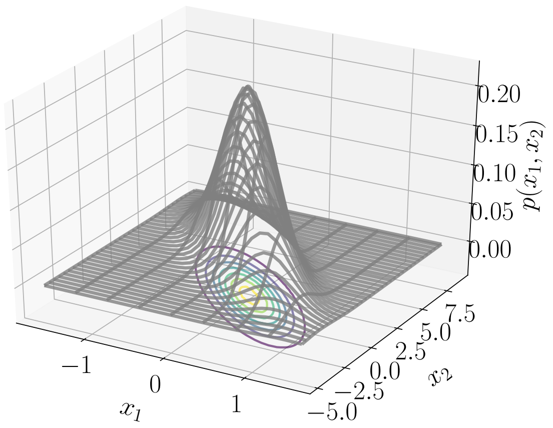


Figure 6.7 Gaussian distribution of two random variables x_1 and x_2.

from the preceding definition of inner products) to compare probability distributions, it is unfortunately not the best way to obtain distances between distributions. Recall that the probability mass (or density) is positive and needs to add up to 1. These constraints mean that distributions live on something called a statistical manifold. The study of this space of probability distributions is called information geometry. Computing distances between distributions are often done using Kullback-Leibler divergence, which is a generalization of distances that account for properties of the statistical manifold. Just like the Euclidean distance is a special case of a metric (Section 3.3), the Kullback-Leibler divergence is a special case of two more general classes of divergences called Bregman divergences and f-divergences. The study of divergences is beyond the scope of this book, and we refer for more details to the recent book by Amari (2016), one of the founders of the field of information geometry. ♢

6.5 Gaussian Distribution

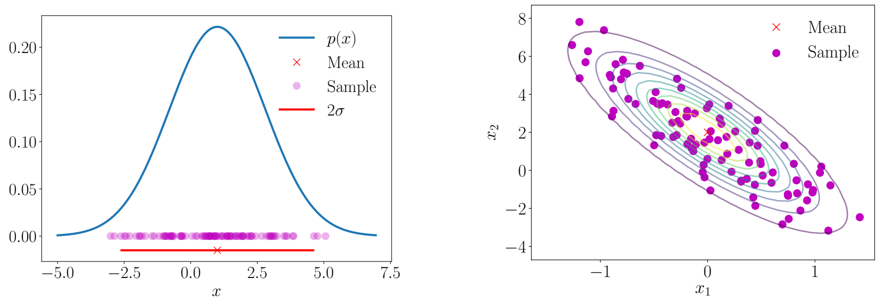
The Gaussian distribution is the most well-studied probability distribution for continuous-valued random variables. It is also referred to as the *normal distribution*. Its importance originates from the fact that it has many computationally convenient properties, which we will be discussing in the following. In particular, we will use it to define the likelihood and prior for linear regression (Chapter 9), and consider a mixture of Gaussians for density estimation (Chapter 11).

normal distribution

The Gaussian distribution arises naturally when we consider sums of independent and identically distributed random variables. This is known as the central limit theorem (Grinstead and Snell, 1997).

There are many other areas of machine learning that also benefit from using a Gaussian distribution, for example Gaussian processes, variational inference, and reinforcement learning. It is also widely used in other application areas such as signal processing (e.g., Kalman filter), control (e.g., linear quadratic regulator), and statistics (e.g., hypothesis testing).

Figure 6.8 Gaussian distributions overlaid with 100 samples. (a) One-dimensional case; (b) two-dimensional case.



(a) Univariate (one-dimensional) Gaussian; The red cross shows the mean and the red line shows the extent of the variance.

(b) Multivariate (two-dimensional) Gaussian, viewed from top. The red cross shows the mean and the colored lines show the contour lines of the density.

