16th European Symposium on Computer Aided Process Engineering
and 9th International Symposium on Process Systems Engineering
W. Marquardt, C. Pantelides (Editors)
© 2006 Published by Elsevier B.V.

A framework for model-based design of parallel experiments in dynamic systems

F. Galvanin[a], M. Barolo[a], F. Bezzo[a] and S. Macchietto[a,b]

[a]*DIPIC – Dipartimento di Principi e Impianti di Ingegneria Chimica, Università di Padova, via Marzolo 9, I-35131, Padova, Italy (fabrizio.bezzo@unipd.it)*
[b]*Department of Chemical Engineering, Imperial College London, South Kensington Campus, SW7 2AZ London, UK (s.macchietto@imperial.ac.uk)*

Abstract

Advanced model-based experiment design techniques are essential for rapid development, refinement and statistical assessment of deterministic process models. One objective of experiment design is to devise experiments yielding the most informative data for use in the estimation of the model parameters. Current techniques assume the multiple experiments are designed in a sequential manner. The concept of model-based design of parallel experiments design is presented in this paper. A novel approach, viable for sequential, parallel and sequential-parallel design is proposed and evaluated through an illustrative case study.

Keywords: model-based experiment design, dynamic modelling, parameter estimation, model validation.

1. Introduction

Model-based experiment design techniques allow selecting conditions for the next experiment that are "best", in the sense of having the maximum information content about the underlying process. Typically, it is desired to establish the most appropriate model structure and the best values of the parameters, so as to provide the best fit to experimental data. Based on earlier work of Espie and Macchietto [1] and Zullo [2], Asprey and Macchietto [3] proposed a general systematic procedure to support the development and statistical verification of dynamic process models for both linear and non-linear dynamic systems described by differential and algebraic equations (DAEs). According to this approach and assuming that no model discrimination is required beforehand, three consecutive steps are needed to determine model parameters:

1. the design of a new set of experiments, based on current knowledge (model structure and parameters, and statistics from prior experiments);
2. the execution of the designed experiment and collection of new data;
3. the estimation of new model parameters and statistical assessment.

The sequential iteration of steps 1, 2 and 3 typically leads to a progressive reduction in the uncertainty region of model parameters, thanks to the new information obtained from the experimental data. The procedure has been successfully demonstrated in several applications, such as crystallisation processes [4], mammalian cell cultures [5] and biofuels production [6]. A similar procedure for optimum experimental design was developed by Bauer et al. [7], who assessed it on an industrial reactive system. They also adopted a sequential approach.

There are a number of research and industrial applications where it is possible to envisage the simultaneous execution of several experiments in parallel rather than

sequentially. Miniaturisation allows the definition of array of modules (e.g. micro-reactors for chemical or biochemical reactions) in which several experimental conditions can be simultaneously set up to carry out parallel experiments. Clear advantages in terms of elapsed time saving are presently offset by the lack of a systematic procedure for model-based design of parallel experiments.
In this work, the possibility of advancing the current techniques to tackle the design of parallel experiments is discussed. Furthermore, a new approach based on a statistical analysis of the variance-covariance matrix of the parameters to be estimated is developed and assessed. It is shown that this can also be applied to develop hybrid sequential-parallel experiment design strategies. Parallel and sequential-parallel techniques are compared to a standard sequential approach and potential advantages/disadvantages are highlighted. The applicability of the new experiment design methods to dynamic systems and their performance are illustrated via an illustrative case study.

2. The methodology

Let us consider a process described by the set of DAEs of the form:

$$f\left(x(t),\dot{x}(t),y(t),u(t),q,\theta\right)=0 \ , \tag{1}$$

where $x(t)$ and $y(t)$ are vectors of the differential and algebraic variables, $u(t)$ and q are vectors of the time-varying and time-invariant control variables, and θ is the set of N_θ unknown model parameters to be estimated. Here it is assumed for simplicity the all the M differential variables x can be measured (the case where only a subset is measured being a trivial extension).
Model-based experiment design for parameter precision aims at determining the optimal vector φ of experimental conditions (initial conditions x^0, control variables u and q and the times when measurements are sampled) required to maximise the expected information content from the measured data generated by these experiments, i.e. to minimise the confidence ellipsoid of the parameters to be estimated. This means that some measure ψ of the variance-covariance matrix $\mathbf{V}_\theta$ of the parameters has to be minimised. If we take into account a number N_{exp} of experiments, the matrix $\mathbf{V}_\theta$ is the inverse of the $N_\theta \times N_\theta$ information matrix $\mathbf{H}_\theta$ [8]:

