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

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.

© 2015 Elsevier Ltd. All rights reserved.

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).

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:

(i) $\epsilon(j) \sim \mathcal{N}(\mathbf{0}, \Sigma_\epsilon)$

(ii) $E[\epsilon(j)\epsilon(k)^{\mathrm{T}}] = \mathbf{0}, \quad \forall \; j \neq k \tag{3}$

(iii) $E[\mathbf{x}(j)\epsilon(j)^{\mathrm{T}}] = \mathbf{0}$

where $E[\cdot]$ denotes the expectation operator. If the error variance–covariance matrix Σ_ϵ 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:

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

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, $\mathbf{W}\Sigma_\epsilon \mathbf{W}^{\mathrm{T}}$.

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 $\hat{\mathbf{A}}\bar{\mathbf{y}}$.
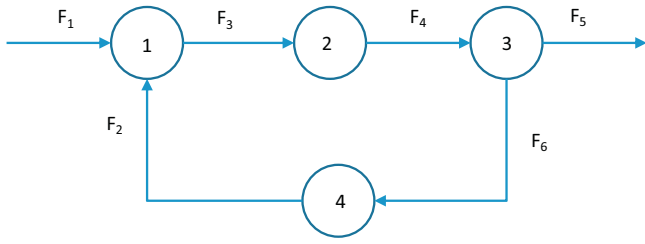
Fig. 1. Schematic of a flow process.

Example 1. The flow process consists of six flow variables and four balance equations, i.e., $n = 6$ and $m = 4$. The flow balance equations for this process can be written as follows:

$$\mathbf{AF} = \mathbf{0}, \text{ with} \tag{8}$$

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 1 & -1 & -1 \\ 0 & -1 & 0 & 0 & 0 & 1 \end{bmatrix}, \tag{9}$$

$$\mathbf{F} = \begin{bmatrix} F_1 & F_2 & F_3 & F_4 & F_5 & F_6 \end{bmatrix}^\mathrm{T}. \tag{10}$$

In order to demonstrate the utility of DR in reducing noise in measurements, we assume that flows of all six streams are measured and noisy measurements are simulated as follows. First noise-free values (true flow rates) that satisfy the constraints are generated, followed by addition of noise to generate the noisy measurements.

1. The flow variables F_1 and F_2 are chosen as independent variables and base values are specified for these variables. For simulating data corresponding to different steady states, normally distributed random fluctuations are added to the base values of flow variables F_1 and F_2.
2. The true values of the dependent flow variables, F_3, F_4, F_5, and F_6 are computed using four flow balance equations in Eq. (8) for all steady states.
3. Noisy measurements are simulated by adding normally distributed random errors to the true values corresponding to different steady states.

The base values, standard deviations of fluctuations for generating different steady states, and the standard deviations of measurement errors are given in Table 1.

A sample of 1000 measurements is simulated and DR is applied to obtain the reconciled estimates corresponding to each sample. The root mean square error (RMSE) values between reconciled and true values over all the 1000 samples for each flow variable is computed and reported in Table 2. The RMSE in the measurements of variables are also reported in the table for a comparison. The RMSE values in the estimates of all variables are less than the RMSE values in the corresponding measurement, clearly indicating that reconciled estimates of variables are more accurate than the measurements.

Table 1
Base values, standard deviation of fluctuation (SDF), and standard deviations of error (SDE) for the flow data.

Variable	Base values	SDF	SDE
F_1	10	1	0.1
F_2	10	2	0.08
F_3	Eq. (8)		0.15
F_4	Eq. (8)		0.2
F_5	Eq. (8)		0.18
F_6	Eq. (8)		0.1

Table 2
Root mean square errors in measurements and reconciled estimates of flow variables for Example 1.

Variable	RMSE before DR	RMSE after DR
F_1	3.2177	2.3318
F_2	2.5750	1.6669
F_3	4.8292	2.5180
F_4	6.5358	2.5180
F_5	5.8500	2.3318
F_6	2.7261	1.6669

2.2. Partially measured systems

For reasons of cost and feasibility, in most processes only a subset of variables is measured. We refer to such a system as a partially measured system. Application of DR to a partially measured system provides reconciled values of measured variables and estimates for unmeasured variables. DR also provides diagnostics on which of the measured variables can be reconciled (known as redundant measured variables) and which of the unmeasured variables can be uniquely estimated (also known as observable unmeasured variables). The details of the redundant and observable variables concepts can be found in Mah (1990). The method for applying DR to a partially measured system is described below.

The constraints for a partially measured system, Eq. (1), can be rewritten in terms of the measured variables (labelled as $\mathbf{x}_k$) and the unmeasured variables (labelled as $\mathbf{x}_u$) as:

$$\mathbf{A}_k \mathbf{x}_k + \mathbf{A}_u \mathbf{x}_u = \mathbf{0}, \tag{11}$$

where $\mathbf{A}_k$ and $\mathbf{A}_u$ are the partitions of $\mathbf{A}$ matrix corresponding to $\mathbf{x}_k$ and $\mathbf{x}_u$, respectively. Then, by constructing a projection matrix $\mathbf{P}$ such that $\mathbf{PA}_u = \mathbf{0}$, and multiplying Eq. (11), we can get a reduced set of constraints involving only the measured variables as (Crowe et al., 1983):

$$\mathbf{PA}_k \mathbf{x}_k = \mathbf{0}. \tag{12}$$

Then, the reduced data reconciliation problem is to find the minimum of the objective defined in Eq. (4) subject to the constraints defined by Eq. (12). The reconciled values $\hat{\mathbf{x}}_k(j)$ obtained from the measurement $\mathbf{y}_k(j)$ are given as follows:

$$\hat{\mathbf{x}}_k(j) = \mathbf{y}_k(j) - \Sigma_\epsilon (\mathbf{PA}_k)^\mathrm{T} (\mathbf{PA}_k \Sigma_\epsilon (\mathbf{PA}_k)^\mathrm{T})^{-1} (\mathbf{PA}_k) \mathbf{y}_k(j) \tag{13}$$

The estimates for the unmeasured variables can be obtained by substituting for the estimates of the measured variables in Eq. (11) and solving these equations. The conditions under which unique estimates for the unmeasured variables can be obtained and the identification of redundant and observable variables using linear algebraic techniques are more completely described in the book by Narasimhan and Jordache (2000). Alternatively, for flow processes, a graph-theoretic approach can be used for obtaining the reduced constraint matrix, and for determining the redundant measured variables and the observable unmeasured variables.

