Preprints of the 10th IFAC International Symposium on Dynamics and Control of Process Systems
The International Federation of Automatic Control
December 18-20, 2013. Mumbai, India

Correntropy-based kernel learning for nonlinear system identification with unknown noise: an industrial case study

Yi Liu*, **Junghui Chen****

* *Engineering Research Center of Process Equipment and Remanufacturing, Ministry of Education, Institute of Process Equipment and Control Engineering, Zhejiang University of Technology, Hangzhou 310014, Zhejiang, P. R. China (Tel: +86-571-8887-1060; e-mail: yliuzju@zjut.edu.cn).*

** *R&D Center for Membrane Technology, Department of Chemical Engineering, Chung-Yuan Christian University, Chung-Li, Taiwan, 320, Republic of China (Tel.: +886-3-2654107; Fax: +886-3-2654199; e-mail: jason@wavenet.cycu.edu.tw).*

Abstract: One significant challenge in nonlinear system identification development for industrial processes is that the modeling samples often contain outliers and unknown noise. In this paper, a novel Correntropy-based Kernel Learning (CKL) method is proposed for identification of nonlinear systems with such uncertainty. Without resort to unnecessary efforts, the CKL identification method can reduce the effects of outliers by the use of a robust nonlinear estimator that maximizes correntropy. The superiority of the proposed CKL method is demonstrated through identification of an industrial process in Taiwan. The benefit of its more accurate and reliable performance indicates that CKL is promising in practice for identification of nonlinear systems with unknown noise.

1. INTRODUCTION

In the past two decades, considerable interest in both the theory and practice for nonlinear system identification has arisen. The popular nonlinear system identification methods include neural networks, support vector machines (SVM), fuzzy systems and other data/rule-based empirical methods (Ljung *et al.*, 2011; Söderström, 2012). Among them, SVM, least-squares SVM (LS-SVM) and a number of kernel learning (KL) modeling methods have found increasing reports recently. Generally, the structure determination of SVM can be implemented in a straightforward manner. Furthermore, SVM and other KL methods can obtain relatively good identification performance when the training data are insufficient. This characteristic is attractive in practice (Liu *et al*., 2010, 2012; Schölkopf and Smola, 2002; Suykens *et al*., 2002a, b).

As for industrial processes, one significant challenge is that the identification samples often contain different kinds of outliers and noise. Outliers are observations which appear to deviate markedly from the typical ranges of other observations. The presence of outliers in the variables affects the quality and reliability of the data, which can result in erroneous interpretations concerning the output variable of interest (Chiang *et al.*, 2003; Khatibisepehr and Huang, 2008; Liu *et al.*, 2004; Pearson, 2002; Söderström, 2012). Without awareness, learning samples with outliers may lead to biased parameter estimation and the overfitting problem because the identification model is corrupted by fitting those unwanted data. Therefore, it becomes very important to remove the effect of outliers.

This work was supported by Ministry of Economic Affairs, Taiwan, R.O.C., for the grant of the Technology Development Program for Academia project and National Science Council, R.O.C.. Corresponding author: Junghui Chen.

There are many outlier detection methods shown to be able to detect obvious outliers (Chiang *et al.*, 2003; Khatibisepehr and Huang, 2008; Liu *et al.*, 2004; Pearson, 2002). However, this issue is currently solved in a rather ad hoc manner, which leads to unnecessarily high costs (Kadlec *et al.*, 2009). Furthermore, none of these methods could detect all the inconspicuous outliers as they are masked by their adjacent outliers. In practice, it is very likely that the refined data set still contains some outliers after an outlier detection method is performed. From a practical viewpoint, it should be more attractive to develop general methods directly from existing nonlinear identification models without resort to unnecessary efforts.

Despite of the good nonlinear modeling ability, traditional SVM-based identification methods are not robust for outliers. Recent studies have shown that improved performance can be obtained using weighted SVM methods (Chuang *et al*., 2002; Suykens *et al*., 2002; Wen *et al*., 2008). However, most of these methods are heuristic as they use some user-defined parameters. They might be difficult to implement for nonlinear identification problems with uncertainty. In this paper, correntropy (Liu *et al*., 2007) is introduced into the area of nonlinear system identification. As a novel statistical measure, correntropy can deal with non-Gaussian noise and impulsive noise (Liu *et al*., 2007). However, to our best knowledge, little work has been reported on the application of correntropy in the field of nonlinear system identification. To this end, a correntropy kernel learning (CKL) method is proposed in this work. The CKL method can reduce the effects of outliers by the use of a robust nonlinear estimator that maximizes correntropy.

