

Data-driven Online Operating Performance Assessment for Multi-Datasets Multivariable Industrial Processes

Yupeng Du^a, Zhenlei Wang^{a,*}, Xin Wang^b

^aKey Laboratory of Advanced Control and Optimization for Chemical Processes, East China University of Science and Technology, Shanghai 200237, China

^bCenter of Electrical & Electronic Technology, Shanghai Jiao Tong University, Shanghai 200240, China
wangzhen_l@ecust.edu.cn

In this study, a novel online operating performance assessment method based on multi-sets two-step basis vector extraction artificial neural networks (MTBVE-ANN) strategy is proposed for industrial applications. The MTBVE-ANN method focuses on finding common and specific information involved in multi-datasets, which improves the accuracy of data nonlinear characterization with artificial neural networks introduced. The optimality related variations are extracted from each operating performance grade by analysing the common and unique variations over online steady performance grades. The online operating performance assessment method is performed based on the similarities between the optimality related variations of the test data and that of historical training data. Previously, total projection to latent structures (T-PLS) operating performance assessment method must be performed based on the availability of both input and output data. The proposed method in this paper takes the artificial neural network to assess the operating performance grade of the online test data without output. The validity and precision of the proposed operating performance assessment method is illustrated with the industrial data of multi-datasets multivariable industrial processes.

1. Introduction

Under the conditions of maximizing the protection of the environment, maximizing economic efficiency has become the focus. Owing to the complexity of the chemical process, maintaining the excellent operating performance assess of chemical process control system is crucial. Its optimization control strategy research is always the hot spot of process control (Hwang et al., 2013), but due to process disturbance, equipment wear, instrument failure and operational environment changes and other uncertain factors, the operating state may gradually deviate from the optimal operating point, resulting in economic performance deterioration and environmental damages. The online process performance monitoring and optimization method is very meaningful. In 1989, Harris proposed the concept of "feedback invariant" to solve the control performance assessment (CPA) problem of single-variable control system. This marks the beginning of CPA theory (Harris et al., 1989). In order to further assess the system operating performance from a data-driven perspective, Qin et al. (2008) proposed a data-based covariance benchmark for control performance monitoring. Therefore, data-driven performance assessment methods gradually become a hot spot. To solve the above problem, Zhou et al. (2010) proposed total projection to latent structures (T-PLS) for process monitoring to reduce false alarm and missing alarm rates of faults related to output data and improve the method for nonlinear processes performance monitoring. Liu et al. (2014) applied the T-PLS algorithm to the field of offline operating performance assessment by comparing the online data and offline model similarity according to the actual needs.

However, the immediacy of online operating performance assessment is limited by the scarcity of the online output data and the nonlinearity of the data degrades the accuracy of assessment. In this study, a novel online operating performance assessment method called MTBVE-ANN is proposed to solve the problems above. In data feature extraction, the paper takes a multi-sets two-step basis vector extraction strategy (MTBVE) on input data, which focuses on finding their common subspace and specific subspace over sets. In online assessment, sliding time window and the artificial neural network approach is carried out for performance

grade classifier. The similarity indices between test data and those of each online performance grade are measured between 0 and 1, which reflect the grade membership.

The remainder of this article proceeds as follows. In Section 2, multi-sets two-step basis vector extraction strategy is briefly summarized. In Section 3, the offline extraction modeling and online operating performance assessment process is demonstrated. In Section 4, MTBVE-ANN method is taken into operation on one chemical factory real online input data. Finally, the main conclusions and acknowledgements are presented at the end.

2. Summary of Multi-sets Two-step Basis Vector Extraction Strategy

To analyse the common variable correlation shared among multi-datasets, Zhao et al. (2011) proposed the method. This method solves the problem of finding the basis vector by the means of mathematical solution to constrained optimization. Assume there are C input datasets, and the i -th dataset is denoted as

$$\mathbf{X}_i = [\mathbf{x}_{i,1}, \mathbf{x}_{i,2}, \dots, \mathbf{x}_{i,N_i}]^T \in R^{N_i \times J} \quad (1)$$

Where $i = 1, 2, \dots, C$, N_i and J represent the numbers of samples and process variable respectively. According to industrial production experience, the input data is divided into different datasets. Each dataset is represented by some sub-basis vectors which are similar or close related with each other over various performance grade. The paper defined each sub-basis vector as a linear combination of the original sample subspace as

$$\mathbf{p}_{i,j} = \sum_{n=1}^{N_i} a_{i,j,n} \mathbf{x}_{i,n} = \mathbf{X}_i^T \mathbf{a}_{i,j} \quad (2)$$

