Increase Efficiency and Model Applicability Domain When Testing Options That Are at First Glance Multilevel Categorical Factors - Discovery Summit Europe 2017

Following an overview of the major tasks performed during the live presentation with JMP. The software used for this presentation is JMP PRO 13.

The used data file is available for download on the conference web page called “MV Example.jmp” and “MV Example Solution.jmp”.

Introduction

When testing options of e.g. different raw materials or formulation ingredients, common practice is to vary them as multilevel categorical variable e.g. A, B, C…. in an experiment. Hence, for identifying the best option all of them have to be tested. A consequence of this is:

- time consuming physical testing is required
- and the resulting model is only applicable to predict the tested options but cannot predict options with changed physical/chemical properties

A much more efficient approach is to design the experiment based on the physical/chemical properties of each option. This,

- significantly decreases the number of required experimental conditions and
- results in a sustainable empirical model that can predict options not tested before.

1. Data

ID	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X14	X15	X16	X17	X18	X19
1	-0.281	-0.739	0.659	0.113	1	2.318	0.327	0.279	1.212	-0.368	-0.097	0.601	0.114	-0.024	0.891	-2.209	0.449	-0.543	0.116
2	1.023	-0.229	1.087	0.807	-1	-0.096	0.751	-0.796	-0.257	0.340	-0.700	1.200	0.054	-1.522	-0.270	0.568	-0.752	0.672	1.489
3	-1.044	0.364	0.395	-1.234	1	-0.967	-0.657	1.822	0.022	-2.830	-0.094	0.149	1.553	1.574	0.192	0.985	0.722	-1.627	-1.053
4	-0.236	0.592	-0.227	-0.467	-1	2.200	0.092	0.226	-0.384	0.340	-0.941	-0.258	-1.212	-0.235	1.533	-2.004	0.061	-0.949	-0.882
5	1.591	-0.543	-0.513	0.389	-1	0.937	0.452	-1.572	-1.488	0.340	0.332	-0.139	-1.305	-1.756	-0.430	-0.538	-1.522	1.669	1.833
6	-0.736	0.500	-1.402	0.536	1	-0.179	-0.844	0.720	0.195	0.340	-0.305	-1.403	-1.289	-0.332	-0.142	-0.149	0.839	0.109	-0.641
7	1.539	0.624	-0.056	1.065	1	-0.729	2.009	-1.421	-0.906	1.723	-0.099	0.260	1.258	-0.267	-1.005	-0.056	-1.051	1.251	1.933
8	2.556	-0.542	1.287	2.048	-1	-0.310	1.572	-2.605	-1.634	0.340	0.820	-1.845	1.532	-0.790	-1.540	-1.155	-1.590	0.462	2.004

The data set is consisting of:

- 45 raw material options, where each is a potential candidate for being the active ingredient in a product formulation
- Material options can be described by 19 physical and chemical descriptors (all variables are coded)
- Two responses to be optimized (Y1 max, Y2 min)

Test Procedure:

- Test (each) raw material option in base formulation

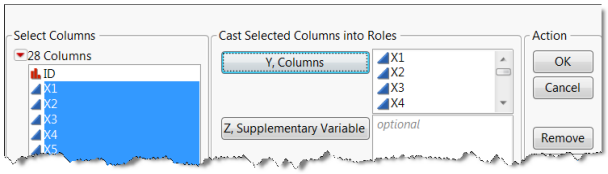
Objective:

- Predict overall optimum raw material option (even optimum may not be physically available at the moment)
- Identify available material options closest to overall optimum
- Establish sustainable empirical model being able to predict new raw material options

2. Identify most prominent dimensions of variation via Principle Component Analysis

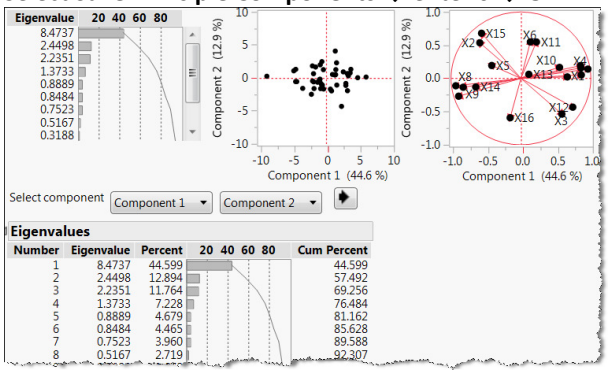
Open data file “MV Example.jmp”.

