., $\mathbf{E}_\nu$ fails to be a metric.

Let us rewrite our prediction problem in terms of metric projections with respect to $\mathbf{E}_\nu$. For that, we will need further properties of $\mathbf{E}_\nu$, which involve the notion of a *copulas.*

3.5.3 Copulas

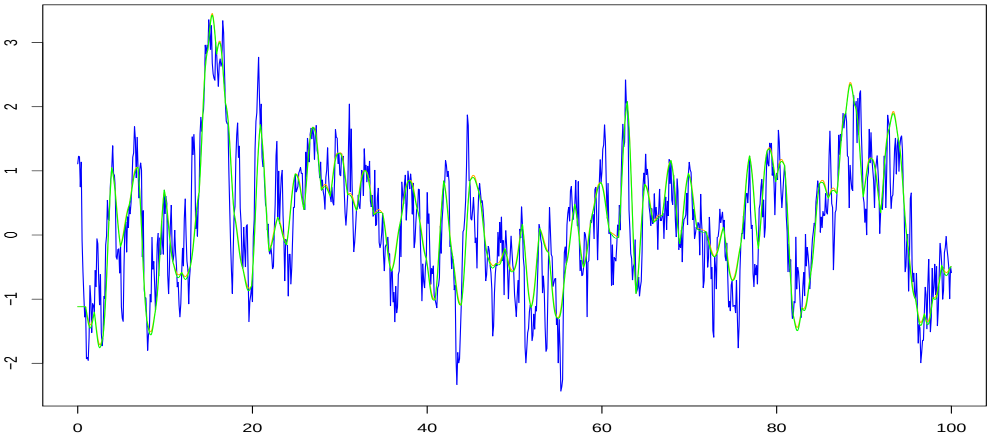
Definition 3.54 A fuction $C : [0, 1]^2 \to [0, 1]$ is a *(bivariate) copula* if it is a cumulative distribution function of a random vector (U_1, U_2), where $U_1, U_2 \sim U[0, 1]$.

Copulas measure the dependence between random variables, which finds its reflections in the following result.

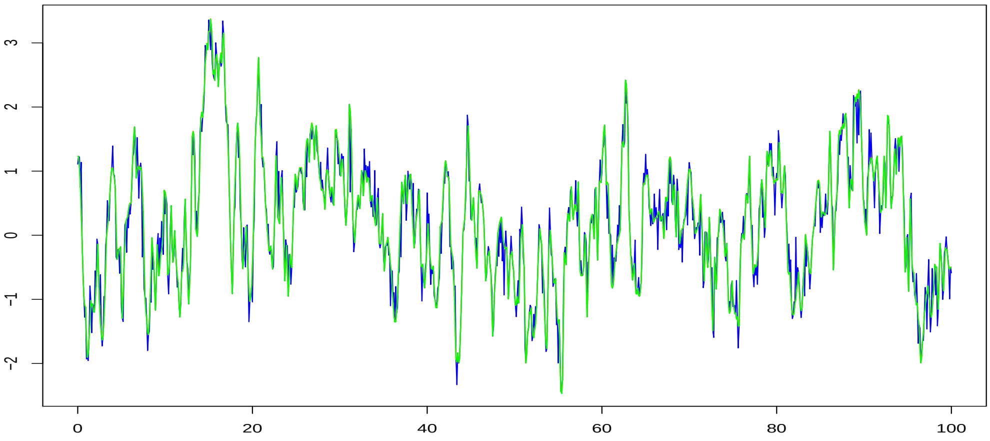
Theorem 3.55 (Sular, 1959): Let $Y = (Y_1, Y_2)$ be a random vector on $(\Omega, \mathcal{F}, \mathbb{P})$ with joint distribution function $F_Y(x_1, x_2) = \mathbb{P}(Y_1 \leq x_1, Y_2 \leq x_2)$, $x_1, x_2 \in \mathbb{R}$, and marginal distribution functions $F_{Y_j}(x) = \mathbb{P}(Y_j \leq x)$, $j = 1, 2$. Then, there exists a copula C such that

$$F_Y(x_1, x_2) = C(F_{Y_1}(x_1), F_{Y_2}(x_2)), \quad x_1, x_2 \in \mathbb{R}.$$

This copula is unique on the set $F_{Y_1}(\bar{\mathbb{R}}) \times F_{Y_2}(\bar{\mathbb{R}})$, where $F_{Y_j}(\bar{\mathbb{R}})$ is the image of $\bar{\mathbb{R}} = \mathbb{R} \cup \{\pm\infty\}$ under the mapping $F_{Y_1} : \bar{\mathbb{R}} \to [0, 1]$, $j = 1, 2$.



(a) X with exponential covariance $C(t) = e^{-|t|}$, observed at $t_j = j$, $j = 1, \ldots, 100$.



(b) X with exponential covariance $C(t) = e^{-|t|}$, observed at $t_j = 0.2j$, $j = 1, \ldots, 500$.

Fig. 3.10: Comparison of trajectories of Gaussian process X with exponential covariance (blue) and its predictor $\hat{X}$ with unknown mean (green) observed at $t_j = j$, $j = 1, \ldots, 100$ (3.10a) and $t_j = 0.2j$, $j = 1, \ldots, 500$ (3.10b).

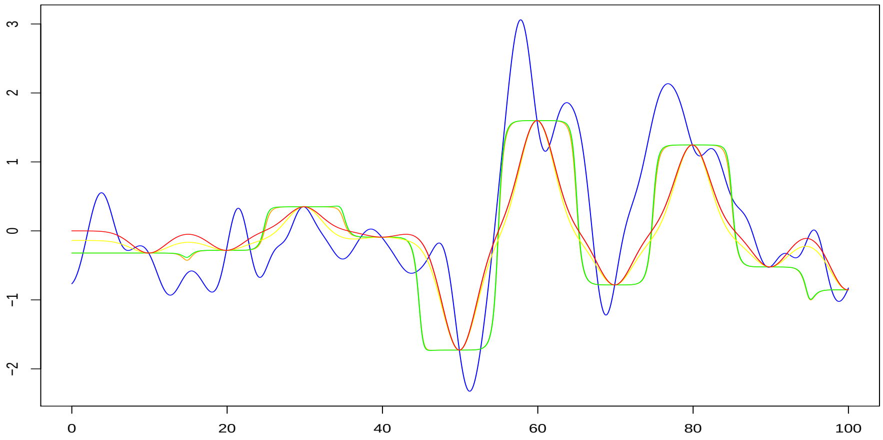


Fig. 3.11: A path of a Gaussian process X (blue) with gaussian covariance function $C(t) = e^{-t^2/2}$ is compared to our new linear predictor $\hat{X}$ with known mean $\mu = 0$ (orange), with unkonwn mean (green), simple kriging (red) and ordinary kriging (yellow).

Remark 3.56 Note that, if the marginal distribution functions F_{Y_j} are continuous on $\mathbb{R}$, then $F_{Y_j}(\bar{\mathbb{R}}) \subset [0,1]$ and hence the copula C in Theorem 3.55 is unique.

Example 3.57 Let $U = (U_1, U_2)$ be a random vector with its joint distribution function being a copula C and $U_j \sim U([0,1]),\ j = 1,2$.

(a) *Independence copula:* If U_1, U_2 are stochastically independent, then

$$C(x_1,x_2) = x_1 x_2, \quad x_1, x_2 \in [0,1].$$

(b) *Comonotonicity copula:* If $U_1 = U_2 = U_0$ a.s., then

$$C(x_1,x_2) = \mathbb{P}(U_0 \le x_1, U_0 \le x_2) = \mathbb{P}(U_0 \le \min\{x_1,x_2\}) = \min\{x_1,x_2\}, \quad x_1,x_2 \in [0,1].$$

Notation: $M_2(x,y) = \min\{x,y\},\ x,y \in [0,1]$.

(c) *Linear dependence copula:* If $U_1 = U_0,\ U_2 = 1 - U_0,\ U_0 \sim U([0,1])$, then

$$\begin{aligned} C(x_1,x_2) &= \mathbb{P}(U_0 \le x_1, 1 - U_0 \le x_2) \\ &= \mathbb{P}(1 - x_2 \le U_0 \le x_1) = \max\{0, x_1 + x_2 - 1\}, \quad x_1,x_2 \in [0,1]. \end{aligned}$$

Notation: $W_2(x,y) := \max\{0, x+y-1\},\ x,y \in [0,1]$.

Theorem 3.58 (Hoeffding-Fréchet bounds (1940, 1951)): For any copula $C : [0,1]^2 \to [0,1]$, it holds that

$$W_2(x,y) \le C(x,y) \le M_2(x,y), \quad x,y \in [0,1].$$

Proof (i) We first show the upper bound. Using the monotonicity of probability measures, it follows that

$$\begin{aligned} C(x,y) &= \mathbb{P}\left(\{U_1 \le x\} \cap \{U_2 \le y\}\right) \le \min \underbrace{\mathbb{P}(U_1 \le x)}_{=x}, \underbrace{\mathbb{P}(U_2 \le x)}_{=y} \\ &= \min\{x,y\} = M_2(x,y), \quad x,y \in [0,1], \end{aligned}$$

where $C(\cdot,\cdot)$ is a joint distribution function of (U_1,U_2), $U_j \sim U([0,1])$, $j=1,2$.

(ii) For the lower bound, write

$$\begin{aligned} C(x,y) &= 1 - \mathbb{P}(\{U_1 > x\} \cup \{U_2 > y\}) \ge 1 - \mathbb{P}(U_1 > x) - \mathbb{P}(U_2 > y) \\ &= 1 - (1-x) - (1-y) = x + y - 1, \quad x,y \in [0,1]. \end{aligned}$$

Since $C(x,y) \ge 0$ as it is a cumulative distribution function, we have $C(x,y) \ge W_2(x,y)$. □

3.5.4 Excursion metric and its properties

For any $Y_1, Y_2 \in C^0(\Omega, \mathcal{F}, \mathbb{P})$ we introduce the notation

$$Y_1 \wedge Y_2 := \min\{Y_1, Y_2\}, \quad Y_1 \vee Y_2 := \max\{Y_1, Y_2\}.$$

Then, $\Delta_{Y_1,Y_2}(u) = \mathbb{P}(\{Y_1 > u\}\Delta\{Y_2 > u\})$ rewrites as

$$\begin{aligned} \Delta_{Y_1,Y_2}(u) &= \mathbb{P}(Y_1 > u) + \mathbb{P}(Y_2 > u) - 2\mathbb{P}(Y_1 > u, Y_2 > u) \\ &= \mathbb{P}(Y_1 \le u) + \mathbb{P}(Y_2 \le u) - 2\mathbb{P}(Y_1 \le u, Y_2 \le u) \\ &= F_{Y_1}(u) + F_{Y_2}(u) - 2C(F_{Y_1}(u), F_{Y_2}(u)) \end{aligned}$$

by Theorem 3.55, where F_{Y_j}, $j = 1,2$, are the marginal distribution functions of (Y_1, Y_2) and C is a copula. Moreover, note that

$$\Delta_{Y_1,Y_2}(u) = \mathbb{P}(Y_1 \vee Y_2 > u) - \mathbb{P}(Y_1 \wedge Y_2 > u) = \mathbb{P}(Y_1 \wedge Y_2 \le u) - \mathbb{P}(Y_1 \vee Y_2 \le u), \qquad (3.62)$$

which leads to the following result.

Lemma 3.59 Let ν be a probability law of a random variable $U : \Omega \to \mathbb{R}$ representing the random choice of an excursion level. Then it holds for any $Y_1, Y_2 \in L^0(\Omega, \mathcal{F}, \mathbb{P})$ that

(a) $\mathbf{E}_\nu(Y_1, Y_2) = \mathbb{E}\left[|F_U(Y_2-) - F_U(Y_1-)|\right]$, where F_U is the cumulative distribution function of U and $F_U(x-) = \lim\limits_{y \to x-0} F_U(y)$ for any $x \in \mathbb{R}$.

(b) $\mathbf{E}_\nu(Y_1, Y_2) = \mathbb{P}(Y_1 \wedge Y_2 \le U \le Y_1 \vee Y_2)$.

Proof (a) Applying Equation (3.62) yields

$$\begin{aligned} \mathbf{E}_\nu(Y_1, Y_2) &= \int_{\mathbb{R}} \mathbb{E}\left[\mathbf{1}(u < Y_1 \vee Y_2) - \mathbf{1}(u < Y_1 \wedge Y_2)\right] \nu(du) \\ &= \mathbb{E}\left[F_U(Y_1 \vee Y_2-) - F_U(Y_1 \wedge Y_2-)\right] = \mathbb{E}\left[|F_U(Y_2-) - F_U(Y_1-)|\right]. \end{aligned} \qquad (3.63)$$

(b) Equivalently, (3.63) can be rewritten as $\mathbf{E}_\nu(Y_1, Y_2) = \mathbb{P}(Y_1 \wedge Y_2 \leq U \leq Y_1 \vee Y_2)$.

□

Note that in part (b), the random level U separates Y_1 and Y_2, which motivated M. Taylor (1984) to call $\mathbf{E}_\nu(Y_1, Y_2)$ a *separation pseudo-metric*. Furthermore, if F_U is continuous, it follows that $F_U(x-) = F_U(x)$ for all $x \in \mathbb{R}$ leading to $\mathbf{E}_\nu(Y_1, Y_2) = \mathbb{E}[|F_U(Y_1) - F_U(Y_2)|]$, which is called F_U*-madogram* in geostatistics.

Denote by $\mathcal{X}_S \subseteq L^0(\Omega, \mathcal{F}, \mathbb{P})$ the subspace of random variables with support $S \subset \mathbb{R}$.

Theorem 3.60 If F_U is strictly increasing on S, then $\mathbf{E}_\nu$ is a metric on $\mathcal{X}_S \times \mathcal{X}_S$.

Proof (i) The symmetry of $\mathbf{E}_\nu$ is trivial.

(ii) For the triangle inequality, we compute

$$\begin{aligned}
\mathbf{E}_\nu(Y_1, Y_2) &\overset{(*)}{=} \mathbb{E}\left[|F_U(Y_1-) - F_U(Y_2-)|\right] \\
&\leq \mathbb{E}\left[|F_U(Y_1-) - F_U(Y_3-)|\right] + \mathbb{E}\left[|F_U(Y_2-) - F_U(Y_3-)|\right] \\
&= \mathbf{E}_\nu(Y_1, Y_3) + \mathbf{E}_\nu(Y_2, Y_3)
\end{aligned}$$

for any $Y_1, Y_2, Y_3 \in \mathcal{X}_S$, where the equality $(*)$ follows from Lemma 3.59.

