Some Properties Of the Gaussian Distribution

Jianxin Wu
GVU Center and College of Computing
Georgia Institute of Technology

April 22, 2004

Contents

1 Introduction

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

In this short paper[1], I try to organize the basic facts about the Gaussian distribution. There is no advanced theory in this paper. However, in order to understand these facts, some linear algebra and multivariate analysis are needed, which are not always covered sufficiently in undergraduate texts. The attempt of this paper is to pool these facts together, and hope that it will be useful for new researchers entering this area.

2 Definition

2.1 Univariate Gaussian

The probability density function of a univariate Gaussian distribution has the following form:

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

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

We have to first verify that eq. (1) is a valid density. It is obvious that $p(x) \geq 0$ always holds for $\boldsymbol{x} \in \mathcal{R}$. From eq. (96) in Appendix A we know that $\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{t}\right) dx = \sqrt{t\pi}$. Applying this equation, we have

$$\begin{aligned}
\int_{-\infty}^{\infty} p(x)\, dx &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) dx && (2)\\
&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{2\sigma^2}\right) dx && (3)\\
&= \frac{1}{\sqrt{2\pi}\sigma} \sqrt{2\sigma^2\pi} = 1, && (4)
\end{aligned}$$

which means that $p(x)$ is a valid density.

The density $\frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right)$ is called the standard normal density. In Appendix A, it is showed that the mean value and standard deviation of the standard normal distribution are 0 and 1 respectively. By doing a change of variables, it is easy to show that $\mu = \int xp(x)\, dx$ and $\sigma^2 = \int (x-\mu)^2 p(x)\, dx$.

[1] I planned to write a short note listing some properties of Gaussian. However, somehow I decided to keep this paper self-containing. The result is that it becomes very fat.

2.2 Multivariate Gaussian

The probability density function of a multivariate Gaussian distribution 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 vector, 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 eq. (5) is a valid probability density function. It is obvious that $p(\boldsymbol{x}) \geq 0$ always holds for $\boldsymbol{x} \in \mathcal{R}^d$. Next we diagonalize Σ as $\Sigma = U^T \Lambda U$ in which U is an orthogonal matrix containing the eigenvectors of Σ, $\Lambda = [\lambda_1, \ldots, \lambda_d]$ is a diagonal matrix containing the eigenvalues of Σ in its diagonal entries and $|\Lambda| = |\Sigma|$. Let's define a new random vector as

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

If we treat eq. (6) as a change of variables, the determinant of the Jacobian matrix will be $\left|\Lambda^{-1/2}\right| = |\Sigma|^{-1/2}$. Now we are ready to calculate the integral

$$\begin{aligned}
\int p(\boldsymbol{x})\, d\boldsymbol{x} &= \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right) d\boldsymbol{x} && (7)\\
&= \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} |\Sigma|^{1/2} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right) d\boldsymbol{y} && (8)\\
&= \prod_{i=1}^{d} \left(\int \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2}\right) dy_i\right) && (9)\\
&= \prod_{i=1}^{d} 1 = 1 && (10)
\end{aligned}$$

in which y_i is the ith component of $\boldsymbol{y}$, i.e. $\boldsymbol{y} = (y_1, \ldots, y_d)$. This equation gives the validity of the multivariate Gaussian density.

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

$$\begin{aligned}
p_{\boldsymbol{y}}(\boldsymbol{y}) &= p\left(\boldsymbol{\mu} + U^T\Lambda^{1/2}\boldsymbol{y}\right)\left|U^T\Lambda^{1/2}\right| && (11)\\
&= \frac{\left|U^T\Lambda^{1/2}\right|}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}\left(U^T\Lambda^{1/2}\boldsymbol{y}\right)^T \Sigma^{-1}\left(U^T\Lambda^{1/2}\boldsymbol{y}\right)\right) && (12)\\
&= \frac{1}{(2\pi)^{d/2}} \exp\left(-\frac{1}{2}\boldsymbol{y}^T\boldsymbol{y}\right) && (13)
\end{aligned}$$

The density defined by

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

is called a spherical Gaussian distribution. Let $\boldsymbol{z}$ be a random vector formed by a subset of the components of $\boldsymbol{y}$. By marginalization it is clear that $p(\boldsymbol{z}) = \frac{1}{(2\pi)^{|\boldsymbol{z}|/2}} \exp\left(-\frac{1}{2}\boldsymbol{z}^T\boldsymbol{z}\right)$, and specifically $p(y_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{y_i^2}{2}\right)$. Using this fact, it is straightforward to show that the mean vector and covariance matrix of a spherical Gaussian are $\mathbf{0}$ and I respectively.

