

MATH 590: Meshfree Methods

Chapter 4: The Connection to Kriging

Greg Fasshauer

Department of Applied Mathematics
Illinois Institute of Technology

Fall 2014



Outline

- 1 Brief Introduction to Kriging
- 2 Random Fields and Random Variables
- 3 RKHSs vs. Spaces Generated by Zero-Mean GRFs
- 4 Modeling and Prediction via Kriging
- 5 Karhunen–Loève Expansions and Polynomial Chaos



Kriging

The kriging approach comes from [geostatistics](#) and is named after the South African mining engineer Danie Krige, who used this method in his work creating gold ore distributions from a collection of ore samples [Kri51].

The method was given a solid mathematical foundation and seen to be an [optimal linear prediction method](#) by the French mathematician and geostatistician Georges Matheron [Mat65].

Part of our exposition follows the paper [SWMW89] in which the kriging method was introduced into the [design of experiments](#) context.



Kriging (cont.)

Main task:

- Use the values y_1, \dots, y_N sampled at locations $\mathbf{x}_1, \dots, \mathbf{x}_N$
- to predict the unknown value $y(\mathbf{x})$ at a location \mathbf{x} (which is not among the sampling locations).

This looks just like the scattered data interpolation/approximation problem.

But now we take the stochastic point of view, i.e., we assume that the data are observed as realizations $y_{\mathbf{x}_i}$, $i = 1, \dots, N$, of the random variables $Y_{\mathbf{x}_i}$ belonging to a random field Y .

Thus the prediction $\hat{y}_{\mathbf{x}}$ will also be a realization of a random variable.



Remark

The *data might contain a random measurement error* (such as the gold ore samples in the original application of Krige), *or it might not* (such as in the computer experiments of [SWMW89]).

Even though we use a stochastic framework, the method can be applied to deterministic problems.



Random Fields and Random Variables

The definition of a random field/stochastic process on the following slide requires the notion of a

probability space $(\mathcal{W}, \mathcal{A}, P)$,

where

\mathcal{W} : sample space containing all possible outcomes,

\mathcal{A} : σ -algebra containing the collection of all events, i.e., a set of subsets of \mathcal{W} ,

P : probability measure.

Remark

In the probability literature the sample space is usually denoted by Ω , but we already use Ω as the domain for the spatial variables (the “parameter space” in probability, see next slide).

Definition (Random field)

Given a probability space $(\mathcal{W}, \mathcal{A}, P)$ and an underlying parameter space Ω , a function $Y : \Omega \times \mathcal{W} \rightarrow \mathbb{R}$, i.e., $(\mathbf{x}, \omega) \mapsto Y(\mathbf{x}, \omega)$, is called a **random field** if, for every fixed $\mathbf{x} \in \Omega$, Y is an \mathcal{A} -measurable function of $\omega \in \mathcal{W}$.

Remark

- *Stochastic process (instead of random field) is common terminology — especially when the parameter space is viewed as “time” or “space-time”.*
- *In the statistics literature the parameter space is often denoted by T , and the random variables by X_t .*
- *We use Ω and $Y_{\mathbf{x}}$ since this agrees better with our usual (numerical analysis) notation.*



Notation:

random field: $Y = \{Y_{\mathbf{x}}\}_{\mathbf{x} \in \Omega}$

random variable: $Y_{\mathbf{x}} = Y(\mathbf{x}, \cdot)$, where $\mathbf{x} \in \Omega$ is fixed. A random variable is a **function** of the random argument $\omega \in \mathcal{W}$.

sample path: $y(\cdot) = Y(\cdot, \omega)$, where $\omega \in \mathcal{W}$ is fixed. A sample path is a **deterministic function** of $\mathbf{x} \in \Omega$, also called a **realization** of the random field.

