Multivariate Normal Distribution

1 Definition of multivariate normal

Recall that a random variable is completely described by its probability density function (PDF).
$X = (X_1, X_2, \ldots, X_n) \in \mathbb{R}^n$ is a multivariate normal random variable if its PDF is

$$\rho_X(x; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}(\det \Sigma)^{1/2}} \exp\left(\frac{-1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

where $x = (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$ is the independent variable (vector) of the PDF, $\mu = (\mu_j) \in \mathbb{R}^n$ is the mean vector, and $\Sigma = (\sigma_{ij}) \in \mathbb{S}_{++}^n$ is the covariance matrix. Here $\mathbb{S}_{++}^n$ represents the set of all real symmetric positive definite matrices. We need to justify several items.

- We need to connect it to the 1D normal distribution we are familiar with.
- We need to justify the name of density: $\displaystyle\int_{\mathbb{R}^n} \rho_X(x; \mu, \Sigma)dx = 1$.
- We need to justify the names of mean vector and covariance matrix.

$$E(X_j) = \mu_j, \qquad E\Big((X_i - \mu_i)(X_j - \mu_j)\Big) = \sigma_{ij}$$

2 Connection to 1D independent Gaussians

Review of linear algebra

A real square matrix U is called orthogonal if $U^T U = UU^T = I$.
For an orthogonal matrix $U \in O(n)$, we have $U^{-1} = U^T$ and $(U^T)^{-1} = U$.
Any real symmetric matrix A is orthogonally diagonalizable. That is, for $A \in \mathbb{S}^n$, there exists an orthogonal matrix $U \in O(n)$ such that

$$A = U \Lambda U^T, \qquad \Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n)$$

Meaning of PDF

Recall the connection between PDF and probability in an infinitesimal region.

$$\Pr(X \in \delta V) = \int_{\delta V} \rho_X(x)dx \approx \text{Vol}(\delta V)\rho_X(x)$$

Here $\text{Vol}(\delta V)$ is the volume of δV.

PDF of a transformed X

Since Σ is symmetric and positive definite, we write Σ and Σ^{-1} as

$$\begin{aligned}\Sigma &= U\Lambda U^T, \qquad \Lambda = \mathrm{diag}(d_1^2, d_2^2, \ldots, d_n^2)\\ \Sigma^{-1} &= U\Lambda^{-1}U^T\end{aligned}$$

Note that since $\Sigma \in \mathbb{S}_{++}^n$ (positive definite), we can write eigenvalues as $\{d_j^2\}$. Let $Y \equiv U^T(X-\mu)$ where U is from the diagonalization of Σ. We write $X = UY + \mu$ and

$$\begin{aligned}\Pr(Y \in \delta V) &= \Pr\Big(X \in U(\delta V){+}\mu\Big)\\ \mathrm{Vol}(\delta V)\rho_Y(y) &= \mathrm{Vol}(U\delta V{+}\mu)\,\rho_X(x)\Big|_{x=Uy+\mu}\end{aligned}$$

Note that the volume is invariant under a rigid body transformation. We obtain

$$\begin{aligned}&\mathrm{Vol}(U\delta V{+}\mu) = \mathrm{Vol}(\delta V)\\ &\rho_Y(y) = \rho_X(x)\Big|_{x=Uy+\mu} = \frac{1}{(2\pi)^{n/2}(\det\Sigma)^{1/2}}\exp\Big(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\Big)\Big|_{x=Uy+\mu}\end{aligned}$$

In the PDF of Y above, we have

$$\det\Sigma = (\det U)(\det\Lambda)(\det U^T) = \det\Lambda = \prod_{j=1}^n d_j^2$$

$$(x-\mu)^T\Sigma^{-1}(x-\mu)\Big|_{x=Uy+\mu} = (Uy)^T U\Lambda^{-1}U^T(Uy) = y^T\Lambda^{-1}y = \sum_{j=1}^n \frac{y_j^2}{d_j^2}$$

Using these results, we write out the PDF of Y.

$$\rho_Y(y) = \frac{1}{(2\pi)^{n/2}(\det\Lambda)^{1/2}}\exp\Big(\frac{-1}{2}y^T\Lambda^{-1}y\Big) = \prod_{j=1}^n \frac{1}{\sqrt{2\pi d_j^2}}\exp\Big(\frac{-y_j^2}{2d_j^2}\Big)$$

