Chemometrics and Intelligent Laboratory Systems 135 (2014) 76–89

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemolab

New contributions to non-linear process monitoring through kernel partial least squares

José L. Godoy [a,*], David A. Zumoffen [b,1], Jorge R. Vega [a,2], Jacinto L. Marchetti [a]

[a] Institute of Technological Development for the Chemical Industry (INTEC-CONICET-UNL), Güemes 3450, 3000 Santa Fe, Argentina
[b] French-Argentine International Center for Information and Systems Sciences (CIFASIS-CONICET-UNR-AMU), Rosario, Santa Fe, Argentina

ARTICLE INFO

Article history:
Received 13 September 2013
Received in revised form 31 March 2014
Accepted 1 April 2014
Available online 12 April 2014

Keywords:
KPLS modeling
Fault detection
Fault diagnosis
Prediction risk assessment
Non-linear processes

ABSTRACT

The kernel partial least squares (KPLS) method was originally focused on soft-sensor calibration for predicting online quality attributes. In this work, an analysis of the KPLS-based modeling technique and its application to non-linear process monitoring are presented. To this effect, the measurement decomposition, the development of new specific statistics acting on non-overlapped domains, and the contribution analysis are addressed for purposes of fault detection, diagnosis, and prediction risk assessment. Some practical insights for synthesizing the models are also given, which are related to an appropriate order selection and the adoption of the kernel function parameter. A proper combination of scaled statistics allows the definition of an efficient detection index for monitoring a non-linear process. The effectiveness of the proposed methods is confirmed by using simulation examples.

1. Introduction

The design of monitoring systems for supervising the operation of industrial processes has acquired great relevance in the last decade. This fact is essentially due to the need of more demanding operating conditions related to security for equipments and personnel, operating costs, and environmental restrictions. Furthermore, the increasing complexity observed in the interactions between energy — and mass — transfer processes, and their corresponding control policies, require more sophisticated monitoring systems in aspects such as detection rate, robustness, user friendliness, easiness of understanding, modeling and data storage requirements, and adaptability, among others [1,2].

The multivariate statistical process monitoring is a well-known research topic where several strategies based on projection to latent structures have successfully been developed. Moreover, they are of great interest in industrial applications because of their excellent properties for handling noisy and highly correlated measurements, and large data sets [2–4]. Some of these approaches are summarized in [4–10] where the principal component analysis (PCA), independent component analysis (ICA), and partial least squares regression (PLSR) methodologies were addressed. There are also several modifications to these tools for including issues such as dynamics, adaptation, and non-linearity [2,8,11–16].

In this work, a non-linear version of the partial least squares (PLS) approach — called kernel PLS (KPLS) — is addressed. KPLS is a powerful statistical tool for obtaining multivariate non-linear relationships from historical data. In fact, it is a non-linear regression method that computes the regression coefficients in a high-dimensional space; the input data are mapped via non-linear functions in this space and then they are linearly related to the measured outputs. Hence, the KPLS approach represents a suitable methodology for predicting online unmeasured quality variables in complex non-linear processes. The overall procedure relies on classic linear algebra, similar to the linear projection methods, and the non-linearity degree is mainly given by the selected kernel function associated to the mapping functions [17]. Ever since the KPLS approach appeared, some modifications as well as applications have been published in the process monitoring area. For example, a kernel-based PLS system linked to orthogonal signal correction has been proposed for data preprocessing and prediction purposes [12]; and a modified PLS method of independent component regression has been used for complex processes monitoring [8]. An application of non-linear multivariate quality prediction based on KPLS has also been presented [14]. In this context, new publications addressing the fault

* Corresponding author. Also with Universidad Tecnológica Nacional — FRP, Paraná, Argentina. Tel.: +54 342 4559174; fax: +54 342 4550944.
E-mail address: jlgodoy@santafe-conicet.gov.ar (J.L. Godoy).
[1] Also with Universidad Tecnológica Nacional — FRRo, Rosario, Argentina.
[2] Also with Universidad Tecnológica Nacional — FRSF, Santa Fe, Argentina

http://dx.doi.org/10.1016/j.chemolab.2014.04.001
0169-7439/

detection tasks based on KPLS have also appeared [18,19]. In the last decade, KPLS or variants thereof have been applied for composition analysis of agricultural materials [20] and foods [21], process analysis [22], determination of structure–activity relationships [23], studies on drug metabolism [24], and quality-related monitoring [25], among others.

