WHICH MODEL SHOULD I CHOOSE?

de Brauwere Anouk[1], Fjo De Ridder[2], Rik Pintelon[2], Marc Elskens[1], Johan Schoukens[2] and Willy Baeyens[1]

[1] Department of Analytical and Environmental Chemistry, Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium
E-mail: adebrauw@vub.ac.be

[2] Department of Electricity and Instrumentation, Vrije Universiteit Brussel
Pleinlaan 2, B-1050 Brussels, Belgium

After collecting a set of data, often the difficulty arises of explaining the observed patterns. Which mechanisms generated these numbers? A priori, many theories could account for the observations, but the question is which of them is closest to the true machinery underlying the measurements. To differentiate between theories in an objective way, it is necessary to translate them into mathematical models. Only then, these models can be quantitatively compared to the numerical data.

Yet, the problem remains to choose exactly how this comparison should be made. In other words, which "quantity" determines how appropriate a model is to describe the given data? Obviously, the optimal model should fit the data well. So goodness-of-fit quantifies, at least partly, the suitability of a model. On the other hand, any measurement is subject to some random error. Consequently, a model that fits the observations too well is not acceptable because it is actually partly modelling the errors. Furthermore, this kind of model will be highly inefficient to account for future or replicate data, since these will be subject to different random noise. To summarize, the best model should exhibit a subtle balance between goodness-of-fit and robustness.

This problem of "model selection" is of present importance in many fields. For instance, think of the climate or ocean models, which are made increasingly complex, sometimes without obvious proof that this complexity is supported by the data.

We propose a method to objectively choose the most appropriate model given a certain dataset. In brief, the Weighted Least Squares cost function is a sample of a known χ^2 distribution. This enables an assessment of how "probable" and thus acceptable a given model is. This approach is combined with the principle of parsimony by stating that the simplest of all acceptable models should be selected. It is very intuitive and easy to implement. The only requirement is the availability of the measurement uncertainties. Although the simplicity of the method, it performs well in very distinct situations, as will be shown on the poster.