ECE 581: Gaussian Processes: From Finite Dimensional to Fields and Reproducing Kernel Hilbert Spaces

Henry D. Pfister
Duke University

December 1, 2025

Contents

1 Introduction

Why Gaussians? Gaussian models are ubiquitous in modern statistics, signal processing, and machine learning. They enable *moderate-complexity* algorithms with *optimal inference* properties under quadratic loss: linear estimators become optimal, posteriors stay Gaussian, and key computations reduce to linear algebra. Beyond convenience, the *central limit theorem* (CLT) explains why sums of many small, independent effects are approximately Gaussian, making Gaussian assumptions broadly reasonable in practice.

What this tutorial covers. We begin with one-dimensional and i.i.d. Gaussian basics, lift to random vectors, and derive the general multivariate Gaussian pdf via affine transformations. We then develop linear–Gaussian inference, showing that conditionals and posteriors remain Gaussian with closed-form mean and covariance. Finally, we connect *Gaussian processes* (GPs) and *Gaussian random fields* (GRFs) to *reproducing kernel Hilbert spaces* (RKHS), highlighting kernels, Mercer decompositions, and the Karhunen–Loève (KL) connection.

2 Background

2.1 The Gaussian Distribution

A (scalar) Gaussian random variable X with mean $\mu \in \mathbb{R}$ and variance $\sigma^2 > 0$ is denoted by $X \sim \mathcal{N}(\mu, \sigma^2)$ and defined by its pdf

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) = \frac{1}{\sigma} \phi\left(\frac{x-\mu}{\sigma} \right), \tag{1}$$

where $\phi(x) \coloneqq (2\pi)^{-1/2} e^{-x^2/2}$ is the pdf of a standard Gaussian. Its cdf is given by

$$\Phi_{\mu,\sigma}(x) = \Pr\{X \le x\} = \int_{-\infty}^{x} f_X(z)\, dz = \Phi\left(\frac{x-\mu}{\sigma} \right),$$

where $\Phi(x) \coloneqq \int_{-\infty}^{x} \phi(z)\, dz$ is the cdf of a standard Gaussian.

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ be a sequence of i.i.d. standard Gaussians. It follows from independence that the joint pdf has the product form given by

$$f_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^{n} \phi(x_i) = (2\pi)^{-n/2} \exp\big(-\tfrac{1}{2} \|\mathbf{x}\|^2 \big).$$

We write $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_n)$ to denote an n-dimensional standard normal.

2.2 Random Vectors

The theory of random variables extends naturally to vectors. Let $\mathbf{X} = (X_1, \ldots, X_n)^\top$ be a vector where each entry is a random variable. Then, $\mathbf{X}$ is called a random vector in $\mathbb{R}^n$. Standard operations such as addition and scalar multiplication are naturally inherited. Also, the expectation operator acts componentwise and is defined by

$$\mathbb{E}[\mathbf{X}] \coloneqq (\mathbb{E}[X_1], \ldots, \mathbb{E}[X_n])^\top.$$

For a matrix $A \in \mathbb{R}^{m\times n}$ and vector $\mathbf{b} \in \mathbb{R}^n$, the linearity of expectation implies that $\mathbb{E}[A\,\mathbf{X} + \mathbf{b}] = A\,\mathbb{E}[\mathbf{X}] + \mathbf{b}$. We write I_n for the $n \times n$ identity matrix and drop the subscript when the dimension is clear from context.

2.3 Mean and Covariance

For a random vector $\mathbf{X}$ in $\mathbb{R}^n$, the *mean vector* is defined to be $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}[\mathbf{X}] \in \mathbb{R}^n$. Similarly, the *covariance matrix* of $\mathbf{X}$ is defined by

$$\Sigma_{\mathbf{X}} = \mathrm{Cov}(\mathbf{X}) \coloneqq \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top\big] \in \mathbb{R}^{n\times n}, \tag{2}$$

where subscripts are dropped when they are clear from the context. We also note that the total variance of a random vector is given by

$$\begin{aligned}\mathbb{E}\big[\|\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}}\|^2\big] &= \mathbb{E}\big[\,\mathrm{Tr}\left((\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top\right)\big] \\ &= \mathrm{Tr}\left(\mathbb{E}\big[\left((\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})^\top\right)\big]\right) \\ &= \mathrm{Tr}\left(\Sigma\right).\end{aligned}$$

Example 2.1 (Linear Form)**.** For the random vector $\mathbf{X}$ in $\mathbb{R}^n$ and any $c \in \mathbb{R}^n$, the linear form $Y = \mathbf{c}^\top\mathbf{X}$ defines a scalar random variable. It is easy to verify that Y has mean $\mathbb{E}[Y] = \mathbf{c}^\top\boldsymbol{\mu}_{\mathbf{X}}$ and variance $\mathrm{Var}(Y) = \mathbf{c}^\top\Sigma_{\mathbf{X}}\,\mathbf{c}$.

We note that Σ is symmetric and positive semidefinite. For example, one can see that Σ is positive semidefinite just by noting that $\mathbf{c}^\top\Sigma_{\mathbf{X}}\,\mathbf{c} = \mathrm{Var}(Y) \geq 0$ for all $\mathbf{c} \in \mathbb{R}^n$. Likewise, it is symmetric because it equals the average of symmetric matrices given by $(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T$ for $\mathbf{x} \in \mathbb{R}^n$.

3 Multivariate Gaussian Distributions

Definition 3.1 (Multivariate Gaussian Distribution)**.** A random vector $\mathbf{Y} \in \mathbb{R}^n$ is *(multivariate) Gaussian* with mean $\boldsymbol{\mu} \in \mathbb{R}^n$ and covariance $\Sigma \in \mathbb{R}^{n\times n}$ (with Σ positive definite) if it has a PDF of the form

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\Big(-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\Big). \tag{3}$$

This statement can be written in shorthand as $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$.

Proposition 3.2. *The moment generating function associated with* (3) *is given by*

$$M_{\mathbf{Y}}(\mathbf{t}) \coloneqq \mathbb{E}\Big[e^{\mathbf{t}^\top\mathbf{Y}}\Big] = \exp\Big(\mathbf{t}^\top\boldsymbol{\mu} + \tfrac{1}{2}\,\mathbf{t}^\top\Sigma\,\mathbf{t}\Big). \tag{4}$$

Proof. Starting from the definition and using the density in (3), we can write

$$\begin{aligned}M_{\mathbf{Y}}(\mathbf{t}) &= \int_{\mathbb{R}^n}\exp(\mathbf{t}^\top\mathbf{y})\,\frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\exp\Big(-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})^\top\Sigma^{-1}(\mathbf{y}-\boldsymbol{\mu})\Big)\,\mathrm{d}\mathbf{y} \\ &= \frac{1}{(2\pi)^{n/2}\sqrt{\det\Sigma}}\int_{\mathbb{R}^n}\exp\Big(-\tfrac{1}{2}\big[\mathbf{y}^\top\Sigma^{-1}\mathbf{y} - 2(\Sigma^{-1}\boldsymbol{\mu}+\mathbf{t})^\top\mathbf{y} + \boldsymbol{\mu}^\top\Sigma^{-1}\boldsymbol{\mu}\big]\Big)\,\mathrm{d}\mathbf{y}.\end{aligned}$$

We complete the square in $\mathbf{y}$ by defining $\mathbf{m} \coloneqq \boldsymbol{\mu} + \Sigma\mathbf{t}$ and writing

$$\mathbf{y}^\top\Sigma^{-1}\mathbf{y} - 2(\Sigma^{-1}\boldsymbol{\mu}+\mathbf{t})^\top\mathbf{y} = (\mathbf{y}-\mathbf{m})^\top\Sigma^{-1}(\mathbf{y}-\mathbf{m}) - \mathbf{t}^\top\Sigma\mathbf{t} - 2\,\mathbf{t}^\top\boldsymbol{\mu}.$$

Plugging in and collecting constants gives

$$M_{\mathbf{Y}}(\mathbf{t}) = \exp\Big(\mathbf{t}^\top \boldsymbol{\mu} + \tfrac{1}{2}\,\mathbf{t}^\top \Sigma\,\mathbf{t}\Big)\,\frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \int_{\mathbb{R}^n} \exp\Big(-\tfrac{1}{2}(\mathbf{y}-\mathbf{m})^\top \Sigma^{-1}(\mathbf{y}-\mathbf{m})\Big)\,\mathrm{d}\mathbf{y}.$$

The integral equals 1 because it is the integral of an n-dimensional Gaussian density with mean $\mathbf{m}$ and covariance Σ. Hence $M_{\mathbf{Y}}(\mathbf{t}) = \exp\big(\mathbf{t}^\top \boldsymbol{\mu} + \frac{1}{2}\,\mathbf{t}^\top \Sigma\,\mathbf{t}\big)$. □

A probability distribution whose mgf exists on an open set containing 0 is uniquely determined by that mgf. Thus, we equations (3) and (4) both imply each other. See Theorem B.1.

Theorem 3.3 (Characterization of a Gaussian)**.** *The following are equivalent for a random vector* $\mathbf{Y}$ *in* $\mathbb{R}^n$ *with mean* $\boldsymbol{\mu}$ *and positive-definite covariance* Σ*:*

(i) $\mathbf{Y}$ *has the multivariate Gaussian distribution whose pdf is given by* (3)*.*

(ii) For every $\mathbf{c} \in \mathbb{R}^n$*, the real random variable* $\mathbf{c}^\top \mathbf{Y}$ *is a scalar Gaussian.*

(iii) $\mathbf{Y} = A\mathbf{X} + \boldsymbol{\mu}$ *for invertible* A *with* $\Sigma = AA^\top$ *and an iid standard Gaussian* $\mathbf{X} \sim \mathcal{N}(0, I_n)$*.*

Proof. (i) ⇒ *(ii).* For any $\mathbf{c} \in \mathbb{R}^n$ and $s \in \mathbb{R}$, we can use (3) to write the moment generating function of $\mathbf{c}^\top \mathbf{Y}$ as

$$M_{\mathbf{c}^\top \mathbf{Y}}(s) = \mathbb{E}\left[e^{s\mathbf{c}^\top \mathbf{Y}}\right] = M_{\mathbf{Y}}(s\mathbf{c}) = \exp\Big(s\,\mathbf{c}^\top \boldsymbol{\mu} + \tfrac{1}{2}s^2\,\mathbf{c}^\top \Sigma\,\mathbf{c}\Big).$$

This equals the MGF of $\mathcal{N}(\mathbf{c}^\top \boldsymbol{\mu},\, \mathbf{c}^\top \Sigma\,\mathbf{c})$ and thus $\mathbf{c}^\top \mathbf{Y}$ has the same distribution by Theorem B.1.