Select **Analyze → Multivariate Methods → Principle Components → Columns** X1 – X19 (variables) **→ Y, Columns → OK**



Inspect principle components, check how good they are covering the variable space. In this case it seems to be reasonable to take 5 – 6 Principle Components to sufficiently capture the variation in the data.

Select **Save Principle Components →** enter 6 **→ OK**

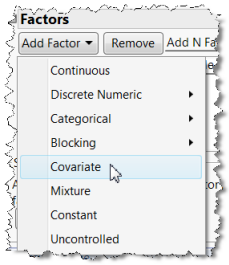


3. Design the Experiment

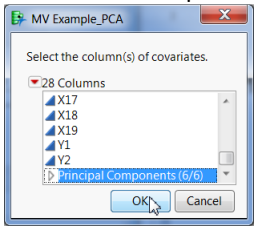
Design the experiment based on the derived inputs i.e. the Principal Components (dimension reduction).

Select **DOE → Custom Design**

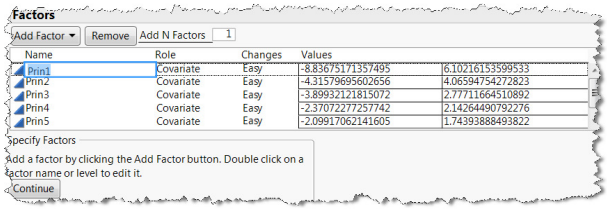
Select **Add Factor → Covariate**



Select all 6 Principal Components **→ OK**

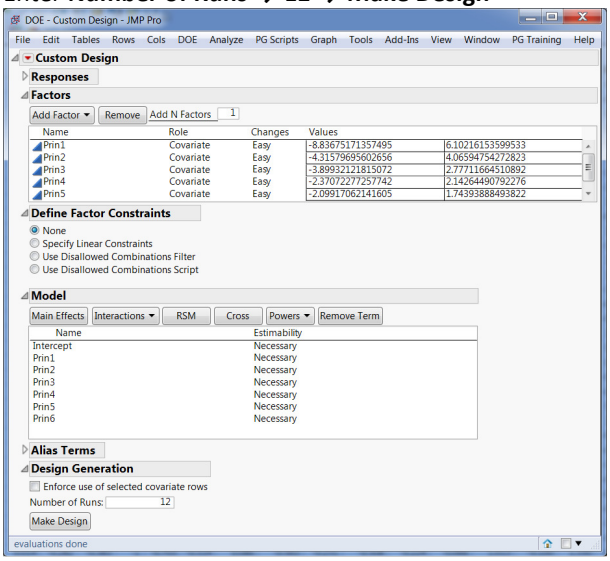


Select **Continue**

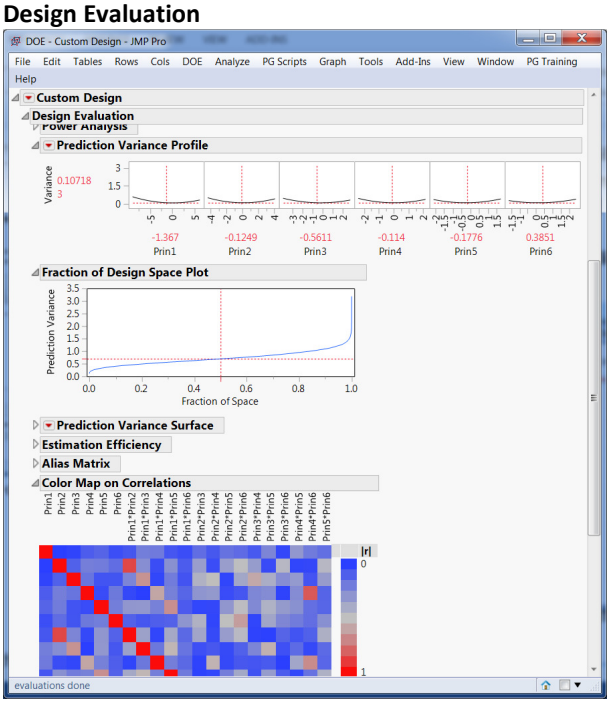


Silvio Miccio

Enter **Number of Runs → 12 → Make Design**



