Computers and Chemical Engineering 77 (2015) 74–84

Contents lists available at ScienceDirect

Computers and Chemical Engineering

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

Deconstructing principal component analysis using a data reconciliation perspective

Shankar Narasimhan*, Nirav Bhatt

Systems & Control Group, Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600036, India

ARTICLE INFO

Article history:
Received 2 October 2014
Received in revised form 10 March 2015
Accepted 24 March 2015
Available online 1 April 2015

Keywords:
Data reconciliation
Principal component analysis
Model identification
Estimation
Denoising

ABSTRACT

Data reconciliation (DR) and principal component analysis (PCA) are two popular data analysis techniques in process industries. Data reconciliation is used to obtain accurate and consistent estimates of variables and parameters from erroneous measurements. PCA is primarily used as a method for reducing the dimensionality of high dimensional data and as a preprocessing technique for denoising measurements. These techniques have been developed and deployed independently of each other. The primary purpose of this article is to elucidate the close relationship between these two seemingly disparate techniques. This leads to a unified framework for applying PCA and DR. Further, we show how the two techniques can be deployed together in a collaborative and consistent manner to process data. The framework has been extended to deal with partially measured systems and to incorporate partial knowledge available about the process model.

1. Introduction

Data reconciliation (DR) is a technique that was proposed in the early 1950s to derive accurate and consistent estimates of process variables and parameters from noisy measurements. This technique has been refined and developed over the past 50 years. Several books and book chapters have been written on this and related techniques (Romagnoli and Sanchez, 1999; Veverka and Madron, 1997; Narasimhan and Jordache, 2000; Hodouin, 2010; Bagajewicz, 2001). The technique is now an integral part of simulation software packages such as ASPEN PLUS®. Several standalone software packages for data reconciliation such as VALI, DATACON®, are also available and deployed in chemical and mineral process industries. The main benefit derived from applying DR is accurate estimates of all process variables and parameters which satisfy the process constraints such as material and energy balances. The derived estimates are typically used in retrofitting, optimization and control applications. In order to apply DR, the following information is required.

(i) The constraints that have to be obeyed by the process variables and parameters must be defined. These constraints are usually derived from first principles model using process knowledge, and consist of material and energy conservation equations including property correlations, and can also include equipment design equations, and thermodynamic constraints.
(ii) The set of process variables that are measured must be specified. Additionally, inaccuracies in these measurements must be specified in terms of the variances and covariances of errors. This information is usually derived from sensor manuals or from historical data.

Another multivariate data processing technique that has become very popular in recent years is principal component analysis (PCA) (Jolliffe, 2002). This method is primarily used for reducing the dimensionality of data and to denoise them. It is also used in developing regression models, when there is collinearity in the regressors variables (Davis et al., 1999). In chemical engineering, it has been used for process monitoring and fault detection, and diagnosis (Kourti and MacGregor, 1995; Yoon and MacGregor, 2001). Generally, PCA has been regarded as a data-driven multivariate statistical technique. In a recent paper, PCA was interpreted as a model identification technique that discovers the linear relationships between process variables (Narasimhan and Shah, 2008). This interpretation of PCA is not well known, although other authors have previously alluded to it.

The purpose of this article is to establish the close connection between PCA and DR. Specifically, it is shown that PCA is a technique that discovers the underlying linear relationships between process

* Corresponding author. Tel.: +91 4422574165; fax: +91 4422574152.
E-mail addresses: naras@iitm.ac.in (S. Narasimhan), niravbhatt@iitm.ac.in (N. Bhatt).

http://dx.doi.org/10.1016/j.compchemeng.2015.03.016
0098-1354/© 2015 Elsevier Ltd. All rights reserved.

variables while simultaneously reconciling the measurements with respect to the identified model. Exploring this connection further, it is shown that iterative PCA (IPCA) is a method which simultaneously extracts the linear process model, error-covariance matrix and reconciles the measurements (Narasimhan and Shah, 2008). Several benefits accrue from this interpretation:

(i) It shows that data reconciliation can be applied to a process purely using measured data, even if it is difficult to obtain a model and measurement error variances using *a priori* knowledge. It thus expands the applicability of data reconciliation and related techniques.
(ii) PCA and IPCA can be used as techniques for obtaining a process model and measurement error-covariance matrix from data. Since these are the two essential information required to apply DR, it is now possible to apply the rigorous and well developed companion technique such as gross error detection (GED) for fault diagnosis. This will eliminate the difficulties and deficiencies present in the current approach of using PCA for fault diagnosis.

Additional useful results presented in this paper include the interpretation of the process model obtained using PCA, when only a subset of the process variables is measured. Modification of the PCA and IPCA techniques to incorporate partial knowledge of some of the process constraints is also proposed. The impact of incorrectly estimating the model order (the actual number of linear constraints) on the reconciled estimates is also discussed, leading to a recommendation for practical application of PCA and combining it with tools of DR and GED.

The paper is organized as follows. Sections 2 and 3 introduce the background on DR and PCA, respectively. Model identification and data reconciliation using PCA for the case of known error-covariance matrix is described in Section 4. For unknown error-covariances case, Section 5 describes a procedure for simultaneous model identification, estimation of error-covariances, and data reconciliation using IPCA. Section 6 extends PCA (IPCA) to partially measured systems, and known constraint matrix. Further, it discusses selection criteria of model order when the model order is not known. Section 7 concludes the paper. The developed concepts are illustrated via a simulated flow process.

2. Basics of data reconciliation

In this section, the application of DR to linear steady-state processes is discussed, including the case when a subset of the process variables is measured (also known as partially measured systems).