Where $\mathbf{p}_{i,j} \in R^{J \times 1}$ is denoted as the j th sub-basis vector of \mathbf{X}_i and $\mathbf{a}_{i,j} = [a_{i,j,1}, \dots, a_{i,j,N_i}]^T$ as the combination coefficient. In order to simplify the interrelationships over sets, $\mathbf{p}_g \in R^{J \times 1}$ is denoted as pseudo $(C+1)$ -th sub-basis vector. Sustained, the extraction problem is transferred into a constrained optimization problem as

$$\begin{aligned} \max R^2 = \max \sum_{i=1}^C (\mathbf{p}_g^T \mathbf{X}_i^T \mathbf{a}_i)^2 \\ \text{s.t.} \begin{cases} \mathbf{p}_g^T \mathbf{p}_g = 1 \\ \mathbf{a}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{a}_i = 1 \end{cases} \end{aligned} \quad (3)$$

In order to obtain the solution of the optimization problem, construct Lagrange operator is introduced as

$$F(\mathbf{p}_g, \mathbf{a}_i, \lambda_g, \lambda_i) = \sum_{i=1}^C (\mathbf{p}_g^T \mathbf{X}_i^T \mathbf{a}_i)^2 - \lambda_g (\mathbf{p}_g^T \mathbf{p}_g - 1) - \sum_{i=1}^C \lambda_i (\mathbf{a}_i^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{a}_i - 1),$$

in which λ_g and λ_i are construct scalars, and the optimization problem is degenerated into an unconstructed extremum problem. The optimization problem is finally converted to an analytical solution as

$$\sum_{i=1}^C (\mathbf{X}_i^T (\mathbf{X}_i \mathbf{X}_i^T)^{-1} \mathbf{X}_i) \mathbf{p}_g = \lambda_g \mathbf{p}_g \quad (4)$$

Therefore, the sub-basis vector in each dataset can be calculated by

$$\mathbf{p}_i = \mathbf{X}_i^T \mathbf{a}_i = \frac{1}{\sqrt{\lambda_i}} \mathbf{X}_i^T (\mathbf{X}_i \mathbf{X}_i^T)^{-1} \mathbf{X}_i \mathbf{p}_g \quad (5)$$

Where $\lambda_i = \mathbf{p}_g^T \mathbf{X}_i^T (\mathbf{X}_i \mathbf{X}_i^T)^{-1} \mathbf{X}_i \mathbf{p}_g$ and eigenvector corresponding to the maximum eigenvalue is the first common basis vector \mathbf{p}_g . In the final, two-step basis vector extraction calculation comes into being. In the first step, the problem is formulated as

$$\max R^2 = \max \sum_{i=1}^C (\bar{\mathbf{p}}_i^T \mathbf{X}_i^T \bar{\mathbf{a}}_i)^2, \text{s.t.} \begin{cases} \bar{\mathbf{p}}_i^T \bar{\mathbf{p}}_i = 1 \\ \bar{\mathbf{a}}_i^T \bar{\mathbf{a}}_i = 1 \end{cases} \quad (6)$$

The problem is simplified by the Lagrangian operator as

$$\sum_{i=1}^C (\mathbf{X}_i^T \mathbf{X}_i) \bar{\mathbf{p}}_i = \bar{\lambda}_i \bar{\mathbf{p}}_i \quad (7)$$

Correspondingly, the sub-basis vector can be calculated by

$$\bar{\mathbf{p}}_i = \mathbf{X}_i^T \bar{\mathbf{a}}_i = \frac{1}{\sqrt{\bar{\lambda}_i}} \mathbf{X}_i^T \mathbf{X}_i \bar{\mathbf{p}}_i \quad (8)$$

where $\bar{\lambda}_i = \bar{\mathbf{p}}_i^T \mathbf{X}_i^T \mathbf{X}_i \bar{\mathbf{p}}_i$ a new subspace $\bar{\mathbf{P}}_i^T = [\bar{\mathbf{p}}_{i,1}, \bar{\mathbf{p}}_{i,2}, \dots, \bar{\mathbf{p}}_{i,\bar{R}}]^T \in R^{\bar{R} \times J}$ is spanned.