For a univariate random variable, the Gaussian distribution has a density that is given by

$$p(x \,|\, \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right). \tag{6.62}$$

multivariate Gaussian distribution

mean vector

covariance matrix

The *multivariate Gaussian distribution* is fully characterized by a *mean vector* $\boldsymbol{\mu}$ and a *covariance matrix* $\boldsymbol{\Sigma}$ and defined as

$$p(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{D}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\big), \tag{6.63}$$

Also known as a multivariate normal distribution.

where $\boldsymbol{x} \in \mathbb{R}^D$. We write $p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x} \,|\, \boldsymbol{\mu}, \boldsymbol{\Sigma})$ or $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Figure 6.7 shows a bivariate Gaussian (mesh), with the corresponding contour plot. Figure 6.8 shows a univariate Gaussian and a bivariate Gaussian with corresponding samples. The special case of the Gaussian with zero mean and identity covariance, that is, $\boldsymbol{\mu} = \boldsymbol{0}$ and $\boldsymbol{\Sigma} = \boldsymbol{I}$, is referred to as the *standard normal distribution*.

standard normal distribution

Gaussians are widely used in statistical estimation and machine learning as they have closed-form expressions for marginal and conditional distributions. In Chapter 9, we use these closed-form expressions extensively for linear regression. A major advantage of modeling with Gaussian random variables is that variable transformations (Section 6.7) are often not needed. Since the Gaussian distribution is fully specified by its mean and covariance, we often can obtain the transformed distribution by applying the transformation to the mean and covariance of the random variable.

6.5.1 Marginals and Conditionals of Gaussians are Gaussians

In the following, we present marginalization and conditioning in the general case of multivariate random variables. If this is confusing at first reading, the reader is advised to consider two univariate random variables instead. Let X and Y be two multivariate random variables, that may have different dimensions. To consider the effect of applying the sum rule of probability and the effect of conditioning, we explicitly write the Gaussian distribution in terms of the concatenated states $[\boldsymbol{x}^\top \ \boldsymbol{y}^\top]^\top$ so that

$$p(\boldsymbol{x}, \boldsymbol{y}) = \mathcal{N}\left(\begin{bmatrix}\boldsymbol{\mu}_x \\ \boldsymbol{\mu}_y\end{bmatrix}, \begin{bmatrix}\boldsymbol{\Sigma}_{xx} & \boldsymbol{\Sigma}_{xy} \\ \boldsymbol{\Sigma}_{yx} & \boldsymbol{\Sigma}_{yy}\end{bmatrix}\right), \tag{6.64}$$

where $\boldsymbol{\Sigma}_{xx} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{x}]$ and $\boldsymbol{\Sigma}_{yy} = \mathrm{Cov}[\boldsymbol{y}, \boldsymbol{y}]$ are the marginal covariance matrices of $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively, and $\boldsymbol{\Sigma}_{xy} = \mathrm{Cov}[\boldsymbol{x}, \boldsymbol{y}]$ is the cross-covariance matrix between $\boldsymbol{x}$ and $\boldsymbol{y}$.

The conditional distribution $p(\boldsymbol{x} \mid \boldsymbol{y})$ is also Gaussian (illustrated in Figure 6.9(c)) and given by (derived in Section 2.3 of Bishop, 2006)

$$p(\boldsymbol{x} \mid \boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{\mu}_{x \mid y}, \, \boldsymbol{\Sigma}_{x \mid y}\big) \tag{6.65}$$

$$\boldsymbol{\mu}_{x \mid y} = \boldsymbol{\mu}_x + \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}(\boldsymbol{y} - \boldsymbol{\mu}_y) \tag{6.66}$$

$$\boldsymbol{\Sigma}_{x \mid y} = \boldsymbol{\Sigma}_{xx} - \boldsymbol{\Sigma}_{xy}\boldsymbol{\Sigma}_{yy}^{-1}\boldsymbol{\Sigma}_{yx} \, . \tag{6.67}$$

Note that in the computation of the mean in (6.66), the $\boldsymbol{y}$-value is an observation and no longer random.

Remark. The conditional Gaussian distribution shows up in many places, where we are interested in posterior distributions:

- The Kalman filter (Kalman, 1960), one of the most central algorithms for state estimation in signal processing, does nothing but computing Gaussian conditionals of joint distributions (Deisenroth and Ohlsson, 2011; Särkkä, 2013).
- Gaussian processes (Rasmussen and Williams, 2006), which are a practical implementation of a distribution over functions. In a Gaussian process, we make assumptions of joint Gaussianity of random variables. By (Gaussian) conditioning on observed data, we can determine a posterior distribution over functions.
- Latent linear Gaussian models (Roweis and Ghahramani, 1999; Murphy, 2012), which include probabilistic principal component analysis (PPCA) (Tipping and Bishop, 1999). We will look at PPCA in more detail in Section 10.7.

