Soft-Sensor Modeling for Semi-Batch Chemical Process Using Limited Number of Sampling

Shinichiro Aoshima,† Tomoyuki Miyao,†,‡ and Kimito Funatsu*†,‡,§

†Graduate School of Science and Technology, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan
‡Data Science Center, Nara Institute of Science and Technology, 8916-5 Takayama-cho, Ikoma, Nara, 630-0192, Japan
§Department of Chemical System Engineering, School of Engineering, The University of Tokyo, 7-3-1 Hongo. Bunkyo-ku, Tokyo 113-8656, Japan

(Received November 5, 2019; Accepted December 23, 2019)

Batch or semi-batch processes have been of great use in various industrial chemical plants. For efficiently monitoring such processes, soft-sensor models can be employed. Many of previously proposed soft-sensor models assumed that objective variable values for model construction can be available at any time during process operation. However, in many chemical plants, it is difficult to sample product from the ongoing process due to such extreme reaction conditions as high pressure and temperature. Therefore, understanding the relationship between time-series soft-sensor model’s predictability and the number of sampling points is important. In the present work, we clarified this relationship using simulation datasets, which can be easily reproduced. When sampling points were scarce, data augmentation strategy was also found to be effective. Soft-sensor models can be effectively built using sampling points in the early phase of the process. These findings were applied to build a soft-sensor model of an industrial semi-batch process.

Key Words: soft sensor, batch processes, process monitoring, quality control, a limited number of samplings

Introduction

Batch or semi-batch processes are widely used in pharmaceutical, materials, and chemical industries due to their operational flexibility. Thanks to rapid advances in computational power and sensors, many in-silico techniques for online monitoring processes and controlling product quality have been developed. Among the techniques, multiway principal component analysis [1] and multiway partial least squares are robust methods, focusing on phases of these processes [2]. A phase in these processes is defined as a single processing unit caused by operational procedures, including chemical reactions. Modeling methods for each phase are also developed [3]. These methods are able to deal with the different durations of batches [4-6], and to identify the phases based on the process variables using ordinary clustering techniques [7]. When building soft-sensor models for these processes, most modeling methods assume that objective variable values, which usually represent product quality, can be obtained at any point during the ongoing process. In some chemical batch process, however, these objective variable values are not always available at any time. This is because sampling itself is sometimes dangerous under extreme operating conditions, such as high pressure and high temperature. Even when sampling is possible, the product quality may not be obtained without the time delay of the measurement. These chemical processes

* funatsu@chemsys.t.u-tokyo.ac.jp
require soft-sensor or data-driven statistical models because of the high sampling cost. Furthermore, the constructed soft-sensor models should be highly predictive enough for the time-series prediction of the product quality. The attempt of constructing soft-sensor models of a batch process with a limited number of data points is fundamentally challenging because extrapolation of the models should be handled properly [8]. Only using the final product quality values, which are always accessible after the process terminates, apparently would fail to construct highly predictive time-series prediction models. Therefore, incorporating data points of the product quality during the ongoing process is inevitable for the purpose.

In this study, two topics were investigated to obtain useful insights on sampling protocol for soft-sensor modeling for batch or semi-batch chemical processes. First, we investigated in which phase and how many data points should be obtained to build soft-sensor models with high predictability. Predictability of soft-sensor models for time-series prediction of a batch process depends on the number and sampling timing of intermediate points (i.e. data points during the ongoing process). The second topic is applying a data augmentation strategy for compensating the data sparseness of intermediate points.

### Materials and Methods

#### Penicillin fermentation process

Simulated penicillin fermentation process implemented in PenSim (version 2.0) [9] was chosen as a benchmark for understanding the relationship between generalization ability of soft-sensor models and incorporation of intermediate data points including data augmentation. Flow diagram of the fed-batch penicillin fermentation process is presented in Figure 1. In this process, two Proportional-Integral-Differential controllers (PID controller) maintained pH and temperature values in the batch at some levels by adjusting related process variables. As simulation conditions, initial substrate concentration (g/L) was set to 20, pH (-) to five, feed substrate concentration (g/L) to 0.045 and feed substrate temperature (K) to 296. Gaussian noises were added to pH and temperature values measurements. The rest of the settings remained default. Simulation duration was set to 600 h by 0.5 h time step.

