Selected Topics of Spatial Statistics

Jyunyi Liao

Preface

This is a collection of lecture notes for the spatial statistics part of *Time Series and Spatial Statistics (DATA130013)*, a wonderful course instructed by Nan Zhang in Fudan University.

Contents

1 Random Field **3**
1.1 Definition and Properties . . . 3
1.1.1 Mean and Covariance . . . 3
1.1.2 Properties of Covariance Function . . . 3
1.2 Staionarity and Isotropy of Random Field . . . 4
1.3 Variogram . . . 4
1.3.1 Properties of Semivariogram . . . 4
1.3.2 Nugget Effect . . . 5

2 Kernel Functions **7**
2.1 Inner Product and Hilbert Space . . . 7
2.2 Kernel Functions . . . 7
2.3 Properties of Kernels . . . 8
2.4 Exapmles of Kernels . . . 9

3 Gaussian Process **11**
3.1 Definition and properties . . . 11
3.2 Hierarchical model for geostatistical process . . . 12
3.3 Analytical Examples . . . 12
3.3.1 Brownian Motion . . . 12
3.3.2 Ornstein–Uhlenbeck process (O-U process) . . . 13

4 Kriging **16**
4.1 Simple Kriging . . . 16
4.1.1 Noise-free observations . . . 16
4.1.2 Observations with additive noise . . . 17
4.1.3 Overlapping observations . . . 17
4.2 Universal Kriging . . . 19
4.3 Ordinary Kriging . . . 20

5 Reproducing Kernel Hilbert Space **21**
5.1 Definition and Properties . . . 21
5.2 Construction of RKHS . . . 21
5.2.1 From Pre-Hilbert Space to its Completion: Moore-Aronszajn Theorem . . . 21
5.2.2 Eigenanalysis: Mercer's Theorem . . . 22

6 Kernel Ridge Regression **26**
6.1 Nonparametric Regression . . . 26
6.1.1 Penalized least squares . . . 26
6.1.2 Generalized loss . . . 26
6.2 Representer Theorem . . . 27
6.3 Equivalence between KRR and Kriging . . . 28

7 Karhunen-Loève Expansion **29**
7.1 Kosambi-Karhunen-Loève Theorem . . . 29
7.2 Function Approximation with Orthogonal Bases . . . 30
7.2.1 Orthogonal Basis . . . 31
7.2.2 Optimality of truncated Karhunen-Loève expansion . . . 32
7.3 Example: Brownian Motion . . . 33

1 Random Field

1.1 Definition and Properties

Definition 1.1. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a random field is a family or collection of random variables indexed by elements in a topological space $\mathcal{T}$. That is, a random field $Z(\cdot)$ is a collection of random variables

$$\{Z(s) : \Omega \to \mathbb{R} \mid s \in \mathcal{T}\}, \tag{1.1}$$

where each $Z(s)$ is a random variable indexed by s. Some examples are shown below.

- Time Series: $X(t),\ t \in \mathbb{Z}$, like random walk;
- Stochastic Process: $X(t),\ t \in \mathbb{R}$, like Poisson process, Brownian process, etc.;
- Spatial Process: $Z(s),\ s \in \mathcal{D} \subseteq \mathbb{R}^d,\ d \geq 2$;
- Spatio-temporal Process: $Z(s,t),\ s \in \mathcal{D},\ t \in \mathbb{R}$.

1.1.1 Mean and Covariance

For a random field $Z(\cdot)$ defined on $\mathcal{D} \subseteq \mathbb{R}^d$, the mean function is defined as a function $\mu_Z : \mathcal{D} \to \mathbb{R}$, given by

$$\mu_Z(s) := \mathbb{E}[Z(s)]; \tag{1.2}$$

the covariance function $K_Z : \mathcal{D} \times \mathcal{D} \to \mathbb{R}$ is defined as

$$K_Z(s_1, s_2) := \mathrm{Cov}\left(Z(s_1), Z(s_2)\right) = \mathbb{E}\left[\left(Z(s_1) - \mu_Z(s_1)\right)\left(Z(s_2) - \mu_Z(s_2)\right)\right]. \tag{1.3}$$

By definition, the covariance function is symmetric. Furthermore, you can verify that it is positive definite, i.e., $\forall n \in \mathbb{N},\ \forall s_1, \cdots, s_n \in \mathcal{D}$, the Gram matrix defined by $\mathbf{K} = \{K_Z(s_i, s_j)\}_{i,j=1}^n$ is positive semidefinite.

1.1.2 Properties of Covariance Function

Suppose that $K(\cdot,\cdot)$ is a valid covariance function defined on $\mathcal{D} \times \mathcal{D}$, where $\mathcal{D} \subseteq \mathbb{R}^d$.

Definition 1.2 (Stationarity). A covariance function K is called stationary if $\forall s_1, s_2 \in \mathcal{D}$, $K(s_1, s_2)$ only depends on $s_1 - s_2$. That is, there exists some $K_1 : \mathbb{R}^d \to \mathbb{R}$ such that

$$K(s_1, s_2) = K_1(s_1 - s_2) \quad \forall s_1, s_2 \in \mathcal{D}. \tag{1.4}$$

A stationary covariance function is invariant to translation.

Definition 1.3 (Isotropy). A covariance function K is called isotropic if $\forall s_1, s_2 \in \mathcal{D}$, $K(s_1, s_2)$ only depends on $\|s_1 - s_2\|$, where $\|\cdot\|$ denotes L^2-norm. That is, there exists some $K_2 : \mathbb{R}_+ \to \mathbb{R}$ such that

$$K(s_1, s_2) = K_2(\|s_1 - s_2\|) \quad \forall s_1, s_2 \in \mathcal{D}. \tag{1.5}$$

Definition 1.4 (Dot product). A covariance function K has rotational invariance if $\forall s_1, s_2 \in \mathcal{D}$, $K(s_1, s_2)$ only depends on their dot product $s_1 \cdot s_2$. That is, there exists some $K_3 : \mathbb{R} \to \mathbb{R}$ such that

$$K(s_1, s_2) = K_3(s_1 \cdot s_2) \quad \forall s_1, s_2 \in \mathcal{D}. \tag{1.6}$$

1.2 Staionarity and Isotropy of Random Field

Definition 1.5. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A random field $Z(\cdot)$ is called stationary (isotropic) if it satisfies the following 3 conditions:

- $Z(\cdot)$ has constant mean, that is, μ_Z is a constant,
- $Z(\cdot)$ has finite 2nd moment, that is, $Z(\cdot) \subseteq L_2(\Omega)$, and
- $Z(\cdot)$ has stationary (isotropic) covariance function.

A random field $Z'(\cdot)$ is called anisotropic if it is not isotropic.

Definition 1.6 (Geometric anisotropy). A random field $Z(\cdot)$ is called geometrically anisotropic if it is isotropic after a linear transformation, i.e. there exists a non-singular matrix $V \in \mathbb{R}^{d\times d}$ such that $Z' : s \mapsto Z(Vs)$ is isotropic.

1.3 Variogram

Definition 1.7 (Variogram). Given a random field $Z(\cdot)$ defined on $\mathcal{D} \subseteq \mathbb{R}^d$, the variogram of $Z(\cdot)$ is defined as the variance of the difference between field values at two locations:

$$2\gamma(s_1, s_2) := \mathrm{Var}\{Z(s_1) - Z(s_2)\}, \tag{1.7}$$

where γ is called the semivariogram.

A stationary variogram and semivariogram can be represented as a function of the difference $h = s_1 - s_2$ between locations only:

$$2\gamma(h) := \mathrm{Var}\{Z(s+h) - Z(s)\}. \tag{1.8}$$

In the case of stationary random field, we have

$$2\gamma(h) = 2K_Z(0) - 2K_Z(h), \tag{1.9}$$

where K_Z is the covariance function of stationary field $Z(\cdot)$. Hence, a stationary random field has a stationary variogram. Analogously, an isotropic variogram and semivariogram can be represented as a function of the distance $\|h\| = \|s_1 - s_2\|$ between locations only:

$$2\gamma(\|h\|) := \mathrm{Var}\{Z(s+h) - Z(s)\}. \tag{1.10}$$

In general, a variogram gives a description of the spatial continuity of our data.

1.3.1 Properties of Semivariogram

Proposition 1.8. A semivariogram γ has the following properties:

- A semivariogram is nonnegative and symmetric, that is, $\gamma(s_1, s_2) = \gamma(s_2, s_1) \geq 0\ \forall s_1, s_2 \in \mathcal{D}$;
- The (isotropic) semivariogram at distance 0 is always 0, since $Z(s) - Z(s) \equiv 0$;
- Let $\mathcal{D} = \mathbb{R}^d$. A function γ is a (isotropic) semivariogram if and only if $\gamma(0) = 0$ and γ is a conditionally negative definite function, i.e. for all $w_1, \cdots, w_n \in \mathbb{R}$ subjected to $w_1 + \cdots + w_n = 0$ and locations $s_1, \cdots, s_n \in \mathbb{R}^d$, it holds
$$\sum_{i=1}^{n}\sum_{j=1}^{n} w_i w_j \gamma(s_i - s_j) \leq 0. \tag{1.11}$$

Proof. We only show the third property.

"Only if" part. Consider a random vector $\mathbf{Z} = (Z(s_1), \cdots, Z(s_n))^\top$ with mean $\mu = \mathbb{E}\mathbf{Z}$ and covariance matrix $\mathbf{K} = \{K(s_i - s_j)\}_{i,j=1}^n$. Define the semivariogram matrix of $\mathbf{Z}$ to be $\Gamma = \{\gamma(s_i - s_j)\}_{i,j=1}^n$. By definition,

$$\Gamma = \frac{1}{2}\text{vdiag}(\mathbf{K})\mathbf{1}_n^\top + \frac{1}{2}\mathbf{1}_n\text{vdiag}(\mathbf{K})^\top - \mathbf{K}, \tag{1.12}$$

where $\text{vdiag}(\mathbf{K}) = (K(0), \cdots, K(0))^\top$ and $\mathbf{1}_n$ is an n-dimensional all-one vector. All diagonal entries of Γ are zero. Hence for any $\mathbf{w} \in \mathbb{R}^n$ such that $\mathbf{1}_n^\top \mathbf{w} = 0$,

$$\mathbf{w}^\top \Gamma \mathbf{w} = -\mathbf{w}^\top \mathbf{K} \mathbf{w} \leq 0, \tag{1.13}$$

which implies that γ is negative definite.

"If" part: This proof is adapted from Matheron ,G. (1972) P-4-1 [1]. Suppose γ is a conditionally negative definite function, i.e. it satisfies equation (1.11), and $\gamma(0) = 0$. Now fix $s_1, \cdots, s_n \in \mathbb{R}^d$. We are going to show that $\forall \boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top \in \mathbb{R}^n$, we have

$$\sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \left(\gamma(s_i) + \gamma(s_j) - \gamma(s_i - s_j)\right) \geq 0. \tag{1.14}$$

Let $\alpha_0 = -(\alpha_1 + \cdots + \alpha_n)$ and s_0. Note that $\gamma(0) = 0$, the inequality (1.14) becomes

$$-\sum_{i=0}^n \sum_{j=0}^n \alpha_i \alpha_j \gamma(s_i - s_j) \geq 0, \tag{1.15}$$

which immediately follows from the conditional negative-definiteness of γ. Then we can construct a Gaussian process $\{Z(s) : s \in \mathbb{R}^d\}$ with mean zero and covariance K such that

$$K(s, s') = \gamma(s) + \gamma(s') - \gamma(s - s'), \ s, s' \in \mathbb{R}^d. \tag{1.16}$$

It can be verified that the semivariogram of $\{Z(s) : s \in \mathbb{R}^d\}$ is γ. □

1.3.2 Nugget Effect

Definition 1.8 (Nugget effect). The nugget effect is the variation of the process at a finer scale than the smallest distance measured:

$$c_0 := \lim_{h \to 0} \gamma(h). \tag{1.17}$$

It can be caused by random noise or measurement errors, and is shown graphically in the variogram plot as a discontinuity at the origin of the function.

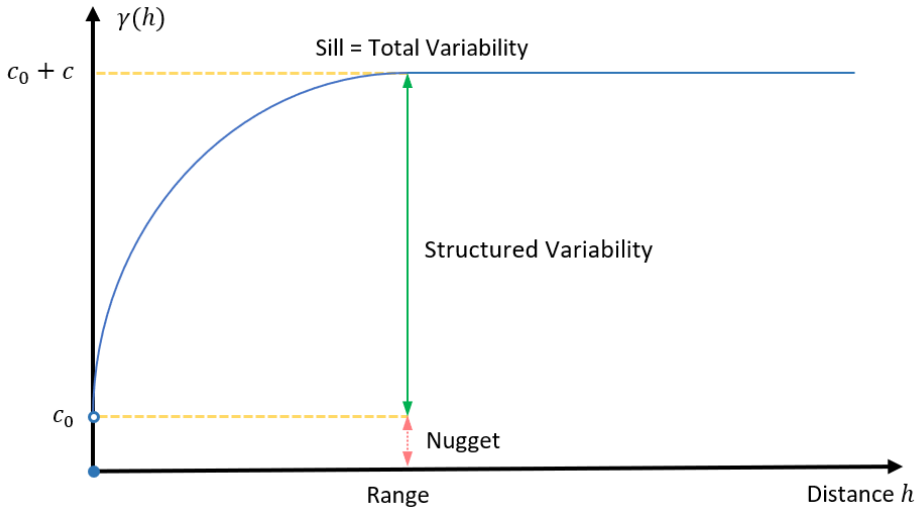

Figure 1: A variogram plot

Definition 1.9 (Empirical Variogram). Suppose that we have observations $z(s_1), \cdots, z(s_n)$, an empirical semivariogram is given by

$$\hat{\gamma}(h) = \frac{1}{2|\mathcal{N}(h)|} \sum_{(i,j)\in\mathcal{N}(h)} \{z(s_i) - z(s_j)\}^2, \tag{1.18}$$

where $\mathcal{N}(h)$ is a set of observation pairs whose distance is close to h.

2 Kernel Functions

2.1 Inner Product and Hilbert Space

Definition 2.1 (Inner product space). Let $\mathcal{H}$ be a vector space over $\mathbb{R}$. A function $\langle\cdot,\cdot\rangle_{\mathcal{H}} : \mathcal{H}\times\mathcal{H}\to\mathbb{R}$ is said to be an inner product on $\mathcal{H}$ if it satisfies the follwing three properties:

- (Non-negativeness) $\langle f,f\rangle_{\mathcal{H}}\geq 0\ \forall f\in\mathcal{H}$, and $\langle f,f\rangle_{\mathcal{H}}=0$ if and only if $f=0$,
- (Symmetry) $\langle f,g\rangle_{\mathcal{H}}=\langle g,f\rangle_{\mathcal{H}},\ f,g\in\mathcal{H}$, and
- (Linearity) $\langle\alpha f_1+\beta f_2,g\rangle_{\mathcal{H}}=\alpha\langle f_1,g\rangle_{\mathcal{H}}+\beta\langle f_2,g\rangle_{\mathcal{H}},\ \alpha,\beta\in\mathbb{R},\ f_1,f_2,g\in\mathcal{H}$.

