Chemometrics and Intelligent Laboratory Systems 95 (2009) 107–121

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

# Improved calibration investigation using phase-wise local and cumulative quality interpretation and prediction

Chunhui Zhao [a], Fuli Wang [a,*], Furong Gao [b]

[a] College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, PR China
[b] Dept. of Chemical Engineering, Hong Kong University of Science and Technology, Clear water bay, Kowloon, Hong Kong SAR, PR China

ARTICLE INFO

*Article history:*
Received 16 June 2008
Received in revised form 25 September 2008
Accepted 25 September 2008
Available online 7 October 2008

*Keywords:*
Bootstrap
Variable selection
O2-PLS
Phase-wise local and cumulative effects
Robust calibration modeling
Multiphase batch processes

ABSTRACT

In the present work, a multiphase calibration modeling and statistical analysis strategy is developed for the improvement of process understanding and quality prediction. Having realized the phase-wise local and cumulative effects on quality interpretation and prediction, the major task lies in how to qualify and quantify them among multiple phases. The proposed scheme is presented on two different levels: On the first level, phase-specific variable selection and O2-PLS are designed focusing on revealing the local effects of individual phases on quality variations, e.g. within the current phase, which part of process variations are responsible for quality variations and which quality variations are dominated. Moreover, bootstrapping technique is employed during the procedure of variable selection and O2-PLS, which can enhance the reliability and robustness of calibration analysis. On the second level, conventional PLS is used to model the quantitative relationship between multiple phases and the final qualities so that the cumulative effects on quality variations are apprehended by additively stacking the local contributions of various phases. The proposed strategy highlights such an idea that in real multiphase processes, each phase may only explain one part of quality variations and the final qualities can only be additively and jointly defined by multiple phases. A benchmark simulation of fed-batch penicillin fermentation production is considered and put into illustration, which demonstrates the efficiency of the proposed algorithm for better process understanding and quality interpretation in multiphase processes.



## 1. Introduction

Over the last few years, multivariate calibration methods have been used to well-establish a quantitative relationship between process measurement (**X**) and quality property (**Y**). For online applications, multivariate calibration can be accomplished with familiar chemometric techniques [1–5], such as multiple linear regression (MLR), principal component regression (PCR), partial least squares (PLS), and so on. Accurate qualitative and quantitative quality interpretation and prediction may help avoiding heavy and costly chemical measurements, which, however, are not satisfactory enough in many practical applications. Because process observations often contain major sources of variations that are of little or no predictive feature, it is often observed that the extracted first several calibration latent components are capturing most of process variations but only explaining very little of qualities. It is known that increasing descriptor information in a regression model will improve the fit to the training reference data, which, however, will often cause a substantial reduction in the predictive ability of the model concerning the test dataset. Calibration regressions in these cases are undesirable since acceptable interpretability and robustness cannot be expected without advance removal of nuisance variables or factors. Therefore, it is a common practice to apply a proper pretreatment before embarking on the calibration modeling to remove those disturbing factors so that the process information that shows the highest correlation with or contributes most to the concerned quality variations can be paid enough attention to.

To address the above problems, various strategies have been developed, in which, variable selection [6–12] and orthogonal signal correction (OSC) [13–19] are the most popular preprocessing means. They can improve the prediction performance of the sequent calibration model because of the increased causal relationship between the predictor and quality variables. Further more, the filtered descriptors provide more accurate process knowledge and understanding in determining the quality performance since only key and relevant information are focused on after removal of disturbing variations. Several strategies for variable selection have been suggested [6–12], which attempted to choose a set of descriptors that produces the most correlated relationship with qualities. The selection result directly condenses the calibration model structure by reducing the dimensionality of regression coefficient matrix. In

* Corresponding author. Tel.: +86 24 83687434; fax: +86 24 23890912.
*E-mail address:* flwang@mail.neu.edu.cn (F. Wang).

0169-7439/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.chemolab.2008.09.003

contrast, the underlying idea of OSC [13–19] suggests that a logical approach would be to pre-treat process measurement by subtracting from it some factors that account for as much as possible of the variation in descriptor matrix (**X**) while remaining orthogonal to property matrix (**Y**). Therefore, the corrected descriptors can explain the quality variations more exactly than the original ones. Consequently, OSC filtering indirectly predigests calibration model by reducing the required number of latent variables, and thus the interpretation of the model becomes easier. OSC algorithm was first introduced by Wold et al. [13] and applied to calibration transfer by Sjöblom et al. [14]. Since then, various versions of OSC algorithms have emerged as a powerful and widely applicable preprocessing tool, in which the filtered data constitute an optimal platform for further calibration modeling according to model complexity, predictive ability and interpretation. Recently, Johan Trygg et al. developed a novel calibration modeling algorithm as an extension of PLS, called orthogonal partial least squares (OPLS) [20–23], which analyzed the disturbing variation in each regular PLS component with objective to improve interpretation of PLS models and reduce model complexity. Further O2-PLS method [24–26] was presented by the same authors, which has an integral OSC filter that separates the structured noise in both **X** and **Y**. It divides the systematic information in **X** and **Y** into three parts (**Y**-orthogonal, joint and **X**-orthogonal) to facilitate an improved process understanding and information extraction since relevant and irrelevant variations are discriminated and analyzed separately. It, thus, enhances the causal relationship between descriptors and response variables better than OPLS and conventional PLS algorithms.

