# APPLICATION OF SOFTWARE SENSORS FOR MONITORING AND PREDICTION IN FERMENTATION PROCESSES

**Mads Thaysen** [*,**,1] **Sten Bay Jørgensen** [**]

[*] *BioProcess Laboratories, Novo Nordisk A/S, Hagedornsvej 1, HAB 2.35, 2800 Gentofte, Denmark*
[**] *CAPEC, Department of Chemical Engineering, Technical University of Denmark*

Abstract: The development of modelbased process software sensors for monitoring of biomass concentration and product concentration in fed-batch and continuous yeast fermentations is presented, followed by a validation of the sensors using data from industrial fermentations. Alternatively, using multiway projection to latent structures (MPLS) algorithm, a model for prediction of one-step ahead and end point product concentrations is developed and demonstrated on industrial process data. The one-step ahead MPLS-predictor is compared to the modelbased product concentration software sensor. The comprison indicates a better performance by the MPLS-predictor. 

Keywords: Process software sensors, MPLS, prediction, industrial process data.

## 1. INTRODUCTION

To improve monitoring and control of industrial fermentation processes it is desirable to include interpreted information of dynamic responses of relevant biological and chemical species to changes in process conditions whenever possible. Fulfilling this desire is however not trivial, since measurements of relevant species often are difficult to conduct and often impossible to obtain at the desired rate. An alternative approach to the direct measurement of species is the development of process software sensors based on mathematical models correlating measurable variables to the desired variables. This work will develop two different types of software sensors, one based upon first principles engineering modelling and another based upon chemometrics.

[1] Partially supported by the Academy of Technical Sciences, Denmark

## 2. PROCESS SOFTWARE SENSORS

First principles engineering models (FPEM) can form the underlying foundation for software sensors. The models infer information of unmeasured entities by using available information from other measured entities. Different frameworks can be used for the model development. First software sensors using FPEMs will be developed and investigated for the prediction of biomass and peptide product concentration in a fermentation broth. Subsequently a chemometric model is used for developing a software sensor for product estimation. Finally the two types of product concentration sensors are compared.

### *2.1 FPEM based Sensor for Biomass Concentration*

Lei (2001) and others demonstrated that it was possible to use a component mass balance on the proton production or consumption rate in a high

performance laboratory setup to obtain a simple on-line estimation of the biomass concentration in batch, fed-batch and continuous fermentation of *Saccharomyces cerevisiae*.

A simplified illustration of the contributions to the proton balance in a bioreactor is shown in figure 1. A component mass balance for the proton concentration $[H^+]$ in the extracellular medium yields:

$$V\frac{d[H^+]}{dt} = F_s[H^+]_{s,in} - F_e[H^+]_{e,out} + F_{H^+,gen} - F_{NH_3} \quad (1)$$

where the dual role of $NH_3$ is i) to maintain a constant pH-level in the medium and ii) to act as the primary nitrogen source for biomass production.

The following assumptions are used for simplification of the mass balance expression:

- Constant pH-level in the bioreactor
- Negligible contribution to proton balance from pH-diff. between feed and medium pH

In the original work the pH of the feed was adjusted to the pH of the medium. In this work estimation of the amount of proton equivalents needed to compensate for this pH-difference indicated that less than 1% of the molar flow of $NH_3$ is needed to balance the pH-difference between feed and medium pH.

The simplified mass balance yields:

$$0 = F_{H^+,gen} - F_{NH_3} \quad (2)$$

The volumetric proton production rate can now be calculated as:

$$r_{H^+} = \frac{F_{H^+,gen}}{V} = \frac{F_{NH_3}}{V} \quad (3)$$

The following assumptions have been made concerning possible sources contributing to the proton production rate from cellular activities during aerobic growth on a complex medium:

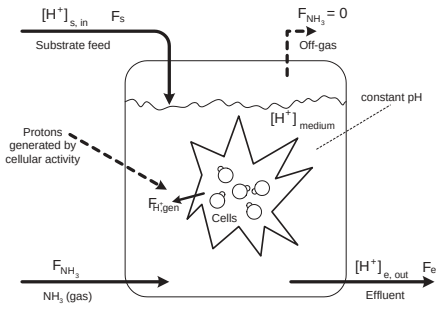



Fig. 1. Simplified schematic illustration of flows and factors that influences the extracellular proton concentration balance in the fermentation medium.

