I&EC research
Industrial & Engineering Chemistry Research

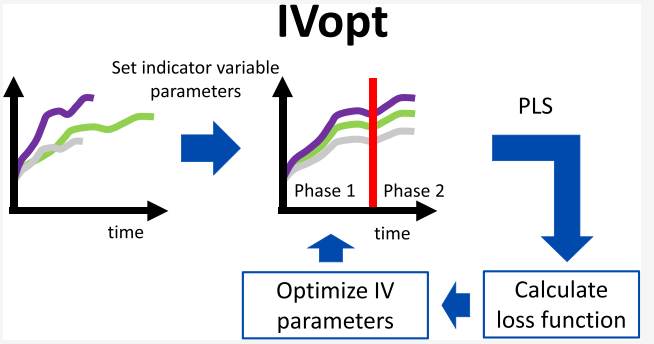
pubs.acs.org/IECR
Article

Optimal Indicator-Variable Approach for Trajectory Synchronization in Uneven-Length Multiphase Batch Processes

Francesco Sartori, Pierantonio Facco, Federico Zuecco, Fabrizio Bezzo, and Massimiliano Barolo*

Cite This: *Ind. Eng. Chem. Res.* 2023, 62, 18511−18525

ABSTRACT: Partial least-squares regression models assessing the end-point product quality in batch processes require that all of the measured variable trajectories across the historical batches have the same length. Most of the conventional and advanced methodologies for batch synchronization need some prior knowledge about the process to carry out one or more of the following activities: partitioning of the batches into phases, selection of an appropriate indicator variable that is then used to synchronize the batches, or selection of a reference batch to which all other batches are matched. We present an optimal indicator-variable approach for phase partitioning and trajectory synchronization in uneven-length multiphase batch processes. The main advantages are that partitioning into phases and selection of the most appropriate indicator variable within each phase are performed automatically rather than manually and are carried out simultaneously rather than disjointly based on a surrogate optimization framework that maximizes the performance of the product quality assessment model under development. Therefore, differently from conventional and advanced synchronization methodologies currently available, the proposed method is completely process-agnostic, which enhances applicability to complex batch processes. Also, in terms of computational times, it scales favorably with the calibration data set size. An industrial fed-batch process for the manufacturing of a specialty chemical and a simulated fed-batch process for the manufacturing of penicillin are used as test beds and demonstrate that the new indicator-variable approach has a superior performance than models built using other synchronization strategies.



1. INTRODUCTION

Many high-value-added products (e.g., specialty chemicals, (bio)pharmaceuticals, food, semiconductors) are obtained by batch processing. Batch processes are run through a recipe, i.e., a sequence of elementary finite-duration processing steps (such as charge, heat up, stir, react, cool down, hold, discharge). Each step is characterized by a given set of operating conditions and is typically triggered by the occurrence of events (e.g., enough reactant has been fed; temperature reaches a given value; torque exceeds a threshold). Flexibility is a key characteristic of batch manufacturing: by adjusting (either directly or indirectly) the length of the processing steps, a batch process can accommodate variability in the raw materials, operating conditions, and status of the equipment and utilities, thus delivering a product that can meet the assigned quality target. As a consequence, a set of batches is often characterized by an uneven duration between batches even if all batches manufacture a product that meets the specification.

From the quality control point of view, most batch processes are run at open loop, meaning that quality is assessed only on the end-product at the end of a batch. Depending on the industrial domain, if an off-spec product is detected, the batch may be rejected, reworked, or progressed with a warning to the downstream process that further processes that product. When product quality cannot be measured conveniently (e.g., because a field sensor is not available or lab analysis takes long to complete), end-point quality assessment is aided by models, which use time-resolved measurements from the plant sensors to either estimate the end-point product quality (models as soft sensors) or simply discriminate between on-spec and off-spec products (models as classifiers). Multivariate statistical methods, such as projection onto latent structures (PLS),[1−3] PLS discriminant analysis (PLS-DA)[4] and their multiway extensions,[5] offer convenient modeling environments in this context because they are interpretable and preserve time resolution in the available data.[6]

A crucial aspect when using multiway-PLS or multiway-PLS-DA as modeling platforms for quality assessment is that they

Received: June 7, 2023
Revised: October 10, 2023
Accepted: October 10, 2023
Published: October 27, 2023

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

typically require data alignment, namely, equalization (all the variables are expressed at the same sampling rate across batches) and synchronization (all the landmarks for the variables trajectories are aligned in time across batches).[7] Ad hoc synchronization techniques, such as truncating the trajectories of all batches to the shortest batch length[8] or extending the length of shorter batches by repeating the last measurement,[9] are simple workarounds that can be set up quickly for preliminary data set screening and analysis but may provide ineffective data modeling.[6] A more effective, yet still simple, synchronization strategy consists of nonlinearly mapping time to an indicator variable (IV),[10] namely, to a measured variable that (*i*) progresses monotonically in time, (*ii*) has a favorable signal-to-noise ratio,[11,12] and (*iii*) has the same starting and ending values for all batches. The IV is to be selected by the user based on process knowledge.[13,14] It may not exist for an entire batch but can exist for single time windows wherein the measured variables have similar correlation structure. Each such window is called a batch phase (not to be confused with a batch processing step). The IV approach for trajectory synchronization is very popular and has been proven effective in a number of applications.[11,15−19] However, when several potential IVs exist, it may not be obvious which one is the most appropriate to choose. Furthermore, partitioning a batch into phases is a challenge in itself[20−23] because phases do not necessarily match the occurrence of physical events in a process (i.e., phases do not necessarily match processing steps). Finally, phase partitioning and batch synchronization have been mostly regarded as two independent activities despite the fact that they are both functional to the model that needs to be developed. Indeed, both phase partitioning and batch synchronization are known to have a strong impact on the model performance.[7,24]

Advanced synchronization techniques exist that do not use an IV for synchronizing batch trajectories. Dynamic time warping (DTW),[25] correlation optimized warping (COW),[26,27] and multisynchro (MS)[28] are the most popular among them. DTW synchronizes two trajectories by translating, compressing, and expanding them so that similar features within them are matched. The method is inherently multivariate because it does not rely on a single variable to perform the synchronization. However, it requires selecting a reference batch the synchronized ones should be matched to; furthermore, its computational burden scales badly with the data set size (namely, with the number of time points characterizing each trajectory).[29] Finally, DTW is known to generate artifacts when some batches are significantly shorter than the chosen reference.[28] COW is based on maximizing the correlation between two trajectories and is less computationally demanding than DTW. However, it is univariate by design because each variable is synchronized separately from the others. Moreover, it can generate artifacts and requires identifying a reference batch and using it for synchronizing all other batches.[30] MS aims not only at minimizing a defined distance between a reference batch and the other batches in a data set but also at removing particular asynchronous behaviors that it can identify among batches (e.g., incomplete batch runs; delayed measurement collection; natural variability). Arguably, it is the most advanced batch synchronization algorithm proposed to date, and it is based upon DTW and can therefore be computationally intensive.

It is to be noted that some multivariate statistical techniques, such as PARAFAC2[21] and GHOPLS-CP,[31] can deal with time-resolved data from uneven-length batch processes without the need for batch synchronization. However, these techniques are computationally inefficient[22,32,33] and more sensitive to noise[34] with respect to multiway PLS and multiway PLS-DA. On the other hand, when retaining time resolution is not a requirement, one can resort to feature-oriented data analysis,[35−37] which does not require batch synchronization.

In this study, we propose an optimal IV approach (IVopt) for trajectory synchronization in uneven-length multiphase batch processes. The ultimate aim is to build an effective multivariate statistical model for end-point quality assessment once a batch is terminated. The idea behind IVopt is that phase partition and trajectory synchronization are carried out simultaneously, rather than disjointly, using an optimization framework based on surrogate modeling with the aim of maximizing the performance of the product quality assessment model that is under development. Within this approach, we resort to surrogate optimization to find the optimal phase partition parameters, and we propose a novel methodology for the automatic identification of the most appropriate IV within each batch phase. We challenge IVopt against standard and advanced synchronization strategies, namely, trajectory truncation (TR), trajectory extension (EXT) with mean values, IV (with *a priori* phase partitioning and IV selection based on engineering judgment), DTW, COW, and MS. We use two case studies as test beds: an industrial fed-batch process for the manufacturing of a specialty chemical and a simulated fed-batch process for the manufacturing of penicillin.[38]

The article is organized as follows. Section 2 provides the mathematical background for both multivariate statistical modeling and automatic phase partitioning. Section 3 illustrates the proposed IVopt methodology. The case studies are presented in Section 4, and the results are discussed in Section 5. Finally, Section 6 summarizes the main conclusions from the study.

2. MODELING BACKGROUND

In this section, we provide the required background on the multivariate statistical modeling methodologies and the automatic phase partition algorithm that are used within the proposed methodology for optimal batch synchronization.

2.1. Projection onto Latent Structures (PLS) and PLS Discriminant Analysis (PLS-DA). In this study, product quality assessment is carried out by means of a soft sensor or a product classifier, which is built using PLS or PLS-DA, respectively. Here, we summarize the basics of these modeling techniques; details can be found in the original references.[1,4]

PLS is a regression methodology to deal with large sets of noisy and collinear data. Let us consider a predictor data set $\mathbf{X}$ $[I \times J]$ and a response data set $\mathbf{Y}$ $[I \times L]$, where I is the number of observations, J is the number of predictors (model inputs), and L is the number of responses (model outputs). Prior to any other operations, the columns of $\mathbf{X}$ and $\mathbf{Y}$ are autoscaled, i.e., mean-centered and scaled to unit variance. The observed data are assumed to be generated by driving forces in a system that can be described by A latent variables, where $A \ll \min(I, J, L)$. The structure of a PLS model is described by the following equations:

$$\mathbf{X} = \mathbf{T}\mathbf{P}^{\mathrm{T}} + \mathbf{E} \tag{1}$$

$$\mathbf{Y} = \mathbf{T}\mathbf{Q}^{\mathrm{T}} + \mathbf{F} \tag{2}$$

$$\mathbf{T} = \mathbf{X}\mathbf{W}(\mathbf{P}^{\mathrm{T}}\mathbf{W})^{-1} \tag{3}$$

where $\mathbf{T}$ $[I \times A]$ is the score matrix, $\mathbf{P}$ $[J \times A]$ and $\mathbf{Q}$ $[L \times A]$ are the $\mathbf{X}$ and $\mathbf{Y}$ loading matrices, $\mathbf{E}$ $[I \times J]$ and $\mathbf{F}$ $[I \times L]$ are the

