Lecture Notes 2: The Multivariate Normal Distributions

Professor: Zhihua Zhang *Scribe:Cheng Chen, Luo Luo, Cong Xie*

1.2.3 Sample Mean and Covariance Matrix

Definition 1.10. *Sample mean is* $\overline{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$ *in which* $\mathbb{E}(\overline{\mathbf{x}}) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(\mathbf{x}_j) = \boldsymbol{\mu}$

Definition 1.11. *Sample covariance is*

$$
\begin{aligned}
\mathbf{S} &= \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{x}_i - \overline{\mathbf{x}})(\mathbf{x}_i - \overline{\mathbf{x}})^T \\
&= \frac{1}{n-1}\mathbf{X}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X} \\
&= \frac{1}{n-1}\mathbf{X}^T(\mathbf{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^T)\mathbf{X} \\
&= \frac{1}{n-1}\mathbf{X}^T\mathbf{H}_n\mathbf{X}
\end{aligned}
$$

where $\mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix}$ *and* $\mathbf{H}_n$ *is the centering matrix. And we can prove that* $\mathbb{E}(\mathbf{S}) = \boldsymbol{\Sigma}$.

Definition 1.12. *With metric* $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ *is p.d, the Mahalanbis distance between* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^p$, *is the square root of* $\triangle^2(\mathbf{x}, \mathbf{y}) = (\mathbf{x}-\mathbf{y})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\mathbf{y})$. *If the* $\boldsymbol{\Sigma} = \mathbf{I}$, *then the resulting distance measure is Euclidean distance.*

For given transformation

$$
\begin{aligned}
\hat{\mathbf{x}} &= \mathbf{B}\mathbf{x} + \mathbf{b} \\
\hat{\mathbf{y}} &= \mathbf{B}\mathbf{y} + \mathbf{b} \\
\hat{\boldsymbol{\Sigma}} &= \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T,
\end{aligned}
$$

where $\mathbf{B}$ is non-singular, the Mahalanbis distance is invariant:

$$
(\mathbf{x} - \mathbf{y})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \mathbf{y}) = (\hat{\mathbf{x}} - \hat{\mathbf{y}})^T\hat{\boldsymbol{\Sigma}}^{-1}(\hat{\mathbf{x}} - \hat{\mathbf{y}}).
$$

If $\mathbf{B}$ is singular, we can use pseudo inverse to replace inverse.

1.2.4 Conditional Expectation

Definition 1.13. *For random variable* X *and* Y, *the conditional expectation is*

$$
\mathbb{E}(X|Y = y) = \begin{cases} \sum_x x f_{X|Y}(x|y) & \textit{for discrete variable} \\ \int x f_{X|Y}(x|y)dx & \textit{for continous variable} \end{cases}.
$$

If $g(x,y)$ is function of x and y then

$$\mathbb{E}(g(X,Y)|Y=y) = \begin{cases} \sum_x g(x,y) f_{X|Y}(x|y) & \text{for discrete variable} \\ \int g(x,y) f_{X|Y}(x|y) dx & \text{for continous variable} \end{cases}.$$

- $\mathbb{E}(\mathbf{x})$ is a constant;
- $\mathbb{E}(X|Y=y)$ is a function with respect to y;
- $\mathbb{E}(X|Y)$ is a random variable with respect to Y.

Example 1.3. *Given $X \sim U(0,1)$, after we observe $X=x$, we draw $Y|X=x \sim U(x,1)$. Then $f(y|x) = \dfrac{1}{1-x}$ for $x<y<1$, $\mathbb{E}(Y|X=x) = \displaystyle\int_x^1 y \frac{1}{1-x} dy = \frac{1+x}{2}$, and $\mathbb{E}(Y|X) = \dfrac{1+X}{2}$.*

Theorem 1.2. *The rule of iterated expectation. Assume X and Y's expectation exist, then $\mathbb{E}(\mathbb{E}(Y|X)) = \mathbb{E}(Y)$ and $\mathbb{E}(\mathbb{E}(X|Y)) = \mathbb{E}(X)$. More generally, for $g(x,y)$, $\mathbb{E}(\mathbb{E}(g(X,Y)|X)) = \mathbb{E}(g(X,Y))$. **(homework)***