Remark

- *A random field is just a collection of random variables.*
- *A random field can also be viewed as a distribution over sample paths.*
- *It is common to omit the dependence on ω from the notation used for the random variable $Y_{\mathbf{x}}$. Unfortunately, random fields are often denoted using the same notation. This makes for confusing notation.*

Remark

- We can also look at a *realization of a random variable*, $y_{\mathbf{x}}$:
 - start with a random field Y
 - fix \mathbf{x} to get the random variable $Y_{\mathbf{x}}$
 - then fix ω .
- Alternatively, we can *evaluate a sample path*, $y(\mathbf{x})$:
 - start with a random field Y
 - fix ω to obtain a deterministic function y
 - then fix \mathbf{x} .

The numbers $y_{\mathbf{x}}$ and $y(\mathbf{x})$ are identical (provided the same \mathbf{x} and ω were used).



Mean and Variance

The moments of a random field Y provide useful information.

The first moment of a random field Y , called **expectation** or **mean**, is a function given by

$$\mu(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}] = \int_{\mathcal{W}} Y_{\mathbf{x}}(\omega) dP(\omega) = \int_{-\infty}^{\infty} y dF_{Y_{\mathbf{x}}}(y),$$

where $F_{Y_{\mathbf{x}}}$ is the **cumulative distribution function (CDF)** of the random variable $Y_{\mathbf{x}}$ with respect to the probability measure P .

If $F_{Y_{\mathbf{x}}}$ has a **density** $p_{Y_{\mathbf{x}}}$ such that $F_{Y_{\mathbf{x}}}(y) = \int_{-\infty}^y p_{Y_{\mathbf{x}}}(z) dz$ then

$$\mu(\mathbf{x}) = \int_{-\infty}^{\infty} yp_{Y_{\mathbf{x}}}(y) dy.$$

The second moment is called the **variance** and is given by

$$\sigma^2(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}^2] - \mathbb{E}[Y_{\mathbf{x}}]^2 = \mathbb{E}[Y_{\mathbf{x}}^2] - \mu^2(\mathbf{x}).$$



Covariance Kernel

More generally, the **covariance kernel** K of Y is defined via

$$\begin{aligned}
 K(\mathbf{x}, \mathbf{z}) &= \text{Cov}(Y_{\mathbf{x}}, Y_{\mathbf{z}}) = \mathbb{E}[(Y_{\mathbf{x}} - \mu(\mathbf{x}))(Y_{\mathbf{z}} - \mu(\mathbf{z}))] \\
 &= \mathbb{E}[(Y_{\mathbf{x}} - \mathbb{E}[Y_{\mathbf{x}}])(Y_{\mathbf{z}} - \mathbb{E}[Y_{\mathbf{z}}])] \\
 &= \mathbb{E}[Y_{\mathbf{x}}Y_{\mathbf{z}} - Y_{\mathbf{x}}\mathbb{E}[Y_{\mathbf{z}}] - \mathbb{E}[Y_{\mathbf{x}}]Y_{\mathbf{z}} + \mathbb{E}[Y_{\mathbf{x}}]\mathbb{E}[Y_{\mathbf{z}}]] \\
 &= \mathbb{E}[Y_{\mathbf{x}}Y_{\mathbf{z}}] - \mathbb{E}[Y_{\mathbf{x}}]\mathbb{E}[Y_{\mathbf{z}}] - \mathbb{E}[Y_{\mathbf{x}}]\mathbb{E}[Y_{\mathbf{z}}] + \mathbb{E}[Y_{\mathbf{x}}]\mathbb{E}[Y_{\mathbf{z}}] \\
 &= \mathbb{E}[Y_{\mathbf{x}}Y_{\mathbf{z}}] - \mathbb{E}[Y_{\mathbf{x}}]\mathbb{E}[Y_{\mathbf{z}}] = \mathbb{E}[Y_{\mathbf{x}}Y_{\mathbf{z}}] - \mu(\mathbf{x})\mu(\mathbf{z}).
 \end{aligned}$$

Therefore, the variance

$$\sigma^2(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}^2] - \mu^2(\mathbf{x})$$

corresponds to the “diagonal” of the covariance, i.e.,

$$\sigma^2(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}).$$



Gaussian Random Fields

There are many different kinds of stochastic processes. We focus on Gaussian random fields (or Gaussian processes) since they model many natural phenomena and are relatively easy to work with. In particular, a Gaussian random field is completely characterized by its first two moments.

