The normal distribution

Jianxin Wu

LAMDA Group
National Key Lab for Novel Software Technology
Nanjing University, China
wujx2001@gmail.com

March 2, 2020

Contents

The normal distribution is the most widely used probability distribution in statistical pattern recognition, computer vision, and machine learning. The nice properties of this distribution might be the main reason for its popularity.

In this chapter, we try to organize the *basic* facts about the normal distribution. There is no advanced theory in this chapter. However, in order to understand these facts, some linear algebra and mathematical analysis basics are needed, which are not covered sufficiently in undergraduate texts or in Chapter 2. We put this preliminary knowledge at the end of this chapter (Section 9).

1 Definition

We will start by defining the normal distribution.

1.1 Univariate normal

The probability density function (p.d.f.) of a univariate normal distribution has the following form:

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \tag{1}$$

in which μ is the expected value, and σ^2 is the variance. We assume that $\sigma > 0$.

We have to first verify that Equation 1 is a valid probability density function. It is obvious that $p(x) \geq 0$ always holds for $x \in \mathbb{R}$. From Equation 98 in Section 9.1, we know that

$$\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{t}\right) \mathrm{d}x = \sqrt{t\pi}\,.$$

Applying this equation, we have

$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \mathrm{d}x \tag{2}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) \mathrm{d}x \tag{3}$$

$$= \frac{1}{\sqrt{2\pi}\sigma} \sqrt{2\sigma^2\pi} = 1\,, \tag{4}$$

which verifies that $p(x)$ is a valid p.d.f.

The distribution with p.d.f. $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ is called the standard normal distribution (whose $\mu = 0$ and $\sigma^2 = 1$). In Section 9.1, we showed that the

mean and standard deviation of the standard normal distribution are 0 and 1, respectively. By making a change of variables, it is easy to show that

$$\mu = \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x$$

and

$$\sigma^2 = \int_{-\infty}^{\infty} (x-\mu)^2 p(x)\,\mathrm{d}x$$

for a general normal distribution.

1.2 Multivariate normal

The probability density function of a multivariate normal distribution X has the following form:

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{5}$$

in which $\boldsymbol{x}$ is a d-dimensional vector, $\boldsymbol{\mu}$ is the d-dimensional mean, and Σ is the d-by-d covariance matrix. We assume that Σ is a symmetric positive definite matrix.

We have to first verify that Equation 5 is a valid probability density function. It is obvious that $p(\boldsymbol{x}) \geq 0$ always holds for $\boldsymbol{x} \in \mathbb{R}^d$. Next, we diagonalize Σ as $\Sigma = U^T\Lambda U$ in which U is an orthogonal matrix containing the eigenvectors of Σ, $\Lambda = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_d)$ is a diagonal matrix containing the eigenvalues of Σ in its diagonal entries, and their determinants satisfy

$$|\Lambda| = |\Sigma|\,.$$

Let us define a new random vector Y as

$$\boldsymbol{y} = \Lambda^{-1/2}U(\boldsymbol{x}-\boldsymbol{\mu})\,. \tag{6}$$

The mapping from $\boldsymbol{y}$ to $\boldsymbol{x}$ is one-to-one. The determinant of the Jacobian is

$$\left|\frac{\partial \boldsymbol{y}}{\partial \boldsymbol{x}}\right| = |\Lambda^{-1/2}U| = |\Sigma|^{-1/2}$$

because $|U| = 1$ and $|\Lambda| = |\Sigma|$. Now we are ready to calculate the integral

$$\int p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)\,\mathrm{d}\boldsymbol{x} \tag{7}$$

$$= \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} |\Sigma|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right)\,\mathrm{d}\boldsymbol{y} \tag{8}$$

$$= \prod_{i=1}^{d}\left(\int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2}\right)\,\mathrm{d}y_i\right) \tag{9}$$

$$= \prod_{i=1}^{d} 1 \tag{10}$$

$$= 1\,, \tag{11}$$

in which y_i is the i-th component of $\boldsymbol{y}$—i.e., $\boldsymbol{y} = (y_1, y_2, \ldots, y_d)^T$. This equation gives the validity of the multivariate normal density function.

Since $\boldsymbol{y}$ is a random vector, it has a density, which we denote as $p_Y(\boldsymbol{y})$. Using the inverse transform method, we get

$$p_Y(\boldsymbol{y}) = p_X\left(\boldsymbol{\mu} + U^T\Lambda^{1/2}\boldsymbol{y}\right)\left|U^T\Lambda^{1/2}\right| \tag{12}$$

$$= \frac{\left|U^T\Lambda^{1/2}\right|}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(U^T\Lambda^{1/2}\boldsymbol{y}\right)^T\Sigma^{-1}\left(U^T\Lambda^{1/2}\boldsymbol{y}\right)\right) \tag{13}$$

$$= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right)\,. \tag{14}$$

The density defined by

$$p_Y(\boldsymbol{y}) = \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right) \tag{15}$$

is called a spherical normal distribution.

Let $\boldsymbol{z}$ be a random vector formed by a subset of the components of $\boldsymbol{y}$. By marginalizing $\boldsymbol{y}$, it is clear that

$$p_Z(\boldsymbol{z}) = \frac{1}{(2\pi)^{d_z/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right)\,,$$

in which d_z is the dimensionality of $\boldsymbol{z}$. More specifically, we have

$$p_{Y_i}(y_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2}\right)\,.$$

Using this fact, it is straightforward to show that the mean vector and covariance matrix of a spherical normal distribution are $\mathbf{0}$ and I, respectively.

Using the inverse transform of Equation 6, we can easily calculate the mean vector and covariance matrix of the density $p(\boldsymbol{x})$:

$$\mathbb{E}[\boldsymbol{x}] = \mathbb{E}[\boldsymbol{\mu} + U^T\Lambda^{1/2}\boldsymbol{y}] \tag{16}$$

$$= \boldsymbol{\mu} + \mathbb{E}[U^T\Lambda^{1/2}\boldsymbol{y}] \tag{17}$$

$$= \boldsymbol{\mu}\,, \tag{18}$$

$$\mathbb{E}\left[(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T\right] = \mathbb{E}\left[(U^T\Lambda^{1/2}\boldsymbol{y})(U^T\Lambda^{1/2}\boldsymbol{y})^T\right] \tag{19}$$

$$= U^T\Lambda^{1/2}\mathbb{E}[\boldsymbol{y}\boldsymbol{y}^T]\Lambda^{1/2}U \tag{20}$$

$$= U^T\Lambda^{1/2}\Lambda^{1/2}U \tag{21}$$

$$= \Sigma\,. \tag{22}$$

2 Notation and parameterization

When we have a density of the form in Equation 5, it is often written as

$$X \sim N(\boldsymbol{\mu}, \Sigma)\,, \tag{23}$$

or

$$N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma)\,. \tag{24}$$

In most cases we will use the mean vector $\boldsymbol{\mu}$ and the covariance matrix Σ to express a normal density. This is called the *moment parameterization*. There is another parameterization of the normal density. In the *canonical parameterization*, a normal density is expressed as

