STAT 512: Statistical Inference **Autumn 2020**

Lecture 8: Linear models and multivariate normal distributions

Instructor: Yen-Chi Chen

Reference: Casella and Berger Chapter 4.

8.1 Review of linear algebra

An $m \times n$ matrix $A = \{A_{ij}\}$ is an array of nm elements such that

$$A = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix}.$$

In this case, we can write $A \in \mathbb{R}^{m\times n}$. The matrix represents a linear mapping (linear transformation) $A : \mathbb{R}^n \to \mathbb{R}^m$ ($x \mapsto Ax$), where $x \in \mathbb{R}^n$ is written as a column vector (i.e., an $n \times 1$ matrix) and

$$Ax = \begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{22} & \cdots & A_{2n} \\ \vdots & \vdots & \vdots & \vdots \\ A_{m1} & A_{m2} & \cdots & A_{mn} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix} = \begin{pmatrix} \sum_j A_{1j}x_j \\ \sum_j A_{2j}x_j \\ \vdots \\ \sum_j A_{mj}x_j \end{pmatrix}$$

Clearly, the above operation implies the linear addition, i.e., for any $a, b \in \mathbb{R}$ and $x, y \in \mathbb{R}^n$, $A(ax + by) = aAx + bAy$.

For two $m \times n$ matrices A, B, the addition $A + B$ is another $m \times n$ matrix such that $[A + B]_{ij} = A_{ij} + B_{ij}$. For an $m \times n$ matrix A and an $n \times p$ matrix B, the *matrix multiplication* AB is an $m \times p$ matrix such that

$$[AB]_{ij} = \sum_{k=1}^{n} A_{ik}B_{kj}.$$

A very important property is that $AB \neq BA$ in general even if $m = n = p$.

8.1.1 Useful characteristics of a matrix

Rank. The *rank* of a matrix A, denoted as $\mathsf{rank}(A)$, is the dimension of its column space. The column space is the vector space spanned by $A_{+1}, \cdots, A_{+n}$, the column vectors of A, i.e.,

$$A_{+j} = \begin{pmatrix} A_{1j} \\ A_{2j} \\ \vdots \\ A_{mj} \end{pmatrix}.$$

One can easily verify that $\mathsf{rank}(A) \leq \min\{m, n\}$. Also, $\mathsf{rank}(AB) \leq \min\{\mathsf{rank}(A), \mathsf{rank}(B)\}$.

Identity matrix. The $n \times n$ identity matrix $\mathbf{I}_n$ is a matrix that has 1's on its diagonal and 0 elsewhere. Namely, $\mathbb{I}_n = \mathsf{Diag}(1, 1, 1, \cdots, 1)$. One can easily see that for an $m \times n$ matrix A and $n \times m$ matrix B,. $A\mathbf{I}_n = A$ and $\mathbf{I}_n B = B$.

Inverse. The *inverse* of an $n \times n$ (square) matrix A, denoted as A^{-1}, is an $n \times n$ matrix with the property that $AA^{-1} = A^{-1}A = \mathbf{I}_n$. Note: the inverse may not exist. When the inverse of A exists, A is called *regular* otherwise it is called singular. The followings are equivalent of a $n \times n$ square matrix A:

- A is regular/non-singular (i.e., has an inverse matrix).
- A is full rank, i.e., $\mathsf{rank}(A) = n$.
- The determinant of A is not 0 (we will define determinant later).

If both $n \times n$ matrices A, B are regular, then AB is also regular with inverse $(AB)^{-1} = B^{-1}A^{-1}$. For a diagonal matrix $D = \mathsf{Diag}(d_1, \cdots, d_n)$, its inverse is $D^{-1} = \mathsf{Diag}(d_1^{-1}, \cdots, d_n^{-1})$.

Transpose. For an $m \times n$ matrix A, its *transpose*, denoted as A^T, is an $n \times m$ matrix such that $[A^T]_{ij} = A_{ji}$. You can easily verify that $(A + B)^T = A^T + B^T$, $(AB)^T = B^T A^T$, and $(A^{-1})^T = (A^T)^{-1}$.

