Int J Adv Eng Sci Appl Math (March–June 2012) 4(1–2):91–115
DOI 10.1007/s12572-012-0059-x

Data reconciliation and gross error analysis of self powered neutron detectors: comparison of PCA and IPCA based models

Rihab Abdul Razak · Mani Bhushan · Madhu N. Belur · Akhilanand P. Tiwari · Mahendra G. Kelkar · Mahitosh Pramanik

Published online: 26 June 2012
© Indian Institute of Technology Madras 2012

Abstract Self powered neutron detectors (SPNDs) are used for measuring neutron flux in a nuclear reactor and hence are essential for safe operation of the reactor. The main objective in this paper is to develop models on representative groups of SPNDs using principal component analysis (PCA) based techniques. In particular, we identify models using regular PCA and the iterative PCA (IPCA) technique proposed in Narasimhan and Shah (Control Eng Pract 16:146–155, 2008). These models are then used for performing data reconciliation, gross error detection, identification and estimation for the SPNDs. Based on these performances, we compare the PCA and IPCA models. The proposed models can be used in the reactor to ensure that faulty detectors can be detected quickly and the corresponding values estimated in real-time thereby ensuring continuous reactor operation.

Keywords Principal component analysis · Iterative PCA · Model identification · Gross errors · SPND

R. A. Razak · M. N. Belur
Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai, India

M. Bhushan (✉)
Department of Chemical Engineering, Indian Institute of Technology Bombay, Mumbai, India
e-mail: mbhushan@iitb.ac.in

A. P. Tiwari
Reactor Control Division, Bhabha Atomic Research Centre, Mumbai, India

M. G. Kelkar · M. Pramanik
Nuclear Power Corporation of India, Mumbai, India

1 Introduction

Self powered neutron detectors (or SPNDs) are widely used to measure neutron flux in a nuclear reactor [2, 16]. A typical nuclear reactor contains several such detectors for measuring neutron flux at different locations in the reactor core. Measurement of flux is important for various reasons. Neutron flux provides a direct measure of the reactor output power and thus measurements from the flux detectors are used for safety, regulatory and control purposes [16]. Flux measurements from various detectors at different locations in the reactor can be used to compute flux and power map for the reactor which is a continuous map showing the neutron flux variations in the reactor core. Thus, healthy operation of the SPNDs is important for safe reactor operation.

Over a period of time, these detectors can develop faults, for example due to drop in insulation resistance, fracture of the detector sheath [3] etc. Further, hardware or software issues involving the data-collection and archiving systems can also cause faulty readings to be recorded. In such cases, if the fault is not detected, then erroneous inferences about reactor operation may be made. If on the other hand the fault is manually detected, then the usual practice is to rectify the fault as soon as possible and this may involve unscheduled reactor shutdown which will be time consuming and have cost and personnel safety implications. Our objective in the current work is to identify models for groups of detectors using their measurement data which can be used for fault detection and diagnosis. These models are extracted using principal component analysis (PCA) based techniques which rely on the eigen-decomposition of the covariance matrix of the detector measurements. In literature, use of PCA and other data based techniques for fault detection and diagnosis in nuclear systems in general are described in [5, 7, 8, 17, 18, 20, 21]. With regard to SPNDs in particular, use of correlation based methods has been described in [1] in which the authors use the correlation between SPNDs as a basis for validation of SPND signals. Despite the widespread usage and relevance of PCA techniques for fault detection and diagnosis in nuclear reactors, to the best of our knowledge, their use for SPND fault detection and diagnosis has not been reported.

In this work, we develop PCA based linear models relating SPNDs of a large nuclear reactor. This exercise is carried out for two groups of detectors: one group containing detectors with similar dynamics, while the other group contains detectors with distinctly different dynamics. We use two variants of PCA: conventional PCA and iterative PCA (IPCA) for the same. While PCA based modeling techniques have been widely used over last several years, the resulting models are known to vary with the scaling of the data. IPCA [12] is a recent variation of PCA which involves iteratively re-scaling the data so that the underlying models can be correctly identified. The iterative IPCA procedure requires solving a nonlinear optimization problem at every iteration. While theoretically elegant, its implementation for extracting models from real data (which usually does not exactly satisfy various assumptions made by different techniques) has not been reported. Hence, apart from development of useful models for SPND fault detection and diagnosis, another outcome of our work is the implementation of IPCA on real data and identification of related issues. The models obtained by PCA and IPCA in our work are compared with regards to their performance for detecting, identifying and estimating gross errors in individual detectors for various cases.

The paper is structured as follows: in Sect. 2 we summarize the available data based on which the analysis has been performed. In Sect. 3 we summarize relevant techniques namely PCA, IPCA, data reconciliation and gross error analysis, which are used in this work. In Sect. 4 we present results for two groups of sensors. In particular, we build models and benchmark them based on the performances for a variety of gross error scenarios. This is done for the case when the underlying sensors have similar dynamics (Sect. 4.1) and when the sensors have different dynamics (Sect. 4.2). In Sect. 5 we discuss some relevant issues before concluding the paper in Sect. 6.

2 Data description

The principle of flux measurement using an SPND is as follows [2, 15, 19]: on being exposed to neutron flux, certain material (called emitter) in the detector undergoes neutron capture to produce radioactive nuclei which decay to produce more stable nuclei. Electrons are generated during these processes which form the current output of the detectors. Electrons are generated mainly due to the following three processes: (i) Neutron capture of emitter atoms releases γ-rays which produce electrons through secondary processes such as photoelectric/compton effect. This part of the current is prompt since the electrons are generated immediately after the neutron capture. (ii) Neutron capture of emitter atoms produce radioactive isotopes which undergo radioactive decay during which β-rays are emitted. This part of the current is delayed since the β decay occurs with a certain half-life. (iii) External γ flux in the reactor which generates background noise in the SPND signal. SPNDs can be classified into prompt or delayed SPNDs depending on whether the major component of the signal is prompt or delayed. For example, cobalt SPNDs are prompt but vanadium SPNDs have a delayed response. Although the cobalt SPNDs give prompt current signal, vanadium SPNDs are more accurate in steady state.

The data used in our work corresponds to a pressurized heavy water reactor (PHWR) of 540 MWe capacity. The reactor consists of a cylindrical core with 42 cobalt and 102 vanadium SPNDs. The core of PHWR for which data is available to us, is considered to be divided into fourteen control zones with each zone containing three cobalt SPNDs. Data from the sensors sampled at 1 min intervals is available for a period of 10 days sometime during the year 2007. Out of the 14,400 time points thus available, data is not recorded for at least one sensor for 339 time points. These data points have been discarded in our work, thereby resulting in 14,061 time points. The 144 SPNDs are labeled using serial numbers from 1 to 144. The numbers 1–42 represent cobalt SPNDs and 43 to 144 represent the vanadium SPNDs. The data available to us has been preprocessed so that readings from both types of detectors are of similar magnitude. The cobalt data available to us was already scaled (divided by a constant) and was hence specified in terms of percentage of flux at full power level, i.e. values were of the order of 0–100, though sometimes they were more than 100. The available vanadium data on the other hand, were directly the flux values and was hence of the order of 10^{13}. We divided the vanadium data by a constant so as to express it in terms of percentage flux as well and therefore the scaled vanadium data also mostly falls in the range 0–100. As a result of this scaling, the data considered in our work are dimensionless.

To build models relating the detectors, we used the following strategy. We analyzed the data in terms of the correlation coefficients and found that while several pairs of detectors were strongly correlated to each other, there were several other pairs which were not correlated. Hence, we grouped the sensors into 15 clusters based on the correlation matrix. As a result sensors grouped together would

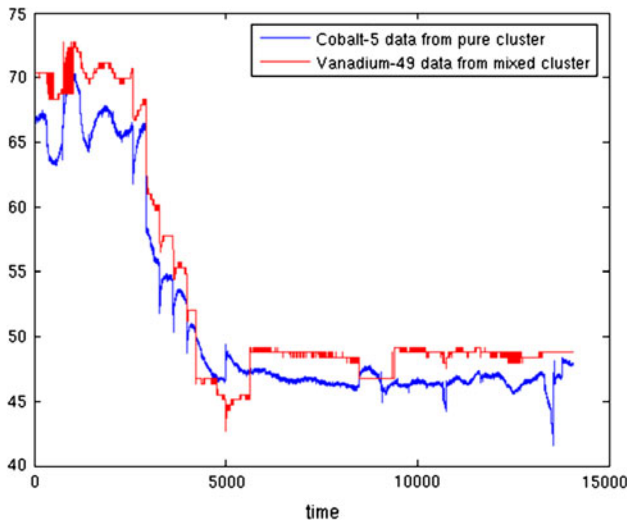

Fig. 1 Time series plots of two sensors—one cobalt (from pure cluster) and one vanadium (from mixed cluster)

be expected to have high correlations with each other as opposed to sensors grouped in different clusters. This grouping was done in several ways (after appropriately transforming the data) to account for differences in dynamics of cobalt and vanadium sensors [13]. In the current work, we perform model building, data reconciliation and gross error analysis on two representative clusters chosen from those reported in [13]: the first cluster contains six cobalt detectors and will hence be labeled pure cluster (containing detectors with same dynamics). The second cluster contains six cobalt and two vanadium detectors and will hence be labeled mixed cluster (detectors with different dynamics). Time-series plots of a detector each from the two clusters are presented in Fig. 1. The flux has a high value initially, then drops and then remains at a lower level for the rest of the time. Further, the vanadium values appear to be quantized at some time points. In our data analysis, we have retained these values as such.

3 Summary of relevant techniques

In this section, we first summarize the PCA and IPCA procedures for obtaining linear, steady state models relating the true values of the n variables $\mathbf{x}(t)$ at any time t. Let the measurements $\mathbf{y}(t)$ be related to the true values as:

$$\mathbf{y}(t) = \mathbf{x}(t) + \epsilon(t) \tag{1}$$

where $\mathbf{y}(t), \mathbf{x}(t), \epsilon(t)$ are column vectors of length n and $\epsilon(t)$ is random measurement noise (or error) assumed to be gaussian with mean 0 and covariance matrix $\mathbf{\Sigma}_\epsilon$. Further, let $\mathbf{Y}$ be the $n \times N$ data matrix (all available measurements) where N is the number of observations of the n variables. Extending Eq. (1), these measurements are assumed to be related to the true unknown values of the variables as: $\mathbf{Y} = \mathbf{X} + \mathbf{E}$, where $\mathbf{X}$ is the corresponding $n \times N$ matrix of the true values and $\mathbf{E}$ is the $n \times N$ matrix of random measurement errors.

It is further assumed that $\mathbf{x}(t)$ satisfies the linear model:

$$\mathbf{A}\mathbf{x}(t) = \mathbf{0} \tag{2}$$

where $\mathbf{A}$ is a $m \times n$ matrix ($m \leq n$) referred to as the model (or constraints) matrix. The main aim in both PCA and IPCA is to estimate the $\mathbf{A}$ matrix from available measurements.

3.1 PCA

A summary of PCA as relevant to our work is presented here. More details can be obtained from literature, for example [4]. PCA allows us to represent noisy data in those directions which capture maximum variability in the data. These principal component directions are the eigenvectors of the measurement data covariance matrix. PCA transforms a set of correlated measurements to a set of uncorrelated measurements. The variance along each direction is given by the corresponding eigenvalue of the covariance matrix. Those directions along which the variance is low contribute little to the variability in the data and represent the relationships between the true values of the variables being measured. These directions can therefore be considered as model equations for the measured data.

Let $\mathbf{P}$ be the $n \times n$ matrix containing the n eigenvectors of the covariance matrix of $\mathbf{Y}$ along the columns. We can write: $\mathbf{P} = [\mathbf{T}|\mathbf{Q}]$ where $\mathbf{T}$ is an $n \times (n - m)$ matrix containing the $n - m$ significant (those corresponding to higher eigenvalues) eigenvectors and $\mathbf{Q}$ is an $n \times m$ matrix with the m least significant (corresponding to lower eigenvalues) eigenvectors along the columns. If we assume that the true variation in the data is only along directions of the $n - m$ eigenvectors contained in $\mathbf{T}$, then we will have for the true data

$$\mathbf{Q}^T\mathbf{X} = \mathbf{0} \tag{3}$$

Thus, if $\mathbf{x}(t)$ is the true vector ($n \times 1$) at some instant t, we have a model relating those n variables as:

$$\mathbf{A}\mathbf{x}(t) = \mathbf{0} \tag{4}$$

where the model matrix $\mathbf{A} = \mathbf{Q}^\mathbf{T}$ contains the least significant m eigenvectors of $\mathbf{P}$ along the rows. An issue here is suitable choice of m. Several methods have been proposed in literature for the same [4].

For any given measurement vector $\mathbf{y}(t)$, $\mathbf{r}(t) = \mathbf{A}\mathbf{y}(t)$ will be small and is referred to as the residual. These residuals $\mathbf{r}(t)$ are central to gross error analysis as explained in Sect. 3.3.

Estimation of noise covariance matrix:

Once the model matrix $\mathbf{A}$ is obtained, the measurement noise covariance matrix $\boldsymbol{\Sigma}_\epsilon$ can be obtained from the residuals as follows: under the assumption that the measurement noises in various detectors are independent of each other, $\boldsymbol{\Sigma}_\epsilon$ will be a diagonal matrix. Thus, the unknowns to be determined are the diagonal elements of $\boldsymbol{\Sigma}_\epsilon$ which are the variances of the measurement noise in the individual measurements. The covariance matrix $\mathbf{S}_r$ of residuals is related to $\boldsymbol{\Sigma}_\epsilon$ as:

$$\mathbf{S}_r = \mathbf{A}\boldsymbol{\Sigma}_\epsilon\mathbf{A}^T \tag{5}$$

$\mathbf{S}_r$ can be computed from the residuals obtained for the given data and the estimated $\mathbf{A}$. The above equations are a linear set of equations in the unknowns (diagonal elements of $\boldsymbol{\Sigma}_\epsilon$) and can be solved for the same. This method for estimating the noise covariance matrix is dealt with in [6]. Also see [14] for more information on solving the above equations. In the results to be presented later, we solve the above equations as an ordinary least squares problem to obtain the diagonal entries of $\boldsymbol{\Sigma}_\epsilon$.

