Prediction of dosing behaviour based on powder properties using Lasso regression and Monte-Carlo techniques.

Jean-Vincent Le Bé[1], Isabelle Castella[2], Vincent Girard[3] & Marie Perrot[4] & Sophie Berçot[5]

[1] *Nestlé System Technology Centre, Orbe, Switzerland*
E-mail : *Jean-Vincent.LeBe@rdor.nestle.com*

[2] *Nestlé System Technology Centre, Orbe, Switzerland*
E-mail : *Isabelle.Castella@rdor.nestle.com*

[3] *Nestlé System Technology Centre, Orbe, Switzerland*
E-mail : *Vincent.Girard@rdor.nestle.com*

[4] *Nestlé Product Technology Centre Beverage, Orbe, Switzerland*
E-mail : *Marie.Perrot1@rdor.nestle.com*

[5] *Nestlé System Technology Centre, Orbe, Switzerland*
E-mail : *Sophie.Bercot@rdor.nestle.com*

Abstract

The behaviour of dry powders in dispensing machines depends on a complex variety of parameters, such as powder properties (physical and chemical characteristics), environmental conditions, canister geometry, etc. In this study, the focus is on the powder properties and their impact on the powder behaviour in a given system. Several physical and chemical characteristics of powders can be measured but a direct link between these properties and their behaviour during dosage remains unknown and largely dependent on the setup. Knowing which are the most influent properties and how they influence the powder behaviour would help in the selection of adequate powders instead of doing extensive tests in machine as well as guide new products developments for a good dosing behaviour.

In this context, there are two main critical points to relate powder properties (predictors) to the dosing behaviour (variables):

- the number of predictors is often larger than the number of trials that are possible at a reasonable time and cost;
- for each predictor, there is some measurement uncertainty inherent to the method.

We therefore propose a four steps approach in order to reveal the most meaningful relations between predictors and variables:

1. simulating predictors datasets with Monte-Carlo technique based on their measurement uncertainty;
2. applying Lasso regression on datasets generated by Monte-Carlo technique to reduce the number of predictors;
3. further reducing the selected predictors through a multi-linear regression with a selection criteria based on the ability to predict variables. This step provides both the predictors selection and the predictive model itself;
4. validating of the model with a validation set (not used to build the model).

The approach is exemplified through a real dataset.

Keywords: Monte-Carlo, Lasso, LARS, Error-in-variables, Multi-linear regression.

1. Introduction

The behaviour of dry powders in dispensing machines depends on a complex variety of parameters, such as powder properties (physical and chemical characteristics), environmental conditions, canister geometry, etc. Several physical and chemical characteristics of powders can be measured but a direct link between these properties and their behaviour during dosage remains unknown and largely dependent on the setup. Knowing which the most influent properties are and how they influence the powder behaviour would help in the selection of adequate powders instead of doing extensive tests in machine as well as guide new products developments for a good dosing behaviour.
In this context, there are two main critical points to relate powder properties (predictors) to the dosing behaviour (variables):

- the number of predictors is often larger than the number of trials that are possible at a reasonable time and cost. In this project we had up to 31 predictors for 19 products measured (19 trials);
- for each predictor, there is some measurement uncertainty inherent to the method.

The following approach was used in this particular project in order to reveal the most meaningful relations between predictors and variables:
1. Generate datasets by taking a random value in a given distribution for the predictors' variation (Monte Carlo simulation).
2. If the number of predictors is higher than the number of observations, run a Lasso regression on each dataset to proceed to a pre-selection with a special selection process. Then run a Lasso regression on each reduced datasets and identify the n number of predictors that is required for the majority of models. Select the n predictors that appear the most often among the models.
3. Further reduce the selected predictors through a multi-linear regression with a selection criteria based on the ability to predict variables.
4. Verify the model with a validation set (not used to build the model).

Each paragraph below details one of the steps above and the approach is exemplified with a real dataset with generalized variable and predictors names.

2. Generation of the Simulated Datasets

The first step of the approach is to generate the simulated datasets. Each measurement of the powder properties must be given with its corresponding error expressed as a standard deviation *sd*. In our case, the *sd* originated from two sources: (1) the standard deviation of repeated measurements on the same sample or (2) if the error is a measurement error ±*e*, the standard deviation used is *sd* = *e/3*. This step generates simulated datasets integrating so-called "error in variables".