Trace. For an $n \times n$ matrix A, its *trace*, denoted as $\mathsf{Tr}(A)$, is $\mathsf{Tr}(A) = \sum_{i=1}^n A_{ii}$. One can easily verify that $\mathsf{Tr}(aA + bB) = a\mathsf{Tr}(A) + b\mathsf{Tr}(A)$ and $\mathsf{Tr}(A) = \mathsf{Tr}(A^T)$. Moreover, for an $m \times n$ matrix A and an $n \times m$ matrix B, $\mathsf{Tr}(AB) = \mathsf{Tr}(BA)$.

Triangular matrix. An $n \times n$ matrix A is upper triangular if $A_{ij} = 0$ for all $i < j$. An $n \times n$ matrix A is lower triangular if A^T is upper triangular. A matrix is called triangular if it is either upper or lower triangular.

Determinant. For an $n \times n$ matrix A, its *determinant*, denoted as $|A|$, is

$$\mathsf{det}(A) = \sum_{\pi} \epsilon(\pi) \prod_{i=1}^n A_{i\pi(i)},$$

where π is all possible permutations of $\{1, 2, 3, \cdots, n\}$ and $\epsilon(\pi) = \pm 1$ according to if the permutation is even or odd permutation. Here are some useful properties of the determinant: $\mathsf{det}(AB) = \mathsf{det}(A) \cdot \mathsf{det}(B)$ when they are both square matrices, $\mathsf{det}(A)^{-1} = \mathsf{det}(A^{-1})$, $\mathsf{det}(A^T) = \mathsf{det}(A)$, $\mathsf{det}(A) = \prod_{i=1}^n A_{ii}$ if A is triangular.

Orthogonal matrix. An $n \times n$ matrix U is *orthogonal* if $U^T U = \mathbf{I}_n$. Namely, its column vectors form an orthonormal basis of $\mathbb{R}^n$. Note that one can easily see that this implies that $U^T = U^{-1}$ so $UU^T = \mathbf{I}_n$ as well.

Eigenvalues and eigenvectors. For an $n \times n$ matrix, its *eigenvalues* are the n roots $\lambda_1, \cdots, \lambda_n$ to the following polynomial equation:

$$\mathsf{det}(A - \lambda \mathbf{I}_n) = 0.$$

For each λ_j, there exists a vector u_j such that $(A - \lambda_j \mathbf{I}_n)u_j = 0$ or $Au_j = \lambda_j u_j$. Such a vector u_j is called the *eigenvector* corresponding to λ_j. Note that if λ_j is distinct from other eigenvalues, then u_j is unique. Also note that the eigenvalues and eigenvector may not be real numbers/vectors.

8.1.2 Symmetric matrices

A square matrix $A \in \mathbb{R}^{n \times n}$ is *symmetric* if $A_{ij} = A_{ji}$, i.e., $A = A^T$. In what follows, we will review some useful properties of a symmetric matrix.

For a symmetric matrix $A \in \mathbb{R}^{n \times n}$, it has the following properties:

- Eigenvalues and eigenvectors are real numbers/vectors.
- For eigenvalues $\lambda_j \neq \lambda_k$, their corresponding eigenvectors u_j, u_k are orthogonal, i.e., $u_j^T u_k = 0$.
- **Spectral decomposition.** Let $\lambda_1, \cdots, \lambda_n$ be the eigenvalues of A and $u_1, \cdots, u_n$ be the corresponding eigenvectors. Let $\Lambda = \mathsf{Diag}(\lambda_1, \cdots, \lambda_n)$ and $U = [u_1, \cdots, u_n]$. Then

  $$A = U\Lambda U^T = \sum_{i=1}^{n} \lambda_i u_i u_i^T.$$

  This is known as the spectral decomposition.
- **Trace.** The trace of A is $\mathsf{Tr}(A) = \sum_{i=1}^{n} \lambda_i$.
- **Determinant.** The determinant of A is $\mathsf{det}(A) = \prod_{i=1}^{n} \lambda_i$

Positive definite matrix. A particular important class of symmetric matrices is the *positive definite (PD) matrices.* A square matrix $A \in \mathbb{R}^{n \times n}$ is *positive semi-definite (PSD)* if

$$x^T A x \geq 0$$

for all $x \in \mathbb{R}^n$. It is *positive definite* if

$$x^T A x > 0$$

for all $x \in \mathbb{R}^n$ and $x^T x > 0$.

Here are some useful properties of PD and PSD matrices.

