Chapter 22
Factor Analysis and Partial Least Squares (PLS) for Complex-Data Reduction (250 Patients)

General Purpose

A few unmeasured factors, otherwise called latent factors, are identified to explain a much larger number of measured factors, e.g., highly expressed chromosome-clustered genes. Unlike factor analysis, partial least squares (PLS) identifies not only exposure (x-value), but also outcome (y-value) variables.

Specific Scientific Question

Twelve highly expressed genes are used to predict drug efficacy. Is factor analysis/PLS better than traditional analysis for regression data with multiple exposure and outcome variables.

This chapter was previously published in "Machine learning in medicine-cookbook 1" as Chap. 7, 2013.

T.J. Cleophas, A.H. Zwinderman, *Machine Learning in Medicine - a Complete Overview*, DOI 10.1007/978-3-319-15195-3_22

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
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 extras.springer.com for the entire data file entitled "optscalingfactorplscanonical").

Factor Analysis

First the reliability of the model was assessed by assessing the test-retest reliability of the original predictor variables using the correlation coefficients after deletion of one variable: all of the data files should produce at least by 80 % the same result as that of the non-deleted data file (alphas > 80 %). SPSS 19.0 is used. Start by opening the data file.

Command:

Analyze….Scale….Reliability Analysis….transfer original variables to Variables box….click Statistics….mark Scale if item deleted….mark Correlations …. Continue….OK.

Item-total statistics					
	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Squared multiple correlation	Cronbach's alpha if item deleted
Geneone	80,8680	276,195	,540	,485	,902
Genetwo	80,8680	263,882	,700	,695	,895
Genethree	80,7600	264,569	,720	,679	,895
Genefour	80,7960	282,002	,495	,404	,904

(continued)

Item-total statistics					
	Scale mean if item deleted	Scale variance if item deleted	Corrected item-total correlation	Squared multiple correlation	Cronbach's alpha if item deleted
Genesixteen	81,6200	258,004	,679	,611	,896
Geneseventeen	80,9800	266,196	,680	,585	,896
Geneeighteen	81,5560	263,260	,606	,487	,899
Genenineteen	82,2040	255,079	,696	,546	,895
Genetwentyfour	81,5280	243,126	,735	,632	,893
Genetwentyfive	81,2680	269,305	,538	,359	,902
Genetwentysix	81,8720	242,859	,719	,629	,894
Genetwentyseven	81,0720	264,501	,540	,419	,903

None of the original variables after deletion reduce the test-retest reliability. The data are reliable. We will now perform the principal components analysis with three components, otherwise called latent variables.

Command:

Analyze….Dimension Reduction….Factor….enter variables into Variables box…. click Extraction….Method: click Principle Components….mark Correlation Matrix, Unrotated factor solution….Fixed number of factors: enter 3….Maximal Iterations plot Convergence: enter 25….Continue….click Rotation….Method: click Varimax…. mark Rotated solution….mark Loading Plots….Maximal Iterations: enter 25…. Continue….click Scores…. mark Display factor score coefficient matrix ….OK.

Rotated component matrix[a]			
	Component		
	1	2	3
Geneone	,211	,810	,143
Genetwo	,548	,683	,072
Genethree	,624	,614	,064
Genefour	,033	,757	,367
Genesixteen	,857	,161	,090
Geneseventeen	,650	,216	,338
Geneeighteen	,526	,297	,318
Genenineteen	,750	,266	,170
Genetwentyfour	,657	,100	,539
Genetwentyfive	,219	,231	,696
Genetwentysix	,687	,077	,489
Genetwentyseven	,188	,159	,825

Extraction method: Principal component analysis
Rotation method: Varimax with Kaiser normalization
[a]Rotation converged in eight iterations

The best fit coefficients of the original variables constituting three new factors (unmeasured, otherwise called latent, factors) are given. The latent factor 1 has a very strong correlation with the genes 16–19, the latent factor 2 with the genes 1–4, and the latent factor 3 with the genes 24–27.

When returning to the data file, we now observe, that, for each patient, the software program has produced the individual values of these novel predictors.

In order to fit these novel predictors with the outcome variables, the drug efficacy scores (variables O1-4), multivariate analysis of variance (MANOVA) should be appropriate. However, the large number of columns in the design matrix caused integer overflow, and the command was not executed. Instead we will perform a univariate multiple linear regression with the add-up scores of the outcome variables (using the Transform and Compute Variable command) as novel outcome variable.