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

$$\begin{aligned}
\mathcal{E}\boldsymbol{x} &= \mathcal{E}\left(\boldsymbol{\mu} + U^T\Lambda^{1/2}\boldsymbol{y}\right) = \boldsymbol{\mu} + \mathcal{E}\left(U^T\Lambda^{1/2}\boldsymbol{y}\right) = \boldsymbol{\mu} && (15)\\
\mathcal{E}(\boldsymbol{x}-\boldsymbol{\mu})(\boldsymbol{x}-\boldsymbol{\mu})^T &= \mathcal{E}\left(U^T\Lambda^{1/2}\boldsymbol{y}\right)\left(U^T\Lambda^{1/2}\boldsymbol{y}\right)^T && (16)\\
&= U^T\Lambda^{1/2}\mathcal{E}\left(\boldsymbol{y}\boldsymbol{y}^T\right)\Lambda^{1/2}U && (17)\\
&= U^T\Lambda^{1/2}\Lambda^{1/2}U && (18)\\
&= \Sigma && (19)
\end{aligned}$$

3 Notation and Parameterization

When we have a density of the form in eq. (5), it is often written as

$$\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \tag{20}$$

or,

$$\mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \tag{21}$$

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

$$p(\boldsymbol{x}) = \exp\left(\alpha + \boldsymbol{\eta}^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\Lambda\boldsymbol{x}\right), \tag{22}$$

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. The parameters in these two representations are related by

$$\begin{aligned}
\Lambda &= \Sigma^{-1} && (23)\\
\boldsymbol{\eta} &= \Sigma^{-1}\boldsymbol{\mu} && (24)\\
\Sigma &= \Lambda^{-1} && (25)\\
\boldsymbol{\mu} &= \Lambda^{-1}\boldsymbol{\eta}. && (26)
\end{aligned}$$

Notice that there is a confusion in our notation: Λ has different meanings in eq. (22) and eq. (6). In eq. (22), Λ is a parameter in the canonical parameterization of a Gaussian density, which is not necessarily diagonal. In eq. (6), Λ is a diagonal matrix formed by the eigenvalues of Σ. It is straightforward to show that the moment parameterization and canonical parameterization of the Gaussian distribution are equivalent. In some cases the canonical parameterization is more convenient to use than the moment parameterization.

4 Linear Operation and Summation

4.1 Univariate case

Suppose $x_1 \sim \mathcal{N}\left(\mu_1, \sigma_1^2\right)$ and $x_2 \sim \mathcal{N}\left(\mu_2, \sigma_2^2\right)$ are two independent univariate Gaussian densities. It is obvious that $ax_1 + b \sim \mathcal{N}\left(a\mu_1 + b, a^2\sigma_1^2\right)$, in which a and b are two scalars.

Now consider a random variable $z = x_1 + x_2$. The density of z is calculated as:

$$
\begin{aligned}
p_z(z) &= \int\int_{z=x_1+x_2} p_{x_1}(x_1)\, p_{x_2}(x_2)\, dx_1 dx_2 && (27)\\
&= \int\int_{z=x_1'+x_2'-\mu_1-\mu_2} p_{x_1}(x_1+\mu_1)\, p_{x_2}(x_2+\mu_2)\, dx_1' dx_2' && (28)\\
&= \int_{x_1'} p_{x_1}(x_1'+\mu_1)\, p_{x_2}(z-x_1'-\mu_1)\, dx_1' && (29)\\
&= \frac{1}{2\pi\sigma_1\sigma_2}\int_x \exp\left(-\frac{x^2}{2\sigma_1^2} - \frac{(z-x-\mu_1-\mu_2)^2}{2\sigma_2^2}\right) dx && (30)\\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(\frac{(z-\mu_1)^2}{\sigma_1^2+\sigma_2^2}\right)\int_x \exp\left(-\frac{\left(x-\frac{(z-\mu_1)\sigma_1^2}{\sigma_1^2+\sigma_2^2}\right)^2}{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}\right) dx && (31)\\
&= \frac{1}{2\pi\sigma_1\sigma_2}\exp\left(\frac{(z-\mu_1)^2}{\sigma_1^2+\sigma_2^2}\right)\sqrt{2\frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\pi} && (32)\\
&= \frac{1}{\sqrt{2\pi}\sqrt{\sigma_1^2+\sigma_2^2}}\exp\left(\frac{(z-\mu_1)^2}{\sigma_1^2+\sigma_2^2}\right), && (33)
\end{aligned}
$$

in which the step from eq. (31) to eq. (32) used the result of eq. (96). The sum of two univariate Gaussian random variables is again a Gaussian random variable, with the mean value and variance summed up respectively, i.e. $z \sim \mathcal{N}\left(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2\right)$. The summation rule is easily generalized to n Gaussians.

4.2 Multivariate Case

Suppose $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_1)$ is a d-dimensional Gaussian density, A is a q-by-d matrix and $\boldsymbol{b}$ is a q-dimensional vector, then $\boldsymbol{z} = A\boldsymbol{x} + \boldsymbol{b}$ is a q-dimensional Gaussian density: $\boldsymbol{z} \sim \mathcal{N}\left(A\boldsymbol{\mu} + \boldsymbol{b}, A\Sigma A^T\right)$.

This fact is proved using the characteristic function tool (see Appendix B).

