# Soft Sensor Development in Fermentation Processes Using Recursive Gaussian Mixture Regression Based on Model Performance Assessment

Yuhan Ding, Yong Su, Congli Mei*

Jiangsu University, Zhenjiang, 212013, Jiangsu, China
clmei@ujs.edu.cn

The soft sensor modeling method based on moving window method (MW) and Gaussian mixture regression (GMR) has been developed in fermentation processes. However, the MW method always results in low computational efficiency because of frequency updating. GMR has the multi-model structure. In order to reduce calibration frequency of recursive GMR modeling methods, a recursive GMR soft sensor based on model performance assessment (MPA) is developed. According to the results of the model performance assessment, the model updating is selectively activated. The developed model was investigated to estimate biomass concentration in an industrial Erythromycin fermentation process. Compared with the GMR model, the prediction accuracy is improved obviously.

## 1. Introduction

Accurate online estimation and prediction of key product quality variables are vital for continuously characterizing the dynamic behavior of chemical processes. For the past decades, soft sensors have been widely used to handle these problems, which provide frequent estimations of key process variables through those that are easy to be measured online (Jin et al., 2015 and Liu et al., 2010 and Luttmann et al., 2012). The most popular soft sensor methods are partial least squares (PLS) (Sharmin et al., 2006 and Wang et al., 2015), artificial neural networks (ANN) (Cui et al., 2012 and Sun et al., 2014), and support vector machines (SVM) (Liu et al., 2010 and Kaneko et al., 2014). Usually, soft sensors are constructed based on process measurements easy to measure online. From the measurement point of view, the performance of these traditional soft sensors may decrease even fail due to the actual working conditions, the processing of the nonlinear characteristics and the influence of the external disturbance of the system, which restricts above-mentioned soft sensors in practical cases.

The recursive PLS algorithm (RPLS) is firstly proposed for the on-line calibration of model parameters. Subsequently, the RPLS algorithm is modified and extended, and it is applied to the modeling and quality prediction of chemical processes. The online calibration method of LS-SVM model (Li et al., 2009) is also applied to a chemical reaction process, hydro-isomerization process of C8-aromatics. Recently, a relative new machine learning method, i.e., Gaussian mixture regression (GMR), has been developed, and began to be applied in soft sensor modeling (Yuan et al., 2014). GMR updates the model usually by using moving window (MW) algorithm (Kadlec et al., 2011). However, the problem is that it does not have a policy to update the model simply by using data accumulation.

In this paper, MPA-GMR modeling method based on model performance evaluation is proposed to solve these problems. The method eliminates the influence of the old samples to the new samples by the sliding window method, selects the model based on the model performance evaluation and adapts the model's confidence limits along with the changes in process characteristics and model performance evaluation results.

## 2. Principle of GMR modeling based on model performance assessment

### 2.1 Principle of GMR modelling

Assume $\mathbf{X}$ represents the space of the explanatory variables and $\mathbf{Y}$ is the space of the response variables. $x$ is the input of training data ( $x \in \mathbf{X}$ ) and $y$ is the ideal output data ( $y \in \mathbf{Y}$ ). For the given $x$ and $y$, the joint probability density is given as (Sung, 2004)

$$f_{XY}(x, y) = \sum_{j=1}^{K} \pi_j \phi(x, y; \mu_j, \Sigma_j) \tag{1}$$

Where, subsequently, the mean $\mu_j$ and covariance $\Sigma_j$ can be divided into the input and output parts like the following

$$\mu_j = \begin{bmatrix} \mu_{jx} \\ \mu_{jy} \end{bmatrix}, \quad \Sigma_j = \begin{bmatrix} \Sigma_{jXX} & \Sigma_{jXY} \\ \Sigma_{jYX} & \Sigma_{jYY} \end{bmatrix}$$

Eq. (1) shows that the relationship between the explanatory variables and the prediction value can be described by Gaussian mixture models (GMMs). Where $\phi(x, y; \mu_j, \Sigma_j)$ denotes the probability density function of the multivariate GMM. The parameters of this model include the number of the mixture components $K$, the prior $\pi_j$, the mean value $\mu_j$, and the variance of each Gaussian component $\Sigma_j$, which are represented as

$$\theta = (\theta_1, \theta_2, \cdots, \theta_k) \text{ with } \theta_j = (\pi_j, \mu_j, \Sigma_j) \text{ and the constraint } \sum_{j=1}^{k} \pi_j = 1 \, (0 \le \pi_j \le 1).$$

