Lab 4 - Multivariate Statistical Methods

Gustav Sternelöv

9 december 2015

A)

To test whether there exist a significant association at 5 % level between the primary and secondary set of variables a likelihood ratio test is conducted. The null hypothesis is that all canonical correlations is zero. This hypothesis is rejected at 5 % signifiance level if

$$-(n-1-0.5(p+q+1))ln\prod(1-\hat{\rho_i^*}) > \chi^2_{pq}(\alpha)$$

$$-(46-1-0.5(3+2+1))ln[(1-0.5173449^2)(1-0.1255082^2)] = 13.7494849$$

and

$$\chi^2_6(0.05) = 12.5915872$$

Since 13.7494849 > 12.5915872, the null hypothesis is rejected. With a significance level of 5 % it is concluded that there is an association between the primary and secondary variables.

B)

To test if also the second canonical correlation is significantly seperated from zero a test similar to the one in *A)* is conducted. The null hypothesis now is that $\rho_2^* \doteq 0$. Again I start by computing the test statistic

$$-(46-1-0.5(3+2+1))ln[(1-0.1255082^2)] = 0.6668632$$

and then also the critical value

$$\chi^2_2(0.05) = 5.9914645$$

The observed test statistic is lower than the critical value, hence the null hypothesis cannot be rejected. This implies that only the first pair of canonical variates is significant.

C)

The squared canonical correlation that was significant according to the test was ρ_1^{*2} which is equal to 0.517ˆ2 (0.268). The interpretation of this value is that 26.8% of the variance of canonical variate U_1 is explained by the secondary set of variables. Reversly this value also could be interpreted as the proportion of variance of canonical variate V_1 that is explained by the primary set of variables.

D)

The canonical variates and the correlations between the variables and their canonical variates:

$$\hat{U_1} = 0.4356829z_1^{(1)} - 0.7046696z_2^{(1)} + 1.0814622z_3^{(1)}$$

$$\hat{V_1} = -1.0202235z_1^{(2)} + 0.160936z_2^{(2)}$$

Table 1: Correlation between U1 and the primary set

Glucose	0.3397282
Insulin	-0.0501787
Insulres	0.7551136

Table 2: Correlation between V1 and the secondary set

Weight	-0.9875069
Fasting	-0.0464645

The canonical variates describes how influent the respective variables are. It can be concluded that the two insulin variables dominates $\hat{U}_1$ and that $\hat{V}_1$ mainly consists of the variable *Weight*.

A high value for the correlation between the variables and their canonical variates indicates that the variable is closely associated with the canonical variate. For $\hat{U}_1$ *Insulres* has a strong correlation and *Glucose* a moderately strong correlation. The last variable in the first set, *Insuline*, is an influent variable for the coefficients in $\hat{U}_1$ but is not closely associated to the variate. Regarding $\hat{V}_1$ *Weight*, the variable who dominated the variate, is very closely associated to the canonical variate and the second variable *Fasting* has a very weak correlation with the variate.

E)

For the first canonical variate, $\hat{U}_1$, only one of the variables is very closely associated to $\hat{U}_1$. This canonical variate is therefore not thought to be a good summary measure of the primary data set. Regarding the secondary data set one of the two variables is closely associated to the $\hat{V}_1$, but the second variable is barely associated at all with the canonical variate. Since this is the case neither the of the significant canonical variates are concluded to be good summary measures of their respective data set.

F)

A rather low amount of the variation is explained, 26.8 % It can be questioned how well the canonical variates works as summary measures of the respective data set. The analysis is below decent, the canonical correlation did not work very vell on this data sets.