IX Lecture

Canonical Correlation Analysis

9.1. Introduction. The basics of the procedure

Assume we are interested in the association between two **sets** of random variables. Typical examples include: relation between set of governmental policy variables and a set of economic goal variables; relation between college "performance" variables (like grades in courses in five different subject matter areas) and precollege "achievement" variables (like high-school gradepoint averages for junior and senior years, number of high-school extracurricular activities) etc.

The way the above problem of measuring association is solved in Canonical Correlation Analysis, is to consider the largest possible correlation between *linear combination of the variables in the first set* and a *linear combination of the variables in the second set*. The pair of linear combinations obtained through this maximization process is called **first canonical variables** and their correlation is called **first canonical correlation**. The process can be continued (similarly to the principal components procedure) to find a second pair of linear combinations having the largest correlation among all pairs that are uncorrelated with the initially selected pair. This would give us the second set of canonical variables with their second canonical correlation etc. The maximization process that we are performing at each step reflects our wish (again like in principal components analysis) to concentrate the initially high dimensional relationship between the 2 sets of variables into a few pairs of canonical variables only. Often, even only **one** pair is considered. The rationale in canonical correlation analysis is that when the number of variables is large, interpreting the whole set of correlation coefficients between pairs of variables from each set is hopeless and in that case one should concentrate on a *few* carefully chosen representative correlations. Finally, we should note that the traditional (simple) correlation coefficient and the multiple correlation coefficient (lecture V) are *special cases* of canonical correlation in which one or both sets contain a single variable.

9.2. Application in testing for independence of sets of variables

Besides being interesting in its own right (see 9.1) , calculating canonical correlations turns out to be important for the sake of **testing independence of sets of random variables**. Let us remember that testing for independence and for uncorrelatedness in the case of multivariate normal are equivalent problems. Assume now that that $\mathbf{X} \sim N_p(\mu, \Sigma)$. Furthermore, let $\mathbf{X}$ be partitioned into r, q components $(r + q = p)$ with $\mathbf{X^{(1)}} \in \mathbf{R^r}, \mathbf{X^{(2)}} \in \mathbf{R^q}$ and correspondingly, the covariance matrix

$$\mathbf{\Sigma} = E(\mathbf{X} - \mu)(\mathbf{X} - \mu)' = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \dots & \dots & \dots & \dots \\ \dots & \dots & \dots & \dots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} \in \mathscr{M}_{\mathbf{p,p}}$$

has been also partitioned into $\mathbf{\Sigma} = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}$, accordingly. We shall assume for simplicity that the matrices Σ, Σ_{11} and Σ_{22} are nonsingular. To test $H_0 : \Sigma_{12} = 0$ against a general alternative, a sensible way to go would be the following: for fixed vectors

$a \in R^r, b \in R^q$ let $Z_1 = a'\mathbf{X^{(1)}}$ and $Z_2 = b'\mathbf{X^{(2)}}$ giving $\rho_{a,b} = Corr(Z_1, Z_2) = \frac{a'\Sigma_{12}b}{\sqrt{a'\Sigma_{11}ab'\Sigma_{22}b}}$. H_0 is equivalent to $H_0 : \rho_{a,b} = 0$ for all $a \in R^r, b \in R^q$. For a particular pair a, b , H_0 would be accepted if $|r_{a,b}| = \frac{|a'S_{12}b|}{\sqrt{a'S_{11}ab'S_{22}b}} \leq k$ for certain positive constant k. (Here S_{ij} are the corresponding data based estimators of Σ_{ij}). Hence an appropriate acceptance region for H_0 would be given in the form $\{\mathbf{X} \in \mathscr{M}_{\mathbf{p,n}} : \max_{a,b} r_{a,b}^2 \leq k^2\}$. But maximizing $r_{a,b}^2$ means to find the maximum of $(a'S_{12}b)^2$ under constraints $a'S_{11}a = 1, b'S_{22}b = 1$ and this is exactly the data-based version of the optimization problem to be solved in 9.1. For the goals in 9.1 and 9.2 to be achieved, we need to solve problems of the following type:

9.3. Precise mathematical formulation and solution to the problem.

Canonical variables are the variables $Z_1 = a'\mathbf{X^{(1)}}$ and $Z_2 = b'\mathbf{X^{(2)}}$ where $a \in R^r, b \in R^q$ are obtained by maximizing $(a'\Sigma_{12}b)^2$ under the constraints $a'\Sigma_{11}a = 1, b'\Sigma_{22}b = 1$. To solve the above maximization problem, we construct

$$Lag(a, b, \lambda_1, \lambda_2) = (a'\Sigma_{12}b)^2 + \lambda_1(a'\Sigma_{11}a - 1) + \lambda_2(b'\Sigma_{22}b - 1)$$

Partial differentiation with respect to the vectors a and b gives:

$$2(a'\Sigma_{12}b)\Sigma_{12}b + 2\lambda_1\Sigma_{11}a = 0 \in R^r \tag{9.1}$$

$$2(a'\Sigma_{12}b)\Sigma_{21}a + 2\lambda_2\Sigma_{22}b = 0 \in R^q \tag{9.2}$$

We multiply (9.1) by the vector a' from left and equation (9.2) by b' from left and after subtracting the two equations obtained we get $\lambda_1 = \lambda_2 = -(a'\Sigma_{12}b)^2 = -\mu^2$. Hence:

$$\Sigma_{12}b = \mu\Sigma_{11}a \tag{9.3}$$

and

$$\Sigma_{21}a = \mu\Sigma_{22}b \tag{9.4}$$

hold. Now we first multiply (9.3) by $\Sigma_{21}\Sigma_{11}^{-1}$ from left, then both sides of (9.4) by the scalar μ and after finally adding the two equations we get:

$$(\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \mu^2\Sigma_{22})b = 0 \tag{9.5}$$

The homogeneous equation system (9.5) having a nontrivial solution w.r. b means that

$$|\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \mu^2\Sigma_{22}| = 0 \tag{9.6}$$

must hold. Then, of course,

$$|\Sigma_{22}^{-\frac{1}{2}}|.|\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} - \mu^2\Sigma_{22}|.|\Sigma_{22}^{-\frac{1}{2}}| = |\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}} - \mu^2 I_q| = 0$$

must hold. This means that μ^2 has to be an eigenvalue of the matrix $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$. Also, $b = \Sigma_{22}^{-\frac{1}{2}}\hat{b}$ where $\hat{b}$ is the eigenvector of $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ corresponding to this eigenvalue (WHY (?!)). (Note, however, that this representation is good mainly for theoretical purposes, the main advantage being that one is dealing with eigenvalues of a symmetric matrix. If doing calculations by hand, it is usually easier to calculate b directly as the solution of the linear equation (9.5), i.e. find the largest eigenvalue of the (non-symmetric) matrix $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$ and then find the eigenvector b that correponds to it.

Besides, we also see from the definition of μ that $\mu^2 = (a'\Sigma_{12}b)^2$ holds. Since we wanted to **maximize** the right hand side, it is obvious that μ^2 must be chosen to be the **largest eigenvalue** of the matrix $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ (or, which is the same thing, the largest eigenvalue of the matrix $\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}$). Finally, we can obtain the vector a from (9.3): $a = \frac{1}{\mu}\Sigma_{11}^{-1}\Sigma_{12}b$. That way , the **first** canonical variables $Z_1 = a'\mathbf{X^{(1)}}$ and $Z_2 = b'\mathbf{X^{(2)}}$ are determined and the value of the first canonical correlation is just μ. The orientation of the vector b is chosen such that the sign of μ should be positive.

Now , it is easy to see that if we want to extract a second pair of canonical variables we need to repeat the same process by starting with the **second largest** eigenvalue μ^2 of the matrix $\Sigma_{22}^{-\frac{1}{2}}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-\frac{1}{2}}$ (or of the matrix $\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}$). This will automatically ensure that the second pair of canonical variables is uncorrelated with the first pair. The process can theoretically be continued until the number of pairs of canonical variables equals the number of variables in the smaller group. But in practice, much fewer canonical variables will be needed. Each canonical variable is uncorrelated with all the other canonical variables of either set except for the one corresponding canonical variable in the opposite set.

