Chapter 10

Canonical Correlation Analysis

10.1 Introduction

Consider a situation where the variables divide naturally into two groups and the aim is to investigate relationships between the two groups.

This is an extension of *regression*, where one of the groups contains a single variable Y and the other group contains the *regressor* variables $X_1, \ldots, X_p$. It is assumed that the variable Y can be expressed in terms of $X_1, \ldots, X_p$ through an equation

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon,$$

where $\epsilon \sim N(0, \sigma^2)$ and the errors for different observations are independent, identically distributed. Canonical correlation analysis considers the situation where there are two *groups* of variables.

For example, consider the butterfly data set, with 16 colonies of the butterfly *Euphydryas editha* in California and Oregon. There are 10 variables; 4 environmental variables and 6 gene frequency variables. An obvious question to be considered is what relationships, if any, exist between the gene frequencies and the environmental variables. The technique of *canonical correlation analysis* is designed to investigate such questions.

The subject was introduced by Hotelling in 1936 and, in the first example of a canonical correlation analysis, he considered the four variables: reading speed (X_1), reading power (X_2), arithmetic speed (Y_1) and arithmetic power (Y_2) for seventh grade school children. The specific question addressed was whether or not reading ability (as measured by X_1 and X_2) is related to arithmetic ability (as measured by Y_1 and Y_2).

Canonical correlation analysis looks for linear combinations

$$U = a_1 X_1 + a_2 X_2$$

$$V = b_1 Y_1 + b_2 Y_2$$

which make the correlation of U and V as high as possible. This is similar to Principal Components Analysis based on the *correlation*; it is the *correlation* that is maximised, not the variance.

In his example, Hotelling found that

$$U = -2.78X_1 + 2.27X_2$$

$$V = -2.44Y_1 + 1.00Y_2$$

gave the best correlation, which was 0.62. U measures the *difference* between reading power and speed, while V measures something that looks like the *difference* (modulo some scaling) between arithmetic power and speed. It is this aspect of reading and arithmetic that seems to have the most in common, or strongest correlation.

Often it is convenient to calculate more than one pair of canonical correlation variables. If $\underline{X}$ and $\underline{Y}$ are p and q vectors respectively, then one can compute k- vectors $\underline{U}$ and $\underline{V}$ where $k \leq \min(p, q)$,L_1 is $k \times p$, L_2 is $k \times q$ and

$$\underline{U} = L_1\underline{X}, \qquad \underline{V} = L_2\underline{Y}.$$

These variables are chosen so that the correlation between U_j and V_j is maximised subject to the constraint that U_j and V_j are both uncorrelated with any of the variables $U_1, \ldots, U_{j-1}, V_1, \ldots, V_{j-1}$. Each pair of variables represents an independent dimension in the relationship between $\underline{X}$ and $\underline{Y}$. The first pair (U_1, V_1) has the highest correlation and is therefore the most important. The next (U_2, V_2) the second highest and so on.

10.2 Setting up the Canonical Correlation Analysis

Firstly, since it is *correlations* that are under consideration here, start by standardising the variables (for each, subtract the means and divide through by the standard deviations).

Secondly, for the standardised variables, obtain the covariance matrix C for the $p + q$ vector $(\underline{X}^t, \underline{Y}^t)^t$. Let C_{11} denote the $p \times p$ correlation matrix of $\underline{X}$ and C_{22} the $q \times q$ correlation matrix of $\underline{Y}$ and C_{12} the matrix with entries $C_{ij} = \mathrm{Cov}(X_i, Y_j)$. Then the $p + q \times p + q$ correlation matrix C of $(\underline{X}^t, \underline{Y}^t)^t$ may be expressed as

$$C := \begin{pmatrix} C_{11} & C_{12} \\ C_{12}^t & C_{22} \end{pmatrix}.$$

Check that C_{11} and C_{22} are non-singular, since they have to be inverted, otherwise remove the smallest number of variables to ensure that they are non singular. Ensure that $p \leq q$. If it isn't, reverse the roles of $\underline{X}$ and $\underline{Y}$. Let k be the rank of C_{12}.
The aim is to find a $\tilde{L}_1$ and $\tilde{L}_2$ such that the random vector

$$\begin{pmatrix} \underline{U} \\ \underline{V} \end{pmatrix} = \begin{pmatrix} \tilde{L}_1 & 0 \\ 0 & \tilde{L}_2 \end{pmatrix} \begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix}$$

has covariance matrix

$$\begin{pmatrix} I_p & P \\ P^t & I_q \end{pmatrix}$$

where $P_{ii} = \rho_i$ for $i = 1, \ldots, k$ and $P_{ij} = 0$ for other (i, j). Clearly, $\tilde{L}_1$ and $\tilde{L}_2$ satisfy:

$$\tilde{L}_1 C_{11} \tilde{L}_1^t = I_p \qquad \tilde{L}_2 C_{22} \tilde{L}_2^t = I_q \qquad \tilde{L}_1 C_{12} \tilde{L}_2^t = P.$$

We have to construct $\widetilde{L}_1$ and $\widetilde{L}_2$ satisfying these properties.

We use the following basic result from linear algebra. Let A be a $p \times q$ matrix. Then we can construct a $p \times p$ orthonormal matrix H and a $q \times q$ orthonormal matrix Q and a matrix P satisfying $P_{ii} = \rho_i$ for $1 \leq i \leq k \leq \min(p, q)$ and $P_{ij} = 0$ otherwise, where $\rho_1^2 \geq \ldots \geq \rho_k^2$, such that $A = H'PQ$.