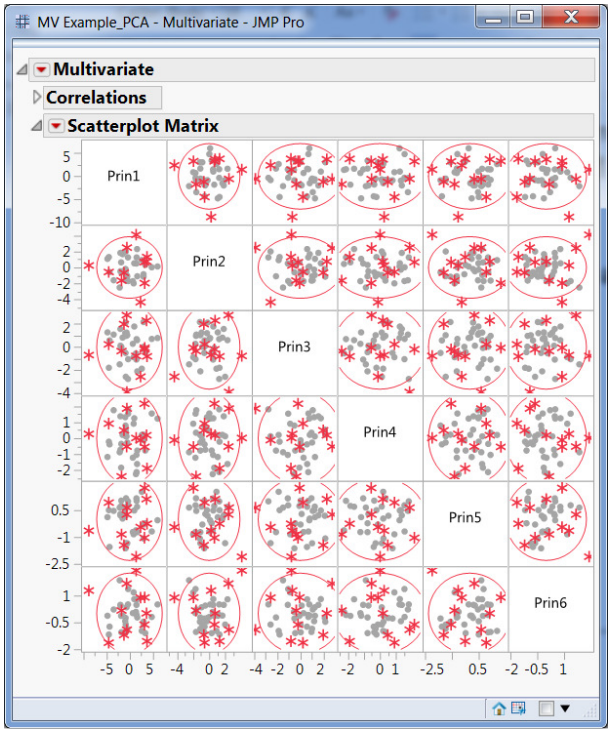
Design Evaluation



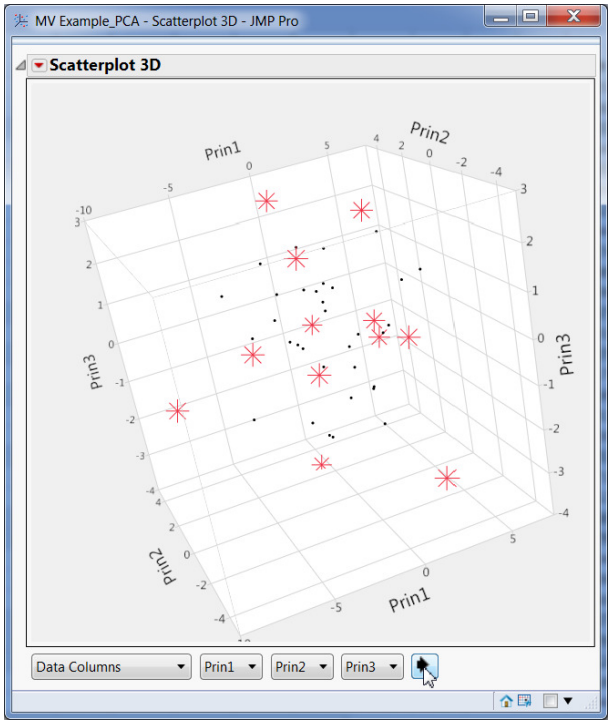
When design diagnostics is ok save the script to data table and close DOE platform. Runs proposed for the experiment are selected in the data sheet. Color and mark the selected rows.

<u>Visualize Experimental Design:</u>
Select **Analyze → Multivariate →** select all 6 principle components **→ Y, Columns → OK**

Colored and marked points are showing how good the selected options are covering the model space (scatterplot matrix on the left)



Alternative also a 3D Scatter Plot can be created

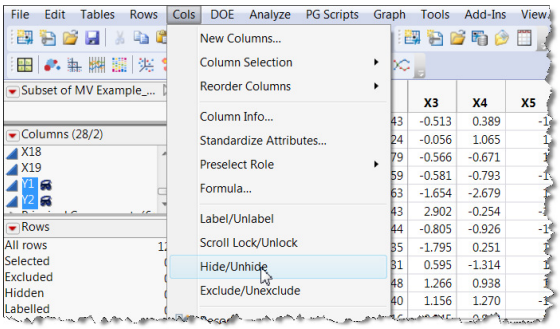


Scroll through the variance components by pushing the arrow to see how the selected options are covering the model space. When being satisfied with the design experiment is run, if not number of runs or number of principle components has to be changed and previous steps area repeated.