3.2 IPCA

For model identification using PCA, the covariance matrix of the available measurements was directly considered. IPCA, on the other hand, involves appropriate scaling of the data to account for the effects of the measurement error [12]. The model is then estimated based on the covariance matrix of appropriately scaled data. This method essentially combines PCA with a maximum likelihood estimation procedure for simultaneously obtaining an estimate of the measurement error covariance matrix $\boldsymbol{\Sigma}_\epsilon$ along with the model $\mathbf{A}$ [12]. The specific algorithm for the overall procedure is summarized as follows [12]:

(1) Set the iteration counter $k = 1$, $\lambda^{(0)} = 0$, $\mathbf{Y}_s = \mathbf{Y}$, $\mathbf{S}_{y_s} = (1/N)\mathbf{Y}_s\mathbf{Y}_s^T$.

(2) Set estimates of the nonzero elements of $\boldsymbol{\Sigma}_\epsilon^{(k)}$ (estimate of $\boldsymbol{\Sigma}_\epsilon$ at iteration k) to be a small fraction of the corresponding elements of $\mathbf{S}_{y_s}$.

(3) Obtain the transformed matrix $\mathbf{Y}_s = \mathbf{L}^{-1}\mathbf{Y}$ where $\mathbf{L}\mathbf{L}^T = \boldsymbol{\Sigma}_\epsilon^{(\mathbf{k})}$.

(4) Let $\mathbf{U}$ be the eigenvector matrix (eigenvectors as columns in order of decreasing eigenvalues) of the covariance matrix of $\mathbf{Y}_s$. Obtain estimate $\mathbf{A}^{(k)} = \mathbf{U}^T_{n-m+1,\ldots,n}\mathbf{L}^{-1}$, where $\mathbf{U}^T_{n-m+1,\ldots,n}$ is the transpose of the submatrix of $\mathbf{U}$ corresponding to the last m columns.

(5) Let $\lambda^{(k)}$ be the sum of the last m eigenvalues. Stop if the relative change in λ is less than the specified tolerance; else continue.

(6) Obtain the solution for the nonzero elements of $\boldsymbol{\Sigma}_\epsilon$ by minimizing the following function over $\boldsymbol{\Sigma}_\epsilon$: $N\log|\mathbf{A}^{(k)}\boldsymbol{\Sigma}_\epsilon(\mathbf{A}^{(k)})^T| + \sum_{t=1}^N(\mathbf{r}(t)(\mathbf{A}^{(k)}\boldsymbol{\Sigma}_\epsilon(\mathbf{A}^{(k)})^T)^{-1}\mathbf{r}(t))$, denote the solution as $\boldsymbol{\Sigma}_\epsilon^{(k+1)}$.

(7) Increment iteration counter k and return to step 3.

As shown in [12], if the model order m is correctly chosen, the m smallest eigenvalues of the covariance matrix (step 4) converge to unity. However, in general it may be possible to obtain only a local optima of the optimization problem in step 6. Further, in the work of [12], it was assumed that the model order is such that $m(m+1) \geq 2n$. This condition is required to ensure a non-degenerate estimate for $\boldsymbol{\Sigma}_\epsilon$ [12]. For the case of unknown m (as is the case with our data), [12] has presented an iterative procedure for identifying m. Essentially, several m values in increasing or decreasing order can be tried. The identified model order is the maximum value of m for which m eigenvalues of $\boldsymbol{\Sigma}_\epsilon$ converge to 1.

We now briefly summarize data reconciliation and gross error detection, identification and estimation algorithms for linear, steady state models.

3.3 Review of data reconciliation and gross error analysis

3.3.1 Data reconciliation

Given noisy measurements related to true values as in Eq. (1) and the model matrix $\mathbf{A}$, data reconciliation is the process of obtaining more precise (compared to the measurements) or reconciled estimates of the true values which would satisfy the process model (constraints). The problem can be posed as a constrained optimization problem as:

$$\min_{\mathbf{x}(t)} (\mathbf{y}(t) - \mathbf{x}(t))^T\mathbf{W}(\mathbf{y}(t) - \mathbf{x}(t)) \tag{6}$$

$$\text{with the constraint } \mathbf{A}\mathbf{x}(t) = 0 \tag{7}$$

where $\mathbf{W}$ is an $n \times n$ weighting matrix usually taken as (and same will be considered in our work) $\boldsymbol{\Sigma}_\epsilon^{-1}$. The solution to the above optimization problem (see [10]) is given by

$$\hat{\mathbf{x}}(t) = \mathbf{y}(t) - \mathbf{W}^{-1}\mathbf{A}^T(\mathbf{A}\mathbf{W}^{-1}\mathbf{A}^T)^{-1}\mathbf{A}\mathbf{y}(t) \tag{8}$$

where $\hat{\mathbf{x}}(t)$ are the reconciled estimates of the true values at time t.

3.3.2 Gross error analysis

Errors in measurement data can be classified as random errors and systematic or gross errors [10]. Random errors generally consist of measurement noise and are always present in measured variables as considered in Eq. (1).

They are stochastic in nature and assumed to be zero mean and Gaussian. Gross errors are non-random and usually occur as a result of faults in the sensors or process. In our work, we consider gross errors to be constant biases in the sensors. Consider a gross error scenario (labeled scenario k), where we assume the gross error to be constant biases in p_k number of sensors, with the indices of these p_k sensors stored in set $\mathbf{P}_k$. Hence, the measurement vector becomes:

$$\mathbf{y}(t) = \mathbf{x}(t) + \epsilon(t) + \sum_{j\in\mathbf{P}_k} b_j\mathbf{e}_j \tag{9}$$

where b_j is the gross error in the jth variable, $\mathbf{e}_j$ is a unit vector with 1 at the jth location and 0s elsewhere and $\epsilon(t)$ is the random error component of the measurements as before. Two methods for gross error analysis (i.e. detection, identification and estimation) are described next: based on individual observations and based on aggregated observations.

Gross error analysis based on individual observations:

The method summarized here is taken from [10]. Given the model constraint matrix $\mathbf{A}$ and the measurements $\mathbf{y(t)}$, the residuals are obtained as

$$\mathbf{r}(t) = \mathbf{A}\mathbf{y}(t) \tag{10}$$

The covariance matrix of these residuals is given by

$$\mathbf{V} = \mathbf{A}\mathbf{\Sigma}_\epsilon\mathbf{A}^T \tag{11}$$

For gross error detection, the following test statistic is considered:

$$\gamma(t) = \mathbf{r}(t)^T\mathbf{V}^{-1}\mathbf{r}(t) \tag{12}$$

In absence of any gross error $\gamma(t)$ follows a χ^2-distribution with m degrees of freedom. Choosing a confidence level α, we then obtain a threshold value as $\chi^2_{1-\alpha,m}$ where $\chi^2_{1-\alpha,m}$ is the $1-\alpha$ upper quantile (i.e. area to the left is $1-\alpha$) from a χ^2 distribution with m degrees of freedom. A gross error is detected if $\gamma \geq \chi^2_{1-\alpha,m}$.

A technique for identifying the variables with gross errors along with estimates of the gross errors is the generalized likelihood ratio method [10]. The residuals $\mathbf{r}(t)$ in the absence of gross error have mean $\mathbf{0}$. When biases are present in p_k variables corresponding to scenario k (Eq. 9), then the residuals have mean $\mu(t) = \mathbf{F}_k\mathbf{b}$ where $\mathbf{F}_k = \mathbf{A}[\mathbf{e}_{j_1}, \mathbf{e}_{j_2}, \ldots, \mathbf{e}_{j_{p_k}}]$, with $j_1, j_2, \ldots, j_{p_k}$ being the indices of variables with gross errors (i.e. set $\mathbf{P}_k$) and $\mathbf{b}$ is the column vector of unknown magnitudes of the corresponding p_k gross errors. Under the assumption that residuals are normally distributed with mean $\mathbf{0}$ for no gross error case and mean $\mathbf{F}_k\,\mathbf{b}$ for the kth gross error scenario, the maximum likelihood estimate of $\mathbf{b}$ for the kth gross error scenario is given by

$$\mathbf{b}^* = (\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{F}_k)^{-1}(\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{r}(t)) \tag{13}$$

i.e. the entries of the vector $\mathbf{b}^*$ are estimates of the gross errors in sensors $j, j \in \mathbf{P}_k$ corresponding to scenario k. Amongst competing k scenarios (each with same number of gross errors), the most likely scenario is the one with the highest L_k (two times the log likelihood ratio) value computed as:

$$L_k = \left(\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{r}(t)\right)^T\left(\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{F}_k\right)^{-1}\left(\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{r}(t)\right) \tag{14}$$

More information on the above method can be found in [10].

Gross error analysis based on aggregated observations:

The strategy for gross error detection and identification based on an individual observation may lead to different estimates of gross errors (both magnitudes and sensors) at different times. An alternate strategy of combining all available observations at any time T to detect, identify and estimate gross errors can also be implemented. The null and alternative hypotheses for this case are:

$$H_0 : \mu(t) = \mathbf{0}; \quad t = 1, 2, \ldots, T \tag{15}$$

$$H_1 : \mu(t) = \mathbf{F_k b}; \quad t = 1, 2, \ldots, T \tag{16}$$

where $\mu(t)$ is the mean of residuals $\mathbf{r}(t)$ at time t when the kth gross error scenario is being considered. The likelihood ratio test statistic is:

$$\lambda(T) = \sup \quad \frac{Pr\{\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(T)|H_1\}}{Pr\{\mathbf{r}(1), \mathbf{r}(2), \ldots, \mathbf{r}(T)|H_0\}} \tag{17}$$

where the supremum (sup) is computed over all possible values of the hypotheses. With the assumption that the residuals are serially independent Gaussian variables with mean 0 and covariance $\mathbf{V}$ under the null hypothesis, and only change their mean as given in Eq. (16) in the presence of a gross error, the likelihood ratio becomes:

$$\lambda(T) = \sup_{\mathbf{b},\mathbf{F}_k} \frac{\exp\left\{\sum_{t=1}^T -0.5(\mathbf{r}(t) - \mathbf{F}_k\mathbf{b})^T\mathbf{V}^{-1}(\mathbf{r}(t) - \mathbf{F}_k\mathbf{b})\right\}}{\exp\left\{\sum_{t=1}^T -0.5\mathbf{r}(t)^T\mathbf{V}^{-1}\mathbf{r}(t)\right\}} \tag{18}$$

For maximizing the above form of likelihood ratio λ, equivalently maximization of $L(T) = 2\ln(\lambda(T))$ can also be considered. For the kth gross error scenario, the optimal value of $\mathbf{b}$ which maximizes this quantity can be obtained to be (by differentiating $L(T)$ with respect to $\mathbf{b}$ and setting the derivative to 0):

$$\mathbf{b}(T)^* = \frac{(\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{F}_k)^{-1}}{T}\sum_{t=1}^{T}\mathbf{F}_k^T\mathbf{V}^{-1}\mathbf{r}(t) \tag{19}$$

where $\mathbf{b}(T)^*$ indicates the optimal bias estimate for scenario k after considering all measurements up to time T. The corresponding value of twice the log likelihood ratio for the kth gross error is:

$$L_k(T) = \left\{ \sum_{t=1}^{T} \mathbf{F}_k^T \mathbf{V}^{-1} \mathbf{r}(t) \right\}^T \frac{(\mathbf{F}_k^T \mathbf{V}^{-1} \mathbf{F}_k)^{-1}}{T} \times \left\{ \sum_{t=1}^{T} \mathbf{F}_k^T \mathbf{V}^{-1} \mathbf{r}(t) \right\} \tag{20}$$

At time T, the gross error is assigned to the $\hat{k}$ scenario that has the largest value of $L_k(T)$ amongst all the considered gross error scenarios. The corresponding $L(T)$ value is then $L(T) = L_{\hat{k}}(T)$. Equations (19) and (20) can be taken as simplifications of the corresponding expressions presented in [11] for dynamic processes. Further, it can be noted that the optimal estimate of the bias for kth scenario at time T (Eq. 19) is simply the average of the estimates of the biases for kth scenarios obtained at each of the times $t = 1, 2, \ldots, T$ considering only individual observations (Eq. 13). However, the corresponding log likelihood ratio obtained at time T using all the measurements till time T (Eq. 20) is not an average of the log likelihood ratios obtained at individual time points (Eq. 14). The overall gross error detection, identification and estimation procedure for this case then is:

(i) Gross error detection: Gross error is detected at time T if $\Psi(T) \geq \chi^2(1-\alpha, mT)$ where

$$\Psi(T) = \sum_{t=1}^{T} \left\{ \mathbf{r}(t)^T \mathbf{V}^{-1} \mathbf{r}(t) \right\} \tag{21}$$

is the test statistic corresponding to the null hypothesis given data till time T, and has a χ^2 distribution with Tm degrees of freedom.

(ii) Once a gross error is detected, the sensor corresponding to highest $L_k(T)$ value is identified as the erroneous sensor with bias of amount $\mathbf{b(T)^*}$. The above procedure is implemented in a growing window framework (T increases from $1, 2, \ldots, N$). In this case, the identified gross errors and their estimates at each time can vary.