$$p(\boldsymbol{x}) = \exp\left(\alpha + \boldsymbol{\eta}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\Lambda\boldsymbol{x}\right)\,, \tag{25}$$

in which

$$\alpha = -\frac{1}{2}\left(d\log(2\pi) - \log(|\Lambda|) + \boldsymbol{\eta}^T\Lambda^{-1}\boldsymbol{\eta}\right)$$

is a normalization constant which does not depend on $\boldsymbol{x}$. The parameters in these two representations are related to each other by the following equations:

$$\Lambda = \Sigma^{-1}\,, \tag{26}$$

$$\boldsymbol{\eta} = \Sigma^{-1}\boldsymbol{\mu}\,, \tag{27}$$

$$\Sigma = \Lambda^{-1}\,, \tag{28}$$

$$\boldsymbol{\mu} = \Lambda^{-1}\boldsymbol{\eta}\,. \tag{29}$$

Notice that there is a conflict in our notation: Λ has different meanings in Equation 25 and Equation 6. In Equation 25, Λ is a parameter in the canonical parameterization of a normal density, which is not necessarily diagonal. In Equation 6, Λ is a diagonal matrix formed by the eigenvalues of Σ.

It is straightforward to show that the moment parameterization and the canonical parameterization of the normal distribution are equivalent to each other. In some cases the canonical parameterization is more convenient to use than the moment parameterization; an example of this case will be shown later in this chapter.

3 Linear operation and summation

In this section, we will touch on some basic operations among several normal random variables.

3.1 The univariate case

Suppose $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ are two *independent* univariate normal variables. It is obvious that

$$aX_1 + b \sim N(a\mu_1 + b, a^2\sigma_1^2)\,,$$

in which a and b are two scalars.

Now consider a random variable $Z = X_1 + X_2$. The density of Z can be calculated by a convolution, i.e.,

$$p_Z(z) = \int_{-\infty}^{\infty} p_{X_1}(x_1) p_{X_2}(z - x_1)\,\mathrm{d}x_1\,. \tag{30}$$

Define $x_1' = x_1 - \mu_1$; we get

$$\begin{aligned}
p_Z(z) &= \int p_{X_1}(x_1' + \mu_1) p_{X_2}(z - x_1' - \mu_1)\,\mathrm{d}x_1' && (31)\\
&= \frac{1}{2\pi\sigma_1\sigma_2} \int \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{(z - x - \mu_1 - \mu_2)^2}{2\sigma_2^2}\right)\mathrm{d}x && (32)\\
&= \frac{\exp\left(-\frac{(z-\mu_1-\mu_2)^2}{2(\sigma_1^2+\sigma_2^2)}\right)}{2\pi\sigma_1\sigma_2} \int \exp\left(-\frac{\left(x - \frac{(z-\mu_1-\mu_2)\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2}{\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}\right)\mathrm{d}x && (33)\\
&= \frac{1}{2\pi\sigma_1\sigma_2} \exp\left(-\frac{(z - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \sqrt{\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2 + \sigma_2^2}\pi} && (34)\\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2 + \sigma_2^2}} \exp\left(-\frac{(z - \mu_1 - \mu_2)^2}{2(\sigma_1^2 + \sigma_2^2)}\right)\,, && (35)
\end{aligned}$$

in which the transition from the third last to the second last line used the result of Equation 98.

In short, the sum of two univariate normal random variables is again a normal random variable, with the mean value and variance summed respectively, i.e.,

$$Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\,.$$

The summation rule is easily generalized to n independent normal random variables. However, this rule cannot be used if x_1 and x_2 are dependent.

3.2 The multivariate case

Suppose $X \sim N(\boldsymbol{\mu}, \Sigma)$ is a d-dimensional normal random variable, A is a q-by-d matrix and $\boldsymbol{b}$ is a q-dimensional vector; then $Z = AX + \boldsymbol{b}$ is a q-dimensional normal random variable:

$$Z \sim N(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T)\,.$$

This fact is proved using the characteristic function (see Section 9.2). The characteristic function of Z is:

$$\begin{aligned}
\varphi_Z(\boldsymbol{t}) &= \mathbb{E}_Z[\exp(i\boldsymbol{t}^T\boldsymbol{z})] && (36)\\
&= \mathbb{E}_X\left[\exp\left(i\boldsymbol{t}^T(A\boldsymbol{x}+\boldsymbol{b})\right)\right] && (37)\\
&= \exp(i\boldsymbol{t}^T\boldsymbol{b})\mathbb{E}_X\left[\exp\left(i(A^T\boldsymbol{t})^T\boldsymbol{x}\right)\right] && (38)\\
&= \exp(i\boldsymbol{t}^T\boldsymbol{b})\exp\left(i(A^T\boldsymbol{t})^T\boldsymbol{\mu} - \frac{1}{2}(A^T\boldsymbol{t})^T\Sigma(A^T\boldsymbol{t})\right) && (39)\\
&= \exp\left(i\boldsymbol{t}^T(A\boldsymbol{\mu}+\boldsymbol{b}) - \frac{1}{2}\boldsymbol{t}^T(A\Sigma A^T)\boldsymbol{t}\right), && (40)
\end{aligned}$$

in which the transition to the last line used Equation 108 in Section 9.2.

Section 9.2 also states that if a characteristic function $\varphi(\boldsymbol{t})$ is of the form $\exp(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t})$, then the underlying density is normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ. Applying this fact to Equation 40, we immediately get

$$Z \sim N(A\boldsymbol{\mu}+\boldsymbol{b}, A\Sigma A^T). \tag{41}$$

Suppose $X_1 \sim N(\boldsymbol{\mu}_1, \Sigma_1)$ and $X_2 \sim N(\boldsymbol{\mu}_2, \Sigma_2)$ are two independent d-dimensional normal random variables, and define a new random vector $Z = X_1 + X_2$. We can calculate the probability density function $p_Z(\boldsymbol{z})$ using the same method as we used in the univariate case. However, the calculation is complex, and we have to apply the matrix inversion lemma in Section 9.3.

The characteristic function simplifies the calculation. Using Equation 111 in Section 9.2, we get

$$\begin{aligned}
\varphi_Z(\boldsymbol{t}) &= \varphi_X(\boldsymbol{t})\varphi_Y(\boldsymbol{t}) && (42)\\
&= \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu}_1 - \frac{1}{2}\boldsymbol{t}^T\Sigma_1\boldsymbol{t}\right)\exp\left(i\boldsymbol{t}^T\boldsymbol{\mu}_2 - \frac{1}{2}\boldsymbol{t}^T\Sigma_2\boldsymbol{t}\right) && (43)\\
&= \exp\left(i\boldsymbol{t}^T(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2) - \frac{1}{2}\boldsymbol{t}^T(\Sigma_1+\Sigma_2)\boldsymbol{t}\right), && (44)
\end{aligned}$$

which immediately gives us

$$Z \sim N(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2, \Sigma_1+\Sigma_2).$$

The summation of two independent multivariate normal random variables is as easy to compute as in the univariate case: sum the mean vectors and covariance matrices. This rule remains the same for summing several multivariate normal random variables.

