Sample Geometry and Random Sampling

Shyh-Kang Jeng
Department of Electrical Engineering/
Graduate Institute of Communication/
Graduate Institute of Networking and Multimedia

Outline

- The Geometry of the Sample
- Random Samples and the Expected Values of the Sample Mean and Covariance Matrix
- Generalized Variance
- Sample Mean, Covariance, and Correlation as Matrix Operations
- Sample Values of Linear Combinations of Variables

Outline

- The Geometry of the Sample
- Random Samples and the Expected Values of the Sample Mean and Covariance Matrix
- Generalized Variance
- Sample Mean, Covariance, and Correlation as Matrix Operations
- Sample Values of Linear Combinations of Variables

Questions

- How to represent a sample of size *n* from a *p*-variate population?
- What is the geometrical representation of sample mean and deviation?
- How to calculate lengths and angles of deviation vectors?
- What is the geometric meaning of the correlation coefficient?

Array of Data

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix}$$

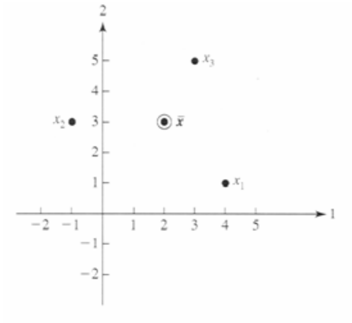
***a sample of size *n* from a *p*-variate population**

Row-Vector View

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{x}_1' \\ \mathbf{x}_2' \\ \vdots \\ \mathbf{x}_j' \\ \vdots \\ \mathbf{x}_n' \end{bmatrix}$$

Example 3.1

$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$



7

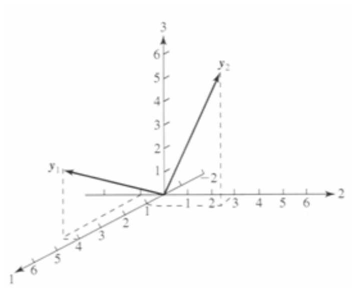
Column-Vector View

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1k} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2k} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{j1} & x_{j2} & \cdots & x_{jk} & \cdots & x_{jp} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nk} & \cdots & x_{np} \end{bmatrix} = [\mathbf{y}_1 \,|\, \mathbf{y}_2 \,|\, \cdots \,|\, \mathbf{y}_p]$$

8

Example 3.2

$$X = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$



Fi
X

9

Geometrical Interpretation of Sample Mean and Deviation

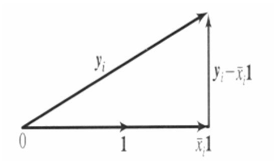
$$\mathbf{1}' = [1, 1, \cdots, 1]$$

$$(\mathbf{y}_i' \frac{1}{\sqrt{n}} \mathbf{1}) \frac{1}{\sqrt{n}} \mathbf{1} = \frac{x_{1i} + x_{2i} + \cdots + x_{ni}}{n} \mathbf{1} = \bar{x}_i \mathbf{1}$$

$$\bar{x}_i = \mathbf{y}_i' \frac{1}{n} \mathbf{1}$$

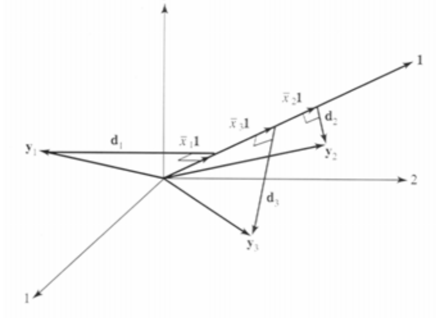
$$\mathbf{d}_i = \mathbf{y}_i - \bar{x}_i \mathbf{1}$$

$$= \begin{bmatrix} x_{1i} - \bar{x}_i \\ x_{2i} - \bar{x}_i \\ \vdots \\ x_{ni} - \bar{x}_i \end{bmatrix}$$



10

Decomposition of Column Vectors



11

Example 3.3

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}, \bar{x}_1 = 2, \bar{x}_2 = 3$$

$$\bar{x}_1 \mathbf{1} = 2[1, 1, 1]' = [2, 2, 2]'$$

$$\bar{x}_2 \mathbf{1} = 3[1, 1, 1]' = [3, 3, 3]'$$

$$\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1 \mathbf{1} = [4, -1, 3]' - [2, 2, 2]' = [2, -3, 1]'$$

$$\mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2 \mathbf{1} = [1, 3, 5]' - [3, 3, 3]' = [-2, 0, 2]'$$