Further, as noted in [10], gross error scenarios with different number of gross errors in each scenario cannot be directly compared on the basis of the likelihood ratio values since the likelihood ratios are distributed as χ^2 variables with different degrees of freedom. They can be instead compared based on the p value (area to the right from the corresponding distribution) obtained for the likelihood ratio of each scenario. In our work however, for simplicity we do not compare scenarios with different number of gross errors. Hence for a single gross error case, all competing scenarios will correspond to single gross errors while for a double gross error case, all competing scenarios will consist of two gross errors.

4 Results

We now present the results of PCA and IPCA application to the two SPND clusters discussed in Sect. 2. In particular, we perform the following:

- Obtain linear models (including model order) using: (i) regular PCA with covariance matrix and (ii) IPCA. Wherever possible, the measurement noise covariance is estimated as well. For other cases, appropriate assumptions are made.
- For both the clusters, the IPCA based model order turns out to be less than that obtained using PCA. Hence we also consider PCA with same model order as IPCA as a third model for comparison. This would allow us to compare PCA and IPCA when they have the same model order.
- Perform data reconciliation using the above three models.
- Perform gross error detection, identification and estimation with gross error introduced in each measurement, one at a time. This is done for bias magnitudes of 1, 3, 5 and 7 % of the mean value (across time) of the corresponding measurement. This analysis is carried out using the two approaches discussed in Sect. 3.3.2, namely based on individual observations and based on aggregated observations
- Perform gross error detection, identification and estimation with gross errors introduced in two measurements simultaneously. Results for a few scenarios are presented. Once again, individual and aggregated observation based strategies are implemented.

For all the bias cases listed above, the corresponding biases are introduced for the entire data length ($t = 1$ to $t = N$). As mentioned earlier (Sect. 2), 14,061 observations sampled at 1 min interval were available for the detectors. The first 7,000 observations were used as training data for model identification for each of the techniques, while the performance characterization was on the entire dataset. For both the clusters, data was mean centred using the mean of the training data before applying any of the above approaches.

For both types of gross error (single and double) cases, the following quantities will be used to compare performances for various methods for the individual and aggregated observations based strategies.

Individual observation analysis:

Detection rate: percentage of the times when a bias (or biases for multiple bias case) is detected (i.e. null hypothesis of no bias is correctly rejected).

Identification rate: percentage of the times when a bias is correctly identified. For multiple bias case, each of the biased detectors should be correctly identified.

Mean square estimation error (MSE): it is defined as: $\sqrt{\left(\sum_{t \in C_I} (\mathbf{b} - \mathbf{b}_t^*)^T (\mathbf{b} - \mathbf{b}_t^*)\right) / (N_{\text{correct}})}$ where $\mathbf{b}$ is the vector of the actual bias introduced, $\mathbf{b}_t^*$ is the estimate of biases using the time t measurement, C_I is the set of time points at which the biases are correctly identified and $N_{correct}$ is the cardinality of C_I.

Aggregated observation analysis:

Rise time (RT): the first time instant (as percentage of the total data length) at which the estimated bias comes within ±5 % band around the true bias and the bias is correctly identified at that time. For multiple bias case, each individual bias estimate should be within the corresponding band.

Settling time (ST): the first time instant (as percentage of the total data length) at which the estimation bias comes within ±5 % band around the true bias such that it does not leave the band later. Once again, for multiple bias case, ST corresponds to the time when each (i.e. "AND" condition) of the individual bias estimate has settled.

Confirmed identification time (CIT): the earliest time (expressed as percentage of the total data length) at which the biases are correctly identified such that they are correctly identified at all subsequent time points as well.

Estimation error (EE): it is defined as: $\sqrt{(\mathbf{b}-\mathbf{b}(N)^*)^T(\mathbf{b}-\mathbf{b}(N)^*)}$ where $\mathbf{b}(N)^*$ is the bias estimate when all the available N measurements have been used. EE is computed only if bias is correctly identified at the Nth instant.

4.1 Results for the pure cluster

The pure cluster consists of six cobalt detectors, namely 5, 12, 19, 26, 33 and 40. It can be noticed that these sensor numbers are separated by seven and represent sensors which are in two adjacent zones in the reactor and hence are strongly correlated.

4.1.1 Model identification using PCA

Eigenvalues of the covariance matrix corresponding to the training data are shown in Fig. 2. It is seen that there is a sharp knee at the second eigenvalue. Hence, the number of significant eigenvalues is taken to be one. Then, the number of independent model equations are five. These equations (constraint model matrix $\mathbf{A}$) are displayed below with the equations listed in order of decreasing eigenvalues.

$$\mathbf{A}=\begin{bmatrix} 0.182 & -0.346 & 0.434 & -0.468 & 0.601 & -0.276 \\ 0.738 & 0.203 & -0.146 & -0.149 & -0.347 & -0.500 \\ -0.231 & 0.699 & -0.367 & -0.417 & 0.374 & -0.085 \\ -0.400 & -0.298 & -0.499 & 0.308 & -0.115 & -0.627 \\ -0.058 & 0.224 & 0.511 & -0.558 & -0.490 & 0.365 \end{bmatrix} \tag{22}$$

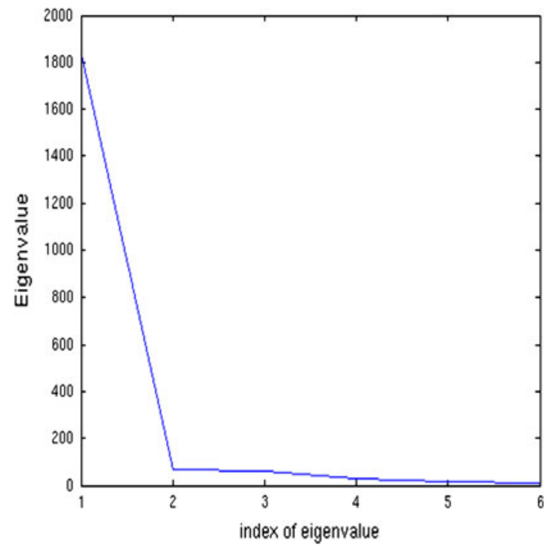

Fig. 2 Plot of the six eigenvalues of covariance matrix for pure cluster

The residuals $\mathbf{r}(t) = \mathbf{A}\mathbf{y}(t)$ for the entire data set are shown in Fig. 3. It can be seen that all of the residuals show a sudden change in the region 7,000–8,500. This can be related to one of the detectors (detector 19) showing a jump in the same region. This issue is discussed later when data reconciliation is performed.

Estimation of noise covariance matrix:

Given the model matrix and the residuals, the measurement noise covariance matrix $\mathbf{\Sigma}_\epsilon$ can be estimated as discussed in Sect. 3.1 (Eq. 5). In particular, under the assumption that $\mathbf{\Sigma}_\epsilon$ is diagonal, these diagonal entries (measurement noise variances in individual detectors) are estimated to be: [0.3875, 0.1817, 2.3930, 0.1904, 0.1991, 0.0799]. From these estimated variances, it is seen that the noise variance of detector 19 is much greater than variances of other detectors. This is related to the fact that detector 19 showed an unusual jump in the 7,000–8,500 time interval as mentioned above. This estimated $\mathbf{\Sigma}_\epsilon$ is used later for performing data reconciliation.

4.1.2 Model identification using IPCA

IPCA as discussed in Sect. 3.2 is used to obtain both the constraint model and the noise covariance matrix simultaneously. Since the true model order is unknown, as suggested in [12] and discussed in Sect. 3.2, we compute the IPCA based model for model orders $m = 5$, 4, 3, and 2. For the true model order, the last m eigenvalues of the covariance matrix of the scaled data should turn out to be unity. The converged eigenvalues for various model orders are shown in Fig. 4. From this figure it is seen that for $m = 2$, the last m eigenvalues converge to 1. However for $m = 3$, 4 and 5, the number of unity eigenvalues is less

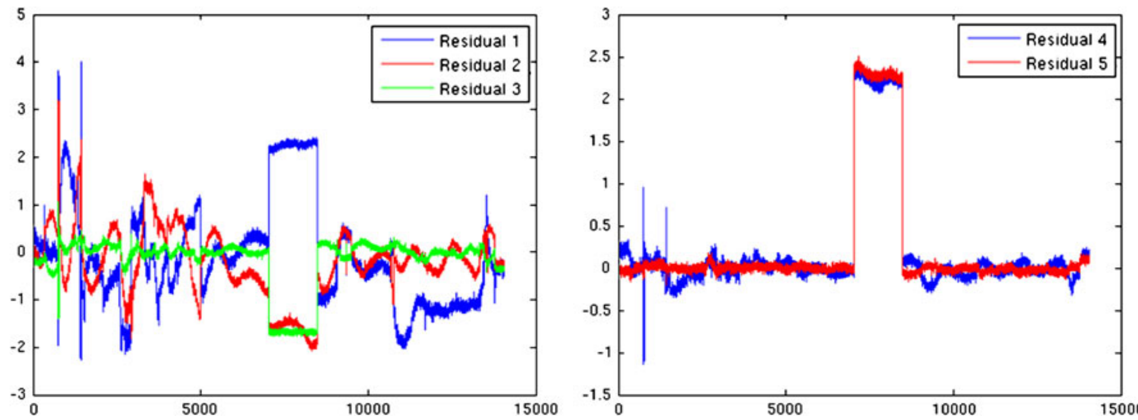

Fig. 3 Plot of residuals with PCA model (m = 5) for pure cluster

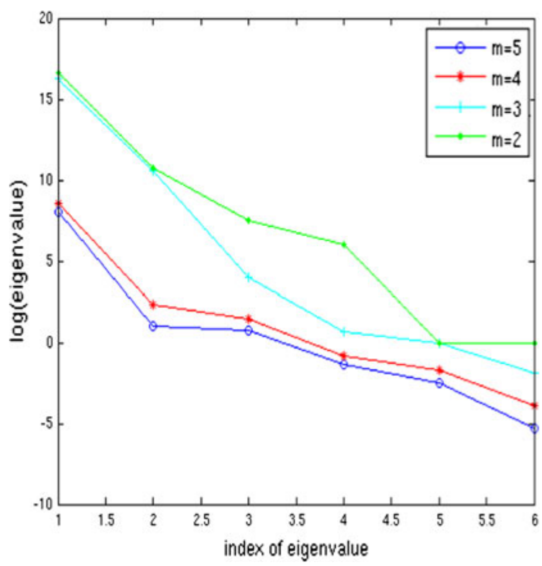

Fig. 4 Plot of eigenvalues for different model orders with IPCA for pure cluster

than m. Hence, for our data, we take $m = 2$ as the model order obtained by IPCA. The corresponding converged model constraint matrix **A** is

$$\mathbf{A} = \begin{bmatrix} 3.184 & -1.959 & 0.213 & -5.124 & -2.993 & 6.934 \\ -2.398 & 6.956 & 13.727 & -14.785 & -12.658 & 9.049 \end{bmatrix} \tag{23}$$

Estimation of noise covariance matrix:

As discussed in [12], for the case when $m(m + 1) < 2n$, the measurement noise variances in individual sensors cannot be obtained uniquely without any additional assumptions. Since this condition is satisfied for the current case ($m = 2, n = 6$), we do not use noise covariance matrix obtained from IPCA. Instead we use the measurement noise covariance matrix obtained by PCA earlier, when performing data reconciliation with the IPCA model.

4.1.3 Comparison of PCA and IPCA models

From the above results, it is noted that different model orders are obtained for PCA and IPCA. Further, while the PCA model equations are scaled (have unit length), IPCA model equations do not have unit length. For the sake of comparison, we scale the IPCA equations as well (premultiply **A** by a diagonal matrix) so that they have unit length. The IPCA constraint matrix then is:

$$\mathbf{A}_{scaled} = \begin{bmatrix} 0.323 & -0.199 & 0.021 & -0.519 & -0.303 & 0.703 \\ -0.090 & 0.262 & 0.517 & -0.557 & -0.477 & 0.341 \end{bmatrix} \tag{24}$$

The residuals obtained from the above scaled equations and those obtained from the last two equations of PCA (corresponding to the two smallest eigenvalues) are plotted together in Fig. 5. For the scaled IPCA model, the first residual does not have the sharp change observed for the PCA model. The second residual for both cases are similar. As discussed in the beginning in Sect. 4, we will also consider the model obtained by PCA for $m = 2$ (same model order as obtained by IPCA) when presenting various results. The model for this case is simply the last two (corresponding to the two smallest eigenvalues) equations of the **A** matrix obtained by PCA.