(ii) ⇒ *(i).* Assume that, for every $\mathbf{c} \in \mathbb{R}^n$, the random variable $\mathbf{c}^\top \mathbf{Y}$ is a scalar Gaussian. Since $\mathbf{Y}$ has mean $\boldsymbol{\mu}$ and covariance Σ, linearity of expectation implies that $\mathbf{c}^\top \mathbf{Y}$ has mean $\mathbf{c}^\top \boldsymbol{\mu}$ and variance $\mathbf{c}^\top \Sigma\,\mathbf{c}$. Thus, we have

$$M_{\mathbf{c}^\top \mathbf{Y}}(s) = \exp\Big(s\,\mathbf{c}^\top \boldsymbol{\mu} + \tfrac{1}{2}s^2\,\mathbf{c}^\top \Sigma\,\mathbf{c}\Big).$$

But, for any random vector $\mathbf{Y}$, we know that

$$M_{\mathbf{c}^\top \mathbf{Y}}(s) = \mathbb{E}\left[e^{s\mathbf{c}^\top \mathbf{Y}}\right] = M_{\mathbf{Y}}(s\mathbf{c}).$$

Thus, choosing $s = 1$ shows that the mgf of $\mathbf{Y}$ equals the mgf of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance Σ. By Theorem B.1, $\mathbf{Y}$ is a Gaussian with these parameters.

(iii) ⇒ *(ii),(i).* For any $\mathbf{c} \in \mathbb{R}^n$ and $s \in \mathbb{R}$, we can use properties of linear forms of Gaussians to write the moment generating function of $\mathbf{c}^\top \mathbf{Y} = \mathbf{c}^\top (A\mathbf{X} + \boldsymbol{\mu})$ as

$$M_{\mathbf{c}^\top \mathbf{Y}}(s) = \mathbb{E}\left[e^{s\mathbf{c}^\top (A\mathbf{X}+\boldsymbol{\mu})}\right] = \mathbb{E}\left[e^{s\mathbf{c}^\top \boldsymbol{\mu} + s(\mathbf{c}^\top \mathbf{A})\mathbf{X}}\right] = \exp\Big(s\,\mathbf{c}^\top \boldsymbol{\mu} + \tfrac{1}{2}s^2\,\mathbf{c}^\top A \underbrace{\Sigma_{\mathbf{X}}}_{I_n} A^\top \mathbf{c}\Big).$$

This equals the MGF of $\mathcal{N}(\mathbf{c}^\top \boldsymbol{\mu},\, \mathbf{c}^\top AA^\top \mathbf{c})$ and thus $\mathbf{c}^\top \mathbf{Y}$ is a scalar Gaussian and we have *(ii)*. For *(i)*, recall that

$$M_{\mathbf{c}^\top \mathbf{Y}}(s) = \mathbb{E}\left[e^{s\mathbf{c}^\top \mathbf{Y}}\right] = M_{\mathbf{Y}}(s\mathbf{c})$$

for any random vector $\mathbf{Y}$. Thus, choosing $s = 1$ shows that the mgf of $\mathbf{Y}$ equals the mgf of a Gaussian random vector with mean $\boldsymbol{\mu}$ and covariance $\Sigma = AA^\top$. □

3.1 Generating Gaussian Random Variables

By Theorem 3.3, one can generate a Gaussian random vector $\mathbf{Y}$ with mean $\boldsymbol{\mu}$ and covariance $\Sigma = AA^T$ simply by computing

$$\mathbf{Y} = A\mathbf{X} + \boldsymbol{\mu},$$

where $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_n)$. The only missing element is how to find A from Σ. There are two standard approaches:

1. Eigenvalue decomposition: For a symmetric positive semidefinite matrix Σ, an orthogonal set of eigenvectors is guaranteed and we have the decomposition $\Sigma = Q\Lambda Q^\top$, where Λ is a diagonal matrix containing the non-negative eigenvalues of Σ and Q is an orthogonal matrix satisfying $Q^\top Q = QQ^\top = I$. Choosing $A = Q\sqrt{\Lambda}$ gives $AA^\top = \Sigma$. For covariance matrices, this idea is closely related to the Karhunen-Loéve transform.

2. LDL$^\top$ factorization: For a symmetric matrix, the standard LU decomposition implied by Gaussian elimination is easily modified to give $\Sigma = LDL^\top$, where L is lower triangular with ones on the diagonal and D is a diagonal matrix. If Σ is positive semidefinite, then D has non-negative entries and we can choose $A = L\sqrt{D}$ to see that $AA^\top = \Sigma$. For $\Sigma \succ 0$, this is the classic Cholesky factorization $\Sigma = AA^T$.

3.2 Deriving the PDF via Change of Variables

We derive the density of $\mathbf{Y} = A\mathbf{X} + \boldsymbol{\mu}$ where $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, I_n)$ and $A \in \mathbb{R}^{n\times n}$ is invertible. The mapping $h : \mathbf{x} \mapsto A\mathbf{x} + \boldsymbol{\mu}$ is a bijection with inverse $h^{-1}(\mathbf{y}) = A^{-1}(\mathbf{y} - \boldsymbol{\mu})$. Let $\mathcal{C} = [-1/2, 1/2]^n$ be the unit cube in $\mathbb{R}^n$ centered at $\mathbf{0}$. Then, we have

$$\begin{aligned}
f_{\mathbf{Y}}(\mathbf{y}) &= \lim_{\delta\to 0} \frac{\Pr(\mathbf{Y} \in \delta\mathcal{C} + \mathbf{y})}{\mathrm{Vol}(\delta\mathcal{C} + \mathbf{y})} \\
&= \lim_{\delta\to 0} \frac{\Pr(h(\mathbf{X}) \in \delta\mathcal{C} + \mathbf{y})}{\delta^n} \\
&= \lim_{\delta\to 0} \frac{\Pr(\mathbf{X} \in h^{-1}(\delta\mathcal{C} + \mathbf{y}))}{\mathrm{Vol}(h^{-1}(\delta\mathcal{C} + \mathbf{y}))} \frac{\mathrm{Vol}(h^{-1}(\delta\mathcal{C} + \mathbf{y}))}{\delta^n} \\
&= \lim_{\delta\to 0} \frac{\Pr(\mathbf{X} \in h^{-1}(\delta\mathcal{C} + \mathbf{y}))}{\mathrm{Vol}(h^{-1}(\delta\mathcal{C} + \mathbf{y}))} \frac{\delta^n \mathrm{Vol}(A^{-1}\mathcal{C})}{\delta^n} \\
&= \lim_{\delta\to 0} \frac{\Pr(\mathbf{X} \in h^{-1}(\delta\mathcal{C} + \mathbf{y}))}{\mathrm{Vol}(h^{-1}(\delta\mathcal{C} + \mathbf{y}))} \left|\det A^{-1}\right| \\
&= f_{\mathbf{X}}\big(h^{-1}(\mathbf{y})\big) \left|\det A^{-1}\right| \\
&= f_{\mathbf{X}}\big(h^{-1}(\mathbf{y})\big) \,/\, \sqrt{\det \Sigma} \\
&= \frac{1}{\sqrt{(2\pi)^n \det \Sigma}} \exp\Big(-\tfrac{1}{2} \left\| A^{-1}(\mathbf{y} - \boldsymbol{\mu}) \right\|^2 \Big),
\end{aligned}$$

where $\left|\det A^{-1}\right| = 1/\sqrt{\det \Lambda} = 1/\sqrt{\det \Sigma}$ is the volume of the parallelepiped $A^{-1}\mathcal{C}$.

4 Estimation and Inference

4.1 Multiple Random Vectors

Consider joint random vectors $\mathbf{X}$ in $\mathbb{R}^n$ and $\mathbf{Y}$ in $\mathbb{R}^m$ with finite second-moments (i.e., $\mathbb{E}[\|\mathbf{X}\|^2] < \infty$ and $\mathbb{E}[\|\mathbf{Y}\|^2] < \infty$). This implies the existence of their expectations, $\boldsymbol{\mu}_{\mathbf{X}} = \mathbb{E}[\mathbf{X}] \in \mathbb{R}^n$ and

$\boldsymbol{\mu}_{\mathbf{Y}} = \mathbb{E}[\mathbf{Y}] \in \mathbb{R}^m$, and their cross-covariance

$$\Sigma_{\mathbf{XY}} \coloneqq \mathrm{Cov}(\mathbf{X}, \mathbf{Y}) = \mathbb{E}\big[(\mathbf{X} - \boldsymbol{\mu}_{\mathbf{X}})(\mathbf{Y} - \boldsymbol{\mu}_{\mathbf{Y}})^\top\big] \in \mathbb{R}^{n \times m}.$$

The minimum mean-squared error (MMSE) when estimating $\mathbf{X}$ from $\mathbf{Y}$ is defined to be

$$\begin{aligned}\mathrm{mmse}(\mathbf{X} \mid \mathbf{Y}) &\coloneqq \inf_{g\colon \mathbb{R}^m \to \mathbb{R}^n} \mathbb{E}\big[\|\mathbf{X} - g(\mathbf{Y})\|^2\big] \\ &= \mathbb{E}\big[\|\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Y}]\|^2\big],\end{aligned}$$

where the conditional expectation $\mathbb{E}[\mathbf{X} \mid \mathbf{Y}] = g(\mathbf{Y})$ is defined by any optimal g. If the conditional pdf $f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y})$ exists, then we have

$$\mathbb{E}[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = g(\mathbf{y}) = \int \mathbf{x}\, f_{\mathbf{X}|\mathbf{Y}}(\mathbf{x}|\mathbf{y}) d\mathbf{x}.$$

The law of nested conditional expectation says that $\mathbb{E}\big[\mathbb{E}[\mathbf{X} \mid \mathbf{Y}]\big] = \mathbb{E}[\mathbf{X}]$.

Using the conditional expectation, we can define the conditional covariance matrix with

$$\mathrm{Cov}(\mathbf{X} \mid \mathbf{Y}) \coloneqq \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Y}])(\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Y}])^\top \mid \mathbf{Y}\big].$$

In this case, the law of nested conditional expectation implies that

$$\mathbb{E}\big[\,\mathrm{Tr}\,\big(\,\mathrm{Cov}(\mathbf{X} \mid \mathbf{Y})\big)\big] = \mathbb{E}\big[\mathbb{E}\big[\|\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Y}]\|^2 \mid \mathbf{Y}\big]\big] = \mathrm{mmse}(\mathbf{X} \mid \mathbf{Y}).$$

In terms of notation, we now introduce notation similar to $\mathbf{Y} \sim \mathcal{N}(\mathbf{0}, I_n)$ to define conditional distributions. To specify the conditional distribution of a random variable $\mathbf{Y}$ given the event $\mathbf{X} = \mathbf{x}$, one can use the notation $\mathbf{Y}|\mathbf{X} = \mathbf{x} \sim \mathrm{Dist}(\mathbf{x})$, where $\mathrm{Dist}(\mathbf{x})$ represents some distribution whose parameters depend on $\mathbf{x}$. For example, if $\mathbf{Y}$ is an observation $\mathbf{X}$ in standard Gaussian noise, then we would write

$$\mathbf{Y}|\mathbf{X} = \mathbf{x} \;\sim\; \mathcal{N}(\mathbf{x}, I_n).$$

4.2 Conditioning for a Bivariate Gaussian

Often, one would like to infer the posterior distribution X_1 from an observation of X_2 when (X_1, X_2) are jointly Gaussian. The following lemma addresses this situation.