Now we use the tool of linear transformation, and revisit Equation 6. For convenience we retype the equation here: $X \sim N(\boldsymbol{\mu}, \Sigma)$, and get Y by

$$\boldsymbol{y} = \Lambda^{-1/2}U(\boldsymbol{x}-\boldsymbol{\mu}). \tag{45}$$

Using the properties of linear transformations on a normal density, Y is indeed normal (in Section 1.2 we painfully calculated $p_Y(\boldsymbol{y})$ using the inverse transform method), and has mean vector $\mathbf{0}$ and covariance matrix I.

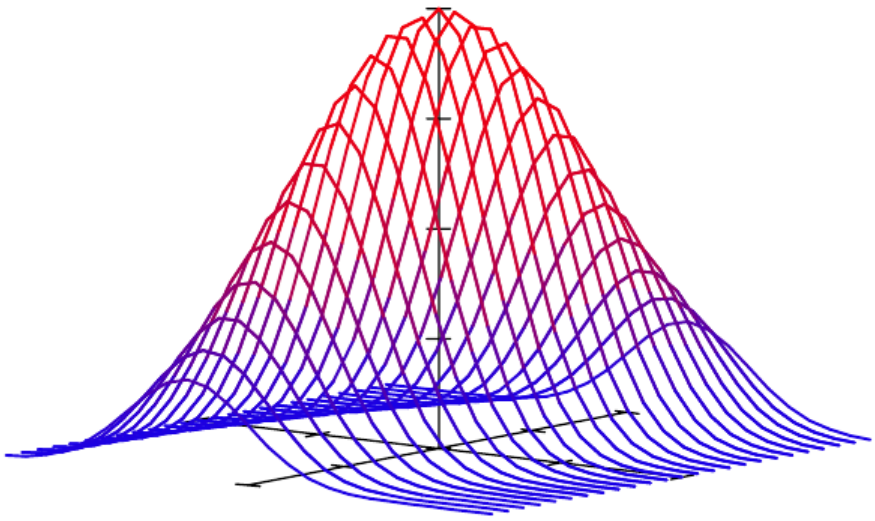

Figure 1: Bivariate normal p.d.f.

The transformation of applying Equation 6 is called the whitening transformation, because the transformed density has an identity covariance matrix and zero mean (cf. Chapter 5).

4 Geometry and the Mahalanobis distance

Figure 1 shows a bivariate normal density function. Normal density has only one mode, which is the mean vector, and the shape of the density is determined by the covariance matrix.

Figure 2 shows the equal probability contour of a bivariate normal random variable. All points on a given equal probability contour must have the following term evaluated to a constant value:

$$r^2(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c\,. \tag{46}$$

$r^2(\boldsymbol{x}, \boldsymbol{\mu})$ is called the Mahalanobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$, given the covariance matrix Σ. Equation 46 defines a hyperellipsoid in d dimensional space, which means that the equal probability contour is a hyperellipsoid in d-dimension space. The principal component axes of this hyperellipsoid are given by the eigenvectors of Σ, and the lengths of these axes are proportional to the square roots of the eigenvalues associated with these eigenvectors (cf. Chapter 5).

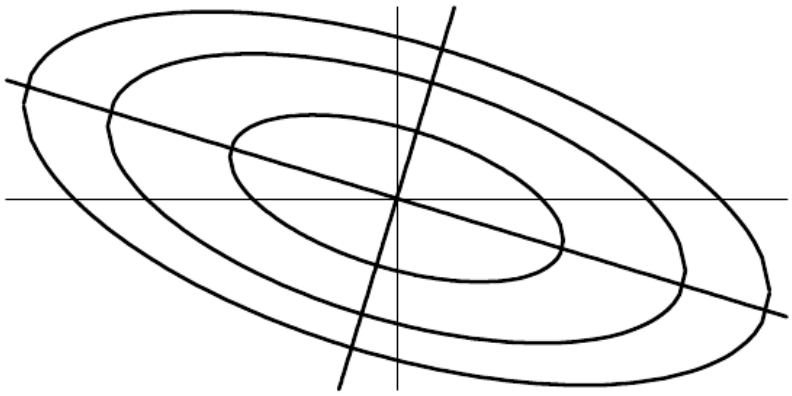

Figure 2: Equal probability contour of a bivariate normal distribution.

5 Conditioning

Suppose X_1 and X_2 are two multivariate normal random variables, which have a joint p.d.f.

$$p\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}\right) = \frac{1}{(2\pi)^{(d_1+d_2)/2}|\Sigma|^{1/2}} \cdot \exp\left(-\frac{1}{2}\begin{bmatrix} \boldsymbol{x}_1-\boldsymbol{\mu}_1 \\ \boldsymbol{x}_2-\boldsymbol{\mu}_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_1-\boldsymbol{\mu}_1 \\ \boldsymbol{x}_2-\boldsymbol{\mu}_2 \end{bmatrix}\right),$$

in which d_1 and d_2 are the dimensionality of X_1 and X_2, respectively; and

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}.$$

The matrices Σ_{12} and Σ_{21} are covariance matrices between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, satisfying

$$\Sigma_{12} = (\Sigma_{21})^T.$$

The marginal distributions $X_1 \sim N(\boldsymbol{\mu}_1, \Sigma_{11})$ and $X_2 \sim N(\boldsymbol{\mu}_2, \Sigma_{22})$ are easy to get from the joint distribution. We are interested in computing the conditional probability $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$.

We will need to compute the inverse of Σ, and this task is completed by using the Schur complement (see Section 9.3). For notational simplicity, we denote the Schur complement of Σ_{11} as S_{11}, defined as

$$S_{11} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}.$$

Similarly, the Schur complement of Σ_{22} is

$$S_{22} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}.$$

Applying Equation 121 and noticing that $\Sigma_{12} = (\Sigma_{21})^T$, we get (writing $\boldsymbol{x}_1 - \boldsymbol{\mu}_1$ as $\boldsymbol{x}'_1$, and $\boldsymbol{x}_2 - \boldsymbol{\mu}_2$ as $\boldsymbol{x}'_2$ for notational simplicity)

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} S_{22}^{-1} & -S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{12}^T\Sigma_{22}^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix}, \quad (47)$$

and

$$\begin{aligned} &\begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \\ &= \boldsymbol{x}'_1 S_{22}^{-1} \boldsymbol{x}'_1 + \boldsymbol{x}_2'^T (\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1})\boldsymbol{x}'_2 \\ &= (\boldsymbol{x}'_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2)^T S_{22}^{-1}(\boldsymbol{x}'_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2) + \boldsymbol{x}_2'^T \Sigma_{22}^{-1}\boldsymbol{x}'_2 \,. \end{aligned} \quad (48)$$

Thus, we can split the joint distribution as

$$\begin{aligned} &p\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}\right) \\ &= \frac{1}{(2\pi)^{d_1}|S_{22}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}'_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2)^T S_{22}^{-1}(\boldsymbol{x}'_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2)}{2}\right) \\ &\quad \cdot \frac{1}{(2\pi)^{d_2}|\Sigma_{22}|^{1/2}} \exp\left(-\frac{1}{2}\boldsymbol{x}_2'^T \Sigma_{22}^{-1}\boldsymbol{x}'_2\right), \end{aligned} \quad (49)$$

in which we used the fact that

$$|\Sigma| = |\Sigma_{22}||S_{22}|\,,$$

a fact that is obvious from Equation 117 in Section 9.3.

