1. CONTEXT

Multivariate Stasticial Process Monitoring (MSPM) is a common data-driven method for monitoring complex processes where mathematical models may be difficult to develop.

Statistical process control methods such as PCA/PLS have been successfully applied to continuous processes in many cases. Batch processes offer a unique challenge due to factors such as (Undey & Cinar, 2002):

1. Transient behaviour
2. Highly non-linear and non-Guassian process behaviour
3. Batch-to-batch variations
4. Multiphase behaviour

The argument for batch processes is cost-effective manufacturing of low-volume, high-value products which would otherwise be uneconomical for continuous processes.

(1) Transient Behaviour

MSPC with PCA/PLS, as applied to continuous processes, consider the signals to be statistically stationary and not change with time – which is typically not satisfied by batch/semi-batch processes. These were extended to batch processes with the pioneering work of Multiway-PCA (and later MPLS) by McGregor (1995).

(2) Non-linear and Non-gaussian Process

Subsequently various improvements were proposed to account for various factors such as dynamic characteristics, nonlinearity, non-gaussian distributions etc. (DPCA, KPCA, HPCA, BDPCA etc.)

(4) Multiphase Behaviour

Most batch processes exhibit multiphase behaviour arising from multiple steps in a single processing unit, chemical or phenomenological actions (chemical reactions, microbial activity) which affect the underlying process correlations.

Monolithic models based on traditional MPCA/MPLS cannot effectively capture the multi-phase behaviour. Each phase usually has its own characteristic dynamics. Furthermore, not all variables were present over the course of the batch meaning models either had to be simplified or missing data had to be generated leading to reduced performance.

An improvement was made by using a multi-model approach, where a nonlinear process can be split into multiple phases conforming to a mostly-linear behaviour. Each phase would have a corresponding local model based on inputs available at that point in time.

1.1. Multi-Model Approach

The multi-model approach to batch process monitoring has various branches for consideration:

1.1.1. Phase Detection

The main challenge in monitoring multiphase processes is the number and manner in which it is divided.

- Early work relied on specific expert process knowledge to divide the process logically. This is not practical in many real-world scenarios.
- Later, changes in 'indicator variables' (e.g. conversion, colour) marked by Singular Points (e.g. discontinuities, inflection points, minima/maxima) were used to divide phases
- Modern approach essentially use a clustering exercise to divide the process into phases that are approximately linear (e.g. K-means, GMM). Improvements such as Finite-GMM, FJ-GMM, GMM-PSD, VB-GMM were able to automatically specify the number of clusters during the EM phase.

1.1.2. Phase Assigment

For a given sample, the corresponding local phase model needs to be invoked with an accurate state estimation approach. Inaccurate state estimation leads to incorrect quality predictions and/or false positives.

- With a GMM-based phase division approach, phase assignment can be done by calculating the posterior probability of a new data sample w.r.t. each phase. The phase with the highest probability is taken as the current state.

1.1.3. Unequal Batch Lengths

Unlike a continuous process, the duration of each batch is typically defined as having reached a quality-goal instead of a specific duration of time. In much of the literature batch lengths are assumed to be equal length which simplifies the modelling procedure.

In practise, the length of each batch is unequal due to:

- Batch-to-batch variations (e.g. initial conditions)
- Process conditions and disturbances
- Operator intervention

The effect of unequal batch lengths is that key landmarks in the process occur at different times during the batch (e.g. switching from batch->fed-batch mode). The Singular Points that define phases become unsynchronized leading to incorrect state estimation in the transitions between phases.

