QUALIFYING AND SEGMENTATION OF HISTORICAL PROCESS DATA USING OPTIMAL EXPERIMENT DESIGN TECHNIQUES FOR SUPPORTING PARAMETER ESTIMATION

László DOBOS, Zoltán BANKÓ, János ABONYI
Department of Process Engineering, University of Pannonia,
Egyetem Street 10, 8200 Veszprém, Hungary, tel.: +36 88 624 770,
e-mail: dobosl@fmt.uni-pannon.hu, bankoz@tvn.hu, abonyij@fmt.uni-pannon.hu

ABSTRACT

With the wide-spread application of process models and simulators, estimation of model parameters becomes a crucial project. In chemical industry the processes are mostly highly non-linear which makes the identification of model parameters difficult. In the practice the process simulators are not just for design but optimization of operating plants in numerous cases various sets of process data are available to determine the necessary model parameters. With further examination of the historical process data, a new possibility becomes applicable: some time-series segments can provide more information about the estimated model parameters than other parts of the recorded time-series. Since the tools of Optimal Experiment Design (OED) are for maximizing the information content of the experiments regarding to the unknown model parameters, the applicability of these tools in qualifying the recorded process data is obvious. In this paper the connection of classical time-series segmentation and OED tools will be examined throughout a simple polymerization example to prove the efficiency of integration of these tools to support the parameter identification of process models.

Keywords: *Optimal Experiment Design, Parameter Identification, Segmentation, Time Series*

1. INTRODUCTION

Process models play important role in computer aided process engineering since most of advanced process monitoring, control, and optimization algorithms rely on the process model. Unfortunately, often some of the parameters of these models are not known a priori, so they must be estimated from experimental data. The accuracy of these parameters largely depends on the information content of the experimental data presented to the parameter identification algorithm [1].

Optimal Experiment Design (OED) can maximize the confidence on model parameters through optimization of the input profile of the system. For parameter identification of different dynamic systems and models, this approach has been already utilized in several studies [2–6]. OED uses an iterative algorithm where the optimal conditions of the experiments or the optimal input of the system depends on the current model, which parameters were estimated based on the result of the previously designed experiment. Consequently, experiment design and parameter estimation are solved iteratively, and both of them are based on nonlinear optimization of cost functions.

That means in practice, the applied nonlinear optimization algorithms have great influence on the whole procedure of OED, because design of experiments for nonlinear dynamical models is a difficult task. This problem is usually solved by several gradient-based methods e.g. nonlinear least squares method or sequential quadratic programming. A review of these methods can be found in [7], while [8] describes the extended maximum likelihood theory for optimizing the experiment conditions.

In this paper, the problem of creating identification support algorithm is investigated. We present a new and intuitive segmentation based method, which makes possible to identify each parameter in the most appropriate time frame of the experimental data. With the help of this method, it becomes possible to reduce the number and cost of experiments and at the same time reduce the time consumption of parameter estimation. It may be caused by the fact that a considered time segment is useless in a certain point of view, but from another aspect the same time series segment can be applicable to determine other parameters.

The rest of the paper is organized as follows. In Section 2, the previous works related to OED are reviewed. Section 3 and Section 4 present the theoretical background of our work, i.e. the applied segmentation method and the combination of classical OED and segmentation techniques, while Section 5 conducts our approach through a case study. Finally, we present our conclusions and suggestions for future work.

2. CLASSICAL OPTIMAL EXPERIMENT DESIGN

The case study considered in this paper belongs to the following general class of process models:

$$\frac{dx(t)}{dt} = f(x(t), u(t), p) \tag{1}$$

$$y(t) = g(x(t)), \tag{2}$$

where u is the vector of the manipulated inputs, y is the vector of the output, x represents the state of the system and p denotes the model parameters. The p parameters are unknown and should be estimated using the data taken from experiments. The estimation of these parameters is based on the minimization of the square error between the output of the system and the output of the model:

$$\min_{p}\left[J_{mse}(u(t),p) = \frac{1}{t_{\exp}} \int_{t=0}^{t_{\exp}} (e^{T}(t)\cdot Q(t)\cdot e(t))dt \right] \quad (3)$$

$$e(t) = \tilde{y}(u(t)) - y(u(t),p)\,, \quad (4)$$

where $\tilde{y}(u(t))$ is the output of the system for a certain u(t) input profile, and $y(u(t))$ is the output of the model for the same $u(t)$ input profile with p parameters. Q is a user supplied square weighting matrix that represents the variance measurement error. The basic element of the experiment design methodology is the Fisher information matrix F, which combines information on the output measurement error and the sensitivity of the model outputs y with respect to the model parameters:

$$F(p^0,u(t)) = \frac{1}{t_{\exp}} \int_{t=0}^{t_{\exp}} \left(\frac{\partial y}{\partial p}(u(t),p)_{p=p^0}\right)^T \cdot Q(t) \cdot \left(\frac{\partial y}{\partial p}(u(t),p)_{p=p^0}\right) dt \quad (5)$$