The remainder of this paper is structured as follows. Correntropy and the maximum correntropy (MC) criterion are introduced in Section 2. The CKL identification method

for nonlinear systems is proposed in Section 3. In Section 4, the method is evaluated in an industrial process in Taiwan. Comparison studies with other methods are also investigated. Finally, concluding remarks are made in Section 5.

2. MAXIMUM CORRENTROPY CRITERION

2.1 Correntropy as a Novel Similarity

Generally, the concept of correntropy is a generalized similarity measure between two arbitrary random variables *W* and *Y* with the same dimensions, defined by (Liu *et al*., 2007):

$$V(W,Y)=E\left[\kappa(W,Y)\right]=\int\kappa(w,y)dF_{WY}(w,y) \tag{1}$$

where V is correntropy; $E[\cdot]$ denotes the mathematic expectation; $F_{WY}(w,y)$ denotes the joint distribution function of (W,Y); and $\kappa(\cdot,\cdot)$ is a shift-invariant Mercer kernel (Liu *et al*., 2007; Príncipe, 2010). The most popular kernel used in correntropy is the Gaussian kernel with its kernel width $\sigma>0$, as defined below.

$$\kappa_\sigma(w,y)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{\|w-y\|^2}{2\sigma^2}\right) \tag{2}$$

When the joint distribution of *W* and *Y* is unknown and only a finite N number of samples $\{(w_i,y_i)\}_{i=1}^N$ are given, the correntropy estimator of samples $\hat{V}_N(W,Y)$ can be defined and calculated as follows (Liu *et al*., 2007):

$$\begin{aligned}\hat{V}_N(W,Y)&=\frac{1}{N}\sum_{i=1}^N\kappa_\sigma(w_i,y_i)\\&=\frac{1}{\sigma\sqrt{2\pi}}\frac{1}{N}\sum_{i=1}^N\exp\left[-\frac{(w_i-y_i)^2}{2\sigma^2}\right]\end{aligned} \tag{3}$$

Intuitively, correntropy is closely related to the similarity between *W* and *Y*. That is, if *W* is similar to *Y,* then the difference between *W* and *Y* should have a large value of correntropy (Liu *et al*., 2007).

2.2 Maximum Correntropy Criterion for Model Estimation

Now the concept of correntropy can be extended for the model estimation issue. The variable *W* can be considered as a mathematical expression of the unknown function $f(\mathbf{X};\boldsymbol{\theta})$ with an input set $\mathbf{X}=\{\mathbf{x}_i\in R^m\}_{i=1,\cdots,N}$ and the model parameters $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_M]^T$, which approximates the dependence of an output set $\mathbf{Y}=\{y_i\in R\}_{i=1,\cdots,N}$. As a new measure, correntropy can be used to describe how well $f(\mathbf{X};\boldsymbol{\theta})$ fits the data set **Y**. Consequently, the maximum of correntropy of the difference between $f(\mathbf{X};\boldsymbol{\theta})$ and **Y** is called the maximum correntropy (MC) criterion for model estimation (Liu *et al*., 2007). That is, for a modeling set of $\mathbf{S}=\{\mathbf{X},\mathbf{Y}\}$, the MC criterion can be formulated as

$$\begin{aligned}\hat{\boldsymbol{\theta}}_{\max\text{ correntropy}}&=\arg\max_{\boldsymbol{\theta}}\ \frac{1}{N}\sum_{i=1}^N\kappa_\sigma\left(f(\mathbf{x}_i;\boldsymbol{\theta}),y_i\right)\\&=\arg\max_{\boldsymbol{\theta}}\ \frac{1}{N}\sum_{i=1}^N\kappa_\sigma(e_i)\end{aligned} \tag{4}$$

where the difference is the fitting error, i.e., $e_i=f(\mathbf{x}_i;\boldsymbol{\theta})-y_i, i=1,\cdots,N$, produced by the model during supervised learning. Note that the following properties always exist in Eq. (4),

$$\begin{cases}\lim\limits_{e_i\to0}\dfrac{1}{N}\sum\limits_{i=1}^N\kappa_\sigma(e_i)=1\\ \lim\limits_{|e_i|\to\infty}\dfrac{1}{N}\sum\limits_{i=1}^N\kappa_\sigma(e_i)=0\\ 0\le\dfrac{1}{N}\sum\limits_{i=1}^N\kappa_\sigma(e_i)\le1\end{cases} \tag{5}$$

