Optimal Measurement Design for Monitoring Batch Processes

Olja Stanimirovic, Huub C. J. Hoefsloot, and Age K. Smilde
Biosystems Data Analysis Group, Swammerdam Institute for Life Sciences, Universiteit van Amsterdam, Amsterdam 1018 WV, The Netherlands

DOI 10.1002/aic.11968
Published online September 29, 2009 in Wiley InterScience (www.interscience.wiley.com).

Keywords: batch processes, sensor, PCA, variable selection, optimal time

Introduction

Recently a method is published[1] to find optimal sensor positions in distributed parameter systems. Using the well known analogy between plug flow reactors and batch reactors this method can easily be adopted to find optimal time-points to perform measurements on batch processes. The approach is data-driven, but can also work with data generated by an existing batch model. As a result it gives the minimum required number of measurements to characterize a batch and the best points in time to do those measurements. Measurements that characterize a batch adequately could be used to reconstruct the batch trajectory or to predict quality parameters of the end product in an early stage of the batch process. If the measurements are to be used for a single predefined task, a PLS approach with internal measurement point selection will be superior.[1] But the same paper shows that the selected points following the lines of this paper are performing almost as well to the points selected by PLS. The approach taken in this note is useful if the measurements have multiple objectives, like monitoring and prediction.

The approach consists of four operations on the available batch data that will be discussed in the theory section. Although the steps involved in the analysis show resemblance to multivariate statistical process control (MSPC) for batch processes it is a different procedure. In the batch MSPC literature,[2,3] the topic when to do measurements on a batch and which measurements to use, is not discussed.

Other approaches that try to find optimal times to measure for batch processes exist, but also depend on knowledge about the kinetic model. For example, Boelens et al.[4] present an optimal sampling strategy using a D-optimal design criterion for estimation of the rate constant k_1 of a (pseudo)first order reaction. The result is the best point in time to do a measurement of the reactant in a batch process with first order kinetics. Westerhuis et al.[5] extend this approach to a second-order reaction. To our knowledge, our approach is unique in the sense that it only requires the availability of measured batch variables of some relevant batch runs.

Correspondence concerning this article should be addressed to H. C. J. Hoefsloot at h.c.j.hoefsloot@uva.nl

Theory

The problem of finding the best sensor location in a distributed parameter system can be transformed into the problem of finding the best point(s) in time to do a measurement that characterizes a batch. Two main contributions to the variation of the data collected from several batch runs are the time evolution within a batch and the differences between the batches. Assume, for example, that a reactant concentration is measured and that the reaction kinetics, i.e., the reaction rate constants, are slightly changing over the batch runs. The data from (repeated) runs of the batch reaction can be put in the rows of a matrix X. In a row of X the time evolution of the batch can be found. The values in a column of X vary due to the different rates. Performing a PCA on X, the difference due to the rates at a certain point in time would turn up directly in the loading vectors.[1] The best points in time to monitor these types of differences over the batches are found by analyzing the loading space of a PCA applied to the column mean centered matrix X.

General approach for finding the best time to measure for a batch reaction

Measured or simulated data of the batch process is collected into a three-way matrix $\underline{X}$. This matrix has dimensions $I \times J \times K$ where: I = Number of batches, $i = 1 \ldots I$; J = Number of process variables, $j = 1 \ldots J$; K = Number of time points, $k = 1 \ldots K$. The number of batches (I) represents the number of (repeated) runs of the same batch reaction. In the case of simulated data, the number of batches represents the number of simulations. A process variable (j) can be an engineering variable

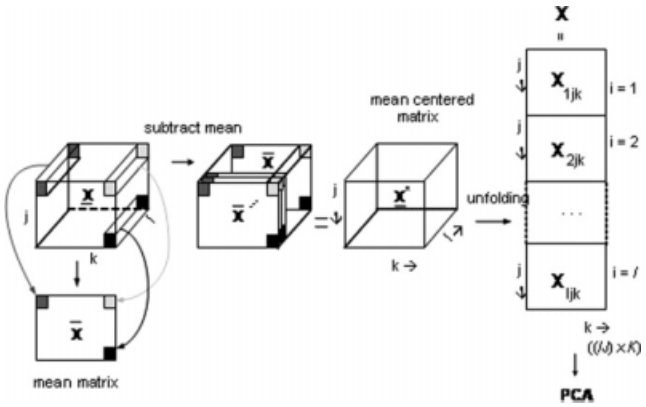

Figure 1. Mean centering over batches and the way to unfold the batch data.

such as: temperature or pressure or a result from an analytical measurement on the batch process, such as a compound concentration, or a channel of a spectroscopic measurement done on the batch reaction mixture.