![Figure 1. Flow diagram of fed-batch penicillin fermentation process.](image)

#### Data preparation

In this study, various timings of sampling penicillin concentration were tested. In all the PenSim simulations, a batch cluster has been grown first, followed by a fed-batch operation by the depletion of glucose. The initial batch-process part usually took about 40 h. Therefore, sampling was conducted after the fed-batch process started. As shown in Figure 2., the sampling points during the ongoing process (termed “intermediate points”) were categorized into four phases by equally splitting the duration: very early, early, middle and late phases.
Duration for the very early phase (VEP) was from 55h to 105h, the early phase (EP) from 125h to 175h, the middle phase (MP) from 185h to 235h, and the late phase (LP) from 245h to 295h from the beginning of the process operation. For each batch process, one intermediate point was randomly sampled in each of the four phases, meaning that four intermediate points and the one final penicillin concentration (termed “end point”) were sampled for a batch. An end point for a batch was defined as the first point where penicillin concentration did not change for five consecutive time steps. For some batches, simulation reached the end point before sampling all four intermediate points. The number of batches used in this study as well as the number of intermediate points classified into the four phases are presented in Table 1. The objective variable was the concentration of penicillin, and independent variables were nine process variables, reported in Table 2. Data distributions for all the process variables including the concentration of penicillin are reported in Figure 3.

![Figure 2. Division of four phases and sampling timing.](image)

**Table 1. Penicillin data set profile.**

| No. batches | No. intermediate points |
|-------------|-------------------------|
|             | Very early phase | Early phase | Middle phase | Late phase |
| 243         | 243              | 238         | 223          | 158        |

**Table 2. Process variables in the penicillin process.**

| Variable description (Unit)         | pH (-)  | Temperature (K) | Generated heat (kcal) | Cooling water flow rate (L/h) |
|------------------------------------|---------|----------------|-----------------------|-------------------------------|
| Aeration rate (L/h)                |         |                |                       |                               |
| Substrate feed temperature (L/h)   |         |                |                       |                               |
| Dissolved oxygen concentration (g/L) |         |                |                       |                               |
| Volume (L)                         |         |                |                       |                               |
| Carbon dioxide concentration (g/L) |         |                |                       |                               |
Figure 3. Process variable distributions. Process variables of data points for model construction as well as validation are reported. Each plot represents one process variable including the concentration of penicillin as the objective variable. Data points were sorted in the descending order of the concentration of penicillin.

Soft-sensor Modeling Methods

Times-series prediction of an objective variable by soft-sensor models built with limited sampling points is fundamentally difficult because majority of data points for the prediction are not similar to the ones used for model construction. For handling these heterogeneous data properly, and understanding for which intermediate points the quality of product can be predicted with high accuracy or not, many machine learning techniques were tested including Gaussian process regression (GPR) [10]. In GPR, Gaussian distributions are assumed for any combinations of prediction functions evaluated at arbitrary set of data points, making it possible to derive a Gaussian distribution as a prediction output. Consequently, prediction uncertainty can be measured based on the width of the Gaussian (standard deviations). This feature seems suitable for fairly evaluating prediction result of product qualities for data points that are sometimes not similar to training data points. For comparing purpose, partial least squares regression (PLS) [11], and support vector regression (SVR) [12] in combination with radial basis function (RBF) as the kernel function, random forest regression (RF) [13] and gradient boosting regression (GBR) [14] were also used. PLS is a linear regression method, known for a standard soft-sensor modeling method. Applications with PLS ranged from industrial process system modeling [15,16], to fault diagnosis, [1,17] due to its capability of handling large quantity of process data with highly correlated variables. SVR is a variant of support vector machine, [18,19] which is characterized its robust prediction due to its objective function for model construction. Although SVR is originally a linear model, it can be extended to non-linear regression method through mapping by a kernel function. In this study, ν-SVR was used that employs a ν parameter representing an upper bound on the fraction of training error [20]. RF is an ensemble learning method, capable of both classification and regression tasks. In RF, a large number of decision trees are built, forming a forest, and the mean value of all the tree predictions becomes the final output of the model [21]. GBR is also an ensemble learning method, but different from RF in the way of using ensemble. GBR makes use of the boosting technique instead of bagging, which is the basis of RF. GBR is getting a standard method in the machine learning community and known for its high generalization ability [22].

