

# Modern Methods of Data Analysis

## Fourth winter symposium on chemometrics

Report for the award of an international travel grant by the RAEng  
Attendee: Jaap van der Weerd, Imperial College London

### *Scope of the conference*

The winter symposium on chemometrics is a small conference on data-analysis procedures. The original aim of these conference series was to improve knowledge and skills on chemometrics among young Russian scientists. Advanced data-mining forms a cheap way to optimise the results obtained in experiments or in manufacturing plants, without the major (and often unavailable) expenses needed to buy new equipment. The first meetings were organised as courses, but this has gradually evolved towards a more normal conference, where scientists discuss current work along with a normally extensive introduction. The WSC conferences are still small, but the organisers manage every year to invite a number of leading people in the field. This combination is excellent to speak to people and discuss issues encountered in daily life.

Some of the well known scientists I met are John Kalivas, Vladimir Palyulin, Kurt Varmuza, and Alexey Pomerantsev. Unfortunately, Kim Esbensen could not come this year due to illness. This year's most pleasant encounter was with Dr. Roma Tauler, (Barcelona). He is well-known for his work on multivariate Curve Resolution (MCR) and the software he distributes via his website.

Most lecturers are aware that most of the audience (including several of the Russian attendees) are beginners, and start with a thorough introduction. This made conference was particularly useful to me: I do not have any formal education in the discussed field, but apply data-mining techniques on a very regular basis. The lectures are really beneficial to acquire a broader overview of the field of chemometrics.

The conferences are organised in low seasons, away from tourist spots and big cities to keep the costs low. This year's conference was organised in Chernogolovka, a small town close to a forest and beautifully covered by a layer of snow. The town basically consists of a number of scientific institutes and their employees.

### *Lectures*

The small scope of the conference implied that there were no parallel sessions. As always, some of the lectures are completely incomprehensible, while with some others leave you wondering where they do fit in into the bigger framework of the field. The most interesting contributions for me were presented by:

**Alexey Pomerantsev**, who discussed with his characteristic wit the technique of simple interval calibration (SIC). I once before heard a lecture on this (by someone else), but didn't understand it at all then. This time, there was an excellent introduction. The characteristic element of this form of calibration is the assumption that measurements have a finite error. This seems logical, but is in fact unlike the standard error normally assumed. With a standard (Gaussian) error, it is assumed normal that a small number of measurements have very large errors (might be >4 times the standard deviation). In SIC such large deviations are considered errors, and thus ignored. A neat way was shown to use this as a starting point to establish a calibration model and determine confidence intervals.

**Vladimir Palyulin** discussed his work in regression of chemical properties, basically trying to predict chemical properties of a compound from its molecular formula (QSAR/QSPR). Based on the molecular formula, a number of 'descriptors' are formulated, such as number of atoms, bonds, connectivity and many others. Descriptors of molecules with known properties (such as toxicity, activity as a drug, or simply boiling point) are used to formulate a calibration model, enabling the theoretical investigation of many other compounds. In this way, many compounds can be screened before their actual synthesis, enabling a smart search for components with required properties.

**Roma Tauler** discussed different forms of looking at data. He discerned the 'white' models, which fit data to a physical model, 'black' models, which try to formulate models based on the data only. He proposed 'grey' models, which do not require an explicit physical model, but start with a number of restrictions which a solution should satisfy. An obvious restriction is non-negativity for absorption spectra or concentrations, but several other restrictions can be formulated. This approach might be very useful in my current work: I normally use classical least squares (CLS), which should be classified as 'white'. I do find cases where my model is not adequate for reality. I did find some ways around, but a 'grey' approach might be more stable.

**John Kalivas** discussed the effective rank of different calibration models, i.e. principal component regression (PCR), partial least squares (PLS) and Riggs regression (RR). Though his talk was theory oriented, more than a humble experimentalist could understand, it was good to get a feel of the deeper workings of these models, and the pitfalls that users should be aware of. The same was true for the presentation of **Juan Pierna**, who discussed the various forms of uncertainty in multivariate calibration.

A number of interesting talks discussed electronic noses and tongues, photochemical reactions, image processing (Fourier transform, wavelets, and AMT), outlier detection, and archaeometry. These lectures presented some interesting applications, but are applied and did not really add new insights. One exception are the neural networks (NN), which came up in a number of these presentations and the concluding discussions. I had heard few things about NNs, but nothing very practical. It was very interesting to hear

knowledgeable people discuss on the relation between NNs and other techniques, such as PCR and PLS. The take-home message for me was that NNs are useful for calibrations where the experimental response is not linear with the sought value. The downside is a tendency to over-fit (i.e. noise is used for calibration rather than the relevant measurement signals). A very good validation of the model is thus crucial. It is good to know that NNs are available, and when they are beneficial. I am quite confident that I would recognise cases where the classical or inverse methods I use now become inadequate, and I'll have to go to NNs.

I presented my own work on drug release from tablets in an oral presentation. I got quite a number of compliments afterwards, and conclude that things went well. It was most informative to discuss afterwards with Dr Tauler on the possibility to apply his grey modelling approach to my data. I hope some collaboration between the research groups will be established.

#### *Downsides*

The few downsides of the conference were the occasionally very cold and windy weather outside and the sub-optimal catering. However, these kept us indoors and focussed on the science. Amazing how the organisers exploited these details.... The poster session was small and of poor quality, consisting of posters with the amount of text you would expect in an ordinary article. I find it impossible to get through these texts in public. In a few cases, the authors explained their work, but generally, this session didn't really work out.

#### *Finally*

However, these downsides are minor, and overall the conference was well worth visiting. I enjoyed the many good introductory lectures, the casual setting, and the ease with which it was possible to talk to established scientists. I am confident that these encounters can be beneficial in my future career. I thank you truly for your support.

Sincerely,

Jaap van der Weerd