12

Lengths and Angles of Deviation Vectors

$$L_{\mathbf{d}_i}^2 = \mathbf{d}_i'\mathbf{d}_i = \sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2 = ns_{ii}$$

$$\mathbf{d}_i'\mathbf{d}_k = \sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)\left(x_{jk} - \bar{x}_k\right) = ns_{ik}$$

$$= L_{\mathbf{d}_i} L_{\mathbf{d}_k} \cos\theta_{ik}$$

$$= \sqrt{\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2}\sqrt{\sum_{j=1}^{n}\left(x_{ji} - \bar{x}_i\right)^2}\cos\theta_{ik}$$

$$\cos\theta_{ik} = \frac{s_{ik}}{\sqrt{s_{ii}}\sqrt{s_{kk}}} = r_{ik}$$

Example 3.4

$$\mathbf{X} = \begin{bmatrix} 4 & 1 \\ -1 & 3 \\ 3 & 5 \end{bmatrix}$$

$$\mathbf{d}_1 = [2, -3, 1]', \quad \mathbf{d}_2 = [-2, 0, 2]'$$

$$\mathbf{d}_1'\mathbf{d}_1 = 14 = 3s_{11}, \quad \mathbf{d}_2'\mathbf{d}_2 = 8 = 3s_{22}$$

$$\mathbf{d}_1'\mathbf{d}_2 = -2 = 3s_{12}$$

$$r_{12} = \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} = -0.189$$

$$\mathbf{S}_n = \begin{bmatrix} 14/3 & -2/3 \\ -2/3 & 8/3 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & -0.189 \\ -0.189 & 1 \end{bmatrix}$$

Outline

- The Geometry of the Sample
- Random Samples and the Expected Values of the Sample Mean and Covariance Matrix
- Generalized Variance
- Sample Mean, Covariance, and Correlation as Matrix Operations
- Sample Values of Linear Combinations of Variables

Questions

- What are random samples?
- What is the geometric interpretation of randomness?
- Result 3.1

Random Matrix

$$\mathbf{X} = \begin{bmatrix} X_{11} & X_{12} & \cdots & X_{1p} \\ X_{21} & X_{22} & \cdots & X_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ X_{n1} & X_{n2} & \cdots & X_{np} \end{bmatrix} = \begin{bmatrix} \mathbf{X}_1' \\ \mathbf{X}_2' \\ \vdots \\ \mathbf{X}_n' \end{bmatrix}$$

Random Sample

- Row vectors $\mathbf{X}_1'$, $\mathbf{X}_2'$, …, $\mathbf{X}_n'$ represent independent observations from a common joint distribution with density function $f(\mathbf{x})=f(x_1, x_2, \ldots, x_p)$
- Mathematically, the joint density function of $\mathbf{X}_1'$, $\mathbf{X}_2'$, …, $\mathbf{X}_n'$ is

Random Sample

- Measurements of a single trial, such as $\mathbf{X}_j' = [X_{j1}, X_{j2}, \ldots, X_{jp}]$, will usually be correlated
- The measurements from different trials must be independent
- The independence of measurements from trial to trial may not hold when the variables are likely to drift over time

19

Geometric Interpretation of Randomness

- Column vector $\mathbf{Y}_k' = [X_{1k}, X_{2k}, \ldots, X_{nk}]$ regarded as a point in n dimensions
- The location is determined by the joint probability distribution $f(\mathbf{y}_k) = f(x_{1k}, x_{2k}, \ldots, x_{nk})$
- For a random sample, $f(\mathbf{y}_k) = f_k(x_{1k}) f_k(x_{2k}) \ldots f_k(x_{nk})$
- Each coordinate x_{jk} contributes equally to the location through the same marginal distribution $f_k(x_{jk})$

20

Result 3.1

$\mathbf{X}_1, \mathbf{X}_2, \cdots, \mathbf{X}_n$ are a random sample from a joint distribution that has mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, then

$$E(\overline{\mathbf{X}}) = \boldsymbol{\mu}, \quad (\overline{\mathbf{X}} \text{ as an unbiased point estimate of } \boldsymbol{\mu})$$

$$\text{Cov}(\overline{\mathbf{X}}) = \frac{1}{n}\boldsymbol{\Sigma}, \quad E(\mathbf{S}_n) = \frac{n-1}{n}\boldsymbol{\Sigma}$$

$$E\left(\frac{n}{n-1}\mathbf{S}_n\right) = \boldsymbol{\Sigma}$$

$$\left(\mathbf{S} = \frac{n}{n-1}\mathbf{S}_n \text{ as an unbiased point estimate of } \boldsymbol{\Sigma}\right)$$

21

Proof of Result 3.1

$$E(\overline{\mathbf{X}}) = E\left(\frac{1}{n}\mathbf{X}_1 + \frac{1}{n}\mathbf{X}_2 + \cdots + \frac{1}{n}\mathbf{X}_n\right)$$

$$= \frac{1}{n}E(\mathbf{X}_1) + \frac{1}{n}E(\mathbf{X}_2) + \cdots + \frac{1}{n}E(\mathbf{X}_n) = \boldsymbol{\mu}$$

$$(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})' = \left(\frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})\right)\left(\frac{1}{n}\sum_{\ell=1}^{n}(\mathbf{X}_\ell - \boldsymbol{\mu})\right)'$$

$$= \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_\ell - \boldsymbol{\mu})'$$

$$\text{Cov}(\overline{\mathbf{X}}) = E(\overline{\mathbf{X}} - \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu})' = \frac{1}{n^2}\sum_{j=1}^{n}\sum_{\ell=1}^{n}E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_\ell - \boldsymbol{\mu})'$$

