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 \( \mu \) and \( \sigma^2 \) specify the mean and variance of the distribution, respectively:

\[ \mu = \mathbb{E}[x]; \quad \sigma^2 = \text{var}[x]. \]

Figure 1 plots the probability density function for several sets of parameters \((\mu, \sigma^2)\). 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 probability density function of the multivariate Gaussian distribution is

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

The normalization constant \( Z \) is

\[ Z = \sqrt{\det(2\pi\Sigma)} = (2\pi)^{d/2}(\det \Sigma)^{1/2}. \]
Examining these equations, we can see that the multivariate density coincides with the univariate density in the special case when $\Sigma$ is the scalar $\sigma^2$.

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

$$\Sigma = \text{cov}(x, x) = E[(x - \mu)(x - \mu)^\top].$$

Covariance matrices are necessarily symmetric and positive semidefinite, which means their eigenvalues are nonnegative. Note that the density function above requires that $\Sigma$ 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 $x$ is entirely through the value of the quadratic form

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

The value $\Delta$ (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 $\Sigma$ 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 $x$ with a joint multivariate Gaussian distribution, but only be interested in reasoning about a subset of these variables. Suppose $x$ has a multivariate Gaussian distribution:

$$p(x \mid \mu, \Sigma) = \mathcal{N}(x, \mu, \Sigma).$$
Let us partition the vector into two components:

\[ x = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}. \]

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

\[ \mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}. \]

Now the marginal distribution of the subvector \( x_1 \) has a simple form:

\[ p(x_1 | \mu, \Sigma) = \mathcal{N}(x_1; \mu_1, \Sigma_{11}), \]

so we simply pick out the entries of \( \mu \) and \( \Sigma \) corresponding to \( x_1 \).

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 \( x \) 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 \( x \) has a multivariate Gaussian distribution:

\[ p(x | \mu, \Sigma) = \mathcal{N}(x; \mu, \Sigma), \]

and that we have partitioned as before: \( x = [x_1, x_2]^T \). Suppose now that we learn the exact value of the subvector \( x_2 \). Remarkably, the posterior distribution

\[ p(x_1 | x_2, \mu, \Sigma) \]
is a Gaussian distribution! The formula is 
\[ p(x_1 \mid x_2, \mu, \Sigma) = \mathcal{N}(x_1; \mu_{1|2}, \Sigma_{11|2}), \]
with 
\[
\mu_{1|2} = \mu_1 + \Sigma_{12} \Sigma_{22}^{-1} (x_2 - \mu_2); \\
\Sigma_{11|2} = \Sigma_{11} - \Sigma_{12} \Sigma_{22}^{-1} \Sigma_{21}.
\]

So we adjust the mean by an amount dependent on: (1) the covariance between \(x_1\) and \(x_2, \Sigma_{12}\), (2) the prior uncertainty in \(x_2, \Sigma_{22}\), and (3) the deviation of the observation from the prior mean, \((x_2 - \mu_2)\). Similarly, we reduce the uncertainty in \(x_1, \Sigma_{11}\), 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 \(x_1\) and \(x_2\) are independent, then \(\Sigma_{12} = 0\), and the conditioning operation does not change the distribution of \(x_1\), as expected.

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

**Pointwise multiplication**

Another remarkable fact about multivariate Gaussian density functions is that pointwise multiplication gives another (unnormalized) Gaussian PDF:

\[
\mathcal{N}(x; \mu, \Sigma) \mathcal{N}(x; \nu, P) = \frac{1}{Z} \mathcal{N}(x; \omega, T),
\]
where 
\[
T = (\Sigma^{-1} + P^{-1})^{-1} \\
\omega = T(\Sigma^{-1} \mu + P^{-1} \nu) \\
Z^{-1} = \mathcal{N}(\mu; \nu, \Sigma + P) = \mathcal{N}(\nu; \mu, \Sigma + P).
\]
**Convolutions**

Gaussian probability density functions are closed under convolutions. Let \( x \) and \( y \) be \( d \)-dimensional vectors, with distributions

\[
p(x | \mu, \Sigma) = \mathcal{N}(x; \mu, \Sigma); \quad p(y | \nu, P) = \mathcal{N}(y; \nu, P).
\]

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

\[
f(y) = \int \mathcal{N}(y - x; \nu, P) \mathcal{N}(x; \mu, \Sigma) \, dx = \mathcal{N}(y; \mu + \nu, \Sigma + P),
\]

where the mean and covariances add in the result.

If we assume that \( x \) and \( y \) are independent, then the distribution of their sum \( z = x + y \) will also have a multivariate Gaussian distribution, whose density will precisely the convolution of the individual densities:

\[
p(z | \mu, \nu, \Sigma, P) = \mathcal{N}(z; \mu + \nu, \Sigma + P).
\]

These results will often come in handy.

**Affine transformations**

Consider a \( d \)-dimensional vector \( x \) with a multivariate Gaussian distribution:

\[
p(x | \mu, \Sigma) = \mathcal{N}(x; \mu, \Sigma).
\]

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

\[
p(y | \mu, \Sigma, A, b) = \mathcal{N}(y; A\mu + b, A\Sigma A^\top).
\]
Figure 5 illustrates an affine transformation of the vector $x$ with the joint distribution shown in Figure 2(c), for the values

$$A = \begin{bmatrix} \frac{1}{5} & -\frac{3}{5} \\ \frac{1}{2} & \frac{3}{10} \end{bmatrix}; \quad b = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$  

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

**Selecting parameters**

The $d$-dimensional multivariate Gaussian distribution is specified by the parameters $\mu$ and $\Sigma$. Without any further restrictions, specifying $\mu$ requires $d$ parameters and specifying $\Sigma$ 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 $\Sigma$ in some way to reduce the number of parameters.

Common choices are to set $\Sigma = \text{diag} \, \tau$, where $\tau$ is a vector of marginal variances, and $\Sigma = \sigma^2 I$, a constant diagonal matrix. Both of these options assume independence between the variables in $x$. The former case is more flexible, allowing a different scale parameter for each entry, whereas the latter assumes an equal marginal variance of $\sigma^2$ 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)).