- The identity matrix is PD.
- A diagonal matrix D is PD if $D_{ii} > 0$ for all i and is PSD if $D_{ii} \geq 0$ for all i.
- If $S \in \mathbb{R}^{n \times n}$ is PSD and $A \in \mathbb{R}^{m \times n}$ be any matrix, then ASA^T is PSD.
- If $S \in \mathbb{R}^{n \times n}$ is PD and $A \in \mathbb{R}^{m \times n}$ be any matrix with $\mathsf{rank}(A) = m \leq n$, then ASA^T is PD.
- AA^T is PSD for any $m \times n$ matrix A.
- AA^T is PD for any $m \times n$ matrix A with $\mathsf{rank}(A) = m \leq n$.
- A is PD $\Rightarrow$ A is full rank $\Rightarrow$ A^{-1} exists $\Rightarrow$ $A^{-1} = A^{-1}AA^{-1}$ is PD.
- A symmetric matrix A is PSD (PD) if all its eigenvalues $\lambda_j \geq 0$ (> 0).
- If $A \in \mathbb{R}^{n \times n}$ is PD, then let its spectral decomposition be $A = U\Lambda U^T$. Then the square root of A, a matrix C such that $CC^T = A$, is $C = U\sqrt{\Lambda}U^T$, where $\sqrt{\Lambda} = \mathsf{Diag}(\sqrt{\Lambda_{11}}, \cdots, \sqrt{\Lambda_{nn}})$.

Partitioned PD matrix. Suppose that $A \in \mathbb{R}^{n \times n}$ is a PD matrix and we suppose that it can be decomposed into 4 submatrices

$$A = \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix},$$

where $S_{ij} \in \mathbb{R}^{n_i \times n_j}$ with $i, j = 1, 2$ and $n = n_1 + n_2$. Then we have the follow properties:

- S_{11} and S_{22} are both PD.

- Let $S_{11,2} = S_{11} - S_{12}S_{22}^{-1}S_{21}$. Then

$$\begin{pmatrix} \mathbf{I}_{n_1} & -S_{12}S_{22}^{-1} \\ 0 & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{n_1} & 0 \\ -S_{22}^{-1}S_{21} & \mathbf{I}_{n_2} \end{pmatrix} = \begin{pmatrix} S_{11,2} & 0 \\ 0 & S_{22} \end{pmatrix}$$

so $S_{11,2}$ is PD as well.

- Following from the above result, we have

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{n_1} & S_{12}S_{22}^{-1} \\ 0 & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} S_{11,2} & 0 \\ 0 & S_{22} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{n_1} & 0 \\ S_{22}^{-1}S_{21} & \mathbf{I}_{n_2} \end{pmatrix}$$

$$\begin{pmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{I}_{n_1} & 0 \\ -S_{22}^{-1}S_{21} & \mathbf{I}_{n_2} \end{pmatrix} \begin{pmatrix} S_{11,2}^{-1} & 0 \\ 0 & S_{22}^{-1} \end{pmatrix} \begin{pmatrix} \mathbf{I}_{n_1} & -S_{12}S_{22}^{-1} \\ 0 & \mathbf{I}_{n_2} \end{pmatrix}$$

- Further, the above implies that

$$A \text{ is PD } \Leftrightarrow S_{11,2}, S_{22} \text{ are PD } \Leftrightarrow S_{22,1}, S_{11} \text{ are PD }.$$

- For any vector $x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \in \mathbb{R}^n$ such that $x_1 \in \mathbb{R}^{n_1}$ and $x_2 \in \mathbb{R}^{n_2}$,

$$xA^{-1}x = (x_1 - S_{12}S_{22}^{-1}x_2)S_{11,2}^{-1}(x_1 - S_{12}S_{22}^{-1}x_2) + x_2S_{22}^{-1}x_2.$$

Later we will see that the above results are very useful in analyzing the conditional normal distribution.

8.1.3 Projection matrices

An $n \times n$ matrix P is called a *projection* matrix if it is symmetric and idenpotent ($P^2 = P$).

P is a projection matrix if and only if there exists orthogonal matrix U such that

$$P = U \begin{pmatrix} \mathbf{I}_m & 0 \\ 0 & 0 \end{pmatrix} U^T.$$

In this case $\mathsf{rank}(P) = m$.