22

Proof of Result 3.1

$E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_\ell - \boldsymbol{\mu}) = 0$ for $j \neq \ell$ because of independence.

$$\text{Cov}(\overline{\mathbf{X}}) = \frac{1}{n^2}\sum_{j=1}^{n}E(\mathbf{X}_j - \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu})' = \frac{1}{n^2}n\boldsymbol{\Sigma} = \frac{1}{n}\boldsymbol{\Sigma}$$

$$E(\mathbf{S}_n) = E\left(\frac{1}{n}\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'\right)$$

$$\sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})(\mathbf{X}_j - \overline{\mathbf{X}})'$$

$$= \sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})\mathbf{X}_j' - \sum_{j=1}^{n}(\mathbf{X}_j - \overline{\mathbf{X}})\overline{\mathbf{X}}' = \sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\overline{\mathbf{X}}\overline{\mathbf{X}}'$$

$$E(\mathbf{X}_j\mathbf{X}_j') = E\left((\mathbf{X}_j - \boldsymbol{\mu} + \boldsymbol{\mu})(\mathbf{X}_j - \boldsymbol{\mu} + \boldsymbol{\mu})'\right) = \boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'$$

23

Proof of Result 3.1

$$E(\overline{\mathbf{X}}\overline{\mathbf{X}}') = E\left((\overline{\mathbf{X}} - \boldsymbol{\mu} + \boldsymbol{\mu})(\overline{\mathbf{X}} - \boldsymbol{\mu} + \boldsymbol{\mu})'\right) = \frac{1}{n}\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}'$$

$$E(\mathbf{S}_n) = \frac{1}{n}E\left(\sum_{j=1}^{n}\mathbf{X}_j\mathbf{X}_j' - n\overline{\mathbf{X}}\overline{\mathbf{X}}'\right)$$

$$= \frac{1}{n}\left(n(\boldsymbol{\Sigma} + \boldsymbol{\mu}\boldsymbol{\mu}') - n\left(\frac{1}{n}\boldsymbol{\Sigma} - \boldsymbol{\mu}\boldsymbol{\mu}'\right)\right) = \frac{n-1}{n}\boldsymbol{\Sigma}$$

24

Some Other Estimators

The expectation of the $(i,k)th$ entry of $\frac{n}{n-1}\mathbf{S}_n$

$$E(\frac{n}{n-1}s_{ik}) = E(\frac{1}{n-1}\sum_{j=1}^{n}\left(X_{ji}-\overline{X}_i\right)\left(X_{jk}-\overline{X}_k\right)) = \sigma_{ik}$$

$$E(\sqrt{s_{ii}}) \neq \sqrt{\sigma_{ii}}, \quad E(r_{ik}) \neq \rho_{ik}$$

Biases $E(\sqrt{s_{ii}}) - \sqrt{\sigma_{ii}}$ and $E(r_{ik}) - \rho_{ik}$ can usually be ignored if size n is moderately large

Outline

- The Geometry of the Sample
- Random Samples and the Expected Values of the Sample Mean and Covariance Matrix
- Generalized Variance
- Sample Mean, Covariance, and Correlation as Matrix Operations
- Sample Values of Linear Combinations of Variables

Questions

- How to define a generalized sample variance?
- What is the geometric interpretation of a generalized sample variance for bivariate cases?
- What is the geometric interpretaion of a generalized sample variance for multivariate cases?

Questions

- What is the equation for points within a constant statistical distance c from the sample mean?
- Example 3.8
- Result 3.2
- Example 3.9
- Examples causing zero generalized variance

Questions

- Example 3.10
- Result 3.3
- Result 3.4
- Generalized Sample Variance of Standardized Variables
- Example 3.11
- Total Sample Variance

Generalized Sample Variance

Generalized Sample Variance $= |\mathbf{S}|$

Example 3.7 : Employees and profits per employee for 16 largest publishing firms in US

$$\mathbf{S} = \begin{bmatrix} 252.04 & -68.43 \\ -68.43 & 123.67 \end{bmatrix}$$

$$|\mathbf{S}| = 26.487$$

Geometric Interpretation for Bivariate Case

Area generated by two deviation vectors $\mathbf{d}_1 = \mathbf{y}_1 - \bar{x}_1\mathbf{1}, \quad \mathbf{d}_2 = \mathbf{y}_2 - \bar{x}_2\mathbf{1}$