This means a larger value of correntropy can lead to a smaller fitting error of the model, and vice versa. And the value of correntropy is in the range of [0, 1]. In the above Eq. (4), $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_M]^T$ is a set of *M* adjustable model parameters to get the maximum correntropy. Consequently, as for model identification, this general estimation is also equivalent to an optimization problem as follows:

$$\hat{\boldsymbol{\theta}}_{\max\text{ correntropy}}=\arg\min_{\boldsymbol{\theta}}\ \frac{1}{\sigma\sqrt{2\pi}}\sum_{i=1}^N\left[1-\exp\left(-\frac{e_i^2}{2\sigma^2}\right)\right] \tag{6}$$

Then, Eq. (6) is differentiated with respect to $\boldsymbol{\theta}=[\theta_1,\cdots,\theta_M]^T$. The derivatives are set to zero and a system of *M* equations can be obtained,

$$\sum_{i=1}^N e_i\rho(e_i)\frac{\partial e_i}{\partial\theta_j}=0, j=1,\cdots,M \tag{7}$$

where $\rho(e_i)=\dfrac{\exp\left(-\dfrac{e_i^2}{2\sigma^2}\right)}{\sigma^3\sqrt{2\pi}}, i=1,\cdots,N$ can be regarded as the weighted terms (Liu *et al*., 2007). The kernel width σ plays an important role in the smoothing process. As recently proved by Chen and Príncipe (2012), the MC estimation is essentially used to smooth and maximize posteriori estimation. Some approaches can be utilized to determine the kernel width σ for $\rho(e_i)$ (Liu *et al*., 2007; Munoz and Chen, 2012; Príncipe, 2010). Here, the kernel width can be simply computed as (Munoz and Chen, 2012)

$$\sigma=\frac{\max|e_i|}{2\sqrt{2}}, i=1,\cdots,N \tag{8}$$

Therefore, the optimal estimation problem in Eq. (6) is equivalent to a weighted least squares problem as:

$$\hat{\boldsymbol{\theta}}_{\max\ \text{correntropy}} = \arg\ \min_{\boldsymbol{\theta}}\ \sum_{i=1}^{N}\rho\left(e_i\right)e_i^2 \tag{9}$$

This problem was first proposed for signal processing by Liu *et al*. (2007) (See Eq. (50) in Liu *et al*. (2007)). The weighted terms $\rho\left(e_i\right)$ mean that large errors get larger attenuation, so the estimation is resistant to outliers (Liu *et al*., 2007; Príncipe, 2010). Recently, Munoz and Chen (2012) applied MC-based wavelet modeling method to fitting batch data according to the time, that is, $e_i = f\left(t_i;\boldsymbol{\theta}\right) - y_i, i = 1,\cdots,N$, where t_i denotes the *i*th time instance. Compared with the traditional criteria adopted for data-driven process modeling, such as the well-known minimum MSE, the MC criterion has several advantages: (1) it is always bounded for any distribution; (2) it contains all even-order moments and is useful for nonlinear and non-Gaussian signal processing; (3) it is a local similarity measure and is robust to outlier samples (Chen and Príncipe, 2012; Liu *et al*., 2007; Príncipe, 2010).

3. CORRENTROPY KERNEL LEARNING (CKL) FOR NONLINEAR SYSTEM IDENTIFICATION

The main objective of this work is to develop a simple and general KL framework for robust identification of nonlinear processes. The central idea of the proposed CKL method is to integrate the MC criterion and KL into a unified framework. Without resort to unnecessary efforts, the effects of outliers can be reduced once the CKL model is obtained.

For simplicity, consider single-input−single-output (SISO) nonlinear systems using the nonlinear autoregressive with exogenous input (NARX) form governed by the following relationship (Ljung *et al.*, 2011; Söderström, 2012):

$$\begin{aligned} y_i &= f\left(y_{i-1},\cdots,y_{i-n_y},u_{i-1},\cdots,u_{i-n_u}\right)+e_i \\ &= f\left(\mathbf{x}_i;\boldsymbol{\theta}\right)+e_i,\quad i=1,\cdots,N \end{aligned} \tag{13}$$

where $f(\cdot)$ is the wanted nonlinear model; i is the time instance; y_i, u_i, and e_i are the system output, the system input and the noise vector at instance i (n_y and n_u are the corresponding lags of the output and the input), respectively. Correspondingly, $\mathbf{x}_i$ is a general input vector that is usually composed of y_i and u_i combined with their corresponding delayed forms at time i.

