Scientific Machine Learning 16
Gaussian Process

Donghyun Ko

January 4, 2026

What you will learn.

Contents

1 Backgrounds

This section builds a bridge from basic Gaussian theory to the marginal and conditional laws that power Gaussian Process (GP) modeling. We begin with the univariate Gaussian random variable, extend to multivariate Gaussians, prove key properties of the covariance matrix, and then work through the bivariate case in full detail, including diagonal-covariance, iso–contours, and closed–form conditional distributions as an example of multivariate Gaussian.

1.1 Gaussian random variable

The *Gaussian* (or *normal*) distribution is central to statistics because it is (i) analytically tractable, enabling closed-form calculations and elegant theory; (ii) shaped like a symmetric bell curve, which makes it a convenient and interpretable modeling choice across many applied settings; and (iii) justified by the Central Limit Theorem (CLT), which implies that, under mild conditions, aggregates such as sample means are approximately normal for large samples, allowing the Gaussian to approximate a wide variety of phenomena. A univariate normal distribution is fully determined by two parameters, the mean μ and variance σ^2, and is denoted by $\mathcal{N}(\mu, \sigma^2)$. A real-valued random variable X is *Gaussian* with mean μ and variance $\sigma^2 > 0$, written $X \sim \mathcal{N}(\mu, \sigma^2)$, if its probability density function (PDF) is

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{(x-\mu)^2}{2\sigma^2}\right], \quad \text{where} \quad x \in \mathbb{R}. \tag{1}$$

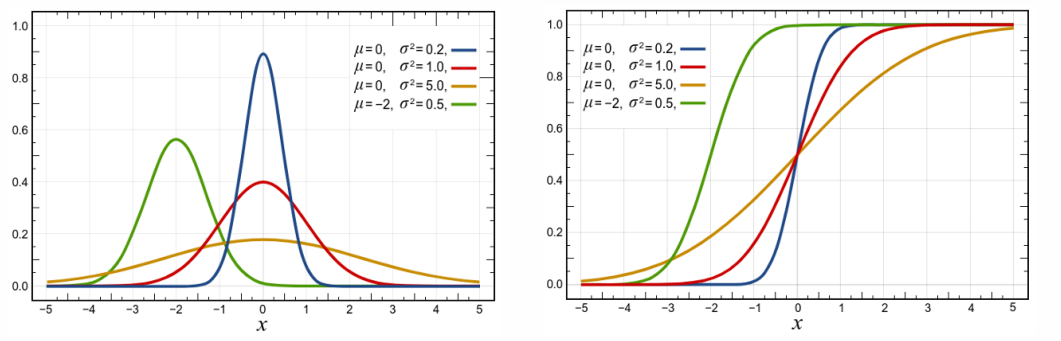
Standardization by $Z = (X - \mu)/\sigma$ yields $Z \sim \mathcal{N}(0, 1)$ with density

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \text{where} \quad z \in \mathbb{R}. \tag{2}$$

For $Z \sim \mathcal{N}(0, 1)$ with PDF $\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$, the cumulative distribution function is

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)\, dt \;=\; \frac{1}{2}\left[1 + \operatorname{erf}\left(\tfrac{z}{\sqrt{2}}\right)\right], \quad \text{where} \quad \operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt. \tag{3}$$

(In the figure, varying μ and σ^2 shifts and scales the bell-shaped PDF and its S-shaped CDF.)



Derivation. Let $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$. Split the integral at the origin:

$$\Phi(z) = \int_{-\infty}^{z} \phi(t)\,dt = \underbrace{\int_{-\infty}^{0} \phi(t)\,dt}_{(*)} + \int_{0}^{z} \phi(t)\,dt.$$

Since $e^{-t^2/2}$ is even, $\phi(-t) = \phi(t)$ and the lower half–mass equals the upper half–mass, so $(*) = \frac{1}{2}$. Hence,

$$\Phi(z) = \frac{1}{2} + \int_{0}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt.$$

With $u = t/\sqrt{2}$, $dt = \sqrt{2}\,du$ and $e^{-t^2/2} = e^{-u^2}$,

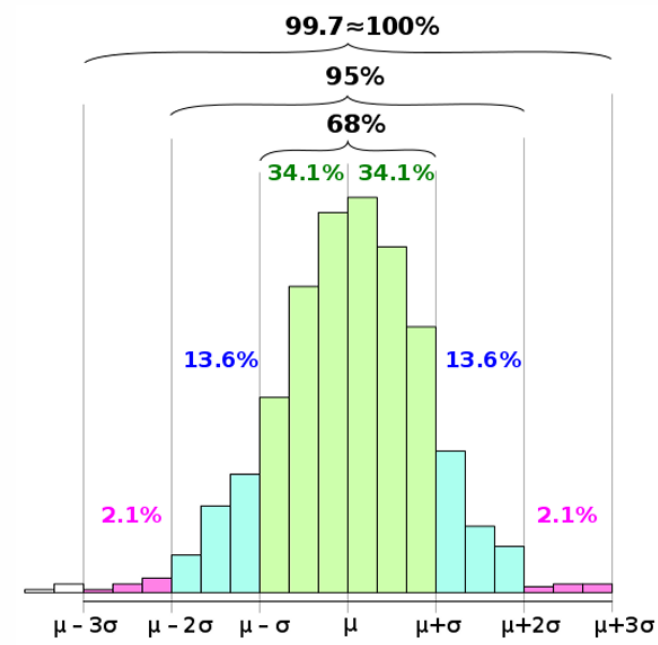
$$\int_{0}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\,dt = \int_{0}^{z/\sqrt{2}} \frac{1}{\sqrt{\pi}} e^{-u^2}\,du = \frac{1}{2}\left(\frac{2}{\sqrt{\pi}} \int_{0}^{z/\sqrt{2}} e^{-u^2}\,du\right) = \frac{1}{2}\,\operatorname{erf}\left(\tfrac{z}{\sqrt{2}}\right),$$

which gives (3).

