Tutorial 2

Using Correlated Component Regression with a Dichotomous Y and Many Correlated Predictors

Dataset for running Correlated Component Regression

This tutorial is based on data simulated according to the assumptions of Linear Discriminant Analysis (LDA) with 2 groups (ZPC1=1,0). The number of available predictors is P = 84 including 28 valid predictors (listed in Table 1 with their ***true coefficients***), some with high within-group correlation, and 56 irrelevant predictors 'INDPT1' – 'INDPT28' and 'extra1' – 'extra28' (with true coefficients equal to 0). We generated 100 simulated samples, each consisting of N=50 cases, with equal group sizes N1 = N2 = 25.

Table 1: True LDA Logit Model Coefficients

	Coefficients			Importance
Predictors	Unstandardized	Standardized*	Importance	Rank
SP1	-9.55	-5.72	5.72	1
GSK3B	4.56	2.48	2.48	2
RB1	-3.82	-2.30	2.30	3
IQGAP1	3.35	2.13	2.13	4
BRCA1	-2.13	-1.36	1.36	5
TNF	2.24	1.32	1.32	6
CDKN1A	2.33	1.29	1.29	7
MAP2K1	2.75	1.20	1.20	8
MYC	-1.81	-1.19	1.19	9
EP300	-1.78	-1.15	1.15	10
CD44	1.85	1.03	1.03	11
CD97	1.44	0.92	0.92	12
SIAH2	1.15	0.87	0.87	13
MAPK1	1.64	0.79	0.79	14
RP5	1.94	0.76	0.76	15
S100A6	1.22	0.74	0.74	16
ABL1	1.44	0.73	0.73	17
NFKB1	1.22	0.70	0.70	18
MTF1	-1.01	-0.62	0.62	19
CDK2	1.20	0.61	0.61	20
IL18	-0.79	-0.56	0.56	21
PTPRC	-0.98	-0.53	0.53	22
SMAD3	-0.57	-0.35	0.35	23
C1QA	-0.29	-0.30	0.30	24
TP53	0.45	0.26	0.26	25
CDKN2A	-0.31	-0.23	0.23	26
CCNE1	-0.21	-0.19	0.19	27
ST14	-0.18	-0.14	0.14	28

*Standardized coefficient = Unstandardized coefficient multiplied by standard deviation of predictor

An Excel sheet containing both the data and the results for use in this tutorial can be downloaded by clicking here.

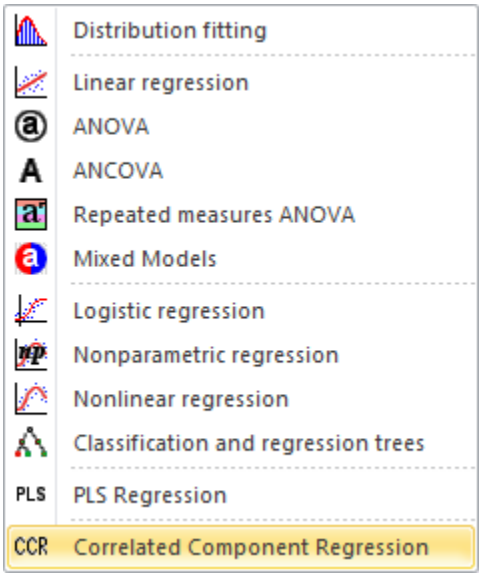
Goal of CCR for this example

CCR will apply the proper amount of regularization (K components) to reduce the confounding effects of high predictor correlation, and the CCR step-down algorithm will be used to exclude irrelevant and weak predictors, resulting in a model with a relatively small number of predictors P*. This results in a sparse model that provides better prediction (better classification) and coefficient estimates closer to the true values than traditional stepwise LDA, which imposes no regularization at all.

For illustration, this tutorial[1] focuses on simulation #1 (N=50). A summary of the results from all 100 simulations can be found in Magidson (2010).

Setting up a Correlated Component Regression

To activate the Correlated Component Regression dialog box, first start XLSTAT by clicking on the button in the Excel toolbar, then select the **XLSTAT / Modeling data / Correlated Component Regression** command in the Excel menu or click the corresponding button on the **Modeling data** toolbar.



[1] To reproduce the results shown in this tutorial exactly, you will need to fix the seed to '123456789'. To fix the seed in XLSTAT, go to Options, then click on the Advanced tab. Check the box to activate the option 'Fix the seed to:', and change the seed to 123456789.

Once you have clicked the button, the **Correlated Component Regression** dialog box is displayed with the Method=CCR.LM (linear regression model) selected by default. In the **Method** section, select the CCR.LDA (linear discriminant analysis model) option.