Similarly, the marginal probability density can be given as (H. G. Sung, 2004)

$$f_X(x) = \int f_{X,Y}(x, y) dy = \sum_{j=1}^{K} \pi_j \phi(x; \mu_{jX}, \Sigma_{jX}) \tag{2}$$

The global GMR function can be deduced by combining Eq. (1) and (2)

$$f_{Y/X}(y/x) = \frac{f_{XY}(x, y)}{f_X(x)} = \sum_{j=1}^{K} w_j(x)\phi(y; m_j(x), \sigma_j^2) \tag{3}$$

with the mixing weight

$$w_j(x) = \frac{\pi_j \phi(x; \mu_{jX}, \Sigma_{jX})}{\sum_{j=1}^{K} \pi_j \phi(x; \mu_{jX}, \Sigma_{jX})} \tag{4}$$

The mean and the variance of the conditional distribution can be acquired in closed form by

$$m_j(x) = \mu_{jX} + \sum_{jYX} \Sigma_{jX}^{-1}(x - \mu_{jX}) \tag{5}$$

$$\sigma_j^2 = \sum_{jYY} - \sum_{jYX} \Sigma_{jX}^{-1} \sum_{jXY} \tag{6}$$

The prediction given a new input can be obtained by computing the expectation over the conditional distribution $f_{Y/X}(y/x)$ (Sung, 2004)

$$E\left[f_{Y/X}(y/x)\right] = \sum_{j=1}^{K} w_j(x) m_j(x) \tag{7}$$

### 2.2 Moving window algorithm for GMR

The strategy of the local GPR models adaptation proposed is the moving window approach (Kadlec et al., 2011). As the new complete input–output samples are acquired, the window slides along the samples belonging to the local model such that the oldest sample is discarded and the new one is added to the

window. In that way, not only is the current process behavior tracked, but also computational and memory requirements are constrained by the fixed window size. At some time instant t during online operation, the samples in a window of size N of the $k^{th}$ local model can be written as:

$$D_k^{(t)} = \{[x_{t+1-N}, \cdots, x_t], [y_{t+1-N}, \cdots, y_t]\}$$ (8)

with an appropriate regularized covariance matrix $\tilde{K}_t = (K_t + \sigma_j^2 I)$. The new complete input-output sample $\{x_{t+1}, y_{t+1}\}$, GPR model can be simply updated by removing first column and first row of covariance matrix due to marginalization property of the Gaussian process (Rasmussen, 2004)

$$\tilde{K}_{temp} = \left[\tilde{K}_t\right]_{-1,-1}$$ (9)

The new observation is added to the model by adding an appropriate row and column to the matrix $\tilde{K}_{temp}$:

$$\tilde{K}_{t+1} = \begin{bmatrix} \tilde{K}_{temp} & K_{t+1} \\ K_{t+1}^T & k_{t+1} + \sigma_j^2 \end{bmatrix}$$ (10)

where $[K_{t+1}]_i = K(x_i, x_{t+1})$, $x_i \in D_k^{(t)}$ and $k_{t+1} = k(x_{t+1}, x_{t+1})$. Then, the inverse of the matrix $\tilde{K}_{t+1}$ is calculated and used for prediction.

However, each model update requires the inverse of the $N \times N$ regularized covariance matrix $\tilde{K}_{t+1}$ which is computationally and memory demanding. This inverse can be efficiently calculated using the matrix inversion formula with the previous inverse $\tilde{K}_t^{-1}$. Generally, if the sample is removed from the window and a new sample added, the inverse of the new regularized covariance matrix can be efficiently calculated by the following expressions (Liu et al., 2011)

$$\tilde{K}_{temp}^{-1} = \left[\tilde{K}_t^{-1}\right]_{-i,-i} - \frac{\left(\left[\tilde{K}_t^{-1}\right]_{-i,i}\left[\tilde{K}_t^{-1}\right]_{-i,i}^T\right)}{\left(\left[\tilde{K}_t^{-1}\right]_{i,i}\right)}$$ (11)

$$\tilde{K}_{t+1}^{-1} = \begin{bmatrix} \tilde{K}_{temp}^{-1} & 0 \\ 0 & 0 \end{bmatrix} + \frac{1}{\sigma_{t+1}^2}\begin{bmatrix} \alpha\alpha^T & -\alpha \\ -\alpha^T & 1 \end{bmatrix}$$ (12)

where $\sigma_{t+1}^2 = k_{t+1} + \sigma_j^2 - k_{t+1}^T \tilde{K}_{temp}^{-1} k_{t+1}$ and $\alpha = \tilde{K}_{temp}^{-1} k_{t+1}$.