Suppose that we can partition $U = [U_1, U_2]$, where $U_1 \in \mathbb{R}^{n \times m}$ and $U_2 \in \mathbb{R}^{n \times (n-m)}$. Then the above result implies that $P = U_1U_1^T$ and $PU_1 = U_1$ and $PU_2 = 0$. This means that P project any vector in $\mathbb{R}^n$ into the column space of U_1 and is orthogonal to the column space of U_2. An interesting property is that $\mathsf{rank}(P) = \mathsf{Tr}(P) = m$.

Also, the matrix $\mathbf{I}_n - P$ is another projection matrix that projects any vector in $\mathbb{R}^n$ to the space orthogonal to the column space of U_1. To see this, $P(\mathbf{I}_n - P) = P - P^2 = 0$.

8.2 Transforming multiple continuous random variables

In lecture 2, we have learned techniques to deal with transforming a single continuous random variable, i.e., investigating the distribution of $U = f(X)$ when we know the distribution of X. In this section, we will study a more general problem where we are transforming two or more (continuous) random variables.

We start with a simple case where we have two random variables X, Y and we know their joint PDF. Consider two random variables $U = f(X, Y)$ and $V = g(X, Y)$, where u, v are two known functions.

We now study the joint PDF of (U, V). By definition,

$$\begin{aligned} p_{U,V}(u, v) &= \frac{\partial^2}{\partial u \partial v} P(U \leq u, V \leq v) \\ &= \frac{\partial^2}{\partial u \partial v} P(f(X, Y) \leq u, g(X, Y) \leq v) \\ &= \frac{\partial^2}{\partial u \partial v} P((X, Y) \in R(u, v)) \\ &= \frac{\partial^2}{\partial u \partial v} \int_{R(u,v)} p_{X,Y}(x, y) dxdy, \end{aligned}$$

where

$$R(u, v) = \{(x, y) : f(x, y) \leq u, g(x, y) \leq v\}.$$

In some simple scenarios, this region $R(u, v)$ has a nice form so that the probability $P((X, Y) \in R(u, v))$ has an analytical expression that we can take derivatives easily. However, this expression might still be hard to compute in general.

Example 1. Let $X, Y \sim \mathsf{Unif}[0, 1]$. Consider $U = \max\{X, Y\}, V = \min\{X, Y\}$. Note that there is an implicit constraint on $f_{U,V}$ that $f_{U,V}(u, v) = 0$ if $v > u$. So we consider any pair $(u, v) : v \leq u$. By a direct computation,

$$\begin{aligned} P(U \leq u, V \leq v) &= P(U \leq u) - P(U \leq u, V > v) \\ &= P(X \leq u, Y \leq u) - P(X \leq u, Y \leq u, X > v, Y > v) \\ &= P(X \leq u)P(Y \leq u) - P(v < X \leq u)P(v < Y \leq u) \\ &= u^2 - (u - v)^2 \end{aligned}$$

when $0 \leq v \leq u \leq 1$. Thus,

$$p_{U,V}(u, v) = \frac{\partial^2}{\partial u \partial v} P(U \leq u, V \leq v) = 2I(0 \leq v \leq u \leq 1).$$

Example 2. Consider $X, Y \sim \mathsf{Exp}(1)$ and let $U = X + Y$ and $V = \frac{X}{X+Y}$. Note that $(U, V) \in [0, \infty) \times [0, 1]$. So we consider any $u \geq 0$ and $v \in [0, 1]$. The joint CDF is

$$\begin{aligned} P(U \leq u, V \leq v) &= P(X + Y \leq u, X \leq v(X + Y)) \\ &= P\left(Y \leq u - X, Y \geq \frac{1 - v}{v} X\right) \\ &= \mathbb{E}\left[I\left(Y \leq u - X, Y \geq \frac{1 - v}{v} X\right)\right] \\ &= \mathbb{E}\left[\mathbb{E}\left[I\left(Y \leq u - X, Y \geq \frac{1 - v}{v} X\right) |X\right]\right] \\ &= \mathbb{E}\left[P\left(Y \leq u - X, Y \geq \frac{1 - v}{v} X|X\right)\right]. \end{aligned}$$

Note that $I(E)$ is the indicator function such that it returns 1 if the event E is true and 0 otherwise; one can easily see that $\mathbb{E}[I(E)] = P(E)$. Condition on X, the probability

$$\begin{aligned}
P\left(Y \leq u - X, Y \geq \frac{1-v}{v}X|X\right) &= P(\frac{1-v}{v}X \leq Y \leq u - X|X) \\
&= \int_{y=\frac{1-v}{v}X}^{u-X} e^{-y}dy \\
&= e^{-\frac{1-v}{v}X} - e^{X-u}.
\end{aligned}$$

