This document contains the **post-print pdf version** of the refereed paper:

“Quality assessment of a variance estimator for Partial Least Squares prediction of batch-end quality”

by *Jef Vanlaer, Geert Gins, and Jan F.M. Van Impe*

which has been archived on the university repository Lirias (<https://lirias.kuleuven.be/>) of the KU Leuven.

The content is identical to the content of the published paper, but without the final typesetting by the publisher.

When referring to this work, please cite the full bibliographic info:

J. Vanlaer, G. Gins and J.F.M. Van Impe (2013). Quality assessment of a variance estimator for Partial Least Squares prediction of batch-end quality, Computers and Chemical Engineering, 52, 230-239.

The journal and the original published paper can be found at:

<http://www.journals.elsevier.com/computers-and-chemical-engineering>

<http://dx.doi.org/10.1016/j.compchemeng.2013.01.012>

The corresponding author can be contacted for additional info.

Conditions for open access are available at:

<http://www.sherpa.ac.uk/romeo/>

Quality assessment of a variance estimator for Partial Least Squares prediction of batch-end quality

Jef Vanlaer^a, Geert Gins^a, Jan F.M. Van Impe^{a,*}

^a*Chemical and Biochemical Process Technology and Control (BioTeC), Department of
Chemical Engineering, KU Leuven, W. de Croylaan 46, B-3001 Heverlee (Leuven),
Belgium*

Abstract

This paper studies batch-end quality prediction using Partial Least Squares (PLS). The applicability of the zeroth-order approximation of Faber & Kowalski (1997) for estimation of the PLS prediction variance is critically assessed. The estimator was originally developed for spectroscopy calibration and its derivation involves a local linearization under specific assumptions, followed by a further approximation. Although the assumptions do not hold for batch process monitoring in general, they are not violated for the selected case study. Based on extensive Monte Carlo simulations, the influence of noise variance, number of components and number of training batches on the bias and variability of the variance estimation is investigated. The results indicate that the zeroth-order approximation is too restrictive for batch process data. The development of a variance estimator based on a full local linearization is required to obtain more reliable variance estimations for the development of prediction intervals.

Keywords: Batch processes, Partial Least Squares, end-quality prediction, variance estimation

*Corresponding author, Tel: +32.16.32.14.66, Fax: +32.16.32.29.91

Email addresses: jef.vanlaer@cit.kuleuven.be (Jef Vanlaer),
geert.gins@cit.kuleuven.be (Geert Gins), jan.vanimpe@cit.kuleuven.be (Jan F.M. Van Impe)

1. Introduction

Today's chemical and biochemical production processes and plants are equipped with numerous sensors that measure various flow rates, temperatures, pressures, pH, concentrations, etc. These online and readily available measurements are a major source of information to monitor and control the processes. However, despite the frequent use of low-level controllers (e.g., PID control for valve opening and closing), detection of and response to abnormal events often remains a manual operation. Process operators visually inspect measurement profiles and compare these to measurements from previous process runs. This task is often hampered by the large size and complexity of modern interconnected process plants. Therefore, the development of automated monitoring systems to assist human process operators in their decisions still remains an important challenge for the chemical and life sciences industries (Venkatasubramanian et al., 2003).

For batch processes in particular, close process monitoring is of utmost importance because of the following two reasons. Firstly, batch processes are commonly used for the production of goods with high added value (e.g., medicines, high performance polymers and enzymes), which leads to high costs associated to lost batches. Secondly, an assessment of product quality is often only possible by offline measurements of a parameter (e.g., purity, enzyme concentration, polymer viscosity) after completion of the batch run. However, the development of adequate techniques for monitoring and control is complicated by the dynamic nature of batch processes.

Multivariate modelling techniques exploit the information in historical databases for process monitoring. Examples include Principal Component Analysis (PCA; Jolliffe, 1986) and PCA-based techniques such as Principal Component Regression (PCR; Geladi & Kowalski, 1986) and Iterative PCA (IPCA; Narasimhan & Shah, 2008), or Partial Least Squares (PLS; Geladi & Kowalski, 1986). These techniques are not only suited for the detection of process faults (e.g., Gregersen & Jørgensen (1999); Nomikos & MacGregor (1995b)), but also for the development of inferential sensors to estimate quality variables that are not measured online, for instance the final quality of a batch process (García-Munoz et al., 2004; Nomikos & MacGregor, 1995a; Ündey et al., 2003). Since PLS-based techniques take information in the outputs into account during model training, they are more suited for regression purposes than PCA-based models. Although PLS has been proven to be a valuable tool in several applications (Facco et al., 2009; Faggian et al., 2009;

Gins et al., 2012; Lopes & Menezes, 2003; Marjanovic et al., 2006), practical implementations in the process industry are limited.

A PLS model for batch-end quality prediction describes a linear relation between online process measurements and the final product quality. The PLS algorithm to obtain the model coefficients, however, has highly non-linear internal properties. As a consequence, knowledge about the statistical properties of PLS models is limited. Moreover, distributional knowledge about the PLS predictor is completely lacking. Hence, the construction of prediction intervals for PLS estimations is nearly impossible, which constitutes one of the method's major drawbacks (Denham, 1997). While it is commonly known that PLS filters noise from correlated measurements in large data sets, no distinct relation has been identified to describe the influence of measurement noise on PLS predictions of final batch quality.

The variance of PLS model predictions and the construction of prediction intervals has been the topic of several research papers (see Section 3 for a brief overview). However, most research has been performed in the field of spectroscopic calibration. The obtained results are not fully transferable to batch process monitoring due to some specific assumptions. First, Lambert-Beer's law for spectroscopy guarantees a linear relation between spectroscopy measurements and analyte concentration. In contrast, the relation between online process measurements and batch product quality might contain (small) non-linear contributions that cannot be captured by a linear PLS relation. Second, it is generally assumed that spectroscopy measurements contain all necessary information for prediction of the analyte concentration, while important explanatory variables are often missing in batch process monitoring because they cannot be measured online. Nevertheless, relations from spectroscopy research are an interesting starting point in the investigation of the variance of PLS batch-end quality predictions.

