STT 843 Key to Homework 3 Spring 2018

Due date: April 16, 2018

8.4. The covariance matrix is

$$\Sigma = \sigma^2 \begin{pmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix} = \sigma^2 R.$$

The eigenvalues of the above matrix could be found by solving $\det(R - \lambda I_3) = 0$. That is

$$\det(R - \lambda I_3) = (1-\lambda)\{(1-\lambda)^2 - \rho^2\} - \rho^2(1-\lambda) = (1-\lambda)\{(1-\lambda)^2 - 2\rho^2\}.$$

Then, the three eigenvalues of R are

$$\lambda_1 = 1 + \sqrt{2}|\rho|, \quad \lambda_2 = 1, \quad \lambda_3 = 1 - \sqrt{2}|\rho|.$$

Due to the symmetry of the eigenvalues, without loss of generality, assume that $\rho \geq 0$. Otherwise, switching λ_1 and λ_3. The corresponding eigenvectors $e_i = (e_{i1}, e_{i2}, e_{i3})'$ satisfy the following equations

$$\begin{pmatrix} 1-\lambda_i & \rho & 0 \\ \rho & 1-\lambda_i & \rho \\ 0 & \rho & 1-\lambda_i \end{pmatrix} \begin{pmatrix} e_{i1} \\ e_{i2} \\ e_{i3} \end{pmatrix} = 0.$$

Therefore, the corresponding eigenvectors are, respectively,

$$e_1 = \begin{pmatrix} 1/2 \\ 1/\sqrt{2} \\ 1/2 \end{pmatrix} \quad e_2 = \begin{pmatrix} 1/\sqrt{2} \\ 0 \\ -1/\sqrt{2} \end{pmatrix} \quad \text{and} \quad e_3 = \begin{pmatrix} 1/2 \\ -1/\sqrt{2} \\ 1/2 \end{pmatrix}.$$

The total variance is $3\sigma^3$. The first principal component is $Y_1 = e_1^T X$, which has variance $\sigma^2(1+\sqrt{2}\rho)$ and explains $(1+\sqrt{2}\rho)/3$ portion of the total population variance. The second principal component is $Y_2 = e_2^T X$, which has variance σ^2 and explains $1/3$ portion of the total population variance. The third principal component is $Y_3 = e_3^T X$, which has variance $\sigma^2(1-\sqrt{2}\rho)$ and explains $(1-\sqrt{2}\rho)/3$ portion of the total population variance.

8.10. (a) Let $X_1, \cdots, X_5$ represent, respectively, the stock price of JP Morgan, Citibank, Wells Fargo, Royal Dutch Shell and Exxon Mobil. Then, the sample covariance S is given by

	X_1	X_2	X_3	X_4	X_5
X_1	$4.33e-04$	$2.75e-04$	$1.59e-04$	$6.41e-05$	$8.89e-05$
X_2	$2.75e-04$	$4.38e-04$	$1.79e-04$	$1.81e-04$	$1.23e-04$
X_3	$1.59e-04$	$1.79e-04$	$2.24e-04$	$7.34e-05$	$6.05e-05$
X_4	$6.41e-05$	$1.81e-04$	$7.34e-05$	$7.22e-04$	$5.08e-04$
X_5	$8.89e-05$	$1.23e-04$	$6.05e-05$	$5.08e-04$	$7.65e-04$

The sample principal components are

$$\begin{aligned}
Y_1 &= 0.22X_1 + 0.31X_2 + 0.15X_3 + 0.64X_4 + 0.65X_5;\\
Y_2 &= 0.63X_1 + 0.57X_2 + 0.34X_3 - 0.25X_4 - 0.32X_5;\\
Y_3 &= 0.33X_1 - 0.25X_2 - 0.04X_3 - 0.64X_4 + 0.65X_5;\\
Y_4 &= 0.66X_1 - 0.41X_2 - 0.50X_3 + 0.31X_4 - 0.22X_5;\\
Y_5 &= 0.12X_1 - 0.59X_2 + 0.78X_3 + 0.15X_4 - 0.09X_5.
\end{aligned}$$