Example 2. The following example illustrates the application of DR to a partially measured flow process. The graph theoretical approach is used for observability and redundancy classification of variables, because it aids in visualization and ease of understanding. Consider again the flow process given in Fig. 1, with only the flows of streams 1, 2, and 5 being measured. In order to apply the graph theoretic procedure, the process is represented as a graph shown in Fig. 2(b), in which an environment node is added to which all process inflows (stream 1) and process outflows (stream 5) are connected. In order to obtain the graph corresponding to the reduced reconciliation problem, we merge nodes which are connected by streams whose flows are not measured. Thus, merging nodes 1 and 4, 2 and 3, 3 and 4, we get the reduced graph shown in Fig. 2(c). The
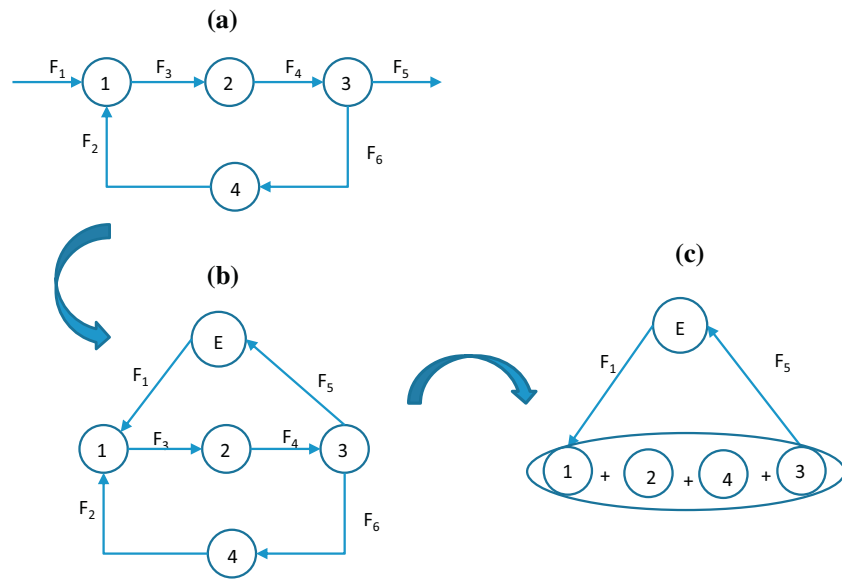
(a)

(b)

(c)

Fig. 2. Flow network after merging variables for Example 2.

reduced graph contains only streams 1 and 5 (whose flows are measured) for which one flow balance can be written (either around node 1 or node E). The reduced DR problem is to obtain reconciled estimates of these two flows subject to the flow balance constraint. The same reduced reconciliation problem can be obtained using the projection technique described earlier. The measurements of streams that appear in the reduced graph are redundant, while the measured flow of stream 2 that was eliminated during the merging process is non-redundant. It can also be deduced that the flows of streams 3, 4, and 6 are observable because the original process graph does not contain any cycle which consists solely of unmeasured flows.

The RMSE values in all the variables after DR are reported in Table 3. It shows that the reconciled estimates of the redundant variables F_1 and F_5 are more accurate than the measurements even with the partial measurements, while there is no improvement in the non-redundant flow F_2. Compared with the RMSE values of estimates reported in Table 2 for the fully measured case, the estimates of all variables obtained are less accurate due to reduced information available about the process.

3. Basics of principal component analysis

Principal component analysis (PCA) is one of the widely used multivariate statistical techniques. The traditional view of PCA as a technique for dimensionality reduction is explained for the sake of continuity, before we describe how PCA can be used as a technique for identifying a steady state linear model from the data.

PCA is a linear transformation of a set of variables to a new set of uncorrelated variables called principal components (PCs)

Table 3
Root mean square errors in the reconciled estimates of flow variables for Example 2.

Variable	RMSE after DR
F_1	2.7899
F_2	2.5750
F_3	3.9033
F_4	3.9033
F_5	2.7899
F_6	2.5750

(Jolliffe, 2002). The first PC is a new variable with the highest variance among all linear transformations of the original variables. The second PC is a new variable orthogonal to the first PC (and hence uncorrelated with it) that has the next highest variance among all linear transformations of the original variables and so on. For reducing the dimensionality of multivariate data, the PCs with the highest variances are retained while the remaining PCs are discarded. The PCs can be obtained from the eigenvectors of the covariance matrix of the data for the given set of variables. The eigenvectors of the covariance matrix of the data can in turn be obtained using the singular value decomposition (SVD) of the data matrix as described below.

As defined in the preceding section, let $\mathbf{Y}$ be the $(n \times N)$-dimensional data matrix. Let $\mathbf{S_y}$ be the covariance matrix of $\mathbf{Y}$ defined by

$$\mathbf{S_y} = \frac{1}{N}\mathbf{Y}\mathbf{Y}^{\mathrm{T}}. \tag{14}$$

The SVD of the scaled data matrix can be written as

$$svd\left(\frac{\mathbf{Y}}{\sqrt{N}}\right) = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^{\mathrm{T}} + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^{\mathrm{T}}, \tag{15}$$

where $\mathbf{U}_1$ are the orthonormal eigenvectors corresponding to the p largest eigenvalues of $\mathbf{S_y}$ while $\mathbf{U}_2$ are the orthonormal eigenvectors corresponding to the remaining $(n - p)$ smallest eigenvalues of $\mathbf{S_y}$. The matrices $\mathbf{S}_1$ and $\mathbf{S}_2$ are diagonal matrices, whose diagonal elements are the square root of the eigenvalues of $\mathbf{S_y}$. It is assumed that the eigenvalues and corresponding eigenvectors are ordered in decreasing order of the magnitudes of the eigenvalues. It can be proved that the first p PCs are given by $\mathbf{U}_1^{\mathrm{T}}\mathbf{y}(j)$, and the variances of these PCs are the corresponding eigenvalues.

If the objective is to reduce the dimensionality of the data, then the values of the first p PCs only need to be computed corresponding to each observation $\mathbf{y}(j)$ and stored. Several heuristics have been suggested in the literature to choose the number of PCs p to be retained, such as percentage of total variance in data captured, or the SCREE plot which looks for sharp changes in the eigenvalues. More details on such heuristics can be found in the book by Jolliffe (2002).

Alternatively, PCA can be used as a denoising technique. The denoised estimates of the measurements obtained using PCA corresponding to p retained PCs are given by

$$\hat{\mathbf{X}} = \sqrt{N}\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^{\mathrm{T}}. \tag{16}$$

4. Model identification and data reconciliation using PCA

The viewpoint that is of importance in this article is that PCA is a method for discovering the underlying linear relationships between variables, while at the same time obtaining the reconciled estimates of variables that satisfy the identified linear relations. The use of PCA as a method for identifying a linear model is less well known although it has been alluded to by different authors (Gertler et al., 1999; Huang, 2001; Jolliffe, 2002). This viewpoint is explored thoroughly in this work. We first analyze the case when all the process variables are measured and the measurement error-covariance matrix is known. The generalization to partially measured systems and estimation of error-covariance matrix simultaneously along with the linear constraint model from data will be elucidated in the following sections. We also make the following additional assumptions regarding the observations. The reasons for these additional assumptions will be explained later in this section.

(A1) The data are drawn from at least n distinct steady states.
(A2) The data used for model building do not contain any outliers or gross errors.

If there are m linear relations between the measured process variables as defined by Eq. (1), then as stated earlier, the true values of variables lie in an $\mathbb{R}^{n-m}$ dimensional subspace of $\mathbb{R}^n$. Therefore, the denoised estimates of variables are also chosen to lie in an $(n-m)$-dimensional subspace. If we assume that the number of linear constraints m relating the variables is known *a priori*, then we can apply PCA, choose the number of retained PCs p to be equal to $n-m$, and obtain the estimates $\hat{\mathbf{X}}$ given by Eq. (16). Based on the orthogonality of the PCs, it immediately follows that the eigenvectors corresponding to the smallest m eigenvalues are orthogonal to these estimates, and therefore they provide an estimate of the rows of the constraint matrix $\mathbf{A}$ (whose m rows lie in an R^m dimensional subspace orthogonal to the true values of process variables), that is, the estimate of the constraint matrix is given by

$$\hat{\mathbf{A}} = \mathbf{U}_2^{\mathrm{T}}. \tag{17}$$

While data compression and denoising focuses on the eigenvectors corresponding to the largest eigenvalues that need to be retained, model identification is concerned with the eigenvectors corresponding to the smallest eigenvalues. The main questions to be answered are whether the estimate of the constraint matrix obtained using PCA is optimal, and whether the number of constraints can also be estimated from data, without any prior knowledge.