KPLS method, as well as other kernel based modeling methods [26], is often used as a black box approach. However, in contrast to kernel PCA (KPCA) [11,16], KPLS is able to properly determine the predictive importance of each input variable onto the final regression model. This result can then be used for reducing the number of inputs and therefore the complexity of the model. For instance, Postma et al. [27] propose a method based on the principle of pseudo-sampled trajectories (representing the original variables) that help visualize and determine the most important variables for regression purposes. This method is able to detect poor predictor variables, providing the chance for improving the KPLS structure by eliminating interfering variables from the pre-selected inputs. The advantage of the KPLS modeling lays in the fact that only the outputs of interest are chosen, while the inputs are determined by their predictive importance, thus limiting the group of variables to be monitored.

The main objective of this article is to provide a deep analysis of the KPLS-based modeling technique and its application to non-linear process monitoring. Initially, the classic KPLS modeling is here extended by adding the projections of the outputs onto the latent space. The underlying structure of the KPLS modeling is highlighted in order to describe the functional relationships between the spaces induced by the KPLS procedure. Moreover, some practical insights are given for the proper selection of the number of latent variables and for setting the kernel function parameter. In fact, the latent space dimension is here defined by using a new balanced index designed to efficiently quantify the squared prediction error in both the input and output spaces. This approach is compared with the standard output prediction error via the Wold's R criterion [7,14]. To deal with non-linear processes, the kernel method is first embedded into the PLS algorithm. Then, new specific statistics (that act on non-overlapped domains) are combined into a single index able to detect process anomalies. Finally, the statistics pattern is used for diagnosing faults or process anomalies. In this regard, the present monitoring technique of non-linear processes is an extension of our PLS-based strategy originally developed for monitoring linear processes [28]. Besides, contribution plots are frequently used to isolate the detected faulty variables without using historical fault patterns [26,29]. However, it is difficult to build a contribution plot for a kernel based model [29]. In this paper, a new contribution plot based on the KPLS model is proposed for identifying faulty variables. The proposed supervision approach puts together the abnormal event detection, the diagnosis, and the isolation in a single method. Besides, a risk assessment index is also developed for online quantification of the predictive capabilities of the KPLS inferential model. The effectiveness of the proposed method is tested through simulated examples taken from the literature.

The article is organized as follows: Section 2 presents the basic background of the KPLS regression. Some details about the KPLS-based modeling approach are given in Section 3. The main contributions of this work are presented in Sections 4 and 5, where we analyze the KPLS model calibration (Section 4), the process monitoring and the statistics for fault detection (Section 5.1), the diagnosis method through the pattern of statistics (Section 5.2), the fault isolation via a contribution analysis (Section 5.3), and the prediction risk assessment (Section 5.4). Section 6 summarizes the simulation results and the overall conclusions are given in Section 7.

2. Basic concepts on KPLS

Consider a process with m measured input variables plus p measured output variables which are arranged in the vectors $\mathbf{x} = [x_1 \dots x_m]'$ and $\mathbf{y} = [y_1 \dots y_p]'$, respectively. Assume that N measurements of each variable are collected while the process is operating under normal conditions. In order to build a KPLS regression model, let us consider the calibration data set consisting of N centered and scaled samples for the input vector (predictor), i.e., $\{\mathbf{x}_j \in \mathbb{R}^m\}_{j=1}^N$, and the corresponding centered and scaled samples for the response vector, $\{\mathbf{y}_j \in \mathbb{R}^p\}_{j=1}^N$.

The key idea of the KPLS approach is to map the input data $\mathbf{x}_j \in \mathbb{R}^m$ to a high-dimensional space $\mathbb{R}^c$ that corresponds to a reproducing kernel Hilbert space, where the non-linear structure in the input space is more likely to be linear, and thus a linear PLSR can be applied [17]. The non-linear mapping is not implemented through an explicit function, $\varphi(\cdot) : \mathbb{R}^m \to \mathbb{R}^c$, instead a kernel function $k(\cdot,\cdot)$ is proposed for computing the following inner products,

$$k\left(\mathbf{x}_j, \mathbf{x}_r\right) = \varphi\left(\mathbf{x}_j\right)' \varphi(\mathbf{x}_r), \quad \text{with} \quad j = 1, \dots, N \quad r = 1, \dots, N. \tag{1}$$

Thus, by replacing each inner product $\varphi(\mathbf{x}_j)'\varphi(\mathbf{x}_r)$ with $k(\mathbf{x}_j, \mathbf{x}_r)$, both the explicit non-linear mapping and the inner product computation can be avoided [17]. The kernel function $k(\cdot,\cdot)$ cannot arbitrarily be selected, but it must satisfy the Mercer's theorem conditions [17]. A specific choice of the kernel function implicitly determines the associated mapping $\varphi(\cdot)$ and the space $\mathbb{R}^c$. Note that the dimension c may be arbitrarily large and can even be infinite.

