MULTIVARIATE STATISTICS, Lesson 2.
Theorems based on multivariate normal distribution

- **Proposition.** Let $(\mathbf{Y}^T, \mathbf{X}^T)^T \sim \mathcal{N}_{q+p}(\mathbf{0}, \mathbf{C})$, where the covariance matrix $\mathbf{C}$ is a hypermatrix:

$$\mathbf{C} = \begin{pmatrix} \mathbf{C}_{22} & \mathbf{C}_{21} \\ \mathbf{C}_{12} & \mathbf{C}_{11} \end{pmatrix}.$$

  Here $\mathbf{C}_{11}$ and $\mathbf{C}_{22}$ are covariance matrices of $\mathbf{X}$ and $\mathbf{Y}$, respectively, while $\mathbf{C}_{21} = \mathbf{C}_{12}^T$ is the cross-covariance matrix. Suppose that $\mathbf{C}_{11}$, $\mathbf{C}_{22}$, and $\mathbf{C}$ are not singular. Then the conditional distribution of $\mathbf{Y}$ conditioned on $\mathbf{X} = \mathbf{x}$ is $\mathcal{N}_q(\mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{x}, \mathbf{C}_{22.1})$-distribution, where

$$\mathbf{C}_{22.1} = \mathbf{C}_{22} - \mathbf{C}_{21}\mathbf{C}_{11}^{-1}\mathbf{C}_{12}$$

- **Lemma (Multidimensional Steiner Theorem).** Let $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^p$ be given vectors, $\overline{\mathbf{x}} = \frac{1}{n}\sum_{k=1}^n \mathbf{x}_k$, and $\mathbf{v} \in \mathbb{R}^p$ be an arbitrary vector. Then

$$\sum_{k=1}^n (\mathbf{x}_k - \mathbf{v})(\mathbf{x}_k - \mathbf{v})^T = \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T + n(\bar{\mathbf{x}} - \mathbf{v})(\bar{\mathbf{x}} - \mathbf{v})^T.$$

  Especially, with $\mathbf{v} = \mathbf{0}$: $\quad \sum_{k=1}^n (\mathbf{x}_k - \bar{\mathbf{x}})(\mathbf{x}_k - \bar{\mathbf{x}})^T = \sum_{k=1}^n \mathbf{x}_k\mathbf{x}_k^T - n\bar{\mathbf{x}}\bar{\mathbf{x}}^T$.

- *Independent sums.* Let $\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n$ be i.i.d. random vectors with expectation vector $\mathbf{m}$ and covariance matrix $\mathbf{C}$. Calculate the expectation vector and covariance matrix of $\overline{\mathbf{X}}$.

- **Multivariate Central Limit Theorem (MCLT).** Let $\mathbf{X}_1, \mathbf{X}_2, \dots$ be i.i.d. p-dimensional random vectors with existing expectation vector $\mathbf{m}$ and (positive semidefinite) covariance matrix $\mathbf{C}$. Let $\mathbf{S}_n := \mathbf{X}_1 + \dots + \mathbf{X}_n$, $n = 1, 2, \dots$. Then for the sequence of the standardized partial sums: $\quad \frac{1}{\sqrt{n}}(\mathbf{S}_n - n\mathbf{m}) \to \mathcal{N}_p(\mathbf{0}, \mathbf{C})$ in distribution as $n \to \infty$.

- *Lemma.* Let $\mathbf{X} = (X_1, \dots, X_k)^T \sim \mathcal{N}_k(\mathbf{0}, \mathbf{C})$ with $\mathbf{C}$ positive semidefinite. Then $\sum_{i=1}^k X_i^2$ can be decomposed as $\sum_{i=1}^k \lambda_i Y_i^2$, where $\mathbf{Y} = (Y_1, \dots, Y_k)^T \sim \mathcal{N}_k(\mathbf{0}, \mathbf{I}_k)$ and the nonnegative real numbers $\lambda_1, \dots, \lambda_k$ are eigenvalues of $\mathbf{C}$.

- *Revisiting the χ^2-test.* Let $A_1, \dots, A_k$ be a complete set of mutually exclusive events. Check

$$H_0 : \mathbb{P}(A_i) = p_i \quad (i = 1, \dots, k).$$

  Denote by $\nu_1, \dots, \nu_k$ the frequencies of $A_1, \dots, A_k$ in n independent trials ($\sum_{i=1}^k \nu_i = n$). Then under the zero-hypothesis

$$\underline{\nu} = (\nu_1, \dots, \nu_k)^T \sim \mathcal{P}oly_n(p_1, \dots, p_k).$$

  (Recall that it is a deformed k-dimensional distribution concentrated on a $(k-1)$-dimensional hyperplane of $\mathbb{R}^k$ because of the linear relation $\nu_1 + \dots + \nu_k = n$ between its components.)

- **Theorem.** If $\underline{\nu} = (\nu_1, \dots, \nu_k)^T$ follows polynomial distribution with parameters n and $p_1, \dots, p_k \quad (p_i > 0, \sum_{i=1}^k p_i = 1)$, then

$$\sum_{i=1}^k \frac{(\nu_i - np_i)^2}{np_i} \to \chi^2(k-1)$$

  in distribution as $n \to \infty$.