I&EC research
Industrial & Engineering Chemistry Research

pubs.acs.org/IECR Article

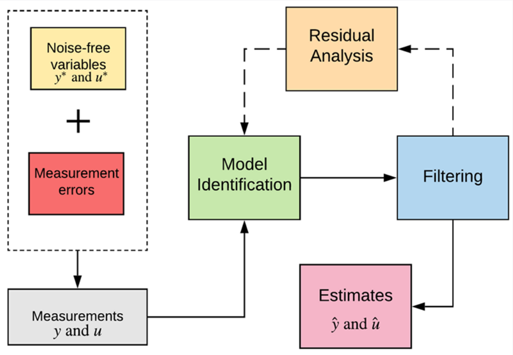
Optimal Filtering and Residual Analysis in Errors-in-Variables Model Identification

Vipul Mann,* Deepak Maurya,* Arun K. Tangirala,* and Shankar Narasimhan*

Cite This: https://dx.doi.org/10.1021/acs.iecr.9b04561 Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: Dynamic model identification from time series data is a critical component of process control, monitoring, and diagnosis. An important adjunct of model identification is the derivation of filtered estimates of the variables and consequent one-step-ahead prediction errors (residuals) which are very useful for model assessment and iterative model identification. In this work, we present an optimal filtering and residual generation method for the errors-in-variables (EIV) scenario, wherein both the input and output variable measurements are contaminated with errors. The main idea is to combine an EIV-identification strategy with the EIV-Kalman filter (EIV-KF) that is known to provide optimal filtered estimates and residuals of both inputs and outputs for a linear dynamical process in the EIV case. In this work, we combine the EIV-KF with the dynamic iterative principal component analysis (DIPCA) approach that has been recently developed for EIV model identification. This work assumes prominence in that the optimally generated residuals are critical to the tasks of model assessment, fault detection, and diagnosis. The use of residuals in model assessment and reidentification is illustrated in this article, while pointing out that the use of DIPCA alone leads to nonunique filtered estimates and hence nonunique residuals. We remark that the proposed method can be used with any other EIV identification technique.



1. INTRODUCTION

Measurements, especially those obtained from an operating process, inevitably contain errors. These errors could be either random or systematic. Random errors or noises arise due to fluctuations in ambient conditions or power, or due to the transmission and data acquisition characteristics,[1] while systematic errors arise due to biases or occasional calibration errors that may be present owing to aging or poor maintenance of sensors. Filtering is one of the most common techniques for reducing the noise content in measurements. This is important for better process monitoring,[2,3] fault detection,[4] diagnosis,[5] and control.[6] A prerequisite, however, for performing filtering is a knowledge of the underlying data generating process. If a good understanding of the process has been obtained, then a dynamic process model can be developed from first principles. However, in practice, owing either to the complexity of processes or for demands of the application (e.g., control), dynamic models are usually identified from input−output data.

The subject of model identification (or system identification) from data has been well researched, and associated techniques have been described in several books.[7−9] The commonly used techniques for solving this problem assume that the inputs do not contain any errors. However, for many practical systems, both the input and output measurements contain errors. Identifying a model under conditions where both input and output measurements are corrupted with errors is known as an errors-in-variables (EIV) identification problem. Several methods have also been developed for solving the EIV model identification problem; read reports by Juricek et al.,[10] Qin,[11] Soderstrom,[12] Söderström,[13] and Wu et al.[14] for insightful reviews of the existing literature. Among the different methods, subspace-based model identification using the instrumental variable (IV) approach and its variants[15] are currently the most preferred methods. A relatively less popular method for identifying dynamic models in the EIV case is the dynamic principal components approach (DPCA) originally proposed by Ku et al.[16] However, as pointed out by Maurya et al.[17] and Vijaysai et al.,[18] DPCA can be used only under restrictive assumptions. DPCA requires the process order to be known *a priori* and gives optimal results only when the error variances in input and output measurements are equal. Li and Qin[19] proposed a consistent DPCA method to estimate a state space model for the EIV case. More recently, Maurya et al.[17] proposed a novel dynamic iterative PCA method (DIPCA) that fully overcomes the shortcomings of DPCA in the sense that it can simultaneously estimate the model order, the error

Received: August 18, 2019
Revised: December 20, 2019
Accepted: January 6, 2020
Published: January 6, 2020

variances corrupting the measurements, and the model parameters of a dynamic *difference equation* model. Although DIPCA was proposed for identifying the model of a single-input single-output (SISO) system, it has the potential to be extended to multi-input multi-output systems.

In general, model identification is an iterative procedure, especially due to the difficulty in estimating the process order correctly. Once a preliminary model is identified, an important post-mortem exercise performed is the analysis of the model residuals (or one-step ahead model prediction errors) to verify whether the process order and parameters have been estimated precisely. Under standard assumptions, the autocorrelation function of the model residuals should be white, and the residuals should be uncorrelated with the inputs. Otherwise, the model order and parameters are re-estimated until these conditions are satisfied. While this procedure is well established in prediction error methods (PEM) used for identifying models in the non-EIV case, hitherto the use of a similar procedure for EIV model identification has not been proposed or discussed in the research literature. In order to perform such an exercise, filtered estimates of both input and output variables have to be obtained. Filtered estimates and one-step ahead prediction errors are useful not only in model reidentification, but also in other applications such as predictive control, fault detection,[20] and process monitoring.[21,22] The aim of this work is to present a method for optimal generation of filtered estimates and residuals for the EIV class of methods.

Subspace-based model identification, DPCA, or DIPCA all use PCA as a core method for estimating the model parameters in the EIV case. PCA was originally developed for identifying a linear static model.[23] Simultaneously, PCA also provides optimal estimates of the variables.[24] Extension of PCA to (only) model development for the dynamic case, as mentioned previously, led to DPCA and DIPCA. In this paper, we show that a naive extension of using PCA for obtaining (filtered) estimates in the dynamic case results in nonunique estimates of the inputs and outputs. The problem arises due to the use of lagged variables in constructing the data matrix and conducting a PCA dynamic model identification to an equivalent static model identification problem on lagged variables. In order to overcome this problem, we propose the use of an EIV Kalman filter[25] to obtain optimal estimates of both inputs and outputs as well as the one step ahead prediction errors or model residuals. We demonstrate the usefulness of these residuals in model reidentification. Although we make use of DIPCA as the EIV model identification method in this work, the EIV-KF can be used in conjunction with any other EIV model identification method for model validation and assessment. The proposed approach thus extends the use of residual analysis to EIV model identification, in general.

The rest of the paper is organized as follows: In Section 2, we formally define our problem, objectives, and the EIV framework that we will be working with. In Section 3, we review Dynamic Iterative Principal Component Analysis (DIPCA) for EIV model identification. In Section 4.1, we discuss the limitations in PCA-based filtering under dynamic conditions and the need for a technique for optimal filtering with an example. The main contribution of this paper, which is an approach for generating model residuals in the EIV class of scenarios, is presented in Section 4.2 with a focus on the DIPCA-based identification method. The utility of the developed method to perform residual analysis for model validation and adequacy check is presented in Section 5. A physical case study is presented in Section 6. Finally, a summary of the useful contributions of this work with a few concluding remarks appears in Section 7.

2. PROBLEM FORMULATION AND OBJECTIVES

Given noisy measurements of input and output variables of a linear dynamic time-invariant, single-input single-output (SISO) system, the main objective is to identify the dynamic model of the system. The model identification (including the filtering step) involves

- Determining the order of the system
- Estimating the error variances that corrupt the input and output measurements
- Estimating the parameters of the model
- Obtaining filtered estimates of the variables and one-step ahead residuals useful for validating the model and iterative identification of the model parameters, if required.

A schematic of the iterative EIV model identification plus filtering procedure that is envisioned is depicted in Figure 1.

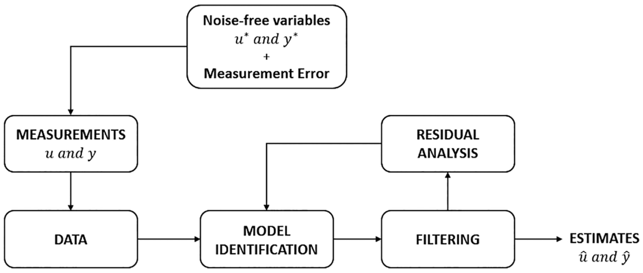


Figure 1. Schematic depicting EIV identification and filtering.

We assume that the process can be described using the following difference equation (DE) model

$$y^*[k] + \sum_{i=1}^{n_a} a_i y^*[k-i] = \sum_{j=D}^{n_b} b_j u^*[k-j] \tag{1}$$

where $u^*[k]$, $y^*[k]$ are the noise-free input and output variables at time k, respectively. The parameters a_i and b_i are the coefficients of the DE model, D is process delay, and n_a, n_b are the orders with respect to input and output, respectively. The maximum of the input and output orders, $\eta = \max(n_a, n_b)$, is denoted as the *process order*.