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

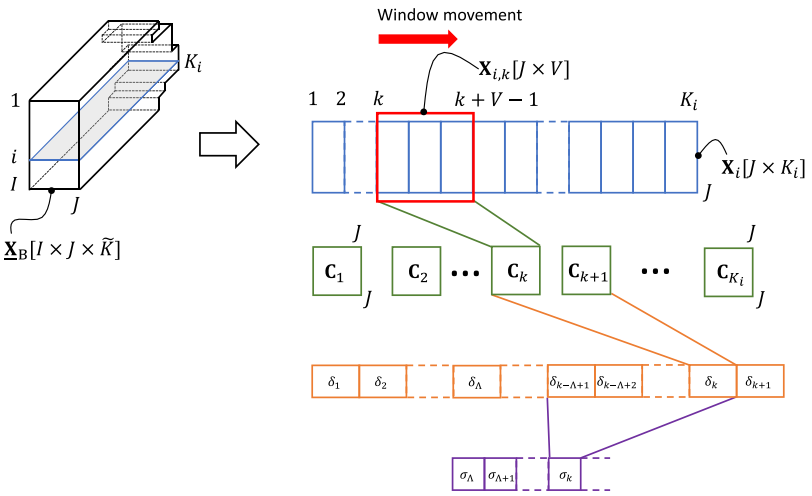


Figure 1. Procedure performed for automatic phase partition of multiphase, uneven-length batches.

matrices of the residuals of $\mathbf{X}$ and $\mathbf{Y}$, and $\mathbf{W}[J \times A]$ is the weight matrix, through which the data in $\mathbf{X}$ are projected onto the latent space to give $\mathbf{T}$ according to eq 3. In this study, PLS models are built using the nonlinear iterative partial least-squares algorithm.[1]

When a new observation $\mathbf{x}_p[1 \times J]$ becomes available, its score $\mathbf{t}_p[1 \times A]$ is calculated as

$$\mathbf{t}_p = \mathbf{x}_p\mathbf{W}(\mathbf{P}^T\mathbf{W})^{-1} \tag{4}$$

and therefore, the predicted response $\hat{\mathbf{y}}_p[1 \times L]$ is given by

$$\hat{\mathbf{y}}_p = \mathbf{t}_p\mathbf{Q}^T \tag{5}$$

PLS-DA is a classification methodology based on the PLS regression algorithm. Consider a response data set $\mathbf{Y}_c[I \times L]$, where the element in row i and column l of the data set is equal to 1 if the ith observation belongs to the lth class (e.g., an on-spec product) or 0 otherwise (e.g., an off-spec product). PLS-DA works by building a PLS model (eqs 1−3) on data sets $\mathbf{X}$ and $\mathbf{Y}_c$. The estimated class attributions for the calibration data set are used to fit a cumulative density function to identify the probability of belonging to a specific class.[39] Once the prediction on a new observation is calculated through eqs 5, the above-mentioned cumulative density functions are used to calculate the probability of attributing the new observation to the relevant class.

Typically, the data from a batch process are arranged in a tensor $\underline{\mathbf{X}}_{\mathcal{B}}[I \times J \times \tilde{K}]$, where I is the number of batches, J is the number of measured variables, and $\tilde{K}$ is the number of time points over which the measurements have been made available through the batches. Notice that because the batches may have different lengths, $\tilde{K}$ changes across batches. The response matrix $\mathbf{Y}_B[I \times L]$ contains the data regarding the L attributes describing the end-point product quality for every batch. To use these data for product quality assessment, multiway PLS is adopted.[5] The batches are first synchronized to a common length $\bar{\tilde{K}}$; then, PLS is applied to $\mathbf{X}_B[I \times J\bar{\tilde{K}}]$ and $\mathbf{Y}_B$, where $\mathbf{X}_B$ is the batchwise-unfolded matrix obtained by concatenating horizontally each vertical slice of size $[I \times J]$ of the synchronized version of $\underline{\mathbf{X}}_B$. Extension to multiway PLS-DA is straightforward.

The selection of the number of latent variables to be retained is carried out by cross-validation, which consists of iteratively removing a subset of the calibration samples, calibrating a model over the remaining samples, and predicting the response of the removed samples. Once all samples are excluded and the relevant responses predicted, a loss function is calculated. For a PLS model, a convenient loss function is the root-mean squared error of cross-validation (RMSECV):

$$\text{RMSECV} = \sqrt{\frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{I - 1}} \tag{6}$$

During PLS model cross-validation, the coefficient of determination in validation Q^2 is also calculated to describe how much of the variation of the response from its mean $\bar{\mathbf{y}}$ is predicted by the regression model:

$$Q^2 = 1 - \frac{\sum_{i=1}^{I}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{I}(y_i - \bar{\mathbf{y}})^2} \tag{7}$$

When using PLS-DA, a convenient loss function is the ratio between the number C_c of samples classified correctly and the total number I of samples to be classified:

$$\text{Accuracy} = \frac{C_c}{I} \tag{8}$$

Accuracy is equal to 1 when all samples (i.e., all batches) are classified correctly, whereas it is equal to 0 when all samples are classified incorrectly.

2.2. Automatic Phase Partition of Uneven-Length Multiphase Batches. Guo and Jin[23] proposed a model-agnostic, phase partition methodology that automatically returns the number Φ_i of phases into which one entire nonsynchronized batch i can be partitioned, together with the time point at which each phase onsets. The methodology (Figure 1) is based on analyzing the change in the correlation structure of the measured data across multiple consecutive time points. Here, we summarize the original methodology proposed by Guo and Jin;[23] further considerations will be made in Section 3.1.

Consider a horizontal slice of $\underline{\mathbf{X}}_{\mathcal{B}}$, including all measurements taken from batch i across all K_i time points along the batch, and arrange the relevant data in matrix $\mathbf{X}_i[J \times K_i]$. A moving window $\mathbf{X}_{i,k}[J \times V]$ of data in $\mathbf{X}_i$ (V being the moving window width) is slid along time, one time point at a time across all measurements, from $k = 1$ up to $k = K_i - V + 1$, so that each time point is

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

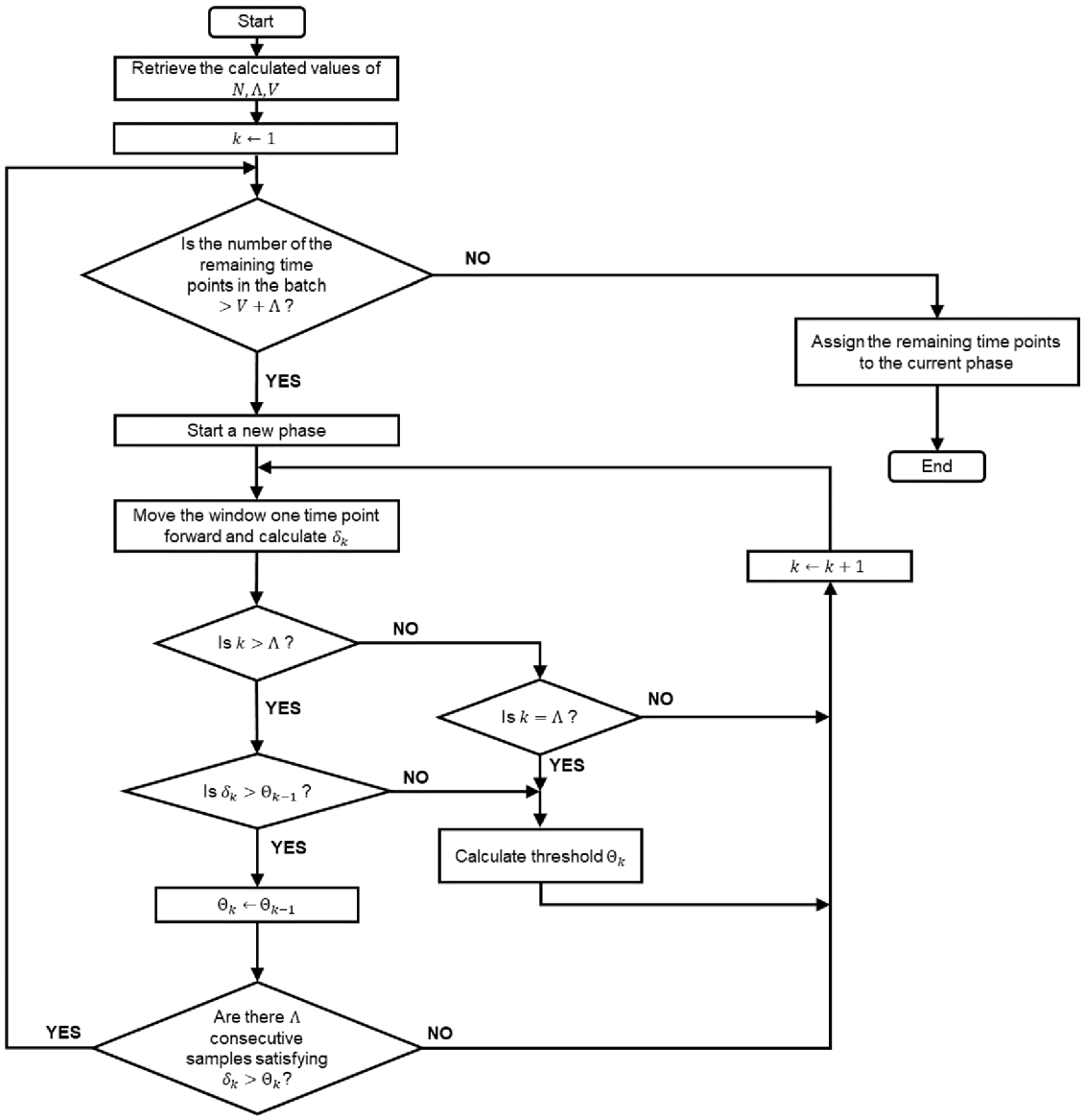


Figure 2. Flowchart of the original automatic phase partition methodology for one entire batch.

included in at least one of the windows. Consider the correlation matrix $\mathbf{C}_k\,[J \times J]$ of $\mathbf{X}_{i,k}$; its generic (p, q) element is calculated as

$$\mathbf{C}_k(p, q) = \frac{\mathrm{cov}(\mathbf{X}_{i,k}^{p}, \mathbf{X}_{i,k}^{q})}{\sigma(\mathbf{X}_{i,k}^{p})\sigma(\mathbf{X}_{i,k}^{q})} \tag{9}$$

where $\mathbf{X}_{i,k}^{p}$ and $\mathbf{X}_{i,k}^{q}$ are the pth and qth rows in $\mathbf{X}_{i,k}$ (respectively), $\mathrm{cov}(\mathbf{X}_{i,k}^{p}, \mathbf{X}_{i,k}^{q})$ is the covariance between the previously mentioned rows, and $\sigma(\mathbf{X}_{i,k}^{p})$ and $\sigma(\mathbf{X}_{i,k}^{q})$ are the standard deviations of the pth and qth rows in $\mathbf{X}_{i,k}$, respectively.

