UNIT 26 CANONICAL CORRELATION ANALYSIS

Structure **Page No.**

26.1 INTRODUCTION

Canonical correlation is a technique to identify and quantify the association between two sets of variables. Each set can contain several variables. Simple and multiple correlations are special cases of canonical correlation in which one or both sets contain a single variable. This technique was given by H. Hotelling in 1935-36, for relating the arithmetic speed and arithmetic power to reading speed and reading power based on a sample data received from 140 seventh grade students. Other examples where canonical correlations can be helpful are: relating governmental policy variables with economic goal variables; relating college performance variables (grades in courses in different subjects) with pre-college achievement variables (percentage of marks in high school, number of extracurricular activities in height school, etc.); relating yield attributing parameters (test weight, plant height, number of grains per panicle, etc.) and quality parameters (protein content, carbohydrate content, etc.) in case of a certain crop; relating job satisfaction variables (supervisor satisfaction, workload satisfaction, general satisfaction, etc.) and job characteristic variables (feedback, task identity, task variety, etc.); relating physiological variables (weight in kg, waist in inches, pulse rate, etc.) with exercise variables (number of sit ups, jumps, etc.) and many others such pairs.

Canonical correlation analysis actually focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in the second set. The idea is first to determine the pair of linear combinations having the largest correlation. Next we determine the pair of linear combinations having the largest correlation among all pairs uncorrelated with the initially selected pair. This process continues until the number of pairs of canonical variables equals the number of variables in the smaller group. The pairs of linear combinations are called the **canonical variables** and their correlations are called **canonical correlations**. The canonical correlations measure the strength of association between the two sets of variables. The maximization aspect of the technique represents an attempt to concentrate a high-dimensional relationship between two sets of variables into a few pair of canonical variables.

The purpose of canonical correlation is to explain the relation of the two sets of variables, not to model the individual variables.

Analogous with ordinary correlation, canonical correlation squared is the percent of variance in the dependent set explained by the independent set of variables along a given dimension (there may be more than one). In addition to asking how strong the relationship is between two latent variables, canonical correlation is useful in determining how many dimensions are needed to account for that relationship. Canonical correlation finds the linear combination of variables that produces the largest correlation with the second set of variables. This linear combination, or "root," is extracted and the process is repeated for the residual data, with the constraint that the second linear combination of variables must not correlate with the first one. The process is repeated until a successive linear combination is no longer significant.

Canonical correlation is a member of the multiple general linear hypothesis (MLGH) family and shares many of the assumptions of mutliple regression such as linearity of relationships, homoscedasticity (same level of relationship for the full range of the data), interval or near-interval data, untruncated variables, proper specification of the model, lack of high multicollinearity, and multivariate normality for purposes of hypothesis testing.

Often in applied research, scientists encounter variables of large dimensions and are faced with the problem of understanding dependency structures, reduction of dimensionalities, construction of a subset of good predictors from the explanatory variables, etc. Canonical correlation Analysis (CCA) provides us with a tool to attack these problems. However, its appeal and hence its motivation seem to differ from the theoretical statisticians to the social scientists. We deal here with the various motivations of CCA as mentioned above and related statistical inference procedures. We shall begin the unit by discussing the canonical correlation in Section 26.2. In Section 26.3, we shall focus on non linear canonical correlation

.

Objectives

After reading this unit, you should be able to

- understand the meaning and concept of canonical correlation
- interpret the results of a canonical correlation analysis
- make use of canonical correlation

26.2 MATHEMATICAL FORMULATION AND COMPUTATIONS

Consider two groups of variables. The first set of *p* variables is represented by a $(p\times 1)$ random vector **X** and the second set of *q* variables is represented by the $(q\times 1)$ random vector **Y**. Without loss of generality, we assume that $p \le q$. For two random vectors **X** and **Y**, let $E(X) = \mu_X$; $D(\mathbf{X}) = \Sigma_{11}$; $E(Y) = \mu_Y$; $D(\mathbf{Y}) = \Sigma_{22}$; Cov $(\mathbf{X}, \mathbf{Y}) = \Sigma_{12}$. Here E (.) denotes expectation and D(.) denotes the variance-covariance matrix. For the sake of convenience, let us consider **X** and **Y** jointly as the random vector