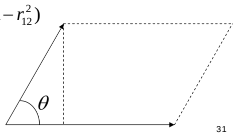
is $area = L_{d_1}L_{d_2}\sin\theta = L_{d_1}L_{d_2}\sqrt{1-\cos^2\theta}$

$$L_{d_1} = \sqrt{\sum_{j=1}^{n}\left(x_{j1}-\bar{x}_1\right)^2} = \sqrt{(n-1)s_{11}}, \quad L_{d_2} = \sqrt{\sum_{j=1}^{n}\left(x_{j2}-\bar{x}_2\right)^2} = \sqrt{(n-1)s_{22}}$$

$$\cos\theta = r_{ik}, \quad area = (n-1)\sqrt{s_{11}s_{22}(1-r_{12}^2)}$$

$$|\mathbf{S}| = \left|\begin{bmatrix} s_{11} & \sqrt{s_{11}}\sqrt{s_{22}}r_{12} \\ \sqrt{s_{11}}\sqrt{s_{22}}r_{12} & s_{22} \end{bmatrix}\right| = s_{11}s_{22}(1-r_{12}^2)$$

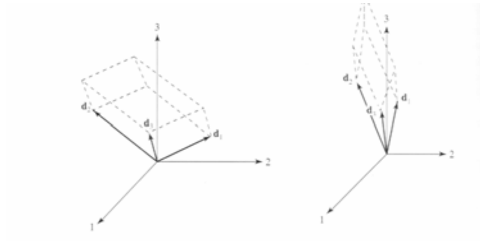
$$= (area)^2/(n-1)^2$$



31

Generalized Sample Variance for Multivariate Cases

$$|\mathbf{S}| = (n-1)^{-p}(volume)^2$$



32

Interpretation in *p*-space Scatter Plot

- Equation for points within a constant distance c from the sample mean

$$(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \le c^2$$

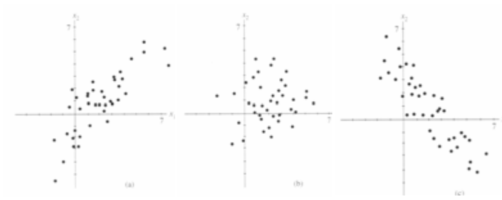
Volume of $\left\{(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \le c^2\right\}$

$$= k_p|\mathbf{S}|^{1/2}c^p$$

A large volume corresponds to a large generalized variance

33

Example 3.8: Scatter Plots



34

Example 3.8: Sample Mean and Variance-Covariance Matrices

$$\mathbf{S} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, r = 0.8$$

$$\mathbf{S} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, r = 0$$

$$\mathbf{S} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, r = -0.8$$

$\bar{\mathbf{x}}' = [2, 1]$, $|\mathbf{S}| = 9$ for all three cases

35

Example 3.8: Eigenvalues and Eigenvectors

$$\begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix} : \lambda_1 = 9, \lambda_2 = 1$$

$$\mathbf{e}_1' = [1/\sqrt{2}, \quad 1/\sqrt{2}], \mathbf{e}_2' = [1/\sqrt{2}, \quad -1/\sqrt{2}]$$

$$\begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix} : \lambda_1 = 3, \lambda_2 = 3$$

$$\mathbf{e}_1' = [1, \quad 0], \mathbf{e}_2' = [0, \quad 1]$$

$$\begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix} : \lambda_1 = 9, \lambda_2 = 1$$

$$\mathbf{e}_1' = [1/\sqrt{2}, \quad -1/\sqrt{2}], \mathbf{e}_2' = [1/\sqrt{2}, \quad 1/\sqrt{2}]$$

36

Example 3.8: Mean-Centered Ellipse

$(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) \le c^2$

$(\mathbf{x}-\bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}) = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}$

$\mathbf{S}^{-1}$: eigenvalues $\frac{1}{\lambda_1}, \frac{1}{\lambda_2}$; eigenvectors $\mathbf{e}_1, \mathbf{e}_2$

$(\because \mathbf{Se} = \lambda\mathbf{e}, \quad \mathbf{e} = \lambda\mathbf{S}^{-1}\mathbf{e})$

$$\begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1' \\ \mathbf{e}_2' \end{bmatrix} \begin{bmatrix} x_1 - \bar{x}_1 \\ x_2 - \bar{x}_2 \end{bmatrix}$$

Choose $c^2 = 5.99$ to cover approximately 95% observations

37

Example 3.8: Semi-major and Semi-minor Axes

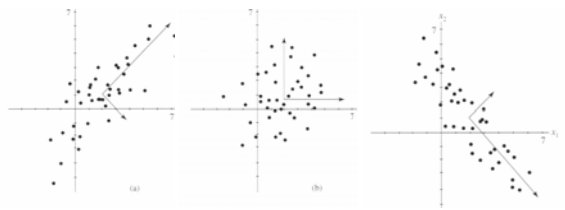
$$\mathbf{S} = \begin{bmatrix} 5 & 4 \\ 4 & 5 \end{bmatrix}, a = 3\sqrt{5.99}, b = \sqrt{5.99}$$

$$\mathbf{S} = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}, a = \sqrt{3}\sqrt{5.99}, b = \sqrt{3}\sqrt{5.99}$$

$$\mathbf{S} = \begin{bmatrix} 5 & -4 \\ -4 & 5 \end{bmatrix}, a = 3\sqrt{5.99}, b = \sqrt{5.99}$$