2.1. Linear steady-state processes

The objective of data reconciliation is to obtain better estimates of process measurements by reducing the effect of random errors in measurements. For this purpose, the relationships between different variables as defined by process constraints are exploited. We restrict our attention to linearly constrained processes which are operating under steady state. An example of such a process is a water distribution network, or a steam distribution network with flows of different streams being measured. We first describe the data reconciliation methodology for the case when the flows of all streams are measured.

Let $\mathbf{x}(j) \in \mathbb{R}^n$ be an n-dimensional vector of the true values of the n process variables corresponding to a steady-state operating point for each sample j. The samples $\mathbf{x}(j), j = 1, 2, \ldots, N$ can be drawn from the same steady state or from different steady states. These variables are related by the following linear relationships[1]:

$$\mathbf{A}\mathbf{x}(j) = \mathbf{0}_{m\times 1} \tag{1}$$

where $\mathbf{A}$ is an $(m \times n)$-dimensional matrix, and $\mathbf{0}$ is an m-dimensional vector with elements being zero. In data reconciliation, $\mathbf{A}$ is labelled as a "*constraint matrix*". Note that the rows of $\mathbf{A}$ span an m-dimensional subspace of $\mathbb{R}^n$, while $\mathbf{x}(j)$ lies in an $(n - m)$-dimensional subspace (orthogonal to the row space of $\mathbf{A}$) of $\mathbb{R}^n$. Let $\mathbf{y}(j) \in \mathbb{R}^n$ be the measurements of the n variables. The measurements are usually corrupted by random errors. Hence, the measurement model can be written as follows:

$$\mathbf{y}(j) = \mathbf{x}(j) + \epsilon(j), \tag{2}$$

where $\epsilon(j)$ is an n-dimensional random error vector at sampling instant j. The following assumptions are made about the random errors:

$$\begin{aligned} &\text{(i)} \quad \boldsymbol{\epsilon}(j) \sim \mathcal{N}(\mathbf{0}, \Sigma_\epsilon) \\ &\text{(ii)} \quad E[\boldsymbol{\epsilon}(j)\boldsymbol{\epsilon}(k)^{\mathrm{T}}] = \mathbf{0}, \quad \forall \;\; j \neq k \\ &\text{(iii)} \quad E[\mathbf{x}(j)\boldsymbol{\epsilon}(j)^{\mathrm{T}}] = \mathbf{0} \end{aligned} \tag{3}$$

where $E[\cdot]$ denotes the expectation operator. If the error variance–covariance matrix $\mathbf{\Sigma}_\epsilon$ is known, then the reconciled estimates for $\mathbf{x}(j)$ (denoted as $\hat{\mathbf{x}}(j)$) can be obtained by minimizing the following objective function:

$$\begin{aligned} \min_{\mathbf{x}(j)} \quad & (\mathbf{y}(j) - \mathbf{x}(j))^{\mathrm{T}} \Sigma_\epsilon^{-1} (\mathbf{y}(j) - \mathbf{x}(j)) \\ \text{s.}t. \quad & \mathbf{A}\mathbf{x}(j) = \mathbf{0}. \end{aligned} \tag{4}$$

The reconciled values of the variables are given by:

$$\hat{\mathbf{x}}(j) = \mathbf{y}(j) - \Sigma_\epsilon \mathbf{A}^{\mathrm{T}} (\mathbf{A}\Sigma_\epsilon\mathbf{A}^{\mathrm{T}})^{-1} \mathbf{A}\mathbf{y}(j) = \mathbf{W}\mathbf{y}(j), \tag{5}$$

where $\mathbf{W} = \mathbf{I} - \Sigma_\epsilon \mathbf{A}^{\mathrm{T}} (\mathbf{A}\Sigma_\epsilon\mathbf{A}^{\mathrm{T}})^{-1} \mathbf{A}$. Under the assumptions made regarding the measurements errors, it can be shown that the reconciled estimates obtained using the above formulation are maximum likelihood estimates. It can also be verified that the estimates $\hat{\mathbf{x}}(j)$ satisfy the imposed constraints and are normally distributed with mean, $\mathbf{x}(j)$, and covariance, $\mathbf{W}\mathbf{\Sigma}_\epsilon\mathbf{W}^{\mathrm{T}}$.

If all the measured samples are drawn from the same steady state operating point, then DR can be applied to the average of the measured samples. However, if the samples are from different steady states, then DR is applied to each sample independently. For ease of comparison with PCA, we consider a set of N samples (which could correspond to different steady state operating periods) to which DR is applied. The set of N samples is arranged in the form of an $(n \times N)$-dimensional data matrix, $\mathbf{Y}$ as

$$\mathbf{Y} = [\mathbf{y}(1), \mathbf{y}(2), \ldots, \mathbf{y}(N)] = \mathbf{X} + \mathbf{E}, \tag{6}$$

where $\mathbf{X}$ and $\mathbf{E}$ are $(n \times N)$-dimensional matrices of the true values and the errors, respectively. The matrix $\hat{\mathbf{X}}$ of reconciled estimates for the N samples is given by

$$\hat{\mathbf{X}} = \mathbf{W}\mathbf{Y}. \tag{7}$$

The following example illustrates DR on the flow process shown in Fig. 1.

[1] For flow processes considered in this paper, the constraint model given by Eq. (1) is appropriate. In general if $\mathbf{A}\mathbf{x}(j) = \mathbf{b}$, then PCA and other related methods described in the paper can be used after subtracting the sample mean $\bar{\mathbf{y}}$ from the measurements. The estimate of $\mathbf{b}$ is given by $\hat{\mathbf{A}}\bar{\mathbf{y}}$.