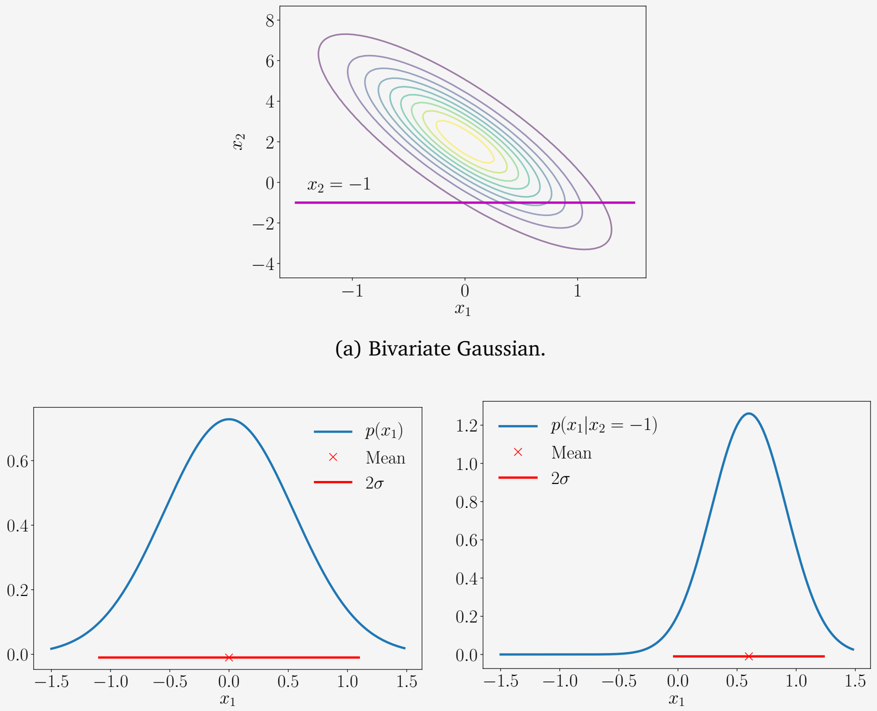
◊

The marginal distribution $p(\boldsymbol{x})$ of a joint Gaussian distribution $p(\boldsymbol{x}, \boldsymbol{y})$ (see (6.64)) is itself Gaussian and computed by applying the sum rule (6.20) and given by

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x}, \boldsymbol{y})\mathrm{d}\boldsymbol{y} = \mathcal{N}\big(\boldsymbol{x} \mid \boldsymbol{\mu}_x, \, \boldsymbol{\Sigma}_{xx}\big) \, . \tag{6.68}$$

The corresponding result holds for $p(\boldsymbol{y})$, which is obtained by marginalizing with respect to $\boldsymbol{x}$. Intuitively, looking at the joint distribution in (6.64), we ignore (i.e., integrate out) everything we are not interested in. This is illustrated in Figure 6.9(b).

Example 6.6



(a) Bivariate Gaussian.

(b) Marginal distribution.

(c) Conditional distribution.

Figure 6.9 (a) Bivariate Gaussian; (b) marginal of a joint Gaussian distribution is Gaussian; (c) the conditional distribution of a Gaussian is also Gaussian.

Consider the bivariate Gaussian distribution (illustrated in Figure 6.9):

$$p(x_1, x_2) = \mathcal{N}\left(\begin{bmatrix}0\\2\end{bmatrix}, \begin{bmatrix}0.3 & -1\\-1 & 5\end{bmatrix}\right). \tag{6.69}$$

We can compute the parameters of the univariate Gaussian, conditioned on $x_2 = -1$, by applying (6.66) and (6.67) to obtain the mean and variance respectively. Numerically, this is

$$\mu_{x_1 \,|\, x_2=-1} = 0 + (-1) \cdot 0.2 \cdot (-1 - 2) = 0.6 \tag{6.70}$$

and

$$\sigma^2_{x_1 \,|\, x_2=-1} = 0.3 - (-1) \cdot 0.2 \cdot (-1) = 0.1\,. \tag{6.71}$$