If $X \sim \mathcal{N}(\mu, \sigma^2)$ and $Z = \frac{X-\mu}{\sigma} \sim \mathcal{N}(0,1)$, then the central mass within one, two, and three standard deviations is

$$\boxed{\begin{aligned} \mathbb{P}(|X-\mu| \le \sigma) &= \mathbb{P}(|Z| \le 1) = 2\Phi(1) - 1 \approx 0.6826, \\ \mathbb{P}(|X-\mu| \le 2\sigma) &= \mathbb{P}(|Z| \le 2) = 2\Phi(2) - 1 \approx 0.9544, \\ \mathbb{P}(|X-\mu| \le 3\sigma) &= \mathbb{P}(|Z| \le 3) = 2\Phi(3) - 1 \approx 0.9974. \end{aligned}} \tag{4}$$

Consequently, the two–sided tail areas are about 31.73%, 4.56%, and 0.27% for 1σ, 2σ, and 3σ respectively, by matching the bell–curve sketch.



As a last peace of basic knowledge, the statistical moments are immediate from (1) s.t:

$$\mathbb{E}[X] = \mu, \qquad \mathbb{E}[(X-\mu)^2] = \mathrm{Var}[X] = \sigma^2, \qquad \mathbb{E}[X^2] = \mu^2 + \sigma^2. \tag{5}$$

Central Limit Theorem (CLT). Let $X_1, \ldots, X_N$ be i.i.d. with population mean μ and variance $\sigma^2 \in (0, \infty)$, and set $\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i$. Then,

$$\frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \xRightarrow[N\to\infty]{} \mathcal{N}(0,1).$$

Equivalently, the standardized sum satisfies

$$\frac{\sum_{i=1}^{N}(X_i - \mu)}{\sigma\sqrt{N}} \xRightarrow[N\to\infty]{} \mathcal{N}(0,1).$$

Consequences. (i) Large–sample procedures that assume normality (e.g., z–tests, Wald intervals) are justified for a wide class of data with finite variance. (ii) Replacing σ by the sample standard deviation S_N yields

$$\frac{\bar{X}_N - \mu}{S_N/\sqrt{N}} \xRightarrow[N\to\infty]{} \mathcal{N}(0,1) \quad \text{(by Slutsky's theorem)},$$

which underpins the usual large N $(1-\alpha)$ confidence interval $\bar{X}_N \pm z_{1-\alpha/2}\, S_N/\sqrt{N}$. The CLT legitimizes normal-based methods for many non-normal problems: sums or averages of many finite-variance variables behave approximately Gaussian, so standard normal tools (e.g., z-intervals/tests) are broadly applicable.

1.2 Multivariate Gaussian: Gaussian Random Vector

A d–dimensional random vector $X = (X_1, \ldots, X_d)^\top$ is *multivariate Gaussian*, representing the distribution of a multivariate random vector that is made up of multiple Gaussian RVs that can be correlated with each other, if there exist a mean vector $\boldsymbol{\mu} = E(X) = (E[X_1], E[X_2], \ldots, E[X_d])^T \in R^d$ and a covariance matrix $\Sigma \in R^{d\times d}$ such that

$$X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \qquad \Longleftrightarrow \qquad f(x \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^d\,|\Sigma|}} \exp\Big[-\frac{1}{2}(x-\boldsymbol{\mu})^\top \Sigma^{-1}(x-\boldsymbol{\mu})\Big], \tag{6}$$

where $|\Sigma|$ denotes the determinant of Σ. The entries of Σ are the pairwise covariances s.t:

$$\Sigma_{ij} = Cov(X_i, X_j) = E\Big[(X_i - \mu_i)(X_j - \mu_j)\Big], \qquad 1 \le i, j \le d. \tag{7}$$

Compactly,

$$\Sigma = E\Big[(X-\boldsymbol{\mu})(X-\boldsymbol{\mu})^\top\Big] = E[XX^\top] - \boldsymbol{\mu}\boldsymbol{\mu}^\top. \tag{8}$$

Why Σ is symmetric positive semidefinite (PSD). Write the covariance entries as $\Sigma_{ij} = Cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)]$. Symmetry is immediate from $\Sigma_{ij} = \Sigma_{ji}$. For any $z \in R^d$, expand the quadratic form in indices:

$$z^\top \Sigma z = \sum_{i=1}^{d}\sum_{j=1}^{d} z_i\, \Sigma_{ij}\, z_j = \sum_{i=1}^{d}\sum_{j=1}^{d} z_i\; \mathbb{E}\Big[(X_i - \mu_i)(X_j - \mu_j)\Big]\, z_j.$$

By linearity of expectation, we may move $\mathbb{E}[\cdot]$ outside the finite sum:

$$z^\top \Sigma z = \mathbb{E}\left[\sum_{i=1}^{d}\sum_{j=1}^{d} z_i\,(X_i - \mu_i)(X_j - \mu_j)\, z_j\right] = \mathbb{E}\left[\left(\sum_{i=1}^{d} z_i\,(X_i - \mu_i)\right)\left(\sum_{j=1}^{d} z_j\,(X_j - \mu_j)\right)\right].$$

Let $Y = X - \boldsymbol{\mu}$. Then, the inner expression is $\left(\sum_{i=1}^{d} z_i Y_i\right)^2 = (Y^\top z)^2$, hence

$$\boxed{z^\top \Sigma z \;=\; \mathbb{E}\Big[(Y^\top z)^2\Big] \;\geq\; 0}$$

Therefore, Σ is always PSD. Moreover, if $z^\top \Sigma z > 0$ for every nonzero z (equivalently, if $\mathbb{E}[(Y^\top z)^2] > 0$ for all $z \neq 0$), then Σ is *positive definite* (PD), which implies full rank and the existence of Σ^{-1} and $|\Sigma|^{-1/2}$ in (6). Conversely, if Σ is only PSD but singular, Σ^{-1} does not exist and the Gaussian density in (6) is undefined. Since any full rank symmetric positive semidefinite matrix is necessarily symmetric positive definite, it follows that Σ must be symmetric positive definite.