Note. It is important to point out that already by definition the canonical correlation is at least as large as the multiple correlation between any variable and the opposite set of variables. It is in fact possible for the first canonical correlation to be *very large* while all the multiple correlations of each separate variable with the opposite set of canonical variables are small. This once again underlines the importance of Canonical Correlation analysis.

9.4. Estimating and testing canonical correlations

The way to estimate the canonical variables and canonical correlation coefficients is based on the plug-in technique. One follows the steps outlined in 9.3, by each time substituting S_{ij} instead of Σ_{ij}.

Let us now discuss the independence testing issue outlined in 9.2. The acceptance region of the independence test of H_0 in 9.2. would be $\{\mathbf{X} \in \mathscr{M}_{\mathbf{p,n}}$: largest eigenvalue of $S_{22}^{-\frac{1}{2}}S_{21}S_{11}^{-1}S_{12}S_{22}^{-\frac{1}{2}} \leq k_\alpha\}$ where k_α has been worked out and is given in the so called **Hecks charts**. This distribution depends on three parameters: $s = min(r,q), m = \frac{|r-q|-1}{2}, N = \frac{n-r-q-2}{2}, n$ being the sample size. Besides using the charts, one can also use good F-distribution-based approximations for a (transformations of) this distribution like Wilk's lambda, Pillai's trace, Hotelling Trace and Roy's greatest root. Here we shall only mention that all these statistics and their P-values (using suitable F-distribution-based approximations) are readily available as an output in the SAS program CANCORR so that performing the test is really easy-one can read out directly the p-value from the SAS output.

9.5. Some important computational issues

Note that calculating $\mathbf{X}^{-\frac{1}{2}}$ and $\mathbf{X}^{\frac{1}{2}}$ for a symmetric positive definite matrix $\mathbf{X}$ according to the theoretically attractive spectral decomposition method may be numerically unstable. This is especially the case when some of the eigenvalues are close to zero. In many numerical packages including SAS, the so called **Cholesky decomposition** is being calculated instead. For a symmetric positive definite matrix $\mathbf{X} \in \mathscr{M}_{\mathbf{p,p}}$ its Cholesky decomposition is defined as an upper triangular matrix $\mathbf{U}$ such that $\mathbf{U'U} = \mathbf{X}$ holds. Note

that in SAS/IML , when you write e.g. $\mathbf{xroot} = \mathbf{root}(\mathbf{x})$ you do **not** get as a result the theoretical $\mathbf{X}^{\frac{1}{2}}$ but you rather get the matrix $\mathbf{U}$ defined above. Looking back at 9.5, we see that if $\mathbf{U}'\mathbf{U} = \mathbf{\Sigma}_{\mathbf{22}}^{-\mathbf{1}}$ gives the Cholesky decomposition of the matrix $\mathbf{\Sigma}_{\mathbf{22}}^{-\mathbf{1}}$ then μ^2 is the eigenvalue of the matrix $\mathbf{A} = \mathbf{U}\mathbf{\Sigma}_{\mathbf{21}}\mathbf{\Sigma}_{\mathbf{11}}^{-\mathbf{1}}\mathbf{\Sigma}_{\mathbf{12}}\mathbf{U}'$. Indeed, by multiplying from left by $\mathbf{U}$ and from right by $\mathbf{U}'$ in (9.6) we get:

$$|\mathbf{A} - \mu^{\mathbf{2}}\mathbf{U}\mathbf{\Sigma}_{\mathbf{22}}\mathbf{U}'| = 0$$

But $\mathbf{U}\mathbf{\Sigma}_{\mathbf{22}}\mathbf{U}' = \mathbf{U}(\mathbf{U}'\mathbf{U})^{-\mathbf{1}}\mathbf{U}' = \mathbf{U}\mathbf{U}^{-\mathbf{1}}(\mathbf{U}')^{-\mathbf{1}}\mathbf{U}' = \mathbf{I}$ holds.

