20th European Symposium on Computer Aided Process Engineering – ESCAPE20
S. Pierucci and G. Buzzi Ferraris (Editors)

Hybrid semi-parametric modeling of biological systems: Application to spectroscopic data for the estimation of concentrations

von Stosch M.,a Oliveira R.,b Peres J.,a Feyo de Azevedo S.,a

a LEPAE, Departmento de Engenharia Quimica, Faculdade de Engenharia, Universidade do Porto, Rua Dr. Roberto Frias s/n, 4200–465 Porto, Portugal.
b REQUIMTE, Departamento de Quimica, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, 2829-516 Caparica, Portugal.

Abstract

In this work, bioprocess monitoring based on spectral data is improved when compared to commonly applied chemometric tools, by merging nonparametric modeling, biological and process *a priori* knowledge into a hybrid semi-parametric model. This particular semi-parametric structure comprises a nonparametric submodel inspired by a NPLS structure, as NPLS has been reported to be successful for dealing with massive numbers of highly correlated spectral data. The method was applied to Bordetella pertussis cultivations equipped with a Near-InfraRed (NIR) probe, showing that estimates of metabolite concentrations are improved when compared to those obtained through classical chemometric modeling, as expressed by lower mean square errors, better calibration properties and a higher statistical confidence.
Keywords: Dynamic modeling, Hybrid modeling, NIR, dynamic nonlinear PLS

1.Introduction

Many biopharmaceutical industries are implementing the new Process Analytic Technology (PAT) guidelines. The first steps therein aim at a better on-line characterization of the process state by implementing advanced monitoring techniques. Spectroscopy techniques such as Near-InfraRed, mid-InfraRed, Terahertz, Raman InfraRed, Fourier-Transform InfraRed or Fluorescence spectroscopy have been widely reported in the context of PAT since they are fast, non invasive, non destructive, and amenable to chemically complex multiphase reaction media. The availability of such analytical devices provides extensive spectral data sets holding complex molecular scale information. To date, chemometric modeling tools such as Principal Component Analysis (PCA), Partial Least Square(PLS) and its nonlinear counterparts (nonlinear-PLS and Kernel-PLS) are applied to deconvolve the complex spectra and to correlate with target state variables. These chemometric techniques present, however, the limitation that they do not incorporate a priori knowledge about the target biological system, where in contrast it is envisaged that such information rich spectral data will in future be integrated with Systems Biology models and macroscopic process operation (Teixeira et al, 2007b).

The integration of spectral data and fundamental biological models is hindered by the fact that such data has no direct physical meaning. Hybrid semi-parametric systems can provide the ideal mathematical framework for bringing together biological and process mechanisms along with the data from such analytical devices. In general the knowledge

can be arranged in parallel or serial, the latter being particularly suitable for complex systems for which large data sets are available without direct physical interpretation (Teixeira et al. 2007b). Biological constraints such as metabolic reactions connectivity in the form of elementary modes (Teixeira et al., 2007a) can also be included in the hybrid structures, paving the way for on-line characterization of the fluxome.
In traditional hybrid models usually Artificial Neural Networks (ANN) found application (Oliveira 2004, Thomson & Kramer 1994, Psichogios & Ungar 1992). However when high-dimensional data are considered as inputs to the ANNs then the number of ANN parameters quickly exceeds the number of measured target state variables, which leads to a clearly underdetermined system of equations. On the other hand (N)PLS models cannot be directly integrated into the serial hybrid structure as for the training of such the kinetic rates would have to be known, and their estimation from noisy and sparse data is prone to errors (e.g., Oliveira 2004). In this article a NPLS like structure is integrated into a serial hybrid model structure and an algorithm for its parameter identification is developed. For the evaluation against a dynamic PLS model, the hybrid structure is applied to experimental data, namely NIR and concentration of target metabolites of a *Bordetella pertussis* cultivation.