In this work, the applicability of the zeroth-order approximation of Faber & Kowalski (1997) for estimation of the PLS prediction variance in case of errors in the dependent variables to MPLS predictions of batch-end quality is critically assessed. This relation is of particular importance since it forms the basis for further research on PLS prediction intervals (see Section 3). Moreover, the case of noisy dependent variables is especially relevant to batch process monitoring since batch quality measurements are often obtained from offline laboratory measurements with relatively high uncertainty. Computer-generated data of an industrial-scale fermentation process for penicillin production are used as a case study. Two aspects of the equa-

tion derived by Faber & Kowalski (1997) are studied based on extensive Monte Carlo simulations: (i) the validity of the assumptions made in the derivation of the relation and (ii) the bias and variability of the variance estimation. The goal of this study is to detect the conditions under which the relation from spectroscopy can be transferred to applications in batch process monitoring.

The structure of the paper is as follows. In Section 2 batch-end quality prediction using Multiway PLS is briefly explained. Section 3 provides a (non-exhaustive) overview of research on PLS prediction variance. Next, Section 4 describes the case study used in this work to assess the variance estimation of batch-end quality predictions. The results of this study are presented and discussed in Sections 5 and 6 and final conclusions are drawn in Section 7.

2. Multiway Partial Least Squares for batch-end quality prediction

To deal with the particular structure of batch process data in predicting final batch quality, general Partial Least Squares (PLS) was extended to Multiway Partial Least Squares (MPLS) by Nomikos & MacGregor (1995b). The modelling procedure consists of two steps. In the first step, the data matrix containing historical data of normal process operation, which has a three-dimensional structure, is unfolded to a two-dimensional matrix, as explained in Section 2.1. Next, this unfolded data matrix, together with the quality measurements, forms the input for the construction of a general Partial Least Squares model. Section 2.2 provides the details for this second step.

2.1. Data matrix unfolding

When historical data of I batches, containing measurements of J different sensors over K time points are available, this information is generally presented as a three-dimensional data tensor $\underline{\mathbf{X}}$ of size $I \times J \times K$. For the construction of a PLS model, a two-dimensional data matrix is needed. Therefore, the tensor $\underline{\mathbf{X}}$ has to be unfolded to a two-dimensional structure. Several unfolding techniques are available, of which variable-wise unfolding (Wold et al., 1987) and batch-wise unfolding (Nomikos & MacGregor, 1994, 1995a) are the most commonly used. Since batch-wise unfolding preserves the batch direction by placing K slices of size $I \times J$ of the original tensor $\underline{\mathbf{X}}$ next to each other, this technique is most suited for batch-end quality

Figure 1: Illustration of batch-wise data matrix unfolding.

prediction. Every row of the unfolded matrix $\mathbf{X}$ corresponds to one complete batch so that the complete batch history is taken into account to predict the final batch quality. Fig. 1 illustrates the procedure.

2.2. Multiway Partial Least Squares (MPLS)

After data matrix unfolding, a standard two-dimensional PLS model is constructed to relate the unfolded input data matrix $\mathbf{X}$ to the output matrix $\mathbf{Y}$ ($I \times L$), which contains L quality measurements for each batch in each row (Geladi & Kowalski, 1986).

$$\begin{cases} \mathbf{X} = \mathbf{T}\mathbf{P}^T + \mathbf{E}_\mathbf{X} \\ \mathbf{Y} = \mathbf{T}\mathbf{Q}^T + \mathbf{E}_\mathbf{Y} \end{cases} \quad (1)$$

In PLS, R PLS components or latent variables are computed as linear combinations of the original measurements. The components are chosen in such a way that as much information (covariance) as possible from the original measurements is retained. The input and output matrices are projected onto a lower-dimensional space of which the dimensions are defined by the components. The loading matrices $\mathbf{P}$ ($JK \times R$) and $\mathbf{Q}$ ($L \times R$) specify the projections of $\mathbf{X}$ and $\mathbf{Y}$, respectively. The scores matrix $\mathbf{T}$ ($I \times R$) is the representation of the data matrices in the reduced space. The residuals or modelling errors are contained in the matrices $\mathbf{E}_\mathbf{X}$ and $\mathbf{E}_\mathbf{Y}$.

Since the matrix $\mathbf{P}$ is not invertible and has no orthonormal columns, a weight matrix $\mathbf{W}$ ($JK \times R$) with orthonormal columns is introduced to compute the scores $\mathbf{T}$ and quality prediction $\mathbf{Y}$ for a new measurement set $\mathbf{X}_{\text{new}}$. $\mathbf{P}^T\mathbf{W}$ is invertible so that the projection of the input measurements $\mathbf{X}$ on the scores space $\mathbf{T}$ and the corresponding regression matrix $\mathbf{B}$ ($JK \times R$)

are calculated as follows.

$$\hat{\mathbf{T}} = \mathbf{X}_{\text{new}} \mathbf{B} \quad (2)$$

$$\mathbf{B} \triangleq \mathbf{W} (\mathbf{P}^T \mathbf{W})^{-1} \quad (3)$$

Combining Eqs. 1 and 2 leads to the following relation between the estimated output matrix $\hat{\mathbf{Y}}$ and the new input matrix $\mathbf{X}_{\text{new}}$.

$$\hat{\mathbf{Y}} = \hat{\mathbf{T}} \mathbf{Q}^T = \mathbf{X}_{\text{new}} \mathbf{B} \mathbf{Q}^T \quad (4)$$

In most applications of batch-end quality prediction, only one quality parameter is considered. In this case, the output matrix $\mathbf{Y}$ reduces to a quality vector of size $(I \times 1)$ containing the final quality for all I batches.