This is a product of n functions, each a 1D normal density. We conclude

$$\begin{aligned}&Y \sim N(0,\Lambda) \in \mathbb{R}^n, \qquad \Lambda = \mathrm{diag}(d_1^2, d_2^2, \ldots, d_n^2)\\ &Y_j \sim N(0, d_j^2) \in \mathbb{R}, \quad Y_i \text{ and } Y_j \text{ are independent } (i \neq j).\end{aligned}$$

This leads to $\displaystyle\int_{\mathbb{R}^n}\rho_X(x;\mu,\Sigma)dx = \int_{\mathbb{R}^n}\rho_Y(y)dy = 1$, which justifies the name of density.

Standard isotropic normal

$Z \sim N(0, I_n) \in \mathbb{R}^n$ is called the standard isotropic normal, in which

$$Z_j \sim N(0,1) \in \mathbb{R}, \quad Z_i \text{ and } Z_j \text{ are independent } (i \neq j).$$

In terms of standard isotropic normal, we write $Y \equiv U^T(X-\mu) \sim N(0,\Lambda)$ as

$$Y = \Lambda^{1/2}Z, \qquad Z \sim N(0, I_n), \qquad \Lambda^{1/2} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$$

Finally, we write X in terms of standard isotropic normal: $X = U\Lambda^{1/2}Z + \mu$.

Theorem 1. *(Multivariate Gaussian as an affine mapping of standard isotropic normal)*
For $X \sim N(\mu, \Sigma) \in \mathbb{R}^n$, we can write is as

$$\boxed{X = U\Lambda^{1/2}Z + \mu, \qquad Z \sim N(0, I_n), \qquad \Sigma = U\Lambda U^T}$$

We make a few observations:

- Any multivariate normal $X \sim N(\mu, \Sigma)$ can be viewed as an affine mapping of a standard isotropic normal Z.
- This makes sense even when $\Sigma \in \mathbb{S}_+^n$ (when it is only positive semi-definite). When $d_j = 0$, we simply take the limit as $d_j \to 0_+$; everything makes sense.

3 Partition function and a key result

$$\rho_X(x) \propto \exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \quad \longleftarrow \quad \text{energy form of density}$$

$$Z \equiv \int_{\mathbb{R}^n} \exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))dx \quad \longleftarrow \quad \text{definition of partition function}$$

Theorem 2. *(a key result on partition function)*

$$\boxed{Z \equiv \underbrace{\int_{\mathbb{R}^n} \exp(\frac{-1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))dx = (2\pi)^{n/2}(\det\Sigma)^{1/2}}_{\textit{key result}}}$$

This result is valid even when μ is a complex vector.

4 Characteristic function of a multivariate normal

For $X \sim N(\mu, \Sigma) \in \mathbb{R}^n$, its characteristic function (CF) is

$$\begin{aligned}\phi_X(\xi) &= E\Big(\exp(i\xi^T X)\Big), \qquad \xi \in \mathbb{R}^n \\ &= \frac{1}{Z}\int_{\mathbb{R}^n} \exp(i\xi^T x - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu))dx\end{aligned}$$

In the exponent, we complete the square (homework).

$$\begin{aligned}&i\xi^T x - \frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu) \\ &= -\frac{1}{2}(x-\mu-i\Sigma\xi)^T\Sigma^{-1}(x-\mu-i\Sigma\xi) + \underbrace{(i\xi^T\mu - \frac{1}{2}\xi^T\Sigma\xi)}_{\text{does not contain } x}\end{aligned}$$

Apply the result of completing the square in the expression of CF, we obtain

$$\phi_X(\xi) = \underbrace{[\frac{1}{Z}\int_{\mathbb{R}^n} \exp(-\frac{1}{2}(x-\mu-i\Sigma\xi)^T\Sigma^{-1}(x-\mu-i\Sigma\xi))dx]}_{=1, \text{ from the key result}} \exp(i\xi^T\mu - \frac{1}{2}\xi^T\Sigma\xi)$$

Theorem 3. *(Characteristic function of multivariate Gaussian)*

$$\boxed{\begin{aligned} X \sim N(\mu, \Sigma) \quad &\Longleftrightarrow \quad \phi_X(\xi) = \exp(i\xi^T \mu - \frac{1}{2}\xi^T \Sigma \xi) \\ &\Longleftrightarrow \quad \phi_{(X-\mu)}(\xi) = \exp(-\frac{1}{2}\xi^T \Sigma \xi) \end{aligned}}$$

Below, we use the expression of CF to derive other results.