Define the multidimensional average gain index δ_k between two consecutive correlation matrices as

$$\delta_k = \frac{\sum_{p=1}^{J}\sum_{q=1}^{J}|\mathbf{C}_{k+1}(p, q) - \mathbf{C}_k(p, q)|}{J^2} \tag{10}$$

The gain captures the variation of the process characteristics (i.e., the change in correlation structure) between consecutive time points as the batch time progresses. A necessary condition to be fulfilled at time point k to trigger the switch from the current phase to a new one is that Λ consecutive values of δ_k exceed a threshold value Θ_k. If the last calculated value of δ_k does not exceed the threshold Θ_{k-1} calculated at the previous window slide, the threshold is calculated from the switch control limit σ_k, defined for a given phase as

$$\sigma_k = \frac{1}{\Lambda}\sum_{z=k-\Lambda+1}^{k}\delta_z \tag{11}$$

where Λ is the number of time points over which δ_k is averaged. The threshold Θ_k is calculated as

$$\Theta_k = N\sigma_k \tag{12}$$

where N is a parameter called tolerance factor. Otherwise, if $\delta_k > \Theta_{k-1}$, then the threshold Θ_k takes the same value as Θ_{k-1}.

To make phase switch actually occur, condition $\delta_k > \Theta_k$ must be satisfied for Λ consecutive time points. It can be shown that the minimum length of a phase that can be detected by this method is $(V + \Lambda)$ time points. A flowchart of the phase partition mechanism for a generic batch is shown in Figure 2.

For a given set of batches, the method requires assigning three adjustable parameters, namely, the moving window width V, the

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

width Λ over which the gains are averaged for each phase, and the tolerance factor N. The achieved phase partition strongly depends on the values assigned to the parameters; therefore, their search is best done by optimization.

3. PROPOSED OPTIMAL INDICATOR-VARIABLE SYNCHRONIZATION METHODOLOGY

Figure 3 shows a flowchart of the proposed IVopt approach for trajectory synchronization in uneven-length multiphase batch processes. The methodology iteratively adjusts the set $\boldsymbol{\xi} = [N, \Lambda, V]$ of parameters defining the partitioning of the available batches into phases until the resulting quality assessment model is optimal in some sense (to be discussed later). The distinguished features of the IVopt algorithm are the following:

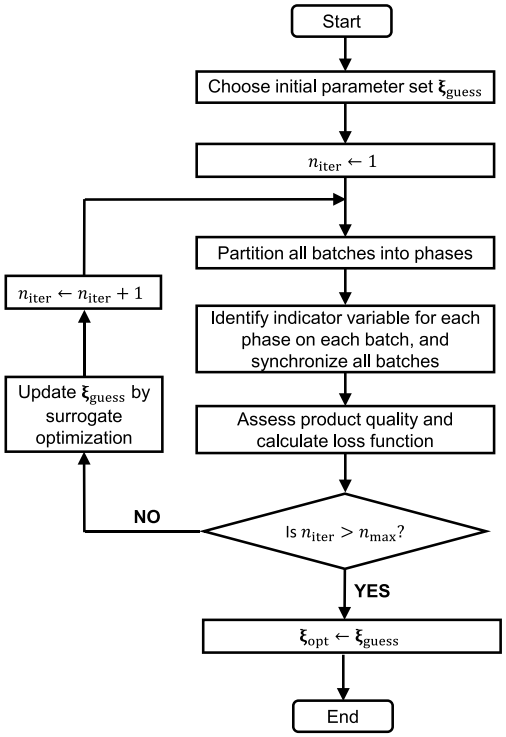


Figure 3. IVopt: flowchart of the proposed optimal indicator-variable approach for trajectory synchronization in uneven-length multiphase batch processes.

- Phase partition is targeted to optimal model-based quality assessment; namely, phase partition, batch synchronization, and quality assessment are not disjoint activities.
- The most appropriate IV for trajectory synchronization within each phase is identified automatically.
- A surrogate optimization approach is used to iteratively update the phase partition parameters.

Eventually, IVOpt returns both the optimal set $\boldsymbol{\xi}_{opt}$ of phase partition parameters and the most appropriate IVs to be used for batch synchronization. New batches can then be synchronized using this information. Next, we discuss the main steps along which the proposed methodology develops.

3.1. Automatic Phase Partition Revisited. Automatic phase partition is performed following the methodology illustrated in Section 2.2. However, we found that the use of a switch control limit as defined in eq 11 may suffer from noise when the signal-to-noise ratio is not high enough. To attenuate the impact of noise in phase identification, the information from the values of δ_k calculated after Λ movements of the moving window is included in the calculation of an adaptive control limit:

$$\sigma_k = \frac{1}{k}\sum_{z=1}^{k} \delta_z \tag{13}$$

which therefore includes not only the last Λ calculated values of δ_k but all the values from the beginning of the current phase. Upon application of the phase partition methodology to all batches in $\underline{\mathbf{X}}_{\mathcal{B}}$, the optimal set $\boldsymbol{\xi}_{opt}$ of phase partition parameters is found, from which the distribution of the number of phases identified across all batches is obtained. The mode of this distribution is set as the actual number $\bar{\Phi}$ of phases to be used for all batches. If, for a given batch i, the number of identified phases is $\Phi_i \neq \bar{\Phi}$, then that batch is forced to partitioning into $\bar{\Phi}$ phases by assigning phase switch time points equal to the average of the phase switch time points obtained for all the batches for which $\Phi_i = \bar{\Phi}$ is found. Note that this typically occurs for a limited number of batches only.

3.2. Automatic Indicator Variable Identification and Batch Synchronization. We propose a methodology that automatically returns an appropriate IV for each of the $\bar{\Phi}$ phases identified across the entire $\underline{\mathbf{X}}_{\mathcal{B}}$ data set. For a given phase, the methodology identifies a given process variable as a candidate IV if it simultaneously fulfills the following three conditions: (*i*) it is monotonic, (*ii*) it has a sufficiently high signal-to-noise ratio, and (*iii*) it has approximately the same initial and final values across all batches in $\underline{\mathbf{X}}_{\mathcal{B}}$. Next, we discuss how fulfillment of the conditions is assessed for a given phase and a given variable.

Condition (*i*) is assessed by performing a Mann−Kendall test for monotonicity.[40] The test returns a yes/no condition (at 0.05 significance level) to the null hypothesis that the trend of the given variable is nonmonotonic.

To assess if condition (*ii*) is fulfilled, the variable is first detrended by subtracting the best straight-line fit from the variable (as implemented in Matlab R2020a); then, the standard deviation of the detrended variable is compared to the range of the nondetrended one: if the standard deviation is smaller than the range, then the signal-to-noise ratio of the variable is deemed acceptable. An F test (at 0.05 significance level) is carried out to verify that the variance of the detrended variable and the variance of the nondetrended variable are statistically different.

Finally, condition (*iii*) is met for the variable under investigation if both the following inequalities are satisfied: $R > \sigma_i$ and $R > \sigma_e$, where R is the range of the variable values in the phase and σ_i and σ_e are the standard deviations of the phase initial and end points, respectively.

The actual IV among all of the identified candidate IVs for a given phase is selected as the one for which the Mann−Kendall test is satisfied more strongly (smallest average p value across all batches). If no process variable is identified as a candidate IV using the above approach for a given phase, time is used as the IV for that phase because time always fulfills the first two conditions and, in most cases, also the third one (at least to some approximation).

Once an IV is identified for each phase, the batches are synchronized in a phase-by-phase fashion using the IV approach.[11]

3.3. Product Quality Assessment and Loss Function Calculation. Product quality assessment is done through a soft sensor or a product classifier by building a PLS or a PLS-DA model (respectively) on the synchronized batches. The relevant

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

loss functions in 10-fold cross-validation (RMSECV and (1 − Accuracy), respectively) are then calculated.

3.4. Phase Partition Parameter Update by Surrogate Optimization. As noted earlier, the resulting phase partition (hence, the performance of the quality assessment model) strongly depends on the set $\boldsymbol{\xi} = [N, \Lambda, V]$ of parameters used within the phase partition methodology. The IVopt algorithm optimizes the selection of $\boldsymbol{\xi}$ by minimizing the loss function $\mathcal{L}(\xi)$ associated with the quality assessment model. The optimization problem can be formulated as

$$\min_{\xi} \mathcal{L}(\xi)$$

subject to:

$$\mathrm{lb}_n \le N \le \mathrm{ub}_N,\ \mathrm{lb}_\Lambda \le \Lambda \le \mathrm{ub}_\Lambda,\ \mathrm{lb}_\Lambda \le \Lambda \le \mathrm{ub}_\Lambda \tag{14}$$

where lb_z and ub_z denote the lower bound and the upper bound of parameter z.

We solve the optimization problem using surrogate optimization.[42] Surrogate optimization is a global optimization methodology especially useful when the objective functions are nonsmooth,[43] as occur for example when the optimization variables are discrete. One significant advantage of surrogate optimization is that it can be applied with an unknown symbolic form of the objective function and unknown exact derivatives of the function itself.[44]

A surrogate $\hat{\mathcal{L}}(\xi)$ is obtained that approximates the loss function, has a known analytical form, and is cheaper to evaluate with respect to the true objective function $\mathcal{L}(\xi)$. In this study, we use the radial basis function (RBF) interpolator, which has the form

$$\hat{\mathcal{L}}(\xi) = \sum_{h=1}^{H} \lambda_h \Psi(\|\xi - \bar{\xi}_h\|_2) + \mathcal{P}(\xi) \tag{15}$$

where H is the number of parameter sets for which the value of $\mathcal{L}$ is known and upon which the interpolation is made, $\bar{\boldsymbol{\xi}}_h$ is one such parameter set, the λ_i's are weights to be determined by calibration, $\Psi(\bullet)$ is the RBF, $\mathcal{P}$ is a polynomial whose coefficients are to be determined, and $\| \bullet \|_2$ is the Euclidean norm.[45] The RBF chosen in this study is the cubic one, which has been shown to outperform other surrogate models,[46] whereas the polynomial $\mathcal{P}$ has degree 1. This RBF has also been proven to minimize a measure of bumpiness.[42] Further information about RBFs and the solution of eq 15 is available in Appendix.

The optimization algorithm alternates between two phases: surrogate construction and minimum search.

The surrogate construction consists of these steps:

1. A_1 quasirandom input vectors (i.e., parameter sets) are sampled within the bounds.
2. $\mathcal{L}$ is evaluated on these points. The minimum value of the objective function among these points is identified as the "incumbent".
3. $\hat{\mathcal{L}}$ is calibrated on the values of $\mathcal{L}$ obtained at point 2.

After these steps, the minimum search starts:

4. A_2 input vectors are sampled in the input space close to the incumbent.
5. $\hat{\mathcal{L}}$ is evaluated on the points identified in step 4.
6. The point with minimum $\hat{\mathcal{L}}$ in step 5 is identified. This point is added to the initial points of step 1, and the algorithm iterates back to step 2 until an assigned number of iterations is reached.

The minimum search problem involves both real (N) and integer (Λ; V) variables, and the optimization step must therefore be adjusted to account for this. In this study, we use a variant of the branch-and-bound mixed-integer optimization algorithm proposed by Achterberg et al.[47] as implemented in the `surrogateopt` function available in the global optimization toolbox included in Matlab R2020a.[41]

4. CASE STUDIES

Two case studies are considered to test the proposed IVopt framework: an industrial batch process for the manufacturing of a specialty chemical and a simulated fed-batch process for the manufacturing of penicillin. Next, we provide details about them.