(iii) Let $\mathbf{E}_\nu(Y_1, Y_2) = 0$ for some $Y_1, Y_2 \in \mathcal{X}_S$. By Lemma 3.59, we have $F_U(Y_1-) = F_U(Y_2-)$ a.s., and since F_U is monotonically increasing on S, it follows that $\mathbb{P}(Y_1 = Y_2) = 1$, i.e. $Y_1 = Y_2$ a.s. Ultimately, it holds that $\mathbf{E}_\nu$ is a metric on $\mathcal{X}_S \times \mathcal{X}_S$.

□

Question: Which choice of U or $\nu(\cdot)$ is preferable from the practical point of view for the prediction of random variables?

Consider $Y_1, Y_2 \in \mathcal{X}_S$. If for example $S \subset \mathbb{R}_+$, $\text{supp}(U) \subset \mathbb{R}$, then

$$F_U(Y_1 \vee Y_2-) = F_U(Y_1 \wedge Y_1-) = 1,$$

which leads to a degenerate metric. Thus, we may require $\mathsf{supp}(U) \cap S \neq \emptyset$, or, ideally, $\mathsf{supp}(U) = S$. This allows for the choices $F_U = F_{Y_1}$ or $F_U = F_{Y_2}$. Without loss of generality, assume $F_U = F_{Y_1}$ in the sequel. Then, relation (3.63) can be stated as

$$\begin{aligned}
\mathbf{E}_\nu(Y_1, Y_2) &= \mathbb{E}[F_{Y_1}(Y_1 \vee Y_2)] - \mathbb{E}[F_{Y_1}(Y_1 \wedge Y_2)] \\
&= 2\mathbb{E}[F_{Y_1}(Y_1 \vee Y_2)] - \mathbb{E}[F_{Y_1}(Y_1)] - \mathbb{E}[F_{Y_1}(Y_2)] \\
&= 2\mathbb{E}[F_{Y_1}(Y_1 \vee Y_2)] - \frac{1}{2} - \mathbb{E}[F_{Y_1}(Y_2)], \qquad (3.64)
\end{aligned}$$

since $F_{Y_1}(Y_1) \sim U([0,1])$ and therefore $\mathbb{E}[F_{Y_1}(Y_1)] = \frac{1}{2}$ in the above.

Now consider the case $Y_1 \overset{d}{=} Y_2$, where $F_{Y_1}(x) := F_1(x)$ is strictly increasing, e.g. if Y_1 is absolutely continuous on S. Then, Equation (3.64) simplifies to $\mathbf{E}_\nu(Y_1, Y_2) = 2\mathbb{E}F_1(Y_1 \vee Y_2) - 1$. By Theorem 3.60, it is a metric on the space $\{Y \in \mathcal{X}_S : \mathbb{P}(Y \leq x) = F_1(x)\} \subset L^0(\Omega, \mathcal{F}, \mathbb{P})$.

Lemma 3.61 Let $Y_1, Y_2 \in L^0(\Omega, \mathcal{F}, \mathbb{P})$ have an absolutely continuous cumulative distribution function F_1. Then, the excursion metric $\mathbf{E}_\nu$ with $\nu \sim F_1$, i.e. $\nu = dF_1$, has the representation

$$\mathbf{E}_\nu(Y_1, Y_2) = 1 - 2\int_0^1 C(x,x)dx,$$

where C is a copula of (Y_1, Y_2).

Proof We know that

$$\Delta_{Y_1,Y_2}(u) = 2(F_1(u) - C(F_1(u), F_1(u))), \quad u \in \mathbb{R}.$$

Substituting $x = F_1(u)$ leads to

$$\mathbf{E}_\nu(Y_1, Y_2) = 2\int\limits_{\mathbb{R}} (F_1(u) - C(F_1(u), F_1(u)))\, dF_1(u) = 2\int\limits_0^1 (x - C(x,x))dx = 1 - 2\int\limits_0^1 C(x,x)dx.$$

□

We arrive at the following definition.

Definition 3.62 Let $Y_1, Y_2 \in L_{F_1}$, where L_{F_1} is the space of random variables with absolutely continuous cumulative distribution function F_1. The excursion metric $G = \mathbf{E}_\nu$ with $\nu \sim F_1$ given by

$$G(Y_1, Y_2) = 1 - 2\int_0^1 C(x,x)dx$$

is called *Gini metric*. Here, the function C is the copula of the random vector (Y_1, Y_2).

If the set $\{C(x,x), x \in [0,1]\}$ were convex, the term $2\int_0^1 (x - C(x,x))dx$ equals to the *Gini coefficient* of the *Lorenz curve* $\{(x, C(x,x)), x \in [0,1]\}$ used in econometrics for example to measure the concentration of wealth in the society. Hence, the name “Gini metric“.

Remark 3.63 Taking into account that $x = \min\{x,x\} = M_2(x,x)$, we can rewrite $G(Y_1, Y_2)$ as

$$G(Y_1, Y_2) = 2\int\limits_0^1 (x - C(x,x))dx = 2\|M_2(x,x) - C(x,x)\|_{L^1[0,1]},$$

where $\|h\|_{L^1[0,1]} = \int_0^1 |h(x)|dx$ is the L^1-norm of h. Hence, $G(Y_1, Y_2)$ can be interpreted as the L^1-distance between the diagonal of the complete dependence copula to the diagonal of the copula of (Y_1, Y_2).

Lemma 3.64 For any $Y_1, Y_2 \in L_{F_1}$, it holds that $G(Y_1, Y_2) \in \left[0, \frac{1}{2}\right]$. Furthermore, $G(Y_1, Y_2) = \frac{1}{2}$ implies $Y_2 = f(Y_1)$ a.s., where f is a decreasing function such that $F_1(x) = 1 - F_1(f^{-1}(x))$, $x \in \mathbb{R}$.

Proof By Theorem 3.58, we have

$$\begin{aligned}\frac{1}{4} = \underbrace{\int_0^1 \max\{0, 2x-1\}dx}_{=\int_{1/2}^1 (2x-1)dx} &= \int_0^1 W_2(x,x)dx \le \int_0^1 C(x,x)dx \\ &\le \int_0^1 M_2(x,x)dx = \int_0^1 xdx = \frac{1}{2},\end{aligned}$$

which yields

$$0 = 1 - 2\frac{1}{2} \leq G(Y_1, Y_2) = 1 - 2\int\limits_0^1 C(x,x)dx \leq 1 - 2\frac{1}{4} = \frac{1}{2}.$$

The upper bound is attained whenever $Y_2 = f(Y_1)$ for a decreasing function f, which means that

$$F_1(x) = \mathbb{P}(Y_2 \leq x) = \mathbb{P}(f(Y_1) \leq x) = \mathbb{P}(Y_1 \geq f^{-1}(x)) = 1 - \mathbb{P}(Y_1 < f^{-1}(x)) = 1 - F_1(f^{-1}(x)),$$

since $Y_1 \overset{d}{=} Y_2$ and Y_2 has an absolutely continuous distribution. □

Example 3.65 (a) Assume that the distribution F_1 is symmetric around $\mu \in \mathbb{R}$, i.e. $F_1(x) = 1 - F_1(\mu - x), \quad x \in \mathbb{R}$. Then, it holds that $G(Y_1, Y_2) = \frac{1}{2}$ if $Y_1 + Y_2 = \mu$ a.s. with $f(x) = \mu - x$, $x \in \mathbb{R}$.

(b) If Y_1, Y_2 are stochastically independent, then $G(Y_1, Y_2) = 1 - 2\int_0^1 x^2 dx = \frac{1}{3}$, since $C(x,y) = x \cdot y$ for $x, y \in [0,1]$.

3.5.5 Forecasting via excursion metric

Let X be a random variable which has to be predicted based on “observations“ $X_1, \ldots, X_N$ such that $X_j \overset{d}{=} X$, $j = 1, \ldots, N$. Assume that X has a continuous distribution function F_X. We consider the predictor $\hat{X}_\lambda$ of X to be of the form

$$\hat{X}_\lambda = g(X_1, \ldots, X_n, \lambda),$$

where $g : \mathbb{R}^N \times \mathbb{R}^N \mapsto \mathbb{R}$ is a Borel-measurable function of the sample $X_1, \ldots, X_N$ and $(\lambda_1, \ldots, \lambda_N) = \lambda \in \Lambda \subset \mathbb{R}$ are the prediction parameters. Here, Λ is the set of admissible parameter values, i.e.

$$\Lambda = \Lambda_g := \left\{\lambda \in \mathbb{R}^N : \hat{X}_\lambda \overset{d}{=} X\right\}.$$

Since $F_X \in C(\mathbb{R})$, we may rewrite the condition $\hat{X}_\lambda \overset{d}{=} X$ as $F_X(\hat{X}_\lambda) \overset{d}{=} F_X(X) \sim U([0,1])$.

The main idea of an *excursion-based forecast* is to look for $\hat{X}_\lambda = g(X_1, \ldots, X_N, \hat{\lambda})$, where

$$\hat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{arginf}}\, \mathbf{E}_{F_X}(X, \hat{X}_\lambda).$$

Let us give some examples of g and Λ_g depending on the distribution class of F_X. The function g has to be chosen such that $\Lambda_g \neq \emptyset$.

Example 3.66 (a) If $(X, X_1, \ldots, X_N)$ is infinitely divisible, then

$$g(X_1, \ldots, X_N, \lambda) = \sum_{j=1}^N \lambda_j X_j. \tag{3.65}$$

(b) If $(X, X_1, \ldots, X_N)$ is max-stable, then

$$g(X_1, \ldots, X_N, \lambda) = \max_{j=1,\ldots,N} \lambda_j X_j. \tag{3.66}$$

Example 3.67 (a) Let $(X, X_1, \dots, X_N)$ be Gaussian with marginal distribution $N(\mu, \sigma^2)$ and $\Sigma = (\mathbf{cov}(X_i, X_j))_{i,j=1}^N$ and let g be as in (3.65). Then,

$$\Lambda_g = \{\lambda \in \mathbb{R}^N : \lambda^\intercal \Sigma \lambda = \sigma^2, \lambda^\intercal e = 1\}$$

is an ellipsoid of dimension $N-1$, see also Section 3.5.1.

(b) Let $(X, X_1, \dots, X_N)$ be a subgaussian random vector with stability index $\alpha \in (0,2)$ and underlying i.i.d. standard Gaussian components. Moreover, let g be as in (3.65). Then,

$$\Lambda_g = \{\lambda \in \mathbb{R}^N : \|\lambda\|_2 = 1\} = S^{n-1}.$$

(c) Let $(X, X_1, \dots, X_N)$ be a $S\alpha S$ random vector with stability index $\alpha \in (0,2)$, spectral measure Γ of $(X_1, \dots, X_N)$ and scale parameter 1 for the marginal distributions. For g as in (3.65), we have

$$\Lambda_g = \left\{ \lambda \in \mathbb{R}^N : \int\limits_{S^{N-1}} |\langle s, \lambda \rangle|^\alpha \Gamma(ds) = 1 \right\},$$

which is a closed subset of $\mathbb{R}^N$ by the dominated convergence theorem. However, the structure of this set may be quite complex.

(d) Let $(X, X_1, \dots, X_N)$ be a max-stable random vector with Fréchet(α)-marginals and tail dependence function l_N of $(X_1, \dots, X_N)$. For g as in (3.66), it holds that

$$\Lambda_g = \{\lambda \in \mathbb{R}_+^N : l_N(\lambda_1^\alpha, \dots, \lambda_N^\alpha) = 1\}.$$

This is true, since

$$\mathbb{P}(\max_{j=1,\dots,N} \lambda_j X_j \le x) = \exp\left\{-x^{-\alpha} l_N(\lambda_1^\alpha, \dots, \lambda_N^\alpha)\right\}, \quad \lambda_1, \dots, \lambda_N \ge 0.$$

The prediction problem $\mathbf{E}_{F_X}(X, \hat{X}_\lambda) \to \inf_{\lambda \in \Lambda}$ can be rewritten in terms of the Gini metric, i.e.

$$G(X, \hat{X}_\lambda) = 1 - 2 \int\limits_0^1 C_{X, \hat{X}_\lambda}(x, x) dx \to \inf_{\lambda \in \Lambda_g},$$

where $C_{X, \hat{X}_\lambda}(\cdot, \cdot)$ is the copula of $(X, \hat{X}_\lambda)$. Consequently, this yields

$$\hat{\lambda} = \operatorname*{argsup}_{\lambda \in \Lambda_g} \int_0^1 C_{X, \hat{X}_\lambda}(x, x) dx. \tag{3.67}$$

Example 3.68 (a) Let $(X, X_1, \dots, X_N)$ be Gaussian as in Example 3.67 (a) with mean $\mu = 0$ and variance $\sigma^2 = 1$. In view of Section 3.5.1, the copula diagonal $C_{X, \hat{X}_\lambda}(x, x)$ is equal to

$$C_{X, \hat{X}_\lambda}(x, x) = x^2 + \frac{1}{2\pi} \int\limits_0^{\sin^{-1}(c_\lambda)} \exp\left\{ -(\varphi^{-1}(x))^2 \frac{1 - \sin\theta}{\cos^2\theta} \right\} d\theta,$$

where $\varphi^{-1}(x)$ is the quantile function of $N(0,1)$ and

$$c_\lambda = \mathbf{corr}(X, \hat{X}_\lambda) = \sum_{j=1}^N \lambda_j \mathbf{cov}(X, X_j).$$

(b) Let $(X, X_1, \dots, X_N)$ be max-stable as in example 3.67 (d). Then, $C_{X,\hat{X}}$ is an extreme-value copula with diagonal given by

$$C_{X,\hat{X}}(x,x) = x^{\theta_\lambda}, \quad x \in [0,1],$$

where θ_λ is the extremal coefficient of $(X, \hat{X}_\lambda)$, i.e. $\theta_\lambda = l_2(1,1)$ with l_2 being the tail dependence function of $(X, \hat{X}_\lambda)$. It follows that

$$\int\limits_0^1 C_{X,\hat{X}_\lambda}(x,x)dx = \int\limits_0^1 x^{\theta_\lambda} dx = \frac{1}{\theta_\lambda + 1},$$

hence the maximization problem

$$\int_0^1 C_{X,\hat{X}_\lambda}(x,x)dx \to \sup_{\lambda \in \Lambda_g}$$

is equivalent to the minimization problem

$$\theta_\lambda \to \inf_{\lambda \in \Lambda_g}.$$

As it was seen in Example 3.67, the structure of the admissible parameter set Λ_g may be quite complex leading to non-linear non-convex optimization for finding $\hat{\lambda}$ in Equation (3.67), which is difficult to solve. A possible way out would be to replace the rigid condition $F_X(\hat{X}_\lambda) \sim U([0,1])$ in the definition of Λ_g by an "approximate" condition

$$\rho\left(\mathbb{P}(F_X(\hat{X}_\lambda) \leq \cdot), F_{U[0,1]}(\cdot)\right) \leq \varepsilon$$

for a small fixed $\varepsilon > 0$, where $\rho(\cdot,\cdot)$ is any handy metric on the space of cumulative distribution functions on $\mathbb{R}$.