The estimates of process variables (that lie in a p-dimensional subspace) obtained using Eq. (16) and the corresponding estimate of the constraint matrix (orthogonal to the estimates) obtained using Eq. (17), can be shown to be optimal in the least-squares sense (Rao, 1964). In other words, given a sample of observations, PCA identifies an optimal p-dimensional subspace in which the estimates lie such that the sum squared differences between measured and estimated values is minimum. If we assume that the errors in measurements obey the assumptions listed in Eq. (3) with the additional condition that the error-covariance matrix is a scale of the identity matrix, i.e., $\mathbf{\Sigma}_\epsilon = \sigma^2\mathbf{I}$, then the following result can be proved

$$\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N} = E[\mathbf{S_y}|\mathbf{X}, N] = \mathbf{M_x} + \sigma^2\mathbf{I}, \tag{18}$$

where $\mathbf{M_x} = (1/N)\mathbf{XX}^{\mathrm{T}}$ is an $(n \times n)$-dimensional matrix, and $\mathbf{I}$ is an n-dimensional identity matrix. Note that the notation $E[\mathbf{S_y}|\mathbf{X}, N]$ is used to emphasize that the expected value of the sample covariance matrix is considered under the assumption that the samples are drawn from the same set of steady states $\mathbf{X}$, and the only variability in $\mathbf{S_y}$ is due to the measurement errors. Eq. (18) leads to the following properties:

Property 1. *The eigenvectors of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$ and $\mathbf{M_x}$ are identical.*

Property 2. *The smallest m eigenvalues of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$ are all equal to σ^2.*

Assumption (A2) made regarding the observations implies that the observations do not contain any systematic error, and, therefore, the true values corresponding to each sample must satisfy the true process constraints. Furthermore, Assumption (A1) ensures that observations from different steady states are obtained. If data corresponding to $(n-m)$ linearly independent steady states are obtained, then the rank of the matrix $\mathbf{M_x}$ will be $(n-m)$ with the last m eigenvalues being equal to zero.[2] The eigenvectors corresponding to the m zero eigenvalues will be orthogonal to the true values and is a basis for the row space of the constraints. We can therefore use Property 1 to conclude that the eigenvectors corresponding to the m smallest eigenvalues of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$ is a basis for the row space of the constraint matrix. Furthermore, Property 2 clearly provides a way of identifying the number of constraints, without any *a priori* knowledge by examining the eigenvalues of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$. The number of constraints (also referred to model-order) is estimated to be equal to the number of small eigenvalues all of which should be equal. Thus, if $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$ is known, we can derive the number of constraints and the constraint matrix by applying PCA to this matrix. However, we can only obtain an unbiased estimate of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$ from the sample covariance matrix $\mathbf{S_y}$. It has been proved that the eigenvectors of $\mathbf{S_y}$ are asymptotically unbiased estimates of the eigenvectors of $\mathbf{\Sigma}_{\mathbf{Y}|\mathbf{X},N}$, if the errors and (hence the measurements) are normally distributed (Jolliffe, 2002). Using this result, we can therefore conclude that if the errors in all measurements are normally distributed with identical variances (homoscedastic errors) and are mutually independent, then PCA can be used to derive an asymptotically unbiased estimate of the constraint matrix. It should be noted that the estimate of the constraint matrix derived using PCA can differ from the true constraint matrix form (we may desire) by a rotation matrix. In other words, the estimated and true constraint matrices are related as

$$\hat{\mathbf{A}} = \mathbf{QA}, \tag{19}$$

where $\mathbf{Q}$ is some non-singular matrix.

It can also be shown that the denoised estimates obtained using PCA given by Eq. (16) are the reconciled values of the data matrix $\mathbf{Y}$ corresponding to the constraint matrix $\hat{\mathbf{A}}$ identified using PCA, under the assumption that $\mathbf{\Sigma}_\epsilon = \sigma^2\mathbf{I}$. By substituting the estimated constraint model given by Eq. (17) in Eq. (5), we obtain

$$\hat{\mathbf{X}} = (\mathbf{I} - \mathbf{U}_2(\mathbf{U}_2^{\mathrm{T}}\mathbf{U}_2)^{-1}\mathbf{U}_2^{\mathrm{T}})\mathbf{Y}. \tag{20}$$

Substituting for the SVD of the data matrix and using the property that the eigenvectors are orthonormal, we get

$$\hat{\mathbf{X}} = \sqrt{N}(\mathbf{I} - \mathbf{U}_2\mathbf{U}_2^{\mathrm{T}})(\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^{\mathrm{T}} + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^{\mathrm{T}}). \tag{21}$$

Expanding the above equation and using the orthonormal property of the eigenvectors we finally get

$$\hat{\mathbf{X}} = \sqrt{N}\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^{\mathrm{T}}. \tag{22}$$

---

[2] Since the number of constraints may not be known *a priori*, it is recommended that data be obtained from as many distinct steady state operating conditions as the number of variables n.

Based on the above analysis, it can be concluded that PCA is a method that simultaneously identifies the constraint model and obtains reconciled estimates with respect to the identified model purely from data under the assumption that measurement errors in different variables are independent and have the same variance. The number of constraints can also be obtained purely from data. It should, however, be noted that while the estimates obtained using PCA satisfy the identified constraint matrix, they will not satisfy the true process constraints due to inaccuracies in the estimated constraint matrix. The larger the sample size of the data set, the more accurate will be the estimated constraint matrix, and more closely will the PCA estimates match the reconciled estimates derived by applying DR using the true process constraints. It may also be verified that the reconciled estimates are invariant to a rotation of the constraint matrix. Thus, the fact that the constraint matrix estimated using PCA differs from the desired form of the true constraint matrix (derived from first principles) by a rotation, will not have an effect on the reconciled estimates.

4.1. Error-covariance matrix known

The results in the preceding subsection were obtained under the assumption that the errors in measurements of different variables have identical variances. In this subsection, we describe the identification of constraint model using PCA when the measurement errors have different variances and may also, in general, be correlated. It is however assumed that the error-covariance matrix Σ_ϵ is known. Furthermore, the errors in measurements of different samples are assumed to be independent and identically distributed with the same known covariance matrix. It may be noted that these assumptions regarding the measurement errors are the same as those made in DR.