Lemma 4.1 (Bivariate Gaussian conditioning)**.** *Let* $(X_1, X_2)^\top \sim \mathcal{N}\left(\begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \begin{bmatrix}\sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2\end{bmatrix}\right)$ *with* $\sigma_1^2 > 0$, $\sigma_2^2 > 0$, *and* $\rho \in [-1, 1]$. *Then, we have*

$$X_1 \mid X_2 = x_2 \;\sim\; \mathcal{N}\Big(\mu_1 - J_{12}(x_2 - \mu_2)/J_{11},\; 1/J_{11}\Big).$$

Proof. Let $\Delta = \det(\Sigma) = \sigma_1^2\sigma_2^2 - \rho^2\sigma_1^2\sigma_2^2 > 0$. Then, we have

$$J = \Sigma^{-1} = \frac{1}{\Delta}\begin{bmatrix}\sigma_2^2 & -\rho\sigma_1\sigma_2 \\ -\rho\sigma_1\sigma_2 & \sigma_1^2\end{bmatrix} \;\Rightarrow\; J_{11} = \frac{\sigma_2^2}{\Delta}, \quad J_{12} = -\frac{\rho\sigma_1\sigma_2}{\Delta}.$$

Since the conditional density of X_1 given X_2 equals

$$f_{X_1|X_2}(x_1|x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{\sqrt{2\pi\sigma_2^2}}{2\pi\sqrt{\det(\Sigma)}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})\Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu}) + \frac{1}{2\sigma_2^2}(x_2 - \mu_2)^2\right)$$

$$= \frac{1}{\sqrt{2\pi(\Delta/\sigma_2^2)}} \exp\left(-\frac{J_{11}}{2}(x_1-\mu_1)^2 + J_{12}(x_1-\mu_1)(x_2-\mu_2) - \frac{J_{22}}{2}(x_2-\mu_2)^2 + \frac{1}{2\sigma_2^2}(x_2-\mu_2)^2 \right),$$

one can simplify the exponential to verify that

$$X_1 \mid X_2 = x_2 \;\sim\; \mathcal{N}\Big(\mu_1 - J_{12}(x_2-\mu_2)/J_{11},\; 1/J_{11}\Big).$$

A simpler method is to observe that

$$\begin{aligned} f_{X_1|X_2}(x_1|x_2) &\propto \exp\big(-\tfrac{1}{2}\big[J_{11}x_1^2 - 2(J_{11}\mu_1 - J_{12}(x_2-\mu_2))x_1\big] + \text{const in } x_1\big) \\ &\propto \exp\big(-\tfrac{1}{2}(x_1-\mu_{1|2})^2/\sigma_{1|2}^2 + \text{const in } x_1\big), \end{aligned}$$

where $\sigma_{1|2}^2$ is the posterior variance and $\mu_{1|2}$ is the posterior mean. From this, one can easily read off these parameters from the terms in the exponential involving x_1. In particular, the posterior variance is half the reciprocal of the coefficient x_1^2 (i.e., $\sigma_{1|2}^2 = 1/J_{11}$) and the mean is $\sigma_{1|2}^2$ times the coefficient of x_1 (i.e., $\mu_{1|2} = \mu_1 - J_{12}(x_2-\mu_2)/J_{11}$). Thus, we have

$$\sigma_{1|2}^2 = \mathrm{Var}(X_1 \mid X_2) = \frac{\Delta}{\sigma_2^2} = \sigma_1^2(1-\rho^2), \qquad \mu_{1|2} = \mathbb{E}[X_1 \mid X_2 = x_2] = \mu_1 + \rho\sigma_1 \frac{x_2-\mu_2}{\sigma_2}.$$

Conceptually, the formula for $\mu_{1|2}$ is natural because $(x_2-\mu_2)/\sigma_2$ is the normalized deviation in x_2 and the scale factor $\rho\sigma_1$ maps this to its linear effect on x_1. □

Example 4.2. Under the conditions of the previous lemma, let $\mu_1 = 1$, $\mu_2 = -2$, $\sigma_1^2 = 4$, $\sigma_2^2 = 9$, and $\rho = 0.5$. For the observation $X_2 = 1.5$, the lemma yields

$$X_1 \mid X_2 = 1.5 \;\sim\; \mathcal{N}\Big(\mu_1 + \rho\,\sigma_1 \tfrac{1.5-\mu_2}{\sigma_2},\; \sigma_1^2(1-\rho^2)\Big) = \mathcal{N}\Big(1 + \tfrac{1}{2}\cdot 2 \cdot \tfrac{3.5}{3},\; 4(1-0.25)\Big) = \mathcal{N}(\mu_{1|2}, \sigma_{1|2}^2),$$

where the posterior mean and variance are $\mu_{1|2} \approx 2.1667$ and $\sigma_{1|2}^2 = 3$.

4.3 Conditioning on Part of a Gaussian Vector

In many cases, one would like to infer the posterior distribution of some elements of a Gaussian vector given the other elements. If $\mathbf{X}$ is Gaussian, then we can split $\mathbf{X}$ into jointly Gaussian random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$. Thus, we define

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \mathbb{E}[\mathbf{X}] = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \Sigma = \mathrm{Cov}(\mathbf{X}) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

Theorem 4.3 (Gaussian conditioning)**.** *Assume Σ is positive definite. Then, for any fixed $\mathbf{x}_2$, the conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2$ is Gaussian:*

$$\mathbf{X}_1 \,\big|\, \mathbf{X}_2 = \mathbf{x}_2 \;\sim \mathcal{N}\Big(\boldsymbol{\mu}_{1|2},\, \Sigma_{1|2}\Big), \tag{5}$$

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \tag{6}$$

$$\Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}. \tag{7}$$

Moreover, the minimum mean-squared error of $\mathbf{X}_1$ given $\mathbf{X}_2$ satisfies

$$\mathrm{mmse}(\mathbf{X}_1 \mid \mathbf{X}_2) \coloneqq \mathbb{E}\big[\|\mathbf{X}_1 - \mathbb{E}[\mathbf{X}_1 \mid \mathbf{X}_2]\|^2\big] = \mathrm{Tr}(\Sigma_{1|2}).$$

Proof. The joint density satisfies

$$f_{\mathbf{X}}(\mathbf{x}) \propto \exp\big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\big),$$

and we write the inverse covariance (or precision) matrix in blocks with

$$\Sigma^{-1} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix},$$

where J_{11}, J_{22} are positive definite (because they are principal submatrices of Σ^{-1} which is positive definite) and $J_{21} = J_{12}^\top$. The conditional pdf of $\mathbf{X}_1$ given $\mathbf{X}_2$ satisfies

$$\begin{aligned} f_{\mathbf{X}_1|\mathbf{X}_2}(\mathbf{x}_1|\mathbf{x}_2) &= \frac{f_{\mathbf{X}_1,\mathbf{X}_2}(\mathbf{x}_1,\mathbf{x}_2)}{f_{\mathbf{X}_2}(\mathbf{x}_2)} \\ &\propto \exp\big(-\tfrac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \tfrac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top \Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\big) \\ &\propto \exp\big(-Q(\mathbf{x}_1)\big), \end{aligned}$$

where (since pdfs normalize to 1) we can drop all terms that do not involve $\mathbf{x}_1$ to get

$$Q(\mathbf{x}_1) = \tfrac{1}{2}\mathbf{x}_1^\top J_{11}\mathbf{x}_1 - \mathbf{x}_1^\top(J_{11}\boldsymbol{\mu}_1 - J_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)) + \text{const in } \mathbf{x}_1.$$

Since pdfs must normalize to 1, it is sufficient to focus on proportionality and only keep only terms that involve $\mathbf{x}_1$. The conditional distribution of $\mathbf{X}_1$ given $\mathbf{X}_2 = \mathbf{x}_2$ is Gaussian because it matches a Gaussian on all terms that depend on $\mathbf{x}_1$ (i.e., terms that are constant with respect to $\mathbf{x}_1$ are determined by normalization). In addition, we can determine the parameters of the conditional distribution by matching terms with

$$\begin{aligned} f_{\mathbf{X}_1|\mathbf{X}_2=\mathbf{x}_2}(\mathbf{x}_1) &\propto \exp\Big(-\tfrac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_{1|2})^\top \Sigma_{1|2}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_{1|2})\Big) \\ &\propto \exp\Big(-\tfrac{1}{2}\mathbf{x}_1^\top \Sigma_{1|2}^{-1}\mathbf{x}_1 + \mathbf{x}_1^\top \Sigma_{1|2}^{-1}\boldsymbol{\mu}_{1|2} + \text{const in } \mathbf{x}_1\Big). \end{aligned}$$

Thus, we can write $\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$ with covariance $\Sigma_{1|2} = J_{11}^{-1}$ and mean $\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - J_{11}^{-1}J_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2)$.

Equivalently, one could observe that

$$Q(\mathbf{x}_1) = \tfrac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^\top J_{11}(\mathbf{x}_1-\boldsymbol{\mu}_1) + (\mathbf{x}_1-\boldsymbol{\mu}_1)^\top J_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2) + \text{const in } \mathbf{x}_1$$

is a quadratic function of $\mathbf{x}_1$ for fixed $\mathbf{x}_2$ and completing the square yields

$$Q(\mathbf{x}_1) = \tfrac{1}{2}\big(\mathbf{x}_1-\boldsymbol{\mu}_{1|2}\big)^\top J_{11}\big(\mathbf{x}_1-\boldsymbol{\mu}_{1|2}\big) + \text{const}, \quad \text{with} \quad \boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - J_{11}^{-1}J_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2).$$

Now, we use block-inversion identities in Appendix A to compute the precision blocks explicitly. Applying the alternative block inverse formula to $\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ gives

$$\Sigma_{1|2}^{-1} = J_{11} = (\Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})^{-1}.$$

For J_{12}, the alternative block inverse formula gives $J_{12} = -J_{11}\Sigma_{12}\Sigma_{22}^{-1}$. Therefore, we have

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 - J_{11}^{-1}J_{12}(\mathbf{x}_2-\boldsymbol{\mu}_2) = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2).$$

Conceptually, this formula makes sense because $\Sigma_{22}(\mathbf{x}_2 - \boldsymbol{\mu}_2)$ gives the whitened $\mathbf{x}_2$ deviation and multiplication by Σ_{12} linearly maps the whitened $\mathbf{x}_2$ deviation to its effect on $\mathbf{x}_1$.