4.1. Case Study #1: Industrial Fed-Batch Manufacturing of a Specialty Chemical. Figure 4 shows the simplified process flow diagram of the industrial fed-batch process under investigation, where product P (actually, an intermediate used in the manufacturing of a polymer stabilizer) is obtained in jacketed reactor R1 (6.5 m^3 volume) from the following catalytic reaction:

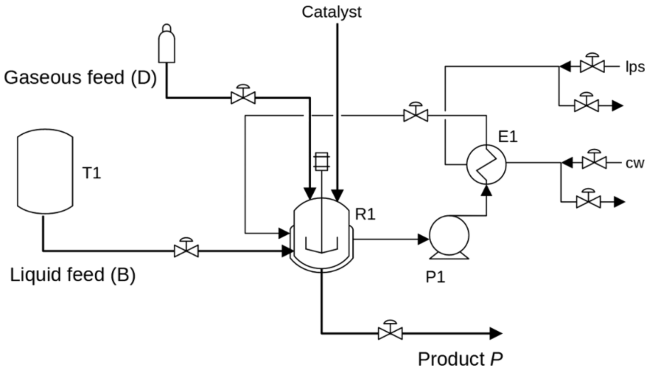


Figure 4. Case study #1: simplified process flow diagram of the industrial process for the manufacturing of a specialty chemical.

$$\mathrm{B} + \mathrm{D} \xrightarrow{\mathrm{Cat.}} \mathrm{G} \tag{16}$$

where species B is a liquid reactant, species D is a gaseous reactant, and G is the desired species. Product P is mainly made of G, traces of unreacted B, and other subproducts. The manufacturing recipe is quite complex and can be summarized by the following finite-length operating steps:

1. Reactor R1 is set up for a new batch.
2. Reactant B and catalyst are loaded into R1.
3. R1 is blanketed with nitrogen.
4. Reactant D is fed to R1 and pressurizes it until an assigned pressure is reached; after that, the feed is stopped, and the reaction is allowed to proceed for an assigned amount of time. The profile through which B is fed depends on several factors and is quite complex, resulting in a very strong variability of this phase.
5. R1 is vented.
6. R1 is blanketed with nitrogen.
7. Product P is discharged from R1 to a downstream plant section, where it is further processed.

Too large an amount of unreacted B in P can be an issue for quality because P is used as a reactant in a downstream unit and

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

an excess of B can downgrade the optical properties of the final product. A lab assay of *P* is taken for some batches only. Real-time measurements of some process variables are available, as listed in Table 1.

Table 1. Case Study #1: Variables Measured in Real Time

variable no.	variable name
1	totalized reactant D fed
2	reactant D flow rate
3	reactant D flow rate controller output
4	R1 internal absolute pressure
5	R1 internal pressure controller output
6	R1 internal pressure controller 2 output
7	R1 internal temperature
8	R1 internal temperature controller output
9	R1 internal temperature difference controller output
10	time

From the data historian, a set of 52 batches (completed across years 2020 and 2021) is collected for which the end-point quality (in terms of concentration of B in *P*) is measured. Within this data set, the batch length ranges between 7 and 19 h. The measured variables are down-sampled to one every 2 min. The data set is split into 36 calibration batches (24 batches ending up in a "good" product and 12 batches ending up in a "bad" product) and 16 validation batches (10/6 good/bad).

The quality assessment model is required to classify the quality of product *P* as either good or bad, once a batch has come to an end, using the time-resolved measurements of the variables listed in Table 2. A multiway PLS-DA model is developed for this purpose.

Table 2. Case Study #2: Variables Measured in Real Time[a]

variable no.	variable name	units	*stdev*
1	dissolved oxygen	g/L	0.0067
2	bulk volume	L	0.033
3	pH	[-]	0.0167
4	temperature	K	0.17
5	glucose feed rate	L/h	0.17
6	aeration rate	L/h	0.0834
7	agitator power	W	0.17
8	glucose feed temperature	K	0.17
9	jacket water flow rate	L/h	0.83
10	cumulated base flow	L	3.33×10^{-6}
11	cumulated acid flow	L	3.33×10^{-7}

[a]*stdev* is the standard deviation of a zero-mean normal distribution of random numbers.

4.2. Case Study #2: Simulated Fed-Batch Manufacturing of Penicillin. We consider a fed-batch fermentation process that manufactures penicillin. The process is simulated using Pensim,[38] a software used in several process control and monitoring studies.[38,48,49] Figure 5 shows a simplified piping and instrumentation diagram of the process.

The penicillin manufacturing recipe is based on two processing steps, as follows:

1. A batch culture step, where the reactor is initially loaded with *Penicillium chrysogenum* and glucose from tank T4 and the reaction starts. This step ends when the concentration of glucose in reactor R2 drops below an assigned threshold.

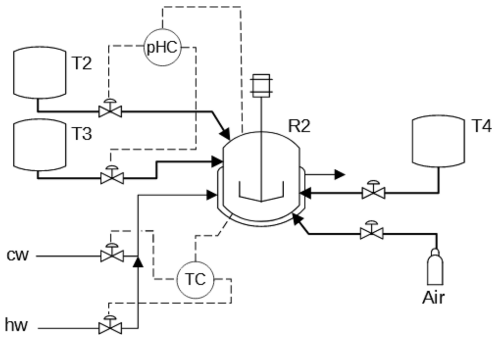


Figure 5. Case study #2: simplified piping and instrumentation of the simulated process for the manufacturing of penicillin.

2. A fed-batch step, where pH is automatically controlled through the addition of acid from tank T2 and base from tank T3. During this step, glucose and air are fed constant rates. The end-point condition is reached when the total volume of glucose fed to R2 during this step reaches 14 L.[50]

Real-time measurements of some process variables are available, as listed in Table 2.

Measurement noise is simulated in the form of additive random numbers sampled from a normal distribution with zero mean and standard deviation (*stdev*) as indicated in Table 2.[51] Process variability is generated by randomly changing the values of some initial conditions and some operating variables, as detailed in Table 3. Further variability is generated by assuming

Table 3. Case Study #2: Nominal Initial Conditions, Nominal Operating Variables, and Variability around Them (ϵ Is Sampled from a Standard Normal Distribution)

initial condition	units	nominal value
glucose concentration	g/L	$15 + \epsilon$
dissolved oxygen	%	1.16
biomass concentration	g/L	0.1
penicillin concentration	g/L	0
culture volume	L	$150 + 10\epsilon$
CO_2 concentration	mmol/L	$0.75 + 0.05\epsilon$
hydrogen ion concentration	mol/L	$10^{-5+0.1\epsilon}$
fermentor temperature	K	298
generated heat	kcal/h	0
operating variable	**units**	**nominal value**
aeration rate	L/h	8
agitator power	W	$30 + \epsilon$
glucose feed rate	L/h	$0.04 + 0.0025\epsilon$
glucose feed temperature	K	296
culture volume	L	$150 + 10\epsilon$
pH	[-]	5
fermentor temperature	K	298

that the threshold glucose concentration determining the switch between operating steps 1 and 2 randomly varies between 0.3 and 7 g/L.

A set of 300 batches is generated. Within this data set, the batch length ranges between 345 and 479 h. The variables measured in real time are sampled every 0.5 h. The data set is split into 250 calibration batches and 50 validation batches.

The quality assessment model for this case study is required to estimate, at the end of a batch, the end-point penicillin

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

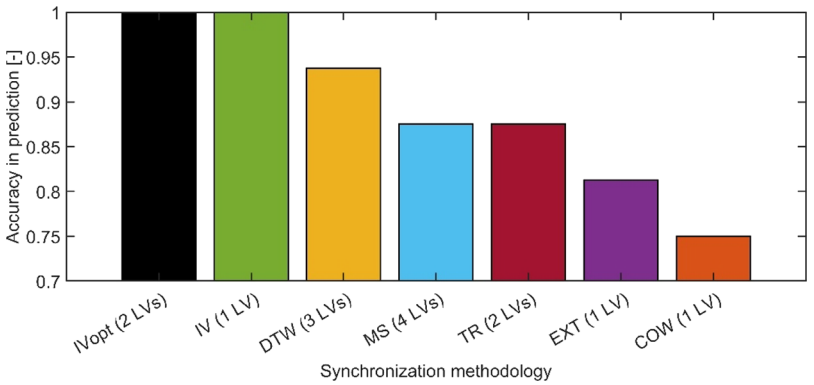


Figure 6. Case study #1. Classification accuracy obtained in prediction using the validation data set for different batch synchronization methodologies (the numbers in parentheses indicate the optimal number of latent variables as determined by cross-validation using the calibration data set).

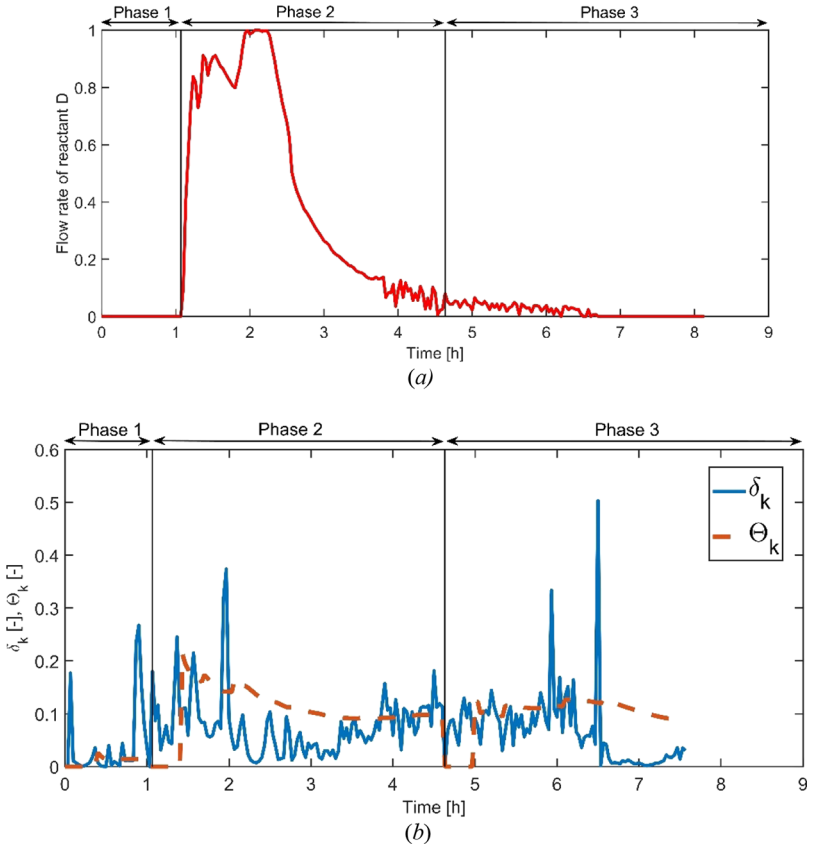


Figure 7. Case study #1, representative batch. (a) Phase partition obtained by the IVopt methodology: time profiles of the (dimensionless) flow rate of reactant D; (b) time evolution of the δ_k and Θ_k parameters.

concentration by using the time-resolved measurements in Table 2. A multiway PLS model is developed to this purpose.