2.The semi-parametric hybrid model

The framework for the serial semi-parametric hybrid model structure are the reactor material balances

$$\frac{dc}{dt} = f = r(L_x, w_A) - D \cdot c + u \ , \qquad (1)$$

where c is the vector of concentrations, r is the vector of kinetic rate functions which depend on the inputs L_x and the parameters w_A, D is the dilution rate and u is the vector of volumetric control inputs. The vector of the kinetic rate functions, r, is the representative of the biological system, which is defined by the following semi-parametric model

$$r(L_x, w_A) = K \cdot \left\langle \phi_j(c) \times \rho_j(L_x, w_A) \right\rangle_{j=1,\ldots,m} \ , \qquad (2)$$

with K being a $n \times m$ matrix of yield coefficients, ϕ being m known kinetic functions and where $\rho(L_x, w_A)$ are unknown kinetic functions which are obtained from a nonparametric model comprising L_x and w_A. Many times these nonparametric models are ANNs (see Oliveira 2004, Thomson & Kramer 1994, Psichogios & Ungar 1992) but these cannot be applied along with spectral data as mentioned above. Due to the outstanding features of the (N)PLS for high-dimensions of redundant inputs, a structure equivalent to a NPLS structure which moreover restores the NPLS features is adopted here. Further a suitable parameter identification algorithm funding on the sensitivities approach is proposed, avoiding the estimation of the kinetic rates from sparse and noisy measurements. The structure which is proposed consists of o independent submodels such that

$$\rho_{1..m}(L_x, w_A) = \sum_{i=1}^{o} \rho_{1..m,i}(L_{i,1..k}, w_A) \ . \qquad (3)$$

Each submodel $\rho_{1..m,i}(L_{i,1..k}, w_A)$ further can be divided into two parts, an outer and an inner model. The outer model linearly compresses the high number of dimensions of the

inputs and outputs by the use of input loadings, $W_{x,i}$, and output loadings, $W_{y,i}$, to one inner and one outer latent variable, respectively. The inner model then links (non)linearly the input latent variable with the output latent variable, for details see (Qin & McAvoy 1992, Baffi et al 2000). In this study the inner models are chosen to be ANNs, which proved to be successful in Baffi et al (2000). The complete nonparametric model can then be written as

$$\rho_{1..m}(L_x, w_A) = \sum_{i=1}^{o} W_{y,i} \cdot \left(w_{2,i} \cdot g(w_{1,i} \cdot h(W_{x,i} \cdot L_{i,1..k}) + b_{1,i}) + b_{2,i} \right) , \quad (4)$$

where the ANN of submodel i comprises the weights, $w_{1,i}$ and $w_{2,i}$, the biases, $b_{1,i}$ and $b_{2,i}$, and the transfer functions, h and g, which are linear and tangential, respectively. The vector of inputs $L_{i,1..k}$ with dimensions $1..k$ is for $i=1$ identical to L_x and can comprise the model estimates of concentrations and/or additional experimental data, such as spectral data. When $i>1$ then, equivalently to the PLS models the input is calculated as

$$L_{i,1..k} = L_{i-1,1..k} - W_{x,i-1} \cdot L_{i-1,1..k} \cdot W_{x,i-1} , \quad (5)$$

such that information which is gathered in prior input latent variables is not processed again, avoiding redundancy of the latent variables. While the structure of the nonparametric model is the prerequisite, the success of the model depends on the parameter identification. The following approach restores the idea of the NIPALS algorithm by application of a twofold objective function. On the one hand this objective consists of the minimization of a least square error of the residual in the concentrations

$$\min_{w_A} \left\{ E_1 = \frac{1}{P \times n} \sum_{l=1}^{P} \sum_{j=1}^{n} \frac{\left(c_{experimental,j}(t) - c_j(t, w_A) \right)^2}{c_{max,j}} \right\} , \quad (6)$$

where P is the number of time events, $c_{experimental,1..n}$ are the experimentally measured concentrations and $c_{max,1..n}$ are the respective standard variances, which account for statistic properties of the data.
On the other hand the objective consists of the maximization of the captured variance of the inputs to the model, which is analogous to minimizing the least square error of

$$\min_{w_A} \left\{ E_2 = \frac{1}{o \times k} \sum_{i=1}^{o} \left(W_{x,i,lin} - W_{x,i} \right)^2 \right\} . \quad (7)$$

The first term in Eq. (7), $W_{x,i,lin}$, is the vector norm of $W_{x,i,lin,un}$, which is calculated as following

$$W_{x,i,lin,un} = \frac{L_{i,1..k} \cdot t_i}{t_i^T \cdot t_i} . \quad (8)$$

Therein the input scores t_i are obtained from the inputs times the input loadings, i.e.

$$t_i = W_{x,i} \cdot L_{i,1..k} . \quad (9)$$

This NIPALS inspired calculation of $W_{x,i,lin}$ restores some basic features of the PLS, such as independence of the latent variables and minimization of redundant information in the latent variables.

