Lecture 3 Random Sampling and Multivariate Normal Distribution

Shiwei Lan[1]

[1]School of Mathematical and Statistical Sciences
Arizona State University

STP533 Multivariate Analysis
Spring 2025

Table of Contents

Radom Samples

- Recall the data array $\mathbf{X}$ is arranged as an $n \times p$ matrix

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} X_1^T \\ X_2^T \\ \vdots \\ X_n^T \end{bmatrix} = \begin{bmatrix} X_1 & X_2 & \cdots & X_p \end{bmatrix}$$

- Each row $X_i^T = [X_{i1}, X_{i2}, \cdots, X_{ip}]$ represents a *independent observation* from a joint distribution p-dimensional random vector.
- Each column $X_j = [X_{1j}, X_{2j}, \cdots, X_{nj}]^T$ represents a *random sample* (collection of observations) of a random variable X_j.

Radom Samples

- Random sample is often assumed to be a collection of *independently identically distributed (i.i.d.)* observations.
- Assume the p-dimensional distribution has a density function $f(\mathbf{x}) = f(x_1, x_2, \cdots, x_p)$. We denote random sample $\{X_i\}_{i=1}^n \overset{iid}{\sim} f(\mathbf{x})$.
- For the joint distribution of all the samples, based on the iid assumption, we have

$$f(\mathbf{X}) = \prod_{i=1}^{n} f(\mathbf{x}_i).$$

- Note, in general $f(\mathbf{x}) \neq \prod_{j=1}^{p} f(x_j)$ where each $f(x_j)$ is the marginal density of random variable X_j.

Expectation of Sample Statistics

- Now we assume a random sample $\{X_i\}_{i=1}^n \overset{iid}{\sim} f(\mathbf{x})$ from a joint distribution with mean $\boldsymbol{\mu} \in \mathbb{R}^p$ and covariance $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$.
- Previously we had sample mean $\bar{\mathbf{X}} = \frac{1}{n}\sum_{i=1}^n X_i$ and

$$\mathrm{E}[\bar{\mathbf{X}}] = \boldsymbol{\mu}, \quad \mathrm{Cov}(\bar{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}$$

- Then we have for sample covariance $\mathbf{S}_n = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{\mathbf{X}})(X_i - \bar{\mathbf{X}})^T$

$$\mathrm{E}[\mathbf{S}_n] = \frac{n-1}{n}\boldsymbol{\Sigma}$$

- Therefore, we often consider the unbiased sample covariance matrix $\mathbf{S} = \frac{n}{n-1}\mathbf{S}_n$.

Generalized Variance

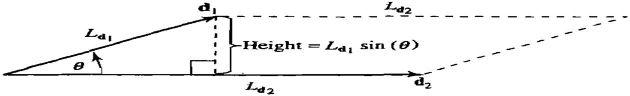
- For p-dimensional random sample $\mathbf{X}_{n\times p}$, the *generalized sample variance* is defined as the determinant of sample covariance $\mathbf{S}$:

$$\text{generalized sample variance} = |\mathbf{S}| = (n-1)^p \mathrm{vol}^2$$

  where vol is the volume generated by p residual (deviation) vectors $\{\mathbf{x}_j - \bar{\mathbf{x}}_j\}_{j=1}^p$.
- It can be shown that $\mathrm{vol}\{\mathbf{x} : (\mathbf{x}-\bar{\mathbf{x}})^T \mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \le c^2\} = k_p |\mathbf{S}|^{\frac{1}{2}} c^p$.
- This quantity measures the variability of the random sample of size n.
- It can be used to detect *multi-colinearity*, i.e. $X_1, X_2, \cdots X_p$ are linearly dependent when $|\mathbf{S}| = 0$.
- If $n \le p$, then $|\mathbf{S}| = 0$ for all samples.

