The Influence of Input and Output Measurement Noise on Batch-End Quality Prediction with Partial Least Squares

Jef Vanlaer, Pieter Van den Kerkhof, Geert Gins, and Jan F.M. Van Impe

BioTeC, Department of Chemical Engineering, KU Leuven
W. de Croylaan 46 PB 2423, B-3001 Heverlee (Leuven), Belgium
{jef.vanlaer,pieter.vandenkerkhof,geert.gins,jan.vanimpe}@cit.kuleuven.be

Abstract. In this paper, the influence of measurement noise on batch-end quality prediction by Partial Least Squares (PLS) is discussed. Realistic computer-generated data of an industrial process for penicillin production are used to investigate the influence of both input and output noise on model input and model order selection, and online and offline prediction of the final penicillin concentration. Techniques based on PLS show a large potential in assisting human operators in their decisions, especially for batch processes where close monitoring is required to achieve satisfactory product quality. However, many (bio)chemical companies are still reluctant to implement these monitoring techniques since, among other things, little is known about the influence of measurement noise characteristics on their performance. The results of this study indicate that PLS predictions are only slightly worsened by the presence of measurement noise. Moreover, for the considered case study, model predictions are better than offline quality measurements.

Keywords: Partial Least Squares, batch-end quality prediction, measurement noise statistics.

1 Introduction

The development of automated monitoring systems to assist human process operators in their decisions is an important challenge for the chemical and life sciences industry [13]. Chemical and biochemical production processes and plants are equipped with numerous sensors that measure various flow rates, temperatures, pressures, pH, concentrations, ... Despite the frequent use of sensor measurements for automated low-level control (e.g., PID control for valve opening and closing), most information in these measurements remains unexploited as responding to abnormal events –one of the most important control tasks– most often remains a manual operation. Human operators investigate the information arising from sensors in the process and compare this information to measurements from previous process runs to detect a departure from normal operation. However, the size and complexity of modern interconnected process plants (e.g., the very high number of sensors) largely complicate this task.

P. Perner (Ed.): ICDM 2012, LNAI 7377, pp. 121–135, 2012.

A lot of research effort in the area of data-driven process monitoring has been directed towards fault detection using techniques based on *Principal Components Analysis* (PCA [3,8,11]). These techniques exploit the information in historical databases to detect deviations from nominal process behavior during a new process run. Techniques based on *Partial Least Squares* (PLS [5]) take process output (quality) measurements into account, which makes them suited not only for detection of process faults, but also for estimation of quality variables that are not measured online. Examples include the final quality of a batch process.

Batch processes are commonly used for the manufacture of products with a high added value (e.g., medicines, enzymes, high-performance polymers). Since the loss of a batch due to process faults is very costly, close monitoring of these processes is of utmost importance. Batch runs that deviate from normal process behavior should be detected as soon as possible so that corrective actions can be taken. However, due to their dynamic nature and the unavailability of final batch quality measurements while the process is running (e.g., batch-end product purity or concentration), monitoring and control of batch processes to achieve a satisfactory product quality is even more complicated. The use of multivariate PLS models to obtain batch-end quality predictions (e.g., [4,10,12]) offers a solution to this problem.

PLS has been developed to deal with large datasets of correlated measurements and to filter noise from these measurements. However, noise present on both online sensor measurements and offline quality measurements will never be removed completely and will hence negatively influence the predictive performance of the PLS models. In addition, the presence of measurement noise in the data has an influence on the selection of model inputs and the optimal model order. As these effects cause many industrial companies to be reluctant in implementing PLS techniques, this work aims at investigating the influence of input and output measurement noise characteristics, more specifically the standard deviation of Gaussian distributed noise, on PLS-based batch-end quality prediction. As a case study, an extensive dataset from a computer simulator for industrial penicillin production [2] is selected.

The paper is structured as follows. Section 2 provides a brief explanation of *Multiway Partial Least Squares* modelling. Next, Section 3 explains how this technique is implemented for online batch-end quality prediction. In Section 4, the techniques for model order and input variable selection are discussed, after which Section 5 presents the selected case study. The results are shown and discussed in Section 6 and final conclusions are drawn in Section 7.

2 Multiway Partial Least Squares Modelling

To predict the final quality of a batch process, a Multiway Partial Least Squares (MPLS [9]) model is trained on historical data of normal process operation.

The modelling consists of two steps. In a first step, the data matrix containing the sensor measurements, which has a three-dimensional structure, is unfolded to a two-dimensional matrix (Section 2.1). A general Partial Least Squares (PLS [5]) model is constructed based on this two-dimensional data matrix in the second step, as explained in Section 2.2.