The noisy measurements of input and output variables at each time are assumed to be related to the corresponding true values by the following additive noise model

$$u[k] = u^*[k] + e_u[k] \quad e_u[k] \sim \mathcal{N}(0, \sigma_{e_u}^2) \tag{2}$$

$$y[k] = y^*[k] + e_y[k] \quad e_y[k] \sim \mathcal{N}(0, \sigma_{e_y}^2) \tag{3}$$

Finally, we assume that the errors in input and output measurements are independent of time and also mutually uncorrelated; that is,

$$E\{e_u[j]e_u[k]\} = 0 \quad \forall\, j \neq k \tag{4}$$

$$E\{e_y[j]e_y[k]\} = 0 \quad \forall\, j \neq k \tag{5}$$

$$E\{e_u[j]e_y[k]\} = 0 \quad \forall\, j,\ k \tag{6}$$

In essence, given noisy data, the objective is to recover process order η, noise variances $\sigma_{e_y}^2$, $\sigma_{e_u}^2$, model coefficients $\{a_i\}$, $\{b_i\}$, and filtered estimates $\hat{u}[k|k]$, $\hat{y}[k|k]$ of input and output variables at every time instant.

3. EIV MODEL IDENTIFICATION USING DIPCA

In this section, we provide an overview of DIPCA that has been recently proposed by Maurya et al.[17] to identify the process order, error variances, and model parameters of a DE model for a SISO process. The DIPCA method can be thought to consist of three steps: (i) stacking of lagged variables as in DPCA (that can handle only homoskedastic errors), up to a large lag L; (ii) estimating the process order and noise variances based on the ideas of Iterative PCA (IPCA),[23] which can determine number of underlying relations in the presence of heteroskedastic errors; and (iii) restacking the matrix with exact order, scaling the data matrix appropriately (with the noise covariance matrix), and identifying the single relation using standard PCA. The *steady-state* version of the second step, based on IPCA, was devised by Narasimhan and Shah[23] as an extension of PCA to determine the number of linear relations from measurements with heteroskedastic errors. The last step can be viewed as a modification of last principal component analysis (LPCA), which determines the single relation by using the right singular vector corresponding to the smallest singular value of the data matrix.[26] While LPCA can handle only homoskedastic errors, the modified LPCA proposed in the last step accounts for heteroskedastic errors. The major contribution of DIPCA is therefore the second step, which is that of simultaneously determining the process order and noise covariance matrix, under *dynamic* conditions. The aforementioned three steps are briefly summarized in the sections below.

3.1. Mapping Dynamic Model Identification to a Static Model Identification Problem.

Noting that the DE model described by eq 1 is an equation relating the lagged variables of input and output, but of unknown order, a lagged data matrix $\mathbf{Z}_L$ of size $(N - L) \times (2L + 2)$ is constructed by stacking input and output measurements up to a user-specified lag L as follows:

$$\mathbf{Z}_L = \begin{bmatrix} y[k] & \dots & y[k-L] & u[k] & \dots & u[k-L] \\ y[k+1] & \dots & y[k-L+1] & u[k+1] & \dots & u[k-L+1] \\ \vdots & & & & & \vdots \\ y[N] & \dots & y[N-L] & u[N] & \dots & u[N-L] \end{bmatrix}$$

where N is the sample size and $k = L + 1$.

Using eqs 2 and 3, the lagged measurement matrix can be related to the corresponding lagged matrix of true values as

$$\mathbf{Z}_L = \mathbf{Z}*_L + \mathbf{E}_L \tag{7}$$

where $\mathbf{Z}*_L$ and $\mathbf{E}_L$ are defined as

$$\mathbf{Z}*_L = \begin{bmatrix} y^*[k] & \dots & y^*[k-L] & y^*[k] & \dots & u^*[k-L] \\ y^*[k+1] & \dots & y^*[k-L+1] & u^*[k+1] & \dots & u^*[k-L+1] \\ \vdots & & & & & \vdots \\ y^*[N] & \dots & y^*[N-L] & u^*[N] & \dots & u^*[N-L] \end{bmatrix}$$

$$\mathbf{E}_L = \begin{bmatrix} e_y[k] & \dots & e_y[k-L] & e_u[k] & \dots & e_u[k-L] \\ e_y[k+1] & \dots & y[k-L+1] & e_u[k+1] & \dots & e_u[k-L+1] \\ \vdots & & & & & \vdots \\ e_y[N] & \dots & e_y[N-L] & e_u[N] & \dots & e_u[N-L] \end{bmatrix}$$

The dynamical model given by eq 1 can be viewed as linear constraints relating the columns of $\mathbf{Z}*_L$ which can be written as

$$\mathbf{Z}*_L \mathbf{A}_L^T = 0 \tag{8}$$

where $\mathbf{A}_L^T$ is a $(2L + 2) \times d$ matrix given by

$$\mathbf{A}_L^T = [\boldsymbol{\theta}_1 \, \boldsymbol{\theta}_2 \dots \boldsymbol{\theta}_d] \tag{9}$$

with the parameter vector $\boldsymbol{\theta}_i$ defined as

$$\boldsymbol{\theta}_i = [\mathbf{0}_{i-1} \; 1 a_1 \dots a_{n_a} \; \mathbf{0}_{L-n_a+D} - b_D \dots - b_{n_b} \; \mathbf{0}_{L-n_b-i+1}]^T \tag{10}$$

Ideally, if L is chosen equal to η, the matrix $\mathbf{A}^T$ contains a single column because only one relation exists among the lagged vector of variables. However, since η is not known *a priori*, L is assumed to be chosen as a large value greater than η. This gives rise to additional $(L - \eta)$ relations among the lagged variables. Thus, the number of relations that exist among the variables of $\mathbf{Z}*_L$ is given by

$$d = L - \eta + 1 \tag{11}$$

Equation 11 is central to the recovery of process order. If the errors contaminating the output and input are of identical variances (the *homoskedastic* case), the number of linear constraints d can be identified using standard PCA and hence the process order η.

3.2. Estimating Process Order and Noise Variance Simultaneously.

When the variances of errors corrupting the input and output measurements are not identical (the heteroskedastic case), the idea of iterative PCA is adopted. The approach in IPCA is to scale the dynamic matrix using the noise covariance $\boldsymbol{\Sigma}_\mathbf{e}$ followed by a standard PCA. However, in a generic situation, $\boldsymbol{\Sigma}_\mathbf{e}$ is unknown. IPCA essentially solves this problem in an iterative fashion. First, for a guessed (given) $\boldsymbol{\Sigma}_\mathbf{e}$, the process order and constraint matrix are estimated. Second, from the estimated η and $\hat{\mathbf{A}}$, the noise covariance $\boldsymbol{\Sigma}_\mathbf{e}$ is re-estimated. This iteration is continued until convergence. A short review of these two steps is presented below.

3.2.1. Estimating Process Order with Known Noise Covariance.

Define the error covariance matrix of lagged measurements in $\mathbf{Z}_L$ as

$$\boldsymbol{\Sigma}_\mathbf{e} = \begin{bmatrix} \sigma_{e_y}^2 \mathbf{I}_{L+1} & \mathbf{0} \\ \mathbf{0} & \sigma_{e_u}^2 \mathbf{I}_{L+1} \end{bmatrix} \tag{12}$$

The lagged data matrix is then scaled as follows:

$$\mathbf{Z}_{L,s} = \mathbf{Z}_L \boldsymbol{\Sigma}_\mathbf{e}^{-1/2} \tag{13}$$

The purpose of scaling the data matrix is to ensure that all measurements of the scaled data matrix will now have identical error variance equal to unity. Essentially, the heteroskedastic case is transformed to a homoskedastic case.

The number of linear constraints that exist among the scaled variables is obtained by examining the singular values of the lagged data matrix. In the limit as sample size goes to infinity, the smallest d singular values of the scaled lagged data matrix should all *converge to unity*.[17] This implies that the number of

https://dx.doi.org/10.1021/acs.iecr.9b04561

linear relations can be consistently estimated from noisy measurements, if the error variances are known. By virtue of eq 11, the process order can also be therefore estimated consistently as

$$\hat{\eta} = L - \hat{d} + 1 \tag{14}$$

Accordingly, an estimate of the linear constraint matrix $\hat{\mathbf{A}}_L$ that relates the lagged variables can be obtained using the standard PCA, i.e., singular value decomposition (SVD) of $\mathbf{Z}_{L,s}$ as follows:

$$\mathbf{Z}_{L,s} = \mathbf{USV}^T = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T + \mathbf{U}_2\mathbf{S}_2\mathbf{V}_2^T \tag{15}$$

where $\mathbf{S}_1$ is a diagonal matrix corresponding to the first $2L + 2 - \hat{d}$ largest singular values, and $\mathbf{S}_2$ is a diagonal matrix corresponding to the smallest $\hat{d}$ singular values all of which should be close to unity. The columns of matrices $\mathbf{U}_1$, $\mathbf{V}_1$, are the singular vectors of appropriate dimensions corresponding to the first largest $2L + 2 - \hat{d}$ singular values, while columns of $\mathbf{U}_2$ and $\mathbf{V}_2$ are the singular vectors corresponding to the smallest $\hat{d}$ singular values. An estimate of the constraint matrix is obtained as

$$\hat{\mathbf{A}}_L = \mathbf{V}_2^T\boldsymbol{\Sigma}_{\mathbf{e}}^{-1/2} \tag{16}$$

Note that the estimated matrix in eq 16 differs from $\mathbf{A}_L$ in eq 8 by an unknown rotation matrix.

We now turn our attention to the second step of IPCA that deals with recovering the unknown $\boldsymbol{\Sigma}_{\mathbf{e}}$ given an estimate of the constraint matrix.

