F.A.Q.

SIMCA –P and Multivariate Analysis

Frequently Asked Questions

Contents:

1. General MVA p2
2. Data Input p3
3. Validation p4
4. Models p6
5. Alternatives to PLS p7
6. SIMCA-P Tips p9
1. General

1.1 I get lost with all the plots in PLS Where do I start?
TIP: Try to progress down the Analysis menu.

1. Firstly model overview plots are useful both by component and by variable

2. Examine the inner relation plots t1/u1 t2/u2 to examine the underlying relationship between X and Y.

3. Next look at t1/t2 to understand the structure of X and u1/u2 the structure of Y and look for outliers. The DmodX plot shows mild outliers.

4. The w*c or loadings plot gives a graphical summary of the correlation between X and Y. (Remember the See-Saw method here) but for more detail refer to the coefficients plot.

5. The VIP plot shows which are the most important variables over the model as a whole.

6. Finally the Validate Plot should be used to check you have a unique model that could not have arisen by chance.

1.2 Do you have a flowchart to help me?
Yes. Please see page 494 of the text book Multi- and Megavariate Data Analysis: Principles and Applications.

1.3 How does MVA separate out useful information from the noise?
The assumption in Multivariate Analysis is that only part of the data contains useful information. This useful information is described in terms of underlying trends or Latent Variables. These are found by finding the correlation patterns in the data. Latent variables are extracted until the amount of useful information diminishes. Beyond a certain point extracting more components will only be modelling noise. The point at which you stop is determined by cross validation.

1.4 What is a loading?
A loading describes the correlation that the Principal or PLS component has with the original variable. This is done by measuring the angle the component makes with the original variable axis and taking its cosine. A high value (max=1) means that the component is aligned with the original variable, a close to zero value shows that it has no influence. A low value (min -1) indicates an opposite influence.

1.5 In PCA what is the score plot showing me?
The scores plot shows correlations between observations. For example are my observations related to each other, are there any groups or trends?

1.6 In PCA what is the loadings plot showing me?
The loadings plot shows correlations between variables. Comparing the loadings plot to the scores plot enables you to understand how the variables relate to the observations.

1.7 What scale are the scores and loadings on?
The scores plot scale results from the projection of the data onto the principal component, therefore the scaling of the axis depend upon the pre-treatment of the data. The loadings plot has a scale of +1 to -1.
2. Data input

2.1 How can I input average values with error bars?
With Multivariate Analysis you should use the original variables if at all possible. In this way you get all the information. Averaging data may lead to a loss of information.

2.2 What scaling method should I use
If your variables are all on the same scale such as spectroscopic data then centring only is recommended. If the variables are on different scales (i.e. you are comparing chalk with cheese) the UV scaling is recommended. If medium and small features in the data are important (such as NMR data) Pareto scaling is often useful.

2.3 When should I use a transformation?
PCA and PLS work best with normally distributed data. Transformations should be used to make the data normally distributed in cases of skewed distributions. For example in drug screening it is common to have many low activity compounds but a small number of high activity ones. You can check the effect of transformations using the quick info function in the spreadsheet. Also in the transformation menu when both the Skew and Min/Max are highlighted in red a transformation is recommended. Often the first step is to construct a PCA model on untransformed data. Significant bunching of data in one area with a few more disperse points elsewhere may be improved with a Log Transform.

2.4 Should I scale my Y data?
If your Y data are on the same scale then centring the data is recommended. If they are on different scales then use UV scaling.

2.5 When importing a secondary dataset does SIMCA match variable names or variable order?
The order does not matter. SIMCA-P matches by name.

2.6 Should I Derivatise my spectra?
In general derivatisation is not necessary with Multivariate analysis as it tends to add noise and leads to no advantage apart from a baseline correction, which may be done in other ways.
3. Validation

3.1 What is Q2 again?
Q2 is an estimate of the predictive ability of the model. It is calculated by cross-validation. The data are divided into 7 parts (by default) and each 1/7th in turn is removed. A model is built on the 6/7th data left in and the left out data are predicted from the new model. This is repeated with each 1/7th of the data until all the data have been predicted. The predicted data are then compared with the original data and the sum of squared errors calculated for the whole dataset. This is then called the Predicted Residual Sum of Squares. The better the predictability of the model the lower this value will be. For convenience we then convert PRESS into Q2 to resemble the scale of the R2. PRESS is divided by the initial sum of squares and subtracted from 1. Good predictions will have low PRESS and so high Q2.