$\mathcal{H}$ is called an inner product space. A norm is induced by the inner product: $\|f\|_{\mathcal{H}}=\sqrt{\langle f,f\rangle_{\mathcal{H}}}$.

Remark. An inner product space is also called a pre-Hilbert space. Moreover, For a vector space over $\mathbb{C}$, the second property is modified as conjugate symmetry: $\langle f,g\rangle_{\mathcal{H}}=\overline{\langle g,f\rangle_{\mathcal{H}}},\ f,g\in\mathcal{H}$.

Definition 2.2 (Hilbert Space). A Hilbert space is a complete inner product space. In particular, every Hilbert space is a Banach space with respect to the norm induced by its inner product.

Remark. Recall the definition of Cauchy sequence and Banach space:

- A sequence $\{f_n\}_{n\in\mathbb{N}^*}\subset\mathcal{H}$ is called a Cauchy sequence with respect to norm $\|\cdot\|_{\mathcal{H}}$, if $\forall\varepsilon>0,\ \exists N\in\mathbb{N}^*$ such that $\forall m,n>N,\ \|f_n-f_m\|_{\mathcal{H}}\leq\varepsilon$.
- A normed vector space $(\mathcal{H},\|\cdot\|_{\mathcal{H}})$ is called a Banach space if any Cauchy sequence $\{f_n\}_{n\in\mathbb{N}^*}\subset\mathcal{H}$ converges to some $f_\infty\in\mathcal{H}$, i.e. $\lim_{n\to\infty}\|f_n-f_\infty\|_{\mathcal{H}}=0$.

Then, the completeness of a Hilbert space $\mathcal{H}$ states that every Cauchy sequence in $\mathcal{H}$ converges with respect to $\|\cdot\|_{\mathcal{H}}$ to an element in $\mathcal{H}$.

2.2 Kernel Functions

Definition 2.3 (Kernel). Let $\mathcal{X}$ be a non-empty set. A function $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ is called a kernel on $\mathcal{X}$ if there exists an $\mathbb{R}$-Hilbert space $\mathcal{H}$ and a map $\phi:\mathcal{X}\to\mathcal{H}$ such that $\forall x,y\in\mathcal{X}$,

$$k(x,y)=\langle\phi(x),\phi(y)\rangle_{\mathcal{H}}. \tag{2.1}$$

It is evident that $k(\cdot,\cdot)$ is symmetric, i.e. $k(x,y)=k(y,x)\ \forall,x,y\in\mathcal{X}$.

Remark. In machine learning, ϕ is called a feature map, and $\mathcal{H}$ is called a feature space of k.

Proposition 2.4. All kernel functions are positive definite. More specifically, let $\mathcal{X}$ be a non-empty set and $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ a kernel on it, then $\forall n\in\mathbb{N},\ \alpha_1,\cdots,\alpha_n\in\mathbb{R},\ x_1,\cdots,x_n\in\mathcal{X}$,

$$\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i k(x_i,x_j)\alpha_j\geq 0. \tag{2.2}$$

Proof. Fix $\alpha_{1:n}$ and $x_{1:n}$. Since $k(\cdot,\cdot)$ is a kernel on $\mathcal{X}$, there exists an $\mathbb{R}$-Hilbert space $\mathcal{H}$ and a map $\phi:\mathcal{X}\to\mathcal{H}$,

$$\begin{aligned}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i k(x_i,x_j)\alpha_j&=\sum_{i=1}^{n}\sum_{j=1}^{n}\langle\alpha_i\phi(x_i),\alpha_j\phi(x_j)\rangle_{\mathcal{H}}\\&=\left\langle\sum_{i=1}^{n}\alpha_i\phi(x_i),\sum_{j=1}^{n}\alpha_j\phi(x_j)\right\rangle_{\mathcal{H}}=\left\|\sum_{i=1}^{n}\alpha_i\phi(x_i)\right\|_{\mathcal{H}}^2\geq 0.\end{aligned} \tag{2.3}$$

Then we complete the proof. □

Theorem 2.5 (Symmetric, positive definite functions are kernels). A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a kernel if and only if it is symmetric and positive definite.

Proof. In view of the discussion above, it suffices to show that a symmetric and positive function $k(\cdot, \cdot)$ is a kernel. Let

$$\mathcal{H}_0 = \left\{ \sum_{i=1}^{n} x_i k(\cdot, x_i) : \ n \in \mathbb{N}, \ c_1, \cdots, c_n \in \mathbb{R}, \ x_1, \cdots, x_n \in \mathcal{X} \right\}, \tag{2.4}$$

and define $\langle \cdot, \cdot \rangle_{\mathcal{H}_0} : \mathcal{H} \times \mathcal{H} \to \mathbb{R}$ such that for

$$f := \sum_{i=1}^{m} \alpha_i k(\cdot, x_i) \in \mathcal{H}_0, \ g := \sum_{j=1}^{n} \beta_j k(\cdot, y_j) \in \mathcal{H}_0, \quad x_1, \cdots, x_m, y_1, \cdots, y_n \in \mathcal{X}, \tag{2.5}$$

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} \alpha_i \beta_j k(x_i, y_j). \tag{2.6}$$

You can verify that $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$ is an inner product on $\mathcal{H}_0$. Now let $\mathcal{H}$ be a completion of $\mathcal{H}_\prime$ with respect to $\langle \cdot, \cdot \rangle_{\mathcal{H}_0}$, then we have

$$\langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}} = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}_0} = k(x, y) \ \forall x, y \in \mathcal{X}. \tag{2.7}$$

Hence $\phi : x \mapsto k(\cdot, x)$ defines a feature map of k. □

2.3 Properties of Kernels

Proposition 2.6. The linear combination, limit and pointwise product of kernels are kernels.

- (Linearity) If k_1, k_2 are kernels, then $\forall \alpha, \beta \geq 0$, $\alpha k_1 + \beta k_2$ is a kernel.
- (Limit) For any kernel series $\{k_n\}_{n \in \mathbb{N}^*}$, if $\lim_{n \to \infty} k_n = k$ uniformly, then k is a kernel.
- (Pointwise product) If k_1, k_2 are kernels, then $k_1 \cdot k_2$ is a kernel.

Proof. We only show the third property. This is an immediate corollary of Schur's product theorem. Fix a positive integer n and $x_1, \cdots, x_n \in \mathcal{X}$, and denote matrices

$$\mathbf{K}_l = \{k_l(x_i, x_j)\}_{i,j=1}^{n}, \ l = 1, 2. \tag{2.8}$$

It suffices to show the Hadamard product $\mathbf{K} = \mathbf{K_1} \circ \mathbf{K_2}$ is positive semidefinite. For any $\mathbf{a} \in \mathbb{R}^n$, denote $\mathbf{A}$ to be the diagonal matrix such that $\mathbf{A}_{ii} = \mathbf{a}_i$, then

$$\begin{aligned} \mathbf{a}^\top \mathbf{K} \mathbf{a} &= \operatorname{trace} \left\{ (\mathbf{A}\mathbf{K}_1)^\top \mathbf{K}_2 \mathbf{A} \right\} \\ &= \operatorname{trace} \left\{ \mathbf{K}_1 \mathbf{A} \mathbf{K}_2 \mathbf{A} \right\} \\ &= \operatorname{trace} \left\{ \mathbf{K}_1^{1/2} \mathbf{A} \mathbf{K}_2^{1/2} \mathbf{K}_2^{1/2} \mathbf{A} \mathbf{K}_1^{1/2} \right\} \\ &= \left\| \mathbf{K}_2^{1/2} \mathbf{A} \mathbf{K}_1^{1/2} \right\|_{\mathrm{F}}^2 \geq 0, \end{aligned} \tag{2.9}$$

which concludes the proof. □

Below are some immediate corollaries of Proposition 2.6.

Corollary 2.7 (Polynomial kernels). Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ for $d \geq 1$, and let m be a positive integer and $c \geq 0$ be a non-negative real. Then

$$k(\mathbf{x}, \mathbf{y}) := (c + \langle \mathbf{x}, \mathbf{y} \rangle)^m \tag{2.10}$$

is a valid kernel.

Corollary 2.8 (Taylor series). Assume the Taylor series

$$f(z) = \sum_{n=0}^{\infty} a_n z^n, \quad |z| < r, z \in \mathbb{R} \tag{2.11}$$

converges for some $r \in (0, \infty]$, with $a_n \geq 0$ for all $n \geq 0$. Define $\mathcal{X}$ to be the $\sqrt{r}$-ball in $\mathbb{R}^d$. Then

$$k(\mathbf{x}, \mathbf{y}) = f(\langle \mathbf{x}, \mathbf{y} \rangle) = \sum_{n=0}^{\infty} a_n \langle \mathbf{x}, \mathbf{y} \rangle^n, \ \mathbf{x}, \mathbf{y} \in \mathcal{X} \tag{2.12}$$

defines a valid kernel.

Corollary 2.9 (Exponential). The exponential kernel on $\mathbb{R}^d$ is defined as

$$k(\mathbf{x}, \mathbf{y}) := \exp(\langle \mathbf{x}, \mathbf{y} \rangle), \ \mathbf{x}, \mathbf{y} \in \mathbb{R}^d, \tag{2.13}$$

and this is a valid kernel.

2.4 Exapmles of Kernels

Definition 2.10 (Gaussian RBF [radial basis function] kernels). Given $\sigma > 0$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the Gaussian RBF kernel is defined as

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\frac{1}{2\sigma^2} \|\mathbf{x} - \mathbf{y}\|^2 \right). \tag{2.14}$$

Now we are going to show that (2.14) is a valid kernel:

Proof. Let Z be a d-dimensional Gaussian vector such that $Z \sim N(0, \sigma^{-2} I_d)$. Then the characteristic function of Z can be calculated:

$$\varphi_Z(\lambda) = \mathbb{E}\left[\exp\left(\mathrm{i} \lambda^\top Z \right) \right] = \exp\left(-\frac{\lambda^\top \lambda}{2\sigma^2} \right), \quad \mathrm{i} = \sqrt{-1}. \tag{2.15}$$

Fix $n \in \mathbb{N}^*$ and $\alpha_1, \cdots, \alpha_n \in \mathbb{R},\ \mathbf{x}_1, \cdots, \mathbf{x}_n \in \mathbb{R}^d$, we have

$$\begin{aligned}
\sum_{s=1}^{n}\sum_{t=1}^{n} \alpha_s \exp\left(-\frac{\|\mathrm{x}_s - \mathrm{x}_t\|^2}{2\sigma^2}\right)\alpha_t &= \sum_{s=1}^{n}\sum_{t=1}^{n} \alpha_s \mathbb{E}\left[\exp(\mathrm{i}\mathbf{x}_s^\top Z - \mathrm{i}\mathbf{x}_t^\top Z)\right]\alpha_t \\
&= \mathbb{E}\left[\sum_{s=1}^{n}\sum_{t=1}^{n} \alpha_s \exp(\mathrm{i}\mathbf{x}_s^\top Z)\exp(-\mathrm{i}\mathbf{x}_t^\top Z)\alpha_t\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{n} \alpha_s \exp(\mathrm{i}\mathbf{x}_s^\top Z)\sum_{t=1}^{n}\alpha_t \exp(-\mathrm{i}\mathbf{x}_t^\top Z)\right] \\
&= \mathbb{E}\left[\sum_{s=1}^{n} \alpha_s \exp(\mathrm{i}\mathbf{x}_s^\top Z)\cdot\sum_{t=1}^{n}\alpha_t \overline{\exp(\mathrm{i}\mathbf{x}_t^\top Z)}\right] \\
&= \mathbb{E}\left[\left|\sum_{s=1}^{n} \alpha_s \exp(\mathrm{i}\mathbf{x}_s^\top Z)\right|^2\right] \geq 0. \qquad (2.16)
\end{aligned}$$

Therefore the Gaussian RBF is positive definite, hence is a valid kernel. □

Definition 2.11 (Laplacian kernels). Given $\alpha > 0$, for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, the Laplacian kernel is defined as

$$k(\mathbf{x}, \mathbf{y}) = \exp\left(-\alpha \sum_{i=1}^{n} |\mathbf{x}_i - \mathbf{y}_i|\right). \qquad (2.17)$$

The proof of validity is analogous to Gaussian RBF, with the Gaussian random variable replaced by Cauchy variable. Note the characteristic function of Cauchy distribution with location parameter 0 and scale parameter α is $\hat{\mu}(\lambda) = \exp(-\alpha|\lambda|)$.

Definition 2.12 (Matérn kernels). The form of the Matérn class of functions is given by

$$k(x, y) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{2\nu}|x-y|}{\ell}\right)^{\nu} K_\nu\left(\frac{\sqrt{2\nu}|x-y|}{\ell}\right), \qquad (2.18)$$

where $\ell > 0$ is a length-scale parameter, $\nu > 0$ is a smoothing parameter, and K_ν is the modified Bessel function of second type.

Remark. In equation (2.18), the parameter ν controls the smoothness of our kernel, and ℓ is the band width.

3 Gaussian Process

3.1 Definition and properties

The Gaussian process can be seen as an extension of multivariate Gaussian distribution from finite-dimensional case to infinite-dimensional case. It is a stochastic process or a random field (a collection of random variables indexed by time or space) such that the joint distribution of every finite collection of those random variables is Gaussian.

Definition 3.1 (Gaussian process). Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an index set $\mathcal{D}$, a random field $\{Z(s) \in L_2(\Omega) : \ s \in \mathcal{D}\}$ is called a Gaussian process, if for every finite set of indices $\{s_1, \cdots, s_n\} \subset \mathcal{D}$, the joint distribution of random variables $Z(s_1), \cdots, Z(s_n)$ is a multivariate Gaussian distribution.

The mean and covariance function of a Gaussian process is defined by

$$\mu(s) := \mathbb{E}[Z(s)], \ \ k(s,t) := \mathrm{Cov}\left\{Z(s), z(t)\right\}, \tag{3.1}$$

and they completely determine a Gaussian process $\mathcal{GP}\left(\mu(\cdot), k(\cdot,\cdot)\right)$.