Therefore, the conditional Gaussian is given by

$$p(x_1 \,|\, x_2 = -1) = \mathcal{N}\big(0.6,\, 0.1\big)\,. \tag{6.72}$$

The marginal distribution $p(x_1)$, in contrast, can be obtained by applying (6.68), which is essentially using the mean and variance of the random variable x_1, giving us

$$p(x_1) = \mathcal{N}\big(0,\, 0.3\big)\,. \tag{6.73}$$

6.5.2 Product of Gaussian Densities

For linear regression (Chapter 9), we need to compute a Gaussian likelihood. Furthermore, we may wish to assume a Gaussian prior (Section 9.3). We apply Bayes' Theorem to compute the posterior, which results in a multiplication of the likelihood and the prior, that is, the multiplication of two Gaussian densities. The *product* of two Gaussians $\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{a},\, \boldsymbol{A}\big)\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{b},\, \boldsymbol{B}\big)$ is a Gaussian distribution scaled by a $c \in \mathbb{R}$, given by $c\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{c},\, \boldsymbol{C}\big)$ with

The derivation is an exercise at the end of this chapter.

$$\boldsymbol{C} = (\boldsymbol{A}^{-1} + \boldsymbol{B}^{-1})^{-1} \tag{6.74}$$

$$\boldsymbol{c} = \boldsymbol{C}(\boldsymbol{A}^{-1}\boldsymbol{a} + \boldsymbol{B}^{-1}\boldsymbol{b}) \tag{6.75}$$

$$c = (2\pi)^{-\frac{D}{2}} |\boldsymbol{A} + \boldsymbol{B}|^{-\frac{1}{2}} \exp\big(-\tfrac{1}{2}(\boldsymbol{a} - \boldsymbol{b})^\top (\boldsymbol{A} + \boldsymbol{B})^{-1}(\boldsymbol{a} - \boldsymbol{b})\big)\,. \tag{6.76}$$

The scaling constant c itself can be written in the form of a Gaussian density either in $\boldsymbol{a}$ or in $\boldsymbol{b}$ with an "inflated" covariance matrix $\boldsymbol{A} + \boldsymbol{B}$, i.e., $c = \mathcal{N}\big(\boldsymbol{a} \,|\, \boldsymbol{b},\, \boldsymbol{A} + \boldsymbol{B}\big) = \mathcal{N}\big(\boldsymbol{b} \,|\, \boldsymbol{a},\, \boldsymbol{A} + \boldsymbol{B}\big)$.

Remark. For notation convenience, we will sometimes use $\mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{m},\, \boldsymbol{S}\big)$ to describe the functional form of a Gaussian density even if $\boldsymbol{x}$ is not a random variable. We have just done this in the preceding demonstration when we wrote

$$c = \mathcal{N}\big(\boldsymbol{a} \,|\, \boldsymbol{b},\, \boldsymbol{A} + \boldsymbol{B}\big) = \mathcal{N}\big(\boldsymbol{b} \,|\, \boldsymbol{a},\, \boldsymbol{A} + \boldsymbol{B}\big)\,. \tag{6.77}$$

Here, neither $\boldsymbol{a}$ nor $\boldsymbol{b}$ are random variables. However, writing c in this way is more compact than (6.76). ◇

6.5.3 Sums and Linear Transformations

If X, Y are independent Gaussian random variables (i.e., the joint distribution is given as $p(\boldsymbol{x}, \boldsymbol{y}) = p(\boldsymbol{x})p(\boldsymbol{y})$) with $p(\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{x} \,|\, \boldsymbol{\mu}_x,\, \boldsymbol{\Sigma}_x\big)$ and $p(\boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{y} \,|\, \boldsymbol{\mu}_y,\, \boldsymbol{\Sigma}_y\big)$, then $\boldsymbol{x} + \boldsymbol{y}$ is also Gaussian distributed and given by

$$p(\boldsymbol{x} + \boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{\mu}_x + \boldsymbol{\mu}_y,\, \boldsymbol{\Sigma}_x + \boldsymbol{\Sigma}_y\big)\,. \tag{6.78}$$

Knowing that $p(\boldsymbol{x} + \boldsymbol{y})$ is Gaussian, the mean and covariance matrix can be determined immediately using the results from (6.46) through (6.49). This property will be important when we consider i.i.d. Gaussian noise acting on random variables, as is the case for linear regression (Chapter 9).