38

Example 3.8: Scatter Plots with Major Axes



39

Result 3.2

- The generalized variance is zero when the columns of the following matrix are linear dependent

$$\begin{bmatrix} \mathbf{x}_1' - \bar{\mathbf{x}}' \\ \mathbf{x}_2' - \bar{\mathbf{x}}' \\ \vdots \\ \mathbf{x}_n' - \bar{\mathbf{x}}' \end{bmatrix} = \begin{bmatrix} x_{11} - \bar{x}_1 & x_{12} - \bar{x}_2 & \cdots & x_{1p} - \bar{x}_p \\ x_{21} - \bar{x}_1 & x_{22} - \bar{x}_2 & \cdots & x_{2p} - \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} - \bar{x}_1 & x_{n2} - \bar{x}_2 & \cdots & x_{pp} - \bar{x}_p \end{bmatrix} = \mathbf{X} - \mathbf{1}\bar{\mathbf{x}}'$$

40

Proof of Result 3.2

$0 = a_1 \,\mathrm{col}_1(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}') + \cdots + a_p \,\mathrm{col}_p(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$

$= (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a}, \quad \mathbf{a} \ne 0$

$(n-1)\mathrm{S} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$

$\because (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')$

$$= \begin{bmatrix} \mathbf{x}_1' - \bar{\mathbf{x}}' & \mathbf{x}_2' - \bar{\mathbf{x}}' & \cdots & \mathbf{x}_p' - \bar{\mathbf{x}}' \end{bmatrix} \begin{bmatrix} \mathbf{x}_1' - \bar{\mathbf{x}}' \\ \mathbf{x}_2' - \bar{\mathbf{x}}' \\ \vdots \\ \mathbf{x}_n' - \bar{\mathbf{x}}' \end{bmatrix}$$

$$= \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})$$

41

Proof of Result 3.2

$(n-1)\mathbf{S}a = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = 0$

$a_1 \,\mathrm{col}_1(\mathbf{S}) + \cdots + a_p \,\mathrm{col}_p(\mathbf{S}) = 0 \Rightarrow |\mathbf{S}| = 0$

if $|\mathbf{S}| = 0, \exists \mathbf{a}$ such that $\mathbf{Sa} = 0$

$0 = (n-1)\mathbf{Sa} = (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a}$

$\mathbf{a}'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = 0$

$L^2_{(\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a}} = 0 \Rightarrow (\mathbf{X} - \mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = 0$

42

Example 3.9

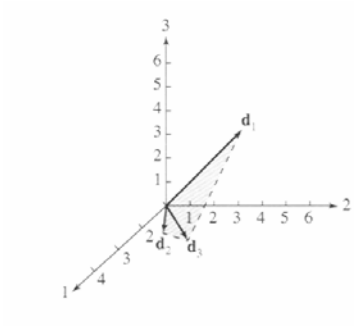
$$\mathbf{X}=\begin{bmatrix}1&2&5\\4&1&6\\4&0&4\end{bmatrix},\bar{\mathbf{x}}'=[3,\quad 1,\quad 5],\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}'=\begin{bmatrix}-2&1&0\\1&0&1\\1&-1&-1\end{bmatrix}$$

$$\mathbf{d}_1'=[-2,1,1],\mathbf{d}_2'=[1,0,-1],\mathbf{d}_3'=[0,1,-1]$$

$$\mathbf{d}_3=\mathbf{d}_1+2\mathbf{d}_2\Rightarrow|\mathbf{S}|=0$$

$$\text{check}:\mathbf{S}=\begin{bmatrix}3&-3/2&0\\-3/2&1&1/2\\0&1/2&1\end{bmatrix}\Rightarrow|\mathbf{S}|=0$$

Example 3.9



Examples Cause Zero Generalized Variance

- Example 1
  - Data are test scores
  - Included variables that are sum of others
  - e.g., algebra score and geometry score were combined to total math score
  - e.g., class midterm and final exam scores summed to give total points
- Example 2
  - Total weight of chemicals was included along with that of each component

Example 3.10

$$\mathbf{X}=\begin{bmatrix}1&9&10\\4&12&16\\2&10&12\\5&8&13\\3&11&14\end{bmatrix},\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}'=\begin{bmatrix}-2&-1&-3\\1&2&3\\-1&0&-1\\2&-2&0\\0&1&1\end{bmatrix},\mathbf{S}=\begin{bmatrix}2.5&0&2.5\\0&2.5&2.5\\2.5&2.5&5.0\end{bmatrix}$$