For the MMSE claim, note that for fixed $\mathbf{x}_2$,

$$\mathbb{E}\big[\|\mathbf{X}_1 - \mathbb{E}[\mathbf{X}_1 \mid \mathbf{X}_2]\|^2 \,\big|\, \mathbf{X}_2 = \mathbf{x}_2\big] = \mathrm{Tr}\left(\mathrm{Cov}(\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2)\right) = \mathrm{Tr}(\Sigma_{1|2}),$$

and $\Sigma_{1|2}$ is independent of $\mathbf{x}_2$. Taking expectation over $\mathbf{X}_2$ yields $\mathrm{mmse}(\mathbf{X}_1 \mid \mathbf{X}_2) = \mathrm{Tr}(\Sigma_{1|2})$. $\square$

Example 4.4. Consider the setup of the previous theorem where $\mathbf{X} = (\mathbf{X}_1, \mathbf{X}_2) \in \mathbb{R}^4$ is jointly Gaussian with

$$\boldsymbol{\mu}_1 = \begin{bmatrix}0\\0\end{bmatrix}, \quad \boldsymbol{\mu}_2 = \begin{bmatrix}1\\-1\end{bmatrix}, \quad \Sigma_{11} = \begin{bmatrix}2 & 0.3\\0.3 & 1.5\end{bmatrix}, \; \Sigma_{22} = \begin{bmatrix}1.2 & 0.4\\0.4 & 2.0\end{bmatrix}, \; \Sigma_{12} = \begin{bmatrix}0.5 & 0.2\\0.1 & 0.3\end{bmatrix}.$$

For the observation $\mathbf{x}_2 = \begin{bmatrix}1.2\\-0.6\end{bmatrix}$, Theorem 4.3 gives $\mathbf{X}_1 \mid \mathbf{X}_2 = \mathbf{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_{1|2}, \Sigma_{1|2})$ with

$$\boldsymbol{\mu}_{1|2} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \qquad \Sigma_{1|2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Numerically, we have

$$\Sigma_{22}^{-1} \approx \begin{bmatrix}0.8929 & -0.1786\\-0.1786 & 0.5357\end{bmatrix}, \quad \Sigma_{12}\Sigma_{22}^{-1} \approx \begin{bmatrix}0.4107 & 0.0179\\0.0357 & 0.1429\end{bmatrix}.$$

Thus, with $\mathbf{x}_2 - \boldsymbol{\mu}_2 = [0.2,\ 0.4]^\top$, we obtain

$$\boldsymbol{\mu}_{1|2} \approx \begin{bmatrix}0.0893\\0.0643\end{bmatrix}, \qquad \Sigma_{1|2} \approx \begin{bmatrix}1.7911 & 0.2536\\0.2536 & 1.4536\end{bmatrix}.$$

Remark 4.5. It is quite common to ignore the overall normalization constant when analyzing distributions by using statements like $f(x) \propto g(x)$ which means that $f(x) = a\,g(x)$ for some unspecified constant a. For distributions, this implies equality because the difference will vanish when both are normalized to integrate to 1. Moreover any function proportional to $e^{-\mathbf{z}^\top J\mathbf{z} + \mathbf{c}^\top \mathbf{z}}$ will normalize to a Gaussian when J is positive definite.

4.4 Conditioning on Noisy Linear Observations

Theorem 4.6 (Noisy Linear Observations)**.** *Consider the model*

$$\mathbf{Y} = H\mathbf{X} + \mathbf{Z}, \qquad \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_\mathbf{X}, \Sigma_\mathbf{X}), \quad \mathbf{Z} \sim \mathcal{N}(\mathbf{0}, \Sigma_\mathbf{Z}) \;\; \textit{with } \Sigma_\mathbf{Z} \succ 0, \quad \mathbf{X} \perp\!\!\!\perp \mathbf{Z}.$$

Then, the posterior of $\mathbf{X}$ *given* $\mathbf{Y}$ *is the Gaussian* $f_{\mathbf{X}|\mathbf{Y}} \sim \mathcal{N}(\boldsymbol{\mu}', \Sigma')$ *with*

$$\boldsymbol{\mu}' = \boldsymbol{\mu}_\mathbf{X} + \Sigma' H^\top \Sigma_\mathbf{Z}^{-1}\big(\mathbf{y} - H\boldsymbol{\mu}_\mathbf{X}\big), \tag{8}$$

$$\Sigma' = \Sigma_\mathbf{X} - \Sigma_\mathbf{X} H^\top \big(H\Sigma_\mathbf{X} H^\top + \Sigma_\mathbf{Z}\big)^{-1} H\Sigma_\mathbf{X}. \tag{9}$$

Also, the minimum mean-squared error of $\mathbf{X}$ *given* $\mathbf{Y}$ *satisfies*

$$\mathrm{mmse}(\mathbf{X} \mid \mathbf{Y}) \coloneqq \mathbb{E}\big[\|\mathbf{X} - \mathbb{E}[\mathbf{X} \mid \mathbf{Y}]\|^2\big] = \mathrm{Tr}(\Sigma').$$

Proof. Using this setup, we define a new Gaussian random vector $\mathbf{W}$ satisfying

$$\mathbf{W} = \begin{bmatrix} \mathbf{X} \\ \mathbf{Y} \end{bmatrix} = \begin{bmatrix} I \\ H \end{bmatrix} \mathbf{X} + \begin{bmatrix} \mathbf{0} \\ \mathbf{Z} \end{bmatrix}, \qquad \boldsymbol{\mu}_{\mathbf{W}} = \begin{bmatrix} \boldsymbol{\mu}_{\mathbf{X}} \\ H\boldsymbol{\mu}_{\mathbf{X}} \end{bmatrix}, \qquad \Sigma_{\mathbf{W}} = \begin{bmatrix} \Sigma_{\mathbf{X}} & \Sigma_{\mathbf{X}} H^\top \\ H\Sigma_{\mathbf{X}} & H\Sigma_{\mathbf{X}} H^\top + \Sigma_{\mathbf{Z}} \end{bmatrix}.$$

Then, we apply Theorem 4.3 with $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_{\mathbf{X}}$, $\boldsymbol{\mu}_2 = H\boldsymbol{\mu}_{\mathbf{X}}$, $\Sigma_{11} = \Sigma_{\mathbf{X}}$, $\Sigma_{12} = \Sigma_{\mathbf{X}} H^\top$ and $\Sigma_{22} = H\Sigma_{\mathbf{X}} H^\top + \Sigma_{\mathbf{Z}}$ to obtain the formulas

$$\begin{aligned} \boldsymbol{\mu}_{1|2} &= \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2) = \boldsymbol{\mu}_{\mathbf{X}} + \Sigma_{\mathbf{X}} H^\top (H\Sigma_{\mathbf{X}} H^\top + \Sigma_{\mathbf{Z}})^{-1}(\mathbf{x}_2 - H\boldsymbol{\mu}_{\mathbf{X}}), \\ \Sigma_{1|2} &= \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \left(\Sigma_{\mathbf{X}}^{-1} + H^\top \Sigma_{\mathbf{Z}}^{-1} H\right)^{-1}. \end{aligned}$$

The Woodbury identity is used to rewrite the stated expressions above for $\boldsymbol{\mu}'$ and Σ'. Also, we have $\mathrm{mmse}(\mathbf{X} \mid \mathbf{Y}) = \mathrm{Tr}(\Sigma_{1|2})$. □

Example 4.7 (Two correlated sensors)**.** Define $X \sim \mathcal{N}(0,1)$, sensors $Y_1 = X + Z_1$, $Y_2 = X + Z_2$ with $\mathrm{Var}(Z_i) = \sigma_i^2$ and $\mathrm{Cov}(Z_1, Z_2) = \rho\sigma_1\sigma_2$. Stacking $\mathbf{Y} = (Y_1, Y_2)^\top$ and defining $H = \begin{bmatrix} 1 & 1 \end{bmatrix}^\top$, one obtains $\Sigma' = (1 + \mathbf{1}^\top \Sigma_{\mathbf{Z}}^{-1} \mathbf{1})^{-1}$ and $\mu' = \Sigma' \, \mathbf{1}^\top \Sigma_{\mathbf{Z}}^{-1} \mathbf{y}$. This estimate reduces the relative contribution of the noisier sensor.

For a concrete numeric illustration, let $\sigma_1^2 = 0.25$, $\sigma_2^2 = 1$, $\rho = 0.3$, and observe $\mathbf{y} = (0.8, -0.2)^\top$. Then

$$\Sigma_{\mathbf{Z}} = \begin{bmatrix} 0.25 & 0.15 \\ 0.15 & 1 \end{bmatrix}, \quad \Sigma_{\mathbf{Z}}^{-1} = \frac{1}{0.2275} \begin{bmatrix} 1 & -0.15 \\ -0.15 & 0.25 \end{bmatrix}.$$

Hence $1 + \mathbf{1}^\top \Sigma_{\mathbf{Z}}^{-1} \mathbf{1} = \; 1 + \frac{0.95}{0.2275} \approx 5.1758$, so $\Sigma' \approx 0.1932$. Moreover,

$$\mu' = \Sigma' \, \mathbf{1}^\top \Sigma_{\mathbf{Z}}^{-1} \mathbf{y} = \Sigma' \, \frac{[\,0.83,\, -0.17\,]\,\mathbf{1}}{0.2275} \approx 0.1932 \times 2.9022 \approx 0.5607.$$

Remark 4.8 (MMSE and MAP estimates). Under Gaussian priors/likelihoods with quadratic loss, the posterior mean is both the MMSE estimate due to symmetry and the MAP estimate due to unimodality. Indeed, if $X \mid Y = \mathbf{y} \sim \mathcal{N}(\boldsymbol{\mu}', \Sigma')$, then

$$\log p(\mathbf{x} \mid \mathbf{y}) = \text{const} - \tfrac{1}{2}(\mathbf{x} - \boldsymbol{\mu}')^\top \Sigma'^{-1} (\mathbf{x} - \boldsymbol{\mu}'),$$

whose gradient vanishes uniquely at $\mathbf{x} = \boldsymbol{\mu}'$. Thus the MAP estimate is $\hat{\mathbf{x}}_{\mathrm{MAP}} = \boldsymbol{\mu}'$. Under squared error, the MMSE estimator is the conditional mean and $\hat{\mathbf{x}}_{\mathrm{MMSE}} = \mathbb{E}[\mathbf{X} \mid \mathbf{Y} = \mathbf{y}] = \boldsymbol{\mu}'$.

5 Gaussian Processes

5.1 Introduction

A Gaussian process (GP) is a stochastic process with a single index such that joint distributions of samples are always Gaussian. Let the index set $\mathcal{T}$ be an arbitrary discrete or continuous set. Note that, from now on, we will use the shorthand $[N] \coloneqq \{1, 2, \ldots, N\} \subset \mathbb{N}$.

Definition 5.1. A stochastic process $\{X_t\}_{t \in \mathcal{T}}$ is a *Gaussian process* if for every finite set of indices $t_1, \ldots, t_m \in \mathcal{T}$, the vector $(X_{t_1}, \ldots, X_{t_m})$ is jointly Gaussian.