In the second step, the common basis vector extraction turns into a simple analytical solution as

$$\sum_{i=1}^C \left(X_i^T X_i \bar{P}_g \left(\bar{P}_g^T X_i^T X_i X_i^T X_i \bar{P}_g \right)^{-1} \left(X_i^T X_i \bar{P}_g \right)^T \right) P_g = \bar{\lambda}_g P_g \quad (9)$$

The common subspace $P_g = [p_{g,1}, p_{g,2}, \dots, p_{g,R^c}] \in R^{J \times R^c}$ is spanned of multiple p_g , which characterizes all the common data information over sets.

3. The Offline Extraction Modelling and Online Artificial Neural Networks Operating Performance Assessment

With the expert's experience and process knowledge, the training data can be easily divided into C species corresponding to different performance grades. According to the idea of just in time learning, similar input will produce similar output (Cheng et al., 2005). In this paper, industrial field input data is organized as training datasets, which represent implied information of different performance grades.

3.1 The Offline Extraction Modelling on Training Data

Assume that the chemical production process contains a total of C steady performance grades, and the training data corresponding to grade i th is denoted as $i=1,2,\dots,C$. X_i should be normalized into zero mean with unit variance. After the data pre-processing, X_i is divided into two orthogonal subspaces as

$$X_i = X_i^c + X_i^s = X_i P_g P_g^T + X_i (I - P_g P_g^T) \quad (10)$$

Where X_i^c includes the process variations determined by the common variable correlation among the whole process, while X_i^s contains differentiate variable correlation specific to grade i . The common subspace is spanned by V basis vectors as $P_g = [p_{g,1}, p_{g,2}, \dots, p_{g,V}] \in R^{J \times V}$. It should further analyse the amplitudes of the variations on each of the basis vectors orderly, and then the basis vectors with approximately equal variations amplitudes are separated from P_g and constitute a subspace which contains the common variable correlation and amplitude. In vectors selection, the variations can be calculated in score matrix manner as $t_{i,v} = X_i p_{g,v}$, $i=1,2,\dots,C$. If the score matrix $\|t_{1,v}\|_\infty, \|t_{2,v}\|_\infty, \dots, \|t_{C,v}\|_\infty$ are approximately equal, it certifies that $p_{g,v}$ is of no use to identify the operating performance, otherwise the $p_{g,v}$ will be left behind as a column vector for operating performance identification. Therefore, an index is denoted as $\eta_i = \|t_{1,v}\|_\infty / \|t_{C,v}\|_\infty$, $i=1,2,\dots,C-1$ to complete the screening. If $\eta_1, \eta_2, \dots, \eta_{C-1}$ are all within $1 \pm \delta$, it indicates that $\|t_{1,v}\|_\infty, \|t_{2,v}\|_\infty, \dots, \|t_{C,v}\|_\infty$ are almost equal one by one. δ is a relaxation factor that can be selected by trial and adapted to the specific training data. On the contrary, the $p_{g,v}$ will be left behind while the corresponding η_i is out of range. As a result, the basis vectors in P_g can be divided into two new subspace \tilde{P}_g and \tilde{P}_g^* , with $\tilde{P}_g = [p_{g,1}, p_{g,2}, \dots, p_{g,\tilde{V}}] \in R^{J \times \tilde{V}}$ reflecting the common variable correlation and common amplitude while $\tilde{P}_g^* = [p_{g,1}, p_{g,2}, \dots, p_{g,\tilde{V}^*}] \in R^{J \times \tilde{V}^*}$ denoting the subspace in which the common variable correlation but unequal amplitudes present in different performance grades with $\tilde{V} = V - \tilde{V}^*$. Finally, the industrial training data is separated into three parts as follows:

$$X_i = X_i^c + X_i^s = X_i \tilde{P}_g \tilde{P}_g^T + X_i \tilde{P}_g^* \tilde{P}_g^{*T} + X_i (I - P_g P_g^T) = \tilde{X}_i^c + \tilde{X}_i^s \quad (11)$$

Where $\tilde{X}_i^c = X_i \tilde{P}_g \tilde{P}_g^T$, representing the optimality unrelated variations and unrelated to operating performance assessment, while $\tilde{X}_i^s = X_i \tilde{P}_g^* \tilde{P}_g^{*T} + X_i (I - P_g P_g^T)$ covering specific operating performance identification information to grade i .