$|\mathbf{S}|=0\Rightarrow\mathbf{Sa}=0$

Eigenvector corresponding to zero eigenvalues of S

$\Rightarrow\mathbf{a}'=[1,1,-1]$

$\therefore 1(x_{j1}-\bar{x}_1)+1(x_{j2}-\bar{x}_2)-(x_{j3}-\bar{x}_3)=0$

Result 3.3

- If the sample size is less than or equal to the number of variables ($n\le p$) then $|\mathbf{S}|$ = 0 for all samples

Proof of Result 3.3

The n row vectors of $\mathbf{X}$-$\mathbf{1}\bar{\mathbf{x}}'$ sum to the zero vector

because $\sum_{j=1}^{n}x_{jk}=\sum_{j=1}^{n}\bar{x}_k$

Thus the rank of $\mathbf{X}$-$\mathbf{1}\bar{\mathbf{x}}'$ is less than or equal to $n-1$,

i.e., less than or equal to $p-1$, because of $n\le p$

Since $(n-1)\mathbf{S}=(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')$,

$$(n-1)\,\mathrm{col}_k(\mathbf{S})=(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'\mathrm{col}_k(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')$$
$$=(x_{1k}-\bar{x}_k)\,\mathrm{row}_1(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'+\cdots+(x_{nk}-\bar{x}_k)\,\mathrm{row}_n(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'$$

Proof of Result 3.3

$\therefore \text{row}_1(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')'$ is a linear combination of the remaining row vectors

$\text{col}_k(S)$ is a linear combination of at most $n-1$ linear independent of transpose of row vectors

The rank of $\mathbf{S}$ is thus less than or equal to n-1, i.e., less than or equal to p-1.

Since $\mathbf{S}$ is a p by p matrix, $|\mathbf{S}|=0$

Result 3.4

- Let the p by 1 vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$, where $\mathbf{x}_j'$ is the jth row of the data matrix $\mathbf{X}$, be realizations of the independent random vectors $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.
- If the linear combination $\mathbf{a}'X_j$ has positive variance for each non-zero constant vector $\mathbf{a}$, then, provided that $p < n$, $\mathbf{S}$ has full rank with probability 1 and $|\mathbf{S}| > 0$
- If, with probability 1, $\mathbf{a}'\mathbf{X}_j$ is a constant c for all j, then $|\mathbf{S}| = 0$

Proof of Part 2 of Result 3.4

$\mathbf{a}'\mathbf{X}_j = a_1X_{j1} + a_2X_{j2} + \cdots + a_pX_{jp} = c$ with probability 1,

$\mathbf{a}'\mathbf{x}_j = c$ for all j. The sample mean for it is

$$\sum_{j=1}^{n}\left(a_1x_{j1} + a_2x_{j2} + \cdots + a_px_{jp}\right)/n = \mathbf{a}'\bar{\mathbf{x}} = c$$

$$(\mathbf{X}-\mathbf{1}\bar{\mathbf{x}}')\mathbf{a} = a_1\begin{bmatrix} x_{11}-\bar{x}_1 \\ \vdots \\ x_{n1}-\bar{x}_1 \end{bmatrix} + \cdots + a_p\begin{bmatrix} x_{1p}-\bar{x}_p \\ \vdots \\ x_{np}-\bar{x}_p \end{bmatrix}$$

$$= \begin{bmatrix} \mathbf{a}'\mathbf{x}_1 - \mathbf{a}'\bar{\mathbf{x}} \\ \vdots \\ \mathbf{a}'\mathbf{x}_n - \mathbf{a}'\bar{\mathbf{x}} \end{bmatrix} = \begin{bmatrix} c-c \\ \vdots \\ c-c \end{bmatrix} = 0 \Rightarrow |\mathbf{S}| = 0$$

Generalized Sample Variance of Standardized Variables

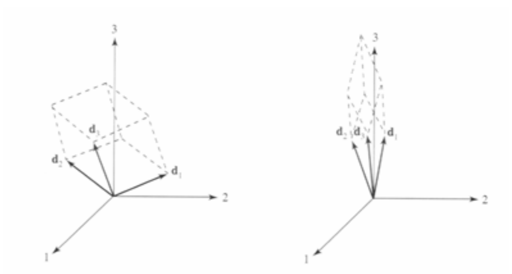
Generalized sample variance of the standardized variables $= |\mathbf{R}|$

$$\frac{y_i - \bar{x}_i 1}{\sqrt{s_{ii}}} = \begin{bmatrix} \dfrac{x_{1i}-\bar{x}_i}{\sqrt{s_{ii}}} & \dfrac{x_{2i}-\bar{x}_i}{\sqrt{s_{ii}}} & \cdots & \dfrac{x_{ni}-\bar{x}_i}{\sqrt{s_{ii}}} \end{bmatrix}',$$