Definition

The random field $Y = \{Y_{\mathbf{x}}\}_{\mathbf{x} \in \Omega}$ is called a Gaussian random field if, for any given choice of finitely many distinct points $\{\mathbf{x}_i\}_{i=1}^N \subset \Omega$, the vector of random variables $\mathbf{Y} = (Y_{\mathbf{x}_1}, \dots, Y_{\mathbf{x}_N})^T$ has a multivariate normal distribution with mean vector $\boldsymbol{\mu} = \mathbb{E}[\mathbf{Y}]$ and covariance matrix $\mathbf{K} = (\text{Cov}(Y_{\mathbf{x}_i}, Y_{\mathbf{x}_j}))_{i,j=1}^N$.

Notation:

$\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{K})$: \mathbf{Y} is a vector of Gaussian random variables

$Y \sim \mathcal{N}(\boldsymbol{\mu}, K)$: Y is a Gaussian random field



The **multivariate normal distribution** has the density function

$$p_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{\sqrt{(2\pi)^N \det \mathbf{K}}} \exp \left(-\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right),$$

where $\boldsymbol{\mu}$ and \mathbf{K} are defined as above.

As long as K is a strictly positive definite kernel, \mathbf{K} will be a positive definite matrix, and \mathbf{K}^{-1} will exist.

Remark

The quadratic form $(\mathbf{y} - \boldsymbol{\mu})^T \mathbf{K}^{-1} (\mathbf{y} - \boldsymbol{\mu})$ is called (the square of) the **Mahalanobis distance**. If $\boldsymbol{\mu} = \mathbf{0}$, then the kernel-based interpolant^a s is defined by the linear system $\mathbf{K}\mathbf{c} = \mathbf{y}$ and $\mathbf{y}^T \mathbf{K}^{-1} \mathbf{y} = \mathbf{c}^T \mathbf{K} \mathbf{c}$ (cf. $\|s\|_{\mathcal{H}_K}^2$ computed in Chapter 2, Part 3).

^aThis also holds for realizations of the kriging predictor (see later).

\mathcal{H}_K vs. \mathcal{H}_Y

We now have **two different ways to view scattered data**:

- 1 values $y_i = f(\mathbf{x}_i)$ of a **deterministic function** $f \in \mathcal{H}_K$, or
- 2 values $y_i = y(\mathbf{x}_i)$ of a **sample path** $y(\cdot)$ of a zero-mean random field Y .

This gives rise to a **duality** that goes back to Parzen's work [Par61, Par70].

The restriction to zero-mean random fields is not necessary but it simplifies much of the following discussion.



Remark


To understand (and differentiate between) these two view points we need to note that (see, e.g., [Wah90, Chapter 1])

- $f \in \mathcal{H}_K$ is generally a **smooth function**,
- a sample path $y(\cdot)$ of a stochastic process Y is in general **not smooth**.

So these are indeed **different** view points.

Example (Brownian motion)

- $\mathcal{H}_K(\Omega) = H^1(0, 1)$, the standard Sobolev space of **functions whose first derivative is square integrable**, i.e., f is differentiable except on a set of measure zero.
- A typical sample path is **nowhere differentiable**.

The **process** Y (as opposed to its sample paths) **does have certain smoothness properties** that are tied to the smoothness of the kernel  (see [BTA04] for more details).

In Chapter 2 we saw that the RKHS \mathcal{H}_K is the set of all linear combinations of “translates” of K together with their native space norm limits.

Now we define a Hilbert space \mathcal{H}_Y as the set of all linear combinations of random variables $Y_{\mathbf{x}}$ of the zero-mean random field $Y = \{Y_{\mathbf{x}}\}_{\mathbf{x} \in \Omega}$ together with their $L_2(\mathcal{W}, \mathcal{A}, P)$ -limits.



Loève's Representation Theorem

In spite of the differences mentioned above, Loève's representation theorem provides an **isometry between the RKHS \mathcal{H}_K and the Hilbert space \mathcal{H}_Y** generated by the zero-mean Gaussian random field Y (see, e.g., [BTA04, Chapter 2]).