Definition 5.2 (Mean and covariance)**.** For a Gaussian process, the *mean function* is $\mu(t) \coloneqq \mathbb{E}[X_t]$ for $t \in \mathcal{T}$ and the *covariance function* is $k(s,t) \coloneqq \mathrm{Cov}(X_s, X_t) = \mathbb{E}\big[(X_s - \mu(s))(X_t - \mu(t))\big]$ for $s, t \in \mathcal{T}$. It is called *centered* if $\mu(\cdot) \equiv 0$.

Proposition 5.3 (Characterization)**.** *For a Gaussian process with $\mu(t) = \mathbb{E}[X_t]$ and $k(s,t) = \mathrm{Cov}(X_s, X_t)$ and any $t_1, \ldots, t_m \in \mathcal{T}$, we have $(X_{t_1}, \ldots, X_{t_m})^\top \sim \mathcal{N}\big([\mu(t_i)]_{i\in[m]}, [k(t_i,t_j)]_{i,j\in[m]}\big)$. Conversely, any pair (μ, k) where $K = [k(t_i,t_j)]_{i,j\in[m]}$ is positive definite for every finite set of times defines a GP.*

Proof. For the forward direction, each finite subvector is multivariate normal by definition. For the converse, any family of finite-dimensional Gaussians with moments specified by a positive-definite kernel is consistent under marginalization. Thus, one can apply Kolmogorov's Extension Theorem [2]. □

Theorem 5.4 (Gaussian Process Regression)**.** *Let X_t be a GP with mean $\mu(t) = 0$ and covariance $k(s,t)$. Let $t_1, \ldots, t_n \in \mathcal{T}$ be n distinct indices and define $Y_i = X_{t_i} + Z_i$ with $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0, \sigma^2)$ for $i \in [n]$. Then, for any $t' \in \mathcal{T} \setminus \{t_1, \ldots, t_n\}$, the conditional distribution of $X_{t'}$ given $\mathbf{Y} = \mathbf{y}$ is Gaussian with mean and variance*

$$m' = \mathbb{E}[X_{t'} \mid \mathbf{Y} = \mathbf{y}] = \mathbf{r}^\top (K + \sigma^2 I)^{-1} \mathbf{y}, \tag{10}$$

$$v' = \mathrm{Var}(X_{t'} \mid \mathbf{Y} = \mathbf{y}) = k(t',t') - \mathbf{r}^\top (K + \sigma^2 I)^{-1} \mathbf{r}, \tag{11}$$

where $K \in \mathbb{R}^{n\times n}$ with $K_{ij} = k(t_i, t_j)$ is the covariance matrix of $(X_{t_1}, \ldots, X_{t_n})$ and $\mathbf{r} \in \mathbb{R}^n$ is the cross-correlation vector with $X_{t'}$ defined by $[\mathbf{r}]_i = k(t_i, t')$.

Proof. For this construction, the joint distribution of $(\mathbf{Y}, X_{t'})$ is Gaussian with mean 0 and covariance

$$K = \begin{bmatrix} K + \sigma^2 I & \mathbf{r} \\ \mathbf{r}^\top & k(t',t') \end{bmatrix}.$$

Thus, we can apply Theorem 4.3 to get formulas for m' and v'. Notice that the posterior mean is a linear combination of the observed values in $\mathbf{y}$. □

5.2 Examples and Spectral Methods

If the index set forms an additive group, then the correlation may be a function only of the difference between s and t.

Definition 5.5 (Stationary)**.** A stochastic process is strictly stationary if all its finite-dimensional distributions are invariant under translation. It is (wide-sense) stationary if $\mu(t)$ is constant and the covariance depends only on the difference between indices: $k(s,t) = r(t-s)$ for all $s, t \in \mathcal{T}$ and some autocovariance function $r(\tau)$.

Example 5.6 (Discrete time)**.** Let $X_t = aX_{t-1} + \sqrt{1-a^2}\, W_t$ for $t \in \mathcal{T} = \mathbb{Z}$ with $|a| < 1$ and $W_t \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. This is the unique stationary Gaussian process with $\mathbb{E}[X_t] = 0$, $\mathrm{Var}(X_t) = 1$, and autocovariance $r(\tau) = \mathrm{Cov}(X_t, X_{t+\tau}) = a^{|\tau|}$.

To see this, we iterate the defining equation to get $X_t = \sqrt{1-a^2} \sum_{k=0}^{\infty} a^k W_{t-k}$ for all $t \in \mathcal{T}$, where the sum converges almost surely because $|a| < 1$. Thus, X_t is a linear combination of standard Gaussians and it follows that X_t is Gaussian with mean $\mu(t) = 0$ and variance $\mathrm{Var}(X_t) = (1-a^2)\sum_{k\geq 0} a^{2k} = 1$. For $\tau \geq 0$, the covariance function is given by

$$\begin{aligned} k(t, t+\tau) &= \mathrm{Cov}(X_t, X_{t+\tau}) \\ &= (1-a^2) \sum_{k\geq 0} \sum_{\ell \geq 0} a^{k+\ell}\, \mathrm{Cov}(W_{t-k}, W_{t+\tau-\ell}) \end{aligned}$$

$$= (1-a^2)\sum_{k\geq 0} a^{2k+\tau}$$
$$= a^{\tau}.$$

Since $k(t, t+\tau)$ is independent of t, we can define the autocovariance $r(\tau) = k(t, t+\tau)$ and note that choosing $t = -\tau$ shows that $r(-\tau) = r(\tau)$.

The following example connects this material with earlier results. First, we know the Wiener or linear MMSE estimate of one random variable given others follows directly from the inner product space viewpoint and normal equations. Second, recall from Section 4 and Theorem 4.3 that, for jointly Gaussian variables under squared loss, the posterior mean also equals the linear MMSE estimator.

Example 5.7 (Optimal Wiener filter)**.** For the autoregressive process $X_t = aX_{t-1} + \sqrt{1-a^2}\, W_t$ with $\mathrm{Var}(X_t) = 1$, the covariance of (X_{-1}, X_0, X_{+1}) is

$$K = \begin{bmatrix} 1 & a & a^2 \\ a & 1 & a \\ a^2 & a & 1 \end{bmatrix}.$$

Thus, the linear MMSE estimate of X_0 given X_{-1} and X_1 equals $\mathbf{w}^\top \mathbf{x}$ with $\mathbf{w} = R^{-1}\mathbf{r}$ where $R = \begin{bmatrix} 1 & a^2 \\ a^2 & 1 \end{bmatrix}$ is the covariance of $\mathbf{x} = (X_{-1}, X_{+1})$ and $\mathbf{r} = [a,\ a]^\top$ is the cross-covariance with X_0.

Likewise, by Theorem 4.3, letting $\mathbf{X}_1 = X_0$ and $\mathbf{X}_2 = (X_{-1}, X_{+1})^\top$ with zero mean and blocks $\Sigma_{11} = 1$, $\Sigma_{22} = R$, and $\Sigma_{12} = \mathbf{r}^\top$, we obtain

$$\mathbb{E}[X_0 \mid (X_{-1}, X_1) = \mathbf{x}] = \Sigma_{12}\Sigma_{22}^{-1}\mathbf{x} = \mathbf{r}^\top R^{-1}\mathbf{x} = (R^{-1}\mathbf{r})^\top \mathbf{x} = \mathbf{w}^\top \mathbf{x},$$
$$\mathrm{Var}(X_0 \mid \mathbf{X}_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - \mathbf{r}^\top R^{-1}\mathbf{r}.$$

Since $R^{-1}\mathbf{r} = \dfrac{a}{1+a^2}[1,\ 1]^\top$ and $1 - \mathbf{r}^\top R^{-1}\mathbf{r} = 1/(1+a^2)$, conditioning on the neighbors yields

$$\mathbb{E}[X_0 \mid X_{-1} = x_{-1}, X_{+1} = x_{+1}] = \frac{a}{1+a^2}\big(x_{-1} + x_{+1}\big), \quad \mathrm{Var}(X_0 \mid X_{-1}, X_{+1}) = \frac{1}{1+a^2}.$$

For a numerical example with $a = 0.8$, $x_{-1} = 0.5$, $x_{+1} = -0.1$, we get $\hat{x}_0 = \frac{0.8}{1+0.64}(0.5 - 0.1) \approx 0.195$, which matches the formulas above. The final form of the conditional expectation illustrates the "matched-filter" perspective where the optimal estimator is shift invariant because the signal is shift invariant.

Example 5.8 (Continuous time)**.** Let $W(t)$ denote a standard zero-mean Gaussian white noise with covariance $k_W(t, t+\tau) = \delta(\tau)$, where $\delta(\cdot)$ denotes the Dirac delta. For $h(t)$ satisfying $\int_{\mathbb{R}} |h(t)|^2 dt < \infty$, we can define the random process $X(t)$ via convolution[1]

$$X(t) = (h * W)(t) \ := \ \int_{\mathbb{R}} h(\tau)\, W(t-\tau)\ \mathrm{d}\tau.$$

Then, $X(t)$ is a stationary Gaussian stationary process with autocovariance

$$r(\tau) = k(t, t+\tau)$$

[1]The process $X(t)$ is well-defined and can be constructed in a fully rigorous manner even though the construction given here based on $W(t)$ is a less formal shortcut.

$$
\begin{aligned}
&= \operatorname{Cov}\big(X(t), X(t+\tau)\big) \\
&= \mathbb{E}\left[\iint h(u)\, h(v)\, W(t-u) W(t+\tau-v)\, du\, dv\right] \\
&= \iint h(u)\, h(v)\, \mathbb{E}\left[W(t-u) W(t+\tau-v)\right]\, du\, dv \\
&= \iint h(u)\, h(v)\, \delta(\tau-(v-u))\, du\, dv \\
&= \int_{\mathbb{R}} h(u)\, h(u+\tau)\, du.
\end{aligned}
$$

Power spectral density (PSD). For a zero-mean, wide-sense stationary (WSS) process $X(t)$ with autocovariance $r_X(\tau) = \mathbb{E}[X(t)X(t+\tau)]$, define the finite-time Fourier transform

$$
X_T(\omega) \;\coloneqq\; \int_{-T}^{T} X(t)\, e^{-\mathrm{i}\omega t}\, dt
$$

and the expected finite-time normalized power spectral density is given by

$$
S_X^{(T)}(\omega) \;\coloneqq\; \frac{1}{2T}\, \mathbb{E}\big[\, |X_T(\omega)|^2 \,\big].
$$

If the autocorrelation function is absolutely integrable, then the following limit exists and gives the two-sided power spectral density:

$$
S_X(\omega) \;=\; \lim_{T\to\infty} S_X^{(T)}(\omega).
$$

Theorem 5.9 (Wiener–Khinchin)**.** *If X is zero-mean WSS and r_X is absolutely integrable, then*

$$
S_X(\omega) = \int_{\mathbb{R}} r_X(\tau)\, e^{-\mathrm{i}\omega\tau}\, d\tau, \qquad r_X(\tau) = \frac{1}{2\pi} \int_{\mathbb{R}} S_X(\omega)\, e^{\mathrm{i}\omega\tau}\, d\omega.
$$