$$\mathbf{V}_\theta(\theta,\varphi)=\mathbf{H}_\theta^{-1}(\theta,\varphi)=\left[\sum_{k=1}^{N_{exp}}\mathbf{H}^*_{\theta|k}+\left(\mathbf{\Sigma}_\theta\right)^{-1}\right]^{-1}=\left[\sum_{k=1}^{N_{exp}}\sum_{i=1}^{M}\sum_{j=1}^{M}\sigma_{ij|k}\mathbf{Q}_{i|k}\mathbf{Q}_{j|k}+\left(\mathbf{\Sigma}_\theta\right)^{-1}\right]^{-1}, \tag{2}$$

where $\mathbf{H}^*_{\theta|k}$ is the information matrix after the k-th experiment, σ_{ij} is the ij-th element of the inverse of the estimated variance-covariance matrix of the residuals $\mathbf{\Sigma}$=cov(x_i, x_j), $\mathbf{Q}_i$ is the i-state matrix of the sensitivity coefficients at each of the n_{sp} sampling points:

$$\mathbf{Q}_i=\left[\frac{\partial x_{il}}{\partial \theta_m}\right] \qquad l=1,...,n_{sp} \qquad m=1,...,N_\theta \ , \tag{3}$$

and $\mathbf{\Sigma}_\theta$ is an approximate variance-covariance matrix of the parameters. Prior information on the parameters can be ignored by dropping the dependency of equation (2) on $\mathbf{\Sigma}_\theta$ [9]. A common choice for the measure ψ is the E-optimality criterion [10], which aims at minimising the largest eigenvalue λ_1 of matrix $\mathbf{V}_\theta$. Note that the definition of matrix $\mathbf{V}_\theta$ and the E-optimality criterion are quite general and do not

depend on whether the experiments are run sequentially or simultaneously. If a sequential approach is considered, the information matrix is defined as:

$$\mathbf{H}_\theta = \sum_{k=1}^{N_{exp}-1} \mathbf{H}^*_{\theta_i|k} + \mathbf{H}^*_{\theta|N_{exp}}(\theta,\varphi) = \mathbf{K} + \mathbf{H}^*_{\theta|N_{exp}}(\theta,\varphi) \quad , \tag{4}$$

where $\mathbf{K}$ is a constant matrix defined by the previous (N_{exp}–1) experiments. In the above information matrix, only the vector φ of the experimental conditions for the new experiment, $N_{exp,}$ is available for optimisation.
On the other hand, N_{exp} new experiments can be designed simultaneously. In this case, the information matrix becomes:

$$\mathbf{H}_\theta = \sum_{k=1}^{N_{exp}} \mathbf{H}^*_{\theta,k}(\theta,\varphi_k) \quad . \tag{5}$$

Here, all vectors φ_k , one for each experiment k are optimized simultaneously, using, as before, the largest eigenvalue λ_1 of the overall matrix $\mathbf{V}_\theta$ (E-optimality) as objective function to be minimised. It is noted that, as the inversion of $\mathbf{H}_\theta$ is a nonlinear operation, the optimum $\mathbf{V}_\theta$ thus obtained will not be the same as the sum of the $\mathbf{V}_\theta$ obtained by optimizing each individual experiment N_{exp} times. In other words, the N_{exp} new optimal experiments will normally be distinct. The main drawback of this approach is that a much larger optimisation problem needs solving.
An alternative method is also proposed here. According to this novel approach each experiment is designed *a-priori* to deliver a vector of experimental conditions producing information which is totally different (orthogonal) from the other ones. In mathematical terms, that means that the information matrix $\mathbf{H}_\theta$ is split into its singular values identified by its N_θ eigenvalues λ_i : the new optimisation criterion, called SV-optimality, aims at maximising the information linked to the N_{exp} largest singular values of $\mathbf{V}_\theta$. Thus, the overall optimisation problem is split into N_{exp} separate optimisation problems, where the k-th measure ψ_k is defined as:

$$\psi_k = \lambda_k(\mathbf{V}_\theta) \qquad k = 1,...,N_{exp} \le N_\theta \qquad \lambda_1 > \lambda_2 > ... > \lambda_{N_{exp}} \quad . \tag{6}$$

The obvious advantage of SV-optimality is that it is easier to solve N_{exp} small optimisation problems rather than a single large one. The second potential advantage is that we do not design the experiments to maximise the information content of the overall set, but each experiment is designed to maximise a specific component of the available information. Note that this approach can also be applied for sequential experiment design: the first experiment will aim at minimising the largest eigenvalue of the variance-covariance matrix, the second will minimise the second largest eigenvalue, and so on.

3. Case study

The methodology discussed in the previous section is applied to a biomass fermentation process that appeared in several papers on the subject [1,3,8]. Assuming Monod-type kinetics for biomass growth and substrate consumption, the system is described by the following set of DAEs:

$$\frac{\mathrm{d}x_1}{\mathrm{d}t} = (y - u_1 - \theta_4)x_1 \quad , \qquad \frac{\mathrm{d}x_2}{\mathrm{d}t} = -\frac{yx_1}{\theta_3} + u_1(u_2 - x_2) \quad , \qquad y = \frac{\theta_1 x_2}{\theta_2 + x_2} \quad , \tag{7}$$

where x_1 is the biomass concentration (g/L), x_2 is the substrate concentration (g/L), u_1 is the dilution factor (h^{-1}), and u_2 is the substrate concentration in the feed (g/L). The experimental condition that characterise a particular experiment are the initial biomass concentration x_1^0 (range 1-10 g/L), the dilution factor u_1 (range 0.05-0.20 h^{-1}), and the substrate concentration in the feed u_2 (range 5-35 g/L). The initial substrate concentration x_2^0 is set to 0 g/L. Both x_1 and x_2 can be measured during the experiment.
The objective is to design a set of experiments to yield the best possible information for the estimation of the four parameters θ_i. The total duration of a single experiment is set equal to 40 h. It is assumed that each experimental run involves 5 sampling intervals. A piecewise-constant profile over 5 switching intervals is assumed for both controls. A total of 15 variables are optimised in each experiment. The elapsed time between any two sampling points is allowed to be between 1 and 20 h and the duration of each control interval between 2 and 20 h. "Experimental data" are obtained by simulation with θ =[0.310, 0.180, 0.550, 0.050]^T as the "true" parameters and by adding multivariate normally distributed noise with a mean of zero; two possible $M \times M$ covariance matrix $\mathbf{\Sigma}$ of the simulated measurements error will be considered:

$$\mathbf{\Sigma}_A = \begin{bmatrix} 0.01 & 0 \\ 0 & 0.05 \end{bmatrix} \qquad \mathbf{\Sigma}_B = \begin{bmatrix} 0.05 & 0 \\ 0 & 0.08 \end{bmatrix}. \tag{8}$$

The initial guess for the parameters' values is set to θ=[0.313, 0.202, 0.551, 0.050]^T.

3.1. Proposed experiment designs and results

Different experiment design approaches are compared assuming that we wish to design the same number of new experiments. Initially, the following designs are implemented:

1. D1: sequential experiment design (E-optimality), 2 experiments
2. D2: parallel experiment design (E-optimality), 2 experiments
3. D3: sequential experiment design (SV-optimality), 2 experiments
4. D4: parallel experiment design (SV-optimality), 2 experiments