It follows from this theorem that the **values of the two corresponding inner products are identical and coupled by K** :

$$\langle Y_{\mathbf{x}}, Y_{\mathbf{z}} \rangle_{\mathcal{H}_Y} = \mathbb{E}[Y_{\mathbf{x}} Y_{\mathbf{z}}] = K(\mathbf{x}, \mathbf{z}) = \langle K(\cdot, \mathbf{x}), K(\cdot, \mathbf{z}) \rangle_{\mathcal{H}_K}.$$



A mapping from \mathcal{H}_K to \mathcal{H}_Y is established by identifying

$$k_j(\cdot) = K(\cdot, \mathbf{x}_j), \quad \mathbf{x}_j \in \Omega, \quad j = 1, 2, \dots$$

with eigenfunctions φ_n , $n = 1, 2, \dots$, of the covariance kernel K of Y .

Using Mercer's theorem and the KL theorem, we can use φ_n and λ_n to represent both functions $f \in \mathcal{H}_K$ and sample paths $y \in \mathcal{H}_Y$ as

$$f(\mathbf{x}) = \langle f, K(\cdot, \mathbf{x}) \rangle_{\mathcal{H}_K} = \sum_{n=1}^{\infty} \hat{f}_n \varphi_n(\mathbf{x}), \quad \text{with } \hat{f}_n = \frac{\langle f, \varphi_n \rangle_{\mathcal{H}_K}}{\|\varphi_n\|_{\mathcal{H}_K}^2} = \lambda_n \langle f, \varphi_n \rangle_{\mathcal{H}_K}$$

$$y_{\mathbf{x}} = Y_{\mathbf{x}}(\omega) = \sum_{n=1}^{\infty} Z_n(\omega) \sqrt{\lambda_n} \varphi_n(\mathbf{x}), \quad \text{with } Z_n(\omega) = \frac{1}{\sqrt{\lambda_n}} \int_{\Omega} Y_{\mathbf{x}}(\omega) \varphi_n(\mathbf{x}) d\mathbf{x}.$$



Remark

The above duality is *not limited to point evaluation* and can be extended in various ways.

- It can be extended to *arbitrary linear functionals* (see [Wah90] or [See04])
 - This provides a *stochastic interpretation of numerical analysis problems such as Hermite interpolation, or collocation solution of PDEs.*
 - For example, [MKGL96] discusses kriging when derivative information is available.
- It can be extended to cover *non-centered, non-Gaussian processes* (see [BTA04, Chapter 2]).



Flavors of Kriging

Many different variants of kriging can be found in the literature, e.g.

simple kriging: for zero-mean or **centered random fields** Y , i.e.,

$\mu(\mathbf{x}) = \mathbb{E}[Y_{\mathbf{x}}] = 0$; uses positive definite covariance kernel.

- If the process/data is not centered, then one can either center the data in a preprocessing step and use simple kriging, or use

universal kriging: also uses a positive definite covariance kernel and adds a deterministic polynomial term to model the **trend** (or mean)

- If the mean is modeled by only a constant, then the method is called **ordinary kriging**.

intrinsic kriging: uses **intrinsic random functions** and **generalized covariance functions** [Mat73] (similar to conditionally positive definite translation-invariant kernels)

We focus on simple kriging.



Kriging: a regression approach

Assume: the approximate value of a realization of a **zero-mean** (Gaussian) random field is given by a **linear predictor** of the form

$$\hat{Y}_{\mathbf{x}} = \sum_{j=1}^N Y_{\mathbf{x}_j} w_j(\mathbf{x}) = \mathbf{w}(\mathbf{x})^T \mathbf{Y},$$

where $\hat{Y}_{\mathbf{x}}$ and $Y_{\mathbf{x}_j}$ are **random variables**, $\mathbf{Y} = (Y_{\mathbf{x}_1}, \dots, Y_{\mathbf{x}_N})^T$, and $\mathbf{w}(\mathbf{x}) = (w_1(\mathbf{x}), \dots, w_N(\mathbf{x}))^T$ is a vector of **weight functions** at \mathbf{x} . Since all of the $Y_{\mathbf{x}_j}$ have zero mean the predictor $\hat{Y}_{\mathbf{x}}$ is automatically **unbiased**.