Proof sketch. Compute

$$
S_X^{(T)}(\omega) = \frac{1}{2T} \int_{-T}^{T}\!\!\int_{-T}^{T} \mathbb{E}[X(t)X(s)]\, e^{-\mathrm{i}\omega(t-s)}\, dt\, ds = \frac{1}{2T} \int_{-T}^{T}\!\!\int_{-T}^{T} r_X(t-s)\, e^{-\mathrm{i}\omega(t-s)}\, dt\, ds.
$$

Let $u = (t+s)/2$ and $\tau = t-s$. The u–integration yields $(2T - |\tau|)_+$, where $(a)_+ \coloneqq \max(a, 0)$. Dividing by $2T$ and taking $T \to \infty$, dominated convergence (using absolute integrability of r_X) gives $S_X(\omega) = \int_{\mathbb{R}} r_X(\tau)\, e^{-\mathrm{i}\omega\tau}\, d\tau$. The inverse formula follows from Fourier inversion under the stated convention. □

Filtered white noise. If $X(t) = (h * W)(t)$ with W unit white noise ($\mathbb{E}[W(t)W(s)] = \delta(t-s)$), then $r_X(\tau) = \int_{\mathbb{R}} h(u)\, h(u+\tau)\, du$. By Wiener–Khinchin and the correlation theorem,

$$
S_X(\omega) = \int_{\mathbb{R}} r_X(\tau)\, e^{-\mathrm{i}\omega\tau}\, d\tau \;=\; |H(\omega)|^2\,, \tag{12}
$$

$$
H(\omega) = \int_{\mathbb{R}} h(t)\, e^{-\mathrm{i}\omega t}\, dt. \tag{13}
$$

These expressions use the Fourier transform convention adopted in the Wiener–Khinchin section.

Example 5.10 (Wiener filter for filtered white noise)**.** If $Y = X + N$ with X WSS and N independent WSS, the optimal LTI estimator in frequency is

$$G^\star(\omega) = \frac{S_{XY}(\omega)}{S_{YY}(\omega)} = \frac{S_X(\omega)}{S_X(\omega) + S_N(\omega)}.$$

For $X = h * W$ driven by unit white noise (so $S_X(\omega) = |H(\omega)|^2$) and white measurement noise with PSD σ_N^2, this gives $G^\star(\omega) = \frac{|H(\omega)|^2}{|H(\omega)|^2 + \sigma_N^2}$. In this setting, the posterior mean (and linear MMSE estimator) is obtained by passing Y through the LTI filter with frequency response $G^\star(\omega)$ above.

5.3 Gaussian Random Fields

In the previous section, we considered GPs with a single index. But, this was only to simplify the description at first. All the previous statements actually apply immediately to Gaussian processes indexed by arbitrary sets. Only the examples are limited to one dimension. Thus, we can index by vectors and the resulting objects are called Gaussian random fields (GRFs). To lighten notation, we will still use the indices $s, t \in \mathcal{T}$ though in some cases they may represent vectors.

Definition 5.11. A collection $\{X_t\}_{t\in\mathcal{T}}$ is a *Gaussian random field* if, for any finite locations $t_1, \dots, t_m \in \mathcal{T}$, the vector $\mathbf{X} = (X_{t_1}, \dots, X_{t_m})$ is jointly Gaussian $\sim \mathcal{N}\big([\mu(t_i)]_{i\in[m]}, [k(t_i, t_j)]_{i,j\in[m]}\big)$.

If the index set $\mathcal{T}$ is equipped with the structure of an abelian group, then Definition 5.5 can be used to define stationarity without change. A stationary GRF is called *isotropic* if $\mathcal{T}$ is equipped with the structure of a normed vector space and the covariance function $k(s, t)$ is a function of $\|s - t\|$. For example, a GRF defined by *radial basis functions* (RBFs) has, for some $a, b \in (0, \infty)$, the covariance function

$$k(s, t) = a \exp\left(-\frac{\|s - t\|^2}{2b^2}\right).$$

Remark 5.12. These ideas also extend naturally to vector-valued GRFs, $\mathbf{X}_t \in \mathbb{R}^p$, where the covariance function $K(s, t) \in \mathbb{R}^{p\times p}$ becomes matrix-valued and is defined to be the cross-covariance $K(s, t) = \mathbb{E}[(\mathbf{X}_s - \boldsymbol{\mu}(s))(\mathbf{X}_t - \boldsymbol{\mu}(t))^\top]$. This is similar to choosing $\mathcal{T} = \mathcal{T}' \times [p]$ so that each $t = (t', j) \in \mathcal{T}$ indexes the j-th element located at the point t'.

6 Reproducing Kernel Hilbert Spaces

6.1 Introduction

Consider the problem of estimating an unknown function $f\colon \mathcal{X} \to \mathbb{R}$ from input-output observations $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ satisfying $y_i = f(x_i)$ for $i \in \mathbb{N}$. For any candidate $g\colon \mathcal{X} \to \mathbb{R}$, we consider the risk (or loss) associated with using g instead of f. Using $\ell(\hat{y}, y)$ for the loss due to estimating $\hat{y}$ when y is correct, the *empirical risk* for the first N data points is

$$L_N(g, f) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \ell(g(x_i), f(x_i)) = \int_{\mathcal{X}} \ell(g(x), f(x))\, d\nu_N(x),$$

where ν_N is the empirical distribution of $\{x_1, \dots, x_N\}$. If we assume that ν_N converges weakly to ν as $N \to \infty$, then we can also define

$$L(g, f) = \lim_{N\to\infty} L_N(g, f) \coloneqq \int_{\mathcal{X}} \ell(g(x), f(x))\, d\nu(x).$$

One can think of ν as the distribution from which the evaluation points are drawn. In this work, we will focus exclusively on the squared-error loss $\ell(\hat{y}, y) = (\hat{y} - y)^2$.

To compute an estimate of $f(x)$, a standard approach is to use a smoothing kernel $k(x, x')$ to give the estimate

$$g_\theta(x) = \sum_{i=1}^{N} \theta_i \, k(x, x_i),$$

where the parameters $\theta_1, \ldots, \theta_N$ are chosen to minimize the overall loss function

$$\mathcal{L}(\theta) = L_N(g_\theta, f) + R(g_\theta).$$

The first term in the loss is the empirical risk and the second term is a regularization term, which is often associated with a prior distribution on the set of functions.

Another common approach to fitting data is to first choose a non-linear feature map and then learn a linear function of the features. For example, we can let $\phi\colon \mathcal{X} \to \ell^2$ map $\mathcal{X}$ to the standard Hilbert space ℓ^2 of square summable sequences whose inner product is the dot product. Then, the Riesz representation theorem implies that any linear functional (i.e., a mapping from ℓ^2 to $\mathbb{R}$) of the feature vector $\phi(x)$ can be written as an inner product $\langle \phi(x),\, h\rangle_{\ell^2}$, where $h \in \ell^2$ defines the linear functional. Thus, this approach leads to

$$g_h(x) = \langle \phi(x),\, h\rangle_{\ell^2}\,.$$

If the feature map is matched to the covariance kernel, then we can estimate f by minimizing the loss

$$\mathcal{L}(h) = L_N(g_h, f) + \frac{1}{2}\|h\|_{\ell^2}^2.$$

This will match the previous solution if the regularization term R is chosen correctly.

Lastly, one can also treat the data, $(x_i, y_i) \in \mathcal{X} \times \mathbb{R}$ for $i \in [N]$, as being generated by a Gaussian random field with covariance function $k(x, x')$. In this case, one can use ideas from GRFs to estimate the posterior mean $g(x)$ of $f(x)$ given the observed data.

Surprisingly, all these perspectives are closely related and their optimal estimates are equal.

6.2 Kernel functions and feature maps

Definition 6.1 (Positive-definite kernel)**.** For any set $\mathcal{X}$, a function $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is symmetric positive semidefinite if $k(x, x') = k(x', x)$ for all $x, x' \in \mathcal{X}$ and, for any $x_1, \ldots, x_n \in \mathcal{X}$ and $\mathbf{c} \in \mathbb{R}^n$, we have $\sum_{i,j} c_i c_j k(x_i, x_j) \geq 0$.

Remark 6.2. While the above kernel is positive semidefinite, this is the standard naming convention.

Theorem 6.3 (Mercer's theorem)**.** *If k is continuous, symmetric, and positive semidefinite on a compact domain $(\mathcal{X}, \nu)$ with ν a finite measure, then the integral operator*

$$(Tf)(x) = \int_{\mathcal{X}} k(x, x') f(x') \, \mathrm{d}\nu(x')$$

is self-adjoint, positive, and compact. Thus, the eigenvectors $\{\psi_i\colon \mathcal{X} \to \mathbb{R}\}_{i \in \mathbb{N}}$ form a complete orthonormal basis for the range of k with

$$\langle \psi_i,\, \psi_j\rangle_{L^2(\nu)} = \int_{\mathcal{X}} \psi_i(x)\psi_j(x) \, \mathrm{d}\nu(x) = \delta_{i,j}.$$

Thus, we can write

$$k(x, x') = \sum_{i=1}^{\infty} \lambda_i \, \psi_i(x)\psi_i(x'),$$

where convergence is uniform on $\mathcal{X} \times \mathcal{X}$*.*

Proof Idea. If $\mathcal{X} = \{x_1, \ldots, x_M\}$ is a finite set, then this follows directly from the eigenvalue decomposition of the covariance matrix K with entries $K_{i,j} = k(x_i, x_j)$ for $i, j \in [M]$. In particular, the matrix is guaranteed to have an orthonormal set of eigenvectors (due to symmetry) with non-negative eigenvalues (due to positivity) that span the space.

For the continuous case, the same idea generalizes naturally to integral operators that are compact, self-adjoint, and positive. In this case, a compact operator has a countable set of eigenvalues, a self-adjoint operator (the infinite-dimensional analogue of a symmetric matrix) has orthonormal eigenfunctions that span the range, and a positive operator has non-negative eigenvalues. □

Remark 6.4. Let us take a moment to consider the effect of the background measure ν. While one gets an orthonormal basis and decomposition of the kernel for any positive ν, choosing the proper ν does matter. For example, if one computes $\|f - g\|^2_{L^2(\nu)}$ with ν equal to the true data distribution, then this norm equals the expected mean-squared error of the approximation when the x values are drawn i.i.d. from ν.

Definition 6.5. The *feature map* $\phi\colon \mathcal{X} \to \ell^2$ associated with a Mercer kernel $k(x, x')$ is defined by its eigenvalue decomposition (i.e., $\{\lambda_i\}_{i\in\mathbb{N}}$ and $\{\psi_i\}_{i\in\mathbb{N}}$) using

$$\phi(x) \coloneqq (\phi_1(x), \phi_2(x), \phi_3(x) \ldots), \tag{14}$$

where $\phi_i(x) \coloneqq \sqrt{\lambda_i}\psi_i(x)$ for $i \in \mathbb{N}$.