9.6. Numerical examples

The SAS procedure for performing Canonical Correlation Analysis is called CANCORR. The following numerical example will be discussed at lectures:

- **Canonical Correlation Analysis of the Fitness Club Data**. Three physiological and three exercise variables were measured on twenty middle aged men in a fitness club. The CANCORR procedure is being used to determine if the physiological variables are related in any way to the exercise variables.

- **Example 10.4, p. 552 in Johnston and Wichern**. Studying canonical correlations between leg and head bone measurements: X_1, X_2 are skull length and skull breadth respectively; X_3, X_4 are leg bone measurements: femur and tibia length, respectively. Observations have been taken on $n = 276$ White Leghorn chicken. The example is chosen to also illustrate how a canonical correlation analysis can be performed when the original data is not given but the empirical correlation matrix (or empirical covariance matrix) is available. The following SAS file implements the calculations.


```
options linesize=70; title 'Canonical Correlation Analysis';
data chicken (type=corr);
input _type_ $ _name_ $ x1 x2 x3 x4;
cards;
 n      .   276     276    276    276
CORR    x1  1.0      .      .      .
CORR    x2   .505   1.0     .      .
CORR    x3   .569   .422   1.0     .
CORR    x4   .602   .467   .926   1.0
;
proc cancorr data=chicken  vprefix=head wprefix=leg;
var x1 x2; with x3 x4;
run;
```

If covariance matrix was available instead and the analysis was performed using the covariance matrix then the modification in the above example would have been that one would need to write COV instead of CORR above and enter the elements of the covariance matrix instead. After running the example, we get that $U_1 = .781Z_1 + .345Z_2; V_1 = .060Z_3 + .944Z_4$ is the first pair of canonical variables. Here $Z_i, i = 1, 2, 3, 4$ are the standardized versions of $X_i, i = 1, 2, 3, 4$. Their correlation (the first canonical correlation) is .631 and is quite significant.

Lecture IX-supplement

Computations used in the MANOVA tests

In standard (univariate) Analysis of Variance, with usual normality assumptions on the errors, testing about effects of the factors involved in the model description is based on the F test. The F tests are derived from the ANOVA decomposition SST=SSA+SSE. The argument goes as follows:
i) SSE and SSA are independent, (up to constant factors involving the variance σ^2 of the errors) χ^2 distributed;
ii) by proper norming to account for degrees of freedom, from SSE and SSA one gets statistics that have the following behaviour: the normed SSE always delivers an unbiased estimator of σ^2 no matter if the hypothesis or alternative is true; the normed SSA delivers an unbiased estimator of σ^2 under hypothesis but delivers an unbiased estimator of a "larger" quantity under the alternative.
The above observation is crucial and motivates the F-testing: F statistics are (**suitably normed to account for degrees of freedom**) ratios of SSA/SSE. When taking the ratio, the factors involving σ^2 **cancel out** and σ^2 does not play any role in the distribution of the ratio. Under H_0 their distribution is F. When the null hypothesis is violated, then the same statistics will tend to have "larger" values as compared to the case when H_0 is true. Hence significant (w.r.to the corresponding F-distribution) values of the statistic lead to rejection of H_0.
Aiming at generalising these ideas to the Multivariate ANOVA (MANOVA) case, we should note that instead of χ^2 distributions we now have to deal with **Wishart** distributions and we need to properly define (a proper functional of) the SSA/SSE ratio which would be a "ratio" of matrices now. Obviously, there are more ways to define suitable statistics in this context! It turns out that such functionals are related to the eigenvalues of the (properly normed) Wishart distributed matrices that enter the decomposition SST=SSA+SSE in the multivariate case.

II.1. Roots distributions.