Goal: to compute “optimal” weights $w_j^*(\cdot)$, $j = 1, \dots, N$. To this end, consider the **mean-squared error (MSE)** of the predictor, i.e.,

$$\text{MSE}(\hat{Y}_{\mathbf{x}}) = \mathbb{E} \left[\left(Y_{\mathbf{x}} - \mathbf{w}(\mathbf{x})^T \mathbf{Y} \right)^2 \right].$$

We now present details (see also, e.g., [SWMW89, BTA04]).



Let's work out the MSE:

$$\begin{aligned} \text{MSE}(\hat{Y}_{\mathbf{x}}) &= \mathbb{E} \left[\left(Y_{\mathbf{x}} - \mathbf{w}(\mathbf{x})^T \mathbf{Y} \right)^2 \right] \\ &= \mathbb{E}[Y_{\mathbf{x}} Y_{\mathbf{x}}] - 2\mathbb{E}[Y_{\mathbf{x}} \mathbf{w}(\mathbf{x})^T \mathbf{Y}] + \mathbb{E}[\mathbf{w}(\mathbf{x})^T \mathbf{Y} \mathbf{Y}^T \mathbf{w}(\mathbf{x})] \end{aligned}$$

Now use $\mathbb{E}[Y_{\mathbf{x}} Y_{\mathbf{z}}] = K(\mathbf{x}, \mathbf{z})$ (since \mathbf{Y} is centered):

$$\text{MSE}(\hat{Y}_{\mathbf{x}}) = K(\mathbf{x}, \mathbf{x}) - 2\mathbf{w}(\mathbf{x})^T \mathbf{k}(\mathbf{x}) + \mathbf{w}(\mathbf{x})^T \mathbf{K} \mathbf{w}(\mathbf{x}),$$

where

$\mathbf{k}(\mathbf{x}) = (k_1(\mathbf{x}), \dots, k_N(\mathbf{x}))^T$: with $k_j(\mathbf{x}) = K(\mathbf{x}, \mathbf{x}_j) = \mathbb{E}[Y_{\mathbf{x}} Y_{\mathbf{x}_j}]$
 \mathbf{K} : the covariance matrix has entries $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbb{E}[Y_{\mathbf{x}_i} Y_{\mathbf{x}_j}]$

Finding the minimum MSE is straightforward.¹ Differentiation and equating to zero yields

$$-2\mathbf{k}(\mathbf{x}) + 2\mathbf{K} \mathbf{w}(\mathbf{x}) = 0,$$

and so the optimum weight vector is

$$\mathbf{w}^*(\mathbf{x}) = \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}).$$

¹This is a quadratic form in $\mathbf{w}(\mathbf{x})$ and thus convex.



We have shown that the (simple) kriging predictor

$$\hat{Y}_{\mathbf{x}} = \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{Y}$$

is the **best** (in the MSE sense) **linear unbiased predictor** (BLUP).

Since we are given the observations \mathbf{y} as realizations of \mathbf{Y} we can compute the **prediction**

$$\hat{y}_{\mathbf{x}} = \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{y}.$$

Notice that this is **formally identical to the kernel interpolant!**



The MSE of the kriging predictor with optimal weights $\hat{\mathbf{w}}^*(\cdot)$,

$$\mathbb{E} \left[\left(Y_{\mathbf{x}} - \hat{Y}_{\mathbf{x}} \right)^2 \right] = K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x}),$$

is known as the **kriging variance**.

Remark

- 1 We will see later that the *kriging variance is formally identical to the square of the power function at \mathbf{x}* , an important quantity in error estimates for kernel-based interpolation.
- 2 For Gaussian random fields the BLUP is also the best *nonlinear unbiased predictor* (see, e.g., [BTA04, Chapter 2]).



