Designation

Document title

**Literature**

Latest Issue date

**24 July 2020**

Synopses/Key words

Distribution

Jedd Orsmond

# PCA & PLS Based On-line Monitoring Techniques

[14] D. Wang and R. Srinivasan, **“Multi-model based real-time final product quality control strategy for batch processes,”** *Comput. Chem. Eng.*, vol. 33, no. 5, pp. 992–1003, 2009, doi: 10.1016/j.compchemeng.2008.10.022.

### Applicability to NC plant

For NC plant: easily adjusts to differences in time of batches (which in my case t most likely would have ended up a significant MV)

Also accounts for the raw material feed impurity which is a significant batch-to-batch variation, not overshadowed by process conditions which is the assumption of most MCC techniques.

Reference trajectory method wont work since we already know that the deviation in raw material purity is a major contributor to final quality.

Also obviates imputation data that is not available, multiple models better/easier for the overall process.

### Technique Overview

Quality control is achieved using a data-driven strategy, periodically predicting the final product quality and adjusting process variables at pre-specified decision points. Instead of at every sample point. Real-time control is obtained from online control optimization similar to MPC, which ensures that the calculated changes in process variable trajectories are feasible. Employs multiple models, one for each decision point, to capture time-varying relationships. The models combine real-time batch information from process variables and initial conditions with information from prior batches.

Control action is taken at pre-specified decision points where MVs are calculated by solving an opitimal control problem similar to MPC. A key benefit of this strategy is that missing data imputation is obviated.

The use of PCA and PLS for batch process monitoring has been investigated extensively. In these approaches the behaviour of the process is characterized using a statistical model derived through multi-way analysis of online measurements obtained when the process is in a state of statistical control.

Offline QC strategy – identify the cause of a failed batch and adjust the conditions for subsequent batches.

Reference Trajectory strategy- control objective is to keep the operating conditions as close to a reference trajectory (process model determined a priori using 1st principles or historical data). This is effective in many cases but can produce off-spec product even with perfect tracking of the reference trajectory as it does not consider batch-to-batch variations (raw material impurities, heat transfer, kinetic effects).

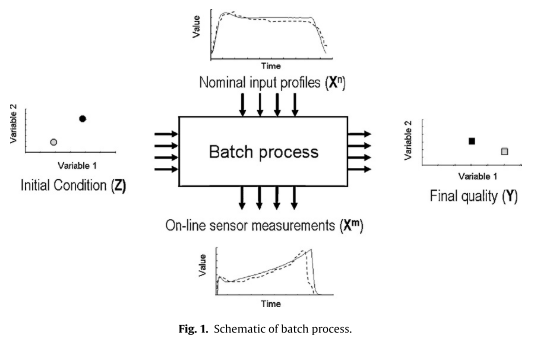
MCC recognize that process conditions during batch evolution will tend to dominate systematic batch-to-batch variations. At some mid-point points, the quality is predicted (inferential model using historical data) if the prediction falls outside the “in control” zone, a model is used to calculate a control move to bring back into statistical control. Training data needs to contain sufficient input variability and disturbance information to allow proper model identification.

The problem with MCC is that the inferential model for quality prediction is identified using historical training data, which are from previously completed successful/unsuccessful batches. T for QC purposes, the training set needs to contain sufficient input variation and disturbance information to allow proper model identification. I.e. the model identification requires persistence of excitation. Often this is not provided sufficiently.

### Proposed approach

In this article, data-driven real-time batch quality control strategy similar to MPC is developed. Process inputs are manipulated online to steer the process so that final quality is on spec. The main difference to other approaches is that control actions are taken at discrete, pre-specified time- points instead of at every sample. At each time point , a data-based model is used for quality prediction and control calculations. Different models are used at different points, each model exploits all online data available up to that point. This mult-model strategy eliminates the need for data imputation and the consequent uncertainties.

The approach deals with 3 data sources: 1) initial conditions (Z), trajectories of online measurmeents (X – partitioned into MVs Xn and measured variables Xm sampled at K sampling instants) 3) end point product quality (Y)



Z = IxL matrix of initial conditions (L=different quantities measured, I different batches)

X= IxJxK recording of trajectories (J process variables, K sampling times) -> partitioned into Xn (MV inputs) and Xm (other measurements).

Y = IxM end point product quality

The desire is to understand how Y is influenced by combination of Z (initial conditions) and batch evolution (X) and then to investigate how to adjust the process conditions in real-time such that Y is within spec.

To do this, 2 related problems must be solved: 1) estimate the final product quality at the end of the current batch 2) design suitable corrective actions as necessary

A finite set of models, each applicable to a specific point during the batch (decision points) are used instead of a complex monnolithic model.

