Multivariate Statistical Analysis

Fall 2011

C. L. Williams, Ph.D.

Lecture 6 for Applied Multivariate Analysis

Outline

I. Normality of linear combinations of the variables in $\mathbf{y}$:

(a) If $\mathbf{a}$ is a vector of constants, the linear function $\mathbf{a}'\mathbf{y} = a_1 y_1 + a_2 y_2 + \cdots + a_p y_p$ is univariate normal:

If $\mathbf{y}$ is $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, then $\mathbf{a}'\mathbf{y}$ is $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

The mean and variance of $\mathbf{a}'\mathbf{y}$ given previously as $\text{E}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\mu}$ and $\text{var}(\mathbf{a}'\mathbf{y}) = \mathbf{a}'\boldsymbol{\Sigma}\mathbf{a}$ for any random vector $\mathbf{y}$. We now have the additional attribute that $\mathbf{a}'\mathbf{y}$ has a (univariate) normal distribution if $\mathbf{y}$ is $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$.

(b) If $\mathbf{A}$ is a constant $q \times p$ matrix of rank q, where $q \leq p$, the q linear combinations in $\mathbf{Ay}$ have a multivariate normal distribution:

If $\mathbf{y}$ is $\text{MVN}_p\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$, then $\mathbf{Ay}$ is $MVN_q\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'\right)$.

$\text{E}(\mathbf{Ay}) = \mathbf{A}\boldsymbol{\mu}$ and $\text{var}(\mathbf{Ay}) = \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}'$ But we now have the additional feature that the q variables in $\mathbf{Ay}$ have a multivariate normal distribution.

Standardized Variables

II. Standardized Variables

A standardized vector $\mathbf{z}$ can be obtained in two ways. The first of which is:

$$\mathbf{z} = \left(\mathbf{T}'\right)^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

where $\mathbf{T}'$ is derived from the Cholesky decomposition of the variance-covariance matrix $\boldsymbol{\Sigma} = \mathbf{T}'\mathbf{T}$.

Second method

Or by using finding the symmetric square root matrix of the variance-covariance matrix $\mathbf{\Sigma} = \mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2}$, giving the standardized variable:

$$\mathbf{z} = \left(\mathbf{\Sigma}^{1/2}\right)^{-1}(\mathbf{y} - \boldsymbol{\mu})$$

In either case it follows that: If $\mathbf{y}$ is $\text{MVN}_p(\boldsymbol{\mu}, \mathbf{\Sigma})$, then $\mathbf{z}$ is $\text{MVN}_p(\mathbf{0}, \mathbf{I})$.

Chi-square distribution

A chi-square random variable with p degrees of freedom is defined as the sum of squares of p independent standard normal random variables. Thus, if $\mathbf{z}$ is the standardized vector defined in (4.4) or (4.5), then this bring us to our definition of Mahalanobis distance, so that if If $\mathbf{y}$ is $\text{MVN}_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$,

$$(\mathbf{y}_i - \boldsymbol{\mu})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}) = \mathbf{z}'\mathbf{z}\ \chi^2_p \tag{1}$$

Normality of marginal distributions

(a) Any subset of the y's in $\mathbf{y}$ has a multivariate normal distribution, with mean vector consisting of the corresponding subvector of $\boldsymbol{\mu}$ and covariance matrix composed of the corresponding submatrix of $\boldsymbol{\Sigma}$.

Let $\mathbf{y}_1 = (y_1, y_2, \ldots, y_r)'$ denote the subvector containing the first r elements of $\mathbf{y}_1$ and $\mathbf{y}_2 = (y_{r+1}, y_{r+2}, \ldots, y_p)'$ consist of the remaining p-r elements. Thus $\mathbf{y}$, $\boldsymbol{\mu}$, and $\boldsymbol{\Sigma}$ are partitioned as

$$\mathbf{y} = \begin{pmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \end{pmatrix},$$

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix},$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

In the next three properties, let the observation vector be partitioned into two subvectors denoted by $\mathbf{y}$ and $\mathbf{x}$, where $\mathbf{y}$ is $p \times 1$ and $\mathbf{x}$ is $q \times 1$. Or, alternatively, let $\mathbf{x}$ represent some additional variables to be considered along with those in $\mathbf{y}$. Then, as in (3.45) and (3.46),

$$E\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\mu}_{\mathbf{y}} \\ \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}$$

and

$$cov\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{yy}} & \boldsymbol{\Sigma}_{\mathbf{yx}} \\ \boldsymbol{\Sigma}_{\mathbf{xy}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{pmatrix}$$

