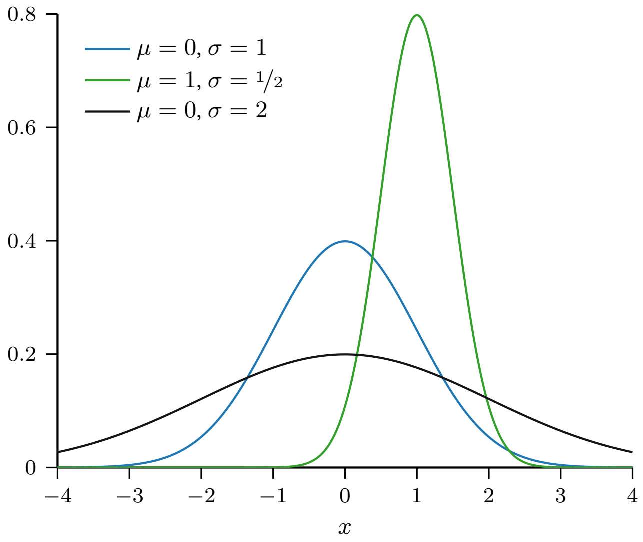


Figure 1: Examples of univariate Gaussian PDFS $\mathcal{N}(x; \mu, \sigma^2)$.

The Gaussian distribution

Probably the most-important distribution in all of statistics is the *Gaussian distribution,* also called the *normal distribution.* The Gaussian distribution arises in many contexts and is widely used for modeling continuous random variables.

The probability density function of the univariate (one-dimensional) Gaussian distribution is

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

The normalization constant Z is

$$Z = \sqrt{2\pi\sigma^2}.$$

The parameters μ and σ^2 specify the mean and variance of the distribution, respectively:

$$\mu = \mathbb{E}[x]; \qquad \sigma^2 = \operatorname{var}[x].$$

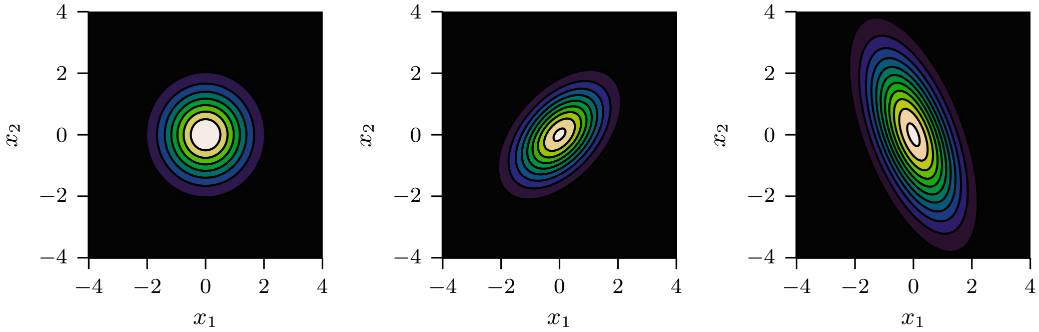
Figure 1 plots the probability density function for several sets of parameters (μ, σ^2). The distribution is symmetric around the mean and most of the density ($\approx 99.7\%$) is contained within $\pm 3\sigma$ of the mean.

We may extend the univariate Gaussian distribution to a distribution over d-dimensional vectors, producing a multivariate analog. The probablity density function of the multivariate Gaussian distribution is

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{Z} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right).$$

The normalization constant Z is

$$Z = \sqrt{\det(2\pi\boldsymbol{\Sigma})} = (2\pi)^{d/2} (\det \boldsymbol{\Sigma})^{1/2}.$$



(a) $\mathbf{\Sigma} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (b) $\mathbf{\Sigma} = \begin{bmatrix} 1 & 1/2 \\ 1/2 & 1 \end{bmatrix}$ (c) $\mathbf{\Sigma} = \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix}$

Figure 2: Contour plots for example bivariate Gaussian distributions. Here $\boldsymbol{\mu} = \mathbf{0}$ for all examples.

Examining these equations, we can see that the multivariate density coincides with the univariate density in the special case when $\mathbf{\Sigma}$ is the scalar σ^2.