Let $A = C_{11}^{-1/2} C_{12} C_{22}^{-1/2}$, then

$$C_{11}^{-1/2} C_{12} C_{22}^{-1/2} = H'PQ,$$

where P satisfies the conditions above. Set

$$L_1 = H C_{11}^{-1/2}, \qquad L_2 = Q C_{22}^{-1/2},$$

then clearly

$$L_1 C_{11} L_1' = H C_{11}^{-1/2} C_{11} C_{11}^{-1/2} H' = I_p,$$
$$L_2 C_{22} L_2' = I_q,$$
$$L_1 C_{12} L_2' = P,$$

which solves the problem.

Set $\underline{U} = L_1 \underline{X}$ and $\underline{V} = L_2 \underline{Y}$, then

$$\begin{pmatrix} \underline{U} \\ \underline{V} \end{pmatrix} = \begin{pmatrix} L_1 & 0 \\ 0 & L_2 \end{pmatrix} \begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix} = L \begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix},$$

where $L = \mathrm{diag}(L_1, L_2)$ (a block diagonal matrix). Then

$$\mathrm{Cov} \begin{pmatrix} \underline{U} \\ \underline{V} \end{pmatrix} = LCL' = \begin{pmatrix} I_p & P \\ P' & I_q \end{pmatrix}.$$

It is therefore clear that the first k (U_j, V_j) pairs from $(\underline{\tilde{U}}, \underline{\tilde{V}})$ constructed in this way satisfy the criteria. Firstly, only the first k components of $\underline{\tilde{U}}$ and the first k components of $\underline{\tilde{V}}$ are relevant; the others are uncorrelated.

10.3 Significance Testing

An approximate relationship between the $\underline{X}$ variables as a whole and the $\underline{Y}$ variables as a whole was proposed by Bartlett.

Theorem 10.1. *Let X^2 denote the random variable*

$$X^2 = -n\sum_{j=1}^{r}\ln(1-\rho_j^2),$$

where $\rho_1,\ldots,\rho_r$ are the sample canonical correlations. Under $H_0:\rho_1=\ldots=\rho_r=0$ (i.e. $\underline{X}$ and $\underline{Y}$ are independent) then approximately, for a sample size n,

$$X^2 \sim \chi^2_{pq}.$$

Large values suggest that at least one of the r canonical correlations is significant. A lack of significance indicates that even the largest canonical correlation can be accounted for by sampling variation only.

Proof This is a likelihood ratio test. Let

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}$$

The assumption is that $\binom{X}{Y}$ has covariance C, where X and Y are the two sets of canonical variables and we have a null hypothesis: $H_0 : C_{12} = 0$. Assume we have a sample size of n, there are p X-variables and q Y-variables; it follows that C_{12} contains pq parameters, so that the difference in dimension between H_0 and the full parameter space is pq. An asymptotic likelihood ratio test statistic will therefore have χ^2_{pq} distribution, under the null hypothesis. It remains to show that $-2\log\Lambda$, where Λ is the likelihood ratio test statistic, has the required form.

Letting $Z = \binom{X}{Y}$ denote the whole collection of variables, the likelihood function for n independent instantiations (say $\mathbf{Z} = (Z_1,\ldots,Z_n)$) is

$$\begin{aligned} L(\mu,C;\mathbf{Z}) &= \frac{1}{(2\pi)^{pq/2}|C|^{n/2}}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(Z_i-\mu)'C^{-1}(Z_i-\mu)\right\} \\ &= \frac{1}{(2\pi)^{pq/2}|C|^{n/2}}\exp\left\{-\frac{1}{2}\left(n(\overline{Z}-\mu)'C^{-1}(\overline{Z}-\mu)+\sum_{i=1}^{n}(Z_i-\overline{Z})'C^{-1}(Z_i-\overline{Z})\right)\right\} \end{aligned}$$

Let A be the matrix with entries:

$$A_{ij} = \sum_{k=1}^{n}(Z_{ki}-\overline{Z}_{.i})(Z_{kj}-\overline{Z}_{.j})$$

then

$$L(\underline{\mu}, C; \mathbf{Z}) = \frac{1}{(2\pi)^{(p+q)n/2}|C|^{n/2}} \exp\left\{-\frac{1}{2}\mathrm{tr}C^{-1}\left(A + n(\overline{\underline{Z}} - \underline{\mu})(\overline{\underline{Z}} - \underline{\mu})^t\right)\right\}$$

This expression is maximised for $\widehat{\mu} = \overline{Z}$. To maximise over C, we have to maximise

$$\begin{aligned} g(C) &= -\frac{n}{2}\log|C| - \frac{1}{2}\mathrm{tr}(C^{-1}A) && (10.1)\\ &= \frac{n}{2}\log|C^{-1}A| - \frac{n}{2}\log|A| - \frac{1}{2}\mathrm{tr}(C^{-1}A) && (10.2)\\ &= \frac{n}{2}\sum_{j=1}^{p+q}(n\log\lambda_j - \lambda_j) - \frac{n}{2}\log A && (10.3) \end{aligned}$$

The function $f(\lambda) = n\log\lambda - \lambda$ has a unique maximum at $\lambda = n$. The maximum is $n\log n - n$. Therefore

$$g(C) \leq \frac{n(p+q)}{2}\log n - \frac{n(p+q)}{2} - \frac{n}{2}\log|A|$$

with equality if and only if $\lambda_j = n$ for all $j = 1, \ldots, p+q$. Hence

$$\widehat{C}^{-1}A = nI_{p+q}$$

so that $\widehat{C} = \frac{1}{n}A$ and:

$$\sup_{\mu,C} L(\mu, C) = L(\overline{Z}, \widehat{C}) = \frac{1}{(2\pi)^{(p+q)n/2}|\widehat{C}|^{n/2}} \exp\left\{-\frac{n(p+q)}{2}\right\}$$

where $\widehat{C}$ is the *maximum likelihoood estimator* which is $\widehat{C} = \frac{1}{n}A$ (same notation as before).