(b) Because the sample variance explained by Y_1, Y_2 and Y_3 are, respectively, $\lambda_1 = 0.0013676780$, $\lambda_2 = 0.0007011596$ and $\lambda_3 = 0.0002538024$, the proportion of variance explained by the first three components are

$$\text{Proportion} = \frac{\lambda_1 + \lambda_2 + \lambda_3}{\lambda_1 + \lambda_2 + \lambda_3 + \lambda_4 + \lambda_5} = 0.899.$$

The proportions explained by the first three components are, respectively, 0.529, 0.271, and 0.098.

The coefficients of X_4 and X_5 both are positive and relative large in Y_1, which suggests that the stock prices of gasoline companies contribute more to the first principal component. and the stock prices of gasoline companies vary together. This component may be viewed as a representation of gasoline companies. For the second principal component Y_2, the coefficients on financial companies are larger when it compared with the coefficients of gasoline companies. This suggests that financial companies' stock prices contribute more to the second principal component. This component maybe viewed as a representation of gasoline companies. In the third component, the magnitudes of coefficients of X_4 and X_5 are larger than the other coefficients. The coefficients of X_4 and X_5 are in opposite direction, which might suggests the competition between two gasoline companies and the stock prices of them are negatively correlated in this component.

(c) Because of the asymptotic normality of the sample eigenvalues, the individual confidence intervals for λ_i are given by

$$(\hat{\lambda}_i - z_{\alpha/2}\sqrt{2\hat{\lambda}_i^2/n}, \hat{\lambda}_i + z_{\alpha/2}\sqrt{2\hat{\lambda}_i^2/n}).$$

Then, the Bonferroni simultaneous confidence intervals for λ_i are given by

$$(\hat{\lambda}_i - z_{\alpha/6}\sqrt{2\hat{\lambda}_i^2/n}, \hat{\lambda}_i + z_{\alpha/6}\sqrt{2\hat{\lambda}_i^2/n}).$$

Thus, the simultaneous confidence intervals for λ_1, λ_2 and λ_3 are, respectively, (9.621124e-04, 0.0017732437), (4.932406e-04, 0.0009090785) and (1.785409e-04, 0.0003290640).

Note that, another type of asymptotic simultaneous confidence intervals for for λ_i are given by

$$(\hat{\lambda}_i/\{1 + z_{\alpha/6}\sqrt{2/n}\}, \hat{\lambda}_i/\{1 - z_{\alpha/6}\sqrt{2/n}\}).$$

(d) The simultaneous confidence intervals given in part (c) indicate that the eigenvalues λ_1 and λ_2 are significantly larger than the eigenvalue λ_3, hence it is also larger than λ_4 and λ_5. This suggests that the most of total variance can be explained by the first two principal components. The proportion of variance explained by the first two components is 80.06%.

8.28. (a) The scatter plots of Family versus DistRd and DistRD versus Cattle are given in Figure 1. The outliers are labeled in the scatter plots. In the first scatter plot, the 25, 69, and 72-th data points are obvious outliers. In the second scatter plot, the 34, 69, and 72-th data points are obvious outliers.

