Chapter 27
Canonical Regression for Overall Statistics of Multivariate Data (250 Patients)

General Purpose

To assess in datasets with multiple predictor and outcome variables, whether canonical analysis, unlike traditional multivariate analysis of variance (MANOVA), can provide overall statistics of combined effects.

Specific Scientific Question

The example of the Chaps. 22 and 23 is used once again. Twelve highly expressed genes are used to predict four measures of drug efficacy. We are more interested in the combined effect of the predictor variables on the outcome variables than we are in the separate effects of the different variables.

G1	G2	G3	G4	G16	G17	G18	G19	G24	G25	G26	G27	O1	O2	O3	O4
8	8	9	5	7	10	5	6	9	9	6	6	6	7	6	7
9	9	10	9	8	8	7	8	8	9	8	8	8	7	8	7
9	8	8	8	8	9	7	8	9	8	9	9	9	8	8	8
8	9	8	9	6	7	6	4	6	6	5	5	7	7	7	6
10	10	8	10	9	10	10	8	8	9	9	9	8	8	8	7
7	8	8	8	8	7	6	5	7	8	8	7	7	6	6	7
5	5	5	5	5	6	4	5	5	6	6	5	6	5	6	4
9	9	9	9	8	8	8	8	9	8	3	8	8	8	8	8
9	8	9	8	9	8	7	7	7	7	5	8	8	7	6	6

(continued)

This chapter was previously published in "Machine learning in medicine-cookbook 1" as Chap. 12, 2013.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_27

(continued)

10	10	10	10	10	10	10	10	10	8	8	10	10	10	9	10
2	2	8	5	7	8	8	8	9	3	9	8	7	7	7	6
7	8	8	7	8	6	6	7	8	8	8	7	8	7	8	8
8	9	9	8	10	8	8	7	8	8	9	9	7	7	8	8

Var G1-27 highly expressed genes estimated from their arrays' normalized ratios
Var O1-4 drug efficacy scores (the variables 20–23 from the initial data file)

The data from the first 13 patients are shown only (see extra.springer.com for the entire data file entitled "optscalingfactorplscanonical"). First, MANOVA (multivariate analysis of variance) was performed with the four drug efficacy scores as outcome variables and the twelve gene expression levels as covariates. We can now use SPSS 19.0. Start by opening the data file.

Canonical Regression

Command:

click Analyze….click General Linear Model….click Multivariate….Dependent Variables: enter the four drug efficacy scores….Covariates: enter the 12 genes…. OK.

	Effect value	F	Hypothesis df	Error df	p-value
Intercept	0.043	2.657	4.0	234.0	0.034
Gene 1	0.006	0.362	4.0	234.0	0.835
Gene 2	0.27	1.595	4.0	234.0	0.176
Gene 3	0.042	2.584	4.0	234.0	0.038
Gene 4	0.013	0.744	4.0	234.0	0.563
Gene 16	0.109	7.192	4.0	234.0	0.0001
Gene 17	0.080	5.118	4.0	234.0	0.001
Gene 18	0.23	1.393	4.0	234.0	0.237
Gene 19	0.092	5.938	4.0	234.0	0.0001
Gene 24	0.045	2.745	4.0	234.0	0.029
Gene 25	0.017	1.037	4.0	234.0	0.389
Gene 26	0.027	1.602	4.0	234.0	0.174
Gene 27	0.045	2.751	4.0	234.0	0.029

The MANOVA table is given (F = F-value, df = degrees of freedom). It shows that MANOVA can be considered as another regression model with intercepts and regression coefficients. We can conclude that the genes 3, 16, 17, 19, 24, and 27 are significant predictors of all four drug efficacy outcome scores. Unlike ANOVA, MANOVA does not give overall p-values, but rather separate p-values for separate

covariates. However, we are, particularly, interested in the combined effect of the set of predictors, otherwise called covariates, on the set of outcomes, rather than we are in modeling the separate variables. In order to asses the overall effect of the cluster of genes on the cluster of drug efficacy scores canonical regression is performed.

Command:

click File….click New….click Syntax….the Syntax Editor dialog box is displayed….enter the following text: "manova" and subsequently enter all of the outcome variables….enter the text "WITH"….then enter all of the gene-names…. then enter the following text: /discrim all alpha(1)/print = sig(eigen dim)….click Run.