A general form of the kernelized nonlinear model for process modeling can be formulated as (Liu *et al.*, 2010, 2012):

$$\begin{aligned} y_i &= f\left(\mathbf{x}_i;\boldsymbol{\theta}\right)+e_i = f\left(\mathbf{x}_i;\mathbf{w},b\right)+e_i \\ &= \mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right)+b+e_i,\quad i=1,\cdots,N \end{aligned} \tag{14}$$

where $\boldsymbol{\phi}$ is the mapping and $\boldsymbol{\theta}=\left[\mathbf{w}^T,b\right]^T$; that is, the symbols $\mathbf{w}$ and b are the model parameter vector and the bias term, respectively. When the MC criterion (Liu *et al*., 2007) and the KL framework with regularization (Schölkopf and Smola, 2002; Suykens *et al*., 2002a) are applied to Eq. (14), the proposed method seeks the nonlinear identification model by solving the following optimization problem:

$$\begin{cases} \min\ J\left(\mathbf{w},b,\rho\right)=\dfrac{\gamma}{2}\displaystyle\sum_{i=1}^{N}\rho\left(e_i\right)e_i^2+\dfrac{1}{2}\left\|\mathbf{w}\right\|^2 \\ \text{s.t. } y_i-\mathbf{w}^T\boldsymbol{\phi}\left(\mathbf{x}_i\right)-b-e_i=0,\ i=1,\cdots,N \end{cases} \tag{15}$$

where the user-defined regularization parameter γ ($\gamma > 0$) determines the trade-off between the model's complexity and approximation accuracy. Here, the same regularization term $\left\|\mathbf{w}\right\|^2$ in LS-SVM (Suykens *et al*., 2002a) is adopted because it is used to further compare Eq. (15) with LS-SVM later in this paper. With a similar introduction of the MC criterion (Liu *et al*., 2007), formulation of other optimization problems (Liu *et al.*, 2010, 2012; Schölkopf and Smola, 2002) can also be straightforward.

The above problem cannot be solved directly because the weighted terms $\rho\left(e_i\right)$ depend on the model coefficients $\boldsymbol{\theta}=\left[\mathbf{w}^T,b\right]^T$. Here, a two-level iterative procedure is suggested as follows. In the first level, the weighted terms $\rho\left(e_i\right)$ can be fixed, which indicates that a weighted KL problem is formulated. In the second level, the weighted terms $\rho\left(e_i\right)$ can be updated using the obtained model coefficients $\boldsymbol{\theta}=\left[\mathbf{w}^T,b\right]^T$. A detailed training algorithm of the proposed CKL method is described below.

Level 1. Initialization and update of the CKL model

First, the weighted terms $\rho\left(e_i\right)$ is set to be fixed but not considered as the function of e_i. The initial value of $\rho\left(e_i\right)$ can be set as 1; i.e., $\rho\left(e_i\right)=1, i=1,\cdots,N$, which means all training samples are weighted equally. To solve the optimization problem, the Lagrangian method can be constructed below:

$$\begin{aligned} L = &\left(\left\|\mathbf{w}\right\|^2+\gamma\sum_{i=1}^{N}\rho\left(e_i\right)e_i^2\right)\Big/2+ \\ &\sum_{i=1}^{N}\alpha_i\left[y_i-\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)-b-e_i\right], i=1,\cdots,N \end{aligned} \tag{16}$$

where $\boldsymbol{\alpha}=\left[\alpha_1,\cdots,\alpha_N\right]^T$ are Lagrange multipliers. The optimality conditions are shown as follows:

$$\begin{cases} \dfrac{\partial L}{\partial \mathbf{w}}=0\rightarrow\mathbf{w}=\displaystyle\sum_{i=1}^{N}\alpha_i\boldsymbol{\phi}\left(\mathbf{x}_i\right) \\ \dfrac{\partial L}{\partial e_i}=0\rightarrow\alpha_i=\gamma\rho\left(e_i\right)e_i, \qquad i=1,\cdots,N \\ \dfrac{\partial L}{\partial \alpha_i}=0\rightarrow y_i-\mathbf{w}^T\boldsymbol{\phi}(\mathbf{x}_i)-b-e_i=0, i=1,\cdots,N \\ \dfrac{\partial L}{\partial b}=0\rightarrow\displaystyle\sum_{i=1}^{N}\alpha_i=0 \end{cases} \tag{17}$$

