Gaussian Identities with Examples

Abstract

This note reviews a collection of useful facts about Gaussian distributions in both the scalar and multivariate settings. It presents marginal and conditional laws for partitioned Gaussians, affine transformations, basic scalar identities, and truncated Gaussians, along with low-dimensional examples that illustrate how these results are used in practice, including in Bayesian linear regression.

1 Gaussian Probability Density Functions

1.1 One-dimensional Gaussian

A real-valued random variable x is Gaussian with mean μ and variance σ^2 if its density is

$$p(x) = \mathcal{N}(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right). \tag{1}$$

This is the standard scalar normal pdf.

1.2 Multivariate Gaussian

A random vector $\boldsymbol{z} \in \mathbb{R}^D$ is multivariate Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ if

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}\sqrt{\det \boldsymbol{\Sigma}}} \exp\left(-\frac{1}{2} (\boldsymbol{z}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\boldsymbol{z}-\boldsymbol{\mu}) \right). \tag{2}$$

This general form appears in the original formula sheet.

1.3 Illustration

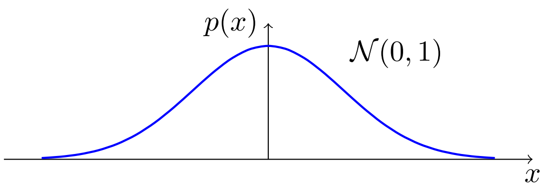

Figure 1: Example of a one-dimensional Gaussian density.

2 Partitioned Gaussian in Terms of x and y

Consider a joint Gaussian vector formed by stacking two random vectors:

$$\boldsymbol{z} = \begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy} \end{pmatrix} \right), \tag{3}$$

where $\boldsymbol{x}$ and $\boldsymbol{y}$ can be multivariate and are generally correlated. (to visualize go to Chris Boucher (2007), "The Bivariate Normal Distribution" Wolfram Demonstrations Project. `https://demonstrations.wolfram.com/TheBivariateNormalDistribution/`)

2.1 Theorem 1: Marginal of x

The marginal distribution of $\boldsymbol{x}$ is again Gaussian and simply uses the corresponding block of the mean and covariance:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\, d\boldsymbol{y} = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{xx}). \tag{4}$$

This is Theorem 1 (marginalization) in the original notes, rewritten with (x, y).

2D example using Theorem 1. Let

$$\boldsymbol{z} = \begin{pmatrix} x \\ y \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right), \tag{5}$$

with correlation parameter $\rho \in (-1, 1)$. Here $\boldsymbol{x} = x$, $\boldsymbol{y} = y$, so $\boldsymbol{\mu}_x = 0$ and $\boldsymbol{\Sigma}_{xx} = 1$. By Theorem 1,

$$p(x) = \mathcal{N}(x; 0, 1), \tag{6}$$

which shows that marginalizing out y only picks out the mean and variance from the x-block of the covariance matrix.

2.2 Theorem 2: Conditional of x given y

The conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}$ is also Gaussian:

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}), \tag{7}$$

$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y), \tag{8}$$

$$\boldsymbol{\Sigma}_{x|y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx}. \tag{9}$$

This is Theorem 2 (conditioning) expressed with (x, y).

2D example using Theorem 2. Reusing the previous 2D Gaussian with

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad \boldsymbol{\mu} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \tag{10}$$

we identify

$$\boldsymbol{\Sigma}_{xy} = \rho, \quad \boldsymbol{\Sigma}_{yy} = 1, \quad \boldsymbol{\mu}_x = 0, \quad \boldsymbol{\mu}_y = 0. \tag{11}$$

Applying Theorem 2 yields

$$\mu_{x|y} = 0 + \rho \cdot 1^{-1}(y - 0) = \rho y, \tag{12}$$

$$\sigma^2_{x|y} = 1 - \rho \cdot 1^{-1}\rho = 1 - \rho^2. \tag{13}$$

Thus

$$x \mid y \sim \mathcal{N}(\rho y, 1 - \rho^2), \tag{14}$$

which shows that conditioning a joint Gaussian corresponds to a linear regression of x on y plus Gaussian noise.