Let $Y_i, i = 1, 2, .., p \sim N_p(\mu_i, \Sigma)$. Then the following data matrix:

$$Y = Y_{(nxp)} = \begin{pmatrix} Y_1' \\ Y_2' \\ . \\ . \\ Y_n' \end{pmatrix} = \begin{pmatrix} Y_{11} & Y_{12} & .. & Y_{1p} \\ Y_{21} & Y_{22} & .. & Y_{2p} \\ . & . & .. & . \\ . & . & .. & . \\ Y_{n1} & Y_{n2} & .. & Y_{np} \end{pmatrix} = [U_1|U_2|\dots|U_p] \text{ is a } nxp \text{ matrix}$$

containing n p-dimensional (transposed) vectors. Denote: $E(Y) = M, Var(\overrightarrow{Y}) = \Sigma \otimes I_n$. Let A, B be projectors and such that $Q_1 = Y'AY$ and $Q_2 = Y'BY$ are two **independent** $W_p(v, \Sigma), W_p(q, \Sigma)$, respectively. Although the theory is general, to keep you on track, you could always think about a linear model example:

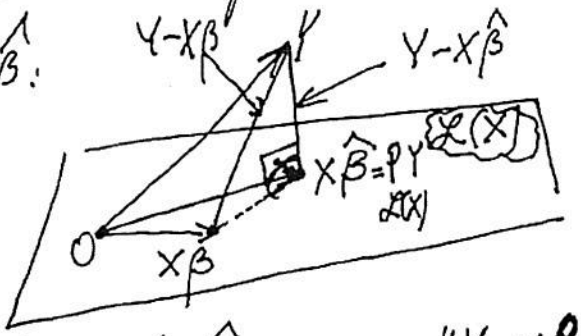
$$Y = X\beta + E, \hat{Y} = X\hat{\beta}, A = I_n - X(X'X)^- X', B = X(X'X)^- X' - \overrightarrow{1}_n(\overrightarrow{1}'_n \overrightarrow{1}_n)^{-1}\overrightarrow{1}'_n$$

and the corresponding decomposition

$$Y[I_n - \overrightarrow{1}_n(\overrightarrow{1}'_n \overrightarrow{1}_n)^{-1}\overrightarrow{1}'_n]Y = Y'BY + Y'AY = Q_2 + Q_1$$

Geometric interpretation:

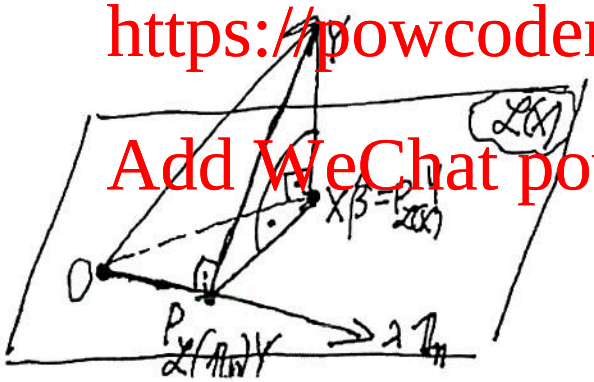
a) LS estimator $\hat{\beta}$:



$\|Y - X\hat{\beta}\| < \|Y - X\beta\|$ and $\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|$

$\|Y - X\hat{\beta}\|^2 = \|Y - P_{\mathcal{L}(X)} Y\|^2 = Y'(I_n - P_{\mathcal{L}(X)})Y$

b) Projection on two linear spaces (one imbedded in the other): $\mathcal{L}(\mathbb{1}_n)$ and $\mathcal{L}(X)$, $\mathbb{1}_n = \begin{pmatrix} 1 \\ \vdots \\ 1 \end{pmatrix}$



$P_{\mathcal{L}(X)} = X(X'X)^{-1}X'$

$P_{\mathcal{L}(\mathbb{1}_n)} = \mathbb{1}_n(\mathbb{1}_n'\mathbb{1}_n)^{-1}\mathbb{1}_n' = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n'$

Pythagoras: $\|Y - P_{\mathcal{L}(\mathbb{1}_n)} Y\|^2 = \|P_{\mathcal{L}(X)} Y - P_{\mathcal{L}(\mathbb{1}_n)} Y\|^2 + \|Y - P_{\mathcal{L}(X)} Y\|^2$