Properties of Gaussian processes

- A Gaussian process can be seen as a distribution over real-valued functions $\{f : \mathcal{D} \to \mathbb{R}\}$. Suppose that $f \sim \mathcal{GP}\left(\mu(\cdot), k(\cdot,\cdot)\right)$, then $\forall n \in \mathbb{N}^*, \ x_1, \cdots, x_n \in \mathcal{D}$, the joint distribution of $f(x_1), \cdots, f(x_n)$ is given by
$$\begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} \sim N\left\{ \begin{pmatrix} \mu(x_1) \\ \vdots \\ \mu(x_n) \end{pmatrix}, \begin{pmatrix} k(x_1,x_1) & \cdots & k(x_1,x_n) \\ \vdots & \ddots & \vdots \\ k(x_n,x_1) & \cdots & k(x_n,x_n) \end{pmatrix} \right\}. \tag{3.2}$$
- The covariance function k is a kernel which is symmetric and positive definite. A wide range of kernels can be selected as the covariance function, such as Gaussian RBF and Matérn kernels.
- If a Gaussian process $Z(\cdot)$ is (weakly) stationary, i.e. it has constant mean μ, and the covariance $k(s,t)$ depends on only the difference between locations $s - t$, then $\forall n \in \mathbb{N}^*, \ s_1, \cdots, s_n \in \mathcal{D}$, and $\forall h$ such that $s_1 + h, \cdots, s_n + h \in \mathcal{D}$, then
$$(Z(s_1), \cdots, Z(s_n)) \overset{d}{=} (Z(s_1 + h), \cdots, Z(s_n + h)). \tag{3.3}$$
In a nutshell, weak stationarity implies strict stationarity in Gaussian processes.

Some Gaussian processes with zero mean and different covariance kernels are visualized below.

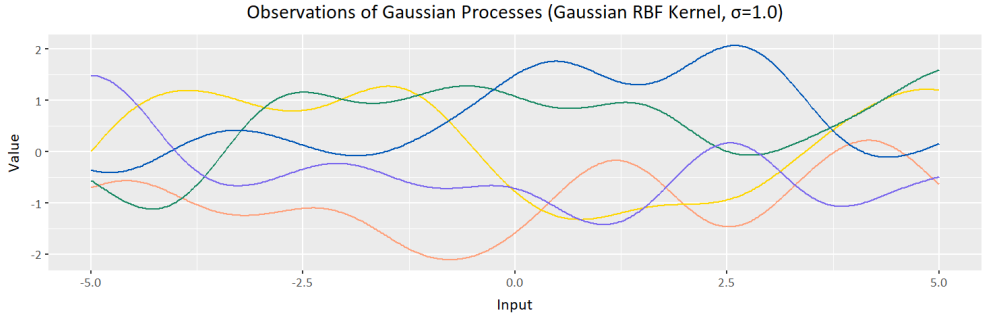

Figure 2: Sample paths of Gaussian processes with Gaussian RBF kernel as its covariance

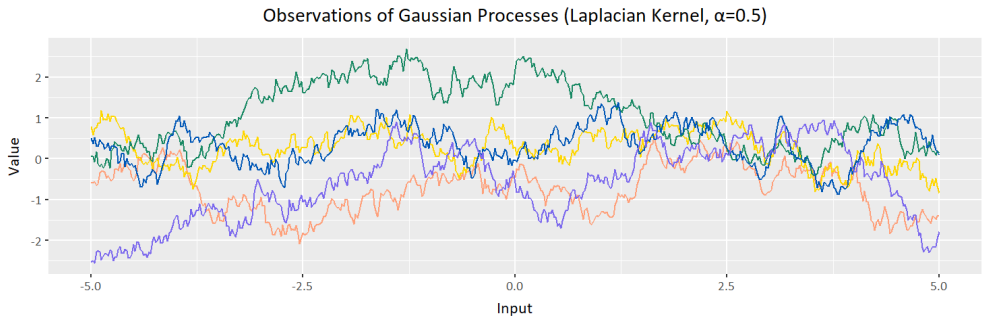

Figure 3: Sample paths of Gaussian processes with Laplacian RBF kernel as its covariance

3.2 Hierarchical model for geostatistical process

In practice, we often apply a hierarchical model to simulate spatial or geostatistical data. A hierarchical model has two components: a process model, which sets up some latent variable with covariance structure, and a data model, which generates observations from latent variables with measurement error.

Process Model. Suppose that $\mathbf{m}(\cdot) \in \mathbb{R}^p$ is a column vector of covariates related to the mean function. Given regression coefficients $\beta \in \mathbb{R}^p$ and kernel $k(\cdot, \cdot)$, let

$$Z(\cdot) \sim \mathcal{GP}\left(\mathbf{m}(\cdot)^\top \beta, k(\cdot, \cdot)\right). \tag{3.4}$$

Data Model. Conditioning on process $Z(\cdot)$, the observation at location $s \in \mathcal{D}$ is

$$Y(s)|Z(\cdot) \sim N(Z(s), \sigma^2), \ \sigma^2 > 0; \tag{3.5}$$

Equivalently, for n locations $s_1, \cdots, s_n$,

$$Y(s_i) = Z(s_i) + e_i, \ e_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \ i = 1, \cdots, n. \tag{3.6}$$

Hence we can use an error process $\epsilon(\cdot)$ to model our observational data:

$$Y(s_i) = \mathbf{m}(s_i)^\top \beta + \epsilon(s_i), \ \begin{pmatrix} \epsilon(s_1) \\ \vdots \\ \epsilon(s_n) \end{pmatrix} \sim N(0, \mathbf{K} + \sigma^2 \mathbf{I}_n), \tag{3.7}$$

where $\mathbf{K} = \{k(s_i, s_j)\}_{i,j=1}^n$.

3.3 Analytical Examples

3.3.1 Brownian Motion

Definition 3.2 (Brownian motion/Wiener process). Given a probability space $(\Omega, \mathcal{F}, \mathbb{R})$, a stochastic process $\{W(t) : \Omega \to \mathbb{R}, \ t \geq 0\}$ is a Brownian motion if it satisfies the following:

- $\mathbb{P}\{W(0) = 0\} = 1$;
- The sample path of $W(t)$ is continuous;
- $W(t)$ has independent and stationary increments;
- $W(t) \sim N(0, \sigma^2 t), \ \sigma > 0$.

For convenience of our discussion, we will consider standard Brownian motion in which $\sigma^2 = 1$. Let $W(\cdot)$ be a standard Brownian motion, let's investigate the properties of $W(\cdot)$.

Proposition 3.3 (Mean and Covariance). $\mathbb{E}[W(t)] = 0,\ \mathrm{Cov}\{W(s), W(t)\} = \min(s,t)$.

Proof. The mean of $W(t)$ is zero by definition, so we only consider the covariance. Without loss of generality, suppose $s < t$ (the case $s = t$ is trivial). Note that $W(t)$ has independent increments, we have

$$W(s) \perp W(t) - W(s),\ \mathbb{E}\left[W(s)\left(W(t) - W(s)\right)\right] = 0. \tag{3.8}$$

Hence

$$\begin{aligned}\mathrm{Cov}\{W(s), W(t)\} &= \mathbb{E}[W(s)W(t)] \\ &= \mathbb{E}[W(s)^2] + \mathbb{E}[W(s)(W(t) - W(s))] = s,\end{aligned} \tag{3.9}$$

which concludes the proof. □

Proposition 3.4 (Gaussianity). $W(\cdot)$ is a Gaussian process with zero mean and kernel $k(s,t) := \min(s,t)$.

Proof. It suffices to show that $\forall n \geq 1,\ 0 < t_1 < \cdots < t_n$, the joint distribution of corresponding variables $\mathbf{W} = (W(t_1), \cdots, W(t_n))^\top$ is Gaussian. Denote Brownian increments

$$\mathbf{Z} = \left(W(t_1), W(t_2) - W(t_1), \cdots, W(t_n) - W(t_{n-1})\right)^\top, \tag{3.10}$$

By stationarity and independence of increments, we have $\mathbf{Z} \sim N(0, \mathbf{D})$, where $\mathbf{D} = \mathrm{diag}\{t_1, t_2 - t_1, \cdots, t_n - t_{n-1}\}$. Denote lower triangular matrix

$$\mathbf{L} = \left\{\mathbb{1}_{\{i \geq j\}}\right\}_{i,j=1}^{n} = \begin{pmatrix} 1 & 0 & \cdots & 0 & 0 \\ 1 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & 1 & \cdots & 1 & 0 \\ 1 & 1 & \cdots & 1 & 1 \end{pmatrix}, \tag{3.11}$$

$$\mathbf{LZ} \sim N(0, \mathbf{LDL}^\top) \ \Rightarrow\ \mathbf{W} \sim N(0, \mathbf{K}), \tag{3.12}$$

where $\mathbf{K} = \{\min(t_i, t_j)\}_{i,j=1}^{n}$. Hence $\mathbf{W}$ is Gaussian, and $W(t)$ is a Gaussian process. □

3.3.2 Ornstein–Uhlenbeck process (O-U process)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{R})$. A Ornstein–Uhlenbeck process $\{X(t) : \Omega \to \mathbb{R},\ t \geq 0\}$ is defined by the following stochastic differential equation (SDE):

$$\mathrm{d}X(t) = -\theta X(t)\mathrm{d}t + \sigma \mathrm{d}W(t), \tag{3.13}$$

where $\theta > 0, \sigma > 0$ and $W(t)$ is a standard Brownian motion. Let's derive the solution of the SDE above:

$$\begin{aligned} \mathrm{d}X(t) + \theta X(t)\mathrm{d}t &= \sigma \mathrm{d}W(t), \\ \mathrm{d}\left(X(t)\mathrm{e}^{\theta t}\right) &= \sigma \mathrm{e}^{\theta t}\mathrm{d}W(t), \\ X(t)\mathrm{e}^{\theta t} - X(0) &= \sigma \int_0^t \mathrm{e}^{\theta s}\mathrm{d}W(s). \end{aligned} \tag{3.14}$$

So analytical form of $X(t)$ is

$$X(t) = X(0)\mathrm{e}^{-\theta t} + \sigma \int_0^t \mathrm{e}^{-\theta(t-s)}\mathrm{d}W(s). \tag{3.15}$$

Now we are going to investigate the properties of Ornstein-Uhlenbeck process.

Proposition 3.4 Let $X(\cdot)$ be an Ornstein–Uhlenbeck process defined above. Conditional on $X(0)$, the mean and variance of $X(\cdot)$ is given by

$$\begin{aligned} \mathbb{E}[X(t)|X(0)] &= X(0)\mathrm{e}^{-\theta t}, \\ \mathrm{Cov}\left\{X(s), X(t)|X(0)\right\} &= \frac{\sigma^2}{2\theta}\left(\mathrm{e}^{\theta|t-s|} + \mathrm{e}^{-(s+t)}\right). \end{aligned} \tag{3.16}$$

Proof. The conditional expectation of $X(t)$ is its deterministic part, since the stochastic part is an integral of Brownian increments with zero mean. It remains to derive the covariace.

$$\mathrm{Cov}\left\{X(s), X(t)|X(0)\right\} = \sigma^2\mathrm{e}^{-(s+t)}\mathbb{E}\left[\int_0^s \mathrm{e}^{\theta u}\mathrm{d}W(u)\int_0^t \mathrm{e}^{\theta u}\mathrm{d}W(u)\right]. \tag{3.17}$$

By the independence of Brownian increments and the Itô isometry, we have

$$\begin{aligned} \mathbb{E}\left[\int_0^s \mathrm{e}^{\theta u}\mathrm{d}W(u)\int_0^t \mathrm{e}^{\theta u}\mathrm{d}W(u)\right] &= \mathbb{E}\left[\left(\int_0^{\min(s,t)} \mathrm{e}^{\theta u}\mathrm{d}W(u)\right)^2\right] \\ &= \int_0^{\min(s,t)} \mathrm{e}^{2\theta u}\mathrm{d}u = \frac{1}{2\theta}\left(\mathrm{e}^{2\theta\min(s,t)} - 1\right). \end{aligned} \tag{3.18}$$

Therefore

$$\mathrm{Cov}\left\{X(s), X(t)|X(0)\right\} = \frac{\sigma^2}{2\theta}\left(\mathrm{e}^{-\theta|t-s|} - \mathrm{e}^{-\theta(t+s)}\right), \tag{3.19}$$

and we complete the proof. □

Proposition 3.5 (Gaussianity). We impose a Gaussian distribution $N(0, \frac{\sigma^2}{2\theta})$ on $X(0)$ which is independent of $W(\cdot)$, then

$$\mathbb{E}[X(t)] = 0, \quad \mathrm{Cov}\left\{X(s), X(t)\right\} = \frac{\sigma^2}{2\theta}\mathrm{e}^{\theta|t-s|}, \tag{3.20}$$

which implies that an (unconditioned) Ornstein–Uhlenbeck process is a stationary Gaussian process with zero mean and Laplacian kernel.

Proof. Since $X(0)$ is independent of $W(\cdot)$, we only need to add the term $\mathrm{e}^{-\theta(s+t)}\mathrm{Var}(Z_0)$ to the conditional covariance. To show that the Ornstein–Uhlenbeck process is a Gaussian process, it remains to consider the

stochastic part of $X(t)$:

$$\int_0^t \mathrm{e}^{-\theta(t-s)}\mathrm{d}W(s) = \lim_{n\to\infty} \sum_{[t_{i-1},t_i]\in\pi_n} \mathrm{e}^{-\theta(t-s_{i-1})}\left(W(s_i) - W(s_{i-1})\right), \tag{3.21}$$

where $\{\pi_n\}$ is a sequence of partitions of $[0,t]$ with the length of maximum sub-interval going to zero. Since all Brownian increments $W(s_i) - W(s_{i-1})$ are Gaussian, the linear combination of finite Brownian increments is Gaussian. Furthermore, the Itô integral as the L^2-limit of a sequence of Gaussian variables is still Gaussian. Hence $X(\cdot)$ is a Gaussian process. □

Remark. A more general form of Ornstein–Uhlenbeck processes has an additional drift term μ:

$$\mathrm{d}X(t) = \theta(\mu - X(t))\mathrm{d}t + \sigma\mathrm{d}W(t). \tag{3.22}$$

The drift term does not change the form of covariance, but the conditional mean need to be modified:

$$\mathbb{E}[X(t)|X(0)] = X(0)\mathrm{e}^{-\theta t} + \mu(1 - \mathrm{e}^{-\theta t}). \tag{3.23}$$

4 Kriging

Nomenclature. Kriging is a method of interpolation based on Gaussian process, and is also known as Gaussian process regression (GPR). Georges Matheron established the theoretic basis of this spatial interpolation technique in 1960, and named it kriging (*French: krigeage*) to honor the pioneering work of Danie G. Krige in geostatistics.