From the original dataset of measured values X_{ij} of P predictors in S samples (or products), each simulated dataset reads:
$\tilde{X}_{ij} = X_{ij} + \varepsilon_{ij}$, where $i = 1, \dots, P$; $j = 1, \dots, S$

and $\varepsilon_{ij} \sim \mathcal{N}(0, sd_i)$, a random number from the normal distribution of average 0 and standard deviation sd_i (which is the *sd* for predictor i). In this study the number of simulated datasets is N = 1'000.

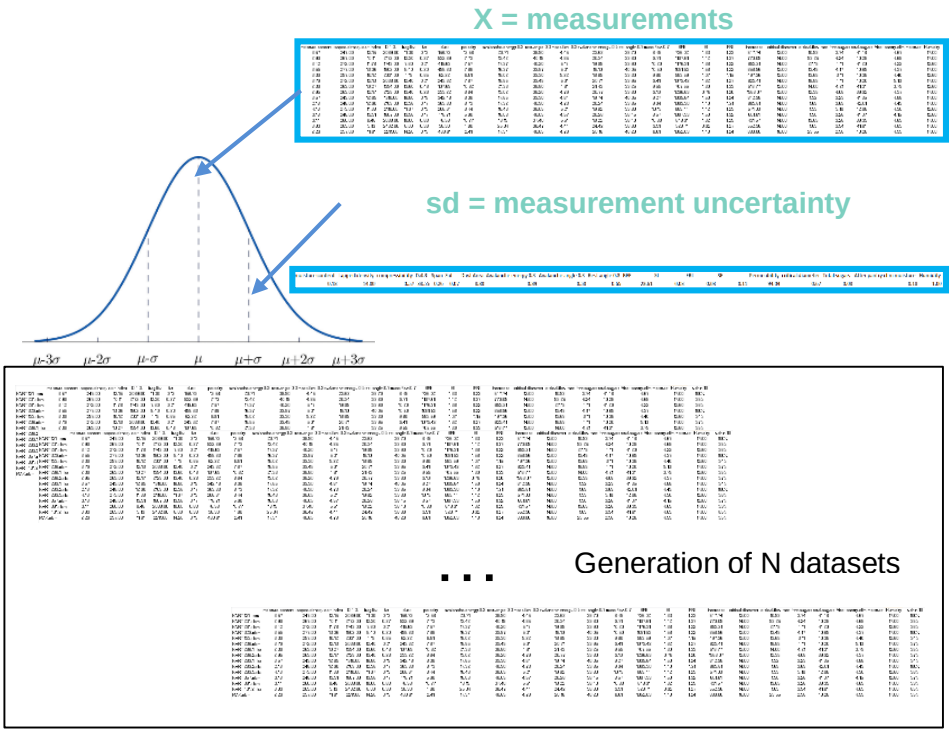

Figure 1: Illustration of the simulated datasets generation.

3. Lasso regressions

3.1. Reduction of the number of predictors to have less predictors than observations

If the number of predictors is larger than the number of observations, the Lasso regression cannot compute a Cp[1] which minimum is taken to identify the Lasso step resulting in the relevant predictors. In order to do a first selection of predictors, the Lasso regression is run using the LARS algorithm on each simulated dataset and the number of predictors to continue with is the number obtained when the R^2 reaches a plateau[2] (Figure 2, detected when the second derivative of the R^2 =f(steps) is minimum). This is done for the top 25% of the maximum R^2 obtained in the process (top 25% are steps with R^2 > 0.75 * range(R^2)). Indeed, when additional predictors do not increase the R^2 further, the model starts to be overfitted and the use of this rule for the steps in the top 25% R^2 prevents an over-reduction of the number of predictors. The plateau indicates the Lasso step to stop at and previous steps can have added or removed a predictor from the list to consider (Figure 3, see predictor X1 removed at step 10 for example).
With this step, the number of predictors was reduced from 16 to between 5 and 12 depending on the random dataset generated.