In Table 1 we list the distance between the row spaces of the models for these three models: PCA with $m = 5$, PCA with $m = 2$ and IPCA ($m = 2$, the scaled **A** matrix). These models are referred to as PCAm5, PCAm2 and IPCAm2 respectively in all subsequent discussions. As discussed in [12], for two full row rank matrices **B** (size

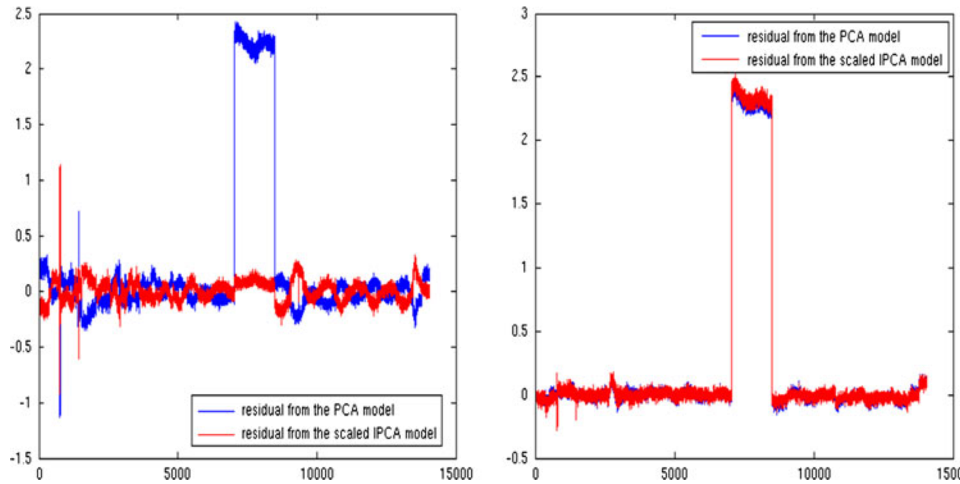

Fig. 5 Residuals for PCA and IPCA models—*right plot* has residual from the last equation (2^{nd} for IPCA and 5th for PCA), while the *left plot* has the residuals from the second last equation (1st for IPCA and 4th for PCA)

$b \times n$) and $\mathbf{C}$ (size $c \times n$) such that each row of $\mathbf{B}$ and $\mathbf{C}$ is normalized to length one, the distance of matrix $\mathbf{B}$ from $\mathbf{C}$ can be computed as:

$$\theta = \sum_{i=1}^{b} \theta_i \tag{25}$$

where θ_i is the minimum distance of $\mathbf{B}_i$ (ith row of $\mathbf{B}$ matrix) from the subspace spanned by the rows of $\mathbf{C}$ and is given by:

$$\theta_i = \| \mathbf{B}_i - \mathbf{B}_i\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C} \|_2 \tag{26}$$

In the above equation, the vector $\mathbf{B}_i\mathbf{C}^T(\mathbf{C}\mathbf{C}^T)^{-1}\mathbf{C}$ is the projection of vector $\mathbf{B}_i$ on the subspace spanned by rows of matrix $\mathbf{C}$. Hence, θ_i will be 0 if $\mathbf{B}_i$ belongs to the subspace of rows of $\mathbf{C}$. Note that the resulting distances are not symmetric. These distances are listed pairwise for the three models in Table 1. An entry in this table corresponds to distance of the model listed in the corresponding row from the model listed in the corresponding column. It is seen from this table that as expected PCAm2 model lies within the PCAm5 model (hence distance is 0) and the distance between PCAm5 and PCAm2 model is 3 since the constraint vectors in PCA are orthonormal and 2 of the 5 vectors of PCAm5 model are included in PCAm2 model. It is to be noted that the distance of PCAm2 model is slightly more from the IPCAm2 model compared to the distance of IPCAm2 model from PCAm2 model. Further, distance of IPCAm2 model from PCAm5 model is very small, though not 0 thereby indicating that the two dimensional row space of IPCAm2 model is not exactly a subspace of the 5 dimensional row space of PCAm5 model. Hence, in this sense, IPCA gives us a fundamentally different model from PCA.

Table 1 Distance between models for pure cluster

Model	PCAm5	PCAm2	IPCAm2
PCAm5	0	3	3.25
PCAm2	0	0	0.29
IPCAm2	3.05E−5	0.20	0

4.1.4 Data reconciliation

In this section, we present the results of data reconciliation of SPNDs based on the three models obtained in previous section. The plots for various detectors after reconciliation along with the raw measurements are presented in Fig. 6.

The test statistic $\gamma(t)$ (Eq. 12) is also plotted in Figs. 7 and 8 for the three models along with the threshold value corresponding to $\alpha = 0.05$. False alarm rates obtained for the entire dataset for the three cases are: 11.87 % for PCAm5, 12.35 % for PCAm2 and 11.31 % for IPCAm2 while for the training data these rates are 4.99, 1.72, 0.53 % for the three models respectively. Hence, only for the full PCA model is the false alarm rate for the training data close to the designed false alarm rate. Further, it can seen from the plot of detector 19 (Fig. 6) that there is a sudden jump in the measured value between the sample numbers 7,040–8,480. As a result of this, the test statistic γ also jumps in this interval for all the three models thereby resulting in high false alarm rate overall. However, the reconciled values for this detector obtained by the three models show that they are able to by and large ignore this jump. Just by looking at this data, one can assume that there may have been a bias in detector 19 in this interval, and as such this bias can be estimated using the gross error related techniques presented earlier. However, in absence of any independent confirmation about this bias, we chose to not do this. Instead in the bias detection, identification and estimation analysis presented later, we will consider cases where known biases have been artificially introduced.

Qualitatively, the performances of the three models for the various detectors are quite similar (Fig. 6). The differences (labeled as adjustments) between the original and

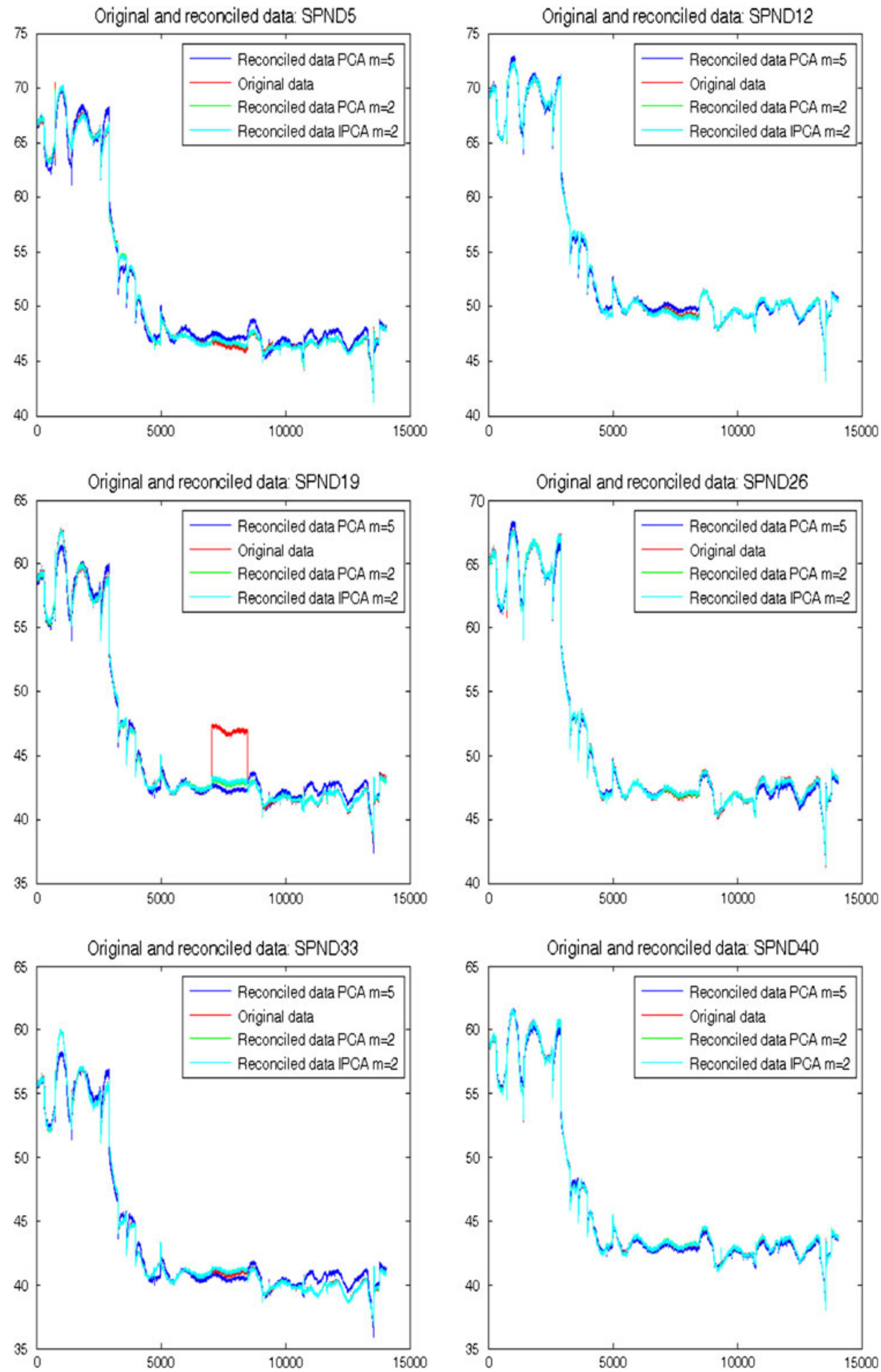

Fig. 6 Original and reconciled data: pure cluster

reconciled data are shown in Table 2 for the three models. In this table, the mean and variances of the adjustments are shown for each of the six detectors. It is seen that the variance of the adjustments for detector 19 are the largest for each model as there is a large adjustment for this detector during the 7,000–8,500 time period. Further, overall adjustments (in terms of their variances) are much higher for PCAm5 than for the other models.

Table 2 Data reconciliation for pure cluster

	Detector	5	12	19	26	33	40
PCAm5	Adj. mean	−0.252	−0.026	0.289	0.075	−0.165	0.091
	Adj. var.	0.394	0.086	2.346	0.072	0.337	0.052
IPCAm2	Adj. mean	−0.038	0.023	0.398	−0.016	−0.020	−0.002
	Adj. var.	0.019	0.005	1.454	0.006	0.005	0.001
PCAm2	Adj. mean	−0.031	0.016	0.408	−0.009	−0.021	−0.003
	Adj. var.	0.013	0.002	1.570	0.004	0.005	0.001

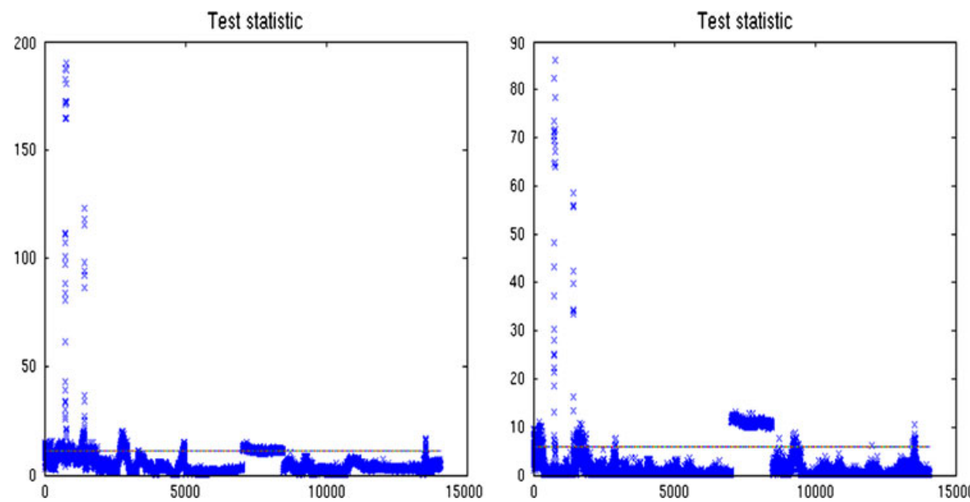

Fig. 7 Test statistic for PCAm5 (*left*) and PCAm2 (*right*) model (*horizontal line* is the threshold value corresponding to $\alpha = 0.05$)

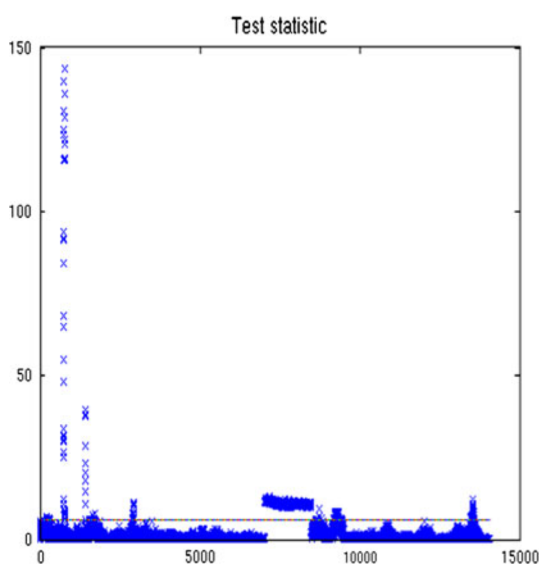

Fig. 8 Test statistic for IPCAm2 model (*horizontal line* is the threshold value corresponding to $\alpha = 0.05$)

4.1.5 Gross error related analysis

The gross error related analysis namely detection, identification and estimation is performed using two approaches: individual observation based analysis and multiple observation based analysis for the following gross error scenarios:

– *Single gross error*: We introduce a bias of 1, 3, 5 or 7 % of the corresponding mean value of the training data in one measurement at a time as discussed in the beginning of Sect. 4.
– *Two simultaneous gross errors*: Three cases are considered for this scenario: (a) bias of 5 % each in detectors 5 and 26, (ii) bias of 5 % each in detectors 12 and 33, and (iii) bias of 5 % each in detectors 19 and 40. For this case, since IPCAm2 and PCAm2 consist of only two equations, it is not possible to identify two gross errors using these two models. Hence in the results presented only gross error detection rates for these two models are given.