Mahalanobis distance. When Σ is PD, the *Mahalanobis distance* of a point x from the mean $\boldsymbol{\mu}$ is

$$D_M(x) = \sqrt{(x - \boldsymbol{\mu})^\top \Sigma^{-1} (x - \boldsymbol{\mu})}.$$

It measures distance in the geometry induced by Σ: directions with large variance count less, and correlated directions are de–weighted jointly. In the univariate case ($d = 1$, $\Sigma = \sigma^2$), $D_M(x) = \left|\frac{x-\mu}{\sigma}\right|$, the absolute standardized score.

Important properties (used later in GP modeling). The multivariate normal family is closed under common operations:

- *Affine transformations:* If $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ and $Y = AX + b$ with $A \in R^{k\times d}$, $b \in R^k$, then $Y \sim \mathcal{N}(A\boldsymbol{\mu},\ A\Sigma A^\top)$.
- *Marginals:* Any subvector of X is Gaussian. Equivalently, if $X = (X_1^\top, X_2^\top)^\top$, then $X_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$ and $X_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_{22})$.
- *Linear combinations:* For any $a \in R^d$, the scalar $a^\top X$ is univariate normal with mean $a^\top \boldsymbol{\mu}$ and variance $a^\top \Sigma a$.

These facts, together with the conditional formulas developed in the next subsection, are the algebraic backbone of Gaussian processes and Kriging.

1.3 The bivariate Gaussian

A special and important case of the multivariate normal is $d = 2$, with a random vector $X = (X_1, X_2)^\top$ having mean $\boldsymbol{\mu} = (\mu_1, \mu_2)^\top$, standard deviations $\sigma_1, \sigma_2 > 0$, and a correlation coefficient $\rho \in (-1, 1)$. Its probability density function (PDF) is

$$f(x_1, x_2) = \frac{1}{2\pi\,\sigma_1\sigma_2\sqrt{1-\rho^2}}\,\exp\!\left\{-\frac{1}{2(1-\rho^2)}\left[\frac{(x_1-\mu_1)^2}{\sigma_1^2} + \frac{(x_2-\mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1-\mu_1)(x_2-\mu_2)}{\sigma_1\sigma_2}\right]\right\}.$$

Equivalently, $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad |\Sigma| = \sigma_1^2\sigma_2^2(1-\rho^2).$$

Marginals and linear combinations. The marginals are univariate normals such that:

$$X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2), \qquad X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2).$$

Any linear combination $aX_1 + bX_2$, with any constants 'a' and 'b', also yields the Gaussian distribution such that:

$$aX_1 + bX_2 \sim \mathcal{N}\Big(a\mu_1 + b\mu_2,\ a^2\sigma_1^2 + b^2\sigma_2^2 + 2ab\,\rho\,\sigma_1\sigma_2\Big).$$

Diagonal covariance and independence. Assume the covariance is diagonal:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \quad \boldsymbol{\mu} = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}.$$

In this case, the correlation is $\rho = \frac{Cov(X_1,X_2)}{\sigma_1\sigma_2} = 0$, and therefore the joint density factors into the product of its marginals:

$$f(x_1, x_2 \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2}\exp\!\left[-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right] = \underbrace{\frac{1}{\sqrt{2\pi}\sigma_1}e^{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2}}}_{f_{X_1}(x_1)}\ \underbrace{\frac{1}{\sqrt{2\pi}\sigma_2}e^{-\frac{(x_2-\mu_2)^2}{2\sigma_2^2}}}_{f_{X_2}(x_2)}.$$

Therefore, X_1 and X_2 are independent. Conversely, within the Gaussian family, independence implies zero covariance; hence,

$$\boxed{X_1 \perp X_2} \iff \boxed{\rho = 0} \iff \boxed{\Sigma \text{ is diagonal}}$$

Iso–contours and ellipses. Fix a constant $c > 0$ and consider the level set with d=2 such that $\{(x_1, x_2) : f(x_1, x_2) = c\}$.

- **Axis–aligned case ($\rho = 0$).** With $\rho = 0$, the bivariate normal density reduces to

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right].$$

  Set $f(x_1, x_2) = c$ and take logs:

$$c = \frac{1}{2\pi\sigma_1\sigma_2} \exp\left[-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}\right] \iff \log\Big(2\pi\sigma_1\sigma_2\, c\Big) = -\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \frac{(x_2-\mu_2)^2}{2\sigma_2^2}.$$

  Equivalently,

$$\log\Big(\frac{1}{2\pi\sigma_1\sigma_2\, c}\Big) = \frac{(x_1-\mu_1)^2}{2\sigma_1^2} + \frac{(x_2-\mu_2)^2}{2\sigma_2^2}.$$

  Define the semi–axes

$$r_1 = \sqrt{2\sigma_1^2 \log\Big(\frac{1}{2\pi\sigma_1\sigma_2\, c}\Big)}, \qquad r_2 = \sqrt{2\sigma_2^2 \log\Big(\frac{1}{2\pi\sigma_1\sigma_2\, c}\Big)}.$$

  Then, the level set is the axis–aligned ellipse such that:

