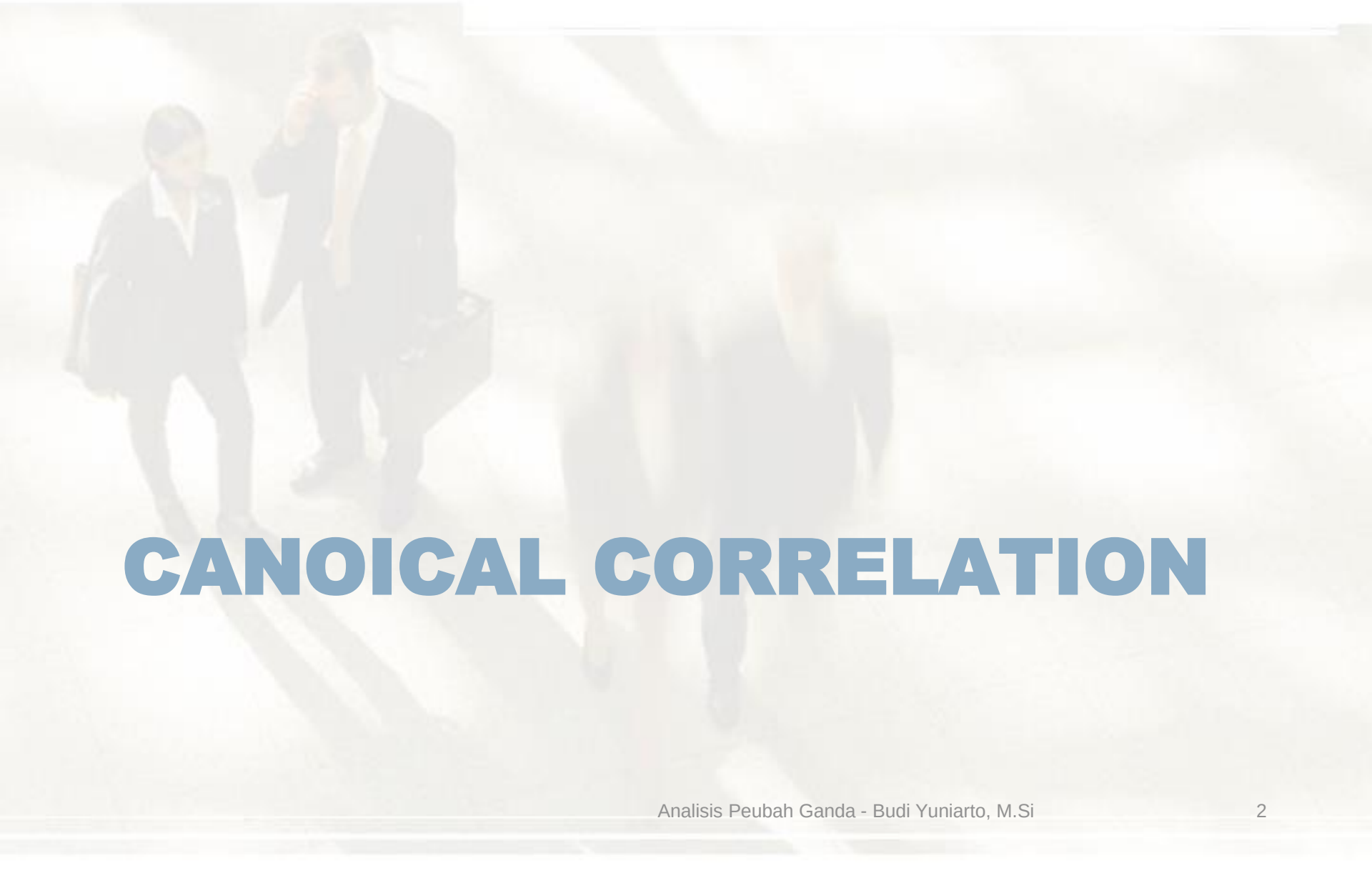
# Analisis Peubah Ganda

Pertemuan XVI



# CANOICAL CORRELATION

# introduction

- We often measure two types of variables on each research unit—for example:
  - a set of aptitude variables and a set of achievement variables,
  - a set of personality variables and a set of ability measures,
  - a set of price indices and a set of production indices,
  - a set of student behaviors and a set of teacher behaviors,
  - a set of psychological attributes and a set of physiological attributes,
  - a set of ecological variables and a set of environmental variables,
  - a set of academic achievement variables and a set of measures of job success

- Canonical correlation analysis is concerned with the amount of (linear) relationship between two sets of variables.
- Canonical correlation analysis seeks to identify and quantify the associations between two sets of variables.
- Canonical correlation analysis focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in another set.
- The pairs of linear combinations are called the canonical variables, and their correlations are called canonical correlations.