Each design is applied first assuming "clean" measurements (Case A: matrix $\mathbf{\Sigma}_A$) and then noisy ones (case B: matrix $\mathbf{\Sigma}_B$).
Results, in terms of the *a-posteriori* statistics obtained after the optimally designed experiments were executed and model parameters re-estimated with the new data, are summarised in Table 1. In all cases, the model responses with the estimated parameters give a statistically good fit of the data derived from the designed experiments, as expressed by the χ^2 value, which is in all cases less than χ^2_{ref} based on a Student distribution. It should be noted that the χ^2 values for the different cases cannot be compared to each other, since each represents the capability of the model to fit the data from the experiments of that specific design. Here, the different designs could be assessed by comparing the estimated parameter values to the true ones. However, in "real life", this test is not possible since the true values are of course not known. Therefore, the best approach is to evaluate the accuracy of the design by observing for each parameter either the interval of estimation confidence or the *t*-value statistics. For a set of experiments to produce a reliable parameter estimation the *t*-value must be greater than a computed reference value derived from a Student distribution (*t*-test).

3.1.1. Case A – Clean measurements

All designs provide statistically sound results (all *t*-values are above the reference threshold). Note, that from this point of view, parallel design is a sensible alternative to save time since the experimental session requires half the time as either D1 or D3 (but, of course, double equipment is needed). One drawback of design D2 is that, as

previously stated, it requires the solution of a larger optimisation problem (30 variables) and, therefore, it may be more upset by convergence issues and, more importantly, by a larger number of local minima. This issue is overcome by design D4.

Table 1. Comparison of sequential and parallel approaches for model-based experiment design (two experiments). Superscript * indicates *t*-values failing the *t*-test

Design	Param. estimate	Conf. interval (95%)	*t*-value (t_{ref}=1.75)	χ^2 (χ^2_{ref} = 26.30)
D1-A	θ = [0.305, 0.164, 0.541, 0.046]^T	[±0.0110, ±0.0518, ±0.0243, ±0.0101]^T	[27.87, 3.17, 22.29, 4.52]^T	21.46
D2-A	θ = [0.299, 0.145, 0.512, 0.042]^T	[±0.0137, ±0.0582, ±0.0474, ±0.0097]^T	[21.80, 2.50, 10.79, 4.32]^T	19.17
D3-A	θ = [0.305, 0.163, 0.542, 0.046]^T	[±0.0107, ±0.0520, ±0.0221, ±0.0096]^T	[28.43, 3.14, 24.60, 4.82]^T	21.63
D4-A	θ = [0.305, 0.269, 0.521, 0.041]^T	[±0.0134, ±0.1431, ±0.0384, ±0.0120]^T	[22.80, 1.88, 13.58, 3.41]^T	15.35
D1-B	θ = [0.300, 0.185, 0.523, 0.038]^T	[±0.0390, ±0.1202, ±0.1138, ±0.0387]^T	[7.69, 1.54*, 4.60, 0.98*]^T	22.19
D2-B	θ = [0.320, 1.189, 0.474, 0.032]^T	[±0.0443, ±1.283, ±0.0769, ±0.0182]^T	[7.22, 0.93*, 6.16, 1.73*]^T	17.12
D3-B	θ = [0.292, 0.151, 0.513, 0.040]^T	[±0.026, ±0.1084, ±0.0564, ±0.0188]^T	[11.20, 1.40*, 9.10, 2.15]^T	20.48
D4-B	θ = [0.300, 0.132, 0.536, 0.044]^T	[±0.0278, ±0.1122, ±0.0627, ±0.0287]^T	[10.78, 1.17*, 8.55, 1.53*]^T	22.80

The best parameter estimation in terms of confidence interval and *t*-values is obtained by means of design methods D1 e D3, i.e. the two sequential ones. This is as expected, since the second experiment is designed using the information content from the first experiment. It is interesting to note that approach D3 performs slightly better than D1. In particular, D3 produces a more confident estimation of parameter θ_3, hinting that some of the information content related to that parameter belong to a different direction in the variance-covariance matrix. Although less precise, a similar behaviour can be detected by comparing D2 and D4. D4 is less precise as far as the estimation of parameters θ_2 and θ_4 is concerned. Nonetheless, a better estimation of θ_3 is obtained.