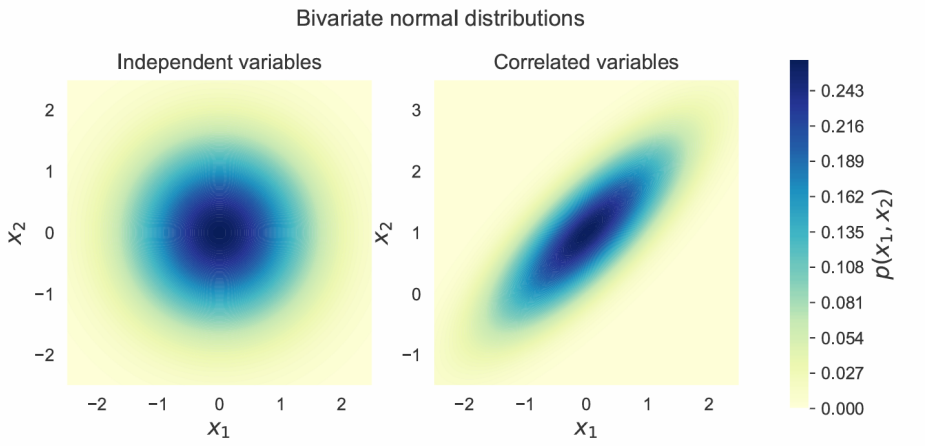
$$1 = \frac{(x_1-\mu_1)^2}{r_1^2} + \frac{(x_2-\mu_2)^2}{r_2^2} = \Big(\frac{x_1-\mu_1}{r_1}\Big)^2 + \Big(\frac{x_2-\mu_2}{r_2}\Big)^2, \tag{9}$$

  centered at (μ_1, μ_2) whose x_1–axis length is $2r_1$ and x_2–axis length is $2r_2$.

- **Correlated case ($\rho \neq 0$).** In general, $f(x) \propto \exp\Big(-\frac{1}{2}(x-\boldsymbol{\mu})^\top \Sigma^{-1}(x-\boldsymbol{\mu})\Big)$. Setting $f(x) = c$ gives the quadratic form $(x-\boldsymbol{\mu})^\top\Sigma^{-1}(x-\boldsymbol{\mu}) = \kappa$ for some $\kappa > 0$. With the Eigen-decomposition such that $\Sigma = Q\Lambda Q^\top$ (orthogonal Q, diagonal $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2)$),

$$(x-\boldsymbol{\mu})^\top\Sigma^{-1}(x-\boldsymbol{\mu}) = \sum_{j=1}^{2} \frac{y_j^2}{\lambda_j}, \qquad y = Q^\top(x-\boldsymbol{\mu}), \tag{10}$$

  so the level sets are *rotated ellipses* with principal axes along the eigenvectors (columns of Q) and semi–axes proportional to $\sqrt{\lambda_1}$ and $\sqrt{\lambda_2}$.



Parameter constraints and interpretation. The parameters satisfy $\sigma_1 > 0$, $\sigma_2 > 0$, and $|\rho| < 1$ (so that $|\Sigma| > 0$). The correlation ρ controls the tilt of the elliptical contours: $\rho > 0$ produces ellipses elongated along the $x_1 \approx x_2$ direction, while $\rho < 0$ tilts them along $x_1 \approx -x_2$. The Mahalanobis geometry $D_M^2(x) = (x-\boldsymbol{\mu})^\top \Sigma^{-1}(x-\boldsymbol{\mu})$ determines both the shape of these contours and the exponent in (1.3), linking covariance, orientation, and concentration of the probability mass.

1.4 Marginal and Conditional distributions of multivariate Gaussian random vectors

Partition a d–vector $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ into subvectors $X_1 \in \mathbb{R}^p$ and $X_2 \in \mathbb{R}^q$ with $p + q = d$:

$$X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \qquad \Sigma_{21} = \Sigma_{12}^\top.$$

The marginals are $X_i \sim \mathcal{N}(\boldsymbol{\mu}_i, \Sigma_{ii})$. The conditional law of X_i given X_j is a multivariate Gaussian with

$$\boxed{\begin{aligned} \mu_{i|j} &= \mathbb{E}[X_i \mid X_j] = \boldsymbol{\mu}_i + \Sigma_{ij}\Sigma_{jj}^{-1}(x_j - \boldsymbol{\mu}_j), \\ \Sigma_{i|j} &= \mathrm{Cov}(X_i \mid X_j) = \Sigma_{ii} - \Sigma_{ij}\Sigma_{jj}^{-1}\Sigma_{ij}^{T}. \end{aligned}} \tag{11}$$

Derivation. Let $X \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$ be partitioned as above and write the joint density

$$f(x_1, x_2 \mid \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^{p+q}\,|\Sigma|}} \exp\Big(-\tfrac{1}{2}\,Q(x_1, x_2)\Big), \quad Q(x_1, x_2) \coloneqq (x-\boldsymbol{\mu})^\top \Sigma^{-1}(x-\boldsymbol{\mu}). \tag{D1}$$

Define the Schur complement of Σ_{11} such that:

$$A \coloneqq \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \ \ (\succ 0). \tag{D2}$$

Then, the block inverse is

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}A^{-1}\Sigma_{21}\Sigma_{11}^{-1} & -\Sigma_{11}^{-1}\Sigma_{12}A^{-1} \\ -A^{-1}\Sigma_{21}\Sigma_{11}^{-1} & A^{-1} \end{bmatrix}. \tag{D3}$$

Substituting (D3) into (D1) and grouping terms yields

$$\begin{aligned} Q(x_1, x_2) &= (x_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1) \\ &\quad + \Big[(x_2 - \boldsymbol{\mu}_2) - \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1)\Big]^\top A^{-1}\Big[(x_2 - \boldsymbol{\mu}_2) - \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1)\Big] \\ &=: Q_1(x_1) + Q_2(x_1, x_2). \end{aligned} \tag{D4}$$

(We used the identity $u^\top M u - 2u^\top M v + v^\top M v = (u-v)^\top M (u-v)$ for symmetric M.)
Block determinant. The determinant factorizes as

$$|\Sigma| = |\Sigma_{11}|\,|A|. \tag{D5}$$

Combining (D4)–(D5) gives

$$f(x_1, x_2 \mid \boldsymbol{\mu}, \Sigma) = \underbrace{\frac{1}{\sqrt{(2\pi)^p |\Sigma_{11}|}} \exp\Big(-\tfrac{1}{2}(x_1 - \boldsymbol{\mu}_1)^\top \Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1)\Big)}_{\mathcal{N}(x_1 | \boldsymbol{\mu}_1, \Sigma_{11})} \times \underbrace{\frac{1}{\sqrt{(2\pi)^q |A|}} \exp\Big(-\tfrac{1}{2}(x_2 - b)^\top A^{-1}(x_2 - b)\Big)}_{\mathcal{N}(x_2 | b, A)}, \tag{D6}$$

where

$$b \coloneqq \boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1), \qquad A \coloneqq \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \tag{D7}$$

From (D6) we obtain the marginals $X_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$ and the conditional

$$X_2 \mid X_1 = x_1 \;\sim\; \mathcal{N}\Big(\boldsymbol{\mu}_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \boldsymbol{\mu}_1),\; \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Big). \tag{D8}$$

By symmetry (swap indices $1 \leftrightarrow 2$), this yields the general formulas in (11).