Numbers variables (covariates v outcome variables)							
	Canon cor	Sq cor	Wilks L	F	Hypoth df	Error df	p
12 v 4	0.87252	0.7613	0.19968	9.7773	48.0	903.4	0.0001
7 v 4	0.87054	0.7578	0.21776	16.227	28.0	863.2	0.0001
7 v 3	0.87009	0.7571	0.22043	22.767	21.0	689.0	0.0001

The above table is given (cor = correlation coefficient, sq = squared, L = lambda, hypoth = hypothesis, df = degree of freedom, p = p-value, v = versus). The upper row, shows the result of the statistical analysis. The correlation coefficient between the 12 predictor and 4 outcome variables equals 0.87252. A squared correlation coefficient of 0.7613 means that 76 % of the variability in the outcome variables is explained by the 12 covariates. The cluster of predictors is a very significant predictor of the cluster of outcomes, and can be used for making predictions about individual patients with similar gene profiles. Repeated testing after the removal of separate variables gives an idea about relatively unimportant contributors as estimated by their coefficients, which are kind of canonical b-values (regression coefficients). The larger they are, the more important they are.

Canon Cor			
Raw Model	12 v 4	7 v 4	7 v 3
Outcome 1	−0.24620	−0.24603	0.25007
Outcome 2	−0.20355	−0.19683	0.20679
Outcome 3	−0.02113	−0.02532	
Outcome 4	−0.07993	−0.08448	0.09037
Gene 1	0.01177		
Gene 2	−0.01727		
Gene 3	−0.05964	−0.08344	0.08489
Gene 4	−0.02865		
Gene 16	−0.14094	−0.13883	0.13755
Gene 17	−0.12897	−0.14950	0.14845

(continued)

(continued)

Canon Cor			
Raw Model	12 v 4	7 v 4	7 v 3
Gene 18	–0.03276		
Gene 19	–0.10626	–0.11342	0.11296
Gene 24	–0.07148	–0.07024	0.07145
Gene 25	–0.00164		
Gene 26	–0.05443	–0.05326	0.05354
Gene 27	0.05589	0.04506	–0.04527
Standardized			
Outcome 1	–0.49754	–0.49720	0.50535
Outcome 2	–0.40093	–0.38771	0.40731
Outcome 3	–0.03970	–0.04758	
Outcome 4	–0.15649	–0.16539	0.17693
Gene 1	0.02003		
Gene 2	–0.03211		
Gene 3	–0.10663	–0.14919	0.15179
Gene 4	–0.04363		
Gene 16	–0.30371	–0.29918	0.29642
Gene 17	–0.23337	–0.27053	0.26862
Gene 18	–0.06872		
Gene 19	–0.23696	–0.25294	0.25189
Gene 24	–0.18627	–0.18302	0.18618
Gene 25	–0.00335		
Gene 26	–0.14503	–0.14191	0.14267
Gene 27	0.12711	0.10248	–0.10229

The above table left column gives an overview of raw and standardized (z transformed) canonical coefficients, otherwise called canonical weights (the multiple b-values of canonical regression), (Canon Cor = canonical correlation coefficient, v = versus, Model = analysis model after removal of one or more variables). The outcome 3, and the genes 2, 4, 18 and 25 contributed little to the overall result. When restricting the model by removing the variables with canonical coefficients smaller than 0.05 or larger than –0.05 (the middle and right columns of the table), the results were largely unchanged. And so were the results of the overall tests (the 2nd and 3rd rows). Seven versus three variables produced virtually the same correlation coefficient but with much more power (lambda increased from 0.1997 to 0.2204, the F value from 9.7773 to 22.767, in spite of a considerable fall in the degrees of freedom). It, therefore, does make sense to try and remove the weaker variables from the model ultimately to be used. The weakest contributing covariates of the MANOVA were virtually identical to the weakest canonical predictors, sug-

gesting that the two methods are closely related and one method confirms the results of the other.

Conclusion

Canonical analysis is wonderful, because it can handle many more variables than MANOVA, accounts for the relative importance of the separate variables and their interactions, provides overall statistics. Unlike other methods for combining the effects of multiple variables like factor analysis/partial least squares (chap. 8), canonical analysis is scientifically entirely rigorous.

Note

More background, theoretical and mathematical information of canonical regression is given in Machine learning in medicine part one, Chap. 18, Canonical regression, pp 225–240, Springer Heidelberg Germany, 2013, from the same authors.