After elimination of the variables, $\mathbf{w}$ and e_i, the following solution can be obtained:

$$\begin{bmatrix} \mathbf{K}+\boldsymbol{\Omega} & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ b \end{bmatrix} = \begin{bmatrix} \mathbf{y} \\ 0 \end{bmatrix} \tag{18}$$

where $\mathbf{y} = [y_1, \cdots, y_N]^T$ and $\mathbf{1} \in R^{N\times 1}$ is a vector of ones; $\mathbf{K} \in R^{N\times N}$ is a kernel matrix whose element is denoted as $K_{ij} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_j) \rangle, \forall i, j = 1, \cdots, N$ using the kernel trick (Schölkopf and Smola, 2002; Suykens *et al*., 2002a); $\boldsymbol{\Omega}$ is a diagonal matrix whose diagonal element $\Omega_i = \frac{1}{\gamma \rho(e_i)}, i = 1, \cdots, N$. Note that $\mathbf{K}+\boldsymbol{\Omega}$ is symmetric positive-definite, so it is invertible. For simplicity, the quantity is defined as $\mathbf{H} = \mathbf{K}+\boldsymbol{\Omega}$ and then its inverse can be computed as

$$\mathbf{P} = \mathbf{H}^{-1} = (\mathbf{K}+\boldsymbol{\Omega})^{-1} \tag{19}$$

As shown in Eq. (17), the solution of model coefficients $\boldsymbol{\theta} = [\mathbf{w}^T, b]^T$ can be transformed to $\boldsymbol{\theta} = [\boldsymbol{\alpha}^T, b]^T$ and then can be expressed as

$$\begin{cases} \boldsymbol{\alpha} = \mathbf{P}\left[\mathbf{y} - \dfrac{\mathbf{1}\mathbf{1}^T\mathbf{P}\mathbf{y}}{\mathbf{1}^T\mathbf{y}\mathbf{1}}\right] \\ b = \dfrac{\mathbf{1}^T\mathbf{P}\mathbf{y}}{\mathbf{1}^T\mathbf{y}\mathbf{1}} \end{cases} \tag{20}$$

Level 2. Iterative weighting

Correspondingly, the predicted values and a new set of values for the weighted terms can be obtained, respectively

$$\begin{aligned} e_i &= y_i - \hat{y}_i \\ &= y_i - \langle \mathbf{w}, \phi(\mathbf{x}_i) \rangle - b \\ &= y_i - \sum_{j=1}^{N} \alpha_j \langle \phi(\mathbf{x}_j), \phi(\mathbf{x}_i) \rangle - b \\ &= y_i - \sum_{j=1}^{N} \alpha_j k_{ij} - b \\ &= y_i - \boldsymbol{\alpha}^T \mathbf{k}_i - b, \quad i = 1, \cdots, N \end{aligned} \tag{21}$$

$$\rho(e_i) = \frac{\exp\left(-\dfrac{e_i^2}{2\sigma^2}\right)}{\sigma^3\sqrt{2\pi}}, i = 1, \cdots, N \tag{22}$$

where $\mathbf{k}_i = [k_{i1}, \cdots, k_{iN}]^T \in R^{N\times 1}$ is a kernel vector.

After the weighted terms $\rho(e_i)$ are updated, $\boldsymbol{\Omega}$ can also be updated with its diagonal element $\Omega_i = \frac{1}{\gamma \rho(e_i)}, i = 1, \cdots, N$. Then, the new values of the coefficient $\boldsymbol{\theta} = [\boldsymbol{\alpha}^T, b]^T$ can be obtained using Eqs. (19) and (20). The iterative procedure can be implemented until the weighted terms $\rho(e_i)$ are almost unchanged. Thus, the predicted model is

$$\hat{y}_t = f(\mathbf{w}, b; \mathbf{x}_t) = \sum_{i=1}^{N} \alpha_i \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_t) \rangle + b = \boldsymbol{\alpha}^T \mathbf{k}_t + b \tag{23}$$

where $\mathbf{k}_t = [k_{t1}, \cdots, k_{tN}]^T \in R^{N\times 1}$ is a kernel vector with its element $k_{ti} = \langle \phi(\mathbf{x}_i), \phi(\mathbf{x}_t) \rangle, \forall i = 1, \cdots, N$.

After a CKL identification model is trained, the outlier samples can be detected simultaneously, because the outlier samples have relatively small weights $\rho(e_i)$, which can be shown in Eq. (22). Due to the advantage, the outliers can be removed by this post-identified method (Munoz and Chen, 2012). Although these outliers are kept in the identification model, they cannot affect the performance of the model because of their small weights. Therefore, despite the outlier samples, a robust CKL identification model can be obtained.