4. Run the experiment

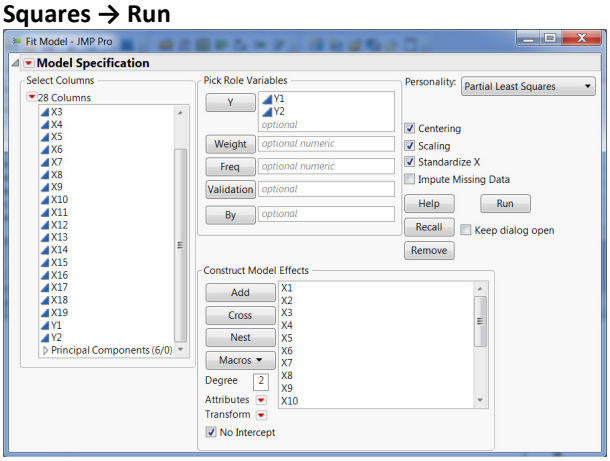
Make a subset table of colored and marked rows, resulting in a data table with 12 rows, the experimental conditions to be tested.

Unhide columns with responses called Y1 and Y2 (in reality the experiment would now been run according experimental design). In fact selected material options have been put into a base formulation and have been tested.

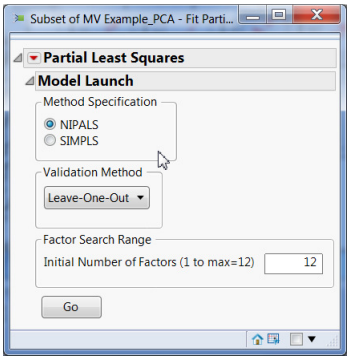


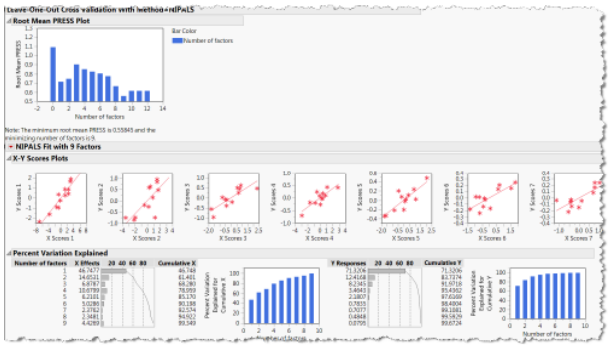
5. Run a PLS Model (details of PLS not explained)

Run a PLS model with the original variables (not with Principle Components). Select **Analyze → Fit Model → Columns** X1 – X19 (predictors) **→ Add →** Select **Columns** Y1 and Y2 (responses) **→ Y → Personality → Partial Least Squares → Run**



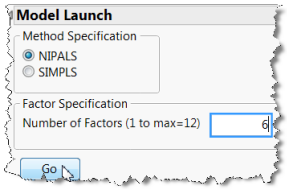
Select **NIPALS → Leave-One-Out → Go**



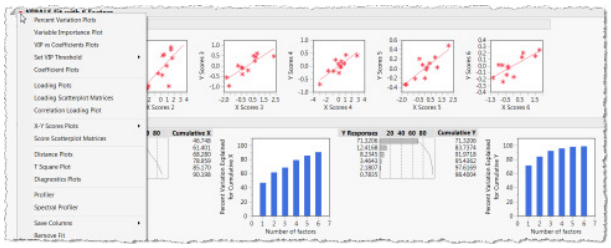


JMP suggests 9 latent factors (which is a valid estimate), but 6 latent factors also seem to be ok (matches the number of dimensions identified in the PCA).

Re-run the model with 6 latent factors.



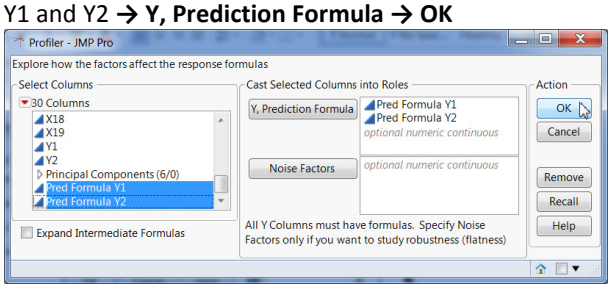
Check model with 6 latent factors. When being satisfied with the model save prediction formulas to the data table.



Important:
Copy the prediction formula (including column properties) to the initial data table with all 45 cases!