Implementation

All soft-sensor models except GPR were constructed with the help of Scikit-learn (version 0.20.3) modules. For GPR, GPy (version 1.9.8) [23] was used. Hyper-
parameters for the models were optimized by 10-fold cross validation based on mean absolute error (MAE). For PLS, the number of components is the hyper-parameter, for GPR, the length scale of matérn covariance function (Matérn form with a \( \nu \) parameter of 5/2) as the kernel function and the variance of Gaussian noise for GPR, for SVR, \( \gamma \), \( \nu \) and \( C \), for RF, the maximum depth of a tree, the maximum number of features and the number of trees, and for GBR, the type of loss function, the maximum depth of a tree, the maximum number of features and the number of estimators.

Performance Measures for Time Series Prediction

In penicillin fermentation process, soft-sensor models were evaluated based on the mean absolute error (MAE) of the time-series prediction for each test batch. Batch-based MAE is defined as follow:

\[
MAE = \frac{1}{T-t_{st}+1} \sum_{t=t_{st}}^{T} |y_{t,obs} - y_{t,pred}|
\]

where \( y_{t,obs} \) denotes the observed \( y \) value at time \( t \); \( y_{t,pred} \) the predicted \( y \) value; \( T \) the time of end point; and \( t_{st} \) the time to start time-series predicting. All the batches were randomly split into test and training data sets. Proportion for training and test was chosen to 1:1. Models were evaluated on the basis of median MAE values and standard deviations for all the test batches.

Results and Discussion

Study Design

For industrial chemical batch- or semi-batch processes, sampling product in query during the ongoing process is sometimes difficult due to its extreme conditions [24]. In that case, only limited number of data for product quality (objective variable) can be acquired during operation, as well as one data point for the final product quality. This scarceness of intermediate data points makes it difficult to estimate objective variable values during batch process operation. Nevertheless, product quality should be time-series predicted during batch process operation to understand the batch and to control the termination of the process. Only simulation data enabled us to understand the relationship between the scarceness of intermediate data points for the objective variable and soft-sensor modeling predictability. This was because a fed-batch simulation process was employed in this study. Using the penicillin fermentation, fed-batch process was implemented in PenSim [9].

Two aspects of soft-sensor modeling for semi-batch processes were investigated: number of intermediate data points and sampling timings for reliable time-series prediction of the objective variable (1), and the effect of data augmentation to improve soft-sensor models’ generalization ability (2). As a demonstrative case study, a polymerization semi-batch process from the industry was analyzed. Our findings and protocols established using the penicillin fed-batch process were applied to the time-series prediction of the polymer quality for intermediate data points in the polymerization semi-batch process.

Incorporation of Intermediate Data

The first aspect of soft-sensor modeling for semi-batch process, i.e. the relationship between number of intermediate points and timings and model performance (1), was investigated with the data sets generated by PenSim. The proportion of training and test data sets was 1:1. The category of intermediate data points as well as the number of batches are presented in Table 1 according to the description in the Materials and Methods section. The ratio of intermediate data points over the final product quality points (end points) in training data sets was varied from 0% (no intermediate points) to 100% (approximately the same number of intermediate points as that of the end points) shown in Table 3. Sampling timings for these intermediate points were simultaneously altered in the four phases (VEP, EP, MP and LP). Although all the combinations of the ratio of intermediate points were tested, models incorporating VEP were generally worse in performance (not shown here). Therefore, ratios of the categorized number of intermediate points over the total number of intermediate points on the phase basis (VEP: EP: MP: LP) were set to 1:0:0:0, 0:1:0:0, 0:0:1:0, 0:0:0:1, 0:1/2:1/2:0, 0:1/2:0:1/2, 0:0:1/2:1/2, and 0:1/3:1/3:1/3.