Again, the vector $\boldsymbol{\mu}$ specifies the mean of the multivariate Gaussian distribution. The matrix $\mathbf{\Sigma}$ specifies the *covariance* between each pair of variables in $\mathbf{x}$:

$$\mathbf{\Sigma} = \operatorname{cov}(\mathbf{x}, \mathbf{x}) = \mathbb{E}\big[(\mathbf{x} - \boldsymbol{\mu})(\mathbf{x} - \boldsymbol{\mu})^\top\big].$$

Covariance matrices are necessarily symmetric and *positive semidefinite,* which means their eigenvalues are nonnegative. Note that the density function above requires that $\mathbf{\Sigma}$ be *positive definite,* or have strictly positive eigenvalues. A zero eigenvalue would result in a determinant of zero, making the normalization impossible.

The dependence of the multivariate Gaussian density on $\mathbf{x}$ is entirely through the value of the quadratic form

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^\top \mathbf{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}).$$

The value Δ (obtained via a square root) is called the *Mahalanobis distance,* and can be seen as a generalization of the Z score $\frac{(x-\mu)}{\sigma}$, often encountered in statistics.

To understand the behavior of the density geometrically, we can set the Mahalanobis distance to a constant. The set of points in $\mathbb{R}^d$ satisfying $\Delta = c$ for any given value $c > 0$ is an ellipsoid with the eigenvectors of $\mathbf{\Sigma}$ defining the directions of the principal axes.

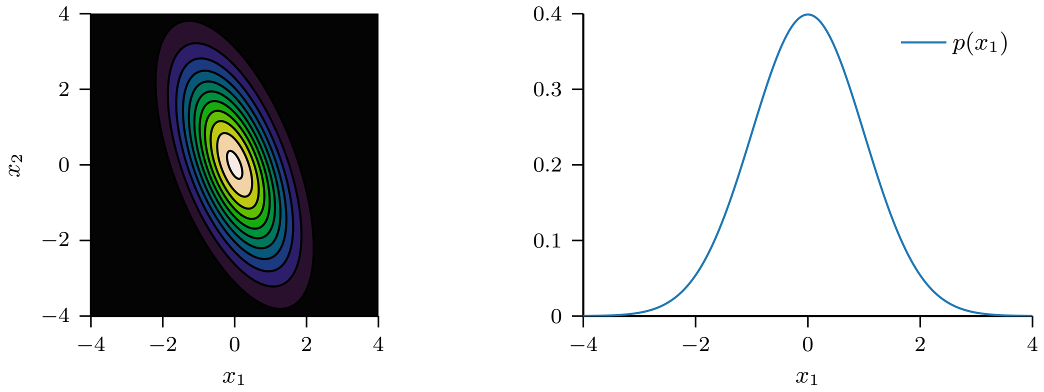
Figure 2 shows contour plots of the density of three bivariate (two-dimensional) Gaussian distributions. The elliptical shape of the contours is clear.

The Gaussian distribution has a number of convenient analytic properties, some of which we describe below.

Marginalization

Often we will have a set of variables $\mathbf{x}$ with a joint multivariate Gaussian distribution, but only be interested in reasoning about a subset of these variables. Suppose $\mathbf{x}$ has a multivariate Gaussian distribution:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \mathbf{\Sigma}) = \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}, \mathbf{\Sigma}).$$



(a) $p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

(b) $p(x_1 \mid \mu_1, \Sigma_{11}) = \mathcal{N}(x_1; 0, 1)$

Figure 3: Marginalization example. (a) shows the joint density over $\mathbf{x} = [x_1, x_2]^\top$; this is the same density as in Figure 2(c). (b) shows the marginal density of x_1.

Let us partition the vector into two components:

$$\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}.$$

We partition the mean vector and covariance matrix in the same way:

$$\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix} \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}.$$

Now the marginal distribution of the subvector $\mathbf{x}_1$ has a simple form:

$$p(\mathbf{x}_1 \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}_1, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}),$$

so we simply pick out the entries of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ corresponding to $\mathbf{x}_1$.

Figure 3 illustrates the marginal distribution of x_1 for the joint distribution shown in Figure 2(c).