The characteristic function of $\mathbf{z}$ is

$$\begin{aligned}
\phi_{\boldsymbol{z}}(\boldsymbol{t}) &= \mathcal{E}_{\boldsymbol{z}}\left[\exp\left(i\boldsymbol{t}^T\boldsymbol{z}\right)\right] && (34)\\
&= \mathcal{E}_{\boldsymbol{x}}\left[\exp\left(i\boldsymbol{t}^T(A\boldsymbol{x}+\boldsymbol{b})\right)\right] && (35)\\
&= \exp\left(i\boldsymbol{t}^T\boldsymbol{b}\right)\mathcal{E}_{\boldsymbol{x}}\left[\exp\left(i\left(A^T\boldsymbol{t}\right)^T\boldsymbol{x}\right)\right] && (36)\\
&= \exp\left(i\boldsymbol{t}^T\boldsymbol{b}\right)\exp\left(i\left(A^T\boldsymbol{t}\right)^T\boldsymbol{\mu}-\left(A^T\boldsymbol{t}\right)^T\Sigma\left(A^T\boldsymbol{t}\right)\right) && (37)\\
&= \exp\left(i\boldsymbol{t}^T(A\boldsymbol{\mu}+\boldsymbol{b})-\boldsymbol{t}^T\left(A\Sigma A^T\right)\boldsymbol{t}\right), && (38)
\end{aligned}$$

in which the step from eq. (37) to eq. (38) used the eq. (107) in Appendix B. Appendix B states that if a characteristic function $\phi(\boldsymbol{t})$ is of the form $\exp\left(i\boldsymbol{t}^T\boldsymbol{\mu}-\boldsymbol{t}^T\Sigma\boldsymbol{t}\right)$, then the underlying density $p(\boldsymbol{x})$ is a Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ. Applying this fact to eq. (38), we immediately get

$$\boldsymbol{z}\sim\mathcal{N}\left(A\boldsymbol{\mu}+\boldsymbol{b},A\Sigma A^T\right). \tag{39}$$

Suppose $\boldsymbol{x}_1\sim\mathcal{N}(\boldsymbol{\mu}_1,\Sigma_1)$ and $\boldsymbol{x}_2\sim\mathcal{N}(\boldsymbol{\mu}_2,\Sigma_2)$ are two independent d-dimensional Gaussian densities, and define a new random vector $\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{y}$. We can calculate the probability density function $p(\boldsymbol{z})$ 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 Appendix C.

Characteristic function simplifies the calculation. Using eq. (110) in Appendix B, we get

$$\begin{aligned}
\phi_{\boldsymbol{z}}(\boldsymbol{t}) &= \phi_{\boldsymbol{x}}(\boldsymbol{t})\,\phi_{\boldsymbol{y}}(\boldsymbol{t}) && (40)\\
&= \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu}_1-\boldsymbol{t}^T\Sigma_1\boldsymbol{t}\right)\exp\left(i\boldsymbol{t}^T\boldsymbol{\mu}_2-\boldsymbol{t}^T\Sigma_2\boldsymbol{t}\right) && (41)\\
&= \exp\left(i\boldsymbol{t}^T(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2)-\boldsymbol{t}^T(\Sigma_1+\Sigma_2)\boldsymbol{t}\right), && (42)
\end{aligned}$$

which immediately gives us $\boldsymbol{z}\sim\mathcal{N}(\boldsymbol{\mu}_1+\boldsymbol{\mu}_2,\Sigma_1+\Sigma_2)$. The summation of two multivariate Gaussian random variables is as easy to compute as in the univariate case: sum up the mean vectors and covariance matrices. The rule is same for summing up several multivariate Gaussians.

Now we have the tool of linear transformation and let's revisit eq. (6). For convenience we retype the equation here: $\boldsymbol{x}\sim\mathcal{N}(\boldsymbol{\mu},\Sigma)$, and

$$\boldsymbol{y}=\Lambda^{-1/2}U(\boldsymbol{x}-\boldsymbol{\mu}). \tag{43}$$

Using the properties of linear transformations on a Gaussian density, $\boldsymbol{y}$ is indeed a Gaussian (in section 2.2 we painfully calculate $p(\boldsymbol{y})$ using the inverse transform method), and has mean vector $\mathbf{0}$ and covariance matrix I.

The transformation of applying eq. (6) is called the whitening transformation since the transformed density has an identity covariance matrix.

5 Geometry and Mahalanobis Distance

Figure 1 shows a bivariate Gaussian density function. Gaussian density has only one mode, which is the mean vector, and the shape of the density is determined

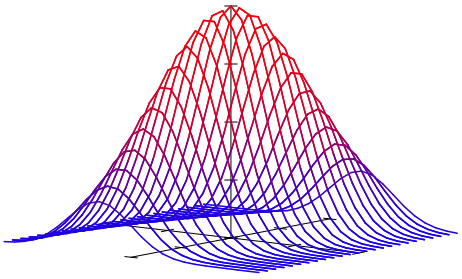

Figure 1: Bivariate Gaussian density

by the covariance matrix.

Figure 2 showed the equal probability contour of a bivariate Gaussian density. All points on a given equal probability contour must have the following term evaluated to a constant value:

$$r^2(\boldsymbol{x}, \boldsymbol{\mu}) = (\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu}) = c \tag{44}$$

$r^2(\boldsymbol{x}, \boldsymbol{\mu})$ is called the Mahalanobis distance from $\boldsymbol{x}$ to $\boldsymbol{\mu}$, given the covariance matrix Σ. Eq. (44) defines a hyperellipsoid in d-dimensional, which means that the equal probability contour of a Gaussian density is a hyperellipsoid in d-dimensional. The principle axes of this hyperellipsoid are given by the eigenvectors of Σ, and the lengths of these axes are proportional to square root of the eigenvalues associated with these eigenvectors.