Since the second term in the right hand side of Equation 49 is the marginal $p(\boldsymbol{x}_2)$ and $p(\boldsymbol{x}_1, \boldsymbol{x}_2) = p(\boldsymbol{x}_1|\boldsymbol{x}_2)p(\boldsymbol{x}_2)$, we now get the conditional probability $p(\boldsymbol{x}_1|\boldsymbol{x}_2)$ as

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \frac{1}{(2\pi)^{d_1}|S_{22}|^{1/2}} \exp\left(-\frac{(\boldsymbol{x}'_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2)^T S_{22}^{-1}(\boldsymbol{x}'_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2)}{2}\right), \quad (50)$$

or

$$\begin{aligned} \boldsymbol{x}_1|\boldsymbol{x}_2 &\sim N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}'_2, S_{22}) & (51) \\ &\sim N(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\boldsymbol{x}_2 - \boldsymbol{\mu}_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21})\,. & (52) \end{aligned}$$

6 Product of Gaussians

Suppose $X_1 \sim p_1(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)$ and $X_2 \sim p_2(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)$ are two independent d-dimensional normal random variables. Sometimes we want to compute the density, which is proportional to the product of the two normal densities, i.e.,

$$p_X(\boldsymbol{x}) = \alpha p_1(\boldsymbol{x})p_2(\boldsymbol{x})\,,$$

in which α is a proper normalization constant to make $p_X(\boldsymbol{x})$ a valid density function.

In this task, the canonical parameterization (see Section 2) will be extremely helpful. Writing the two normal densities in the canonical form:

$$p_1(\boldsymbol{x}) = \exp\left(\alpha_1 + \boldsymbol{\eta}_1^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \Lambda_1 \boldsymbol{x}\right) \tag{53}$$

$$p_2(\boldsymbol{x}) = \exp\left(\alpha_2 + \boldsymbol{\eta}_2^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \Lambda_2 \boldsymbol{x}\right), \tag{54}$$

the density $p_X(\boldsymbol{x})$ is then easy to compute, as

$$\begin{aligned} p_X(\boldsymbol{x}) &= \alpha p_1(\boldsymbol{x}) p_2(\boldsymbol{x}) \\ &= \exp\left(\alpha' + (\boldsymbol{\eta}_1 + \boldsymbol{\eta}_2)^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T (\Lambda_1 + \Lambda_2)\boldsymbol{x}\right), \end{aligned} \tag{55}$$

in which α' summarizes all terms that are not dependent on $\boldsymbol{x}$. This equation states that in the canonical parameterization, in order to compute the product of two Gaussians, we just sum the parameters.

This result is readily extendable to the product of n normal densities. Suppose we have n normal distributions $p_i(\boldsymbol{x})$, whose parameters in the canonical parameterization are $\boldsymbol{\eta}_i$ and Λ_i, respectively ($i = 1, 2, \ldots, n$). Then, $p_X(\boldsymbol{x}) = \alpha \prod_{i=1}^n p_i(\boldsymbol{x})$ is also a normal density, given by

$$p_X(\boldsymbol{x}) = \exp\left(\alpha' + \left(\sum_{i=1}^n \boldsymbol{\eta}_i\right)^T \boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T \left(\sum_{i=1}^n \Lambda_i\right)\boldsymbol{x}\right). \tag{56}$$

Now let us go back to the moment parameterization. Suppose we have n normal distributions $p_i(\boldsymbol{x})$, in which $p_i(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}_i, \Sigma_i)$, $i = 1, 2, \ldots, n$. Then, $p_X(\boldsymbol{x}) = \alpha \prod_{i=1}^n p_i(\boldsymbol{x})$ is normal,

$$p(\boldsymbol{x}) = N(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma), \tag{57}$$

where

$$\Sigma^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1} + \cdots + \Sigma_n^{-1}, \tag{58}$$

$$\Sigma^{-1}\boldsymbol{\mu} = \Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2 + \cdots + \Sigma_n^{-1}\boldsymbol{\mu}_n. \tag{59}$$

7 Application I: Parameter estimation

Now we have listed some properties of the normal distribution. Next, let us show how these properties are applied.

The first application is parameter estimation in probability and statistics.

7.1 Maximum likelihood estimation

Let us suppose that we have a d-dimensional multivariate normal random variable $X \sim N(\boldsymbol{\mu}, \Sigma)$, and n i.i.d. (independent and identically distributed) samples $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ sampled from this distribution. The task is to estimate the parameters $\boldsymbol{\mu}$ and Σ.

The log-likelihood function of observing the dataset $\mathcal{D}$ given parameters $\boldsymbol{\mu}$ and Σ is:

$$\ell\ell(\boldsymbol{\mu}, \Sigma|\mathcal{D}) \tag{60}$$

$$= \log \prod_{i=1}^{n} p(\boldsymbol{x}_i) \tag{61}$$

$$= -\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\,. \tag{62}$$

Taking the derivative of the log-likelihood with respect to $\boldsymbol{\mu}$ and Σ^{-1} gives (see Section 9.4):

$$\frac{\partial \ell\ell}{\partial \boldsymbol{\mu}} = \sum_{i=1}^{n}\Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu})\,, \tag{63}$$

$$\frac{\partial \ell\ell}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T\,, \tag{64}$$

in which Equation 63 used Equation 126 and the chain rule, and Equation 64 used Equations 133 and 134, and the fact that $\Sigma = \Sigma^T$. The notation in Equation 63 is a little bit confusing. There are two Σs in the right hand side: the first represents a summation and the second represents the covariance matrix.

In order to find the maximum likelihood solution, we want to find the maximum of the likelihood function. Setting both Equation 63 and Equation 64 to $\mathbf{0}$ gives us the solution:

$$\boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i\,, \tag{65}$$

$$\Sigma_{ML} = \frac{1}{n}\sum_{i=1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})^T\,. \tag{66}$$

These two equations clearly state that the maximum likelihood estimation of the mean vector and the covariance matrix are just the sample mean and the sample covariance matrix, respectively.

7.2 Bayesian parameter estimation

In this Bayesian estimation example, we assume that the covariance matrix Σ is known. Let us suppose that we have a d-dimensional multivariate normal

density $X \sim N(\boldsymbol{\mu}, \Sigma)$, and n i.i.d. samples $\mathcal{D} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ sampled from this distribution. We also need a prior on the parameter $\boldsymbol{\mu}$. Let us assume that the prior is $\boldsymbol{\mu} \sim N(\boldsymbol{\mu}_0, \Sigma_0)$. The task is then to estimate the parameters $\boldsymbol{\mu}$.

Note that we assume $\boldsymbol{\mu}_0$, Σ_0, and Σ are all known. The only parameter to be estimated is the mean vector $\boldsymbol{\mu}$.