The number and location of decision points is an open question (specified by process knowledge, using indiciator variable)

Any change in Xn will affect the remainder of the batch, cumulative effect described as:

Y=f(Z, Xki)

Where Xki represents all process measurements collected from the start to deciion point ki. A model is established for each decision point as well as for process control.

### Process modelling

Based on PLS

Most variables can be correlated to one another. The first task is to build an emprical model based on the representatives of these three data sets. MLR is used to relate the predictor variable (Z and X) with the response variable. Y = pred\*theta +F

Where theta is the regression coefficient and F is the residual. PLS is used to determine theta isntead of OLS (high degree of correlation or noise in predictor varaibles causes it to fail). Sever ways exist for calculating PLS model parameters, the one uses is NIPALS (non-iterative partial least squares apprach). NIPALS calculates LVs (T, P) and additional set of vectors known as weight (W) such that:

Theta = W(PTW)^-1(T^TT)%-1 T^TY

To build a model the three way data Xki is first unfolded. Batch-wise unfolding forms a two way matrix Xi (IxJki with mode J entities neseted within mode ki entitiites). With the information available at decision point ki, the PLS model is developed on the auto-scaled data sets.the multi-block and multi-way PLS (MBMWPLS) regresssion technique is implemented to model the relationship bewtween predictor and Y. At each decision point ki, a PLS model is built by relating the data avaialbe from the start of the batch up to ki, to the final product quality.

An important feature of this approach is there is no need to estimate the missing data, as is the case with single model for a whole batch.

### Control strategy

During the batch evolution, process measurements are collected up to the specified decision point, at which a decision is made on if or not a control action is expected. The control action includes the adjustment of the setpoint trajectories of CVs. By adjusting these trajectories the remainder of the batch will progress ot achieve the desired quality at the end.If a faulty condition is deterected during process monitoring using measures such as T^2 or SPE, the diagnosis shows that recovery is needed, then control action is triggered.

Once control action is triggered, the model corresponding to this decision point is employed to predict the final product quality using all previous information up to that ppoint. A set of MVs are then calculated in order ot minimize the difference between the prediction from the model and the target, subject to the constraints on the control or other process variables (if any).

At ki, the Y^hat,i is calculated. I corresponds to the ki decision point, theta\_i corresponds to the parameters of the model at decision point ki and

With the quality predicted , the values of control variables can be manipulated in order to achieve quality control. This is achieved by solving an optimization problem where the new adjustment , is obtained by minimizing an objective function as a linear quadratic Gaussian regulatior (LQG).

### Data input/manipulations required / Results

The technique is tested on a simulated batch process (consecutive homongenous liquid phase reaction of A and B forming desirable product C that reacts further to form waste product D). The objective is produce a specified number of moles of product C and keep product D at a specified range for each batch. Measurments of C and D are only available at the end of the batch, other online measurements are available. These include (reactor temp, reactank feeding tank temps, flow rate of B, coolant flow rate [6]. The MVs are selected as the trajectory of treactor temperature setpoint (Tsp) and the B feed flow rate (Fb)..

Scenario 1: raw material impurity (of specific interest)

A disturbance is introduced in the concentration of reactant A (0.9 -> 0.88) compared to the nominal case. The small variation exerts a large influence on the reaction kinetics consequently the final product quality. With no control, spec of C,D = 1238/558. Two sampling points are taken (30 and 5)) with C\_30/D\_30 = 1236, 546 || C\_50/D\_50=1235,548. Corrective action at the two decision points recovered the batch to C/D = 1235, 554) which is in the acceptable range.

### Advantages/Disadvantages

Most online batch supervision approaches require imputation of values of future measurmeents that will affect the final product quality. They also resort to an explicit method for synchrozing batches of different lengths.

The proposed method does not need all variables to be available at all decision points since different models are used at each point. Only data that would be available at that juncture has to be incorporated thus obviating the missing data imputation step essential when using only a single model.

The approach is model-agnostic, although a linear PLS model has been used here the proposed multi-model strategy is not limited and other structures such as nonlinear PLS or ANNs can be used. Can also use Dynamic models that account for time lag in predictor if they offer a better representation of the underlying dynamics furing a phase.

### Extension

[5] D. Wang, **“Robust data-driven modeling approach for real-time final product quality prediction in batch process operation,”** *IEEE Trans. Ind. Informatics*, vol. 7, no. 2, pp. 371–377, 2011, doi: 10.1109/TII.2010.2103401.

The data-driven, multimodel approach is extended with focus on making the models robust in order to eliminate the effect of noise, especially , outliers in the data. This is achieved with a model-based outlier detection method.