3. Variance of PLS predictions

The development of prediction intervals for PLS estimations has been the subject of several research papers, especially in the field of spectroscopic calibration. Since very little is known on the distributional properties of the PLS estimator, non-parametric methods of sample reuse, such as bootstrapping, cross-validation and noise addition, are often used (Faber, 2002; Reis & Saraiva, 2012). In these computationally intensive methods, information on the variance of the PLS estimator for the construction of prediction intervals is gained from the training data. However, these methods do not always yield accurate results because often only few observations are available for training (Denham, 1997). Moreover, the results are case specific and no general information on PLS prediction variance can be extracted.

An alternative approach is the derivation of an (approximate) analytical relation for the prediction variance by linearization of the PLS estimator. Early research in this field was conducted by – amongst others – Höskuldsson (1988) and Phatak et al. (1993). Based on this work, Faber & Kowalski (1997) constructed approximate relations for the prediction variance and prediction intervals using error propagation. In the case where measurement errors are present in the dependent variables only, the following relation is obtained:

$$\hat{V}(\hat{y}_u) = (N^{-1} + h_u) \sigma_{\Delta y}^2. \quad (5)$$

While Eq. (5) is exact for ordinary least squares regression and PCR, it corresponds to a zeroth-order approximation for PLS. The equation relates

the variance $\hat{V}(\hat{y}_u)$ of a prediction for an unknown sample to the variance $\sigma_{\Delta y}^2$ of the noise on the output measurements. An influence of the number of training measurements N seems logical since the availability of more training data should lead to more precise predictions. The leverage h_u is a measure for the distance of the unknown sample to the training data in the low-dimensional PLS hyperplane. Including it in the equation leads to sample-specific prediction errors. For zero-intercept models, the leverage is calculated as

$$h_u = \mathbf{t}^T (\mathbf{T}^T \mathbf{T})^{-1} \mathbf{t} \quad (6)$$

with $\mathbf{t}$ the unknown sample score vector and $\mathbf{T}$ the training scores matrix (Faber & Bro, 2002).

Eq. (5) is referred to as a zeroth-order approximation because it involves an additional approximation to the local linearization of the PLS regression vector (Faber & Kowalski, 1997). The equation is easy and interpretable, and provides a clear view on the important aspects that influence prediction variance. However, due to the linearization and subsequent approximation, it should be used with great care.

The zeroth-order approximation of Faber & Kowalski (1997) is particularly interesting because it forms the basis for many other PLS variance equations in literature. In their work, Faber & Kowalski (1997) do not only take noise on measurements of the dependent variables into account. Homoscedastic measurement errors in the independent variables are included in extensions to Eq. 5. Application of the resulting equation on several near-infrared data sets yields promising results (Fernández Pierna et al., 2003). The obtained relations were further extended to include variance due to modelling errors (residuals) (Faber & Kowalski, 1996) and heteroscedastic noise in the independent variables (Bro et al., 2005; Faber & Bro, 2002). Recently, a first-order approximation for prediction intervals using trilinear PLS, developed in the prediction error framework, was proposed (Serneels et al., 2011).

Next to the importance of Eq. 5 as a basis for many PLS variance equations in literature, the case of noisy dependent variables is also especially relevant to batch process monitoring. Batch quality measurements (the outputs $\mathbf{Y}$ used in PLS model training) are most often obtained from offline laboratory measurements with relatively high uncertainty. The model inputs $\mathbf{X}$ on the other hand are often online measurements of pressures, temperatures and flows, which can be measured more accurately.

However, Faber & Kowalski (1997) warn not to use the equation in other areas of data analysis than spectroscopic calibration without careful consideration. Apart from the applied approximations, the derivation of Eq. 5 involves some important assumptions. First, an exact linear relation is assumed between the dependent and independent variables. This assumption is motivated by Beer's law in spectroscopy, which prescribes a linear relation between analyte concentration and spectral responses. Further, the derivation assumes that the data set is complete, which means that no important explanatory variables are missing.

While these assumptions hold for PLS predictions in spectroscopic calibrations, they are likely to be violated when dealing with (bio)chemical batch processes. Non-linear reaction kinetics, non-linear behaviour of equipment, non-homogeneity of the reactors, etc. may have a (possibly small) impact on the relation between online measurements and final batch quality. Furthermore, although these processes are equipped with many sensors, measurements of certain variables that are expected to be important for batch-end quality prediction (e.g., biomass concentration for biochemical processes or monomer concentration and melt index in batch polymerization) are generally not available online. Nonetheless, Eq. (5) may still serve as a starting point for the development of prediction intervals in MPLS batch-end quality prediction.

In this work, the applicability of the relation proposed by Faber & Kowalski (1997) for estimation of MPLS batch-end quality prediction variance is assessed critically. Based on extensive Monte Carlo simulations, the validity of the above assumptions is checked and the bias and variability of the variance estimations are studied to detect the conditions under which Eq. 5 can be applied to batch process monitoring.

4. Case study

The Monte Carlo simulations in this work are performed on a computer-generated data set of an industrial-scale fermentation process for penicillin production. Section 4.1 describes the data set, while Section 4.2 provides the details on the Monte Carlo simulations.

4.1. Data

Extensive Monte Carlo simulations require a large set of batch process data with adjustable noise levels. Therefore, computer-generated data orig-

Table 1: Properties of the initial conditions of the state variables in the case study.