5. RESULTS

We benchmark IVopt against DTW, COW, TR, EXT, IV, and MS for the two case studies illustrated in the previous section. The batch synchronization methodologies are compared in terms of the performance of the relevant quality assessment model and computational cost. For a given batch, IVopt is applied using the automatic phase partition methodology as discussed in Section 3.1; DTW, COW, MS, TR, and EXT are applied without any phase partition, and IV is applied by partitioning a batch into processing steps rather than into phases. The computation time refers to the use of a laptop computer equipped with an Intel Core i7-9750H 2.60 GHz CPU and 32 GB of RAM.

5.1. Results for Case Study #1. With reference to IVopt, the initial phase partition parameter guesses are $N = 1.08$, $\Lambda = 11$, and $V = 10$, and the following constraints are enforced in the surrogate optimization algorithm: $1 \leq N \leq 4$, $2 \leq \Lambda \leq 15$, and $5 \leq V \leq 35$. The maximum number of iterations is set to 100. The algorithm returns a 10-fold cross-validation accuracy of 92% (with two latent variables) in the calibration data set at $N = 2.3$, $\Lambda = 5$, and $V = 15$.

Using the optimal phase partition parameter set, the classification accuracy for the validation data set is 100%, meaning that all validation batches are classified correctly. A comparison of the product quality classification results obtained

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

for the validation data set for the batch synchronization methods considered in this study is shown in Figure 6 (the reported optimal number of latent variables is determined by cross-validation using the calibration data set). IVopt outperforms all other synchronization methodologies except IV (which, however, requires manually assigning both the phase partitioning and the indicator variable within each phase). Some synchronization methods (namely, EXT and COW) lead to poor classification accuracy.

Despite the fact that as many as seven operating stages exist, IVopt returns a batch partitioning into only three phases. Figure 7a shows the time profile of the flow of reactant D to reactor R1 for a representative batch together with the phase partitioning returned by IVopt. The partitioning is physically meaningful: the phase switching points correspond roughly to the time points when the flow rate of D starts to be greater than zero and then returns close to zero. The automatically selected indicator variables are time for phase 1 and the totalized amount of reactant D fed to R1 for both phases 2 and 3. The time profiles of the gain index δ_k and its threshold value Θ_k are illustrated in Figure 7b. It can be seen that both of them get adjusted as the batch progresses; when δ_k values are consistently greater than the corresponding values of Θ_k, a phase switch occurs.

Figure 8 compares the computer time required to perform batch synchronization for all methods. Although the time required by IVopt is significantly greater than the one required by any of the other methods, it is nevertheless very short (5 min). Computationally demanding methods like DTW and MS require less than 1 min to run (hence, much less than IVopt) because the data set to be analyzed has quite a small size (on average, ~250 samples per measured variable in a batch). However, these methods scale badly with the data set size, as will be shown for case study #2.

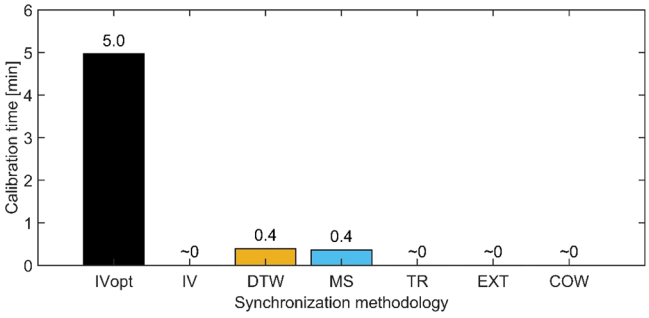


Figure 8. Case study #1. Time required to carry out the synchronization of the calibration data set trajectories using different techniques.

The synchronization results are illustrated in Figure 9 for all synchronization methods with reference to the profiles of the flow rate of reactant D (variable no. 2) across the calibration batches. The unsynchronized trajectories are shown in Figure 9a. It can be seen that the COW (Figure 9b) struggles to obtain an effective synchronization for some batches. DTW (Figure 9c) and MS (Figure 9d) effectively minimized the differences between the trajectories. However, to achieve this, they introduce distortions in some trajectory segments, particularly when a strong compression is applied; these distortions appear as horizontal segments for several trajectories, approximately located between synchronized time 50 and 100. IV (Figure 9e) and IVopt (Figure 9f) work differently from the other synchronization methods. Recall that IVopt selects the totalized volume of reactant D as the indicator variable during phases 2 and 3. This indicator variable is basically the time integral of the variable shown in Figure 9a. Therefore, the portions of a trajectory with larger values of the reactant D flow rate within phases 2 and 3 are expanded in time (thus magnifying the trajectory differences across batches) to maximize the classification accuracy; on the other hand, the portions with smaller values are contracted (minimizing such differences), as they have less impact on the classification. Therefore, when using IVopt (Figure 9f), time is substituted with an indicator variable that is nonlinearly related to time itself and is more descriptive of the progress of the process. The IV method (Figure 9e) works somewhat similarly to IVopt, resulting in a similar classification performance. Yet, IVopt performs all operations (phase partition and indicator variable selection) automatically.

5.2. Results for Case Study #2. The optimization is carried out using $N = 1.8$, $\Lambda = 2$, and $V = 30$ as initial guesses and the following box constraints: $1 \le N \le 3$, $1 \le \lambda \le 5$, and $5 \le V \le 100$. The surrogate optimization algorithm iterates 300 times, yielding the following optimal parameter set: $N = 1.02$, $\Lambda = 1$, and $V = 98$.

Figure 10 shows that IVopt provides the best validation results among all synchronization methods: the root-mean squared error of prediction is 0.0057 g/L, slightly better than with DTW and COW and considerably better than with MS and EXT.

IVopt identifies three phases (Figure 11a). In the (very short) first phase, time is selected as the indicator variable, whereas in the second and third phases, the cumulated base flow rate is selected as the indicator variable. Figure 11 suggests that abrupt changes in pH correspond to large variations in the correlation structure of the measured variables (hence, to phase switch), as captured by the time profiles of δ_k and Θ_k.

Figure 12 clarifies that, whereas the computational time required by IVopt is not negligible (~30 min), it is slightly smaller than the one required by DTW (~40 min) and much smaller than that required by MS (~39 h). Comparing Figures 8 and 12, in the face of a data set size increase from ~90k data entries (case study #1) to ~2M data entries (case study #2), the computational time required by IVopt increased by ~6 times, whereas DTW increased by ~100 times and MS by ~5900 times. Therefore, IVopt scales with the data set size much better than these other two advanced synchronization methodologies.

We assessed the sensitivity of the model performance to the number of calibration batches for all of the synchronization methods. Figure 13 clarifies that using a smaller number of calibration batches leads to a minor loss of performance for all methods, unless very few (namely, 20) calibration batches are used. However, IVopt also outperforms the other methods in this limiting case.

To evaluate the robustness of the phase partition parameters against the inherent variability in the data, we tested the optimization results on 50 different (random) splits of the available data into calibration/validation data sets. The distribution of the optimal parameters is illustrated in Figure 14.

It is observed that N (a real number) has a very narrow distribution of around 1.0. Parameters Λ and V (natural numbers) exhibit slightly greater variability, but in most cases, Λ is either 1 or 2, and V is around 100. We conclude that the results obtained by IVopt are robust to typical fluctuations in the data.

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

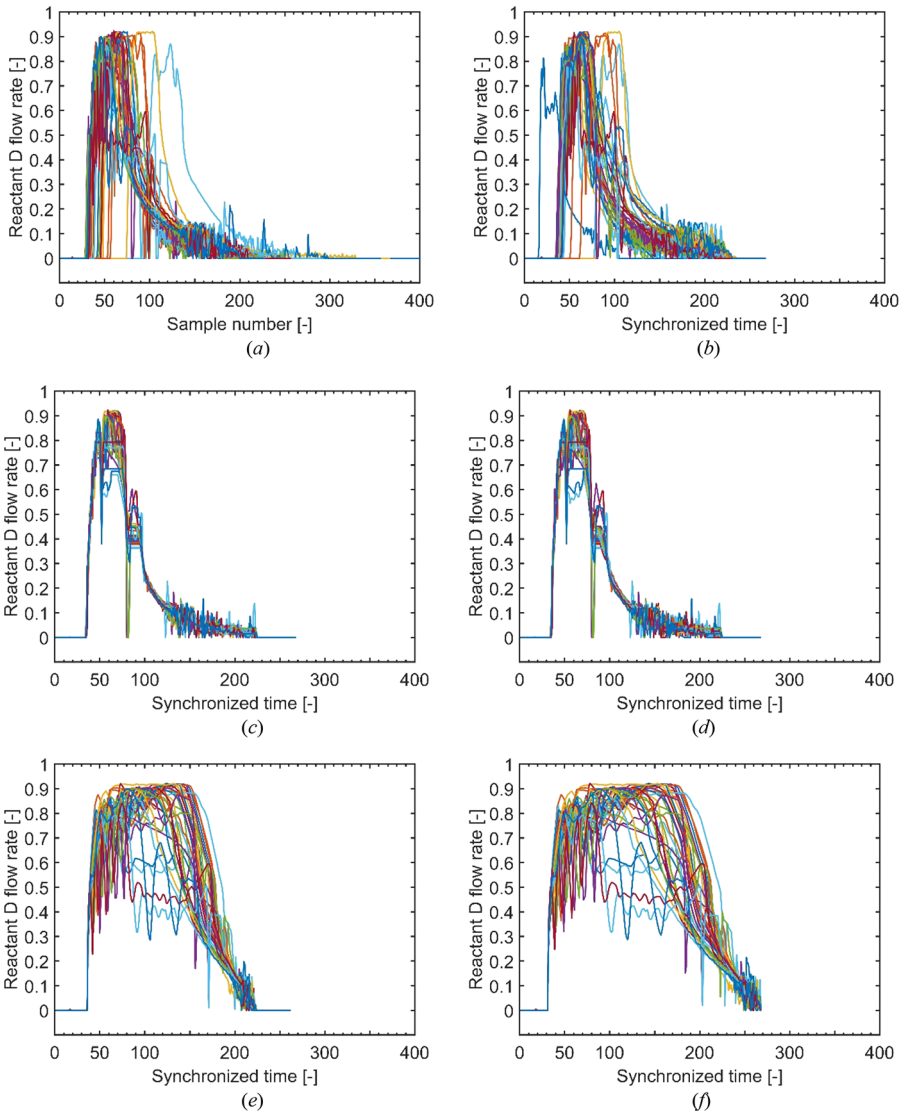


Figure 9. Case study #1. Time trajectories of the (dimensionless) flow rate of reactant D (a) without synchronization and synchronized using (b) COW, (c) DTW, (d) MS, (e) IV, and (f) IVopt.