Generally, traditional identification methods, e.g., LS-SVM, are sensitive to outliers because they are based on the mean squared error (MSE) loss function which is only optimal when the underlying noises obey Gaussian distribution (Suykens *et al*., 2002a; Wen *et al*., 2008). Different from the MSE criterion, the non-Gaussian noise and outliers in the process can be suitably treated by the MC criterion (Chen and Príncipe, 2012; Liu *et al*., 2007; Príncipe, 2010). Additionally, the nonlinear relationship between the process input and output can be identified using the kernel trick and the regularization technique. Consequently, as expected, the proposed CKL method can achieve better identification performance for industrial systems because the quality of modeling samples is not always good.

In essence, the proposed CKL method can be considered as a nonlinear robust estimator. At the first glance, CKL is somewhat similar to the weighted LS-SVM approaches (Suykens *et al*., 2002b; Wen *et al*., 2008). However, most of the traditional weighted methods adopt different heuristic weighting strategies to reduce the effect of outlier samples. Actually, it is difficult to check whether these weighting schemes are suitable to the complicated industrial data set beforehand. Because of the correntropy-based weighting strategy, CKL has one main advantage in its adaptive scheme for more general identification problems with outliers and unknown noise rather than heuristic schemes for special problems.

4. AN INDUSTRIAL CASE IN TAIWAN

In this section, an industrial example in Taiwan is explored to validate the effect of the CKL identification method. A simplified flowchart of the process is shown in Fig. 1. The main purpose of this reboiler is to produce pure ethylene chloride. Based on the fundamental knowledge and the past experience in this production line, one important control loop for the final product is manipulation of the liquid level of NC-103B as shown in Fig. 1. Advanced control strategies can be designed once a suitable identification model for this loop

is obtained. Therefore, the proposed CKL identification method is validated using industrial data in Dec., 2011.

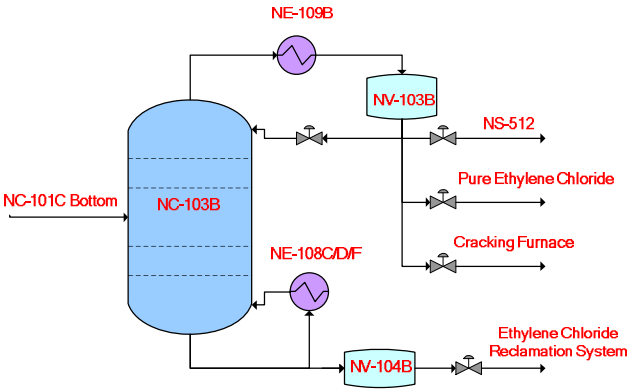

Fig. 1. A simplified flowchart of the reboiler in the production of ethylene chloride

In this process, the sampling time is 30 seconds. All of the training and test data are not pre-processing. Samples collected in the first 100 minutes are for training and in the rest 100 minutes are for testing, respectively. Without any prior knowledge, the general input vector of the CKL model consists of u_k and y_k, as well as their one step delayed terms, respectively, i.e., $\mathbf{x}_k = [y_k, y_{k-1}, u_k, u_{k-1}]^T$. The cross-validation approach is adopted to train the model.

Here, to show the characteristics of CKL, it is compared with a well-known weighted LS-SVM (WLS-SVM) method proposed by Suykens et al. (2002b).

$$\rho(e_i) = \begin{cases} 1, & \text{if } |e_i/\hat{s}| \le c_1 \\ \dfrac{c_2 - |e_i/\hat{s}|}{c_2 - c_1}, & \text{if } c_1 < |e_i/\hat{s}| \le c_2 \\ 10^{-4}, & \text{otherwise} \end{cases} \tag{24}$$

where c_1 and c_2 are user-defined parameters; $\hat{s}$ is a robust estimate of the standard deviation of the LS-SVM error variables e_i. In the estimate of $\hat{s}$, one takes into account how much the estimated error distribution deviates from a Gaussian distribution (Suykens *et al.*, 2002b). However, the WLS-SVM method is heuristic but not general. Additionally, it is not in an adaptive weighting manner because several user-defined parameters in Eq. (24) should be determined. Those parameters cannot be used straightforward for many practical problems.

The training results of CKL and WLS-SVM identification models and their weighted terms $\rho(e_i)$ can be shown in Fig. 2. The fitting results of CKL and WLS-SVM are almost the same. The weighted terms of CKL are continuous. The outliers tend to have smaller weights and thus they show less effect on the model, although they are kept in the model. They can be adaptively determined by their training errors. However, the weights of WLS-SVM are discontinuous and most of them are equal to 1. This indicates that the heuristic weighting strategy of WLS-SVM (Suykens *et al.*, 2002b) is not suitable for many practical problems.