Generalized Variance

Consider the area generated within the plane by two deviation vectors $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1}$ and $\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1}$. Let $L_{\mathbf{d}_1}$ be the length of $\mathbf{d}_1$ and $L_{\mathbf{d}_2}$ the length of $\mathbf{d}_2$. By elementary geometry, we have the diagram



and the area of the trapezoid is $|L_{\mathbf{d}_1}\sin(\theta)|L_{\mathbf{d}_2}$. Since $\cos^2(\theta) + \sin^2(\theta) = 1$, we can express this area as

$$\text{Area} = L_{\mathbf{d}_1}L_{\mathbf{d}_2}\sqrt{1 - \cos^2(\theta)}$$

From (3-5) and (3-7),

$$L_{\mathbf{d}_1} = \sqrt{\sum_{j=1}^{n}(x_{j1} - \bar{x}_1)^2} = \sqrt{(n-1)s_{11}}$$

$$L_{\mathbf{d}_2} = \sqrt{\sum_{j=1}^{n}(x_{j2} - \bar{x}_2)^2} = \sqrt{(n-1)s_{22}}$$

and

$$\cos(\theta) = r_{12}$$

Therefore,

$$\text{Area} = (n-1)\sqrt{s_{11}}\sqrt{s_{22}}\sqrt{1 - r_{12}^2} = (n-1)\sqrt{s_{11}s_{22}(1 - r_{12}^2)} \qquad (3\text{-}13)$$

Also,

$$|\mathbf{S}| = \left|\begin{bmatrix} s_{11} & s_{12} \\ s_{12} & s_{22}\end{bmatrix}\right| = \left|\begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}r_{12} & s_{22}\end{bmatrix}\right|$$

$$= s_{11}s_{22} - s_{11}s_{22}r_{12}^2 = s_{11}s_{22}(1 - r_{12}^2) \qquad (3\text{-}14)$$

Generalized Variance

$$\mathbf{X} = \begin{bmatrix} 1 & 9 & 10 \\ 4 & 12 & 16 \\ 2 & 10 & 12 \\ 5 & 8 & 13 \\ 3 & 11 & 14 \end{bmatrix}$$

Other Generalization of Variance

- Consider the *generalized sample variance of the standardized variables*

$$|\mathbf{R}| = (n-1)^p \mathrm{vol}^2$$

where vol is the volume generated by p standardized vectors $\left\{ \frac{\mathbf{x}_j - \bar{\mathbf{x}}_j}{\sqrt{s_{jj}}} \right\}_{j=1}^p$.

- What is the relationship between $|\mathbf{R}|$ and $|\mathbf{S}|$?

- Another generalization of variance is *total sample variance* defined as $\mathrm{tr}(\mathbf{S})$.

Sample Statistics of Linear Combinations of Variables

- Recall we had the following matrix representation of sample statistics:

$$\bar{\mathbf{X}} = \frac{\mathbf{1}_n^T}{n}\mathbf{X}, \quad \mathbf{S} = \frac{1}{n-1}\mathbf{X}^T(\mathbf{I}_n - \mathbf{J})\mathbf{X}, \quad \mathbf{J} = \frac{\mathbf{1}\mathbf{1}_n^T}{n}$$

- Now suppose we have two linear combinations $\mathbf{Xb}$ and $\mathbf{Xc}$. Then we have

$$\overline{\mathbf{Xb}} = \bar{\mathbf{X}}\mathbf{b}, \quad s_{\mathbf{Xb},\mathbf{Xc}} = \mathbf{b}^T\mathbf{S}\mathbf{c}$$

- For example, $\mathbf{X} = \begin{bmatrix} 1 & 2 & 5 \\ 4 & 1 & 6 \\ 4 & 0 & 4 \end{bmatrix}$, $\mathbf{b} = [2, 2, -1]^T$ and $\mathbf{c} = [1, -1, 3]^T$.

Table of Contents

Multivariate Normal Density

- The read data are not exactly multivariate normal, but normal density can serve as a good approximation.
- The density of multivariate normal random vector $\mathbf{X} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is

$$f(\mathbf{x}) = (2\pi)^{-\frac{p}{2}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}} \exp\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$$

where the covariance matrix $\boldsymbol{\Sigma}$ is PSD.

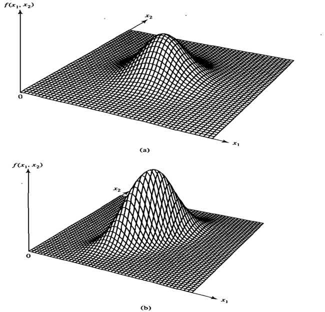


Figure 4.2 Two bivariate normal distributions. (a) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = 0$. (b) $\sigma_{11} = \sigma_{22}$ and $\rho_{12} = .75$.