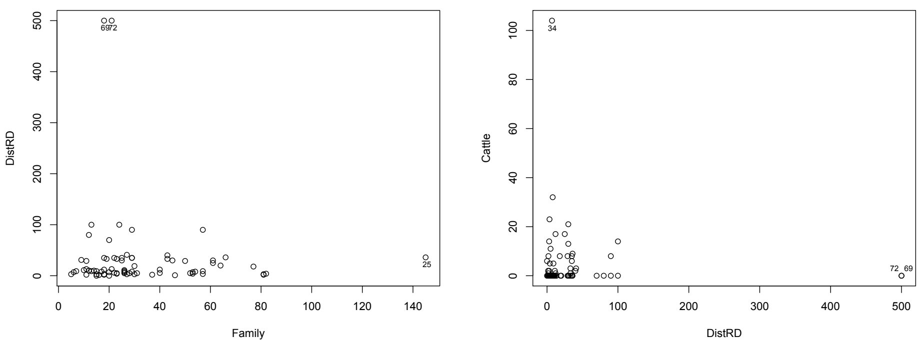

Figure 1: Left Panel: scatter plot of Family versus DistRD. Right Panel: scatter plot of DistRD versus Cattle

(b) The principal component analysis using correlation matrix was done in R (see code part). A scree plot of the eigenvalues is given in Figure 2. It is clear that there is a significant drop at the first to second eigenvalue but the first principal component only explains about 46.5% variation. We calculate the drop of the eigenvalues in the following:

-2.747 -0.353 -0.292 -0.187 -0.238 -0.126 -0.068 -0.053

Based on the above information, the second significant elbow point is at the fifth to the sixth eigenvalue. Thus, using five components is appropriate. With five principal components, the total proportion of variation explained by the first five components is about 90%.

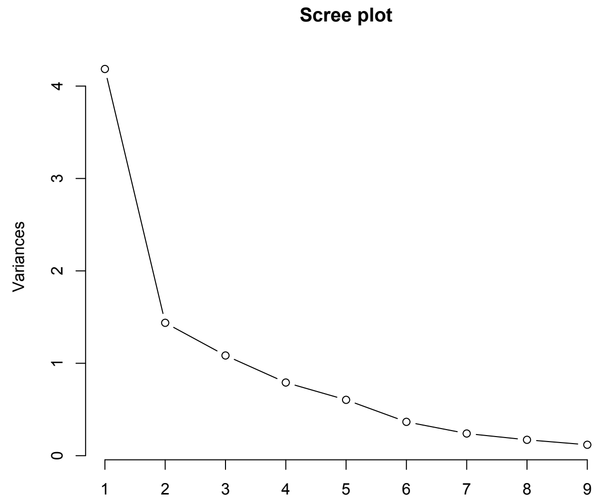

Figure 2: A scree plot of eigenvalues of correlation matrix

(c) The coefficients of the first five components are given in the following matrix

	$PC1$	$PC2$	$PC3$	$PC4$	$PC5$
$Family$	0.433	−0.065	0.098	−0.171	0.011
$DistRD$	0.007	0.496	−0.568	−0.495	−0.377
$Cotton$	0.446	0.008	0.132	0.027	−0.218
$Maize$	0.352	0.352	0.388	−0.240	−0.079
$Sorg$	0.203	−0.603	−0.111	0.058	−0.644
$Millet$	0.240	−0.415	−0.115	−0.616	0.526
$Bull$	0.445	0.068	−0.030	0.145	−0.028
$Cattle$	0.355	0.284	0.013	0.372	0.217
$Goats$	0.254	−0.048	−0.686	0.350	0.248

For the first component, all the coefficients are comparable except the coefficient for DistRD. This component may be considered as a farm size component. The third component has relatively large coefficients on DistRD and Goats, which might be called “goats and distant to road” component. The second component has largest coefficients on DistRD, Maize, Sorg and Millet. This component might be interpreted as the arable farming component. The fourth component has large coefficients on DistRD, Millet, Cattle and Goats, and the coefficients for Millet and Cattle and Goats are opposite, which might be a "competition" between arable versus pastoral farming. The fifth component has large coefficients on Sorg and Millet, and both have opposite signs. This might means that these two crops are typically not planted in the same farm.