- Uptake of $NH_4^+$ as primary nitrogen source
- Negligible production or consumption of organic acids
- Negligible consumption of amino acids from complex medium
- No acidification of the medium due to production of $CO_2$

During aerobic growth on glucose only negligible amounts of organic acids are produced; $CO_2$ and biomass being the primary carbon-products formed. Contribution to the proton balance by the solution and dissociation of $H_2CO_3$ to carbonate can be disregarded when the pH-level is significantly below pH 7. By further assuming only negligible consumption of organic N-sources from the complex medium, the uptake of $NH_4^+$ is the only contributor to the proton production rate and the only significant nitrogen source. A 1:1 ratio between proton production rate and the $NH_4^+$ uptake rate (using has been observed indicating that the biomass production rate is proportional to the proton production rate, under the assumption that the nitrogen content of the biomass to be constant during balanced growth.

Based on the above comments and assumptions the volumetric biomass production rate, $r_x^{H^+}$, can be calculated from volumetric $NH_3$ addition rate:

$$r_x^{H^+} = \frac{M_{DW} \cdot r_{H^+}}{Y_{xH}} = \frac{M_{DW} \cdot F_{NH_3}}{Y_{xH} \cdot V} \quad (4)$$

with $M_{DW}$ as the molar weight of dry weight biomass and $Y_{xH}$ is the yield coefficient of mole protons produced per mole biomass i.e. the molar content of nitrogen in biomass based on the overall growth stoichiometry:

$$CH_xO_yN_z + aCO_2 + Y_{xH}H^+ - bCH_2O - cNH_4^+ - dO_2 = 0 \quad (5)$$

From the stoichiometric equation it can be seen that $Y_{xH}$ is constant, since $NH_4^+$ is the only proton source and $z = c\ (= Y_{xH})$ since $NH_4^+$ is the only nitrogen source. Combination of the above expressions with a dynamic mass balance for biomass $(x)$:

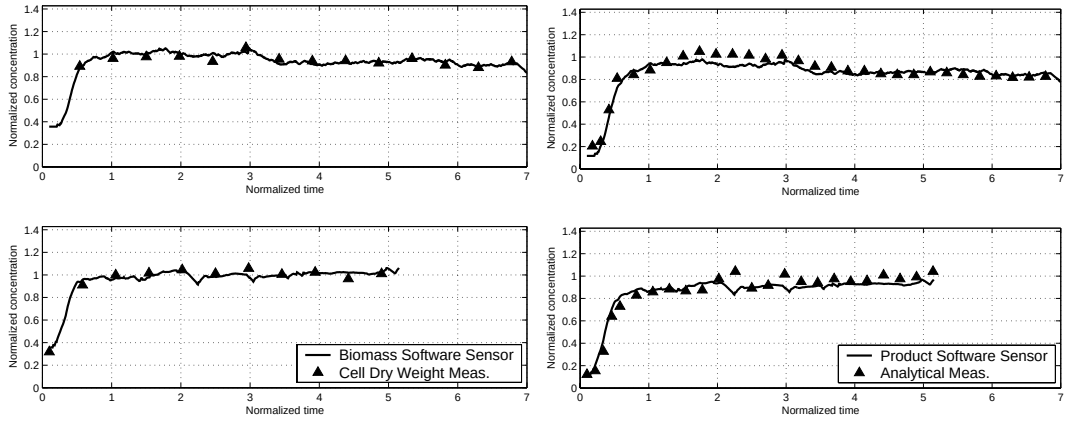
$$\frac{dx}{dt} = r_x - Dx \quad (6)$$

yields a simple biomass predictor:

$$x_{k+1} = x_k \cdot exp\left(\left(\frac{r_{x_k}^{H^+}}{x_k} - D_k\right)(t_{k+1} - t_k)\right) \quad (7)$$

where $D_k$ is the dilution rate at time point $t_k$.