Thus, using the fact that $U \leq u, V \leq v \Rightarrow X \leq uv$, we have

$$\begin{aligned}
P(U \leq u, V \leq v) &= \mathbb{E}\left[P\left(Y \leq u - X, Y \geq \frac{1-v}{v}X|X\right)\right] \\
&= \int_0^{uv} [e^{-\frac{1-v}{v}x} - e^{x-u}]e^{-x}dx \\
&= \int_0^{uv} [e^{-\frac{x}{v}} - e^{-u}]dx \\
&= v(1 - e^{-u} - ue^{-u}).
\end{aligned}$$

By taking the derivative, we obtain

$$p_{U,V}(u,v) = ue^{-u}I(0 \leq v \leq 1) = \underbrace{ue^{-u}}_{p_U(u)} \cdot \underbrace{I(0 \leq v \leq 1)}_{p_V(v)}.$$

Thus, we conclude that $U \sim \mathsf{Gamma}(2,1)$ and $V \sim \mathsf{Uni}[0,1]$ and $U \perp V$.

8.2.1 Jacobian method

The Jacobian method is an elegant approach for substituting variables (change of varibales) in an integration. Consider $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^n$ and assume that there is a 1-1 and onto mapping $T : \mathbb{R}^n \to \mathbb{R}^n$ for almost all x such that $y = T(x)$. We define the Jacobian matrix

$$J_T(x) = \left(\frac{\partial T(x)}{\partial x}\right) = \left(\frac{\partial y}{\partial x}\right) = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_2}{\partial x_1} & \cdots & \frac{\partial y_n}{\partial x_1} \\ \frac{\partial y_1}{\partial x_2} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_2} \\ \vdots & \vdots & \cdots & \vdots \\ \frac{\partial y_1}{\partial x_n} & \frac{\partial y_2}{\partial x_n} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix} \in \mathbb{R}^{n\times n}.$$

The *Jacobian* is the absolute value of the determinant of this matrix, i.e., $|\mathsf{det}(J_T(x))| = |\left(\frac{\partial y}{\partial x}\right)| = \left|\frac{\partial y}{\partial x}\right|$.

Theorem 8.1 *Assume that $y = T(x)$, where T is 1-1 and onto for almost all x and the Jacobian $\mathsf{det}(J_T(x)) \neq 0$ for all x. Let $A, B \subset \mathbb{R}^n$ be two subsets such that $B = \{T(x) : x \in A\}$. Let f be an integrable function. Then*

$$\int_A f(x)dx = \int_B f(T^{-1}(y))\left|\frac{\partial x}{\partial y}\right| dy.$$

Under the same condition, suppose X is a random variable with a PDF $p_X(x)$ and $Y = T(X)$. Then the PDF of Y is

$$p_Y(y) = p_X(T^{-1}(y))\left|\frac{\partial x}{\partial y}\right|.$$

The Jacobian has a nice chain rule that if $z = S(y)$ and $y = T(x)$ such that S, T are both 1-1 and onto. Then

$$\left|\frac{\partial z}{\partial x}\right| = \left|\frac{\partial z}{\partial y}\right|\left|\frac{\partial y}{\partial x}\right|.$$

Also, we have the inverse rule:

$$\left|\frac{\partial y}{\partial x}\right| = \left|\frac{\partial x}{\partial y}\right|^{-1}.$$

Example: Gamma distributions. Consider X, Y are independently from Gamma distribution with parameter α, λ. Recall that the PDF of a Gamma (α, λ) is

$$p(x) = \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x} I(x \geq 0).$$

Now we consider $U = X + Y$ and $W = \frac{X}{X+Y}$. In this case, the mapping $T(x, y) = (u, w)$ such that $T = (T_1, T_2)$ with $T_1(x, y) = x + y$ and $T_2(x, y) = \frac{x}{x+y}$. Thus, the inverse mapping $T^{-1}(u, w) = (x, y)$ will be $T_1^{-1}(u, w) = uw$ and $T_2^{-1}(u, w) = u - uw$. The Jacobian

$$\begin{aligned}
\left|\frac{\partial(x, y)}{\partial(u, w)}\right| &= \left|\frac{\partial T^{-1}(u, w)}{\partial(u, w)}\right| \\
&= \left|\mathsf{det}\left(\begin{pmatrix} w & 1-w \\ u & -u \end{pmatrix}\right)\right| \\
&= u.
\end{aligned}$$