3.2.2. Estimating Error Variances for a Given Process Order. The foregoing section discussed the procedure of obtaining an estimate of constraint matrix for a guessed $\boldsymbol{\Sigma}_{\mathbf{e}}$. Given an estimate of the constraint matrix relating the lagged variables, the error variances can be estimated by solving the following minimization problem

$$\min_{\boldsymbol{\Sigma}_{\mathbf{e}}} \quad N \log|\hat{\mathbf{A}}_L\boldsymbol{\Sigma}_{\mathbf{e}}\hat{\mathbf{A}}_L^T| + \sum_{k=L+1}^{N} \mathbf{r}^T[k](\hat{\mathbf{A}}_L\boldsymbol{\Sigma}_{\mathbf{e}}\hat{\mathbf{A}}_L^T)^{-1}\mathbf{r}[k] \tag{17}$$

where the constraints residuals $\mathbf{r}[k]$ are generated from the lagged measurement vector $\mathbf{z}_L[k]$ (transpose of rows of lagged matrix $\mathbf{Z}_L$) as

$$\mathbf{r}[k] = \hat{\mathbf{A}}\mathbf{z}_L[k] \tag{18}$$

The above optimization problem amounts to solving a maximum likelihood estimation (MLE) problem which is solved using an iterative procedure. It may be noted that the only two decision variables in the above optimization problem are $\sigma_{e_y}^2$ and $\sigma_{e_u}^2$, since $\boldsymbol{\Sigma}_{\mathbf{e}}$ is completely determined by these parameters (see eq 12).

Thus, the second stage of DIPCA iterates between the steps of estimating noise covariance matrix and the constraint matrix (process order) as follows. An initial estimate of the constraint matrix is obtained assuming the error variances are equal, i.e., using the standard PCA on $\mathbf{Z}_L$. This estimated constraint matrix is then used in the above optimization problem to obtain new estimates of the error variances. Subsequently these revised estimates are used to rescale the data to obtain an updated estimate of the constraint matrix. The procedure is iterated until convergence of the estimated variances and constraint matrix. This algorithm, as previously stated, termed as Iterative PCA (IPCA) was developed for simultaneously estimating the error covariance matrix and linear constraint matrix by Narasimhan and Shah[23] for steady-state model identification. DIPCA, in essence, suitably modifies and adapts IPCA for the identification of dynamical models.

The iterative procedure described above rests on the determination of number of constraints d. In order to estimate the number of constraints, constraint matrix, and error variances simultaneously, we embed the above procedure in an outer iterative loop as follows. We start with a guess of the number of constraints $\hat{d} = L$, which corresponds to a first-order model and obtain the converged estimates $\hat{\mathbf{A}}_L$ and $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$. If the smallest $\hat{d}$ singular values of scaled lagged data matrix, $\mathbf{Z}_s$ (scaled using estimated $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$), are not all equal to unity, we decrement $\hat{d}$ by one and repeat the procedure. If the smallest $\hat{d}$ singular values of $\mathbf{Z}_s$ are all equal to unity, then the guessed number of constraints is correct. From this converged estimated number of constraints, the order of the process is estimated using eq 14.

Remark 1: Unfortunately, the dynamic model cannot be easily extracted from the estimated constraint matrix $\hat{\mathbf{A}}_L$ because it differs from $\mathbf{A}_L$ defined by eq 9 by an unknown rotation matrix. Therefore, the matrix $\mathbf{Z}_L$ is restacked with $L = \hat{\eta}$ and the idea of LPCA is used as described in the following subsection.

Remark 2: Due to the availability of only finite-sized samples in practice, the eigenvalues would never be exactly equal to unity. In such cases, a hypothesis test for establishing the *equality* of the eigenvalues can be set up using the distributional results as proposed in ref 27. It may be reiterated that this test is aimed at determining the equality of the smallest d eigenvalues, but not whether they are unity valued.

3.3. Estimating Dynamic Model Parameters. Once the process order is estimated, the dynamic model parameters are estimated simply by reconfiguring the data matrix using the lag $L = \hat{\eta}$ and scaling it using $\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}$. The model parameters are then directly obtained from the right singular vector corresponding to the smallest singular value of this reconfigured scaled lagged data matrix. The estimate of the parameter vector $\boldsymbol{\theta}_1$ defined in eq 10 is obtained as

$$\hat{\boldsymbol{\theta}}_1 = \hat{\boldsymbol{\Sigma}}_{\mathbf{e}}^{-1/2}\mathbf{v}_{L+1}/v_{L+1,1} \tag{19}$$

A detailed flowchart of DIPCA algorithm is given in Figure 2.

It may be noted from the flowchart that the starting guess of d is set to L, essentially amounting to starting with a first order model. If $\eta = 1$ is unsuitable (as deemed by the unity eigenvalues condition), the guess (of d) is decreased to $d = L - 1$. The procedure is continued until a value of d that yields the smallest d eigenvalues which are all equal. The strategy of starting with a low η is in line with the general principle of parsimonious modeling. Complete technical details and proof of consistency (of estimating d) are available in ref 17. As is the case with all estimation problems, nonzero estimates of parameters may be obtained even for parameters in $\boldsymbol{\theta}_1$ which are truly zero-valued. A significance test on parameter estimates is therefore required. In order to conduct this test, a bootstrap approach can be used to obtain 95% confidence intervals (CIs) for all parameters, and only the parameters whose CIs do not contain zero may be retained.

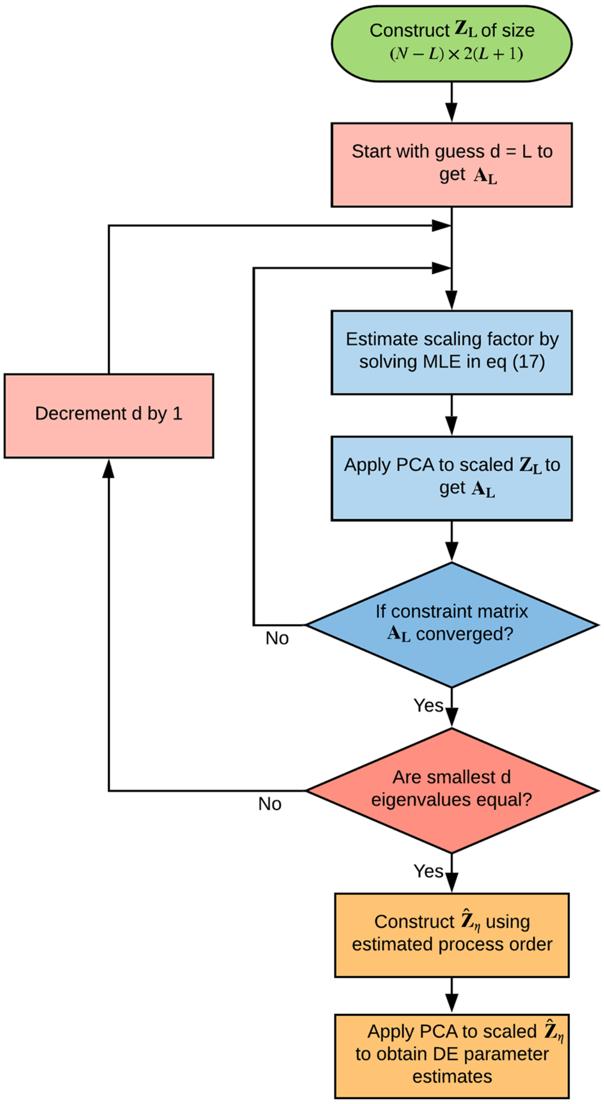

Construct $\mathbf{Z}_L$ of size $(N-L)\times 2(L+1)$
Start with guess d = L to get $\mathbf{A}_L$
Estimate scaling factor by solving MLE in eq (17)
Apply PCA to scaled $\mathbf{Z}_L$ to get $\mathbf{A}_L$
If constraint matrix $\mathbf{A}_L$ converged?
No
Yes
Are smallest d eigenvalues equal?
No
Yes
Decrement d by 1
Construct $\hat{\mathbf{Z}}_\eta$ using estimated process order
Apply PCA to scaled $\hat{\mathbf{Z}}_\eta$ to obtain DE parameter estimates

Figure 2. Flowchart of the DIPCA algorithm.

4. FILTERING AND RESIDUAL GENERATION FOR EIV MODELS

An important and useful tool used in identification methods is residual analysis based on one-step ahead prediction errors. These residuals offer crucial model diagnostics since they can be used to check whether the estimated order is acceptable, or whether the model has to be reidentified for a different estimate of the process order. Residual analysis is a well-studied and widely implemented technology in the non-EIV identification class of problems. Surprisingly, in EIV model identification literature there is neither a well-established method to generate unique residuals nor any report of residual analysis. In this section, we present the main contribution of this work, which is a generic method for generating residuals in EIV model identification. In delineating the proposed method, we use the DIPCA-based EIV-identification method for modeling purposes. However, the method is fairly generic in the sense that it can accommodate any other EIV-identification technique.

A crucial first step in residual generation is the computation of one-step ahead predictions (estimates). We first discuss the fact that direct estimates obtained from PCA for *dynamic* processes are nonunique, and hence this issue has to be suitably addressed.

