Predicting complex industrial processes using Sparse Regression Models

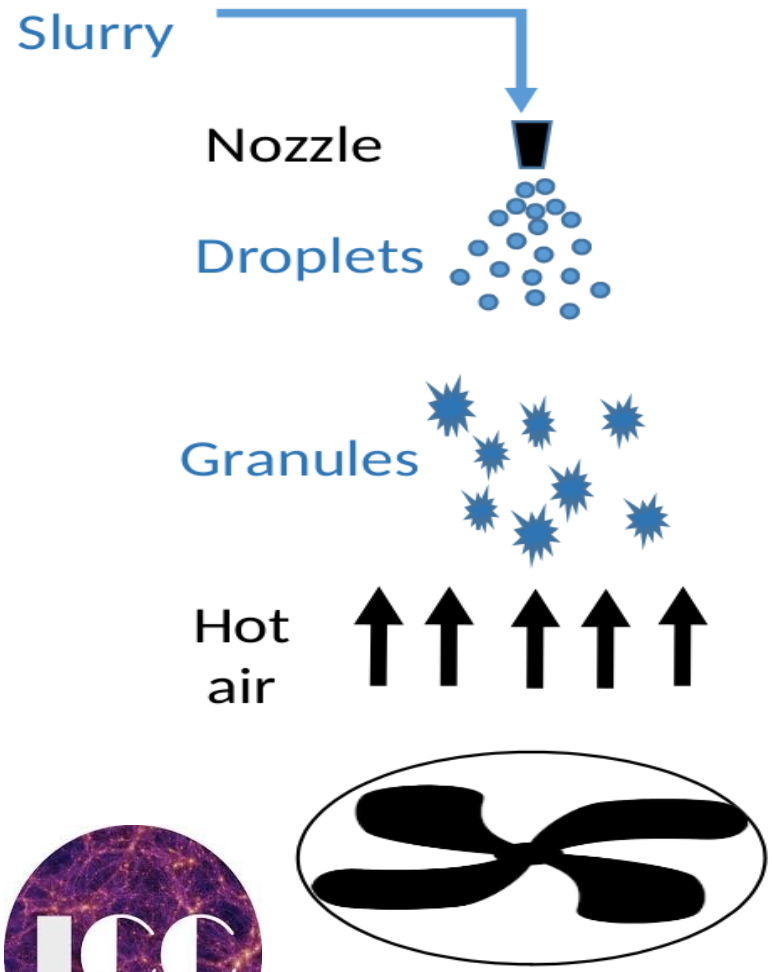
Project done under the supervision of Prof. Richard Bower. Mr. Stefan Egan and Mr. Arturo Martinez.

Miguel de Icaza Lizaola (ICC, DU).

29 September 2019

Durham University

The problem: Predicting physical characteristics of a spray drying process for laundry powder.



Durham University

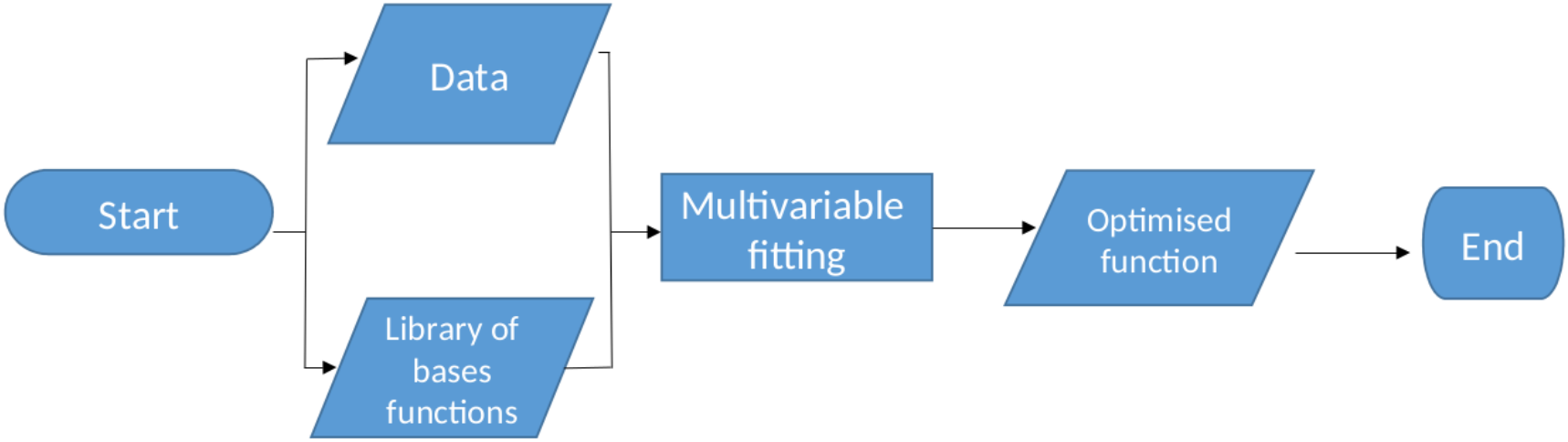
ICC

Our task: Finding the optimal parameters.