$\mathbf{Z} = \begin{bmatrix} \mathbf{X} \\ \cdots \\ \mathbf{Y} \end{bmatrix}$ with mean vector $\mu = E(Z) = \begin{bmatrix} \mu_X \\ \cdots \\ \mu_Y \end{bmatrix}$ and variance covariance matrix as

$D(Z) = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$. The pq elements in $\mathbf{\Sigma}_{12}$ gives the measures of association between *p* variables of first set and *q* variables of second set. The canonical correlation analysis summarizes the associations between **X** and **Y** in terms of few correlations rather than the pq elements in $\mathbf{\Sigma}_{12}$.

Consider $U = a'X$ and $V = b'Y$ as the linear combinations of variables within the two sets for some coefficient vectors $\mathbf{a}$ and $\mathbf{b}$. Now, using simple statistical computations, one can see that

$$\text{Var}(U) = a'\Sigma_{11}a\,;\ \text{Var}(V) = b'\Sigma_{22}b \ \text{ and } \ \text{Cov}\big(U,V\big) = \mathbf{a'\Sigma_{12}b}\,.$$

As a consequence

$\text{Corr}(U,V) = \dfrac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}a}\sqrt{b'\Sigma_{22}b}}$. Through canonical correlation analysis we want to find $\mathbf{a}$ and $\mathbf{b}$ such that $\text{Corr}(U,V)$ is as large as possible.

The first linear combination of two sets of variables known as first pair of canonical variables is

$U_1 = a_{11}x_1 + a_{21}x_2 + ... + a_{p1}x_p = \mathbf{a}_1'\mathbf{X}$

$V_1 = b_{11}y_1 + b_{21}y_2 + ... + b_{q1}y_q = \mathbf{b}_1'\mathbf{Y}$.

The coefficients are so chosen that the two canonical variables, U_1 and V_1 have unit variances and have the largest possible correlation. This maximized correlation between the two canonical variables is the first canonical correlation. The coefficients of the linear combinations are canonical coefficients or canonical weights. Let the maximum correlation attained in first pair of canonical variables be $\text{Corr}(U_1, V_1) = \rho_1$.

The second set of canonical variables, uncorrelated with the first pair of canonical variables is

$U_2 = a_{12}x_1 + a_{22}x_2 + ... + a_{p2}x_{p=} \ \mathbf{a}_2'\mathbf{X}$

$V_2 = b_{12}y_1 + b_{22}y_2 + ... + b_{q2}y_q = \mathbf{b}_2'\mathbf{Y}$.

The coefficients are so chosen that the two canonical variables, U_2 and V_2; U_2 is uncorrelated with U_1 and V_1, V_2 is uncorrelated with U_1 and V_1, and U_2 and V_2 have the largest possible correlation subject to these constraints. Let the maximum correlation attained in second pair of canonical variables be $\text{Corr}(U_2, V_2) = \rho_2$.

This process continues until the number of pairs of canonical variables is equal to the number of variables in the set containing smaller number of variables. In this case number of canonical variables will be p as $p \le q$. Let the maximum correlation attained in p^th^ pair of canonical variables be $\text{Corr}(U_p, V_p) = \rho_p$.

Now the question arises, how to compute $\rho_1, \rho_2, \ldots, \rho_p$.