The general procedure to analyze the data is: (1) Mean center the three-way matrix ($\underline{X}$) across the batch mode (mode *I*), (2) Unfold the resulting matrix, (3) Apply principal component analysis to the unfolded matrix, (4) Analyze the results using a variable selection technique (Figure 1).

The way of unfolding the data determines the type of information that is obtained in the data analysis that follows the unfolding.[6] The unfolding method used in this article can be used to find the informative time points if all variables are measured simultaneously. The obtained time points are not the points that are the most informative for a single variable, but gives time points where on average the variables are most informative. Other unfolding methods can be used in conjunction with the PCA steps as described in our article, but this is beyond the scope of this present article.

Orthogonal Variables in Loading space (OVL)[1] is used as the variable selection method. There are other possible variable selection methods available that could be incorporated as our selection method. We use OVL here because in the paper on distributed systems it outperformed the other methods under consideration.[1]

The first and most important time-point selected by OVL is the point that is furthest away from the origin in the loading space of the PCA solution. This time-point contains the maximum amount of variation between the batches. The next time point is the time-point in the loading space that is the furthest away from the line through the origin and the first point that is chosen. The next point is the time-point furthest away from the (hyper) plane defined by the previous points. By following this procedure the information in the selected points is a trade off between the amount of information in the selected time-point and the amount of correlation with the previously selected points.

Studied Batch Reaction

The batch reaction between 3-chlorophenylhydrazonopropane dinitrile and β-mercaptoethanol is taken as an example.[7] The basic reaction scheme is:

$$\mathrm{A} + \mathrm{B} \rightarrow \mathrm{C} \text{ reaction constant } k_1 \quad (\mathrm{M}^{-1}\ \mathrm{min}^{-1})$$

$$\mathrm{C} \rightarrow \mathrm{D} + \mathrm{E} \text{ reaction constant } k_2 \quad (\mathrm{min}^{-1})$$

Only species A, B are present at the start of the reaction. An excess of reactant B is used to create pseudo first order conditions for the first step of the reaction. The corresponding first order reaction rate constant k_{1B} (min^{-1}) is defined as $k_1.c_{B0}$ (in which c_{B0} is the initial concentration of species B). The reaction was monitored by UV–vis spectroscopy. Only A, C, and D are spectroscopically active. Ten repeats of the batch reaction were performed. The spectra are collected at a time interval of 10 s. The batch reaction was stopped after 45 min. At this point in time not all intermediate reactant (C) had disappeared (Figure 2).

Pretreatment of UV–vis spectral data

From the measured spectra during the batch reaction and the measured pure spectra of the compounds A, C, and D the concentration time profile of the each compound is calculated using the Lambert-Beer law: $\mathbf{S} = \mathbf{C}\ \mathbf{S}_{\mathrm{pure}}'$ $\mathbf{C} = \mathbf{S} \cdot \mathbf{S}_{\mathbf{pure}} \cdot (\mathbf{S}_{\mathbf{pure}}' \cdot \mathbf{S}_{\mathbf{pure}})^{-1}$ in which $\mathbf{S}$ is the $(K \times J)$ matrix that contains measured UV–vis mixture spectra, $\mathbf{S}_{\mathrm{pure}}$ is the $(J \times 3)$ matrix that contains measured pure spectra of A, C, and D and $\mathbf{C}$ is the $(K \times 3)$ matrix that contains concentration profiles of A, C, and D. Although these concentration profiles are not directly measured during the process, they are calculated using measured pure and measured mixture spectra. We will therefore refer to them also as the measured concentration profiles.

Simulations

It is assumed that the main batch to batch differences are caused by slight changes in reaction rate constants. Simulations are done in which the kinetic rate constants are perturbed. Also, other batch to batch differences might be present, but it is assumed that they are not dominant.

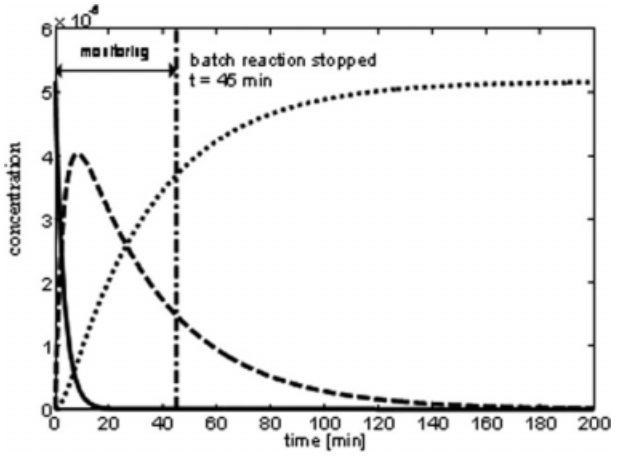

Figure 2. Concentration profiles of reactant A (solid), intermediate C (dashed), and product D (dotted).