$Y'(I_n - P_{\mathcal{L}(\mathbb{1}_n)})Y = Y'(P_{\mathcal{L}(X)} - P_{\mathcal{L}(\mathbb{1}_n)})Y + Y'(I_n - P_{\mathcal{L}(X)})Y$

(not depend on model)

$= Y'BY + Y'AY = Q_2 + Q_1 = H_{(ypothesis)} + E_{(rror)}$

"Small" E (relative to H) means that $Y = X\beta + \varepsilon$ is a good model. Can look at: small $|Q_1|/|Q_1+Q_2|$ (WILKS), large $Q_1^{-1}Q_2$ (Lawley), large $(Q_1+Q_2)^{-1}Q_2$ (Pillai), small $Q_1Q_2^{-1}$, largest eigenvalue of $Q_1^{-1}Q_2$ etc.

of $SST = SSA + SSE = Q_2 + Q_1$ where Q_2 is the "hypothesis matrix" and Q_1 is the "error matrix".

Lemma II.1 Let $Q_1, Q_2 \in \mathscr{M}_{p,p}$ be two positive definite symmetric matrices . Then the roots of the determinant equation $|Q_2 - \theta(Q_1 + Q_2)| = 0$ are related to the roots of the equation $|Q_2 - \lambda Q_1| = 0$ by: $\lambda_i = \frac{\theta_i}{1-\theta_i}$ (or $\theta_i = \frac{\lambda_i}{1+\lambda_i}$).

Lemma II.2 Let $Q_1, Q_2 \in \mathscr{M}_{p,p}$ be two positive definite symmetric matrices . Then the roots of the determinant equation $|Q_1 - v(Q_1 + Q_2)| = 0$ are related to the roots of the equation $|Q_2 - \lambda Q_1| = 0$ by: $\lambda_i = \frac{1-v_i}{v_i}$ (or $v_i = \frac{1}{1+\lambda_i}$).

We can employ the above two lemmas to see that:

If λ_i, v_i, θ_i are the roots of :

$$|Q_2 - \lambda Q_1| = 0, |Q_1 - v(Q_1 + Q_2)| = 0, |Q_2 - \theta(Q_1 + Q_2)| = 0$$

then:

$$\Lambda = |Q_1(Q_1 + Q_2)^{-1}| = \prod_{i=1}^{p}(1 + \lambda_i)^{-1}$$

(Wilks' Criterion statistic) or

$$|Q_2 Q_1^{-1}| = \prod_{i=1}^{p} \lambda_i = \prod_{i=1}^{p} \frac{1 - v_i}{v_i} = \prod_{i=1}^{p} \frac{\theta_i}{1 - \theta_i}$$

or

$$|Q_2(Q_1 + Q_2)^{-1}| = \prod_{i=1}^{p} \theta_i = \prod_{i=1}^{p} \frac{\lambda_i}{1 + \lambda_i} = \prod_{i=1}^{p}(1 - v_i)$$

and other functional transformations of these products of (random) roots would have a distribution that would only depend on p, q, v. There are various ways to choose such functional transformations (statistics) and many have been suggested like:

- Λ (Wilks' Lambda)
- $tr(Q_2 Q_1^{-1}) = tr(Q_1^{-1} Q_2) = \sum_{i=1}^{p} \lambda_i$ (Lawley-Hotelling trace)
- $max_i \lambda_i$ (Roy's criterium)
- $V = tr[Q_2(Q_1 + Q_2)^{-1}] = \sum_{i=1}^{p} \frac{\lambda_i}{1+\lambda_i}$ (Pillai statistic)

Tables and charts for their exact or approximate distributions are available. Also, P-values for these statistics are readily calculated in SAS, SPSS, Minitab and other statistical packages. In these applications, the meaning of Q_1 is of the "error matrix" (also denoted by E sometimes) and the meaning of Q_2 is that of a "hypothesis matrix" (also denoted by H sometimes). The distribution of the statistics defined above depends on the following three parameters:

- $p =$ the number of responses
- $q = \nu_h =$ degrees of freedom for the hypothesis
- $v = \nu_e =$ degrees of freedom for the error