Definition 1.14. *The conditional variance is $\mathrm{Var}(Y|X=x) = \int (y-\mu(x))^2 f(y|x) dy$, where $\mu(x) = \mathbb{E}(Y|X=x)$.*

Theorem 1.3. *X and Y is random variable, $\mathrm{Var}(Y) = \mathbb{E}(\mathrm{Var}(Y|X)) + \mathrm{Var}(\mathbb{E}(Y|X)) \Rightarrow \mathrm{Var}(Y) \geq \mathrm{Var}(\mathbb{E}(Y|X))$. **(homework)***

1.3 Moment Generating Function

Definition 1.15. *The Moment Generating Function (MGF) of X with CDF $F_X(s)$ is $\psi_X(t) \triangleq \mathbb{E}(e^{tX}) = \int e^{tx} dF_X(x)$.*

- $\mathbb{E}(e^{-tX}) = \int e^{-tx} dF_X(x)$ is Laplace transformation.
- $\mathbb{E}(e^{itX}) = \int e^{itx} dF_X(x)$ is characteristic function.
- $\psi'_X(t) = \int x e^{tx} dF_X(x)$, $\psi(0) = \int x dF(x) = \mathbb{E}(x)$, $\mathbb{E}(x^k) = \psi^{(k)}(0)$.

Theorem 1.4. *The function ψ on $(0,+\infty)$ is the Laplace transform of CDF F iff it is completely monotone and $\lim\limits_{\lambda \to 0} \phi(\lambda) = 1$.*

Definition 1.16. *L is completely monotone if $L^{(n)}$ exists and $(-1)^n L^{(n)}(\lambda) \geq 0$, $\lambda > 0$.*

Example 1.4.

$$\frac{1}{1+\lambda} = \int_0^{+\infty} e^{-\lambda x} e^{-x} dx$$

$$\exp(-\lambda) = \int_0^{+\infty} e^{-\lambda x} \delta(x) dx$$

2 The Multivariate Normal Distributions

$X = (X_1, \ldots, X_m)^T$ with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^T$ and $\boldsymbol{\Sigma}$ is $p \times p$ positive definite, its p.d.f of multivariate normal distributions is

$$\mathcal{N}_m(X = \mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{\frac{m}{2}}|\boldsymbol{\Sigma}|^{\frac{1}{2}}} \exp\{\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\}.$$

We have $\mathbb{E}(X) = \boldsymbol{\mu}$, $\mathrm{Var}(X) = \boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}^{-1}$ is concentration matrix (precision matrix).

Theorem 2.1. *If $X \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{b}$ is $k \times 1$ and $\mathbf{B}$ is an $k \times m$ matrix such that $\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T$ is non-singular, (imply $k \leq m$ and $\mathbf{B}$ is full rank) then $Y = \mathbf{B}X + \mathbf{b} \sim \mathcal{N}_k(\mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^T)$. And $Z = \boldsymbol{\Sigma}^{-\frac{1}{2}}(X - \boldsymbol{\mu}) \sim \mathcal{N}(0, \mathbf{I})$.*

Let $\mathbf{x} = \begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{bmatrix}$, $\boldsymbol{\Sigma}_{22.1} = \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}$ where $\mathbf{x}_1, \boldsymbol{\mu}_1 \in \mathbb{R}^p$, $\mathbf{x}_2, \boldsymbol{\mu}_2 \in \mathbb{R}^q$, $\boldsymbol{\Sigma}_{11} \in \mathbb{R}^{p\times p}$, $\boldsymbol{\Sigma}_{12} \in \mathbb{R}^{p\times q}$, $\boldsymbol{\Sigma}_{21} \in \mathbb{R}^{q\times p}$, $\boldsymbol{\Sigma}_{22} \in \mathbb{R}^{q\times q}$ and $m = p + q$.

Lemma 2.1. *If $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{x}$ are independent iff $\mathbf{A}\boldsymbol{\Sigma}\mathbf{B}^T = 0$.*

Theorem 2.2. *If $\mathbf{x} \sim \mathcal{N}_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\boldsymbol{\Sigma} \succ 0$, then $\mathbf{x}_1$ and $\mathbf{x}_{2.1} = \mathbf{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1$ are statistically independent and*

$$\mathbf{x}_1 \sim \mathcal{N}_p(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}), \quad \mathbf{x}_{2.1} \sim \mathcal{N}_q(\boldsymbol{\mu}_{2.1}, \boldsymbol{\Sigma}_{22.1}),$$

where $\boldsymbol{\mu}_{2.1} = \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1$.