[Narasimhan and Shah (2008)](#) described an approach for applying PCA for model identification from data when the error-covariance matrix is known. This approach is based on transforming the data using an appropriate matrix before applying PCA. The measurement errors corrupting the transformed data are independent and identically distributed (i.i.d.), and PCA can be applied to the transformed data matrix. The approach is as follows. The Cholesky decomposition of Σ_ϵ is given by

$$\Sigma_\epsilon = \mathbf{L}\mathbf{L}^{\mathrm{T}}, \tag{23}$$

where $\mathbf{L}$ is an $(n \times n)$-dimensional lower triangular matrix. Then, the noisy data matrix can be transformed as follows:

$$\mathbf{Y}_s = \mathbf{L}^{-1}\mathbf{Y} = \mathbf{L}^{-1}\mathbf{X} + \mathbf{L}^{-1}\mathbf{E}, \tag{24}$$

where $\mathbf{E}$ is an $(n \times N)$-dimensional error matrix. Let $\mathbf{S}_{\mathbf{y}_s}$ be the covariance matrix of the transformed data defined by

$$\mathbf{S}_{\mathbf{y}_s} = \frac{1}{N}\mathbf{Y}_s\mathbf{Y}_s^{\mathrm{T}} \tag{25}$$

By taking expectation we can show that

$$\Sigma_{\mathbf{Y}_s|\mathbf{X},N} = E[\mathbf{S}_{\mathbf{y}_s}|\mathbf{X}, N] = \mathbf{M}_{\mathbf{x}_s} + \mathbf{I}, \tag{26}$$

where $\mathbf{M}_{\mathbf{x}_s} = \frac{1}{N}\mathbf{L}^{-1}\mathbf{X}\mathbf{X}^{\mathrm{T}}\mathbf{L}^{-\mathrm{T}}$. Eq. (26) is similar Eq. (18) and Properties 1 and 2 hold for the matrix $\Sigma_{\mathbf{Y}_s|\mathbf{X},N}$. Therefore, PCA can be applied to the transformed data in order obtain the constraint model and reconciled estimates corresponding to the transformed data. The reconciled estimates and constraint model corresponding to the original data can be derived using an inverse transformation as described below. Let the SVD of $\mathbf{Y}_s$ be decomposed corresponding to the first largest $n - m$ singular values and remaining m singular values as

$$\mathbf{Y}_s = \underbrace{\sqrt{N}\mathbf{U}_{1s}\mathbf{S}_{1s}\mathbf{V}_{1s}^{\mathrm{T}}}_{\hat{\mathbf{x}}_s} + \sqrt{N}\mathbf{U}_{2s}\mathbf{S}_{2s}\mathbf{V}_{2s}^{\mathrm{T}}. \tag{27}$$

Table 4
Root mean square errors in the reconciled estimates of flow variables using PCA for Example 3.

Variable	RMSE using PCA
F_1	2.3383
F_2	1.6758
F_3	2.5225
F_4	2.5288
F_5	2.3290
F_6	1.6682

The first part on the right-hand side of above equation corresponds to the reconciled estimates of the transformed data, while the second part contains the information about the constraints which relate the transformed variables. The reconciled estimates and the constraint matrix corresponding to the original data are given by

$$\hat{\mathbf{X}} = \mathbf{L}\hat{\mathbf{X}}_s \tag{28}$$

$$\hat{\mathbf{A}} = \mathbf{U}_{2s}^{\mathrm{T}}\mathbf{L}^{-1}. \tag{29}$$

Since the smallest m eigenvalues of $\mathbf{S}_{\mathbf{y}_s}$ matrix are all equal to 1, it provides a systematic method to determine the constraint model order. Due to finite sample sizes and numerical inaccuracies, the equality of smallest m eigenvalues can be numerically checked by testing whether the average of the smallest m eigenvalues approximately equals unity.

Example 3. The flow process of [Example 1](#) is again considered along with the simulated data set of 1000 samples. We assume that the error-covariance matrix (used to simulate the measurements) is known, but neither the number of constraints nor the constraint matrix is assumed to be known. It may be noted that the error-covariance matrix is diagonal, with the diagonal elements being the error variances as given in [Table 1](#). The data matrix is transformed as defined by Eq. (24) and its SVD obtained. The six ordered singular values obtained are 263.16, 21.83, 1.05, 1.04, 1.01, and 0.98. Examination of the singular values clearly indicates that there are four constraints, because the smallest four singular values are all nearly equal. The denoised estimates of the transformed data are obtained using the first two PCs, and the reconciled estimates of the original data are obtained using Eq. (28). The RMSE in reconciled estimates of each variable is computed and reported in [Table 4](#). Comparing the RMSE of estimates obtained using PCA with those reported in [Table 2](#) which are obtained by applying DR using the known process model, it is clear that they are nearly equal. Although the PCA estimates do not exactly satisfy the true flow balances given by Eq. (8), the maximum constraint residual is of the order of 10^{-12}. These indirectly indicate that the constraint model obtained using PCA is a good estimate of the true process constraints.

4.2. Comparison of PCA estimated constraint model and true process constraint model

In [Example 3](#), the eigenvectors of the transformed data covariance matrix corresponding to the smallest four eigenvalues are used to derive an estimate of the process constraint matrix using Eq. (29), and is given below

$$\hat{\mathbf{A}} = \begin{bmatrix} -2.4832 & 0.2270 & -2.4441 & 4.3450 & 0.5926 & -2.1390 \\ 1.5738 & -5.0224 & 1.4541 & 1.4337 & -4.4687 & 2.1376 \\ 5.3860 & 6.5976 & -2.9594 & -0.3256 & -2.0764 & -3.3396 \\ 2.9833 & -3.4863 & -3.8879 & -0.1440 & 1.0598 & 7.5021 \end{bmatrix}. \tag{30}$$

The elements of the estimated constraint matrix do not seem to correspond with those of the constraint matrix derived from the first principles given by Eq. (9). However, it should be noted that an element by element comparison between the estimated and true constraint matrices cannot be done, because the estimated

constraint matrix may differ from the first principles model by a rotation matrix (see Eq. (19)). Thus, only the row spaces of the estimated and true process constraint matrices can be compared. Two criteria were proposed by Narasimhan and Shah (2008) for making such a comparison (i) the subspace angle between the row subspaces of the estimated and true constraint matrices, and (ii) the sum of orthogonal distances of the row vectors of the estimated constraint matrix from the subspace defined by the rows of the true constraint matrix. This metric (denoted by 'α') is computed as follows:

$$\alpha = \Sigma_i \alpha_i, \quad \text{with} \tag{31}$$

$$\alpha_i = \|\hat{\mathbf{A}}_{i.} - \hat{\mathbf{A}}_{i.}\mathbf{A}^{\mathrm{T}}(\mathbf{A}\mathbf{A}^{\mathrm{T}})^{-1}\mathbf{A}\|, \tag{32}$$

where $\hat{\mathbf{A}}_{i.}$ is the ith row of the estimated constraint matrix. Hence, α value near to zero indicates that $\hat{\mathbf{A}}$ is a good approximation of $\mathbf{A}$.

For Example 3, the subspace angle between the row spaces of estimated and true constraint matrices is $0.2377°$ and $\alpha = 0.0555$. These values suggest that a good estimate of the constraint matrix has been obtained using PCA. Our experience with larger examples indicates that subspace angle is not a good criterion for comparison, because it does not clearly indicate the quality of the estimated constraint matrix, especially as the number of constraints increases.