We already know the joint PDF $p_{XY}(x, y)$ since they are independent Gamma. Thus,

$$\begin{aligned}
p_{UW}(u, w) &= p_{XY}(T_1^{-1}(u, w), T_2^{-1}(u, w))uI(0 \leq w \leq 1, u \geq 0) \\
&= p_X(T_1^{-1}(u, w))p_Y(T_2^{-1}(u, w))uI(0 \leq w \leq 1, u \geq 0) \\
&= \frac{\lambda^{2\alpha}}{\Gamma^2(\alpha)}(uw)^{\alpha-1}e^{-\lambda uw}(u - uw)^{\alpha-1}e^{-\lambda(u-uw)}uI(0 \leq w \leq 1, u \geq 0) \\
&= \frac{\lambda^{2\alpha}}{\Gamma^2(\alpha)}u^{2\alpha-1}e^{-\lambda u}I(u \geq 1)w^{\alpha-1}(1-w)^{\alpha-1}I(0 \leq w \leq 1) \\
&= p_U(u)p_W(w)
\end{aligned}$$

such that $U \sim \mathsf{Gamma}(2\alpha, \lambda)$ and $W \sim \mathsf{Beta}(\alpha, \alpha)$.

Example: Polar coordinate. A common reparametrization of two variable X, Y is via the polar coordinate R, Θ. Specifically, we choose $R = \sqrt{X^2 + Y^2}$ and $\Theta \in [0, 2\pi]$ such that

$$X = R\cos(\Theta), \quad Y = R\sin(\Theta).$$

In this case, $T(x, y) = (r, \theta)$ is 1-1 and onto for almost all points (x, y) except $(0, 0)$ so we can still apply the Jacobian trick. You can easily work out that

$$\left|\frac{\partial(x, y)}{\partial(r, \theta)}\right| = r$$

so if we know the PDF of X, Y as $p_{X,Y}(x, y)$, then

$$p_{R,\Theta}(r, \theta) = p_{X,Y}(r\cos(\theta), r\sin(\theta))r.$$

If the joint PDF of (X, Y) is radial, i.e., $p_{X,Y}(x, y) = g(x^2 + y^2)$, then $p_{R,\Theta}(r, \theta) = g(r^2)r$ so $R \perp \Theta$ and $\Theta \sim \mathsf{Uni}[0, 2\pi]$.

8.3 Random vector and covariance matrix

A random vector is a vector of random variables. Let $X \in \mathbb{R}^n$ be a random vector. We often express X as a column vector, i.e.,

$$X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_n \end{pmatrix}.$$

The expectation/expected value of X is the elementwise expectation:

$$\mathbb{E}[X] = \begin{pmatrix} \mathbb{E}[X_1] \\ \mathbb{E}[X_2] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}.$$

Similar to random variables, the expectation is an linear operation of random vectors. Namely, for two random vectors $X, Y \in \mathbb{R}^n$ and two real numbers a, b,

$$\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y].$$

An important characteristic of a random vector is the *variance-covariance* matrix (often we just called it the covariance matrix):

$$\begin{aligned} \mathsf{Cov}(X) &= \mathbb{E}[(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T] \\ &= \begin{pmatrix} \mathsf{Var}(X_1) & \mathsf{Cov}(X_1, X_2) & \mathsf{Cov}(X_1, X_3) & \cdots & \mathsf{Cov}(X_1, X_n) \\ \mathsf{Cov}(X_2, X_1) & \mathsf{Var}(X_2) & \mathsf{Cov}(X_2, X_3) & \cdots & \mathsf{Cov}(X_2, X_n) \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ \mathsf{Cov}(X_n, X_1) & \mathsf{Cov}(X_n, X_2) & \mathsf{Cov}(X_n, X_3) & \cdots & \mathsf{Var}(X_n) \end{pmatrix}. \end{aligned}$$

Using the fact that $\mathsf{Var}(X_i) = \mathsf{Cov}(X_i, X_i)$, elements in the above matrix can be written as $\mathsf{Cov}(X)_{ij} = \mathsf{Cov}(X_i, X_j)$.

Here are some nice properties of the covariance matrices.