Proof. Let $\mathbf{B}_1 = [\mathbf{I}_p, 0]$, then $\mathbf{B}_1\mathbf{x} = \mathbf{x}_1$ and $\mathbf{B}_1\boldsymbol{\Sigma}\mathbf{B}_1^T = \boldsymbol{\Sigma}_{11}$. Let $\mathbf{B}_2 = [-\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}, \mathbf{I}_q]$, then $\mathbf{x}_{2.1} = \mathbf{x}_2 - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\mathbf{x}_1 = \mathbf{B}_2\mathbf{x}$ and $\mathbf{B}_2\boldsymbol{\Sigma}\mathbf{B}_2^T = \boldsymbol{\Sigma}_{22.1}$. $\mathbf{x}_1$ and $\mathbf{x}_{2.1}$ are independent by lemma 2.1, $\mathbf{x}_1 \perp\!\!\!\perp \mathbf{x}_{2.1}$. □

We also have $p(\mathbf{x}) = p(\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix}) = p(\mathbf{x}_1, \mathbf{x}_{2.1}) = p(\mathbf{x}_1)p(\mathbf{x}_{2.1})$.

Proof. By LDU decomposition

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix} = \begin{bmatrix} \mathbf{I}_p & 0 \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_q \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22.1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_p & \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ 0 & \mathbf{I}_q \end{bmatrix}.$$

Then we have

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{11}||\boldsymbol{\Sigma}_{22.1}|$$

and

$$\begin{aligned} \boldsymbol{\Sigma}^{-1} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}^{-1} &= \begin{bmatrix} \mathbf{I}_p & \boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ 0 & \mathbf{I}_q \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\Sigma}_{11} & 0 \\ 0 & \boldsymbol{\Sigma}_{22.1} \end{bmatrix}^{-1} \begin{bmatrix} \mathbf{I}_p & 0 \\ \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_q \end{bmatrix}^{-1} \\ &= \begin{bmatrix} \mathbf{I}_p & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} \\ 0 & \mathbf{I}_q \end{bmatrix} \begin{bmatrix} \boldsymbol{\Sigma}_{11}^{-1} & 0 \\ 0 & \boldsymbol{\Sigma}_{22.1}^{-1} \end{bmatrix} \begin{bmatrix} \mathbf{I}_p & 0 \\ -\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1} & \mathbf{I}_q \end{bmatrix} \end{aligned}$$

Hence, consider exponents of these Normal distribution

$$\begin{aligned}
p(\mathbf{x}) &= \frac{1}{(2\pi)^{\frac{m}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\}, \\
p(\mathbf{x}_1) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}_1|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1)\}, \\
p(\mathbf{x}_{2.1}) &= \frac{1}{(2\pi)^{\frac{q}{2}}|\mathbf{\Sigma}_{22.1}|^{\frac{1}{2}}} \exp\{-\frac{1}{2}(\mathbf{x}_{2.1}-\boldsymbol{\mu}_{22.1})^T\mathbf{\Sigma}_{22.1}^{-1}(\mathbf{x}_{2.1}-\boldsymbol{\mu}_{2.1})\}.
\end{aligned}$$