“the focus here is on how to improve the modeling accuracy for prediction in case of “noisy” training data which was not discussed in the previous work”

Improtant modelling isusue is to extract appropriate information from “noisy” data. Which prevents application on real industrial data. Most assume that data is normally distrubted, which is seldom the case. Some robust PLS approaches consider outliers (e.g. NIPALS).

Detecting outliers is done using PCA approach based on the estimated scores (T), loadings (P) and residuals (E). Leverage statistics (which are derived from the scores and loadings) are used to reflect the extremeness of samples in the PCA model.

In contrast to the detection in T^2 space, the statistics based on the model residual can be used to detect samples that havea high amount of variance outside the PCA model.

A method using PLS is also proposed. It is necessary since the outliers in response data for Y cannot be detected using PCA as the information in Y idata is not being used in PCA model.

The same semi-batch process simulation is used for evaluation.

### Additional Literature

#list of other papers, 1-2 sentences describing each one.

[11] J. Chen and K. C. Liu, **“On-line batch process monitoring using dynamic PCA and dynamic PLS models,”** *Chem. Eng. Sci.*, vol. 57, no. 1, pp. 63–75, 2002, doi: 10.1016/S0009-2509(01)00366-9.

Development of batch dynamic PCA (BDPCA) and batch dynamic PLS (BDPLS).

[13] M. Jia, F. Chu, F. Wang, and W. Wang, **“On-line batch process monitoring using batch dynamic kernel principal component analysis,”** *Chemom. Intell. Lab. Syst.*, vol. 101, no. 2, pp. 110–122, 2010, doi: 10.1016/j.chemolab.2010.02.004.

MPCA and MPLS is a simple and powerful method for SPC applies to batch processes and ghas been widely applied to industrial batch monitoring. This paper proposes a new dynamic and nonlinear batch processing monitoring bethod (BDKPCA) which integrates kernal PCA and ARMAX time series model by estimateingthe Average Kernal Matrix (AKM) of all batch runs. This contains the information of the stochastic variations and deviations among batches. BDKPCA calculates the Hotellings T^2 statistic and Q statistic for every time point which enhances sensitivity to fhe faults. Interestingly, the kernal function selection is the most important aspect in KPCA since the degree to which non-linearities are captured is determined by it. Three representative kernal functions (Polynomial, Sigmoid and Radial) are given but selection is an open problem. Radial is chosen in this paper.

A batch dynamic PCA is developed for the process variables and batch dytnamic PLS for the process variables and quality variables. These two methods incorporate both static and dynamic process characteristics to extend the static MPCA and MPLS for dynamic multivariate batch processes.

[1] R. Jia, Z. Mao, F. Wang, and D. He, **“Self-tuning final product quality control of batch processes using kernel latent variable model,”** *Chem. Eng. Res. Des.*, vol. 94, pp. 119–130, 2015, doi: 10.1016/j.cherd.2014.12.013.

MCC strategy emplying K-PLS models developed using batch-wise unfolding data set to capture the relationship between process variables and final quality. The estimators to predict future trajectories are based on Latent variables with missing data imputation method based on muilti-PCA models. The control strategy is constrained to lie in the Kernal latent variable space defined by historical batch data set.

[8] M. Largoni, P. Facco, D. Bernini, F. Bezzo, and M. Barolo, **“Quality-by-Design approach to monitor the operation of a batch bioreactor in an industrial avian vaccine manufacturing process,”** *J. Biotechnol.*, vol. 211, pp. 87–96, 2015, doi: 10.1016/j.jbiotec.2015.07.001.

Multivariate statistical models are built from historical databases of batches already completed. A sample of 38 batches (33 normal [23 in spec, 10 offspec], 5 faulty). Multivarate latent variables methods are used to produce a smaller set of artificial variables to be used in multiway versions of PCA and PLS. Abnormality detection using a multiway PCA model is proposed. Batch quality classification using multiway PLS model inversion is proposed that can also provide an indication on how the batch is evolving in terms of endpoint product quality.

# Support Vector On-line Monitoring Techniques

[12] Z. Ge, F. Gao, and Z. Song, **“Batch process monitoring based on support vector data description method,”** *J. Process Control*, vol. 21, no. 6, pp. 949–959, 2011, doi: 10.1016/j.jprocont.2011.02.004.

### Applicability to NC process

Although the difference is sometimes subtle, the NC plant does run in effectively different modes for different grades of NC. If a single dataset can be used to develop for all grades that would be awesome.

The NC process is very likely to show non-linear behaviour (although I have not confirmed this) and having inherent capability to overcome this is better than methods to deal with it.

### Process Overview

An efficient one-class classification method for batch processing monitoring called support vector data description (SVDD) is proposed. Different from traditional methods such as PCA and PLS, SVDD has no Gaussian assumption of the process data and is effective for nonlinear process modelling. Furthermore, SVDD only incorporates a quadration optimization step which makes it easy for practical implementation, it is also extended to multiphase batch process and multimode.