Table 1. Best Times to Measure Using Concentrations as an Input

Ranking	Measured	Simulated
1	01:50	03:20
2	45:00	28:10
3	08:00	11:00

The reaction rate constants and their standard deviation that are used as input for simulations are taken from Bijlsma.[7] Five hundred sets of concentration profiles are generated using these rate constants. From the simulated concentration profiles and the measured spectra of the pure compounds, spectra of the batch reaction mixture at time t could be calculated. These spectra are referred to as simulated spectra. Such spectra were calculated until $t = 45$ min to be comparable with actual batch experiments done or till $t = 200$ min, extrapolating the batch time. At 200 min the reactions are almost finished.

Results

Results for the measured concentration profiles

The strategy is applied to the measured concentration profiles, resulting in a data cube of size $10 \times 3 \times 271$. Results are presented in Table 1. Using a leave-one out cross-validation[8] it is determined that three PCs are sufficient, indicating that measurements at three points in time are enough to characterize the batch trajectories.

We use smoothed PCA[9] to analyze the data to obtain smoother profiles. Figure 3A shows the curves that were used to find the times listed in Table 1. The distances in the loading space are plotted. The solid line represents the distance of each time point from the origin of the loading space. The time point that has the maximum distance is selected as first time to do a measurement. The dashed line represents orthogonal distances of all remaining time points from the line that goes through the origin and first selected time point. The best time point is the one that has the largest orthogonal distance.

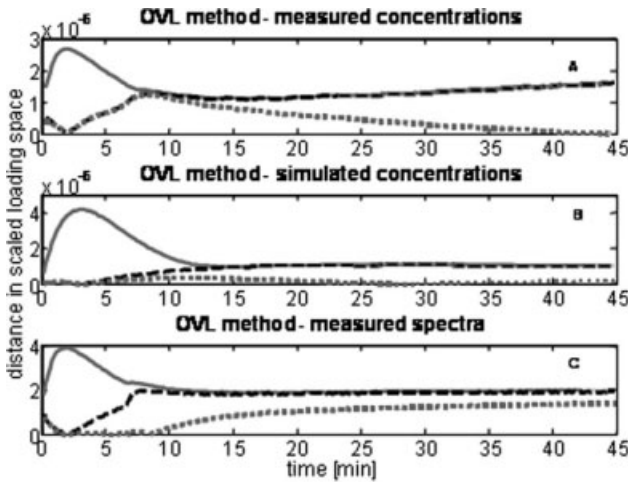

Figure 3. The OVL results for: A: measured concentrations; B: simulated concentrations; and C: measured spectra. In each subfigure there are 3 OVL components, first (solid), second (dashed), and third (dotted).

The OVL method selects a time point at the beginning (01:50 min), at the end of monitoring time (45:00 min) and around a batch time of 8 min. These results will be further discussed below. The solid line in Figure 3A shows that the OVL has a clear maximum around 2 min. For the OVL method, the second distance curve (dashed curve, Figure 3B) indeed has a low value around 2 min and it has a maximum approximately at 45 min. The flatness of the lines for more then 7 min in Figure 3 indicates that the timing is not very crucial. This is because the information in the neighboring points form the optimal contains only a slightly lower amount of information.

Comparison between the results for measured and simulated concentration data

Columns 2 and 3 in Table 1 show that best points in time found for simulated and measured concentrations follow the same pattern. The most important time to measure is found when the first step of the reaction (A $\rightarrow$ C) dominates and hardly any product is formed yet. The next point in time is selected when second step of reaction (C $\rightarrow$ D) is proceeding: at this stage the amount of A is very small and considerable amount of the intermediate C and the product D are present. Finally, a point in time is selected in the transition stage of the reaction around the point in time that the intermediate C has its maximum concentration. Both steps of the reactions are then active.

Figure 3 shows the OVL distance curves for the measured (Figure 3A) and the simulated concentrations (Figure 3B). The fact that the first two distance curves in Figure 3A have roughly the same shape as the curves in Figure 3B confirms our assumption that kinetics is the main source of variation over the batches. The curves for the measured concentration, however, slightly increase towards higher batch times (Figure 3A). Other batch to batch differences present in measured data (such as instrumental drift and noise or disturbances in inlet concentration etc.) may cause this. It could also be that the proposed kinetic model used in the simulations is not fully adequate for higher batch times.

Comparison between results for measured and simulated spectral data

Instead of using the concentration it is also possible to use a whole spectrum as a measurement input. Table 2 gives the results when simulated and measured spectra are used in matrix $\underline{X}$ instead of compound concentrations for the OVL method. A good agreement is found between the obtained time points. There is, however, a ranking difference of 2nd and 3rd time point found for simulated and measured data.

