Gaussian Distribution

Chunpai Wang

The Gaussian distribution is widely used model for the distribution of continuous variables, and it arises in many different context and can be motivated from a variety of different perspective. For example,

- for a single random variable, the distribution that maximizes the entropy is the Gaussian. This property applies also to the multivariate Gaussian.
- when we consider the sum of multiple random variables, which is of course itself a random variable, has a distribution that becomes increasingly Gaussian as the number of terms in the sum increase, which is also known as *central limit theorem*. N random variables $x_1, ..., x_N$ follows the $Uniform(0, 1)$ distribution,

  $$x_1, x_2, ..., x_N \sim Uniform(0, 1)$$

  then by central limit theorem for large N

  $$\frac{(x_1 + x_2 + ... + x_N)}{N} \sim Gaussian(\mu, \sigma)$$

1 Probability Density Functions

For a single random variable x,

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} \tag{1}$$

where μ is mean, and σ^2 is variance. For a D-dimensional vector $\mathbf{x}$, the multivariate Gaussian distribution takes the form

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\} \tag{2}$$

where $\boldsymbol{\mu}$ is a D-dimensional mean vector, $\boldsymbol{\Sigma}$ is a $D \times D$ covariance matrix and $\boldsymbol{\Sigma} \in S^D_{++}$, and $|\boldsymbol{\Sigma}|$ denotes the determinant of $\boldsymbol{\Sigma}$. $\boldsymbol{\Sigma}$ must be positive definite, which implies the determinant $|\boldsymbol{\Sigma}| > 0$and normalization constant is positive, and ensures the density function integral to 1.

2 Covariance Matrix and Ellipsoids

If entries of a column vector X are random variables, each with finite variance, then the covariance matrix is the matrix whose (i, j) entry is the covariance of random variable X_i and X_j

$$\boldsymbol{\Sigma}_{ij} = cov(X_i, X_j) = E[(X_i - \mu_i)(X_j - \mu_j)] = E(X_i X_j) - \mu_i \mu_j$$

and thus we can generalize it as

$$\boldsymbol{\Sigma} = E[(X - E(X)(X - E(X))^\top)]$$

Here are some properties of covariance matrix:

- Every covariance matrix is symmetric and positive semidefinite
- $\boldsymbol{\Sigma} = E(\boldsymbol{X}\boldsymbol{X}^\top) - \boldsymbol{\mu}\boldsymbol{\mu}$
- $cov(AX + a) = Acov(X)A^\top$
- $cov(X, Y) = cov(Y, X)^\top$
- $cov(X_1 + X_2, Y) = cov(X_1, Y) + cov(X_2, Y)$

Canonical form of ellipsoid is

$$\frac{x_1^2}{\sigma_1^2} + \frac{x_2^2}{\sigma_2^2} + \cdots + \frac{x_n^2}{\sigma_n^2} = 1$$

and we can represent it with a matrix form as

$$\boldsymbol{x}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{x} = 1$$

where $\boldsymbol{\Sigma}^{-1}$ is a diagonal matrix with entries are $\frac{1}{\sigma_i^2}$. We can see that σ_i cannot be zero, and that is the reason why Σ must be positive definite.

Shifted Ellipsoids have the form

$$\mathcal{E} = \{x | (x - x_c)^\top \Sigma^{-1} (x - x_c) \leq 1\}$$

where $\Sigma = \Sigma^\top \succ 0$, i.e., Σ is symmetric and positive definite, and x_c is the center of the ellipsoid. We also find that Σ can be decomposed as

$$\Sigma = U \Lambda U^\top = U \Lambda^{1/2} \Lambda^{1/2} U^\top = A A^\top$$

where $A = U\Lambda^{1/2}$. Now, if we know that a random variable $X \sim Gaussian(0, I)$, then $AX + \mu \sim Gaussian(\mu, \Sigma)$. Following are some geometric interpretation of affine transformation on Gaussian random variable:

3 Geometric Form

First, let's denote by

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

where Δ is known as *Mahalanobis Distance* from $\boldsymbol{\mu}$ to $\mathbf{x}$. Note that, when $\boldsymbol{\Sigma}$ is the identity matrix, Mahalanobis Distance reduces to Euclidean distance, which quadratic form becomes a **constant**.