Example 6.7

Since expectations are linear operations, we can obtain the weighted sum of independent Gaussian random variables

$$p(a\boldsymbol{x} + b\boldsymbol{y}) = \mathcal{N}\big(a\boldsymbol{\mu}_x + b\boldsymbol{\mu}_y,\, a^2\boldsymbol{\Sigma}_x + b^2\boldsymbol{\Sigma}_y\big)\,. \tag{6.79}$$

Remark. A case that will be useful in Chapter 11 is the weighted sum of Gaussian densities. This is different from the weighted sum of Gaussian random variables. ♢

In Theorem 6.12, the random variable x is from a density that is a mixture of two densities $p_1(x)$ and $p_2(x)$, weighted by α. The theorem can be generalized to the multivariate random variable case, since linearity of expectations holds also for multivariate random variables. However, the idea of a squared random variable needs to be replaced by $\boldsymbol{x}\boldsymbol{x}^\top$.

Theorem 6.12. *Consider a mixture of two univariate Gaussian densities*

$$p(x) = \alpha p_1(x) + (1-\alpha)p_2(x)\,, \tag{6.80}$$

where the scalar $0 < \alpha < 1$ is the mixture weight, and $p_1(x)$ and $p_2(x)$ are univariate Gaussian densities (Equation (6.62)*) with different parameters, i.e., $(\mu_1, \sigma_1^2) \neq (\mu_2, \sigma_2^2)$.*

Then the mean of the mixture density $p(x)$ is given by the weighted sum of the means of each random variable:

$$\mathbb{E}[x] = \alpha\mu_1 + (1-\alpha)\mu_2\,. \tag{6.81}$$

The variance of the mixture density $p(x)$ is given by

$$\mathbb{V}[x] = \left[\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2\right] + \left(\left[\alpha\mu_1^2 + (1-\alpha)\mu_2^2\right] - \left[\alpha\mu_1 + (1-\alpha)\mu_2\right]^2\right). \tag{6.82}$$

Proof The mean of the mixture density $p(x)$ is given by the weighted sum of the means of each random variable. We apply the definition of the mean (Definition 6.4), and plug in our mixture (6.80), which yields

$$\begin{aligned}
\mathbb{E}[x] &= \int_{-\infty}^{\infty} x p(x)\mathrm{d}x && \text{(6.83a)}\\
&= \int_{-\infty}^{\infty} \left(\alpha x p_1(x) + (1-\alpha) x p_2(x)\right)\mathrm{d}x && \text{(6.83b)}\\
&= \alpha\int_{-\infty}^{\infty} x p_1(x)\mathrm{d}x + (1-\alpha)\int_{-\infty}^{\infty} x p_2(x)\mathrm{d}x && \text{(6.83c)}\\
&= \alpha\mu_1 + (1-\alpha)\mu_2\,. && \text{(6.83d)}
\end{aligned}$$

To compute the variance, we can use the raw-score version of the variance from (6.44), which requires an expression of the expectation of the squared random variable. Here we use the definition of an expectation of a function (the square) of a random variable (Definition 6.3),

$$\begin{aligned}
\mathbb{E}[x^2] &= \int_{-\infty}^{\infty} x^2 p(x)\mathrm{d}x && \text{(6.84a)}\\
&= \int_{-\infty}^{\infty} \left(\alpha x^2 p_1(x) + (1-\alpha) x^2 p_2(x)\right)\mathrm{d}x && \text{(6.84b)}
\end{aligned}$$

$$= \alpha \int_{-\infty}^{\infty} x^2 p_1(x)\mathrm{d}x + (1-\alpha)\int_{-\infty}^{\infty} x^2 p_2(x)\mathrm{d}x \tag{6.84c}$$

$$= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2)\,, \tag{6.84d}$$

where in the last equality, we again used the raw-score version of the variance (6.44) giving $\sigma^2 = \mathbb{E}[x^2] - \mu^2$. This is rearranged such that the expectation of a squared random variable is the sum of the squared mean and the variance.