Results for single gross error scenarios:

(i) Individual observation based approach: results are presented in Table 3. In this table as well as all subsequent tables, a "–" means that the corresponding event does not take place or the corresponding quantity cannot be calculated. From this table, it can be seen that PCAm5 performs better than the other models for almost all cases. IPCAm2 is better than PCAm2 for detectors 5 and 12. For detector 19 they

Table 3 Single gross error: individual observations approach for pure cluster

	1 % (Bias 0.515)			3 % (Bias 1.545)			5 % (Bias 2.575)			7 % (Bias 3.604)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 5												
Detection rate	27.62	24.74	28.98	99.91	98.53	99.79	100	99.89	99.94	100	99.99	100
Ident. rate	7.89	13.94	18.41	96.55	88.09	89.34	99.89	89.65	89.64	99.94	89.74	89.71
MSE	0.631	0.606	0.492	0.270	0.322	0.293	0.274	0.337	0.296	0.275	0.345	0.290
	1 % (Bias 0.544)			3 % (Bias 1.632)			5 % (Bias 2.721)			7 % (Bias 3.809)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 12												
Detection rate	78.42	12.52	19.47	100	18.74	91.57	100	41.37	99.99	100	77.28	100
Ident. rate	67.15	0	0.16	99.89	0.11	86.35	99.91	26.02	99.69	100	68.15	99.79
MSE	0.132	–	0.617	0.157	1.558	0.314	0.157	1.050	0.409	0.163	0.951	0.414
	1 % (Bias 0.464)			3 % (Bias 1.393)			5 % (Bias 2.322)			7 % (Bias 3.251)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 19												
Detection rate	15.10	12.47	11.40	17.86	13.50	12.38	28.38	17.57	16.43	61.50	56.92	56.23
Ident. rate	10.36	10.24	10.24	12.03	10.24	10.24	22.27	10.24	10.24	56.87	47.92	51.28
MSE	3.337	4.494	4.509	3.145	4.494	4.509	2.386	4.494	4.509	1.544	2.079	2.018
	1 % (Bias 0.514)			3 % (Bias 1.543)			5 % (Bias 2.571)			7 % (Bias 3.599)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 26												
Detection rate	98.34	92.89	68.76	100	100	100	100	100	100	100	100	100
Ident. rate	91.96	72.31	47.45	99.68	89.50	84.23	99.94	89.61	89.37	100	89.67	89.60
MSE	9.815E−2	9.71E−2	0.130	0.103	0.131	0.157	0.104	0.134	0.178	0.104	0.136	0.180
	1 % (Bias 0.439)			3 % (Bias 1.317)			5 % (Bias 2.194)			7 % (Bias 3.072)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 33												
Detection rate	21.43	22.40	21.97	99.57	95.87	95.53	100	99.94	99.99	100	99.98	100
Ident. rate	5.45	0.87	7.1119−3	86.91	81.32	67.86	99.61	89.22	84.52	99.85	89.52	88.60
MSE	0.450	0.368	0.376	0.256	0.239	0.208	0.256	0.296	0.279	0.257	0.300	0.296
	1 % (Bias 0.467)			3 % (Bias 1.401)			5 % (Bias 2.334)			7 % (Bias 3.268)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 40												
Detection rate	96.52	95.65	85.11	100	100	99.96	100	100	100	100	100	100
Ident. rate	82.09	78.98	72.51	99.41	89.61	89.65	99.85	89.69	89.69	99.98	89.73	89.72
MSE	9.962E−2	0.104	0.111	0.105	0.119	0.139	0.108	0.120	0.141	0.109	0.121	0.141

have similar performances. For detector 33, PCAm2 is better for all magnitudes while for detectors 26 and 40, PCAm2 is significantly better for 1 % magnitude. For all models, performance improves as the bias magnitude increases.

(ii) Aggregated observations based approach: results for this case are presented in Table 4. In this as well as subsequent tables in the paper, "Inst." indicates instantly (i.e. in the first observation). Once again, results for PCAm5 are much better compared to the other two models for low and medium bias magnitudes (1, 3, and 5 %) for all detectors. For high (7 %) bias, results of PCAm5 and PCAm2 are comparable and are better than IPCAm2. In almost all cases (detectors and

Table 4 Single gross error: aggregated observations approach for pure cluster

	1 % (Bias 0.515)			3 % (Bias 1.545)			5 % (Bias 2.575)			7 % (Bias 3.604)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 5												
Det. rate	99.99	94.64	96.72	100	99.99	100	100	100	100	100	100	100
CIT	9.86	–	–	Inst.	–	–	Inst.	95.44	–	Inst.	70.02	46.68
RT	12.35	14.05	49.31	4.05	12.83	48.40	3.72	12.02	47.53	3.45	11.36	46.70
ST	65.40	–	–	11.14	–	–	3.72	95.40	–	3.45	70.02	46.70
EE	8E−4	–	–	8E−4	–	–	8E−4	5E−3	–	8E−4	5E−3	2.7E−2
	1 % (Bias 0.544)			3 % (Bias 1.632)			5 % (Bias 2.721)			7 % (Bias 3.809)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 12												
Det. rate	99.99	77.05	100	100	100	100	100	100	100	100	100	100
CIT	5.59	86.29	–	Inst.	6.94	–	Inst.	3.76	–	Inst.	Inst.	–
RT	15.15	14.71	40.62	9.94	14.10	14.66	8.81	13.65	12.89	7.54	13.28	12.04
ST	15.15	–	–	9.94	–	–	8.81	–	–	7.54	96.52	–
EE	5.5E−3	0.181	–	5.5E−3	0.181	–	5.5E−3	0.181	–	5.5E−3	0.181	–
	1 % (Bias 0.464)			3 % (Bias 1.393)			5 % (Bias 2.322)			7 % (Bias 3.251)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 19												
Det. rate	99.99	74.61	99.99	99.99	100	99.99	100	100	99.99	100	100	100
CIT	14.07	52.93	–	10.33	11.01	–	5.26	6.56	2.87	7.8E−2	4.13	Inst.
RT	17.08	12.96	49.38	10.33	11.01	48.52	6.29	6.56	47.57	2.8E−2	4.13	46.55
ST	–	–	–	–	–	–	–	–	–	–	–	–
EE	0.383	0.446	–	0.383	0.446	–	0.383	0.446	0.915	0.383	0.446	0.915
	1 % (Bias 0.514)			3 % (Bias 1.543)			5 % (Bias 2.571)			7 % (Bias 3.599)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 26												
Det. rate	100	100	100	100	100	100	100	100	100	100	100	100
CIT	Inst.	–	–	Inst.	–	–	Inst.	73.96	–	Inst.	Inst.	–
RT	14.46	13.08	47.71	11.50	11.24	43.99	9.55	7.40	40.58	7.10	4.44	37.75
ST	14.46	–	–	11.50	–	–	9.55	73.96	–	7.10	4.44	–
EE	8.6E−3	–	–	8.6E−3	–	–	8.6E−3	3.8E−3	–	8.6E−3	3.8E−3	–
	1 % (Bias 0.439)			3 % (Bias 1.317)			5 % (Bias 2.194)			7 % (Bias 3.072)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 33												
Det. rate	100	100	100	100	100	100	100	100	100	100	100	100
CIT	7.32	–	61.37	Inst.	–	47.19	Inst.	81.05	45.61	Inst.	Inst.	44.08
RT	17.15	14.11	48.87	15.25	12.94	47.19	13.43	12.15	45.61	12.26	11.53	44.08
ST	–	–	85.68	–	–	47.19	13.43	81.05	45.61	12.26	11.53	44.08
EE	6.7E−2	–	1.5E×2	6.7E−2	–	1.5E−2	6.7E−2	2E−2	1.5E−2	6.7E−2	2E−2	1.5E−2

Table 4 continued

	1 % (Bias 0.467)			3 % (Bias 1.401)			5 % (Bias 2.334)			7 % (Bias 3.268)		
	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2	PCAm5	PCAm2	IPCAm2
Detector 40												
Det. rate	98.02	99.93	99.99	100	100	100	100	100	100	100	100	100
CIT	6.81	–	–	Inst.	–	–	Inst.	71.23	–	Inst.	Inst.	–
RT	14.68	13.03	47.25	12.24	11.15	42.76	10.82	7.09	38.90	6.80	4.31	35.50
ST	14.68	–	–	12.24	–	–	10.82	71.23	–	6.80	4.31	–
EE	1.6E−2	–	–	1.6E−2	–	–	1.6E−2	1.6E−3	–	1.6E−2	1.6E−3	–

bias magnitudes), PCAm5 is able to confirm the correct bias instantly while the other two models are not always able to confirm the correct bias. As expected, as bias magnitude increases for a given detector, the performance of each model gets better.

Results for two simultaneous gross error scenarios:

(i) Individual observation based approach: the results are presented in Table 5.

(ii) Aggregated observations based approach: the results are again presented in Table 5.

For both (i) and (ii) above, all three models result in almost 100 % gross error detection rate. PCAm5 does much better for scenarios 1 and 2 compared to scenario 3 for both approaches. The reason is probably the presence of detector 19 for which there was an unexpected jump for some time as discussed earlier.

4.2 Results for the mixed cluster

The mixed cluster consists of eight detectors in total—six cobalt detectors and two vanadium detectors, namely 4, 11, 18, 25, 32, 39, 49 and 139.

4.2.1 *Model identification using PCA*

Eigenvalues of the covariance matrix corresponding to the training data are shown in Fig. 9. It is seen that there is a sharp knee at the second eigenvalue. The number of significant eigenvalues is thus taken to be one. Then, the number of independent model equations are seven. These equations (constraint model matrix $\mathbf{A}$) are displayed below with the equations listed in order of decreasing eigenvalues.

$$\mathbf{A} = \begin{bmatrix} -0.114 & -0.095 & -0.091 & -0.077 & -0.112 & 0.869 & -0.434 & -0.084 \\ -0.469 & -0.265 & -0.291 & -0.054 & -0.037 & 0.244 & 0.749 & 0.010 \\ 0.584 & 0.072 & 0.178 & -0.384 & -0.401 & 0.162 & 0.365 & -0.392 \\ -0.045 & 0.459 & -0.164 & 0.295 & -0.702 & 0.024 & 0.042 & 0.421 \\ 0.178 & -0.323 & 0.029 & -0.556 & -0.020 & -0.009 & -0.039 & 0.743 \\ -0.185 & -0.449 & 0.758 & 0.298 & -0.311 & 0.007 & 0.023 & 0.036 \\ 0.512 & -0.547 & -0.424 & 0.504 & -0.060 & 0.001 & -0.005 & 0.032 \end{bmatrix} \quad (27)$$

In the above simultaneous gross error scenarios, we had introduced gross errors in both the sensors to be of same % magnitude of the corresponding sensor mean. To check if the results would vary significantly if these percentages were to be different, we also performed gross error analysis for an additional scenario involving bias of 5 % in sensor 5 along with a simultaneous bias of 3 % in sensor 26. The results (not presented here) indicated that once again PCAm5 model performs better both for individual and aggregated observations cases.

The residuals $\mathbf{r}(t) = \mathbf{A}\mathbf{y}(t)$ for the entire data set are shown in Fig. 10. It can be seen that the residual corresponding to the first equation (largest eigenvalue) appears to be significantly different from 0 for several time points and is consistently above 0 for the latter half of the data. This probably indicates that model order $m = 6$ could also have been tried, but we have not explored this option in our work. Further, it can be seen that residual 1 shows a sharp drop in the region of 5,600–7,000 time points. This is due to a

Table 5 Double gross error: results for pure cluster

Individual observations approach				Aggregate observations approach			
	PCAm5	PCAm2	IPCAm2		PCAm5	PCAm2	IPCAm2
Scenario 1: (sensors 5 and 26) (5 % magnitude of actual bias)							
Detection rate	100	100	99.58	Detection rate	100	100	100
Ident. rate	94.76	–	–	CIT	Inst.	–	–
MSE	0.314	–	–	RT	9.70	–	–
				ST	9.70	–	–
				EE	0.014	–	–
Scenario 2: (sensors 12 and 33) (5 % magnitude of actual bias)							
Detection rate	100	100	100	Detection rate	100	100	100
Ident. rate	95.74	–	–	CIT	Inst.	–	–
MSE	0.296	–	–	RT	13.40	–	–
				ST	13.40	–	–
				EE	0.067	–	–
Scenario 3: (sensors 19 and 40) (5 % magnitude of actual bias)							
Detection rate	100	100	100	Detection rate	100.00	100.00	100.00
Ident. rate	73.39	–	–	CIT	Inst.	–	–
MSE	1.347	–	–	RT	10.82	–	–
				ST	10.00	–	–
				EE	0.385	–	–

sudden drop in detector 39 which dominates residual 1. Also some of the other residuals show a sudden change in the region 7,000–8,500 time points (for example, residual 6 increases while 7 decreases). This is due to a sudden increase in detector 18 in this region.

Estimation of noise covariance matrix:

For the above PCA model, assuming it to be a diagonal matrix, the diagonal entries of $\mathbf{\Sigma}_\epsilon$ are estimated as: [0.476, 0.158, 0.25, 0.092, 0.386, 8.627, 1.962, 0.317].

4.2.2 Model Identification using IPCA

We compute the IPCA based model for model orders $m = 7$, 6, 5 and 4. The converged eigenvalues for various model orders are shown in Fig. 11. From this figure it is seen that for $m = 4$, the last m eigenvalues converge to 1. However for $m = 5$, 6 and 7, the number of unity eigenvalues is less than m. Hence, for this case, we take $m = 4$ as the model order obtained by IPCA. The converged model constraint matrix $\mathbf{A}$ is:

$$\mathbf{A} = \begin{bmatrix} 1.000 & 2.651 & -4.104 & -3.477 & 1.437 & 0.236 & 0.180 & 1.140 \\ 1.400 & 1.592 & -4.014 & 0.391 & 1.717 & 0.036 & 0.063 & -2.147 \\ -4.945 & 6.144 & 3.156 & -5.812 & 1.573 & -0.158 & -0.179 & -0.376 \\ -2.794 & 1.043 & 4.618 & -0.980 & -1.164 & 0.309 & 0.539 & -0.889 \end{bmatrix} \quad (28)$$

In this case since $m = 4$, the condition $m(m + 1) \geq 2n$ is satisfied [12], and hence it is possible to obtain an estimate of the measurement error variances (assuming $\mathbf{\Sigma}_\epsilon$ to be a diagonal matrix) from IPCA. These converged measurement error variances are: [0.001, 0.001, 0.013, 0.016, 0.042, 5.528, 0.001, 0.147].