2.1 Data Matrix Unfolding

When for I batches, measurements of J different variables are available over K time points, a three-dimensional data matrix $\underline{\mathbf{X}}$ of size $I \times J \times K$ is obtained. To deal with this specific three-dimensional structure, the dimensionality of the matrix $\underline{\mathbf{X}}$ is reduced by means of *batch-wise data matrix unfolding* [8,10]. The matrix $\underline{\mathbf{X}}$ is divided in K slices of size $I \times J$ and these slices are placed side by side. This way, an unfolded data matrix $\mathbf{X}$ of size $I \times JK$ is obtained. The technique preserves the batch direction: every row of the unfolded matrix corresponds to one complete batch. Figure 1 illustrates the procedure.

Other techniques for data matrix unfolding are available (e.g., variable-wise unfolding [14]). However, since batch-end quality is related to the complete batch history, batch-wise unfolding is used for prediction of the final product quality.

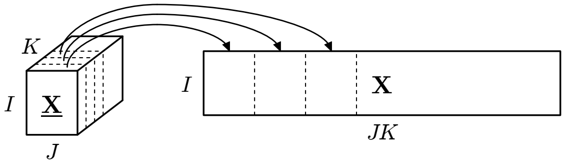

Fig. 1. Illustration of batch-wise data matrix unfolding.

2.2 Multiway Partial Least Squares (MPLS)

After data matrix unfolding, a regression model is constructed between the unfolded (input) data matrix $\mathbf{X}$ and the (output) matrix $\mathbf{Y}$ ($I \times L$), which contains L quality measurements for each batch in its columns, using standard two-dimensional Partial Least Squares (PLS [5]).

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

In the PLS procedure, the input and output matrices are projected onto a lower-dimensional space, each dimension of which is defined by one of the R principal components or latent variables. These principal components are computed as linear combinations of the original measurements in such a way that they contain as much information (covariance) about the original input and output measurements as possible. The projections of $\mathbf{X}$ and $\mathbf{Y}$ are defined by the loading matrices $\mathbf{P}$ ($JK \times R$) and $\mathbf{Q}$ ($L \times R$) respectively. The scores matrix $\mathbf{T}$ ($I \times R$)

represents the data matrices in the reduced space. The matrices $\mathbf{E_X}$ and $\mathbf{E_Y}$ contain the residuals or modelling errors.

The matrix $\mathbf{P}$ is not invertible and its columns are not orthonormal. Therefore, a $JK \times R$ weight matrix $\mathbf{W}$ with orthonormal columns is introduced to calculate the scores matrix $\mathbf{T}$ and quality prediction $\mathbf{Y}$ for a given measurement set $\mathbf{X}$. $\mathbf{P}^T\mathbf{W}$ is invertible so that the projection of the inputs $\mathbf{X}$ on the scores space $\mathbf{T}$ and the corresponding regression matrix $\mathbf{B}$ $(JK \times R)$ are computed as follows:

$$\mathbf{T} = \mathbf{XB} \tag{2}$$

$$\mathbf{B} \overset{\triangle}{=} \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1} . \tag{3}$$

The relation between the quality variables $\mathbf{Y}$ and the input measurements $\mathbf{X}$ then becomes

$$\mathbf{Y} = \mathbf{TQ}^T = \mathbf{XBQ}^T . \tag{4}$$

3 Online Batch-End Quality Prediction

During a new batch process run, only the measurements up until the current time k are known. Missing data techniques are used to compensate for the unknown future measurements. Several techniques were investigated in [4]. The best performance was obtained with *Trimmed Scores Regression* (TSR [1]).

A major advantage of TSR is that it only requires a single PLS model to predict the batch-end quality online at every sample instance throughout the batch instead of K different models. Moreover, previous research by the authors has shown that it exhibits similar performance to the training of a new PLS model for every instance at which a prediction is asked, both for noiseless data and in industrial practice [6].

In TSR, the known part of the data matrix $\mathbf{X}_{\text{new}}$ for a new batch (the first k columns of this data matrix, referred to as $\mathbf{X}_{\text{new},k}$) is multiplied with a matrix $\mathbf{B}_k$, consisting of the first k rows of the PLS regression matrix $\mathbf{B}$, to obtain the trimmed scores $\mathbf{T}^*_{\text{new},k}$.

$$\mathbf{T}^*_{\text{new},k} = \mathbf{X}_{\text{new},k}\mathbf{B}_k \tag{5}$$

Subsequently, a regression model is used to estimate the final scores $\mathbf{T}_{\text{new},k}$ of the new batch based on these trimmed scores. The time-varying regression matrix $\mathbf{A}_k$ that links the estimated final scores to the trimmed scores is computed by means of a least-squares regression on the training data, for which both the complete scores $\mathbf{T}_{\text{train}}$ and the trimmed scores $\mathbf{T}^*_{\text{train},k}$ are known.