Using these definitions and results, we observe that

$$\big\langle \phi(x),\, \phi(x') \big\rangle_{\ell^2} = \sum_{i=1}^{\infty} \lambda_i \, \psi_i(x)\psi_i(x') = k(x, x').$$

Moreover, this implies that

$$\begin{aligned} g_\theta(x) &= \sum_{i=1}^{N} \theta_i \, k(x, x_i) \\ &= \sum_{i=1}^{N} \theta_i \, \langle \phi(x),\, \phi(x_i) \rangle_{\ell^2} \\ &= \left\langle \phi(x),\, \sum_{i=1}^{N} \theta_i \phi(x_i) \right\rangle_{\ell^2} \\ &= \langle \phi(x),\, g_\theta \rangle_{\ell^2} \,, \end{aligned}$$

where we abuse notation by defining $g_\theta = \sum_{i=1}^{N} \theta_i \phi(x_i) \in \ell^2$. A key observation is that the vector g_θ is a representation of the function $g_\theta\colon \mathcal{X} \to \mathbb{R}$ where evaluation at x is given by an inner product (i.e., $g_\theta(x) = \langle \phi(x),\, g_\theta \rangle$). This also implies that all such weighted averages of kernels can be written as the inner product between the feature vector $\phi(x)$, which depends only on the evaluation point x, and another ℓ^2 vector that determines g_θ.

Remark 6.6. From a signal processing point of view, the change of variables associated with the feature map is a decorrelating transform that whitens the process.

Definition 6.7 (Reproducing kernel Hilbert space)**.** A vector space of functions $f : \mathcal{X} \to \mathbb{R}$ is called an RKHS $\mathcal{H}_k$ with kernel k if the:

1. kernel satisfies $k(\cdot, x') \in \mathcal{H}_k$ for all $x' \in \mathcal{X}$,
2. inner product is *reproducing*, $\langle f,\, k(\,\cdot\,, x')\rangle_{\mathcal{H}_k} = f(x')$ for all $f \in \mathcal{H}_k$ and $x' \in \mathcal{X}$.

Theorem 6.8. *If $k(x, x')$ is a kernel satisfying the conditions of Theorem 6.3, then it defines an RKHS spanned by the eigenfunctions $\{\psi_i\}_{i\in\mathbb{N}}$ with inner product defined by*

$$\langle \psi_i,\, \psi_j\rangle_{\mathcal{H}_k} = \frac{1}{\lambda_i}\delta_{i,j}. \tag{15}$$

Proof. From Theorem 6.3, we know that the space $\mathcal{H}_k$ equals the range of k and is spanned by $\{\psi_i\}_{i\in\mathbb{N}}$. Since $\psi_i \in \mathcal{H}_k$, we can use the reproducing property to observe that

$$\begin{aligned}
\psi_i(x') &= \big\langle \psi_i(\,\cdot\,),\, k(\,\cdot\,, x')\big\rangle_{\mathcal{H}_k} \\
&= \left\langle \psi_i(\,\cdot\,),\, \sum_{j=1}^{\infty} \lambda_j \psi_j(\,\cdot\,)\psi_j(x') \right\rangle_{\mathcal{H}_k} \\
&= \sum_{j=1}^{\infty} \lambda_j \psi_j(x')\, \big\langle \psi_i(\,\cdot\,),\, \psi_j(\,\cdot\,)\big\rangle_{\mathcal{H}_k}.
\end{aligned}$$

Taking the $L^2(\nu)$ inner product (with respect to x') of both sides with ψ_ℓ gives

$$\delta_{i,\ell} = \lambda_j \delta_{j,\ell}\, \big\langle \psi_i(\,\cdot\,),\, \psi_j(\,\cdot\,)\big\rangle_{\mathcal{H}_k}.$$

If $i = j = \ell$, then this gives $\langle \psi_i(\,\cdot\,),\, \psi_i(\,\cdot\,)\rangle_{\mathcal{H}_k} = 1/\lambda_i$. If $i \neq j = \ell$, then this gives $\langle \psi_i(\,\cdot\,),\, \psi_j(\,\cdot\,)\rangle_{\mathcal{H}_k} = 0$. Together, these establish (15). □

Remark 6.9. From this, we can better understand the RKHS inner product. Recall that k is diagonal in the orthonormal basis $\{\psi_i\}_{i\in\mathbb{N}}$ with eigenvalues $\{\lambda_i\}_{i\in\mathbb{N}}$. By computing the RKHS inner product between pairs of basis vectors in this orthonormal basis, we see that the implied Gram matrix is diagonal in that basis but its eigenvalues are the reciprocals $\{1/\lambda_i\}_{i\in\mathbb{N}}$. Thus, it is a standard inner product weighted by the inverse of the covariance function. Such an inner product naturally induces the Mahalanobis distance $d(x, x') = \sqrt{(x - x')^\top \Sigma^{-1}(x - x')}$ that appears in the exponent of the Gaussian pdf when x' is chosen to be the mean.

6.3 Connection to Gaussian Random Fields

In Section 3.1, we saw how an eigenvalue decomposition of the covariance matrix can allow one to transform a Gaussian with identity covariance into a Gaussian with general covariance. The same idea extends to GRFs with the Mercer decomposition of the covariance function.

Theorem 6.10 (Karhunen–Loève expansion)**.** *Consider a centered GRF $\{X_t\}_{t\in\mathcal{T}}$ with mean $\mu(t) = 0$ and covariance function $k(s, t)$. If $k(s, t)$ satisfies the conditions of Theorem 6.3, then there exist i.i.d. standard Gaussians $Z_i \sim \mathcal{N}(0, 1)$ such that*

$$X(t) = \sum_{i=1}^{\infty} \sqrt{\lambda_i}\, Z_i\, \psi_i(t),$$

with convergence in $L^2(\mathcal{X} \times \Omega)$ and covariance $\mathbb{E}[X(s)X(t)] = k(s, t)$.

Proof. Let $\{(\lambda_i, \psi_i)\}_{i\geq 1}$ be the Mercer eigenpairs of k on $(\mathcal{X}, \nu)$, so that $k(s,t) = \sum_{i=1}^{\infty} \lambda_i \psi_i(s)\psi_i(t)$ with $\{\psi_i\}$ orthonormal in $L^2(\nu)$ and $\lambda_i > 0$. Define the random coefficients

$$Z_i \;\coloneqq\; \frac{1}{\sqrt{\lambda_i}} \int_{\mathcal{X}} X(t)\, \psi_i(t)\, d\nu(t).$$

This is well-defined because $\mathbb{E}[X(t)^2] = k(t,t) < \infty$ is integrable under the Mercer assumptions (i.e., k is continuous on a compact domain). Then, $\mathbb{E}[Z_i] = 0$ and, for $i, j \in \mathbb{N}$, we have

$$\begin{aligned}
\operatorname{Cov}(Z_i, Z_j) &= \frac{1}{\sqrt{\lambda_i \lambda_j}} \iint \mathbb{E}[X(s)X(t)]\, \psi_i(s)\psi_j(t)\, d\nu(s)\, d\nu(t) \\
&= \frac{1}{\sqrt{\lambda_i \lambda_j}} \iint k(s,t)\, \psi_i(s)\psi_j(t)\, d\nu(s)\, d\nu(t) \\
&= \frac{1}{\sqrt{\lambda_i \lambda_j}} \sum_{\ell=1}^{\infty} \lambda_\ell \Big(\int \psi_\ell(s)\psi_i(s)\, d\nu(s) \Big) \Big(\int \psi_\ell(t)\psi_j(t)\, d\nu(t) \Big) \\
&= \frac{1}{\sqrt{\lambda_i \lambda_j}} \sum_{\ell=1}^{\infty} \lambda_\ell\, \delta_{\ell,i}\, \delta_{\ell,j} = \delta_{i,j}.
\end{aligned}$$

Thus $\{Z_i\}$ are uncorrelated standard Gaussians. Since $X(t)$ is a Gaussian random field and Z_i are linear functionals on $X(t)$, they are jointly Gaussian. Together, these imply that $Z_1, Z_2, \ldots$ are independent.

Consider the m-term approximation

$$X_m(t) \;\coloneqq\; \sum_{i=1}^{m} \sqrt{\lambda_i}\, Z_i\, \psi_i(t).$$

It has mean zero and covariance $\mathbb{E}[X_m(s)X_m(t)] = \sum_{i=1}^{m} \lambda_i\, \psi_i(s)\psi_i(t)$. Moreover,

$$\begin{aligned}
\mathbb{E} \int_{\mathcal{X}} \big(X(t) - X_m(t)\big)^2\, d\nu(t) &= \mathbb{E} \int_{\mathcal{X}} X(t)^2\, d\nu(t) - 2\,\mathbb{E} \int X(t) X_m(t)\, d\nu(t) + \mathbb{E} \int X_m(t)^2\, d\nu(t) \\
&= \sum_{i\geq 1} \lambda_i - 2 \sum_{i\leq m} \lambda_i + \sum_{i\leq m} \lambda_i = \sum_{i>m} \lambda_i \;\xrightarrow[m\to\infty]{}\; 0,
\end{aligned}$$

where we used Parseval's Theorem with the Mercer basis and the definitions above. Hence $X_m(t) \to X(t)$ in $L^2(\mathcal{X} \times \Omega)$. Taking limits of the covariance also establishes $\mathbb{E}[X(s)X(t)] = k(s,t)$. □

This highlights the connection between kernel methods and GPs. Any positive-definite kernel can be used to define a GP. Observations of that GP at fixed locations allow optimal inference of the process at other locations. Moreover, the Karhunen–Loéve expansion shows any centered GRF with kernel k can be written as a linear combination of orthogonal Gaussian coordinates.

Theorem 6.11 (Representer theorem for GP posterior mean)**.** *Consider data (x_i, y_i) with $y_i = f(x_i) + \sigma Z_i$ with $Z_i \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, where f has a GP prior with mean 0 and covariance $k(x,x')$. Then, the posterior mean is the unique minimizer of*

$$\min_{f \in \mathcal{H}_k} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \sigma^2\, \|f\|_{\mathcal{H}_k}^2,$$

and has the finite expansion $f_\alpha(x) = \sum_{i=1}^{n} \alpha_i k(x, x_i)$ with $\boldsymbol{\alpha} = (K + \sigma^2 I)^{-1}\mathbf{y}$.

Proof. From a Bayesian view, we have a GP prior and the independent conditional distributions $Y_i \mid f \sim \mathcal{N}(f(x_i), \sigma^2)$. Thus, the negative log-posterior (up to an additive constant) equals

$$\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - f(x_i))^2 \;+\; \frac{1}{2}\,\|f\|_{\mathcal{H}_k}^2 .$$

Multiplying by $2\sigma^2$ yields exactly the stated objective and the MAP estimator minimizes it. For Gaussians, the posterior is Gaussian and unimodal, so by Section 4 the MAP equals the posterior mean.