A practically useful method for comparing the estimated and true constraint matrix is proposed in this work. For this purpose, the process variables are partitioned into a set of dependent variables $\mathbf{x}_D$ and a set of independent variables $\mathbf{x}_I$, based on process knowledge. The number of dependent variables should be chosen equal to the number of constraints. The constraints given by Eq. (1) can be rewritten as

$$\mathbf{A}_D\mathbf{x}_D + \mathbf{A}_I\mathbf{x}_I = 0, \tag{33}$$

where $\mathbf{A}_D$ and $\mathbf{A}_I$ are the sub-matrices of $\mathbf{A}$ corresponding to dependent and independent variables, respectively. From Eq. (33), the regression matrix relating the dependent variables to the independent variables can be obtained as

$$\mathbf{x}_D = -(\mathbf{A}_D)^{-1}\mathbf{A}_I\mathbf{x}_I = \mathbf{R}\mathbf{x}_I. \tag{34}$$

In a similar manner, the estimated regression matrix can be obtained from the estimated constraint matrix as

$$\hat{\mathbf{R}} = -\hat{\mathbf{A}}_D^{-1}\hat{\mathbf{A}}_I \tag{35}$$

An element by element comparison can be made between $\mathbf{R}$ and $\hat{\mathbf{R}}$.

For Example 3, if the flow variables F_3 to F_6 are chosen as the dependent variables, then the true regression matrix and the estimated regression matrix derived from the estimated constraint matrix are given by

$$\mathbf{R} = \begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{36}$$

$$\hat{\mathbf{R}} = \begin{bmatrix} 1.0057 & 0.9936 \\ 1.0013 & 0.9988 \\ 1.0013 & -0.0022 \\ 0.0013 & 0.9991 \end{bmatrix}. \tag{37}$$

An element by element comparison between the two regression matrices shows that the maximum absolute difference between the elements is 0.0064. This again clearly shows that an accurate estimate of the constraint matrix is obtained using PCA.

5. Model identification, error variance estimation and data reconciliation using IPCA

In Section 4.1, it was established that PCA is a technique for obtaining the steady-state constraint model and reconciled estimates from measured data when the error-covariance matrix is known. In practice, the error variances and covariances are not easily available and may also change with time. If replicate measurements are available corresponding to one or more steady states, then the error-covariance matrix can be directly estimated from the data, provided the measurements corresponding to each steady state operating period are clearly identified. However, identification of steady operating periods from data is itself a challenging problem. Given these practical difficulties, the question is whether it is possible to simultaneously estimate the error-covariance matrix, the process constraint model, and the reconciled estimates purely from data, without the need for replicate measurements corresponding to a steady state. Surprisingly, it is possible to extract all this information from data. A method for this purpose known as iterative PCA was recently proposed by Narasimhan and Shah (2008). A brief description of this method follows.

In Section 4.1, it was shown how PCA can be used to obtain an estimate of the constraint matrix and reconciled estimates if the error-covariance matrix is given. In IPCA, this approach is iteratively combined with an algorithm for estimating the error-covariance matrix from data, given the constraint matrix, by solving the following optimization problem.

$$\min_{\Sigma_\epsilon} N \log |\hat{\mathbf{A}}\Sigma_\epsilon\hat{\mathbf{A}}^{\mathrm{T}}| + \sum_{k=1}^{N}[\mathbf{r}^{\mathrm{T}}(k)(\hat{\mathbf{A}}\Sigma_\epsilon\hat{\mathbf{A}}^{\mathrm{T}})^{-1}\mathbf{r}(k)] \tag{38}$$

Under the assumptions made regarding the measurement errors Eq. (3), the objective function used for the estimating the error-covariance matrix is identical to maximizing the likelihood function of the constraints residuals $\mathbf{r}(k)$ defined by

$$\mathbf{r}(k) = \hat{\mathbf{A}}\mathbf{y}(k) \tag{39}$$

An updated estimate of the constraint matrix can be obtained using the estimated error-covariance matrix as described in Section 4.1, and the procedure repeated until the estimates of the error variance matrix (and constraint matrix) converges. At convergence, the smallest m eigenvalues should be all equal to unity. Thus, the convergence test can be applied to check if the average of the smallest m eigenvalues is close to unity.

The method described in this section for estimating the measurement error-covariance matrix from constraints residuals is known as an indirect method in the area of data reconciliation. Using indirect methods, the maximum number of elements of the symmetric error-covariance matrix that can be estimated is $m(m+1)/2$. Almasy and Mah (1984) were the first to propose an indirect method, and obtained a solution which minimizes the sum of the off-diagonal elements of the error-covariance matrix. Later, Keller et al. (1992) obtained an analytical solution for the least squares solution for the elements of error-covariance matrix. A robust indirect approach for estimating the error variances was also proposed by Chen et al. (1997). The approach described in this section is a maximum likelihood estimation procedure. Besides, a significant difference between the procedure described here and other indirect approaches proposed earlier in the field of DR, is that the constraint matrix is assumed to be known in earlier approaches, whereas the procedure described here is combined with PCA for simultaneous estimation of the constraint matrix and error-covariance matrix. The reconciled estimates of variables corresponding to these estimated constraint and error-covariance matrix are also simultaneously obtained using Eq. (28).

Table 5
Standard deviation (SD) and root mean square errors in the reconciled estimates of flow variables using IPCA (Example 4).

Variable	Estimated SD	RMSE using IPCA
F_1	0.1006	2.3439
F_2	0.0848	1.6753
F_3	0.1507	2.5246
F_4	0.2128	2.5309
F_5	0.1873	2.3344
F_6	0.0866	1.6689

Example 4. The flow process of Example 1 is considered to demonstrate IPCA. It is assumed that the error-covariance matrix is diagonal, but its elements are unknown. We need to estimate six elements of Σ_ϵ, along with the constraint matrix and reconciled estimates. The number of constraints for this example is four, and it is possible to estimate a maximum of $4(4+1)/2 =10$ elements of the error-covariance matrix. Since we are estimating only 6 diagonal elements of Σ_ϵ, the problem is identifiable. IPCA is applied to the data matrix and the estimated standard deviations of errors are given in Table 5. The results show that IPCA provides accurate estimates of error variances. Further, the six ordered singular values obtained are 258.47, 21.53, 1.02, 1, 0.99, and 0.99, which clearly indicates that there are four constraints. The reconciled values and the constraint matrix can be estimated using Eqs. (28) and (29). The subspace angle between the row spaces of estimated and true constraint matrices is 0.2373° and $\alpha = 0.0509$. The maximum constraint residual is of the order of 10^{-12}. The estimated regression matrix ($\hat{\mathbf{R}}_{ipca}$) computed by considering the flow variables F_3 to F_6 as dependent variables is as follows:

$$\hat{\mathbf{R}}_{ipca} = \begin{bmatrix} 1.0056 & 0.9937 \\ 1.0012 & 0.9990 \\ 1.0012 & -0.0022 \\ 0.0012 & 0.9993 \end{bmatrix}. \tag{40}$$

The maximum absolute difference between the elements of $\mathbf{R}$ and $\hat{\mathbf{R}}_{ipca}$ is 0.0063. The comparison of the estimated constraint matrix using the subspace angle (α), and $\hat{\mathbf{R}}_{ipca}$ shows that IPCA provides an accurate estimate of the constraint matrix without knowledge of the error-covariance matrix. Further, the RMSE values in the reconciled estimates of each variable are computed and reported in Table 5. The RMSE values show that the reconciled estimates are comparable to the one obtained by applying DR using the known constraint model. The simulation results show that IPCA is a reliable data-driven method for obtaining the reconciled values, a constraint matrix, and estimates of standard deviations from the data without any *a priori* knowledge.