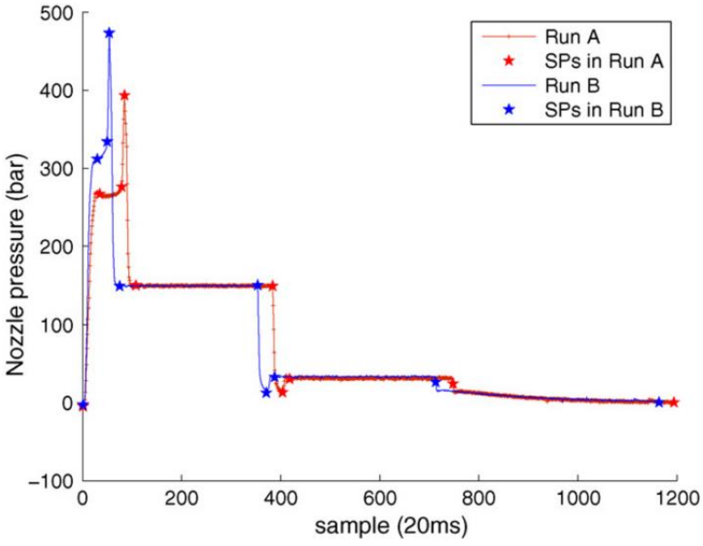


Fig. 2. Typical injection molding profiles showing location of Singular points

1.1.4. Transition Phases

The incorrect state estimation in the transition between phases will cause inaccurate quality predictions or false positives.

- Synchronizing of key events between different batches has been successfully implemented using dynamic time warping (DTW) (Doan & Srinivasan, 2008). This method will translate, compress and expand a pair of signals in such a way that the diffence between them is minimized.
    - Symmetric DTW will minimize the difference of two singals onto a new time domain
    - Assymetric DTW will map the test signal onto that of a reference signal
        - This reference signal is often referred to as the 'golden' reference trajectory
        - Can be generated by simulating process under ideal conditions
        - Alternatively, if this is unavailable reference trajectory can be derived from historical data
- In addition, a Baysian Model Average (BMA) approach has been shown effective This technique will use the posterior probability of adjacent models (or all global models) as a dynamic weighting of the final output. (Liu et al., 2018; Yu, Chen, & Rashid, 2013).

2. PROJECT GOAL

(D. Wang & Srinivasan, 2009) proposed a real-time product quality control strategy for batch processes. This strategy is achieved by periodically predicting the final product quality and adjusting process variables at pre-specified 'decision points'. This data-driven methodology employs multiple models (one for each decision point) to overcome multiphase behaviour and avoid missing data imputation.

The control action is triggered if the final product quality predicted at a decision point is beyond an acceptable range. At this point, the manipulated variables values are calculating by solving an optimal control problem similar to model predictive control (MPC).

Quality Prediction:

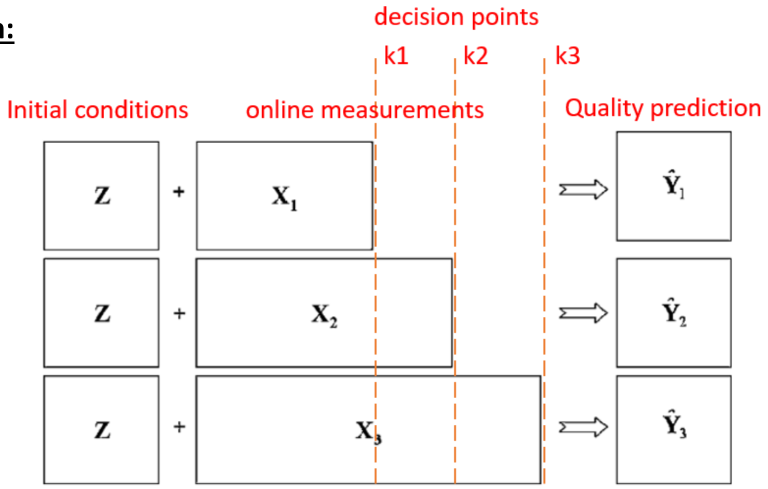


Fig. 5. Interpretation of proposed multi-model strategy in terms of multi-way multi-block PLS.