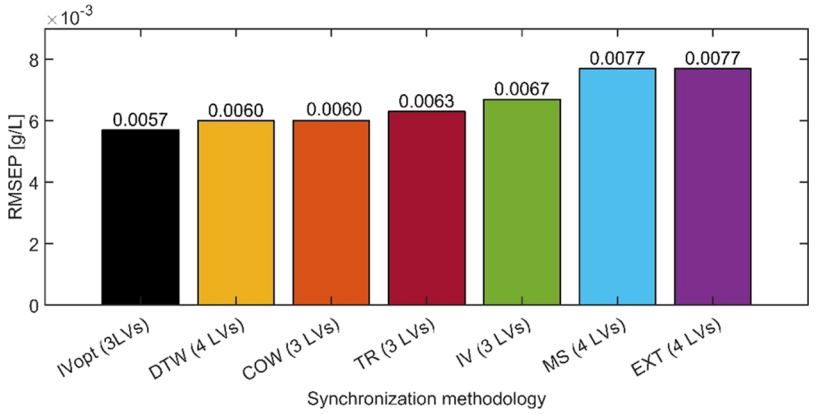


Figure 10. Case study #2. Root-mean squared prediction error using the validation data set for different batch synchronization methodologies (the numbers in parentheses indicate the optimal number of latent variables as determined by cross-validation using the calibration data set).

6. CONCLUSIONS

This paper presents a novel methodology (called IVopt) for phase partitioning and trajectory synchronization in uneven-length multiphase batch processes. The methodology retains the effectiveness of a simple trajectory synchronization methodology like the classic indicator variable (IV) approach but improves it in two directions, namely, partitioning into phases and selection of the most appropriate IV within each phase (*i*)

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

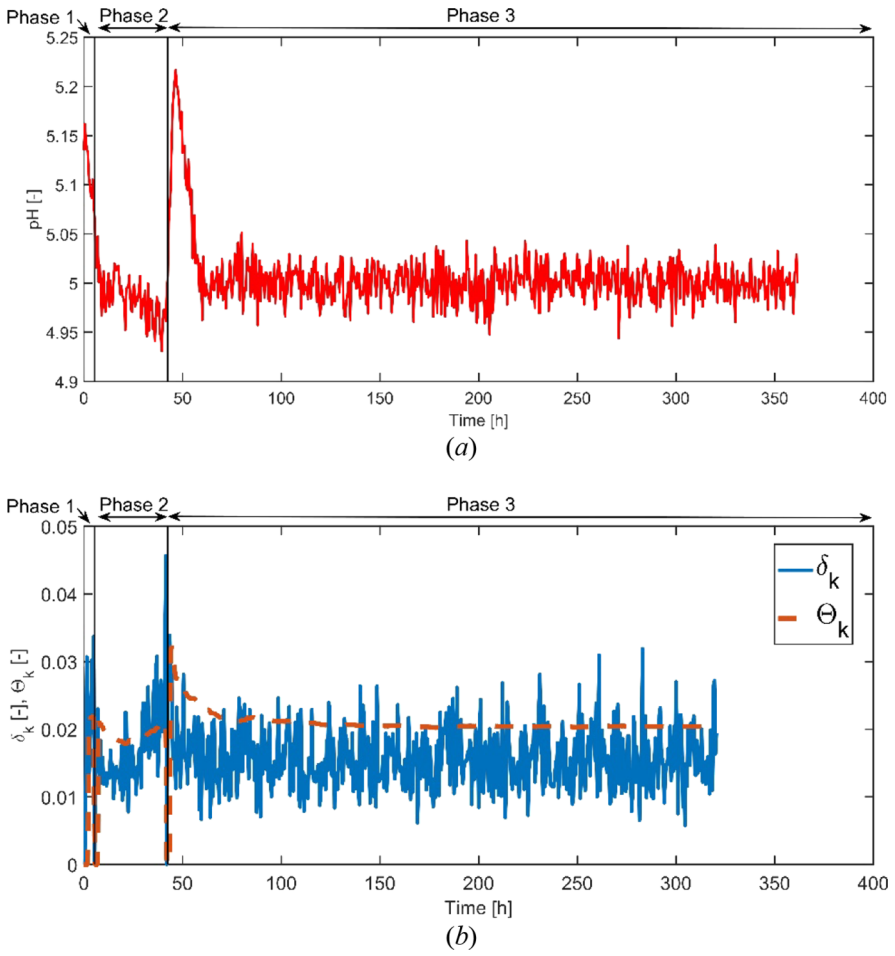


Figure 11. Case study #2, representative batch. (a) Phase partition obtained by the IVopt methodology: time profiles of pH; (b) time evolution of the δ_k and Θ_k parameters.

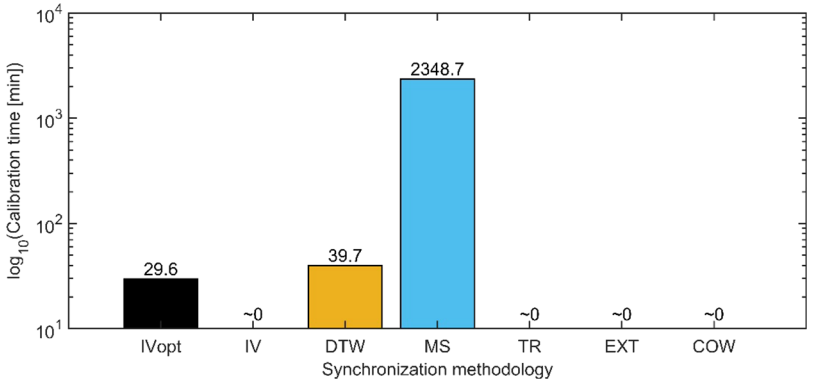


Figure 12. Case study #2. Computation time required to carry out the synchronization of the calibration data set trajectories using different synchronization techniques.

are performed automatically rather than manually and (*ii*) are carried out simultaneously rather than disjointly based on an optimization framework that maximizes the performance of a model for product quality assessment that is to be built using the available data sets. Differently from classic IV and from advanced synchronization methodologies like dynamic time warping (DTW), correlation optimized warping (COW), and multi-synchro (MS), the proposed methodology is process-agnostic; i.e., it does not require identifying either a reference batch or a reference variable.

To test the proposed data preprocessing methodology, we considered two data sets: one from an industrial process and one from a simulated process. We compared the performance of the resulting product quality assessment model when the available data were preprocessed with IVopt and with other synchronization strategies, namely, trajectory truncation (TR), trajectory extension (EXT) with mean values, classic IV, DTW, COW, and MS. IVopt always led to the best model performance both when the model was used as a soft sensor to estimate the product end-point quality and when the model was used as a classifier to

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

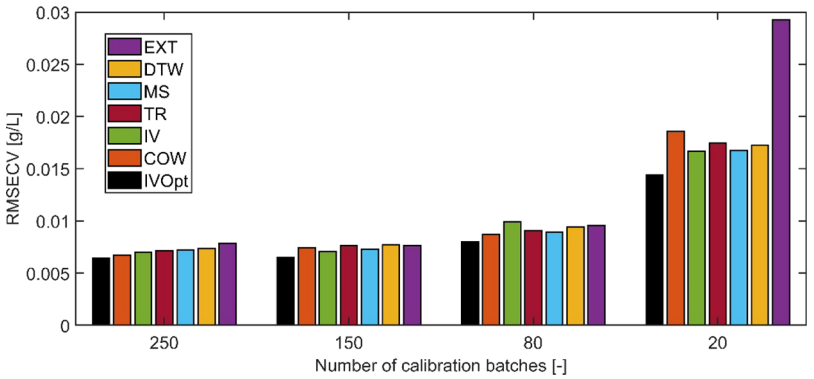


Figure 13. Case study #2. Impact of the number of calibration batches on the root-mean squared prediction error using the validation data set for different batch synchronization methodologies.

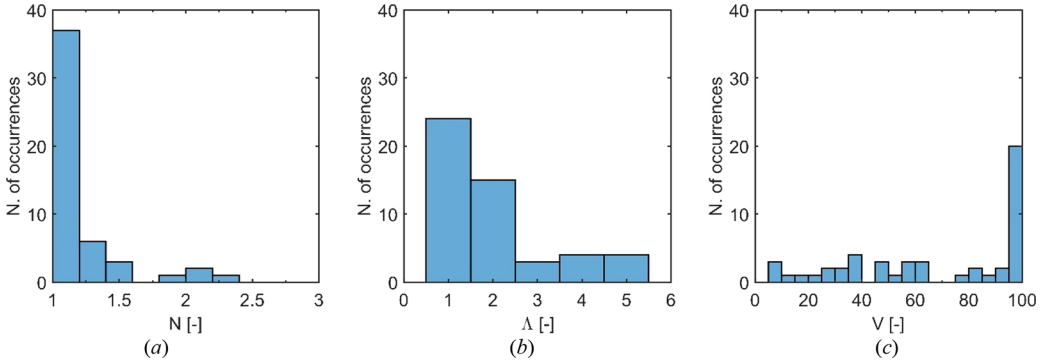


Figure 14. (a) Case study #2. Distribution of the optimal phase portioning parameters for 50 different splits of the calibration/validation data sets: (a) N, (b) Λ, and (c) V.

discriminate between on-spec and an off-spec products. In this latter case, the classic IV also led to excellent classification performance, but partitioning into phases and selection of the most appropriate IV within each phase had to be done manually based on engineering judgment.

From the computational side, IVopt is more demanding than simple (and less effective) strategies like IV, TR and EXT and also than COW. Compared to other advanced methodologies, whenever DTW and MS become computationally intensive, IVopt outperforms them, more so as the number of samples per batch increases.

APPENDIX A

Data Interpolation with Radial Basis Functions

In this Appendix, we provide a short overview of radial basis functions (RBFs) and some details on how to obtain the solution of eq 15. More details can be found in specialized references.[52,53]

RBFs have found applications in several domains, such as computer graphics,[54] predictive maintenance,[55] and chemometrics,[56] most frequently for scattered data interpolation The data interpolation problem can be stated as follows: given H multidimensional data points $\bar{\boldsymbol{\xi}}_h$ (with h = 1,2, ..., H), with corresponding scalar values $\mathcal{L}(\bar{\boldsymbol{\xi}}_h)$, compute a function $\hat{\mathcal{L}}(\boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a generic point belonging to the same space to which the data points $\bar{\boldsymbol{\xi}}_h$ belong, that smoothly interpolates the data points and for which $\mathcal{L}(\bar{\boldsymbol{\xi}}_h) = \hat{\mathcal{L}}(\bar{\boldsymbol{\xi}}_h)$ for all the values of h.

