Chapter 3 Moving to Higher Dimensions

Applied Multivariate Statistical Analysis

Course Instructor: Shuen-Lin Jeng

Department of Statistics
National Cheng Kung University

March, 2, 2021

3.1 Covariance

Covariance is a measure of dependency between random variables. Given two (random) variables X and Y the (theoretical) covariance is defined by

$$\sigma_{XY} = \mathbf{Cov}(X, Y) = \mathbf{E}(XY) - (\mathbf{E}X)(\mathbf{E}Y) \tag{3.1}$$

the covariance matrix is:

$$\Sigma = \begin{pmatrix} \sigma_{X_1X_1} & \cdots & \sigma_{X_1X_p} \\ \vdots & \ddots & \vdots \\ \sigma_{X_pX_1} & \cdots & \sigma_{X_pX_p} \end{pmatrix}.$$

3.2 Correlation

The correlation between two variables X and Y is defined from the covariance as the following:

$$\rho_{XY} = \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{Var}(X)\mathbf{Var}(Y)}} \tag{3.7}$$

The empirical version of ρ_{XY} is as follows:

$$r_{XY} = \frac{s_{XY}}{\sqrt{s_{XX}s_{YY}}} \tag{3.8}$$

Theorem 3.1 *if X and Y are independent, then $\rho(X,Y) = \mathbf{Cov}(X,Y)=0$.*

3.2 Correlation

Example 3.4 Consider a standard normally distributed random variable X and a random variable $Y = X^2$, which is surely not independent of X. Here we have

$$\mathbf{Cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y) = \mathbf{E}(X^3) = 0$$

(because $\mathbf{E}(X) = 0$ and $\mathbf{E}(X^2) = 1$). Therefore ρ(X,Y)=0 ,as well. This example also shows that correlations and covariances measure only linear dependence. The quadratic dependence of $Y = X^2$ on X is not reflected by these measures of dependence.

3.2 Correlation

Fisher's Z-transformation,

$$W = \frac{1}{2}\log(\frac{1+r_{XY}}{1-r_{XY}}), \tag{3.11}$$

we obtain a variable that has a more accessible distribution. Under the hypothesis that ρ=0, W has an asymptotic normal distribution.

$$\mathbf{E}(W) \approx \frac{1}{2}\log(\frac{1+\rho_{XY}}{1-\rho_{XY}})$$

$$\mathbf{Var}(W) \approx \frac{1}{(n-3)} \tag{3.12}$$

Theorem 3.2

$$Z = \frac{W - \mathbf{E}(W)}{\sqrt{\mathbf{Cov}(W)}} \xrightarrow{\mathcal{L}} N(0,1) \tag{3.13}$$

3.2 Correlation

Remark 3.2 The normalizing and variance stabilizing properties of W are asymptotic. In addition, the use of W in small samples (for n ≤ 25) is improved by Hotelling's transform Hotelling (1953):

$$W^* = W - \frac{3W + \tanh(W)}{4(n-1)} \text{ with } \mathbf{Var}(W^*) = \frac{1}{n-1}.$$

Remark 3.4 Under the assumptions of normality of X and Y, we may text their independence ($\rho_{XY} = 0$) using the exact t-distribution of the statistic

$$T = r_{XY}\sqrt{\frac{n-2}{1-r_{XY}^2}} \overset{\rho_{XY}=0}{\sim} t_{n-2}$$

Setting the probability of the first error type to α, we reject the null hypothesis $\rho_{XY} = 0$ if $|T| \geq t_{1-\alpha/2;n-2}$.

3.3 Summary Statistic

$$\bar{x} = \begin{pmatrix} \bar{x_1} \\ \vdots \\ \bar{x_p} \end{pmatrix} = n^{-1}\mathcal{X}^T I_n \tag{3.17}$$

The empirical covariances are :

$$\mathcal{S} = n^{-1}\mathcal{X}^T\mathcal{X} - \bar{x}\bar{x}^T = n^{-1}(\mathcal{X}^T\mathcal{X} - n^{-1}\mathcal{X}^T 1_n 1_n^T \mathcal{X}) \tag{3.18}$$

Note that this matrix is equivalently defined by

$$\mathcal{S} = \frac{1}{n}\sum_{i=1}^{N}(x_i - \bar{x})(x_i - \bar{x})^T$$

3.3 Summary Statistic

The covariance formula (3.18) can be rewritten as $\mathcal{S} = n^{-1}\ \mathcal{X}^T\mathcal{H}\mathcal{X}$ with the *centering matrix*

$$\mathcal{H} = \mathcal{I}_n - n^{-1}1_n1_n^T. \tag{3.19}$$

Note that the centering matrix is symmetric and idempotent.