6 Conditioning

Suppose $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ are two multivariate Gaussian densities, which have a joint density function

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

in which d_1 and d_2 are the dimensionality of $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ 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 $\boldsymbol{x}_2$, and satisfying $\Sigma_{12} = \Sigma_{21}^T$.

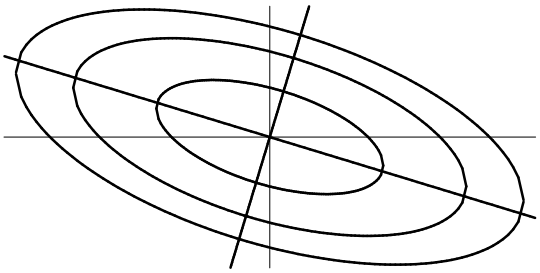

Figure 2: Equal probability contour of a bivariate Gaussian density

The marginal distributions $\boldsymbol{x}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \Sigma_{11})$ and $\boldsymbol{x}_2 \sim \mathcal{N}(\boldsymbol{\mu}_2, \Sigma_{22})$ are easy to get from the joint distribution. We are interested in computing the conditional probability density $p(x_1|x_2)$.

We will need to compute the inverse of Σ, and this task is completed by using the Schur complement (see Appendix C). 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 eq. (120) 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 $\boldsymbol{x}_2'$)

$$\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} = \begin{bmatrix} S_{22}^{-1} & -S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \\ -\Sigma_{22}^{-1}\Sigma_{12}^T S_{22}^{-1} & \Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1} \end{bmatrix} \tag{45}$$

and

$$\begin{aligned} & \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix}^T \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{x}_1 - \boldsymbol{\mu}_1 \\ \boldsymbol{x}_2 - \boldsymbol{\mu}_2 \end{bmatrix} \\ = & \boldsymbol{x}_1'^T S_{22}^{-1}\boldsymbol{x}_1' + \boldsymbol{x}_2'^T\left(\Sigma_{22}^{-1} + \Sigma_{22}^{-1}\Sigma_{12}^T S_{22}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\right)\boldsymbol{x}_2' \\ = & \left(\boldsymbol{x}_1' + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2'\right)^T S_{22}^{-1}\left(\boldsymbol{x}_1' + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2'\right) + \boldsymbol{x}_2'^T\Sigma_{22}^{-1}\boldsymbol{x}_2'. \end{aligned} \tag{46}$$

Thus we can split the joint distribution as

$$\begin{aligned} & p\left(\begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \end{bmatrix}\right) \\ = & \frac{1}{(2\pi)^{d_1}|S_{22}^{-1}|^{1/2}} \exp\left(-\frac{\left(\boldsymbol{x}_1' + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2'\right)^T S_{22}^{-1}\left(\boldsymbol{x}_1' + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2'\right)}{2}\right) \\ & \cdot\frac{1}{(2\pi)^{d_2}|\Sigma_{22}^{-1}|^{1/2}} \exp\left(\frac{1}{2}\boldsymbol{x}_2'^T\Sigma_{22}^{-1}\boldsymbol{x}_2'\right) \end{aligned} \tag{47}$$

in which we used the fact that $|\Sigma| = |\Sigma_{22}|\,|S_{22}|$ (from eq. (119) in Appendix C).

Since the second term in the right hand side of eq. (47) 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

$$p(\boldsymbol{x}_1|\boldsymbol{x}_2) = \frac{1}{(2\pi)^{d_1} |S_{22}^{-1}|^{1/2}} \exp\left(-\frac{\left(\boldsymbol{x}_1' + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2'\right)^T S_{22}^{-1} \left(\boldsymbol{x}_1' + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2'\right)}{2}\right), \tag{48}$$

or

$$\begin{aligned} \boldsymbol{x}_1|\boldsymbol{x}_2 &\sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\boldsymbol{x}_2', S_{22}^{-1}\right) && (49) \\ &\sim \mathcal{N}\left(\boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}\left(\boldsymbol{x}_2 - \boldsymbol{\mu}_2\right), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right) && (50) \end{aligned}$$

7 Product of Gaussians

Suppose $p_1(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_1, \Sigma_1)$ and $p_2(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_2, \Sigma_2)$ are two independent d-dimensional Gaussian densities. Sometimes we want to compute the density which is proportional to the product of the two Gaussian densities, i.e. $p(\boldsymbol{x}) = \alpha p_1(\boldsymbol{x})\, p_2(\boldsymbol{x})$, in which α is a proper normalization constant to make $p(\boldsymbol{z})$ a valid density function.