Given a random field $\{Z(s) : s \in \mathcal{D}\}$ with its observation at location $s_1, \cdots, s_n$, the goal of kriging is to predict $Z(\cdot)$ at an arbitrary location $s \in \mathcal{D}$.

4.1 Simple Kriging

In simple kriging, the mean of spatial field is known. Without loss of generality, set it to be zero. Let $f : \mathcal{X} \to \mathbb{R}$ be the field of interest and $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$, we will interpolate the value of f at an arbitrary location using its finite observations. Please note that f is not a random field or a Gaussian process itself since it is deterministic. Instead, f is a realization of the aforementioned Gaussian process. In the context of Bayesian regression, $\mathcal{GP}(0, k(\cdot,\cdot))$ is the prior distribution of f.

We first introduce some notations. Given locations $x_1, \cdots, x_N \in \mathcal{X}$, we use $\mathbf{K}$ to denote the covariance matrix of $(f(x_1), \cdots, f(x_N))^\top$, i.e., $\mathbf{K} = \{k(x_i, x_j)\}_{i,j=1}^N \in \mathbb{R}^{n\times n}$. We define $\mathbf{k}(\cdot) : \mathcal{X} \to \mathbf{R}^N$ to be such that

$$\mathbf{k}(x) = (k(x, x_1), \cdots, k(x, x_N))^\top. \tag{4.1}$$

4.1.1 Noise-free observations

We suppose that our observations are faithful, that is, the observation at location x_i is the true value of $f(x_i)$. If we have observations at N distinct locations $x_1, \cdots, x_N \in \mathcal{X}$, then our dataset is $\{(x_i, f(x_i))\}_{i=1}^N$. Denote $\mathbf{f} = (f(x_1), \cdots, f(x_N))^\top$, the posterior of f at any location is given by the following theorem.

Theorem 4.1 (Posterior of f conditioning on noise-free observations). Under the notations above plus $\mathbf{K}$ is nonsingular, the posterior of f at an arbitrary location $x \in \mathcal{X}$ is Gaussian. The mean and variance is given by

$$\mathbb{E}[f(x)|\mathbf{f}] = \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{f}, \tag{4.2}$$

$$\mathrm{Var}\{f(x)|\mathbf{f}\} = k(x,x) - \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{k}(x). \tag{4.3}$$

Moreover, $\forall x, x' \in \mathcal{X}$,

$$\mathrm{Cov}\{f(x), f(x')|\mathbf{f}\} = k(x,x') - \mathbf{k}(x)^\top \mathbf{K}^{-1}\mathbf{k}(x'). \tag{4.4}$$

Proof. $\forall x, x' \in \mathcal{X}$, the joint distribution of $(f(x), f(x'), f(x_1), \cdots, f(x_n))^\top$ can be derived:

$$\begin{pmatrix} f(x) \\ f(x') \\ \mathbf{f} \end{pmatrix} \sim N\left\{ \mathbf{0}, \begin{pmatrix} k(x,x) & k(x,x') & \mathbf{k}(x)^\top \\ k(x',x) & k(x',x') & \mathbf{k}(x')^\top \\ \mathbf{k}(x) & \mathbf{k}(x') & \mathbf{K} \end{pmatrix} \right\}. \tag{4.5}$$

Then conditional on $\mathbf{f}$, the distribution of $(f(x), f(x'))^\top$ is still Gaussian:

$$\begin{pmatrix} f(x) \\ f(x') \end{pmatrix} \Bigg| \, \mathbf{f} \sim N\left\{ \begin{pmatrix} \mathbf{k}(x)^\top \\ \mathbf{k}(x')^\top \end{pmatrix} \mathbf{K}^{-1}\mathbf{f}, \begin{pmatrix} k(x,x) & k(x,x') \\ k(x',x) & k(x',x') \end{pmatrix} - \begin{pmatrix} \mathbf{k}(x)^\top \\ \mathbf{k}(x')^\top \end{pmatrix} \mathbf{K}^{-1} \left(\mathbf{k}(x), \mathbf{k}(x')\right) \right\}. \tag{4.6}$$

Thus we complete the proof. □

Remark. (I) When x is some $x_i \in \{x_1, \cdots, x_N\}$, the posterior of $f(x_i)$ reduces to the Dirac measure centered on $\mathbf{f}_i$. To see this, note that

$$\mathbf{e}_i = \mathbf{K}^{-1}\mathbf{K}\mathbf{e}_i = \mathbf{K}^{-1}\mathbf{k}(x_i), \tag{4.7}$$

then

$$\mathbb{E}\left[f(x_i)|\mathbf{f}\right] = \mathbf{e}_i\mathbf{f} = f(x_i),\ \mathrm{Var}\left\{f(x)|\mathbf{f}\right\} = k(x_i, x_i) - \mathbf{k}(x_i)^\top \mathbf{e}_i = 0. \tag{4.8}$$

(II) Define $\mu^\perp(x) = \mathbf{f}^\top\mathbf{K}^{-1}\mathbf{k}(x),\ k^\perp(x, x') = k(x, x') - \mathbf{k}(x)^\top\mathbf{K}^{-1}\mathbf{k}(x')$, the posterior of f is a Gaussian process:

$$f \,|\, \mathbf{f} \sim \mathcal{GP}\left(\mu^\perp(\cdot),\ k^\perp(\cdot,\cdot)\right). \tag{4.9}$$

To show this you need to verify that $k^\perp(\cdot,\cdot)$ is a valid kernel, i.e., $k^\perp(\cdot,\cdot)$ is positive definite. You may use Schur's complement to derive the form of $k^\perp(\cdot,\cdot)$.

4.1.2 Observations with additive noise

In a real scenario, the true process f is not accessible due to random fluctuation or measurement error. Like in hierarchical model, assume that there exists additive noise in our observational data:

$$y_i = f(x_i) + \varepsilon_i,\ \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \tag{4.10}$$

then the posterior of f at any location need to be modified.

Theorem 4.2 (Posterior of f conditioning on observations with additive noise). Under the notations above with observations $Y = (y_1, \cdots, y_N)^\top$, the posterior of f at an arbitrary location $x \in \mathcal{X}$ is Gaussian. The mean and variance is given by

$$\mathbb{E}\left[f(x) \,|\, Y\right] = \mathbf{k}(x)^\top \left(\mathbf{K} + \sigma^2\mathbf{I}_N\right)^{-1} Y, \tag{4.11}$$

$$\mathrm{Var}\left\{f(x) \,|\, Y\right\} = k(x, x) - \mathbf{k}(x)^\top \left(\mathbf{K} + \sigma^2\mathbf{I}_N\right)^{-1} \mathbf{k}(x). \tag{4.12}$$

Moreover, $\forall x, x' \in \mathcal{X}$,

$$\mathrm{Cov}\left\{f(x), f(x') \,|\, Y\right\} = k(x, x') - \mathbf{k}(x)^\top \left(\mathbf{K} + \sigma^2\mathbf{I}_N\right)^{-1} \mathbf{k}(x'). \tag{4.13}$$

Proof. Analogous to Theorem 1, use the joint distribution of $(f(x), f(x'), Y)$ to derive the conditional (posterior) distribution. □

Remark. The posterior of f is a Gaussian process with the mean and covariance function given by Theorem 4.2. A point estimator is the expectation:

$$\hat{f}(\cdot) = Y^\top(\mathbf{K} + \sigma^2\mathbf{I}_N)^{-1}\mathbf{k}(\cdot), \tag{4.14}$$

which is a linear combination of $\{k(\cdot, x_i)\}_{i=1}^N$.

4.1.3 Overlapping observations

In previous discussions, locations $x_1, \cdots, x_N$ are supposed to be distinct. In other words, we have only one observation for a unique location. Now we suppose that we have N observations at n unique locations

$\bar{x}_1, \cdots, \bar{x}_n$ with $n < N$. At each location $\bar{x}_j$, we have a_j observations $y_{j,1}, \cdots, y_{j,a_j}$, with $j = 1, \cdots, n,\ a_1 + \cdots + a_n = N$ and $\max_{1 \le j \le n} a_j > 1$.

We first introduce some notations.

$$(x_1, \cdots, x_N) = (\underbrace{\bar{x}_1, \cdots, \bar{x}_1}_{a_1}, \underbrace{\bar{x}_2, \cdots, \bar{x}_2}_{a_2}, \cdots, \underbrace{\bar{x}_n, \cdots, \bar{x}_n}_{a_n}) \in \mathcal{X}^N, \tag{4.15}$$

$$Y = (y_{1,1}, \cdots, y_{1,a_1}, y_{2,1}, \cdots, y_{2,a_2}, \cdots, y_{n,1}, \cdots, y_{n,a_n}) \in \mathbb{R}^N, \tag{4.16}$$

$$\bar{Y} = (\bar{y}_1, \cdots, \bar{y}_n) \in \mathbb{R}^n, \quad \bar{y}_j = \frac{1}{a_j} \sum_{l=1}^{a_j} y_{j,a_j}. \tag{4.17}$$

Moreover, denote

$$\bar{\mathbf{k}}(x) = (k(x, \bar{x}_1), \cdots, k(x, \bar{x}_n))^\top \in \mathbb{R}^n, \quad \bar{\mathbf{K}} = \{k(\bar{x}_i, \bar{x}_j)\}_{i,j=1}^n \in \mathbb{R}^{n \times n}. \tag{4.18}$$

Let $\mathbf{U} \in \mathbb{R}^{N \times n}$ be a block diagonal matrix such that $\mathbf{U} = \mathrm{diag}\{\mathbf{1}_{a_1}, \cdots, \mathbf{1}_{a_n}\}$, where $\mathbf{1}_{a_j}$ is an a_j-dimensional all-one vector, and let $\mathbf{A} = \mathrm{diag}\{a_1, \cdots, a_n\}$.

Proposition 4.3. With the notations above, we have

$$\bar{Y} = \mathbf{A}^{-1}\mathbf{U}^\top Y,\ \bar{\mathbf{k}}(x) = \mathbf{A}^{-1}\mathbf{U}^\top \mathbf{k}(x),\ \mathbf{k}(x) = \mathbf{U}\bar{\mathbf{k}}(x),\ \mathbf{K} = \mathbf{U}\bar{\mathbf{K}}\mathbf{U}^\top. \tag{4.19}$$

These equalities can be verified by direct calculation.

Proposition 4.4. Suppose $\sigma^2 > 0$, then $\mathbf{U}^\top \left(\sigma^2 \mathbf{I}_N + \mathbf{K}\right)^{-1} \mathbf{U} = \left(\sigma^2 \mathbf{A}^{-1} + \bar{\mathrm{K}}\right)^{-1}$.

Proof. Recall the Sherman-Morrison-Woodbury formula:

$$(\mathbf{A} + \mathbf{UCV})^{-1} = \mathbf{A}^{-1} - \mathbf{A}^{-1}\mathbf{U}\left(\mathbf{C}^{-1} + \mathrm{V}\mathbf{A}^{-1}\mathrm{U}\right)^{-1}\mathbf{V}\mathbf{A}^{-1}, \tag{4.20}$$

we have

$$\left(\sigma^2 \mathbf{I}_N + \mathbf{K}\right)^{-1} = \left(\sigma^2 \mathbf{I}_N + \mathbf{U}\bar{\mathbf{K}}\mathbf{U}^\top\right)^{-1} = \sigma^{-2}\mathbf{I}_N - \sigma^{-4}\mathrm{U}\left(\bar{\mathbf{K}}^{-1} + \sigma^{-2}\mathbf{U}^\top\mathbf{U}\right)^{-1}\mathbf{U}^\top. \tag{4.21}$$

Note that $\mathbf{U}^\top\mathbf{U} = \mathbf{A}$, then

$$\mathbf{U}^\top \left(\sigma^2 \mathbf{I}_N + \mathbf{K}\right)^{-1} \mathbf{U} = \sigma^{-2}\mathbf{A} - \sigma^{-4}\mathbf{A}\left(\bar{\mathbf{K}}^{-1} + \sigma^{-2}\mathbf{A}\right)^{-1}\mathbf{A}. \tag{4.22}$$

Using the Sherman-Morrison-Woodbury formula, we have

$$\sigma^{-2}\mathbf{A} + \sigma^{-4}\mathbf{A}\left(\bar{\mathbf{K}}^{-1} + \sigma^{-2}\mathbf{A}\right)^{-1}\mathbf{A} = \left(\sigma^2\mathbf{A}^{-1} + \bar{\mathbf{K}}\right)^{-1}, \tag{4.23}$$

which concludes the proof. □

Theorem 4.5 (Posterior of f conditioning on overlapping observations). Under the notations above, the posterior of f at an arbitrary location $x \in \mathcal{X}$ is Gaussian. The mean and variance is given by

$$\mathbb{E}\left[f(x) \,|\, \bar{Y}\right] = \bar{\mathbf{k}}(x)^\top \left(\bar{\mathbf{K}} + \sigma^2\mathbf{A}^{-1}\right)^{-1} \bar{Y}, \tag{4.24}$$

$$\mathrm{Var}\left\{f(x) \,|\, \bar{Y}\right\} = k(x,x) - \bar{\mathbf{k}}(x)^\top \left(\bar{\mathbf{K}} + \sigma^2\mathbf{A}^{-1}\right)^{-1} \bar{\mathbf{k}}(x). \tag{4.25}$$

Moreover, $\forall x, x' \in \mathcal{X}$,

$$\mathrm{Cov}\left\{f(x), f(x') \,|\, \bar{Y}\right\} = k(x,x') - \bar{\mathbf{k}}(x)^\top \left(\bar{\mathbf{K}} + \sigma^2 \mathbf{A}^{-1}\right)^{-1} \bar{\mathbf{k}}(x'). \tag{4.26}$$

Proof. Plug in Proposition 1 and 2 to the posterior of f given in Theorem 2. □

4.2 Universal Kriging

In universal kriging, the mean of spatial field is unknown. To apply regression, we choose a group of basis functions $\{m_1(\cdot), \cdots, m_p(\cdot)\}$ from $L_2(\mathcal{X})$ and denote $\mathbf{m}(\cdot) = \left(m_1(\cdot), \cdots, m_p(\cdot)\right)^\top : \mathcal{X} \to \mathbb{R}^p$. Assume that $f \sim \mathcal{GP}\left(\mathbf{m}(\cdot)^\top \beta,\ k(\cdot,\cdot)\right)$, where $\beta \in \mathbb{R}^p$ is unknown, and that our observations $Y(x)|\, f \sim N(f(x), \sigma^2)$. Note these assumptions implies the independence of $\{Y(x)|\, x \in \mathcal{X}\}$ given f. Then, we can define the error process

$$\epsilon(x) := Y(x) - \mathbf{m}(x)^\top \beta, \tag{4.27}$$

which implies the following distribution of observations at $x_1, \cdots, x_n$:

$$\epsilon(x_{1:n}) \sim N(0, \Sigma), \quad \Sigma = \mathbf{K} + \sigma^2 \mathbf{I}_n, \quad \mathbf{K} = \{k(x_i, x_j)\}_{i,j=1}^n. \tag{4.28}$$

Furthermore, the joint distribution of our observations can be induced:

$$\mathbf{Y} := \left(Y(x_1), \cdots, Y(x_n)\right)^\top \sim N\left(\mathbf{M}\beta, \Sigma\right), \quad \mathbf{M} = \left(\mathbf{m}(x_1), \cdots, \mathbf{m}(x_n)\right)^\top \in \mathbb{R}^{n \times p}. \tag{4.29}$$

Now we derive the best unbiased linear prediction (BLUP) of $Y(x)$ for arbitrary $x \in \mathcal{X}$. Generally, we have the following result.