5 Justifying the names of μ and Σ

We show $E(X_j) = \mu_j$ and $E\big((X_i - \mu_i)(X_j - \mu_j)\big) = \sigma_{ij}$.
Differentiating $\phi_{(X-\mu)}(\xi)$ with respect to ξ_j gives

$$E\big(i(X_j - \mu_j)\big) = \frac{\partial \phi_{(X-\mu)}(\xi)}{\partial \xi_j}\Big|_{\xi=0} = \frac{\partial \exp(-\frac{1}{2}\xi^T \Sigma \xi)}{\partial \xi_j}\Big|_{\xi=0} = 0$$

$$\Longrightarrow \quad E(X_j) = \mu_j$$

Differentiating $\phi_{(X-\mu)}(\xi)$ with respect to ξ_i and ξ_j leads to

$$E\big(-(X_i - \mu_i)(X_j - \mu_j)\big) = \frac{\partial^2 \phi_{(X-\mu)}(\xi)}{\partial \xi_i \partial \xi_j}\Big|_{\xi=0} = \frac{\partial^2 \exp(-\frac{1}{2}\xi^T \Sigma \xi)}{\partial \xi_i \partial \xi_j}\Big|_{\xi=0} = -\sigma_{ij}$$

$$\Longrightarrow \quad E\big((X_i - \mu_i)(X_j - \mu_j)\big) = \sigma_{ij}$$

6 Affine mapping of a Gaussian

Theorem 4. *(An affine mapping of a Gaussian is a Gaussian)*
Let $X \sim N(\mu, \Sigma) \in \mathbb{R}^n$. Consider $Y \equiv AX + b$ where $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^m$. We have

$$\boxed{Y \sim N(\mu_Y, \Sigma_{YY}), \quad \mu_Y = A\mu + b, \quad \Sigma_{YY} = A\Sigma A^T}$$

Proof. We write $Y = A(X-\mu) + A\mu+b$ and find its CF.

$$\begin{aligned} \phi_Y(\xi) &= E\Big(\exp(i\xi^T Y)\Big) = E\Big(\exp[i\xi^T A(X-\mu) + i\xi^T (A\mu+b)]\Big), \qquad \xi \in \mathbb{R}^m \\ &= \exp[i\xi^T (A\mu+b)] E\Big(\exp[i(A^T\xi)^T (X-\mu)]\Big), \qquad \tilde{\xi} \in \mathbb{R}^n \\ &= \exp[i\xi^T (A\mu+b)] \phi_{(X-\mu)}(\tilde{\xi})\Big|_{\tilde{\xi}=A^T\xi} = \exp[i\xi^T (A\mu+b)] \exp(-\frac{1}{2}\tilde{\xi}^T \Sigma \tilde{\xi})\Big|_{\tilde{\xi}=A^T\xi} \\ &= \exp[i\xi^T \underbrace{(A\mu+b)}_{\mu_Y} - \frac{1}{2}\xi^T \underbrace{(A\Sigma A^T)}_{\Sigma_Y Y} \xi] = \exp[i\xi^T \mu_Y - \frac{1}{2}\xi^T \Sigma_{YY} \xi] \end{aligned}$$

Since the CF is reversible, we conclude $Y \sim N(\mu_Y, \Sigma_{YY})$. $\square$

Special case 4.1 (Sum of independent Gaussians). Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{bmatrix}\right), \qquad X, Y \in \mathbb{R}^n$$

Then we have

$$(X+Y) \sim N(\mu_X + \mu_Y,\ \Sigma_{XX} + \Sigma_{YY})$$

Derivation: In Theorem 4, pick $A = [I \ \ I]$ and $b = 0$.

$$A \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} = \mu_X + \mu_Y, \qquad A \begin{bmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{bmatrix} A^T = \Sigma_{XX} + \Sigma_{YY}$$