In Bayesian estimation, instead of finding a point $\hat{\boldsymbol{\mu}}$ in the parameter space that gives maximum likelihood, we calculate $p(\boldsymbol{\mu}|\mathcal{D})$, the posterior density for the parameter. And we use the entire distribution of $\boldsymbol{\mu}$ as our estimation for this parameter.

Applying Bayes' rule, we get

$$p(\boldsymbol{\mu}|\mathcal{D}) = \alpha p(\mathcal{D}|\boldsymbol{\mu}) p_0(\boldsymbol{\mu}) \tag{67}$$

$$= \alpha p_0(\boldsymbol{\mu}) \prod_{i=1}^{n} p(\boldsymbol{x}_i)\,, \tag{68}$$

in which α is a normalization constant that does not depend on $\boldsymbol{\mu}$.

Applying the result in Section 6, we know that $p(\boldsymbol{\mu}|\mathcal{D})$ is also normal, and

$$p(\boldsymbol{\mu}|\mathcal{D}) = N(\boldsymbol{\mu}; \boldsymbol{\mu}_n, \Sigma_n)\,, \tag{69}$$

where

$$\Sigma_n^{-1} = n\Sigma^{-1} + \Sigma_0^{-1}\,, \tag{70}$$

$$\Sigma_n^{-1}\boldsymbol{\mu}_n = n\Sigma^{-1}\boldsymbol{\mu} + \Sigma_0^{-1}\boldsymbol{\mu}_0\,. \tag{71}$$

Both $\boldsymbol{\mu}_n$ and Σ_n can be calculated from known parameters and the dataset. Thus, we have determined the posterior distribution $p(\boldsymbol{\mu}|\mathcal{D})$ for $\boldsymbol{\mu}$.

We choose the normal distribution to be the prior family. Usually, the prior distribution is chosen such that the posterior belongs to the same functional form as the prior. A prior and posterior chosen in this way are said to be *conjugate*. We have just observed that the normal distribution has the nice property that both the prior and the posterior are normal—i.e., the normal distribution is auto-conjugate.

After $p(\boldsymbol{\mu}|\mathcal{D})$ is determined, a new sample is classified by calculating the probability

$$p(\boldsymbol{x}|\mathcal{D}) = \int_{\boldsymbol{\mu}} p(\boldsymbol{x}|\boldsymbol{\mu}) p(\boldsymbol{\mu}|\mathcal{D})\,\mathrm{d}\boldsymbol{\mu}\,. \tag{72}$$

Equation 72 and Equation 31 have the same form. Thus, we can guess that $p(\boldsymbol{x}|\mathcal{D})$ is normal again, and

$$p(\boldsymbol{x}|\mathcal{D}) = N(\boldsymbol{x}; \boldsymbol{\mu}_n, \Sigma + \Sigma_n)\,. \tag{73}$$

This guess is correct, and is easy to verify by repeating the steps in Equation 31 through Equation 35.

8 Application II: Kalman filter

The second application is Kalman filtering.

8.1 The model

The Kalman filter addresses the problem of estimating a state vector $\boldsymbol{x}$ in a discrete time process, given a linear dynamic model

$$\boldsymbol{x}_k = A\boldsymbol{x}_{k-1} + \boldsymbol{w}_{k-1}\,, \tag{74}$$

and a linear measurement model

$$\boldsymbol{z}_k = H\boldsymbol{x}_k + \boldsymbol{v}_k\,. \tag{75}$$

Note that in this example we use lower case letters to denote random variables.

The process noise $\boldsymbol{w}_k$ and measurement noise $\boldsymbol{v}_k$ are assumed to be normal:

$$\boldsymbol{w} \sim N(\boldsymbol{0}, Q)\,, \tag{76}$$

$$\boldsymbol{v} \sim N(\boldsymbol{0}, R)\,. \tag{77}$$

These noises are assumed to be independent of all other random variables.

At time $k-1$, assuming that we know the distribution of $\boldsymbol{x}_{k-1}$, the task is to estimate the posterior probability of $\boldsymbol{x}_k$ at time k, given the current observation $\boldsymbol{z}_k$ and the previous state estimation $p(\boldsymbol{x}_{k-1})$.

From a broader perspective, the task can be formulated as estimating the posterior probability of $\boldsymbol{x}_k$ at time k, given all the previous state estimates and all the observations up to time step k. Under certain Markovian assumptions, it is not hard to prove that these two problem formulations are equivalent.

In the Kalman filter setup, we assume that the prior is normal—i.e., at time $t = 0$, $p(\boldsymbol{x}_0) = N(\boldsymbol{x}; \boldsymbol{\mu}_0, P_0)$. Instead of using Σ, here we use P to represent a covariance matrix, in order to match the notations in the Kalman filter literature.

8.2 The estimation

Now we are ready to see that with the help of the properties of Gaussians we have obtained, it is quite easy to derive the Kalman filter equations. The derivation in this section is neither precise nor rigorous, and mainly provides an intuitive way to interpret the Kalman filter.

The Kalman filter can be separated in two (related) steps. In the first step, based on the estimation $p(\boldsymbol{x}_{k-1})$ and the dynamic model (Equation 74), we get an estimate $p(\boldsymbol{x}_k^-)$. Note that the minus sign means the estimation is done before we take the measurement into account.

In the second step, based on $p(\boldsymbol{x}_k^-)$ and the measurement model (Equation 75), we get the final estimation $p(\boldsymbol{x}_k)$. However, we want to emphasize that this estimation is in fact conditioned on the observation $\boldsymbol{z}_k$ and previous state $\boldsymbol{x}_{t-1}$, although we omitted these dependencies in our notations.

First, let us estimate $p(\boldsymbol{x}_k^-)$. Assume that at time $k-1$, the estimation we already obtained is a normal distribution

$$p(\boldsymbol{x}_{k-1}) \sim N(\boldsymbol{\mu}_{k-1}, P_{k-1})\,. \tag{78}$$

This assumption coincides well with the prior $p(\boldsymbol{x}_0)$. We will show that, under this assumption, after the Kalman filter updates $p(\boldsymbol{x}_k)$ will also become normal, and this makes the assumption a reasonable one.

Applying the linear operation equation (Equation 41) on the dynamic model (Equation 74), we immediately get the estimation for $\boldsymbol{x}_k^-$:

$$\boldsymbol{x}_k^- \sim N(\boldsymbol{\mu}_k^-, P_k^-)\,, \tag{79}$$
$$\boldsymbol{\mu}_k^- = A\boldsymbol{\mu}_{k-1}\,, \tag{80}$$
$$P_k^- = AP_{k-1}A^T + Q\,. \tag{81}$$

The estimate $p(\boldsymbol{x}_k^-)$ conditioned on the observation $\boldsymbol{z}_k$ gives $p(\boldsymbol{x}_k)$, the estimation we want. Thus the conditioning property (Equation 52) can be used.