The KPLS approach only uses the inner product values for performing the non-linear regression. From Eq. (1) the so-called Gram kernel matrix, $\mathbf{K} \in \mathbb{R}^{N \times N}$, can be obtained:

$$\mathbf{K} = \Phi\, \Phi', \quad \text{with} \quad \Phi = [\varphi(\mathbf{x}_1), \dots, \varphi(\mathbf{x}_N)]' \in \mathbb{R}^{N \times c}. \tag{2}$$

Similar to PLSR, the non-linear KPLS model includes zero-mean variables. The mapped input vectors $\varphi(\mathbf{x}_j)$ are centered as follows:

$$\overline{\varphi}\left(\mathbf{x}_j\right) = \varphi\left(\mathbf{x}_j\right) - \Phi' \mathbf{e} \tag{3}$$

where e is a column vector with all its entries equal to $1/N$ [17]. In this way, $\overline{\Phi} = [\overline{\varphi}(\mathbf{x}_1), \dots, \overline{\varphi}(\mathbf{x}_N)]'$ is the centered version of Φ. Now the centered Gram kernel matrix is given by

$$\overline{\mathbf{K}} = \overline{\Phi}\, \overline{\Phi}' = (\mathbf{I} - \mathbf{E})\mathbf{K}(\mathbf{I} - \mathbf{E}) \tag{4}$$

where $\mathbf{E}$ is a $(N \times N)$ matrix with all its entries equal to $1/N$ [17] and $\overline{k}\left(\mathbf{x}_j, \mathbf{x}_r\right) = \overline{\varphi}\left(\mathbf{x}_j\right)' \overline{\varphi}(\mathbf{x}_r)$ is the element (j,r) of $\overline{\mathbf{K}}$.

From the centered data matrices $\overline{\mathbf{K}}$ and $\mathbf{Y} = [\mathbf{y}_1, \dots, \mathbf{y}_N]'$, a KPLS calibration algorithm can be developed by modifying the steps of the NIPALS algorithm [17] as shown in Algorithm 1. Specific details about the parameter setting for the kernel function and the optimal selection of the number of latent variables, A, are given in Section 4.

The prediction of the response variables by using the calibration data is given by [17]:

$$\hat{\mathbf{Y}} = \overline{\Phi}\mathbf{B}_{\mathrm{PLS}} = \overline{\Phi}\,\overline{\Phi}'\mathbf{U}(\mathbf{T}'\overline{\mathbf{K}}\mathbf{U})^{-1}\mathbf{T}'\mathbf{Y} = \overline{\mathbf{K}}\mathbf{U}(\mathbf{T}'\overline{\mathbf{K}}\mathbf{U})^{-1}\mathbf{T}'\mathbf{Y} = \mathbf{T}\mathbf{T}'\mathbf{Y} = \mathbf{T}\mathbf{C}' \tag{5}$$

where the matrices $\mathbf{T} = [\mathbf{t}_1, \dots, \mathbf{t}_A]$ and $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_A]$ are orthonormal by columns. Note that, although the regression coefficients matrix $\mathbf{B}_{\mathrm{PLS}}$ might exist (for $\overline{\varphi}(.) \in \mathbb{R}^c$ when $c \neq \infty$), the KPLS algorithm does not calculate these values explicitly, i.e. the kernel substitution avoids this evaluation.

Eq. (5) shows that the response variables (outputs) can be obtained from the inner products of the mapped vectors. Hence, for a new observation $\mathbf{x}$ of the predictor vector, the outputs are estimated by

$$\hat{\mathbf{y}} = \mathbf{B}'_{\mathrm{PLS}}\overline{\varphi}(\mathbf{x}) = \mathbf{Y}'\mathbf{T}\left[\mathbf{U}(\mathbf{T}'\overline{\mathbf{K}}\mathbf{U})^{-1}\right]'\overline{\mathbf{k}}(\mathbf{x}) = \mathbf{C}\mathbf{V}'\overline{\mathbf{k}}(\mathbf{x}) \tag{6}$$