Elliptic Contours of MVN Density

- The contour of MVN density is determined by

$$(\mathbf{x}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2$$

- This is an ellipsoid centered at $\boldsymbol{\mu}$ and having axes $\pm\sqrt{\lambda_i}\mathbf{v}_i$ with eigen-paris $\{\lambda_i, \mathbf{v}_i\}$ of $\boldsymbol{\Sigma}$.

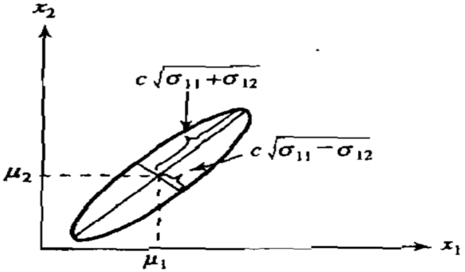


Figure 4.3 A constant-density contour for a bivariate normal distribution with $\sigma_{11} = \sigma_{22}$ and $\sigma_{12} > 0$ (or $\rho_{12} > 0$).

Properties of MVN

- The linear combination of MVN is another MVN. Let $\mathbf{A} \in \mathbb{R}^{q\times p}$. Then

$$\mathbf{AX} \sim N_q(\mathbf{A}\boldsymbol{\mu}, \mathbf{A\Sigma A}^T)$$

- The marginal of MVN is also MVN. Consider $\mathbf{A} = \begin{bmatrix}\mathbf{I}_q & \mathbf{0}\end{bmatrix}$.
- The conditional pdf of MVN is also MVN. Let $\mathbf{X} = \begin{bmatrix}\mathbf{X}_1 \\ \mathbf{X}_2\end{bmatrix}$, $\boldsymbol{\mu} = \begin{bmatrix}\boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2\end{bmatrix}$, $\boldsymbol{\Sigma} = \begin{bmatrix}\boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22}\end{bmatrix}$. Then

$$\mathbf{X}_1|\mathbf{X}_2 = \mathbf{x}_2 \sim N_{p_1}(\boldsymbol{\mu}_1 - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}(\mathbf{x}_2 - \boldsymbol{\mu}_2), \boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21})$$

- Note $\mathbf{X}_1 \perp \mathbf{X}_2$ if and only if $\boldsymbol{\Sigma}_{12} = 0$. What is the caveat?
- What is the distribution of $(\mathbf{x} - \boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})$?

The Multivariate Normal Likelihood

- Consider the random sample $\mathbf{X}_{n\times p}$. The *likelihood* of the sample is the joint density

$$L_{\mathbf{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \prod_{i=1}^n f(\mathbf{x}_i) = (2\pi)^{-\frac{np}{2}} |\boldsymbol{\Sigma}|^{-\frac{n}{2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) \right\}$$

- We notice the sum of quadratic form can be rewritten as

$$\sum_{i=1}^n (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu}) = \mathrm{tr}\left[(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right]$$

- The maximum likelihood estimation (MLE) is to maximize the following log-likelihood with respect to $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$:

$$\ell_{\mathbf{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \log L_{\mathbf{X}}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{tr} \left[(\mathbf{X} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{X} - \boldsymbol{\mu}) \right]$$

- Setting $\frac{\partial \ell}{\partial \boldsymbol{\mu}} = 0$ and $\frac{\partial \ell}{\partial \boldsymbol{\Sigma}} = 0$, we obtain the MLE for $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ as

$$\hat{\mu} = \bar{\mathbf{X}}, \quad \hat{\boldsymbol{\Sigma}} = \frac{n-1}{n} \mathbf{S}$$

- Note that $\bar{\mathbf{X}}$ and $\mathbf{S}$ are also *sufficient statistics*.