Initial conditions				
Variable	Mean	Stand.dev.	Lower limit	Upper limit
Substrate conc. [g/L]	17.5	1	15	20
Biomass conc. [g/L]	0.125	0.03	0.05	0.20
Volume [L]	102.5	5	90	115
DO conc. [g/L]	1.16	—	—	—
Penicillin conc. [g/L]	0	—	—	—
CO ₂ conc. [g/L]	0.4487	—	—	—
pH [g/L]	5	—	—	—
Reactor temp. [K]	298	—	—	—
Reaction heat [cal]	0	—	—	—

inating from *Pensim* (Birol et al., 2002) are used as a case study. *Pensim* is based on a detailed unstructured model for penicillin production in an industrial-scale biochemical reactor. Penicillin is produced as a secondary metabolite during fermentation by filamentous microorganisms. The process consists of two phases. During the initial batch phase, the microorganisms grow on the glucose (the main substrate) available in the broth. Once the substrate concentration drops below 0.3 g/L, the reactor is switched to fed-batch mode. A continuous feed stream containing additional substrate is introduced in the reactor during this phase. Due to the low substrate concentration in the reactor, the microorganisms produce penicillin. When a total of 25 L of substrate feed has been added to the reactor, the process is terminated. Reactor temperature and pH are controlled at their respective setpoints of 298 K and 5 by standard PID controllers during both phases. For the model equations, parameter values and controller tunings, the reader is referred to the original *Pensim* paper of Birol et al. (2002).

Data of 150 batches are generated to serve as a training set to build MPLS models for the prediction of the final penicillin concentration. The initial substrate concentration, biomass concentration, and culture volume are subject to random variations for each batch by sampling them from a standard normal distribution. The properties of the initial conditions are summarized in Table 1. Additionally, the profiles of several process inputs exhibit small fluctuations around their setpoints to represent process variability, modelled

Table 2: Setpoints and size of the PRBS signal of the fluctuations on the process inputs in the case study.

Process inputs		
Input variable	Setpoint	PRBS deviation
Substrate feed rate [L/h]	0.06	0.005
Aeration rate [L/h]	8	0.3
Agitator power [W]	30	1
Feed temperature [K]	296	0.5

Table 3: Online measurements available from Pensim.

Dissolved oxygen concentration	Agitator power
Volume	Feed temperature
pH	Water flow rate
Reactor temperature	Base flow rate
Substrate feed rate	Acid flow rate
Aeration rate	

by averaging a pseudo-random binary sequence (PRBS) over 1000 samples. A summary of the input setpoints and the size of the PRBS signals is given in Table 2. *Pensim* provides measurements of 17 process variables, 11 of which are generally acquired from online sensors and thus practically available as model inputs for online prediction of the batch-end penicillin concentration. Table 3 provides an overview of these online measurements. To focus on the validity of Eq. 5 (which was derived for noise on the dependent variables only) and avoid interference of the uncertainty on the independent variables with the results of the study, no noise was added to the measurements of the model inputs. 50 extra batches are generated to validate the results of the Monte Carlo simulations.

The measured signals are aligned and resampled via *indicator variables*, following the procedure of Ündey et al. (2003). Employing the culture volume and percent substrate fed as indicator variables for the batch and fed-batch phase respectively, the length of the measurement profiles is equalized to 602 samples for all batches (101 samples for the batch phase and 501 for the fed-batch phase). The time signal is added to the input measurements after alignment, so that 12 online measurement signals are available for every batch. This yields a training data matrix $\underline{\mathbf{X}}_{\text{train}}$ of size $150 \times 12 \times 602$ and a validation data matrix $\underline{\mathbf{X}}_{\text{val}}$ of size $50 \times 12 \times 602$.

4.2. Monte Carlo simulations

Extensive Monte Carlo simulations are performed to critically assess the applicability of Eq. 5 for estimation of the variance of MPLS batch-end quality predictions. Three factors are varied in the Monte Carlo simulations: (i) the output noise standard deviation, (ii) the number of training batches and (iii) the number of PLS components. For every combination of these factors, 1000 predictions of the final penicillin concentration are made with the output noise each time being resampled from a Gaussian distribution with zero mean, both for the training and the validation batches. Noise standard deviation ranges from 1 to 10% of the mean final penicillin concentration of the training batches. 15 levels of the number of training batches N_{tr} are chosen in such a way that approximately equidistant ranges of $1/N_{tr}$ are obtained ($N_{tr} \in \{10, 11, 12, 13, 14, 15, 17, 19, 21, 25, 30, 38, 50, 75, 150\}$). Finally, models with 1 to N_{tr} latent variables are trained. To save computation time only multiples of 5 and 10 are taken into account for the higher component numbers.

5. Predictive performance

As explained in Section 3, the assumptions of (i) a linear relation between measurements and final batch quality and (ii) fully informative data, made in the derivation of Eq. 5, are not necessarily met for batch-end quality prediction. On simulation level however, it is possible to check the validity of these assumptions by training an MPLS model based on noiseless data. This is done in Section 5.1 for the present case study. Section 5.2 illustrates and discusses the predictive performance of MPLS models in the case of noisy output measurements.

An assessment of the predictive performance of MPLS models is performed by comparing the actual final penicillin concentration of the validation batches, the quality measurements and the MPLS predictions based on the Mean Squared Error (MSE) and the correlation coefficient R^2 , defined by the following equations:

$$MSE = \frac{SSE}{N_{val}} \quad (7)$$

$$R^2 = 100 \left(1 - \frac{SSE}{SSTO} \right) \quad (8)$$

with N_{val} the number of validation batches, SSE the Sum of Squared Errors, and $SSTO$ the Total Sum of Squares:

$$SSE = \sum_{i=1}^{N_{val}} (y_i - y_{i,ref})^2 \quad (9)$$

$$SSTO = \sum_{i=1}^{N_{val}} (y_{i,ref} - \bar{y}_{ref})^2 \quad (10)$$

where $\bar{y}_{ref}$ corresponds to the mean of the reference data $y_{i,ref}$.