In this task the canonical notation (see section 3) will be extremely helpful. Writing the two Gaussians in canonical form

$$\begin{aligned} p_1(\boldsymbol{x}) &= \exp\left(\alpha_1 + \boldsymbol{\eta}_1^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\Lambda_1\boldsymbol{x}\right) && (51) \\ p_2(\boldsymbol{x}) &= \exp\left(\alpha_2 + \boldsymbol{\eta}_2^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\Lambda_2\boldsymbol{x}\right), && (52) \end{aligned}$$

the density $p(\boldsymbol{z})$ is easy to compute, as

$$\begin{aligned} p(\boldsymbol{z}) &= \alpha\alpha p_1(\boldsymbol{x})\, p_2(\boldsymbol{x}) \\ &= \exp\left(\alpha' + \left(\boldsymbol{\eta}_1 + \boldsymbol{\eta}_{12}\right)^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\left(\Lambda_1 + \Lambda_2\right)\boldsymbol{x}\right). && (53) \end{aligned}$$

This equation states that in the canonical parameterization, in order to compute product of Gaussians, we just sum the parameters.

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

$$p(\boldsymbol{x}) = \exp\left(\alpha' + \left(\Sigma_{i=1}^n\boldsymbol{\eta}_i\right)^T\boldsymbol{x} - \frac{1}{2}\boldsymbol{x}^T\left(\Sigma_{i=1}^n\Lambda_i\right)\boldsymbol{x}\right). \tag{54}$$

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

$$p(\boldsymbol{x}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) \tag{55}$$

where

$$\Sigma^{-1} = \Sigma_1^{-1} + \Sigma_2^{-1} + \ldots + \Sigma_n^{-1} \quad (56)$$

$$\Sigma^{-1}\boldsymbol{\mu} = \Sigma_1^{-1}\boldsymbol{\mu}_1 + \Sigma_2^{-1}\boldsymbol{\mu}_2 + \ldots + \Sigma_n^{-1}\boldsymbol{\mu}_n \quad (57)$$

Now we have listed some properties of the Gaussian distribution. Next let's see how these properties are applied.

8 Application I: Parameter Estimation

8.1 Maximum Likelihood Estimation

Let us suppose that we have a d-dimensional multivariate Gaussian density $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, and n i.i.d (independently, 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 data set $\mathcal{D}$ given parameters $\boldsymbol{\mu}$ and Σ is:

$$l(\boldsymbol{\mu}, \Sigma|\mathcal{D}) \quad (58)$$

$$= \log \prod_{i=1}^{n} p(\boldsymbol{x}_i) \quad (59)$$

$$= -\frac{nd}{2}\log(2\pi) + \frac{n}{2}\log\left|\Sigma^{-1}\right| - \frac{1}{2}\sum_{i-1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}). \quad (60)$$

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

$$\frac{\partial l}{\partial \boldsymbol{\mu}} = \sum_{i-1}^{n} \Sigma^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) \quad (61)$$

$$\frac{\partial l}{\partial \Sigma^{-1}} = \frac{n}{2}\Sigma - \frac{1}{2}\sum_{i-1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})^T, \quad (62)$$

in which eq. (61) used eq. (125) and the chain rule, and eq. (62) used eq. (133), eq. (134) and the fact that $\Sigma = \Sigma^T$. The notation in eq. (61) is a little bit confusing. There are two Σ 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 eq. (61) and eq. (62) to 0 gives us the solution:

$$\boldsymbol{\mu}_{ML} = \frac{1}{n}\sum_{i-1}^{n}\boldsymbol{x}_i \quad (63)$$

$$\Sigma_{ML} = \frac{1}{n}\sum_{i-1}^{n}(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})(\boldsymbol{x}_i - \boldsymbol{\mu}_{ML})^T \quad (64)$$

Eq. (63) and eq. (64) clearly states that the maximum likelihood estimation of the mean vector and covariance matrix are just the sample mean and sample covariance matrix respectively.

8.2 Bayesian Parameter Estimation

In Bayesian estimation, we assume that the covariance matrix is known. Let us suppose that we have a d-dimensional multivariate Gaussian density $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, 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's assume that the prior is that $\boldsymbol{\mu} \sim \mathcal{N}(\boldsymbol{\mu}_0, \Sigma_0)$. The task is 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 find a point $\hat{\boldsymbol{\mu}}$ in the parameter space that gives maximum likelihood, we calculate the posterior density for the parameter $p(\boldsymbol{\mu}|\mathcal{D})$, and use the entire distribution of $\boldsymbol{\mu}$ as our estimation for this parameter.

Applying the Bayes' law,

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D}) &= \alpha p(\mathcal{D}|\boldsymbol{\mu})\, p(\boldsymbol{\mu}) && (65) \\ &= \alpha \prod_{i=1}^{n} p(\boldsymbol{x}_i)\, p_0(\boldsymbol{\mu}) && (66) \end{aligned}$$

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

Apply the result in section 7, we know that $p(\boldsymbol{\mu}|\mathcal{D})$ is also a Gaussian, and

$$p(\boldsymbol{\mu}|\mathcal{D}) = \mathcal{N}(\boldsymbol{\mu}; \boldsymbol{\mu}_n, \Sigma_n) \tag{67}$$

where

$$\begin{aligned} \Sigma_n^{-1} &= n\Sigma^{-1} + \Sigma_0^{-1} && (68) \\ \Sigma_n^{-1}\boldsymbol{\mu}_n &= n\Sigma^{-1}\boldsymbol{\mu} + \Sigma_0^{-1}\boldsymbol{\mu}_0 && (69) \end{aligned}$$

Both $\boldsymbol{\mu}_n$ and Σ_n can be calculated from know parameters. So we have determined the posterior distribution $p(\boldsymbol{\mu}|\mathcal{D})$ for $\boldsymbol{\mu}$ given the data set $\mathcal{D}$.