4.1. Nonuniqueness of PCA Estimates for Dynamic Processes.

The pitfalls in using the standard PCA-based denoised estimates for dynamic processes are discussed. Subsequently, we propose the use of the EIV Kalman filter (EIV-KF) for obtaining the optimal estimates of input and output variables as well as corresponding residuals. Although EIV-KF has already been proposed[25] for deriving filtered estimates when both the model and error covariance matrices are specified, to our knowledge, it has not been used in EIV model identification before. We demonstrate the use of residual analysis as a validation tool in EIV model identification for assessing the quality of the identified model, and reidentification of the model, if necessary, based on an improved estimate of the process order. It must be noted that although the description of residual analysis is presented in the context of using DIPCA as the model identification method, it is also applicable to other EIV model identification approaches.

In the steady-state case, PCA can be used to obtain not only the number of linear constraints relating the variables but also the denoised estimates.[24] The denoised estimates are given by the first term $\mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T$ in the right-hand side of eq 15. Since SVD is applied to scaled measurements, the estimates obtained are for the scaled variables, and therefore, the estimates of the original (lagged) variables are obtained as

$$\hat{\mathbf{Z}}_L = \mathbf{U}_1\mathbf{S}_1\mathbf{V}_1^T\hat{\boldsymbol{\Sigma}}_{\mathbf{e}}^{1/2} \tag{20}$$

If the same method is applied to derive the estimates for the EIV dynamic model, then multiple estimates are obtained for each variable at each instant. This is due to the fact that in the lagged data matrix (eq 7), the same variable appears in different columns across different rows. For example, the variable corresponding to $y(k)$ occurs in the first row of the first column, second row of second column and so on. In fact, for a chosen lag L, all input and output variables at time instants k, ..., $N-L+1$ occur L times, whereas input and output variables at time instants $N-L+j$, ..., N occur $L-j+1$ times for $j = 2, ..., L$. However, PCA (and other subspace-based approaches) treats the lagged outputs and inputs as distinct variables giving rise to as many nonunique estimates for a variable as the number of times it occurs in the lagged data matrix. To illustrate this standpoint, consider the simulation of the following second order data generating process (DGP):

$$\begin{aligned}\text{DGP1: } & y^*[k] + 0.2y^*[k-1] + 0.6y^*[k-2] \\ & = 1.2u^*[k-1] + 1.6u^*[k-2]\end{aligned} \tag{21}$$

The process is excited with a full-band white PRBS input signal of length $N = 1023$ observations. White noise is added to the resulting input and output vectors to generate the noisy measurements. The error variances for the input and output variables used in the simulation are 0.0998 and 0.5591, respectively, which corresponds to a signal-to-noise ratio (SNR) of 10. DIPCA is applied to these data to estimate the order, error variances, and model parameters. In the first step, a lag of $L = 15$ is used to estimate the error variances and process order. Starting with a guess value of $d = 15$ (which corresponds to a first order model), the value of d is gradually decremented until the number of unity singular values obtained is equal to the guess value of d. In order to avoid subjectivity, a hypothesis test as described in Appendix B is applied for each guess of d to test whether the smallest d singular values are all equal or not. If the null hypothesis is rejected, then the guess value of d is decremented. The procedure stops when the null hypothesis is not rejected.

The singular values obtained using DIPCA for a guess value of d = 15 and d = 14 are shown in Figures 3 and 4, respectively.

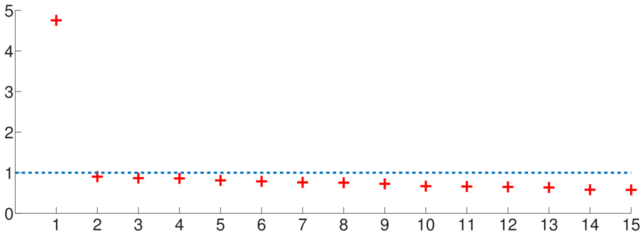


Figure 3. Smallest 15 singular values of lagged data matrix for d = 15.

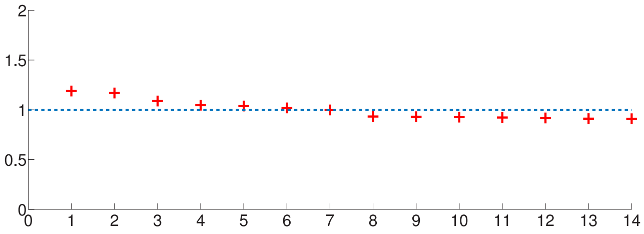


Figure 4. Smallest 14 singular values of lagged data matrix for d = 14.

It is clearly seen that the smallest 15 singular values are not all equal to unity for a guess value of d = 15, whereas the smallest 14 singular values are close to unity for a guess value of d = 14. This is also confirmed by the hypothesis test results shown in Table 1. The null hypothesis is rejected for a guess of d = 15 while it is not rejected for d = 14. Based on this we infer that the correct estimate of d is 14, from which the process order is correctly estimated to be 2 (eq 14). The error variances estimated using DIPCA corresponding to input and output measurements are 0.0905 and 0.6022, respectively, which are also found to be close to their true values.

Table 1. Hypothesis Test Results for DGP1

Guess of d	Degrees of freedom (ν)	Test statistic	Test criterion (α = 0.05)
15	119	2988.4	145.5
14	104	57.1	128.8

In order to estimate the dynamic model parameters, we construct the lagged data matrix using L = 2, since the process order has been estimated to be equal to 2. This lagged data matrix ($\mathbf{Z}_2$) is scaled using the estimated error variances, and the eigenvector corresponding to the smallest singular value of scaled lagged matrix is used to obtain the dynamic model parameters. The estimated dynamic model is as follows:

$$y[k] + \underset{\pm(0.0191)}{0.1994}y[k-1] + \underset{\pm(0.0175)}{0.5999}y[k-2] = \underset{\pm(0.0544)}{0.0043}u[k] + \underset{\pm(0.0540)}{1.1987}u[k-1] + \underset{\pm(0.0558)}{1.6025}u[k-2] \quad (22)$$

The 95% CI given below each parameter estimate in the above equation is obtained using bootstrap simulations of 100 runs. The CIs indicate that the coefficient of $u(k)$ is insignificant, while all other parameter estimates are significant. The CIs for all parameters contain the true values, indicating that the parameter estimates obtained are unbiased.

Following the IPCA procedure outlined in,[24] if we obtain estimates of the input and output variables from the singular value decomposition of $\mathbf{Z}_{2,s}$, we obtain three different estimates for both the variables at all time instants from k = 3 to k = N − 2. Figure 5 shows the plot of two different estimates obtained for the output variable, from the first two columns of the lagged data matrix. Besides the fact that nonunique estimates are obtained, the estimates obtained using the PCA-based approach cannot be considered as one step ahead predicted estimates, because they are derived by using all the measurements in the entire sample. These estimates can at best be regarded as smoothed estimates. It may also be noted that a standard Kalman filter cannot be used as in the non-EIV based model identification procedure to generate the one step ahead predictions and residuals, since the inputs are also noisy in the EIV case. We, therefore, need a technically correct way of generating optimal filtered unique estimates of input and output variables for the EIV dynamic model.

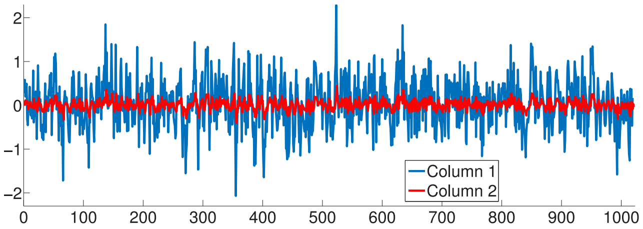


Figure 5. Nonunique estimates of output variable obtained using IPCA procedure.

4.2. EIV Kalman Filter for Deriving Optimal Estimates.

Given a linear EIV dynamic model with known error variances, an EIV Kalman filter (EIV-KF) was proposed in[25] for obtaining the optimal filtered estimates. With the dynamic model and error variances estimated using DIPCA, we propose the use of EIV-KF for generating the filtered estimates as well as corresponding prediction error or residuals.

In order to apply the EIV Kalman filter, a state-space representation of the dynamic model has to be used. A generic causal dynamic process described by the DE model

$$y^*[k] + \sum_{i=1}^{\eta} a_i y^*[k-i] = \sum_{j=0}^{\eta} b_j u^*[k-j] \quad (23)$$

together with the measurement equation for $y[k]$ in eq 3 can be represented in state-space form as follows:

$$\mathbf{x}[k] = \mathbf{A}\mathbf{x}[k-1] + \mathbf{B}u^*[k-1] \quad (24)$$

$$y[k] = \mathbf{C}\mathbf{x}[k] + \mathbf{D}u^*[k] + e_y[k] \quad (25)$$

For a controller canonical form, the state space matrices are defined as

$$\mathbf{A} = \begin{bmatrix} -a_1 & -a_2 & \dots & -a_\eta \\ 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 \\ \vdots & \vdots & \dots & \vdots \\ 0 & \dots & 1 & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} b_1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

$$\mathbf{C} = [b_1 - a_1 b_0 \quad b_2 - a_2 b_0 \quad \dots \quad b_\eta - a_\eta b_0] \quad \mathbf{D} = b_0$$

Since the true inputs are unknown, we rewrite the state-space model of eq 24 in terms of the known measurements using eq 2 to obtain the state equation in standard form:

$$\mathbf{x}[k] = \mathbf{A}\mathbf{x}[k-1] + \mathbf{B}u[k-1] + \mathbf{w}[k-1] \tag{26}$$

$$y'[k] = y[k] - Du[k] = \mathbf{C}\mathbf{x}[k] + e_{y'}[k] \tag{27}$$

where the process noise vector $\mathbf{w}[k-1]$ in eq 26 is defined as $\mathbf{w}[k-1] = -\mathbf{B}e_u[k-1]$ and the measurement noise in modified measurements of eq 27 is defined as $e_{y'}[k] = -\mathbf{D}e_u[k] + e_y[k]$. It may be noted that the above procedure can also be applied to the estimated dynamic model using the estimated model parameters to obtain a state space model in standard form.