Bivariate Example. For $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix}$ with the mean $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, standard deviations (σ_1, σ_2) and correlation ρ,

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho\,\sigma_1\sigma_2 \\ \rho\,\sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}, \qquad \Sigma_{11} = \sigma_1^2, \quad \Sigma_{22} = \sigma_2^2, \quad \Sigma_{12} = \Sigma_{21}^\top = \rho\,\sigma_1\sigma_2.$$

Plugging these blocks into (11) for the conditional of X_2 given $X_1 = x_1$:

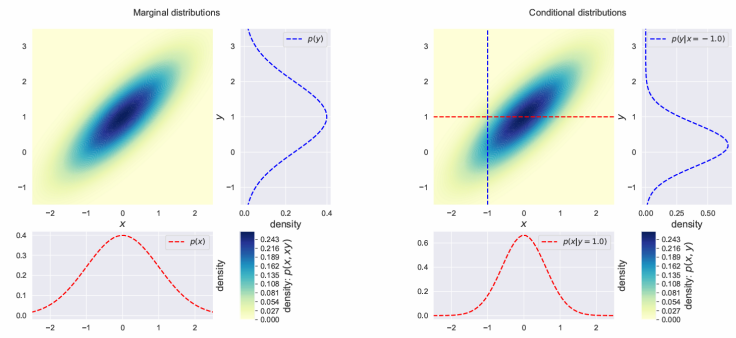
$$\mu_{2|1} = \mu_2 + \Sigma_{21}\Sigma_{11}^{-1}(x_1 - \mu_1) = \mu_2 + \frac{\rho\,\sigma_1\sigma_2}{\sigma_1^2}(x_1 - \mu_1) = \boxed{\mu_2 + \frac{\rho\,\sigma_2}{\sigma_1}(x_1 - \mu_1)}$$

$$\Sigma_{2|1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} = \sigma_2^2 - \frac{(\rho\,\sigma_1\sigma_2)^2}{\sigma_1^2} = \boxed{(1 - \rho^2)\,\sigma_2^2}$$

By symmetry, the conditional of X_1 given $X_2 = x_2$ is

$$\mu_{1|2} = \mu_1 + \frac{\rho\,\sigma_1}{\sigma_2}(x_2 - \mu_2), \qquad \Sigma_{1|2} = (1 - \rho^2)\,\sigma_1^2.$$

When $\rho = 0$, both conditionals reduce to the corresponding marginals, reflecting independence.



1.5 Random Process

Probability space (triple). A random process is built on a probability space (a probability triple)

$$(\Omega, \mathcal{F}, \mathbb{P}),$$

where Ω is the *sample space*, $\mathcal{F}$ is a σ–algebra (event space), and $\mathbb{P}$ is a probability measure on $(\Omega, \mathcal{F})$. Examples of sample spaces:

$$\text{coin toss: } \Omega = \{H, T\}, \qquad \text{die throw: } \Omega = \{1, 2, 3, 4, 5, 6\}.$$

For finite or countable Ω, the canonical choice is the *power set* $\mathcal{F} = 2^{\Omega}$ (all subsets, including $\emptyset$ and Ω). In general, a σ–algebra $\mathcal{F}$ satisfies:

1. $\emptyset \in \mathcal{F}$;
2. closed under complementation: $A \in \mathcal{F} \Rightarrow A^c = \Omega \setminus A \in \mathcal{F}$;
3. closed under countable unions: $A_i \in \mathcal{F}$ $(i \geq 1) \Rightarrow \bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$.

A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ obeys:

$$\mathbb{P}(A) \geq 0, \quad \mathbb{P}(\Omega) = 1, \quad A_i \text{ pairwise disjoint } \Rightarrow \mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i \Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Random (stochastic) process. A *random process* is a family of random variables $X = (X_t : t \in \mathcal{T})$ on a common $(\Omega, \mathcal{F}, \mathbb{P})$. Typical index sets $\mathcal{T}$ are time domains:

$$\mathcal{T} = \mathbb{Z} \text{ (discrete time)}, \qquad \mathcal{T} \subseteq \mathbb{R} \text{ (continuous time)}.$$

Three equivalent views of X: (1) for fixed t, X_t is an RV on Ω; (2) X is a function $X : \mathcal{T} \times \Omega \to \mathbb{R}$, $(t, \omega) \mapsto X_t(\omega)$; (3) for fixed $\omega \in \Omega$, $t \mapsto X_t(\omega)$ is a *sample path*.

Moments and finite–dimensional laws. The *mean*, *correlation*, and *covariance* of X are

$$\mu_X(t) = \mathbb{E}[X_t], \qquad R_X(s, t) = \mathbb{E}[X_s X_t], \qquad C_X(s, t) = \operatorname{Cov}(X_s, X_t) = R_X(s, t) - \mu_X(s)\mu_X(t).$$

The nth–order CDF collects a finite set of marginals:

$$F_{X,n}(x_1, t_1; \ldots; x_n, t_n) = \mathbb{P}\{X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n\}.$$

A *second–order* process satisfies $\mathbb{E}[X_t^2] < \infty$ for all t, so μ_X, R_X, and C_X are well defined and finite.