The sensitivities are calculated based on the partial derivatives of the model parameters. As the true parameters p^* are unknown during experiment design, the derivatives are calculated near to the so-called nominal parameters p^0, which can be given by some initial guess, extracted from literature or estimated from the previous experiments. The optimal design criterion aims the minimization of a scalar function of the F matrix. Several optimal criterion exist, we present D-optimal and E-optimal criterion suggested by Bernaerts et al. [1]:

$$J_D = \min_{u(t)}(\det(F)) \quad (6)$$

$$J_E = \min_{u(t)}\left(\frac{\lambda_{\max}}{\lambda_{\min}}\right) \quad (7)$$

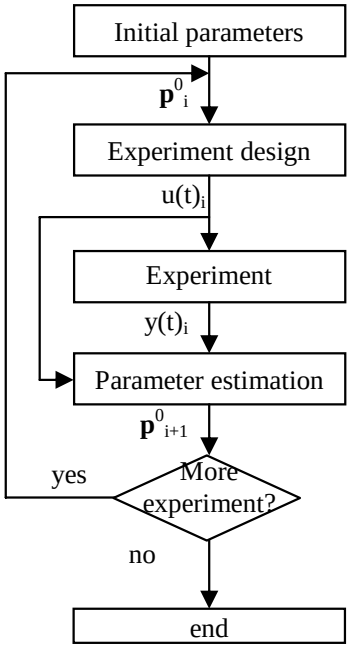

Fig. 1 Classical scheme of optimal experiment design

If the p^0 nominal parameters are far from the p^* true parameters, convergence cannot be guaranteed after the first optimal design. Thus, an iterative design scheme is needed to obtain convergence from p^0 to p^* (Fig. 1).

Both the parameter estimation and the experiment design steps of this iterative scheme represent a complex nonlinear optimization problem, hence the effectiveness of the applied optimization algorithms have great influence on the performance of the whole procedure. The classical solution is to use nonlinear least squares (NLS) algorithm for parameter estimation (3), and sequential quadratic programming (SQP) for the experiment design (7).

3. TIME SERIES SEGMENTATION

Real-life time-series can be taken from business, physical, social and behavioral science, economics, engineering etc. Depending on the application, the goal of the segmentation of a time-series is to locate stable periods of time, to identify change points, or to simply compress the original time-series into a more compact representation.

A univariate, m-element time series, $x = [x(1),x(2),\ldots,x(m)]$, is a column vector, where $x(i)$ is the ith element. The ith segment of x is a set of consecutive time points, $S_i(a,b) = [x(a), x(a+1),\ldots, x(b)]$, while the c-segmentation of x is a partition of x to c non-overlapping segments, $S^c_x = [S_1(1,a), S_2(a+1,b), \ldots, S_c(k+1,m)]$. In other words, a c-segmentation splits x to c disjoint time intervals, where $1 \le a$ and $k \le m$.

The simplest but yet powerful segmentation technique for univariate time series is Piecewise Aggregate Approximation (PAA) algorithm. In this case, to reduce the m-length data from N, the time series are simply divided into N similar sized frames and each frame is represented by its mean value. Assuming that N is a factor of m, we get:

$$\underline{x}(i) = \frac{N}{m} \sum_{j=\frac{m}{N}(i-1)+1}^{\frac{m}{N}i} x(j)\,, \quad (8)$$

where $x(i)$ represents the ith PAA segment of x. Please note, although PAA is not the most sophisticated segmentation method it is perfectly suits for our case study as it can be seen in Section 5.