Since eqs 26 and 27 are in the standard form, a Kalman filter can be applied to obtain the optimal filtered state estimates, $\hat{\mathbf{x}}[k|k]$ and its error covariance matrix $\mathbf{P}[k|k]$ in a recursive manner. For ease of reference, the EIV Kalman filter equations are given in Appendix A.

The EIV Kalman filter innovations which represent the one step ahead prediction errors are obtained as

$$\nu[k] = y'[k] - \mathbf{C}\hat{\mathbf{x}}[k|k-1] \tag{28}$$

Equation 28 together with the state space model and DIPCA provide the complete package for generation of optimal residuals, estimates of model parameters and noise variances.

In order to demonstrate the proposed DIPCA-EIV Kalman filter, consider again the process described by eq 21 for which measurements are simulated as described earlier. The dynamic model estimated using DIPCA is given by eq 22. From this estimated model, we eliminate the term corresponding to $u[k]$, since its coefficient is insignificant. The state-space model corresponding to this estimated DE model is defined by the following matrices

$$\mathbf{A} = \begin{bmatrix} -0.1994 & -0.5999 \\ 1.0 & 0 \end{bmatrix} \quad \mathbf{B} = \begin{bmatrix} 1.0 \\ 0 \end{bmatrix}$$

$$\mathbf{C} = [1.1987 \quad 1.6025] \quad \mathbf{D} = [0]$$

The covariance matrix of the process noise vector is obtained by using the estimated input error variance and is given by $0.0905\mathbf{B}\mathbf{B}^T$, and the measurement noise variance obtained from the estimated output error variance is equal to 0.6022.

EIV-KF is applied to the above state space model. For this purpose, an initial estimate of the state vector and its error covariance matrix has to be specified. The initial estimates of all states are assumed to be unity (which are significantly different from the true values which are zero), and the initial estimate error covariance matrix ($\mathbf{P}_0$) is taken as an identity matrix (which is an order of magnitude larger than the process noise and measurement noise variances to ensure that the effect of an incorrect initial state estimate dies down quickly). The EIV-KF is applied to obtain the one step ahead predicted estimate of the output at each time instant. Figure 6 shows the filtered estimate and corresponding measured values of the output, while Figure 7 shows the residuals corresponding to these estimates.

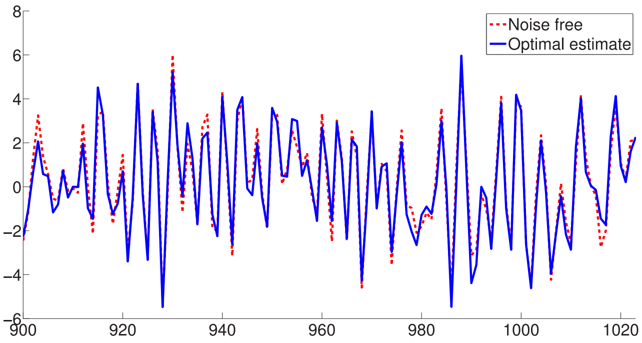


Figure 6. True value of output and its filtered estimate using EIV KF.

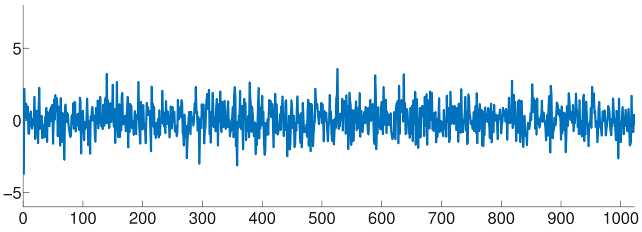


Figure 7. Residuals (innovations) obtained from the EIV KF.

The filtered estimates reported in Figure 6 clearly demonstrate the effectiveness of the proposed DIPCA-EIVKF in estimating the residuals for the EIV case. It is useful to contrast these results with those in Figure 5. We obtain estimates that are not only unique (unlike those from the DIPCA) but also optimal by virtue of the Kalman filter properties.

The residuals, reported in Figure 7, can be used in a variety of applications. Two widely encountered applications are in system identification and fault detection. In the ensuing section we demonstrate the use of residual analysis to the former application, i.e., model development.

5. RESIDUAL ANALYSIS USING DIPCA-EIV KALMAN FILTER

This section demonstrates the use of residual analysis (using DIPCA-EIV Kalman filter) in testing model adequacy. One of the major challenges in such an exercise is the estimation of the process order. As discussed in Section 3, DIPCA uses a theoretical criterion based on singular value analysis to estimate the process order. Further, it has been proven[17] that if the error variances are known, the process order is consistently estimated in DIPCA. In the case when error variances are also simultaneously estimated along with process order and model parameters, simulation studies seem to indicate that these estimates converge to their respective true values as the sample size increases. However, theoretical proof of this is still lacking. Furthermore, for small sample sizes and for high order processes, it has been observed that the smallest singular values of the scaled data matrix (which are supposed to be unity) depart significantly from unity, which can cause ambiguity in estimating the process order correctly.

Model adequacy testing including order determination is a standard critical step in non-EIV or classical model identification methods (e.g., prediction error minimization (PEM) methods) as well. This holds for both scenarios, whether the order is user-specified or generated through some other semiautomatic method. In this regard, two standard diagnostic tools, namely, the cross-correlation between residuals and inputs, and the autocorrelation of the residuals, are used for determining the adequacy of plant (deterministic) and noise models, respectively. However, if the noise corrupting the output is known to be white and the input has colored noise characteristics, it is sufficient to test for whiteness of residuals, i.e., examine the autocorrelation function of residuals, for determining the adequacy of the plant model. If the residuals fail the whiteness test, the model is reidentified for a different choice of guess parameters such as model order or delay.

In EIV identification, especially using DIPCA, all model parameters including order, delay, error variances etc. are estimated from the data, and the user need not specify a guess of these parameters. The most important parameter is the process order which is estimated by examining the singular values of the scaled lagged data matrix in the first step of DIPCA. For large sample sizes, the order determination step involving examination of the d singular values of the lagged matrix is fairly accurate. Nevertheless, the residual analysis step is recommended as a corroborative tool. On the other hand, if the sample size is small, then it is possible that, in the first step of DIPCA, the smallest d singular values of the lagged matrix may not be all be close to unity, even if the number of constraints d is guessed correctly. This may cause an ambiguity in estimating the correct number of constraints and, consequently, lead to an ambiguity in estimating the process order. In such cases, we demonstrate below, by way of residual analysis, that the autocorrelation of the residuals obtained using eq 28 can be used to resolve the ambiguity.

The autocorrelation function of the residuals along with the CI is shown in Figure 8. From the significance levels and estimates, the null hypothesis that the residuals are white is not rejected and hence the estimated process order and parameters are deemed satisfactory. In this scenario, if one were to proceed systematically starting with a small guess for model order (preferably 1) and incrementally increase the guessed order in steps of 1, the EIV-DIPCA Kalman filter serves as a confirmatory tool for correct order determination based on the singular values of the lagged data matrix.

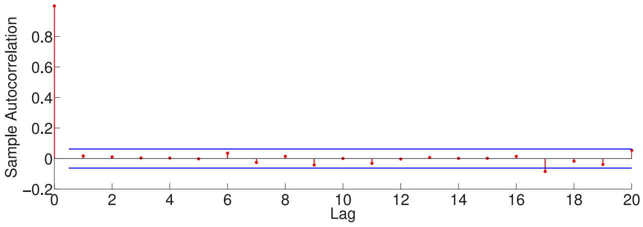


Figure 8. Autocorrelation function of residuals obtained using EIV-KF.

If the sample size is small and/or the process order is high, it is possible that the process order cannot be unambiguously estimated by analyzing the singular values of the lagged data matrix. In order to demonstrate this, consider the following fifth order data generating process (DGP):

$$\begin{aligned}\text{DGP2: } & y^*[k] + 0.2y^*[k-1] + 0.6y^*[k-5] \\ & = 1.2u^*[k-1] + 1.6u^*[k-2] \end{aligned} \tag{29}$$

This process is simulated using a full-band white PRBS input signal of length $N = 256$. Measurements are generated by adding white noise to the true input and output sequences such that the signal-to-noise ratio (SNR) is maintained at 10. DIPCA is applied to the data using a lag of 15 in the first step. The value of d is initially guessed as $d = L = 15$ (corresponding to a first order model) and is gradually decremented. For each guess of d the hypothesis test is applied to verify whether the smallest d eigenvalues are equal. The results of the hypothesis test are reported in Table 2. The null hypothesis is rejected for guesses of d = 15, 14, 13, 12, but is not rejected for d = 11. Based on the hypothesis test the estimated d is 11, and hence the estimated process order is $\hat{\eta} = 15{-}11 + 1 = 5$, which is the true process order. A plot of the smallest 11 eigenvalues obtained for a guess value of d = 11 is shown in Figure 9. The figure shows that the smallest 11 eigenvalues are not all equal to unity. However, it may be noted that by using the hypothesis test we are still able to estimate the correct process order. We can further confirm this by performing residual analysis using the EIV-KF step as described earlier in this section.