Conditioning

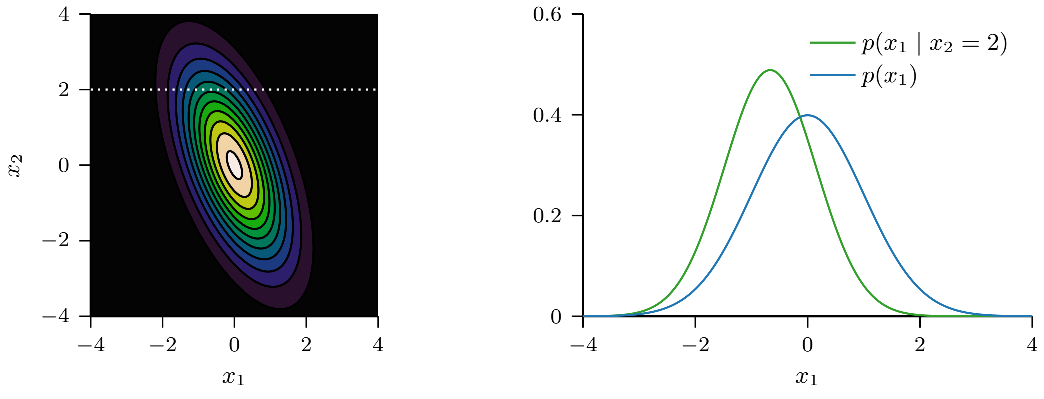
Another common scenario will be when we have a set of variables $\mathbf{x}$ with a joint multivariate Gaussian prior distribution, and are then told the value of a subset of these variables. We may then condition our prior distribution on this observation, giving a posterior distribution over the remaining variables.

Suppose again that $\mathbf{x}$ has a multivariate Gaussian distribution:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}),$$

and that we have partitioned as before: $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2]^\top$. Suppose now that we learn the exact value of the subvector $\mathbf{x}_2$. Remarkably, the posterior distribution

$$p(\mathbf{x}_1 \mid \mathbf{x}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma})$$



(a) $p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

(b) $p(x_1 \mid x_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\big(x_1; -2/3, (2/3)^2\big)$

Figure 4: Conditioning example. (a) shows the joint density over $\mathbf{x} = [x_1, x_2]^\top$, along with the observation value $x_2 = 2$; this is the same density as in Figure 2(c). (b) shows the conditional density of x_1 given $x_2 = 2$.

is a Gaussian distribution! The formula is

$$p(\mathbf{x}_1 \mid \mathbf{x}_2, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}_1; \boldsymbol{\mu}_{1|2}, \boldsymbol{\Sigma}_{11|2}),$$

with

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

So we adjust the mean by an amount dependent on: (1) the covariance between $\mathbf{x}_1$ and $\mathbf{x}_2$, $\boldsymbol{\Sigma}_{12}$, (2) the prior uncertainty in $\mathbf{x}_2$, $\boldsymbol{\Sigma}_{22}$, and (3) the deviation of the observation from the prior mean, $(\mathbf{x}_2 - \boldsymbol{\mu}_2)$. Similarly, we reduce the uncertainty in $\mathbf{x}_1$, $\boldsymbol{\Sigma}_{11}$, by an amount dependent on (1) and (2). Notably, the reduction of the covariance matrix does *not* depend on the values we observe.

Notice that if $\mathbf{x}_1$ and $\mathbf{x}_2$ are independent, then $\boldsymbol{\Sigma}_{12} = \mathbf{0}$, and the conditioning operation does not change the distribution of $\mathbf{x}_1$, as expected.

Figure 4 illustrates the conditional distribution of x_1 for the joint distribution shown in Figure 2(c), after observing $x_2 = 2$.

