Increase Efficiency and Model Applicability Domain When Testing Options That Are at First Glance Multilevel Categorical Factors

Multivariate Characterization

Design of Experiments

Projection to Latent Structures (PLS)

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

Introduction

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.

Example Data Set

Consisting of:

- 45 raw material options, which are a potential improvement of one ingredient of the product formulation
- Material options can be described by 19 physical and chemical descriptors (one categorical two level factor was coded as -1/1)
- Two responses to be optimized (Y1 max, Y2 min)

Test Procedure:

- Test (each) raw material option in base formulation

Objective:

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

Steps

1. Transform multilevel categorical factors into continuous factors (or low level categorical factors) based on their physical/chemical properties
   → major effort and challenge of whole procedure!
2. Identify most prominent dimensions of variation in the data with Principle Components Analysis
3. Create an experimental design using the principle components as factors (covariates)
4. Run the experiment
5. Model the results via PLS (or other methods being able to handle correlated factors)
6. Determine the overall optimum solution
7. Identify physical available option which is closest to the optimum

How Many Principle Components/Runs?

How Many Principle Components?

- Most critical is to extract enough principle components
- In practice it has proven to be on the save side when extracting enough principle components to explain more than 80 % of the variation in the data.

How Many Runs?

- Number of runs depends to certain extend on the number of principle components that have to be extracted
- The number of runs is by far not that critical as extracting the right number of principle components
- In practice often 20 % – 25 % of the original number of options is sufficient

Advanced Experimental Designs

It is also possible to run more complex experiments by considering additional variables e.g:

- Additional ingredients
- Add on levels or proportions in a mixture
- Process settings
- Locations/production systems
- Operators
- ….

For doing this there are two options

Option 1: Test Additional Variables in a Custom Design

Model

Main Effects | Interactions ▾ | RSM | Cross | Powers ▾ | Remove Term

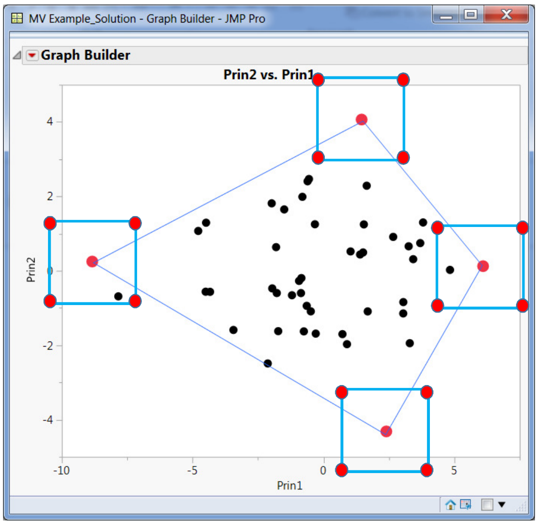
Name	Estimability
Intercept	Necessary
Prin1	Necessary
Prin2	Necessary
Prin3	Necessary
Prin4	Necessary
Prin5	Necessary
Prin6	Necessary
Temperature	Necessary
Pressure	
Time	
Temperature*Pressure	
Temperature*Time	
Pressure*Time	

Design

Run	Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Temperature	Pressure	Time
1	1.459774	4.065948	-0.76688	0.017331	-2.09917	2.40423	-1	-1	-1
2	3.300438	-1.9443	-0.05761	-0.53561	1.743939	0.894154	1	1	1
3	3.809148	1.299812	-0.80438	-1.86969	-0.25649	-1.55175	-1	1	1
4	-0.326	1.251898	-2.77008	1.796158	0.995466	0.50825	-1	1	1
5	-7.84013	-0.68618	-3.00377	2.572315	-0.04456	0.19769	1	1	-1
6	-0.80426	1.989223	-2.58127	0.123937	-1.34175	-0.61985	1	-1	1
7	3.257499	0.659277	-0.19093	2.142645	-1.0416	-0.624	1	1	-1
8	2.666051	0.912149	-1.8766	0.20434	1.530053	-0.97627	-1	-1	-1
9	-1.72799	-1.62117	-0.34643	-2.37072	-0.85292	0.167288	1	-1	-1
10	-1.20257	-0.65523	1.999371	0.490963	-1.45326	-1.20309	1	-1	-1
11	3.701522	0.746091	2.307331	0.186431	1.119849	0.007948	1	-1	1
12	-4.3332	-0.56435	0.584605	-0.38235	-1.55367	-0.73627	1	1	1
13	-8.83675	0.249583	-0.69157	0.267327	-0.64525	1.280862	1	1	1
14	1.654851	2.287485	0.603641	-1.60039	0.692916	0.791342	1	1	-1
15	4.838594	0.02524	0.711142	1.536039	-0.68963	-0.19554	-1	-1	1
16	-0.2936	-1.68681	-1.04904	1.526093	0.773614	1.365872	1	-1	-1
17	-4.48578	1.294069	-1.79564	-0.21418	0.503752	-1.17608	-1	1	-1
18	-4.50055	-0.56213	0.238305	0.90977	0.925168	-1.63407	-1	-1	-1
19	-0.60545	2.406317	2.777117	-0.5178	0.25677	1.671139	-1	1	-1
20	-1.97812	1.814463	1.57958	1.362908	0.212072	0.480226	-1	1	1
21	-2.1275	-2.48916	-1.49697	-0.87037	0.647153	1.853874	-1	-1	1
22	-4.78556	1.071942	1.281411	-1.16382	0.110738	-0.12794	-1	-1	1
23	-0.55374	2.467006	-3.89932	1.84563	0.641503	-0.91093	1	-1	1
24	-3.43978	-1.58949	1.086962	1.741127	0.462491	0.223086	1	-1	1
25	3.049363	-1.14474	2.179191	0.221778	-1.08074	-0.46454	-1	1	-1
26	2.40405	-4.3158	-2.57101	-0.11582	-0.44633	0.887965	-1	1	-1
27	1.546012	1.248633	0.694786	-2.10376	1.023652	-0.08342	1	1	-1
28	6.102162	0.125295	0.792199	1.222359	-0.06133	0.506054	1	1	1

- When it is indented to test additional variables, Principle Components and additional variables are entered in the DOE platform
- In this case more different options have to be tested since more variables require more runs
- Notice that each option can be only run once, which may result in an experimental design not being well conditioned

Option 2: Run Experiments at the "Corners of the Box"



- Example on the left shows an experimental design with two principle components as covariates + two additional variables tested in a 2^2 full factorial design at each "corner of the box"
- For creating a design like this
  - first a covariate design is made based on principle components to identify the options suggested to be tested
  - in the second step a full factorial or custom design is made where the options identified in the previous step are added as factor levels of a categorical variable plus the additional variables
- Notice, modeling is done based on the original variables (physical/chemical properties) e.g. via PLS

Conclusion

- Compress the available information regarding physical and chemical properties of different options via principal components
- Select the "corners of the box" for testing representative raw materials based on Design of Experiments
- Model the data via PLS (or other appropriate modeling methods)
- Find the overall optimum solution
- Identify physical available options closest to the calculated optimum solution

→ **Highly efficient experimentation**

→ **Sustainable empirical model based on physical/chemical properties**

→ **Predicts optimum raw material properties although material not tested i.e. not yet available**