For simplicity, we consider the *2-Wasserstein distance* in place of ρ, which is defined as follows.

Definition 3.69 Let Y_1, Y_2 be random variables with quantile functions F_1^{-1} and F_2^{-1}. For $p \geq 1$, the *p-Wasserstein distance* of Y_1, Y_2 is defined as

$$W_p(Y_1, Y_2) := \left(\int_0^1 \left|F_1^{-1}(x) - F_2^{-1}(x)\right|^p dx\right)^{1/p}.$$

Note that, the Wasserstein distance acts on the space of distributions of random variables rather than on $L^0(\Omega, \mathcal{F}, \mathbb{P})$ itself. For $p = 2$ and $Y_2 \sim U([0,1])$, we may rewrite

$$\begin{aligned} W_2^2(Y_1, Y_2) &= \int_0^1 \left(F_1^{-1}(x) - x\right)^2 dx \\ &= \int_0^1 F_1^{-1}(x)^2 dx - 2\int_0^1 x F_1^{-1}(x) dx + \underbrace{\int_0^1 x^2 dx}_{=1/3} \\ &\overset{(*)}{=} \int_0^1 y^2 dF_1(y) - \int_0^1 y dF_1^2(y) + \frac{1}{3} \\ &= \frac{1}{3} + \mathbb{E}Y_1^2 - \mathbb{E}(Y_1 \vee Y), \end{aligned} \tag{3.68}$$

where $\mathbb{P}(Y_1 \vee Y \leq y) = F_1^2(y)$ and Y is an independent copy of Y_1, and for the equality $(*)$ we substituted $y = F_1^{-1}(x)$. Integration by parts then yields

$$W_2^2(Y_1, Y_2) = \frac{1}{3} + \int_0^1 F_1(y)(F_1(y) - 2y)dy. \tag{3.69}$$

This allows us to rewrite the prediction problem (3.67) using the form of (3.64) of the excursion metric and the approximative constraint $W_2^2(X, \hat{X}) \leq \varepsilon$ as

$$\hat{\lambda} = \operatorname*{arginf}_{\lambda \in \Lambda} \left\{ 2\mathbb{E}\left[F_X(X \vee \hat{X}_\lambda)\right] - \mathbb{E}\left[F_X(\hat{X}_\lambda)\right] + \gamma W_2^2(X, \hat{X}_\lambda) \right\}, \tag{3.70}$$

where $-\frac{1}{2}$ in (3.64) is ommited and the Wasserstein distance $W_2^2(\cdot,\cdot)$ is of the form (3.68) or (3.69), where the set $\Lambda \subset \mathbb{R}^N$ does not depend on $g(.,.)$ (it may be $\mathbb{R}^N$, $[-M, M]^N$ or $\mathbb{R}_+^N$). The factor $\gamma \geq 0$ weighs the significance of how close F_X has to be to $F_{\hat{X}_\lambda}$. Under certain conditions, which are to be specified later, the infimum in (3.70) is attained, thus turning $\operatorname*{arginf}_{\lambda \in \Lambda}$ to an $\operatorname*{argmin}_{\lambda \in \Lambda}$. It follows that

$$W_2^2(F_X(X), F_X(\hat{X}_\lambda)) = \begin{cases} \mathbb{E}\left[F_X^2(\hat{X}_\lambda)\right] - \mathbb{E}\left[F_X(\hat{X}_\lambda) \vee Y\right] + \frac{1}{3}, \\ \int_0^1 F_{F_X(\hat{X}_\lambda)}(y) \left(F_{F_X(\hat{X}_\lambda)}(y) - 2y\right) dy + \frac{1}{3}, \end{cases}$$

where Y is an independent copy of $F_X(\hat{X}_\lambda)$ and $F_{F_X(\hat{X}_\lambda)}$ is the cumulative distribution function of $F_X(\hat{X}_\lambda)$.

Let us examine the existence of a solution to (3.70).

Theorem 3.70 Let the joint distribution of $(X, X_1, \ldots, X_N)$ be absolutely continous. If the following conditions are met

(I) Λ is compact in $\mathbb{R}^N$,

(II) $C_{X,\hat{X}_\lambda}(x, x)$ is uniformly continuous on $\lambda \in \Lambda$,

(III) for each $\lambda \in \Lambda$, the distribution of $\hat{X}_\lambda$ is absolutely continuous with probability density function $f_{\hat{X}_\lambda}$ such that $f_{\hat{X}_\lambda} : \Lambda \to L^1(\mathbb{R})$ is continuous on Λ with respect to the L^1-norm,

then there exists a solution to minimization problem (3.70).

Proof Using Lemma 3.61, the target functional in (3.70) can be rewritten as

$$\bar{\phi}(\lambda) := 2 - 2\int_0^1 C_{X,\hat{X}_\lambda}(x, x)dx - \mathbb{E}(Z_\lambda) - \gamma \int_0^1 F_{Z_\lambda(y)}dy, \quad \lambda \in \Lambda, \tag{3.71}$$

where $Z_\lambda := F_X(\hat{X}_\lambda)$ and $\mathbb{E}[Z_\lambda] = \int_0^1 \mathbb{P}(Z_{\lambda > y})dy = 1 - \int_0^1 F_{\hat{X}_\lambda}(F_X^{-1}(y))dy$. As an integral with parameter λ, i.e $\lambda \mapsto \int_0^1 C_{X,\hat{X}_\lambda}(x, x)dx$, continuity on Λ follows from condition (II).

Furthermore, for any sequence $\{\lambda_k\} \subset \Lambda$ with $\lambda_k \to \lambda_0 \in \Lambda$ as $k \to \infty$, we have

$$\begin{aligned} \sup_{x \in \mathbb{R}} \left| F_{\hat{X}_{\lambda_k}}(x) - F_{\hat{X}_{\lambda_0}}(x) \right| &= \sup_{x \in \mathbb{R}} \left| \int_{-\infty}^{x} f_{\hat{X}_{\lambda_k}}(y) - f_{\hat{X}_{\lambda_0}}(y)dy \right| \\ &\leq \int_{\mathbb{R}} \left| f_{\hat{X}_{\lambda_k}}(y) - F_{\hat{X}_{\lambda_0}}(y) \right| dy \to 0, \end{aligned}$$

as $k \to \infty$ by condition (III). Therefore, $F_{\hat{X}_{\lambda_k}}(F_X^{-1}(y))$ is uniformly continuous on Λ with respect to $y \in [0,1]$. Applying the theorem on the continuity of integrals with parameters proves the continuity of $\mathbb{E}[Z_\lambda]$ on Λ.

Similarly, the term $F_{Z_\lambda}(y)(2y - F_{Z_\lambda}(y))$ is uniformly continuous on Λ with respect to $y \in [0,1]$, so that also the third term in (3.71) is lies in $C(\Lambda)$. Hence, the target functional $\bar{\phi} \in C(\Lambda)$ attains its minimum on the compact set Λ. □

Remark 3.71 The L^1- continuity of $f_{\hat{X}_\lambda}$ in condition (III) means that

$$\left\| f_{\hat{X}_{\lambda_k}} - f_{\hat{X}_{\lambda_0}} \right\|_1 = \int\limits_{\mathbb{R}} \left| f_{\hat{X}_{\lambda_k}}(y) - f_{\hat{X}_{\lambda_0}}(y) \right| dy \to 0, \quad k \to \infty,$$

for any sequence $\{\lambda_k\} \subset \Lambda$ with $\lambda_k \to \lambda_0$ as $k \to \infty$. However, due to

$$\frac{1}{2} \int\limits_{\mathbb{R}} \left| f_{\hat{X}_{\lambda_k}}(y) - f_{\hat{X}_{\lambda_0}}(y) \right| dy = d_{TV}(\hat{X}_{\lambda_k}, \hat{X}_{\lambda_0}),$$

where d_{TV} is the total variation distance, we see that this is equivalent to $\hat{X}_{\lambda_k} \xrightarrow{TV} \hat{X}_{\lambda_0}$ as $k \to \infty$.

Example 3.72 In the following we show that Theorem 3.70 holds true for any Gaussian random vector $(X, X_1, \ldots, X_N)$ with $N(0,1)$-distributed marginals. Since Λ_g in Example 3.67 (a) is an ellipsoid in $\mathbb{R}^{N-1}$, it is sufficient to consider $\Lambda = [-M, M]^N \supset \Lambda_g$ for $M > 0$ large enough. Using the exact form of the copula diagonal $C_{X,\hat{X}_\lambda}(x,x)$ form Example 3.68, we may see that

$$\left| C_{X,\hat{X}_{\lambda_1}}(x,x) - C_{X,\hat{X}_{\lambda_2}}(x,x) \right| \leq \frac{1}{2\sigma} \left| \sin^{-1}(c_{\lambda_1}) - \sin^{-1}(c_{\lambda_2}) \right|$$

uniformly on $x \in [0,1]$ due to the inequality

$$\exp\left\{ -\left(\varphi^{-1}(x)\right)^2 \frac{1 - \sin(\theta)}{\cos^2(\theta)} \right\} \leq 1.$$

This shows condition (II) of Theorem 3.70.

To show the validity of condition (III), assume the covariance matrix Σ of $(X_1, \ldots, X_N)$ to be positive definite such that $\hat{X}_\lambda \sim N(0, \lambda^\intercal \Sigma \lambda)$ has a density for all $\lambda \neq 0$. This density is obviously continuous on $\mathbb{R}^N \backslash \{0\}$ with respect to the L^1-norm.

In the next result, we show that it is often sufficient to consider bounded spaces Λ only, e.g. $\Lambda = \left\{ \lambda \in \mathbb{R}^N : \|\lambda\|_2 \leq \mu \right\}$ with $\mu > 0$.

Lemma 3.73 Assume that there exists a $\lambda_0 \in \Lambda$ such that $\bar{\phi}(\lambda_0) < 1 + \frac{\gamma}{3}$. Let $\hat{X}_{\lambda_k} \xrightarrow{p} \infty$ as $k \to \infty$ for any sequence $\{\lambda_k\} \subset \Lambda$ with$\|\lambda_k\|_2 \to \infty$ as $k \to \infty$. Then, there exists a constant $M > 0$ such that

$$\min_{\lambda \in \Lambda} \bar{\phi}(\lambda) = \min_{\lambda \in \Lambda : \|\lambda\|_2 \leq M} \bar{\phi}(\lambda).$$

Proof Consider the target functional $\bar{\phi}$ in its form

$$\bar{\phi}(\lambda) = 2\mathbb{E}\left[F_X(X \vee \hat{X}_\lambda)\right] - \mathbb{E}\left[F_X(\hat{X}_\lambda)\right] + \gamma\left(\mathbb{E}\left[F_X^2(\hat{X}_\lambda)\right] - \mathbb{E}\left[F_X(\hat{X}_\lambda \vee Y)\right] + \frac{1}{3}\right),$$

where Y is an independent copy of $\hat{X}_\lambda$. The sequences $\{F_X(\hat{X}_\lambda)\}$, $\{F_X^2(\hat{X}_\lambda)\}$, $\{F_X(X\hat{X}_\lambda)\}$ and $\{F_X(\hat{X}_\lambda \vee Y)\}$ are uniformly integrable since they are a.s. bounded by 0 and 1. Hence, their expectations tent to 1 as $\lambda_k \to \infty$, $k \to \infty$, while $\hat{X}_{\lambda_k} \xrightarrow{p} \infty$, $k \to \infty$, and since $F_X(y) \to 1$, $y \to \infty$. Then, it follows that

$$\bar{\phi}(\lambda_k) \to 1 + \frac{\gamma}{3}, \quad k \to \infty.$$

Choosing $M > 0$ such that $\bar{\phi}(\lambda_k) > \bar{\phi}(\lambda_0)$ for all $k \in \mathbb{N}$ such that $\|\lambda_k\|_2 > M$ concludes the proof of this lemma. □

Example 3.74 Assume there exists a $\lambda_0 \in \Lambda$ such that

(a) $\hat{X}_{\lambda_0} = X$ a.s. This may happen for some prediction function g, if $X = X_{j_0}$ for some $j_0 \in \{1, \dots, N\}$. Then,

$$\bar{\phi}(\lambda_0) = 2\frac{1}{2} - \frac{1}{2} + \gamma\left(\frac{1}{3} - \frac{1}{2} + \frac{1}{3}\right) = \frac{1}{2} + \gamma\frac{1}{6} < 1 + \frac{\gamma}{3}$$

for all $\gamma \geq 0$.

(b) $\hat{X}_{\lambda_0}$ and X are stochastically independent. Then,

$$\begin{aligned}\mathbb{P}\left(F_X(X \vee \hat{X}_{\lambda_0}) \leq x\right) &= \mathbb{P}\Big(F_X(x) \vee \underbrace{F_X(\hat{X}_{\lambda_0})}_{=Y_1} \leq x\Big)\\ &= \mathbb{P}\left(F_X(X) \leq x\right)\mathbb{P}\left(Y_1 \leq x\right) = xF_{Y_1}(x)\end{aligned}$$

and

$$\begin{aligned}2\mathbb{E}\left[F_X(X \vee \hat{X}_{\lambda_0})\right] - \mathbb{E}\left[Y_1\right] &= 2\int_0^1 xd(xF_{Y_1}(x)) - \int_0^1 xdF_{Y_1}(x)\\ &= 2\left(1 - \int_0^1 xF_{Y_1}(x)dx\right) - 1 + \int_0^1 dF_{Y_1}(x)\\ &= 1 + \int_0^1 F_{Y_1}(x)(1-2x)dx\\ &= 1 + \int_0^1 \underbrace{\left(F_{Y_1}(x) - F_{Y_1}\left(\frac{1}{2}\right)\right)(1-2x)}_{\leq 0} dx < 1\end{aligned}$$

by integration by parts, since $F_{Y_1}(\frac{1}{2})\int_0^1(1-2x)dx = 0$. Since

$$\mathbb{E}\left[F_X^2(\hat{X}_{\lambda_0})\right] - \mathbb{E}\left[F_X(\hat{X}_{\lambda_0}) \vee Y\right] < 0,$$

we get $\bar{\phi}(\lambda_0) < 1 + \frac{\gamma}{3}$ for all $\gamma \geq 0$.