First of all, we note that the matrix $\boldsymbol{\Sigma}$ can be taken to be *symmetric*, without loss of generality, because any antisymmetric component would disappear from the exponent[5]. Now consider eigen pair of covariance matrix,

$$\boldsymbol{\Sigma} \boldsymbol{v}_i = \lambda_i \boldsymbol{v}_i \tag{2}$$

where $i = 1, ..., D$.

- $\boldsymbol{\Sigma}$ is real and symmetric, therefore its eigenvalues are real
- its eigenvectors can be chosen to form an orthonormal set, so that

  $$\boldsymbol{v}_i^\top \boldsymbol{v}_j = I_{ij} \tag{3}$$

  where

  $$I_{ij} = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

- $\boldsymbol{V} = [\boldsymbol{v}_1, ..., \boldsymbol{v}_D]^\top$ is called orthogonal, it's square and $\boldsymbol{V}^\top \boldsymbol{V} = \boldsymbol{I}$, and $\boldsymbol{V}\boldsymbol{V}^{-1} = \boldsymbol{I}$
- The covariance matrix can be expressed as

  $$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top \tag{5}$$

  and

  $$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top \tag{6}$$

- We can rewrite

  $$\begin{aligned} \Delta^2 &= (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= (\mathbf{x} - \boldsymbol{\mu})^\top \left(\sum_{i=1}^{D} \frac{1}{\lambda_i} \boldsymbol{v}_i \boldsymbol{v}_i^\top \right) (\mathbf{x} - \boldsymbol{\mu}) \\ &= \sum_{i=1}^{D} \frac{(\boldsymbol{v}_i (\mathbf{x} - \boldsymbol{\mu}))^2}{\lambda_i} \end{aligned}$$

  where we can denote $\boldsymbol{v}_i(\mathbf{x} - \boldsymbol{\mu})$ as y_i

- We can interpret $\{y_i\}$ as a new coordinate system defined by the orthonormal eigenvectors $\boldsymbol{v}_i$ that are shifted and rotated with respect to the original $\{x_i\}$ coordinates. Let $\mathbf{y} = (y_1, ..., y_D)^\top$, we have

$$\mathbf{y} = \boldsymbol{V}(\mathbf{x} - \boldsymbol{\mu}) \tag{7}$$

  where $\boldsymbol{V}$ is a matrix whose rows are given by $\boldsymbol{v}_i^\top$.

- The quadratic form Δ^2, and hence the Gaussian distribution, will be constant on surfaces for which y_i is constant.(???) If all the eigenvalues are positive, then these surfaces represent ellipsoid, with their centers at $\boldsymbol{\mu}$ and their axes oriented along $\boldsymbol{\mu}_i$, and with *scaling factor* in the directions of axes given by $\lambda_i^{1/2}$.

The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.

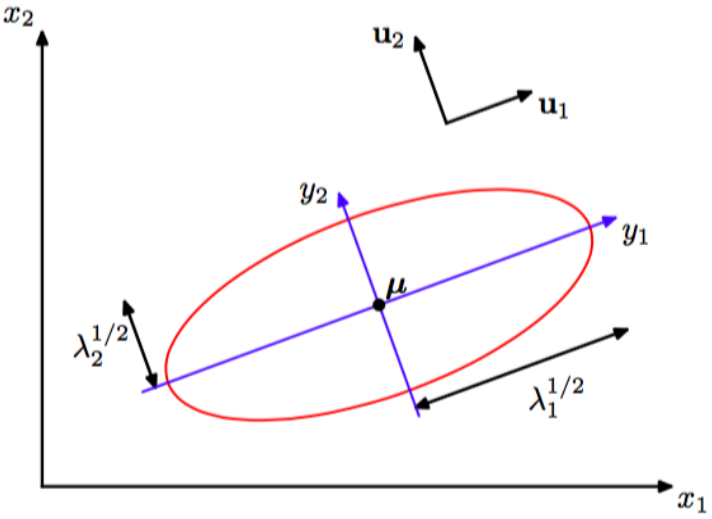

Figure 1: The Gaussian Distribution

3.1 P.D.F of Gaussian Distribution in The Eigen-Coordinates

In order to change the probability density function based on different coordinates, we use