3.3 Summary Statistic

As a consequence $\mathcal{S}$ is positive semidefinite, i.e.,

$$\mathcal{S} \geq 0 \tag{3.20}$$

The sample correlation coefficient between the i-th and j-th variables is $r_{X_iX_j}$,see (3.8). If $\mathcal{D} = \text{diag}(s_{X_iX_i})$, then the correlation matrix is

$$\mathcal{R} = \mathcal{D}^{-1/2}\mathcal{S}\mathcal{D}^{-1/2}, \tag{3.21}$$

where $\mathcal{D}^{-1/2}$ is a diagonal matrix with elements $(s_{X_iX_i})^{-1/2}$ on its main diagonal.

3.3 Summary Statistic

Linear Transformation

Let $\mathcal{A}$ be a $q \times p$ matrix and consider the transformed data matrix

$$\mathcal{Y} = \mathcal{X}\mathcal{A}^T = (y_1, \ldots, y_n)^T. \tag{3.22}$$

The row $y_i = (y_{i1}, \ldots, y_{iq}) \in \mathbb{R}^q$ can be viewed as the i-th observation of a q-dimensional random variable $Y = \mathcal{A}X$. In fact we have $y_i = x_i\mathcal{A}^T$.

$$\bar{y} = \frac{1}{n}\mathcal{Y}^{\mathsf{T}} I_n = \frac{1}{n}\mathcal{A}\mathcal{X}^{\mathsf{T}} I_n = \mathcal{A}\bar{x} \tag{3.23}$$

$$\mathcal{S}_{\mathcal{Y}} = \frac{1}{n}\mathcal{Y}^{\mathsf{T}}\mathcal{H}\mathcal{Y} = \frac{1}{n}\mathcal{A}\mathcal{X}^{\mathsf{T}}\mathcal{H}\mathcal{X}\mathcal{A}^{\mathsf{T}} = \mathcal{A}\mathcal{S}_{\mathcal{X}}\mathcal{A}^{\mathsf{T}} \tag{3.24}$$

3.3 Summary Statistic

Note that if the linear transformation is nonhomogeneous, i.e.,

$$y_i = \mathcal{A}x_i + b$$

only (3.23) changes: $\bar{y} = \mathcal{A}\bar{x} + b$.
q = 1, i.e., y $=\mathcal{X}a$, i.e., $y_i = a^T x_i$; i $= 1, \ldots, n$:

$$\bar{y} = a^T \bar{x}$$

$$\mathcal{S}_y = a^T \mathcal{S}_{\mathcal{X}} a.$$

3.3 Summary Statistic

Example 3.9 Suppose that $\mathcal{X}$ is the pullover data set. The manager wants to compute his mean expenses for advertisement (X_3) and sales assistant(X_4).

$$Y = X_3 + 10X_4$$

and $\mathcal{A}(4 \times 1)$ as

$$\mathcal{A} = \begin{pmatrix} 0, & 0, & 1, & 10, \end{pmatrix}.$$

$$\overline{y} = \mathcal{A}\overline{x} = (0, 0, 1, 10) \begin{pmatrix} 172.7 \\ 104.6 \\ 104.0 \\ 93.8 \end{pmatrix} = 1042.0$$

$$\mathcal{S}_{\mathcal{Y}} = \mathcal{A}\mathcal{S}_{\mathcal{X}}\mathcal{A}^{\top} = (0, 0, 1, 10) \begin{pmatrix} 1152.5 & -88.9 & 1589.7 & 301.6 \\ -88.9 & 244.3 & 102.3 & -101.8 \\ 1589.7 & 102.3 & 2915.6 & 233.7 \\ 301.6 & -101.8 & 233.7 & 197.1 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \\ 1 \\ 10 \end{pmatrix}$$

$$= 2915.6 + 4674 + 19710 = 27299.6.$$

3.3 Summary Statistic

Mahalanobis Transformation

$$z_i = \mathcal{S}^{-\frac{1}{2}}(x_i - \bar{x}),\ \mathrm{i} = 1, \ldots,\ \mathrm{n}. \tag{3.25}$$

Note that for the transformed data matrix $\mathcal{Z} = (z_1, \ldots, z_n)^T$,

$$\mathcal{S}_{\mathcal{Z}} = n^{-1}\mathcal{Z}^T\mathcal{H}\mathcal{Z} = \mathcal{I}_p. \tag{3.26}$$

So the Mahalanobis transformation eliminates the correlation between the variables and standardizes the variance of each variable.