Under the null hypothesis that $C_{12} = 0$,

$$L(\mu, C) = L(\mu_1, C_{11})L(\mu_2, C_{22})$$

where μ_1 and μ_2 are the mean vectors for the X and Y variables respectively. Note that the maximum likelihood estimators for C_{11} and C_{22} remain the same under the restriction of the null hypothesis, so that

$$\prod_{i=1}^{2}\sup_{\mu_i, C_{ii}} L(\mu_i, C_{ii}) = \frac{1}{(2\pi)^{(p+q)n/2}} e^{-(p+q)n/2} n^{n/2} |A_{11}|^{-n/2} |A_{22}|^{-n/2}.$$

Hence the likelihood ratio test statistic is:

$$\Lambda = \frac{|A|^{n/2}}{|A_{11}|^{n/2}|A_{22}|^{n/2}}.$$

Now,

$$\begin{aligned}\frac{|A|}{|A_{11}||A_{22}|} &= \left|\begin{pmatrix} A_{11} & 0 \\ 0 & A_{22} \end{pmatrix}^{-1}\begin{pmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{pmatrix}\right| \\ &= \left|\begin{array}{cc} I & A_{11}^{-1}A_{12} \\ A_{22}^{-1}A_{21} & I \end{array}\right| = |I - A_{11}^{-1}A_{12}A_{22}^{-1}A_{21}| = \prod_{i=1}^{p}(1-\rho_j)^2.\end{aligned}$$

Recall (Statistics 1) that $2\times$ the log of the likelihood ratio test statistic is asymptotically χ^2, with degrees of freedom equal to the difference in number of parameters between null and alternative; in this case pq (the number of components of C_{12}). Hence

$$-2\log\Lambda \sim_{n\to+\infty} \chi^2_{pq}.$$

□

This test can be modified to allow each of the canonical functions to be tested individually, but the results have to be treated with caution. There are two basic ideas:

1. Compare the jth contribution of the right hand side, namely

$$-\left(n - \frac{p+q+3}{2}\right)\ln(1-\rho_j^2)$$

with a χ^2_{p+q-2j} distribution.

2. Compare

$$-\left(n - \frac{p+q+3}{2}\right)\sum_{k=j+1}^{r}\ln(1-\rho_k^2)$$

with a $\chi^2_{(p-j)(q-j)}$ distribution.

The first tests the jth function directly, while the second tests the functions $j+1,\ldots,r$ as a whole. These tests are unreliable for the same reason as discussed for related tests in discriminant function analysis: it can happen, by chance, that that the ith *sample* canonical correlation function is not a random observation from the ith *population* canonical correlation function. This gives a larger probability of wrongly rejecting the hypothesis that the function is insignificant.

10.4 Interpreting Canonical Variates

If

$$U_i = a_{i1}X_1 + \ldots + a_{ip}X_p$$

and

$$V_i = b_{i1}Y_1 + \ldots + b_{iq}Y_q$$

then it is usual to interpret U_i in terms of the X variables with large coefficients a_{ij} and V in terms of the Y variables with large coefficients b_{ij}. Large means either large positive or large negative.

Large correlations within the collection of X variables and large correlations within the collection of Y variables can upset this interpretation process. Hence if X_1 is highly correlated with X_2, the coefficient a_{il} can be substantially negative even if U_i and X_1 have a high positive correlation. This will be compensated for with the coefficient a_{i2}.

If one of the X variables is almost a linear combination of the other X variables, there will be a whole family of linear combinations, giving very different a_{ij} values, that give virtually the same U_i value.

The interpretation problems that arise with high correlation within the X variables or within the Y variables should be familiar from study of multiple linear regression; high correlation within one of the sets of variables leads to *variance inflation* so that *significant* canonical correlation functions are not detected by any significance test.

It is therefore *necessary* that there is no substantial correlation *within* the $\underline{X}$ variables and *within* the $\underline{Y}$ variables; that C_{11} and C_{22} are *non* singular. Otherwise, subtantially different linear combinations may indicate different canonical variables that represent almost the same information, in which case the results may not be very meaningful. Worse than this, *variance inflation* effects may prevent significant correlations from being detected.

10.5 Canonical Correlation Analysis and Factor Analysis

Just as with factor analysis, we can consider the U variables as factors. Recall that $U = L_1X$, $V = L_2Y$ and hence $X = L_1^{-1}U$, $Y = L_2^{-1}V$. Just as with factor analysis, if there are k significant factors, we set $U_{k+j} = 0$ for $j \geq 1$ to obtain $\widetilde{U}$ and similarly for Y;

$$X = L_1^{-1}\widetilde{U} + \epsilon_X, \qquad Y = L_2^{-1}\widetilde{V} + \epsilon_Y.$$

From this, we can obtain various things; $\widetilde{X} = L_1^{-1}\widetilde{U}$ gives us the part of X which is useful for explaining Y, similarly $\widetilde{Y} = L_2^{-2}\widetilde{V}$ the information in Y that is correlated with the X variables.

10.6 Example: Environmental and Genetic Correlations for Colonies of a Butterfly

The data in the butterfly file can be used to illustrate the procedure. There are 16 colonies of the butterfly *Euphydryas editha* in California and Oregon. These vary with respect to four environmental variables (altitude, annual precipitation, annual maximum temperature and annual minimum temperature) and six genetic variables (percentages of six phosphoglucose-isomerase (Pgi) genes as determined

by electrophoresis). The data may be found in `butterfly.dat` on the course home page. Significant relationships are of interest, because they may indicate that the butterfly has adapted to local environments.

Since there are fewer environment variables, the environmental variables have to be treated as X and the gene variables as Y. But all the gene frequencies *cannot* be used, since they are percentages and add up to 100.