Shortcomings of this approach:

1. Decision points are specified based on expert process knowledge
2. Simulated batch data are all of equal length
3. Uncertainty during transition phases are not considered

Project Goal:

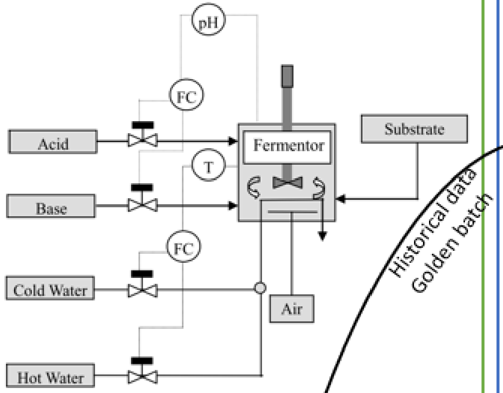
Improve the quality control framework by extending:

1. No process knowledge required – use data-driven phase detection based on GMMs to determine the process phases and use these as the corresponding decision points since they define changes in underlying process variable correlations.
2. Unequal batch lengths – simulated historical data will be of varying length with key landmarks occurring at different times. Batch trajectories will be synchronized with assymetric DTW based on a reference batch generated under optimal conditions.
3. Transition regions – implement a BMA strategy to dynamically weight the predictions of adjacent local models. This same strategy can be used to dynamically weight the calculated control action during the optimization step.

   Alternatively: define transition region intervals explicitly and build dedicated models for these points.

APPROACH

Simulation



Fed-batch Penicillin Process

- Multiphase behaviour
- Highly nonlinear
- Widely used

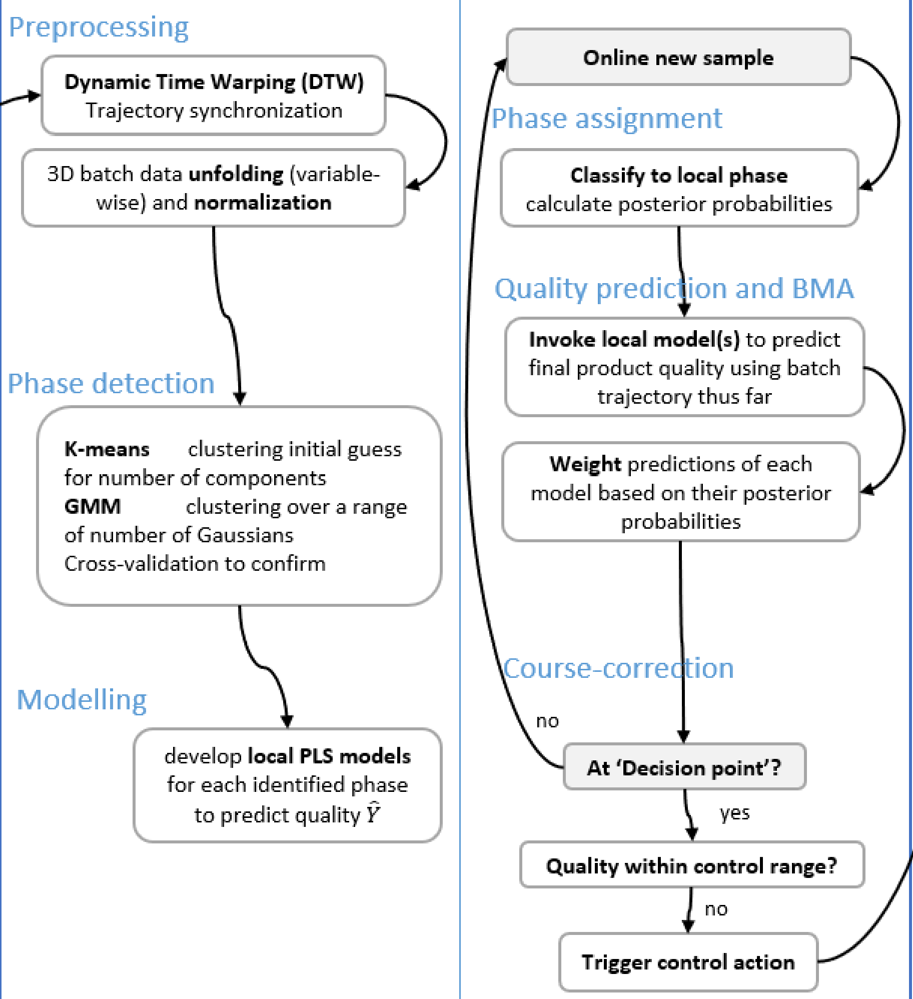
Data Generation

- Variations in initial conditions (within literature ranges)
- Noise on input and output signals