3.2 Why is my Q2 negative? How can Q2 be negative
Q2 is not really a square (see 3.1). If you have negative Q2 your model is not at all predictive.

3.3 What R2 Q2 should I expect
The R2 Q2 you can expect is highly application dependent. In general R2 should not exceed Q2 by more than 2 units. As an approximate guide see the list below:

<table>
<thead>
<tr>
<th>Application</th>
<th>R2</th>
<th>Q2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Spectroscopic Calibration</td>
<td>0.99</td>
<td>0.98</td>
</tr>
<tr>
<td>Good QSAR model</td>
<td>0.78</td>
<td>0.65</td>
</tr>
<tr>
<td>Biological PCA model</td>
<td>0.5</td>
<td>0.4</td>
</tr>
<tr>
<td>PCA Stable Process</td>
<td>0.3</td>
<td>0.2</td>
</tr>
<tr>
<td>PCA Market research</td>
<td>0.3</td>
<td>0.2</td>
</tr>
</tbody>
</table>

3.4 Does an R2 < 0.5 indicate no correlation and therefore no model?
Because Multivariate Analysis separates out useful information from noise a low R2 indicates a large amount of noise or irrelevant information in the data. The model can still be usable. In the case of PLS you need to asses the value of the model by using a validation set. The correlation coefficient of the Obs. vs. Predicted for your validation set give you an "external Q2" which you can use to asses the value of your model.

3.2 How can I get a leave one out estimate
Leave one out is a valid statistical test with a low number of observations, however as n increases leaving one observation out is not a sufficiently vigorous test of the model. SIMCA-P by default leaves out 1/7th of the data, which is a more stringent test. In cases of low n (< say 20) you may change the number of groups used for cross validation to equal the number of observations. In this way the cross validation will be equivalent to “leave one out”.

3.3 How can I be sure my model will work?
The Q2 is a reasonable first guess as to how your model will perform on new data but the real test is to use an external validation set.

3.4 Is removing outliers just cheating?
Outliers are always interesting and worth studying. Reasons for the observation being an outlier can be found using the contribution plot. You then need to apply your scientific expertise to make a judgement on whether it is valid to remove that point. It is often found that an outlier is a transcription error so a check back to the original data is always the first step.

3.6 OSC to me seems like "fiddling the data"
Indeed the OSC procedure on the training set is removing information in X that is uncorrelated to Y so you will get a better model on the training set. With OSC it is vital to use an external validation set to see if the OSC procedure improves the model compared with the original non OSC PLS model. If you get a model which predicts new data then the technique is valid to use.
3.7 How can I have an outlier in DMOD X but an inlier in Hotellings T2?
Strong explainable outliers will be outside the Hotellings T2 ellipse. Weak outliers will be seen in DmodX. It is possible for a point to lie a long way from the model plane but be projected into the centre of the model.

3.8 What is a residual?
A residual is the difference between the model and the original data. The way SIMCA-P works is to subtract the explained variation from the original data with every component leaving a residual matrix E. This represents all the errors or unexplained variation. If we know what is unexplained we can conversely calculate what is explained. This is the basis for the R2 parameter.

3.9 What is the difference between Scores and DModX contribution plots?
The contribution plot of scores shows variation which is explained, i.e. which variables contribute to the model. The contribution plot of DMod X shows variation which is not explained, i.e which variables contribute to the high DmodX.
4. Models

4.1 Can I take out the calibration model from SIMCA-P and use it elsewhere?
Yes. Go to the analysis menu, click coefficients list. Right click on the list and select properties on the Coefficients tab select Unscaled and apply. Click the top left hand cell to select the data and copy and paste (for example) into excel.

By making predictions outside SIMCA-P you will lose the ability to see diagnostic information like DModX and Probability of membership, which flags unusual, possibly outlying, candidates for predictions.

For Online predictions consider using Umetrics SIMCA-P 4000, SIMCA-Batch On Line or SIMCA-QP. In this way models from SIMCA-P may be used online while retaining valuable validation information.

4.2 How does SIMCA cope with missing data?
Put simply the NIPALS algorithm interpolates the missing point using a least squares fit but give the missing data no influence on the model. Successive iterations refine the missing value by simply multiplying the score and the loading for that point. Many different methods exist for missing data, such as estimation but they generally converge to the same solution. Missing data is acceptable if they are randomly distributed. Systematic blocks of missing data are problematic.