4.2.3 Comparison of PCA and IPCA models

As in the case of the pure cluster, it is noted that different model orders are obtained for PCA and IPCA. Since the IPCA model equations did not have unit length, for the sake of comparison, we scale the IPCA equations as well (premultiply $\mathbf{A}$ by a diagonal matrix) so that they have unit length. The equations then are:

$$\mathbf{A}_{scaled} = \begin{bmatrix} 0.157 & 0.417 & -0.645 & -0.547 & 0.226 & 0.037 & 0.028 & 0.179 \\ 0.263 & 0.299 & -0.754 & 0.073 & 0.322 & 0.006 & 0.012 & -0.403 \\ -0.474 & 0.589 & 0.303 & -0.558 & 0.151 & -0.015 & -0.017 & -0.036 \\ -0.481 & 0.179 & 0.795 & -0.169 & -0.200 & 0.053 & 0.093 & -0.153 \end{bmatrix} \quad (29)$$

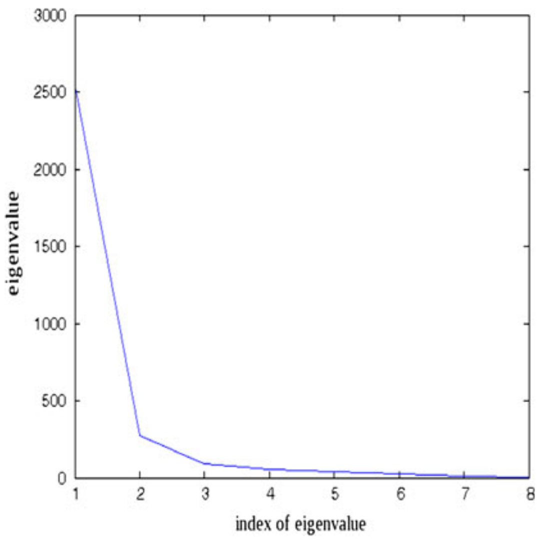

Fig. 9 Plot of the eigenvalues of covariance matrix for mixed cluster

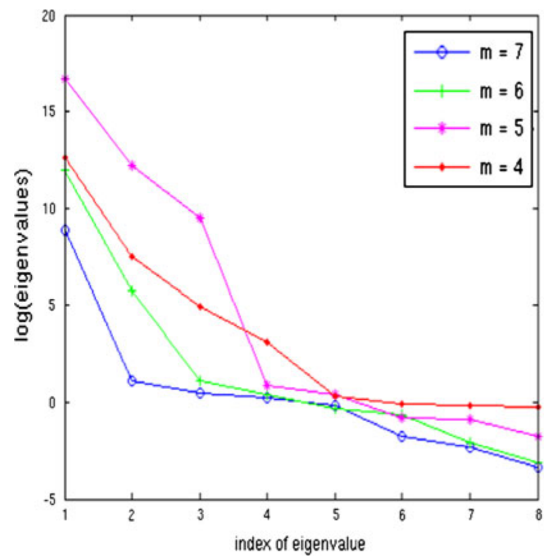

Fig. 11 Plot of eigenvalues for different model orders with IPCA for mixed cluster

The residuals obtained from the above scaled equations and those obtained from the last four equations of PCA (corresponding to the four smallest eigenvalues) are plotted together in Fig. 12. They appear to have different behaviours, specially in the 7,000–8,500 time points interval. Once again, the sudden changes in residuals for IPCA model in this region can be attributed to changes in detector 18.

We also list the distance between the row spaces of the models obtained by PCA with $m = 7$, PCA with $m = 4$ and IPCA ($m = 4$) in Table 6. These models are labeled as PCAm7, PCAm4 and IPCAm4 respectively. Similar to the results for the pure cluster case, distance of IPCAm4 from PCAm4 is smaller than vice versa. Also, distance of IPCAm4 is non-zero (though small) from PCAm7 thereby indicating that these two are indeed different models. Comparison of the noise variances estimated by the two approaches is also quite interesting. For both PCA and

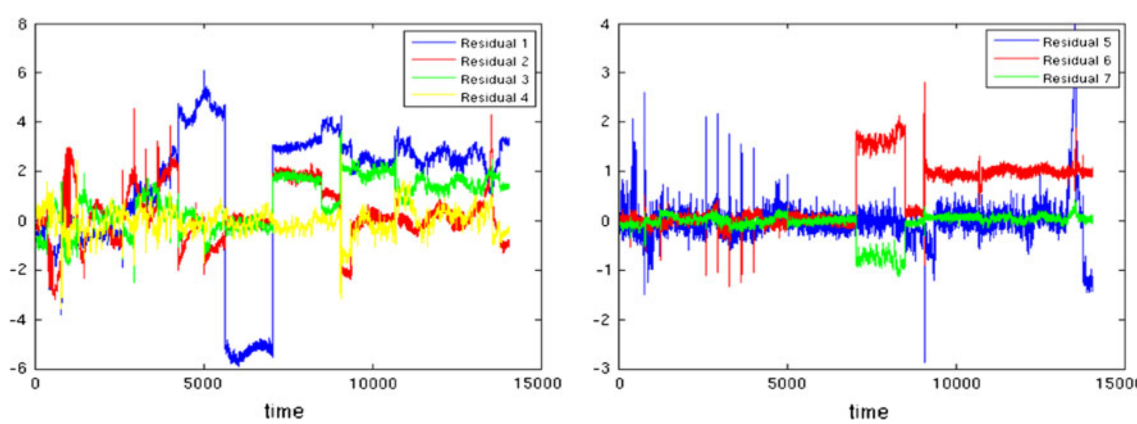

Fig. 10 Plot of residuals for PCA model (m = 7) for mixed cluster

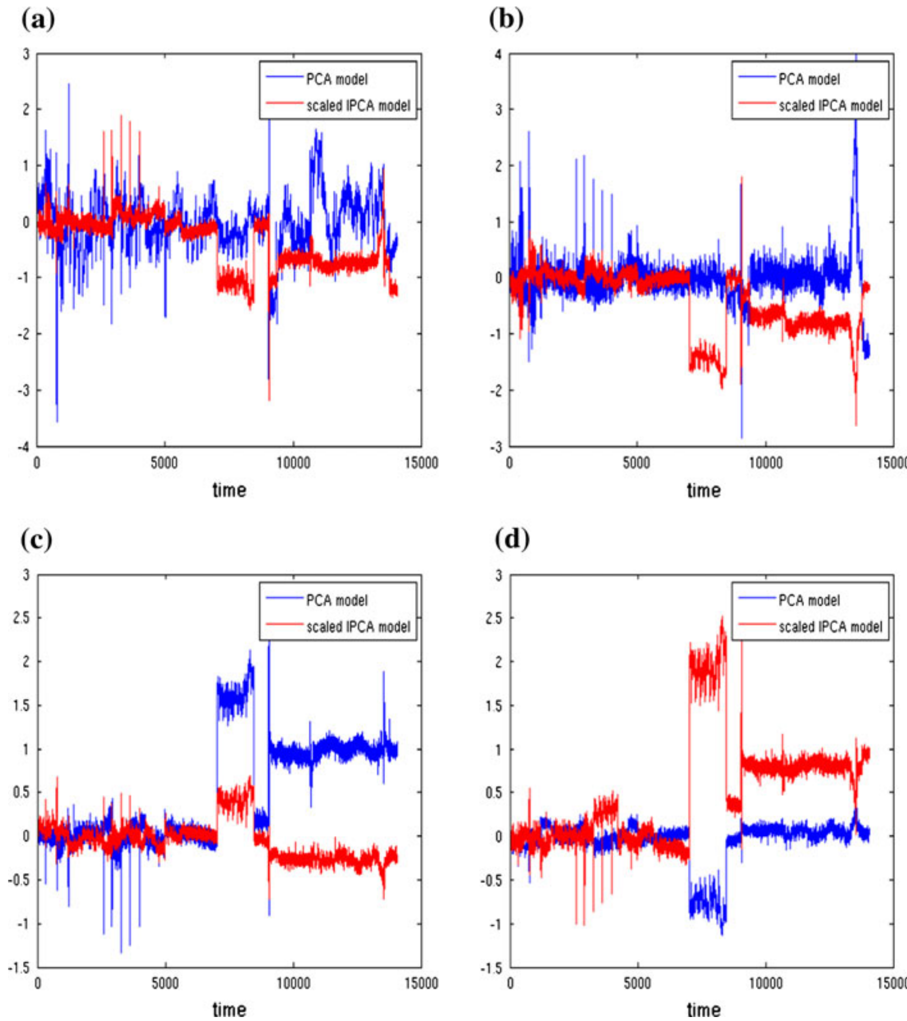

Fig. 12 Residuals for PCA and IPCA models for mixed cluster—in the decreasing order of eigenvalues. **a** Residual 1 (corresponding to the larger eigenvalue), **b** residual 2, **c** residual 3, and **d** residual 4 (corresponding to the smallest eigenvalue)

Table 6 Distance between models for mixed cluster

Model	PCAm7	PCAm4	IPCAm4
PCAm7	0	3	3.66
PCAm4	0	0	1.25
IPCAm4	2.58E−4	0.3	0

IPCA, detector 39 has the highest measurement noise variance amongst all detectors. However, the variances estimated by IPCA are much lower than those estimated by PCA.

4.2.4 Data reconciliation

In this section, we present the results of data reconciliation of SPNDs based on the models obtained in previous section. The plots for various detectors are given in Fig. 13. In these plots, reconciled values as obtained by PCAm7, PCAm4 and IPCAm4 models are compared with the measured values.

The test statistic $\gamma(t)$ (Eq. 12) is also plotted in Figs. 14 and 15 for the three models along with the threshold value corresponding to $\alpha = 0.05$. False alarm rates obtained for the entire dataset for the three cases are: 9.82 % for PCAm7, 16.76 % for PCAm4 and 13.32 % for IPCAm4 while for the training data these rates are 3.31, 1.25, 0.86 % for the three models respectively. In this case, once again PCAm7 is closest to the designed value of false alarm rate of 5 %. It can be seen from the plot of detector 18 (Fig. 13) that there is a sudden jump in the measured value between the sample numbers 7,040 and 8,480. A similar behaviour was observed for the pure cluster case for detector 19. Hence, this further gives us a reason to not consider these deviations as biases. However, by just looking at the data from the mixed cluster, one can conclude that this is quite likely a bias. As a result of this jump, the test statistic also increases in this interval for

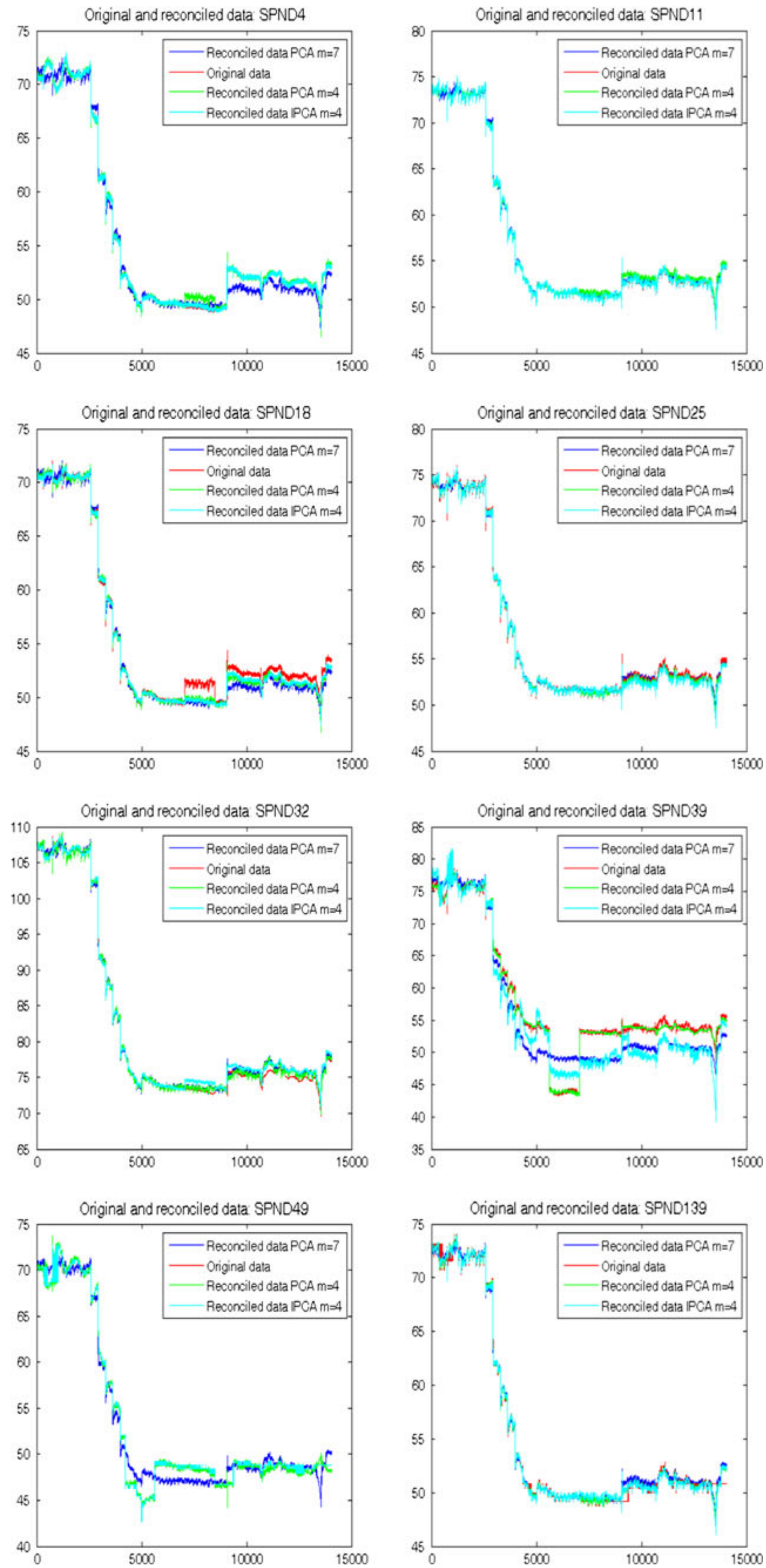

Fig. 13 Original and reconciled data: mixed cluster

all the three models (Figs. 14, 15) resulting in high false alarm rate overall for each model. However, similar to the pure cluster case, all the three models are able to by and large ignore this jump when data reconciliation is performed (Fig. 13). Detector 39 also has a sudden drop in the 5,600–7,000 time points interval. While PCAm7 is able to essentially ignore this drop, estimates from other models are affected by it: the drop is more in PCAm4 estimates than IPCAm4 estimates. Interesting thing to note is that none of the models reject the null hypothesis of no bias in this interval (Figs. 14, 15).