## 2.3 Online calibration model

Firstly, Initialize the prior $\pi_k$, mean $\mu_k$, covariance $\sum_k$ and meanwhile the sample is standardized. Assume $\pi_k^{(t)}$, $\mu_k^{(t)}$, and $\sum_k^{(t)}$ are the prior, mean, and covariance of the t step, respectively. When new samples $[x_{t+1} \quad y_{t+1}]$ are available, a new sample is added to the training sample set, and the most used samples are removed from the training samples. Modeling sample length and the number of components remain unchanged during online operation, then recursive update equations can be used in order to update the mean and covariance matrix of k-th Gaussian component and the related mixing coefficient. These are obtained from Zivkovic (2004) by ignoring Dirichlet prior on a number of components:

$$\pi_k^{(t+1)} = \pi_k^{(t)} + \beta(w_j(x_{t+1}) - \pi_k^{(t)})$$ (13)

$$\mu_k^{(t+1)} = \mu_k^{(t)} + \beta\frac{w_j(x_{t+1})}{\pi_k^{(t)}}(x_{t+1} - \mu_k^{(t)})$$ (14)

$$\sum_{k}^{(t+1)} = \sum_{k}^{(t)} + \beta \frac{w_j(x_{t+1})}{\pi_k^{(t)}} ((x_{t+1} - \mu_k^{(t)})(x_{t+1} - \mu_k^{(t)})^T - \sum_{k}^{(t)}) \tag{15}$$

where $x_{t+1}$ is the sample acquired at time instant t + 1. The influence of the samples is controlled by parameter $\beta$ which is usually set to a value of 1/T where T can be understood as the number of already acquired samples that are used in the mixture update. In the process of correction of Gauss parameters, the model performance evaluation criteria and confidence limits are needed to be set up. Generally, the model performance can be measured by the prediction error. The root mean square error (RMSE) is commonly used a measure, which can be calculated as (Mu et al., 2006)

$$RMSE = \sqrt{\frac{1}{N_s} \sum_{i=1}^{N_s} (y_i - \hat{y}_i)^2} \tag{16}$$

where $N_s$ denotes the number of samples in the test set, $y_i$ and $\hat{y}_i$ are the actual and predicted measurements respectively.

### 2.4 Online implementation of MPA-GMR

The implementation steps of MPA-GMR modeling method are as follows:

Step1. Determine the length of training sample $N$ and Gaussian component $K$, initialize the Gaussian mixture model parameters: the prior, mean and covariance.

Step2. The sample is standardized, constituting the training sample set [X, Y].

Step3. According to the GMR model, to estimate the prediction error of the modeling sample; according to eq.(18) to calculate the standard deviation, and regarded it as the confidence limit of forecast.

Step4. The sliding window moves one step forward to collect a new sample $\begin{bmatrix} X_{new} & Y_{new} \end{bmatrix}$. According to equation (4) calculating the mixing weight. The prior, the mean and the covariance of the training sample set are derived by the eq.(13), (14) (15) to take recursive correction.

Step5. Calculate the predicted measurements and RMSE using eq.(7) and eq.(16), respectively.

Step6. If $RMSE > \sigma_e$, turn to step2, otherwise turn to step4.

## 3. Industrial case study

In this work, the industrial data were collected from a practical erythromycin fermentation process in Zhenjiang medicine company, P. R. China (Liu et al., 2010). In the process, 15 process variables can be measured online: time, pH, dissolved oxygen, dextrin flow, propanol flow, soybean oil flow, water flow, air flow, soybean oil volume, propanol volume, dextrin volume, water volume, temperature, speed, and relative pressure. In each batch, 180 groups of data points can be collected by automatic instruments. The one batch is used as training set, and the other one batch is used as test set. Before constructing soft sensors, experimental data have to be normalized. Moreover, their dimensions should be reduced by variable selection methods for simplification. In this research, PCA method was used to select variables (King et al., 1999). Finally, five highly related variables selected as input variables are tabulated in Table 1. Biomass concentration, difficult to measure online, was chosen as the primary variable in soft sensing modeling.