Command:

Transform….Compute Variable….transfer outcomeone to Numeric Expression box….click + ….outcometwo idem….click + ….outcomethree idem….click + …. outcomefour idem….Target Variable: enter "summaryoutcome"….click OK.

In the data file the summaryoutcome values are displayed as a novel variable.

Command:

Analyze….Regression….Dependent: enter summaryoutcome….Independent: enter Fac 1, Fac 2, and Fac 3….click OK.

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. error	Beta	t	Sig.
1	(Constant)	27,332	,231		118,379	,000
	REGR factor score 1 for analysis 1	5,289	,231	,775	22,863	,000
	REGR factor score 2 for analysis 1	1,749	,231	,256	7,562	,000
	REGR factor score 3 for analysis 1	1,529	,231	,224	6,611	,000

[a]Dependent variable: summaryoutcome

All of the three latent predictors were, obviously, very significant predictors of the summary outcome variable.

Partial Least Squares Analysis (PLS)

Because PLS is not available in the basic and regression modules of SPSS, the software program R Partial Least Squares, a free statistics and forecasting software available on the internet as a free online software calculator was used (www.wessa.

net/rwasp). The data file is imported directly from the SPSS file entitled "optscalingfactorplscanonical" (cut/past commands).

Command:

List the selected clusters of variables: latent variable 2 (here G16-19), latent variable 1 (here G24-27), latent variable 4 (here G1-4), and latent outcome variable 3 (here O 1-4).

A square boolean matrix is constructed with "0 or 1" values if fitted correlation coefficients to be included in the model were "no or yes" according to the underneath table.

	Latent variable	1	2	3	4
Latent variable	1	0	0	0	0
	2	0	0	0	0
	3	1	1	0	0
	4	0	0	1	0

Click "compute". After 15 s of computing the program produces the results. First, the data were validated using the GoF (goodness of fit) criteria. GoF= $\sqrt{}$ [mean of r-square values of comparisons in model * r-square overall model], where * is the sign of multiplication. A GoF value varies from 0 to 1 and values larger than 0.8 indicate that the data are adequately reliable for modeling.

GoF value	
Overall	0.9459
Outer model (including manifest variables)	0.9986
Inner model (including latent variables)	0.9466.

The data are, thus, adequately reliable. The calculated best fit r-values (correlation coefficients) are estimated from the model, and their standard errors would be available from second derivatives. However, the problem with the second derivatives is that they require very large data files in order to be accurate. Instead, distribution free standard errors are calculated using bootstrap resampling.

Latent variables	Original r-value	Bootstrap r-value	Standard error	t-value
1 versus 3	0.57654	0.57729	0.08466	6.8189
2 versus 3	0.67322	0.67490	0.04152	16.2548
4 versus 3	0.18322	0.18896	0.05373	3.5168

All of the three correlation coefficients (r-values) are very significant predictors of the latent outcome variable.

Traditional Linear Regression

When using the summary scores of the main components of the three latent variables instead of the above modeled latent variables (using the above Transform and Compute Variable commands), the effects remained statistically significant, however, at lower levels of significance.

Command:

Analyze….Regression….Linear….Dependent: enter summaryoutcome…. Independent: enter the three summary factors 1-3….click OK.

Coefficients[a]

Model		Unstandardized coefficients		Standardized coefficients		
		B	Std. Error	Beta	t	Sig.
1	(Constant)	1,177	1,407		,837	,404
	Summaryfac1	,136	,059	,113	2,316	,021
	Summaryfac2	,620	,054	,618	11,413	,000
	Summaryfac3	,150	,044	,170	3,389	,001

[a]Dependent variable: summaryoutcome

The partial least squares method produces smaller t-values than did factor analysis (t = 3.5–16.3 versus 6.6–22.9), but it is less biased, because it is a multivariate analysis adjusting relationships between the outcome variables. Both methods provided better t-values than did the above traditional regression analysis of summary variables (t = 2.3–11.4).

Conclusion

Factor analysis and PLS can handle many more variables than the standard methods, and account the relative importance of the separate variables, their interactions and differences in units. Partial least squares method is parsimonious to principal components analysis, because it can separately include outcome variables in the model.

Note

More background, theoretical and mathematical information of the three methods is given in Machine learning in medicine part one, Chaps. 14 and 16, Factor analysis pp 167–181, and Partial least squares, pp 197–212, Springer Heidelberg Germany 2013, from the same authors.