Convolutions

Gaussian probability density functions are closed under convolutions. Let $\mathbf{x}$ and $\mathbf{y}$ be d-dimensional vectors, with distributions

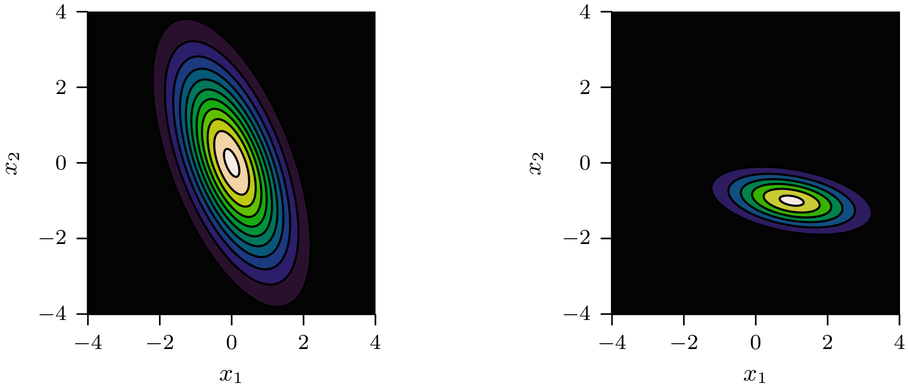
$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma}); \qquad p(\mathbf{y} \mid \boldsymbol{\nu}, \mathbf{P}) = \mathcal{N}(\mathbf{y}; \boldsymbol{\nu}, \mathbf{P}).$$

Then the convolution of their density functions is another Gaussian PDF:

$$f(\mathbf{y}) = \int \mathcal{N}(\mathbf{y} - \mathbf{x}; \boldsymbol{\nu}, \mathbf{P})\,\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})\,\mathrm{d}\mathbf{x} = \mathcal{N}(\mathbf{y}; \boldsymbol{\mu} + \boldsymbol{\nu}, \boldsymbol{\Sigma} + \mathbf{P}),$$

where the mean and covariances add in the result. (This can be proven by considering the Fourier transform of a Gaussian, which is another Gaussian.)

This result will often come in handy.



(a) $p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$

(b) $p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{b})$

Figure 5: Affine transformation example. (a) shows the joint density over $\mathbf{x} = [x_1, x_2]^\top$; this is the same density as in Figure 2(c). (b) shows the density of $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$. The values of $\mathbf{A}$ and $\mathbf{b}$ are given in the text. The density of the transformed vector is another Gaussian.

Affine transformations

Consider a d-dimensional vector $\mathbf{x}$ with a multivariate Gaussian distribution:

$$p(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Suppose we wish to reason about an affine transformation of $\mathbf{x}$ into $\mathbb{R}^D$, $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{b}$, where $\mathbf{A} \in \mathbb{R}^{D \times d}$ and $\mathbf{b} \in \mathbb{R}^D$. Then $\mathbf{y}$ has a D-dimensional Gaussian distribution:

$$p(\mathbf{y} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}, \mathbf{A}, \mathbf{b}) = \mathcal{N}(\mathbf{y}, \mathbf{A}\boldsymbol{\mu} + \mathbf{b}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top).$$

Figure 5 illustrates an affine transformation of the vector $\mathbf{x}$ with the joint distribution shown in Figure 2(c), for the values

$$\mathbf{A} = \begin{bmatrix} 1/5 & -3/5 \\ 1/2 & 3/10 \end{bmatrix}; \qquad \mathbf{b} = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

The density has been rotated and translated, but remains a Gaussian.

Selecting parameters

The d-dimensional multivariate Gaussian distribution is specified by the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Without any further restrictions, specifying $\boldsymbol{\mu}$ requires d parameters and specifying $\boldsymbol{\Sigma}$ requires a further $\binom{d}{2} = \frac{d(d-1)}{2}$. The number of parameters therefore grows quadratically in the dimension, which can sometimes cause difficulty. For this reason, we sometimes restrict the covariance matrix $\boldsymbol{\Sigma}$ in some way to reduce the number of parameters.

Common choices are to set $\boldsymbol{\Sigma} = \operatorname{diag} \boldsymbol{\tau}$, where $\boldsymbol{\tau}$ is a vector of marginal variances, and $\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$, a constant diagonal matrix. Both of these options assume independence between the variables in $\mathbf{x}$. The former case is more flexible, allowing a different scale parameter for each entry, whereas the latter assumes an equal marginal variance of σ^2 for each variable. Geometrically, the densities are axis-aligned, as in Figure 2(a), and in the latter case, the isoprobability contours are spherical (also as in Figure 2(a)).