$$p(\mathbf{y}) = p(\mathbf{x})|\boldsymbol{J}| \tag{8}$$

where $\boldsymbol{J}$ is the Jacobian matrix with elements given by

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = V_{ji} \tag{9}$$

where V_{ji} are the elements of the matrix $\boldsymbol{V}^\top$. Based on orthonormality, we have the square of *determinant* of Jacobian matrix is

$$|\boldsymbol{J}|^2 = |\boldsymbol{V}^\top|^2 = |\boldsymbol{V}^\top||\boldsymbol{V}| = |\boldsymbol{V}^\top\boldsymbol{V}| = |\boldsymbol{I}| = 1 \tag{10}$$

Also, the determinant of covariance matrix can be written as a product of eigenvalues, hence

$$|\boldsymbol{\Sigma}|^{1/2} = \prod_{i=1}^{D} \lambda_i^{1/2} \tag{11}$$

Therefore, we can write the distribution based on y_i coordinate as

$$p(\mathbf{y}) = p(\mathbf{x})|\boldsymbol{J}| = \prod_{i=1}^{D} \frac{1}{(2\pi\lambda_i)^{1/2}} \exp\left\{-\frac{y_i^2}{2\lambda_i}\right\} \tag{12}$$

which is product of D independent univariate Gaussian distribution. *The eigenvectors therefore define a new set of shifted and rotated coordinates with respect to which the joint probability distribution factorizes into a product of independent distributions.*

4 Parametrization

In addition, the parameters in the multivariate Gaussian distribution are $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, which can be interpreted as $E(\mathbf{x})$ and $Cov(\mathbf{x})$, which can be derived by the moment generating function and the canonical parametrization method.

4.1 The Moment Parametrization

The expectation of $\mathbf{x}$ under the Gaussian distribution is given by

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})d\mathbf{z}\end{aligned}$$

Now we consider second order moments of the Gaussian. In the univariate case, we considered the second order moment given by $\mathbb{E}[x^2]$. For the multivariate Gaussian, there are D^2 second order moments given by $\mathbb{E}[x_i x_j]$, which we can group together to form the matrix $\mathbb{E}[\mathbf{x}\mathbf{x}^\top]$,

$$\begin{aligned}\mathbb{E}[\mathbf{x}\mathbf{x}^\top] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\mathbf{x}^\top d\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^\top\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})(\mathbf{z}+\boldsymbol{\mu})^\top d\mathbf{z}\end{aligned}$$

???

Then, we have the important result

$$\boldsymbol{\mu} = \mathbb{E}[\mathbf{x}] \tag{1}$$
$$\boldsymbol{\Sigma} = \mathbb{E}(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^\top \tag{2}$$

4.2 The Canonical Parametrization

Expanding the quadratic form in Eq.(2), defining *canonical parameters*

$$\boldsymbol{\Lambda} = \boldsymbol{\Sigma} \tag{3}$$
$$\boldsymbol{\eta} = \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} \tag{4}$$

and putting the term before exponent into exponent, we obtain

$$p(\mathbf{x}|\boldsymbol{\eta},\boldsymbol{\Lambda}) = \exp\left\{a + \boldsymbol{\eta}^\top\mathbf{x} - \frac{1}{2}\mathbf{x}^\top\boldsymbol{\Lambda}\mathbf{x}\right\} \tag{5}$$

where

$$a = -\frac{1}{2}(n\log(2\pi) - \log|\boldsymbol{\Lambda}| + \boldsymbol{\eta}^\top\boldsymbol{\Lambda}\boldsymbol{\eta}) \tag{6}$$

is the normalizing constant in this representation. The canonical parametrization is also sometimes referred to as the *information parametrization*.