Example 3.75 The condition $\hat{X}_{\lambda_k} \xrightarrow{p} \infty$ as $k \to \infty$ with $\|\lambda\| \to \infty$, $\lambda_k = (\lambda_k(1), \dots, \lambda_k(N))$ is satisfied for $\Lambda = \mathbb{R}_n^+$ and

$$\hat{X}_\lambda = \sum_{j=1}^{N} \lambda_k(j)X_j$$

or

$$\hat{X}_\lambda = \max_{j=1,\ldots,N} \lambda_k(j) X_j$$

if $X_j \geq 0$ a.s.

3.5.6 Excursion-based prediction of stationary random fields

Let $X = \{X(t), t \in \mathbb{R}^d\}$ be a strictly stationary measurable random field with marginal distribution $F_{\theta_0} \in \{F_\theta, \theta \in \Theta\}$, where $\{F_\theta, \theta \in \Theta\}$ is a parametric family of absolutely continuous distributions and $\Theta \subset \mathbb{R}^k$ is its parameter space. The distribution F_θ may be heavy-tailed with no finite moments at all.

Let X be observed on a set of locations $T_0 = W_0 \cap \mathbb{Z}_n$, where $W_0 \subset \mathbb{R}^d$ is a compact set and $\mathbb{Z}_h = h_1\mathbb{Z} \times \cdots \times h_d\mathbb{Z}$ is a d-dimensional grid with mesh sizes $h = (h_1, \ldots, h_d) \in (0, \infty)^d$. We denote the observation of X on T_0 by $X_{T_0} = \{X(t), t \in T_0\}$.

For a location $t \in \mathbb{Z}_n,\ t \in W_0$, the goal is to predict $X(t)$ from the so-called *forecast sample* $X_{T_f} = \{X(t_1), \ldots, X(t_N)\},\ T_f = \{t_1, \ldots, t_N\} \subset \mathbb{Z}_h$. As before, we are looking for a predictor $\hat{X}_\lambda$ of $X(t)$ of the form $\hat{X}_\lambda = g(\lambda, X_{T_f}),\ \lambda \in \Lambda \subset \mathbb{R}^N$, such that

$$\hat{\lambda} = \underset{\lambda \in \Lambda}{\operatorname{argmin}} \Big\{ \overbrace{2\mathbb{E}\left[F_{\theta_0}(X(t) \vee \hat{X}_\lambda)\right] - \mathbb{E}\Big[\underbrace{F_{\theta_0}(\hat{X}_\lambda)}_{=Z_\lambda}\Big] - \frac{1}{2}}^{=\mathbf{E}_{F_{\theta_0}}(X(t),\hat{X}_\lambda)} + \gamma W_2^2(F_{Z_\lambda}, F_U)\Big\}, \tag{3.72}$$

where F_U is the cumulative distribution function of $U \sim U([0,1])$. Furthermore, in the above we have

$$W_2^2(F_{Z_\lambda}, F_U) = \begin{cases} \mathbb{E}\left[Z_\lambda^2\right] - \mathbb{E}\left[Z_\lambda \vee Y\right] + \frac{1}{3} \\ \int\limits_0^1 F_{Z_\lambda}(y)(F_{Z_\lambda}(y) - 2y)dy + \frac{1}{3} \end{cases}, \tag{3.73}$$

where Y is an independent copy of Z_λ with cumulative distribution function F_{Z_λ}.

Theorem 3.76 (Weak consistency): Let the random field X be absolutely continuous. Assume that there exists a $\tilde{\lambda}_k \in \Lambda$ such that $\hat{X}_{\tilde{\lambda}_k} = X(t_k)$ a.s. for all $k = 1, \ldots, N$, and $\min\limits_{j=1,\ldots,N} \|t_j - t\|_2 \to 0$ as $N \to \infty$. Then,

$$\hat{X}_\lambda(t) \xrightarrow{p} X(t)$$

as $N \to \infty$.

Proof Let $\tilde{t}_N = \underset{j=1,\ldots,N}{\operatorname{argmin}} \|t_j - t\|_2$ and $\tilde{\lambda} \in \Lambda$ such that $\hat{X}_{\tilde{\lambda}} = X(\tilde{t}_N)$ a.s. Then,

$$\begin{aligned} &\mathbf{E}_{F_{\theta_0}}\left(X(t), \hat{X}_{\hat{\lambda}}\right) + \gamma W_2^2\left(F_{Z_{\hat{\lambda}}}, F_\eta\right) \\ &\leq \mathbf{E}_{F_{\theta_0}}\left(X(t), \hat{X}_{\tilde{\lambda}}\right) + \gamma W_2^2\left(F_{Z_{\tilde{\lambda}}}, F_\eta\right) = \mathbf{E}_{F_{\theta_0}}\left(X(t), X(\tilde{t}_N)\right)) + \gamma \cdot 0 \to 0 \end{aligned}$$

as $N \to \infty$, since the left-hand side is a minimum of all $\lambda \in \Lambda$ and $\hat{X}_{\tilde{\lambda}} = X(\tilde{t}_N) \overset{d}{=} X(t)$ due to stationarity implies that $W_2^2(F_{Z_{\tilde{\lambda}}}, F_n) = 0$. The last convergence holds since

$$\mathbf{E}_{F_{\theta_0}}\left(X(t), X(\tilde{t}_N)\right) = \mathbb{E}\left[\left|F_{\theta_0}(X(t))) - F_{\theta_0}(X(\tilde{t}_N))\right|\right]$$

and

$$F_{\theta_0}(X(\tilde{t}_N)) \xrightarrow{p} F_{\theta_0}(X(t))$$

due to $(X(\tilde{t}_N)) \xrightarrow{p} (X(t))$ as $N \to \infty$ and $F_{\theta_0} \in C(\mathbb{R})$. The L^1-convergence of $F_{\theta_0}(X(\tilde{t}_N))$ to $F_{\theta_0}(X(t))$ holds since $\{F_{\theta_0}(X(\tilde{t}_N)) - F_{\theta_0}(X(t))\}_{N\geq 1}$ is uniformly integrable due to its a.s. boundedness. □

In order to compute the forecast $\hat{X}_{\bar{\lambda}}$ numerically, we first estimate θ_0 by a statistic $\hat{\theta}$, and then use $F_{\hat{\theta}}$ as a plug-in estimate of F_θ. Next, we discretize all expectations and integrals in (3.72) to minimize the functional

$$\bar{\phi}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} Q_j \to \min_{\lambda\in\Lambda}, \tag{3.74}$$

where

$$\begin{aligned} Q_j(\lambda) = {} & 2F_{\hat{\theta}}\left(X(t+s_j) \vee g(\lambda, X_{T_f+s_j})\right) - F_{\hat{\theta}}\left(g(\lambda, X_{T_f+s_j})\right) \\ & + \gamma\left(F^2_{\hat{\theta}}\left(g(\lambda, X_{T_f+s_j})\right) - F_{\hat{\theta}}\left(g(\lambda, X_{T_f+s_j})\right) \vee Y_j\right), \end{aligned} \tag{3.75}$$

for $j = 1, \ldots, n$, $\gamma \geq 0$. Here, it holds that $\{s_1, \ldots, s_n\} = \{s \in \mathbb{Z}_n : s + T_f \cup \{t\} \subset T_0\}$, and the sets $T_f + s_j$ are called *learning samples*, $j = 1, \ldots, n$, see Figure 3.12.

The random variables Y_j are independent copies of $F_{\hat{\theta}}\left(g(\lambda, X_{T_f+s_j})\right)$. In practice, the sample $Y_1, \ldots, Y_n$ may be obtained from

$$\left(F_{\hat{\theta}}\left(g(\lambda, X_{T_f+s_1}), \ldots, F_{\hat{\theta}}\left(g(\lambda, X_{T_f+s_n})\right)\right)\right)$$

by resampling, e.g. bootstrapping. Note that in (3.75), we used the first variant of the 2-Wasserstein distance from (3.73). Alternatively, the second variant can be used as well, discretizing the integral $\int_0^1 \ldots dy$ therein by a sum.

To guarantee that

$$\frac{1}{n}\sum_{j=1}^{n} Q_j(\lambda) \to \mathbf{E}_{F_{\theta_0}}(X(t), \hat{X}_\lambda) + \gamma \cdot W_2^2(F_{Z_\lambda}, F_n) \quad a.s.$$

as $n \to \infty$, we require the ergodicity of X together with strong consistency of $\hat{\theta}$, i.e. $\hat{\theta} \to \theta_0$ a.s. as $n \to \infty$. The minimization problem in (3.74) may be solved using the stochastic *subgradient descent method*.

3.5.7 Stochastic approximation and gradient descent

Our functional $\bar{\phi}(\lambda) = \frac{1}{n}\sum_{j=1}^{n} Q_j(\lambda)$ in Equation (3.74) can be seen as $\bar{\phi}(\lambda) = \mathbb{E}Q_\xi(\lambda)$, where $\lambda \in \Lambda$, $\xi \sim U(\{1, \ldots, n\})$. The procedures of stochastic optimization (the so-called *procedures of Robbins-Monro* [35]) seek a solution to this problem as

$$\lambda^{(l+1)} = \Pi_\Lambda\left[\lambda^{(l)} - \eta_l \cdot \varphi(j_l, \lambda^{(l)})\right], \quad l \in \mathbb{N}_0, \tag{3.76}$$

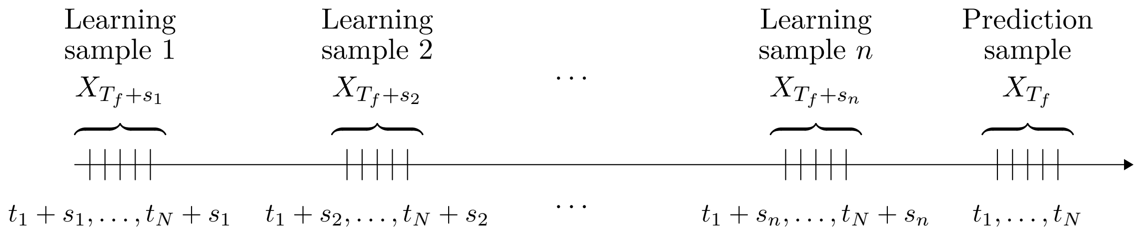


Fig. 3.12: Prediction and learning samples $X_{T_f+s_j} = \{X(t_1+s_j),\dots,X(t_N+s_j)\}$, $j = 0,1,\dots,n$ ($s_0 = 0$ for $d = 1$).

where $\Pi_\Lambda[\cdot]$ is a metric (back) projection onto the space Λ, $\eta_l > 0$ is a *step length factor* such that

$$\sum_{l=1}^{\infty} \eta_l = \infty \quad \text{and} \quad \sum_{l=1}^{\infty} \eta_l^2 < \infty, \tag{3.77}$$

j_l is an independent realization of ξ, i.e., a member chosen randomly from $\{1,\dots,n\}$, and (for the case of stochastic (sub-)gradient of $Q_j(.)$) $\varphi(j,\lambda) = \nabla^* Q_j(\lambda)$ is the (sub-)gradient of $Q_j(\cdot)$ with respect to $\lambda \in \Lambda \subset \mathbb{R}^N$. The difference to the *classical batch (sub-)gradient descent* lies in the use of $\phi(j,\lambda) = \bar{\phi}(\lambda)$ in the classical case, which does not depend on a realization j of ξ.

Note that a sequence $\{\eta_l\}$ satisfying the conditions (3.77) can be for example $\eta_l = \frac{1}{l}$. The iterations in (3.76) stop whenever $|\lambda^{(l^*+1)} - \lambda^{(l^*)}| < \delta$ for some small threshold $\delta > 0$. Then, the solution $\hat{\lambda}$ to the minimization problem (3.74) may be chosen to be either

$$\hat{\lambda} = \operatorname*{argmin}_{l=0,\dots,l^*} \bar{\phi}(\lambda^{(l)})$$

or

$$\hat{\lambda} = \frac{1}{l^* - l_0} \sum_{l=l_0}^{l^*-1} \lambda^{(l)},$$

the so-called *Polyak-Ruppert averaging*, which excludes the *burn-in period* of length l_0.

It is well known that the convergence of all gradient-like descent methods heavily depends on the right choice of the initial value $\lambda^{(0)}$. Hence, we recommend to choose $\lambda^{(0)}$ to be the outcome of another optimization procedure of the minimization problem

$$\bar{\phi}(\lambda) \to \min_{\lambda\in\Lambda},$$

e.g. simulated annealing genetic search, etc. In order to avoid the back-projection Π_Λ for $\Lambda = \mathbb{R}^N_+$, it is reasonable to use $\lambda^{(l)} = \left((\lambda_1^{(l)})^2 \dots (\lambda_N^{(l)})^2\right)$ in the computation of the gradients.

Let us consider the convergence of stochastic (sub-)gradient descent.

Theorem 3.77 Let $\Lambda \subset \mathbb{R}^N$ be compact. Assume that Q_j, $j = 1,\dots,n$, are piecewise $C^2(\Lambda)$ and bounded on Λ, $\bar{\phi}(\lambda) \geq 0$ almost everywhere on Λ, and there exists a unique $\lambda^* \in \Lambda$ such that $\nabla\bar{\phi}(\lambda^*) = 0$, $\bar{\phi}(\lambda^*) = \bar{\phi}(\lambda)$, $\lambda \in \Lambda$. Assume that for $i,j \in \{1,\dots,N\}$ it holds that

$$\langle \nabla Q_i(\lambda), \nabla Q_j(\lambda)\rangle \geq 0 \tag{3.78}$$

almost everywhere on Λ and that this inequality is strict for at least one pair $(i,j) \in \{1,\dots,n\}^2$. Then, the stochastic gradient descent

$$\lambda^{(l+1)} - \lambda^{(l)} = -\eta_l \nabla Q_\xi(\lambda^{(l)}), \quad \xi \sim U(\{1,\dots,n\}) \tag{3.79}$$

with a sequence $\{\eta_l\}$ satisfying (3.77) converges a.s., i.e.

$$\lambda^{(l)} \to \lambda^* \quad a.s.,$$

as $l \to \infty$.