$$\begin{gathered}\mathbf{T}_{\text{train}} = \mathbf{T}^*_{\text{train},k}\mathbf{A}_k + \mathbf{E_T} \\ \Downarrow \\ \mathbf{A}_k = \left(\mathbf{T}^{*T}_{\text{train},k}\mathbf{T}^*_{\text{train},k}\right)^{-1}\mathbf{T}^{*T}_{\text{train},k}\mathbf{T}_{\text{train}}\end{gathered} \tag{6}$$

The final scores of a new batch can be estimated from the trimmed scores using this regression matrix as follows.

$$\hat{\mathbf{T}}_{\text{new},k} = \mathbf{T}^*_{\text{new},k}\mathbf{A}_k \tag{7}$$

Substituting Equation (6) into Equation (7) and exploiting the PLS relations from Equations (2) and (4), the online estimation of the batch-end quality is obtained.

$$\mathbf{Y}^{\text{TSR}}_{\text{new},k} = \mathbf{X}_{\text{new},k}\mathbf{B}_k\left(\mathbf{B}_k^T\mathbf{X}^T_{\text{tr},k}\mathbf{X}_{\text{tr},k}\mathbf{B}_k\right)^{-1}\mathbf{B}_k^T\mathbf{X}^T_{\text{tr},k}\mathbf{X}_{\text{tr}}\mathbf{B}\mathbf{Q}^T \tag{8}$$

4 Model Order and Input Variable Selection

The selection of the optimal number of principal components (i.e., the order of the PLS model) is important to obtain good predictions of the batch-end quality. Section 4.1 explains the procedure for model order selection. Moreover, a selection of the most relevant model inputs may also improve the prediction performance of the model since not all available measurements are necessarily correlated with the final batch quality. The procedure for selecting the most relevant model inputs is explained in Section 4.2.

4.1 Model Order Selection

A leave-one-out cross-validation procedure is employed to select the optimal model order R, which corresponds to the number of principal components of the PLS model. Each batch in the training dataset is left out once and MPLS models of different model orders are trained based on the other available batches. Next, the models are validated on the left out batch and the mean *Sum of Squared Errors* (*SSE*) over all batches in the training dataset is calculated for every model order. An *adjusted Wold's criterion* with a threshold of 0.9, as proposed in [7], is used to select the model order. Instead of taking the number of latent variables corresponding to the observed minimum in the *SSE*-curve, the number of principal components is determined as the smallest model order R for which the following equation holds.

$$\frac{SSE(R+1)}{SSE(R)} > 0.9 \tag{9}$$

$SSE(R)$ is the (crossvalidation) *SSE* of the MPLS model with model order R. According to the adjusted Wold's criterion, the $(R{+}1)^{\text{th}}$ component is only added if it significantly improves the prediction and thus decreases the crossvalidation error.

4.2 Input Variable Selection

Despite the capability of PLS models to deal with noisy data, model predictions can be improved by eliminating useless measurements that are not correlated with the final batch quality. The optimal input set is selected using a *bottom-up* *branch-and-bound* procedure, assuming that the optimal set of j input variables also contains the optimal set of $j - 1$ inputs.

When J (online) measurement variables are available, J single input models are trained in a first step. Each of these models uses one of the available variables as input. The measurement variable that yields the model with the lowest leave-one-out cross-validation *SSE* is selected as the most important input variable. In a second step, $J - 1$ combinations of two inputs are formed by combining the first selected variable with all remaining measurements. These combinations are then used for the training of $J - 1$ two-input PLS models. Once more, the input set that results in the lowest cross-validation *SSE* is selected. This optimal combination of two inputs is combined with the remaining variables in the next step and the procedure continues until a ranking of all available measurements from most to least important is obtained.

Finally, a comparison is made between the cross-validation *SSE* for all selected input combinations. With the addition of extra input variables, the *SSE* will initially decrease. At a certain number of inputs however, the *SSE* curve reaches a minimum after which it starts rising again. The number of model inputs that corresponds to this minimum *SSE* value, is selected as the optimal number of input variables.