Without observing $\boldsymbol{z}_k$ at time k, the best estimate for it is

$$H\boldsymbol{x}_k^- + \boldsymbol{v}_k\,,$$

which has a covariance

$$\mathrm{Cov}(\boldsymbol{z}_k) = HP_k^-H^T + R\,,$$

by applying Equation 41 to Equation 75. In order to use Equation 52, we compute

$$\begin{aligned}\mathrm{Cov}(\boldsymbol{z}_k, \boldsymbol{x}_k^-) &= \mathrm{Cov}(H\boldsymbol{x}_k^- + \boldsymbol{v}_k, \boldsymbol{x}_k^-) &(82)\\ &= \mathrm{Cov}(H\boldsymbol{x}_k^-, \boldsymbol{x}_k^-) &(83)\\ &= HP_k^-\,; &(84)\end{aligned}$$

the joint covariance matrix of $(\boldsymbol{x}_k^-, \boldsymbol{z}_k)$ is

$$\begin{bmatrix} P_k^- & P_k^-H^T \\ HP_k^- & HP_k^-H^T + R \end{bmatrix}. \tag{85}$$

Applying the conditioning property (Equation 52), we get

$$\begin{aligned}p(\boldsymbol{x}_k) &= p(\boldsymbol{x}_k^-|\boldsymbol{z}_k) &(86)\\ &\sim N(\boldsymbol{\mu}_k, P_k)\,, &(87)\\ P_k &= P_k^- - P_k^-H^T(HP_k^-H^T + R)^{-1}HP_k^-\,, &(88)\\ \boldsymbol{\mu}_k &= \boldsymbol{\mu}_k^- + P_k^-H^T(HP_k^-H^T + R)^{-1}(\boldsymbol{z}_k - H\boldsymbol{\mu}_k)\,. &(89)\end{aligned}$$

The two sets of equations (Equation 79 to Equation 81, and Equation 86 to Equation 89) are the Kalman filter updating rules.

The term $P_k^- H^T (H P_k^- H^T + R)^{-1}$ appears in both Equation 88 and Equation 89. Defining

$$K_k = P_k^- H^T (H P_k^- H^T + R)^{-1} , \tag{90}$$

these equations are simplified as

$$P_k = (I - K_k H) P_k^- , \tag{91}$$

$$\boldsymbol{\mu}_k = \boldsymbol{\mu}_k^- + K_k (\boldsymbol{z}_k - H \boldsymbol{\mu}_k^-) . \tag{92}$$

The term K_k is called the Kalman gain matrix, and the term $\boldsymbol{z}_k - H\boldsymbol{\mu}_k^-$ is called the innovation.

9 Useful math in this chapter

Although Chapter 2 provides some useful mathematical results, for convenience we supplement this chapter with a few mathematical facts at the end.

9.1 Gaussian integral

We will compute the integral of the univariate normal p.d.f. in this section. The trick in doing this integration is to consider two independent univariate Gaussians at one time.

$$\int_{-\infty}^{\infty} e^{-x^2} \, \mathrm{d}x = \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2} \, \mathrm{d}x \right) \left(\int_{-\infty}^{\infty} e^{-y^2} \, \mathrm{d}y \right)} \tag{93}$$

$$= \sqrt{\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)} \, \mathrm{d}x \, \mathrm{d}y} \tag{94}$$

$$= \sqrt{\int_{0}^{\infty} \int_{0}^{2\pi} r e^{-r^2} \, \mathrm{d}r \mathrm{d}\theta} \tag{95}$$

$$= \sqrt{2\pi \left[-\frac{1}{2} e^{-r^2} \right]_0^{\infty}} \tag{96}$$

$$= \sqrt{\pi} , \tag{97}$$

in which a conversion to polar coordinates is performed in Equation 95, and the extra r that appears inside the equation is the determinant of the Jacobian.

The above integral can be easily extended as

$$f(t) = \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{t} \right) \mathrm{d}x = \sqrt{t\pi} , \tag{98}$$

in which we assume $t > 0$. Then, we have

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{t} \right) \mathrm{d}x \tag{99}$$

$$= \int_{-\infty}^{\infty} \frac{x^2}{t^2} \exp\left(-\frac{x^2}{t}\right) \mathrm{d}x\,, \tag{100}$$

and

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{t}\right) \mathrm{d}x = \frac{t^2}{2}\sqrt{\frac{\pi}{t}}\,. \tag{101}$$

As a direct consequence, we have

$$\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = \frac{1}{\sqrt{2\pi}}\frac{4}{2}\sqrt{\frac{\pi}{2}} = 1\,. \tag{102}$$

And, it is obvious that

$$\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) \mathrm{d}x = 0\,, \tag{103}$$

since $x \exp\left(-\frac{x^2}{2}\right)$ is an odd function.

The last two equations prove that the mean and variance of a standard normal distribution are 0 and 1, respectively.

9.2 Characteristic functions

The characteristic function of a random variable with p.d.f. $p(\boldsymbol{x})$ is defined as its Fourier transform

$$\varphi(\boldsymbol{t}) = \mathbb{E}[e^{i\boldsymbol{t}^T\boldsymbol{x}}]\,, \tag{104}$$

in which $i = \sqrt{-1}$.

Let us compute the characteristic function of a normal random variable:

$$\begin{aligned}
&\varphi(\boldsymbol{t}) && (105)\\
&= \mathbb{E}[\exp(i\boldsymbol{t}^T\boldsymbol{x})] && (106)\\
&= \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + i\boldsymbol{t}^T\boldsymbol{x}\right) \mathrm{d}\boldsymbol{x} && (107)\\
&= \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t}\right)\,. && (108)
\end{aligned}$$

Since the characteristic function is defined as a Fourier transform, the inverse Fourier transform of $\varphi(\boldsymbol{t})$ will be exactly $p(\boldsymbol{x})$—i.e., a random variable is completely determined by its characteristic function. In other words, when we see that a characteristic function $\varphi(\boldsymbol{t})$ is of the form

$$\exp(i\boldsymbol{t}^T\boldsymbol{\mu} - \frac{1}{2}\boldsymbol{t}^T\Sigma\boldsymbol{t})\,,$$

we know that the underlying density is normal with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

Suppose X and Y are two *independent* random vectors with the same dimensionality, and we define a new random vector $Z = X + Y$. Then,

$$p_Z(\boldsymbol{z}) = \iint_{\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{y}} p_X(\boldsymbol{x}) p_Y(\boldsymbol{y}) \, \mathrm{d}\boldsymbol{x} \, \mathrm{d}\boldsymbol{y} \tag{109}$$

$$= \int p_X(\boldsymbol{x}) p_Y(\boldsymbol{z} - \boldsymbol{x}) \, \mathrm{d}\boldsymbol{x} \,, \tag{110}$$

which is a convolution. Since convolution in the function space is a product in the Fourier space, we have

$$\varphi_Z(\boldsymbol{t}) = \varphi_X(\boldsymbol{t}) \varphi_Y(\boldsymbol{t}) \,, \tag{111}$$

which means that the characteristic function of the sum of two independent random variables is just the product of the characteristic functions of the summands.

9.3 Schur complement & Matrix inversion lemma

The Schur complement is very useful in computing the inverse of a block matrix.

Suppose M is a block matrix expressed as

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix} , \tag{112}$$

in which A and D are non-singular square matrices. We want to compute M^{-1}.