Proof For $\bar{\phi} \in C^2(\Lambda)$ a.e. on Λ its Taylor expansion is given by

$$\bar{\phi}\left(\lambda^{(l+1)}\right) - \phi\left(\lambda^{(l)}\right) = \left\langle \nabla\bar{\phi}\left(\lambda^{(l)}\right), \lambda^{(l+1)} - \lambda^{(l)} \right\rangle + \frac{1}{2}\left(\lambda^{(l+1)} - \lambda^{(l)}\right)^{\intercal} H\bar{\phi}\left(\lambda^{(l)}\right)\left(\lambda^{(l+1)} - \lambda^{(l)}\right),$$

where $\langle\cdot,\cdot\rangle$ denotes the Euclidean scalar product in $\mathbb{R}^N$ and $\tilde{\lambda}^{(l)} = \lambda^{(l)} + \delta(\lambda^{(l+1)} - \lambda^{(l)})$, $\delta \in (0,1)$. Furthermore, $H\bar{\phi}(\tilde{\lambda}^{(l)}) = \frac{1}{n}\sum_{j=1}^{n} HQ_j(\tilde{\lambda}^{(l)})$ is the Hessian matrix of $\bar{\phi}$ at $\tilde{\lambda}^{(l)}$.

Then, the above in combination with (3.79) yields

$$\begin{aligned}&\bar{\phi}\left(\lambda^{(l+1)}\right) - \phi\left(\lambda^{(l)}\right)\\ &= \frac{1}{n}\sum_{j=1}^{n}\eta_l\left(-\left\langle \nabla Q_j\left(\lambda^{(l)}\right), \nabla Q_\xi\left(\lambda^{(l)}\right)\right\rangle + \frac{\eta_l}{2}\nabla Q_\xi\left(\lambda^{(l)}\right) H\nabla Q_j\left(\tilde{\lambda}^{(l)}\right)\nabla Q_\xi\left(\lambda^{(l)}\right)\right). \end{aligned} \tag{3.80}$$

Due to the condition $\sum_{l=1}^{\infty}\eta_l^2 < \infty$ we have $\eta_l \to 0$ as $l \to \infty$. Additionally, since $Q_j \in C^2(\Lambda)$ piecewise, the gradient and Hessian matrix of Q_j, i.e. ∇Q_j and HQ_j, are bounded on the compact set Λ for all $j = 1, \dots, n$.

From Equation (3.78) we get

$$\bar{\phi}\left(\lambda^{(l+1)}\right) - \bar{\phi}\left(\lambda^{(l)}\right) \leq 0$$

as $l \to \infty$. Since $\bar{\phi}(\lambda) \geq 0$ a.e. on Λ, it follows that there exists an a.s. limit, which we denote by $\lim_{l\to\infty}\bar{\phi}(\lambda^{(l)}) := \bar{\phi}_\infty$ a.s.

It remains to show that

$$\bar{\phi}_\infty = \min_{\lambda\in\Lambda}\phi(\lambda).$$

Since $\xi \in U\{1,\dots,n\}$, it holds that $\bar{\phi}_\infty$ is a random variable with $\mathbb{E}[\bar{\phi}_\infty] < \infty$ by boundness on Λ. Thus,

$$\mathbb{E}\left[\bar{\phi}_\infty\right] = \sum_{l=0}^{\infty}\mathbb{E}\Big[\underbrace{\bar{\phi}(\lambda^{(l+1)}) - \bar{\phi}(\lambda^{(l)})}_{\leq 0 \text{ for sufficiently large } l}\Big] + \bar{\phi}\left(\lambda^{(0)}\right),$$

which implies

$$\sum_{l=0}^{\infty}\mathbb{E}\left[\bar{\phi}\left(\lambda^{(l)}\right) - \bar{\phi}\left(\lambda^{(l+1)}\right)\right] < \infty.$$

Plugging-in (3.80) and dissolving the expectation with respect to ξ yields

$$\begin{aligned}&\sum_{l=0}^{\infty}\mathbb{E}\left[\bar{\phi}\left(\lambda^{(l)}\right) - \phi\left(\lambda^{(l+1)}\right)\right]\\ &= \frac{1}{n^2}\sum_{l=0}^{\infty}\sum_{i,j=1}^{n}\left(\eta_l\left\langle \nabla Q_i\left(\lambda^{(l)}\right), \nabla Q_j\left(\lambda^{(l)}\right)\right\rangle - \frac{\eta_l^2}{2}\nabla Q_j\left(\lambda^{(l)}\right)^{\intercal} HQ_i\left(\tilde{\lambda}^{(l)}\right)\nabla Q_j\left(\lambda^{(l)}\right)\right)\\ &= \sum_{l=0}^{\infty}\left(\eta_l\left\|\nabla\bar{\phi}\left(\lambda^{(l)}\right)\right\|_2^2 - \frac{\eta_l^2}{2}\nabla\bar{\phi}\left(\lambda^{(l)}\right)^{\intercal} H\bar{\phi}\left(\tilde{\lambda}^{(l)}\right)\nabla\bar{\phi}\left(\lambda^{(l)}\right)\right).\end{aligned}$$

Since $\nabla\bar{\phi}$ and $H\bar{\phi}$ are bounded on Λ and $\sum_{l=1}^{\infty}\eta_l^2 < \infty$, the second term in the sum is bounded a.s. Hence, also the first term has to be bounded a.s., and since $\sum_{l=1}^{\infty}\eta_l = \infty$, it has to hold

$$\|\nabla\bar{\phi}(\lambda^{(l)})\|_2 \to 0 \quad a.s.$$

as $l \to \infty$.

It follows that $\lambda^* = \lim_{l\to\infty}\lambda^{(l)}$ is a point of extremum of $\bar{\phi}$. Since $\bar{\phi}(\lambda^{(l)}) \downarrow \bar{\phi}_\infty$ and there exists a unique minimum of $\bar{\phi}$ by assumption, we have $\bar{\phi}_\infty = \min_{\lambda\in\Lambda}\bar{\phi}(\lambda)$. □

Now let us turn to the computation of the forecast $\hat{X}_\lambda$. Assume that the marginal distribution F_θ has a density f_θ, and that the random vectors $(X(t+s_j), X(t_1+s_j), \dots, X(t_N+s_j))$, $j = 1,\dots,n$ have joint densities. Then, it is reasonable to assume

$$\mathbb{P}(X(t+s_j) = g(\lambda, X_{T_f+s_j})) = 0 \quad \text{and} \quad \mathbb{P}(Y_j = f_{\hat{\theta}} g(\lambda, X_{T_f+s_j})) = 0$$

for $j = 1,\dots,n$, as well as

$$\mathbb{P}(g(\lambda, X(T_f+s_i)) = g(\lambda, X_{T_f+s_j})) = 0$$

for $i, j = 1,\dots,n$.

As a consequence, it becomes easy to calculate the sub-gradient $\nabla^* Q_j$, i.e.

$$\begin{aligned}\nabla^* Q_j = &\Big(2\cdot\mathbf{1}(X(t+s_j) < g(\lambda, X_{T_f+s_j})) - 1\Big) f_{\hat{\theta}}(g(\lambda, X_{T_f+s_j}))\cdot\nabla^* g(\lambda, X_{T_f+s_j})\\ &+\gamma\Big(2F_{\hat{\theta}}(g(\lambda, X_{T_f+s_j})) - \mathbf{1}(g(\lambda, Y_{T_f+s_j}) < g(\lambda, X_{T_f+s_j}))\Big)\cdot f_{\hat{\theta}}(g(\lambda, X_{T_f+s_j}))\nabla^* g(\lambda, X_{T_f+s_j})\\ &-\gamma\cdot\mathbf{1}(g(\lambda, Y_{T_f+s_j}) \geq g(\lambda, X_{T_f+s_j}))f_{\hat{\theta}}(g(\lambda, Y_{T_f+s_j}))\nabla^* g(\lambda, Y_{T_f+s_j}),\end{aligned}$$

where $Y_{T_f+s_j}$ is an independent copy of $X_{T_f+s_j})$ for all $j = 1,\dots,n$. The subgradients $\nabla^* g$ depend on the form of g, i.e. we have $\nabla^* g(\lambda, X_{T_f+s_j}) = X_{T_f+s_j}$ for $g(\lambda, X_{T_f}) = \sum_{i=1}^{N}\lambda_i X(t_i)$ and

$$\nabla^* g(\lambda, X_{T_f+s_j}) = \Big(X(t_i+s_j)\mathbf{1}(\lambda_i X(t_i+s_j) = \max_{k=1\dots N}\lambda_k X(t_k+s_j)),\ i = 1,\dots,N\Big)$$

for $g(\lambda, X_{T_f}) = \max_{i=1,\dots,N}\lambda_i X(t_i)$.

3.5.8 Excursion-based prediction of max-stable random fields

Let us apply the above theory to the extrapolation of stationary max-stable random fields from Section 1.3. Let $X = \{X(t), t \in \mathbb{R}^d\}$ be a stationary ergodic max-stable random field with Fréchet(α)-marginal distributions, $\alpha > 0$. Let

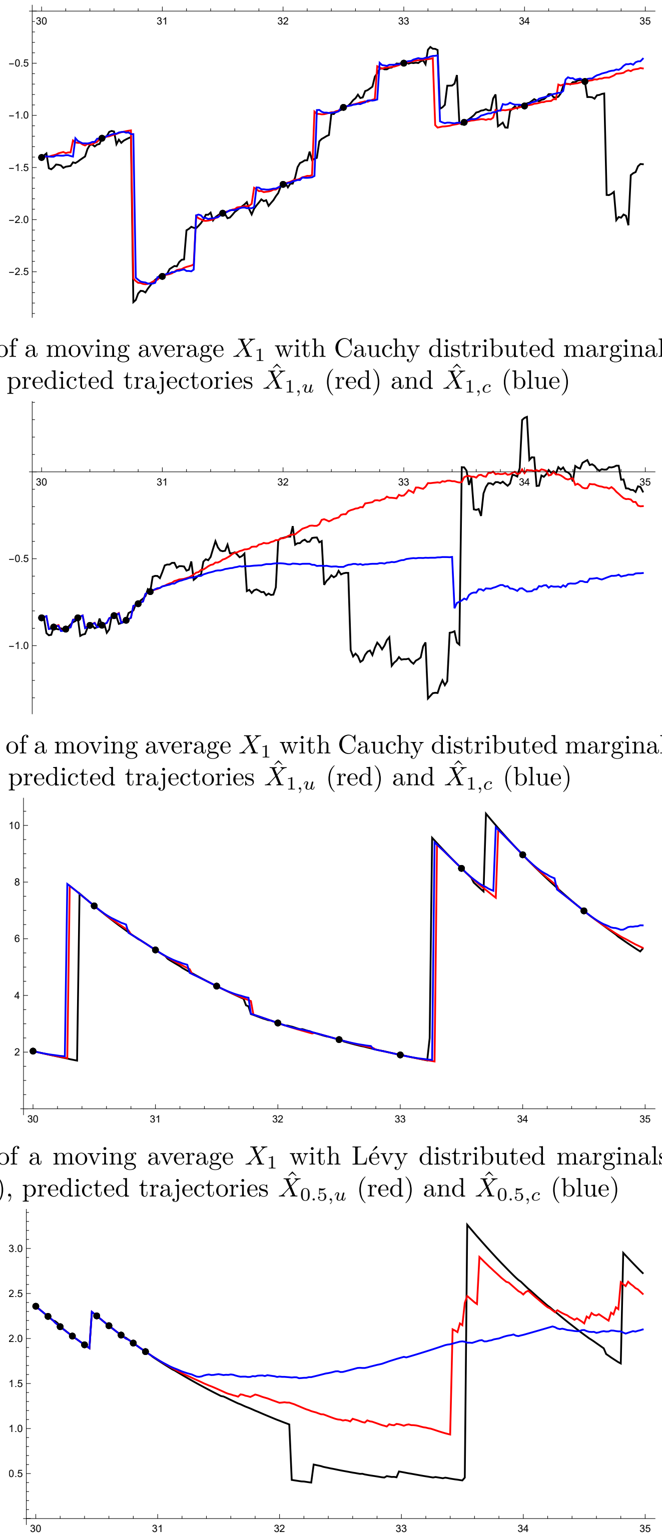
$$\hat{X} = \max_{j=1,\dots,N}\lambda_j X(t_j)$$

be the predictor of $X(t)$ with $\lambda_1,\dots,\lambda_N \geq 0$.

Lemma 3.78 The pairwise extremal coefficient θ_λ of $(X(t), \hat{X}(t))$ is given by

$$\theta_\lambda = l_{N+1}(1, \lambda_1^\alpha, \dots, \lambda_N^\alpha),$$

where $l_{N+1}(\cdot)$ is the tail dependence function of $(X(t), X(t_1), \dots, X(t_N))$.



(a) Interpolation of a moving average X_1 with Cauchy distributed marginals: True trajectory $X_1(t)$ (black), predicted trajectories $\hat{X}_{1,u}$ (red) and $\hat{X}_{1,c}$ (blue)

(b) Extrapolation of a moving average X_1 with Cauchy distributed marginals: True trajectory $X_1(t)$ (black), predicted trajectories $\hat{X}_{1,u}$ (red) and $\hat{X}_{1,c}$ (blue)

(c) Interpolation of a moving average X_1 with Lévy distributed marginals: True trajectory $X_{0.5}(t)$ (black), predicted trajectories $\hat{X}_{0.5,u}$ (red) and $\hat{X}_{0.5,c}$ (blue)

(d) Extrapolation of a moving average X_1 with Lévy distributed marginals: True trajectory $X_{0.5}(t)$ (black), predicted trajectories $\hat{X}_{0.5,u}$ (red) and $\hat{X}_{0.5,c}$ (blue)

Fig. 3.13: Excursion based interpolation and extrapolation.