The prediction results for the test samples of all the CKL, WLS-SVM and LS-SVM identification methods are shown in Fig. 3. As shown in Figs. 2-3, although the fitting results of CKL and WLS-SVM approaches are almost the same for training, the prediction results on the test set are different. Generally, the CKL method can achieve better prediction performance than the other two methods because more samples have relatively small prediction errors.

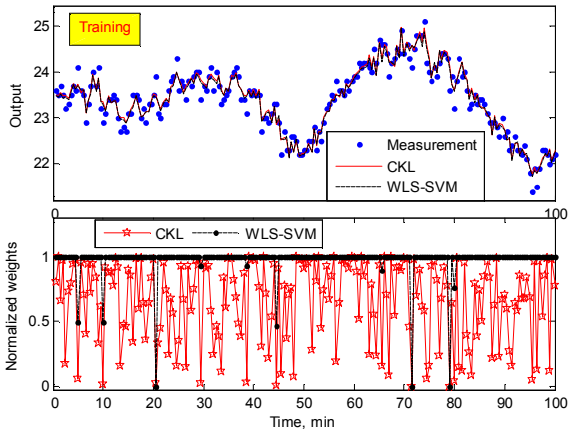

Fig. 2. The training results of CKL and WLS-SVM identification models and their weighted terms $\rho(e_i)$.

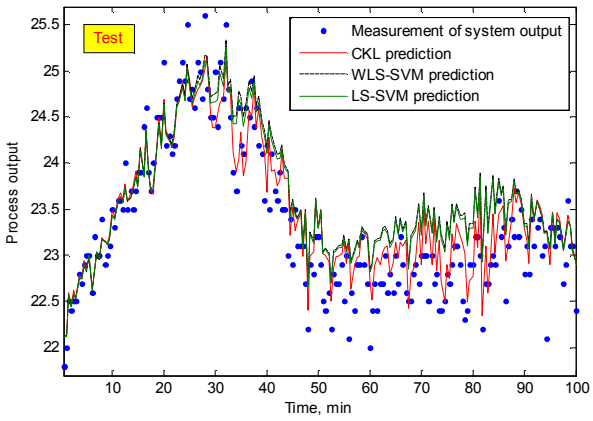

Fig. 3. The validation results of CKL, WLS-SVM and LS-SVM identification methods for the industrial process

The RMSE index ($\text{RMSE} = \sqrt{\frac{1}{N_{\text{tst}}}\sum_{i=1}^{N_{\text{tst}}} e_i^2}$, N_{tst} is the number of test samples) is adopted to evaluate the performance quantitatively. The obtained identified results of all the CKL, WLS-SVM and LS-SVM methods with different model orders are also summarized in Table 1. As for nonlinear system identification, there is still relatively little work on determining the proper model orders. In Table 1, whatever model orders are chosen, the CKL method achieves superior performance to WLS-SVM and LS-SVM. And as for this case, the suitable model orders can be roughly determined as $\mathbf{x}_k = [y_k, y_{k-1}, u_k, u_{k-1}]^T$. If the model orders are not enough, i.e., $\mathbf{x}_k = [y_k, u_k]^T$, CKL is only a little better than

WLS-SVM and LS-SVM. If the model orders are a little larger, e.g., $\mathbf{x}_k = [y_k, y_{k-1}, y_{k-2}, u_k, u_{k-1}, u_{k-2}]^T$, CKL can still obtain much more than WLS-SVM and LS-SVM because error-in-variables can be accumulated with more orders. Therefore, from the obtained results in Table 1, as for offline identification, the model orders can be simply determined using the CKL method.

Table 1. The performance evaluation of CKL, WLS-SVM and LS-SVM identification methods for the industrial case

General input vector (Model orders)	Method	RMSE
$\mathbf{x}_k = [y_k, u_k]^T$	CKL	**0.42**
	WLS-SVM	0.43
	LS-SVM	0.43
$\mathbf{x}_k = [y_k, u_k, u_{k-1}]^T$	CKL	**0.42**
	WLS-SVM	0.53
	LS-SVM	0.52
$\mathbf{x}_k = [y_k, y_{k-1}, u_k]^T$	CKL	**0.39**
	WLS-SVM	0.46
	LS-SVM	0.44
$\mathbf{x}_k = [y_k, y_{k-1}, u_k, u_{k-1}]^T$	CKL	**0.34**
	WLS-SVM	0.45
	LS-SVM	0.43
$\mathbf{x}_k = [y_k, y_{k-1}, y_{k-2}, u_k, u_{k-1}, u_{k-2}]^T$	CKL	**0.39**
	WLS-SVM	0.50
	LS-SVM	0.48