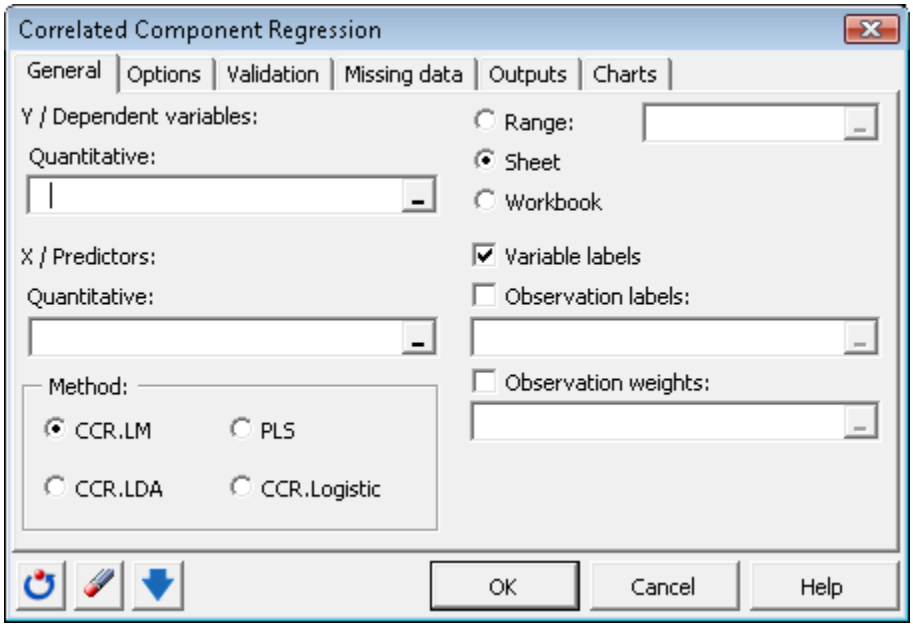


Figure 1. General Tab

In the **Y/ Dependent variables** field, use your mouse to select the (column A) variable 'ZPC1' (see the tutorial on Selecting data for more information on this topic).

The ZPC1 values are the "Ys" of the model as we want to predict the probability of being in group ZPC1=1 as a function of the 84 predictors. Specifically, Logit(Y) is determined as a linear function of the predictors, where Logit(Y)=exp(Prob[Y=1|X])/(1+exp(Prob[Y=1|X])

In the **X/ Predictors** field, select the 84 predictors.

The case ID of the subjects (ID) has also been selected as **Observation labels**.

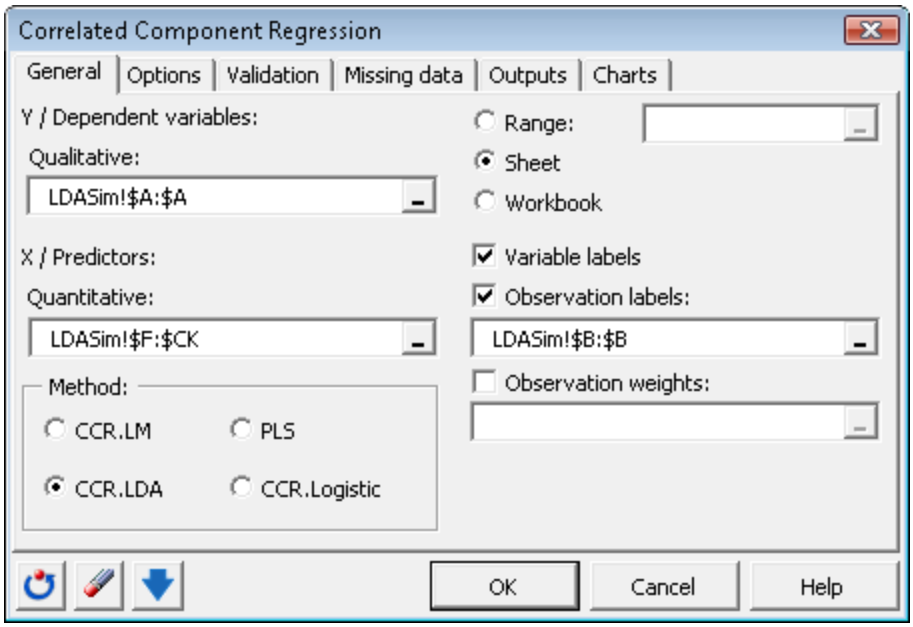


Figure 2. General Tab

In the **Options** tab of the dialog box, enter '5' as the number of components and activate the Step-down option. Make sure that the **settings** are **as shown below.**