10.2. (a) The canonical correlations ρ_1^* and ρ_2^* can be found by computing the eigenvalues of

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21} = \begin{pmatrix} 0.27704678 & -0.02412281 \\ -0.04239766 & 0.26754386 \end{pmatrix}$$

The eigenvalues are 0.3046268 and 0.2399638. Therefore, the canonical correlations are $\rho_1^* = 0.552$ and $\rho_2^* = 0.489$.

(b) The canonical pairs (U_1, V_1) and (U_2, V_2) could be found through finding the eigenvectors of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ and $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$. The eigenvectors of $\Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\Sigma_{11}^{-1/2}$ are

$$\begin{pmatrix} -0.742 & -0.670 \\ 0.670 & -0.742 \end{pmatrix}$$

and the eigenvectors of $\Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22}^{-1/2}$ are

$$\begin{pmatrix} -0.919 & -0.393 \\ 0.393 & -0.919 \end{pmatrix}.$$

Thus, the first canonical pair is $U_1 = -0.3168X_1^{(1)} + 0.3622X_2^{(1)}$ and $V_1 = -0.3647X_1^{(2)} + 0.09506X_2^{(2)}$. The second canonical pair is $U_2 = -0.1962X_1^{(1)} - 0.3017X_2^{(1)}$ and $V_2 = -0.2262X_1^{(2)} - 0.3858X_2^{(2)}$.

(c) We can write U, V as linear combinations of $X^{(1)}$ and $X^{(2)}$. That is

$$\begin{pmatrix} U_1 \\ U_2 \\ V_1 \\ V_2 \end{pmatrix} = A \begin{pmatrix} X_1^{(1)} \\ X_2^{(1)} \\ X_1^{(2)} \\ X_2^{(2)} \end{pmatrix}.$$

where

$$A = \begin{pmatrix} -0.3168 & 0.3622 & 0.0000 & 0.0000 \\ -0.1962 & -0.3017 & 0.0000 & 0.0000 \\ 0.0000 & 0.0000 & -0.3647 & 0.0951 \\ 0.0000 & 0.0000 & -0.2263 & -0.3858 \end{pmatrix}.$$

Then, the expectations of canonical pairs are given by

$$E\begin{pmatrix} U_1 \\ U_2 \\ V_1 \\ V_2 \end{pmatrix} = A \begin{pmatrix} \mu_1^{(1)} \\ \mu_2^{(1)} \\ \mu_1^{(2)} \\ \mu_2^{(2)} \end{pmatrix} = \begin{pmatrix} 1.6749 \\ -0.0146 \\ 0.0950 \\ -0.3858 \end{pmatrix}$$

and the covariance is given by

$$\text{Cov}\begin{pmatrix} U_1 \\ U_2 \\ V_1 \\ V_2 \end{pmatrix} = A\Sigma A' = \begin{pmatrix} 1.0000 & 0.0000 & 0.5519 & 0.0000 \\ 0.0000 & 1.0000 & 0.0000 & 0.4899 \\ 0.5519 & 0.0000 & 1.0000 & 0.0000 \\ 0.0000 & 0.4899 & 0.0000 & 1.0000 \end{pmatrix}.$$

Comparing the above covariance matrix with the properties in Result 10.1 in textbook, all the properties about covariances between U and V are verified.

10.10. (a) The sample canonical correlations can be found through the eigenvalues of the matrix $R_{11}^{-1}R_{12}R_{22}^{-1}R_{21}$, which are 0.10668190 and 0.02926479. Therefore, the canonical correlations are 0.3266219 and 0.1710696.