# canonical correlations and canonical variates

- Canonical correlation is an extension of multiple correlation, which is the correlation between one *y* and several *x*'s.
- Canonical correlation analysis is often a useful complement to a multivariate regression analysis.
- Assume that two sets of variables $\mathbf{y} = (y_1, y_2, \ldots, y_p)$ and $\mathbf{x} = (x_1, x_2, \ldots, x_q)$ are measured on the same sampling unit.
- We first review multiple correlation.

- The sample covariances and correlations among $y$, $x_1$, $x_2$, . . . , $x_q$ can be summarized in the matrices

$$\mathbf{S} = \begin{pmatrix} s_y^2 & \mathbf{s}_{yx}' \\ \mathbf{s}_{yx} & \mathbf{S}_{xx} \end{pmatrix},$$

$$\mathbf{R} = \begin{pmatrix} 1 & \mathbf{r}_{yx}' \\ \mathbf{r}_{yx} & \mathbf{R}_{xx} \end{pmatrix},$$

- From multiple regression analysis we know that the squared multiple correlation between $y$ and the $x$'s can be computed from the partitioned covariance matrix or correlation matrix above as follows:

$$R^2 = \frac{\mathbf{s}_{yx}' \mathbf{S}_{xx}^{-1} \mathbf{s}_{yx}}{s_y^2} = \mathbf{r}_{yx}' \mathbf{R}_{xx}^{-1} \mathbf{r}_{yx}.$$

- In $R^2$, the $q$ covariances between $y$ and the $x$'s in $\mathbf{s}_{yx}$ or the $q$ correlations between $y$ and the $x$'s in $\mathbf{r}_{yx}$ are channeled into a single measure of linear relationship between $y$and the $x$'s.

$$\mathbf{R} = \left(\begin{array}{c|ccc} 1 & r_{y1} & r_{y2} & \cdots & r_{yq} \\ \hline r_{1y} & 1 & r_{12} & \cdots & r_{1q} \\ \vdots & \vdots & \vdots & & \vdots \\ r_{qy} & r_{q1} & r_{q2} & \cdots & 1 \end{array}\right) = \begin{pmatrix} 1 & \mathbf{r}'_{yx} \\ \mathbf{r}_{yx} & \mathbf{R}_{xx} \end{pmatrix}.$$

- The multiple correlation $R$ can be defined alternatively as the maximum correlation between $y$ and a linear combination of the $x$'s; that is, $R = \max_{\mathbf{b}} r_{y,\mathbf{bx}}$.

- We now return to the case of several *y*'s and several *x*'s.
- The covariance structure associated with two subvectors **y** and **x**. The overall sample covariance matrix for $(y_1, \ldots, y_p, x_1, \ldots, x_q)$ can be partitioned as

$$\mathbf{S} = \begin{pmatrix} \mathbf{S}_{yy} & \mathbf{S}_{yx} \\ \mathbf{S}_{xy} & \mathbf{S}_{xx} \end{pmatrix},$$

- where $\mathbf{S}_{yy}$ is the $p \times p$ sample covariance matrix of the *y*'s, $\mathbf{S}_{yx}$ is the $p \times q$ matrix of sample covariances between the *y*'s and the *x*'s, and $\mathbf{S}_{xx}$ is the $q \times q$ sample covariance matrix of the *x*'s.

- In multiple regression analysis, we have several measures of association between the *y*'s and the *x*'s. The first of these is defined as

$$R_M^2 = |\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}|/|\mathbf{S}_{yy}|$$
$$R_M^2 = |\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}|.$$

- For any square matrix **A** with eigenvalues $\lambda_1, \lambda_2, \ldots, \lambda_n$, we have $|\mathbf{A}| = \prod_{i=1}^{n} \lambda_i.$
- Then $R_M^2 = |\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}_{xy}| = \prod_{i=1}^{s} r_i^2,$
- where $s = \min(p, q)$ and $r_1^2, r_2^2, \ldots, r_s^2$ are the eigenvalues of $\mathbf{S}_{yy}^{-1}\mathbf{S}_{yx}\mathbf{S}_{xx}^{-1}\mathbf{S}xy$.
- The square roots of the eigenvalues, $r_1, r_2, \ldots, r_s$ , are called *canonical correlations*

# alternative approach

- Suppose *u* is a linear combination of the *y*'s, *u* = **a'y**, and *v* is a linear combination of the *x*'s, *v* = **b'x**;
- Then

  Var(U) = **a'** Cov(**Y**) **a** = $\mathbf{a'S_{11}a}$

  Var(V) = **b'** Cov(**X**) **b** = $\mathbf{b'S_{22}b}$

  Cov(U,V) = **a'** Cov(**Y,X**) **b** = $\mathbf{a'S_{12}b}$

- We shall seek coefficient vectors **a** and **b** such that

$$Corr(U,V) = \frac{\boldsymbol{a'S_{12}b}}{\sqrt{\boldsymbol{a'S_{11}a}}\sqrt{\boldsymbol{b'S_{22}b}}}$$

  is as large as possible.