Therefore, the 1.30 gene frequency may be omitted. It also seemed a good idea to combine the 0.40 and 0.60 gene fequencies. Thus the X variables considered are X_1 altitude, X_2 annual precipitation, X_3 annual maximum temperature, X_4 annual minimum temperature and the Y variables considered are Y_1 frequency of 0.4 and 0.6 gene, Y_2 frequency of 0.8 gene, Y_3 frequency of 1.0 gene, Y_4 frequency of the 1.16 gene.

```
> install.packages("CCA")
> library("CCA")
> install.packages("yacca")
> library("yacca")
> www =
"https://www.mimuw.edu.pl/~noble/courses/MultivariateStatistics/data/
butterfly.dat"
> butterfly <- read.table(www, header=T, quote="\"")
>   View(butterfly)
> a <- butterfly[,5]+butterfly[,6]
> y <- cbind(a,butterfly[,7:9])
> y <- simplify2array(y)
> x <- butterfly[,1:4]
> x<-simplify2array(x)
> canon <- cca(x,y,xcenter=TRUE, ycenter=TRUE, xscale=TRUE, yscale=TRUE,
  standardize.scores=TRUE)
```

FIRST these columns are *standardised*, because it is only *correlations* that are of interest here, and the rest of the analysis is performed using the standardised variables.

NEXT the correlation matrix C for the standardised variables is obtained (which is the same as the correlation matrix for the raw variables) and is partitioned into C_{11}, C_{22}, C_{12} as described earlier.

NEXT find the eigenvalues $\lambda_1, \lambda_2, \lambda_3, \lambda_4$ and eigenvectors $\underline{b}_1$, $\underline{b}_2$, $\underline{b}_3$, $\underline{b}_4$ which solve the eigenvalue problem

$$(C_{22}^{-1/2} C_{21} C_{11}^{-1} C_{12} C_{22}^{-1/2} - \lambda_j I_4)\underline{b}_j = 0$$

where I_4 denotes the 4×4 identity matrix. Then

$$C_{22}^{-1}Q^t = L_2^t = (\underline{b}_1, \underline{b}_2, \underline{b}_3, \underline{b}_4).$$

Here $\lambda_j = \rho_j^2$.
The canonical correlations are

```
> canon$corr
      CV 1       CV 2       CV 3       CV 4
0.86187348 0.45026476 0.38594833 0.08846899
```

so let

$$\hat{P} = \text{diag}(\frac{1}{0.86}, \frac{1}{0.45}, \frac{1}{0.39}, \frac{1}{0.09})$$

and set

$$H = \hat{P}QC_{22}^{-1/2}C_{21}C_{11}^{-1/2}.$$

Finally, the canonical variables $\underline{U} = HC_{11}^{-1/2}\underline{X}$ and $\underline{V} = QC_{22}^{-1/2}\underline{Y}$ are found by:

```
> canon

Canonical Correlation Analysis

Canonical Correlations:
      CV 1       CV 2       CV 3       CV 4
0.86187348 0.45026476 0.38594833 0.08846899

X Coefficients:
              CV 1       CV 2       CV 3        CV 4
alt     -0.1243327 -2.4315539 -2.9501097 -1.36853772
prec    -0.2931481  0.6752071 -1.3581107 -0.24164723
maxtemp  0.4682769 -0.4785208 -0.5772789 -1.70143588
mintemp  0.2597280 -1.4033550 -3.5314287  0.08820545

Y Coefficients:
               CV 1     CV 2     CV 3       CV 4
a         0.5479728 1.766016 3.483095 -0.6632469
gene0p8   0.4217912 2.256306 1.295853  1.4087920
gene1p0  -0.0885412 3.850887 3.747371  0.5044550
gene1p16  0.8256279 2.848238 2.745531 -0.6358737
```

```
Structural Correlations (Loadings) - X Vars:
              CV 1        CV 2        CV 3       CV 4
alt     -0.9214647 -0.33881431  0.06238689 -0.1794870
prec    -0.7708669  0.51342567 -0.26736578 -0.2658457
maxtemp  0.8983191  0.20232579 -0.02405527 -0.3892409
mintemp  0.9193910  0.05251474 -0.22853336  0.3158084

Structural Correlations (Loadings) - Y Vars:
               CV 1       CV 2        CV 3       CV 4
a         0.3841633 -0.6200365  0.60266987  0.3236701
gene0p8   0.7395508 -0.1493935  0.08335961  0.6509972
gene1p0  -0.9610740  0.2508416  0.00196467 -0.1158075
gene1p16  0.4753453  0.5147425 -0.44237008 -0.5598175

Aggregate Redundancy Coefficients (Total Variance Explained):
 X | Y: 0.6017311
 Y | X: 0.4024482
```

This gives:

$$\begin{cases} U_1 = -0.12X_1 - 0.29X_2 + 0.47X_3 + 0.26X_4 \\ V_1 = +0.55Y_1 + 0.42Y_2 - 0.08Y_3 + 0.83Y_4, \end{cases}$$

$$\begin{cases} U_2 = 2.43X_1 - 0.68X_2 + 0.48X_3 + 1.40X_4 \\ V_2 = -1.76Y_1 - 2.26Y_2 - 3.85Y_3 - 2.85Y_4, \end{cases}$$

$$\begin{cases} U_3 = -2.95X_1 - 1.36X_2 - 0.58X_3 - 3.53X_4 \\ V_3 = 3.48Y_1 + 1.30Y_2 + 3.75Y_3 + 2.75Y_4, \end{cases}$$

$$\begin{cases} U_4 = -1.37X_1 - 0.24X_2 - 1.70X_3 - 0.09X_4 \\ V_4 = -0.66Y_1 - 1.41Y_2 - 0.50Y_3 - 0.64Y_4, \end{cases}$$

The correlations are between observed variables and canonical variables are known as the *canonical loadings*.