To carry out this task, a function $\Psi(\bullet)$ of the distance between $\bar{\boldsymbol{\xi}}_h$ and $\boldsymbol{\xi}$, called RBF, is used for generalizing the concept that the closer we get to a certain data point $\bar{\boldsymbol{\xi}}_h$, the closer the value of $\hat{\mathcal{L}}$ should get to $\mathcal{L}(\bar{\boldsymbol{\xi}}_h)$. A cubic RBF Ψ is defined as

$$\Psi(\|\xi - \bar{\boldsymbol{\xi}}_h\|_2) = \|\xi - \bar{\boldsymbol{\xi}}_h\|_2^3 \tag{A1}$$

where $\| \bullet \|_2$ is the Euclidean norm. Thus, the RBF-based interpolator takes the form

$$\hat{\mathcal{L}}(\xi) = \sum_{h=1}^{H} \lambda_h \Psi(\bullet) \tag{A2}$$

Solving this equation consists in solving the following linear system:

$$\begin{pmatrix} \Psi_{1,1} & \Psi_{1,2} & \cdots & \Psi_{1,H} \\ \Psi_{2,1} & \Psi_{2,2} & \cdots & \Psi_{2,H} \\ \vdots & \vdots & \ddots & \\ \Psi_{H,1} & \cdots & & \Psi_{H,H} \end{pmatrix} \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_H \end{pmatrix} = \begin{pmatrix} \mathcal{L}(\bar{\boldsymbol{\xi}}_1) \\ \mathcal{L}(\bar{\boldsymbol{\xi}}_2) \\ \vdots \\ \mathcal{L}(\bar{\boldsymbol{\xi}}_H) \end{pmatrix} \Rightarrow \Psi\lambda = \boldsymbol{\mathcal{L}} \tag{A3}$$

where the element in position (i, j) of matrix $\boldsymbol{\Psi}$ is $\Psi_{i,j} = \Psi(\|\bar{\boldsymbol{\xi}}_i - \bar{\boldsymbol{\xi}}_j\|_2)$ and the unknowns are the weights λ_h's.

One disadvantage of the RBF interpolator in eq A2 is that it is unable to represent polynomial functions. To make it able to approximate polynomial functions, a polynomial function $\mathcal{P}(\boldsymbol{\xi})$ is appended to the right-hand side of eq A2. For example, a linear polynomial can be used:

$$\mathcal{P}(\xi) = c_1 + c_2\xi \tag{A4}$$

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

where c_1 and c_2 are the parameters of the polynomial. Thus, we obtain the final form of the RBF-based interpolator:

$$\hat{\mathcal{L}}(\xi) = \sum_{h=1}^{H} \lambda_i \Psi + \mathcal{P}(\xi) \tag{A5}$$

Let $\mathcal{B}$ be the basis of $\mathcal{P}$:

$$\mathcal{B} = \begin{pmatrix} 1 & \overline{\xi}_1 \\ 1 & \overline{\xi}_2 \\ \vdots & \vdots \\ 1 & \overline{\xi}_H \end{pmatrix} \tag{A6}$$

The linear system to be solved becomes

$$\Psi\lambda + \mathcal{B}\mathbf{c} = \mathcal{L} \tag{A7}$$

where $\mathbf{c} = (c_1 \; c_2)^{\mathrm{T}}$. However, the system is now underdetermined. To be able to solve eq A7, we constrain the weights $\boldsymbol{\lambda}$ to be zero if the polynomial terms match the data points exactly with the coefficients $\mathbf{d}$:

$$\Psi\lambda + \mathcal{B}\mathbf{c} = \mathcal{B}\mathbf{d} \tag{A8}$$

After a few algebraic manipulations, we end up with the following linear system:[52]

$$\begin{pmatrix} \Psi & \mathcal{B} \\ \mathcal{B}^{\mathrm{T}} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \lambda \\ \mathbf{c} \end{pmatrix} = \begin{pmatrix} \mathcal{L} \\ \mathbf{0} \end{pmatrix} \tag{A9}$$

which can easily be solved by linear algebra.

AUTHOR INFORMATION

Corresponding Author

Massimiliano Barolo − *CAPE-Lab−Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, 35131 Padova, PD, Italy*; orcid.org/0000-0002-8125-5704; Email: max.barolo@unipd.it

Authors

Francesco Sartori − *CAPE-Lab−Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, 35131 Padova, PD, Italy*

Pierantonio Facco − *CAPE-Lab−Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, 35131 Padova, PD, Italy*

Federico Zuecco − *BASF Italia SpA, 40037 Pontecchio Marconi, BO, Italy*

Fabrizio Bezzo − *CAPE-Lab−Computer-Aided Process Engineering Laboratory, Department of Industrial Engineering, University of Padova, 35131 Padova, PD, Italy*; orcid.org/0000-0003-1561-0584

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.3c01897

Funding

The research reported in this paper was funded by BASF Italia S.p.A. under contract "APP 4.0−Approaching process reliability through the smart use of process sensor data in the Industry 4.0 era".

Notes

The authors declare no competing financial interest.

LIST OF SYMBOLS

A	number of latent variables
A_1, A_2	sets of ξ quasirandomly sampled during surrogate optimization
Accuracy	accuracy in classification
B	reactant in case study 1
C	number of samples in the proposed version of the phase partition threshold
C_c	correctly classified samples
$\mathbf{C}_k$	correlation matrix
D	reactant in case study 1
$\mathbf{E}$	$\mathbf{X}$ residual matrix
$\mathbf{F}$	$\mathbf{Y}$ residual matrix
G	product in case study 1
H	number of data points for building the radial basis function interpolant
I	number of batches
J	number of predictor variables
K	number of time samples
L	number of response variables
$\mathcal{L}$	objective function
$\hat{\mathcal{L}}$	surrogate objective function
lb	lower bounds
N	phase partition threshold adjustable parameter
n_{out}	number of δ_k values exceeding Θ_{k-1}
n_{iter}	number of iterations
P	product mixture in case study 1
$\mathbf{P}$	$\mathbf{X}$ loading matrix
p	row index
$\mathcal{P}$	radial basis function interpolant polynomial
$\mathbf{Q}$	$\mathbf{Y}$ loading matrix
q	column index
Q^2	coefficient of determination in cross-validation
R^2	coefficient of determination
$\mathbf{T}$	$\mathbf{X}$ score matrix
t_{p}	predicted scores for a new sample
$\mathbf{U}$	$\mathbf{Y}$ score matrix
ub	upper bounds
V	Window width for phase partition algorithm
$\mathbf{W}$	weight matrix
$\mathbf{X}$	regressor matrix
$\underline{\mathbf{X}}_{\mathcal{B}}$	trensor of regressors (batch processes)
$\mathbf{X}_i$	matrix of data from batch i
$\mathbf{X}_{i,k}$	data window at time point k in batch i
$\boldsymbol{x}_{\mathrm{p}}$	new sample regressors
$\mathbf{Y}$	response matrix
$\hat{y}_{\mathrm{p}}$	vector of PLS predicted responses
$\mathbf{Y}_{\mathrm{B}}$	matrix of responses (batch responses)
$\mathbf{Y}_{\mathrm{c}}$	class response matrix
δ_k	multidimensional average gain index
Θ	switch control limit threshold
Θ_k	adaptive witch control limit threshold
Λ	optimizable parameter
λ	radial basis function interpolant weights
$\underline{\xi}$	set containing N, Λ, V
$\underline{\xi}_h$	set N, Λ, V on which the value of $\mathcal{L}$ is known and that is used for building the radial basis function interpolant
ξ_{guess}	initial guess ξ set
ξ_{opt}	optimal ξ set
σ	switch control limit
σ_k	adaptive switch control limit

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

$\bar{\Phi}$	number of phases on the overall data set
Φ_i	number of phases recognized in each batch
Ψ	basis function used for building the radial basis function interpolant
COW	correlation optimized warping
DTW	dynamic time warping
EXT	extension with mean values
IV	indicator variable
IVopt	indicator variable optimization
MS	multisynchro
PLS	projection onto latent structures
PLS-DA	projection onto latent structures and discriminant analysis
RBF	radial basis function
RMSECV	root-mean-square error of cross-validation
RMSEP	root-mean-square error in prediction
TR	truncation
cw	cooling water
lps	low pressure steam

REFERENCES

(1) Geladi, P.; Kowalski, B. R. Partial Least-Squares Regression: A Tutorial. *Anal. Chim. Acta* **1986**, *185*, 1−17.

(2) Wold, S.; Ruhe, A.; Wold, H.; Dunn, W. J., III The Collinearity Problem in Linear Regression, the Partial Least Squares (PLS) Approach to Generalized Inverses. *SIAM J. Sci. Stat. Comput.* **1984**, *5*, 735−743.

(3) MacGregor, J. F.; Jaeckle, C.; Kiparissides, C.; Koutoudi, M. Process Monitoring and Diagnosis by Multiblock PLS Methods. *AIChE J.* **1994**, *40* (5), 826−838.

(4) Barker, M.; Rayens, W. Partial Least Squares for Discrimination. *J. Chemom.* **2003**, *17* (3), 166−173.

(5) Nomikos, P.; MacGregor, J. F. Multi-Way Partial Least Squares in Monitoring Batch Processes. *Chemom. Intell. Lab. Syst.* **1995**, *30* (1), 97−108.

(6) Rendall, R.; Chiang, L. H.; Reis, M. S. Data-Driven Methods for Batch Data Analysis − A Critical Overview and Mapping on the Complexity Scale. *Comput. Chem. Eng.* **2019**, *124*, 1−13.

(7) González-Martínez, J. M.; Camacho, J.; Ferrer, A. MVBatch: A Matlab Toolbox for Batch Process Modeling and Monitoring. *Chemom. Intell. Lab. Syst.* **2018**, *183* (July), 122−133.

(8) Rothwell, S. G.; Martin, E. B.; Morris, A. J. Comparison of Methods for Dealing with Uneven Length Batches. *IFAC Proc.* **1998**, *31* (8), 387−392.

(9) Lakshminarayanan, S.; Gudi, R. D.; Shah, S. L.; Nandakumar, K. Monitoring Batch Processes Using Multivariate Statistical Tools: Extensions and Practical Issues. *IFAC Proc.* **1996**, *29* (1), 6037−6042.

(10) Nomikos, P.; MacGregor, J. F. Monitoring Batch Processes Using Multiway Principal Component Analysis. *AIChE J.* **1994**, *40* (8), 1361−1375.

(11) García-Muñoz, S.; Kourti, T.; MacGregor, J. F.; Mateos, A. G.; Murphy, G. Troubleshooting of an Industrial Batch Process Using Multivariate Methods. *Ind. Eng. Chem. Res.* **2003**, *42* (15), 3592−3601.

(12) Ündey, C.; Ertunç, S.; Çinar, A. Online Batch/Fed-Batch Process Performance Monitoring, Quality Prediction, and Variable-Contribution Analysis for Diagnosis. *Ind. Eng. Chem. Res.* **2003**, *42* (20), 4645−4658.