Proof It is known that

$$\lim_{x\to\infty} \mathbb{P}\left(\hat{X}(t) > x \mid X(t) > x\right) = 2 - \theta_\lambda.$$

Since $\mathbb{P}(X(t) > x) = 1 - e^{-x^{-\alpha}},\ x \in \mathbb{R}$ and

$$\begin{aligned}\mathbb{P}\left(\hat{X}(t) > x, X(t) > x\right) &= 1 - \mathbb{P}(\hat{X}(t) \le x) - \mathbb{P}(X(t) \le x) + \mathbb{P}(\hat{X}(t) \le x, X(t) \le x)\\ &= 1 - 2e^{-x^{-\alpha}} + e^{-l_{N+1}(1,\lambda_1^\alpha,\ldots,\lambda_N^\alpha)x^{-\alpha}},\end{aligned}$$

we have

$$\begin{aligned}2 - \theta_\lambda &= \lim_{x\to\infty} \frac{\mathbb{P}(\hat{X}(t) > x, X(t) > x)}{\mathbb{P}(X(t) > x)}\\ &= \lim_{x\to\infty} \frac{1 - 2e^{-x^{-\alpha}} + e^{-l_{N+1}(1,\lambda_1^\alpha,\ldots,\lambda_N^\alpha)x^{-\alpha}}}{1 - e^{-x^{-\alpha}}}\\ &\overset{(*)}{=} \lim_{x\to\infty} \frac{e^{-x^{-\alpha}}\left(2 - l_{N+1}(1,\lambda_1^\alpha,\ldots,\lambda_N^\alpha)e^{-(l_{N+1}(1,\lambda_1^\alpha,\ldots,\lambda_N^\alpha)-1)x^{-\alpha}}\right)}{e^{-x^{-\alpha}}}\\ &= 2 - l_{N+1}(1,\lambda_1^\alpha,\ldots,\lambda_N^\alpha)\end{aligned}$$

where the equality $(*)$ follows from l'Hopital's rule. The assertion follows immediately. □

Denote by l_N the tail-dependence function of $(X(t_1),\ldots,X(t_N))$.

Corollary 3.79 Let $\Lambda_g = \{\lambda \in \mathbb{R}_+^N : l_N(\lambda_1^\alpha),\ldots,\lambda_N^\alpha) = 1\}$, see Example 3.67. Then, the optimization problem $\theta_\lambda \to \inf\limits_{\lambda\in\Lambda_g}$ from Example 3.68 reads

$$l_{N+1}(1,x) \to \min_{x\in\mathbb{R}_+^N:\ l_{N+1}(0,x)=1}, \tag{3.81}$$

where $x = (x_1,\ldots,x_N)$ with $x_i = \lambda_i^\alpha,\ i = 1,\ldots,N$.

Proof It is clear that $l_N(x) = l_{N+1}(0,x)$ for all $x \in \mathbb{R}^d$. The constraint $l_N(x) = 1$ is equivalent to $\hat{X}(t) \overset{d}{=} X(t)$. The remaining part is trivial. □

For $\alpha > 1$ the above result can be interpreted geometrically. To do so, the so-called *D-norm* needs to be introduced.

Definition 3.80 Let $(Y_1,\ldots,Y_N)$ be a random vector with components $Y_j \ge 0$ a.s. and $\mathbb{E}[Y_j] = 1$, $j = 1,\ldots,N$. Then,

$$\|x\|_D := \mathbb{E}\left[\max_{j=1,\ldots,n} |x_j|Y_j\right], \quad x \in \mathbb{R}^N,$$

defines the so-called *D-norm* with a generator $(Y_1,\ldots,Y_N)$.

Exercise Show that $\|\cdot\|_D$ is a norm on the space $\mathbb{R}^N$.

For $Y_1,\ldots,Y_N \sim$ Fréchet(α), $\alpha > 1$, the D-norm with generator $(\mathbb{E}[Y_1])^{-1}(Y_1,\ldots,Y_N)$ is given by

$$\|x\|_D = \mathbb{E}\left[\max_{j=1,\ldots,N} x_j \frac{Y_j}{\mathbb{E}(Y_j)}\right] = \Gamma^{-1}\left(1 - \frac{1}{\alpha}\right)\mathbb{E}\left[x_1Y_1 \vee \cdots \vee x_NY_N\right].$$

Lemma 3.81 Let $(Y_1, \ldots, Y_N)$ be a max-stable random vector such that $Y_j = X(t_j)$, $j = 1, \ldots, N$, generating the above D-norm. For all $\lambda = (\lambda_1, \ldots, \lambda_N) \in \mathbb{R}^N_+$, it holds that

(a) $\|\lambda\|_D = l_N^{1/\alpha}(\lambda_1^\alpha, \ldots, \lambda_N^\alpha)$

(b) $\|.\|_D$ is convex.

(c) If $X(t_1), \ldots, X(t_N)$ are stochastically independent, then $\|\lambda\|_D = \|\lambda\|_\alpha = \left(\sum_{j=1}^N \lambda_j^\alpha\right)^{1/\alpha}$.

Proof (a) By homogeneity of l_N, it holds that

$$\mathbb{P}(\lambda_1 Y_1 \vee \cdots \vee \lambda_N Y_N) = \mathbb{P}(\lambda_1 Y_1 \le x, \ldots, \lambda_N Y_N \le x) = \exp\{-x^{-\alpha} l_n(\lambda_1^\alpha, \ldots, \lambda_N^\alpha)\},$$

i.e.

$$\max_{j=1,\ldots,N} \lambda_j Y_j \sim \text{Fréchet}(\alpha, 0, l_N(\lambda_1^\alpha, \ldots, \lambda_N^\alpha)).$$

Then,

$$\|\lambda\|_D = \Gamma\left(1 - \frac{1}{\alpha}\right)^{-1} \mathbb{E}\left[\lambda_1 Y_1 \vee \cdots \vee \lambda_N Y_N\right] = l_N^{1/\alpha}(\lambda_1^\alpha, \ldots, \lambda_N^\alpha).$$

(b) This follows from convexity of l_N, see Proposition 1.6.

(c) This follows from Exercise 1.7.

□

Remark 3.82 For $\lambda \in \mathbb{R}^N_+$, let $\tilde{\lambda} = (1, \lambda) \in \mathbb{R}^{N+1}_+$. If $\alpha > 1$, then the minimization problem (3.81) rewrites as

$$\|\tilde{\lambda}\|_D \to \min_{\lambda \in \mathbb{R}^N_+ : \|(0,\lambda)\|_D = 1},$$

which means that the D-norm on $\mathbb{R}^{N+1}$ generated by $(X(t), X(t_1), \ldots, X(t_N))$ is minimized on the positive part of N-dimensional unit ball $\|(0, \lambda)\|_D = 1$.

In this case the predictor (3.72) can be rewritten as

$$\hat{\lambda} = \operatorname*{argmin}_{\lambda \in \mathbb{R}^N_+} \left\{ 2\mathbb{E}\left[e^{-(X(t) \vee \hat{X}(t))^{-\alpha}}\right] - \mathbb{E}\left[e^{-\hat{X}(t)^{-\alpha}}\right] + \gamma\left(\frac{1}{3} - \mathbb{E}\left[e^{-\hat{X}(t)^{-\alpha}}\right] + \mathbb{E}\left[e^{-2\hat{X}(t)^{-\alpha}}\right]\right) - \frac{1}{2} \right\},$$

with $\gamma \ge 0$, where Y is an independent copy of $e^{-\hat{X}(t)^{-\alpha}}$. Taking the Wasserstein metric from Equation (3.73) into account, it follows that

$$\hat{\lambda} = \operatorname*{argmin}_{\lambda \in \mathbb{R}^N_+} \left\{ 2\mathbb{E}\left[e^{-(X(t) \vee \hat{X}(t))^{-\alpha}}\right] - \mathbb{E}\left[e^{-\hat{X}(t)^{-\alpha}}\right] - \frac{1}{2} + \gamma\left(\frac{1}{3} + \int_0^1 F_{Z_\lambda}(u)(F_Z(u) - 2u)du\right) \right\}, \tag{3.82}$$

where $Z_\lambda = e^{-(\hat{X}(t))^{-\alpha}}$. Using a substitution $x_j = \lambda_j^\alpha$, $j = 1, \ldots, N$, again, we may conclude with the following lemma.

Lemma 3.83 The minimization problem in (3.82) is equivalent to

$$\frac{2 l_{N+1}(1, x)}{l_{N+1}(1, x) + 1} - \frac{l_{N+1}(0, x)}{l_{N+1}(0, x) + 1} + \frac{\gamma}{3} \frac{(l_{N+1}(0, x) - 1)^2}{(l_{N+1}(0, x) + 1/2)(l_{N+1}(0, x) + 2)} \to \min_{x \in \mathbb{R}^N_+},$$

where $x = (x_1, \ldots, x_N)$ and l_{N+1} is the tail dependence function of $(X(t), X(t_1), \ldots, X(t_N))$.

Proof Set $t_0 = t$ and $\lambda_0 = 1$. Then, clearly it holds that

$$X(t) \vee \hat{X}(t) = \max_{j=0,\dots,N} \lambda_j X(t_j).$$

Computing the cumulative distribution function of Z_λ yields

$$F_{Z_\lambda}(u) = \mathbb{P}\left(e^{-X(t)^{-\alpha}} \leq u\right) = \mathbb{P}\left(\hat{X}(t) \leq (-\log u)^{-1/\alpha}\right) = u^{l_N(\lambda_1^\alpha,\dots,\lambda_N^\alpha)}, \quad u \in [0,1].$$

Similarly, we have

$$\mathbb{P}\left(e^{-(X(t)\vee\hat{X}(t))^{-\alpha}} \leq u\right) = u^{l_{N+1}(1,\lambda_1^\alpha,\dots,\lambda_N^\alpha)}, \quad u \in [0,1].$$

As a consequence

$$\mathbb{E}\left[e^{-(X(t)\vee\hat{X}(t))^{-\alpha}}\right] = \int\limits_0^1 \left(1 - u^{l_{N+1}(1,\overbrace{x_1,\dots,x_N}^{=x})}\right) du = \frac{l_{N+1}(1,x)}{l_{N+1}(1,x)+1},$$

and similarly

$$\mathbb{E}\left[e^{-\hat{X}_\lambda^\alpha}\right] = \frac{l_{N+1}(0,x)}{l_{N+1}(0,x)+1}.$$

The squared W_2-distance rewrites as

$$\begin{aligned} W_2^2(F_{Z_\lambda}, F_u) &= \frac{1}{3} + \int\limits_0^1 \left(u^{2l_{N+1}(0,x)} - 2u^{l_{N+1}(0,x)+1}\right) du \\ &\overset{(*)}{=} \frac{1}{3} + \frac{1}{2l_{N+1}(0,x)+1} - \frac{2}{l_{N+1}(0,x)+2} \\ &= \frac{1}{3}\frac{(l_{N+1}(0,x)-1)^2}{(l_{N+1}(0,x)+1/2)(l_{N+1}(0,x)+2)}, \end{aligned}$$

where the equality $(*)$ follows from $U \sim U([0,1])$. □

Let us show the existence of the forecast $\hat{\lambda}$ from Equation (3.82).

Theorem 3.84 The minimization problem (3.82) has a solution.

Proof The random vector $(X(t_0), X(t_1), \dots, X(t_N))$ is max-stable with Fréchet(α)-marginals, and hence absolutely continuously distributed. Since $\lim_{x\to 0} e^{-x^{-\alpha}} = 0$ and $\lim_{\|\lambda\|_2\to 0} \hat{X}(t) = 0$ a.s., there exists a $\lambda_0 \in \mathbb{R}_+^N$ such that

$$\begin{aligned} \Psi(\lambda) &= 2\mathbb{E}\left[e^{-(X(t)\vee\hat{X}(t))^{-\alpha}}\right] - \mathbb{E}\left[e^{-\hat{X}(t)^{-\alpha}}\right] + \gamma\left(\frac{1}{3} + E\left[e^{-(2\hat{X}(t))^{-\alpha}}\right] - \mathbb{E}\left[e^{-(\hat{X}(t))^{-\alpha}} \vee Y\right]\right) \\ &< 1 + \frac{\gamma}{3} \end{aligned}$$

holds.

Together with the a.s. divergence $\lim_{\|\lambda\|_2\to\infty} \hat{X}(t) = \infty$ a.s. and Lemma 3.73, the existence of $M > 0$ follows, i.e. $\min_{\lambda\in\mathbb{R}_+^N} \Psi(\lambda) = \min_{\lambda\in[0,M]^N} \Psi(\lambda)$, and thus condition (I) of Theorem 3.70 is satisfied.

Since

$$\hat{X}(t) \sim \text{Fréchet}(\alpha, 0, \underbrace{l_N^{1/\alpha}(\lambda_1^{1/\alpha}, \dots, \lambda_N^{1/\alpha})}_{=\sigma(\lambda,\alpha)}),$$

its density is given by

$$f_{\hat{X}(t)}(x) = \frac{\alpha\sigma^\alpha(\lambda,\alpha)}{x^{\alpha+1}} \exp\left\{-\left(\frac{\sigma(\lambda,\alpha)}{x}\right)^\alpha\right\}.$$

Note that $f_{\hat{X}(t)}$ is bounded from above uniformly in $\lambda \in [0, M]^N$.

By the dominated convergence theorem, we get (denoting $\hat{X}_\lambda := \hat{X}(t)$)

$$\lim_{\|h\|_2 \to 0} \left\| f_{\hat{X}_\lambda} - f_{\hat{X}_{\lambda+h}} \right\|_{L_1} = \lim_{\|h\|_2 \to 0} \int\limits_{\mathbb{R}} \left| f_{\hat{X}_\lambda} - f_{\hat{X}_{\lambda+h}} \right| dx = \int\limits_{\mathbb{R}} \lim_{\|h\|_2 \to 0} \left| f_{\hat{X}_\lambda} - f_{\hat{X}_{\lambda+h}} \right| = 0.$$

Since the tail dependence function $l_N(\cdot)$ is convex, it is also continuous, and consequently $f_{\hat{X}_\lambda}$ is continuous in λ. Thus, condition (III) of Theorem 3.70 is satisfied.