To perform tests of the significance of Canonical Correlation, try:

```
> canon$chisq
      CV 1       CV 2       CV 3       CV 4
18.4140962  4.1550757  1.7760659  0.0825043
```

```
> canon$df
CV 1 CV 2 CV 3 CV 4
  16    9    4    1
```

To 'look up' the chi squared table,

```
> pchisq(canon$chisq,canon$df,ncp=0)
      CV 1       CV 2       CV 3       CV 4
0.69978802 0.09908763 0.22314179 0.22606809
```

Although the canonical correlations seem quite large, Bartlett's test does not reject the null hypothesis (that they are insignificant) because the sample size is rather small. It is found that $X^2 = 18.41$ with $16df$. The probability of obtaining a value greater than this is 0.30. This is not sufficiently small to reject the null hypothesis.

Nevertheless, the canonical correlation provides useful pointers at a descriptive level. U_1 is mainly a contrst between the maximum and minimum temperatures on the one hand and precipitation on the other. V_1 has moderate to large coefficients for Y_1, Y_2 and Y_4, with a small negative coefficient for Y_3. It appears that the $0.4, 0.6, 0.8$ and 1.16 genes tend to be frequent in colonies with high temperatures and low precipitation.

If the *correlations* are studied instead of the *coefficients*, a slightly different picture emerges. The correlations between the environmental variables and U_1 are

	X_1	X_2	X_3	X_4
U_1	-0.92	-0.77	0.90	0.92

and the correlations between V_1 and the gene variables are

	Y_1	Y_2	Y_3	Y_4
V_1	0.38	0.74	-0.96	0.48

This suggests that U_1 is best interpreted as a measure of high temperatures and low altitude and low precipitation.

V_1 comes out clearly indicating a lack of 1.00 genes.

Note that the interpretations differ when made on the basis of the coefficients and when made on the basis of correlations. For U_1 the difference is not substantial, but for V_1 the importance of the 1.00 gene is very different. On the whole, for this data set, the interpretations based on correlations seem best. For example, the colony GL, which has highest altitude, high precipitation and low temperatures has the highest frequency of 1.00 genes. The colony UO with low altitude, low precipitation, high temperatures has the lowest frequency of 1.00 genes.

Try plotting the values of V_1 against U_1.

```
> u1<- x%*%canon$xcoef[,1]
> v1 <- y%*%canon$ycoef[,1]
```

There is one outlier; one of the colonies is somewhat unusual compared with the other colonies. From the interpretations given for U_1 and V_1, it would seem that the frequency of 1.00 genes is unusually high for a colony with this environment.

One final remark should be made: if the colonies are located close to each other, then it may be naive to consider them as 16 *independent* observations.

10.7 Exercises

1. Consider the covariance matrix

$$\text{Cov}\left(\begin{array}{c} X_1 \\ X_2 \\ \hline Y_1 \\ Y_2 \end{array}\right) = \left(\begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 100 & 0 & 0 & 0 \\ 0 & 1 & 0.95 & 0 \\ \hline 0 & 0.95 & 1 & 0 \\ 0 & 0 & 0 & 100 \end{array}\right)$$

   Verify that the first pair of canonical variates are $U_1 = X_2$, $V_1 = Y_1$ with canonical correlation $\rho_1 = 0.95$.

2. Suppose $\underline{X}$ and $\underline{Y}$ are standardised variables (i.e. each with mean zero and unit variance), with correlation

$$R = \left(\begin{array}{cc} R_{11} & R_{12} \\ R_{21} & R_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 1.0 & 0.4 & 0.5 & 0.6 \\ 0.4 & 1.0 & 0.3 & 0.4 \\ \hline 0.5 & 0.3 & 1.0 & 0.2 \\ 0.6 & 0.4 & 0.2 & 1.0 \end{array}\right).$$

   You may assume that

$$R_{11}^{-1/2} = \left(\begin{array}{cc} 1.0681 & -0.2229 \\ -0.2229 & 1.0681 \end{array}\right)$$

$$R_{22}^{-1} = \left(\begin{array}{cc} 1.0417 & -0.2083 \\ -0.2083 & 1.0417 \end{array}\right)$$

$$R_{11}^{-1/2} R_{12} R_{22}^{-1} R_{21} R_{11}^{-1/2} = \left(\begin{array}{cc} 0.4371 & 0.2178 \\ -0.2178 & 0.1096 \end{array}\right).$$

   Determine the canonical correlations ρ_1 and ρ_2 and the pairs of canonical variates (U_1, V_1) and (U_2, V_2).

3. The 2-random vectors $\underline{X}$ and $\underline{Y}$ have joint mean vector and joint covariance matrix

$$\underline{\mu} = \left(\begin{array}{c} \underline{\mu}_X \\ \underline{\mu}_Y \end{array}\right) = \left(\begin{array}{c} -3 \\ 2 \\ \hline 0 \\ 1 \end{array}\right),$$

$$C = \left(\begin{array}{cc} C_{11} & C_{12} \\ C_{21} & C_{22} \end{array}\right) = \left(\begin{array}{cc|cc} 8 & 2 & 3 & 1 \\ 2 & 5 & -1 & 3 \\ \hline 3 & -1 & 6 & -2 \\ 1 & 3 & -2 & 7 \end{array}\right)$$

   (a) Calculate the canonical correlations ρ_1 and ρ_2.

(b) Determine the canonical variate pairs (U_1, V_1) and (U_2, V_2).

4. (a) Show that the canonical correlations are invariant under nonsingular linear transformations of $(\underline{X}, \underline{Y})$ of the form $(M_1\underline{X}, M_2\underline{Y})$.