Depending on the data (y_i and $y_{i,ref}$) used to calculate the SSE and $SSTO$, different types of MSE and R^2 are obtained. Comparing measured to actual final penicillin concentrations ($y_i = y_{i,meas}$, $y_{i,ref} = y_{i,act}$) results in the measurement performance metrics MSE_{meas} and R^2_{meas} . A comparison of MPLS predictions to the measurements ($y_i = \hat{y}_i$, $y_{i,ref} = y_{i,meas}$) yields the apparent metrics MSE_{app} and R^2_{app} . Finally, the actual performance metrics MSE_{act} and R^2_{act} compare model predictions to the actual penicillin concentrations ($y_i = \hat{y}_i$, $y_{i,ref} = y_{i,act}$). For noiseless data ($y_{i,meas} = y_{i,act}$), MSE_{app} and MSE_{act} are equal and simply referred to as MSE . The same applies to the correlation coefficient.

5.1. Noiseless data

An MPLS model is trained on data of the 150 training batches without output noise to analyze the predictive performance of MPLS models for batch-end quality prediction. The measurements are centered around the average profile and scaled to unit variance before applying batch-wise unfolding. The number of latent variables is selected using an adjusted Wold's R criterion (Li et al., 2002). According to this criterion, the optimal number of PLS components corresponds to the smallest number of components R for which the following equation holds:

$$\frac{MSE(R+1)}{MSE(R)} > \alpha \quad (11)$$

with $MSE(R)$ the (crossvalidation) Mean Squared Error of the MPLS model with R components. Employing this criterion, the $(R+1)^{\text{th}}$ component is only added if it significantly improves the prediction and thus decreases the crossvalidation error according to a chosen significance value α . The adjusted Wold's R criterion is similar to other criteria that also use the ratio of subsequent MSE or SSE values to identify the optimal number of latent variables, such as the F-test proposed by Osten (1988). A comparison of selection procedures by Li et al. (2002) leads to the conclusion that the adjusted Wold's R criterion with a selection threshold $\alpha = 0.90$ or $\alpha = 0.95$ yields parsimonious models with the highest probability of selecting the true model. The threshold α is determined empirically. Employing $\alpha = 0.95$ results in the selection of 7 components for the present case study.

Fig. 2 shows a very good agreement between the predicted batch-end penicillin concentration and the actual final concentration for the 50 validation batches. A validation MSE of 3.45×10^{-6} and a correlation coefficient $R^2 = 99.94\%$ also indicate an excellent prediction performance. These results suggest that a linear PLS relation is adequate to predict the batch-end quality and that the online measurements contain all relevant information for the present case study. This means that Eq. 5 should be applicable since the assumptions made during its derivation are not violated. However, this result is case-specific and should not be generalized.

5.2. Noisy data

Fig. 3 illustrates the power of MPLS models for batch-end quality predictions with noisy output measurements. White Gaussian noise with a standard deviation of 10% of the mean final penicillin concentration was added to

Figure 2: Predicted versus actual final penicillin concentration for 50 validation batches in the noiseless case. An MPLS model including 7 components was used.

Figure 3: Measured versus actual final penicillin concentration (a), predicted versus measured final penicillin concentration (b) and predicted versus actual final penicillin concentration (c) for 50 validation batches with white Gaussian output noise with a standard deviation of 10% of the mean final concentration. An MPLS model including 1 component was used.

both the 150 training batches and the 50 validation batches. The penicillin concentration measurements of the validation batches are plotted against their actual values in Fig. 3(a). It is clear that the measurements deviate considerably from the real concentration values, so that it is nearly impossible to distinguish high quality from low quality batches based on these measurements. The high noise level results in an MSE_{meas} of 3.53×10^{-2} for the validation batches, which is obviously comparable to the noise variance (3.45×10^{-2}). The correlation coefficient $R_{meas}^2 = -544\%$ also indicates very bad measurements.

Using an adjusted Wold's R criterion (Eq. 11) with a significance value 0.95 only 1 component is selected. Based on a plot of the predicted penicillin concentration versus the measured concentration (Fig. 3(b)), one could assume a very bad prediction performance. The apparent correlation coefficient ($R_{app}^2 = 6.33\%$) seems to corroborate this conclusion. However, the apparent MSE (MSE_{app}) equals 3.53×10^{-2} , which is equal to MSE_{meas} and very close to the noise variance. Hence, the apparent bad prediction could be caused by the bad measurements.

On a simulation level, it is possible to plot the model predictions as a function of the actual penicillin concentrations. Fig. 3(c) shows a very good correspondence between predictions and real values in validation. This is corroborated by an MSE_{act} of 7.97×10^{-4} and a correlation coefficient R_{act}^2 of 85.5%. Despite the high levels of noise present in both training and validation data, the MPLS model with only 1 principal component achieves a good prediction of the final penicillin concentration. Obviously, the model is better suited to distinguish between high and low quality batches than the measurements. Comparing model predictions to actual quality variables is not possible in industrial practice, where – in general – only quality measurements including measurement noise are available. Care should be taken when apparent MSE values for validation batches are comparable to or only slightly higher than the noise variance, since model predictions may be substantially better than expected.

Fig. 4 generalizes the above conclusions to other noise levels. The figure depicts the actual prediction MSE for the *worst case scenario* (i.e., the Monte Carlo simulation that yields the highest MSE_{act}), divided by the noise variance, as a function of the number of components for noise standard deviations ranging from 1 to 10% of the mean quality. A value of this scaled worst case prediction MSE of 1 indicates a prediction of the same quality as the measurements. From the figure, it is clear that a range of component

Figure 4: Worst case MSE over 1000 Monte Carlo simulations, scaled with the noise variance, in function of number of components for noise standard deviations ranging from 1 to 10% of the mean quality.

numbers can be found where the prediction outperforms the measurement for every noise level, even in the worst case scenario. This corroborates the fact that PLS models capture the most important information in the measurements, remove irrelevant information and discard most of the measurement noise.