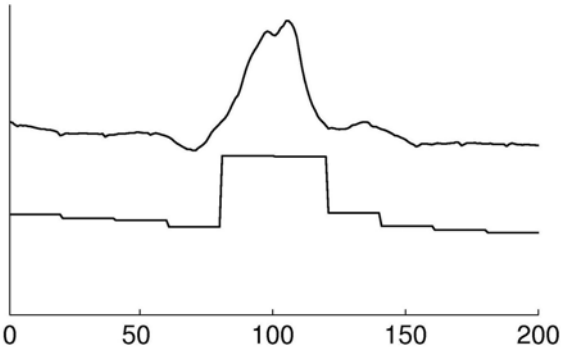

Fig. 2 The original signal (top) and its PAA representation (bottom) using 10 segments

4. THE APPLIED ALGORITHM USING TIME SERIES SEGMENTATION INTEGRATED WITH OED TOOLS

The aim of our work is basically to determine those time series segments recorded during the operation of the plant that are appropriate for estimation of model parameters. This way it becomes possible to reduce the number of the necessary experiments for parameter estimation by substituting the appropriate, previously recorded historical process data segments. In some simple cases it is sufficient to handle and examine the recorded data sets separately during parameter identification thus the application of the algorithm introduced below can provide sufficient result. In case of complex identification problems the proposed method may provide insufficient results that is why more advanced methods e. g. based on dynamic principle component analysis shall be applied.

In Section 2 the tools of Optimal Experiment Design were introduced. As the first step of our algorithm with the help of *E* and *D* criteria (6 - 7) it becomes possible to qualify numerically the information content of an input signal applied during the operation of the process respect to the estimated parameter. Both criteria are based on the Fischer information matrix which contains the sensitivity of the outputs respect to estimated model parameters. Since this information matrix is valid only a certain time interval ($t = [0\ t_{exp}]$) it is possible to calculate the *F* matrix – and throughout this the *E* or *D* criteria – in a moving time window with fixed length of time. This way it is possible to record an "evaluating time-series" which characterizes the information content of the examined operational time-series respect to the chosen parameter.

As a second step the segmentation of the "evaluating time-series" shall be done. This way it is easy to segregate the time-series segments with different information content. As the third step by choosing the time-series segment with the lowest value of *E* or *D* criteria the most appropriate time series segment is chosen for determine the value of the certain model parameters (*p*).

5. APPLICATION EXAMPLE

5.1. Process description

As an application example of the previously presented algorithm polymerization reactor

The reactor what have been studied is a SISO (single input-single output) process, a continuously stirred tank reactor (CSTR) where a free radical polymerization reaction of methyl-metacrylate is considered using azobisisobutironitil (AIBN) as initiator, and toulene as solvent The aim of the process is to produce different kinds of product grades. The number-average molecular weight is used for qualifying the product and process state, and it can be influenced by the inlet initiator flow rate. When this assumption is considered, and the effect of the temperature is neglected, the multi input-multi output model could be reduced to a SISO process. Because of the isothermal assumption, a four-state model can be obtained [9].

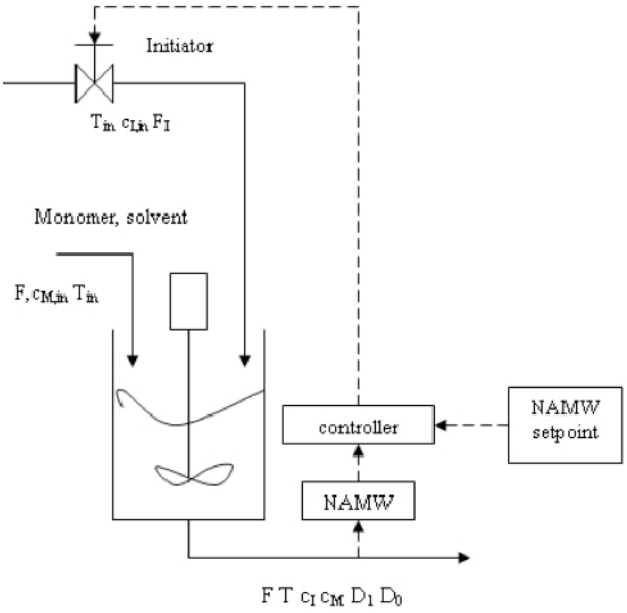

Fig. 3 The configuration of the SISO process

$$\frac{dC_m}{dt} = -(k_p + k_{fm})C_m P_0 + \frac{F(C_{m,in} - C_m)}{V} \tag{9}$$

$$\frac{dC_I}{dt} = -k_I C_I + \frac{F_I C_{I,in} - FC_I}{V} \tag{10}$$

$$\frac{dD_0}{dt} = (0.5k_{Tc} + k_{Td})P_0^2 + k_{f_m} C_m P_0 - \frac{FD_0}{V} \tag{11}$$

$$\frac{dD_1}{dt} = M_m(k_p + k_{fm})P_0 C_m - \frac{FD_1}{V} \tag{12}$$

$$y = \frac{D_1}{D_0}, \tag{13}$$

where:

C_m – concentration of the monomer in the reactor
$C_{m,in}$ – monomer concentration in feed
C_I – initiator concentration in the reactor
$C_{I,in}$ – initiator concentration in feed
k_p, k_{fm}, k_I, k_{Tc}, k_T – kinetic parameters and

$$P_0 = \left[\frac{2f \cdot k_I C_I}{k_{Td} + k_{Tc}}\right]^{0.5} \tag{14}$$

D_0 is the zero order moment of the chain length distribution of the inactive polymer chain, which represents the length of inactive chains. D_1 is the first order moment of inactive polymer chains, which means the distribution of molecular weight of inactive chains. The number-average molecular weight, represented by *y*, cannot be measured, but it is calculated, as can be seen in (13).

5.2. Example for using OED tools and segmentation

In this following the combination of OED tools and time series segmentation has been presented for support parameter identification through a case study of the previously presented polymerization reactor.

The model of the reactor is used as the operating plant and at the same the model also represents the process model that needs some of its parameters to be identified. Imagine that k_p and k_i kinetic parameters are not known properly and previously an experiment was carried out to determine the parameters.

Expression (7) was applied as the basis of extracting more information from these time series. It means that lower value the cost function E has, indirectly the considered time series segment is more and more appropriate for identification purposes. Directly the value of the E criteria can express the potential information content of the examined input signal segment regarded to the considered parameters. That is why important to examine the value of E optimal criteria as function of time over the period of the experiment.

Performing the presented PAA method for segmentation of time series of E – and indirectly throughout this the original experimental time series also – we have the possibility to separate the useful time series segments from the time series segments with less information content. The result of the segmentation is shown by Fig. 4.

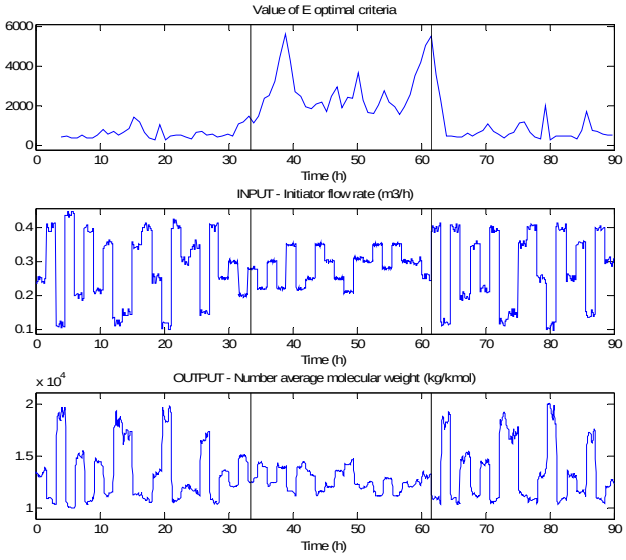

Fig. 4 Segmentation of time series of E criteria and the segments of the experimental time series

As it can be seen, the experiment can be divided into 3 parts. The first and the last segments have lower E value than the middle one. This means that input signal of these segments have potentially more information content that the middle segment possess. That is why middle segment can be neglected during the process of parameter identification.

In Fig. 5 the result of the identification is shown. During identification, the first and last time series segments were applied. The parameter fitting for the model was pretty successful since the output of the model is equal to the experimental data.

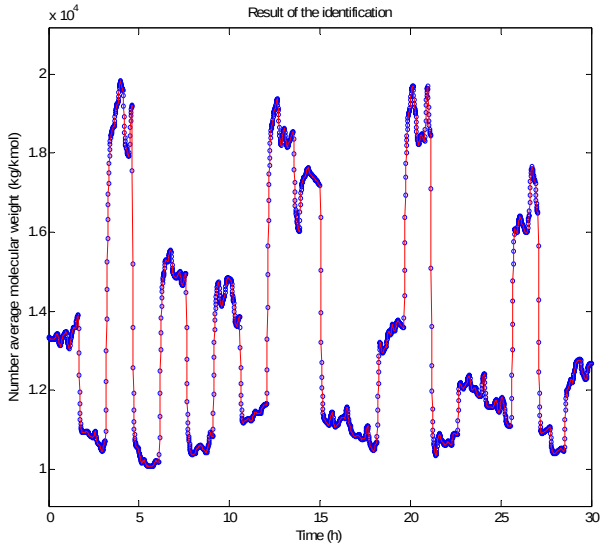

Fig. 5 Result of the identification (circles – experimental data, full line – model output)

ACKNOWLEDGMENTS

The financial support from the TAMOP-4.2.2-08/1/2008-0018 (Livable environment and healthier people – Bioinnovation and Green Technology research at the University of Pannonia, MK/2) project is gratefully acknowledged.