$$Property(x_{1,\ldots},x_n) = a_0 \cdot f_0(x_{1,\ldots},x_n) + \ldots + a_m \cdot f_m(x_{1,\ldots},x_n)$$

Ingredients and Process Conditions

Parameters from the linear combination of bases functions



Durham University

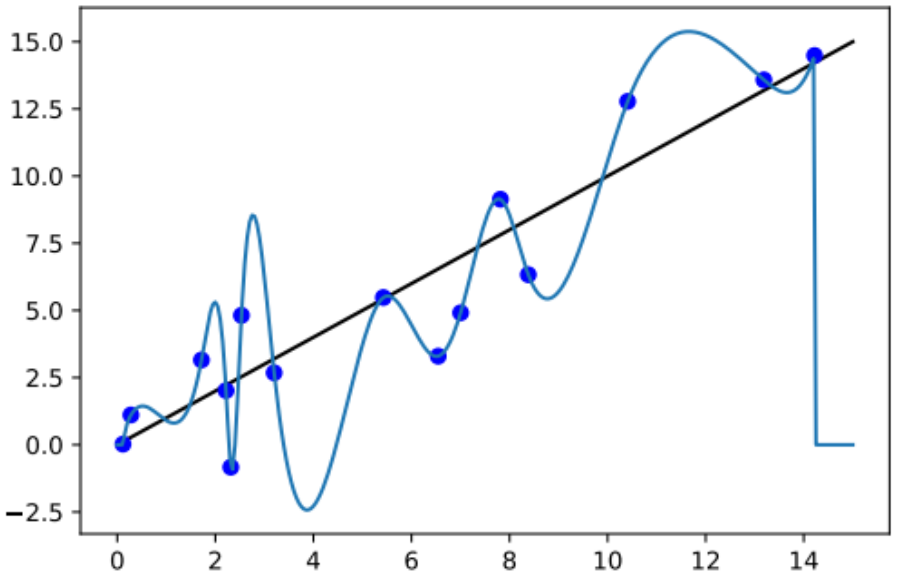
ICC

Model: 13-20 Independent variables combined in hundreds of functional forms.

100-1000 Parameters!!!!!

Data: ~300 points!!!

We have to be very careful with overfitting.



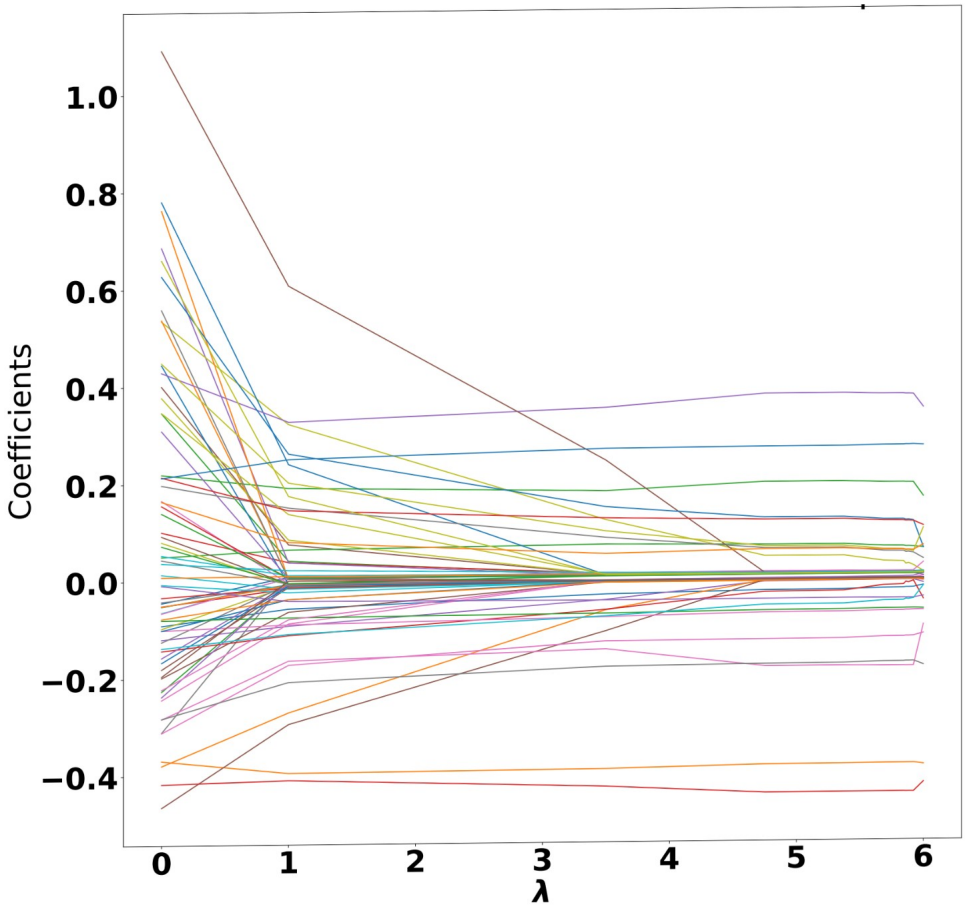
The blue points are fitted by two different models the turquoise line is a perfect fit to the data but it is difficult to justify. The black line while being the worst fit is simple and more realistic.