6. Quality of the variance estimator

Eq. 5 describes a linear relation between prediction variance and noise variance. This linear relation is confirmed by calculating the empirical mean prediction variance from the results of the Monte Carlo simulations, as depicted in Fig. 5. However, a good estimator for batch-end quality prediction variance exhibits low bias and low variability¹. As evidenced by Fig. 6, which compares the estimated to the observed variance of the prediction over 1000 Monte Carlo simulations, Eq. 5 exhibits considerable bias and variability. This section investigates the quality of the variance estimator in terms of bias and variability. The influence of the noise variance, the number of components and the number of training batches is examined based on results of the Monte Carlo simulations.

¹To avoid confusion with the PLS prediction variance, the variation in the variance estimations obtained by employing Eq. 5 over the different Monte Carlo simulations will consistently be referred to as variability.

Figure 5: Empirical mean prediction variance over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of noise variance for MPLS models incorporating 1–10 components (bottom curve to top curve).

Figure 6: 95% range of batch-end quality prediction variance estimations from Eq. 5 for 50 validation batches employing an MPLS model with 1 component (*left*) and 10 components (*right*) using 150 batches for training and at a noise standard deviation of 10% of the mean final quality over 1000 simulations. The mean variance estimation is indicated with a bullet ($\bullet$). The observed prediction variance over the 1000 simulations is indicated with a diamond ($\diamond$).

Figure 7: Mean relative bias of Eq. 5 over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of noise variance for MPLS models incorporating 1 (○), 2 (□), 3 (△) and 4 (◇) components.

6.1. Estimator bias

The bias b_i of the variance estimation by Eq. 5 is calculated as the absolute value of the difference between the mean variance estimation $\hat{\bar{V}}(\hat{y}_i)$ for validation batch i (with $\hat{V}(\hat{y}_i)$ calculated from Eq. 5 for every simulation) and the observed prediction variance $V(\hat{y}_i)$ for the same batch over 1000 Monte Carlo simulations.

$$b_i = \left| \hat{\bar{V}}(\hat{y}_i) - V(\hat{y}_i) \right| \quad (12)$$

with

$$\hat{\bar{V}}(\hat{y}_i) = \frac{1}{1000} \sum_{mc=1}^{1000} \hat{V}(\hat{y}_i)_{mc} \quad (13)$$

$$V(\hat{y}_i) = \frac{1}{999} \sum_{mc=1}^{1000} ((\hat{y}_i)_{mc} - \overline{\hat{y}_i})^2 \quad (14)$$

A relative bias, that allows a comparison of the importance of bias under different conditions, is obtained by dividing the estimation bias by the observed prediction variance.

$$b_{i,rel} = \frac{b_i}{V(\hat{y}_i)} \quad (15)$$

Figure 8: Mean relative bias of Eq. 5 over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of the number of components with an output noise standard deviation of 10% of the mean final quality for models trained on 21, 30, 50, 75 and 150 training batches.

6.1.1. Influence of the noise variance

The bias of Eq. 5 increases with increasing noise level. The relative bias however is mostly constant with the noise variance. This observation is illustrated in Fig. 7, which shows the evolution of the relative bias (averaged over the 50 validations batches) with the noise variance for MPLS models incorporating 1–4 components when data of 150 batches are used for training. For 1 component, a quasi constant relative bias of 12% is obtained. Only for models incorporating 2 or 3 components, the relative bias increases with increasing noise variance. At 4 components the relative bias is again constant with the noise variance. Data for models with more components show the same trend. Similar observations are made when using fewer training batches, although the increasing trend is already observed for models incorporating only 1 component when the number of training batches becomes lower than 25.

6.1.2. Influence of the number of components

Fig. 7 already indicates a profound influence of the number of components in the model on the relative bias of the variance estimator. Fig. 8 illustrates the relation between the number of components and the relative bias for different number of training batches when the output noise standard deviation equals 10% of the final penicillin concentration. Numerical instabilities that lead to very high biases are sometimes observed for the highest component

Figure 9: Mean relative bias of Eq. 5 over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of the number of training batches for MPLS models incorporating 1 component (*left*) and 10 components (*right*) with an output noise standard deviation of 10% of the mean final quality.

numbers. For clarity reasons, these high component numbers are omitted from the figure. In practice, very high numbers of latent variables are never used because they also correspond to very high prediction variance.

For all numbers of training batches, the relation between relative estimator bias and the number of components is characterized by a fast increase at low component numbers after which a maximum relative bias is quickly reached. A slow decrease is observed at the higher component numbers. Acceptable bias is only obtained for either one or a very high number of components. In practice, high component numbers are never selected due to the high prediction variance. The reason for this evolution can be found in the neglecting of a first-order term in the derivation of Eq. 5, which is small only for very low and very high component numbers (Faber & Kowalski, 1997). The results of this case study indicate that this neglected term may take on large values when dealing with batch process data, resulting in high bias for the variance estimator. As clearly observed in Fig. 6 for the case of 10 components, the zeroth-order approximation of Eq. 5 severely underestimates the prediction variance when more than 1 PLS component is used in the model.

6.1.3. Influence of the number of training batches

The influence of the number of training batches is strongly linked with the number of components included in the model. Fig. 9 compares the relation between relative bias and the number of training batches for MPLS models incorporating 1 and 10 latent variables at a noise standard deviation of 10% of the mean final quality.

tion of 10% of the mean final penicillin concentration. For models with 1 component, which is evidently the lowest possible number of components for every number of training batches, relative bias decreases with the number of training batches. Including more training batches leads to better performance of the variance estimator and sufficient training batches should be available to obtain acceptable bias. For models incorporating 10 components on the other hand, relative bias increases with increasing number of training batches. As batch-wise unfolding leads to training matrices containing more columns than rows, the number of training batches (which corresponds to the number of rows) determines the maximal number of PLS components. Since relatively low bias is obtained at the high component numbers (as illustrated in Fig. 8), the lowest relative bias for 10 components is obtained for 10 training batches. When more training batches are available, the relative bias shifts closer to the maximum in the graph of bias vs. number of components.