Theorem 4.6 (Universal kriging). Suppose the covariance matrix of $\mathbf{Y} = (Y(x_1), \cdots, Y(x_n))$ is nonsingular, and $\mathbf{M} = \left(\mathbf{m}(x_1), \cdots, \mathbf{m}(x_n)\right)^\top$ is of full column rank. Then for $x \in \mathcal{X}$, the BLUP of $Y(x)$ is given by

$$\hat{Y}(x) = \mathbf{m}(x)^\top \hat{\beta} + \mathbf{k}(x)^\top \Sigma^{-1} \left(\mathbf{Y} - \mathbf{M}\hat{\beta}\right), \tag{4.30}$$

where $\hat{\beta} = \left(\mathbf{M}^\top \Sigma^{-1} \mathbf{M}\right)^{-1} \mathbf{M}^\top \Sigma^{-1} \mathbf{Y}$.

Proof. The derivation of BLUP is straight forward but tedious. We first express the prediction of $Y(x)$ as a linear combination:

$$\hat{Y}(x) = \lambda_0(x) + \sum_{i=1}^n \lambda_i(x) Y(x_i) = \lambda_0(x) + \lambda(x)^\top \mathbf{Y}. \tag{4.31}$$

The unbiasedness of BLUP implies that

$$\mathbf{m}(x)^\top \beta = \mathbb{E}\left[\hat{Y}(x)\right] = \lambda_0(x) + \lambda(x)^\top \mathbf{M}\beta \quad \forall \beta \in \mathbb{R}^p. \tag{4.32}$$

Set $\beta = 0$, we have $\lambda_0(x) = 0$. To solve the BLUP, which has the minimum variance across all predictions of the form above, one need to consider the following optimization problem:

$$\min_{\lambda(x) \in \mathbb{R}^p} \quad \mathbb{E}\left[\left(Y(x) - \lambda(x)^\top \mathbf{Y}\right)^2\right] \quad \text{subjected to} \quad \mathbf{M}^\top \lambda(x) = \mathbf{m}(x). \tag{4.33}$$

The mean square error of BLUP is

$$\mathbb{E}\left[\left(Y(x)-\lambda(x)^\top \mathbf{Y}\right)^2\right] = k(x,x) - 2\mathbf{k}(x)^\top\lambda(x) + \lambda(x)^\top\Sigma\lambda(x), \tag{4.34}$$

then we can construct the Lagrangian function as:

$$L(\lambda(x),\nu(x)) = \lambda(x)^\top\Sigma\lambda(x) - 2\mathbf{k}(x)^\top\lambda(x) + 2\nu(x)^\top\left(\mathbf{m}(x)-\mathbf{M}^\top\lambda(x)\right). \tag{4.35}$$

Apply Karush-Kuhn-Tucker conditions, the optimal solutions of primal and dual problems satisfy:

$$\begin{cases} \frac{\partial L}{\partial \lambda^*(x)} = 2\Sigma\lambda^*(x) - 2\mathbf{k}(x) - 2\mathbf{M}\nu^*(x) = 0 \\ \frac{\partial L}{\partial \nu^*(x)} = \mathbf{M}^\top\lambda^*(x) - \mathbf{m}(x) = 0, \end{cases} \tag{4.36}$$

and it can be solved that

$$\begin{cases} \nu^*(x) = \left(\mathbf{M}^\top\Sigma^{-1}\mathbf{M}\right)^{-1}\left(\mathbf{m}(x) - \mathbf{M}^\top\Sigma^{-1}\mathbf{k}(x)\right) \\ \lambda^*(x) = \Sigma^{-1}\mathbf{k}(x) + \Sigma^{-1}\mathbf{M}\left(\mathbf{M}^\top\Sigma^{-1}\mathbf{M}\right)^{-1}\left(\mathbf{m}(x) - \mathbf{M}^\top\Sigma^{-1}\mathbf{k}(x)\right). \end{cases} \tag{4.37}$$

Let $\hat{\beta} = \left(\mathbf{M}^\top\Sigma^{-1}\mathbf{M}\right)^{-1}\mathbf{M}^\top\Sigma^{-1}\mathbf{Y}$, then

$$\hat{Y}(x) = \lambda^*(x)^\top\mathbf{Y} = \mathbf{m}(x)^\top\hat{\beta} + \mathbf{k}(x)^\top\Sigma^{-1}\left(\mathbf{Y}-\mathbf{M}\hat{\beta}\right). \tag{4.38}$$

Thus we complete the proof. □

Remark. The BLUP satisfies the following properties:

- $\mathbb{E}\hat{\beta} = \beta,\ \mathrm{Cov}(\hat{\beta}) = \left(\mathbf{M}^\top\Sigma^{-1}\mathbf{M}\right)^{-1}$.
- For any $x, x' \in \mathcal{X}$,

$$\mathbb{E}\hat{Y}(x) = \mathbf{m}(x)^\top\beta, \tag{4.39}$$

$$\mathrm{Cov}\left\{\hat{Y}(x),\hat{Y}(x')\right\} = \left(\mathbf{m}(x)^\top - \mathbf{k}(x)^\top\Sigma^{-1}\mathbf{M}\right)\left(\mathbf{M}^\top\Sigma^{-1}\mathbf{M}\right)^{-1}\left(\mathbf{m}(x) - \mathbf{M}^\top\Sigma^{-1}\mathbf{k}(x)\right). \tag{4.40}$$

4.3 Ordinary Kriging

In ordinary kriging, the spatial field is supposed to have a constant mean, which is unknown. Ordinary kriging can be handled as a special case of universal kriging where the regressor $\mathbf{m}(\cdot) \equiv 1$ is one-dimensional.

Suppose the mean of our spatial field of interest is β. Inherited from the discussion in universal kriging, the BLUP of process $Y(\cdot)$ at an arbitrary location $x \in \mathcal{X}$ is given by

$$\hat{Y}(x) = \hat{\beta} + \mathbf{k}(x)^\top\Sigma^{-1}(\mathbf{Y} - \beta\mathbf{1}_n),\ \text{where}\ \ \hat{\beta} = \frac{\mathbf{1}_n^\top\Sigma^{-1}\mathbf{Y}}{\mathbf{1}_n^\top\Sigma^{-1}\mathbf{1}_n}. \tag{4.41}$$

5 Reproducing Kernel Hilbert Space

5.1 Definition and Properties

Definition 5.1 (Reproducing kernel Hilbert space, RKHS). Let $\mathcal{X}$ be a non-empty set and $\mathcal{H}$ be a Hilbert space of real-valued functions defined on $\mathcal{X}$ and equipped with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$. A symmetric and positive definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a reproducing kernel, and $\mathcal{H}$ is a reproducing kernel Hilbert space, if k satisfies the following two conditions:

- $\forall x \in \mathcal{X},\ k(\cdot, x) \in \mathcal{H}$, and
- (Reproducing property) $\forall x \in \mathcal{X},\ \forall f \in \mathcal{H},\ f(x) = \langle f, k(\cdot, x) \rangle_{\mathcal{H}}$.

By definition, for any $x, y \in \mathcal{X}$, it holds $k(x, y) = \langle k(\cdot, x), k(\cdot, y) \rangle_{\mathcal{H}}$.

Proposition 5.2 Define δ_x to be the evaluation functional at x, i.e.,

$$f(x) = \delta_x f = \int_{\mathcal{X}} \delta_x(\mathrm{d}t) f(t), \tag{5.2}$$

then $\forall x \in \mathcal{X}$, δ_x is a bounded operator on $\mathcal{H}$, i.e., there exists a corresponding $\lambda_x \geq 0$ such that

$$|f(x)| = |\delta_x f| \leq \lambda_x \|f\|_{\mathcal{H}}. \tag{5.3}$$

Proof. Given a Hilbert space $\mathcal{H}$ with reproducing kernel $k(\cdot, \cdot)$, it holds

$$|\delta_x[f]| = |f(x)| = |\langle f, k(\cdot, x) \rangle_{\mathcal{H}}| \leq \|k(\cdot, x)\|_{\mathcal{H}} \|f\|_{\mathcal{H}} = \sqrt{k(x, x)} \cdot \|f\|_{\mathcal{H}}. \tag{5.4}$$

Hence $\delta_x : \mathcal{H} \to \mathbb{R}$ is a bounded operator with $\lambda_x = \sqrt{k(x, x)}$. □

Remark. An equivalent definition of RKHS is stated as follows: $\mathcal{H}$ is a RKHS if the evaluation functional δ_x is bounded for all $x \in \mathcal{X}$. We have proven that the first definition implies the second. The proof of the other direction uses the Riesz representation theorem.

5.2 Construction of RKHS

5.2.1 From Pre-Hilbert Space to its Completion: Moore-Aronszajn Theorem

Theorem 5.3 (Moore-Aronszajn). RKHS and positive definite kernel are one-to-one correspondent, i.e., for each positive definite kernel $k(\cdot, \cdot)$, there exists a unique RKHS with $k(\cdot, \cdot)$ as its reproducing kernel.

Proof. Let $\mathcal{X}$ be a non-empty set and $k(\cdot, \cdot)$ be a positive definite kernel on $\mathcal{X} \times \mathcal{X}$. Define

$$\mathcal{H}_0 = \operatorname{span}\{k(\cdot, x) : x \in \mathcal{X}\} = \left\{ \sum_{i=1}^{n} c_i k(\cdot, x_i) : n \in \mathbb{N},\ c_1, \cdots, c_n \in \mathbb{R},\ x_1, \cdots, x_n \in \mathcal{X} \right\}, \tag{5.5}$$

then $\mathcal{H}_0$ is a pre-Hilbert space with inner product

$$\langle f, g \rangle_{\mathcal{H}_0} = \sum_{i=1}^{m} \sum_{j=1}^{n} a_i b_j k(x_i, y_j), \tag{5.6}$$

where $f, g \in \mathcal{H}_0$ have representations

$$f = \sum_{i=1}^{m} a_i k(\cdot, x_i),\ g = \sum_{j=1}^{n} b_j k(\cdot, y_j),\ x_1, \cdots, x_m, y_1, \cdots, y_n \in \mathcal{X}. \tag{5.7}$$

We need to show that $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ is indeed an inner product. It is easy to verify that $\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$ is a semi-inner product. Then it remains to show that $\langle f,f\rangle_{\mathcal{H}_0}=0$ only if $f=0$. Set $\|f\|_{\mathcal{H}_0}=0$, and fix $g\in\mathcal{H}_0$. Then $\forall t\in\mathbb{R}$,

$$0\le\|f-tg\|_{\mathcal{H}_0}^2=-2t\langle f,g\rangle_{\mathcal{H}_0}+t^2\langle g,g\rangle_{\mathcal{H}_0}. \tag{5.8}$$

If $|\langle f,g\rangle_{\mathcal{H}_0}|>0$, then set $t=\frac{\langle f,g\rangle_{\mathcal{H}_0}}{\langle g,g\rangle_{\mathcal{H}_0}}$ yields a contradiction of (5.8). Since $g\in\mathcal{H}_0$ is arbitrarily chosen, we have $\langle f,g\rangle_{\mathcal{H}_0}=0\ \forall g\in\mathcal{H}_0$. Then for every $x\in\mathcal{X}$, $f(x)=\langle f,k(\cdot,x)\rangle_{\mathcal{H}_0}=0$. Hence $f=0$.

Let $\mathcal{H}$ be the completion of $\mathcal{H}_0$ under $\|\cdot\|_{\mathcal{H}_0}=\langle\cdot,\cdot\rangle_{\mathcal{H}_0}$, so $\mathcal{H}$ contains all equivalence classes of Cauchy sequences in $\mathcal{H}_0$. Then $\mathcal{H}$ has all functions $f\in\mathbb{R}^{\mathcal{X}}$ of the form

$$f=\sum_{i=1}^{\infty}c_ik(\cdot,x_i),\ \text{and}\ \left\|\sum_{i=n+1}^{n+m}c_ik(\cdot,x_i)\right\|_{\mathcal{H}_0}\to 0\ \text{as}\ n,m\to\infty. \tag{5.9}$$

We claim that $\mathcal{H}$ is a RKHS with $k(\cdot,\cdot)$ as its reproducing kernel. To show this it suffices to check the reproducing property: $\forall x\in\mathcal{X}$,

$$\langle f,k(\cdot,x)\rangle_{\mathcal{H}}=\left\langle\sum_{i=1}^{\infty}c_ik(\cdot,x_i),k(\cdot,x)\right\rangle_{\mathcal{H}}=\sum_{i=1}^{\infty}c_i\left\langle k(\cdot,x_i),k(\cdot,x)\right\rangle_{\mathcal{H}_0}=\sum_{i=1}^{\infty}c_ik(x,x_i)=f(x). \tag{5.10}$$

Now we prove that $\mathcal{H}$ is unique. Suppose that $\tilde{\mathcal{H}}$ is another RKHS with $k(\cdot,\cdot)$ as its reproducing kernel. Recall the properties of RKHS, $\mathcal{H}_0\subset\tilde{\mathcal{H}}$, and $\langle\cdot,\cdot\rangle_{\mathcal{H}}=\langle\cdot,\cdot\rangle_{\tilde{\mathcal{H}}}$ on $\mathcal{H}_0$. Since $\tilde{\mathcal{H}}$ is complete, it contains the completion of $\mathcal{H}_0$, i.e. $\mathcal{H}\subseteq\tilde{\mathcal{H}}$. it remains to prove that $\mathcal{H}\supseteq\tilde{\mathcal{H}}$. Let $\tilde{\mathcal{H}}=\mathcal{H}\oplus\mathcal{H}^{\perp}$, $\forall f\in\tilde{\mathcal{H}}$, it can be decomposed as $f=f^*+f^{\perp}$, where $f^*\in\mathcal{H}, f^{\perp}\in\mathcal{H}^{\perp}$, then $\forall x\in\mathcal{X}$,

$$f(x)=\langle f,k(\cdot,x)\rangle_{\tilde{\mathcal{H}}}=\langle f^*,k(\cdot,x)\rangle_{\tilde{\mathcal{H}}}+\langle f^{\perp},k(\cdot,x)\rangle_{\tilde{\mathcal{H}}}=\langle f^*,k(\cdot,x)\rangle_{\tilde{\mathcal{H}}}=\langle f^*,k(\cdot,x)\rangle_{\mathcal{H}}=f^*(x), \tag{5.11}$$

which implies $f^{\perp}=0$ on $\mathcal{X}$. Hence $f=f^*\in\mathcal{H}$, which concludes the proof. □

5.2.2 Eigenanalysis: Mercer's Theorem

Assumption 5.4 (Regular conditions). In the following discussion, we fix the domain $\mathcal{X}$ as a compact set, and we make the following five assumptions of our kernel function $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$.