[1] Cp is a corrected measure of the model error, as defined in James *et al.*, 2013.
[2] R^2 = 1 – RSS / TSS for the fit with the coefficients active and calculated at the given LARS step.

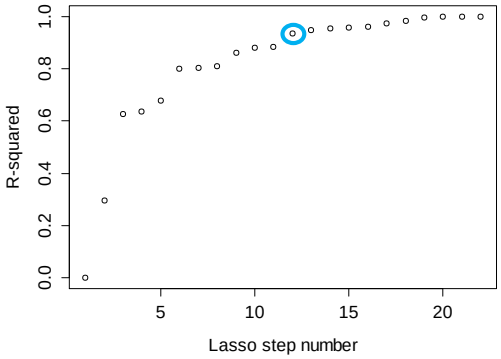

Figure 2: R^2 evolution as a function of Lasso steps. The blue circle indicates the start of the R^2 plateau and the corresponding step for parameter cutting (step 12 in this example, which does not necessarily correspond to 12 predictors as some predictors can be removed by a step).

```
Sequence of LASSO moves:
     X4 X14 X7 X5 X3 X1 X6 X9 X10 X1 X8 X2 X13 X1 X6 X13 X16 X11 X2 X15 X12
Var   4  14  7  5  3  1  6  9  10 -1  8  2  13  1 -6 -13  16  11 -2  15  12
Step  1   2  3  4  5  6  7  8   9 10 11 12  13 14 15  16  17  18 19  20  21
```

Figure 3 Sequence of Lasso steps corresponding to Figure 2. Note that the step 10 removes X1 from the list of predictors to consider.

3.2. Get Models with minimum Cp

Once the selected number of predictors is lower than the number of observations, a Lasso regression is run on the same but reduced simulated dataset and this time, the step resulting in the lowest Cp is selected. All active predictors at this step are recorded. This is done for each simulated dataset.

Once the 1'000 simulations are completed, the number of selected predictors finally considered is the number occurring the most frequently in the simulations (7 on Figure 4 left). The selected predictors are the most occurring in the simulations until the optimal number of predictors is reached (Figure 4 right). The number of predictors was then reduced from 16 to 7 in this example.

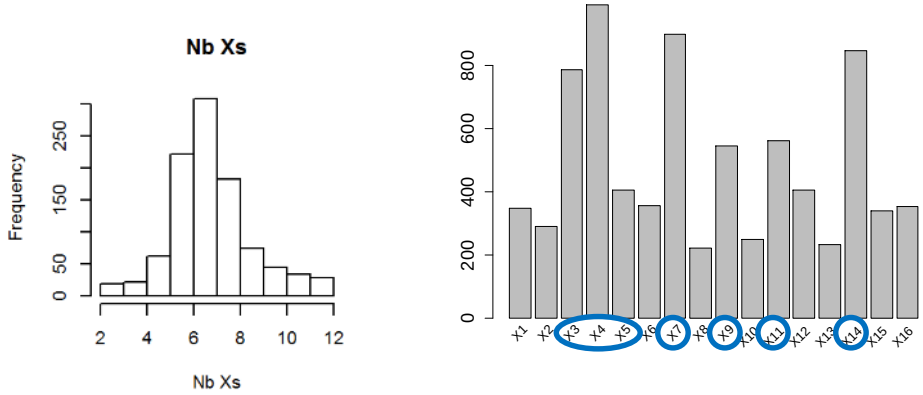

Figure 4 Left: Frequency of numbers of predictors in the best models for the 1000 simulations. Right: Frequency of occurrence of each predictor of the original dataset in the 1000 models. Selected predictors are circled in blue.

4. Multi-linear regression and final model calculation

The objective of this step is to reduce as much as possible the number of predictors without reducing too much the adjusted R^2. The final combination of predictors may not be the one resulting in the strictly smallest error but it may be more useful in practice. In addition, it is very important for this step to have experts in the field who are involved in the process to select predictors that can be interpreted. It will also help choosing to eliminate a meaningless predictor and keep a meaningful predictor even if the former would give a mathematically better prediction (but for only a few percent or even less in the adjusted R^2).