Recently, various work [10,27–31] have been reported focusing on the local time-specific effects of process variations on process comprehension and quality prediction. In the proposed phase-based quality prediction method by Lu and Gao [31], they checked the critical-to-quality phases and developed the phase-representative quality inferential models, which can get earlier quality prediction without having to wait until the end of process operation for online application. However, their phase-representative calibration models were derived by averaging all time-slice regression relationships within the same phase, which are prone to induce overfitting problem. The disadvantage may result from such a cognition that there might be no regular and stable correlation existing between process variations and quality properties at each time, whereas some so-called regression relationship is toughly derived and is excessively expected to provide satisfying prediction capability. It overlooks the time-cumulative contributions of process variations to quality interpretation within the same phase, in which credible and stable regression relationship can't be favorably derived until the input descriptors spanning some segment of time region are taken into consideration and involved in calibration modeling. In real multiphase industrial processes, various phases generally operate orderly under the domination of different physical phenomena, revealing different effects on the final qualities, both local and cumulative. Therefore, product qualities depend on not only the overall level of each phase but also the operation statuses of all phases within the entire process duration, in which different phases may contribute to different part of quality variations with different extents. Considering that multiplicity of phase is an inherent nature of many batch processes, therefore, it is necessary to divide a process into several meaningful process phases, analyze the phase-specific local and cumulative phenomenon in detail and thus develop the corresponding quality prediction algorithm for complex chemical processes. However, up to now, few literatures have been reported to comprehensively capture and analyze both aspects simultaneously although multiphase strategy has been given great insights as the flexible modeling approach.

In this paper, we concentrate on multiphase knowledge extraction and investigation, including phase-specific process analysis, calibration modeling and quality interpretation and prediction, which is developed on two levels respectively using phase-specific O2-PLS and conventional PLS technique. The proposed algorithm highlights the fact that in real chemical processes, each phase may only explain one part of quality property and none but all phases work together to describe the overall quality variations. The phenomenon is called "phase-specific local and cumulative effects" here. Once the argument is recognized, it is not advisable to strive for excessive prediction precision localized in an individual phase. So our emphasis has been on how to find the different local contribution powers of various phases and their collective efforts to express the final qualities. First the process duration has to be properly divided into different phases and phase-specific modeling units are formed by data unfolding batch-wise, revealing different process variation characteristics. Then a two-level statistical analysis and modeling technique is formulated. In the first-level procedure, we focus on the local analysis referring to quality interpretation and prediction. Wherein, variable selection technique is adopted to remove uninformative descriptors prior and the phase-wise O2-PLS method is then smartly utilized to figure out within the current phase the responsible descriptor variations and their governed quantity and characteristics of quality variations, which can readily achieve this by dividing the systematic variations in both **X** and **Y** into different parts. Moreover, a variant bootstrapping technique is incorporated into the first-level modeling procedure, which produces multiple results of variable selection and OSC filtering focusing on different bootstrapping subsets. An integrated solution can then be obtained by synthetically analyzing them, which has advantages of reducing dependence on reference modeling samples, and thus enhances the robustness. In the second-level procedure, conventional PLS algorithm is used to consider and conduct a cumulative analysis. Whenever an evolving process completes and quality-relevant process covariations of various phases are derived, conventional PLS can be used to model the relationship between various local phases and the final qualities, in which, the effects of various phases are stacked by regression weights imposed on each phase. Therefore, performing O2-PLS and conventional PLS algorithms respectively on different levels, we can better reveal the details of both process and quality variations and thus enhance one's process understanding and improve the performance of quality interpretation and prediction.

This paper is organized as follows. First, the details of the proposed method are described in Section 2. Then, the effectiveness and feasibility of the proposed predicting method are illustrated by applying it to a benchmark simulation of fed-batch penicillin fermentation production in Section 3. Finally, conclusions are drawn in the last section.

## 2. Methodology

The purposes of phase-wise calibration analysis are twofold: to explore the local effect of each individual phase on quality variations and cumulate their effects to jointly explain the quality variations. In this section, the proposed method will be dwelt on focusing on the above-mentioned two aspects to enhance process understanding and improve the performance of quality prediction.

In each batch run, assume that $J_x$ process variables are measured at $k=1,2,\ldots,K$ time instances throughout the batch and $J_y$ quality variables at the end of each batch. Then vast amount of process observations collected from similar $I$ batches can be organized as a three-way array $\mathbf{X}\,(I\times J_x\times K)$ and a corresponding quality matrix $\mathbf{Y}$ of dimension $I\times J_y$. The means of each column are subtracted to approximately eliminate the main non-linearity due to the dynamic behaviors of the process and look at the deviation from the average trajectory. And each variable is scaled to unit variance to handle different measurement units, thus giving each equal weight. In the present work, the batches are of equal length without special declaration so that the specific process time can be used as an