Through algebraic manipulations, it can easily be seen that the eigenvalues of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ are $\rho_1^2 \ge \rho_2^2 \ge \ldots \ge \rho_p^2$ with corresponding $p \times 1$ eigenvectors as $a_1, a_2, \ldots, a_p$. Similarly the eigenvalues of $\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma_{21}}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}$ are $\rho_1^2 \ge \rho_2^2 \ge \ldots \ge \rho_p^2$ with corresponding $q \times 1$ eigenvectors as $b_1, b_2, \ldots, b_p$. Therefore, the maximum correlation, ρ_1 is the positive square root of the largest eigenvalue of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ or $\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma_{21}}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}$. Similarly the maximum correlation between second pair of canonical variables ρ_2 is the positive square root of the second largest eigenvalue of $\mathbf{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ or $\mathbf{\Sigma}_{22}^{-1/2}\mathbf{\Sigma_{21}}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}\mathbf{\Sigma}_{22}^{-1/2}$. The canonical variables are obtained by using the coefficient vectors as corresponding ortho-normalized eigenvectors. This process continues till we get p^th^ pair of canonical variables and p canonical correlations. In the above discussion, it has been tacitly assumed that Σ_{ii}, i = 1,2 is non-singular. If Σ_{ii}, i=1 or 2 happens to be singular, one can use a g-inverse of Σ_{ii} in place of true inverse of Σ_{ii}.

Remark 1: $p = q = 1 \Rightarrow \rho_1$= usual Pearson's product moment correlation coefficient between the scalar random variables X and Y; p = 1, q >1 $\Rightarrow \rho_1$= Multiple correlation coefficient between the scalar X and the vector random variable **Y**. Sample analogues are trivially defined.

Remark 2: The canonical correlation also reduces the dimensionality. One feature of dimensionality reduction is that we have only 2p canonical variables rather than p+q original variables. Further, reduction in dimensionality can be achieved by retaining only those pair of canonical variables, whose correlation coefficient is significantly different from zero. If the correlation coefficient between (k+1)^th^ pair of canonical variables is not significantly different from zero, then we can retain only first k pairs of canonical variables. This gives only 2k variables rather than p+q original variables.

Remark 3: The canonical correlations measure the linear association between two sets of variables. If the variables are associated in a nonlinear manner, the association can not be captured by canonical correlation. Thus, while canonical correlation analysis is the most generalized multivariate method, it is still constrained to identifying linear associations. Canonical correlation analysis can accommodate any metric variable without the strict assumption of normality. Normality is desirable because it standardizes a distribution to allow for a higher correlation among the variables. But in the strictest sense, canonical correlation analysis can accommodate even non-normal variables if the distributional form (e.g., highly skewed) does not decrease the correlation with other variables. This allows for transformed nonmetric data (in the form of dummy variables) to be used as well. However, multivariate normality is required for the statistical inference test of the significance of each canonical function. Because tests for multivariate normality are not readily available, the prevailing guideline is to ensure that each variable has univariate normality. Thus, although normality is not strictly required, it is highly recommended that all variables be evaluated for normality and transformed if necessary. Homoscedasticity, to the extent that it decreases the correlation between variables, should also be remedied. Finally, multicollinearity among either variable set will confound the ability of the technique to isolate the impact of any single variable, making interpretation less reliable.

Now, we shall discuss nonlinear canonical correlation in the following Section.

26.3 NONLINEAR CANONICAL CORRELATION

Nonlinear canonical correlation analysis corresponds to categorical canonical correlation analysis with optimal scaling. The OVERALS procedure in SPSS (part of SPSS Categories) implements nonlinear canonical correlation. Independent variables can be nominal, ordinal, or interval, and there can be more than two sets of variables (more than one independent set and one dependent set). Whereas ordinary canonical correlation maximizes correlations between the variable sets, in OVERALS the sets are compared to an unknown compromise set defined by the object scores

OVERALS makes use of optimal scaling, which quantifies categorical variables and then treats as numerical variables, including applying nonlinear transformations to find the best-fitting model. For nominal variables, the order of the categories is not retained but values are created for each category such that goodness of fit is maximized. For ordinal variables, order is retained and values maximizing fit are created. For interval variables, order is retained as are equal distances between values.