In summary, as explored in this paper, the proposed CKL method is more general and efficient than LS-SVM and WLS-SVM methods for identification of nonlinear systems with unknown noise. In general, this method can be extended to other KL models, such as state-dependent models, nonlinear autoregressive and moving average models, subspace models, Hammerstein/Wiener models, and so forth. Additionally, as for online identification, the recursive form of CKL model can also be used straightforward. This is because the recursive formulation of **P** in Eq. 19 by increasing/decreasing samples can be implemented in a similar way with an existing recursive KL method proposed by Liu et al (2010).

4. CONCLUSION

The concept of correntropy is introduced in the area of nonlinear system identification. Without much effort to outlier detection, the proposed CKL method can be utilized for identification of nonlinear systems with outliers and unknown noise. The main appealing properties, such as the structural risk minimization principle, the kernel technique, the convex optimization problem, and a few free parameters to be adjusted, are still preserved in this identification framework. The superiority of the proposed CKL method, in terms of more accurate and reliable performance, has been validated through an industrial process. Some interesting future studies have also been highlighted.

REFERENCES

Chen B.D., and Príncipe J.C. (2012). Maximum correntropy estimation is a smoothed MAP estimation. *IEEE Signal Proc. Let.*, **55**(19): 491-494.

Chiang L.H., Pell R.J., Seasholtz M.B. (2003). Exploring process data with the use of robust outlier detection algorithms. *J. Process Control*, **13**(5): 437-449.

Chuang C.C., Su S.F., Jeng J.T., Hsiao C.C. (2002). Robust support vector regression networks for function approximation with outliers. *IEEE Trans. Neural Netw.*, **13**(6): 1322-1330.

Kadlec P., Gabrys B., Strandt S. (2009) Data-driven soft sensors in the process industry. *Comput. Chem. Eng.*, **33**(4): 795-814.

Khatibisepehr S., and Huang B. (2008). Dealing with irregular data in soft sensors: Bayesian method and comparative study. *Ind. Eng. Chem. Res.*, **47**(22): 8713-8723.

Liu H., Shah S.L., Jiang W. (2004). Online outlier detection and data cleaning. *Comput. Chem. Eng.*, **28**(9): 1635-1647.

Liu W.F., Pokharel P.P., Príncipe J.C. (2007). Correntropy: properties and applications in non-Gaussian signal processing. *IEEE Trans. Signal Proc.*, **55**(11): 5286-5298.

Liu Y., Wang H.Q., Yu J., Li P. (2010). Selective recursive kernel learning for online identification of nonlinear systems with NARX form. *J. Process Control*, **20**(2): 181-194.

Liu Y., Gao Z.L., Li P., Wang H.Q. (2012). Just-in-time kernel learning with adaptive parameter selection for soft sensor modeling of batch processes. *Ind. Eng. Chem. Res.*, **51**(11): 4313-4327.

Ljung L., Hjalmarsson H., Ohlsson H. (2011). Four encounters with system identification. *Eur. J. Control*, **17**(5-6): 449-471.

Munoz J.C., and Chen J.H. (2012). Removal of the effects of outliers in batch process data through maximum correntropy estimator. *Chemom. Intell. Lab. Syst.*, **111**: 53-58.

Pearson R.K. (2002). Outliers in process modeling and identification. *IEEE Trans. Control Syst. Technol.*, **10**(1): 55-63.

Príncipe J.C. (2010). *Information Theoretic Learning: Renyi's Entropy and Kernel Perspectives*. Springer Verlag.

Schölkopf B. and Smola A.J. (2002). *Learning with Kernels*. Cambridge, MA: MIT Press.

Söderström T. (2012). System identification for the errors-in-variables problem. *Trans. Inst. Meas. Control*, **34**(7): 780-792.

Suykens J.A.K., Van Gestel T., De Brabanter J., De Moor B., Vandewalle J. (2002a). *Least Squares Support Vector Machines*. Singapore: World Scientific.

Suykens J.A.K., De Brabanter J., Lukas L., Vandewalle J., (2002b). Weighted least squares support vector machines: robustness and sparse approximation. *Neurocomputing*, **48**(1-4): 85-105.

Wen W., Hao Z.F., Yang X.W. (2008). A heuristic weight-setting strategy and iteratively updating algorithm for weighted least-squares support vector regression. *Neurocomputing*, **71**(16-18): 3096-3103.