Special case 4.2 (Marginal distribution of Gaussian). Let

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right), \qquad X \in \mathbb{R}^m, \quad Y \in \mathbb{R}^n$$

Here m and n may be different. Then we have

$$X \sim N(\mu_X,\ \Sigma_{XX}), \qquad Y \sim N(\mu_Y,\ \Sigma_{YY})$$

Derivation: In Theorem 4, pick $A = [I \ \ 0]$ and $b = 0$.

$$A \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix} = \mu_X, \qquad A \begin{bmatrix} \Sigma_{XX} & 0 \\ 0 & \Sigma_{YY} \end{bmatrix} A^T = \Sigma_{XX}$$

Special case 4.3 (Independent Gaussians based on the standard isotropic normal).

Let $A \in \mathbb{R}^{m \times n}$ be a matrix with orthogonal rows. In the matrix form, A satisfies

$$AA^T = \Lambda = \operatorname{diag}(d_1^2, d_2^2, \ldots, d_m^2), \qquad d_i = \|a_{i,:}\|$$

Here we do not require $\|a_{i,:}\| = 1$. Then for $Z \sim N(0, I_n)$, we have

$$X = AZ \sim N(0, \Lambda), \qquad \Lambda = \operatorname{diag}(d_1^2, d_2^2, \ldots, d_m^2)$$

That is, the components of $X = AZ$ are independent Gaussians. This result is practically useful.

7 Conditional distribution of Gaussian

Theorem 5. *(Conditional distribution of X when Y is fixed). Let*

$$\begin{bmatrix} X \\ Y \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix}\right), \qquad X \in \mathbb{R}^m, \quad Y \in \mathbb{R}^n$$

Here m and n may be different. Then we have

$$\boxed{\begin{aligned} &(X|Y=y) \sim N(\mu_{X|Y},\ \Sigma_{X|Y}) \\ &\qquad \mu_{X|Y} = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y - \mu_Y) \\ &\qquad \Sigma_{X|Y} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \end{aligned}}$$

Proof. For finding the conditional distribution, the characteristic function is not very helpful. We work directly with density. The conditional density of $(X|Y=y)$ is

$$\rho_{(X|Y=y)}(x) = \frac{\rho_{(X,Y)}(x,y)}{\rho_Y(y)} \propto \rho_{(X,Y)}(x,y)$$

$$\propto \exp(\frac{-1}{2}[(x-\mu_X)^T \;\; (y-\mu_Y)^T]\begin{bmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_X\\ y-\mu_Y\end{bmatrix})$$

Note that we examine $\rho_{(X|Y=y)}(x)$ as a function of x. The denominator $\rho_Y(y)$ is independent of x and is viewed as a part of the normalizing factor. To proceed, we write Σ^{-1} as

$$\begin{bmatrix}\Sigma_{XX} & \Sigma_{XY}\\ \Sigma_{YX} & \Sigma_{YY}\end{bmatrix}^{-1} = \begin{bmatrix}A & B\\ B^T & C\end{bmatrix}$$

where the needed properties of A, B and C are to be determined. The full expressions of A, B and C are neither necessary nor sufficient! We write $\rho_{(X|Y=y)}(x)$ as

$$\rho_{(X|Y=y)}(x) \propto \exp(\frac{-1}{2}[(x-\mu_X)^T \;\; (y-\mu_Y)^T]\begin{bmatrix}A & B\\ B^T & C\end{bmatrix}\begin{bmatrix}x-\mu_X\\ y-\mu_Y\end{bmatrix})$$

$$\propto \exp(\frac{-1}{2}[(x-\mu_X)^T A(x-\mu_X) + 2(x-\mu_X)^T B(y-\mu_y)])$$

Again, any term independent of x in the exponent contributes only to the normalizing factor. In the exponent, we complete the square (homework).

$$(x-\mu_X)^T A(x-\mu_X) + 2(x-\mu_X)^T B(y-\mu_y)$$
$$= (x-\mu_X + A^{-1}B(y-\mu_Y))^T A(x-\mu_X + A^{-1}B(y-\mu_Y)) + \underbrace{G(y)}_{\text{does not contain } x}$$