Obtain OVERALS from the SPSS menu by selecting Analyze, Data Reduction, Optimal Scaling; Select Multiple sets; Select either Some variable(s) not multiple nominal or All variables multiple nominal; click Define; define at least two sets of variables; define the value range and measurement scale (optimal scaling level) for each selected variable. SPSS output includes frequencies, centroids, iteration history, object scores, category quantifications, weights, component loadings, single and multiple fit, object scores plots, category coordinates plots, component loadings plots, category centroids plots, and transformation plots.

Tip: To minimize output, use the Automatic Recode facility on the Transform menu to create consecutive categories beginning with 1 for variables treated as nominal or ordinal. To minimize output, for each variable scaled at the numerical (integer) level, subtract the smallest observed value from every value and add 1.

Warning: Optimal scaling recodes values on the fly to maximize goodness of fit for the given data. As with any atheoretical, post-hoc data mining procedure, there is a danger of overfitting the model to the given data. Therefore, it is particularly appropriate to employ cross-validation, developing the model for a training dataset and then assessing its generalizability by running the model on a separate validation dataset.

The SPSS manual notes, "If each set contains one variable, nonlinear canonical correlation analysis is equivalent to principal components analysis with optimal scaling. If each of these variables is multiple nominal, the analysis corresponds to homogeneity analysis. If two sets of variables are involved and one of the sets contains only one variable, the analysis is identical to categorical regression with optimal scaling."

Redundancy is the percent of variance in one set of variables accounted for by the variate of the other set. The researcher wants high redundancy, indicating that independent variate accounts for a high percent of the variance in the dependent set of original variables. Note this is not the canonical correlation squared, which the percent of variance in the dependent variate is accounted for by the independent variate.

Applications of Canonical Correlation Analysis

- There could be a situation where some of variables have high structure correlations even though their canonical weights are near zero. This could happen because the weights are partial coefficients whereas the structure correlations (canonical factor loadings) are not: if a given variable shares variance with other independent variables entered in the linear combination of variables used to create a canonical variable, its canonical coefficient (weight) is computed based on the residual variance it can explain after controlling for these variables. If an independent variable is totally redundant with another independent variable, its partial coefficient (canonical weight) will be zero. Nonetheless, such a variable might have a high correlation with the canonical variable (that is, a high structure coefficient). In summary, the canonical weights have to do with the unique contributions of an original variable to the canonical variable, whereas the structure correlations have to do with the simple, overall correlation of the original variable with the canonical variable.
- Canonical correlation is not a measure of the percent of variance explained in the original variables. The square of the structure correlation is the percent of the variance in a given original variable accounted for by a given canonical variable on a given (usually the first) canonical correlation. Note that the average percent of variance explained in the original variables by a canonical variable (the mean of the squared structure correlations for the canonical variable) is not at all the same as the canonical correlation, which has to do with the correlation between the weighted sums of the two sets of variables. Put another way, the canonical correlation does not tell us how much of the variance in the original variables is explained by the canonical variables. Instead, that is determined on the basis of the squares of the structure correlations.
- Canonical coefficients can be used to explain with which original variables a canonical correlation is predominantly associated. The canonical coefficients are standardized coefficients and (like beta weights in regression) their magnitudes can be compared. Looking at the columns in SPSS output which list the canonical coefficients as columns and the variables in a set of variables as rows, some researchers simply note variables with the highest coefficients to determine which variables are associated with which canonical correlations and use this as the basis for inducing the meaning of the dimension represented by the canonical correlation.

However, Levine (1977) argues against the procedure above on the ground that the canonical coefficients may be subject to multi-collinearity, leading to incorrect judgments. Also, because of suppression, a canonical coefficient may even have a different sign compared to the correlation of the original variable with the canonical variable. Therefore, instead, Levine recommends interpreting the relations of the original variables to a canonical variable in terms of the correlations of the original variables with the canonical variables - that is, by structure coefficients. This is now the standard approach.