(b) To obtain the first pair of canonical pairs, we compute eigenvectors of $R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2}$ and eigenvectors of $R_{22}^{-1/2}R_{21}R_{11}^{-1}R_{12}R_{22}^{-1/2}$. Then, the first canonical pair is $\hat{U}_1 = -1.0015898Z_1^{(1)} + 0.002588365Z_2^{(1)}$ and $\hat{V}_1 = 0.6016105Z_1^{(2)} + 0.9768515Z_2^{(2)}$, where Z are standardized version of X.
We observe that $\hat{U}_1$ has large coefficient on $Z_1^{(1)}$ but close to 0 coefficient on $Z_2^{(1)}$. For $\hat{V}_1$, the coefficients on $Z_1^{(2)}$ and $Z_2^{(2)}$ are comparable, but the coefficient on $Z_2^{(2)}$ is relatively large. This means that the certainty of the punishment and severity of punishment in 1970 is highly correlated with the decrease of the 1973 non primary homicides. In particular, the certainty of punishment in 1970 is more closely related to the decrease of the 1973 non primary homicides.

10.13. (a) To find out the significant canonical pairs, we conduct sequential tests. We first test for $H_0 : R_{12} = 0$. The data were standardized for the canonical analysis. The test is equivalent to test for $H_0 : \Sigma_{12} = 0$. The test statistic is given by

$$\Lambda_n = -(n-1-(p+q+1)/2)\log(\prod_{i=1}^{4}(1-\rho_i^{*2})) = 309.9884.$$

where ρ_i^{*2} are eigenvalues of $R_{22}^{-1}R_{21}R_{11}^{-1}R_{12}$. Compare it with the chi-square distribution with pq degrees of freedom, the p-value is 0. Thus, we reject the null hypothesis.
Next, we test for $H_0 : \rho_1^* \neq 0, \rho_2^* = 0, \cdots, \rho_4^* = 0$. The test statistic is

$$\Lambda_n = -(n-1-(p+q+1)/2)\log(\prod_{i=2}^{4}(1-\rho_i^{*2})) = 78.63197.$$

Compare it with the chi-square distribution with $(p-1)(q-1)$ degrees of freedom, the p-value is 7.521317e-12. Thus, we reject the null hypothesis.

Next, we test for $H_0 : \rho_1^* \neq 0, \rho_2^* \neq 0, \rho_3^* = 0, \rho_4^* = 0$. The test statistic is

$$\Lambda_n = -(n-1-(p+q+1)/2)\log(\prod_{i=3}^{4}(1-\rho_i^{*2})) = 10.10658.$$

Compare it with the chi-square distribution with $(p-2)(q-2)$ degrees of freedom, the p-value is 0.120235. Thus, we do not have evidence to reject the null hypothesis. Therefore, the first two canonical correlations are significant at the nominal level 0.01.

(b) By computing the eigenvectors of of $R_{11}^{-1/2}R_{12}R_{22}^{-1}R_{21}R_{11}^{-1/2}$ and eigenvectors of $R_{22}^{-1/2}R_{21}R_{11}^{-1}R_{12}R_{22}^{-1/2}$, we obtain that

$$\begin{aligned}\hat{U}_1 &= -0.215z_1^{(1)} - 0.172z_2^{(1)} + 0.330z_3^{(1)} + 0.264z_4^{(1)} - 0.298z_5^{(1)};\\ \hat{V}_1 &= -0.535z_1^{(2)} - 0.288z_2^{(2)} + 0.457z_3^{(2)} + 0.025z_4^{(2)}.\end{aligned}$$

We notice that $\hat{U}_1$ has large coefficients on $z_3^{(1)}, z_4^{(1)}$ and $z_5^{(1)}$, all represent the quality of wheat. So, $\hat{U}_1$ might be considered as a measure of the quality of wheat. $\hat{V}_1$ has large coefficient on $z_1^{(2)}$. Hence, it may be used as a measure of quality of flour.

(c) The proportion of total sample variance in the first set $Z^{(1)}$ explained by $\hat{U}_1$ is

$$\text{Proportion of variance explained by } \hat{U}_1 \text{ in } Z^{(1)} = 0.6292283.$$

The proportion of total sample variance in the first set $Z^{(2)}$ explained by $\hat{V}_1$ is

$$\text{Proportion of variance explained by } \hat{V}_1 \text{ in } Z^{(2)} = 0.4496485.$$