For visualization, please go to Chris Boucher (2008), "The Bivariate Normal and Conditional Distributions" Wolfram Demonstrations Project. `https://demonstrations.wolfram.com/TheBivariateNormalAndConditionalDistributions/`

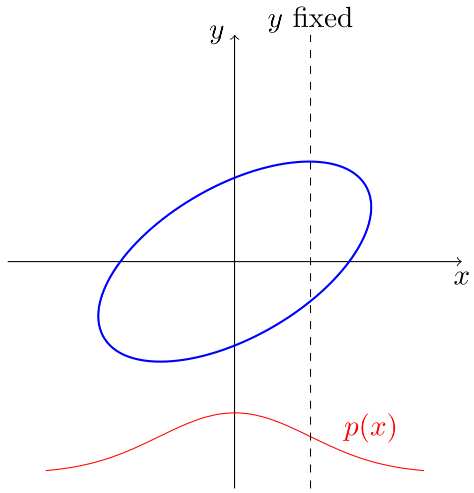

Figure 2: Joint Gaussian over (x, y), with marginal $p(x)$ and conditional slices $p(x \mid y)$.

3 Affine Gaussian Models in Terms of x and y

Affine-Gaussian constructions are central to Bayesian linear regression.

3.1 Theorem 3: Affine Transformation (Forward Model)

Assume

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_x), \tag{15}$$
$$p(\boldsymbol{y} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}, \boldsymbol{\Sigma}_{y|x}), \tag{16}$$

where the conditional $\boldsymbol{y} \mid \boldsymbol{x}$ is a linear function of $\boldsymbol{x}$ plus Gaussian noise. Then the stacked vector $(\boldsymbol{x}, \boldsymbol{y})^\top$ is jointly Gaussian:

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{pmatrix} \boldsymbol{x} \\ \boldsymbol{y} \end{pmatrix}; \begin{pmatrix} \boldsymbol{\mu}_x \\ \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b} \end{pmatrix}, \boldsymbol{R} \right), \tag{17}$$

with block covariance

$$\boldsymbol{R} = \begin{pmatrix} \boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_x \boldsymbol{A}^\top \\ \boldsymbol{A}\boldsymbol{\Sigma}_x & \boldsymbol{\Sigma}_{y|x} + \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top \end{pmatrix}. \tag{18}$$

3.2 Corollary 2: Marginal of y (Predictive Distribution)

Combining Theorem 3 with Theorem 1 yields a closed form for the marginal of $\boldsymbol{y}$:

$$p(\boldsymbol{y}) = \int p(\boldsymbol{y} \mid \boldsymbol{x})\, p(\boldsymbol{x})\, d\boldsymbol{x} = \mathcal{N}(\boldsymbol{y}; \boldsymbol{\mu}_y, \boldsymbol{\Sigma}_y), \tag{19}$$

where

$$\boldsymbol{\mu}_y = \boldsymbol{A}\boldsymbol{\mu}_x + \boldsymbol{b}, \tag{20}$$
$$\boldsymbol{\Sigma}_y = \boldsymbol{\Sigma}_{y|x} + \boldsymbol{A}\boldsymbol{\Sigma}_x\boldsymbol{A}^\top. \tag{21}$$

Example: Predictive distribution in a scalar linear model. Let

$$x \sim \mathcal{N}(0, 1), \quad y \mid x \sim \mathcal{N}(2x + 1, 0.5^2). \tag{22}$$

Here $\boldsymbol{x} = x$, $\boldsymbol{y} = y$, $A = 2$, $b = 1$, $\boldsymbol{\Sigma}_x = 1$, $\boldsymbol{\Sigma}_{y|x} = 0.25$. Corollary 2 gives

$$\mu_y = A\mu_x + b = 2 \cdot 0 + 1 = 1, \tag{23}$$
$$\sigma_y^2 = \Sigma_{y|x} + A^2\Sigma_x = 0.25 + 4 \cdot 1 = 4.25. \tag{24}$$

Thus

$$y \sim \mathcal{N}(1, 4.25), \tag{25}$$

which can be interpreted as the prior predictive distribution in Bayesian linear regression.