Canonical correlation places the fewest restrictions on the types of data on which it operates. Because the other techniques impose more rigid restrictions, it is generally believed that the information obtained from them is of higher quality and may be presented in a more interpretable manner. For this reason, many researchers view canonical correlation as a last-ditch effort, to be used when all other higher-level techniques have been exhausted. But in situations with multiple dependent and independent variables, canonical correlation is the most appropriate and powerful multivariate technique. It has gained acceptance in many fields and represents a useful tool for multivariate analysis, particularly as interest has spread to considering multiple dependent variables.

Now try the following exercises.

E1) Following data was collected on physiological variables (weight in pounds, waist in inches and pulse rate) and exercise variables (chins, situps and jumps) on middle-aged men in a fitness club. Perform the canonical correlation analysis and interpret your results.

Weight	Waist	Pulse	Chins	Situps	Jumps
191	36	50	5	162	60
189	37	52	2	110	60
193	38	58	12	101	101
162	35	62	12	105	37
189	35	46	13	155	58
182	36	56	4	101	42
211	38	56	8	101	38
167	34	60	6	125	40
176	31	74	15	200	40
154	33	56	17	251	250
169	34	50	17	120	38
166	33	52	13	210	115
154	34	64	14	215	105
247	46	50	1	50	50
193	36	46	6	70	31
202	37	62	12	210	120
176	37	54	4	60	25
157	32	52	11	230	80
156	33	54	15	225	73
138	33	68	2	110	43

Source: http://v8doc.sas.com/sashtml/

E2) The variance covariance matrix between 5 yield attributing parameters (X1, X2, X3, X4 and X5) and four quality attributes (Y1, Y2, Y3, Y4) based on a sample of size 200 is given as (after standardizing the variables)

$$
\begin{array}{c|ccccccccc}
 & X1 & X2 & X3 & X4 & X5 & Y1 & Y2 & Y3 & Y4 \\
\hline
X1 & 1.000 & 0.754 & -0.690 & -0.440 & 0.702 & -0.605 & -0.480 & 0.780 & -0.152 \\
X2 & 0.754 & 1.000 & -0.710 & -0.515 & 0.412 & -0.722 & -0.419 & 0.542 & -0.100 \\
X3 & -0.690 & -0.710 & 1.000 & 0.323 & -0.444 & 0.737 & 0.361 & -0.546 & 0.172 \\
X4 & -0.440 & -0.515 & 0.323 & 1.000 & -0.334 & 0.527 & 0.461 & -0.393 & -0.019 \\
X5 & 0.702 & 0.412 & -0.444 & -0.334 & 1.000 & -0.383 & -0.505 & 0.737 & -0.148 \\
Y1 & -0.605 & -0.722 & 0.737 & 0.527 & -0.383 & 1.000 & 0.251 & -0.490 & 0.250 \\
Y2 & -0.480 & -0.419 & 0.361 & 0.461 & -0.505 & 0.251 & 1.000 & -0.434 & -0.079
\end{array}
$$

Y3	0.780	0.542	-0.546	-0.393	0.737	-0.490	-0.434	1.000	-0.163
Y4	-0.152	-0.100	0.172	-0.019	-0.148	0.250	-0.079	-0.163	1.000

Perform canonical correlation analysis and interpret your results.

E3) Consider the following variance-covariance matrix

$$\begin{array}{c|cccc} & X1 & X2 & Y1 & Y2 \\ \hline X1 & 100 & 0 & 0 & 0 \\ X2 & 0 & 1 & 0.95 & 0 \\ Y1 & 0 & 0.95 & 1 & 0 \\ Y2 & 0 & 0 & 0 & 100 \end{array}$$

Obtain the correlation coefficient between first pair of canonical variables.

Now, let us summarize the unit.

26.4 SUMMARY

In this unit, we have covered the following points.

1) **Canonical correlation** is a technique to identify and quantify the association between two sets of variables.

2) Canonical correlation analysis actually focuses on the correlation between a linear combination of the variables in one set and a linear combination of the variables in the second set.

3) Simple and multiple correlations are special cases of canonical correlation in which one or both sets contain a single variable.