The above model has been developed assuming ideal conditions in fermentor. Both for small and large scale fermentations with high cell densities this assumption is unlikely to be valid. To account for these variations the model has been modified as follows:




(a) Biomass concentration (b) Product concentration

Fig. 2. Comparison of signals from software sensors (−) and analytical measurements (▲) for fermentations in two different fermentors.

$$x_{k+1} = x_k \cdot exp\left( \left( \frac{\alpha \cdot F_{NH_3,k}}{V_k \cdot x_k} - D_k \right)(t_{k+1} - t_k) \right) \tag{8}$$

where $\alpha = f(t, M_{DW}, Y_{xH}, \text{vessel properties})$ is determined for the individual fermentor.

Two examples of the application in simulation of the biomass concentration software sensor are illustrated in figure 2(a).

### *2.2 FPEM based Sensor for Product Concentration*

To develop a process software sensor for prediction to the peptide product concentration, physiological knowledge of the recombinant yeast strain is used. It is known that the control of the promotor for transcription of the product gene is linked to the activity of the glycolysis of the recombinant strain. To simplify the model formulation the following assumptions are made:

- Production rate of product ($r_p$) proportional to production rate of biomass ($r_x$)
- High stability of recombinant gene
- No influence from transport and folding in organelles on production rate
- Effective excretion of product

A high stability of the recombinant gene ensures that no decay in specific productivity of the peptide product is experienced over time. Furthermore by assuming that the transport of the peptide product through the organelles of the cell does not seem to have any influence on the production rate, combined with effective folding and excretion of the peptide product to the abiotic phase, the rate limiting step of the cellular production proces becomes transcription of the recombinant gene.

Based on the above assumptions the following model for the production rate of the product ($p$) is proposed:

$$r_p \propto r_x = \frac{M_{DW} \cdot F_{NH_3}}{Y_{xH} \cdot V} \tag{9}$$

Introducing a parameter ($\beta$) accounting for the issues relating to non-ideal process conditions and variations in growth stoichiometry ($Y_{xH}$) and cell composition ($M_{DW}$) a dynamic mass balance on the product becomes:

$$\frac{dp}{dt} = r_p - Dp = \beta \frac{F_{NH_3}}{V} - Dp \tag{10}$$

leading to to the product predictor:

$$p_{k+1} = p_k \cdot exp\left( \left( \frac{\beta \cdot F_{NH_3,k}}{V_k \cdot p_k} - D_k \right)(t_{k+1} - t_k) \right) \tag{11}$$

where $\beta = g(t, M_{DW}, Y_{xH}, \text{vessel properties})$ is determined for the individual fermentor.

Two examples of the application in simulation of the product concentration software sensor are illustrated in figure 2(b) along with signals from biomass concentration software sensors from the same fermentations. The software sensors are activated after the batch fase and used for the fed-batch and continuous phases of the fermentation with constant $\beta$ values.

### *2.3 Multiway Projection to Latent Structures (MPLS)*

Process monitoring and prediction of end quality using MPLS have been illustrated by a number of research groups e.g. Nomikos and MacGregor (1995). The general idea behind MPLS is that an empirical model is build on measurements from reference batches operated under normal operating conditions producing a good quality product in terms of high concentration. This work has focused on the prediction possibilities of the MPLS. The available on-line measurements are used to estimate or predict product quality, which

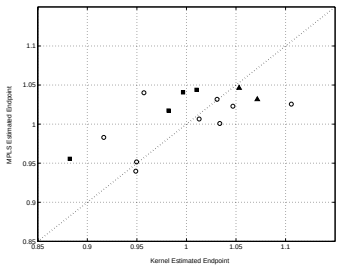



Fig. 3. Product concentrations at end of batch estimated using MPLS and kernel estimators. (○) M-, (▲) V- and (■) A-batches.

is desirable, since a limited number of analytical measurements of the quality variables is available in an off-line fashion. The on-line measurements are arranged in an array $\mathbf{X}$ and the quality measurements are arranged in another array $\mathbf{Y}$.

Using the MPLS-algorithm a regression equation can be formulated:

$$\hat{\mathbf{Y}} = \mathbf{XB}, \text{ with } \mathbf{B} = \mathbf{W}\left(\mathbf{P}^T\mathbf{W}\right)^{-1}\mathbf{Q}^T \quad (12)$$

This regression model $\mathbf{B}$ can then be used for on-line prediction of the end quality of the batch provided that a suitable method for the estimation of future on-line measurements is available (Nomikos and MacGregor, 1995). This work has been applying the method using the $J$ measurements obtained at the last sampling number $k$ to fill in the empty spaces.