You may do this by considering

$$\mathrm{Cov}\begin{pmatrix} M_1\underline{X} \\ M_2\underline{Y} \end{pmatrix} = \begin{pmatrix} M_1C_{11}M_1 & M_1C_{12}M_2^t \\ M_2C_{21}M_1^t & M_2C_{22}M_2^t \end{pmatrix}$$

(b) Let $\underline{X}$ and $\underline{Y}$ be two sets of variables and let

$$E\left[\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix}\right] = \begin{pmatrix} \underline{\mu}_X \\ \underline{\mu}_Y \end{pmatrix}, \qquad \mathrm{Cov}\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix} = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}.$$

Set

$$\underline{Z}_1 = C_{11}^{-1/2}(\underline{X} - \underline{\mu}_X), \qquad \underline{Z}_2 = C_{22}^{-1/2}(\underline{Y} - \underline{\mu}_Y).$$

If $(U_i, V_i) = (L_{i\cdot}^{(1)}\underline{X}, L_{i\cdot}^{(2)}\underline{Y})$ are the canonical variables for $(\underline{X}, \underline{Y})$, $i = 1, \ldots, p$, with canonical correlations $\rho_1, \ldots, \rho_p$, determine the canonical variates and the canonical correlations for the sets $(\underline{Z}_1, \underline{Z}_2)$.

5. (a) Let R denote the correlation matrix of $\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix}$, where

$$R = \left(\begin{array}{cc|cc} 1 & \rho & \rho & \rho \\ \rho & 1 & \rho & \rho \\ \hline \rho & \rho & 1 & \rho \\ \rho & \rho & \rho & 1 \end{array}\right).$$

Determine the canonical coordinates corresponding to non zero canonical correlation.

(b) Generalise the results of the previous part to the case where $\underline{X}$ has p components and $\underline{Y}$ has q components. Assume $\rho > 0$.

Hint: $R_{12} = \rho\mathbf{1}_p\mathbf{1}_q^t$ where $\mathbf{1}_k$ denotes a vector length k with each entry 1 and

$$R_{11}\mathbf{1}_p = (1 + (p-1)\rho)\mathbf{1}_p$$

so that

$$R_{11}^{-1/2}\mathbf{1}_p = (1 + (p-1)\rho)^{-1/2}\mathbf{1}.$$

6. (Correlation for angular measurements) Some observations, such as wind direction are in the form of angles. Let $\underline{Y} = (\cos(\theta), \sin(\theta))^t$.

(a) Show that $\underline{b}^t\underline{Y} = \sqrt{b_1^2 + b_2^2}\cos(\theta - \beta)$, where

$$\frac{1}{\sqrt{b_1^2 + b_2^2}}(b_1, b_2) = (\cos(\beta), \sin(\beta))$$

(use $\cos(\theta - \beta) = \cos(\theta)\cos(\beta) + \sin(\theta)\sin(\beta)$)

(b) Let X have a single component. Show that the single canonical correlation is

$$\rho_1 = \max_{\beta} \mathrm{Corr}(X, \cos(\theta - \beta)).$$

Selecting the canonical variable V_1 therefore amounts to selecting a new origin for the angular variable θ.

(c) Let X denote the amount of ozone (in parts per million) and θ the wind direction measured from the north. Nineteen measurements made in downtown Milwaukee, Wisconsin, give the sample correlation matrix

$$R = \left(\begin{array}{c|cc} 1.0 & 0.166 & 0.694 \\ \hline 0.166 & 1.0 & -0.051 \\ 0.694 & -0.051 & 1.0 \end{array}\right).$$

Compute the sample canonical correlation $\hat{\rho}_1$ and the sample canonical variate $\hat{V}_1$ representing β, the new origin.

(d) Suppose that $\underline{X} = (\cos(\phi), \sin(\phi))$, so that

$$\underline{a}^t\underline{X} = \sqrt{a_1^2 + a_2^2}\cos(\phi - \alpha).$$

Prove that

$$\rho_1 = \max_{\alpha,\beta} \mathrm{Corr}(\cos(\phi - \alpha), \sin(\theta - \beta)).$$

(e) Twenty one observations on the 6.00 a.m. and noon wind directions give the correlation matrix

$$R = \left(\begin{array}{cc|cc} 1.0 & -0.291 & 0.440 & 0.372 \\ -0.291 & 1.0 & -0.205 & 0.243 \\ \hline 0.440 & -0.205 & 1.0 & 0.181 \\ 0.372 & 0.243 & 0.181 & 1.0 \end{array}\right)$$

Find the first sample canonical correlation $\hat{\rho}_1$ and corresponding canonical variates $(\hat{U}_1, \hat{V}_1)$.

Short Answers

1. The correlation matrix is

$$R = \left(\begin{array}{cc|cc} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.95 & 0 \\ \hline 0 & 0.95 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{array}\right)$$

The first pair $U_1 = a_1X_1 + a_2X_2$, $V_1 = b_1Y_1 + b_2Y_2$. They must satisfy $\mathrm{Var}(U_1) = \mathrm{Var}(V_1) = 1$. Since X_1, X_2 are independent and Y_1, Y_2 are independent, it follows that $a_1^2 + a_2^2 = b_1^2 + b_2^2 = 1$ to give highest correlation

$$\begin{aligned} &R(a_1X_1 + a_2X_2, b_1Y_1 + b_2Y_2) \\ &\quad = a_1b_1R(X_1, Y_1) + a_1b_2R(X_1, Y_2) + a_2b_1R(X_2, Y_1) + a_2b_2R(X_2, Y_2) = 0.95a_2b_1 \end{aligned}$$

Maximised if $a_2 = b_1 = 1$ (or $a_2 = b_1 = -1$). Hence (up to sign)

$$U_1 = X_2, \qquad V_1 = Y_1.$$

2. Since $p = q = 2$ (using notation of lectures) we can exchange the roles of the $\underline{X}$ and $\underline{Y}$ variables. Let $L^{(1)}$ and $L^{(2)}$ denote the linear transformations such that $L^{(1)}\underline{X} = \underline{U}$ and $L^{(2)}\underline{Y} = \underline{V}$. Then, from lectures, $L^{(1)} = QR_{11}^{-1/2}$ and $L^{(2)} = HR_{22}^{-1/2}$ where Q and H are 2×2 orthonormal matrices that have to be determined, which satisfy

$$H^tPQ = R_{22}^{-1/2}R_{21}R_{11}^{-1/2}.$$

From this,

$$Q(R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2})Q^t = P^tP$$

where $P^tP = \mathrm{diag}(\rho_1^2, \ldots, \rho_k^2, 0, \ldots, 0)$. It follows that $\rho_j^2 = \lambda_j$ where λ_j is the jth eigenvalue of $M := R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2}$. The eigenvalues are computed as the roots of the equation $|M - \lambda I| = 0$.