5 Case Study

Due to the need for data from a lot of batch runs with many different levels of measurement noise, a simulated process is selected as a case study. A biochemical process for penicillin fermentation at industrial scale is simulated via an extended version of the `Pensim` simulator [2]. To represent (biochemical) process variability, the initial substrate concentration, biomass concentration, and culture volume are subject to random variations for each batch. The process inputs (e.g., the substrate feed rate) exhibit variations around their setpoints as well. The process consists of two phases. Initially, the bioreactor is operated in batch mode. Once the substrate concentration drops below 0.3 g/L, the fed-batch phase is started. During this phase, additional substrate is fed into the reactor. The process is terminated after the addition of 25 L of substrate. The penicillin concentration at the end of the batch is the batch-end quality variable for which an online estimation is needed.

A total of 200 batches is simulated to investigate the influence of input and output measurement noise on the prediction of the final penicillin concentration. 15 concentrations and flows, and the temperature and pH in the bioractor are available from the simulator during the fermentation. Only 11 of these measurements are generally acquired by online sensors and thus practically available as model inputs for online prediction of the batch-end penicillin concentration. To avoid problems with badly tuned PID controllers at higher noise levels, Gaussian noise is added to the measurements of these variables after simulation. Input noise at 20 different levels is considered, which will be denoted with respect to a reference noise level described in Table 1.

Table 1. Overview of available online measurements with their mean nominal values and the standard deviation of the reference noise level $\sigma_{\text{noise,ref}}$ for these measurements.

Variable	**Mean**	$\sigma_{\mathbf{noise,ref}}$	**Variable**	**Mean**	$\sigma_{\mathbf{noise,ref}}$
Time [h]	-	0	Aeration rate [L/h]	8.0	$1.667e^{-1}$
DO [mmol/L]	1.1	$1.333e^{-2}$	Agitator power [W]	30.0	$3.333e^{-1}$
Volume [L]	107.5	$6.667e^{-1}$	Feed temperature [K]	296.0	$3.333e^{-1}$
pH [-]	5.0	$3.333e^{-2}$	Water flow rate [L/h]	64.2	1.667
Reactor temp. [K]	298.0	$3.333e^{-1}$	Base flow rate [L/h]	$2.5e^{-5}$	$6.667e^{-6}$
Feed rate [L/h]	0.05	$1.667e^{-3}$	Acid flow rate [L/h]	$7.9e^{-6}$	$6.667e^{-7}$

After noise addition, the measured signals are aligned and resampled to a length of 602 samples via *indicator variables*, comparable to the procedure in [2]. To obtain a monotonically increasing variable for the alignment of the batch phase, a straight line is fitted through the noisy measurements of the bioreactor volume. The time signal is added to the input measurements as an extra (aligned) variable, so that 12 online measurement signals are available for every batch. Therefore, the size of the training data matrix $\underline{\mathbf{X}}$ is 200 × 12 × 602.

Output measurement noise is added to the value of the final penicillin concentration. Gaussian noise with a standard deviation of 1 to 10 percent of the mean batch-end penicillin concentration is considered. As such, measurements at 10 different levels of output noise are available.

MPLS models to predict the final penicillin concentration are constructed for all combinations of input and output noise. The optimal input variables and the model order are selected according to the procedures in Section 4 to improve the predictions. The leave-one-out cross-validation Root Mean Squared Error (*RMSE*) is calculated both offline (i.e., after conclusion of the batch operation) and online to compare the predictions at different noise levels. To assess the influence of measurement noise on quality predictions without the influence of different model inputs, the prediction performance of models that use all 12 available measurements as inputs is also compared for different input and output noise levels. All calculations are performed thrice with different noise values sampled from the respective Gaussian distributions.

6 Results and Discussion

The next sections present the results of the study. The discussion of the influence of input noise and output measurement noise on batch-end quality prediction are decoupled in Sections 6.1 and 6.2 respectively. In each part, the influence of the noise on input variable and model order selection, offline quality prediction and online quality prediction is discussed.

6.1 Input Measurement Noise

Input Variable and Model Order Selection

After the addition of input measurement noise and alignment of the data, the optimal set of input variables and the model order are selected for every input noise level according to the procedure in Section 4.

In the noiseless case, 6 inputs (Dissolved Oxygen (*DO*), feed rate, time, pH, reactor temperature, and water flow rate) are selected. Several of these variables (e.g., pH and temperature) are PID controlled and vary only slightly. When even low amounts of measurement noise are added to the data, these measurements are rendered uninformative and a lower number of inputs is selected. Up to a noise level of $1/8^{\text{th}}$ of the reference level, the selected number of inputs is mostly 3. *DO*, feed rate, and time remain the most important input variables.

At a noise level of $1/8^{\text{th}}$ of the reference level, the noise has reached the size of the normal variation of the *DO* measurements. 6 inputs are again selected in an attempt to filter out the noise by exploiting the variable correlation.