The adjustments between the original and reconciled data are shown in Table 7. As expected, for detector 39 the variance of the adjustment is highest for PCAm7 and is the lowest for PCAm4. Further, while a significant adjustment is made to detector 49 by PCAm7, the same is not true for IPCAm4. The variance of correction made by PCAm4 are in between the variances of the other two models. As shown in Fig. 13 (and also in Fig. 1), there seems to be some quantization effects in the measurements of this sensor and as such we would expect some corrections in them. Similarly, we would expect corrections in detector 139 (Fig. 13) as well, but all the three models don't correct it by much.

4.2.5 Gross error related analysis

Similar to the pure cluster case, the gross error related analysis namely detection, identification and estimation is performed using individual observation and aggregated observations based approaches. For each approach, the performances for various models are compared for the following two scenarios:

- *Single gross error*: We introduce a bias of 1, 3, 5 or 7 % in one measurement at a time as discussed in the beginning in Sect. 4.

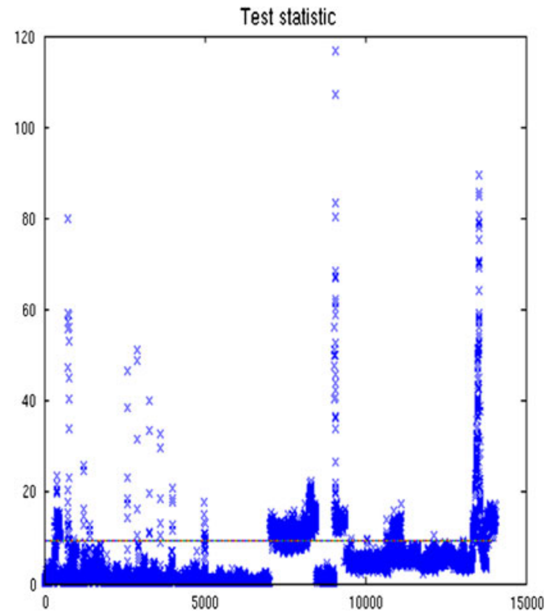

Fig. 15 Test statistic for IPCAm4 model (*horizontal line* is the threshold value corresponding to $\alpha = 0.05$)

- *Two simultaneous gross errors*: We consider three scenarios where simultaneous biases in two measurements are introduced. These scenarios are: (i) bias of 5 % each in sensors 4 and 25, (ii) bias of 5 % each in sensors 11 and 49, and (iii) bias of 5 % each in sensors 49 and 139. Note that sensors 4,11, and 25 are cobalt while sensors 49 and 139 are vanadium sensors. Hence, we investigate performances for scenarios when biases are introduced in only cobalt and vanadium sensors, as well as in both a cobalt and a vanadium sensor.

Results for single gross error scenarios:

(i) Individual observation based approach: the results are presented in Table 8. For low bias magnitudes (1 %),

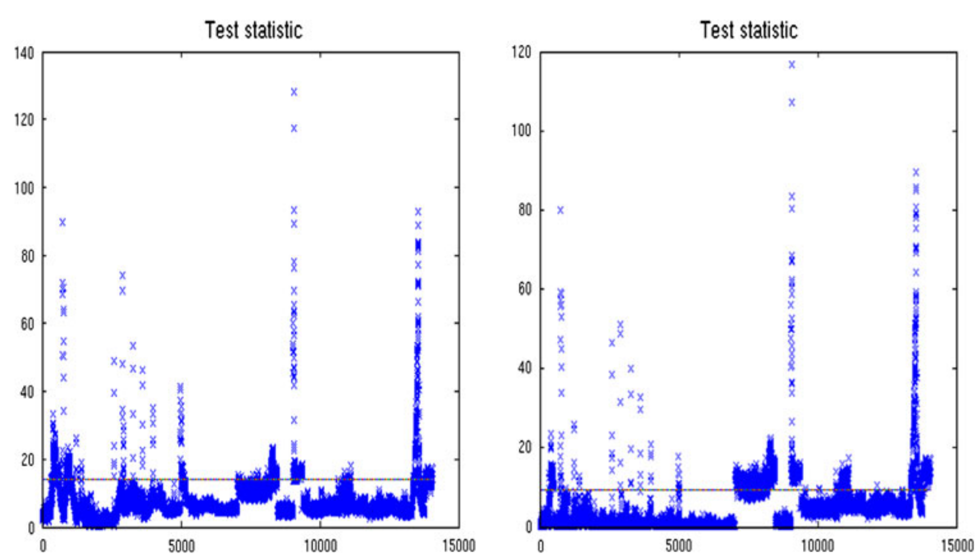

Fig. 14 Test statistic for PCAm7 (*left*) and PCAm4 (*right*) models (*horizontal line* is the threshold value corresponding to $\alpha = 0.05$)

Table 7 Data reconciliation for mixed cluster

	Detector	4	11	18	25	32	39	49	139
PCAm7	Adj. mean	0.284	−0.120	0.589	−0.065	−0.281	1.654	0.116	−0.173
	Adj. var.	0.505	0.072	0.544	0.064	0.292	8.279	1.638	0.302
PCAm4	Adj. mean	−0.107	−0.226	0.420	0.077	−0.196	0.087	0.078	−0.030
	Adj. var.	0.085	0.100	0.277	0.026	0.242	0.188	0.059	0.209
IPCAm4	Adj. mean	−0.011	−0.010	0.395	0.128	−0.489	1.870	0.001	−0.052
	Adj. var.	0.001	0.001	0.273	0.065	0.292	6.988	0.000	0.296

all three models perform quite poorly against all the criteria with results from PCAm7 being comparable (better in some and worse in some) to the other two models. However, for higher magnitudes, results of PCAm7 are clearly superior to other two models. Between PCAm4 and IPCAm4, the results are similar for 1 and 3 % magnitudes, while they are superior for IPCAm4 for most cases for 5 and 7 % magnitudes. For detectors 39 and 49, performance of PCAm4 is quite poor irrespective of the bias magnitudes. This is due to the PCAm4 equations (columns 6 and 7 of the **A** matrix in Eq. 27) being reasonably insensitive (have small coefficients) to these detectors.

As pointed out earlier, detector 39 has a sudden drop in the 5,600–7,000 time points interval. In order to determine if this could have resulted in the poor performances when gross errors were introduced in detector 39, we carried out the following additional analysis (results not shown): we replaced the SPND39 data from 5,600–7,000 points by a constant value equal to the value of SPND39 at 5599th instant, built new models based on this new data, and then performed gross error analysis on this new data. In this case, the performance improved significantly compared to the previous cases. For example, detection rates for PCAm7 for bias magnitudes from 3 % onwards were almost 100 %. This does suggest that the jump in the data is responsible for the bad performance of the identified models when applied to gross error analysis for detector 39. It is thus quite possible that this jump corresponds to bias in the detector, but in absence of any independent confirmation, we have still included this data in the analysis.

(ii) Aggregated observations based approach: the results are presented in Table 9. From this table, for a given sensor and a model, as the magnitude of the introduced bias increases, as expected the performance of that model improves. Across various sensors, for low (1 and 3 %) bias magnitudes, PCAm7 is clearly outperforming other models for almost all sensors. For higher bias magnitudes (5 and 7 %), the performances are relatively comparable across the three models. Overall, PCAm7 seems to be best followed by IPCAm4. PCAm4, in particular, performs quite poorly for detectors 39 and 49.

Results for two simultaneous gross error scenarios:

(i) Individual observation based approach: results are presented in Table 10. From this table, it is seen that PCAm7 easily outperforms the other two models for all the three scenarios. Between PCAm4 and IPCAm4, the performance of the latter is markedly better. For all three methods, detection rates are almost always 100 %.

(ii) Aggregated measurements based approach: results for the aggregated observation based approach for two simultaneous gross errors are presented in Table 11. From this table, it is seen that PCAm7 does much better than the other two models for all the three scenarios. Between PCAm4 and IPCAm4, the latter can be said to perform better for the first two scenarios. For all three methods, the detection rates are once again very high (100 %).

Similar to the pure cluster case, to check if the results would vary significantly if the % magnitudes of the two biases were to be different, we also performed gross error analysis for an additional scenario involving bias of 5 % in detector 11 along with a simultaneous bias of 3 % in detector 49. The results (not presented here) once again indicated that PCAm7 model performs better for both individual and aggregated observations cases.

5 Discussions

Performances of PCA and IPCA models for two types of SPND clusters have been compared in this paper. In both the cases, it was found that IPCA led to a lower model order than PCA. Indeed for the pure cobalt cluster, the model order m was much lower than the minimum required to obtain unique estimate of the measurement error covariance matrix (assuming it is a diagonal matrix). In

Table 8 Single gross error: individual observations approach for mixed cluster

	1 % (Bias 0.555)			3 % (Bias 1.666)			5 % (Bias 2.777)			7 % (Bias 3.888)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 4												
Det. rate	11.97	20.51	16.56	88.97	90.92	92.14	100	100	100	100	100	100
Ident. rate	3.46	0	1.24	81.67	53.86	72.62	99.42	74.85	95.95	99.94	85.7	99.83
MSE	0.555	–	0.601	0.348	0.190	0.344	0.432	0.277	0.408	0.432	0.314	0.457
	1 % (Bias 0.572)			3 % (Bias 1.716)			5 % (Bias 2.861)			7 % (Bias 4.005)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 11												
Det. rate	33.91	18.68	23.66	100	100	100	100	100	100	100	100	100
Ident. rate	17.37	0.88	5.32	98.05	72.41	93.44	99.25	86.11	97.08	99.74	99.42	98.20
MSE	0.289	0.392	0.371	0.224	0.231	0.318	0.227	0.301	0.323	0.230	0.330	0.326
	1 % (Bias 0.558)			3 % (Bias 1.674)			5 % (Bias 2.79)			7 % (Bias 3.906)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 18												
Det. rate	22.73	20.59	18.97	89.11	71.89	94.50	100	100	100	100	100	100
Ident. rate	13.28	10.35	10.91	86.83	68.17	91.17	99.42	99.06	99.27	99.89	99.66	99.82
MSE	0.819	1.892	0.973	0.391	0.758	0.422	0.417	0.635	0.435	0.417	0.634	0.435
	1 % (Bias 0.575)			3 % (Bias 1.726)			5 % (Bias 2.876)			7 % (Bias 4.027)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 25												
Det. rate	48.88	43.94	43.16	100	99.97	99.99	100	100	100	100	100	100
Ident. rate	35.25	24.76	23.21	97.85	96.21	95.39	99.14	99.08	97.09	99.71	99.66	98.24
MSE	0.296	0.509	0.491	0.238	0.327	0.328	0.239	0.331	0.329	0.240	0.333	0.330
	1 % (Bias 0.822)			3 % (Bias 2.467)			5 % (Bias 4.111)			7 % (Bias 5.756)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 32												
Det. rate	13.78	19.69	12.75	75.04	81.99	28.50	99.40	99.79	69.92	100	100	99.28
Ident. rate	4.10	1.66	0.16	70.79	59.74	3.19	98.91	92.91	49.89	99.95	96.44	84.39
MSE	1.275	1.172	3.974	0.454	0.419	0.842	0.629	0.636	0.429	0.650	0.651	1.183
	1 % (Bias 0.580)			3 % (Bias 1.740)			5 % (Bias 2.900)			7 % (Bias 4.061)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 39												
Det. rate	9.84	16.78	13.85	12.57	16.77	19.64	23.18	16.86	32.66	41.39	17.02	53.39
Ident. rate	0.16	0.24	0.14	1.39	0.22	1.61	9.10	0.23	5.33	28.74	0.23	13.61
MSE	4.961	80.700	5.038	4.511	79.850	4.530	3.663	77.989	3.887	2.774	77.870	2.853
	1 % (Bias 0.534)			3 % (Bias 1.603)			5 % (Bias 2.671)			7 % (Bias 3.740)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 49												
Det. rate	10.29	16.89	14.42	27.39	17.23	24.12	70.19	17.88	46.27	91.15	18.57	66.75
Ident. rate	0.43	0.45	6.401E−2	16.12	0.44	8.12	60.47	0.43	30.05	82.59	0.56	48.03
MSE	2.394	37.509	3.835	1.315	37.560	2.521	0.838	36.950	1.805	0.752	32.680	1.487

Table 8 continued

	1 % (Bias 0.554)			3 % (Bias 1.663)			5 % (Bias 2.772)			7 % (Bias 3.881)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 139												
Det. rate	11.82	18.81	14.81	72.09	67.51	78.17	100	98.64	99.96	100	100	99.96
Ident. rate	3.01	3.10	2.90	63.03	54.39	61.30	99.71	97.22	97.63	99.90	99.77	99.90
MSE	1.646	2.168	1.925	0.460	0.569	0.492	0.483	0.504	0.474	0.485	0.551	0.509