- Variations in key landmarks introduced:
    - Batch->fedbatch at S = 0.3 g/mol
    - Batch completion at P>= 1.4 g/mol

- 50 batches for historical data using nominal ranges
- 1 ‘golden’ reference batch using optimal ranges
- 3-5 fault batches for testing of control strategy

Modelling and Prediction



Preprocessing

Dynamic Time Warping (DTW) Trajectory synchronization

3D batch data **unfolding** (variable-wise) and **normalization**

Phase detection

K-means clustering initial guess for number of components
GMM clustering over a range of number of Gaussians
Cross-validation to confirm

Modelling

develop **local PLS models** for each identified phase to predict quality $\hat{Y}$

Online new sample

Phase assignment

Classify to local phase calculate posterior probabilities

Quality prediction and BMA

Invoke local model(s) to predict final product quality using batch trajectory thus far

Weight predictions of each model based on their posterior probabilities

Course-correction

no

At ‘Decision point’?

yes

Quality within control range?

no

Trigger control action

Control Action

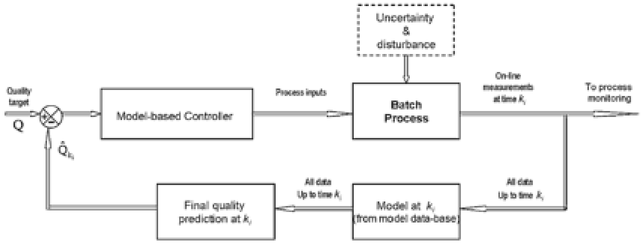

Uncertainty
&
disturbance
Quality target
Q
Model-based Controller
Process inputs
Batch Process
On-line measurements at time k_i
To process monitoring
$\hat{Q}_{k_i}$
Final quality prediction at k_i
All data Up to time k_i
Model at k_i (from model data-base)
All data Up to time k_i

Fig. 2. Proposed real-time quality control framework.

calculate the new adjustment ΔX^n by minimizing an objective function as a linear quadratic Gaussian regulator (LQG)

- Doing global weighting of control calculations is not a good solution
- Instead use adjacent local models and weight two control action calculations by their posterior probabilities

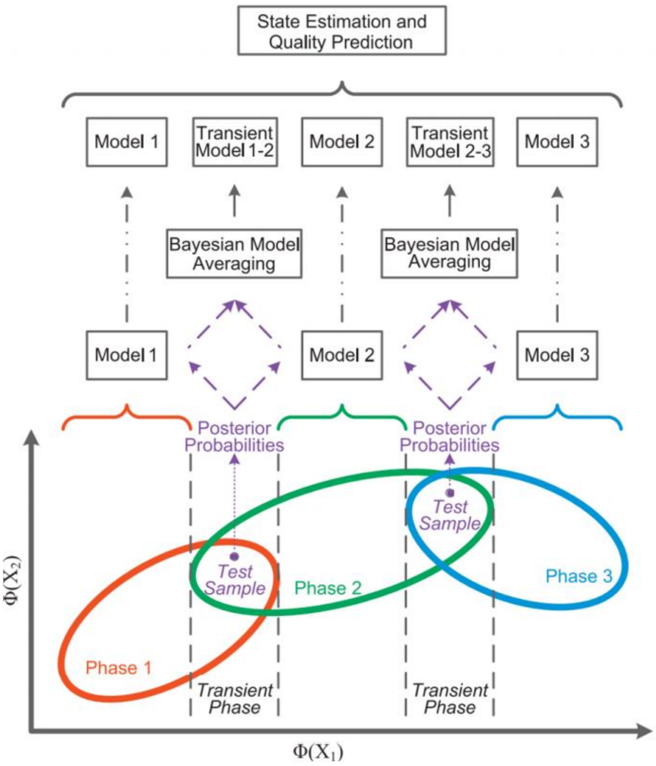


Figure 1: Bayesian model averaging strategy in detail

3. FAULT BATCHES

‘Faulty’ batches will be used to evaluate the quality control action. In (D. Wang & Srinivasan, 2009), 3-scenarios are used to evaluate the control framework performance:

1. Scenario A – raw material impurity
   a. A disturbance is introduced in the feed concentration of reactant A (0.90 -> 0.88M) which would otherwise result in an off-spec final product quality
   b. **Result:** at decision point 1 and 2, the temperature setpoint of the reactor is adjusted which recovers the batch
2. Scenario B – human error in specifying temperature setpoint
   a. An operator error is introduced at the start of the batch with Tsp set at 320K instead of 330K which would otherwise result in an off-spec final product quality
   b. **Result:** at decision point 1 and 2 the flow-rate setpoints are adjusted which recovers the batch
3. Scenario C – process disturbance during operation

a. Flow rate of reactant B is increased to 2.3 (from 1.8) at t=20hr, representing a fault in the flowmeter that would result in an off-spec final product quality.
b. **Result:** at decision point 1, temperature is decreased. At decision point 2, temperature is slightly increased which recovers the batch.

Alternative 'faults' introduced from other literature sources:

1. 2% ramp increase in substrate feed rate (Chen et al., 2010)
2. 15% decrease in agitator power, maintained throughout (Chen et al., 2010)
3. Sudden drop in %O_2 representing batch contamination **(unrecoverable)** (Largoni et al., 2015)
4. ramp change in agitator power (Jiang & Yan, 2019)

4. BRIEF LITERATURE REVIEWS

(Jiang & Yan, 2019) – *"Multimode Process Monitoring Using Variational Bayesian Inference and Canonical Correlation Analysis"*

- As applied to batch-fed penicillin fermentation process
- Used the VB-GMM algorithm to automatically perform phase division from historical data without specifying the number of operation modes…7 phases identified
- Establishes local CCA models (PCA alternative) for monitoring purposes
- During monitoring, the current operating phase is identified by largest posterior probability
- Local CCA model is used to determine the fault residuals, a BIP monitoring index is incorporated
- 100 batches of fixed 400hr length are generated, sampling time of 1hr
- 3 faulty batches were generated
  - 1. Step change introduced in agitator power from t=150hr -> 300hr
  - 2. Ramp change introduced in agitator power from t=150hr -> 300hr
  - 3. Ramp change introduced to aeration rate from t=150hr -> 300hr
- The proposed solution showed effective fault monitoring with the advantage of automatically determining the number of operation modes (which makes it more practical). CCA performance was shown to be superior to PCA for this implementation and incorporation of BIP monitoring index allows for monitoring in a probabilistic manner instead of deterministic.

(Liu et al., 2018) – *“Gaussian Process Regresion and Bayesian Inference Based Operating Performance Assessment for Multiphase Batch Processes”*

- As applied to batch-fed penicillin fermentation process
- Used the GMM-PSD algorithm to automatically perform phase division from historical data without specifying the number of operation modes…5 phases identified
- Established local Gaussian Process Regression (GPR) models for each local phase
- From the results of offline phase division, the boundary intervals of adjacent phases overlap which is defined as ‘fuzzy intervals’. The cut-off is defined using a threshold value for the posterior probability
- During monitoring, the current operating phase is determined from the intervals as determined offline
- If the sample falls within a fuzzy interval, both adjacent local GPR models are invoked and the prediction is weighted by the posterior probability in each phase as calculated (Bayesian Model Averaging)
- An economic index is defined that defines whether a batch is current ‘optimal’ or ‘sub-optimal’. This is determined by the ratio of the current quality prediction to the min/max as determined from historical data. Values >0.5 are defined as ‘optimal’.
- Author further performs ‘non-optimal cause identification’
    - If performance is non-optimal, the possible cause is determined calculating variable contribution values
    - The calculated values are adjusted again into ratios of min/max as per historical data
- 50 training batches generated lengths of 390hr to 420hr, sampling time 1hr
- Sub-optimal/fault batch produced by using a culture volume of 150L – larger volumes experience evaporative loss since temperature is fixed which results in penicillin decline
- Prediction quality was shown to perform well, but from what I see the initial ramp up is not captured
- Sub-optimal batch was identified and variable contribution indicated Culture Volume as the likely cause