4) The canonical correlation also reduces the dimensionality.

5) Nonlinear canonical correlation analysis corresponds to categorical canonical correlation analysis with optimal scaling.

26.5 SOLUTIONS/ANSWERS

E1) Step 1: Obtain the variance covariance matrix of weight, waist, pulse rate, chins, situps and jumps.

Covariance Matrix

	Weight	**Waist**	**Pulse**	**Chins**	**Situps**	**Jumps**
Weight	609.621053	68.800000	-65.115789	-50.863158	-761.715789	-286.505263
Waist	68.800000	10.252632	-8.147368	-9.347368	-129.336842	-31.442105
Pulse	-65.115789	-8.147368	51.989474	5.742105	101.521053	12.915789
Chins	-50.863158	-9.347368	5.742105	27.944737	230.107895	134.384211
Situps	-761.715789	-129.336842	101.521053	230.107895	3914.576316	2146.984211
Jumps	-286.505263	-31.442105	12.915789	134.384211	2146.984211	2629.378947

Now variance-covariance matrix of physiological variables weight, waist and pulse is

$$\boldsymbol{\Sigma}_{11} = \begin{bmatrix} 609.621053 & 68.800000 & -65.115789 \\ 68.000000 & 10.252632 & -8.147368 \\ -65.115789 & -8.147368 & 51.989474 \end{bmatrix}.$$

Variance-covariance matrix of exercise variables chins, situps and jmps is

$$\boldsymbol{\Sigma}_{22} = \begin{bmatrix} 27.944737 & 230.107895 & 134.384211 \\ 230.107895 & 3914.576316 & 2146.984211 \\ 134.384211 & 2146.984211 & 2629.378947 \end{bmatrix} .$$

Covariance matrix between physiological variables and exercise variables is

$$\boldsymbol{\Sigma}_{12} = \begin{bmatrix} -50.863158 & -9.347368 & 5.742105 \\ -761.715789 & -129.336842 & 101.521053 \\ -286.505263 & 101.521053 & 12.915789 \end{bmatrix}.$$

Step 2: To obtain $\boldsymbol{\Sigma}_{11}^{-1/2}$, first obtain eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_{11}$. Let λ_i and γ_i denote ith eigenvalue and ith eigenvector respectively; i = 1,2,3. Now $\Sigma_{11}^{-1/2} = \sum_{i=1}^{3} \lambda_i^{-1/2}\gamma_i\gamma_i'$. The eigenvalues $\boldsymbol{\Sigma}_{11}$ are 624.93238, 44.487838 and 2.4429359. The corresponding eigenvectors are: (0.987172, 0.1120006, -0.113786)'; (0.1150796, -0.005131, 0.993343)'; (-0.110671, 0.9936949, 0.017954)'. Now

$$\boldsymbol{\Sigma}_{11}^{-1/2} = \begin{bmatrix} 0.0488043 & -0.066027 & 0.0113741 \\ -0.066027 & 0.6322628 & 0.0101406 \\ 0.0113741 & 0.0101406 & 0.1486615 \end{bmatrix}.$$

Step 3: Compute $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ =**F** (say). In this case it is

$$\mathbf{F} = \begin{bmatrix} 0.2033438 & 0.2611853 & -0.044758 \\ 0.2611853 & 0.4540921 & -0.083341 \\ -0.044758 & -0.083341 & 0.0210455 \end{bmatrix}.$$

Step 4: Obtain eigenvalues and eigenvectors of **F**. The eigenvalues are 0.6332992, 0.040223, 0.005266 and the corresponding eigenvectors are $\mathbf{a}_1' = (0.524930, 0.837384, -0.152437)'$; $\mathbf{a}_2' = (0.847489, 0.497640, 0.184708)'$; $\mathbf{a}_3' = (0.078813, 0.226147, 0.970900)'$ respectively.