Since the copula $C_{X(t),\hat{X}(t)}(x,x) = x^{\theta_\lambda}$, where $\theta_\lambda = l_2(1,1)$ is the pairwise extremal coefficient of $(X(t), \hat{X}(t))$, θ_λ is continuous on $[0, M]^N$ because of the continuity of l_2. Hence, $\int\limits_0^1 C_{X(t),\hat{X}(t)}(x,x)dx = \frac{1}{\theta_\lambda + 1}$ is continuous as well, and the application of Theorem 3.70 resumes this proof. □

Now let us discretize the expectations in $\Psi(\lambda)$ and write the functional $\bar{\phi}(\lambda) = \frac{1}{n}\sum\limits_{j=1}^n Q_j(\lambda)$ with

$$\begin{aligned} Q_j(\lambda) := {} & 2\exp\{-(X(t+s_j) \vee g(\lambda, X_{T_f+s_j}))^{-\alpha}\} - \exp\{-g^{-\alpha}(\lambda, X_{T_f+s_j})\} \\ & + \gamma\left(\frac{1}{3} - \exp\{-g^{-\alpha}(\lambda, X_{T_f+s_j})\} \vee Y_j + \exp\{-2g^{-\alpha}(\lambda, X_{T_f+s_j})\}\right), \quad j = 1, \dots, n, \end{aligned}$$

where Y_j is an independent copy of $\exp\{-g^{-\alpha}(\lambda, X_{T_f+s_j})\}$. To implement the stochastic subgradient descent, we also need the subgradients

$$\nabla^* Q_j(\lambda) = \left(\nabla q_j^{(l)}(\lambda) - \nabla p_j^{(l)}(\lambda),\ l = 1, \dots, N\right)^\intercal,$$

where

$$\begin{aligned} \nabla q_j^{(l)}(\lambda) = \Bigg(& 2 \cdot \mathbf{1}\left(X(t+s_j) < \lambda_l X(t_l+s_j)\right) - 1 \\ & - \gamma \cdot \mathbf{1}\left(\max_{i=1,\dots,N} \lambda_i \tilde{X}(t_i+s_j) < \max_{i=1,\dots,N} \lambda_i X(t_i+s_j)\right) \\ & + 2\gamma \exp\{-(\lambda_l \tilde{X}(t_l+s_j))\}\Bigg) \\ & \times f(\lambda_l \tilde{X}(t_l+s_j)) \cdot X(t_l+s_j) \cdot \mathbf{1}\left(\lambda_l \tilde{X}(t_l+s_j) = \max_{i=1,\dots,N} \lambda_i X(t_i+s_j)\right) \end{aligned}$$

and

$$\begin{aligned}\nabla p_j^{(l)}(\lambda) = \gamma \cdot \mathbf{1}\left(\max_{i=1,\ldots,N} \lambda_i \tilde{X}(t_i+s_j) > \max_{i=1,\ldots,N} \lambda_i X(t_i+s_j)\right)\\ \times f(\lambda_l \tilde{X}(t_l+s_j)) \cdot \mathbf{1}\left(\lambda_l \tilde{X}(t_l+s_j) = \max_{i=1,\ldots,N} \lambda_i \tilde{X}(t_i+s_j)\right)\end{aligned}$$

for $j = 1, \ldots, n$, $l = 1, \ldots, N$, where $(\tilde{X}(t_1+s_j), \ldots, \tilde{X}(t_N+s_j))$ is an independent copy of $(X(t_1+s_j), \ldots, X(t_N+s_j))$ for all $j = 1, \ldots, n$ and

$$f(x) = \alpha x^{-1-\alpha} e^{-x^{-\alpha}}, \quad x > 0,$$

is the probability density function of the Fréchet(α) distribution.

3.6 Conditional simulation of stationary Gaussian random fields

Let $X = \{X(t), t \in T\}$, $T \subset \mathbb{R}^d$, be a stationary Gaussian random field with mean zero, $\mathbf{var}(X(t)) \equiv 1$ and covariance function $C(t) = \mathbf{cov}(X(s), X(t))$. Our goal is to simulate $X(t)$ at locations $t \notin \{t_1, \ldots, t_N\} \subset W$, where W is a compact observation window, provided that observations $X(t_1) = y_1, \ldots, X(t_N) = y_N$ are given.

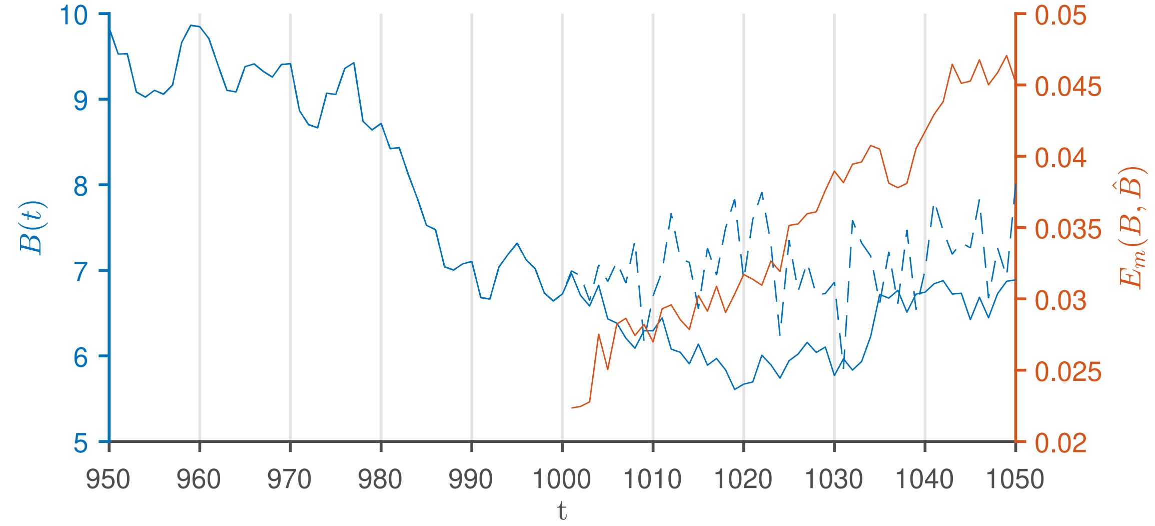
Let $\hat{X} = \{\hat{X}(t), t \in W\}$ be the simple Kriging prediction of X in W, compare Section 3.2.1. By theorem 3.16 (iv), the random fields $\hat{X}$ and $X - \hat{X}$ are stochastically independent. Indeed, in the Gaussian case the orthogonality of $\hat{X}(t)$ and $X(t) - \hat{X}(t)$ means independence, for any $t \in W$, since $\hat{X}(t)$ and $X(t) - \hat{X}(t)$ are jointly Gaussian as linear combinations of observations $X(t_j)$, $j = 1, \ldots, N$. The same reasoning can be applied to arbitrary linear combinations of values of $\hat{X}(t)$ and $X(t) - \hat{X}(t)$, $t \in W$.

Therefore, it holds that $X(t) = (X(t) - \hat{X}(t)) + \hat{X}(t)$, where both components are independent and Gaussian. This leads to the following algorithm to simulate $X(t)$ conditional on $X(t_j) = y_j$, $j = 1, \ldots, N$:

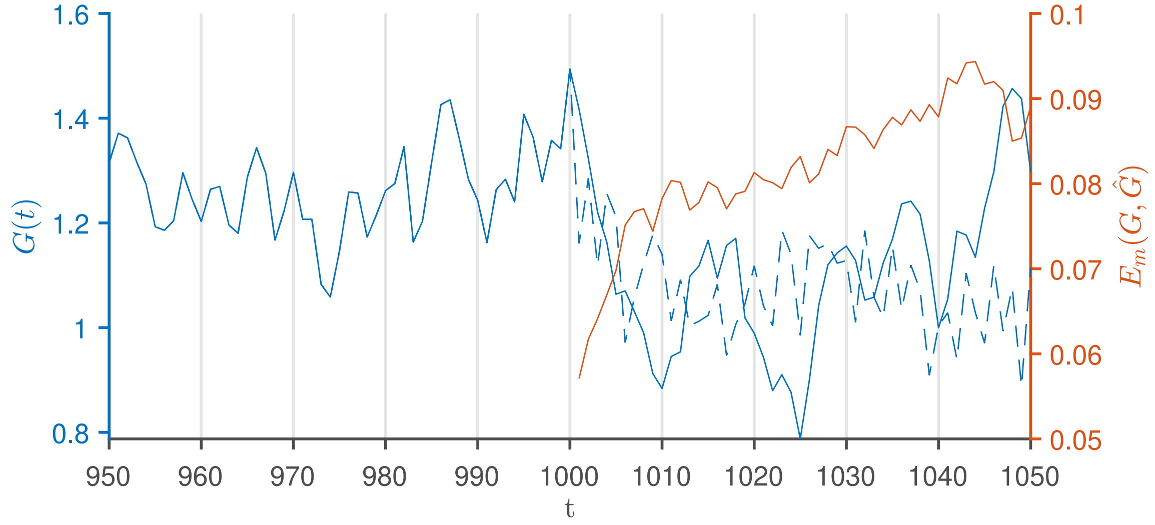
1. Simulate a Gaussian random field with mean 0 and covariance function $C(\cdot)$ at locations $t \in W$. Denote the resulting field by $\{X^*(t), t \in W\}$.
2. Compute the simple Kriging estimates of $\{X^*(t),\ t \in W\}$ on the sample $\{X^*(t_j), j = 1, \ldots, N\}$. Denote the resulting field by $\left\{\hat{X}^*(t) = \sum_{j=1}^N \lambda_j^* X^*(t_j),\ t \in W\right\}$.
3. Return $\tilde{X}(t) = \hat{X}(t) + X^*(t) - \hat{X}^*(t),\ t \in W$.

The field $\{\tilde{X}(t), t \in W\}$ is a conditional simulation of X provided that $X(t_j) = y_j$. Indeed, since simple Kriging is exact, it holds that $\hat{X}^*(t_j) = X^*(t)$ and $\hat{X}(t_j) = X(t_j) = y_j$ for all $j = 1, \ldots, N$. Moreover, $\tilde{X}$ is obviously Gaussian as a sum of two independent Gaussian components $\hat{X}$ and $X^* - \hat{X}^*$.

Remark 3.85 For points t lying far away from $t_1, \ldots, t_N$ one can expect $\tilde{X}(t) \approx X^*(t)$, i.e., the simulation becomes uncondiditonal. Indeed, if $C(t) \to 0$ as $\|t\|_2 \to \infty$, the simple Kriging weights λ_j in $\hat{X}(t) = \sum_{j=1}^N \lambda_j Y_j$ and λ_j^* of $\hat{X}^*(t) = \sum_{j=1}^N \lambda_j^* X^*(t_j)$ are very small, which yields $\hat{X}(t) \approx 0$ and $\hat{X}^*(t) \approx 0$.



(a) A 50 time steps' forecast (dashed blue line), together with its corresponding excursion metric (red line), of a Brown-Resnick process (blue line). After observing 110 values of B, the predictor $\hat{B}$ used ten learning samples of size eleven. Step-sizes for the stochastic gradient descent were given by the harmonic series $\{1/n\}_{n\in\mathbb{N}}$. The underlying process Y of the Brown-Resnick process was a standard Brownian motion.



(b) A 50 time steps' forecast (dashed blue line), together with its corresponding excursion metric (red line), of an extremal Gaussian process (blue line). After observing 110 values of G, the predictor $\hat{G}$ used ten learning samples of size eleven. Step-sizes for the stochastic gradient descent were given by the harmonic series $\{1/n\}_{n\in\mathbb{N}}$. The underlying process Y of G was a Gaussian process with Cauchy covariance function $C(t) = \exp(-|t|0.01)$.