This predictors' reduction is done with simulating again datasets with Monte-Carlo principle. Multilinear models were calculated for each set and the p-values of the predictors selected in step 2 were displayed (Figure 5). The predictors with a p-value ranging from 0 to 1 and being too much uniform among the simulated models were discarded and the simulation was ran again. One predictor was discarded by step until p-values histogram had a peak towards 0 (Figure 5) or the adjusted R^2 of the model was decreasing too much. This indicated that the removed predictor, even though having a high p-value in some simulations is more explanatory than random noise.
A CR-plot is the "Component + Residual" plot against the predictor. The component is the predictor multiplied by the corresponding coefficient. The residual of the corresponding point is then added and the result is plotted. It shows how strongly the predictor affects the variable and how much noise is around this predicted point. The CR-plot was then displayed for each step to help discarding predictors which parameters were too small (Figure 6).

The final model is obtained with the original dataset and the remaining predictors selected based on p-value and relevance given by experts (Table 1). This final step allowed to reduce the number of predictors from 7 to 3.

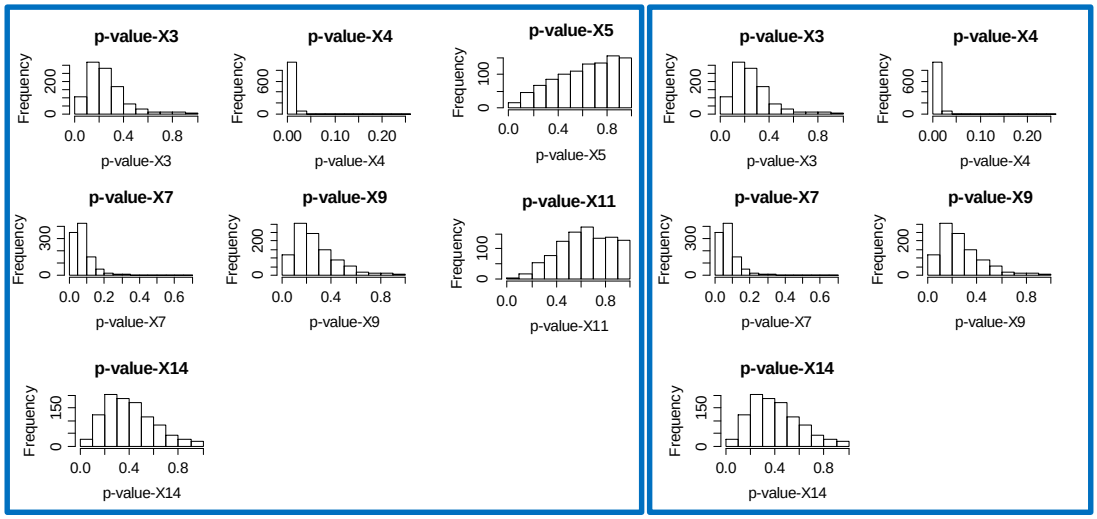

Figure 5 Example of predictor elimination with p-value comparison (elimination of X5 and X11).

	CoeffsMeas
(Intercept)	1.549247e+00
X4	-8.761529e-05
X7	4.914006e-05
X14	-1.516741e-02
R-squared	8.001594e-01
Adj R-squared	7.402073e-01
p-value-X4	6.191382e-04
p-value-X7	2.527405e-02
p-value-X14	4.834882e-02

Table 1 Final model selected calculated directly from the original dataset (non simulated).