For example, a combination of 0:1/2:1/2:0, with the ratio of the intermediate data points over the end points as 15%, indicates that out of total number of 144 training data points, 18 intermediate points were taken into account. And, nine intermediate points were assigned for each of the EP and MP. For each combination of a number of intermediate points and a ratio of phases, five trials were conducted with different splitting of training and test data sets by changing random states. Model performance in terms of the median value of batch-based MAEs for a test data set are reported in Figure 4. Starting time of the test data set was set to 150 h (center of EP) to confirm the accuracy of time-series prediction after reasonable time passed from the beginning of the fed-batch process. Overall, irrespective of modeling methods, model predictability improved with the number of intermediate data points. However, when either of VEP and LP data sets were used, performance improvement was not observed for PLS, RF and GBR by increasing the number
of intermediate points. For combinations of GPR/SVR and VEP/LP data sets, performance improved when more intermediate points were incorporated. Both these methods employed kernel functions based on distances in process variable space. This might be a point of expanding their domain of applicability to a certain extent. Incorporating intermediate data points in the EP improved the predictability of the models when comparing four lines (1:0:0:0, 0:1:0:0, 0:0:1:0 and 0:0:0:1) in each of the plots (Figure 4). When the ratio of intermediate points is less than 10%, incorporating intermediate points in the VEP gave better performance for PLS, GPR and SVR. Using intermediate data points in LP and MP (0:0:1:0, 0:0:1/2:1/2) showed worse performance. For these data sets, model performance continued to increase and saturation was not observed. Combinations of EP and MP (0:1:0:0, 0:1/2:1/2) showed the best performance for all the methods, suggesting that when sampling intermediate points for time-series prediction, the sampling should be conducted in the EP of the time-region of interest. When the percentage of intermediate points exceeded 20%, further significant performance improvement was not observed and performance looked saturated. Therefore, sampling around 20% of the number of end points is a reasonable starting point for constructing soft-sensor models for time-series prediction. Exemplary time-series predictions by models with different modeling methods and phases for intermediate points are reported in Figure 5. After the time of the intermediate points was taken into the training data set, soft-sensor models with GPR and PLS successfully predicted the concentration of penicillin, but for GBR, only the regions where intermediate points were sampled could be predicted. Furthermore, incorporating intermediate points in the EP were sufficient to predict data points in the MP and LP using PLS and GPR. On the other hand, GBR tended to overfit the incorporated points, showing poor generalization ability. GPR predicted the objective variable information as a Gaussian distribution. Shaded regions in the GPR time-series prediction plots (Figure 5) represents the three standard deviations of the distributions.

In contrast to our expectation, standard deviation values did not change much in one batch process after semi-batch operation started, irrespective of the phases of intermediate points. This implied that data points in the phases that were not incorporated in the training data set were very similar to the training data points. Soft-sensor models in a batch or semi-batch process can be used to online monitor (predict) product quality in order to control the process based on the monitored (predicted) quality values. In this case, identifying when the monitored process reaches a specific product quality value is of importance. Therefore, the time when a specific penicillin concentration was reached based on a model was compared to ground truth time for the concentration. If the time difference between model-based and ground truth was large, soft-sensor model-based operation should not be employed. Four concentrations were tested: 0.8, 1.0, 1.2 and 1.4, representative concentrations in VEP, EP, MP and LP, respectively. Proportion of the number of intermediate points to that of end points was set to 0.20. All the test data sets were used for this experiment, and median absolute error values are reported in Table 4. It is obvious that intermediate points in the VEP were necessary to identify proper time for penicillin concentrations of 0.8 and 1.0. Furthermore, for the concentrations in the MP and the LP, incorporating intermediate points in the EP or VEP into the training data penicillin concentration. Although intermediate points in the VEP did not contribute much in reducing the entire error in time-series prediction (Figure 4), it did help in making entire shape of time-series predicted concentration more accurate.