Remark

- 1 The *simple kriging approach just described is precisely how Krige introduced the method*:
 - The unknown value to be predicted is given by a *weighted average of the observed values*, where the *weights depend on the prediction location*.
 - Usually *one assigns a smaller weight to observations further away from \mathbf{x}* .

The latter statement implies that one should be using kernels whose associated weights decay away from \mathbf{x} . *Positive definite translation invariant kernels have this property*.

- 2 The more advanced kriging variants are discussed in papers such as [SWMW89, SSS13], or books such as [Cre93, Ste99, BTA04].



Kriging: a Bayesian approach

The use of a **Bayesian perspective within numerical analysis** was discussed by Persi Diaconis [Dia88], who suggests that such ideas were already entertained by Henri Poincaré [Poi96].

Formulating Gaussian processes within a **Bayesian framework** is attractive because they are hierarchical in nature and relatively easy to implement.

Once we have done this, we can apply powerful statistical methods such as **maximum likelihood estimation**, **confidence intervals**, and **Bayesian inference**.



Bayes' rule

The definition of **conditional density** allows us to express the **joint density** of the predicted value \hat{y}_x and the vector \mathbf{y} of observations as

$$p_{Y_x, \mathbf{Y}}(\hat{y}_x, \mathbf{y}) = p_{Y_x | \mathbf{Y}}(\hat{y}_x | \mathbf{Y} = \mathbf{y}) p_{\mathbf{Y}}(\mathbf{y}) = p_{\mathbf{Y} | Y_x}(\mathbf{y} | Y_x = \hat{y}_x) p_{Y_x}(\hat{y}_x).$$

This immediately gives rise to **Bayes' rule**

$$p_{Y_x | \mathbf{Y}}(\hat{y}_x | \mathbf{Y} = \mathbf{y}) = \frac{p_{\mathbf{Y} | Y_x}(\mathbf{y} | Y_x = \hat{y}_x) p_{Y_x}(\hat{y}_x)}{p_{\mathbf{Y}}(\mathbf{y})},$$

where

$p_{Y_x | \mathbf{Y}}(\hat{y}_x | \mathbf{Y} = \mathbf{y})$: **posterior density** for the prediction

$p_{\mathbf{Y} | Y_x}(\mathbf{y} | Y_x = \hat{y}_x)$: **likelihood** of the prediction, also denoted by $L(\hat{y}_x; \mathbf{y})$
 indicates how compatible the prediction \hat{y}_x is with the data

$p_{Y_x}(\hat{y}_x)$: **prior density** which we assume to be uninformed

$p_{\mathbf{Y}}(\mathbf{y})$: **evidence** (on which we have no influence)



We want to **maximize the likelihood**. Using the definitions of likelihood and joint density we get

$$L(\hat{\mathbf{y}}_{\mathbf{x}}; \mathbf{y}) = p_{\mathbf{Y}|\mathbf{Y}_{\mathbf{x}}}(\mathbf{y} | \mathbf{Y}_{\mathbf{x}} = \hat{\mathbf{y}}_{\mathbf{x}}) = \frac{1}{p_{\mathbf{Y}_{\mathbf{x}}}(\hat{\mathbf{y}}_{\mathbf{x}})} p_{\mathbf{Y}_{\mathbf{x}}, \mathbf{Y}}(\hat{\mathbf{y}}_{\mathbf{x}}, \mathbf{y}) \propto p_{\mathbf{Y}_{\mathbf{x}}, \mathbf{Y}}(\hat{\mathbf{y}}_{\mathbf{x}}, \mathbf{y}),$$

i.e., the **likelihood is proportional to the joint probability**.

Since our **random field is zero-mean Gaussian** we have

$$L(\hat{\mathbf{y}}_{\mathbf{x}}; \mathbf{y}) \propto p_{\mathbf{Y}_{\mathbf{x}}, \mathbf{Y}}(\hat{\mathbf{y}}_{\mathbf{x}}, \mathbf{y}) = \frac{1}{\sqrt{(2\pi\sigma^2)^N \det \tilde{\mathbf{K}}}} \exp\left(-\frac{1}{2\sigma^2} \tilde{\mathbf{y}}^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{y}}\right),$$

where $\tilde{\mathbf{y}} = \begin{pmatrix} \hat{\mathbf{y}}_{\mathbf{x}} \\ \mathbf{y} \end{pmatrix}$, $\tilde{\mathbf{K}} = \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & \mathbf{k}(\mathbf{x})^T \\ \mathbf{k}(\mathbf{x}) & \mathbf{K} \end{pmatrix}$ and σ^2 is the **process variance**.