(Yu, Chen, Mori, et al., 2013) – "*Multi-kernel Gaussian Process Regression and Bayesian Model Averaging Based Nonlinear State Estimation and Quality Prediction of Multiphase batch Processes*"

Similar paper (Yu, Chen, & Rashid, 2013).

- As applied to poylmerization process
- Offline phase division achieved using a kernel mixture model, individual phases are expressed by a local kernel density function (Gaussian density functions are used)
    - Bayesian inference is used to classify measurements into phases based on maximum posterior probability
    - If it satisfies a threshold (0.05) it is taken to be a certain phase, not between-phase transition
- Local GPR models are developed for each identified phase, a kernel function is incorporated to improve performance for nonlinear process
- During monitoring, current operating phase is determined by posterior probability
- If it falls within a transition region, a BMA is used where the prediction of adjacent phases are weighted dynamically based on the posterior probability
- 20 training batches, all fixed length of 3hr, sampling period 1 min
- The approach showed effective quality prediction better or comparable to SVR

(Doan & Srinivasan, 2008) – "*Online monitoring of multi-phase batch processes using phase-based multivariate statistical process control*"

- As applied to batch-fed penicillin fermentation process
- Offline phase division achieved using Singular points in key variables along
    - 5 phases were detected
    - Golden batch was selected as a batch of 400hr length
- Run-to-run variations among different instances of a phase are synchronized using DTW
- Local dynamic-PCA models are developed for monitoring purposes (as an extension of PCA which incorporates a time-lag that is used to capture serial correlation in the process)
- During monitoring, for a given new sample, the sample is checked if it's a Singular point
    - If it is, a phase change is flagged and the corresponding MSPC model is retrieved
    - If it is not, the monitoring statistics T^2 is calculated for key variables and compared with a threshold from historical data to announce a fault if exceeded.
    - A time-lagged data matrix is constructed and a further T2 is calculated for comparison.
- 14 batches of data were generated with random initial condition and set points
- Batch lengths of 380-420hr, sampling time 0.5hr

- Batch lengths were cut-off to the longest common length since DPCA required equal length
- 4 fault batches were generated for testing:
  - 15% step increase in substrate feed rate from t=150hr to end
  - 15% step decrease in substrate feed flow rate t=160hr to end
  - 15% step decrease in agitation power from t=20hr to 40hr
- The Singular Point phase decomposition did not show great performance. A false positive was flagged in a test batch at t=88hr until t=93hr
  - This is a result of variations in initial conditions affecting the location of SPs

(Yu, 2012b) – *"online quality prediction of nonlinear and non-Gaussian chemical processes with shifting dynamics using finite mixture model based Guassian process regression approach"*

This paper takes a continuous process and uses a technique similar to GMM-based phase identified in batch processes to identify different operating **modes** of the continuous process.

- As applied to Tennessee Eastman chemical process
- Offline mode division using finite mixture model (FMM) where number of components is specified manually…6 were identified
- Nonlinear lernel function is selected for Gaussian process regression models (Gaussian kernel function is used).
- During monitoring, for any new input measurement, the data is normalized using the mean and std deviation from the training set
- The posterior probability is calculated with all identified operating modes
- The quality prediction of the test sample within each operating mode is estimated from the local Guassian process regression model
- The overall quality variable predicted is computed by incorporating all localized estimations within different operating modes – weighted by their posterior probabilities
- 2 test cases used:
  - Case 1 – data could belong to any of 3 operating modes
  - Case 2- operated under all 6 modes with random switching between
- Case 1 was identified and quality predicted accurately
- Case 2 showed improved performance compared to LSSVM approach with comparable accuracy as in Case 1

(D. Wang & Srinivasan, 2009) – *"Multi-model based real-time final product quality control strategy for batch processes"*

As described in this document.