(13) Kourti, T. Multivariate Dynamic Data Modeling for Analysis and Statistical Process Control of Batch Processes, Start-Ups and Grade Transitions. *J. Chemom.* **2003**, *17* (1), 93−109.

(14) García-Muñoz, S.; Polizzi, M.; Prpich, A.; Strain, C.; Lalonde, A.; Negron, V. Experiences in Batch Trajectory Alignment for Pharmaceutical Process Improvement through Multivariate Latent Variable Modelling. *J. Process Control* **2011**, *21* (10), 1370−1377.

(15) Barton, M.; Duran-Villalobos, C. A.; Lennox, B. Multivariate Batch to Batch Optimisation of Fermentation Processes to Improve Productivity. *J. Process Control* **2021**, *108*, 148−156.

(16) Brunner, V.; Klöckner, L.; Kerpes, R.; Geier, D. U.; Becker, T. Online Sensor Validation in Sensor Networks for Bioprocess Monitoring Using Swarm Intelligence. *Anal. Bioanal. Chem.* **2020**, *412* (9), 2165−2175.

(17) Kourti, T.; Lee, J.; Macgregor, J. F. Experiences with Industrial Applications of Projection Methods for Multivariate Statistical Process Control. *Comput. Chem. Eng.* **1996**, *20* (SUPPL.1), 745−750.

(18) Krause, D.; Hussein, M. A.; Becker, T. Online Monitoring of Bioprocesses via Multivariate Sensor Prediction within Swarm Intelligence Decision Making. *Chemom. Intell. Lab. Syst.* **2015**, *145*, 48−59.

(19) Neogi, D.; Schlags, C. E. Multivariate Statistical Analysis of an Emulsion Batch Process. *Ind. Eng. Chem. Res.* **1998**, *37* (10), 3971−3979.

(20) Lu, N.; Gao, F.; Yang, Y.; Wang, F. PCA-Based Modeling and on-Line Monitoring Strategy for Uneven-Length Batch Processes. *Ind. Eng. Chem. Res.* **2004**, *43* (13), 3343−3352.

(21) Luo, L.; Bao, S.; Mao, J.; Tang, D. Phase Partition and Phase-Based Process Monitoring Methods for Multiphase Batch Processes with Uneven Durations. *Ind. Eng. Chem. Res.* **2016**, *55* (7), 2035−2048.

(22) Zhang, S.; Zhao, C.; Gao, F. Two-Directional Concurrent Strategy of Mode Identification and Sequential Phase Division for Multimode and Multiphase Batch Process Monitoring with Uneven Lengths. *Chem. Eng. Sci.* **2018**, *178*, 104−117.

(23) Guo, R.; Jin, Y. Phase Identification and Online Monitoring for the Uneven Batch Processes. *IEEE Access* **2019**, *7*, 81351−81363.

(24) Zhao, C. A Quality-Relevant Sequential Phase Partition Approach for Regression Modeling and Quality Prediction Analysis in Manufacturing Processes. *IEEE Trans. Autom. Sci. Eng.* **2014**, *11* (4), 983−991.

(25) Kassidas, A.; MacGregor, J. F.; Taylor, P. A. Synchronization of Batch Trajectories Using Dynamic Time Warping. *AIChE J.* **1998**, *44* (4), 864−875.

(26) Fransson, M.; Folestad, S. Real-Time Alignment of Batch Process Data Using COW for on-Line Process Monitoring. *Chemom. Intell. Lab. Syst.* **2006**, *84* (1−2), 56−61.

(27) Nielsen, N. P. V.; Carstensen, J. M.; Smedsgaard, J. Aligning of Single and Multiple Wavelength Chromatographic Profiles for Chemometric Data Analysis Using Correlation Optimised Warping. *J. Chromatogr. A* **1998**, *805* (1−2), 17−35.

(28) González-Martínez, J. M.; de Noord, O. E.; Ferrer, A. Multisynchro: A Novel Approach for Batch Synchronization in Scenarios of Multiple Asynchronisms. *J. Chemom.* **2014**, *28* (5), 462−475.

(29) Zhou, M.; Wong, M. H. Efficient Online Subsequence Searching in Data Streams under Dynamic Time Warping Distance. *Proc. - Int. Conf. Data Eng.* **2008**, *00*, 686−695.

(30) Lu, B.; Xu, S.; Stuber, J.; Edgar, T. F. Constrained Selective Dynamic Time Warping of Trajectories in Three Dimensional Batch Data. *Chemom. Intell. Lab. Syst.* **2016**, *159* (July), 138−150.

(31) Luo, L.; Bao, S.; Gao, Z. Quality Prediction Based on HOPLS-CP for Batch Processes. *Chemom. Intell. Lab. Syst.* **2015**, *143*, 28−39.

(32) Yu, H.; Augustijn, D.; Bro, R. Accelerating PARAFAC2 Algorithms for Non-Negative Complex Tensor Decomposition. *Chemom. Intell. Lab. Syst.* **2021**, *214* (April), 104312.

(33) Tian, K.; Wu, L.; Min, S.; Bro, R. Geometric Search: A New Approach for Fitting PARAFAC2Models on GC-MS Data. *Talanta* **2018**, *185* (March), 378−386.

(34) Amigo, J. M.; Skov, T.; Bro, R.; Coello, J.; Maspoch, S. Solving GC-MS Problems with PARAFAC2. *TrAC, Trends Anal. Chem.* **2008**, *27* (8), 714−725.

(35) Rendall, R.; Lu, B.; Castillo, I.; Chin, S. T.; Chiang, L. H.; Reis, M. S. A Unifying and Integrated Framework for Feature Oriented Analysis of Batch Processes. *Ind. Eng. Chem. Res.* **2017**, *56* (30), 8590−8605.

(36) He, Q. P.; Wang, J. Statistics Pattern Analysis: A New Process Monitoring Framework and Its Application to Semiconductor Batch Processes. *AIChE J.* **2011**, *57* (1), 107−121.

(37) Rato, T. J.; Blue, J.; Pinaton, J.; Reis, M. S. Translation-Invariant Multiscale Energy-Based PCA for Monitoring Batch Processes in

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525

Semiconductor Manufacturing. *IEEE Trans. Autom. Sci. Eng.* **2017**, *14* (2), 894−904.
(38) Birol, G.; Ündey, C.; Çinar, A. A Modular Simulation Package for Fed-Batch Fermentation: Penicillin Production. *Comput. Chem. Eng.* **2002**, *26* (11), 1553−1565.
(39) Pérez, N. F.; Ferré, J.; Boqué, R. Calculation of the Reliability of Classification in Discriminant Partial Least-Squares Binary Classification. *Chemom. Intell. Lab. Syst.* **2009**, *95* (2), 122−128.
(40) Gilbert, R. O. *Statistical Methods for Environment Pollution Monitoring*; 1st ed.; John Wiley and Sons Inc.: New York, 1987.
(41) The Mathworks. *MATLAB Version 9.8 (R2020a)*; Natick, Massachussetts, 2020. https://www.mathworks.com (accessed 2023−03−31).
(42) Gutmann, H. M. A Radial Basis Function Method for Global Optimization. *J. Global Optim.* **2001**, *19* (3), 201−227.
(43) Queipo, N. V.; Haftka, R. T.; Shyy, W.; Goel, T.; Vaidyanathan, R.; Kevin Tucker, P. Surrogate-Based Analysis and Optimization. *Prog. Aerosp. Sci.* **2005**, *41* (1), 1−28.
(44) Bhosekar, A.; Ierapetritou, M. Advances in Surrogate Based Modeling, Feasibility Analysis, and Optimization: A Review. *Comput. Chem. Eng.* **2018**, *108*, 250−267.
(45) Chen, G.; Zhang, K.; Xue, X.; Zhang, L.; Yao, C.; Wang, J.; Yao, J. A Radial Basis Function Surrogate Model Assisted Evolutionary Algorithm for High-Dimensional Expensive Optimization Problems. *Appl. Soft Comput.* **2022**, *116*, 108353.
(46) Bano, G.; Wang, Z.; Facco, P.; Bezzo, F.; Barolo, M.; Ierapetritou, M. A Novel and Systematic Approach to Identify the Design Space of Pharmaceutical Processes. *Comput. Chem. Eng.* **2018**, *115*, 309−322.
(47) Achterberg, T.; Koch, T.; Martin, A. Branching Rules Revisited. *Oper. Res. Lett.* **2005**, *33* (1), 42−54.
(48) Reis, M. S.; Rendall, R.; Rato, T. J.; Martins, C.; Delgado, P. Improving the Sensitivity of Statistical Process Monitoring of Manifolds Embedded in High-Dimensional Spaces: The Truncated-Q Statistic. *Chemom. Intell. Lab. Syst.* **2021**, *215* (March), 104369.
(49) Wan, J.; Marjanovic, O.; Lennox, B. Uneven Batch Data Alignment with Application to the Control of Batch End-Product Quality. *ISA Trans.* **2014**, *53* (2), 584−590.
(50) Sun, W.; Meng, Y.; Palazoglu, A.; Zhao, J.; Zhang, H.; Zhang, J. A Method for Multiphase Batch Process Monitoring Based on Auto Phase Identification. *J. Process Control* **2011**, *21* (4), 627−638.
(51) Vanlaer, J.; Van der Kerkhof, P.; Gins, G.; Van Impe, J. F. M. The Influence of Input and Output Measurement Noise on Batch-End Quality Prediction with Partial Least Squares. In *Advances in Data Mining. Applications and Theoretical Aspects: 12th Industrial Conference, ICDM 2012, Berlin, Germany, July 13−20, 2012. Proceedings 12*; Springer Berlin Heidelberg, 2012; pp 121−135.
(52) Biancolini, M. E. *Fast Radial Basis Functions for Engineering Applications*; Springer International Publishing: Heidelberg, 2017.
(53) Iske, A. Scattered Data Modelling Using Radial Basis Functions. In *Tutorials on Multiresolution in Geometric Modelling*; Springer-Verlag: Berlin, Heidelberg, 2002; pp 205−242. DOI: 10.1007/978-3-662-04388-2_9.
(54) Zhong, D. Y.; Wang, L. G.; Bi, L. Implicit Surface Reconstruction Based on Generalized Radial Basis Functions Interpolant with Distinct Constraints. *Appl. Math. Model.* **2019**, *71*, 408−420.
(55) She, C.; Wang, Z.; Sun, F.; Liu, P.; Zhang, L. Battery Aging Assessment for Real-World Electric Buses Based on Incremental Capacity Analysis and Radial Basis Function Neural Network. *IEEE Trans. Ind. Inf.* **2020**, *16* (5), 3345−3354.
(56) Xu, Y.; Zomer, S.; Brereton, R. G. Support Vector Machines: A Recent Method for Classification in Chemometrics. *Crit. Rev. Anal. Chem.* **2006**, *36* (3−4), 177−188.

https://doi.org/10.1021/acs.iecr.3c01897
Ind. Eng. Chem. Res. 2023, 62, 18511−18525