| Ratio of No. intermediate points to | Training data points |
|-----------------------------------|----------------------|
| that of end points [%]            | Total | Intermediate |
| 0                                 | 126   | 0            |
| 5                                 | 132   | 6            |
| 10                                | 138   | 12           |
| 15                                | 144   | 18           |
| 20                                | 151   | 25           |
| 25                                | 157   | 31           |
| 30                                | 163   | 37           |
| 100                               | 249   | 123          |
Figure 4. Performance of machine learning models using training data sets containing data points in different sampled timings. Machine learning models were built using training data sets containing data points during process operation (termed “intermediate points”) in various timings. For each machine learning method, median values of MAEs are plotted against the ratio of the number of “intermediate points” in the training data set over the number of data points for the final penicillin concentration (termed “end points”). Different lines represent different ratios of “intermediate points” in the four phases: very early, early, middle and late, according to Table 3.
Figure 5. Exemplary time series predictions of penicillin concentration. For each combination of a machine learning method and a ratio of the “intermediate points” in the four phases: very early, early, middle and late from left to right, time-series predictions of penicillin concentration are presented in blue dots. The red line represents the ground truth and shaded regions for GPR are predicted standard deviation from the mean value. Proportion of the number of “intermediate points” to that of “end points” was set to 0.15.
Table 4. Median value of absolute time difference [h] between ground truth and model-based determination.
For each target penicillin concentration, time difference between ground truth and the point determined by models are reported as the median value of all the test batches. The ratio of the intermediate points to that of end points was set to 20%, and the intermediate points were sampled in one of the very early (VEP), early (EP), middle (MP) and late (LP) phases.

| Target penicillin concentration [g/L] | PLS | GPR | GBR |
|--------------------------------------|-----|-----|-----|
| LP | MP | EP | VEP | LP | MP | EP | VEP | LP | MP | EP | VEP |
| 0.8 | 62.5 | 62.5 | 52.5 | **16.5** | 62.5 | 62.5 | 48.5 | **18** | 62.5 | 62.5 | 62.5 | **28.5** |
| 1.0 | 84.5 | 80.5 | 35.5 | **15.5** | 91 | 83 | 30.5 | **25** | 101 | 101 | 87.5 | **51** |
| 1.2 | 63 | 40 | 15.5 | **12.8** | 66.5 | 40.3 | **18** | 30 | 86 | 52.5 | **23** | 62.5 |
| 1.4 | 0 | 0 | 0 | 0.5 | 0 | 0 | 0.5 | 0 | 0 | 0.5 | 0.25 |

Incorporation of Pseudo Data

When there are not many intermediate points available for model construction, using pseudo data points is an option generally used for the data augmentation strategy. [25] In this strategy, pseudo data points based on a current data set are generated, and these points are incorporated into the training data set to improve the predictability of the model or expanding the domain of applicability of the model. In our study, this second issue was also investigated considering that intermediate sampling is expensive. As shown in Figure 6, pseudo data points were generated by interpolating between an intermediate point of a batch and the end point of the batch. Gaussian noise of variance 0.1 was added. From these generated data points, pre-determined number of data points were selected with an equal interval. Model performances in terms of the median value of batch-based MAEs for a test set were measured by changing the number pseudo data points incorporated in the models.