We choose a Gaussian 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 seen that Gaussian have the nice property that both the prior and posterior are Gaussian, i.e. Gaussian is auto-conjugate.

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

$$p(\boldsymbol{x}|\mathcal{D}) = \int_{\boldsymbol{\mu}} p(\boldsymbol{x}|\boldsymbol{\mu})\, p(\boldsymbol{\mu}|\mathcal{D})\, d\boldsymbol{\mu}. \tag{70}$$

Eq. (70) and eq. (27) has the same form. Thus we can guess that $p(\boldsymbol{x}|\mathcal{D})$ is a Gaussian again, and

$$p(\boldsymbol{x}|\mathcal{D}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_n, \Sigma + \Sigma_n). \tag{71}$$

The guess is correct, and is easy to verify it by repeating the steps in eq.(27) through eq. (33).

9 Application II: Kalman Filter

9.1 The Model

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

$$\boldsymbol{x}_k = A\boldsymbol{x}_{k-1} + \boldsymbol{w}_{k-1}, \tag{72}$$

and a linear measurement model

$$\boldsymbol{z}_k = H\boldsymbol{x}_k + \boldsymbol{v}_k. \tag{73}$$

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

$$\boldsymbol{w} \sim \mathcal{N}(\mathbf{0}, Q) \tag{74}$$

$$\boldsymbol{v} \sim \mathcal{N}(\mathbf{0}, R) \tag{75}$$

At time $k-1$, assume 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 $p(\boldsymbol{x}_{k-1})$.

In a broader point of view, 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 assumption, it is not hard to prove that the two problem formulations are equivalent.

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

9.2 The Estimation

I will show that, with the help of the properties of Gaussians we have obtained, it is quite easy to derive the Kalman filter equations.

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 (72), we get an estimate $p(\boldsymbol{x}_k^-)$. Note that the minus sign means that this estimation is done before we take into account the measurement. In the second step, based on $p(\boldsymbol{x}_k^-)$ and the measurement model (73), we get the final estimation $p(\boldsymbol{x}_k)$.

First, let's estimate $p(\boldsymbol{x}_k^-)$. Assume that at time $k-1$, the estimation we already got is a Gaussian

$$p(\boldsymbol{x}_{k-1}) = \mathcal{N}(\boldsymbol{x}; \boldsymbol{\mu}_{k-1}, P_{k-1}). \tag{76}$$

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

Applying the linear operation equation (39) on the dynamic model (72), we immediately get the estimation for $\boldsymbol{x}_k^-$:

$$\begin{aligned} \boldsymbol{x}_k^- &\sim \mathcal{N}\left(\boldsymbol{\mu}_k^-, P_k^-\right) && (77)\\ \boldsymbol{\mu}_k^- &= \boldsymbol{A}\boldsymbol{\mu}_{k-1} && (78)\\ P_k^- &= AP_{k-1}A^T + Q && (79) \end{aligned}$$

The estimate $p\left(\boldsymbol{x}_k^-\right)$ conditioned on the observation $\boldsymbol{z}_k$ gives $p(\boldsymbol{x}_k)$, the estimation we want. Thus the conditioning property (50) can be used. In order to use eq. (50), we have to build the joint covariance matrix first. Since $Cov(\boldsymbol{z}_k) = HP_k^-H^T + R$ (applying eq. (39) to eq. (73)) and

$$\begin{aligned} Cov\left(\boldsymbol{z}_k, \boldsymbol{x}_k^-\right) &= Cov\left(H\boldsymbol{x}_k^- + \boldsymbol{v}_k, \boldsymbol{x}_k^-\right) && (80)\\ &= Cov\left(H\boldsymbol{x}_k^-, \boldsymbol{x}_k^-\right) && (81)\\ &= HP_k^-, && (82) \end{aligned}$$

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

$$\begin{bmatrix} P_k^- & P_k^-H^T \\ HP_k^- & HP_k^-H^T + R \end{bmatrix}. \quad (83)$$

Applying the conditioning equation (50), we get

$$\begin{aligned} p(\boldsymbol{x}_k) &= p\left(\boldsymbol{x}_k^-|\boldsymbol{z}_k\right) && (84)\\ &\sim \mathcal{N}(\boldsymbol{\mu}_k, P_k) && (85)\\ P_k &= P_k^- - P_k^-H^T\left(HP_k^-H^T + R\right)^{-1}HP_k^- && (86)\\ \boldsymbol{\mu}_k &= \boldsymbol{\mu}_k^- + P_k^-H^T\left(HP_k^-H^T + R\right)^{-1}\left(\boldsymbol{z}_k - H\boldsymbol{\mu}_k^-\right) && (87) \end{aligned}$$

The equations (77∼79) and (84∼87) are the Kalman filter updating rules.