Apply the result of completing the square in conditional density, we obtain

$$\rho_{(X|Y=y)}(x) \propto \exp\Big(\frac{-1}{2}(x-\mu_X + A^{-1}B(y-\mu_Y))^T (A^{-1})^{-1}(x-\mu_X + A^{-1}B(y-\mu_Y))\Big)$$

$$\Longrightarrow \quad (X|Y=y) \sim N(\mu_{X|Y},\, \Sigma_{X|Y}), \quad \mu_{X|Y} = \mu_X - A^{-1}B(y-\mu_Y), \quad \Sigma_{X|Y} = A^{-1}$$

Lemma 1. *(expression of $A^{-1}B$ and A^{-1})*

$$\begin{cases} A^{-1}B = -\Sigma_{XY}\Sigma_{YY}^{-1} \\ A^{-1} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \end{cases}$$

The proof of Lemma is in your homework.
Substituting the result of Lemma into the expression of $\mu_{X|Y}$ and $\Sigma_{X|Y}$, we obtain

$$\begin{cases} \mu_{X|Y} = \mu_X - A^{-1}B(y-\mu_Y) = \mu_X + \Sigma_{XY}\Sigma_{YY}^{-1}(y-\mu_Y) \\ \Sigma_{X|Y} = A^{-1} = \Sigma_{XX} - \Sigma_{XY}\Sigma_{YY}^{-1}\Sigma_{YX} \end{cases}$$

This concludes the proof of Theorem 5. $\square$

Sometimes we can use an ad hoc way to find $(X|Y)$

An ad hoc method (conditional distribution of combinations of standard isotropic normal)

Let $Z \sim N(0, I_n)$. We have

- $(a_1 Z_1 + a_2 Z_2)$ and $(a_2 Z_1 - a_1 Z_2)$ are independent.

$$X \equiv \begin{bmatrix} (a_2 Z_1 - a_1 Z_2) \\ (a_1 Z_1 + a_2 Z_2) \end{bmatrix} = \begin{bmatrix} a_2 & -a_1 \\ a_1 & a_2 \end{bmatrix} \begin{bmatrix} Z_1 \\ Z_2 \end{bmatrix} \sim N(0, \begin{bmatrix} a_1^2 + a_2^2 & 0 \\ 0 & a_1^2 + a_2^2 \end{bmatrix})$$

  Note that the matrix above has orthogonal rows. It follows that

$$\Big(a_2 Z_1 - a_1 Z_2 \Big| a_1 Z_1 + a_2 Z_2 = x_2\Big) \sim \Big(a_2 Z_1 - a_1 Z_2\Big)$$

- Conditional distributions involving (Z_1, Z_2) are independent of $\{Z_j,\ j = 3, \ldots n\}$.

$$\Big(b_1 Z_1 + b_2 Z_2 \Big| a_1 Z_1 + a_2 Z_2 = x_2,\ Z_j = z_j,\ j = 3, \ldots n\Big)$$
$$\sim \Big(b_1 Z_1 + b_2 Z_2 \Big| a_1 Z_1 + a_2 Z_2 = x_2\Big)$$

- An example:

$$\Big(a_1 Z_1 \Big| a_1 Z_1 + a_2 Z_2 = x_2,\ Z_j = z_j,\ j = 3, \ldots n\Big) \sim \Big(a_1 Z_1 \Big| a_1 Z_1 + a_2 Z_2 = x_2\Big)$$
$$\sim \Big(a_1 \underbrace{\frac{1}{a_1^2 + a_2^2}\big[a_2(a_2 Z_1 - a_1 Z_2) + a_1(a_1 Z_1 + a_2 Z_2)\big]}_{Z_1} \Big| a_1 Z_1 + a_2 Z_2 = x_2\Big)$$
$$\sim \Big(\frac{a_1 a_2 X_1 + a_1^2 X_2}{a_1^2 + a_2^2} \Big| X_2 = x_2\Big), \qquad X_1 \sim N(0, a_1^2 + a_2^2)$$
$$\sim N(\frac{a_1^2 x_2}{a_1^2 + a_2^2}, \frac{a_1^2 a_2^2}{a_1^2 + a_2^2})$$

  In particular, for $a_1 = a_2 = a$ we have

$$\boxed{\Big(a Z_1 \Big| a Z_1 + a Z_2 = x_2\Big) \sim N(\frac{x_2}{2}, \frac{a^2}{2})}$$

  This result is very useful in the discussion of constrained Wiener process.