$|\mathbf{R}| = (n-1)^{-p}(volume)^2, |\mathbf{S}| = (s_{11}s_{22}\cdots s_{pp})|\mathbf{R}|$

$|\mathbf{R}|$ is large when all r_{ik} are nearly zero, and is small when one or more r_{ik} are nearly $+1$ or -1

Volume Generated by Deviation Vectors of Standardized Variables



Example 3.11

$$\mathbf{S} = \begin{bmatrix} 4 & 3 & 1 \\ 3 & 9 & 2 \\ 1 & 2 & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 1 & 1/2 & 1/2 \\ 1/2 & 1 & 2/3 \\ 1/2 & 2/3 & 1 \end{bmatrix}$$

$s_{11} = 4, \quad s_{22} = 9, \quad s_{33} = 1$

$$|\mathbf{S}| = 14, \quad |\mathbf{R}| = \frac{7}{18}, \quad |\mathbf{S}| = s_{11}s_{22}s_{33}|\mathbf{R}|$$

Total Sample Variance

$$\text{Total Sample Variance} = s_{11} + s_{22} + \cdots + s_{pp}$$

Pays no attention to the orientation of the residual vectors

$$\text{Example 3.7}: S = \begin{bmatrix} 252.04 & -68.43 \\ -68.43 & 123.67 \end{bmatrix}$$

$$\text{Total sample variance} = 375.71$$

$$\text{Example 3.9}: S = \begin{bmatrix} 3 & -3/2 & 0 \\ -3/2 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}$$

$$\text{Total sample variance} = 5$$

55

Outline

- The Geometry of the Sample
- Random Samples and the Expected Values of the Sample Mean and Covariance Matrix
- Generalized Variance
- Sample Mean, Covariance, and Correlation as Matrix Operations
- Sample Values of Linear Combinations of Variables

56

Questions

- How to compute sample mean by matrix operation?
- How to compute sample covariance matrix by matrix operation?
- How to compute sample correlation coefficient matrix by matrix operation?

Sample Mean as Matrix Operation

$$\bar{\mathbf{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \\ \vdots \\ \bar{x}_p \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1'\mathbf{1}/n \\ \mathbf{y}_2'\mathbf{1}/n \\ \vdots \\ \mathbf{y}_p'\mathbf{1}/n \end{bmatrix} = \frac{1}{n} \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ x_{12} & x_{22} & \cdots & x_{n2} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p} & x_{2p} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

$$= \frac{1}{n}\mathbf{X}'\mathbf{1}$$

58

Covariance as Matrix Operation

$$\begin{bmatrix} \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ \bar{x}_1 & \bar{x}_2 & \cdots & \bar{x}_p \end{bmatrix} = \mathbf{1}\bar{\mathbf{x}}' = \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}$$

$$\begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix} = \mathbf{X} - \mathbf{1}\mathbf{1}'\mathbf{X}$$

59

Covariance as Matrix Operation

$$(n-1)\mathbf{S} = \begin{bmatrix} x_{11}-\bar{x}_1 & x_{21}-\bar{x}_1 & \cdots & x_{n1}-\bar{x}_1 \\ x_{12}-\bar{x}_2 & x_{22}-\bar{x}_2 & \cdots & x_{n2}-\bar{x}_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_{1p}-\bar{x}_p & x_{2p}-\bar{x}_p & \cdots & x_{np}-\bar{x}_p \end{bmatrix} \times$$

$$\begin{bmatrix} x_{11}-\bar{x}_1 & x_{12}-\bar{x}_2 & \cdots & x_{1p}-\bar{x}_p \\ x_{21}-\bar{x}_1 & x_{22}-\bar{x}_2 & \cdots & x_{2p}-\bar{x}_p \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1}-\bar{x}_1 & x_{n2}-\bar{x}_2 & \cdots & x_{np}-\bar{x}_p \end{bmatrix}$$

$$= \left(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right)'\left(\mathbf{X} - \frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right)$$

60

Covariance as Matrix Operation

$$\left(\mathbf{X}-\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right)'\left(\mathbf{X}-\frac{1}{n}\mathbf{1}\mathbf{1}'\mathbf{X}\right)$$

$$=\mathbf{X}'(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')'(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}$$

$$(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')'(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')=\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}'-\frac{1}{n}\mathbf{1}\mathbf{1}'+\frac{1}{n^2}\mathbf{1}\mathbf{1}'\mathbf{1}\mathbf{1}'$$

$$=\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}' \quad (\because \mathbf{1}'\mathbf{1}=n)$$

$$\mathbf{S}=\frac{1}{n-1}\mathbf{X}'(\mathbf{I}-\frac{1}{n}\mathbf{1}\mathbf{1}')\mathbf{X}$$

