—— WORKSHOP ——

Applied Classical and Modern Multivariate Statistical Analysis

Module 2-2: Multivariate Data Summarization

Weixing Song, Juan Du

Department of Statistics
Kansas State University

August 5, 2018

Outline

Outline

Example (Lizard Size Data): A zoologist obtained measurements on $n = 25$ lizards known scientifically as Cophosaurus texanus. The weight, or mass, is given in grams while the snout-vent length (SVL) and hind limb span (HLS) are given in millimeters.

Table 1.3 Lizard Size Data

Lizard	Mass	SVL	HLS	Lizard	Mass	SVL	HLS
1	5.526	59.0	113.5	14	10.067	73.0	136.5
2	10.401	75.0	142.0	15	10.091	73.0	135.5
3	9.213	69.0	124.0	16	10.888	77.0	139.0
4	8.953	67.5	125.0	17	7.610	61.5	118.0
5	7.063	62.0	129.5	18	7.733	66.5	133.5
6	6.610	62.0	123.0	19	12.015	79.5	150.0
7	11.273	74.0	140.0	20	10.049	74.0	137.0
8	2.447	47.0	97.0	21	5.149	59.5	116.0
9	15.493	86.5	162.0	22	9.158	68.0	123.0
10	9.004	69.0	126.5	23	12.132	75.0	141.0
11	8.199	70.5	136.0	24	6.978	66.5	117.0
12	6.601	64.5	116.0	25	6.890	63.0	117.0
13	7.622	67.5	135.0				

Source: Data courtesy of Kevin E. Bonine.

Layouts of Multivariate Data: Matrix

We use a p-dimensional vector $X = (X_1, \dots, X_p)'$ to denote the p features of a population.

A sample of size n draw from X are denoted by

$$\mathbf{x}_1 = \begin{bmatrix} x_{11} \\ x_{12} \\ \vdots \\ x_{1p} \end{bmatrix}, \cdots, \mathbf{x}_j = \begin{bmatrix} x_{j1} \\ x_{j2} \\ \vdots \\ x_{jp} \end{bmatrix}, \cdots, \mathbf{x}_n = \begin{bmatrix} x_{n1} \\ x_{n2} \\ \vdots \\ x_{np} \end{bmatrix}.$$

The vectors $\mathbf{x}_1, \dots, \mathbf{x}_n$ are independent, and the measurements of the p variables in a single trial will usually be correlated.

The entire sample is often placed in an $n \times p$ matrix:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$$

Two Geometric Interpretations

First Geometric Interpretation:

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

$x_1, \ldots, x_n$ can be viewed as n-points in a p-dimensional Euclidean space.

The scatter plot of n points in p-dimensional space provides information on the locations and variability of the points.

Second Geometrical Representation:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix} = [\mathbf{y}_1 \mid \mathbf{y}_2 \mid \cdots \mid \mathbf{y}_p].$$

The j-th point $\mathbf{y}_j$ are the n measurements on the j-th variable.

In the geometrical representations, we depict $\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_p$ as vectors rather than points.

Outline

Sample Mean and Covariance Matrix

- **Sample Mean:** The sample mean vector of a multivariate sample $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is defined as

$$\bar{\mathbf{x}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i = \begin{bmatrix} \frac{1}{n}\sum_{i=1}^{n} x_{i1} \\ \vdots \\ \frac{1}{n}\sum_{i=1}^{n} x_{ip} \end{bmatrix} = \frac{1}{n}\mathbf{X}'\mathbf{1},$$

  where $\mathbf{1}_{p\times 1} = (1, 1, \cdots, 1)'$.

- **Sample Covariance Matrix:** The sample covariance of a multivariate sample $\mathbf{x}_1, \cdots, \mathbf{x}_n$ is defined as

$$S_n = \left(\frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k)\right)_{j,k=1,2,\ldots,p} = \frac{1}{n}\mathbf{X}'\left(I - \frac{1}{n}\mathbf{1}\mathbf{1}'\right)\mathbf{X},$$

Statistical Properties

Result 3.1

Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$ be a random sample from a joint distribution that has mean μ and covariance matrix Σ. Let $\bar{\mathbf{X}}$ be the sample mean and S_n be the sample covariance matrix. Then

$$E\bar{\mathbf{X}} = \mu, \quad \mathrm{Cov}(\bar{\mathbf{X}}) = \frac{1}{n}\Sigma, \quad ES_n = \frac{n-1}{n}\Sigma.$$

Note:

- Result 3.1 implies that $\bar{\mathbf{X}}$ is an unbiased estimator of μ; $nS_n/(n-1)$ is an unbiased estimator of Σ.
- Denote $S = nS_n/(n-1)$. The determinant of S is called the generalized sample variance of $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n$.