Table 9 Single gross error: aggregated observations based approach for mixed cluster

	1 % (Bias 0.555)			3 % (Bias 1.666)			5 % (Bias 2.777)			7 % (Bias 3.888)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 4												
Det. rate	96.74	52.17	56.82	99.99	100.	100.	100.	100.	100.	100.	100.	100.
CIT	9.05	–	–	1.4E−2	Inst.	2.1E−2	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
RT	20.76	11.85	28.19	12.83	10.18	12.37	11.05	1.4E−2	10.33	10.01	Inst.	9.05
ST	–	–	–	69.53	73.31	71.58	61.9	66.79	64.75	10.01	5.75	9.05
EE	6.7E−2	–	–	6.7E−2	5.6E−2	7.3E−2	6.7E−2	5.6E−2	7.3E−2	6.7E−2	5.6E−2	7.3E−2
	1 % (Bias 0.572)			3 % (Bias 1.716)			5 % (Bias 2.861)			7 % (Bias 4.005)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 11												
Det. rate	99.86	99.96	99.37	100.	100.	100.	100.	100.	100.	100.	100.	100.
CIT	2.47	–	8.75	Inst.	Inst.	3.656	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
RT	2.64	11.44	8.75	Inst.	9.41	0.48	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
ST	–	–	–	6.91	–	–	6.26	6.33	2.1E−2	5.41	Inst.	Inst.
EE	5.9E−2	–	0.112	5.9E−2	0.130	0.112	5.9E−2	0.130	0.112	5.9E−2	0.130	0.112
	1 % (Bias 0.558)			3 % (Bias 1.674)			5 % (Bias 2.790)			7 % (Bias 3.906)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 18												
Det. rate	99.92	55.89	68.1	100.	100.	100.	100.	100.	100.	100.	100.	100.
CIT	8.534E−2	55.69	50.15	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
RT	22.9	11.14	50.15	21.24	9.04	21.05	13.99	5.7E−2	12.53	11.43	Inst.	10.47
ST	–	–	–	–	–	–	61.46	–	63.22	11.43	–	10.47
EE	8.8E−2	0.262	9.6E−2	8.8E−2	0.262	9.6E−2	8.8E−2	0.262	9.6E−2	8.8E−2	0.262	9.6E−2
	1 % (Bias 0.575)			3 % (Bias 1.726)			5 % (Bias 2.876)			7 % (Bias 4.027)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 25												
Det. rate	99.99	97.55	99.95	100.	100.	100.	100.	100.	100.	100.	100.	100.
CIT	5.23	5.67	6.29	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
RT	2.84	12.42	6.29	Inst.	5.7E−2	Inst.	Inst.	1.4E−2	Inst.	Inst.	Inst.	Inst.
ST	–	–	–	6.79	–	–	6.10	5.06	Inst.	Inst.	Inst.	Inst.
EE	5.8E−2	0.122	0.104	5.8E−2	0.122	0.104	5.8E−2	0.122	0.104	5.8E−2	0.122	0.104

Table 9 continued

	1 % (Bias 0.822)			3 % (Bias 2.467)			5 % (Bias 4.111)			7 % (Bias 5.756)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 32												
Det. rate	96.82	64.54	41.62	100.	100.	99.99	100.	100.	100.	100.	100.	100.
CIT	7.40	99.82	–	Inst.	Inst.	2.1E−2	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
RT	7.42	7.10	–	6.88	6.64	0.60	6.49	5.93	3.5E−2	Inst.	Inst.	Inst.
ST	–	–	–	–	–	–	–	–	–	5.87	5.43	–
EE	0.232	0.254	–	0.232	0.254	0.689	0.232	0.254	0.689	0.232	0.254	0.689
	1 % (Bias 0.580)			3 % (Bias 1.740)			5 % (Bias 2.900)			7 % (Bias 4.061)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 39												
Det. rate	81.00	20.94	35.47	91.24	22.86	44.73	96.61	24.17	76.67	96.81	24.83	92.21
CIT	52.31	–	–	21.98	–	–	7.45	–	–	6.43	–	–
RT	52.31	–	–	24.08	–	–	23.66	–	23.33	23.34	–	22.97
ST	–	–	–	–	–	–	–	–	–	–	–	–
EE	0.374	–	–	0.374	–	–	0.374	–	–	0.374	–	–
	1 % (Bias 0.534)			3 % (Bias 1.603)			5 % (Bias 2.671)			7 % (Bias 3.740)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 49												
Det. rate	89.82	22.49	38.55	96.59	24.83	62.99	97.20	27.15	82.53	99.95	38.62	96.79
CIT	44.04	–	–	6.89	–	52.90	2.81	99.56	21.13	5.7E−2	66.64	6.51
RT	23.42	–	–	22.74	–	52.90	22.08	–	23.10	21.48	8.06	22.74
ST	–	–	–	–	–	–	78.56	–	–	34.17	–	–
EE	0.123	–	–	0.123	–	0.315	0.123	2.175	0.315	0.123	2.175	0.315
	1 % (Bias 0.554)			3 % (Bias 1.663)			5 % (Bias 2.772)			7 % (Bias 3.881)		
	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4	PCAm7	PCAm4	IPCAm4
Detector 139												
Det. rate	99.44	52.51	57.77	100.	100.	100.	100.	100.	100.	100.	100.	100.
CIT	0.57	59.17	–	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.	Inst.
RT	23.03	11.66	54.10	14.09	7.13	13.40	11.75	6.79	10.67	10.30	6.41	8.84
ST	95.70	98.19	–	30.65	7.13	13.40	11.75	6.79	10.67	10.30	6.41	8.84
EE	2.8E−2	1.9E−3	–	2.8E−2	1.9E−3	3.0E−2	2.8E−2	1.9E−3	3.0E−2	2.8E−2	1.9E−3	3.0E−2

literature [12], IPCA has been implemented only for cases where the model order is sufficiently high. However, in our work, despite the low model order, we obtained convergence of exactly the last m eigenvalues to unity.

Overall, we conclude the following for the SPND data analyzed in this work: between IPCA and PCA model of the same order, IPCA generally does better for gross error related criteria. Overall however, the full PCA model does much better than IPCA model or the reduced PCA model. This is an unexpected result since for the case when the model and measurement noise covariances are unknown, IPCA is a theoretically rigorous procedure for estimating the unknown model as opposed to PCA which involves ad hoc decision making. The factors that may be responsible for this behaviour as well as some other interesting observations are listed below:

(1) Both PCA and IPCA give algebraic equations relating the measurements at a given instant of time and assume that the reactor is operating under steady state conditions. Thus assuming the true data and measurement noise to be serially independent, we expect the residuals to be white. To test this, we considered the autocorrelation values of the residuals obtained

Table 10 Double gross error: individual observations approach for mixed cluster

	PCAm7	PCAm4	IPCAm4
Scenario 1: (sensors 4 and 25) (5 % magnitude of actual bias)			
Detection rate	100	100	100
Ident. rate	98.95	56.59	80.90
MSE	0.531	0.525	0.617
Scenario 2: (sensors 11 and 49) (5 % magnitude of actual bias)			
Detection rate	100	100	100
Ident. rate	78.16	19.58	47.94
MSE	0.818	9.164	1.821
Scenario 3: (sensors 49 and 139) (5 % magnitude of actual bias)			
Detection rate	97.28	98.69	98.58
Ident. rate	77.19	17.88	27.91
MSE	0.996	6.798	2.122

Table 11 Double gross error: aggregated observations approach for mixed cluster

	PCAm7	PCAm4	IPCAm4
Scenario 1: (sensors 4 and 25) (5 % magnitude actual bias)			
Detection rate	100	100	100
CIT	Inst.	4.50	Inst.
RT	11.39	0.014	20.08
ST	62.51	–	70.47
EE	0.058	0.247	0.143
Scenario 2: (sensors 11 and 49) (5 % magnitude actual bias)			
Detection rate	100	100	100
CIT	Inst.	77.44	6.96
RT	21.87	22.83	22.97
ST	–	–	–
EE	0.165	3.032	0.590
Scenario 3: (sensors 49 and 139) (5 % magnitude actual bias)			
Detection rate	100.00	100.00	100.00
CIT	Inst.	–	–
RT	21.85	10.90	23.03
ST	99.79	–	–
EE	0.131	–	–

for various lags and subjected them to whiteness tests as given in [9]. While the corresponding plots are not shown here, we concluded from this test that the residuals are strictly not white. In literature, while IPCA has been applied to simulated data containing serial correlations [12] and was found to obtain a model close to the true model, extensive testing for other data sets has not been performed. Further, techniques such as adding lagged data as an additional variable can also be tried.

(2) We feel that a key issue in IPCA is the nonlinear optimization that is required at each step of the iterative procedure. During the course of implementation we experienced several cases (model orders and initial guess of Σ_ϵ) where IPCA did not converge (we used the fmincon function in Matlab for performing the required optimization). Even when it converges, there is no guarantee of the global optima [12].

(3) As mentioned in Sect. 2, measurements from vanadium detectors are most likely corrupted by quantization effects. Effect of such phenomena on the ability of PCA or IPCA to estimate models needs to be systematically investigated.

(4) Another interesting observation is the estimation of widely different measurements noise variances by the two approaches for the mixed cluster data. As discussed in the results section, the estimates of the variances obtained by IPCA were generally lower than those obtained by PCA. We feel that IPCA estimates may be better since there is an explicit incorporation of measurement noise variances in the iterative procedure. However, an extensive study on several datasets with known variances is required to validate this.

6 Conclusions

In this paper, we used the measurement data to identify models on two groups of SPNDs. PCA and IPCA [12] were used to estimate the models (including model order) from the available data. It was found that model order obtained by IPCA was consistently lower than model order obtained by PCA. Comparison of these models showed that PCA based model performs better overall than IPCA based model for data reconciliation and gross error analysis related criteria. Possible reasons for the same were discussed as well. Currently, we are working on developing a suite of such models which can capture relationships between SPNDs for a wide range of reactor operations.

Acknowledgements This work was supported by the Board of Research in Nuclear Sciences (BRNS), India. Grant number: 2009/39/16-BRNS.

References

1. Adorjan, F., Morita, T.: A correlation-based signal validation method for fixed in-core detectors. Nucl. Technol. **118**, 264–275 (1997)
2. Agu, M.N., Petitcolas, H.: γ-compensated self-powered detectors for reactor instrumentation and control. Meas. Sci. Technol. **1**, 1047–1051 (1990)

3. Banda, L.: Operational experience in rhodium self-powered detectors. IEEE Trans. Nucl. Sci. **26**, 910–915 (1979)
4. Jolliffe, I.T.: Principal Component Analysis. Springer, New York (2002)
5. Kaistha, N., Upadhyaya, B.R.: Incipient fault detection and isolation of field devices in nuclear power systems using principal component analysis. Nucl. Technol. **136**(2), 221–230 (2001)
6. Keller, J.Y., Zasadzinski, M., Darouach, M.: Analytical estimator of measurement error variances in data reconciliation. Comput. Chem. Eng. **16**(3), 185–188 (1992)
7. Li, F., Upadhyaya, B.R., Coffey, L.A.: Model-based monitoring and fault diagnosis of fossil power plant process units using group method of data handling. ISA Trans. **48**(2), 213–219 (2009)
8. Lu, B.: Monitoring and fault diagnosis of the steam generator system of a nuclear power plant using data-driven modeling and residual space analysis. Ann. Nucl. Energy **32**(9), 897–912 (2005)
9. Mehra, R.K., Peschon, J.: An innovations approach to fault detection and diagnosis in dynamic systems. Automatica **7**, 637–640 (1971)
10. Narasimhan, S., Jordache, C.: Data Reconciliation & Gross Error Detection: An Intelligent Use of Process Data. Gulf Publishing, Houston (2000)
11. Narasimhan, S., Mah, R.S.H.: Generalized likelihood ratios for gross error identification in dynamic processes. AIChE J. **34**, 1321–1331 (1998)
12. Narasimhan, S., Shah, S.L.: Model identification and error covariance matrix estimation from noisy data using PCA. Control Eng. Pract. **16**, 146–155 (2008)
13. Razak, R.A.: Data based techniques for fault diagnosis of self powered neutron detectors, MTech Stage I report, Department of Electrical Engineering, Indian Institute of Technology Bombay, Mumbai (2011)
14. Romagnoli, J., Sánchez, M.: Data Processing and Reconciliation for Chemical Process Operations, Process Systems Engineering. Academic Press, San Diego (2000)
15. Sovka, J.A.: Response of cobalt neutron flux detectors, Technical Report, AECL-3368 (1969)
16. Todt, W.H.: Characteristics of self powered neutron detectors used in power reactors. In: Proceedings of the International Nuclear Congress, ENC'98, Nice (1998)
17. Upadhyaya, B.R., Skorska, M.: Fault analysis of in-core detectors in a PWR using time-series models. In: Proceedings of the IEEE Conference on Decision and Control, vol. 2, San-Francisco, pp. 758–763 (1983)
18. Upadhyaya, B.R., Zhao, K., Lu, B.: Fault monitoring of nuclear power plant sensors and field devices. Prog. Nucl. Energy **43**(1–4), 337–342 (2003)
19. Yusuf, S.O., Wehe, D.K.: Analog and digital compensation techniques for delayed self-powered neutron detectors. Nucl. Sci. Eng. **106**, 399–408 (1990)
20. Zhao, K., Upadhyaya, B.R.: Model based approach for fault detection and isolation of helical coil steam generator systems using principal component analysis. IEEE Trans. Nucl. Sci. **53**(4), 2343–2352 (2006)
21. Zhao, K., Upadhyaya, B.R., Wood, R.T.: Robust dynamic sensor fault detection and isolation of helical coil steam generator systems using a subspace identification technique. Nucl. Technol. **153**(3), 326–340 (2006)