The parameter identification is accomplished by application of the sensitivities equation. These sensitivities equations are not presented here for the sake of briefnesses, but can be derived by building the derivative after all parameters w_A, where w_A comprises all input loadings, all output loadings and all inner model ANN parameters. However for the identification of the input and output loadings their normalization to unit length, which is a condition arising from PLS, needs to be taken into account. In any case, the sensitivity equations need to be integrated along with the reactor material balances, i.e. Eq. (1). In general the MATLAB® integration routines could be applied but they are rather time expensive when compared to a linear, Euler integration approximation schema, which therefor found application.

The two least square objectives, namely E_1 and E_2 are sought to be simultaneously minimized using the "lsqnonlin" MATLAB® function which uses a subspace trust region method and is based on the interior-reflective Newton method (MATLAB® Optimization toolbox). Depending on the inputs L_x a deficiency of the presented twofold objective might come into account, namely when model estimates are inputs to the nonparametric model then w_A might be optimized rather towards the second objective than the first. In order to circumvent this deficiency the parameter identification is carried out until the "best" parameters are obtained, which is accomplished as described below, and then a further parameter identification is carried out, in which only a subset of the parameters, namely all ANN parameters and all output loadings, along with only the first objective is applied.

Identification of the best parameters bares two well known challenges. One is due to the fact that gradient based nonlinear optimization does not necessarily identify the global minimum but rather a local minimum of the objective and that the optimization might get stuck there. To overcome this challenge in this study, as done by several other authors, Oliveira (2004), Peres et al. (2008), Teixeira et al. (2007a), at least four runs of the optimization with random initial values of all parameters are performed, where the consistency of the obtained solutions is taken as a measure whether or not to perform further runs. The other challenge is know as "over-training" of the nonparametric model and is due to the fact that after a certain threshold the optimization rather results in modeling the noise of the data then to further identify the underlying function. This challenge is usually overcome by the application of two data sets, one containing about 2/3 of the data for training the model, and the other containing 1/3 of the data referred to as validation data set. Then parameter identification is carried out on the training set and is then stopped when the residual of the validation set is the smallest.

The presented nonparametric structure admits two structural changes, when the number of inputs and outputs are fixed. One change can be made to the number of nodes used in the hidden layer of the ANNs, but in this study is fixed to be one because reported in Qin & McAvoy (1992) and Baffi et al (2000) to have only little influence on the quality of estimates. The other change that can be made is regarding the number of latent variables. These number of latent variables is however not like in PLS or NPLS consecutive increased till a certain amount of variance is captured but instead fixed at the beginning of the parameter identification. Therefore hybrid model structures with different numbers of latent variables are compared to each other in order to find the most "suitable" structure. For the comparison of the structures several facts need to be taken into account such as the model residual, the number of model parameters and the number of data the residual is build on. A criteria reported to be suitable for these model

comparisons is the Bayesian Information Criteria (BIC), see Peres et al. (2008). In the sense of the BIC the model is the most suitable which exhibits the greatest BIC value.

3. Application to experimental data of a *Bordetella pertussis* cultivation

Cultivations of *Bordetella pertussis* are used for the production of a vaccine against whooping cough. The key for controlling the process is the online knowledge about biomass and the specific growth rate, Soons et al. (2008a). Such online knowledge can in principal be derived from the online NIR measurements applying a "suitable" model.

3.1.The process

The experimental data of *Bordetella pertussis* which find application in this study are the ones collected, described and used in Soons et al (2008a) and Soons et al (2008b). The processes were run in batch mode and variations to the process conditions, such as in pH, Temperature and dissolved oxygen, were made as reported in Soons et al (2008a). The recorded NIR data were pretreated as in Soons et al (2008a) by the application of a Savtisky-Golay smoothing with a 45-point window and a second-order polynomial in order to reduce noise and then shifted to zero-mean and scaled by the variance. Measurements of the concentrations of lactate, glutamate and biomass over time for eight batches were recorded Soons et al (2008b). However due to uncertainties in the measurements of the substrates in one of the batches, only the remaining batch data were used, i.e. split into a 5 batches comprising training data set and a 2 batches comprising validation set.