Fig. 3.14: Forecast of a Brown-Resnick and extremal Gaussian process

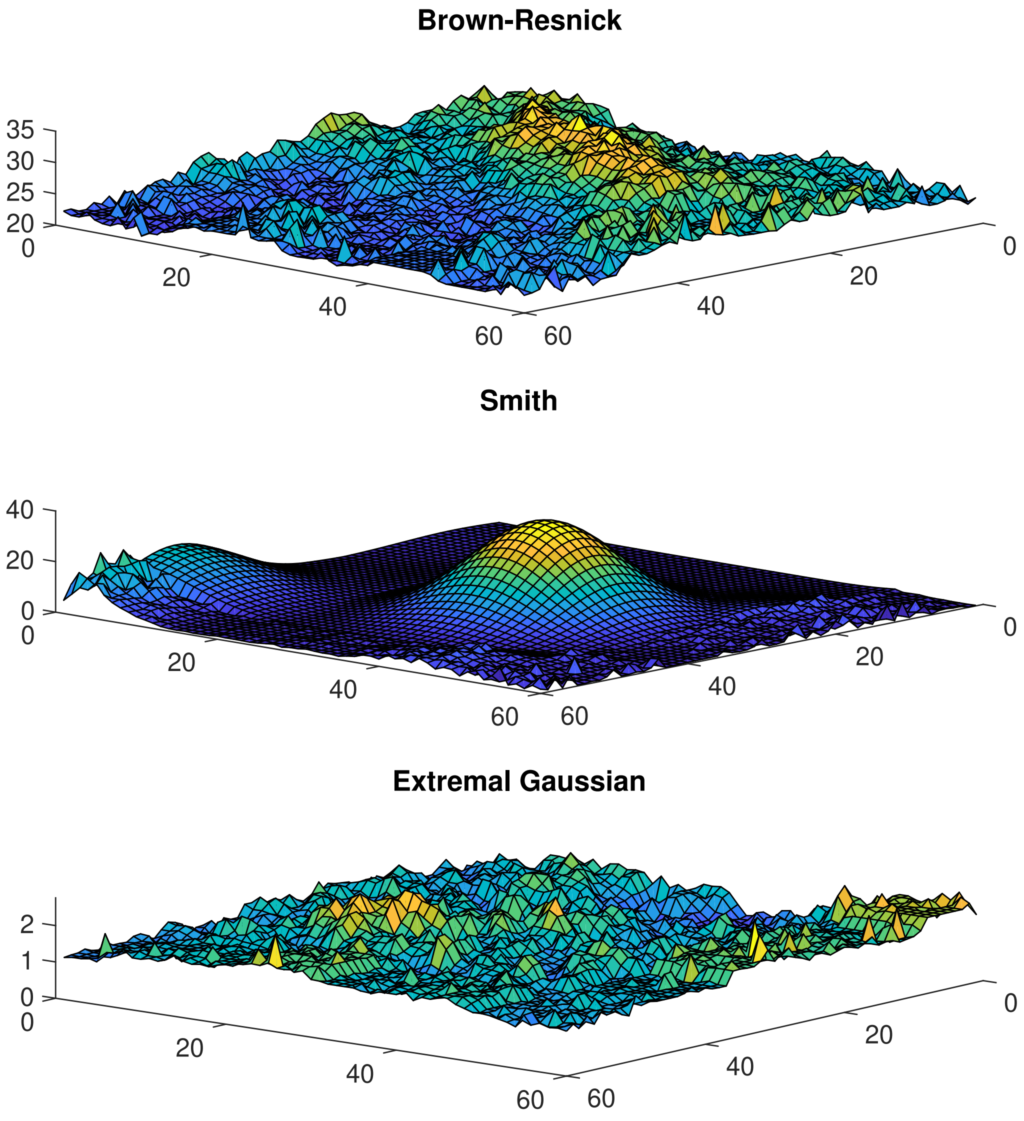
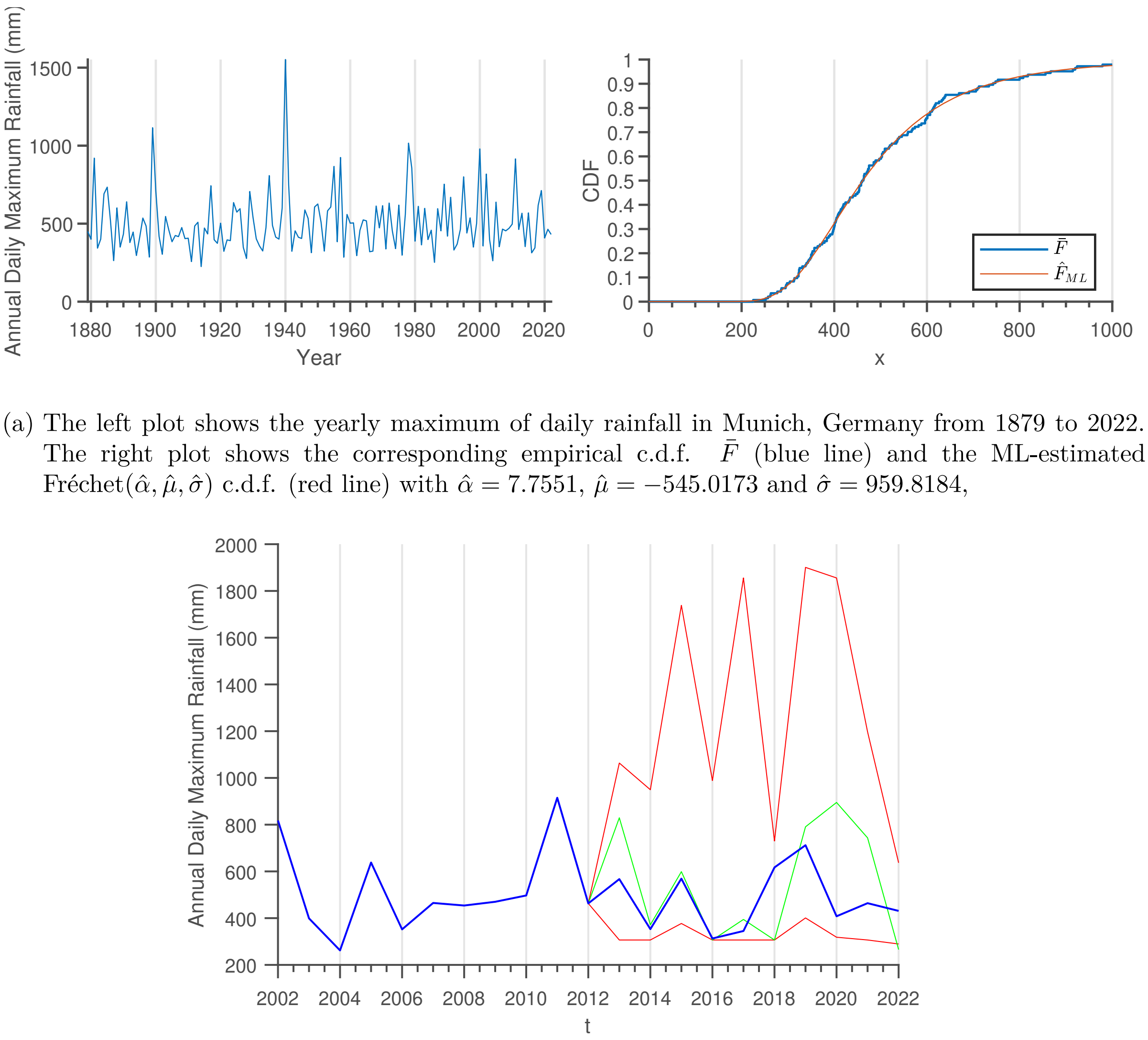


Fig. 3.15: A ten steps' forecast of random fields of each type in both directions t_1 and t_2. After observing true values of the random fields at locations $t \in \{1, \ldots, 50\} \times \{1, \ldots, 50\}$, the predictor extended the random surfaces to $t \in \{1, \ldots, 60\} \times \{1, \ldots, 60\}$.
Step-sizes for the stochastic gradient descent were given by the harmonic series $\{1/n\}_{n \in \mathbb{N}}$.



(a) The left plot shows the yearly maximum of daily rainfall in Munich, Germany from 1879 to 2022. The right plot shows the corresponding empirical c.d.f. $\bar{F}$ (blue line) and the ML-estimated Fréchet$(\hat{\alpha}, \hat{\mu}, \hat{\sigma})$ c.d.f. (red line) with $\hat{\alpha} = 7.7551$, $\hat{\mu} = -545.0173$ and $\hat{\sigma} = 959.8184$,

(b) Forecasts for the annual daily rainfall maxima from 2013 to 2022. All data from 1883-2012 was used in learning samples. The real data is shown by the blue line. The red lines mark the maximum and minimum of 100 forecasts using the max-stable predictor with bootstrap. The green line yields the forecast using the alternative formulation. For every extrapolation 12 learning samples of size 10 containing data from 1883-2002 and a forecast sample containing data from 2003-2012 were used.

Fig. 3.16: Forecast of Munich daily maximums of rainfall

Bibliography

[1] S Bandyopadhyay and S. Lahiri. Asymptotic properties of discrete fourier transforms for spatial data. *Sankhya*, 71:221 – 259, 2009.

[2] S Bandyopadhyay, S.N. Lahiri, and D.J. Nordman. A frequency domain empirical likelihood method for irregularly spaced spatial data. *Ann. Statist.*, 43:519 – 545, 2015.

[3] A. V. Bulinski and A. P. Shashkin. *Limit theorems for associated random fields and related systems.* World Scientific, 2007.

[4] J. P. Chilès and P. Delfiner. *Geostatistics: Modeling Spatial Uncertainty.* Wiley, 2 edition, 2012.

[5] Y.-S. Chow and U. Grenander. A sieve method for the spectral density. *Ann. Statist.*, 13:998–1010, 1985.

[6] R. Dahlhaus and H. Künsch. Edge effects and efficient parameter estimation for stationary random fields. *Biometrika*, 74:877 – 882, 1987.

[7] L. De Haan and A. Ferreira. *Extreme value theory: an introduction.* Springer-Verlag GmbH, New York, 2006.

[8] S. Deb, M. Pourahmadi, and W. B. Wu. An asymptotic theory for spectral analysis of random fields. *Electron. J. Statist.*, 11:4297 –4322, 2017.

[9] T. Donhauser, V. Makogin, and E. Spodarev. *Extrapolation of max-stable random fields with Fréchet margins.* Preprint, 2023.

[10] P. Doukhan. *Mixing: properties and examples.* Springer-Verlag GmbH, 1994.

[11] M. Fuentes. Spectral methods for nonstationary spatial processes. *Biometrika*, 89:197–210, 2002.

[12] M. Fuentes. Approximate likelihood for large irregular spaced spatial data. *J. Am. Statist. Assoc.*, 102:321 – 331, 2007.

[13] A. Genz and F. Bretz. *Computation of multivariate normal and t probabilities.* LNS. Springer-Verlag GmbH, Berlin, 2009.

[14] T. Gneiting and A. E. Raftery. Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Assoc.*, 102:359 – 378, 2007.

[15] J. Guiness. Spectral density estimation for random fields via periodic embeddings. *Biometrika*, 106(2):p.267–286, 2019.

[16] J. Guiness and M. Fuentes. Circulant embedding of approximate covariances for inference from gaussian data on large lattices. *J. Comp. Graph. Statist.*, 26:88 – 97, 2017.

[17] J. Guinness. Permutation and grouping methods for sharpening gaussian process approximations. *Techno-metrics*, 60:415 – 429, 2018.

[18] X. Guyon. Parameter estimation for a stationary process on a d-dimensional lattice. *Biometrika*, 69:95 – 105, 1982.

[19] C. C. Heyde and R. Gay. Smoothed periodogram asymptotics and estimation for processes and fields with possible long-range dependence. *Stochastic Processes and their Applications*, 45(1):169–182, March 1993.

[20] A.V. Ivanov and N.N. Leonenko. *Statistical Analysis of Random Fields.* Kluwer Academic Publishers, 1989.

[21] A. N. Kolmogorow. On the problem of the goodness of empirical statistical forecast formulae. *J. Geophys.*, 3:p.78–82, 1933.

[22] C. Lantuéjoul. *Geostatistical Simulation: Models and Algorithms.* Springer-Verlag GmbH, 2002.

[23] T. C. Lee. A simple span selector for periodofram smoothing. *Biometrika*, 84:965 – 969, 1997.

[24] T. C. Lee. A stabilized bandwidth selection method for kernel smooting of the periodogram. *Sig. Proces.*, 81:419 – 430, 2001.

[25] T. C. Lee and Z. Zhu. Nonparametric spectral density estimation with missing observations. *International Conference of Acoustics, Speech and Signal Processing*, 2009.

[26] C. Y. Lim and M. Stein. Properties of spatial cross-periodograms using fixed-domain asymptotics. *J. Mult. Anal.*, 99:1962–1984, 2008.

[27] R. J. Little and D. B. Rubin. *Statistical Analysis with Missing Data.* John Wiley & Sons, Hoboken, New Jersey, 2014.

[28] Y. Matsuda and Y. Yajima. Fourier analysis of irregularly spaced data on r. *J. R. Statist. Soc. B*, 71:191–217, 2009.

[29] G. Miller. Properties of certain symmetric stable distributions. *Journal of multivariate analysis*, 8:p.346–360, 1978.

[30] D. Nychka, S. Bandyopadhyay, D. Hammerling, F. Lidgren, and S. Sain. A multiresolution gaussian process model for the analysis of large spatial datasets. *J. Comp. Graph. Statist.*, 24:579–599, 2015.

[31] H. C. Ombao, J. A. Raz, R. L. Strawderman, and R. Von Sachs. A simple generalised crossvalidation method. *Biometrika*, 88:1186–1192, 2001.

[32] Y. Pawitan and F. O'Sullivan. Nonparametric spectral density estimation using penelized whittle likelihood. *J. Am. Statist. Assoc.*, 69:600–610, 1994.

[33] Dimitris N. Politis and Joseph P. Romano. Bias-corrected nonparametric spectral estimation. *Journal of Time Series Analysis*, 16(1):67–103, 1995.

[34] S. Resnick. *Extreme values, Regular variation and point processes.* Springer- Verlag GmbH, 1987.

[35] H. Robbins and S. Monroe. A stochastic spproximation method. *Ann. Math. Stat.*, 22:400 – 407, 1951.

[36] M. Rosenblatt. *Stationary sequences and random fields.* Birkhäuser Boston, 1985.

[37] G. Samorodnitsky and M.S. Taqqu. *Stable non-Gaussian random processes.* Chapmann & Hall, 1994.

[38] V. Schmidt. *Stochastic geometry, spatial statistics and random fields. Models and algorithms*, volume 2120 of *Lecture Notes of Mathematics.* Springer-Verlag GmbH, 2015.

[39] A. N. Shiryayev. *Interpolation and Extrapolation of Stationary Random Sequences*, pages 272–280. Springer Netherlands, Dordrecht, 1992.

[40] E. Spodarev. *Stochastic geometry, spatial statistics and random fields. Asymptotic methods*, volume 2068 of *Lecture Notes of Mathematics.* Springer-Verlag GmbH, 2013.

[41] E. Spodarev. *Wahrscheinlichkeitstheorie und stochastische Prozesse.* Lecture Notes. Ulm University, 2020.

[42] E. Spodarev. *Random fields.* Lecture Notes. Ulm University, 2022.

[43] Evgeny Spodarev. *Random Fields.* 2022.

[44] M. L. Stein. Fixed-domain asymptotics for spatial periodograms. *J. Am. Statist. Assoc.*, 90:1277–1288, 1995.

[45] M. L. Stein. *Interpolation of Spatial Data: Some Theory for Kriging.* Springer-Verlag GmbH, New York, 1999.

[46] J. R. Stroud, M. L. Stein, and S. Lysen. Bayesian and maximum likelihood estimation for gaussian processes on an incomplete lattice. *J. Comp. Graph. Statist.*, 26:108–120, 2017.

[47] S. Subba Rao. Statistical inference for spatial statistics defined in the fourier domain. *Ann. Statist.*, 46:466–469, 2018.

[48] A. M. Sykulski, S. C. Olhede, A. P. Guillaumin, J. M. Lilly, and Early; J. J. The debiased whittle likelihood. *Biometrika*, 106:251–266, 2019.

[49] M. A. Tanner and W. H. Wong. The calculation of posterior distributions by data augmentation. *J. Am. Statist.*, 82:528–540, 1987.

[50] A. V. Vecchia. Estimation and model identification for continuous spatial processes. *J. R. Statist. Soc. B*, 50:297–312, 1988.

[51] H. Wackernagel. *Multivariate Geostatistics. An Introduction with Applications.* Springer-Verlag GmbH, 3 edition, 2003.

[52] G. Wahba. Automatic smoothing of the log periodogram. *J. Am. Statist. Assoc.*, 75:122–132, 1980.

[53] P. Whittle. On stationary processes in the plane. *Biometrika*, 41:434–449, 1954.

[54] Y. Zheng, J. Zhu, and A Roy. Nonparametric bayesian inference for the spectral density function of a random field. *Biometrika*, 97:238–245, 2009.

Index