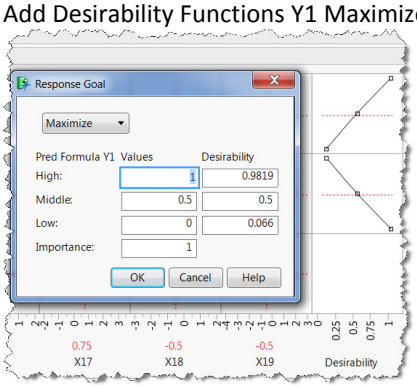
6. Determine the overall optimum solution (from the initial data table with all 45 cases and the prediction formula determined in the previous step)

Select **Graph → Profiler →** select prediction formulas for Y1 and Y2 **→ Y, Prediction Formula → OK**

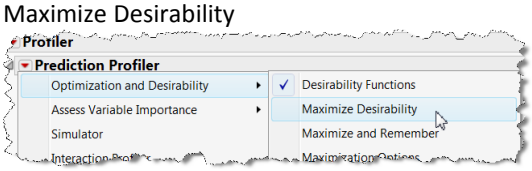


Increase Efficiency and Model Applicability Domain When Testing Options That Are at First Glance Multilevel Categorical Factors - Discovery Summit Europe 2017

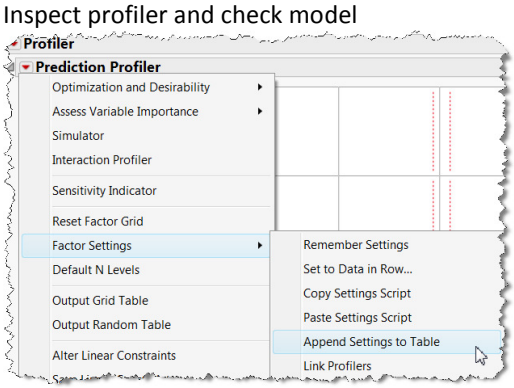
Add Desirability Functions Y1 Maximize, Y2 Minimize



Maximize Desirability

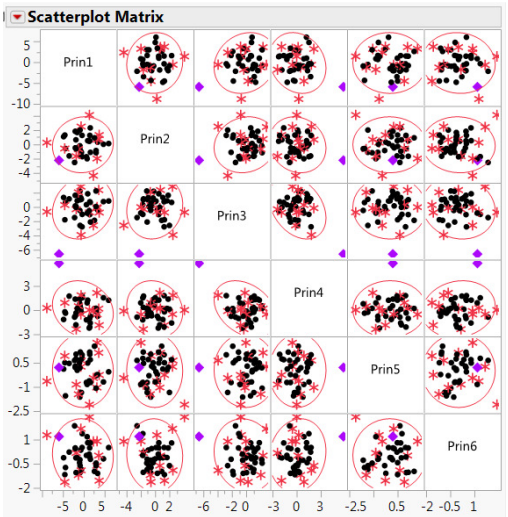


Inspect profiler and check model



Save optimum to data table by selecting **Factor Settings → Append Settings to Table**. New row is added to data table (row 46 containing the set-up and predictions for optimum solution).

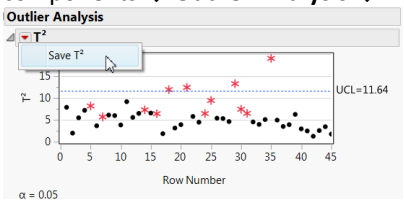
Inspect Optimum Solution e.g. via Scatterplot Matrix:



Optimum solution seems to be outside the model space. Prediction is questionable (extrapolation).

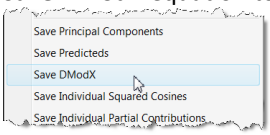
Valid Model Space:

Analyze → Multivariate → select all principle components **→ Outlier Analysis → T2**

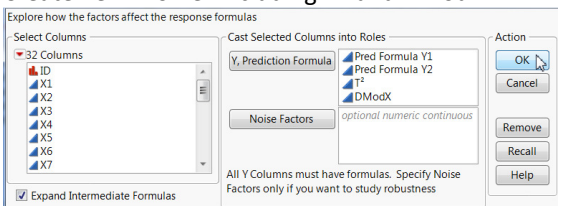


Save T2 equation to data table. T2 quantifies the **deviation of data points <u>on the model plane</u>** (PCA is a multivariate projection method).

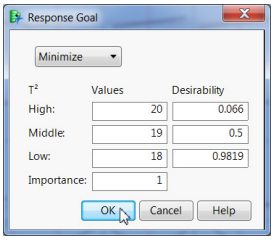
Deviation <u>from the model plane</u> is captured by DModX. For this re-run the PCA with the original predictors and **Save DModX** equation to the data table.