6. Extensions to PCA based data reconciliation

6.1. Partially measured systems

In the preceding sections, the close link between DR and PCA was established. Further, we have demonstrated that it is possible to identify the constraint model and error-covariance matrix, if required, from data using PCA (or IPCA). The estimated model and error-covariance matrix can be subsequently used for obtaining more accurate estimates of variables, by reconciling the measurements with respect to the identified model. In deriving these results, it was assumed that all process variables are measured, which is not valid in general. The connection between PCA and DR applied to a partially measured system is elucidated in this section.

We consider a process defined in Section 2.2. Let n_p be the number of measured variables out of n process variables, and let $\mathbf{y}_p(j)$ be measurements of n_p variables at the jth sample, where $j = 1, \ldots, N$.

The corresponding $(n_p \times N)$-dimensional measurement matrix $\mathbf{Y}_p$ can be obtained by collating $\mathbf{y}_p(j)$ for all N samples. Then, the objective is to identify a linear model relating the measured variables and obtain reconciled estimates of these variables.

If the measurement error-covariance matrix is known, then we can apply PCA as described in Section 4.1 to the data matrix $\mathbf{Y}_p$ and estimate the number of constraints, linear constraint matrix and reconciled values of the measured variables. On the other hand, if the measurement error-covariance matrix is unknown, then we can apply IPCA as described in Section 5 to simultaneously estimate the error-covariance matrix, constraint matrix and reconciled values. The key question is how these estimates derived using PCA or IPCA are related to the estimates obtained by applying DR to the partially measured system described in Section 2.2, which uses knowledge of the true process constraints and error-covariance matrix. It may be noted that by applying PCA or IPCA to the data matrix $\mathbf{Y}_p$, it is possible to identify the linear constraints relating only the measured variables. The constraints relating only the measured variables can also be derived using a projection matrix on the known true process constraints as described in Section 2.2. The reduced data reconciliation problem obtained after projecting out the unmeasured variables is identical to the DR problem for the completely measured process, with the only difference being that the constraint matrix relating the measured variables is the reduced balance matrix $\mathbf{PA}_k$. Therefore, using similar arguments as described in Sections 4.1 and 5, we can derive the following result

$$\mathbf{U}_{2p} = \mathbf{QPA}_x, \tag{41}$$

where $\mathbf{U}_{2p}$ are the eigenvectors of the covariance matrix of $\mathbf{Y}_p$ corresponding to the smallest m_p eigenvalues, $\mathbf{Q}$ is an arbitrary non-singular rotation matrix and $\mathbf{PA}_x$ is the reduced constraint matrix. The number of constraints in the reduced reconciliation problem m_p can be shown to be equal to $(m-t)$ where t is the rank of $\mathbf{A}_u$ (Narasimhan and Jordache, 2000). The number of constraints m_p can also be obtained from the data because the smallest m_p eigenvalues should all be equal. The reconciled estimates corresponding to the identified model are given as before by

$$\hat{\mathbf{X}}_p = \mathbf{U}_{1p}\mathbf{S}_{1p}\mathbf{V}_{1p}. \tag{42}$$

While the above discussion brings out the relationship between PCA and DR as applied to a partially measured system, it should be pointed out that because there is no information regarding the unmeasured variables, PCA (or IPCA) can be used to estimate only the reduced constraint matrix that relate the measured variables. The true constraint matrix that relates both the measured and unmeasured variables cannot obviously be estimated from measured data. Furthermore, using the first principles model, it is also possible to classify unmeasured variables as observable or unobservable based on the given measurement structure, which cannot be derived using the data driven approach. The classification of measured variables as redundant or non-redundant can, however, be performed using the data driven approach, by examining the columns of the estimated reduced constraint matrix. If all the elements of a column of $\hat{\mathbf{A}}_p$ are close to zero, then the corresponding measured variable does not participate in any of the reduced constraints and is therefore a non-redundant measured variable. The following example illustrates the application of PCA to a partially measured flow process.

Example 5. The flow process of Example 2 is considered here. We assume that the error-covariances for these variables are known. Then, PCA approach described in Section 4.1 can be applied to the available flow measurements. The three ordered singular values obtained are 169.84, 18.96, and 1.03. This indicates that there is one constraint among the measured variables. The estimated constraint row is $[-4.8612, 0.0107, 4.8549]$. The estimated constraint is a scale

Table 6
Root mean square errors in the reconciled estimates of flow variables for the partially measured system in Example 5.

Variable	RMSE using PCA with the known SD
F_1	2.7951
F_2	2.5749
F_5	2.7873

Table 7
Partially known constraint matrix: standard deviations (SDs) and root mean square errors in the reconciled estimates of flow variables using IPCA (Example 6).

Variable	Estimated SD	RMSE using IPCA
F_1	0.1007	2.3418
F_2	0.0850	1.6689
F_3	0.1513	2.5189
F_4	0.2128	2.5189
F_5	0.1872	2.3342
F_6	0.0865	1.6692

of the projected constraint obtained from first principles obtained in Example 2. Note that the coefficient corresponding to the flow F_2 is almost zero, which indicates that F_2 is a non-redundant variable and its measurement cannot be reconciled. Comparing the RMSE values of estimates obtained using PCA in Table 6 with the RMSE values of measurements reported in Table 2, it is clear that PCA reconciles the flow variables F_1 and F_5, but not F_2. The maximum constraint residual is of the order of 10^{-13} which indicates that the identified model is close to the true one.

6.2. Partially known constraint matrix

In Section 4, it was shown how PCA can be used for simultaneous model identification and data reconciliation when no knowledge of the constraint matrix is available. In some cases, a subset of constraint matrix (or linear relationships) may be known. In such a situation, it is required to identify only the remaining constraints that relate the measured variables. In this subsection, we extend the approach of PCA and IPCA so that the partial knowledge available about the process constraints can be fully exploited.

Let $\mathbf{A}_g$ be the $(m_g \times n)$-dimensional known constraint matrix. For simplicity we assume that the error-covariance matrix is known and is an identity matrix. The procedure described in this section can be generalized using the approach described in Sections 4 and 5, if the error-covariance matrix is not an identity matrix, or if it is unknown. We are interested in identifying from the measured data only the linear constraints other than those that are specified. We ensure that none of the given constraints or any linear combination of these can be identified from the data by first determining the component of the measured data that is orthogonal to the given constraints. The orthogonal component of the data is given by

$$\mathbf{Y}_{proj} = [\mathbf{I} - \mathbf{A}_g^{\mathrm{T}}(\mathbf{A}_g\mathbf{A}_g^{\mathrm{T}})^{-1}\mathbf{A}_g]\mathbf{Y}. \tag{43}$$

PCA is applied using $\mathbf{Y}_{proj}$ to estimate the remaining $(m - m_g)$ constraints. In order to identify the remaining constraints from the SVD of the projected data matrix, it should be first noted that the projected data matrix satisfies the given constraints and will therefore have a rank equal to $(n - m_g)$. The smallest m_g singular values of the data matrix will be equal to zero and the corresponding left singular vectors (eigenvectors of the covariance matrix of the projected data matrix) will be exact linear combinations of the given constraints. Out of the non-zero singular values, we consider the smallest $(m - m_g)$ singular values (all of which should be almost equal). The SVD of the projected data matrix is therefore partitioned corresponding to the largest $(n - m)$, next largest $(m - m_g)$, and remaining smallest m_g singular values as

$$svd(\mathbf{Y}_{proj}) = \mathbf{U}_{proj,1}\mathbf{S}_{proj,1}\mathbf{V}_{proj,1}^{\mathrm{T}} + \mathbf{U}_{proj,2}\mathbf{S}_{proj,2}\mathbf{V}_{proj,2}^{\mathrm{T}}$$
$$+ \mathbf{U}_{proj,3}\mathbf{S}_{proj,3}\mathbf{V}_{proj,3}^{\mathrm{T}} \tag{44}$$

The transpose of $\mathbf{U}_{proj,2}$ is an estimate of the remaining $(m - m_g)$ constraints. The complete constraint matrix can be constructed as

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{U}_{proj,2}^{\mathrm{T}} \\ \mathbf{A}_g \end{bmatrix} \tag{45}$$

The reconciled estimates are obtained using the estimated constraint matrix in Eq. (5). The following example demonstrates how partial knowledge of constraints can be combined with PCA to obtain improved estimates of variables as well as error-covariance matrix.