61

Sample Standard Deviation Matrix

$$\mathbf{D}^{1/2}=\begin{bmatrix}\sqrt{s_{11}} & 0 & \cdots & 0\\ 0 & \sqrt{s_{22}} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & \sqrt{s_{pp}}\end{bmatrix},\ \mathbf{D}^{-1/2}=\begin{bmatrix}1/\sqrt{s_{11}} & 0 & \cdots & 0\\ 0 & 1/\sqrt{s_{22}} & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 1/\sqrt{s_{pp}}\end{bmatrix}$$

$$\mathbf{R}=\begin{bmatrix}\frac{s_{11}}{\sqrt{s_{11}}\sqrt{s_{11}}} & \frac{s_{12}}{\sqrt{s_{11}}\sqrt{s_{22}}} & \cdots & \frac{s_{1p}}{\sqrt{s_{11}}\sqrt{s_{pp}}}\\ \frac{s_{21}}{\sqrt{s_{22}}\sqrt{s_{11}}} & \frac{s_{22}}{\sqrt{s_{22}}\sqrt{s_{22}}} & \cdots & \frac{s_{2p}}{\sqrt{s_{22}}\sqrt{s_{pp}}}\\ \vdots & \vdots & \ddots & \vdots\\ \frac{s_{p1}}{\sqrt{s_{pp}}\sqrt{s_{11}}} & \frac{s_{p2}}{\sqrt{s_{pp}}\sqrt{s_{22}}} & \cdots & \frac{s_{pp}}{\sqrt{s_{pp}}\sqrt{s_{pp}}}\end{bmatrix}=\mathbf{D}^{-1/2}\mathbf{S}\mathbf{D}^{-1/2}$$

$$\mathbf{S}=\mathbf{D}^{1/2}\mathbf{R}\mathbf{D}^{1/2}$$

62

Outline

- The Geometry of the Sample
- Random Samples and the Expected Values of the Sample Mean and Covariance Matrix
- Generalized Variance
- Sample Mean, Covariance, and Correlation as Matrix Operations
- Sample Values of Linear Combinations of Variables

63

Questions

- Result 3.5
- Result 3.6

Result 3.5

$$\mathbf{b}'\mathbf{X}=b_1X_1+b_2X_2+\cdots+b_pX_p$$

$$\mathbf{c}'\mathbf{X}=c_1X_1+c_2X_2+\cdots+c_pX_p$$

Sample mean of $\mathbf{b}'\mathbf{X}=\mathbf{b}'\bar{\mathbf{x}}$

Sample variance of $\mathbf{b}'\mathbf{X}=\mathbf{b}'\mathbf{S}\mathbf{b}$

Sample covariance of $\mathbf{b}'\mathbf{X}$ and $\mathbf{c}'\mathbf{X}=\mathbf{b}'\mathbf{S}\mathbf{c}$

65

Proof of Result 3.5

$$\mathbf{b}'\mathbf{x}_j=b_1x_{j1}+b_2x_{j2}+\cdots+b_px_{jp}$$

$$\text{Sample mean}=\frac{\mathbf{b}'\mathbf{x}_1+\mathbf{b}'\mathbf{x}_2+\cdots+\mathbf{b}'\mathbf{x}_n}{n}=\mathbf{b}'\bar{\mathbf{x}}$$

$$(\mathbf{b}'\mathbf{x}_j-\mathbf{b}'\bar{\mathbf{x}})^2=\mathbf{b}'(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\mathbf{b}$$

$$\text{Sample variance}=\frac{1}{n-1}\sum_{j=1}^{n}\left(\mathbf{b}'\mathbf{x}_j-\mathbf{b}'\bar{\mathbf{x}}\right)^2$$

$$=\frac{1}{n-1}\mathbf{b}'\sum_{j=1}^{n}(\mathbf{x}_j-\bar{\mathbf{x}})(\mathbf{x}_j-\bar{\mathbf{x}})'\mathbf{b}=\mathbf{b}'\mathbf{S}\mathbf{b}$$

66

Proof of Result 3.5

$$\text{Sample covariance} = \frac{1}{n\text{-}1}\sum_{j=1}^{n}\left(\mathbf{b}'\mathbf{x}_j - \mathbf{b}'\bar{\mathbf{x}}\right)\left(\mathbf{c}'\mathbf{x}_j - \mathbf{c}'\bar{\mathbf{x}}\right)$$

$$= \frac{1}{n-1}\mathbf{b}'\sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{c} = \mathbf{b}'\mathbf{S}\mathbf{c}$$

67

Result 3.6

$$\mathbf{AX} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{q1} & a_{q2} & \cdots & a_{qp} \end{bmatrix}\begin{bmatrix} X_1 \\ X_2 \\ \vdots \\ X_p \end{bmatrix}$$

Sample mean of $\mathbf{AX} = \mathbf{A}\bar{\mathbf{x}}$

Sample covariance matrix $= \mathbf{ASA}'$

68