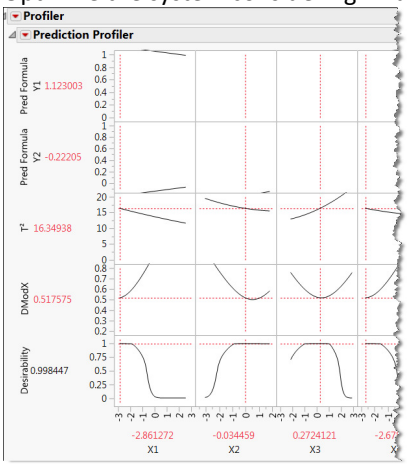
Create new Profiler including T2 and DModX



Set Limits for T2 and DModX



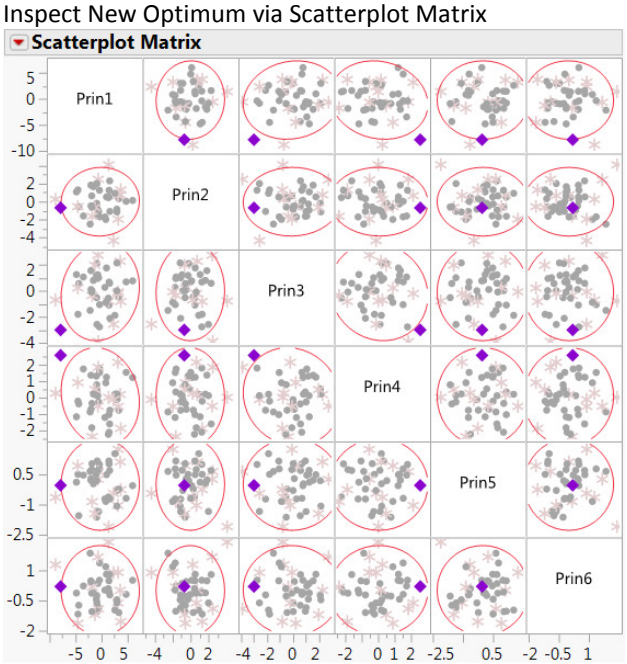
Optimize the system considering T2 and DModX



Silvio Miccio

Save optimum setting to data table (**Append Setting to Table**) and color and mark it.

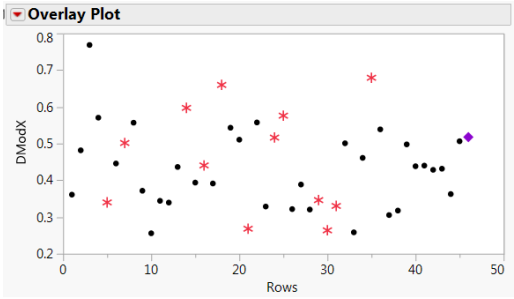
Inspect New Optimum via Scatterplot Matrix



Optimum condition now within the model space.

To check how good the model space coverage is done for deviations from the model plane e.g. create an overlay plot, where plotting DModX by options/rows.

Select **Graph → Overlay Plot → DModX → Y → OK**

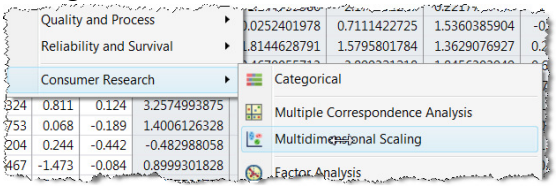


Also in terms of deviations from the model plane the optimum conditions is within the investigated model space.

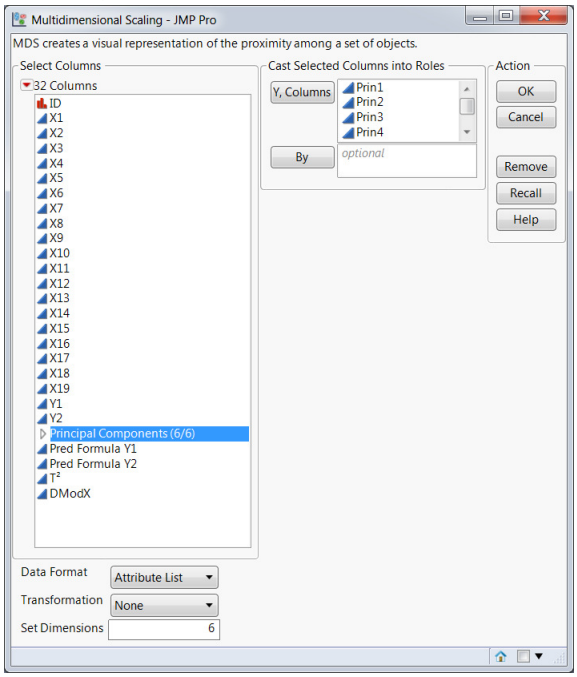
7. Identify physical available option which is closest to the optimum

Overall optimum solution currently not commercially available, therefore it will be interesting to find out which physically available options are closest to the optimum.