(David Wang, 2011) – "*robust Data-driven Modeling Approach for Real-Time final product quality prediction in batch process operation"*

An extension of the above strategy that incorporates a noise reduction technique.

References:

Chen, X., Gao, X., Zhang, Y., & Qi, Y. (2010). Enhanced batch process monitoring and quality prediction based on multi-phase multi-way partial least squares. *Proceedings - 2010 IEEE International Conference on Intelligent Computing and Intelligent Systems, ICIS 2010*, *2*(3), 32–36. https://doi.org/10.1109/ICICISYS.2010.5658834

Doan, X. T., & Srinivasan, R. (2008). Online monitoring of multi-phase batch processes using phase-based multivariate statistical process control. *Computers and Chemical Engineering*, *32*(1–2), 230–243. https://doi.org/10.1016/j.compchemeng.2007.05.010

Jiang, Q., & Yan, X. (2019). Multimode Process Monitoring Using Variational Bayesian Inference and Canonical Correlation Analysis. *IEEE Transactions on Automation Science and Engineering*, *16*(4), 1814–1824. https://doi.org/10.1109/TASE.2019.2897477

Largoni, M., Facco, P., Bernini, D., Bezzo, F., & Barolo, M. (2015). Quality-by-Design approach to monitor the operation of a batch bioreactor in an industrial avian vaccine manufacturing process. *Journal of Biotechnology*, *211*, 87–96. https://doi.org/10.1016/j.jbiotec.2015.07.001

Liu, Y., Wang, X., Wang, F., & Gao, F. (2018). Gaussian Process Regression and Bayesian Inference Based Operating Performance Assessment for Multiphase Batch Processes. *Industrial and Engineering Chemistry Research*, *57*(21), 7232–7244. https://doi.org/10.1021/acs.iecr.8b00234

Undey, C., & Cinar, A. (2002). Statistical Monitoring of Multistage, Multiphase Batch Processes. *IEEE Control Systems*, *22*(5), 40–52. https://doi.org/10.1109/MCS.2002.1035216

Wang, D., & Srinivasan, R. (2009). Multi-model based real-time final product quality control strategy for batch processes. *Computers and Chemical Engineering*, *33*(5), 992–1003. https://doi.org/10.1016/j.compchemeng.2008.10.022

Wang, David. (2011). Robust data-driven modeling approach for real-time final product quality prediction in batch process operation. *IEEE Transactions on Industrial Informatics*, *7*(2), 371–377. https://doi.org/10.1109/TII.2010.2103401

Yu, J. (2012a). A Bayesian inference based two-stage support vector regression framework for soft sensor development in batch bioprocesses. *Computers and Chemical Engineering*, *41*, 134–144.

https://doi.org/10.1016/j.compchemeng.2012.03.004

Yu, J. (2012b). Online quality prediction of nonlinear and non-Gaussian chemical processes with shifting dynamics using finite mixture model based Gaussian process regression approach. *Chemical Engineering Science*, *82*, 22–30. https://doi.org/10.1016/j.ces.2012.07.018

Yu, J., Chen, K., Mori, J., & Rashid, M. M. (2013). Multi-kernel Gaussian process regression and Bayesian model averaging based nonlinear state estimation and quality prediction of multiphase batch processes. *Proceedings of the American Control Conference*, 5451–5456. https://doi.org/10.1109/acc.2013.6580690

Yu, J., Chen, K., & Rashid, M. M. (2013). A Bayesian model averaging based multi-kernel Gaussian process regression framework for nonlinear state estimation and quality prediction of multiphase batch processes with transient dynamics and uncertainty. *Chemical Engineering Science*, *93*, 96–109. https://doi.org/10.1016/j.ces.2013.01.058