Therefore, the variance is given by subtracting (6.83d) from (6.84d),

$$\mathbb{V}[x] = \mathbb{E}[x^2] - (\mathbb{E}[x])^2 \tag{6.85a}$$

$$= \alpha(\mu_1^2 + \sigma_1^2) + (1-\alpha)(\mu_2^2 + \sigma_2^2) - (\alpha\mu_1 + (1-\alpha)\mu_2)^2 \tag{6.85b}$$

$$= \left[\alpha\sigma_1^2 + (1-\alpha)\sigma_2^2\right] + \left(\left[\alpha\mu_1^2 + (1-\alpha)\mu_2^2\right] - \left[\alpha\mu_1 + (1-\alpha)\mu_2\right]^2\right)\,. \tag{6.85c}$$

□

Remark. The preceding derivation holds for any density, but since the Gaussian is fully determined by the mean and variance, the mixture density can be determined in closed form. ♢

For a mixture density, the individual components can be considered to be conditional distributions (conditioned on the component identity). Equation (6.85c) is an example of the conditional variance formula, also known as the *law of total variance*, which generally states that for two random variables X and Y it holds that $\mathbb{V}_X[x] = \mathbb{E}_Y[\mathbb{V}_X[x|y]] + \mathbb{V}_Y[\mathbb{E}_X[x|y]]$, i.e., the (total) variance of X is the expected conditional variance plus the variance of a conditional mean.

law of total variance

We consider in Example 6.17 a bivariate standard Gaussian random variable X and performed a linear transformation $\boldsymbol{Ax}$ on it. The outcome is a Gaussian random variable with mean zero and covariance $\boldsymbol{AA}^\top$. Observe that adding a constant vector will change the mean of the distribution, without affecting its variance, that is, the random variable $\boldsymbol{x} + \boldsymbol{\mu}$ is Gaussian with mean $\boldsymbol{\mu}$ and identity covariance. Hence, any linear/affine transformation of a Gaussian random variable is Gaussian distributed.

Any linear/affine transformation of a Gaussian random variable is also Gaussian distributed.

Consider a Gaussian distributed random variable $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For a given matrix $\boldsymbol{A}$ of appropriate shape, let Y be a random variable such that $\boldsymbol{y} = \boldsymbol{Ax}$ is a transformed version of $\boldsymbol{x}$. We can compute the mean of $\boldsymbol{y}$ by exploiting that the expectation is a linear operator (6.50) as follows:

$$\mathbb{E}[\boldsymbol{y}] = \mathbb{E}[\boldsymbol{Ax}] = \boldsymbol{A}\mathbb{E}[\boldsymbol{x}] = \boldsymbol{A\mu}\,. \tag{6.86}$$

Similarly the variance of $\boldsymbol{y}$ can be found by using (6.51):

$$\mathbb{V}[\boldsymbol{y}] = \mathbb{V}[\boldsymbol{Ax}] = \boldsymbol{A}\mathbb{V}[\boldsymbol{x}]\boldsymbol{A}^\top = \boldsymbol{A\Sigma A}^\top\,. \tag{6.87}$$

This means that the random variable $\boldsymbol{y}$ is distributed according to

$$p(\boldsymbol{y}) = \mathcal{N}(\boldsymbol{y}\,|\,\boldsymbol{A\mu},\, \boldsymbol{A\Sigma A}^\top). \tag{6.88}$$

Let us now consider the reverse transformation: when we know that a random variable has a mean that is a linear transformation of another random variable. For a given full rank matrix $\boldsymbol{A} \in \mathbb{R}^{M \times N}$, where $M \geqslant N$, let $\boldsymbol{y} \in \mathbb{R}^M$ be a Gaussian random variable with mean $\boldsymbol{Ax}$, i.e.,

$$p(\boldsymbol{y}) = \mathcal{N}\big(\boldsymbol{y} \,|\, \boldsymbol{Ax},\, \boldsymbol{\Sigma}\big)\,. \tag{6.89}$$

What is the corresponding probability distribution $p(\boldsymbol{x})$? If $\boldsymbol{A}$ is invertible, then we can write $\boldsymbol{x} = \boldsymbol{A}^{-1}\boldsymbol{y}$ and apply the transformation in the previous paragraph. However, in general $\boldsymbol{A}$ is not invertible, and we use an approach similar to that of the pseudo-inverse (3.57). That is, we premultiply both sides with $\boldsymbol{A}^\top$ and then invert $\boldsymbol{A}^\top\boldsymbol{A}$, which is symmetric and positive definite, giving us the relation

$$\boldsymbol{y} = \boldsymbol{Ax} \iff (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y} = \boldsymbol{x}\,. \tag{6.90}$$