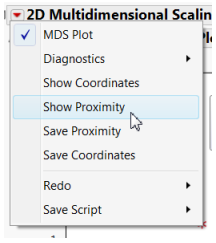
Analyze → Consumer Research → Multidimensional Scaling



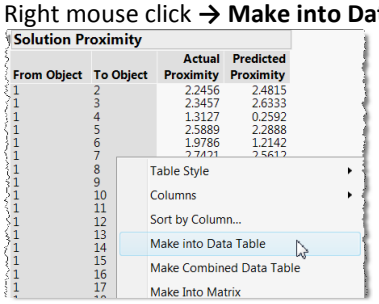
Select Principle Components → **Y, Columns → Data Format → Attribute List → Set Dimensions →** 6



Select **Show Proximity**



Right mouse click → **Make into Data Table**



Select Option 46 in column **To Object → Matching Cells** (in the row menu) **→ Tables → Subset → Tables → Sort → by Actual Proximity → OK**

	From Object	To Object	Actual Proximity	Predicted Proximity
1	18	46	1.5309	0.9688
2	13	46	2.1050	2.2426
3	14	46	2.1886	1.6477
4	9	46	2.4856	1.8103
5	3	46	2.5120	2.1126
6	36	46	2.7347	1.8899
7	11	46	2.9785	2.9717
8	42	46	3.0379	2.8774
9	19	46	3.0883	2.8722
10	20	46	3.1864	3.3490
11			3.2159	3.8592

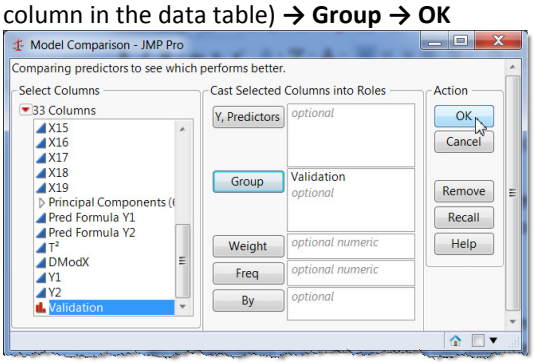
Options 18 is closest to the optimum. Maybe there are other consideration (e.g. commercial) to also consider the next closer options

Validation

Essential for all empirical models is a validation based on an independent test data set (data not used for building or selecting the model). This means that it would have been a good idea to run some more options for being able to test the predictive capability of the model.

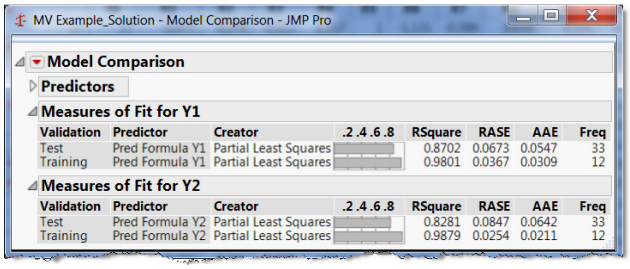
In this particular example in fact information for all options is available. Hence, all options not used for running the experiment can be used for model validation.

Go to the Select **Analyze → Predictive Modeling → Model Comparison → Columns** Validation (hidden column in the data table) **→ Group → OK**



Notice that there is no need to add also the columns with the prediction formulas. JMP usually detects them automatically.

Test data are the data not used for building the model. They provide information about the expected predictive capability of the model. From the table on the left it can be seen that the R-square prediction is bigger than 0.8 for both responses indicating a quite good model, although only 12 from 45 options have been used for modeling i.e. in the experiment.



Further Information

The shown method for designing and analyzing an experiment is partially based on Quantitative Structure - Activity Relationship (QSAR), a method commonly used in Chemometrics. For further information it is suggested to have a look at the related literature.