REFERENCES

[1] BERNAERTS, K. – SERVAES, R. D. – KOOYMAN, S. – VERSYCK, K. J. – VAN IMPE, J.F.: Optimal temperature design for estimation of the Square Root model parameters: parameter accuracy and model validity restrictions, Int. Jour. Of Food Microbiology, Volume 73, 2002, pp. 145-157.

[2] BERNAERTS, K. – VAN IMPE, J. F.: Optimal dynamic experiment design for estimation of microbial growth kinetics at sub-optimal temperatures: Modes of implementation, Simulation Modelling Practice and Theory, Volume 13, 2005, pp. 129-138.

[3] VERSYCK, K. J. – BERNAERTS, K. – GEERAERD, A. H. – VAN IMPE, J. F.: Introducing optimal experimental design in predictive modeling: A motivating example, International Journal of Food Microbiology, Volume 51, 1999, pp. 39-51.

[4] BERNAERTS, K. – GYSEMANS, K. P. M. – MINH, T. N. – VAN IMPE, J. F.: Optimal experiment design for cardinal values estimation: guidelines for data collection, International Journal of Food Microbiology, Volume 100, 2005, pp. 153-165.

[5] CHEN, B. H. – BERMINGHAM, S. – NEUMANN, A. H. – KRAMER, H. J. M. – ASPREY, S. P.: On the Design of Optimally Informative Experiments for

Dynamic Crystallization Process Modeling, Ind. Eng. Chem. Res., Volume 43, 2004, pp. 4889-4902.

[6] GULLO, F. – PONTI, G. – TAGARELLI, A. – GRECO, S.: A time series representation model for accurate and fast similarity detection, in Pattern Recognition, Volume 42, 2009, pp. 2998-3014.

[7] COHN, D. A.: Neural Network Exploration Using Optimal Experiment Design, Neural Networks, Vol. 9, Issue 6, 1996, pp. 1071-1083.

[8] POINT, N. – VADEWOUWER, A. – REMY, M.: Practical Issues in Distributed Parameter Estimation: Gradient Computation and Optimal Experiment Design, Control Engineering Practice, Vol. 4, Issue 11, 1996, pp. 1553-1562.

[9] MANER, B. R. – DOYLE, F. J.: Polymerization reactor control using autoregressive volterra-based MPC, AIChe Journal, Volume 43, 1997, pp. 1763-1784.

Received January 31, 2010, accepted April 19, 2010

BIOGRAPHIES

László Dobos was born on 05.05.1986. In 2009 he graduated (MSc) at the department of Process Engineering at University of Pannonia in Veszprém, Hungary. He wrote his master thesis in the field of modelling and model predictive control of heat exchanger network at Norwegian University of Science and Technology supervised by Sigurd Skogestad. Since his graduation he is working as a PhD student at the department of Process Engineering. His main scientific research field is the application of semi-mechanistic models in process optimization. Related to this field he examines the field of experiment design to develop algorithms to for support efficient parameter estimation.

Zoltán Bankó has graduated at the Department of Information Technology at University of Pannonia, Hungary. He wrote his master thesis in the field of multivariate time series data mining and continued his studies as PhD student at the Department of Process Engineering.

János Abonyi received the MEng and PhD degrees in chemical engineering in 1997 and 2000 from the University of Veszprem, Hungary, respectively. In 2008, he earned his Habilitation in the field of Process Engineering. Currently, he is an associate professor at the Department of Process Engineering at the University of Pannonia. In the period of 1999-2000 he was employed at the Control Laboratory of the Delft University of Technology and worked on adaptive control of bioreactors, developing the ADI 1010 Bio Controller (Applikon). Dr. Abonyi has co-authored more than 90 journal papers and chapters in books and has published two research monographs, Fuzzy Model Identification for Control (Birkhauser Boston, 2003) and Fuzzy Clustering For Data mining and System Identification with B. Feil (Birkhauser, 2007), and one Hungarian textbook about data mining. His research interests include process engineering, quality engineering, data-mining, the use of fuzzy models, genetic algorithms and neural networks in nonlinear system identification, process monitoring and control, and using empirical and first-principle process models in predictive alarm management As a co-founder of Simcont llc., Dr. Abonyi is active in carrying out hazard and operability studies.