Durham University

ICC

Solution: Sparse regression with a penalty function.

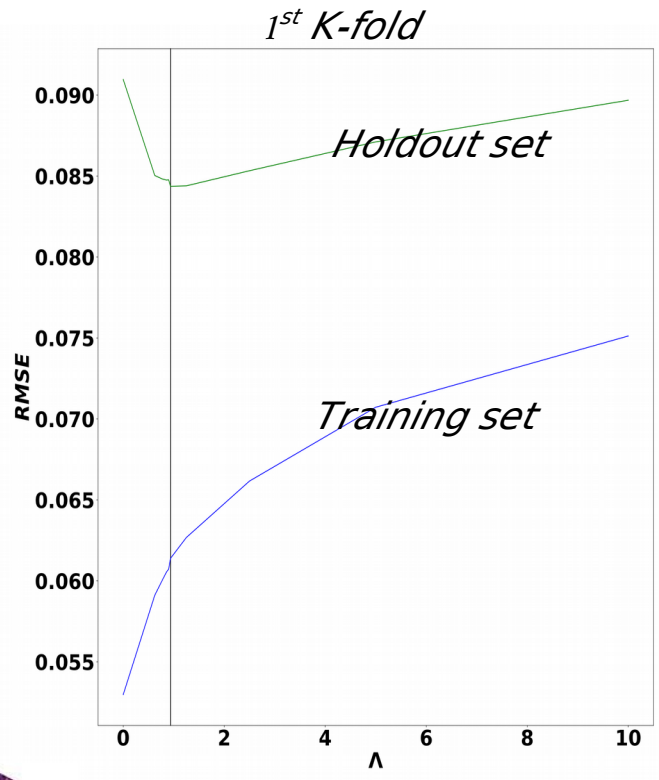
$$L = \frac{1}{N}\sum_{i=0}^{M}\left(y_i - f\left(x_i\right)\right)^2 + \lambda \times Penalty_{NDA}$$