So if $\begin{pmatrix} \mathbf{y} \\ \mathbf{x} \end{pmatrix} \sim MVN_{p+q}\left(\begin{pmatrix} \boldsymbol{\mu}_{\mathbf{y}} \\ \boldsymbol{\mu}_{\mathbf{x}} \end{pmatrix}, \begin{pmatrix} \boldsymbol{\Sigma}_{\mathbf{yy}} & \boldsymbol{\Sigma}_{\mathbf{yx}} \\ \boldsymbol{\Sigma}_{\mathbf{xy}} & \boldsymbol{\Sigma}_{\mathbf{xx}} \end{pmatrix} \right)$

Independence:

(a) The subvectors $\mathbf{y}$ and $\mathbf{x}$ are independent if $\boldsymbol{\Sigma}_{yx} = \mathbf{0}$.

(b) Two individual variables y_j and y_k are independent if $\sigma_{jk} = 0$. Note that this is not true for many nonnormal random variables, as illustrated in Section 3.2.1.

(c) Further any random variables that are independent are pairwise independent, and any pairwise independent random variables are uncorrelated.

(d) Reprise: If the random variables are jointly normally distributed and uncorrelated, then they are independent. Therefore, for jointly normally distributed random variables, independence, pairwise independence, and absence of correlation are all equivalent.

Conditional distribution:

If $\mathbf{y}$ and $\mathbf{x}$ are not independent, then $\mathbf{\Sigma}_{yx} = \mathbf{0}$, and the conditional distribution of $\mathbf{y}$ given $\mathbf{x}$, $f(\mathbf{y}|\mathbf{x})$, is multivariate normal with

$$E(\mathbf{y}|\mathbf{x}) = \boldsymbol{\mu}_y + \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}(\mathbf{x} - \boldsymbol{\mu}_x)$$

and

$$cov(\mathbf{y}|\mathbf{x}) = \mathbf{\Sigma}_{yy} - \mathbf{\Sigma}_{yx}\mathbf{\Sigma}_{xx}^{-1}\mathbf{\Sigma}_{xy}$$

Note that $E(\mathbf{y}|\mathbf{x})$ is a vector of linear functions of $\mathbf{x}$, whereas $cov(y|x)$ is a matrix that does not depend on $\mathbf{x}$. The linear trend in (4.7) holds for any pair of variables. Thus to use (4.7) as a check on normality, one can examine bivariate scatter plots of all pairs of variables and look for any nonlinear trends. In (4.7), we have the justification for using the covariance or correlation to measure the relationship between two bivariate normal random variables.

As noted in Section 3.2.1, the covariance and correlation are good measures of relationship only for variables with linear trends and are generally unsuitable for nonnormal random variables with a curvilinear relationship. The matrix $\boldsymbol{\Sigma}_{yx}\boldsymbol{\Sigma}_{xx}^{-1}$ in (4.7) is called the *matrix of regression coefficients* because it relates $E(\mathbf{y}|\mathbf{x})$ to $\mathbf{x}$. The sample counterpart of this matrix appears in (10.52) when we talk about multivariate regression.

Distribution of the sum of two subvectors:

If **y** and **x** are the same size (both p $\times$ 1) and independent, then

$$\mathbf{y} + \mathbf{x} \sim MVN_p(\boldsymbol{\mu}_y + \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy} + \boldsymbol{\Sigma}_{xx}),$$

$$\mathbf{y} - \mathbf{x} \sim MVN_p(\boldsymbol{\mu}_y - \boldsymbol{\mu}_x, \boldsymbol{\Sigma}_{yy} + \boldsymbol{\Sigma}_{xx}).$$