Gaussian process. X is a *Gaussian process* if every finite vector $(X_{t_1}, \ldots, X_{t_n})^{\top}$ is multivariate Gaussian. In that case, *all* finite–dimensional distributions are completely determined by the pair (μ_X, C_X):

$$(X_{t_1}, \ldots, X_{t_n})^{\top} \sim \mathcal{N}\left(\begin{bmatrix} \mu_X(t_1) \\ \vdots \\ \mu_X(t_n) \end{bmatrix}, \left[C_X(t_i, t_j) \right]_{i,j=1}^{n} \right).$$

Worked Gaussian example (affine plus trend). Let $A, B \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ be independent and define

$$X_t = A + Bt + t^2, \qquad t \in \mathbb{R}.$$

Then each sample path is a parabola (random intercept A and slope B). Its moments are

$$\mu_X(t) = \mathbb{E}[A + Bt + t^2] = t^2, \qquad R_X(s,t) = \mathbb{E}[(A + Bs + s^2)(A + Bt + t^2)] = 1 + st + s^2t^2,$$

$$C_X(s,t) = R_X(s,t) - \mu_X(s)\mu_X(t) = 1 + st.$$

For each fixed t, X_t is Gaussian with

$$X_t \sim \mathcal{N}\big(t^2,\ 1 + t^2\big), \qquad f_{X_t}(x) = \frac{1}{\sqrt{2\pi(1+t^2)}} \exp\left(-\frac{(x - t^2)^2}{2(1+t^2)}\right).$$

For distinct $s \neq t$,

$$\begin{bmatrix} X_s \\ X_t \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} s^2 \\ t^2 \end{bmatrix}, \begin{bmatrix} 1 + s^2 & 1 + st \\ 1 + st & 1 + t^2 \end{bmatrix}\right),$$

whose covariance determinant is

$$\det \begin{bmatrix} 1 + s^2 & 1 + st \\ 1 + st & 1 + t^2 \end{bmatrix} = (1 + s^2)(1 + t^2) - (1 + st)^2 = (s - t)^2 > 0 \ (s \neq t),$$

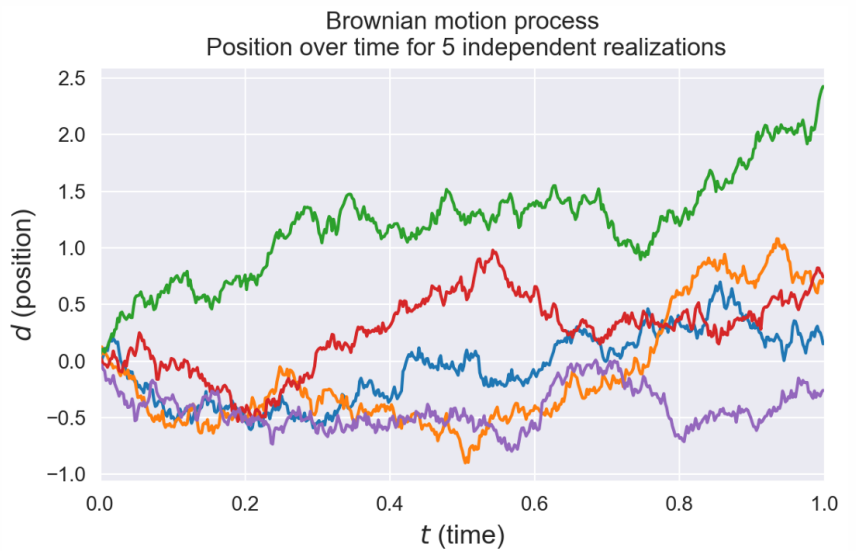
so the joint PDF exists and equals

$$f_{X_s, X_t}(x_s, x_t) = \frac{1}{2\pi\,|s - t|} \exp\left(-\tfrac{1}{2} \begin{bmatrix} x_s - s^2 \\ x_t - t^2 \end{bmatrix}^\top \begin{bmatrix} 1 + s^2 & 1 + st \\ 1 + st & 1 + t^2 \end{bmatrix}^{-1} \begin{bmatrix} x_s - s^2 \\ x_t - t^2 \end{bmatrix}\right).$$

Canonical continuous–time example (Brownian motion). The *Brownian motion* (Wiener process) W_t is a continuous–time process with $W_0 = 0$, independent stationary increments, Gaussian marginals $W_t \sim \mathcal{N}(0, t)$, mean $\mu_W(t) = 0$, and covariance

$$C_W(s,t) = \min(s,t).$$

Its sample paths are almost surely continuous and nowhere differentiable; it is a central building block for stochastic modeling and Gaussian process kernels (e.g., Matérn families).



2 Kriging as GP Regression: BLUP, Interpolation, and Uncertainty

Suppose a deterministic response $y_M(x)$ is observed at design points $X = \{x^{(i)}\}_{i=1}^m \subset \mathbb{R}^d$ with outputs $\boldsymbol{y} = (y^{(1)}, \dots, y^{(m)})^\top$. Kriging (GP regression) models

$$y(x) = f(x)^\top \boldsymbol{\beta} + z(x), \qquad z(\cdot) \sim \mathrm{GP}\Big(0,\ \sigma^2 R_\theta(\cdot,\cdot)\Big), \tag{12}$$

where $f(x) \in \mathbb{R}^n$ encodes a low-order trend (e.g., constant or polynomial), $\boldsymbol{\beta} \in \mathbb{R}^n$ are unknown coefficients, σ^2 is the process variance, and R_θ is a correlation kernel with hyperparameters θ (and, possibly, ARD exponents). Stack the trend rows

$$F = \begin{bmatrix} f(x^{(1)})^\top \\ \vdots \\ f(x^{(m)})^\top \end{bmatrix} \in \mathbb{R}^{m\times n}, \quad R = \Big[R_\theta(x^{(i)}, x^{(j)})\Big]_{i,j=1}^m \in \mathbb{R}^{m\times m}, \quad r(x_*) = \begin{bmatrix} R_\theta(x_*, x^{(1)}) \\ \vdots \\ R_\theta(x_*, x^{(m)}) \end{bmatrix}.$$