Table 2. Hypothesis Test Results for DGP2

Guess of d	Degrees of freedom (ν)	Test statistic	Test criterion (α = 0.05)
15	119	614.7	145.5
14	104	378.5	128.8
13	90	303.5	113.1
12	77	170.1	98.5
11	65	69.6	84.8
10	54	56.7	72.2

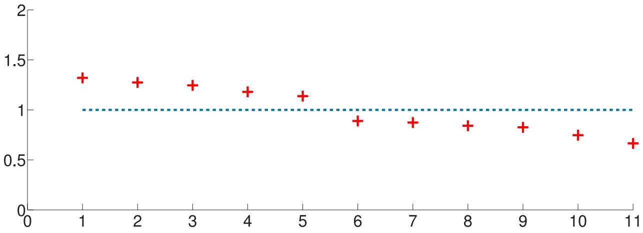


Figure 9. Smallest 11 eigenvalues obtained for DGP2 using a lag of 15 and d = 11.

We estimate the parameters for a 5th order dynamic model in the second step of DIPCA. The following model is obtained.

$$\begin{aligned} & y[k] + 0.2025y[k-1] + 0.0270y[k-2] \\ & \quad + 0.0283y[k-3] + 0.0348y[k-4] + 0.6322y[k-5] \\ & = 0.0148u[k] + 1.1419u[k-1] + 1.5229u[k-2] \\ & \quad - 0.0181u[k-3] + 0.0784u[k-4] - 0.0041u[k-5] \end{aligned}$$

EIV-KF is applied to this model to obtain the one step ahead prediction errors, and the corresponding autocorrelation function is shown in Figure 10. The autocorrelation function clearly shows that the residuals are white. This further confirms that a 5th order process model is acceptable.

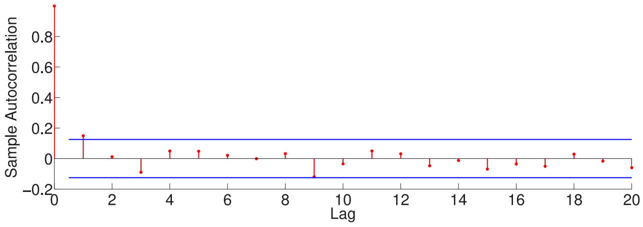


Figure 10. ACF of residuals obtained for DGP2 using an estimated 5^{th} order model.

Residual analysis is especially useful, if the hypothesis test used in DIPCA rejects the null hypothesis prematurely, thereby underestimating the process order. In order to demonstrate the efficacy of residual analysis, suppose (incorrectly) that the null hypothesis is not rejected for a guess of d = 12, and the corresponding process order is therefore underestimated as 15−12 + 1 = 4. The resulting plot of the smallest 12 eigenvalues is shown in Figure 11. A visual examination of this plot also indicates that the highest eigenvalue among them departs from unity significantly, and the guessed number of constraints may be incorrect. Nevertheless, we proceed with the estimated 4th order model and the corresponding model parameters are obtained as follows.

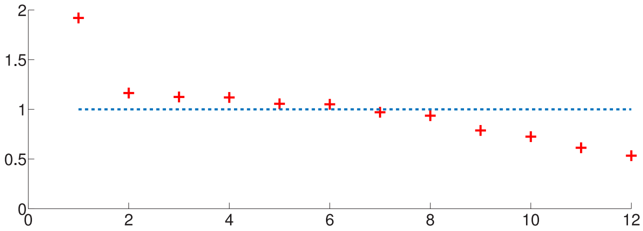


Figure 11. Smallest 12 eigenvalues obtained for DGP2 using a lag of 15 and d = 12.

$$y[k] - 0.7398y[k-1] + 0.6142y[k-2] - 0.3686y[k-3] + 0.5278y[k-4]$$

$$= -0.0245u[k] + 1.1600u[k-1] + 0.4437u[k-2]$$

$$- 0.5630u[k-3] + 0.6447u[k-4]$$

The EIV-KF approach is applied to the estimated model to compute the one step ahead prediction errors (residuals). The autocorrelation function of the residuals is shown in Figure 12. The plot indicates that the residuals are nonwhite which is a signature of the model order being estimated incorrectly. Thus, in case the hypothesis test in DIPCA underestimates the model order, residual analysis can be used to detect this error and model reidentification can be done using a higher order. On the other hand, if the hypothesis test used in DIPCA rejects the null hypothesis for the correct estimate of d, we may end up overestimating the process order. In order to demonstrate this, we again incorrectly assume that for this process the null hypothesis for d = 11 is rejected whereas for d = 10 it is not rejected (Figure 13). The process order is incorrectly estimated as 6 as a consequence. We proceed with DIPCA by reconfiguring the data matrix using the estimated sixth order model and obtain the dynamic model parameters. The overfit model obtained is reported below:

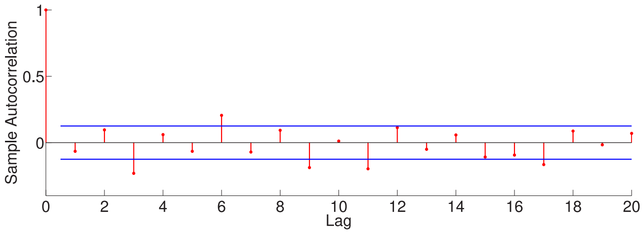


Figure 12. ACF of residuals obtained for DGP2 using an estimated 4^{th} order model.

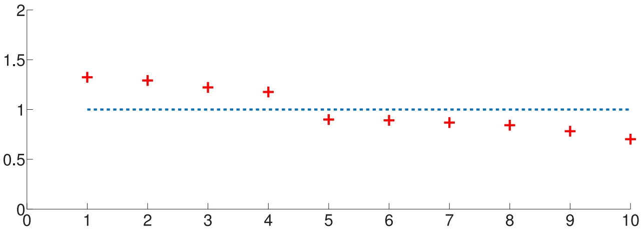


Figure 13. Smallest 10 eigenvalues obtained for DGP2 using a lag of 15 and d = 10.

$$y[k] - 0.8169y[k-1] - 0.1579y[k-2] - 0.0801y[k-3] + 0.0803y[k-4]$$

$$+ 0.5678[k-5] - 0.6083[k-6] = 0.0566u[k] + 1.1622u[k-1] + 0.2708u[k-2]$$

$$- 1.5696u[k-3] + 0.0145u[k-4] - 0.0606u[k-5] + 0.1052u[k-6]$$

EIV-KF is applied to the above estimated model in eq 5, in order to obtain the one step prediction errors of the output. The autocorrelation function of the residuals is shown in Figure 14. The autocorrelation function plot indicates that the residuals are white, a result that corroborates the overfit model. This behavior is consistent with residual analysis for classical (non-EIV) model identification, where an underfit model can be detected using autocorrelation of residuals, whereas an overfit model cannot be detected.

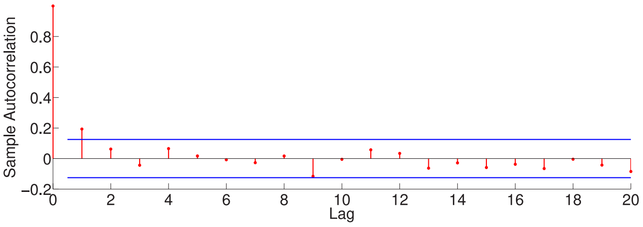


Figure 14. ACF of residuals obtained for DGP2 using an estimated 6th order model.

6. CASE STUDY: TWO NONINTERACTING TANK SYSTEM

In order to test the performance of the proposed method on a physical system, we consider the system of two noninteracting tanks in series that was considered in Maurya et al.[17] A schematic of the system is shown in Figure 15.

An approximate linear dynamic model of the nonlinear system is estimated using DIPCA, and the efficacy of the

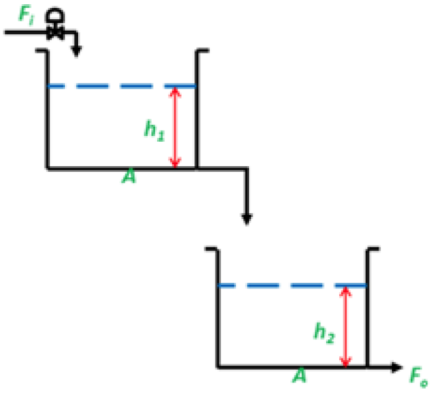


Figure 15. Two tank noninteracting system.

proposed method in obtaining optimal filtered estimates of the output variables and establishing model correctness is shown.

A nonlinear first-principles model of the liquid level system that describes the changes in tank levels $h_2(t)$ with respect to inlet flow rate $F_i(t)$ is given by

$$\frac{dh_1(t)}{dt} + \frac{Cv_1}{A_1}\sqrt{h_1(t)} = \frac{1}{A_1}F_i(t) \tag{30a}$$

$$\frac{dh_2(t)}{dt} + \frac{Cv_2}{A_1}\sqrt{h_2(t)} = \frac{Cv_1}{A_2}\sqrt{h_1(t)} \tag{30b}$$

The system is brought to a steady state before exciting it with the designed input. The operating conditions are as follows:

$$Cv_1 = 1.8 \quad Cv_2 = 0.5 \quad A_1 = 2.4 \quad A_2 = 1.2$$

$$h_{1ss} = 1.23 \quad h_{2ss} = 16$$

The input, $F_i(t)$, used for simulation is PRBS signal around the nominal operating point. Gaussian white noises with variances 0.047 and 0.004 are added to the noise-free output and input data sampled at 1 s intervals, respectively, to generate 2047 noisy measurements. The variances are chosen such that SNR is maintained as 10. A snapshot of the noisy input-output data is shown in Figure 16.