We can solve the optimization by defining $\mathbf{f} = [f(x_i)]_{i=1}^n = K\boldsymbol{\alpha}$ so that the objective over $\boldsymbol{\alpha}$ becomes $\|\mathbf{y} - K\boldsymbol{\alpha}\|_2^2 + \sigma^2\,\boldsymbol{\alpha}^\top K\,\boldsymbol{\alpha}$. Setting the gradient to zero gives $-2K(\mathbf{y} - K\boldsymbol{\alpha}) + 2\sigma^2 K\boldsymbol{\alpha} = 0$, i.e., $(K + \sigma^2 I)\boldsymbol{\alpha} = \mathbf{y}$. Thus $\boldsymbol{\alpha} = (K + \sigma^2 I)^{-1}\mathbf{y}$ and f_α equals the GP posterior mean. □

Remark 6.12. This provides yet another way to see that the posterior mean of any point in a Gaussian process can be written a linear combination of all the observed values. Of course, the coefficients of the linear combination depend on the covariance matrix induced by the locations of the observations and the point of interest. Also, this correlation matrix is essentially given by evaluating the covariance function of the process at all pairs of locations.

6.4 Simple kernels, spaces, and GRFs

- **Linear kernel:** $k(s,t) = s^\top t$.
  - RKHS: linear functions with norm $\|f\| = \|w\|_2$ where $f(x) = w^\top x$.
  - GRF: prior equivalent to Bayesian linear regression.
- **Polynomial kernel:** $k(s,t) = (s^\top t + c)^p$.
  - RKHS: finite-dimensional space of degree-$\leq p$ polynomials in lifted coordinates.
  - GRF: equivalent to Bayesian linear regression in the polynomial feature space. Let $\phi(x)$ collect all monomials up to degree p (with appropriate scaling depending on c). Then, we see that $X(t) = \theta^\top \phi(t)$ with Gaussian coefficients $\theta \sim \mathcal{N}(0, I)$ induces $\mathrm{Cov}(X(s), X(t)) = (s^\top t + c)^p$. Sample paths are a.s. equal to polynomials of degree $\leq p$.
- **RBF kernel:** $k(s,t) = \sigma^2 \exp(-\|s-t\|^2/(2\ell^2))$.
  - RKHS: consists of very smooth (indeed real-analytic) functions. One can show that the norm admits the spectral form

$$\|f\|_{\mathcal{H}_k}^2 = \frac{1}{(2\pi)^d}\int_{\mathbb{R}^d} \frac{|\widehat{f}(\omega)|^2}{S(\omega)}\,d\omega, \quad S(\omega) = (2\pi)^{d/2}\sigma^2\ell^d \exp\Big(-\tfrac{\ell^2}{2}\|\omega\|^2\Big).$$

  - GRF: stationary with spectral density given $S(\omega)$ above. The sample paths are a.s. real-analytic on $\mathbb{R}^d$ and the posterior mean function also inherits this property.

7 Worked examples

Example 7.1 (Deriving a bivariate conditional)**.** Let $(X,Y)^\top \sim \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & \rho\\ \rho & 1\end{bmatrix}\right)$ with $|\rho| < 1$. Apply Theorem 4.3 with $\mu_1 = \mu_2 = 0$, $\Sigma_{11} = \Sigma_{22} = 1$, $\Sigma_{12} = \Sigma_{21} = \rho$:

$$\mathbb{E}[X \mid Y = y] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y - \mu_2) = \rho\, y, \quad \mathrm{Var}(X \mid Y) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = 1 - \rho^2.$$

Thus $X \mid Y = y \sim \mathcal{N}(\rho y,\, 1-\rho^2)$.

Example 7.2 (Posterior for linear regression)**.** Model $\mathbf{y} = X\beta + \varepsilon$, with $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2 I)$ and prior $\beta \sim \mathcal{N}(\beta_0, \Sigma_0)$. The log-posterior (up to a constant) is

$$-\tfrac{1}{2\sigma^2}\|\mathbf{y} - X\beta\|_2^2 - \tfrac{1}{2}(\beta-\beta_0)^\top \Sigma_0^{-1}(\beta-\beta_0).$$

Completing the square in β gives a Gaussian posterior with

$$\Sigma_{\text{post}} = \Big(\Sigma_0^{-1} + \tfrac{1}{\sigma^2}X^\top X\Big)^{-1}, \qquad \beta_{\text{post}} = \Sigma_{\text{post}}\Big(\Sigma_0^{-1}\beta_0 + \tfrac{1}{\sigma^2}X^\top \mathbf{y}\Big).$$

Equivalently, view $(\beta, \mathbf{y})$ as jointly Gaussian and apply Theorem 4.3.

8 Summary

We develop a tutorial path from scalar Gaussians to multivariate normals including core inference tools (conditioning, MMSE/MAP equivalence for Gaussians, linear observation models, and Woodbury/Schur complements). We then lift these ideas to Gaussian processes/fields, showing how covariance functions determine finite-dimensional laws, how GP regression arises from Gaussian conditioning, and how spectral viewpoints (Wiener–Khinchin) connect filtering and power spectra. Through Mercer's theorem we built feature maps and RKHSs, explaining the reproducing property and the RKHS inner product as an inverse-covariance weighting (Mahalanobis geometry). The Karhunen–Loéve expansion is linked to GRFs and orthogonal Gaussian coordinates. Finally, the representer theorem shows that GP posterior means solve a regularized risk in $\mathcal{H}_k$ with closed-form coefficients.

Practical takeaways:

- For linear–Gaussian models, posteriors remain Gaussian with means/covariances computable by linear algebra; MMSE=MAP=posterior mean.
- GP regression is just Gaussian conditioning with kernels playing the role of covariances; RBF kernels yield analytic interpolants and a simple spectral density.
- RKHS methods, GP priors, and Wiener filtering are different faces of the same quadratic, kernel-driven machinery. These principles underpin algorithms from Kalman filters to kernel ridge regression and modern GP modeling.

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, 2nd ed., Wiley, 2006.

[2] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[3] L. L. Scharf, *Statistical Signal Processing*, Addison-Wesley, 1991.

[4] A. Papoulis and S. U. Pillai, *Probability, Random Variables, and Stochastic Processes*, 4th ed., McGraw-Hill, 2002.

[5] A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Springer, 2004.

A Useful matrix identities

Lemma A.1. *The following well-known matrix identities are quite useful:*

$$\textit{(Block inverse)} \quad \begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}CA^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{bmatrix}, \; S = D - CA^{-1}B, \tag{16}$$

$$\textit{(Woodbury)} \quad (A+UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}. \tag{17}$$

$$\textit{(Matrix determinant lemma)} \quad \det(A+UCV) = \det(C^{-1} + VA^{-1}U)\det(C)\det(A). \tag{18}$$

Proof. **Block inverse.** Define the matrices.

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \qquad L = \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix}, \qquad R = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix}.$$

Gaussian elimination generalizes to block matrices and it is easy to verify that LM is zero in the bottom left block while MR is zero in the top right block. Moreover, direct calculation shows LMR is block diagonal with A in the top left position and S (the Schur complement of A) in the bottom right. Thus, we have

$$M^{-1} = R \begin{bmatrix} A^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} L = \begin{bmatrix} A^{-1} + A^{-1}BS^{-1}CA^{-1} & -A^{-1}BS^{-1} \\ -S^{-1}CA^{-1} & S^{-1} \end{bmatrix}.$$

One can also do the elimination in a different order. In particular, we can define alternative matrices L', R' so that R' eliminates C and L' eliminates B. This gives

$$L' = \begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix}, \quad R' = \begin{bmatrix} I & -A^{-1}B \\ 0 & I \end{bmatrix}, \quad M^{-1} = \begin{bmatrix} T^{-1} & -T^{-1}BD^{-1} \\ -D^{-1}CT^{-1} & D^{-1} + D^{-1}CT^{-1}BD^{-1} \end{bmatrix},$$

where $T = A - BD^{-1}C$ is the Schur complement of D.

For block matrices, if all blocks above (or below) the diagonal are zero, then the matrix is called block lower (or upper) triangular. In both cases, the determinant equals the product of the determinants of the blocks on the diagonal. One can prove this via standard cofactor expansion. For the 2 by 2 case, this implies that $\det(L) = \det(R) = 1$ and $\det(M) = \det(L)\det(M)\det(R) = \det(LMR) = \det(A)\det(S)$. Moreover, if the block matrices in any row or column commute, then this reduces to the simple formula $\det(M) = \det(AD - BC)$.

Woodbury. First, we note that the Woodbury formula is equivalent to

$$(I+UV)^{-1} = I - U(I+VU)^{-1}V,$$

which one gets by replacing A and C with identity matrices. To verify this simplified identity, we multiply on the left by $I + UV$ to get

$$\begin{aligned}(I+UV)\big(I - U(I+VU)^{-1}V\big) &= I + UV - U(I+VU)^{-1}V - UVU(I+VU)^{-1}V \\ &= I + UV - U\big(\underbrace{(I+VU)^{-1} + VU(I+VU)^{-1}}_{I}\big)V = I.\end{aligned}$$

The same idea works for multiplication on the right.

Now, we will derive Woodbury's formula from the simpler identity by substituting $U' = A^{-1}U$ and $V' = CV$. To do this, we write

$$(A+UCV)^{-1} = (I + A^{-1}UCV)^{-1}A^{-1} = (I + U'V')^{-1}A^{-1}$$

$$= (I - U'(I + V'U')^{-1}V')A^{-1} = A^{-1} - A^{-1}U(I + CVA^{-1}U)^{-1}CVA^{-1}$$
$$= A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}.$$

Determinant. First, we note that

$$\det(A + UCV) = \det(A)\,\det\big(I + A^{-1}UCV\big).$$

Using the block inverse notes, we can easily establish Sylvester's determinant identity

$$\det(I + AB) = \det\left(\begin{bmatrix} I & A \\ 0 & I \end{bmatrix}\begin{bmatrix} I & -A \\ B & I \end{bmatrix}\right) = \det\left(\begin{bmatrix} I & -A \\ B & I \end{bmatrix}\begin{bmatrix} I & A \\ 0 & I \end{bmatrix}\right) = \det(I + BA).$$

With this, we find that

$$\begin{aligned}\det\big(I + A^{-1}UCV\big) &= \det\big(I + CVA^{-1}U\big) \\ &= \det\big(C(C^{-1} + VA^{-1}U)\big) = \det(C)\,\det\big(C^{-1} + VA^{-1}U\big).\end{aligned}$$

Hence, we have $\det(A + UCV) = \det(C^{-1} + VA^{-1}U)\,\det(C)\,\det(A)$. □

B Ancillary Results from Probability

Theorem B.1 (Uniqueness of the mgf/Laplace transform)**.** *If two real-valued random variables have mgfs that agree on an open interval around* 0*, then their distributions are identical. Equivalently, for integrable densities f, g with Laplace transforms $\mathcal{L}\{f\}$ and $\mathcal{L}\{g\}$ equal on an interval, one has $f = g$ almost everywhere.*