3.2 Online Operating Performance Assessment Based on Artificial Neural Networks

In online operating performance assessment, it is far from characterizing the true operating performance with a single sampling. Liu et al. (2009) carried out an approach of data window to trace the time-varying dynamics of a chemical process. In this paper, test data analysis unit is conducted on a data window with width H whose value is determined by the actual situation of the chemical industrial process. The step size of the sliding window is determined by the test data object. Data-driven operating performance assessment takes the similarity of between the test data and historical training data to indicate the performance grade of online data. Because the simulation data is less perturbed and often lack of non-linear features. The Euclidean distance

similarity index based on the score matrix is feasible where the whole process data is relatively stable and the noise is relatively small (Qin et al., 2012). However, industrial field online data owns numerous input variables and the data may contain a certain degree of nonlinear characteristics. Therefore, artificial neural network classification is carried out to measure the similarity of the online test data to the training data extracted offline. The online operating performance assessment based on artificial neural networks is arranged as follows:

Step1: At moment k , construct online test data as

$$\mathbf{X}_{test,k} = [\mathbf{x}_{test}(k-H+1), \dots, \mathbf{x}_{test}(k)]^T \quad (12)$$

Step2: Assume that $\mathbf{X}_{test,k}$ belongs to one specific grade. There would be C normalized versions of $\mathbf{X}_{test,k}$. A

data window with width H is introduced. Denote the $\mathbf{X}_{test,k}$ and its mean vector as $\bar{\mathbf{x}}_{test,k}^i = \sum_{h=k-H+1}^k \mathbf{x}_{test}^i(h) / H$ respectively.

Step3: Remove the common information over sets from the online data as

$$\tilde{\mathbf{x}}_{test,k}^{i,s} = \bar{\mathbf{x}}_{test,k}^{iT} - \bar{\mathbf{x}}_{test,k}^{iT} \bar{\mathbf{P}}_g \bar{\mathbf{P}}_g^T \quad (13)$$

Where $\tilde{\mathbf{x}}_{test,k}^{i,s}$ indicates the specific variations of the online test data.

Step4: Apply the simplest single hidden layer feed-forward artificial neural network as a classifier, with $\tilde{\mathbf{x}}_{test,k}^{i,s}$ as the input data of the neural network. The number of input neurons is J and the number of output neurons is C , representing the number of input variables and the number of performance grades, respectively. Then, an index matrix will be defined as $\boldsymbol{\gamma}_{test,k}^j = [\gamma_{test,k}^1, \dots, \gamma_{test,k}^q]^T$, $q=1,2,\dots,C$ where $\gamma_{test,k}^q \in [0,1]$ is the label data of the neural network classifier, representing the similarity of the online test data to the grade q . The closer the value of $\gamma_{test,k}^q$ is to 1, the higher the belief that the online test unit belongs to the performance grade q .

Step5: An assessment threshold β ($0 < \beta < 1$) is introduced for distinguishing whether the performance runs in a specific performance grade or the conversion.

Case 1: If $\max_{1 \leq q \leq C} (\gamma_{test,k}^q) \geq \beta$, it indicates that the online test data unit is operating in grade q .

Case 2: If case1 is not satisfied, such as $0 \leq \max_{1 \leq q \leq C} (\gamma_{test,k}^q) < \beta$ and $\gamma_{test,k}^i > \gamma_{test,k}^p$, ($\gamma_{test,k}^p = \arg \max \{\gamma_{test,k}^j\}, i=1,2,\dots, C \& i \neq q$), it can be determined that the current process is converting from grade q to grade p .

Case3: If both case 1 and 2 are not satisfied, influence of noise or uncertainties may answer the door and the assessment result is still.

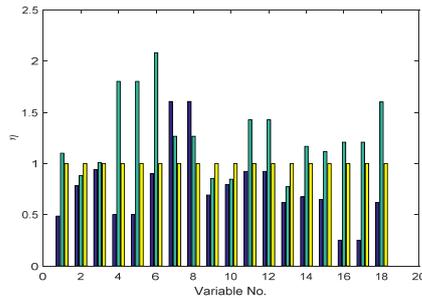


Figure 1: Amplitudes of score matrix corresponding to the variables

4. Case study

4.1 Description of Depth Optimization Process of Ethylene Cracking Furnace

Ethylene is a cornerstone in petrochemical industry, addressing many implications in process safety management in view of sustainable production (Fabiano et al., 2015). In the ethylene production process, the cracking furnace is the core of the production. Its operating performance not only affects the entire ethylene production, but also the downstream process. In this paper, actual process data is collected from a cracking furnace to a domestic ethylene industry. After preheating in the convection section, cracked raw materials flow into the radiation section of the furnace tube for cracking reaction through Venturi flow distributor. Next, the heated high temperature pyrolysis gas flows into the quench heat exchanger, and boiler water heat transfer for a quickly cool and end to the cracking reaction. The pyrolysis gas of the quench heat exchanger enters the