Example 6. The flow system of Example 1 is considered here. It is assumed that the following constraints are known

$$\mathbf{A}_g = \begin{bmatrix} 1 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 \end{bmatrix}. \tag{46}$$

However, the error-covariances, and the number of constraints are unknown. IPCA is applied to the data matrix to estimate the measurement error variances and remaining constraints. The six singular values obtained at convergence are 258.17, 21.51, 1.01, 0.99, 0, and 0. The eigenvectors corresponding to the last two zero singular values correspond to the given constraints. Two singular values are equal to one, and, hence two additional constraints are correctly identified. The standard deviations of measurement errors estimated using IPCA and the RMSE in reconciled estimates are shown in Table 7. Comparing with the results reported in Table 5, it can be inferred that by utilizing knowledge of the two known process constraints, IPCA is able to obtain slightly more accurate reconciled estimates. The subspace angle between the row spaces of estimated and true constraint matrices is $0.0188°$ and $\alpha = 0.0032$. Further, the estimated regression matrix computed by considering the flow variables F_3 to F_6 as dependent variables is as follows:

$$\hat{\mathbf{R}}_{par} = \begin{bmatrix} 1.0000 & 1.0000 \\ 1.0000 & 1.0000 \\ 0.9999 & -0.0007 \\ 0.0002 & 1.0003 \end{bmatrix}. \tag{47}$$

The maximum absolute difference between the elements of $\mathbf{R}$ and $\hat{\mathbf{R}}_{par}$ is 7.4346×10^{-4}. These values indicate that the incorporation of the given constraints into IPCA (PCA) method improves the estimation of the remaining constraints.

6.3. Selection of model order

If the assumptions made regarding the measurement errors hold in practice, and the data are obtained when the process is strictly in a steady state, then the model order can be determined by examination of the singular values as described in Sections 4 and 5. If these assumptions do not hold, then it is difficult to precisely identify the model order from data when PCA or IPCA is applied. In such a scenario, the estimated model order m_e may be greater or lesser than the true model order m. If the estimated model order is greater than the true model order, we refer to it as *overfitting* because the reconciled estimates are forced to satisfy additional constraints that are not valid. Conversely, if the estimated number of constraints is lesser than the actual number, then we refer to it as *underfitting*. In this section, we investigate the effect of underfitting and overfitting the data, due to incorrect estimation of the number of constraints.

Table 8
Singular values for different model order selection.

No.	Singular values		
	$m_e = 5$ (overfitting)	$m_e = 4$ (perfect model order)	$m_e = 3$ (underfitting)
1	30194.45	258.47	57191.37
2	1.98	21.53	41.35
3	1.00	1.02	2.54
4	0.16	0.99	1.01
5	0.11	0.99	0.99
6	0.07	0.99	0.99

Table 9
RMSE values for different model order selection for each flow variable.

Variable	RMSE values		
	$m_e = 5$ (overfitting)	$m_e = 4$ (perfect model order)	$m_e = 3$ (underfitting)
F_1	35.4270	2.3439	2.6043
F_2	35.5942	1.6753	1.6943
F_3	4.8291	2.5246	2.6577
F_4	4.8345	2.5309	2.6314
F_5	35.4262	2.3344	5.8500
F_6	35.5942	1.6689	1.6766

We write the SVD of the data matrix in partitioned form as given by Eq. (15), where $\mathbf{U}_2$ contain the eigenvectors of $\mathbf{S}_y$ corresponding to the smallest m eigenvalues. In the case of underfitting, the eigenvectors corresponding to the smallest $m_e < m$ eigenvalues, denoted as $\mathbf{U}_{2s}$ will be chosen. This implies that $\mathbf{U}_{2s}$ is a subset of $\mathbf{U}_2$, and the columns of $\mathbf{U}_{2s}$ will therefore be an asymptotically unbiased estimate of a subset of linear combinations of the rows of $\mathbf{A}$. Therefore, in the limit as sample size goes to infinity, the estimated constraint matrix will be orthogonal to the true values of the variables. That is,

$$\mathbf{U}_{2s}^T \mathbf{X} = \mathbf{0} \tag{48}$$

The reconciled estimates corresponding to the identified constraint matrix is given by

$$\hat{\mathbf{X}} = \mathbf{Y} - \mathbf{U}_{2s} \mathbf{S}_{2s} \mathbf{V}_{2s}^T \tag{49}$$

$$= (\mathbf{I} - \mathbf{U}_{2s} \mathbf{U}_{2s}^T) \mathbf{Y}. \tag{50}$$

Taking expectation of the above equation and using Eq. (48) we can prove that

$$E[\hat{\mathbf{X}}] = \mathbf{X}. \tag{51}$$

Eq. (51) indicates that the reconciled estimates are unbiased even if the estimated number of constraints is less than the actual number.

If the number of constraints estimated is greater than m, then the estimated constraint matrix is given by

$$\hat{\mathbf{A}} = \begin{bmatrix} \mathbf{U}_2^T \\ \mathbf{U}_{1s}^T \end{bmatrix} \tag{52}$$

where $\mathbf{U}_{1s}$ is a subset of $\mathbf{U}_1$. In the limit as the sample size goes to infinity, the eigenvectors $\mathbf{U}_2$ will be orthogonal to the true values, but the eigenvectors $\mathbf{U}_{1s}$ form a subspace of the true data space. This implies that

$$\begin{bmatrix} \mathbf{U}_2^T \\ \mathbf{U}_{1s}^T \end{bmatrix} \mathbf{X} = \begin{bmatrix} \mathbf{0} \\ \alpha \end{bmatrix} \tag{53}$$

Substituting the above estimate of the constraint matrix in Eq. (20), the reconciled estimates are obtained as

$$\hat{\mathbf{X}} = \mathbf{Y} - [\mathbf{U}_2 \ \mathbf{U}_{1s}] \begin{bmatrix} \mathbf{U}_2^T \\ \mathbf{U}_{1s}^T \end{bmatrix} \mathbf{Y} \tag{54}$$

$$= \mathbf{Y} - (\mathbf{U}_2 \mathbf{U}_2^T + \mathbf{U}_{1s} \mathbf{U}_{1s}^T) \mathbf{Y} \tag{55}$$

Taking the expectation of above equation and using Eq. (53) we get

$$E[\hat{\mathbf{X}}] = \mathbf{X} - \mathbf{U}_{1s} \alpha \tag{56}$$

This shows that overfitting introduces a bias in the reconciled values. Based on the above analysis, we make the recommendation that in practice it is better to be conservative in estimating the number of constraints, and avoid overfitting when using heuristics to determine the model order. The following example demonstrates the effect of underfitting and overfitting the data.