![Figure 6. Generation of pseudo data points.](image-url)
Different sampling timings for the intermediate data points were simultaneously tested as in the previous investigation for intermediate data point effects. These intermediate data points were used for pseudo data generation; hence it could affect model performance. Furthermore, two ratios (5% and 10%) of the number of intermediate data points over that of the end points were tested. Only performances associated with sampling in the early phase are presented in Figure 7 with three modeling methods (a: PLS, b: GPR and c: GBR), as exemplary results. Representation of the proportion of intermediate points is according to Table 3. In Figure 7, all machine learning models were found to reduce the median MAE by incorporating a limited number of pseudo data points, in particular, using less number of intermediate points (5%). Only single pseudo data points per batch drastically reduced the errors for this category. Incorporating more number of pseudo data points, however, did not contribute to model performance improvement. On the contrary, GPR showed performance degradation when 15 pseudo data points per batch were incorporated irrespective of the ratios of intermediate point phases (Figure 7b). This trend was also observed for SVR (not shown here), implying that the distance-based kernel functions might be sensitive to the objective variable credibility because pseudo data points for model construction contained Gaussian noises for avoiding overfitting to the objective variable values for the points. Summarizing this analysis, incorporating pseudo data points in training data enhanced the performance of time-series prediction on the condition that intermediate data points were scarce and the number of data points per batch were a few.

Figure 7a.
Figure 7. Incorporation of Pseudo data points. Model performance was monitored in terms of the median MAE for the test data set. For each one of the machine learning methods (a: PLS, b: GPR and c: GBR) and different proportion of “intermediate points”, model performance was plotted against the number of pseudo data points in the training data set. Representation of the proportion of the “intermediate points” is according to Table 3. The ratios of the number of total intermediate points over that of end points were set to 5% (orange) or 10% (green)
Application to Industrial Semi-Batch Process

As a practical application of our approach to time-series prediction of a semi-batch process, an industrial semi-batch process was selected and scrutinized. The semi-batch process was a polymerization process in a chemical plant at DAIKIN INDUSTRIES, LTD. Due to severe reaction conditions (e.g. high pressure), frequent sampling of the product during the ongoing process was not feasible. Furthermore, required quality level of the product was sometimes changed, meaning that online prediction of the product quality was needed to determine when the reaction should be terminated. Data for this process consisted of 1265 points, including 344 intermediate points. The objective variable (product quality) was melt flow rate (MFR), which represents polymer liquidity. Online monitored process variables were 27 variables, including the pressure of the batch reactor. The goal of this analysis was to build a high-predictive soft-sensor model, which enables time-series prediction of MFR values from monitored process variable values. Since there were few intermediate points in the (very) early phase, the data augmentation strategy was adopted. For evaluating predictability performance focusing on the high MFR regions, one sample with the highest MFR value was traced. This can be rationalized by the fact that true time-series MFR values could not be accessed because there were up to two data points per batch (one end point and one intermediate point) and the higher MFR values implied that the data point was in the earlier phase of the process. Therefore, the data point with the highest MFR value was in the test data set and the data points with the second to the fourth highest MFR values were in the training data set. Pseudo data points were generated, whose MFR values were more than or equal to 40. The number of pseudo data points per batch was limited up to five. Model construction and evaluation trials were repeated 10 times. The absolute prediction error for the highest MFR data point is reported in Figure 8 as box plots. Incorporating pseudo data points into the training data set reduced the absolute error irrespective of modeling methods. Correlation plots using GPR as a modeling method were shown in Figure 9 as an example, and introducing pseudo data points moved the highest MFR point to closer to the diagonal line (Figure 9b). Furthermore, both training and test data points with higher MFR points were moved toward the diagonal line by slightly compensating the data point predictability in the regions with middle or low MFR values.

![Figure 8. Absolute error of prediction for the data point with the highest MFR value.](image-url)
Conclusions

In this paper, we have proposed a scheme for time-series soft-sensor model construction when a limited number of intermediate points were available in a batch or semi-batch process. Which intermediate points should be sampled and incorporated into the training data set for soft-sensor models was investigated. Because only simulated data sets allowed us to investigate this aspect, a fed-batch process for penicillin fermentation implemented in PenSim [9] was investigated with careful experimental settings.