At higher noise levels, *DO* measurements become uninformative due to the noise. The reactor volume is then selected as the most important variable. The number of inputs varies between 2 and 5.

The model order shows a decreasing trend with increasing input noise level, ranging from 9 for the noiseless case to 1-2 for the highest tested noise level (6 times the reference level). Ideally, the model order is a measure for the number of independent underlying phenomena that determine the course of the process. When more noise is added to the data, more and more of these phenomena are masked and fewer latent variables are selected.

Offline Quality Prediction

Offline prediction of the batch-end quality is the estimation of –in this case– the final penicillin concentration at the end of the batch operation. When the batch operation has finished, the complete data matrix $\underline{\mathbf{X}}$ is known, so no compensation for missing variables is needed. Figure 2 shows the average offline prediction *RMSE* as a function of the input noise level when no noise is present in the output measurements for both MPLS models with optimal inputs (*full curve*) and models which employ all 12 available online measurements as input variables (*dashed curve*). The selection of optimal input variables leads to better offline estimations of the batch-end penicillin concentration, evidenced by the lower *RMSE* values. However, both curves exhibit the same trend. As expected, the *RMSE* increases (so the prediction performance decreases) with increasing input noise level. The increase is most obvious at low noise levels, while at higher noise levels, the increase is less pronounced and the *RMSE* saturates. Even at high input noise levels, the *RMSE* is still relatively small and very good quality predictions are obtained.

This is also concluded from Figure 3, which shows a plot of the offline leave-one-out cross-validation prediction against the real final penicillin concentration for both the noiseless case (Figure 3(a)) and an input noise level of 6 times the reference level (Figure 3(b)). Without measurement noise, a nearly perfect

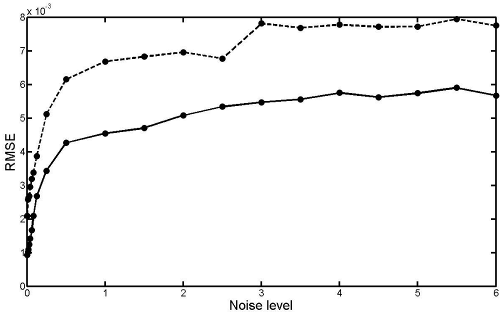

Fig. 2. Leave-one-out cross-validation *RMSE* for offline prediction of the final penicillin concentration in function of the input noise level: with selection of model inputs (—) and with all available inputs (- -). No output measurement noise is present in the data.

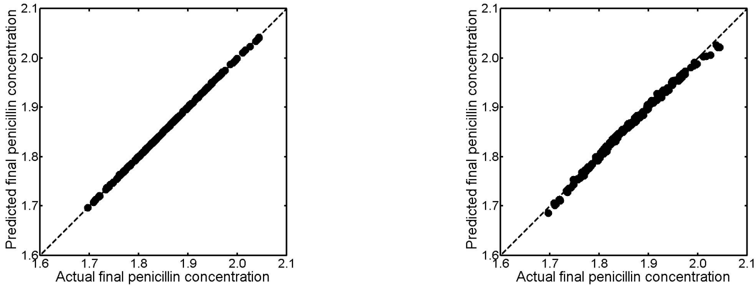

Fig. 3. Optimized offline prediction of the final penicillin concentration versus real penicillin concentration for the noiseless case (*left*) and an input noise level of 6 times the reference level (*right*). No output measurement noise is present in the data.

prediction is obtained (also evidenced by an *RMSE* of 0.001). However, even for the highest noise level under study –much higher than the noise encountered in industry–, the estimated quality approaches the real quality very well. Hence, a very efficient removal of the input noise from the data is achieved.

Online Quality Prediction

Using Trimmed Scores Regression (TSR) to compensate for missing future measurements, online predictions of the final penicillin concentration are obtained as explained in Section 3. The evolution in time of the maximal relative deviation of the online prediction from the real final penicillin concentration is depicted in Figure 4, both for the noiseless case and for noise of 6 times the reference level.

Initially, the predicted penicillin concentration deviates considerably from the real final value for both cases, since very few measurements are available. For the noiseless case, the deviation quickly drops below 1% as the batch progresses and by the end of the process run, the prediction has evolved towards the correct value. For noise of 6 times the reference level, the relative deviation decreases more slowly. Nonetheless, a stable prediction that deviates less than 1% from the real batch-end quality is obtained in about 200 samples.