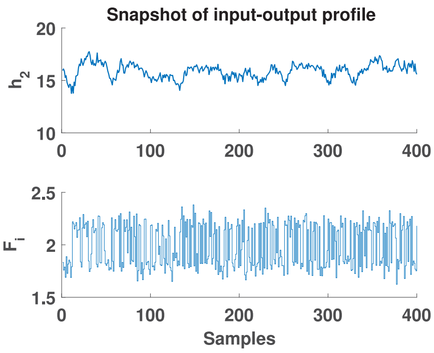


Figure 16. A snapshot of input-output data.

DIPCA algorithm is used for model identification with a stacking order of $L = 5$. We start with a guess of $d = 5$ and apply the hypothesis test as described in Appendix B to estimate its correct value. The results of the hypothesis test are shown in Table 3. The estimated value of d is 4 (since the null hypothesis is rejected for $d = 5$ and not rejected for $d = 4$). The estimated process order is therefore $5 - 4 + 1 = 2$, which is consistent with the true process order.

Table 3. Hypothesis Test Results for Two Tank Process[a]

Guess of d	Degrees of freedom (ν)	Test statistic	Test criterion ($\alpha = 0.05$)
5	14	132.5	23.7
4	9	5.1	16.9

[a] The 2nd order linear model estimated using DIPCA is given by.

$$y^*[k] - 1.3626y^*[k-1] + 0.4116y^*[k-2] = 0.4620u^*[k-1] + 0.3633u^*[k-2] \tag{31}$$

The autocorelation function of the residuals and the filtered estimates for the output obtained corresponding to this model are shown in Figures 17 and 18, respectively.

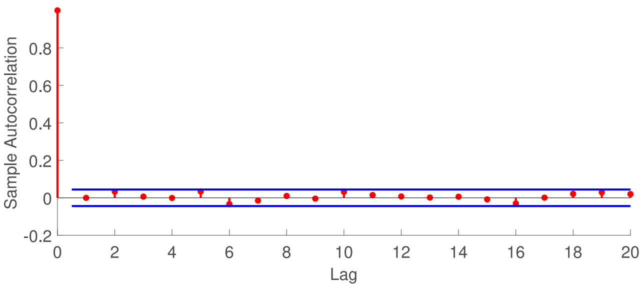


Figure 17. ACF of residuals for model order = 2.

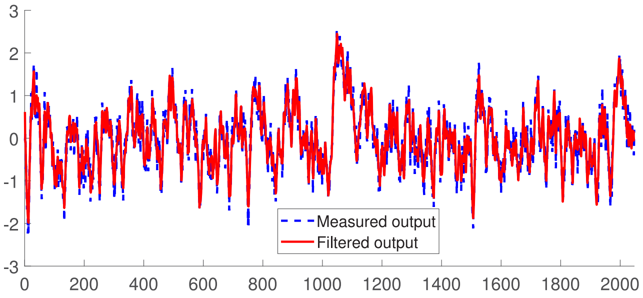


Figure 18. Filtered and measured values for the output for model order = 2.

Since the autocorrelation function is white, the second order model is acceptable. On the other hand, if the model is estimated incorrectly as first order using DIPCA, then we obtain the following first order linear model,

$$y^*[k] - 0.9895y^*[k-1] = 0.5063u^*[k-1] \tag{32}$$

The corresponding autocorrelation function of the residuals and the filtered estimates are shown in Figures 19 and 20, respectively. The autocorrelation function is not white indicating that a first order model is unacceptable. This simulation study clearly demonstrates the efficacy of the proposed method on a realistic process.

7. CONCLUSIONS

The paper addressed the problem of obtaining optimal filtered estimates and hence an optimal method of residual generation from errors-in-variables class of models. For this purpose, we developed a method that combines the EIV-Kalman Filter and EIV model identification method. The DIPCA-based identi-

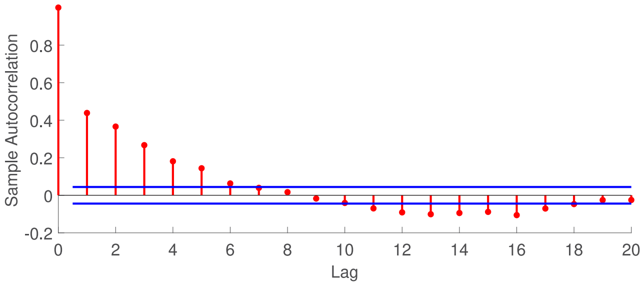


Figure 19. ACF of residuals for model order = 1.

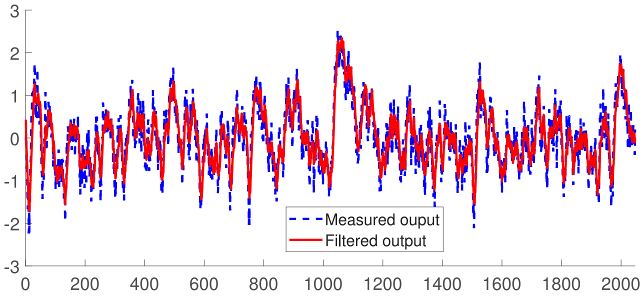


Figure 20. Filtered and measured values for the output for model order = 1.

fication method was specifically chosen for illustrating the ideas. The contributions of this work are to be considered significant in a few important respects. First, the problem of residual generation and analysis in the EIV case has hardly been investigated before. Second, residuals obtained from dynamic PCA-based EIV identification methods (or its variants) are nonunique, thereby giving rise to nonunique estimates. The third aspect stems from the utility of an optimal residual generation method in model validation and fault detection/diagnosis. In particular, the use of EIV-KF residual analysis in EIV model development has been demonstrated. The proposed filter has been shown to be a useful tool for statistically detecting underfitting (under determination of process order), while serving as a confirmatory tool for models that adequately capture the process dynamics. This can be especially useful when the sample size is small and the process order is high. In this respect, the work has essentially led to the development of a tool and a method that can be termed the equivalent of standard model validation tools available for non-EIV identification.

Though we have demonstrated this approach for a SISO case, this approach can be easily extended for the MIMO case as well. The EIV-KF equations for the MIMO case have been provided by Diversi et al.[25] These equations along with a MIMO state-space identification algorithm can be used to obtain filtered estimates of the variables, and the associated residuals. However, in the MIMO case a vector of residuals has to be analyzed for model diagnostics.

■ APPENDIX A

EIV Kalman Filter

The EIV Kalman filter (EIV-KF) equations for a multi-input multi-output (MIMO) process were derived in ref 25. For a single-input single-output (SISO) system, under the additional assumption that the errors in input and output measurements are uncorrelated, the simplified EIV-KF equations are presented here.

For a process described by the state-space model given by the eqs 26 and 27, the optimal one step ahead predicted estimates of the state variables and the corresponding covariance matrix of error in the state estimates are given by

$$\hat{\mathbf{x}}[k|k-1] = \mathbf{A}\hat{\mathbf{x}}[k-1|k-1] + \mathbf{B}u[k-1] \tag{33}$$

$$\mathbf{P}[k|k-1] = \mathbf{A}\mathbf{P}[k-1|k-1]\mathbf{A}^T + \mathbf{Q} \tag{34}$$

where $\hat{\mathbf{x}}[k|k-1]$ is the one step ahead predicted estimates of the states at time k using all measurements until time $k-1$, $\hat{\mathbf{x}}[k-1|k-1]$ is the filtered estimate of the state at time instant $(k-1)$ using all measurements until time $k-1$, and $\mathbf{P}[k-1|k-1]$ is the corresponding covariance matrix of errors in the estimates. The process noise covariance matrix $\mathbf{Q}$ is given by

$$\mathbf{Q} = \sigma_{e_u}^2\mathbf{B}\mathbf{B}^T \tag{35}$$

The optimal filtered estimates of the states $\hat{\mathbf{x}}[k|k]$ at time k and their corresponding error covariance matrix $\mathbf{P}[k|k]$ are given by

$$\hat{\mathbf{x}}[k|k] = \hat{\mathbf{x}}[k|k-1] + \mathbf{K}[k]\nu[k] \tag{36}$$

$$\nu[k] = y[k] - \mathbf{D}u[k] - \mathbf{C}\hat{\mathbf{x}}[k|k-1] \tag{37}$$

$$\mathbf{P}[k|k] = \mathbf{P}[k|k-1] - (\mathbf{A}\mathbf{P}[k|k-1]\mathbf{C}^T + \mathbf{S}) \times (\mathbf{A}\mathbf{P}[k|k-1]\mathbf{C}^T + \mathbf{S})^T/\sigma_\nu^2[k] \tag{38}$$

The cross covariance matrix between the process noise and measurement noise, $\mathbf{S}$, the error variance in modified measurement $\sigma_{y'}^2$, and the error variance of the innovations σ_ν^2 are given by

$$\mathbf{S} = \sigma_u^2\mathbf{B}\mathbf{D}^T \tag{39}$$

$$\sigma_\nu^2[k] = \mathbf{C}\mathbf{P}[k|k-1]\mathbf{C}^T + \sigma_{e_{y'}}^2 \tag{40}$$

$$\sigma_{e_{y'}}^2 = \sigma_{e_y}^2 + \sigma_{e_u}^2\mathbf{D}\mathbf{D}^T \tag{41}$$

The Kalman gain matrix $\mathbf{K}[k]$ is given by

$$\mathbf{K}[k] = (\mathbf{A}\mathbf{P}[k|k-1]\mathbf{C}^T + \mathbf{S})/\sigma_\nu^2[k] \tag{42}$$

Due to the assumptions made on the statistical properties of the measurement errors in our work, the optimal filtered estimates of the input and output variables derived in ref 25 simplify to the following equations

$$\hat{u}[k|k] = u[k] - \frac{\sigma_{e_u}^2}{\sigma_\nu^2}\mathbf{D}^T\nu[k] \tag{43}$$

$$\hat{y}[k|k] = y[k] - \frac{\sigma_{e_y}^2}{\sigma_\nu^2}\nu[k] \tag{44}$$

In processes where there is no direct feedback of the input, that is, $\mathbf{D} = 0$, the above equations reduce to the standard KF equations, with the process noise covariance matrix defined by 35.