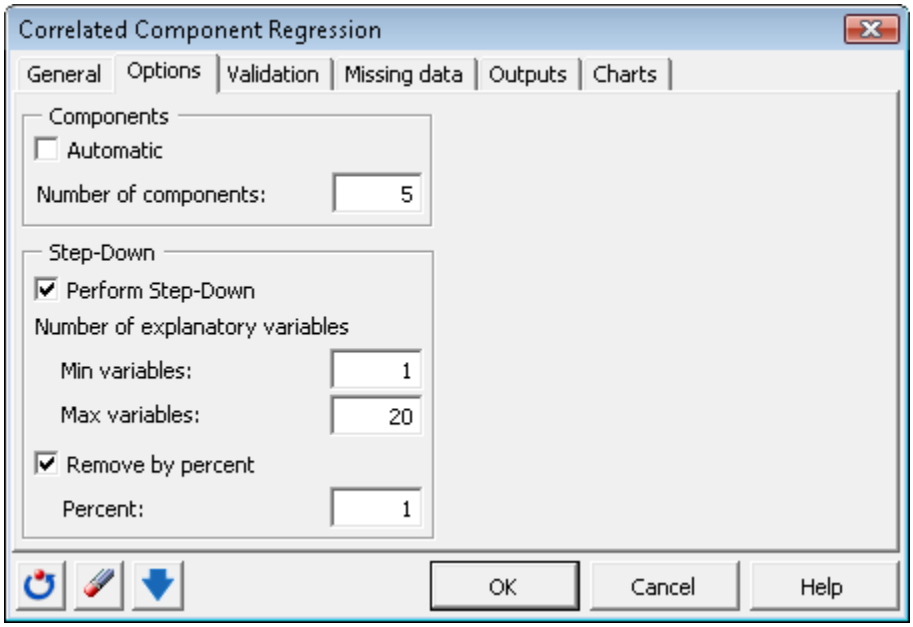


Figure 3. Options Tab

In the **Validation** tab of the dialog box, activate the **Validation** option and select 'N last rows' from the **Validation set** drop down menu. In the **Number of observations** field, type '4950'. We have now specified the 'Training set' as the first 50 rows of the data file (simulation #1) and the last 4,950

rows of the data file will be used as the validation set (simulations #2-100). Activate the **Cross-validation** option and change the default number of folds from '10' to '5'. Activate the 'Stratify' option.

Make sure that the **settings** are **as shown below.**

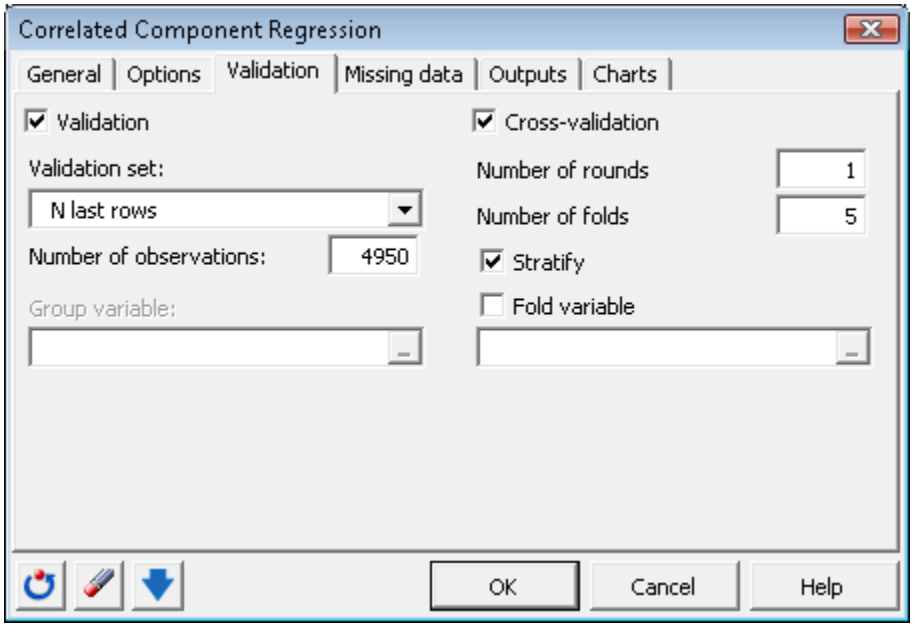


Figure 4. Validation Tab

Estimate the 5-component model

Click OK to estimate the model.

Interpreting the Results of a CCR Model with 10 Predictors

The **Cross-Validation Step-down Plot** shows that for K=5 components the Cross-validation Accuracy (CV-ACC) is best with P=10 predictors.

