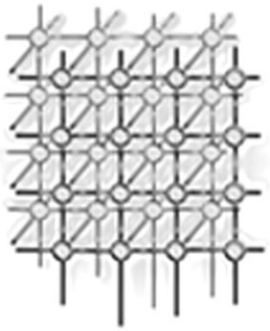
CONCURRENCY AND COMPUTATION: PRACTICE AND EXPERIENCE
Concurrency Computat.: Pract. Exper. 2006; **18**:263–269
Published online 11 October 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/cpe.916

A Problem Solving Environment for interactive modelling of multiway data



Ivo H. M. van Stokkum*,† and Henri E. Bal

Faculty of Sciences, Vrije Universiteit, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands

SUMMARY

A prototype Problem Solving Environment (PSE) is presented for problems in interactive modelling of multiway data. Multiway data result from measurements as a function of two or more independent variables. The PSE comprises a parameter estimation loop and a model adjustment loop. The model can be specified hierarchically using mathematically described building blocks which encapsulate the model assumptions. A typical case study of three-way data illustrates the need for interactive model adjustment. Requirements for interactive problem solving are discussed. Copyright © 2005 John Wiley & Sons, Ltd.

KEY WORDS: Problem Solving Environment; interactive modelling; multiway data; parameter estimation; spectrotemporal

1. INTRODUCTION

A Problem Solving Environment (PSE) is a computer system that provides all the computational facilities necessary to solve a target analysis class of problems [1,2]. In this paper a class of problems will be described which necessitates interactive modelling of multiway data, and a prototype PSE will be presented. The PSE for this application should be suited for collaborative research, enabling distributed interactive modelling, where an expert in modelling in one place can collaborate with a scientist in another place who is an expert in the experimental data and system under study. Furthermore, compute intensive interactive modelling studies will require parallel systems. Studying a complex system quantitatively one can distinguish two problems: finding the proper model to describe the experimental data, and when such a model is available, estimating the parameters of scientific interest. These two problems are illustrated schematically with the terminology in Figure 1 and key concepts and some pictures in Figure 2 (details of these figures will be discussed below).

*Correspondence to: Ivo H. M. van Stokkum, Faculty of Sciences, Vrije Universiteit, De Boelelaan 1081, 1081 HV, Amsterdam, The Netherlands.
†E-mail: ivo@nat.vu.nl

Received 22 July 2003
Revised 9 February 2004
Accepted 3 May 2004

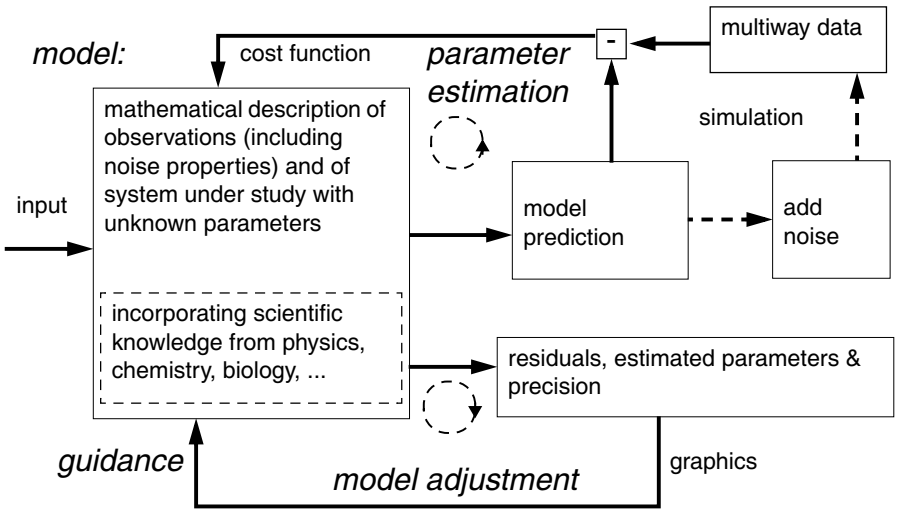


Figure 1. Flow chart of prototype PSE. Using the multiway data and the specified model the parameters are estimated in the upper loop. The bottom model adjustment loop is traversed many times, which is the main motivation for the interactivity of the PSE. Further explanation in text.