The number of PLS-components ($C$) necessary to obtain a desired level of regression can be evaluated using different methods of (cross-)validation techniques. In this work the root mean square error of prediction ($RMSEP$) is used:

$$RMSEP = \sqrt{\frac{1}{K}\sum_{k=1}^{K}\left(\hat{y}_k - y_k\right)^2} \quad (13)$$

For increasing numbers of PLS-components $C$ used for model identification the $RMSEP$ is evaluated on validation data, where the lowest value of $RMSEP$ indicates the number of PLS-components $C$ to be used.

### *2.4 MPLS for On-line Prediction and Estimation*

In the case where quality measurements are taken frequently during the batch run, the MPLS-framework can be used for estimation and prediction of the intra-batch quality. For all the batches considered in this work, both on-line measurements and off-line quality measurements in each batch have been subsampled to the same frequency by applying a kernel estimator for smoothening using a tricubic kernel with a local linear fit of 3 nearest neighbors (Hastie *et al.*, 2001). With the smoothened data a $\mathbf{Y}$ array is obtained. At sample number $k$ in a new batch the full batch profiles of the quality variables $\hat{Y}_k$ can be obtained by filling in the empty spaces in $X_k$ as described above and applying the regression matrice $\mathbf{B}$:

$$\hat{Y}_k = X_k\mathbf{B} \quad (14)$$

Table 1. Exp. var. of $\mathbf{X}$ and $\mathbf{Y}$. Mean $RMSEP$ from the validation.

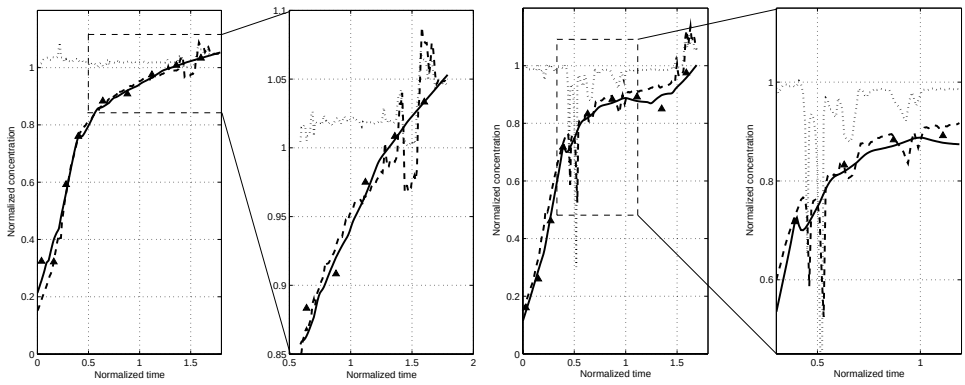
| | No | of | PLS | Comp | |
|---|---|---|---|---|---|
| Expl. var | 1 | 2 | 3 | 4 | 5 |
| $\mathbf{X}$ | 16.3 | 27.4 | 43.9 | 56.3 | 65.2 |
| $\mathbf{Y}$ | 56.9 | 73.1 | 80.7 | 86.5 | 94.4 |
| $RMSEP$ | **0.039** | 0.056 | 0.061 | 0.060 | 0.083 |

### *2.5 MPLS Applied on Industrial Data*

In this work the only quality variable to be regressed was the product concentration. 11 on-line measured variables were sampled 180 times during a fermentation, operating in fed-batch phase followed by a continuous phase. 9 batches conducted under normal operating conditions were used for the model identification (M-batches), while 2 validation batches (V-batches) were used to determine the number of PLS-components to be included in the model evaluated by the $RMSEP$ as describe above. The explained variance and $RMSEP$ for the 5 first PLS-components are shown in table 1. It is interesting to note that the $RMSEP$ evaluation indicates that only 1 PLS-component should be included in the model, explaining 55 % of the variation in $\mathbf{Y}$.