- Continuous: k is continuous on $\mathcal{X}\times\mathcal{X}$;
- Symmetric: $\forall x,y\in\mathcal{X},\ k(x,y)=k(y,x)$;
- Bounded: $\sup_{x\in\mathcal{X}}|k(x,x)|<\infty$;
- Square integrable: $\int_{\mathcal{X}}\int_{\mathcal{X}}|k(x,y)|^2\mathrm{d}x\mathrm{d}y<\infty$;
- Positive-definite: $\forall f\in L_2(\mathcal{X}),\ \int_{\mathcal{X}}\int_{\mathcal{X}}f(x)k(x,y)f(y)\mathrm{d}x\mathrm{d}y\ge 0$.

Definition 5.5 (Fredholm integral operator). Fix a kernel $k:\mathcal{X}\times\mathcal{X}\to\mathbb{R}$ on a compact set $\mathcal{X}\subset\mathbb{R}^d$, the operator $T_k:L_2(\mathcal{X})\to\mathcal{X}$ is defined as

$$(T_kf)(x):=\int_{\mathcal{X}}k(x,y)f(y)\mathrm{d}y,\quad f\in L_2(\mathcal{X}). \tag{5.12}$$

T_k is said to be positive definite if $\forall f\in L_2(\mathcal{X})$, it holds $\langle f,T_kf\rangle_{L_2(\mathcal{X})}\ge 0$, that is,

$$\int_{\mathcal{X}}\int_{\mathcal{X}}f(x)k(x,y)f(y)\mathrm{d}x\mathrm{d}y\ge 0. \tag{5.13}$$

Proposition 5.6 The integral operator given by the definition above have following properties:

- (Linear) $T_k(\alpha f + \beta g) = \alpha T_k f + \beta T_k g,\ \alpha, \beta \in \mathbb{R},\ f, g \in L_2(\mathcal{X})$;
- (Bounded) By Cauchy-Schwarz inequality,

$$|T_k f(x)|^2 \leq \left(\int_{\mathcal{X}} k(x,y)^2 \mathrm{d}y\right)\left(\int_{\mathcal{X}} f(y)^2 \mathrm{d}y\right), \tag{5.14}$$

  furthermore,

$$\|T_k f\|_{L_2(\mathcal{X})} \leq C_k \|f\|^2_{L_2(\mathcal{X})},\ C_k = \left(\int_{\mathcal{X}}\int_{\mathcal{X}} k(x,y)^2 \mathrm{d}x\mathrm{d}y\right)^{1/2}. \tag{5.15}$$

- (Symmetric/Self-adjoint).

$$\langle T_k f, g\rangle_{L_2(\mathcal{X})} = \langle f, T_k g\rangle_{L_2(\mathcal{X})},\ f, g \in L_2(\mathcal{X}). \tag{5.16}$$

- (Eigendecomposition) In functional analysis it is shown that T_k is a self-adjoint compact operator. Riesz-Schauder theorem gives the eigendecomposition of compact and self-adjoint operator T_k:
  (i) T_k has at most countably many eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ such that $\lim_{n\to\infty} \lambda_n = 0$;
  (ii) The corresponding eigenfunctions $\{\phi_n\}$, satisfying $T_k\phi_n = \lambda_n \phi_n$, form an orthogonal basis in $L_2(\mathcal{X})$, i.e. $\int_{\mathcal{X}} \phi_i(x)\phi_j(x)\mathrm{d}x = \delta_{ij},\ i, j \in \mathbb{N}^*$.

Proposition 5.7. For every $f \in L_2(\mathcal{X})$, it has expansion

$$f(x) = \sum_{n=1}^{\infty} \langle f, \phi_n\rangle_{L_2(\mathcal{X})}\, \phi_n(x), \tag{5.17}$$

and the convergence holds in L_2 sense.

Proof. Since $\{\phi_n\}$ form an orthogonal basis in $L_2(\mathcal{X})$, f has some representation of the form

$$f = \sum_{n=1}^{\infty} f_n \phi_n,\ \{f_n\} \subset \mathbb{R}. \tag{5.18}$$

By orthogonality, it can be calculated that

$$f_n = \langle f, \phi_n\rangle_{L_2(\mathcal{X})} = \int_{\mathcal{X}} f(x)\phi_n(x)\mathrm{d}x,\quad \|f\|^2_{L_2(\mathcal{X})} = \sum_{n=1}^{\infty} \langle f, \phi_n\rangle^2_{L_2(\mathcal{X})} = \sum_{n=1}^{\infty} f_n^2 < \infty, \tag{5.19}$$

then we have

$$\begin{aligned}\left\| f - \sum_{j=1}^{n} f_j\phi_j \right\|^2_{L_2(\mathcal{X})} &= \|f\|^2_{L_2(\mathcal{X})} - 2\sum_{j=1}^{n} f_j \langle f, \phi_j\rangle_{L_2(\mathcal{X})} + \left\|\sum_{j=1}^{n} f_j\phi_j\right\|^2_{L_2(\mathcal{X})} \\ &= \|f\|^2_{L_2(\mathcal{X})} - \sum_{j=1}^{n} f_j^2 \overset{n\to\infty}{\longrightarrow} 0.\end{aligned} \tag{5.20}$$

Thus we conclude the proof. □

Theorem 5.8 (Mercer). Under regular conditions, $k(\cdot,\cdot)$ admits the following spectral representation:

$$k(x,y)=\sum_{l=1}^{\infty}\lambda_l\phi_l(x)\phi_l(y), \tag{5.21}$$

Proof. By definition of the integral operator,

$$\langle k(\cdot,y),\phi_l\rangle_{L_2(\mathcal{X})}=(T_k\phi_l)(y)=\lambda_l\phi_l(y). \tag{5.22}$$

Proposition 5.7 implies that

$$k(x,y)=\sum_{l=1}^{\infty}\langle k(\cdot,y),\phi_l\rangle_{L_2(\mathcal{X})}\,\phi_l(x)=\sum_{l=1}^{\infty}\lambda_l\phi_l(x)\phi_l(y). \tag{5.23}$$

Thus we conclude the proof. □

Remark. Further study shows that this converge is absolute and uniform, i.e.,

$$\lim_{n\to\infty}\sup_{x,y\in\mathcal{X}}\left|k(x,y)-\sum_{l=1}^{n}\lambda_l\phi_l(x)\phi_l(y)\right|=0. \tag{5.24}$$

Corollary 5.9 (Trace of Functions). Under regular conditions, the trace of kernel k can be calculated by

$$\int_{\mathcal{X}}k(x,x)\mathrm{d}x=\sum_{l=1}^{\infty}\lambda_l,\ \int_{\mathcal{X}}\int_{\mathcal{X}}k(x,y)^2\mathrm{d}x\mathrm{d}y=\int_{\mathcal{X}}\left(\int_{\mathcal{X}}k(y,x)k(x,y)\mathrm{d}x\right)\mathrm{d}y=\sum_{l=1}^{\infty}\lambda_l^2. \tag{5.25}$$

More generally, extend the matrix multiplication to functions:

$$k^{(1)}(x,y)=k(x,y),\ k^{(n)}(x,y)=\int_{\mathcal{X}}k^{(n-1)}(x,z)k(z,y)\mathrm{d}z,\ n\geq 2, \tag{5.26}$$

then we have

$$k^{(n)}(x,y)=\sum_{l=1}^{\infty}\lambda^n\phi_l(x)\phi_l(y),\quad\int_{\mathcal{X}}k^{(n)}(x,x)\mathrm{d}x=\sum_{l=1}^{\infty}\lambda_l^n,\quad n\geq 1. \tag{5.27}$$

Theorem 5.10 (An alternative construction of RKHS). Let $\mathcal{H}$ to be a Hilbert space defined as

$$\mathcal{H}=\left\{f=\sum_{l=1}^{\infty}f_l\phi_l:\ \|f\|_{\mathcal{H}}^2=\sum_{l=1}^{\infty}\frac{f_l^2}{\lambda_l}<\infty,\ \{f_l\}\subset\mathbb{R}\right\}, \tag{5.28}$$

with the inner product defined as

$$\langle f,g\rangle_{\mathcal{H}}=\sum_{l=1}^{\infty}\frac{f_lg_l}{\lambda_l},\quad\text{where } f,g\in\mathcal{H}\text{ have representations } f=\sum_{l=1}^{\infty}f_l\phi_l,\ g=\sum_{l=1}^{\infty}g_l\phi_l, \tag{5.29}$$

then $\mathcal{H}$ is the RKHS associated with reproducing kernel $k(\cdot,\cdot)$.

Proof. Recall that $k(\cdot,x)=\sum_{l\in\mathbb{N}^*}\lambda_l\phi_l(\cdot)\phi_l(x)$,

$$\|k(\cdot,x)\|_{\mathcal{H}}^2=\sum_{l=1}^{\infty}\frac{\lambda_l^2\phi_l(x)^2}{\lambda_l}=\sum_{l=1}^{\infty}\lambda_l\phi_l(x)^2=k(x,x)\leq\sup_{x\in\mathcal{X}}k(x,x)<\infty, \tag{5.30}$$

hence $k(\cdot, x) \in \mathcal{H},\ \forall x \in \mathcal{X}$.

Now we prove the reproducing property. For an arbitrary $f = \sum_{l \in \mathbb{N}^*} f_l \phi_l \in \mathcal{H}$,

$$\langle f(\cdot), k(\cdot, x) \rangle_{\mathcal{H}} = \left\langle \sum_{l=1}^{\infty} f_l \phi_l(\cdot), \sum_{l=1}^{\infty} \lambda_l \phi_l(x) \phi_l(\cdot) \right\rangle_{\mathcal{H}} = \sum_{l=1}^{\infty} \frac{f_l \lambda_l \phi_l(x)}{\lambda_l} = \sum_{l=1}^{\infty} f_l \phi_l(x) = f(x). \tag{5.31}$$

Therefore $\mathcal{H}$ is a reproducing kernel Hilbert space with reproducing kernel $k(\cdot, \cdot)$. □

Remarks. (I) To see into the structure of the inner product defined above, let's derive it under the original definition of RKHS. Using the reproducing properties, $\forall l \in \mathbb{N}^*$,

$$\phi_l(x) = \left\langle \phi_l(\cdot), \sum_{l'=1}^{\infty} \lambda_{l'} \phi_{l'}(\cdot) \phi_{l'}(x) \right\rangle = \sum_{l'=1}^{\infty} \lambda_{l'} \phi_{l'}(x) \langle \phi_l(\cdot), \phi_{l'}(\cdot) \rangle_{\mathcal{H}},\ \forall x \in \mathcal{X}, \tag{5.32}$$

which implies

$$\langle \phi_l(\cdot), \phi_{l'}(\cdot) \rangle_{\mathcal{H}} = \frac{\delta_{ll'}}{\lambda_{l'}} = \frac{\delta_{ll'}}{\lambda_l}. \tag{5.33}$$

It is seen that the orthogonality of eigenfunctions is preserved in RKHS.

(II) The RKHS norm and L_2-norm are not equivalent. Note that $\lim_{n \to \infty} \lambda_n = 0$, we have

$$\sup_{f \in \mathcal{H}} \frac{\|f\|_{\mathcal{H}}}{\|f\|_{L_2(\mathcal{X})}} \geq \lim_{n \to \infty} \frac{\|\phi_n\|_{\mathcal{H}}}{\|\phi_n\|_{L_2(\mathcal{X})}} = \infty. \tag{5.34}$$

One interpretation of this non-equivalence is that the RKHS norm measures not only the magnitude of a function but also its smoothness.

6 Kernel Ridge Regression

6.1 Nonparametric Regression

6.1.1 Penalized least squares

Suppose that we have a dataset $\{(x_i, y_i)\}_{i=1}^n$ which contains n observations of a covariate $x \in \mathcal{X}$ and a response $y \in \mathbb{R}$. A nonparametric regression model assumes that

$$y_i = f(x_i) + \varepsilon_i,\ i = 1, \cdots, n, \tag{6.1}$$

where $\{\varepsilon_i\}_{i=1}^n$ are iid error terms with mean zero.

Assume that $\mathcal{H}$ is a RKHS with symmetric and positive definite $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ as its reproducing kernel, and $f \in \mathcal{H}$. We estimate the optimal f^* as the solution to the penalized least squares:

$$f^* \in \underset{f \in \mathcal{H}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^n (y_i - f(x_i))^2 + J(f), \tag{6.2}$$

where $J : \mathcal{H} \to \mathbb{R}$ is some differentiable regularization function. The first term (mean square error) measures the goodness-of-fit, and the second penalizes on the complexity of fitted function. Note that $\mathcal{H}$ is usually infinite dimensional, we can find some f which perfectly matches each data point (x_i, y_i). However, such f often suffers from bad generalizability, which is the reason that we add a penalty term to the objective function, like in ridge regression.

6.1.2 Generalized loss

We continue to discuss the nonparametric regression model given by (6.1). Suppose that f is realizable in a RKHS $\mathcal{H}$ with reproducing kernel $k(\cdot,\cdot)$. In the RKHS framework, let the regularization term be $R(f) = g(\|f\|_{\mathcal{H}})$, where $g : [0, \infty) \to \mathbb{R}$ is some strictly monotonically increasing function.

In penalized least squares, we use mean square error as the criterion to evaluate the goodness-of-fit. This can be relaxed by defining an arbitrary loss function $L : (\mathcal{X} \times \mathbb{R}^2)^n \to \mathbb{R} \cup \{\infty\}$. Furthermore, we define the optimal f^* as the minimizer of the regularized empirical risk functional:

$$f^* \in \underset{f \in \mathcal{H}}{\operatorname{argmin}}\ L\left\{(x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n))\right\} + g(\|f\|_{\mathcal{H}}). \tag{6.3}$$

Here L can be the mean square error:

$$L\left\{(x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n))\right\} = \frac{1}{n} \sum_{i=1}^n \{y_i - f(x_i)\}^2, \tag{6.4}$$

Another example of the loss function is the hinge loss:

$$L\left\{(x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n))\right\} = \frac{1}{n} \sum_{i=1}^n \max\{1 - y_i f(x_i), 0\}, \tag{6.5}$$

which is commonly used in support vector machine (SVM).