Step 5: On the similar lines of Steps 2, 3 and 4 obtain $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} = \mathbf{G}$, its eigenvalues and eigenvectors.

$$\mathbf{G} = \begin{bmatrix} 0.2634144 & -0.013700 & -0.001758 \\ -0.013700 & 0.0205081 & -0.008377 \\ -0.001758 & -0.0083771 & 0.0248539 \end{bmatrix}$$

. The eigenvalues are 0.6332992, 0.040223, 0.005266 and the corresponding eigenvectors are $\mathbf{b}_1' = (0.297668, 0.926878, -0.228672)'$; $\mathbf{b}_2' = (-0.247629, 0.3062953, 0.9191642)'$; $\mathbf{b}_3' = (0.9219943, -0.216980, 0.3206965)'$ respectively

Step 6: The first pair of canonical variables is now given by $\mathbf{a}_1'\Sigma_{11}^{-1/2}\mathbf{X}$ and $\mathbf{b}_1'\Sigma_{22}^{-1/2}\mathbf{Y}$. Here **X** denotes the matrix of physiological variables and **Y** denote the matrix of exercise variables. Similarly other two pairs of canonical variables can be obtained.

$\mathbf{a}_1'\Sigma_{11}^{-1/2}$ = [-0.031405 0.4932416 -0.008199];
$\mathbf{a}_2'\Sigma_{11}^{-1/2}$ = [0.0763195 -0.368723 0.032052];
$\mathbf{a}_3'\Sigma_{11}^{-1/2}$ = [-0.007735 0.1580337 0.1457322].

$\mathbf{b}_1'\Sigma_{22}^{-1/2}$ = [0.066114 0.0168462 -0.013972];
$\mathbf{b}_2'\Sigma_{22}^{-1/2}$ = [-0.071041 0.0019737 0.0207141];
$\mathbf{b}_3'\Sigma_{22}^{-1/2}$ =[0.2452754 -0.019768 0.0081675].

The canonical correlations between three pairs of canonical variables can be obtained by taking the square root of the eigenvalues of $\boldsymbol{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ or $\Sigma_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$. Here the canonical correlations are 0.795608, 0.200556 and 0.072570.

E2) Step 1: Variance-covariance matrix of five yield attributing parameters X1, X2, X3, X4 and X5) is $\boldsymbol{\Sigma}_{11}$ and is given by 5×5 cells in left hand upper corner of the given variance covariance matrix. The variance-covariance matrix of four quality parameters Y1, Y2, Y3 and Y4 is $\boldsymbol{\Sigma}_{22}$ and is given by 4×4 cells in right hand bottom corner of the given variance covariance matrix. The covariance matrix between yield attributing parameters and quality attributes is $\boldsymbol{\Sigma}_{12}$ and is the 5×4 cells in right hand upper corner of the given variance covariance matrix.

Step 2: To obtain $\boldsymbol{\Sigma}_{11}^{-1/2}$, first obtain eigenvalues and eigenvectors of $\boldsymbol{\Sigma}_{11}$. Let λ_i and γ_i denote ith eigenvalue and ith eigenvector respectively; i = 1,2,3. Now $\Sigma_{11}^{-1/2} = \sum_{i=1}^{3}\lambda_i^{-1/2}\gamma_i\gamma_i'$. The eigenvalues $\boldsymbol{\Sigma}_{11}$ are 3.1758116, 0.7417691, 0.6572676, 0.275192 and 0.1499598.
The corresponding eigenvectors are:
(0.5157006 0.2095696 0.0655101 0.3701104 -0.740851)';
(0.4876162 -0.15341 -0.3956110 0.5454541 0.5335426)';
(-0.457400 -0.206258 0.5068357 0.7007362 0.0181492)';
(-0.349891 0.8468537 -0.3073330 0.2408609 0.0891511)' and
(0.4057652 0.4157429 0.6984732 -0.128269 0.39773710)'.
Now

$$\boldsymbol{\Sigma}_{11}^{-1/2} = \begin{bmatrix} 1.8839857 & -0.564091 & 0.3180696 & 0.0793549 & -0.576395 \\ -0.564091 & 1.6560546 & 0.4178776 & 0.2766612 & 0.1107616 \\ 0.3180696 & 0.4178776 & 1.4205354 & 0.0207793 & 0.0802538 \\ 0.0793549 & 0.2766612 & 0.0207793 & 1.1490046 & 0.0970125 \\ -0.576395 & 0.1107616 & 0.0802538 & 0.0970125 & 1.334717 \end{bmatrix}.$$