At the highest level two loops can be distinguished. In the top loop it is depicted how a model can be used to describe data. This is known as parameter estimation, or model fitting, or regression. In the bottom loop it is indicated that when there are many candidate models available, they should be tested, and using scientific knowledge a choice should be made. It is this model adjustment loop for which the PSE described below is needed.

2. STATEMENT OF A TYPICAL MODELLING PROBLEM

Multiway data result from measurements across multiple dimensions. An example is measurement of absorption or emission of light as a function of independent variables like time, wavelength and polarization. The multiway data can usually be well described by a superposition model. In the typical modelling problem outlined below $\sim 10^5$ datapoints will be described by $\sim 10^3$ parameters which have to be estimated from the data. The model is based upon physics and chemistry, and the estimated parameters should be meaningful. This provides an important criterion for judging the applicability of the model. The goal of the experiment is to identify the underlying system and estimate its unknown physicochemical parameters. For example, the spectroscopic properties of a mixture of components are a superposition of the spectroscopic properties of the components weighted by their concentration (see Equation (1)). With absorption this is known as the Beer–Lambert law [3]. Measurement of light absorption as a function of time t, wavelength λ and angle of polarization ϕ results in three-way data. At angle of polarization ϕ, the noise-free time resolved spectrum ψ_ϕ is a superposition of the

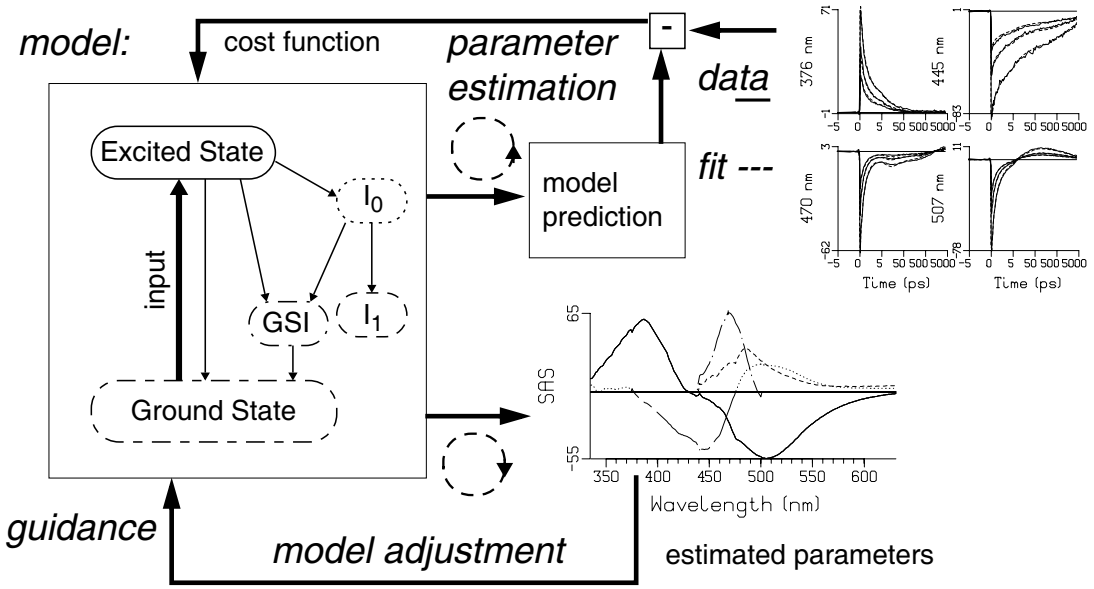


Figure 2. Possible GUI of prototype PSE showing the key concepts and some pictures from a typical case study with real three-way data. Data (solid lines) at four representative wavelengths (indicated by the vertical label) are depicted in the upper right-hand panel. Note that the time axis is linear from -5 to $+5$ ps relative to the IRF maximum, and logarithmic thereafter. Dashed lines represent the fit, calculated from the model prediction after traversing the parameter estimation loop. The compartmental scheme (left-hand panel) contains five different states. The model scheme was adjusted in the model adjustment loop in order to arrive at a satisfactory description, in particular realistic shapes of the estimated spectral parameters, the SAS (bottom right-hand panel). The linetypes of the SAS correspond to the linetypes of the boxes in the compartmental scheme. The units of the SAS and of the data are milliOD. Further explanation in text.