- $\mathsf{Cov}(X) = \mathbb{E}[XX^T] - \mathbb{E}[X]\mathbb{E}[X]^T$
- For a matrix $A \in \mathbb{R}^{m\times n}$ and a vector $b \in \mathbb{R}^m$,

$$\mathsf{Cov}(AX + b) = A\mathsf{Cov}(X)A^T.$$

- For a vector $a \in \mathbb{R}^n$, $\mathsf{Var}(a^T X) = a^T \mathsf{Cov}(X)a$.
- **The covariance matrix is positive semi-definite (PSD)**.
- The covariance matrix is PD if the only vector $a \in \mathbb{R}^n$ such that $\mathsf{Var}(a^T X) = 0$ is $a = 0$.

The covariance matrix immediately implies some useful properties of the sample mean. Suppose $X_1, \cdots, X_n$ are IID with mean u and variance σ^2. Then $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i = a^T X$, where $a_j = \frac{1}{n}$. As a result,

$$\mathsf{Var}(\bar{X}_n) = a^T \mathsf{Cov}(X)a = \frac{1}{n^2}\sum_{i=1}^n \mathsf{Var}(X_i) = \frac{\sigma^2}{n}.$$

Now, suppose that the random variables are not independent but instead, they have correlation $\mathsf{Cov}(X_i, X_j) = \rho$ when $i \neq j$. Then the variance of the sample mean will be

$$\begin{aligned}
\mathsf{Var}(\bar{X}_n) &= a^T \mathsf{Cov}(X) a \\
&= \begin{pmatrix} \frac{1}{n} & \frac{1}{n} & \cdots & \frac{1}{n} \end{pmatrix} \begin{pmatrix} \sigma^2 & \sigma^2\rho & \cdots & \sigma^2\rho \\ \sigma^2\rho & \sigma^2 & \cdots & \sigma^2\rho \\ \vdots & \vdots & \cdots & \vdots \\ \sigma^2\rho & \sigma^2\rho & \cdots & \sigma^2 \end{pmatrix} \begin{pmatrix} \frac{1}{n} \\ \frac{1}{n} \\ \vdots \\ \frac{1}{n} \end{pmatrix} \\
&= \frac{1}{n^2}(n\sigma^2 + n(n-1)\sigma^2\rho) \\
&= \frac{\sigma^2}{n}(1 + (n-1)\rho).
\end{aligned}$$

8.4 The multivariate normal distribution

Recall that for a standard Normal random variable Z_1, its PDF is

$$p_0(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

Thus, for iid random variables $Z_1, \cdots, Z_n$, we can represent them as a random vector Z and its joint PDF will be

$$p(z_1, \cdots, z_n) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi}} e^{-z_i^2/2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}\sum_{i=1}^n z_i^2} = \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}z^T z}.$$

Now we consider a linear transformation that $A \in \mathbb{R}^{n\times n}$ is an invertible square matrix and $\mu \in \mathbb{R}^n$ is a vector and $X = AZ + \mu$. Since Z is a random vector, X will also be a random vector. Using the fact that $Z = A^{-1}(X - \mu)$ and the Jacobian method, you can show that the PDF of X is

$$\begin{aligned}
p(x_1, \cdots, x_n) &= \left(\frac{1}{2\pi}\right)^{n/2} e^{-\frac{1}{2}(x-\mu)^T [A^{-1}]^T A^{-1}(x-\mu)} \frac{1}{\sqrt{\mathsf{det}(AA^T)}} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(AA^T)}} e^{-\frac{1}{2}(x-\mu)^T [AA^T]^{-1}(x-\mu)} \\
&= \left(\frac{1}{2\pi}\right)^{n/2} \frac{1}{\sqrt{\mathsf{det}(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)},
\end{aligned}$$

where $\Sigma = \mathsf{Cov}(X) = AA^T$ is the covariance matrix of X. Note that $\mathbb{E}[X] = \mu$ by construction. In this case, we will say that X is from a *multivariate normal distribution* with a mean (vector) μ and a covariance matrix Σ. For abbreviation, we often write $X \sim N_n(\mu, \Sigma)$.