The conventional multivariate SPC methods widely used for continuous process monitoring have been extended to batch processes. MPCA, MPLS have been developed and a whole host of variations on these. Most of the methods assume data is gassuian distribution and linear correlation between diffrerent process variables. For nonlinear batch process monitoring, traditional MKPCA and MKPLS introduce the non-linear transformation function to incorporate into the monitoring methods. However, the gausian assumption is typically still required. When this is not the case, monitoring based on PCA/PLS may not function well.

By representing all normal process data samples as one class, an SVDD model can be constructed to differentiate the abnormal data samples from the normal. It was used for monitoring of continuous processes but this paper extends it to batch processes.

In order to model the nonlinear process, SVDD first maps the data from the original space to the feature space by nonlinear transformation function. Then a hypersphere with the minimum volume can be found in the feature space.

The dataset of the batch process is collected in a three way manner X( IxJxK…I = batch number, J = variable number, K = total number of data sampels). In MPCA or MPLS, this dat is first unfolded into 2D datasets before PCa or PLS is performed. In this case, the dataset is first unfolded through batch direction for data auto-scaling and then it is re-unfolded through the variable direction. This also avoids the future value estimation step of traditional MPCA.

The technique is extended to multiphase batch process modelling using sub-SVDD models. The batch process is separated into several phases (same open question problem) – 4 methods have been proposed in literature. The dataset X is also divided into different sub-datasets corresponding tho the phases. The sub-SVDD model is built for each phase. The center and radius of each hypersphere can be calculated in the same procedure.

Multiple operation modes – when a batch process runs under multiple operation modes different sub-models should be developed. However, the SVDD technique different types of normal datasets can be labelled as one class and compressed into a single hypersphere in the feature space and a single SVDD model is sufficient to model the multimode batch process which greatly simplifies the modeling and monitoring procedures.

### Advantages/Disadvtanges

The main advtanges of the SVDD based method

1. No gauss assumption
2. Nonlinear relationships
3. Tigh boundary of dat distribution
4. Does not require some monitoring statistics such as T^2 and SPE.

Disadvantages:

Computation complexity increased for high dimensional process variable sets

Cannot inherently take into account the error contained in the variables

A tighter control limit may lead to false alarms rates

Applied to semiconductor process. 17 monitoring variables

### Additional Literature

[3] S. Zhang, F. Wang, D. He, and R. Jia, **“Online quality prediction for cobalt oxalate synthesis process using least squares support vector regression approach with dual updating,”** *Control Eng. Pract.*, vol. 21, no. 10, pp. 1267–1276, 2013, doi: 10.1016/j.conengprac.2013.06.002.

* An effective soft sensor based on least squares support vector regression (LSSVR) with dual updating is developed to predict the average particle size in an industrial cobal oxalte synthesis process. LSSVR requires only a small number of training samples and has been widely used for nonlinear function regression and system indetification problems.

[9] J. Wang, W. Liu, K. Qiu, T. Yu, and L. Zhao, **“Dynamic hypersphere based support vector data description for batch process monitoring,”** *Chemom. Intell. Lab. Syst.*, vol. 172, no. October 2017, pp. 17–32, 2018, doi: 10.1016/j.chemolab.2017.11.002.

* Using nonlinear transformation functions, SVDD constructs an irregular hypersphere in high dimensional space. This paper proposes a dynamic hypersphere based SVDD method. The dynamic hypersphere is built by the important SVs of combined dataset (current test sample and training dataset) compared to static hypersphere of just the training dataset.
* Also mentions other methods such as k-chart SVDD, max limit SVDD and validation limit SVDDD.

# Gaussian Mixture Model Techniques

[6] T. Chen and J. Zhang, **“On-line multivariate statistical monitoring of batch processes using Gaussian mixture model,”** *Comput. Chem. Eng.*, vol. 34, no. 4, pp. 500–507, 2010, doi: 10.1016/j.compchemeng.2009.08.007.

Calculation of the confidence bounds for T2 and SPE in conventional monitoring approaches (e.g. PCA) are done so under the assumption that the PCA/PLS scores and precision errors are Gaussian distributed – this may be invalid for complex processes or when non-linear projection techniques are used.

GMM has been proposed to address this issue to estimate the probability density function (pdf) with improved results reported for continuous processes and *batch-wise* monitoring. This paper extends GMM to *on-line* monitoring of batch processes.

MPCA is first used to extract low dimensional representation of the process. The nominal batches are passed through the monitoring procedure to collect the predicted scores and SPE at each time step, then GMM is employed to estimate the joint pdf of these predicted scores and SPE from MPCA at each time step.