- We define the following:
  - The first pair of canonical variables, or first canonical variate pair, is the pair of linear combinations $U_1$, $V_1$ having unit variances, which maximize the correlation;
  - The second pair of canonical variables, or second canonical variate pair, is the pair of linear combinations $U_2$, $V_2$ having unit variances, which maximize the correlation among all choices that are uncorrelated with the first pair of canonical variables.
  - At the k-th step, the k-th pair of canonical variables, or k-th canonical variate pair, is the pair of linear combinations $U_k$, $V_k$ having unit variances, which maximize the correlation among all choices uncorrelated with the previous k-1 canonical variable pairs.

- Then,

$$max_{a,b}\ Corr(U, V) = \rho_1$$

attained by the linear combinations (first canonical variate pair)

$$U_1 = \underbrace{\mathbf{e'_1 S_{11}^{-1/2}}}_{\mathbf{a'_1}}\mathbf{Y} \qquad V_1 = \underbrace{\mathbf{f'_1 S_{22}^{-1/2}}}_{\mathbf{b'_1}}\mathbf{X}$$

The *k*-th pair of canonical variates, k = 2, 3,..., p,

$$U_k = \underbrace{\mathbf{e'_k S_{11}^{-1/2}}}_{\mathbf{a'_k}}\mathbf{Y} \qquad V_k = \underbrace{\mathbf{f'_k S_{22}^{-1/2}}}_{\mathbf{b'_k}}\mathbf{X}$$

maximizes

$$max_{a,b}\ Corr(U_k, V_k) = \rho_k$$

# properties of canonical correlations

Two interesting properties of canonical correlations are the following:

1. Canonical correlations are invariant to changes of scale on either the *y*'s or the *x*'s. For example, if the measurement scale is changed from inches to centimeters, the canonical correlations will not change (the corresponding eigenvectors will change). This property holds for simple and multiple correlations as well.
2. The first canonical correlation $r_1$ is the maximum correlation between linear functions of **y** and **x**. Therefore, $r_1$ exceeds (the absolute value of) the simple correlation between any *y* and any *x* or the multiple correlation between any *y* and all the *x*'s or between any *x* and all the *y*'s.

# properties of canonical variates

$$
\begin{aligned}
\operatorname{Var}(U_k) &= \operatorname{Var}(V_k) = 1 \\
\operatorname{Cov}(U_k, U_\ell) &= \operatorname{Corr}(U_k, U_\ell) = 0 \quad k \neq \ell \\
\operatorname{Cov}(V_k, V_\ell) &= \operatorname{Corr}(V_k, V_\ell) = 0 \quad k \neq \ell \\
\operatorname{Cov}(U_k, V_\ell) &= \operatorname{Corr}(U_k, V_\ell) = 0 \quad k \neq \ell
\end{aligned}
$$

# tests of significance

basic tests of significance associated with canonical correlations:

1. Tests of No Relationship between the y's and the x's
   - Test of $H0 : \sum_{yx} = \mathbf{O}$, independence of two sets of variables.
   - Test of $H0 : \mathbf{B}_1 = \mathbf{O}$, significance of overall multivariate multiple regression.
   - Test of significance of the canonical correlations.

2. Test of Significance of Succeeding Canonical Correlations after the First

# test of significance of the canonical correlations

- the significance of $r_1, r_2, \ldots, r_s$ can be tested by

$$\Lambda_1 = \frac{|\mathbf{S}|}{|\mathbf{S}_{yy}||\mathbf{S}_{xx}|} = \frac{|\mathbf{R}|}{|\mathbf{R}_{yy}||\mathbf{R}_{xx}|},$$

  which is distributed as $\Lambda_{p,q,n-1-q}$

- We reject $H_0$ if $\Lambda_1 \leq \Lambda_\alpha$.

- If the parameters exceed the range of critical values for Wilks', we can use the $\chi^2$-approximation.

$$\chi^2 = -\left[n - \tfrac{1}{2}(p+q+3)\right] \ln \Lambda_1,$$

- Alternatively, we can use the *F*-approximation

$$F = \frac{1-\Lambda_1^{1/t}}{\Lambda_1^{1/t}} \frac{\mathrm{df}_2}{\mathrm{df}_1},$$

$$\mathrm{df}_1 = pq, \qquad \mathrm{df}_2 = wt - \tfrac{1}{2}pq + 1,$$

$$w = n - \tfrac{1}{2}(p+q+3), \qquad t = \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}}.$$

# test of significance of succeeding canonical correlations after the first

- $\Lambda_1$ is expressible in terms of the squared canonical correlations:

$$\Lambda_1 = \prod_{i=1}^{s}(1 - r_i^2).$$

- If the test significance of the canonical correlations based on all *s* canonical correlations rejects $H_0$, we are not sure if the canonical correlations beyond the first are significant.
- To test the significance of $r_2, \ldots, r_s$ , we delete $r_1^2$ from $\Lambda_1$ in to obtain

$$\Lambda_2 = \prod_{i=2}^{s}(1 - r_i^2).$$

- If this test rejects the hypothesis, we conclude that at least *r*2 is significantly different from zero.

