Parametric identification using global optimization methods

E. Balsa-Canto and J.R. Banga

Model building is an iterative process which starts from the definition of the purpose of the model and the use of the existing knowledge to choose an appropriate framework to propose a first model structure. Given such structure and a set of experimental data, the objective of parameter identification is to estimate the non measurable parameters and, possibly some initial conditions, so as to reproduce the experimental results in the best possible way.

The parameter identification problem is formulated as a nonlinear optimization problem, where the objective is to find the set of parameters to minimize the function quantifying the goodness of the fit subject to the system dynamics.

This apparently simple formulation hides several difficulties that will definitely condition the predictive capabilities of a given model. Particularly relevant are: the multimodal nature of the optimization problem, that is, the presence of several suboptimal solutions; and the lack of (structural or practical) identifiability, that is, the impossibility of giving an unique solution for the parameters.

This talk addresses such difficulties and presents and overview of the methodologies being developed at the (Bio)Process Engineering Group-CSIC to deal with them. Particular emphasis will be paid to:
- The development of new stochastic global optimization methods, such as hybrid stochastic - deterministic or scatter search techniques, which are able to efficiently overcome sub-optimal solutions for multimodal cases.
- The solution of optimal experimental design problems via dynamic optimization.
- The applicability and advantages of the proposed techniques will be illustrated with several examples related to the modeling of biological systems.

Related works

Evaluation and unique identification of unidentifiable models

G. Cedersund

This talk deals with the evaluation, comparison, and ultimately the analysis, of competing biological hypotheses for a given data set.
After formalising a given hypothesis to a mathematical model, and optimising it to estimation data, the agreement with estimation/validation data can be assessed. There are a number of statistical tests available for this purpose, and I will review and discuss some of these. I will also look at methods for comparison of two models, including a rather generally applicable parametric bootstrap approach.

The final part of the talk deals with the problem of evaluating predictions in unidentifiable models. I will talk about some recent approaches we have developed to find the global uncertainty of arbitrary model predictions, and how this can be used to find core predictions, i.e., uniquely identified model predictions even when the model itself is unidentifiable.

The different steps in the model-based data analysis will be illustrated on the analysis of time-series data of insulin signalling in human adipocytes.


----------------------------------------

Sloppy parameter sensitivities in systems biology
Ryan Gutenkunst

Many quantitative modeling efforts, particularly in biology, are hampered by a profusion of unmeasured parameters. We demonstrate that biochemical network models exhibit a universal "sloppy" pattern of sensitivity to parameter variation; different directions in parameter space vary by orders of magnitude in their constraint. We argue that, consequently, predictions of future experiments may be usefully constrained even when the available data only very poorly constrain parameter values. This suggests a change in mindset for some, away from a focus on parameters and toward a focus on predictions.

----------------------------------------

Estimation in Systems Biology under Nonnormality
Chris A.J. Klaassen

When data about a biological system are being collected, two issues arise of a statistical nature. One issue is the construction of a statistical model, the other one is parameter estimation. The statistical model should adequately describe both the biological system with its intrinsic randomness and the measurement errors inherent to the data collection. We will focus on the situation in which the randomness in the model may not be assumed to be normally distributed. The classical estimation technique of least squares does not extract all information from the data then. Efficient estimation will be discussed and illustrated in a statistical model based on a
Experimental Design in Systems Biology

C. Kreutz, J. Timmer

Aspects of experimental planning are discussed for applications in the cell biology, especially in Systems Biology. In Systems Biology, mechanistic models are established describing the interactions and time dependencies of cellular processes. The case of a deterministic nonlinear model, e.g. defined by a system of ODEs, is chosen.

In the first part of the talk, design optimization in a relatively well defined experimental situation, as it is common in technical disciplines like physics, is described. Numerical design optimization methods for the purpose of parameter estimation and for model discrimination are introduced. The outcome of such approaches are optimal sampling times, stimulations and the most informative observables.

In the second part, additional aspects are introduced for more heterogeneous and less controllable settings as they often appear in cell biology. Classical statistical approaches to deal with uncontrollable biological and technical variations are summarized.

In contrast to a specific application, the focus of the talk is rather providing an overview about relevant principles for Systems Biology.