6.1.4. Summary

In most cases, considerable bias is observed when applying the zeroth-order approximation of Faber & Kowalski (1997) to data of the case study. Due to the neglecting of a first-order term in the derivation of the estimator, the number of PLS components incorporated in the model has a large influence on the bias, sometimes attaining values up to 100% of the observed prediction variance. Acceptable bias is only obtained for 1 component or at very high component numbers, the latter of which are never used in practice (Fig. 8). For the more common low component numbers, the estimator bias decreases with the number of training batches. However, appropriate model order selection is important because the selection of even a few too many components leads to increased bias when employing more training batches (Fig. 9). Finally, the noise variance has no significant impact on the relative bias, except for low number of PLS components, where an increasing trend of bias with noise variance is observed, especially when few training batches are available (Fig. 7).

6.2. Estimator variability

The presence of the leverage in Eq. 5 leads to variability in the estimation of the prediction variance because the leverage is dependent on the specific noise sample. In this paper, the estimation variability is quantified by the width of the 95% empiric confidence interval (the 95% range r_i) over the 1000 Monte Carlo simulations, calculated by leaving out the 2.5% highest

Figure 10: Mean relative 95% range of Eq. 5 over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of noise variance for MPLS models incorporating 1 (○), 2 (□), 3 (△) and 4 (◇) components.

and lowest variance estimations and computing the difference between the maximum and the minimum of the remaining values for every batch.

$$r_i = \hat{V}_{97.5\%}(\hat{y}_i) - \hat{V}_{2.5\%}(\hat{y}_i) \quad (16)$$

A relative 95% range is calculated by dividing the 95% range by the mean estimation of the prediction variance. This relative range makes a comparison of the estimation variability under different conditions possible.

$$r_{i,rel} = \frac{r_i}{\hat{V}(\hat{y}_i)} \quad (17)$$

6.2.1. Influence of the noise variance

Fig. 10 illustrates the relation between the relative 95% range, averaged over the 50 validation batches, and the noise variance for models incorporating 1–4 PLS components and trained on 150 training batches. For the first three components, an increase in relative range with the noise variance is observed. The increase is largest for 2 and 3 components. Higher component numbers lead to a quasi constant relative range (although for 4 components, a slight increase is still visible at the lower values of the noise variance). The case of 150 training batches is representative for the cases with fewer training batches.

Figure 11: Mean relative 95% range of Eq. 5 over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of the number of components with an output noise standard deviation of 10% of the mean final quality for models trained on 21, 30, 50, 75 and 150 training batches.

6.2.2. Influence of the number of components

The evolution of the relative range with the number of components in the MPLS model is comparable to the evolution of the relative bias. A fast initial increase is observed, followed by a slower decrease, as shown in Fig. 11 for an output noise standard deviation of 10% of the mean batch-end penicillin concentration. When fewer than 20 training batches are available, the initial increase disappears and the relative range decreases monotonically with the number of PLS components.

6.2.3. Influence of the number of training batches

Finally, Fig. 12 illustrates the relation between the relative 95% range and the number of available training batches. Similar to the relative bias, the dependence of relative range on the number of training batches is linked to the number of components in the MPLS model. For models incorporating 1 component, the relative range decreases from 10 to 150 available training batches. For models containing 10 PLS components however, an increase with the number of training batches is observed, since 10 components corresponds with the maximal number of components and thus low variability in case of 10 available training batches.

Figure 12: Mean relative 95% range of Eq. 5 over 50 validation batches, calculated from Monte Carlo simulations with 1000 repetitions, as a function of the number of training batches for MPLS models incorporating 1 component (*left*) and 10 components (*right*) with an output noise standard deviation of 10% of the mean final quality.

6.2.4. Summary

Conclusions concerning the variability of prediction variance estimation employing Eq. 5 are very similar to those of the bias section. Estimations with low variability are only obtained for models employing either 1 or very high numbers of PLS components at low noise variance and when enough training batches are available. The practical applicability of Eq. 5 is very limited because high component numbers are never used due to the high associated prediction variance.

7. Conclusions

Multiway Partial Least Squares (MPLS) models have a large potential for the prediction of the final quality of batch processes, as was proven by extensive Monte Carlo simulations. However, a reliable estimator for the variance of MPLS batch-end quality predictions is not available. In this paper, the use of a variance estimator developed in the field of spectroscopy calibration – the zeroth-order approximation of Faber & Kowalski (1997) – for batch monitoring applications was evaluated critically. Although quality predictions on noiseless process data indicate that the specific assumptions made in the derivation of the estimator are not violated for the selected case study, high bias and variability of the variance estimations are observed for most combinations of output noise variance, number of PLS components and number of training batches. Acceptable estimations, exhibiting low bias and low variability, are only obtained at low noise variance for MPLS models in-

incorporating very few components when enough training batches are available. Although these results are obtained for simulation studies on a specific case study, the authors believe that using the zeroth-order approximation of Faber & Kowalski (1997) in other applications of batch-end quality prediction will also lead to underestimation of the PLS prediction variance.

The results indicate that the zeroth-order approximation is too restrictive for batch process monitoring applications, and causes highly unreliable estimations of the variance of MPLS predictions for models incorporating more than 1 latent variable. The development of a variance estimator based on the full local linearization, as was recently done for trilinear PLS (Serneels et al., 2011), could solve this problem. However, as long as no reliable estimator for batch-end quality prediction variance is available, variance estimations obtained from non-parametric methods such as bootstrapping or noise addition are to be preferred to estimators from spectroscopy based on the zeroth-order approximation.