contributions of the different components:

$$\psi_\phi(t,\lambda) = \sum_{l=1}^{n_{\text{comp}}} c_l(t,\phi)\epsilon_l(\lambda) \tag{1}$$

where $c_l(t,\phi)$ and $\epsilon_l(\lambda)$ denote, respectively, the concentration and spectrum of component l. Note that according to Equation (1) a separability of time and wavelength properties is possible. Measurement of ψ poses the inverse problem: how can the spectroscopic and kinetic (dynamic) properties of the components be recovered? In a simple case the concentrations can be described by exponential decays $\exp(-k_l t)$ and the rate constants k_l and spectra $\epsilon_l(\lambda)$ are the parameters that have to be estimated from the data.

Table I. Hierarchical modelling of polarized absorption.

Level of modelling	Parametric description of
Linking of experiments	Relative scaling, linkage schemes
Contribution of component l, $c_l(t,\phi)\epsilon_l(\lambda)$	Spectrum of component l, $\epsilon_l(\lambda)$
Convolution of IRF and $a_l(t,\phi)c_l^{\mathrm{MA}}(t)$ resulting in $c_l(t,\phi)$	IRF, dispersion and excitation conditions
Dependence upon angle of polarization ϕ	Anisotropy decay $a_l(t,\phi)$
MA concentration $c_l^{\mathrm{MA}}(t)$ with δ-input	Compartmental scheme with microscopic rates

3. CHOSEN SOLUTION

In practice, various problems can arise: first of all the number of components present in the system is usually unknown. Secondly, in general, neither the concentration profiles $c_l(t,\phi)$ nor the spectra $\epsilon_l(\lambda)$ are known. However, in our case, knowledge is available in the form of a parameterized compartmental model [4] for $c_l(t,\phi)$, in which the dynamics of $c_l(t,\phi)$ are described by ordinary differential equations. Furthermore, the scientist usually has *a priori* knowledge about which shapes of spectra are realistic. This amounts to common statements regarding continuity, non-negativity, unimodality, etc. Implementing such *a priori* knowledge with the help of constraints on the spectral parameters $\epsilon_l(\lambda)$ is termed a spectral model. When no constraints are used the quality of the fit only depends upon the number of components used, and not upon the compartmental scheme. However, different schemes result in different estimated spectral parameters $\epsilon_l(\lambda)$ and the scientist must choose a model based upon the physicochemical plausibility of the parameters (rate constants and spectra). Thus, there are many candidate compartmental and spectral models available. According to the principle of parsimony (Ockham's razor) the preferred model should be as simple as possible. Consequently, the PSE should allow interactive hierarchical modelling to incorporate scientific knowledge and enable flexible testing of candidate models and hypotheses. In the model specification part of the PSE an overall model is constructed from building blocks (see Table I).

Suppose that we want to analyse simultaneously a set of N_ϕ time-resolved spectra $\psi_\phi(t,\lambda)$ measured at angle of polarization ϕ from the system of interest. On the top level of the hierarchy these ψ_ϕ are combined by introducing relevant scaling parameters for measurements done at the same wavelengths. On the bottom level of the hierarchy a model function is built for experiment ψ_ϕ, starting from the candidate compartmental model. First at magic angle (MA) where there is no polarization angle dependence, a concentration $c_l^{\mathrm{MA}}(t)$ is calculated for component l of the compartmental scheme, assuming a perturbation by a unit impulse $\delta(t)$. A simple example is the exponential decay $c_l^{\mathrm{MA}}(t) = \exp(-k_l t)$, which is fully described by the decay rate parameter k_l. On the next level an anisotropy decay function $a_l(t,\phi)$ is associated with component l in order to model the dependence upon the angle of polarization ϕ. A simple example is $a_l(t,\phi) = 1 + (3\cos^2(\phi) - 1)r_l$, which is fully described by the anisotropy parameter r_l. On the next level, this product function is convolved with the appropriate instrument response function (IRF), which takes into account the excitation conditions as well as the detector properties, and which limits the time resolution. This IRF usually depends upon the detection