We can write

$$\begin{aligned}
& (\mathbf{x}-\boldsymbol{\mu})^T\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) \\
= & \begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix}^T \begin{bmatrix}\mathbf{I}_p & -\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} \\ 0 & \mathbf{I}_q\end{bmatrix} \begin{bmatrix}\mathbf{\Sigma}_{11}^{-1} & 0 \\ 0 & \mathbf{\Sigma}_{22.1}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{I}_p & 0 \\ -\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1} & \mathbf{I}_q\end{bmatrix} \begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_2-\boldsymbol{\mu}_2\end{bmatrix} \\
= & \begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ (\mathbf{x}_2-\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{x}_1)-(\boldsymbol{\mu}_2-\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1)\end{bmatrix}^T \begin{bmatrix}\mathbf{\Sigma}_{11}^{-1} & 0 \\ 0 & \mathbf{\Sigma}_{22.1}^{-1}\end{bmatrix} \\
& \begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ (\mathbf{x}_2-\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{x}_1)-(\boldsymbol{\mu}_2-\mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\boldsymbol{\mu}_1)\end{bmatrix} \\
= & \begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_{2.1}-\boldsymbol{\mu}_{2.1}\end{bmatrix}^T \begin{bmatrix}\mathbf{\Sigma}_{11}^{-1} & 0 \\ 0 & \mathbf{\Sigma}_{22.1}^{-1}\end{bmatrix} \begin{bmatrix}\mathbf{x}_1-\boldsymbol{\mu}_1 \\ \mathbf{x}_{2.1}-\boldsymbol{\mu}_{2.1}\end{bmatrix} \\
= & (\mathbf{x}_1-\boldsymbol{\mu}_1)^T\mathbf{\Sigma}_1^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1) + (\mathbf{x}_{2.1}-\boldsymbol{\mu}_{2.1})^T\mathbf{\Sigma}_{22.1}^{-1}(\mathbf{x}_{2.1}-\boldsymbol{\mu}_{2.1}).
\end{aligned}$$

Hence $p(\mathbf{x}) = p(\mathbf{x}_1)p(\mathbf{x}_{2.1})$. □

Theorem 2.3. *The condition distribution has*

$$\mathbf{x}_2|\mathbf{x}_1 \sim \mathcal{N}_q(\boldsymbol{\mu}_2 + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}(\mathbf{x}_1-\boldsymbol{\mu}_1), \mathbf{\Sigma}_{22.1})$$

$$\mathbf{x}_2 = \mathbf{x}_{2.1} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{x}_1, \quad \mathbb{E}(\mathbf{x}_2|\mathbf{x}_1) = \boldsymbol{\mu}_{2.1} + \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{x}_1.$$

Theorem 2.4. *Assume $X = [X_1, \dots X_m]^T \sim \mathcal{N}_m(0, \mathbf{\Sigma})$, $\mathbf{\Sigma} = (\sigma_{ij})$ and $\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = (\theta_{ij})$. Then $X_i \perp\!\!\!\perp X_j$ iff $\sigma_{ij} = 0$ and $X_i \perp\!\!\!\perp X_j | X_{\{1\dots m\}\setminus\{i,j\}}$ iff $\theta_{ij} = 0$.*

(***homework***)Prove $\mathbf{\Sigma}_{11.2} = \mathbf{\Theta}_{11}^{-1}$ and $\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12} = \mathbf{\Theta}_{12}\mathbf{\Theta}_{22}^{-1}$.

Proof. Without loss of generality, consider that $i = 1$ and $j = 2$. Let $\mathbf{y}_1 = [X_1, X_2]^T$ and $\mathbf{y}_2 = [X_3, \dots, X_m]^T$. We have

$$\begin{aligned}
\mathbf{y}_1 &\sim \mathcal{N}(\boldsymbol{\mu}_1, \mathbf{\Sigma}_{11}) \\
\mathbf{y}_1|\mathbf{y}_2 &\sim \mathcal{N}(\boldsymbol{\mu}_1 + \mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1}(\mathbf{y}_2-\boldsymbol{\mu}_2), \mathbf{\Sigma}_{11.2}).
\end{aligned}$$

where the subscript $1, 2$ and 11.2 are with respect to $\mathbf{y}_1$ and $\mathbf{y}_2$. Then we have $X_1 \perp\!\!\!\perp X_2$ iff $\sigma_{12} = 0$. By homework,

$$\mathbf{\Sigma}_{11.2} = \begin{bmatrix}\theta_{11} & \theta_{12} \\ \theta_{21} & \theta_{22}\end{bmatrix}^{-1} = \frac{1}{\theta_{11}\theta_{22} - \theta_{12}\theta_{21}} \begin{bmatrix}\theta_{22} & -\theta_{21} \\ -\theta_{12} & \theta_{11}\end{bmatrix}.$$

Hence we have $X_1 \perp\!\!\!\perp X_2 | X_{\{1\dots m\}\setminus\{1,2\}}$ iff $\theta_{12} = \theta_{21} = 0$ □