The bivariate normal as a special case

Let

$$\mathbf{u} = \begin{pmatrix} y \\ x \end{pmatrix}$$

so that

$$E(\mathbf{u}) = \begin{pmatrix} \boldsymbol{\mu}_y \\ \boldsymbol{\mu}_x \end{pmatrix}$$

and

$$cov(\mathbf{u}) = \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \sigma_{yx} \\ \sigma_{xy}^2 & \sigma_x \end{pmatrix}$$

$$g(x, y) = f(\mathbf{y}|\mathbf{x})\, h(x)$$

$$\begin{aligned} x &= (0 \quad 1)\begin{pmatrix} y \\ x \end{pmatrix} = (0 \quad 1)\,\mathbf{u} = \mathbf{a}'\mathbf{u}, \\ z &= y - \beta x = (1, -\beta)\,\mathbf{u} = \mathbf{b}'\mathbf{u}. \end{aligned}$$

$$
\begin{aligned}
cov\,(x, z) &= cov\left(\mathbf{a}'\mathbf{u}, \mathbf{b}'\mathbf{u}\right) \\
&= \mathbf{a}'\boldsymbol{\Sigma}\mathbf{b} \\
&= \begin{pmatrix} 0 & 1 \end{pmatrix} \begin{pmatrix} \sigma_y^2 & \sigma_{xy}^2 \\ \sigma_{yx}^2 & \sigma_x^2 \end{pmatrix} \begin{pmatrix} 1 \\ -\beta \end{pmatrix} = \begin{pmatrix} \sigma_{yx}^2 & \sigma_x^2 \end{pmatrix} \begin{pmatrix} 1 \\ -\beta \end{pmatrix} \\
&= \sigma_{yx}^2 - \beta\sigma_x^2
\end{aligned}
$$

$$z = y - \frac{\sigma_{yx}}{\sigma_x^2} x$$

$$\begin{aligned} E\left(y - \frac{\sigma_{yx}}{\sigma_x^2} x\right) &= \mu_y - \frac{\sigma_{yx}}{\sigma_x^2} \mu_x \\ var\left(y - \frac{\sigma_{yx}}{\sigma_x^2} x\right) &= var\left(\mathbf{b}'\mathbf{u}\right) = \mathbf{b}'\mathbf{\Sigma}\mathbf{b} \\ &= \sigma_y^2 - \frac{\sigma_{yx}}{\sigma_x^2} \end{aligned}$$

$$
\begin{aligned}
E\left(y|x\right) &= \beta x + E\left(y - \beta x\right) = \beta x + \mu_y - \beta\mu_x \\
&= \mu_y + \frac{\sigma_{yx}}{\sigma_x^2}\left(x - \mu_x\right) \\
var\left(y|x\right) &= \sigma_y^2 - \frac{\sigma_{yx}}{\sigma_x^2}
\end{aligned}
$$

Marginals and Conditionals

Let $\mathbf{Y} = (Y_1, Y_2, Y_3, Y_4)'$ have a multivariate normal distribution with mean vector

$$\boldsymbol{\mu} = E(\mathbf{Y}) = \begin{pmatrix} 1 \\ 2 \\ 3 \\ 4 \end{pmatrix}, \text{and covariance matrix}$$

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 2 & 2 \\ 1 & 1 & 3 & 3 \\ 1 & 2 & 3 & 4 \end{pmatrix}$$

Now let $\mathbf{Q}_1 = (Y_1, Y_2)'$ and $\mathbf{Q}_2 = (Y_3, Y_4)'$. We know that both $\mathbf{Q}_1$ and $\mathbf{Q}_2$ both have bivariate normal distributions.
That is

$$\mathbf{Q}_1 = \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\right)$$

and

$$\mathbf{Q}_2 = \begin{pmatrix} Y_3 \\ Y_4 \end{pmatrix} \sim MVN_2\left(\begin{pmatrix} 3 \\ 4 \end{pmatrix}, \begin{pmatrix} 3 & 3 \\ 3 & 4 \end{pmatrix}\right)$$

So that Y_1 and Y_2 have a bivariate normal distribution with means 1 and 2 and variances 1 and 2, and covariance 1. What is the correlation coefficient r_{Y_1,Y_2} ?

$$r_{Y_1,Y_2} = \sqrt{\frac{1}{1 \times 2}} = \left(\sqrt{\frac{1}{2}}\right) = \frac{1}{\sqrt{2}}$$

So that Y_1 and Y_2 have a bivariate normal distribution with means 1 and 2 and variances 1 and 2, and covariance 1. What is the correlation coefficient r_{Y_1,Y_2} ?

$$r_{Y_1,Y_2} = \sqrt{\frac{1}{1 \times 2}} = \left(\sqrt{\frac{1}{2}}\right) = \frac{1}{\sqrt{2}}$$