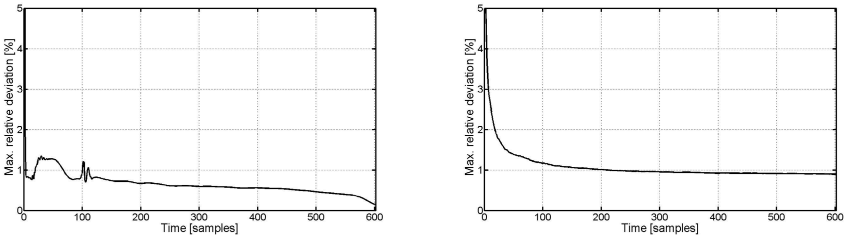

Fig. 4. Maximal relative deviation of the online prediction from the real final penicillin concentration in function of time for the noiseless case (*left*) and input measurement noise of 6 times the reference level (*right*). No output noise is present in the data.

An overview of the maximal relative prediction deviation for different input noise levels and sample times is given in Table 2(a). At first sight, the results are a little unexpected: adding noise improves the online prediction in some situations. This is especially visible during the batch phase (the first 101 samples of the process), where better predictions are obtained with noise of the reference level than at a noise level which is 16 times smaller. The selection of input variables, which aims at improving the offline batch-end quality prediction, does not necessarily guarantee optimal online predictions. Especially when the process consists of different phases, selecting one set of input variables for the complete process may result in a decrease in online prediction performance during certain phases.

As discussed earlier, different model inputs are selected for low and high noise levels. Apparently, the selected inputs for the low noise levels do not contain enough information to obtain good online predictions during the batch phase. This is evidenced by the fact that models that employ all available online measurements as inputs result in better online predictions during the batch phase for lower noise levels, as shown in Table 2(b). Consequently, it is better to use all available model inputs to obtain good online predictions from the start of the process. Another option is the training of different models (with different inputs) for all process phases.

From the results in Table 2, it can be concluded that higher input noise levels result in slightly worse prediction performance. However, the predictions improve with time and good and stable predictions are obtained in fewer than 200 samples for all noise levels.

Table 2. Influence of input measurement noise level on the maximal relative deviation of the online prediction from the real final penicillin concentration for models with (a) selected inputs and (b) all available inputs. No output noise is present in the data.

Time	No noise	Level 1/16	Level 1	Level 3	Level 6
1	12.4%	12.5%	1.7%	4.3%	7.1%
50	1.3%	10.8%	1.0%	1.1%	1.2%
100	0.9%	7.9%	1.0%	1.0%	1.2%
200	0.7%	1.0%	0.9%	1.0%	1.0%
300	0.6%	0.7%	0.8%	0.9%	1.0%
400	0.6%	0.6%	0.8%	0.9%	0.9%
500	0.5%	0.5%	0.7%	0.9%	0.9%
602	0.1%	0.3%	0.7%	0.9%	0.9%

(a)

Time	No noise	Level 1/16	Level 1	Level 3	Level 6
1	0.9%	2.5%	3.1%	5.0%	7.5%
50	0.9%	1.1%	1.1%	1.3%	1.6%
100	1.1%	1.1%	1.2%	1.3%	1.4%
200	0.8%	0.9%	1.2%	1.6%	1.3%
300	0.7%	0.8%	1.1%	1.5%	1.3%
400	0.6%	0.7%	1.1%	1.4%	1.2%
500	0.5%	0.6%	1.1%	1.3%	1.2%
602	0.3%	0.5%	1.0%	1.2%	1.2%

(b)

6.2 Output Measurement Noise

Input Variable and Model Order Selection

When output noise is added to measurements of the final penicillin concentration, the number of selected inputs varies greatly for different combinations of input and output noise levels. However, dissolved oxygen concentration (only at low input noise levels) or reactor volume always remain the most important variables. No real conclusions can be drawn about the importance of the other available measurements since various combinations of variables are selected at different combinations of input and output noise levels.

The optimal model order decreases quickly with the size of the output measurement noise. For output noise with a standard deviation of 1 percent of the mean final penicillin concentration 1 to 3 latent variables are selected at input noise levels smaller than the reference level. At higher input noise levels a model order of 1 is selected. For output noise with a standard deviation of 2 to 10 percent of the mean batch-end quality measurement a model order of 1 is selected in most cases.