Some algebraic manipulations give

$$\begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix} M \begin{bmatrix} I & -A^{-1}B \\ \mathbf{0} & I \end{bmatrix} \tag{113}$$

$$= \begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ \mathbf{0} & I \end{bmatrix} \tag{114}$$

$$= \begin{bmatrix} A & B \\ \mathbf{0} & D - CA^{-1}B \end{bmatrix} \begin{bmatrix} I & -A^{-1}B \\ \mathbf{0} & I \end{bmatrix} \tag{115}$$

$$= \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & D - CA^{-1}B \end{bmatrix} = \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & S_A \end{bmatrix} , \tag{116}$$

in which I and $\mathbf{0}$ are identity and zero matrices of appropriate size, respectively; and the term

$$D - CA^{-1}B$$

is called the *Schur complement of* A, denoted as S_A.

Taking the determinant of both sides of the above equation gives

$$|M| = |A||S_A| \,. \tag{117}$$

Equation $XMY = Z$ implies that $M^{-1} = YZ^{-1}X$ when both X and Y are invertible. Hence, we have

$$M^{-1} = \begin{bmatrix} I & -A^{-1}B \\ \mathbf{0} & I \end{bmatrix} \begin{bmatrix} A & \mathbf{0} \\ \mathbf{0} & S_A \end{bmatrix}^{-1} \begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix} \tag{118}$$

$$= \begin{bmatrix} A^{-1} & -A^{-1}BS_A^{-1} \\ \mathbf{0} & S_A^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix} \tag{119}$$

$$= \begin{bmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{bmatrix} . \tag{120}$$

Similarly, we can also compute M^{-1} by using the Schur complement of D, in the following way:

$$M^{-1} = \begin{bmatrix} S_D^{-1} & -S_D^{-1}BD^{-1} \\ -D^{-1}CS_D^{-1} & D^{-1} + D^{-1}CS_D^{-1}BD^{-1} \end{bmatrix} , \tag{121}$$

$$|M| = |D||S_D| . \tag{122}$$

Equations 120 and 121 are two different representations of the same matrix M^{-1}, which means that the corresponding blocks in these two equations must be equal, for example,

$$S_D^{-1} = A^{-1} + A^{-1}BS_A^{-1}CA^{-1} .$$

This result is known as the *matrix inversion lemma*:

$$S_D^{-1} = (A - BD^{-1}C)^{-1} = A^{-1} + A^{-1}B(D - CA^{-1}B)^{-1}CA^{-1} . \tag{123}$$

The following result, which comes from equating the two upper right blocks, is also useful:

$$A^{-1}B(D - CA^{-1}B)^{-1} = (A - BD^{-1}C)^{-1}BD^{-1} . \tag{124}$$

This formula and the matrix inversion lemma are useful in the derivation of the Kalman filter equations.

9.4 Vector and matrix derivatives

Suppose y is a scalar, A is a matrix, and $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors. The partial derivative of y with respect to A is defined as

$$\left(\frac{\partial y}{\partial A}\right)_{ij} = \frac{\partial y}{\partial a_{ij}} , \tag{125}$$

where a_{ij} is the (i, j)-th component of the matrix A.

Based on this definition, it is easy to get the following rule

$$\frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{x}^T\boldsymbol{y}) = \frac{\partial}{\partial \boldsymbol{x}}(\boldsymbol{y}^T\boldsymbol{x}) = \boldsymbol{y} . \tag{126}$$

For a square matrix A that is n-by-n, the determinant of the matrix defined by removing from A the i-th row and j-th column is called a *minor* of A, and denoted as M_{ij}. The scalar $c_{ij} = (-1)^{i+j}M_{ij}$ is called a *cofactor* of A. The matrix A_{cof} with c_{ij} in its (i, j)-th entry is called the *cofactor matrix* of

A. Finally, the *adjoint* matrix of A is defined as the transpose of the cofactor matrix

$$A_{adj} = A_{cof}^T \,. \tag{127}$$

There are some well-known facts about the minors, determinant, and adjoint of a matrix:

$$|A| = \sum_j a_{ij} c_{ij} \,, \tag{128}$$

$$A^{-1} = \frac{1}{|A|} A_{adj} \,. \tag{129}$$

Since M_{ij} has removed the i-th row, it does not depend on a_{ij}; neither does c_{ij}. Thus, we have

$$\frac{\partial}{\partial a_{ij}} |A| = c_{ij} \,, \qquad \text{or,} \tag{130}$$

$$\frac{\partial}{\partial A} |A| = A_{cof} \,, \tag{131}$$

which in turn shows that

$$\frac{\partial}{\partial A} |A| = A_{cof} = A_{adj}^T = |A| (A^{-1})^T \,. \tag{132}$$

Using the chain rule, we immediately get that for a positive definite matrix A,

$$\frac{\partial}{\partial A} \log |A| = (A^{-1})^T \,. \tag{133}$$

Applying the definition, it is also easy to show that for a square matrix A,

$$\frac{\partial}{\partial A} (\boldsymbol{x}^T A \boldsymbol{x}) = \boldsymbol{x} \boldsymbol{x}^T \,, \tag{134}$$

since $\boldsymbol{x}^T A \boldsymbol{x} = \sum_{i=1}^n \sum_{j=1}^n a_{ij} x_i x_j$, where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^T$.

Exercises

In the exercises for this chapter, we will discuss a few basic properties of the *exponential family*. The exponential family is probably the most important class of distributions, with the normal distribution being a representative of it.

We say a p.d.f. or p.m.f. (for continuous or discrete random vectors) is in the exponential family with parameters $\boldsymbol{\theta}$ if it can be written as the following form:

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x})\right) . \tag{135}$$

Various notations involved in this definition are explained as follows.

- *Canonical parameters.* $\boldsymbol{\theta} \in \mathbb{R}^d$ are the canonical parameters or natural parameters.
- $\boldsymbol{x} \in \mathbb{R}^m$ are the random variables, which can be either continuous or discrete.
- *Sufficient statistics.* $\phi(\boldsymbol{x}) \in \mathbb{R}^d$ is a set of sufficient statistics for $\boldsymbol{x}$. Note that $m = d$ may not (and often does not) hold. Let X be a set of i.i.d. samples from $p(\boldsymbol{x}|\boldsymbol{\theta})$. Loosely speaking, the term "sufficient" means that the set $\phi(X)$ contains all the information (i.e., sufficient) to estimate the parameters $\boldsymbol{\theta}$. Obviously, $\phi(\boldsymbol{x}) = \boldsymbol{x}$ is a trivial set of sufficient statistics with respect to $\boldsymbol{x}$.
- $h(\boldsymbol{x})$ is a scaling function. Note that $h(\boldsymbol{x}) \geq 0$ is required to make $p(\boldsymbol{x}|\boldsymbol{\theta})$ a valid p.d.f. or p.m.f.
- *Partition function.* $Z(\boldsymbol{\theta})$ is called a partition function, whose role is to make $p(\boldsymbol{x}|\boldsymbol{\theta})$ integrate (or sum) to 1. Hence,

  $$Z(\boldsymbol{\theta}) = \int h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x})\right) \, \mathrm{d}\boldsymbol{x}$$

  in the continuous case. In the discrete case, we simply replace the integration with a summation.
- *Cumulant function.* We can define a *log partition function* $A(\boldsymbol{\theta})$, as

  $$A(\boldsymbol{\theta}) = \log(Z(\boldsymbol{\theta})) .$$

  With such a new notation, Equation 135 has an equivalent form:

  $$p(\boldsymbol{x}|\boldsymbol{\theta}) = h(\boldsymbol{x}) \exp\left(\boldsymbol{\theta}^T \phi(\boldsymbol{x}) - A(\boldsymbol{\theta})\right) . \tag{136}$$

  $A(\boldsymbol{\theta})$ is also called a cumulant function, the meaning of which will be made clear soon.