The Sampling Distribution of $\bar{\mathbf{X}}$ and $\mathbf{S}$

Suppose $\mathbf{X}_i \overset{iid}{\sim} N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then we have

1. $\bar{\mathbf{X}} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}/n)$.
2. $(n-1)\mathbf{S} \sim W_{n-1}(\boldsymbol{\Sigma})$, Wishart distribution with degree of freedom $n-1$.
3. $\bar{\mathbf{X}} \perp \mathbf{S}$.

Definition (Wishart distribution)

A square matrix $\mathbf{A} \sim W_m(\boldsymbol{\Sigma})$ Wishart distribution with degree of freedom m if it can be expressed as $\mathbf{A} = \sum_{j=1}^m \mathbf{Z}\mathbf{Z}^T$, where $\mathbf{Z}_j \overset{iid}{\sim} Np(\mathbf{0}, \boldsymbol{\Sigma})$. The density of $\mathbf{A}$ is

$$f_m(\mathbf{A}|\boldsymbol{\Sigma}) = \frac{|\mathbf{A}|^{(m-p-1)/2} \exp\{-\mathrm{tr}(\mathbf{A}\boldsymbol{\Sigma}^{-1})/2\}}{2^{pm/2}\pi^{p(p-1)/4}|\boldsymbol{\Sigma}|^{m/2}\prod_{i=1}^p \Gamma((m+1-i)/2)}$$

- If $\mathbf{A}_1 \sim W_{m_1}(\boldsymbol{\Sigma})$ and $\mathbf{A}_2 \sim W_{m_2}(\boldsymbol{\Sigma})$, then $\mathbf{A}_1 + \mathbf{A}_2 \sim W_{m_1+m_2}(\boldsymbol{\Sigma})$.
- If $\mathbf{A} \sim W_m(\boldsymbol{\Sigma})$, then $\mathbf{C}\mathbf{A}\mathbf{C}^T \sim W_m(\mathbf{C}\boldsymbol{\Sigma}\mathbf{C}^T)$.

Large-Sample Behavior $\bar{\mathbf{X}}$ and S

Suppose $\mathbf{X}_i \overset{iid}{\sim} (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then we have

LLN $\bar{\mathbf{X}} \xrightarrow{P} \boldsymbol{\mu}$, i.e. for any $\epsilon > 0$, $P[|\bar{\mathbf{X}} - \boldsymbol{\mu}| > \epsilon] \to 0$ as $n \to \infty$.

CLT $\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \xrightarrow{L} N_p(\mathbf{0}, \boldsymbol{\Sigma})$, i.e $P[\sqrt{n}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \leq \mathbf{x}] \to p_N(\mathbf{x}; \mathbf{0}, \boldsymbol{\Sigma})$ as $n \to \infty$.

- We also have $n(\bar{\mathbf{X}} - \boldsymbol{\mu})^T \mathbf{S}^{-1}(\bar{\mathbf{X}} - \boldsymbol{\mu}) \dot{\sim} \chi^2_p$.

Evaluating Normality: `qqnorm`, `qqplot`

- For univariate normality:

  The steps leading to a Q–Q plot are as follows:

  1. Order the original observations to get $x_{(1)}, x_{(2)}, \ldots, x_{(n)}$ and their corresponding probability values $\left(1 - \frac{1}{2}\right)/n, \left(2 - \frac{1}{2}\right)/n, \ldots, \left(n - \frac{1}{2}\right)/n$;
  2. Calculate the standard normal quantiles $q_{(1)}, q_{(2)}, \ldots, q_{(n)}$; and
  3. Plot the pairs of observations $(q_{(1)}, x_{(1)}), (q_{(2)}, x_{(2)}), \ldots, (q_{(n)}, x_{(n)})$, and examine the "straightness" of the outcome.

- For bivariate normality:

  To construct the chi-square plot,

  1. Order the squared distances in (4-32) from smallest to largest as $d^2_{(1)} \le d^2_{(2)} \le \cdots \le d^2_{(n)}$.
  2. Graph the pairs $\left(q_{c,p}\left(\left(j - \frac{1}{2}\right)/n\right), d^2_{(j)}\right)$, where $q_{c,p}\left(\left(j - \frac{1}{2}\right)/n\right)$ is the $100\left(j - \frac{1}{2}\right)/n$ quantile of the chi-square distribution with p degrees of freedom.