Linearity. The linear transformation of multivariate normal is still normal. Namely,

$$Y = CX + b \sim N_n(C\mu + b, C\Sigma C^T)$$

for non-singular matrix $C \in \mathbb{R}^{n\times n}$ and any vector $b \in \mathbb{R}^n$. Also, for a vector $a \in \mathbb{R}^n$,

$$a^T X \sim N(a^T\mu, a^T\Sigma a).$$

Independence $\Leftrightarrow$ uncorrelation. You can easily verify that if X follows a multivariate normal, then

$$X_i \perp X_j \Leftrightarrow \mathsf{Cov}(X_i, X_j) \equiv \Sigma_{ij} = 0.$$

Namely, pairwise independent is the same as being uncorrelated.

Marginal is normal. Suppose we partition X into two blocks

$$X = \begin{pmatrix} X_1 \\ X_2 \end{pmatrix},$$

where $X_1 \in \mathbb{R}^{n_1}$ and $X_2 \in \mathbb{R}^{n_2}$. Let μ_1, μ_2 be the mean vector correspond to each of the block and $\Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$. Then you can easily verify that

$$W_1 \sim N_{n_1}(\mu_1, \Sigma_{11}), \quad W_2 \sim N_{n_2}(\mu_2, \Sigma_{22})$$

so the marginals of the random vector are also multivariate normals.

Conditional is normal. Following the partition in the marginal case, the conditional distribution of $X_1|X_2$ is

$$X_1|X_2 \sim N_{n_1}(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2), \Sigma_{11,2}),$$

where $\Sigma_{11,2} = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$. You can compare this to the partitioned of PD matrix in Section 8.1.2.

Regression is linear and covariance is constant. Suppose that we have bivariate normal random vector (X_1, Y_2). Then the regression function (conditional mean) is

$$\mathbb{E}[X_1|X_2] = \mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(X_2 - \mu_2),$$

and the conditional variance

$$\mathsf{Var}(X_1|X_2) = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21},$$

where $\Sigma_{ij} = \mathsf{Cov}(X_i, X_j)$. This follows directly from the properties of conditional normals.

8.4.1 Chi-square distribution

Let $X = (X_1, \cdots, X_n)^T$ be a multivariate normal vector with mean 0 and identity covariance matrix. Then the random variable

$$W_n = \sum_{i=1}^{n} X_i^2 = X^T X = \|X\|^2$$

has a distribution called the χ^2 distribution with a degree of freedom n. In this case, we write $W_n \sim \chi_n^2$. The χ_n^2 is the same as $\Gamma(\frac{n}{2}, \frac{1}{2})$ and $\mathbb{E}(W_n) = n$ and $\mathsf{Var}(W_n) = 2n$.

Normalizing a Gaussian vector. Suppose a random vector $Y \sim N(\mu, \Sigma)$, then

$$Z = \Sigma^{-\frac{1}{2}}(Y - \mu) \sim N(0, \mathbf{I}_n)$$

so

$$Z^T Z = (Y - \mu)^T \Sigma^{-1}(Y - \mu) \sim \chi_n^2.$$

Projection property. Here is an interesting property of a projection matrix. Lete $X \sim N(\mu, \mathbf{I}_n)$ be a multivariate normal vector in $\mathbb{R}^n$. Let $P \in \mathbb{R}^{n\times n}$ be a projection matrix with $\mathsf{rank}(P) = \mathsf{Tr}(P) = m < n$. Then

$$(X - \mu)^T P (X - \mu) \sim \chi_m^2.$$

You can prove the above result using the decomposition in Section 8.1.3.

IID normals. Suppose $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$ form an IID random sample. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ be the sample mean and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ be the sample variance. Then we have the following results:

- $\bar{X}_n$ and S_n^2 are independent.
- $\bar{X}_n \sim N(\mu, \sigma^2/n)$.
- $(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$.

The above results are based on the following insight. Let $X = (X_1, \cdots, X_n)^T$ be a multivariate normal formed by the IID elements. Let $e_n = \frac{1}{\sqrt{n}}(1, 1, \cdots, 1)^T$ be a unit vector. Define two projection matrices $P = e_n e_n^T$ and $Q = \mathbf{I}_n - e_n e_n^T$. One can easily see that $PQ = QP = 0$ so the two projection matrices are orthogonal. This, together with the fact that $\mathsf{Cov}(X) = \sigma^2 \mathbf{I}_n$, implies that PX and QX are independent. Moreover, one can easily see that $\bar{X}_n$ is a function of PX and S_n^2 is a function of QX so they are independent. The last assertion is based on the fact that $S_n^2 = \frac{1}{n-1}[QX]^T QX$.