Table 1: Input variables in the erythromycin fermentation process

| No. | Input variable | Unit |
|---|---|---|
| 1 | Temperature | K |
| 2 | PH | - |
| 3 | Agitator power | W |
| 4 | Dissolved oxygen concentration | g/L |
| 5 | relative pressure | N |

*Table 2: Results of normality test*

| Number of clusters | Meet the normal distribution of variables (p>0.05) | Number of variables |
|---|---|---|
| 1 | 1, 4 | 2 |
| 2 | 1, 2, 3 | 3 |
| 3 | 1, 2, 3, 4, 5 | 5 |
| 4 | 1, 2, 4, 5 | 4 |
| 5 | 1, 3, 4,5 | 4 |

By the proposed normality test method, the optimal number of Gaussian components can be achieved on training set. From Table 2, the optimal number of mixture components is obtained as $K_{opt}$ = 3. Meanwhile, the GMR and MPA-GMR are also constructed for comparisons of the quality prediction performance. The prediction results of the two different soft sensor approaches are shown in Fig. 1 and Fig. 2. It gives predicted values of the different model and original measurements, where the red line denotes the fitted biomass concentration measurements, the blue line represents the prediction results of the dynamic GMR soft sensor and the black dotted line depicts the 95% confidence interval of estimates.
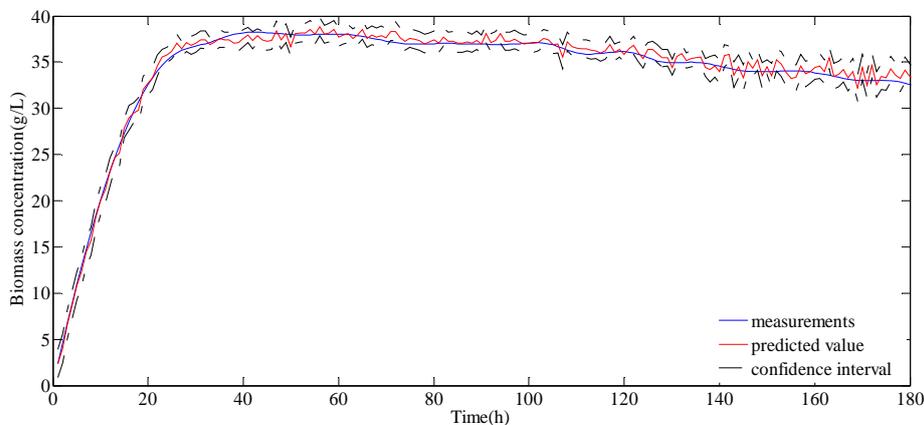


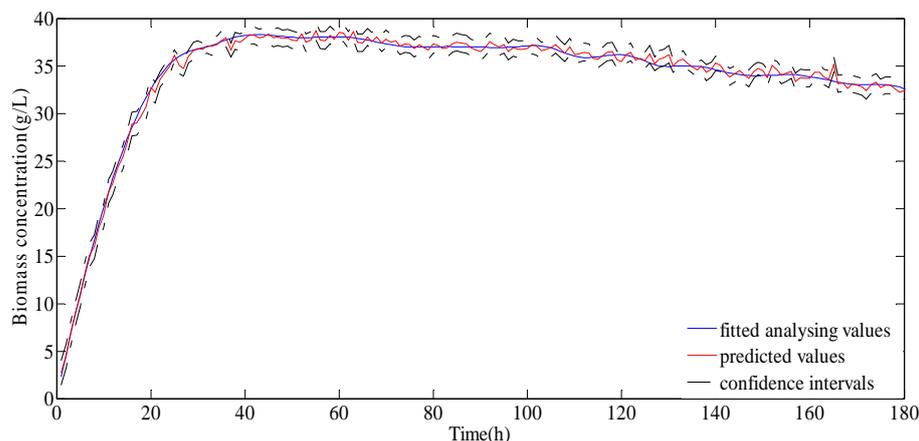*Figure 1: Simulation results of GMR model*



*Figure 2: Simulation results of MPA-GMR model*

## 4. Conclusions

In order to reduce the model calibration frequency of GMR model during modeling, the problems are solved about the confidence limits (or threshold) setting of the model and the difficulty of correction. This paper develops the modeling method based on performance evaluation model of GMR (MPA-GMR) and discusses the effect on improving the performance of the model. The simulation was carried out to study, where erythromycin fermentation reaction process as the research object. This modeling method generates

confidence limits of the model automatically according to initial characteristics process; and results of an evaluation of model performance, selectively activating correction model calibration and confidence limits. Simulation results show that the MPA-GMR modeling method can improve the accuracy of the model and the computational efficiency.