The model performance was then investigated using the 2 V-batches along with 4 additional batches (A-batches), the latters having normal end-point concentrations of the product, but undergoing small process upsets during operation. A comparison between the MPLS estimated and kernel estimated product concentration at the end of the batch is shown in figure 3. The latter of the two estimators is comparable to the analytical measurements. From the figure it is seen how the MPLS estimations at worst are within 10 % of the kernel estimated values for model, validation and A-batches. It is interesting to notice how the MPLS estimations of 4 A-batches all are larger than the kernel estimations.

Figure 4(a) shows the prediction results for a V-batch. A good description of the variations in the kernel estimation and the analytical measurements can be seen by the one- step ahead MPLS-prediction. From time 1.4 and to the end of the batch some variations in both the one-step ahead and end point prediction (starting at coordinates (0,1)) can be noticed. The variations are explained




(a) V-batch (b) A-batch

Fig. 4. MPLS product concentration predictions in a validation batch (a) and in batch with a process operation upset (b). (−) kernel estimation, (▲) analytical measurements, (−−) one-step ahead prediction and (··) batch end prediction.

by a temporary outfall (time 1.4-1.6) of the mass spectroscopy instrument measuring the contents in the off-gas from the fermentor. The effects of the disturbance are seen to have settled at the end of the batch.

The predictions of the MPLS-model in one of the A-batches where a small upset in the process operation occurs are illustrated in figure 4(b). The first upset occurs at time 0.4, where the fermentation is stopped because a fault has occured in the ammonia supply system. A number of actions occur as a consequence of this fault, resulting in large variations in the one-step ahead prediction and the end concentration prediction. A general decrease in the end point concentration is seen until the system is fully returned to normal operating conditions at time 0.7.

At time 0.9 a new disturbance appears, this time the substrate flow is stopped for a while. Both predictions decrease with this change, but are restored to normal after the substrate flow is reinitiated at time 1.0.

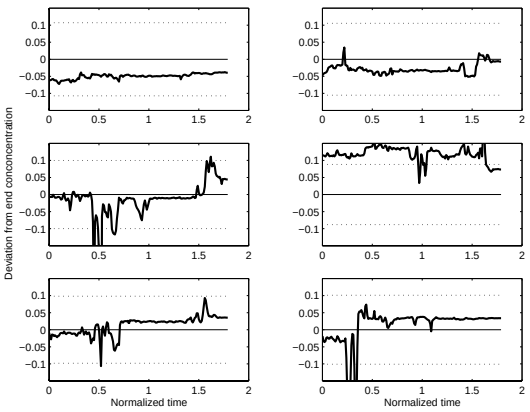



Fig. 5. Prediction error of end point product concentration in 2 V-batches (top) and 4 A-batches. (−) MPLS-prediction and (··) ± 10 % errors of the kernel est. end value.

The one-step ahead predictor is very close for all but one of the analytical measurements. Also the kernel estimated product concentration lies in general close to the one-step ahead predictor. however at the end of the batch both of the predictors are seen to sharply increase their predictions around time 1.6, the reason being a decrease in the dilution rate.

Figure 5 shows prediction error between the kernel estimated concentration and the analytical measurements and the one-step ahead prediction respectively. The prediction errors are illustrated for the 2 V-batches (top left and right) and the 4 A-batches, where the dotted lines represents the ± 10 % value of the kernel estimation. For 5 of 6 batches (not middle right) it is seen how the MPLS-predicted end concentrations are within ± 10 % of the final product concentration thus indicating that even with possible process upsets the predicted end point concentration was good.

## 3. PERFORMANCE COMPARISON OF FPEM- VERSUS MPLS-PREDICTOR

In the above two methods for one-step ahead prediction of the product concentration have been developed and tested. Figure 6 shows the prediction error between the kernel estimated concentration and the analytical measurements, the one-step ahead MPLS-predictor and the FPEM-predictor (product concentration software sensor) respectively. It can be seen that the MPLS-predictor to some degree is able to capture the values of the analytical measurements and the kernel estimations between the data points. However after approximately 20% of the batch time, the predictions are within 10% of the analytical measurements during normal operation. The MPLS- predictor performs better than the FPEM-predictor, which in general can be seen to have big positive errors in the first