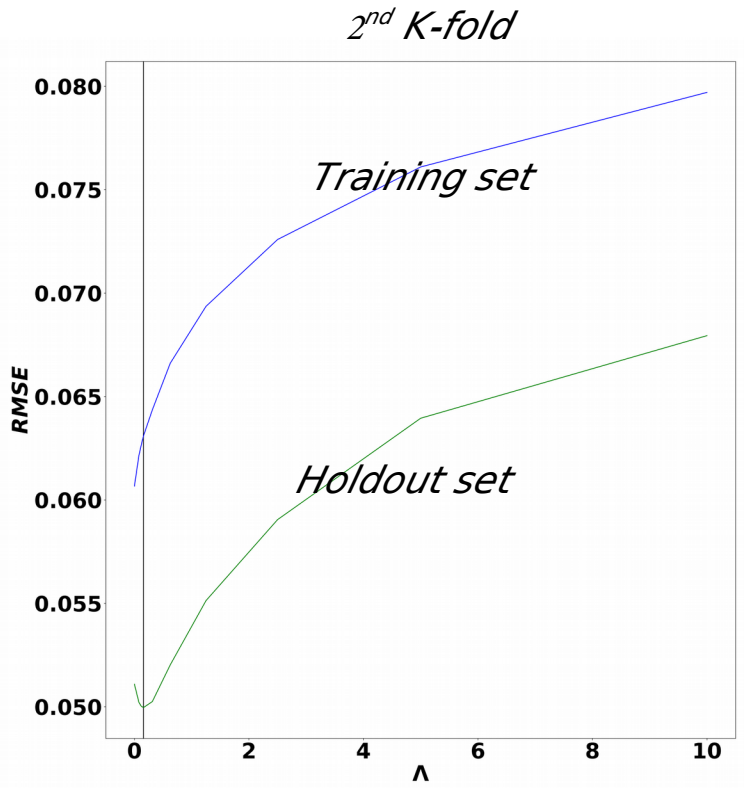


Best fit normalized coefficients as a function of the penalty parameter.

Durham University

Standard Penalty selection method (k-folds) is too unstable.



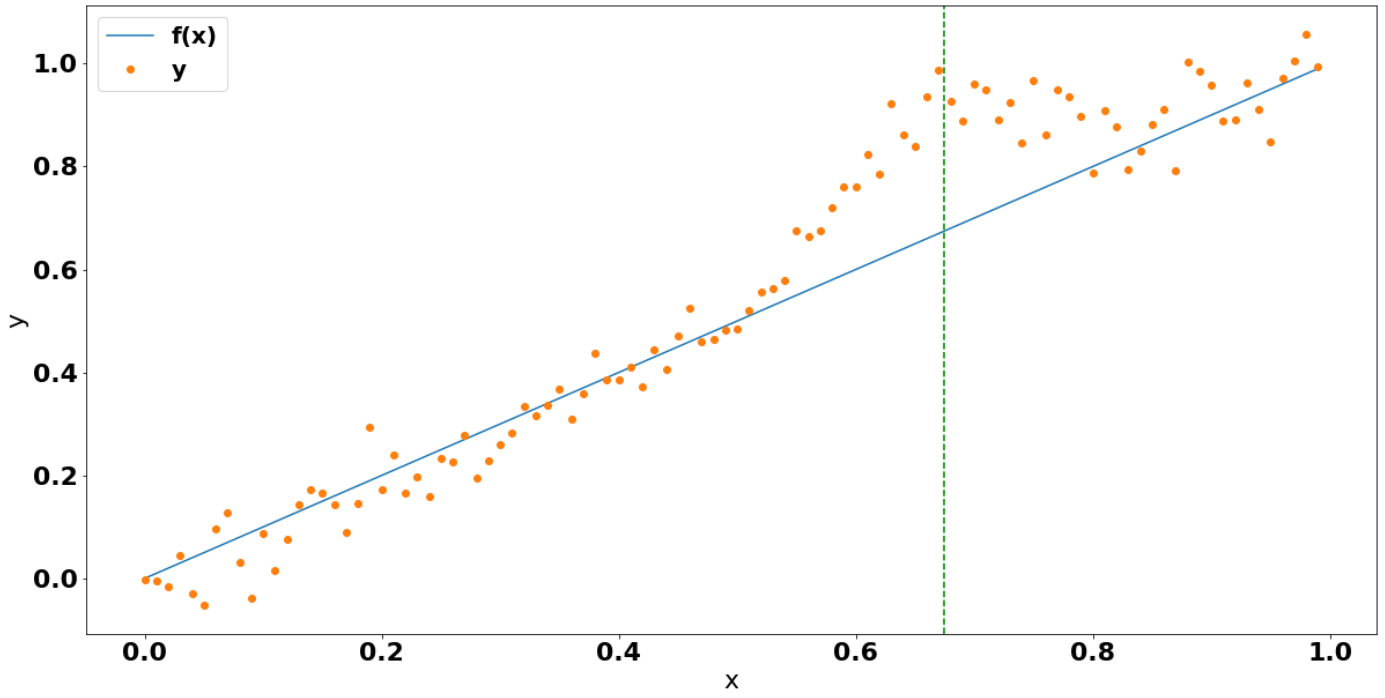


The K-folds method divides the data into N fractions. One of them is selected as a Holdout set and is systematically kept out of the fit.

Solution: Use mock data sets (Icaza-Lizaola et al. In prep).

Durham University

Sparce regression models the global behavior properly, but **Gaussian Emulators** can be used to improve local deviations.