3.2.The Bordetella pertussis hybrid model

The hybrid model in this case contains mechanistic knowledge about the process, which was reported in Soons et al (2008b). The system of model equations reads,

$$\frac{d}{dt}\begin{bmatrix} Lac \\ Glu \\ X \end{bmatrix} = \begin{bmatrix} Lac \cdot X & 0 & 0 \\ 0 & Glu \cdot X & 0 \\ 0 & 0 & X \end{bmatrix} \cdot \begin{bmatrix} r_{Lac} \\ r_{Glu} \\ \mu \end{bmatrix} - D \cdot \begin{bmatrix} Lac \\ Glu \\ X \end{bmatrix} \qquad (10)$$

where Lac, Glu and X are the concentrations of lactate, glutamate and biomass, respectively and r_{Lac}, r_{Glu} and μ are the respective unknown kinetic functions which are obtained by the nonparametric model. The inputs L_x to the nonparametric model in this study contain the estimates of all concentrations, the pretreated NIR data and measured data of pH, Temperature and the percentage of dissolved oxygen. For the hybrid NPLS model the number of latent variables is varied in order to identify the best model structure and in addition for the (N)PLS models the inputs are varied, by the means of number and kind in the sense of AutoRegressive eXogenous (ARX) models.

3.3.Results &Discussion

The best obtained model performance criteria over model methodologies and structural parameters are shown in Tab.1. It therein can be seen that the hybrid model outperforms by far, in the order of one magnitude, the traditional dynamic ARX-PLS for all BIC as for all MSE values. These results are consistent with graphical observations and in-line with the expectations formulated in the introduction. It can be further noticed that the best identified hybrid model structure consists of only 2 latent variables which thus offers a drastically smaller number of involved parameters when compared to the best

identified ARX-PLS which comprises 3 latent variables and there due exhibit higher statistical confidence reflected in the higher BIC values shown in Tab. 1.

Table 1. Model performance criteria, namely the Bayesian Information Criteria (BIC) & the Mean Square Error (MSE) for training & validation sets over model methodologies & structural parameters. Lv stands for the number of latent variables & nt for the number of time series.

Model	Structure	BIC train	MSE train	BIC valid	MSE valid
ARX-PLS	[nt =1, lv =3]	-11615	0.2483	-8418	0.7810
Hybrid model	[lv =2]	-326	0.0991	-83	0.0753

4. Conclusions

The proposed semi-parametric hybrid model when applied to NIR data of *Bordetella pertussis* cultivations clearly improves the estimation of concentrations when compared to those obtained through classical chemometric modeling, namely ARX-PLS, as expressed by the mean square error, which is one order of magnitude lower for the hybrid model, and higher statistical confidence of the estimates. Further, as a result of applying this hybrid structure, the trajectories of the estimated fluxes are directly accessible on-line, opening the possibility for on-line metabolic flux control.

5.Acknowledgment

Sincere thanks for the provided data go to the Netherlands Vaccine Institute and for financial support to the Fundação para a Ciência e a Tecnologia, where the reference number of the provided scholarship is: SFRH / BD / 36990 / 2007.

References

Baffi G., Martin E.B. and Morris A.J., (2000), Non-linear dynamic projection to latent structures modelling, Chemom. Intell. Lab. Syst., 52, 5–22

Oliveira R. (2004), Combining first principles modelling and artificial neural networks: a general framework, Computers & Chemical Engineering, 28, 755-766

Peres J, Oliveira R, Feyo de Azevedo S., (2008), Bioprocess hybrid parametric/nonparametric modelling based on the concept of mixture of experts, Biochemical Eng. J., 39, 190-206

Qin S.J. and McAvoy T.J., (1992), Non-linear PLS modelling using neural networks, Computers & Chemical Engineering, 23, 395–411

Psichogios D.D. Ungar L.H., (1992), Comparison of four neural net learning methods for dynamic system identification, IEEE Transactions on Neural Networks, 3, 1, 122-130

Soons Z.I.T.A., Streefland M., van Straten G., van Boxtel A.J.B., (2008a), Assessment of near infrared and "software sensor" for biomass monitoring and control, Chemometrics and Laboratory Systems, 94, 2, 166-174

Soons Z.I.T.A., Shi J,, Stigter J,D,, van der Pol L.A., van Straten G., van Boxtel A.J.B., (2008b), Observer design and tuning for biomass growth and k(L)a using online and offline measurements, J. of Processcontrol, 18, 7-8, 621-631

Teixeira A.P., Alves C., Alves P.M., Carrondo M.J.T., Oliveira R., (2007a), Hybrid elementary flux analysis/nonparametric modelling: application for bioprocess control, BMC Bioinformatics, 8, 30

Teixeira A.P., Carinha N., Dias J.M.L., Cruz P., Alves P.M., Carrondo M.J.T., Oliveira R., (2007b), Hybrid semi-parametric mathematical systems: Bridging the gap between systems biology and process engineering, J. of Biotechnology, 132, 418-425

Thompson M.L., Kramer M.A., (1994), Modelling chemical processes using prior knowledge and neural networks, AIChE Journal, 40,8, 1328-1340