Note that these functions or statistics are not unique. For example, we can multiply the parameters $\boldsymbol{\theta}$ by a constant $c > 0$, multiply the sufficient statistics ϕ by $1/c$, and change h and Z accordingly to obtain an equivalent $p(\boldsymbol{x}|\boldsymbol{\theta})$. Similarly, we can change the scale h and partition function Z simultaneously. It is often the case that we choose $h(\boldsymbol{x}) = 1$ for any $\boldsymbol{x}$.

1. Answer the following questions.

   (a) Show that the canonical parameterization (Equation 25) is equivalent to the more common moment parameterization in Equation 5.

   (b) Show that a normal distribution is in the exponential family.

   (c) The Bernoulli distribution is a discrete distribution. A Bernoulli random variable X can be either 0 or 1, with $\Pr(X = 1) = \pi$ and $\Pr(X = 0) = 1 - \pi$, in which $0 \leq \pi \leq 1$. Show that the Bernoulli distribution is in the exponential family.

2. (Cumulant function) In statistics, the first cumulant of a random variable X is the expectation $\mathbb{E}[X]$, and the second cumulant is the variance (or covariance matrix) $\mathbb{E}[(X - \mathbb{E}X)^2]$. In the exponential family, the cumulant function $A(\boldsymbol{\theta})$ has close relationships to these cumulants of the sufficient statistics $\phi(\boldsymbol{x})$.

   (a) Prove that

   $$\frac{\partial A}{\partial \boldsymbol{\theta}} = \mathbb{E}[\phi(X)] \,.$$

   (Hint: You can exchange the order of the integration and differentiation operators in this case safely.)

   (b) Prove that

   $$\frac{\partial^2 A}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \mathrm{Var}\left(\phi(X)\right) \,.$$

   (c) Use the above theorems to find the expectation and variance of the Bernoulli distribution. Check the correctness of your calculations using the definition of mean and variance.

3. (Beta distributions) The beta distribution is a continuous distribution. The support of a beta distribution is $[0, 1]$—i.e., its p.d.f. is 0 for values that are negative or larger than 1. For simplicity, we use the range $(0, 1)$ as a beta distribution's support in this problem—i.e., excluding $x = 0$ and $x = 1$.

   A beta distribution has two *shape parameters* $\alpha > 0$ and $\beta > 0$, which determine the shape of the distribution. And, a beta random variable is often denoted as $X \sim \mathrm{Beta}(\alpha, \beta)$ when the two shape parameters are α and β, respectively. Note that $\mathrm{Beta}(\alpha, \beta)$ and $\mathrm{Beta}(\beta, \alpha)$ are two different distributions when $\alpha \neq \beta$.

   (a) The p.d.f. of a beta distribution is

   $$p(x) = \frac{1}{\mathrm{B}(\alpha, \beta)} x^{\alpha - 1} (1 - x)^{\beta - 1} \,, \tag{137}$$

for $0 < x < 1$, in which

$$\mathrm{B}(\alpha, \beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$$

is the *Beta function*. The p.d.f. is zero for other x values. Show that a beta distribution is in the exponential family. What is the partition function?

(b) The Gamma function is defined as

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\,\mathrm{d}t\,.$$

Read the information at https://en.wikipedia.org/wiki/Gamma_function and https://en.wikipedia.org/wiki/Beta_function to pick up a few important properties of the Gamma and Beta functions, especially the following ones (proofs are not required):

i. $\Gamma(0.5) = \sqrt{\pi}$

ii. $\Gamma(-\frac{1}{2}) = -2\sqrt{\pi}$

iii. $\Gamma(n) = (n-1)!$ for any positive integer n, in which ! means the factorial function.

iv. $\mathrm{B}(x, y) = \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)}$

v. $\mathrm{B}(x+1, y) = \mathrm{B}(x, y) \cdot \frac{x}{x+y}$

vi. $\mathrm{B}(x, y+1) = \mathrm{B}(x, y) \cdot \frac{y}{x+y}$

(c) Write your own code to draw curves for the p.d.f. of $\mathrm{Beta}(0.5, 0.5)$, $\mathrm{Beta}(1, 5)$, and $\mathrm{Beta}(2, 2)$. Calculate the p.d.f. values for $x = 0.01n$ using Equation 137 to draw the curves, where $1 \le n \le 99$ enumerates positive integers between 1 and 99.

4. (Conjugate prior) The exponential family is particularly useful in Bayesian analysis, because their conjugate priors exist. Given a distribution $p(\boldsymbol{x}|\boldsymbol{\theta})$ in the exponential family (hence the likelihood function is in the exponential family too), we can always find another distribution $p(\boldsymbol{\theta})$ such that the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{x})$ is in the same family as that of $p(\boldsymbol{\theta})$. We say the prior is a *conjugate prior for the likelihood function* if the prior and the posterior have the same form. In this chapter, we have shown that the normal distribution is conjugate to itself.

   In this problem, we will use the Bernoulli-Beta pair as an example to further illustrate the conjugate priors for exponential family distributions. Similar procedures can be extended to handle other exponential family distributions and the exponential family in general.

   (a) Let $\mathcal{D} = \{x_1, x_2, \ldots, x_n\}$ be i.i.d. samples from a Bernoulli distribution with $\Pr(X = 1) = \pi$. Show that the likelihood function is

$$p(\mathcal{D}|\pi) = (1-\pi)^n \exp\left(\ln(\frac{\pi}{1-\pi})\sum_{i=1}^n x_i\right)\,.$$

(b) Because $\theta^x \cdot \theta^y = \theta^{x+y}$, it is natural to set the prior $p(\pi)$ to the following form:

$$p(\pi|\nu_0, \tau_0) = c(1-\pi)^{\nu_0} \exp\left(\ln(\frac{\pi}{1-\pi})\tau_0\right) ,$$

in which $c > 0$ is a normalization constant, and ν_0 and τ_0 are parameters for the prior distribution. Show that

$$p(\pi|\nu_0, \tau_0) \propto \pi^{\tau_0}(1-\pi)^{\nu_0-\tau_0} ,$$

and further show it is a Beta distribution. What are the parameters of this Beta distribution? And, what is the value of c in terms of ν_0 and τ_0?

(c) Show that the posterior $p(\pi|\mathcal{D})$ is a Beta distribution. What are the parameters of this Beta distribution?

(d) Intuitively explain what the prior does.