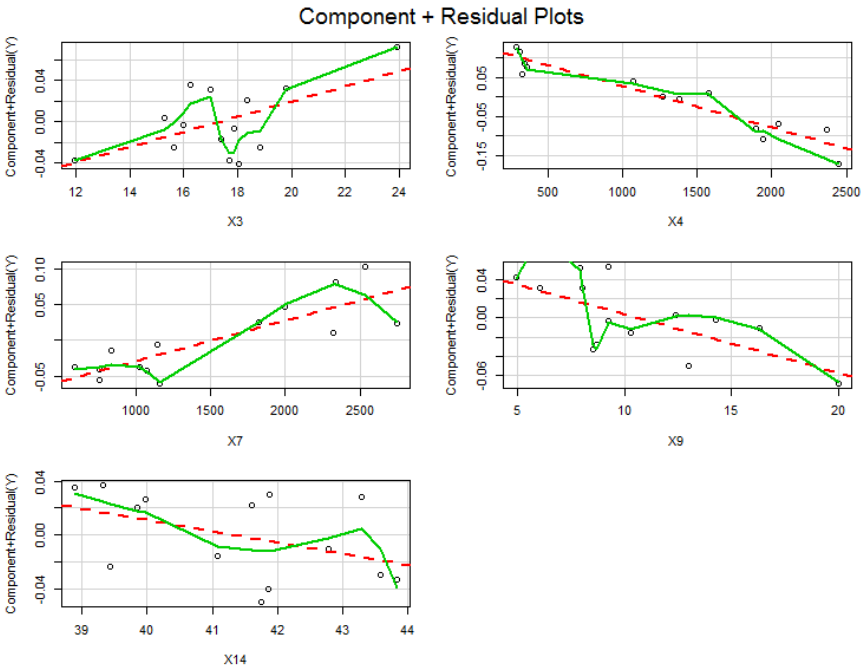

Figure 6 CR plots corresponding to the model in Figure 5 right. Y-axis is the predictor multiplied by its corresponding coefficient to which is added the residual for the corresponding sample. The green curve is a smoothing of the dots and the red dashed curve a linear regression. A regular deviation between the red and green lines indicate that the variable is not linearly dependent on the predictor.

5. Validation

The coefficients obtained with the original dataset and the finally selected predictors were then used to generate a validation graph showing the predicted values against measured values (Figure 7). Additional data (in this case 5) that were not in the original dataset was also plotted the same way. It is then possible to assess the accuracy of the model. The model was considered as validated when the difference between the predicted and observed values was not more than two standard deviations of repeated measurements of the variable for both the construction and validation sets.
In the case exampled here we started with a set of 16 predictors and reduced it to 3 predictors fitted on 14 observations giving an R^2 = 80%.

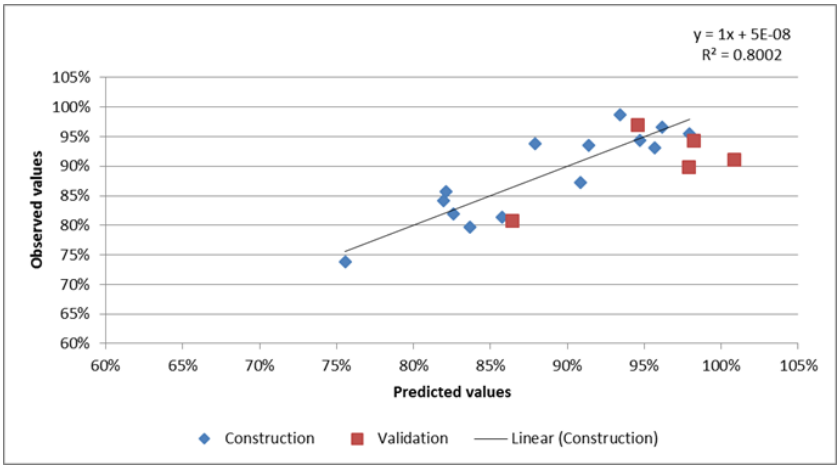

Figure 7 Validation graph

6. Conclusions and improvements

The method proposed here to solve a so-called "error in variables model regression" is the result of successive analysis that started with a direct Lasso regression on the original dataset. Then the variation in the predictors encouraged the use of simulated datasets with a Monte Carlo approach. The original program was not optimized enough and the computation time was relatively long (about 8' per variable and environmental condition), hence the choice of only 1'000 sets repeated 3 times. With the improved program (computation time reduced to 1'25'', i.e. divided by 5.6), a unique simulation of 10'000 or more could be done. The development of a convergence test would also be useful and would help defining the appropriate number of simulations needed.

Another improvement could be to consider as well the variation of the output variable in the datasets generation. In our case, this variation was considered as being included in the residual error of the models. Testing and comparing both techniques would be of interest in a future development of the method.

References

James, G., Witten, D., Hastie, T. & Tibshirani, R. (2013). *An Introduction to Statistical Learning*, Springer, ISBN: 978-1-4614-7137-0.