As results, it was found to be possible to build time-series models with high predictability using only intermediate points in the early phase (EP). Also, performance saturation was observed when the number of intermediate points were around 20% of that of the end points. Although intermediate points in the very early phase (VEP) did not contribute much in reducing error in time-series prediction, it helped in making the entire shape of time-series prediction more reasonable, supported by the fact that models built with intermediate points in the VEP predicted more accurate time for reaching a specific penicillin concentration. We also demonstrated that the data augmentation strategy was useful if there were not enough number of intermediate points. Overall, incorporating a few pseudo intermediate points per batch improved model predictability, in particular when the number of intermediate points in the model were smaller. However, more than 10 pseudo data points per batch did not show any improvement. On the contrary, for GPR, the predictability of the models decreased when incorporating more than 10 pseudo data points per batch.

As a demonstrative application of an actual chemical semi-batch process, a polymerization semi-batch process was analyzed for building a time-series soft sensor model. In this process, incorporating pseudo data points when building soft-sensor models contributed to predictability improvement, in particular predictability in the EP.

Acknowledgments

We thank DAIKIN INDUSTRIES, LTD. for providing the process data for polymerization.

References

[1] P. Nomikos, J. F. MacGregor, AIChE J. 40, 1361-1375 (1994).
[2] P. Nomikos, J.F. MacGregor, Chemom. Intel. Lab. Syst. 30, 97-108 (1995).
[3] C. Undey, A. Cinar, IEEE Contr. Syst. Mag. 22, 40-52 (2002).
[4] A. Kassidas, J.F. MacGregor, Taylor. P.A., AIChE J. 44, 864-875 (1998).
[5] S.G. Rothwell, E.B. Martin, A.J. Morris, IFAC proceedings volumes 31, 67-72 (1998).
[6] D. Neogi, C. Schlags, Ind. Eng. Chem. Res. 37, 3971-3979 (1998).
[7] Jie Yu, S. Joe Qin, Ind. Eng. Chem. Res. 48, 8585-
8594 (2009).
[8] H. Kaneko, M. Arakawa, K. Funatsu, *AIChE J.*, **57**, 1506-1513 (2011).
[9] G. Birol, C. Undey, A. Cinar, *Comput. Chem. Eng.* **26**, 1553-1565 (2002).
[10] C.K. Williams, C.E. Rasmussen, **14**, 69-106 (2006).
[11] P. Geladi, B.R. Kowalski, *Anal. Chim. Acta* **185**, 1-17 (1986).
[12] B.E. Boser, I.M. Guyon, V.N. Vapnik, *Proceedings of the fifth annual workshop on Computational learning theory*, 144-152 (1992).
[13] A. Liaw, M. Wiener, *R News*, **2**, 18-22 (2002).
[14] J.H. Friedman, *Annals of statistics* **29**, 1189-1232 (2001).
[15] P. Kadlec, B. Gabrys, S. Strandt, *Comput. Chem. Eng.* **32**, 795-814 (2009).
[16] S. Wold, M. Sjöströma, L. Eriksson, *Chemom. Intell. Lab. Syst.* **58**, 109-130 (2001).
[17] J. Chen, K. Liu, *Chem. Eng. Sci.* **57**, 63-75 (2002).
[18] Y. Liu, Z. Gao, P. Li, H. Wang, *Ind. Eng. Chem. Res.* **51**, 4313-4327 (2012).
[19] W. Yan, H. Shao, X.Wang, *Comput. Chem. Eng.* **28**, 1489-1498 (2004).
[20] C.C. Chang, C.J. Lin, *Neural Computation* **14**, 1959-1977 (2002).
[21] Lara F.A. Napier, Chris Aldrich, *IFAC-PapersOnline*, **50**, 1175-1180 (2017).
[22] R.S. Zemel, T. Pitassi, *Advances in neural information processing system* **13**, 696-702 (2001).
[23] Gpy. (2012) (URL=http://github.com/SheffieldML/GPy)
[24] R. Javid, P. Jacques, *Macromolecules* **35**, 1489-1490 (2002).
[25] Allen, M. David, *Technometrics* **16**, 125-127 (1974).