Error structure of spectroscopic data (NIR, FTIR etc)

Harald Martens and Achim Kohler

Measurements are usually done in order to quantify an “analyte” – a chemical or physical property of the samples analyzed. But measured data usually reflect several sources of variation, not only the desired chemical or physical property. This is clearly true for spectroscopic measurements. By proper design of the measurements and proper modelling of the data, it is possible to separate the measured signals into the various sources of variation. Of course, different measurement types and different sample types offer different error structures. But most systematic errors have a surprisingly systematic nature, from a mathematical point of view, and can therefore be discovered and corrected for by the same procedures.

This lecture addresses some psychological, technical, mathematical and statistical issues in how to reduce undesired variations in measured spectra.

Keywords:
Remember to address also the undesired phenomena – the interferants, not only the desired analytes! Usually, this requires multi-channel profiling.

Increase the number and type of measurement channels in the profile, so that even the interferants can be quantified and subtracted. This may increase the cost of the measurements a little, but saves you a lot of problems later.

If possible, measure each sample under a set of different conditions – this increases the information value of a given instrument type.

Study the raw spectra graphically, in 1-way, 2-way and 3-way plots, to detect unexpected phenomena. Check them also by PCA etc to look for more “hidden” structures.

Model-based pre-processing should then be applied, building on a combination of approximate causal understanding and empirical covariations. The purpose is to identify and correct for various signal contribution types: random noise, wavelength shifts, multiplicative amplifications (e.g. instrument amplification, sample thickness, effective optical path length), additive contributions (analyte and interferant concentrations), log-additive contributions (stray light) and response non-linearities of various kinds. Interference effects should be removed, but stored for later studies, - not just “filtered” away and lost.

Multivariate calibration models should finally be developed, in order to enhance the selectivity and provide graphical insight into the main structures in the pre-processed data.

Analyte predictions and multivariate scores are in turn obtained by passing new spectra through the calibration models, and can in turn be related to other types of information, e.g. from genetics and genomics.

Automatic outlier warnings of various kinds should be used in order to detect anomalies.

Selectivity enhancement by multivariate analysis has been well known for 25 years in Chemometrics. But the field is still developing, as the lecture will illustrate, e.g. based on high-throughput instrument types (FTIR robotics) and various new uses of the Extended Multiplicative Signal Correction (EMSC).

Measurement Errors in Gene Expression Data from Spotted DNA Microarrays

Peter D. Wentzell

From their earliest use, it has been widely recognized that the utility of DNA microarrays can be limited by measurement quality issues. These issues have led to some approaches to experimental design and data analysis that are unique to microrrays. For example, variation in spot morphology in spotted arrays led to the use of two channel measurements where one channel acts as a kind of internal standard. Systematic variations in the two channels due to differences in sensitivity (caused by a variety of factors) led to a host of different normalization strategies, the most popular being LOWESS (locally weighted scatterplot smoothing) and its variants. Logarithmic transformation of expression ratio data is normally applied to provide a more Gaussian error structure, although this can have ramifications for the undelying structure of the data. Automated and manual “flagging” is also performed to exclude measurements of dubious quality, leading to issues with censored data.
Based on these and similar practices, it is fair to say that microarray data analysis is driven in large part by the error structure of the measurements. It is important, then, to consider this error structure in some detail and develop models that are useful in practice. This presentation will discuss the nature of measurement errors in spotted DNA microarrays, describe methods to estimate and model these errors, and comment on the importance of this information in multivariate data analysis.

-----------------------------------------------------

**Metabolomics data: Structure and Noise**

Johan A. Westerhuis

The analysis of metabolomics data is a complex issue. Multiple analytical platforms are often necessary to obtain a broad overview of the large range of metabolites in a biological system. NMR, GCMS and LCMS are the most used instruments and work in rather different concentration ranges. Each of these platforms has its own problems in delivering a list of measured metabolites and their concentration due to sampling, analytical, technical and data handling issues. In this presentation some of these problems will be discussed as well as their effect in the multivariate analysis of the metabolomics data. Finally, preprocessing methods that can help in reducing the effect of the problems as well as specific experimental design issues to help the multivariate analysis are discussed.