Deriving the universal Kriging predictor

Jointly,

$$\begin{bmatrix} \boldsymbol{y} \\ y(x_*) \end{bmatrix} = \begin{bmatrix} F \\ f(x_*)^\top \end{bmatrix} \boldsymbol{\beta} + \begin{bmatrix} z \\ z_* \end{bmatrix}, \qquad \begin{bmatrix} z \\ z_* \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ 0 \end{bmatrix}, \sigma^2 \begin{bmatrix} R & r(x_*) \\ r(x_*)^\top & 1 \end{bmatrix} \right).$$

By the conditional-Gaussian formula, the *best linear unbiased predictor* (BLUP) is

$$\widehat{y}(x_*) = f(x_*)^\top \widehat{\boldsymbol{\beta}} + r(x_*)^\top R^{-1}\Big(\boldsymbol{y} - F\widehat{\boldsymbol{\beta}}\Big), \tag{13}$$

where the generalized least-squares estimator of the trend is

$$\widehat{\boldsymbol{\beta}} = (F^\top R^{-1} F)^{-1} F^\top R^{-1} \boldsymbol{y}. \tag{14}$$

The mean-squared prediction error (MSPE) follows from Schur complements:

$$\mathrm{MSPE}(x_*) = \sigma^2\Big[1 - r(x_*)^\top R^{-1} r(x_*) + \Delta(x_*)^\top (F^\top R^{-1} F)^{-1} \Delta(x_*)\Big], \tag{15}$$

with the trend-correction $\Delta(x_*) = F^\top R^{-1} r(x_*) - f(x_*)$. Equations (13)–(15) constitute *Universal Kriging* (UK).

Interpolation and zero variance at design points (noise-free case). If observations are exact and R is nonsingular, then for any design input $x^{(i)}$,

$$\widehat{y}(x^{(i)}) = y^{(i)}, \qquad \mathrm{MSPE}(x^{(i)}) = 0.$$

Reason. At $x_* = x^{(i)}$, $r(x_*)$ equals the i-th column of R, so $R^{-1} r(x^{(i)}) = e_i$. Hence $\Delta(x^{(i)}) = F^\top e_i - f(x^{(i)}) = 0$, and (13) reduces to $y^{(i)}$. Substituting into (15) gives zero variance. □

Two useful specializations streamline the trend. If $f(x) \equiv 1$ is an unknown constant, (13) becomes *Ordinary Kriging* (OK); the estimated mean $\widehat{\beta}_0$ is a generalized average that recenters residuals before correlation-based smoothing. If, instead, the mean μ is known, we obtain *Simple Kriging* (SK) with predictor $\mu + r^\top R^{-1}(\boldsymbol{y} - \mu\mathbf{1})$ and MSPE $\sigma^2(1 - r^\top R^{-1} r)$.

Noisy observations and the nugget

When measurements contain i.i.d. noise, $y^{(i)} = y_M(x^{(i)}) + \varepsilon^{(i)}$, $\varepsilon^{(i)} \sim \mathcal{N}(0, \tau^2)$, the correlation matrix is replaced by $R + \eta I$ with $\eta = \tau^2/\sigma^2$. The predictor (13) remains the same with this substitution, but interpolation gives way to smoothing, and predictive variance stays positive at design points. Heteroskedastic noise is handled by replacing ηI with $\mathrm{diag}(\eta_1, \ldots, \eta_m)$.

Posterior predictive distribution (multiple test points)

For test inputs $X_* = \{x_*^{(1)}, \ldots, x_*^{(q)}\}$, stack F_*, R_*, R_{**} analogously. Conditioning yields

$$\boldsymbol{y}_* \mid \boldsymbol{y} \sim \mathcal{N}\Big(\boldsymbol{\mu}_*,\ \Sigma_*\Big), \qquad \boldsymbol{\mu}_* = F_*\widehat{\boldsymbol{\beta}} + R_*^\top R^{-1}\Big(\boldsymbol{y} - F\widehat{\boldsymbol{\beta}}\Big),$$

$$\Sigma_* = \sigma^2\Big(R_{**} - R_*^\top R^{-1} R_* + \Gamma\Big), \quad \Gamma = \Big(F_*^\top - F^\top R^{-1} R_*\Big)^\top (F^\top R^{-1} F)^{-1}\Big(F_*^\top - F^\top R^{-1} R_*\Big),$$

whose diagonal provides pointwise predictive variances that underpin credible intervals and acquisition functions in Bayesian optimization.

3 Kernels, Hyperparameters, Likelihood, and Practical Diagnostics

A kernel must be symmetric positive semidefinite so that every Gram matrix is. Stationary kernels depend only on $r = x - x'$, and automatic relevance determination (ARD) endows each coordinate with its own characteristic length-scale, learning which inputs matter.

Geometry and smoothness of popular kernels

The squared-exponential (SE/RBF) kernel,

$$k_{\mathrm{SE}}(x, x') = \exp\Big(-\tfrac{1}{2} \sum_{j=1}^{d} \frac{(x_j - x'_j)^2}{\ell_j^2} \Big),$$

assumes an infinitely mean-square differentiable field. It yields exceptionally smooth interpolants—often desirable for well-resolved PDE solvers, but sometimes overconfident between sparse measurements. The Matérn family,

$$k_\nu(x, x') = \prod_{j=1}^{d} \frac{2^{1-\nu}}{\Gamma(\nu)} \Big(\frac{\sqrt{2\nu}\,|x_j - x'_j|}{\ell_j} \Big)^{\nu} K_\nu \Big(\frac{\sqrt{2\nu}\,|x_j - x'_j|}{\ell_j} \Big),$$

introduces a smoothness parameter ν: $\nu = \frac{1}{2}$ reproduces an exponential kernel with rough sample paths, while $\nu = \frac{3}{2}, \frac{5}{2}$ produce once- and twice-mean-square differentiable fields, respectively—excellent compromises for engineering surrogates. Periodic kernels encode exact seasonality by replacing $(x_j - x'_j)$ with $\sin\big(\pi(x_j - x'_j)/p_j\big)$ inside an SE envelope. Linear (dot-product) kernels recover global trends through covariance rather than the explicit mean.