Now we want to find the prediction \hat{y}_x that maximizes the likelihood.

It's more convenient to minimize the negative log-likelihood, i.e., find \hat{y}_x such that

$$-\log L(\hat{y}_x; \mathbf{y}) = \frac{N}{2} \log(2\pi\sigma^2) + \frac{1}{2} \log \det \tilde{\mathbf{K}} + \frac{\tilde{\mathbf{y}}^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{y}}}{2\sigma^2} + \text{const.}$$

is minimized.

The only term that depends on \hat{y}_x is the quadratic form

$$Q(\tilde{\mathbf{y}}) = \frac{\tilde{\mathbf{y}}^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{y}}}{2\sigma^2}.$$



We will need to use the Schur complement representation of the inverse of a block matrix:

$$\tilde{K}^{-1} = \begin{pmatrix} K(\mathbf{x}, \mathbf{x}) & \mathbf{k}(\mathbf{x})^T \\ \mathbf{k}(\mathbf{x}) & K \end{pmatrix}^{-1} = \begin{pmatrix} A & B \\ B^T & C \end{pmatrix},$$

where

$$A = \frac{1}{K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1} \mathbf{k}(\mathbf{x})},$$

$$B = -\frac{\mathbf{k}(\mathbf{x})^T K^{-1}}{K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1} \mathbf{k}(\mathbf{x})},$$

$$C = K^{-1} + \frac{K^{-1} \mathbf{k}(\mathbf{x}) \mathbf{k}(\mathbf{x})^T K^{-1}}{K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1} \mathbf{k}(\mathbf{x})}.$$

All blocks of \tilde{K}^{-1} are guaranteed to exist whenever K is a strictly positive definite covariance kernel.



If we substitute the Schur complement formula into the quadratic form

$$Q(\tilde{\mathbf{y}}) = \frac{\tilde{\mathbf{y}}^T \tilde{\mathbf{K}}^{-1} \tilde{\mathbf{y}}}{2\sigma^2}$$

we get

$$Q(\tilde{\mathbf{y}}) = \frac{1}{2\sigma^2} \left(\hat{y}_{\mathbf{x}}^2 \mathbf{A} + 2\hat{y}_{\mathbf{x}} \mathbf{B} \mathbf{y} + \mathbf{y}^T \mathbf{C} \mathbf{y} \right).$$

Differentiating with respect to $\hat{y}_{\mathbf{x}}$ and equating the result to zero yields

$$\begin{aligned} \mathbf{A} \hat{y}_{\mathbf{x}} + \mathbf{B} \mathbf{y} &= 0 \\ \iff \frac{\hat{y}_{\mathbf{x}}}{K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})} - \frac{\mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{y}}{K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{k}(\mathbf{x})} &= 0. \end{aligned}$$

Therefore the optimal prediction is

$$\hat{y}_{\mathbf{x}} = \mathbf{k}(\mathbf{x})^T \mathbf{K}^{-1} \mathbf{y},$$

just as we derived earlier with the regression approach.



Remark

- The *process variance dropped out* when we obtained the optimal prediction. However, the process variance *does play a role in determining optimal parameters in the covariance kernel K* . We will come back to this idea later.
- The Bayesian setting in which we have performed this derivation also tells us that
 - $\hat{y}_{\mathbf{x}}$ is the posterior mean and
 - the posterior variance is given by $K(\mathbf{x}, \mathbf{x}) - \mathbf{k}(\mathbf{x})^T K^{-1} \mathbf{k}(\mathbf{x})$, the kriging variance.



Karhunen–Loève Expansions

The Karhunen–Loève theorem was presented in Chapter 2.

The KL expansion is especially useful for practical applications such as SPDEs since it separates the random component of the process from a deterministic component.