Offline Quality Prediction

The full curve in Figure 5 gives an overview of the leave-one-out cross-validation *RMSE* for offline prediction of the batch-end penicillin concentration in function of the output noise standard deviation at the reference input noise level when optimal input variables are selected. The course of the curve is very similar for other input noise levels and for models that employ all available input variables. The *RMSE* increases linearly with the size of the output noise and its value is approximately equal to the standard deviation of the output measurement noise. Thus, it seems that the size of the output noise has a very big influence on the prediction performance of the PLS models. However, the *RMSE* was calculated by comparing the model predictions to the final penicillin concentration measurements, which contain noise. By comparing the predictions to the real (noiseless) value of the batch-end quality an actual *RMSE* value is obtained. The dashed curve in Figure 5 shows the course of this actual *RMSE* in function of the output noise standard deviation. From this curve it becomes clear that the influence of the output noise on the offline prediction is actually very small. Unlike the measurement *RMSE*, the actual *RMSE* increases only slightly with increasing output noise size and even at an output noise standard deviation of 10 percent, the size of the actual *RMSE* is around 1 percent of the value of the final penicillin concentration.

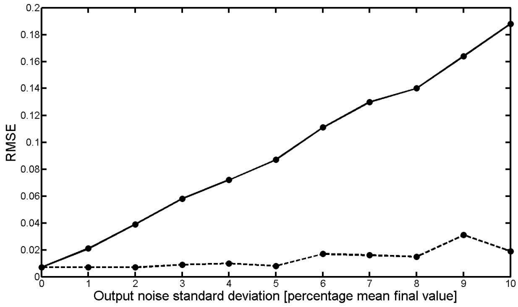

Fig. 5. Measurement (—) and actual (- -) leave-one-out cross-validation *RMSE* for offline prediction of the final penicillin concentration in function of the output noise standard deviation at the reference input noise level with model input selection

A graphical representation of this result is given in Figure 6. The graph on the left shows a plot of the measured final penicillin concentration, which contains noise with a standard deviation of 5% of the mean concentration, versus the actual (noiseless) batch-end penicillin concentration. In the middle graph, the model prediction is plotted against the measured penicillin concentration. Correlation between these variables is small and the size of the deviation of the prediction from the measurements is equal to the size of the measurement noise in the left graph. However, when the predicted penicillin concentration is plotted against the actual value in the graph on the right, a high correlation is obtained.

Of course, perfect (noiseless) quality measurements are never available in industry. However, as illustrated in this case study, PLS model predictions may

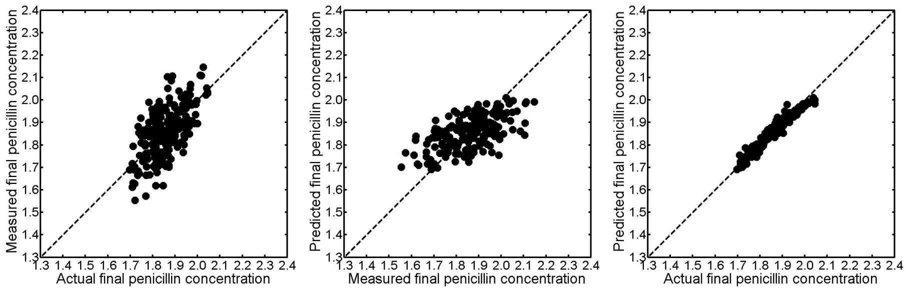

Fig. 6. Results of the offline prediction of the final penicillin concentration with optimized input variables with input noise of the reference level and Gaussian output noise with a standard deviation of 5% of the mean final penicillin concentration: measured vs. actual penicillin concentration (*left*), predicted vs. measured penicillin concentration (*middle*) and predicted vs. actual penicillin concentration (*right*).

be better than offline quality measurements, even when these noisy measurements are used to train the models. It is important to be aware of the size of the noise on the quality measurements, since even perfect predictions result in a measurement *RMSE* of approximately the same size as the standard deviation of the measurement noise σ_{noise}. This is corroborated by the formulas of both variables:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}{N}} \tag{10}$$

$$\sigma_{\text{noise}} = \sqrt{\frac{\sum_{i=1}^{N}(y_{i,\text{real}} - y_i)^2}{N-1}} \tag{11}$$

with $\hat{y}_i$ the model prediction, y_i the measured quality and $y_{i,\text{real}}$ the real (noiseless) quality of batch i, and N the number of training batches. In case of a nearly perfect prediction ($\hat{y}_i \approx y_{i,\text{real}}$) and for a sufficiently high number of training batches, these formulas are approximately the same.

In this case, it is valuable to temporarily invest in some extra quality measurements with higher accuracy to check the prediction performance of the model.

Online Quality Prediction

As for the offline quality prediction, output measurement noise has very little influence on online batch-end quality prediction when the deviation of the prediction from the actual (noiseless) final penicillin concentration is considered. Online predictions are slightly worse than in the case where no output noise is present in the data, but good and stable predictions are still obtained within an acceptable time span.