3.1.2. Case B – Noisy Measurements

These results are rather more interesting. First of all, no design is capable of providing a full set of reliable parameters (D2 produces a particularly bad θ_2 estimation). More experiments are needed. In this case SV-optimality is a better criterion. Both designs D3 and D4 are sensibly more performing. Design D3 is the only one providing a statistically sound estimation of three parameters. However, what is surprising is that D4 is overall a better design than D1. Exploiting the information related to λ_2 is more important than having the chance to design the second experiment by using the information of the first experiment. Once again, it can be seen that SV-optimality leads to a good estimation of parameter θ_3, while E-optimality provide a better estimation of parameter θ_2. This confirms the hypothesis that the direction identified by the second eigenvalue contains some valuable information related to the third parameter.

In view of the above results, it seems reasonable to design a set of 3 experiments aiming first at extracting most of the information related to the first eigenvalue (indeed, the most informative) and then at maximising the information related to the next two largest eigenvalues. Two more design formulations are thus considered:

5. D5: sequential experiment design (E-optimality), 3 experiments
6. D6: sequential-parallel experiment design (E+SV-optimality), 1+(2 parallel) experiments

Results are summarised in Table 2 (from the same initial conditions as before). Design D5 shows that three sequential experiments are still insufficient to reliably estimate all parameters: the estimate of parameter θ_2 is nearly acceptable, but that of θ_4 is not. On the contrary, the results from design D6 are fully satisfactory. Not only is it possible to obtain (in a shorter time period) a statistically precise estimation of the entire set θ (particularly of θ_3), but all parameters are better estimated than in D5. This seems to confirm that valuable information is related to the smaller eigenvalues and that a proper exploitation of such information can produce more effective experimental designs.

Table 2. Comparison of sequential and sequential-parallel approaches for model-based experiment design (three experiments). Superscript * indicates *t*-values failing the *t*-test

Design	Param. estimation	Conf. interval (95%)	*t*-value (t_{ref}=1.70)	χ^2 (χ^2_{ref}= 38.85)
D5-B	θ= [0.305, 0.189, 0.532, 0.041]^T	[±0.0297, ±0.1118, ±0.0920, ±0.0307]^T	[10.28, 1.69*, 5.79, 1.34*]^T	29.78
D6-B	θ= [0.298, 0.158, 0.528, 0.043]^T	[±0.0105, ±0.0364, ±0.0237, ±0.0080]^T	[13.87, 2.11, 10.85, 2.61]^T	27.54

4. Final remarks

A novel procedure based on the decomposition of the variance-covariance matrix has been suggested, which is applicable to the model-based design of both sequential and parallel experiments. Preliminary results on an illustrative application demonstrate the promising potential of this new approach. Future work will assess the applicability of the methods to larger applications and the development of a systematic procedure to help determine the best approach to use for model-based experiment design, whether sequential, parallel, or mixed sequential-parallel.

References

[1] D. Espie and S. Macchietto, AIChE J., 35 (1989) 22.
[2] L. Zullo, PhD Thesis, The University of London, 1991.
[3] S.P Asprey and S. Macchietto, Comput. chem. Engng., 24 (2000) 1261.
[4] B.H. Chen, S. Bermingham, A.H. Neumann, H.J.M. Kramer and S.P. Asprey, S.P, Ind. Eng. Chem. Res., 43 (2004) 4889.
[5] F.R. Sidoli, A. Manthalaris and S.P. Asprey, Ind. Eng. Chem. Res., 44 (2005) 868.
[6] G. Franceschini and S. Macchietto (L. Puigjaner and A. Espuna Eds), ESCAPE –15, CACE Series 20A, Elsevier, Amsterdam, The Netherlands, (2005) 349.
[7] I. Bauer, H.G. Bock, S. Körkel and J.P. Schlöder, J. Comput. Appl. Mathem., 120 (2000) 1.
[8] S.P. Asprey and S. Macchietto, J. Proc. Control, 12 (2002) 545.
[9] G.E.P. Box and H.L. Lucas, Biometrika, 46 (1959) 77.
[10] J. Kiefer and J. Wolfowitz, Ann. Math. Stat., 30 (1959) 271.