In practice, a truncated KL expansion is used. This truncated expansion is optimal in the MSE sense, i.e., an M -term KL expansion is the best M -term approximation of Y in the MSE sense.

On the other hand, usually the random field Y is unknown and so the random variables Z_n must be simulated as i.i.d. random variables following the distribution of Y .

The eigenfunctions φ_n and eigenvalues λ_n of K may also be unknown, in which case a numerical method is required to solve the Hilbert–Schmidt integral eigenvalue problem.



The KL expansion is frequently used to model random coefficients in stochastic systems.

However, the solution itself is generally *not* modeled by KL since the covariance of the solution is assumed to be unknown.

In such cases one frequently uses (generalized) polynomial chaos to approximate functionals of random variables.

According to [NX12], algorithms for obtaining generalized polynomial chaos expansions for the solution of PDEs come in two flavors:

- “intrusive” Galerkin methods, which require lots of effort to modify existing deterministic codes
- “nonintrusive” collocation methods, which require a minimal amount of coding overhead.

If the dependence on the random inputs is smooth one may be able to exploit the spectral accuracy of certain collocation methods (see [CFY12] for similar insights).



References I

- [BTA04] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, Dordrecht, 2004.
- [CFY12] Igor Cialenco, Gregory E. Fasshauer, and Qi Ye, *Approximation of stochastic partial differential equations by a kernel-based collocation method*, International Journal of Computer Mathematics **89** (2012), no. 18, 2543–2561.
- [Cre93] N. Cressie, *Statistics for Spatial Data*, revised ed., Wiley-Interscience, 1993.
- [Dia88] P. Diaconis, *Bayesian numerical analysis*, Statistical decision theory and related topics IV, Papers from the 4th Purdue Symp., West Lafayette/Indiana 1986 (S. S. Gupta and J. O. Berger, eds.), vol. 1, Springer-Verlag, New York, 1988, pp. 163–175.
- [Kri51] D. G. Krige, *A statistical approach to some basic mine valuation problems on the Witwatersrand*, J. Chem. Met. & Mining Soc., S. Africa **52** (1951), no. 6, 119–139.



References II

- [Mat65] G. Matheron, *Les variables régionalisées et leur estimation*, Masson (Paris), 1965.
- [Mat73] _____, *The intrinsic random functions and their applications*, *Advances in Applied Probability* **5** (1973), no. 3, 439–468.
- [MKGL96] K. V. Mardia, J. T. Kent, C. R. Goodall, and J. A. Little, *Kriging and splines with derivative information*, *Biometrika* **83** (1996), no. 1, 207–221.
- [NX12] Akil Narayan and Dongbin Xiu, *Stochastic collocation methods on unstructured grids in high dimensions via interpolation*, *SIAM Journal on Scientific Computing* **34** (2012), no. 3, A1729–A1752.
- [Par61] Emanuel Parzen, *An approach to time series analysis*, *The Annals of Mathematical Statistics* **32** (1961), no. 4, 951–989.
- [Par70] E. Parzen, *Statistical inference on time series by RKHS methods*, *Proceedings 12th Biennial Seminar, Canadian Mathematical Congress*, 1970, pp. 1–37.



References III

- [Poi96] Henri Poincaré, *Calcul des probabilités*, George Carré, Paris, 1896.
- [See04] M. Seeger, *Gaussian processes for machine learning*, International Journal of Neural Systems **14** (2004), no. 2, 69–106.
- [SSS13] M. Scheuerer, R. Schaback, and M. Schlather, *Interpolation of spatial data — a stochastic or a deterministic problem?*, European Journal of Applied Mathematics **24** (2013), no. 04, 601–629.
- [Ste99] M. L. Stein, *Interpolation of Spatial Data: Some theory for Kriging*, Springer-Verlag, New York, 1999.
- [SWMW89] Jerome Sacks, W. J. Welch, T. J. Mitchell, and H. P. Wynn, *Design and analysis of computer experiments*, Statistical Science **4** (1989), no. 4, 409–423.
- [Wah90] Grace Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.