Example 7. In this example, we will demonstrate effect of selection of model order on the estimates of the constraints, standard deviations, and the reconciled values. The settings of Example 4 are considered to demonstrate the concept. IPCA is applied to the data matrix under the assumption of different choices of model order. The singular values obtained at convergence of the method are reported for different model orders in Table 8. Examination of the singular values show that if the model order assumed is 5 which is greater than the true model order of 4, then the last five singular values are not all equal to unity. This clearly indicates that the assumed model order is incorrect. However, if the assumed model order is equal to or less than the true model order, then the number of unity singular values obtained is equal to the assumed model order. Based on these observations, a systematic method can be devised for determining the true model order using IPCA. We start with the least value of the model order that results in an identifiable problem, that is the assumed value of m should be such that $m(m+1)/2 \geq n$ (in this example it is 3 since we have to estimate 6 error variances). If the number of unity singular values obtained at convergence is equal to m, then we cannot conclude that the assumed model order is correct. Instead we increment the assumed model order by unity and apply IPCA again. If the number of unity singular values obtained is not equal to the assumed model order, then this violates the theoretical result expected from IPCA. The true model order is one less that the assumed model order at which this violation is observed. Although, in theory the true model order can be estimated exactly using this systematic procedure with IPCA, it may not work in practice if the noise in measurements do not satisfy the assumptions made, or there is mild nonlinearity in the process constraints which we have not considered. In practical applications when there is an ambiguity in selecting the model order it is better to choose a lower value. This is clearly shown by the

Table 10
Quality of the model identified for different model orders by IPCA.

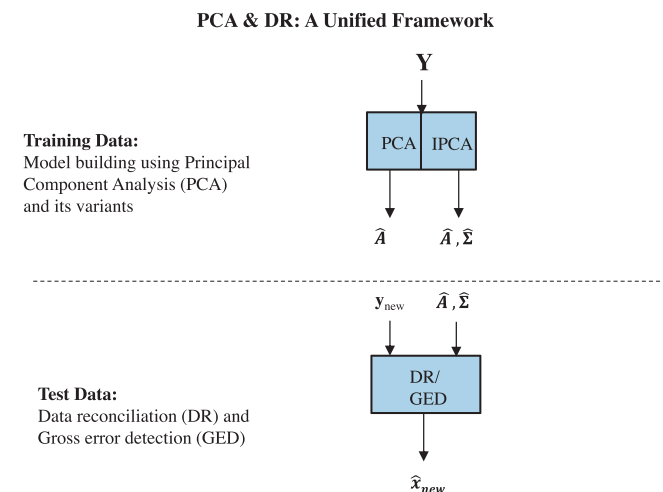| Model order	α
$m_e = 5$ (overfitting)	0.9125
$m_e = 4$ (true model-order)	0.0509
$m_e = 3$ (underfitting)	0.0492

PCA & DR: A Unified Framework

Fig. 3. A unified framework for principal component analysis (PCA) and data reconciliation (DR).

RMSE values of the reconciled estimates obtained for this example for different model-order choices reported in Table 9. Comparing the RMSE values, it can be seen that overfitting leads to poor reconciled estimates due to bias introduced in the estimates. In contrast underfitting only marginally increases the inaccuracy in the reconciled estimates. The α values computed for various orders in Table 10 also indicate that the overfitting leads to poor estimates of the constraint model.

7. Conclusions

PCA has been primarily regarded as a multivariate statistical technique which is useful for reducing dimensionality of data as well as for denoising it. The main message that we have attempted to convey in this paper is that PCA is a method that can be used to derive the steady state constraints of a linear process entirely from data along with reconciliation of the measurements. Iterative PCA which is a recent extension of PCA derives both the error-covariance matrix of measurements and the process constraint model simultaneously from data. Thus, the information necessary to reconcile data, and the reconciled estimates can be extracted from the data itself without any a priori process knowledge using PCA (or IPCA). We have also shown that if partial knowledge of process constraints is available, then they can be exploited to improve the estimates. Further, the effect of model order on the reconciled

values have been studied, and it is shown that it is better to be conservative in fitting the number of constraints (i.e. estimate less constraints) in case of unknown model order. We have used a simple example in this paper to illustrate all these ideas for ease of understanding. The methods can as well be applied to larger realistic examples as demonstrated for IPCA in the second example described in Narasimhan and Shah (2008).

The perspective provided in this paper can be used to seamlessly integrate PCA or IPCA with data reconciliation (DR) and its companion technique of gross error detection (GED). This integration is depicted in Fig. 3. We propose that IPCA should be used on historical (training) data to derive the process model and error-covariance matrix, if they cannot be easily obtained using *a priori* process knowledge. The derived model and error-covariance matrix can be used in the techniques of DR and GED for reconciling future measurements and to detect gross errors in these measurements or the model as required.

References

Almasy GA, Mah RSH. Estimation of measurement error variances from process data. Ind Eng Chem Process Des Dev 1984;23:779–84.
Bagajewicz MJ. Process plant instrumentation: design and upgrade. Lancaster: Technomic Publishing Company; 2001.
Chen J, Bandoni A, Romagnoli JA. Robust estimation of measurement error variance/covariance from process sampling data. Comput Chem Eng 1997;21:593–600.
Crowe CM, Garcia Campos GY, Hrymak A. Reconciliation of process flow rates by matrix projection. AIChE J 1983;29:881–8.
Davis J, Piovoso MJ, Kosanovich KA, Bakshi BR. Process data analysis and interpretation. In: Advances in chemical engineering. Academic Press; 1999. p. 2–97.
Gertler J, Li W, Huang Y, McAvoy T. Isolation enhanced principal component analysis. AIChE J 1999;45:323–34.
Hodouin D. Process observers and data reconciliation using mass and energy balances. In: Sbarbaro D, Villar R, editors. Advanced control and supervision of mineral processing plants. Springer; 2010. p. 15–83.
Huang B. Process identification based on last principal component analysis. J Process Control 2001;11:19–33.
Jolliffe IT. Principal component analysis. 2nd ed. New York: Springer-Verlag; 2002.
Keller JY, Zasadzinski M, Darouach M. Analytical estimator of measurement error variances in data reconciliation. Comput Chem Eng 1992;16:185–8.
Kourti T, MacGregor JF. Process analysis, monitoring and diagnosis, using multivariate projection methods. Chemom Intell Lab Syst 1995;28:3–21.
Mah RSH. Chemical process structures and information flows. USA: Butterworths Publications; 1990.
Narasimhan S, Jordache C. Data reconciliation & gross error detection: an intelligent use of process data. Gulf Publishing Company; 2000.
Narasimhan S, Shah S. Model identification and error covariance matrix estimation from noisy data using PCA. Control Eng Pract 2008;16:146–55.
Rao CR. The use and interpretation of principal component analysis in applied research. Sankhya, Ser A 1964;26:329–58.
Romagnoli J, Sanchez M. Data processing and reconciliation for chemical process operation. Academic Press; 1999.
Veverka VV, Madron F. Material and energy balancing in the process industries: from microscopic balances to large plants. Elsevier; 1997.
Yoon S, MacGregor JF. Fault diagnosis with multivariate statistical models. Part I: using steady state fault signatures. J Process Control 2001;11:387–400.