Sample Correlation Coefficient Matrix

- Let s_{jk} be the (j,k)-th element in S. Then the sample correlation coefficient of $\mathbf{X}_1, \cdots, \mathbf{X}_n$ is defined by

$$R = \left(\frac{s_{jk}}{\sqrt{s_{jj} \cdot s_{kk}}} \right)_{j,k=1,2,\ldots,p}$$

Note that

$$s_{jk} = \frac{1}{n-1} \sum_{i=1}^{n} (x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k).$$

Define $D = \mathrm{Diag}(s_{11}, s_{22}, \ldots, s_{pp})$. Then

$$R = D^{-1/2} S D^{-1/2}.$$

Deviations: The vectors

$$\mathbf{d}_j = \mathbf{y}_j - \bar{\mathbf{x}}_j \mathbf{1} = \begin{bmatrix} x_{1j} - \bar{\mathbf{x}}_j \\ x_{2j} - \bar{\mathbf{x}}_j \\ \vdots \\ x_{nj} - \bar{\mathbf{x}}_j \end{bmatrix}$$

is called the deviation vector of $\mathbf{y}_j$, $j = 1, 2, \ldots, p$.

First Generalized Sample Variance

Let $\mathbf{S}$ be a sample covariance matrix. The first generalized sample variance is defined as the determinant of $\mathbf{S}$, or $|\mathbf{S}|$.

Geometric Meaning of $|\mathbf{S}|$

(1)

$$|\mathbf{S}| = \frac{(\text{Volume})^2}{(n-1)^p},$$

where the Volume is the volume generated by the p deviation vectors $\mathbf{d}_1, \ldots, \mathbf{d}_p$.

(2) Define the hyperellipsoid $V := \{\mathbf{x} : (\mathbf{x} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x} - \bar{\mathbf{x}}) \le c^2\}$.

Then

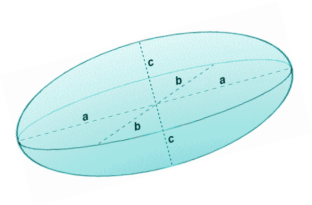
$$\text{Volume of } V = \frac{2\pi^{p/2}|\mathbf{S}|^{1/2}c^p}{p\Gamma(p)}.$$

Fact: An arbitrarily oriented ellipsoid, centered at v, is defined by the solutions x to the equation

$$(\mathbf{x}-\mathbf{v})^T A(\mathbf{x}-\mathbf{v}) = 1,$$

where A is a positive definite matrix and x, v are vectors.

The eigenvectors of A define the principal axes of the ellipsoid and the eigenvalues of A are the reciprocals of the squares of the semi-axes.



An ellipsoid with semi-axes a, b and c

When $|S| = 0$

Result 3.2

The generalized variance is zero when, and only when, at least one deviation vector lies in the (hyper-)plane formed by all linear combinations of the others. That is, when the columns of the matrix of deviations are linearly dependent.

Remark:

- Collinearity!
- If $n \leq p$, then $|\mathbf{S}| = 0$.

Second Generalized Variance

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

Geometric Interpretation: It can be shown that the total sample variance is the sum of the squared lengths of the p deviation vectors $\mathbf{d}_1, \ldots, \mathbf{d}_p$, divided by $n-1$.

Sample Values of Linear Combinations of Variables

For a random vector $X = (X_1, \ldots, X_p)'$, and two real valued vectors $\mathbf{b} = (b_1, \ldots, b_p)$, $\mathbf{c} = (c_1, \ldots, c_p)$, the linear combinations of X with coefficient vector $\mathbf{b}$ and $\mathbf{c}$ is defined as

$$\mathbf{b}'X = b_1X_1 + \cdots + b_pX_p, \quad \mathbf{c}'X = c_1X_1 + \cdots + c_pX_p.$$

The observed value of $\mathbf{c}'X$ on the j-th trial is

$$\mathbf{c}'\mathbf{x}_j = c_1x_{j1} + c_2x_{j2} + \cdots + c_px_{jp}, \quad j = 1, 2, \ldots, n.$$

Sample Mean and (Co)Variance

$$\begin{aligned} \text{Sample Mean of } \mathbf{c}'X &= \mathbf{c}'\bar{\mathbf{x}}, \\ \text{Sample Variance of } \mathbf{c}'X &= \mathbf{c}'\mathbf{S}\mathbf{c}, \\ \text{Sample Covariance of } \mathbf{c}'X \text{ and } \mathbf{b}'X &= \mathbf{c}'\mathbf{S}\mathbf{b}. \end{aligned}$$