Furthermore,

$$H^tPQ = R_{22}^{-1/2}R_{21}R_{11}^{-1/2}$$

gives

$$\begin{aligned} H^tP &= R_{22}^{-1/2}R_{21}R_{11}^{-1/2}Q^t \\ P^tH &= QR_{11}^{-1/2}R_{12}R_{22}^{-1/2} \\ PL^{(2)} &= L^{(1)}R_{12}R_{22}^{-1} \end{aligned}$$

In this example,

$$0 = \begin{vmatrix} 0.4371-\lambda & 0.2178 \\ 0.2178 & 0.1096-\lambda \end{vmatrix} = (0.4371-\lambda)(0.1096-\lambda) - (2.178)^2 = \lambda^2 - 0.5467\lambda + 0.0005$$

$$\rho_1^2 = \lambda_1 = 0.5458, \qquad \rho_2^2 = \lambda_2 = 0.0009.$$

Now, compute Q. The *row* $(Q_{j1}, \ldots, Q_{jq})$ satisfies

$$M\underline{e}_j = \lambda_j \underline{e}_j$$

where

$$\underline{e}_j^t = (Q_{j1}, \ldots, Q_{jq}).$$

It follows that

$$\begin{pmatrix} 0.4371 & 0.2178 \\ 0.2178 & 0.1096 \end{pmatrix} \underline{e}_1 = 0.5458\underline{e}_1$$

giving

$$\underline{e}_1 = \begin{pmatrix} 0.8947 \\ 0.4466 \end{pmatrix}$$

$$(L_{11}^{(1)}, L_{12}^{(1)}) = (Q_{11}, Q_{12})R_{11}^{-1/2} = (0.8561, 0.2776).$$

It follows that

$$U_1 = 0.8561X_1 + 0.2776X_2.$$

For $(L_{11}^{(2)}, L_{12}^{(2)})$,

$$\rho_1(L_{11}^{(2)}, L_{12}^{(2)}) = (L_{11}^{(1)}, L_{12}^{(1)})R_{12}R_{22}^{-1} = (0.8561, 0.2776)\begin{pmatrix} 0.3959 & 0.5209 \\ 0.2292 & 0.3542 \end{pmatrix} = (0.4026, 0.5443)$$

Note that

$$\rho_1^2 = 0.5460$$

$$(L_{11}^{(2)}, L_{12}^{(2)}) = \frac{1}{\sqrt{0.5460}}(0.4026, 0.5443) = (0.5448, 0.7366).$$

$$U_1 = 0.86X_1 + 0.28X_2$$

$$V_1 = 0.54Y_1 + 0.74Y_2$$

Their canonical correlation is $\rho_1 = \sqrt{0.5458} = 0.74$. Computation of (U_2, V_2) is similar.

3. Computation is the same as the previous example - first create a correlation matrix

$$R = \left(\begin{array}{cc|cc} 1 & 1/\sqrt{10} & \sqrt{3}/4 & 1/2\sqrt{14} \\ 1/\sqrt{10} & 1 & -1/\sqrt{30} & 3/\sqrt{35} \\ \hline \sqrt{3}/4 & -1/\sqrt{30} & 1 & -\sqrt{2/21} \\ 1/2\sqrt{14} & 3/\sqrt{35} & -\sqrt{2/21} & 1 \end{array}\right)$$

4. (a) Suppose a canonical correlation analysis is carried out on the *covariance* matrix in the old coordinates, where $\underline{X}$ is a p vector and $\underline{Y}$ is a q vector and suppose the analysis yields that there are k correlated variables, $(\underline{U}, \underline{V}) = (L^{(1)}\underline{X}, L^{(2)}\underline{Y})$, where where $L^{(1)}$ is a $k \times p$ matrix and $L^{(2)}$ is a $k \times q$ matrix. Then there is a $p \times p$ matrix $\tilde{L}^{(1)}$ and a $q \times q$ matrix $\tilde{L}^{(2)}$ where the first k rows of $\tilde{L}^{(1)}$ are $L^{(1)}$ and the first k rows of $\tilde{L}^{(2)}$ are $L^{(2)}$ and such that

$$\tilde{L}^{(1)} C_{11} \tilde{L}^{(1)t} = I_p, \qquad \tilde{L}^{(2)} C_{22} \tilde{L}^{(2)t} = I_q, \qquad \tilde{L}^{(1)} C_{12} \tilde{L}^{(2)t} = P,$$

where

$$P_{ii} = \rho_i, \quad i = 1, \ldots, k, \qquad P_{ij} = 0 \qquad \text{other} \quad (i, j)$$

and $\rho_1, \ldots, \rho_k$ are the correlations.