In addition, we can also convert from canonical parameters to moment parameters:

$$\boldsymbol{\mu} = \boldsymbol{\Lambda}^{-1}\boldsymbol{\eta} \tag{7}$$
$$\boldsymbol{\Sigma} = \boldsymbol{\Lambda}^{-1} \tag{8}$$

5 Joint Distributions

Suppose that we partition the $n\times 1$ vector $\mathbf{x}$ into a $p\times 1$ subvector $\mathbf{x}_1$ and a $q\times 1$ subvector $\mathbf{x}_2$, where $n = p+q$. Form corresponding partitions of the $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ parameters:

$$\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix} \qquad\qquad \boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix} \tag{1}$$

We can write a *joint Gaussian distribution* for $\mathbf{x}_1$ and $\mathbf{x}_2$ using these partitioned parameters:

$$p(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = p(\mathbf{x}_1,\mathbf{x}_2|\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{(2\pi)^{(p+q)/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix}\right\} \tag{2}$$

6 Partitioned Matrices

First, we will show how to block diagonalize a partitioned matrix. Consider a general partitioned matrix:

$$M = \begin{bmatrix} E & F \\ G & H \end{bmatrix} \tag{1}$$

where assume that both E and H are invertible. To *invert* this matrix, we follow a similar procedure to that of diagonalization. In particular, we wish to block diagonalize the matrix. We wish to push a block of zeros in place of G and a block of zeros in place of F.

To zero out the upper-right-hand corner of M, note that is suffices to *premultiply* the second block column of M by a "block row vector" having elements I and $-FH^{-1}$. Similarly, to zero out the lower-left-hand corner of M, it suffices to *postmultiply* the second row of M by a "block column vector" having elements I and $-H^{-1}G$. The magical fact is that these two operations do not interfere with each other; thus we can block diagonalize M by doing both operations. In particular, we have

$$\begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \begin{bmatrix} E & F \\ G & H \end{bmatrix} \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} = \begin{bmatrix} E - FH^{-1}G & 0 \\ 0 & H \end{bmatrix} \tag{2}$$

Now we denoted the term $E - FH^{-1}G$ as M/H. Note that, if $XYZ = W$, then inverting both side yields $Y^{-1} = ZW^{-1}X$, therefore,

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} I & 0 \\ -H^{-1}G & I \end{bmatrix} \begin{bmatrix} (M/H)^{-1} & 0 \\ 0 & H^{-1} \end{bmatrix} \begin{bmatrix} I & -FH^{-1} \\ 0 & I \end{bmatrix} \tag{3}$$

$$= \begin{bmatrix} (M/H)^{-1} & -(M/H)^{-1}FH^{-1} \\ -H^{-1}G(M/H)^{-1} & H^{-1} + H^{-1}G(M/H)^{-1}FH^{-1} \end{bmatrix} \tag{4}$$

and the determinant of M is

$$1 * |M| * 1 = |M/H||H| \tag{5}$$

$$|M| = |M/H||H| \tag{6}$$

6.1 Schur Complement

Based on this result, we can find that the term $M/H = E - FH^{-1}G$ is called the *Schur Complement* of the matrix M with respect to H, and M/H is invertible. Based on *Schur Complement*, we also can decompose the matrix M in terms of E and M/E, yielding the following expression for the inverse:

$$\begin{bmatrix} E & F \\ G & H \end{bmatrix}^{-1} = \begin{bmatrix} E^{-1} + E^{-1}F(M/E)^{-1}GE^{-1} & -E^{-1}F(M/E)^{-1} \\ -(M/E)^{-1}GE^{-1} & (M/E)^{-1} \end{bmatrix} \tag{7}$$

These two expressions for the inverse of M must be the same, thus we can set the corresponding blocks equals to each other. This yields:

$$(E - FH^{-1}G)^{-1} = E^{-1} + E^{-1}F(H - GE^{-1}F)^{-1}GE^{-1} \tag{8}$$

and

$$-(M/H)^{-1}FH^{-1} = -E^{-1}F(M/E)^{-1} \tag{9}$$

$$(E - FH^{-1}G)^{-1}FH^{-1} = E^{-1}F(H - GE^{-1}F)^{-1} \tag{10}$$

where both equations are quite useful in transformations involving Gaussian distribution. They allow expressions involving the inverse of E to be converted into expression involving the inverse of H and vice versa.