3.3 Corollary 1: Posterior of x given y

Combining Theorem 3 with Theorem 2 yields a closed form for the posterior $p(\boldsymbol{x} \mid \boldsymbol{y})$:

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{x|y}, \boldsymbol{\Sigma}_{x|y}), \tag{26}$$

with

$$\boldsymbol{\Sigma}_{x|y} = \left(\boldsymbol{\Sigma}_x^{-1} + \boldsymbol{A}^\top \boldsymbol{\Sigma}_{y|x}^{-1} \boldsymbol{A}\right)^{-1}, \tag{27}$$
$$\boldsymbol{\mu}_{x|y} = \boldsymbol{\Sigma}_{x|y} \left(\boldsymbol{\Sigma}_x^{-1}\boldsymbol{\mu}_x + \boldsymbol{A}^\top \boldsymbol{\Sigma}_{y|x}^{-1}(\boldsymbol{y} - \boldsymbol{b})\right). \tag{28}$$

This is Corollary 1 and is precisely the posterior over regression weights in Bayesian linear regression when $\boldsymbol{x}$ represents weights and $\boldsymbol{y}$ are observed outputs.

Example: Scalar Bayesian update via Corollary 1. Take again

$$x \sim \mathcal{N}(0, 1), \quad y \mid x \sim \mathcal{N}(x, 0.25), \tag{29}$$

and suppose we observe $y = 1$. We fit the affine form with $A = 1$, $b = 0$, $\Sigma_x = 1$, $\Sigma_{y|x} = 0.25$. Corollary 1 gives

$$\sigma_{x|y}^2 = \left(1^{-1} + 1^2 \cdot 0.25^{-1}\right)^{-1} = (1 + 4)^{-1} = \frac{1}{5}, \tag{30}$$
$$\mu_{x|y} = \sigma_{x|y}^2 \left(1^{-1} \cdot 0 + 1 \cdot 0.25^{-1}(y - 0)\right) = \frac{1}{5} \cdot 4y = \frac{4}{5}y. \tag{31}$$

For $y = 1$ we obtain

$$x \mid y = 1 \sim \mathcal{N}(0.8, 0.2), \tag{32}$$

demonstrating the Gaussian posterior that results from a linear-Gaussian likelihood and Gaussian prior, as in Bayesian linear regression.

3.4 Graphical summary of theorem relationships

4 Scalar Gaussian Identities

4.1 Scaling

If $y = ax$ with $a \neq 0$ and y is Gaussian in x with variance σ^2, then as a function of x one can rewrite

$$\mathcal{N}(y; ax, \sigma^2) \propto \mathcal{N}\left(x; \frac{y}{a}, \frac{\sigma^2}{a^2}\right), \tag{33}$$

which is the scaling property noted in the original sheet.

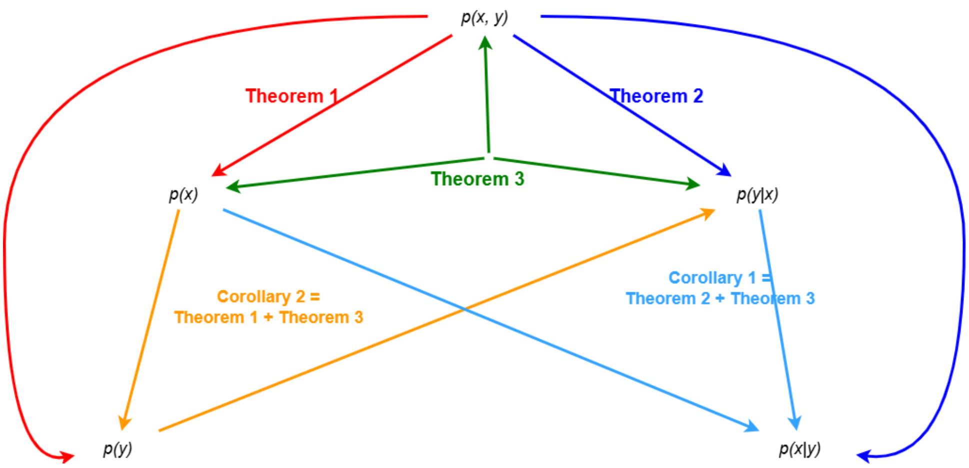

Figure 3: A graphical summary of how Theorem 1, 2, 3, Corollary 1 and Corollary 2 relate to each other.