7 Conclusions

In this paper, the influence of input and output measurement noise characteristics on PLS-based batch-end quality prediction is investigated. As a case study, realistic computer-generated data of a fed-batch process for penicillin production are used. Gaussian noise of different levels (i.e., different size of the noise standard deviation) is added to the process input and output measurements. The effect of the noise level on input and model order selection, and offline and online prediction performance is studied.

The information content of measurements decreases with increasing noise level. While measurements of controlled variables, which vary only slightly, may be informative in the noiseless case, they are soon rendered uninformative when noise is added to the data. When the size of the input noise approaches the normal variation of informative measurements, PLS is no longer able to filter out the noise and a new set of optimal input variables is selected. Since higher noise values mask more and more important underlying phenomena, the model order decreases with both the input and output noise level.

The offline prediction performance of the PLS models decreases only slightly with increasing noise levels. Even for noise levels much higher than those encountered in industry, very good offline quality predictions are obtained. This proves the ability of PLS models to filter the noise from the data. Since no perfect (noiseless) quality measurements are available in industry, it is important to be aware of the size of the measurement noise. As illustrated in the case study, model predictions may be better than the measurements since even perfect predictions result in a measurement *RMSE* of approximately the same size as the standard deviation of the noise on the quality measurement.

When the selection of different model inputs at different noise levels is not taken into account, online predictions of the batch-end quality using Trimmed Scores Regression (TSR) deteriorate slightly with increasing levels of both input and output measurement noise. Despite the slightly lower prediction performance at higher noise levels, accurate and stable online predictions are obtained, even at noise levels much higher than in industrial practice. Future research will investigate the generalization of the obtained results.

Acknowledgements. Work supported in part by Project PVF/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 and P. Van den Kerkhof are 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

1. Arteaga, F., Ferrer, A.: Dealing with Missing Data in MSPC: Several Methods, Different Interpretations, Some Examples. J. Chemometr. 16, 408–418 (2002)
2. Birol, G., Ündey, C., Çinar, A.: A Modular Simulation Package for Fed-Batch Fermentation: Penicillin Production. Comput. Chem. Eng. 26, 1553–1565 (2002)
3. Eriksson, L., Johansson, E., Kettaneh, N., Wold, S.: Multi- and Megavariate Data Analysis: Principles and Applications. Umetrics Academy (2002)
4. García-Munoz, S., Kourti, T., MacGregor, J.: Model Predictive Monitoring for Batch Processes. Ind. Eng. Chem. Res. 43, 5929–5941 (2004)
5. Geladi, P., Kowalski, B.: Partial Least-Squares Regression: a Tutorial. Anal. Chim. Acta 185, 1–17 (1986)
6. Gins, G., Vanlaer, J., Van Impe, J.: Online Batch-End Quality Estimation: Does Laziness Pay Off? In: Quevedo, J., Escobet, T., Puig, V. (eds.) Proceedings of the 7th IFAC International Symposium on Fault Detection, Supervision and Safety of Technical Processes (Safe Process 2009), pp. 1246–1251 (2009)
7. Li, B., Morris, J., Martin, E.: Model Selection for Partial Least Squares Regression. Chemometr. Intell. Lab. 64, 79–89 (2002)
8. Nomikos, P., MacGregor, J.: Monitoring Batch Processes Using Multiway Principal Component Analysis. AIChE J. 40(8), 1361–1375 (1994)
9. Nomikos, P., MacGregor, J.: Multiway Partial Least Squares in Monitoring Batch Processes. Chemometr. Intell. Lab. 30, 97–108 (1995)
10. Nomikos, P., MacGregor, J.: Multivariate SPC Charts for Monitoring Batch Processes. Technometrics 37(1), 41–59 (1995)
11. Simoglou, A., Georgieva, P., Martin, E., Morris, A., de Azevedo, S.: On-line Monitoring of a Sugar Crystallization Process. Comput. Chem. Eng. 29(6), 1411–1422 (2005)
12. Ündey, C., Ertunç, S., Çinar, A.: Online Batch/Fed-Batch Process Performance Monitoring, Quality Prediction, and Variable-Contribution Analysis for Diagnosis. Ind. Eng. Chem. Res. 42, 4645–4658 (2003)
13. Venkatasubramanian, V., Rengaswamy, R., Yin, K., Kavuri, S.: A Review of Process Fault Detection and Diagnosis. Part I: Quantitative Model-Based Methods. Comput. Chem. Eng. 27, 293–311 (2003)
14. Wold, S., Geladi, P., Ebensen, K., Öhman, J.: Multi-way Principal Components- and PLS-Analysis. J. Chemometr. 1(1), 41–56 (1987)