Cross-validation results:

Cross-validation step-down table:

Predictors	AUC	ACC
20	0.762	0.720
19	0.768	0.720
18	0.773	0.760
17	0.787	0.760
16	0.779	0.740
15	0.778	0.760
14	0.776	0.720
13	0.773	0.740
12	0.789	0.760
11	0.806	0.760
10	0.854	0.820
9	0.872	0.780
8	0.874	0.800
7	0.882	0.800
6	0.890	0.760
5	0.874	0.780
4	0.830	0.740
3	0.790	0.700
2	0.789	0.640
1	0.693	0.660

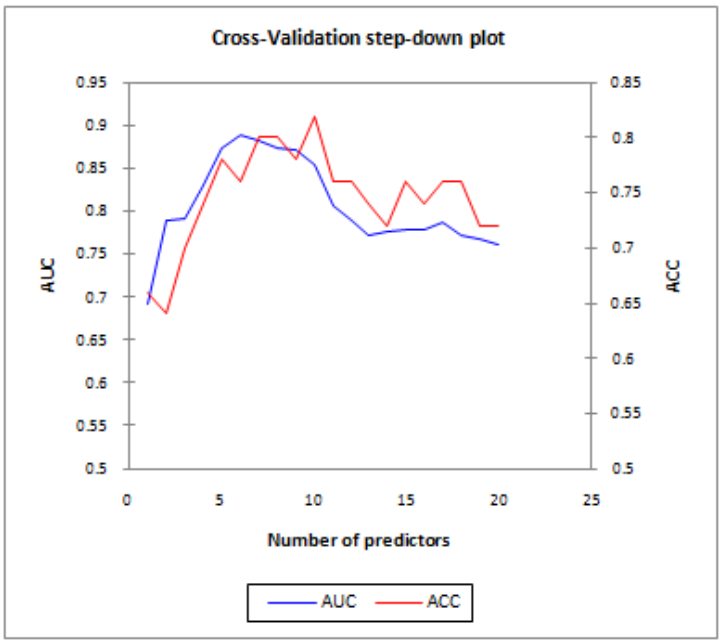


Figure 5. Plot of Cross-validated Area Under ROC Curve (CV-AUC) and Cross-validated Accuracy (CV-ACC) for K=5, N=50

Goodness of fit statistics for the 5-component model with 10 predictors (N=50)

Goodness of fit statistics:

	Value	Validation	Cross-validation
Number of observations	50	4950	
Sum of weights	50	4950	
ACC	0.920	0.836	0.820
AUC	0.981	0.914	0.854

Unstandardized coefficients for the 5-component model with 10 predictors are given below.
Unstandardized coefficients:

Variable	Coefficient
Intercept	-19.614
BRCA1	-5.414
GSK3B	8.311
IQGAP1	5.469
MAPK1	1.871
PTPRC	-3.601
RP5	4.956
SP1	-5.566
IL18	-3.163
EP300	-2.928
CDKN1A	4.129

These results obtained from CCR.LDA outperform step-wise linear discriminant analysis in the following respects:

- More valid predictors included in the model: 10 for CCR.LDA vs. 4 for step-wise LDA.
- Fewer irrelevant predictors included in the model: 0 for CCR.LDA vs. 2 for step-wise LDA.
- Higher accuracy as determined from the validation sample: 83.6% for CCR.LDA vs. 77.8% for step-wise LDA.

The results for step-wise LDA are provided below.

Classification functions and beta:

	0	1	Beta
Intercept	-873.217	-861.751	11.465
BRCA1	82.208	75.488	-6.719
IQGAP1	-89.340	-78.888	10.452
SP1	31.485	25.388	-6.097
CDKN1A	41.848	46.860	5.012
INDPT9	6.106	2.411	-3.695
INDPT23	7.732	5.681	-2.051

Confusion matrix for the validation sample:

from \ to	0	1	Total	% correct
0	2047	428	2475	82.71%
1	673	1802	2475	72.81%
Total	2720	2230	4950	77.76%

Overall, the results based on all simulated samples show that CCR.LDA outperforms step-wise LDA as well as penalized regression on these data (Magidson, 2010: Correlated Component Regression: A Prediction/Classification Methodology for Possibly Many Features. 2010 Proceedings of the American Statistical Association.)

For CCR **Tutorial 1** click here: **Getting Started with Correlated Component Regression (CCR) in XLSTAT**
For CCR **Tutorial 3** click here: **Obtaining Predictions from a 2-class Regression**

For other XLSTAT Tutorials click here