wavelength, giving rise to a higher level of IRF description. Now we come to the level just below the top, where the specified concentration $c_l(t, \phi)$ is combined with a model for the spectrum of component l, $\epsilon_l(\lambda)$. This spectral model can be fairly simple, no constraints, or more specific, ranging from constraints (e.g. zero contribution at particular wavelengths) to a detailed analytical description when a candidate parametric spectral model is available. Linguistic constructs are used for the specification of constraints and of relations between spectra. For this combination of models for dynamics and spectra we have introduced the term spectrotemporal model [5]. At all levels of description parameters are introduced, which can be linked directly between different experiments, or indirectly by functional relations. Fitting parameters can be free, fixed, or subject to constraints. Thus many levels of indirection are discernible in the overall model.

A prerequisite for the model specification part of the PSE is that a language natural for the problem class can be used [6] to specify the building blocks of the model. In our case the natural language is the mathematical description of the model in Equation (1) and the hierarchy of models in Table I. Next to the hierarchy of model construction also a hierarchy of data fitting can be distinguished. Crucial in the nonlinear least squares fitting is the treatment of conditionally linear parameters ($\epsilon_l(\lambda)$ in Equation (1)) by a variable projection algorithm [7,8]. Appropriate weighting must be applied [9], often giving rise to iterative fitting procedures. Linguistic constructs are used to specify the weighting of observations. After convergence, exhaustive search methods [10,11] and profiling (constructing likelihood based confidence intervals [9]) can be applied (each requiring many minimizations) to check for uniqueness and precision of the parameter estimates. Appropriate graphics output is produced, to facilitate interactive data analysis by the human in the loop. Figure 1 illustrates the prototype PSE. Using the multiway data and the specified model the parameters are estimated in the upper loop. The estimation is based upon minimization of a cost function, e.g. the sum of squares of the weighted residuals. The residuals are the difference between the multiway data and the model prediction. Linguistic constructs are used to express both the model and the fitting process. Simulation can be used to check the model identifiability and estimability of parameters, and of course to test the software implementation. Finding a good model is an iterative process (the bottom loop), requiring interaction with the PSE, trying different model assumptions. The output, in particular the estimated kinetic parameters and a graphical representation of the residuals and of the estimated spectra $\epsilon_l(\lambda)$, is fed back to guide the user and suggest possible model improvements.

4. CASE STUDY THREE-WAY DATA

The purpose of Figure 2 is to illustrate a typical case study with real three-way data, and provide the reader with some numbers indicating the size of the problem. A pigment–protein complex was studied by time-resolved polarized difference absorption spectroscopy. The photophysics and photochemistry of this model system are discussed elsewhere [12,13]. Part of the data are depicted in the upper right-hand panel, with different absolute magnitudes corresponding to the different angles of polarization. The quality of the fit can be judged by the small differences between solid and dashed lines. In total, 240 wavelengths were measured at 100 time points and three polarization angles, thus comprising nearly 10^5 data points. The compartmental scheme (left-hand panel) contains five different states. The thick upward arrow represents the excitation from the ground state to an excited state intermediate (ESI). The thin arrows depict transitions between the states. Each transition is described by a rate

constant, and each state is characterized by its species associated spectrum (SAS, $\epsilon_l(\lambda)$ in Equation (1)). The estimated SASs are illustrated in the bottom right-hand panel, with different linetypes indicating the different states. The model is specified by the states and allowed transitions, the IRF, the anisotropies of the states, requiring in total $\sim$25 parameters. The SASs comprise 240 parameters for each of the seven states. The total number of spectral parameters is reduced by spectral equalities and constraints to $\sim 10^3$ free parameters. The linetypes of the SAS correspond to the linetypes of the boxes in the compartmental scheme. Crucial for the shape of the SAS (solid) of the ESI is the decay rate from the ESI state directly to the ground state. Since this rate cannot be estimated from the fit, it was adjusted iteratively in order to produce a satisfactory shape. Note that the negative part of the ESI SAS (solid) resembles the mirror image of the ground state SAS (long dashed). This illustrates that a rate parameter that does not influence the quality of fit of the data can be determined indirectly from the resulting SAS.