We can then find the conditional distribution of Y_3 and Y_4 given Y_1 and Y_2,

$$\mathbf{\Sigma}_{11}^{-1} = \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}^{-1} = \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}$$

and

$$\begin{aligned} \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1} &= \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}^{-1} \begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix} \\ &= \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \end{aligned}$$

So that

$$\begin{aligned}\boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}(\mathbf{Q}_1 - \boldsymbol{\mu}_1) &= \begin{pmatrix} 3 \\ 4 \end{pmatrix} + \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix}\begin{pmatrix} Y_1 - 1 \\ Y_2 - 2 \end{pmatrix} \\ &= \begin{pmatrix} Y_2 + 1 \\ Y_2 + 2 \end{pmatrix}\end{aligned}$$

and

$$\begin{aligned}\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12} &= \begin{pmatrix} 3 & 3 \\ 3 & 4 \end{pmatrix} - \begin{pmatrix} 1 & 2 \\ 1 & 2 \end{pmatrix}\begin{pmatrix} 2 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 & 1 \\ 2 & 2 \end{pmatrix} \\ &= \begin{pmatrix} 1 & 1 \\ 1 & 2 \end{pmatrix}\end{aligned}$$

$$\mathbf{Q}_2|\mathbf{Q}_1 \sim MVN_2\left(\left(\begin{array}{c} Y_2+1 \\ Y_2+2 \end{array}\right),\left(\begin{array}{cc} 1 & 1 \\ 1 & 2 \end{array}\right)\right)$$

That is conditionally on Y_1 and Y_2, Y_3 and Y_4, have a bivariate normal distribution with mean $Y_2{+}1$ and $Y_2{+}2$, variances 1 and 2 and covariance 1.

Note that the previous holds with any rearrangement of the random variables. So that if

$$\begin{pmatrix} Y_1 \\ Y_3 \\ Y_2 \\ Y_4 \end{pmatrix} \sim MVN_4\left(\begin{pmatrix} 1 \\ 3 \\ 2 \\ 4 \end{pmatrix}, \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 3 & 2 & 3 \\ 1 & 2 & 2 & 2 \\ 1 & 3 & 2 & 4 \end{pmatrix}\right)$$

Outline

The multivariate normal likelihood, with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is given by:

$$\mathscr{L}(\mathbf{y}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{y}_i-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i-\boldsymbol{\mu})} \tag{2}$$

which is maximised taking $\hat{\boldsymbol{\Sigma}} = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{y}_i - \bar{\mathbf{y}})(\mathbf{y}_i - \bar{\mathbf{y}})'$ and $\hat{\boldsymbol{\mu}} = \bar{\mathbf{y}}$. A little algebra shows that $\sum_{i=1}^{n}(\mathbf{y}_i - \hat{\boldsymbol{\mu}})'\hat{\boldsymbol{\Sigma}}^{-1}(\mathbf{y}_i - \hat{\boldsymbol{\mu}}) = np$ so we can express the maximum as:

$$max\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\boldsymbol{\Sigma}}|^{n/2}} e^{-np/2} \tag{3}$$

Under the hypothesis $H_0 : \boldsymbol{\mu} = \boldsymbol{\mu}_0$

$$\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\boldsymbol{\Sigma}|^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_0)\boldsymbol{\Sigma}^{-1}(\mathbf{y}_i - \boldsymbol{\mu}_0)'} \qquad (4)$$

so $\boldsymbol{\mu}_0$ is now fixed but $\boldsymbol{\Sigma}$ can be varied to find the most likely value. As before, this can be rearranged to give:

$$max\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{np/2}|\hat{\boldsymbol{\Sigma}}_0|^{n/2}} e^{-np/2} \tag{5}$$

where $\hat{\boldsymbol{\Sigma}}_0 = \frac{1}{n}\sum_{i=1}^{n}(\mathbf{x}_i - \boldsymbol{\mu}_0)(\mathbf{x}_i - \boldsymbol{\mu}_0)'$

To determine whether $\boldsymbol{\mu}_0$ is plausible, we wish to compare $\mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})$ with $\mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, conventionally performed by means of the likelihood ratio statistic:

$$\Lambda = \frac{max \mathscr{L}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma})}{max \mathscr{L}(\boldsymbol{\mu}, \boldsymbol{\Sigma})} = \left(\frac{|\hat{\boldsymbol{\Sigma}}|}{|\hat{\boldsymbol{\Sigma}}_0|} \right)^{n/2} \tag{6}$$