4.3 What is the difference between VIP plot and the Coefficent plot?
The coefficients plot summarises the relationship between the Y variables and the X variables. These are directly analogous to, but not identical to, coefficients obtained from multiple regression. You will have one coefficient plot for each Y variable.

The VIP plot carries similar information to the coefficients plot and in practical terms the two plots often look very similar. The major difference is that the VIP plot describes which X variables characterise the X block well AND which variables correlate with Y. PLS is a dual technique which tries to finds directions in X which both characterise X well and are related to Y. In extreme cases, it is possible for an X variable to have a high VIP but not be related to Y at all.

The VIP values summarise the overall contribution of each X-variable to the PLS model, summed over all components and weighted according to the Y variation accounted for by each component, therefore you only ever get one VIP plot per model.

4.4 Can I get a better model by removing unimportant variables?
In our experience removing variables may lead to an apparently better model but when used predictively you may be misleading yourself as to how good the model actually is. Leaving in all variables usually gives a more 'honest' model.

Use of the jack-knifed confidence intervals in the coefficients plot combined with the VIP plot may allow you to identify important and significant variables.

It is generally discouraged to remove variables though there may be good reasons for doing so such as noisy uninformative regions of spectra, interfering spectral peaks or perhaps to produce a simple interpretable model. If you do produce a pruned model it is good practice to run it in parallel with the original model. Apparently unimportant variables do contribute to DmodX, so leaving variables in may be useful in monitoring situations.

4.5 My PLS model is rubbish, why?
Perhaps your descriptors are not related to your Y variables. In which case go and find better ones. Are you being too ambitious? Local models have the best chance for successful modelling. Are there groups in your data? Try modelling each Y-variable separately. There is not enough variation in your historical data to build a model. Try to use DoE and acquire more data with better variation.

You may be trying to model non-linear behaviour, in which case try GIFI PLS or expanded terms.
4.6 Can I get error bars on my predictions?
No but you can get the values for RMSEE (Root Mean Squared Error of Estimation) or RMSEP (Root Mean Squared Error of Prediction) from the Obs vs Predicted plot. These represent one standard deviation in the metric of the Y variable.

5. Alternatives to SIMCA-P and PLS

5.1 I wish SIMCA could do MLR or PCR. Why doesn’t it?
MLR or Multiple Linear Regression does not work well with correlated data, assumes all the data have no noise, can only model one Y variable at a time and requires more observations than variables. If your data has no correlation, is precise with no noise, you only want to model one Y variable and you have more observations than variables then MLR is fine.

PCR Principal Components Regression involves PCA of the X data followed by regression against a Y variable. The disadvantage of this method is that many directions in X may have no relationship with Y and so which PCA component gives maximum correlation to Y is unpredictable.

In fact SIMCA-P can be forced into producing MLR and PCR as follows:

**Multiple Linear Regression (MLR) in SIMCA-P**
Do PLS but force all components so A=K
If you then produce a coefficient list but right click and unscaled and uncentred, you get the same results as Excel’s MLR (plus you will have fitted all the noise - as the assumption with MLR is that X is precise)

Notes:
(1) MLR must have more Observations than Variables
(2) In PLS Force all components will result in A=K or A=N which ever is smaller. i.e. you cannot have more components than min (obs,vars)

**Principal Components Regression (PCR) in SIMCA-P**
Do PCA on the X block, set as Hierarchical Base Model
Set up PLS as Hierarchical top model
Regress the T scores vs Y
You will only get 1 component as the scores are orthogonal. The unscaled and uncentred coefficients will then be equivalent to PCR

The disadvantage of PCR is that the components that are extracted in the PCA may not have anything to do with Y

The interesting thing is that doing the upper level model you will only get one component, as PLS on orthogonal datasets = MLR. If you think about it this makes sense because in MLR you are expecting X to be precise, independent and have no latent structure.

5.2 Why can’t I use MLR for my data?
MLR or Multiple Linear Regression does not work well with correlated data, assumes all the data have no noise, can only model one Y variable at a time and requires more observations than variables. If your data has no correlation, is precise with no noise, you only want to model one Y variable and you have more observations than variables then MLR is fine.
5.3 Will I get a better fit of my data using a Neural Network?
There is always a trade-off between fit and predictive ability. It is easy to fit data but not so easy to predict. Overfitting or overtraining a model results in modelling noise which results in an over-optimistic view of the predictive ability of a model. PLS models generated with SIMCA-P will generally give an honest appraisal of its predictive ability.