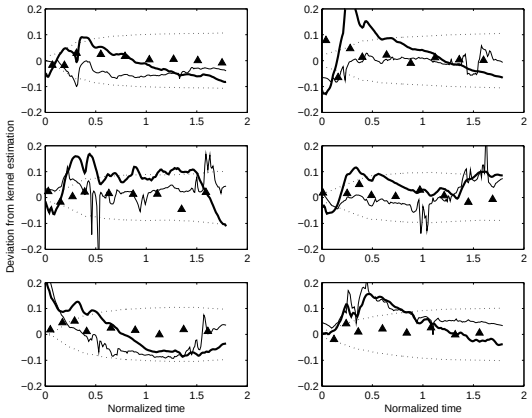



Fig. 6. Prediction errors between kernel est. values and one-step ahead product concentration predictors. MPLS-predictor (−) and FPEM-predictor (−, bold) in 2 V-batches (top) and 4 A-batches. (▲) analytical measurements and $(\cdot\cdot) \pm 10\%$ on the kernel est. values.

quater of the batch, approximately corresponding to the fed-batch phase. This is not desirable, since it is in the non-stationary phases that accurate and precise estimation and prediction is most important from an optimization perspective.

To support the comparison, $RMSEP$ values have been calculated in different phases of the fermentation process and shown in table 2. The FPEM-predictor can be seen to have relatively large errors in the first part of the batch corresponding to the fed-batch phase. While the $RMSEP$ of the MPLS-predictor also is the largest in the fed-batch phase, the predictor still performs well. In the stationary phase (phase 3) the results of the two predictors are approximately the same.

## 4. DISCUSSION AND CONCLUSION

In this paper two different methods for obtaining quantitative information from a fermentation process has been presented and preliminarily compared. The methods have been applied using on-line process data from an industrial fermentation process to illustrate the type and quality of information obtainable with the methods.

A software sensor was developed for monitoring of the biomass concentration and was based on FPEMs using the feed rate of ammonia, volume of broth and the dilution rate as inputs. Application of the software sensor using on-line process data gave a good description of the variations seen in the analytical measurements, leading to the conclusion that the implementation of this device will enable on-line monitoring of the biomass concentration.

Table 2. Mean $RMSEP$ of phases.

| Phase | 1 | 2 | 3 | Total |
|---|---|---|---|---|
| Time | 0.0 - 0.5 | 0.5 - 1.0 | 1.0 - 1.8 | 0.0 - 1.8 |
| FPEM | 0.039 | 0.027 | 0.022 | 0.030 |
| MPLS | 0.025 | 0.021 | 0.021 | 0.023 |

A similar software sensor was then developed for monitoring of the product concentration using the same framework as the biomass concentration software sensor. Although complex cellular processes are involved in the processes for generating the peptide a very simple model was developed by only slightly modifying the FPEM used for modelling the biomass concentration. Applied on the industrial data this simple software sensor was also able to give a good description of the general product concentration trajectory making on-line monitoring of the product concentration possible if implemented.

An alternative approach for monitoring component concentrations in a process is through process chemometrics. A model for predicting the product concentration based on the MPLS-algorithm was developed producing a linear model describing changes around an average trajectory. The model was tested on the industrial data and indicated that both one-step ahead and end point predictions of the product concentrations came within 5-10 % of the kernel estimated values based on analytical measurements.

The MPLS-predictor for the one-step ahead prediction was compared with the simple product concentration software sensor (FPEM-predictor), where the first gave a more accurate description of the variations in the product concentration.

In conclusion this work has provided insight into tools for monitoring a given fermentation process with respect to biomass concentration and product concentration. Areas to address in future work is the trade off between bias and variance in on-line estimators, the use of available analytical measurements for parameter adaption, development of a better description of the dynamics of product formation and extending the application of the MPLS-algorithm.

## REFERENCES

Hastie, Trevor, Robert Tibshirani and Jerome Friedman (2001). *The Elements of Statistical Learning. Data Mining, Inference and Prediction.*. Springer Verlag. New York.

Lei, F. (2001). Dynamics and non-linear phenomena in continuous cultivations of *Saccharomyces cerevisiae.*. PhD thesis. KT, Technical University of Denmark.

Nomikos, Paul and John F. MacGregor (1995). Multivariate SPC charts for monitoring batch processes. *Technometrics* **37**(1), 41–59.