7 Marginalization and Conditioning

An important property of the multivariate Gaussian distribution is that if two sets of variables are jointly Gaussian, then the conditional distribution of one set conditioned on the other is again Gaussian. Similarly, the marginal distribution of either set is also Gaussian. We will derive the marginal Gaussian distribution and conditional Gaussian distribution and corresponding parameters in this section. Our goal is to split the joint distribution into a conditional probability for $\mathbf{x}_1$ and a marginal probability for $\mathbf{x}_2$ according to the factorization:

$$p(\mathbf{x}_1, \mathbf{x}_2) = p(\mathbf{x}_1|\mathbf{x}_2)p(\mathbf{x}_2) \tag{1}$$

We make use of the partition method in previous section to expand the exponential factor in joint Gaussian distribution,

$$-\frac{1}{2}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix}^\top\begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12}\\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}^{-1}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix} \tag{2}$$

$$=-\frac{1}{2}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix}^\top\begin{bmatrix}I & 0\\ -\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21} & I\end{bmatrix}\begin{bmatrix}(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1} & 0\\ 0 & \boldsymbol{\Sigma}_{22}^{-1}\end{bmatrix}\begin{bmatrix}I & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\\ 0 & I\end{bmatrix}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix} \tag{3}$$

$$=-\frac{1}{2}\begin{bmatrix}(\mathbf{x}_1-\boldsymbol{\mu}_1)^\top+(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top(-\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}) & \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix}\begin{bmatrix}(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1} & 0\\ 0 & \boldsymbol{\Sigma}_{22}^{-1}\end{bmatrix}\begin{bmatrix}I & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\\ 0 & I\end{bmatrix}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix} \tag{4}$$

$$=-\frac{1}{2}\begin{bmatrix}(((\mathbf{x}_1-\boldsymbol{\mu}_1)^\top+(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top(-\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}))^\top(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1} & (\mathbf{x}_2-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_{22}^{-1}\end{bmatrix}\begin{bmatrix}I & -\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\\ 0 & I\end{bmatrix}\begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1\\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix} \tag{5}$$

$$=-\frac{1}{2}\left\{(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))^\top(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))+(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right\} \tag{6}$$

Therefore, the exponential term is equal to

$$\exp\left\{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))^\top(\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22})^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1-\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2))\right\}$$
$$\times\exp\left\{-\frac{1}{2}(\mathbf{x}_2-\boldsymbol{\mu}_2)^\top\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2-\boldsymbol{\mu}_2)\right\}$$

The normalization term also can be split into two factors:

$$\frac{1}{(2\pi)^{(p+q)/2}|\boldsymbol{\Sigma}|^{1/2}}=\frac{1}{(2\pi)^{(p+q)/2}(|\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22}||\boldsymbol{\Sigma}_{22}|)^{1/2}} \tag{7}$$

$$=\left\{\frac{1}{(2\pi)^{(p+q)/2}|\boldsymbol{\Sigma}/\boldsymbol{\Sigma}_{22}|^{1/2}}\right\}\left\{\frac{1}{(2\pi)^{(p+q)/2}|\boldsymbol{\Sigma}_{22}|^{1/2}}\right\} \tag{8}$$

8 Maximum Likelihood Estimation

References

[1] (Chaprter 2) Bishop, Christopher M. Pattern Recognition and Machine Learning. springer, 2006.

[2] (Chapter 13) An Introduction to Probabilistic Graphical Models, by M. Jordan. https://people.eecs.berkeley.edu/~jordan/courses/260-spring10/other-readings/chapter13.pdf

[3] Do, Chuong B. "The multivariate gaussian distribution." Section Notes, Lecture on Machine Learning, CS 229 (2008). http://cs229.stanford.edu/section/gaussians.pdf

[4] Do, Chuong B. "More on multivariate gaussians." http://cs229.stanford.edu/section/more_on_gaussians.pdf

[5] https://math.stackexchange.com/questions/2262171/we-note-that-the-matrix-\begingroup\let\relax\relax\endgroup[Pleaseinsert\PrerenderUnicode{Îč}intopreamble]-can-be-taken-to-be-symmetric-without-loss-of-general