Length-scales ℓ_j admit an intuitive reading: small ℓ_j means rapid variation along x_j; large ℓ_j means the process is nearly constant in that direction. After rescaling inputs to $[0, 1]^d$, ARD values become directly comparable and double as a data-driven relevance measure.

Concentrated maximum likelihood

Under model (12) with Gaussian residuals, the marginal likelihood of $\boldsymbol{y}$ is

$$\boldsymbol{y} \sim \mathcal{N}\big(F\boldsymbol{\beta},\ \sigma^2 R_\theta\big).$$

The log-likelihood, up to an additive constant, is

$$\ell(\boldsymbol{\beta}, \sigma^2, \theta) = -\frac{m}{2} \log \sigma^2 - \frac{1}{2} \log\det R_\theta - \frac{1}{2\sigma^2} (\boldsymbol{y} - F\boldsymbol{\beta})^\top R_\theta^{-1} (\boldsymbol{y} - F\boldsymbol{\beta}).$$

Setting derivatives to zero reveals the *concentrated* estimators

$$\widehat{\boldsymbol{\beta}}(\theta) = (F^\top R_\theta^{-1} F)^{-1} F^\top R_\theta^{-1} \boldsymbol{y}, \qquad \widehat{\sigma}^2(\theta) = \frac{1}{m} (\boldsymbol{y} - F\widehat{\boldsymbol{\beta}})^\top R_\theta^{-1} (\boldsymbol{y} - F\widehat{\boldsymbol{\beta}}),$$

and the profile objective

$$\ell_c(\theta) = -\frac{m}{2} \log \widehat{\sigma}^2(\theta) - \frac{1}{2} \log\det R_\theta,$$

which we *maximize* with respect to θ (or equivalently minimize $-2\ell_c$). In practice one optimizes the log-length-scales to enforce positivity, uses multiple random restarts, and evaluates the objective via a Cholesky factorization $R_\theta = LL^\top$:

$$\log\det R_\theta = 2\sum_{i=1}^{m}\log L_{ii}, \qquad R_\theta^{-1}\boldsymbol{v} \text{ via triangular solves } L\boldsymbol{w}=\boldsymbol{v},\ L^\top\boldsymbol{u}=\boldsymbol{w}.$$

A tiny *jitter* εI (e.g., $\varepsilon = 10^{-8}$) stabilizes near-singular kernels.

Noisy data and identifiability

When a nugget η is present, the concentrated likelihood trades off η and short length-scales: very small ℓ can mimic measurement noise. Modelers therefore fix η from instrument specifications when possible, or apply weak priors/penalties to avoid degenerate optima. If the data are truly deterministic (computer codes), a small but nonzero nugget still helps numerical stability without erasing the near-interpolatory character.

Validation: leave-one-out (LOO) and predictive coverage

GPs admit analytic LOO formulas. Let $K = R_\theta$ or $R_\theta + \eta I$, and define the *generalized residual vector* $\alpha = K^{-1}(\boldsymbol{y} - F\widehat{\boldsymbol{\beta}})$. Then, after a single Cholesky solve, the LOO prediction at $x^{(i)}$ and its variance can be expressed using entries of K^{-1}. The standardized LOO residuals

$$e_i^{\mathrm{LOO}} = \frac{y^{(i)} - \widehat{y}_{-i}(x^{(i)})}{\sqrt{\widehat{s}^2_{-i}(x^{(i)})}}$$

should resemble $\mathcal{N}(0,1)$ when the kernel is well specified. Beyond point errors (RMSE, MAE), assess the *empirical coverage* of nominal $100(1-\alpha)\%$ intervals; consistent models yield coverage near target with neither chronic under- nor over-coverage. Spatial plots of residuals against inputs often reveal missing structure (e.g., anisotropy) or the need for a richer mean.

From theory to numerics: a minimal, robust recipe

Rescale each input dimension to $[0,1]$ and center the output to ease conditioning. Start with universal Kriging using a Matérn-5/2 kernel with ARD. Optimize the concentrated likelihood with several random initializations of $\log \ell_j$; add a small jitter before Cholesky. Inspect LOO residuals and coverage; if residuals cluster or coverage is too optimistic, introduce a nugget or allow a lower ν (rougher prior). If extrapolation is required, include a low-order polynomial trend rather than forcing the kernel to shoulder long-range behavior.

Interpretation via two didactic processes

Return to the affine-plus-trend process $X_t = A + Bt + t^2$. The 2×2 covariance for (X_s, X_t) is

$$\Sigma = \begin{bmatrix} 1+s^2 & 1+st \\ 1+st & 1+t^2 \end{bmatrix}, \quad \det\Sigma = (s-t)^2 > 0 \ (s \neq t).$$

Conditioning $X_t \mid X_s$ produces an affine mean in $(X_s - s^2)$ and a variance that shrinks linearly with the correlation $\rho = \frac{1+st}{\sqrt{(1+s^2)(1+t^2)}}$, mirroring (13)–(15). For Brownian motion, the Matérn family with $\nu = \frac{1}{2}$ recovers exponential covariance; increasing ν enforces smoother physics, which is crucial when surrogate gradients feed downstream optimizers.

Scaling and structure (brief outlook)

The cubic cost $O(m^3)$ and quadratic memory $O(m^2)$ motivate approximations for large m: inducing-point methods compress information to $n \ll m$ virtual locations; kernel interpolation on structured grids (SKI/KISS-GP) exploits fast transforms; and localized mixtures of experts split the domain with continuity constraints at boundaries. When physics is known, additive or multiplicative kernel compositions encode symmetries, conservation laws, or periodic forcing without abandoning probabilistic uncertainty.

This is a posting that I summarized with study-purpose and is adapted from lecture notes of NE-795 (Scientific Machine Learning), taught by Prof. Xu Wu, NC State University.