Example of one data set fitted by a model (Blue line) where the global behavior is properly explained, but local deviations would not be modeled.

Durham University

ICC

Conclusions

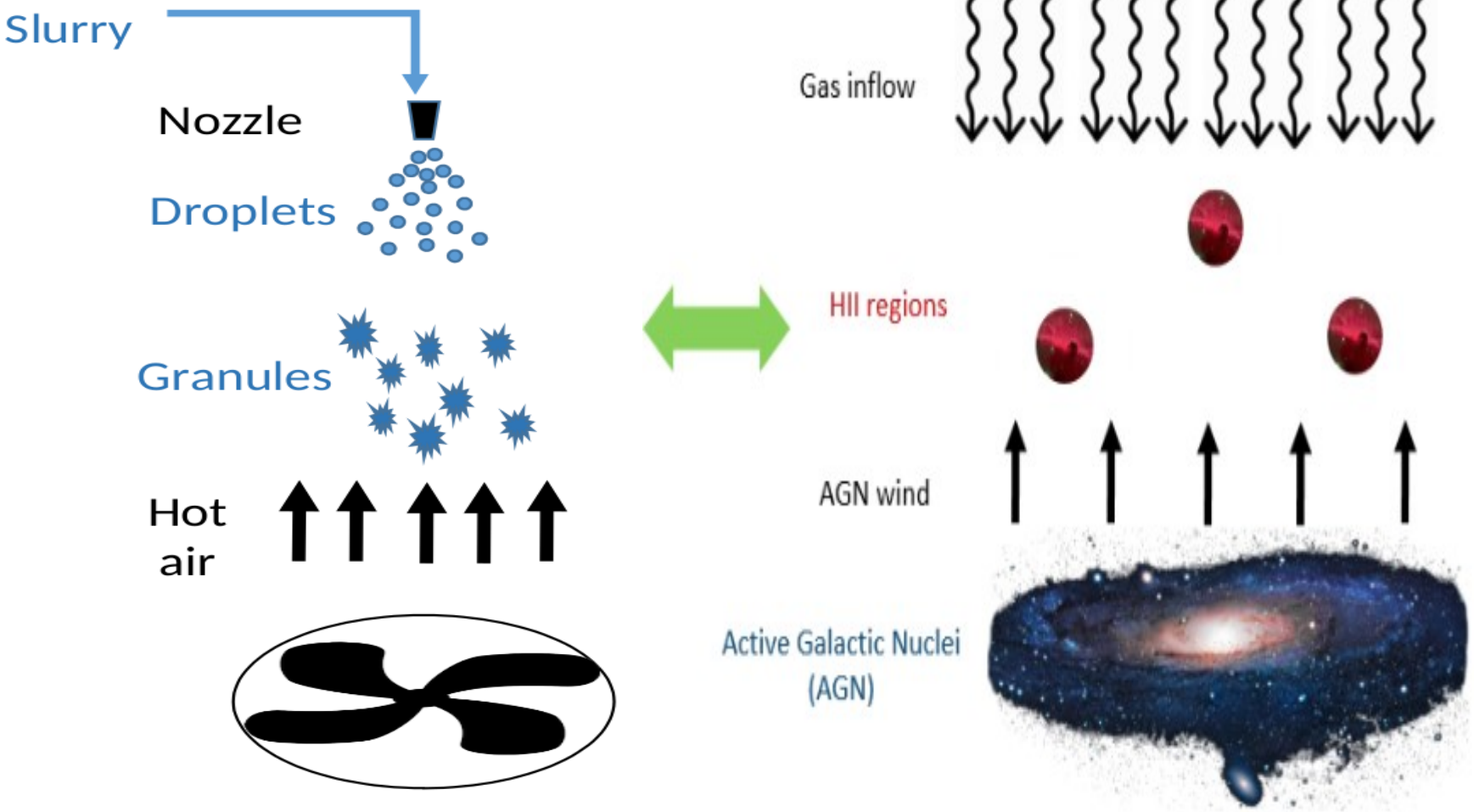
-We developed an automatized methodology that is able to model P&Gs spray drying tower parameters.

-Our sparse regression model has similar accuracy to that obtained by P&G chemical engineers.

-When enough data is included Gaussian Emulators make our code more accurate than previous P&G models. However, the lack of a functional form makes it harder to interpret.

-We are currently using our methodology to try to predict the Mass of a galaxy inside a Dark Mater halo.

Durham University



Durham University

ICC