Now try a canonical correlation on $M_1\underline{X}$ and $M_2\underline{Y}$. There is a $p \times p$ matrix $T^{(1)}$ and a $q \times q$ matrix $T^{(2)}$ such that

$$T^{(1)} M_1 C_{11} M_1^t T^{(1)t} = I_p, \qquad T^{(2)} M_2 C_{22} M_2^t T^{(2)t} = I_q, \qquad T^{(1)} M_1 C_{12} M_2^t T^{(2)} = P_2$$

where P_2 only has non zero elements on the diagonals, which are the canonical correlations. Comparing the first two equations for both pairs gives (easily) that

$$T^{(1)} M_1 = H_1 L^{(1)}, \qquad T^{(2)} M_2 = H_2 L^{(2)}$$

for some orthonormal matrices H_1 and H_2. It follows that $P_2 = H_1 P H_2^t$. Hence $P_2 = P$, hence the canonical correlations are the same and the k transformations are given by $\tilde{T}^{(1)} = \tilde{L}^{(1)} M_1^{-1}$ and $\tilde{T}^{(2)} M_2^{-1}$. It follows that the canonical variables for the new coordinates are $(\underline{U}, \underline{V})$, as before.

IMPORTANT POINT: This means that, after centring the variables, one can apply a canonical correlation analysis to the original (centred) variables using the covariance matrix, or to the standardised variables using the correlation matrix. The resulting pairs $(\underline{U}, \underline{V})$ will be the same.

(b) The canonical correlation analysis is based on the *covariance* matrix; the centring is not taken into consideration. From the preceeding part, if the transformations for the original variables are $\tilde{L}^{(1)}, \tilde{L}^{(2)}$ then for $(\underline{Z}_1, \underline{Z}_2)$, the transformations are $L^{(1)} C_{11}^{1/2}$ and $L^{(2)} C_{22}^{1/2}$. Therefore, for the transformed variables, the correlated variables are

$$(\tilde{U}_i, \tilde{V}_i) = (\underline{a}_i^t C_{11}^{1/2} \underline{Z}_1, \underline{b}_i^t C_{22}^{1/2} \underline{Z}_2) = (U_i - \underline{a}_i^t \underline{\mu}_X, V_i - \underline{b}_i^t \underline{\mu}_Y).$$

5. (a) Consider R_{11}. The eigenvalues are solutions to $(1 - \lambda)^2 - \rho^2 = 0$ giving $\lambda_1 = 1 + \rho$ and $\lambda_2 = 1 - \rho$. Then, since $L^{(1)} R_{11} L^{(1)t} = I_2$, solving

$$(R_{11} - (1 + \rho))\underline{v}_1 = 0, \qquad (R_{11} - (1 - \rho))\underline{v}_2 = 0$$

gives (the rows may be multiplied by ± 1 and they may be swapped around)

$$L^{(1)} = \begin{pmatrix} \frac{1}{\sqrt{2(1+\rho)}} & \frac{1}{\sqrt{2(1+\rho)}} \\ \frac{1}{\sqrt{2(1-\rho)}} & -\frac{1}{\sqrt{2(1-\rho)}} \end{pmatrix}.$$

Similarly (the rows may have to be altered by multiplying by ± 1 and they may be swapped around)

$$L^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2(1+\rho)}} & \frac{1}{\sqrt{2(1+\rho)}} \\ \frac{1}{\sqrt{2(1-\rho)}} & -\frac{1}{\sqrt{2(1-\rho)}} \end{pmatrix}.$$

So that P is a diagonal matrix, with values non negative and in descending order, it follows that (for $\rho > -\frac{1}{3}$) the matrix of correlations is

$$P = \rho L^{(1)} \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix} L^{(2)t} = \begin{pmatrix} \frac{2\rho}{1+\rho} & 0 \\ 0 & 0 \end{pmatrix}$$

and there is only one pair of correlated variables;

$$L^{(1)} = \begin{pmatrix} \frac{1}{\sqrt{2(1+\rho)}} & \frac{1}{\sqrt{2(1+\rho)}} \end{pmatrix} \qquad U = \frac{1}{\sqrt{2(1+\rho)}}(X_1 + X_2)$$

$$L^{(2)} = \begin{pmatrix} \frac{1}{\sqrt{2(1+\rho)}} & \frac{1}{\sqrt{2(1+\rho)}} \end{pmatrix} \qquad U = \frac{1}{\sqrt{2(1+\rho)}}(Y_1 + Y_2)$$

with correlation $\frac{2\rho}{1+\rho}$. Note: the covariance matrix for $\begin{pmatrix} \underline{X} \\ \underline{Y} \end{pmatrix}$ is no longer positive definite and hence no longer a covariance matrix if $\rho < -\frac{1}{3}$.

(b) Following a similar technique: firstly note that $1-\rho$ is an eigenvalue of multiplicity $p-1$ for R_{11}, since $R_{11} - (1-\rho)I_p$ is the matrix with each entry ρ and it is possible to find $p-1$ vectors, $\underline{v}_2, \ldots, \underline{v}_p$, orthonormal to each other and satisfying $\sum_{j=1}^{p} v_{kj} = 0$ such that $(R_{11} - (1-\rho)I_p)\underline{v}_k = 0$. They are orthogonal to $\underline{v}_1 \frac{1}{\sqrt{p}}(1, \ldots, 1)^t$ which is therefore the remaining eigenvector and solving gives the remaining eigenvalue $\lambda_1 = 1 + (p-1)\rho$. Similarly for R_{22}.

It follows (as before) by considering $P = L^{(1)} R_{12} L^{(2)t}$ that there is exactly one pair of correlated variables (because, similarly to before, the other entries of the matrix P are all zero) and the pair is

$$U_1 = c_1 \sum_{j=1}^{p} X_j$$

$$V_1 = c_2 \sum_{j=1}^{q} Y_j$$

for constants c_1 and c_2 to be determined. To ensure that $\mathrm{Var}(U_1) = 1$, it follows that $c_1 = \frac{1}{\sqrt{p(1+(p-1)\rho)}}$ and $c_2 = \frac{1}{\sqrt{q(1+(q-1)\rho)}}$, with canonical correlation

$$\rho_1 = \mathrm{Cov}(U_1, V_1) = \frac{\sqrt{pq}\rho}{\sqrt{(1+(p-1)\rho)(1+(q-1)\rho)}}.$$