Hence, $\boldsymbol{x}$ is a linear transformation of $\boldsymbol{y}$, and we obtain

$$p(\boldsymbol{x}) = \mathcal{N}\big(\boldsymbol{x} \,|\, (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{y},\, (\boldsymbol{A}^\top\boldsymbol{A})^{-1}\boldsymbol{A}^\top\boldsymbol{\Sigma}\boldsymbol{A}(\boldsymbol{A}^\top\boldsymbol{A})^{-1}\big)\,. \tag{6.91}$$

6.5.4 Sampling from Multivariate Gaussian Distributions

We will not explain the subtleties of random sampling on a computer, and the interested reader is referred to Gentle (2004). In the case of a multivariate Gaussian, this process consists of three stages: first, we need a source of pseudo-random numbers that provide a uniform sample in the interval [0,1]; second, we use a non-linear transformation such as the Box-Müller transform (Devroye, 1986) to obtain a sample from a univariate Gaussian; and third, we collate a vector of these samples to obtain a sample from a multivariate standard normal $\mathcal{N}\big(\boldsymbol{0},\, \boldsymbol{I}\big)$.

For a general multivariate Gaussian, that is, where the mean is non zero and the covariance is not the identity matrix, we use the properties of linear transformations of a Gaussian random variable. Assume we are interested in generating samples $\boldsymbol{x}_i, i = 1, \dots, n$, from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. We would like to construct the sample from a sampler that provides samples from the multivariate standard normal $\mathcal{N}\big(\boldsymbol{0},\, \boldsymbol{I}\big)$.

To compute the Cholesky factorization of a matrix, it is required that the matrix is symmetric and positive definite (Section 3.2.3). Covariance matrices possess this property.

To obtain samples from a multivariate normal $\mathcal{N}\big(\boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)$, we can use the properties of a linear transformation of a Gaussian random variable: If $\boldsymbol{x} \sim \mathcal{N}\big(\boldsymbol{0},\, \boldsymbol{I}\big)$, then $\boldsymbol{y} = \boldsymbol{Ax} + \boldsymbol{\mu}$, where $\boldsymbol{AA}^\top = \boldsymbol{\Sigma}$ is Gaussian distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$. One convenient choice of $\boldsymbol{A}$ is to use the Cholesky decomposition (Section 4.3) of the covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{AA}^\top$. The Cholesky decomposition has the benefit that $\boldsymbol{A}$ is triangular, leading to efficient computation.

6.6 Conjugacy and the Exponential Family

Many of the probability distributions "with names" that we find in statistics textbooks were discovered to model particular types of phenomena. For example, we have seen the Gaussian distribution in Section 6.5. The distributions are also related to each other in complex ways (Leemis and McQueston, 2008). For a beginner in the field, it can be overwhelming to figure out which distribution to use. In addition, many of these distributions were discovered at a time that statistics and computation were done by pencil and paper. It is natural to ask what are meaningful concepts in the computing age (Efron and Hastie, 2016). In the previous section, we saw that many of the operations required for inference can be conveniently calculated when the distribution is Gaussian. It is worth recalling at this point the desiderata for manipulating probability distributions in the machine learning context:

"Computers" used to be a job description.

1. There is some "closure property" when applying the rules of probability, e.g., Bayes' theorem. By closure, we mean that applying a particular operation returns an object of the same type.
2. As we collect more data, we do not need more parameters to describe the distribution.
3. Since we are interested in learning from data, we want parameter estimation to behave nicely.

It turns out that the class of distributions called the *exponential family* provides the right balance of generality while retaining favorable computation and inference properties. Before we introduce the exponential family, let us see three more members of "named" probability distributions, the Bernoulli (Example 6.8), Binomial (Example 6.9), and Beta (Example 6.10) distributions.

exponential family

Example 6.8

The *Bernoulli distribution* is a distribution for a single binary random variable X with state $x \in \{0, 1\}$. It is governed by a single continuous parameter $\mu \in [0, 1]$ that represents the probability of $X = 1$. The Bernoulli distribution $\text{Ber}(\mu)$ is defined as

Bernoulli distribution

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x}\,, \quad x \in \{0,1\}\,, \tag{6.92}$$

$$\mathbb{E}[x] = \mu\,, \tag{6.93}$$

$$\mathbb{V}[x] = \mu(1-\mu)\,, \tag{6.94}$$

where $\mathbb{E}[x]$ and $\mathbb{V}[x]$ are the mean and variance of the binary random variable X.

An example where the Bernoulli distribution can be used is when we are interested in modeling the probability of "heads" when flipping a coin.