Recall the Mercer's theorem, if $k(\cdot,\cdot)$ has the expansion $k(x,y) = \sum_{l\in\mathbb{N}^*} \lambda_l \phi_l(x)\phi_l(y)$, then f has the following representation

$$f(\cdot) = \sum_{l=1}^{\infty} f_l \phi_l(\cdot),\ f_l \langle f, \phi_l \rangle_{\mathcal{H}} \tag{6.6}$$

then we rewrite the minimization problem as

$$\min_{\{f_l\}_{l\geq 1}} L\left\{\left(x_1, y_1, \sum_{l=1}^{\infty} f_l\phi(x_1)\right), \cdots, \left(x_n, y_n, \sum_{l=1}^{\infty} f_l\phi(x_n)\right)\right\} + g\left(\sqrt{\sum_{l=1}^{\infty} \frac{f_l^2}{\lambda_l}}\right). \tag{6.7}$$

As we can see, the optimization problem has an infinite dimensional form, which can be very knotty. The solution is given by the following representer theorem.

6.2 Representer Theorem

Theorem 6.1 (Schölkopf-Herbrich-Smola, [4]). Suppose we are given a nonempty set $\mathcal{X}$, an RKHS $\mathcal{H}$ with reproducing kernel $k : \mathcal{X}\times\mathcal{X} \to \mathbb{R}$, a finite training dataset $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X}\times\mathbb{R}$, a strictly monotonically increasing function $g : [0,\infty) \to \mathbb{R}$ and an arbitrary loss function $L : (\mathcal{X}\times\mathbb{R}^2)^n \to \mathbb{R}\cup\{\infty\}$. Then any minimizer $f^* \in \mathcal{H}$ of the regularized empirical risk functional

$$\mathcal{R}[f] := L\left\{(x_1, y_1, f(x_1)), \cdots, (x_n, y_n, f(x_n))\right\} + g(\|f\|_{\mathcal{H}}) \tag{6.8}$$

admits a representation of the form

$$f^*(\cdot) = \sum_{i=1}^{n} \alpha_i k(\cdot, x_i), \quad \alpha_1, \cdots, \alpha_n \in \mathbb{R}. \tag{6.9}$$

Proof. Define the subspace of $\mathcal{H}$:

$$\mathcal{H}_1 = \operatorname{span}\{k(\cdot, x_i) : i = 1, \cdots, n\} := \left\{\sum_{i=1}^{n} \alpha_i k(\cdot, x_i) : \alpha_1, \cdots, \alpha_n \in \mathbb{R}\right\}, \tag{6.10}$$

and let $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$. Then, any $f \in \mathcal{H}$ can be decomposed to two orthogonal components:

$$f = \widehat{f} + \delta,\ \widehat{f} \in \mathcal{H}_1,\ \delta \in \mathcal{H}_2. \tag{6.11}$$

Since δ is orthogonal to $\mathcal{H}_1$, and using the reproducing property of $\mathcal{H}$, we have

$$\delta(x_i) = \langle \delta(\cdot), k(\cdot, x_i)\rangle_{\mathcal{H}} = 0, \quad i = 1, \cdots, n. \tag{6.12}$$

Hence $f(x_i) = \widehat{f}(x_i)$ for $i = 1, \cdots, n$, and the first term of $\mathcal{R}$ does not change between f and $\widehat{f}$.

For the second term, it implies by orthogonality that

$$g(\|f\|_{\mathcal{H}}) = g\left(\sqrt{\|\widehat{f}\|_{\mathcal{H}}^2 + \|\delta\|_{\mathcal{H}}^2}\right) \geq g\left(\|\widehat{f}\|_{\mathcal{H}}\right). \tag{6.13}$$

Hence $\mathcal{R}[f] \geq \mathcal{R}[\widehat{f}]$, and the equality holds if and only if $\delta = 0$. Suppose f^* is a minimizer of $\mathcal{R}$ with expansion $f^* = \widehat{f^*} + \delta^*$. Then we simultaneously have $\mathcal{R}[f^*] \leq \mathcal{R}[\widehat{f^*}]$ (because f^* is a minimizer) and $\mathcal{R}[f^*] \geq \mathcal{R}[\widehat{f^*}]$ (by the conclusion above). Hence $\mathcal{R}[f^*] = \mathcal{R}[\widehat{f^*}]$, and $f^* = \widehat{f^*} \in \mathcal{H}_1$, which concludes the proof. □

Remark. The representer theorem finds a finite dimensional subspace $\mathcal{H}_1$ which contains the optimal solution in an infinite dimensional space $\mathcal{H}$. The solution thereupon can be represented by a finite linear combination of kernel products evaluated on data points in the training set.

6.3 Equivalence between KRR and Kriging

Both kernel ridge regression (KRR) and kriging are nonparametric regression models which are applied in spatial interpolation. KRR finds an optimal solution through minimizing some loss function, analogous to other frequentist methods such as maximum likelihood estimation. Kriging, as a Bayesian approach, imposes a prior distribution (which is a Gaussian process) on the spatial field of interest, and then derives the posterior.

In this part we will reveal the intrinsic connection between the two methods (Frequentist & Bayesian).

Setting of Kriging. Given n observations $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$, consider a nonparametric regression model

$$y_i = f(x_i) + \varepsilon_i,\ \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \tag{6.14}$$

with prior $f \sim \mathcal{GP}(0, k(\cdot,\cdot))$. The posterior of f is a Gaussian process, with

$$\begin{aligned} \mathbb{E}[f(x)|\mathbf{Y}] &= \mathbf{k}(x)^\top (\mathbf{K} + \sigma^2 \mathbf{I})^{-1} \mathbf{Y}, \\ \mathrm{Cov}\{f(x), f(x')|\mathbf{Y}\} &= k(x,x') - \mathbf{k}(x)^\top (\mathbf{K} + \sigma^2 \mathbf{I}_n)^{-1} \mathbf{k}(x'), \end{aligned} \tag{6.15}$$

where $\mathbf{k}(x) = (k(x,x_1), \cdots, k(x,x_n))^\top$, $\mathbf{K} = \{k(x_i,x_j)\}_{i,j=1}^n$, $\mathbf{Y} = (y_1, \cdots, y_n)^\top$.

Setting of KRR. Given n observations $\{(x_i, y_i)\}_{i=1}^n \subset \mathcal{X} \times \mathbb{R}$, and suppose f is realizable in RKHS $\mathcal{H}$ with RK $k(\cdot,\cdot)$. Choose MSE loss as the loss function, and consider the following kernel ridge regression model:

$$\min_{f \in \mathcal{H}} \sum_{i=1}^n \{y_i - f(x_i)\}^2 + \lambda \|f\|_{\mathcal{H}}^2. \tag{6.16}$$

where $\lambda > 0$ is a hyperparameter. According to the representer theorem, the solution has a finite dimensional representation

$$f(\cdot) = \sum_{i=1}^n \alpha_i k(\cdot, x_i), \tag{6.17}$$

then the optimization problem can be reformulated as

$$\min_{\alpha \in \mathbb{R}^n} (\mathbf{Y} - \mathbf{K}\boldsymbol{\alpha})^\top (\mathbf{Y} - \mathbf{K}\boldsymbol{\alpha}) + \lambda \boldsymbol{\alpha}^\top \mathbf{K} \boldsymbol{\alpha}, \tag{6.18}$$

where $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_n)^\top$, $\mathbf{K} = \{k(x_i,x_j)\}_{i,j=1}^n$, $\mathbf{Y} = (y_1, \cdots, y_n)^\top$. When $\mathbf{K}$ is nonsingular, the minimizer is unique:

$$\hat{\boldsymbol{\alpha}} = (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{Y}, \tag{6.19}$$

$$\hat{f}(x) = \sum_{i=1}^n \hat{\alpha}_i k(x, x_i) = \mathbf{Y}^\top (\mathbf{K} + \lambda \mathbf{I}_n)^{-1} \mathbf{k}(x), \tag{6.20}$$

where $\mathbf{k}(x) = (k(x,x_1), \cdots, k(x,x_n))^\top$. As we can see, the point estimate $\hat{f}$ is equivalent to the posterior mean of f in Kriging when we replace λ with σ^2.

7 Karhunen-Loève Expansion

7.1 Kosambi-Karhunen-Loève Theorem

Let $\mathcal{X}$ be a compact space, and let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\{X(s) \in L^2(\Omega) : s \in \mathcal{X}\}$ be a second order process over $(\Omega, \mathcal{F}, \mathbb{P})$ with zero mean and continuous covariance function $k(\cdot, \cdot)$, where $k(\cdot, \cdot)$ is a Mercer's kernel, i.e. there exists $\{\lambda_l\}_{l \in \mathbb{N}}$ such that $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and an orthogonal basis $\{\phi_l\}_{l \in \mathbb{N}^*} \subset L^2(\Omega)$,

$$k(s,t) = \sum_{l=1}^{\infty} \lambda_l \phi_l(s) \phi_l(t), \ \forall s, t \in \mathcal{X}. \tag{7.1}$$

Theorem 7.1 (Kosambi-Karhunen-Loève). The process $X(\cdot)$ defined above admits the following expansion:

$$X(t) = \sum_{l=1}^{\infty} \sqrt{\lambda_l} \xi_l \phi_l(t), \tag{7.2}$$

where the random variables $\{\xi_l\}_{l \in \mathbb{N}}$ are

$$\xi_l = \frac{1}{\sqrt{\lambda_l}} \int_{\mathcal{X}} X(t) \phi_l(t) \mathrm{d}t, \tag{7.3}$$

and the convergence in (7.2) holds uniformly on $\mathcal{X}$ in L^2 sense. Furthermore, $\{\xi_l\}_{l \in \mathbb{N}}$ are uncorrelated, with

$$\mathbb{E}\xi_i = 0, \ \mathrm{Cov}(\xi_i, \xi_j) = \delta_{ij}, \ \forall i, j \in \mathbb{N}^*. \tag{7.4}$$

Proof. By definition, the mean and covariance of $\{\xi_l\}_{l \in \mathbb{N}}$ are

$$\mathbb{E}\xi_i = \frac{1}{\sqrt{\lambda_i}} \int_{\mathcal{X}} \mathbb{E}[X(t)] \phi_i(t) \mathrm{d}t = 0, \tag{7.5}$$

$$\begin{aligned}
\mathrm{Cov}(\xi_i, \xi_j) = \mathbb{E}[\xi_i \xi_j] &= \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}} \int_{\mathcal{X}} \int_{\mathcal{X}} \mathbb{E}[X(s)X(t)] \phi_i(s) \phi_j(t) \mathrm{d}s \mathrm{d}t \\
&= \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}} \int_{\mathcal{X}} \left(\int_{\mathcal{X}} k(s,t) \phi_i(s) \mathrm{d}s \right) \phi_j(t) \mathrm{d}t \\
&= \frac{1}{\sqrt{\lambda_i}} \frac{1}{\sqrt{\lambda_j}} \int_{\mathcal{X}} \lambda_i \phi_i(t) \phi_j(t) \mathrm{d}t \\
&= \frac{\lambda_i}{\sqrt{\lambda_i \lambda_j}} \int_{\mathcal{X}} \phi_i(t) \phi_j(t) \mathrm{d}t \\
&= \frac{\lambda_i}{\sqrt{\lambda_i \lambda_j}} \delta_{ij} = \delta_{ij}.
\end{aligned} \tag{7.6}$$

Also, we have

$$\begin{aligned}
\mathbb{E}[X(t)\xi_l] &= \frac{1}{\sqrt{\lambda_l}} \mathbb{E}\left[\int_{\mathcal{X}} X(t) X(s) \phi_l(s) \mathrm{d}s \right] \\
&= \frac{1}{\sqrt{\lambda_l}} k(t,s) \phi_l(s) \mathrm{d}s = \sqrt{\lambda_l} \phi_l(t).
\end{aligned} \tag{7.7}$$

To show the convergence, we define $X_L(\cdot)$ as

$$X_L(t) = \sum_{l=1}^{L} \sqrt{\lambda_l} \xi_l \phi_l(t). \tag{7.8}$$

Then

$$\begin{aligned}\mathbb{E}\left[(X(t)-X_L(t))^2\right] &= \mathbb{E}\left[X(t)^2\right] - 2\mathbb{E}\left[X(t)X_L(t)\right] + \mathbb{E}\left[X_L(t)^2\right] \\ &= \mathbb{E}\left[X(t)^2\right] - 2\sum_{l=1}^{L}\sqrt{\lambda_l}\phi_l(t)\mathbb{E}\left[X(t)\xi_l\right] + \sum_{l=1}^{L}\lambda_l\phi_l(t)^2\mathbb{E}[\xi_l^2] \\ &= k(t,t) - \sum_{l=1}^{L}\lambda_l\phi_l(t)^2. \end{aligned} \tag{7.9}$$

Note that k is bounded, which implies

$$\sup_{t\in\mathcal{X}}|k(t,t)| < \infty. \tag{7.10}$$

By (5.24), we have

$$\lim_{L\to\infty}\sup_{t\in\mathcal{X}}\left|k(t,t) - \sum_{l=1}^{L}\lambda_l\phi_l(t)^2\right| = 0 \tag{7.11}$$

Hence $X_L(\cdot) \overset{L^2}{\to} X(\cdot)$, which concludes the proof. □

Remarks. (I) The Karhunen-Loève expansion gives the representation of a stochastic process as an infinite linear combination of orthogonal functions. 0.1cm

(II) (Gaussian case). For a Gaussian process $X(\cdot) \sim \mathcal{GP}(0, k(\cdot,\cdot))$, where k admits Mercer's property $k(s,t) = \sum_{l\in\mathbb{N}}\lambda_l\phi_l(s)\phi_l(t)$, it can be represented as

$$X(t) = \sum_{l=1}^{\infty}\sqrt{\lambda_l}Z_l\phi_l(t), \quad Z_l \overset{\text{i.i.d.}}{\sim} N(0,1). \tag{7.12}$$

(III) If $X(\cdot)$ is finite dimensional, i.e. there exists some N such that $\lambda_n = 0$ if $n > N$, then the Karhunen-Loève expansion has a truncated version:

$$X(t) = \sum_{l=1}^{N}\sqrt{\lambda_l}\xi_l\phi_l(t). \tag{7.13}$$

Note that the proof above need to be slightly modified, since ξ_n's with $n > N$ are undefined.