The term $P_k^-H^T\left(HP_k^-H^T + R\right)^{-1}$ appears in both eq. (86) and eq. (87). Defining

$$K_k = P_k^-H^T\left(HP_k^-H^T + R\right)^{-1}, \quad (88)$$

the equations are simplified as

$$\begin{aligned} P_k &= (I - K_kH)P_k^- && (89)\\ \boldsymbol{\mu}_k &= \boldsymbol{\mu}_k^- + K_k\left(\boldsymbol{z}_k - H\boldsymbol{\mu}_k^-\right). && (90) \end{aligned}$$

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

A Gaussian Integral

We will compute the integral of the univariate Gaussian density in this section. The trick in doing this integration is to consider two independent univariate gaussians at one time.

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-x^2} dx &= \sqrt{\left(\int_{-\infty}^{\infty} e^{-x^2} dx\right)\left(\int_{-\infty}^{\infty} e^{-y^2} dy\right)} && (91)\\
&= \sqrt{\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} e^{-(x^2+y^2)} dxdy} && (92)\\
&= \sqrt{\int_{0}^{\infty}\int_{0}^{2\pi} re^{-r^2} drd\theta} && (93)\\
&= \sqrt{2\pi[-\frac{1}{2}e^{-r^2}]_0^{\infty}} && (94)\\
&= \sqrt{\pi}, && (95)
\end{aligned}$$

in eq. (93) the polar coordinates are used, and the extra r appeared inside the integral is the determinant of the Jacobian matrix.

The above integral can be easily extend as

$$f(t) = \int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{t}\right) dx = \sqrt{t\pi} \tag{96}$$

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

$$\begin{aligned}
\frac{df}{dt} &= \frac{d}{dt}\int_{-\infty}^{\infty} \exp\left(-\frac{x^2}{t}\right) dx && (97)\\
&= \int_{-\infty}^{\infty} \frac{x^2}{t^2} \exp\left(-\frac{x^2}{t}\right) dx && (98)
\end{aligned}$$

and

$$\frac{df}{dt} = \frac{d}{dt}\sqrt{t\pi} = \frac{1}{2}\sqrt{\frac{\pi}{t}}. \tag{99}$$

The above two equations give us

$$\int_{-\infty}^{\infty} x^2 \exp\left(-\frac{x^2}{t}\right) dx = \frac{t^2}{2}\sqrt{\frac{\pi}{t}}. \tag{100}$$

Applying eq. (100), we have

$$\int_{-\infty}^{\infty} x^2 \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = \frac{1}{\sqrt{2\pi}}\frac{4}{2}\sqrt{\frac{\pi}{2}} = 1 \tag{101}$$

It is obvious that

$$\int_{-\infty}^{\infty} x \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx = 0 \tag{102}$$

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

Eq. (102) and eq. (101) have proved that the mean and standard deviation of a standard normal distribution are 0 and 1 respectively.

B Characteristic Functions

The characteristic function of a probability density $p(\boldsymbol{x})$ is defined as its Fourier transform

$$\phi(\boldsymbol{t}) = \mathcal{E}\left[\exp\left(i\boldsymbol{t}^T\boldsymbol{x}\right)\right] \tag{103}$$

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

Let's computer the characteristic function of a Gaussian density.

$$\begin{aligned}
& \phi(\boldsymbol{t}) \\
= \; & \mathcal{E}\left[\exp\left(i\boldsymbol{t}^T\boldsymbol{x}\right)\right] && (104)\\
= \; & \int \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) + i\boldsymbol{t}^T\boldsymbol{x}\right) d\boldsymbol{x} && (105)\\
= \; & \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \boldsymbol{t}^T\Sigma\boldsymbol{t}\right) \\
& \cdot \int \frac{\exp\left(-\frac{1}{2}\left(\boldsymbol{x} - \left(\Sigma^{-1}\boldsymbol{\mu} - i\boldsymbol{t}\right)\right)^T\Sigma^{-1}\left(\boldsymbol{x} - \left(\Sigma^{-1}\boldsymbol{\mu} - i\boldsymbol{t}\right)\right)\right)}{(2\pi)^{d/2}|\Sigma|^{1/2}} d\boldsymbol{x} && (106)\\
= \; & \exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \boldsymbol{t}^T\Sigma\boldsymbol{t}\right) && (107)
\end{aligned}$$

Since the characteristic function is defined as a Fourier transform, the inverse Fourier transform of $\phi(\boldsymbol{t})$ will be exactly $p(\boldsymbol{x})$, i.e. a density is completely determined by its characteristic function. When we see a characteristic function $\phi(\boldsymbol{t})$ is of the form $\exp\left(i\boldsymbol{t}^T\boldsymbol{\mu} - \boldsymbol{t}^T\Sigma\boldsymbol{t}\right)$, we know that the underlying density is a Gaussian with mean vector $\boldsymbol{\mu}$ and covariance matrix Σ.