3.4 Linear Model for Two Variables

A slope line is a linear relationship between X and Y :

$$y_i = \alpha + \beta x_i + \varepsilon_i, i = 1, \ldots, n. \tag{3.27}$$

Here, α is the intercept and β is the slope of the line. The errors(or deviations from the line) are denoted as ε_i and are assumed to have zero mean and finite variance σ^2. ($\hat{\alpha}$,$\hat{\beta}$) via graphical techniques. A very common numerical and statistical technique is to use those $\hat{\alpha}$ and $\hat{\beta}$ that minimize:

$$(\hat{\alpha}, \hat{\beta}) = \arg \underset{(\alpha,\beta)}{min} \sum_{i=1}^{n} (y_i - \alpha - \beta x_i)^2 \tag{3.28}$$

$$\hat{\beta} = \frac{s_{XY}}{s_{XX}} \tag{3.29}$$

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x} \tag{3.30}$$

3.4 Linear Model for Two Variables

The variance of $\hat{\beta}$ is

$$\mathbf{Var}(\hat{\beta}) = \frac{\sigma^2}{n \cdot s_{XX}} \tag{3.31}$$

and

$$t = \frac{\hat{\beta}}{SE(\hat{\beta})} \tag{3.33}$$

and rejects the hypothesis at a 5% significance level if $|t| \geq t_{0.975;n-2}$ where the 97.5% quantile of the Student's t_{n-2} distribution is clearly the 95% critical value for the two-sided test.

3.5 Simple Analysis of Variance

The goal of a simple ANOVA is to analyze the observation structure

$$y_{kl} = \mu_l + \varepsilon_{kl} \text{ for } k = 1, \ldots, m, \text{ and } l = 1, \ldots, p. \tag{3.41}$$

$$\sum_{l=1}^{p} \sum_{k=1}^{m} (y_{kl} - \bar{y})^2 = m \sum_{l=1}^{p} (\bar{y}_l - \bar{y})^2 + \sum_{l=1}^{p} \sum_{k=1}^{m} (y_{kl} - \bar{y}_l)^2 \tag{3.42}$$

3.5 Simple Analysis of Variance

$$\text{SS(reduced)} = \sum_{l=1}^{p} \sum_{k=1}^{m} (y_{kl} - \bar{y})^2 \tag{3.43}$$

$$\text{SS(full)} = \sum_{l=1}^{p} \sum_{k=1}^{m} (y_{kl} - \bar{y}_l)^2 \tag{3.44}$$

$$F = \frac{\dfrac{\text{SS(reduced) - SS(full)}}{df(r)\text{ - }df(f)}}{\dfrac{\text{SS(full)}}{df(f)}} \tag{3.45}$$

3.6 Multiple Linear Model

$$y = \mathcal{X}\beta + \varepsilon \tag{3.50}$$

where ε are the errors. The least squares solution is given by $\hat{\beta}$:

$$\hat{\beta} = \arg\min_{\beta} (y - \mathcal{X}\beta)^T (y - \mathcal{X}\beta) = \arg\min_{\beta} \varepsilon^T \varepsilon \tag{3.51}$$

Suppose that $(\mathcal{X}^T\mathcal{X})$ is of full rank and thus invertible. Minimizing the expression (3.51) with respect to β yields

$$\hat{\beta} = (\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T y \tag{3.52}$$

The fitted value $\hat{y} = \mathcal{X}\hat{\beta} = \mathcal{X}(\mathcal{X}^T\mathcal{X})^{-1}\mathcal{X}^T y = \mathcal{P}y$ is the projection of y onto $C(\mathcal{X})$ as computed in (2.47).

3.6 Multiple Linear Model

The least squares residuals are

$$e = y - \hat{y} = y - \mathcal{X}\hat{\beta} = \mathcal{Q}y = (\mathcal{I}_n - \mathcal{P})y.$$

The vector e is the projection of y onto the orthogonal complement of $C(\mathcal{X})$.

Properties of $\hat{\beta}$

$$\mathbf{E}(\hat{\beta}) = \beta$$

$$\mathbf{Var}(\hat{\beta}) = \sigma^2(\mathcal{X}^T\mathcal{X})^{-1}$$

$$\hat{\sigma^2} = \frac{1}{n-(p+1)}(y-\hat{y})^T(y-\hat{y})$$