quench tank and is cooled by the quench oil to about 200 °C. Finally, the cracker gas is separated into the gasoline by the pyrolysis gas. The ethylene cracking process improves the assessment accuracy by maintaining the stability of the input variables to maximize production efficiency. In this study, 18 process variables are taken into the actual process for operating performance assessment. The operating performance of the multivariable system is evaluated by the high absorption rate, which acts as a single output object. According to expert knowledge, the ethylene production process can be divided into three steady performance grades, i.e. poor, general, and optimal. The training datasets is divided into three categories as X_1 , X_2 and X_3 , corresponding to poor, general, and optimal grade historically. There are 1,200 samples in total of historical process. Set the relaxation factor as $\delta = 0.1$.

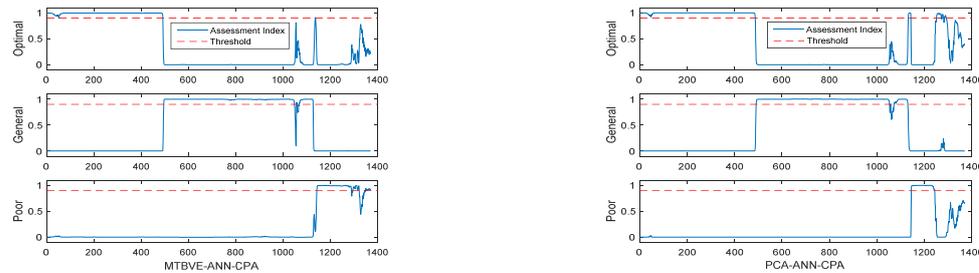


Figure 2: A comparison curve of the assessment results

4.2 Accuracy Analysis and discussion

To clearly illustrate the superiority of the assessment method proposed in this study, the PCA-ANN operating performance assessment method is illustrated for comparison. Assessment results of the two methods are shown as Figure 2. Relevant parameters are set as $H = 30, \beta = 0.9$. Because the optimal operating performance grade is relatively stable, the two methods are not very different. In the general grade, the assessment of the performance is different due to the data disturbance. After the sampling time 1143, the PCA-ANN method has two serious state misjudgments and the whole process assessment coefficient is jittery while the MTBVE-ANN method still owns a relatively good effect in the poor performance grade, and there is no miscarriage of assessment, indicating that MTBVE-ANN method is superior to the traditional feature extraction assessment method. MTBVE-ANN method of operating performance degradation is not only better than traditional methods, but also can be found in the advantage of selecting the relevant input variables to performance degradation, so as to provide a basis for the adjustment of control strategies. According to the index in Figure 2, the process variables, such as F_{feed} , F_{DS} , F_{ss} , T_{ss} , C_{OLE} are screened out, indicating that these variables have less responsibility for performance degradation.

Table 1: Comparison of several operating performance assessment methods

	Actual Grade	PCA-ANN	MTBVE-ANN
Optimal	1-490	1-484	1-491
Conversion	491-502	485-501	492-505
General	503-1,055	502-1,127	506-1,071
Conversion	1,056-1,062	1,128-1,147	1,072-1,142
Poor	1,063-1,371	1,148-1,234	1,143-1,371

Table 2: Variables definitions

Variables	Descriptions	Unit	Variables	Descriptions	Unit
ρ_{NAP}	cracking raw material density	$\text{kg} \cdot \text{m}^{-3}$	F_{Stuel}	side fuel flow	$\text{m}^3 \cdot \text{h}^{-1}$
C_{NP}	concentration of normal paraffin	%	C_{O2}	concentration of flue oxygen	%
C_{IP}	concentration of iso-paraffin	%	T_{g1}	flue gas temperature 1	°C
C_{OLE}	concentration of olefin	%	T_{g2}	flue gas temperature 2	°C
C_{BTX}	concentration of arene	%	F_{ss}	high-pressure steam flow	$\text{kg} \cdot \text{h}^{-1}$
C_{NAP}	concentration of naphthenic	%	T_{ss}	high-pressure steam temperature	°C
F_{feed}	feed flow	$\text{t} \cdot \text{h}^{-1}$	COT	cracking outlet temperature	°C
F_{DS}	dilution steam flow	$\text{t} \cdot \text{h}^{-1}$	T_{HK}	initial boiling point temperature	°C
F_{Bfuel}	bottom fuel flow	$\text{m}^3 \cdot \text{h}^{-1}$	T_{KK}	final boiling point temperature	°C