Suppose $\boldsymbol{x}$ and $\boldsymbol{y}$ are two independent random vectors with the same dimension, and define a new random vector $\boldsymbol{z} = \boldsymbol{x} + \boldsymbol{y}$. Then

$$\begin{aligned}
p_{\boldsymbol{z}}(\boldsymbol{z}) &= \int\int_{\boldsymbol{z}=\boldsymbol{x}+\boldsymbol{y}} p_{\boldsymbol{x}}(\boldsymbol{x}) p_{\boldsymbol{y}}(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} && (108)\\
&= \int_{\boldsymbol{x}} p_{\boldsymbol{x}}(\boldsymbol{x}) p_{\boldsymbol{y}}(\boldsymbol{z}-\boldsymbol{x}) d\boldsymbol{x}. && (109)
\end{aligned}$$

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

$$\phi_{\boldsymbol{z}}(\boldsymbol{z}) = \phi_{\boldsymbol{x}}(\boldsymbol{x}) \phi_{\boldsymbol{y}}(\boldsymbol{y}), \tag{110}$$

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

C Schur Complement and the 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

$$M = \begin{bmatrix} A & B \\ C & D \end{bmatrix}, \tag{111}$$

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

Some algebraic manipulations give

$$\begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix} M \begin{bmatrix} I & -A^{-1}B \\ \mathbf{0} & I \end{bmatrix} \tag{112}$$

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

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

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

in which the term $D - CA^{-1}B$ is called the Schur complement of A, denoted as S_A.

Equation $XMY = Z$ implies that $M^{-1} = YZ^{-1}X$. Hence we have

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

$$= \begin{bmatrix} A^{-1} & -A^{-1}BS_A^{-1} \\ 0 & S_A^{-1} \end{bmatrix} \begin{bmatrix} I & \mathbf{0} \\ -CA^{-1} & I \end{bmatrix} \tag{117}$$

$$= \begin{bmatrix} A^{-1} + A^{-1}BS_A^{-1}CA^{-1} & -A^{-1}BS_A^{-1} \\ -S_A^{-1}CA^{-1} & S_A^{-1} \end{bmatrix} \tag{118}$$

Taking the determinant of both sides of eq. (116), it gives

$$|M| = |A|\,|S_A|\,. \tag{119}$$

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

$$M^{-1} = \begin{bmatrix} S_D^{-1} & -S_D^{-1}BD^{-1} \\ -D^{-1}CS_D^{-1} & D^{-1} + D^{-1}CS_D^{-1}BD^{-1} \end{bmatrix} \tag{120}$$

$$|M| = |D|\,|S_D|\,. \tag{121}$$

Eq. (118) and eq. (120) are two different representations of the same matrix M^{-1}, which means the corresponding blocks in these two equations must be

equal, e.g. $S_D^{-1} = A^{-1} + A^{-1}BS_A^{-1}CA^{-1}$. This result is known as the matrix inversion lemma:

$$S_D^{-1} = \left(A - BD^{-1}C\right)^{-1} = A^{-1} + A^{-1}B\left(D - CA^{-1}B\right)^{-1}CA^{-1}. \tag{122}$$

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

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

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

D 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:

$$\left(\frac{\partial y}{\partial A}\right)_{ij} = \frac{\partial y}{\partial a_{ij}} \tag{124}$$

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

From the definition (124), it is easy to get the following rule.

$$\frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{x}^T\boldsymbol{y}\right) = \frac{\partial}{\partial \boldsymbol{x}}\left(\boldsymbol{y}^T\boldsymbol{x}\right) = \boldsymbol{y} \tag{125}$$

For a square matrix A that is n-by-n, the matrix M_{ij} defined by removing from A the i-th row and j-th column is called a minor of A. 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 transpose of the cofactor matrix

$$A_{adj} = A_{cof}^T. \tag{126}$$

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

$$|A| = \sum_j a_{ij}c_{ij} \tag{127}$$

$$A^{-1} = \frac{1}{|A|}A_{adj}. \tag{128}$$

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

$$\frac{\partial}{\partial a_{ij}}|A| = c_{ij}, \tag{129}$$

$$\text{or, } \frac{\partial}{\partial A}|A| = A_{cof} \tag{130}$$

which in turn shows that

$$= A_{adj}^T \tag{131}$$

$$= |A| \left(A^{-1}\right)^T. \tag{132}$$

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

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

Applying the definition (124), it is easy to show that for a square matrix A,

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

since $\boldsymbol{x}^T A \boldsymbol{x} = \Sigma_i \Sigma_j a_{ij} x_i x_j$ where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$.

References

[1] C. M. Bishop. Neural Networks for Pattern Recognition, Oxford University Press, Oxford, UK, 1996.

[2] R. O. Duda, P. E. Hart, and D. G. Stork. Pattern Classification, second edition, Wiley-Interscience, New York, NY, USA, 2001.

[3] http://mathworld.wolfram.com/

[4] M. I. Jordan. An Introduction to Probabilistic Graphical Models, chapter 13, draft.

[5] G. Welch and G. Bishop. An Introduction to the Kalman Filter, TR 95-041, Department of Computer Science, University of North Carolina at Chapel Hill.