Typically parameter estimation by nonlinear least squares requires $\sim 10^2$ s on a workstation (IBM Power3 II, 375 MHz), when using the variable projection algorithm to eliminate the conditionally linear parameters (SAS, $\epsilon_l(\lambda)$ in Equation (1)). This time is needed for the upper loop in Figure 1. However, the number of model adjustments, the bottom loop in Figure 1, is routinely about 10, and with difficult problems it can easily go up to $\sim 10^2$ or $\sim 10^3$. This is the main motivation for the interactivity of the PSE.

5. FUTURE DIRECTIONS

Lacking in the prototype PSE is a graphical user interface (GUI). Figure 2 shows an example of how a GUI could look. The process of problem solving is visualized, allowing us to zoom in on all steps. Visualization of data, model, and fit results come naturally. Ideally the GUI should support collaborative research, enabling distributed interactive modelling, where an expert in modelling and parameter estimation can analyse the data and the experimental scientist can contribute to the interactive modelling by discarding unrealistic models and suggesting model improvements. Currently, the typical compute time of the parameter estimation loop is $\sim 10^2$ s. To allow for true interactivity this loop needs to be accelerated $\sim 10^2$ times using a parallel system. It is a challenge to incorporate modelling knowledge into the PSE which can provide guidance to the user and help to reduce the number of model adjustments.

ACKNOWLEDGEMENTS

The data for the case study have kindly been provided by Delmar S. Larsen and R. van Grondelle. Zsofia Derzsi is thanked for critical reading of the text.

REFERENCES

1. Gallopoulos E, Houstis EN, Rice JR. Workshop on problem-solving environments: Findings and recommendations. *ACM Computing Surveys* 1995; **27**(2):277–279.
2. Houstis EN, Rice JR, Gallopoulos E, Brambley R (eds.). *Enabling Technologies for Computational Science: Frameworks, Middleware and Environments*. Kluwer Academic: Dordrecht, 2000.

3. Cantor CR, Schimmel PR. *Biophysical Chemistry. Part II: Techniques for the Study of Biological Structure and Function*. Freeman: New York, 1980.
4. Godfrey K. *Compartmental Models and their Application*. Academic Press: London, 1983.
5. Van Stokkum IHM, Scherer T, Brouwer AM, Verhoeven JW. Conformational dynamics of flexibly and semirigidly bridged electron donor-acceptor systems as revealed by spectrotemporal parameterization of fluorescence. *Journal of Physical Chemistry* 1994; **98**:852–866.
6. Houstis EN, Rice JR. Future problem solving environments for computational science. *Mathematics and Computers in Simulation* 2000; **54**:243–257.
7. Golub GH, LeVeque RJ. Extensions and uses of the variable projection algorithm for solving nonlinear least squares problems. *Proceedings of the 1979 Army Numerical Analysis and Computation Conference. ARO Report* **79**(3):1–12.
8. Nagle JF. Solving complex photocycle kinetics. Theory and direct method. *Biophysical Journal* 1991; **59**:476–487.
9. Bates DM, Watts DG. *Nonlinear Regression and its Applications*. Wiley: New York, 1988.
10. Dioumaev AK. Evaluation of intrinsic chemical kinetics and transient product spectra from time-resolved spectroscopic data. *Biophysical Chemistry* 1997; **67**:1–25.
11. Roelofs TA, Lee CH, Holzwarth AR. Global target analysis of picosecond chlorophyll fluorescence kinetics from pea-chloroplasts: A new approach to the characterization of the primary processes in photosystem-II alpha-units and beta-units. *Biophysical Journal* 1992; **61**:1147–1163.
12. Larsen DS, Van Stokkum IHM, Vengris M, Van der Horst MA, De Weerd FL, Hellingwerf KJ, Van Grondelle R. Incoherent manipulation of the photoactive yellow protein photocycle with dispersed pump-dump-probe spectroscopy. *Biophysical Journal* 2004; **87**:1858–1872.
13. Van Stokkum IHM, Larsen DS, Van Grondelle R. Global and target analysis of time resolved spectra. *Biochimica Biophysica Acta* 2004; **1657**:82–104 (erratum: **1657**:262).