■ APPENDIX B

Hypothesis Test for Equality of Eigenvalues

This Appendix provides a brief description of the hypothesis test used for testing the equality of eigenvalues of the covariance matrix of lagged samples at each stage of DIPCA.

The results described in section 3.7.3 of Joliffe's book[27] are utilized for this purpose.

Let $\lambda_{L,i}$, $i = 1, ..., 2L + 2$ be the ordered eigenvalues (from largest to smallest) of the population covariance matrix of the lagged data vector $z^{*}_{L,s}$, and let $l_{L,i}$ be the corresponding ordered eigenvalues obtained from the sample covariance matrix of scaled measurements $z_{L,s}$. It may be noted that the covariance matrix of errors in scaled measurements is an identity matrix.

In DIPCA if the guessed value of d is correct, the smallest d eigenvalues of the covariance matrix of lagged measurements should all be equal to unity. Based on the eigenvalues of the sample covariance matrix of lagged measurements, the equality of the last d eigenvalues is tested using the following hypothesis.

$$H_0 : \lambda_{L,2L+3-d} = \lambda_{L,2L+4-d}\cdot \ldots = \lambda_{L,2L+2}$$

$$H_1 : \text{ at least two of smallest } d \text{ eigenvalues are unequal}$$

The statistic used to test the above hypothesis is given by

$$\tau = n'\left[d \ln \bar{l} - \sum_{j=2L+3-d}^{j=2L+2} \ln l_{L,j}\right] \tag{45}$$

where

$$\bar{l} = \frac{1}{d} \sum_{j=2L+3-d}^{j\,=\,2L\,+\,2} l_{L,j} \tag{46}$$

$$n' = (N - L) - (4L + 15)/6 \tag{47}$$

The null hypothesis is rejected if the above test statistic exceeds $\chi^2_{\nu,\alpha}$, the value at α level of significance drawn from a chi-square distribution with ν degrees of freedom. The degrees of freedom is given by

$$\nu = \frac{1}{2}(d + 2)(d - 1) \tag{48}$$

It may be noted that the above test has been derived under the assumption that the true values of variables follow a normal distribution and the error covariance matrix is known. However, in our case the elements of the error covariance matrix are estimated from the data. Furthermore, the true values are assumed to be arbitrary deterministic quantities and not assumed to follow a multivariate normal distribution. Despite these differences the test has been used in our work to remove subjectivity in determining whether the smallest d eigenvalues are equal are not.

■ AUTHOR INFORMATION

Corresponding Authors

Vipul Mann − *Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600036, India;* Email: vipulmann2310@gmail.com

Deepak Maurya − *Department of Computer Science, Indian Institute of Technology Madras, Chennai 600036, India;* Email: ee11b109@ee.iitm.ac.in

Arun K. Tangirala − *Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600036, India;* orcid.org/0000-0002-7921-5340; Email: arunkt@iitm.ac.in

Shankar Narasimhan − *Department of Chemical Engineering, Indian Institute of Technology Madras, Chennai 600036, India;* orcid.org/0000-0002-0558-0206; Email: naras@iitm.ac.in

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.iecr.9b04561

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

The authors acknowledge partial financial support from Robert Bosch Centre for Data Science and Artificial Intelligence (RBCDSAI), IIT Madras, Chennai for this work. We also wish to thank Lokesh Rajulapati, a project associate at RBCDSAI, for implementing the hypothesis tests.

■ REFERENCES

(1) Hofmann, D. Common sources of errors in measurement systems. *Handbook of measuring system design*; 2005.

(2) Yu, H.; Khan, F.; Garaniya, V. An alternative formulation of PCA for process monitoring using distance correlation. *Ind. Eng. Chem. Res.* **2016**, *55*, 656−669.

(3) Ge, Z.; Song, Z. Process monitoring based on independent component analysis- principal component analysis (ICA- PCA) and similarity factors. *Ind. Eng. Chem. Res.* **2007**, *46*, 2054−2063.

(4) Dunia, R.; Qin, S. J.; Edgar, T. F.; McAvoy, T. J. Identification of faulty sensors using principal component analysis. *AIChE J.* **1996**, *42*, 2797−2812.

(5) Deng, X.; Deng, J. Incipient fault detection for chemical processes using two-dimensional weighted SLKPCA. *Ind. Eng. Chem. Res.* **2019**, *58*, 2280−2295.

(6) Du, Y.; Budman, H.; Duever, T. A.; Du, D. Fault Detection and Classification for Nonlinear Chemical Processes using Lasso and Gaussian Process. *Ind. Eng. Chem. Res.* **2018**, *57*, 8962−8977.

(7) Ljung, L. *System Identification - A Theory for the User*; Prentice Hall International: Upper Saddle River, NJ, USA, 1999.

(8) Soderstrom, T.; Stoica, P. System Identification; *Prentice Hall International Series In Systems And Control Engineering*; Prentice Hall: 1989.

(9) Tangirala, A. K. *Principles of System Identification: Theory and Practice*; CRC Press, Taylor & Francis Group: Boca Raton, FL, USA, 2014.

(10) Juricek, B. C.; Seborg, D. E.; Larimore, W. E. Identification of multivariable, linear, dynamic models: Comparing regression and subspace techniques. *Ind. Eng. Chem. Res.* **2002**, *41*, 2185−2203.

(11) Qin, S. J. An overview of subspace identification. *Comput. Chem. Eng.* **2006**, *30*, 1502−1513.

(12) Soderstrom, T. Errors-in-variables methods in system identification. *Automatica* **2007**, *43*, 939−958.

(13) Söderström, T. *Errors-in-Variables Methods in System Identification*; Springer International Publishing: 2018.

(14) Wu, P.; Pan, H.; Ren, J.; Yang, C. A new subspace identification approach based on principal component analysis and noise estimation. *Ind. Eng. Chem. Res.* **2015**, *54*, 5106−5114.

(15) Hou, J.; Chen, F.; Li, P.; Zhu, Z.; Liu, F. An Improved Consistent Subspace Identification Method Using Parity Space for State-space Models. *International Journal of Control, Automation and Systems* **2019**, *17*, 1167−1176.

(16) Ku, W.; Storer, R. H.; Georgakis, C. Disturbance detection and isolation by dynamic principal component analysis. *Chemom. Intell. Lab. Syst.* **1995**, *30*, 179−196.

(17) Maurya, D.; Tangirala, A. K.; Narasimhan, S. Identification of Linear Dynamic Systems using Dynamic Iterative Principal Component Analysis. *Ind. Eng. Chem. Res.* **2018**, *49*, 1014−1019.

(18) Vijaysai, P.; Gudi, R.; Lakshminarayanan, S. Errors-in-variables-based modeling using augmented principal components. *Ind. Eng. Chem. Res.* **2005**, *44*, 368−380.

(19) Li, W.; Qin, S. J. Consistent dynamic PCA based on errors-in-variables subspace identification. *J. Process Control* **2001**, *11*, 661−676.

(20) Luo, L.; Bao, S.; Tong, C. Sparse robust principal component analysis with applications to fault detection and diagnosis. *Ind. Eng. Chem. Res.* **2019**, *58*, 1300−1309.

(21) Lou, Z.; Shen, D.; Wang, Y. Two-step principal component analysis for dynamic processes monitoring. *Can. J. Chem. Eng.* **2018**, *96*, 160−170.

(22) Chen, J.; Wang, W.-Y. PCA- ARMA-Based Control Charts for Performance Monitoring of Multivariable Feedback Control. *Ind. Eng. Chem. Res.* **2010**, *49*, 2228−2241.

(23) Narasimhan, S.; Shah, S. L. Model identification and error covariance matrix estimation from noisy data using PCA. *Control Engineering Practice* **2008**, *16*, 146−155.

(24) Narasimhan, S.; Bhatt, N. Deconstructing principal component analysis using a data reconciliation perspective. *Comput. Chem. Eng.* **2015**, *77*, 74−84.

(25) Diversi, R.; Guidorzi, R.; Soverini, U. Kalman filtering in extended noise environments. *IEEE Trans. Autom. Control* **2005**, *50*, 1396−1402.

(26) Huang, B. Process identification based on last principal component analysis. *J. Process Control* **2001**, *11*, 19−33.

(27) Jolliffe, I. *Principal component analysis*; Springer Verlag: New York, 2002.