**Acknowledgments**

**References**

Birol G., Undey C., Cinar A., 2002, A modular simulation package for fed-batch fermentation: penicillin production, Computers & Chemical Engineering, 26, 1553-1565.

Cui L., Xie P., Sun J., Yu T., Yuan J., 2012, Data-driven prediction of the product formation in industrial 2-keto-l-gulonic acid fermentation, Computers & Chemical Engineering, 36, 386-391.

Jin H., Chen X., Yang J., Wang L., Wu L., 2015, Online local learning based adaptive soft sensor and its application to an industrial fed-batch chlortetracycline fermentation process, Chemometrics and Intelligent Laboratory Systems, 143, 58-78.

Kadlec P., Grbić R., Gabrys B., 2011, Review of adaptation mechanisms for data-driven soft sensors, Computers & Chemical Engineering, 35, 1-24.

Kaneko H., Funatsu K., 2014, Adaptive soft sensor based on online support vector regression and Bayesian ensemble learning for various states in chemical plants, Chemometrics and Intelligent Laboratory Systems, 137, 57-66.

King J. R., Jackson D.A., 1999, Variable selection in large environmental data sets using principal components analysis, Environmetrics, 10, 67–77.

Li L., Su H., Chu J., 2009, Modeling of isomerization of C8 aromatics by online least squares support vector machine. Chinese Journal of Chemical Engineering, 17(3):437-444.

Liu W., Principe J.C., Haykin S., 2011, Kernel Adaptive Filtering: A Comprehensive Introduction. The 2013 International Joint Conference on Neural Networks (IJCNN), 1 - 6.

Liu G., Zhou D., Xu H., Mei C., 2010, Model optimization of SVM for a fermentation soft sensor, Expert Systems with Applications, 37, 2708-2713.

Luttmann R., Bracewell D. G., Cornelissen G., Gernaey K. V., Glassey J., Hass V. C., et al., 2012, Soft sensors in bioprocessing: a status report and recommendations, Biotechnology Journal, 7(8), 1040.

Mu S., Zeng Y., Liu R., Wu P., Su H., Chu J., 2006, Online dual updating with recursive PLS model and its application in predicting crystal size of purified terephthalic acid (PTA) process, Journal of Process Control, 16, 557-566.

Rasmussen C.E., 2004, Gaussian Processes in Machine Learning, Advanced Lectures on Machine Learning. Springer Berlin Heidelberg.

Sharmin R., Sundararaj U., Shah S., Griend L.V., Sun Y.J., 2006, Inferential sensors for estimation of polymer quality parameters: Industrial application of a PLS-based soft sensor for a LDPE plant, Chemical Engineering Science, 61, 6372-6384.

Sun K., Liu J., Kang J., Jang S., Wong S., Chen D., 2014, Development of a variable selection method for soft sensor using artificial neural network and nonnegative garrote, Journal of Process Control, 24, 1068-1075.

Sung H. G., 2004, Gaussian mixture regression and classification, Rice University.

Wang Z., He Q., Wang J., 2015, Comparison of variable selection methods for PLS-based soft sensor modeling, Journal of Process Control, 26, 56-72.

Yuan X., Ge Z., Song Z., 2014, Soft sensor model development in multiphase/multimode processes based on Gaussian mixture regression, Chemometrics and Intelligent Laboratory Systems, 138, 97-109.

Zhang X., Li Y., Kano M., 2015, Quality prediction in complex batch processes with just-in-time learning model based on non-Gaussian dissimilarity measure, Industrial & Engineering Chemistry Research.

Zivkovic Z., Ferdinand V. D. H. 2004. Recursive unsupervised learning of finite mixture models. IEEE Transactions on Pattern Analysis & Machine Intelligence, 26(5), 651-656.