Step 3: Compute $\boldsymbol{\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}}$ =**F** (say). In this case it is

$$\mathbf{F}=\begin{bmatrix} 0.2863075 & 0.0949999 & -0.119406 & -0.084734 & 0.2775501 \\ 0.0949999 & 0.2120403 & -0.238331 & -0.18475 & 0.0624013 \\ -0.119406 & -0.238331 & 0.2852454 & 0.1846082 & -0.069907 \\ -0.084734 & -0.18475 & 0.1846082 & 0.1997659 & -0.089549 \\ 0.2775501 & 0.0624013 & -0.069907 & -0.089549 & 0.3175887 \end{bmatrix}.$$

Step 4: Obtain first four eigenvalues and eigenvectors of **F**. The eigenvalues are 0.8265343, 0.4002511, 0.0651783, 0.0089843 and the corresponding eigenvectors are

$\mathbf{a}_1' = (0.4743674, 0.4263205, -0.485436, -0.396704, 0.4474416)'$;

$\mathbf{a}_2' = (0.4682176, -0.392238, 0.428314, 0.2902513, 0.599352)'$;

$\mathbf{a}_3' = (0.3723098, 0.0200515, -0.485523, 0.736616, -0.287484)'$;

$\mathbf{a}_4' = (0.6273877, 0.1540741, 0.4326632, -0.259644, -0.572742)'$ respectively.

Step 5: On the similar lines of Steps 2, 3 and 4 obtain $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2} = \mathbf{G}$, its eigenvalues and eigenvectors.

$$\mathbf{G}=\begin{bmatrix} 1.1354777 & -0.073763 & 0.2631705 & -0.124552 \\ -0.073763 & 1.0984645 & 0.2394113 & 0.0889018 \\ 0.2631705 & 0.2394113 & 1.1877782 & 0.0714007 \\ -0.124552 & 0.0889018 & 0.0714007 & 1.0378877 \end{bmatrix}.$$

The eigenvalues are 0.8265343, 0.4002511, 0.0651783, 0.0089843 and the corresponding eigenvectors are

$\mathbf{b}_1' = (0.6527588, 0.4359061, -0.616865, 0.0580444)'$;

$\mathbf{b}_2' = (0.7101565, -0.072348, 0.6880977, -0.13025)'$;

$\mathbf{b}_3' = (-0.258398, 0.8350398, 0.279248, -0.397442)'$;

$\mathbf{b}_4' = (-0.05305, 0.3278111, 0.2608056, 0.90648)'$ respectively.

Step 6: The first pair of canonical variables is now given by $\mathbf{a}_1'\Sigma_{11}^{-1/2}\mathbf{X}$ and $\mathbf{b}_1'\Sigma_{22}^{-1/2}\mathbf{Y}$. Here **X** denotes the matrix of yield attributing parameters and **Y** denote the matrix of quality attributes. Similarly other three pairs of canonical variables can be obtained.

The canonical correlations between four pairs of canonical variables can be obtained by taking the square root of the eigenvalues of $\boldsymbol{\Sigma}_{11}^{-1/2}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1/2}$ or $\boldsymbol{\Sigma}_{22}^{-1/2}\boldsymbol{\Sigma}_{21}\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1/2}$. Here the canonical correlations are 0.9091, 0.6327, 0.2553 and 0.0948.

E3) Proceeding on the same steps as in E2), one can easily see that the canonical correlation between first pair of canonical variables is 0.95. Which is also obvious.