4.2 Shifting

If $y = x + m$ and y is Gaussian in x with variance σ^2, then

$$\mathcal{N}(y; x + m, \sigma^2) \propto \mathcal{N}(y - m; x, \sigma^2), \tag{34}$$

corresponding to a simple shift of the mean.

4.3 Product of two Gaussians

Let

$$\mu_1(x) \propto \mathcal{N}(x; m_1, \sigma_1^2), \tag{35}$$
$$\mu_2(x) \propto \mathcal{N}(x; m_2, \sigma_2^2). \tag{36}$$

Then

$$\mu_1(x)\mu_2(x) \propto \mathcal{N}(x; m, \sigma^2), \tag{37}$$

with

$$\sigma^2 = \left(\frac{1}{\sigma_1^2} + \frac{1}{\sigma_2^2}\right)^{-1}, \tag{38}$$
$$m = \sigma^2 \left(\frac{m_1}{\sigma_1^2} + \frac{m_2}{\sigma_2^2}\right). \tag{39}$$

This is the scalar version of the posterior update used in Gaussian Bayesian inference.

4.4 Ratio of two Gaussians

Similarly, for

$$\mu_1(x) \propto \mathcal{N}(x; m_1, \sigma_1^2), \tag{40}$$
$$\mu_2(x) \propto \mathcal{N}(x; m_2, \sigma_2^2), \tag{41}$$

the ratio where defined is proportional to another Gaussian:

$$\frac{\mu_1(x)}{\mu_2(x)} \propto \mathcal{N}(x; \tilde{m}, \tilde{\sigma}^2), \tag{42}$$

with

$$\tilde{\sigma}^2 = \left(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2} \right)^{-1}, \tag{43}$$

$$\tilde{m} = \tilde{\sigma}^2 \left(\frac{m_1}{\sigma_1^2} - \frac{m_2}{\sigma_2^2} \right), \tag{44}$$

matching the division identity in the formula sheet.

5 Gaussian CDF and Truncation

5.1 Cumulative distribution function

For a scalar Gaussian, the CDF is

$$F(x; \mu, \sigma^2) = \int_{-\infty}^{x} \mathcal{N}(t; \mu, \sigma^2)\, dt. \tag{45}$$

This definition appears in the original document.

5.2 Two-sided truncated Gaussian

For fixed bounds $a < b$, a Gaussian truncated to (a, b) has pdf

$$f(x; \mu, \sigma^2, a, b) = \frac{\mathcal{N}(x; \mu, \sigma^2)}{F(b; \mu, \sigma^2) - F(a; \mu, \sigma^2)}, \quad a < x < b, \tag{46}$$

and zero elsewhere.

5.3 One-sided truncated Gaussian and example

For a single upper bound b, a one-sided truncated Gaussian is

$$f(x; \mu, \sigma^2, b) = \frac{\mathcal{N}(x; \mu, \sigma^2)}{F(b; \mu, \sigma^2)}, \quad x < b, \tag{47}$$

and zero otherwise.

Example: Physical constraint $x < 1$. If $x \sim \mathcal{N}(0, 1)$ but the system enforces $x < 1$, then the effective distribution is

$$f(x; 0, 1, b = 1) = \frac{\mathcal{N}(x; 0, 1)}{F(1; 0, 1)}, \quad x < 1, \tag{48}$$

where $F(1; 0, 1)$ is the standard normal CDF evaluated at 1. This models, for example, a sensor reading that cannot exceed a maximum value due to saturation.

6 Acknowledgements

This document was converted into LaTeX slides and formatted with assistance from ChatGPT (OpenAI). It uses directly the material from "Formula sheet for the Gaussian distribution" from Advanced Probabilistic Machine Learning course, 2020.