- We can continue in this manner, testing each $r_i$ in turn, until a test fails to reject the hypothesis. At the $k$-th step, the test statistic is

$$\Lambda_k = \prod_{i=k}^{s}(1 - r_i^2),$$

  which is distributed as $p-k+1, q-k+1, n-k-q$
- The usual $\chi 2$- and $F$-approximations can also be applied to $\Lambda_k$.

# interpretation

Three common tools for interpretation of canonical variates:

1) standardized coefficients,

   The standardized coefficients show the contribution of the variables in the presence of each other.

2) the correlation between each variable and the canonical variate,

   Such correlations are sometimes referred to as *loadings* or *structure coefficients*, and it is widely claimed that they provide a more valid interpretation of the canonical variates → no information about how the *y*'s contribute jointly

2) rotation of the canonical variate coefficients.

   In an attempt to improve interpretability, the canonical variate coefficients can be rotated to increase the number of high and low coefficients and reduce the number of intermediate ones. → not recommended

# standardized coefficient

- The coefficients in the canonical variates $u_i = \mathbf{a_i}'\mathbf{y}$ and $v_i = \mathbf{b}_i'\mathbf{x}$ reflect differences in scaling of the variables as well as differences in contribution of the variables to canonical correlation.
- To remove the effect of scaling, $\mathbf{a}_i$ and $\mathbf{b}_i$ can be standardized by multiplying by the standard deviations of the corresponding variables as

$$\mathbf{c}_i = \mathbf{D}_y\mathbf{a}_i, \qquad \mathbf{d}_i = \mathbf{D}_x\mathbf{b}_i,$$

where $\mathbf{D}_y = \text{diag}(s_{y1}, s_{y2}, \ldots, s_{yp})$ and $\mathbf{D}_x = \text{diag}(s_{x1}, s_{x2}, \ldots, s_{xq})$

# Example

The results of a planned experiment involving a chemical reaction are given in Table in the next page (Box and Youle 1955).

The input (independent) variables are

$x_1$ = temperature, $x_2$ = concentration, $x_3$ = time.

The yield (dependent) variables are

$y_1$ = percentage of unchanged starting material,

$y_2$ = percentage converted to the desired product,

$y_3$ = percentage of unwanted by-product.

**Table 10.1. Chemical Reaction Data**

| Experiment Number | Yield Variables | | | Input Variables | | |
|---|---|---|---|---|---|---|
| | $y_1$ | $y_2$ | $y_3$ | $x_1$ | $x_2$ | $x_3$ |
| 1 | 41.5 | 45.9 | 11.2 | 162 | 23 | 3 |
| 2 | 33.8 | 53.3 | 11.2 | 162 | 23 | 8 |
| 3 | 27.7 | 57.5 | 12.7 | 162 | 30 | 5 |
| 4 | 21.7 | 58.8 | 16.0 | 162 | 30 | 8 |
| 5 | 19.9 | 60.6 | 16.2 | 172 | 25 | 5 |
| 6 | 15.0 | 58.0 | 22.6 | 172 | 25 | 8 |
| 7 | 12.2 | 58.6 | 24.5 | 172 | 30 | 5 |
| 8 | 4.3 | 52.4 | 38.0 | 172 | 30 | 8 |
| 9 | 19.3 | 56.9 | 21.3 | 167 | 27.5 | 6.5 |
| 10 | 6.4 | 55.4 | 30.8 | 177 | 27.5 | 6.5 |
| 11 | 37.6 | 46.9 | 14.7 | 157 | 27.5 | 6.5 |
| 12 | 18.0 | 57.3 | 22.2 | 167 | 32.5 | 6.5 |
| 13 | 26.3 | 55.0 | 18.3 | 167 | 22.5 | 6.5 |
| 14 | 9.9 | 58.9 | 28.0 | 167 | 27.5 | 9.5 |
| 15 | 25.0 | 50.3 | 22.1 | 167 | 27.5 | 3.5 |
| 16 | 14.1 | 61.1 | 23.0 | 177 | 20 | 6.5 |
| 17 | 15.2 | 62.9 | 20.7 | 177 | 20 | 6.5 |
| 18 | 15.9 | 60.0 | 22.1 | 160 | 34 | 7.5 |
| 19 | 19.6 | 60.6 | 19.3 | 160 | 34 | 7.5 |