Table 2. Best Times to Measure Using Spectra as an Input

Ranking	Measured	Simulated
1	02:00	03:10
2	08:20	36:30
3	44:30	10:20

This is explained by Figure 3C, that shows results for the OVL method and measured spectra. Although the actual maximum of the curve is found at 8:20 min, the level of the OVL distance line for the second point in time is almost constant after 7 min. The flatness of this curve makes the ranking difference of 2nd and 3rd time point for simulated and measured data insignificant. The times given in Table 1 and Table 2 differ due to the different nature of the measurement, but the global picture is similar.

Best point in time to measure: a comparison with the literature for pseudo first order reactions

To validate our approach, a comparison is made with theoretical results on the same system given in Boelens et al. In that article only the first reaction A + B $\rightarrow$ C is considered and optimal measuring time of 3 min and 20 s is found. We simulated this system by putting normal distributed noise on the values of the k_1 with a variance as measured.[7] Performing the four steps of our algorithm we found the optimal measurement time to be 3 min and 10 s which is in good agreement with the theoretical result.

Conclusions

The basic question addressed here is: ''When measurements should be done for optimal characterization of a batch process?'' A general, model-free strategy is proposed that has as input the data of relevant and well-instrumented batch runs or the data that is produced by an already existing process model. There is no restriction on the type of data that can be used. For the studied batch reaction, we showed results based on concentrations and on UV–vis spectra. The method may, however, also be applied on physical measurements (temperature, pressure) done during a batch process. The main idea of the strategy is to select the points in time at which the (state) variables have a large variation between the available batch runs.

If the selected measurements obtained by the method described in this article are to be used for predictions several situations can be distinguished. If it is clear what quantity is to be predicted a PLS approach[10] is superior, although the selected measurements points perform well together if used in a PLS model.[1] The method described here is meant for situation where multiple objectives are present or it is not clear which quantity should be predicted.

The method to analyze the batch data consists of four simple and fast operations: mean centering, unfolding, principal component analysis (PCA), and selecting best points in time using information from the loading space of the PCA. We tested and validated this method for a two-step biochemical batch reaction monitored by UV–vis spectroscopy.

When the first step of this reaction is operated as a pseudo first order reaction there is a good agreement between the best points in time of our method and literature values. It is confirmed that the best time point to measure for a (pseudo)-first order reaction is $1/k$, with k the rate constant. As input only a few repeated batch runs of the biochemical reaction were used. The advantage is that a kinetics model is not needed to obtain this result.

The agreement we found between the best times for measured and data simulated using kinetics model for the reaction shows that the main differences between the repeated batches in this application are caused by changes in the kinetics. The proposed strategy picks up these variations adequately.

The OVL method yields time points having nearly independent information. An additional advantage of OVL method is that inspection of the distance curves directly supplies information about the relative importance of potential monitoring time points found. In this way the method directly supplies insight in the advantage of selecting additional measurement time points.

Literature Cited

1. Stanimirovic O, Hoefsloot HCJ, De Bokx PK, Smilde AK. Variable selection methods as a tool to find sensor locations for distributed parameter systems. *Ind Eng Chem Res*. 2008;47,1184–1191.
2. Nomikos P, MacGregor JF. Monitoring batch processes using multiway principal component analysis. *AIChE J*. 1994;40:1361–1375.
3. Qin SJ. Statistical process monitoring: basics and beyond. *J Chemom*. 2003;17:480–502.
4. Boelens HFM, Iron D, Westerhuis JA, Rothenberg G. Tracking chemical kinetics in high-throughput systems. *Chem Eur J*. 2003;9: 3876–3881.
5. Westerhuis JA, Boelens HFM, Iron D, Rothenberg G. Model selection and optimal sampling in high-throughput experimentation. *Anal Chem*. 2004;76:3171–3178.
6. Camacho J, Pico J, Ferrer A. Bilinear modelling of batch processes. Part I: theoretical discussion. *J Chemom*. 2008;22:299–308.
7. Bijlsma S, Smilde AK, Estimating reaction rate constants from a two-step reaction: a comparison between two-way and three-way methods. *J Chemom*. 2000;14:541–560.
8. Wise BM, Gallagher NB. *PLS_Toolbox Version 4.0*. Wenatchee, WA: Eigenvector Research Inc, 2000.
9. Ramsay JO, Silverman BW. *Functional Data Analysis*, 2nd ed. New York, NY: Springer-Verlag (Springer Series in Statistics), 2005.
10. Camacho J, Pico J, Ferrer A. Bilinear modelling of batch processes. Part II: a comparison of PLS soft-sensors. *J Chemom*. 2008;22:533–547.

Manuscript received Feb. 3, 2009, and revision received Apr. 28, 2009.