Neural nets may be useful for very non-linear or discontinuous systems, however there is a danger they can fit anything without good prediction. Using PCA scores as the input to a neural net has been found useful as the PCA scores are orthogonal and represent the usable variation (minus the noise) in the data. Interpretability of Neural nets may be difficult.

5.4 PLS that’s for linear data only right?
Wrong. Although based on a linear inner relationship PLS may adapt to very mild non linearity in successive PLS components. Also methods of handling moderate non-linearity are available such as GIFI-PLS and X matrix expansion. Both available in SIMCA-P.

### GIFI-PLS Regression in SIMCA-P

GIFI-PLS replaces continuous variables with qualitative variables representing the range of that variable. Using SIMCA-P, under the Data Set menu use the Generate Variables function to create bins for the continuous variables.

For example In the Expression for new variables field type the command Bins(v[3:12],5). This will create five bins for each of the X-variables from 3 to 12

TIP: when selecting the number of bins use the rule of thumb Bins= INT(Range/10).
You will need at least N = 30 and 3 bins, otherwise you cannot reliably search for non-linear relationships in your data.

In this way part of the range of a variable may found to be more important by examining the coefficients plot. Also non-linearity may be detected when looking across the bins for a particular variable.

Check for overfit by using an external test set and compute RMSEP and Q2 ext.

### INLR-PLS Regression in SIMCA-P (Expansion of X variables)

Use the Expansion Tab to add Square, Cubic or Cross terms to the X matrix. Do PLS on the expanded matrix.

Check for overfit by using an external test set and compute RMSEP and Q2 ext.

**WARNING:** Cross terms should only be used when Design of Experiments has been used. This is because without spanning the whole experimental space, inappropriate combinations of variables may be chosen which represent a significant extrapolation. The aim of reliable calibration models is to **interpolate** not **extrapolate.**
Non Linear PLS Regression in SIMCA-P

Using SIMCA-P’s hierarchical modelling feature it is possible to look at non-linear relationships between the PCA or PLS components and the Y variable.

Do a PCA or PLS and set it as a base level model. Set a new PLS model as the upper level model but use the Transform tab to expand the scores. Regress against Y.

Check for overfit by using an external test set and compute RMSEP and Q2 ext.

For PLS and PCA only work on monotonic data, that is something is increasing or decreasing over the experimental range, hence radically non-linear data cannot be handled. In such cases the only way to proceed is to use local models over smaller intervals.

5.5 Why should I choose Umetrics Software?

Umetrics software is aimed at end-user scientists and engineers rather than the dedicated statistician, though many do use SIMCA-P for its ease of use. Model validity is emphasised within SIMCA-P and presented in easy to visualise plots. Data import and manipulation is very easily handled. Umetrics as an organisation backs up its software products with comprehensive training and consultancy services and has an unrivalled background in the practical application of Chemometrics to real world data challenges.

6. SIMCA-P Tips

6.1 When you save a SIMCA-P project, is there any way of saving all the plots corresponding to that project, so you don’t have to re-plot them every time you open the project?

The best way of retrieving the plots corresponding to a particular project is to use Favorites. Go to View/Favorites. At the bottom of the Favorites there is a directory called “Project Favorites”, open that directory. Now open the plots you want to see, right-click in each one and select "Add to Favorites". When you have finished adding all the plots click on "Save". The next time you open the project, right-click on "Project Favorites" and choose "Open All Items in Folder". This will simultaneously open all the saved plots.

SIMCA-P For Multivariate Analysis
SIMCA-P+ For Batch Analysis
SIMCA-4000 For Online Process Modelling
SIMCA-BOL For Online Batch Analysis and Modelling
SIMCA-QP For integrating calibration models into your own software
SIMCA-QM For integrating Umetrics model building algorithms into software

For more information, training, consulting and software please contact;

In all countries: In USA and Canada: In UK:
Umetrics AB Umetrics, Inc. Umetrics UK Ltd.
Box 7960 17 Kiel Ave. Woodside House, Winkfield
SE-90719 Umea Kinnelon, NJ 07405 Windsor, SL4 2DX
Sweden USA United Kingdom
Tel +46 (0)90 184800 Tel: +1 973 492 8355 Tel: +44 (0)1344 885615
Fax: +46 (0)90 184899 Fax: +1 973 492 8359 Fax: +44 (0) 1344 885410
Email: info@umetrics.com E-mail: info.us@umetrics.com E-mail info.uk@umetrics.com

www.umetrics.com