Based on these, the following quantities are calculated:

$s = min(p, q), m = 0.5(|p - q| - 1), n = 0.5(v - p - 1), r = v - 0.5(p - q + 1), u = 0.25(pq - 2)$. Moreover, we define: $t = \sqrt{\frac{p^2q^2-4}{p^2+q^2-5}}$ if $p^2 + q^2 - 5 > 0$ and $t = 1$ otherwise. Let us order the eigenvalues of $E^{-1}H = Q_1^{-1}Q_2$ according to: $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_p$. Then the following distribution results are true: (Note: the F-statistic quoted below is **exact** if $s = 1$ or 2, otherwise the F-distribution is an **approximation**):

- Wilks's test. The test statistics, Wilks's lambda, is $\Lambda = \frac{|E|}{|E+H|} = \prod_{i=1}^{p} \frac{1}{1+\lambda_i}$ Then it holds: $F = \frac{1-\Lambda^{1/t}}{\Lambda^{1/t}} \cdot \frac{rt-2u}{pq} \sim F_{pq,rt-2u}$ df (**Rao's F**).
- Lawley-Hotelling trace Test. The Lawley-Hotelling statistic is $U = tr(E^{-1}H) = \lambda_1 + \ldots + \lambda_p$, and $F = 2(sn + 1)\frac{U}{s^2(2m+s+1)} \sim F_{s(2m+s+1),2(sn+1)}$ df.
- Pillai's test. The test-statistics , Pillai trace, is $V = tr(H(H + E)^{-1}) = \frac{\lambda_1}{1+\lambda_1} + \ldots + \frac{\lambda_p}{1+\lambda_p}$ and $F = \frac{2n+s+1}{2m+s+1} \cdot \frac{V}{s-V} \sim F_{s(2m+s+1),s(2n+s+1)}$ df.
- Roy's maximum root criterium. The test-statistic is just the largest eigenvalue λ_1.

Finally, we shall mention one historically older and very universal approximation to the distribution of the Λ statistic due to Bartlett (1927):

It holds: level of $-[\nu_e - \frac{p-\nu_h+1}{2}]ln\Lambda = c(p, \nu_h, M) *$ level of $\chi^2_{p\nu_h}$ where the constant $c(p, \nu_h, M = \nu_e - p + 1)$ is given in accompanying tables. Such tables are prepared for levels $\alpha = 0.10, 0.05, 0.025$ etc.

In the context of testing the hypothesis about significance of the first canonican correlation, we have:

$$K = S_{22} - S_{21}S_{11}^{-1}S_{12}, M = S_{21}S_{11}^{-1}S_{12}.$$

The Wilks statistic becomes $\frac{|S|}{|S_{11}||S_{22}|}$(!) **We also see that in this case, if** μ_i^2 **were the squared canonical correlations then** μ_1^2 **was defined as the maximal eigenvalue to** $S_{22}^{-1}H$**, that is, it is a solution to** $|(E+H)^{-1}H - \mu_1^2 I| = 0$ **However, setting** $\lambda_1 = \frac{\mu_1^2}{1-\mu_1^2}$ **we see that:**

$$|(E+H)^{-1}H-\mu_1^2 I| = 0 \rightarrow |H-\mu_1^2(E+H)| = 0 \rightarrow |H-\frac{\mu_1^2}{1-\mu_1^2}E| = 0 \rightarrow |E^{-1}H-\lambda_1 I| = 0$$

holds and λ_1 **is an eigenvalue of** $E^{-1}H$**. Similarly you can argue for the remaining** $\lambda_i = \frac{\mu_i^2}{1-\mu_i^2}$ **values.**

II.1.2. Comparisons, applications.

From all statistics discussed, Wilks's lambda has been most widely applied. One important reason for this is that this statistic has the virtue of being convenient to use and, more importantly, being related to the Likelihood Ratio Test! Despite the above, the fact that so many different statistics exist for the same hypothesis testing problem, indicates that there is no universally best test. Power comparisons of the above tests are almost lacking since the distribution of the statistic under alernatives is hardly known.