5. Conclusion

A novel online operating performance assessment method based on MTBVE-ANN is carried out for industrial processes. The advantages of the MTBVE-ANN assessment method are summarized as follows: (a) Common and specific variations over online steady performance grades are measured respectively, which highlights the distinction of different performance grade. (b) Online assessment with artificial neural network is adaptive to the nonlinear characteristics of the data somehow. (c) The proposed operating performance assessment method is applied to the real industrial production data, which is more credible than the simple simulation data. However, the method still owns some shortcomings, such as the assessment method of the transitional grade targeting, which is the spot for further research.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (61590922, 61533003, 21376077, 61673268), the Key Project of NSFC (61533012), the Shanghai Natural Science Foundation (14ZR1421800), the State Key Laboratory of Synthetical Automation for Process Industries (PAL-N201404).

References

- Cheng C., Chiu M.S., 2005, Nonlinear process monitoring using JITL-PCA, *Chemometrics and Intelligent Laboratory Systems*, 76(1), 1-13.
- Fabiano B., Pistrutto F., Reverberi A., Palazzi E., 2015, Ethylene-air mixtures under flowing conditions: a model-based approach to explosion conditions, *Clean Technologies and Environmental Policy*, 17(5), 1261-1270.
- Guo L., Gao J., Yang J., Kang L., 2009, Criticality evaluation of petrochemical equipment based on fuzzy comprehensive evaluation and a BP neural network, *Journal of Loss Prevention in the Process Industries*, 22(4), 469-476.
- Harris T.J., 1989, Assessment of control loop performance, *The Canadian Journal of Chemical Engineering*, 67(5), 856-861.
- Hwang J.H., Roh M.I., Lee K.Y., 2013, Determination of the optimal operating conditions of the dual mixed refrigerant cycle for the LNG FPSO topside liquefaction process, *Computers & Chemical Eng.*, 49, 25-36.
- Huang B., Shah S.L., 1998, Practical issues in multivariable feedback control performance assessment, *Journal of Process Control*, 8(5-6), 421-430.
- Krämer N., Sugiyama M., Braun M.L., 2009, Lanczos Approximations for the Speedup of Kernel Partial Least Squares Regression, *AISTATS*, 288-295.
- Li G., Qin S.J., Zhou D., 2010, Output relevant fault reconstruction and fault subspace extraction in total projection to latent structures models, *Industrial & Engineering Chemistry Research*, 49(19), 9175-9183.
- Liu Y., Hu N., Wang H., Li P., 2009, Soft chemical analyzer development using adaptive least-squares support vector regression with selective pruning and variable moving window size, *Industrial & Engineering Chemistry Research*, 48(12), 5731-5741.
- Liu Y., Chang Y., Wang F., 2014, Online process operating performance assessment and nonoptimal cause identification for industrial processes, *Journal of Process Control*, 24(10), 1548-1555.
- Peng K., Zhang K., Li G., 2013, Quality-related process monitoring based on total kernel PLS model and its industrial application, *Mathematical Problems in Engineering*, 1-14.
- Qin S.J., 2012, Survey on data-driven industrial process monitoring and diagnosis, *Annual Reviews in Control*, 36(2), 220-234.
- Tarafder A., Rangaiah G.P., Ray A.K., 2007, A study of finding many desirable solutions in multiobjective optimization of chemical processes, *Computers & Chemical Engineering*, 31(10), 1257-1271.
- Yu J., Qin S.J., 2008, Statistical MIMO controller performance monitoring. Part I: Data-driven covariance benchmark, *Journal of Process Control*, 18(3), 277-296.
- Yu J., Qin S.J., 2008, Statistical MIMO controller performance monitoring. Part II: Performance diagnosis, *Journal of Process Control*, 18(3), 297-319.
- Xu Y., Yang J.Y., Jin Z., 2004, A novel method for Fisher discriminant analysis, *Pattern Recognition*, 37(2), 381-384.
- Zhao C., Gao F., Niu D., Wang F., 2011, A two-step basis vector extraction strategy for multiset variable correlation analysis, *Chemometrics and Intelligent Laboratory Systems*, 107(1), 147-154.
- Zhou D., Li G., Qin S.J., 2010, Total projection to latent structures for process monitoring, *AIChE Journal*, 56(1), 168-178.