Acknowledgements

Work supported in part by Project PFV/10/002 (OPTEC Optimization in Engineering Center) of the Research Council of the KU Leuven, Project KP/09/005 (SCORES4CHEM) of the Industrial Research Council of the KU Leuven, and the Belgian Program on Interuniversity Poles of Attraction initiated by the Belgian Federal Science Policy Office. J. Vanlaer is funded by a Ph.D grant of the Agency for Innovation by Science and Technology (IWT). J. Van Impe holds the chair Safety Engineering sponsored by the Belgian chemistry and life sciences federation essenscia. The scientific responsibility is assumed by the authors.

References

- Birol, G., Ündey, C., & Çinar, A. (2002). A modular simulation package for fed-batch fermentation: penicillin production. *Comput Chem Eng*, 26, 1553–1565.
- Bro, R., Rinnan, A., & Faber, N. (2005). Standard error of prediction for multilinear PLS. 2. Practical implementation in fluorescence spectroscopy. *Chemom Intell Lab Syst*, 75, 69–76.

- Denham, M. (1997). Prediction intervals in partial least squares. *J Chemom*, 11, 39–52.
- Faber, K., & Kowalski, B. (1996). Prediction error in least squares regression: Further critique on the deviation used in The Unscrambler. *Chemom Intell Lab Syst*, 34, 283–292.
- Faber, K., & Kowalski, B. (1997). Propagation of measurement errors for the validation of predictions obtained by principal component regression and partial least squares. *J Chemom*, 11, 181–238.
- Faber, N. (2002). Uncertainty estimation for multivariate regression coefficients. *Chemom Intell Lab Syst*, 64, 169–179.
- Faber, N., & Bro, R. (2002). Standard error of prediction for multiway PLS. 1. Background and simulation study. *Chemom Intell Lab Syst*, 61, 133–149.
- Facco, P., Doplicher, F., Bezzo, F., & Barolo, M. (2009). Moving average PLS soft sensor for online product quality estimation in an industrial batch polymerization process. *J Process Control*, 19, 520–529.
- Faggian, A., Facco, P., Doplicher, F., Bezzo, F., & Barolo, M. (2009). Multivariate statistical real-time monitoring of an industrial fed-batch process for the production of specialty chemicals. *Chem Eng Res Des*, 87, 325–334.
- Fernández Pierna, J., Jin, L., Wahl, F., Faber, N., & Massart, D. (2003). Estimation of partial least squares regression prediction uncertainty when the reference values carry a sizeable measurement error. *Chemom Intell Lab Syst*, 65, 281–291.
- García-Munoz, S., Kourti, T., & MacGregor, J. (2004). Model predictive monitoring for batch processes. *Ind Eng Chem Res*, 43, 5929–5941.
- Geladi, P., & Kowalski, B. (1986). Partial least-squares regression: a tutorial. *Anal Chim Acta*, 185, 1–17.
- Gins, G., Van den Kerkhof, P., & Van Impe, J. (2012). Hybrid derivative dynamic time warping for online industrial batch-end quality estimation. *Ind Eng Chem Res*, 51, 6071–6084.

- Gregersen, L., & Jørgensen, S. (1999). Supervision of fed-batch fermentations. *Chem Eng J*, 75, 69–76.
- Höskuldsson, A. (1988). PLS regression methods. *J Chemom*, 2, 211–228.
- Jolliffe, I. (1986). *Principal component analysis*. Springer Verlag, New York.
- Li, B., Morris, J., & Martin, E. (2002). Model selection for partial least squares regression. *Chemom Intell Lab Syst*, 64, 79–89.
- Lopes, J., & Menezes, J. (2003). Industrial fermentation end-product modelling with multilinear PLS. *Chemom Intel Lab Syst*, 68, 75–81.
- Marjanovic, O., Lennox, B., Sandoz, D., Smith, K., & Crofts, M. (2006). Real-time monitoring of an industrial batch process. *Comput Chem Eng*, 30, 1476–1481.
- Narasimhan, S., & Shah, S. (2008). Model identification and error covariance matrix estimation from noisy data using PCA. *Control Eng Pract*, 16, 146–155.
- Nomikos, P., & MacGregor, J. (1994). Monitoring batch processes using multiway principal component analysis. *AIChE J*, 40(8), 1361–1375.
- Nomikos, P., & MacGregor, J. (1995a). Multivariate SPC charts for monitoring batch processes. *Technometrics*, 37(1), 41–59.
- Nomikos, P., & MacGregor, J. (1995b). Multiway partial least squares in monitoring batch processes. *Chemom Intell Lab Syst*, 30, 97–108.
- Osten, D. (1988). Selection of optimal regression models via cross-validation. *J Chemom*, 2, 39–48.
- Phatak, A., Reilly, P., & Penlidis, A. (1993). An approach to interval estimation in partial least squares regression. *Anal Chim Acta*, 277, 495–501.
- Reis, M., & Saraiva, P. (2012). Prediction of profiles in the process industries. *Ind Eng Chem Res*, 51, 4254–4266.
- Serneels, S., Faber, K., Verdonck, T., & Van Espen, P. (2011). Case specific prediction intervals for tri-PLS1: The full local linearisation. *Chemom Intell Lab Syst*, 108, 93–99.

- Ündey, C., Ertunç, S., & Çinar, A. (2003). Online batch/fed-batch process performance monitoring, quality prediction, and variable-contribution analysis for diagnosis. *Ind Eng Chem Res*, 42, 4645–4658.
- Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. (2003). A review of process fault detection and diagnosis. Part I: Quantitative model-based methods. *Comput Chem Eng*, 27, 293–311.
- Wold, S., Geladi, P., Ebensen, K., & Öhman, J. (1987). Multi-way principal components- and PLS-analysis. *J Chemom*, 1(1), 41–56.