7.2 Function Approximation with Orthogonal Bases

We consider the problem of function approximation. Let $\mathcal{X}$ be a compact set. For a space of (real-valued) functions defined on $\mathcal{X}$, we may seek for a basis $\{\psi_i\}_{i\in\mathbb{N}}$ such that the first n can be used to approximate the functions in this space. To evaluate the approximation error, we define the μ-weighted L^2 norm:

$$\|f\|_{L^2(\mu)} = \sqrt{\int_{\mathcal{X}}|f(x)|^2\mathrm{d}\mu}, \tag{7.14}$$

where μ is a σ-finite measure on $\mathcal{X}$. This norm has an associated (real) inner product

$$\langle f, g\rangle_{L^2(\mu)} = \int_{\mathcal{X}} f(x)g(x)\mathrm{d}\mu. \tag{7.15}$$

Consider approximating functions in the $L^2(\mu)$ space

$$L^2(\mathcal{X}, \mu) = \left\{ f : \mathcal{X} \to \mathbb{R}, \|f\|_{L^2(\mu)} < \infty \right\}. \tag{7.16}$$

A basis $\{\psi_i\}_{i\in\mathbb{N}}$ on $L^2(\mathcal{X}, \mu)$ is a set of functions such that every $f \in L^2(\mathcal{X}, \mu)$ can be uniquely represented as

$$f = \sum_{l=1}^{\infty} c_l \psi_l \tag{7.17}$$

in the sense that the series of partial sums converges in norm:

$$\lim_{n\to\infty} \left\| f - \sum_{l=1}^{n} c_l \psi_l \right\|_{L^2(\mu)} = 0. \tag{7.18}$$

7.2.1 Orthogonal Basis

Definition 7.2 (Orthogonality). An *orthogonal basis* for $L^2(\mathcal{X}, \mu)$ is a collection $\{\psi_\alpha\}_{\alpha\in J}$ of integrable functions such that $\langle \psi_\alpha, \psi_\beta \rangle_{L^2(\mu)} = 0$ for all $\alpha \neq \beta$, and that $L^2(\mathcal{X}, \mu) = \overline{\text{span}}\{\psi_\alpha\}_{\alpha\in J}$. Additionally, if $\|\psi_\alpha\|_{L^2(\mu)} = 1$ for all $\alpha \in J$, then $\{\psi_\alpha\}_{\alpha\in J}$ is said to be an *orthonormal* basis for $L^2(\mathcal{X}, \mu)$.

Proposition 7.3. Suppose $\{\psi_i\}_{i\in\mathbb{N}}$ is an orthogonal basis over $L^2(\mathcal{X}, \mu)$. For any $f \in L^2(\mathcal{X}, \mu)$, it has the representation $f = \sum_{n\in\mathbb{N}} c_n \psi_n$.

- (Coefficient Matching). The coefficient c_l can be easily found by taking the inner product with a basis function:

$$\langle f, \psi_l \rangle_{L^2(\mu)} = \sum_{n=1}^{\infty} c_n \langle \psi_l, \psi_n \rangle_{L^2(\mu)} = c_l \|\psi_l\|^2_{L^2(\mu)} \implies c_l = \frac{\langle f, \psi_l \rangle_{L^2(\mu)}}{\|\psi_l\|^2_{L^2(\mu)}}. \tag{7.19}$$

- (Best approximation). The approximation $f_N = \sum_{l=1}^{N} c_l \psi_l$ is the best approximation to f in subspace $\mathcal{S}_N = \text{span}\{\psi_1, \cdots, \psi_N\}$ in the sense of minimizing the norm of error, i.e.

$$f_N = \underset{g\in\mathcal{S}_N}{\text{argmin}} \|f - g\|_{L^2(\mu)}. \tag{7.20}$$

- (Parseval's theorem). The norm of error is

$$\|f - f_N\|_{L^2(\mu)} = \sqrt{\sum_{l=N+1}^{\infty} c_l^2 \|\psi_l\|^2_{L^2(\mu)}}. \tag{7.21}$$

Examples. In the following discussion, we consider function approximation in the L^2-space, i.e. the weighting measure μ is Lebesgue measure.

- (Fourier Series). Consider function approximations on the closed interval $[0, 2\pi]$, an orthogonal basis is $\{1\} \cup \{\cos(nx)\}_{n\in\mathbb{N}} \cup \{\sin(nx)\}_{n\in\mathbb{N}}$ Then we can expand a function $f \in L_2([0, 2\pi])$ to its Fourier series:

$$f(x) = A_0 + \sum_{n=1}^{\infty} A_n \cos(nx) + \sum_{n=1}^{\infty} B_n \sin(nx), \tag{7.22}$$

where the coefficients are

$$A_0 = \frac{1}{2\pi}\int_0^{2\pi} f(x)\mathrm{d}x, \tag{7.24}$$

$$A_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\cos(nx)\mathrm{d}x,\ n \geq 1, \tag{7.25}$$

$$B_n = \frac{1}{\pi}\int_0^{2\pi} f(x)\sin(nx)\mathrm{d}x,\ n \geq 1. \tag{7.26}$$

- (Legendre Polynomials). Consider function approximations on the closed interval $[-1, 1]$. Beginning from $P_0(x) = 1,\ P_1(x) = x$, one can use Gram-Schmidt process to find a polynomial basis $\{P_n\}_{n\in\mathbb{N}}$ on $[-1, 1]$. A general solution is given by Rodrigues formula:

$$P_n(x) = \frac{1}{2^n n!}\frac{\mathrm{d}^n}{\mathrm{d}x^n}(x^2 - 1)^n. \tag{7.27}$$

7.2.2 Optimality of truncated Karhunen-Loève expansion

Let $X(t)$, $t \in \mathcal{X}$ be a zero mean second-order stochastic process over $(\Omega, \mathcal{F}, \mathbb{P})$, with covariance $k(\cdot, \cdot)$ a Mercer's kernel. Let $\{\psi_l\}_{l\in\mathbb{N}}$ be a orthonormal basis in $L^2(\mathcal{X})$, i.e. $\int_{\mathcal{X}} \psi_i(t)\psi_j(t) = \delta_{ij}$, Then we can expand $X(t)$ as an infinite series:

$$X(t) = \sum_{l=1}^{\infty} \sqrt{\nu_l}\psi_l(x)\xi_l, \tag{7.28}$$

where

$$\xi_l = \frac{1}{\sqrt{\nu_l}}\int_{\mathcal{X}} X(t)\psi_l(t)\mathrm{d}t. \tag{7.29}$$

Then the truncated version of $X(t)$ at order p and the corresponding term can be written as

$$\begin{aligned} X_p(t) &= \sum_{l=1}^{p} \sqrt{\nu_l}\psi_l(x)\xi_l, \\ e_p(t) = X(t) - X_p(t) &= \sum_{l=p+1}^{\infty} \sqrt{\nu_l}\psi_l(t)\xi_l = \sum_{l=p+1}^{\infty}\int_{\mathcal{X}} X(s)\psi_l(s)\psi_l(t)\mathrm{d}s. \end{aligned} \tag{7.30}$$

Furthermore, the integrated mean squared error (IMSE) is defined as

$$\begin{aligned} \mathcal{E}_p &= \int_{\mathcal{X}} \mathbb{E}\left[e_p(t)^2\right]\mathrm{d}t \\ &= \int_{\mathcal{X}} \mathbb{E}\left[\sum_{m=p+1}^{\infty}\sum_{n=p+1}^{\infty}\int_{\mathcal{X}}\int_{\mathcal{X}} X(s)X(u)\psi_m(s)\psi_m(t)\psi_n(u)\psi_n(t)\mathrm{d}u\mathrm{d}s\right]\mathrm{d}t \\ &= \sum_{m=p+1}^{\infty}\sum_{n=p+1}^{\infty}\left(\int_{\mathcal{X}}\psi_m(t)\psi_n(t)\mathrm{d}t\right)\left(\int_{\mathcal{X}}\int_{\mathcal{X}} k(s,u)\psi_m(s)\psi_n(u)\mathrm{d}u\mathrm{d}s\right) \\ &= \sum_{m=p+1}^{\infty}\left(\int_{\mathcal{X}}\int_{\mathcal{X}} k(s,u)\psi_m(s)\psi_m(u)\mathrm{d}u\mathrm{d}s\right). \end{aligned} \tag{7.31}$$

The following therorem states the optimality of truncated Karhunen-Loève expansion in the sense of minimizing IMSE.

Theorem 7.4 (Optimality of truncated Karhunen-Loève expansion). Among all the truncated expansion expressed as a finite linear combination of orthonormal basis, the truncated Karhunen-Loève expansion minimizes the IMSE.

Proof. We fix the truncation order p. Recall the expression of IMSE $\mathcal{E}_p$, consider the following optimization problem:

$$\min_{\{\lambda_l\}_{l>p}} \sum_{l=p+1}^{\infty} \left(\int_{\mathcal{X}} \int_{\mathcal{X}} k(s,u)\psi_l(s)\psi_l(u)\mathrm{d}u\mathrm{d}s \right) \quad \text{s.t.} \int_{\mathcal{X}} \psi_l(t)^2 \mathrm{d}t = 1,\ l \geq p+1. \tag{7.32}$$

Construct the Lagrangian function

$$\mathcal{L} = \sum_{l=p+1}^{\infty} \left\{ \left(\int_{\mathcal{X}} \int_{\mathcal{X}} k(s,u)\psi_l(s)\psi_l(u)\mathrm{d}u\mathrm{d}s \right) + \zeta_l \left(1 - \int_{\mathcal{X}} \psi_l(t)^2 \mathrm{d}t \right) \right\}. \tag{7.33}$$

Differentiate $\mathcal{L}$ with respect to $\psi_l(\cdot),\ l \geq p+1$, we can obtain a functional derivative:

$$\frac{\partial \mathcal{L}}{\partial \psi_l(\cdot)} = 2 \int_{\mathcal{X}} \left(\int_{\mathcal{X}} k(s,t)\psi_l(s)\mathrm{d}s - \zeta_l \psi_l(t) \right) \mathrm{d}t. \tag{7.34}$$

Set this derivative to zero yields a Fredholm integral equation:

$$\zeta_l \psi_l(t) = \int_{\mathcal{X}} k(s,t)\psi_l(s)\mathrm{d}s. \tag{7.35}$$

Hence ψ_l is selected as the eigenfunction of integral operator $T_k(\cdot) = \int_{\mathcal{X}} k(\cdot, s)\psi_l(s)\mathrm{d}s$, which yields Karhunen-Loève expansion. □

7.3 Example: Brownian Motion

Recall that the standard Brownian motion $W(\cdot)$ is a Gaussian process with mean zero and covariance $k(s,t) = \min(s,t)$. For simplicity, the domain of this process is assumed to be $[0,1]$. The integral equation is

$$\int_0^1 \min(s,t)\phi(s)\mathrm{d}s = \lambda\phi(t),\ \lambda > 0. \tag{7.36}$$

Equivalently,

$$\int_0^t s\phi(s)\mathrm{d}s + t\int_t^1 \phi(s)\mathrm{d}s = \lambda\phi(t),\ \lambda > 0. \tag{7.37}$$

Take the derivatives on both sides two times, we get

$$\int_t^1 \phi(s)\mathrm{d}s = \lambda \frac{\mathrm{d}}{\mathrm{d}t}\phi(t), \tag{7.38}$$

$$-\phi(t) = \lambda \frac{\mathrm{d}^2}{\mathrm{d}t^2}\phi(t). \tag{7.39}$$

Solve the second order differential equation (7.39), we got

$$\phi(t) = A\cos\left(\frac{t}{\sqrt{\lambda}}\right) + B\sin\left(\frac{t}{\sqrt{\lambda}}\right). \tag{7.39}$$

Take $t = 0$ in the integral equation (7.36), we have $\phi(0) = 0$, which implies $\phi(t) = B \sin\left(\frac{t}{\sqrt{\lambda}}\right)$. Plug in this to (7.38), we obtain

$$B \cos\left(\frac{t}{\sqrt{\lambda}}\right) - B \cos\left(\frac{1}{\sqrt{\lambda}}\right) = B \cos\left(\frac{t}{\sqrt{\lambda}}\right), \tag{7.40}$$

hence $\cos\left(\frac{1}{\sqrt{\lambda}}\right) = 0$, the eigenvalues are

$$\lambda_l = \frac{1}{\left(l - \frac{1}{2}\right)^2 \pi^2},\ \phi_l(t) = B \sin\big[(l - \frac{1}{2})\pi t\big],\ l = 1, 2, \cdots. \tag{7.41}$$

Moreover, the orthonormality condition implies

$$1 = \int_0^1 \phi_l(t)^2 \mathrm{d}t = \frac{1}{2}B^2 - \frac{1}{4}\sqrt{\lambda_l}B^2 \sin\left((2l-1)\pi\right) = \frac{1}{2}B^2, \tag{7.42}$$

hence $B = \pm\sqrt{2}$. (The sign does not matter since the standard Gaussian distribution is symmetric.)

Therefore, the standard Brownian motion $W(t),\ 0 \le t \le 1$ can be represented as its Karhunen-Loève expansion:

$$W(t) = \sqrt{2}\sum_{l=1}^{\infty} \frac{2}{(2l-1)\pi} \sin\big[(l - \frac{1}{2})\pi t\big] Z_l, \tag{7.43}$$

where $\{Z_l\}_{l\in\mathbb{N}} \overset{\text{i.i.d.}}{\sim} N(0,1)$.

References

[1] Matheron, G. (1972). *Leçon sur les fonctions aléatoire d'ordre 2*, Technical Report C-53, MINES Paristech - Centre de Géosciences.

[2] Steinwart, I. and Christmann, A. (2008). Kernels and Reproducing Kernel Hilbert Spaces. *Support Vector Machines*. Springer, New York.

[3] Ferreira, J.C., Menegatto, V.A. (2009). Eigenvalues of Integral Operators Defined by Smooth Positive Definite Kernels. *Integral equation operator theory* 64, 61–81.

[4] Schölkopf, B., Herbrich, R. and Smola, A. J. (2001). A Generalized Representer Theorem. In Helmbold, D.; Williamson, B. (Eds.): *Computational Learning Theory*. Lecture Notes in Computer Science. Vol. 2111. Berlin, Heidelberg: Springer. pp. 416–426.

[5] Wang, L. (2008). *Karhunen-Loève expansions and their applications.* London School of Economics and Political Science (United Kingdom).

[6] Wikle, C.K., Zammit-Mangion, A., and Cressie, N. (2019). *Spatio-Temporal Statistics with R*. Chapman & Hall/CRC, Boca Raton, FL.