Pergamon

Computers and Chemical Engineering Supplement (1999) S589-S592
© 1999 Elsevier Science Ltd. All rights reserved
PII: S0098-1354/99/00130-1

Performance evaluation of PCA tests for multiple gross error identification

Miguel Bagajewicz[(+)], Qiyou Jiang[(+)] and Mabel Sánchez[(*)]

[(+)] School of Chemical Engineering and Materials Science, University of Oklahoma
The Energy Center, 100 E. Boyd, Norman, Oklahoma 73019, USA
[(*)] Planta Piloto de Ingeniería Química, UNS-CONICET, Camino La Carrindanga km 7,
(8000)-Bahía Blanca – Argentina

Abstract

In this paper, the performance of the Principal Component Test, including both the Principal Component Nodal Test (PCNT) and Principal Component Measurement Test (PCMT), is evaluated when used for the identification of multiple gross errors. A few existing gross error identification techniques are modified to replace the nodal, global and measurement tests they use, by a principal component test. Comparative analysis indicates that PCA tests do not significantly enhance the ability in identification features of these strategies, performing worse in some cases.

Keywords: Principal component analysis, Gross error detection, Data reconciliation

Introduction

Data reconciliation techniques are performed to estimate process variables in such a way that balance constraints are satisfied, but to obtain accurate estimates, some action should be taken to eliminate the influence of gross errors. The presence of biased instruments and leaks, as well as departures from steady state, invalidate the results of data reconciliation techniques, so hypothesis testing has been extensively used for detection and identification purposes. A good survey of this issue can be found in Mah (1990) and Madron (1992).

Recently, Principal Component Tests (PCT) have been proposed by Tong and Crowe (1995, 1996). Industrial applications of PCMT were reported by Tong and Bluck (1998). The authors indicated that tests based on PCA are more sensitive to subtle gross errors, and have greater power to correctly identify the variables in error than conventional Nodal, Measurement and Global Tests. To this date there is no published work assessing the effectiveness of the method, with the exception of a few isolated claims based on only one gross error.

The increasing application of PCA in process monitoring and fault diagnosis, and the lack of performance evaluation studies like those proposed by Iordache (1985) and Serth and Heenan (1986), motivates the present work. For this analysis, the statistical tests applied in the identification step of three collective compensation strategies and the serial elimination strategy are replaced with PCA tests. Performance evaluation results and a comprehensive discussion are provided.

Principal component tests

Principal component tests were first proposed by Tong and Crowe (1995). With principal component analysis (PCA), a set of correlated variables can be transformed into a new set of uncorrelated variables, known as principal components (PCs). Each PC is a linear combination of original variables. The coefficients of each linear combination are obtained from an eigenvector of the covariance matrix of the original variables. Therefore, with PCA one can investigate the PCs rather than the original variables for gross error detection. Measurement Test (MT) and Nodal Test (NT) are two typical statistical tests based on the measurement adjustments and constraint residuals, respectively. The Principal Component Measurement Test (PCMT) and Principal Component Nodal Test (PCNT) were developed on the basis of the principal components of measurement adjustments and constraint residuals.

Three collective compensation and one serial elimination strategies

a) *MUBET*

The Unbiased Estimation Technique (UBET) (Rollins and Davis, 1992) is developed from the balance residuals. It uses a gross error identification strategy, like the nodal strategies reported by Mah et al. (1976) and Serth and Heenan (1986), to isolate the suspect nodes and also construct the candidate bias/leak list θ_1 from the suspect nodes. Then it constructs a constraint matrix C_1 corresponding to the elements in θ_1 with rank equal to the number of constraint equations q and obtains the size estimation for θ_1. Finally use Bonferroni Test for identification.

Bagajewicz et al. (1998) presented a modified version of this strategy, which addresses singularities and uncertainties of the original method (MUBET).

b) *MSEGE*

The Simultaneous Estimation of Gross Error (SEGE) was proposed by Sánchez and Romagnoli (1994). It includes a two-stage procedure. In the first stage, it allows the isolation of a subset of constraints that do not pass the global test. It applies a recursive procedure that has the advantage that only the reciprocal of a scalar has to be computed in each step. In this procedure, equations are added one by one to the least square estimation problem of the vector *x*. After each addition, the objective function of the least square estimation technique is calculated and compared with the critical value τ_c. Stage 1 of the procedure provides a set of measurements and units suspect of being biased or having leaks. In Stage 2, the identification and estimation of gross errors is accomplished by the following procedure: First assume there is only one gross error. Take all combinations of one gross error and run the reconciliation model. If the lowest objective function value is lower than the global test, then one gross error is declared and stop. Otherwise assume one more gross error and repeat this procedure.

Sánchez et al. (1998) modified this strategy to addresses singularities and uncertainties of the original method (MSEGE).

C) SICC

This strategy relies on the MT for gross error detection. It uses the MT to make a list of suspect gross errors and identifies from the list one gross error using a compensation model (Bagajewicz and Jiang, 1998). This error is put in a list of confirmed gross errors. Next a new list of suspects is constructed and the compensation model is run using the confirmed gross errors and a new candidate at a time to determine which should be added to the confirmed gross error list. The procedure is repeated until no gross errors are detected. Leaks are identified using the equivalency theory (Bagajewicz and Jiang, 1998).

D) SEM

Serial elimination strategy based on measurement test (SEM) also relies on the MT. It calculates the MT first with data reconciliation. If no MT flags, declares no gross error and stops. Otherwise it eliminates the stream with the largest MT. It repeats this procedure until no MT flags.

Inclusion of principal component tests

The aforementioned techniques have been modified to use PCA tests. For MUBET, the strategy of Pseudonodes (Serth and Heenan, 1986) has been replaced by PCNT. The same modification is performed to MSEGE: Its Stage 1 is accomplished now by PCNT to obtain a suspect set of measurements and leaks. In the case of SICC and SEM the measurement test was replaced by the PCMT test.

Analysis of gross error identification performance

The performance evaluation procedure used consists of simulating the presence of multiple bias of fixed size. Ten thousand simulation trials with random measurement errors are carried out for each case. To evaluate the identification ability of the strategies, four measures of performance are considered: AVTI (average type I error), OP (overall power), OPF (overall performance of perfect identification) and the recently introduced OPFE (overall performance of equivalent identification) (Bagajewicz et al. 1998).

First, the process flowsheet in Figure 1 is used. It consists of a recycle system with five units and nine streams. The true flow rate values are x = [10. 20. 30. 20. 10. 10. 10. 4. 6.]. The flow rate standard deviations were taken as 2% of the true flow rates.

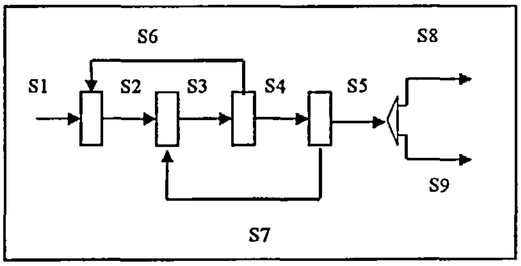

Figure 1: An Example

Measurement values for each simulation trial were taken as the average of ten random generated values. In order to compare results on the same basis, the level of significance of each method was chosen such that it gives an AVTI equal to 0.1 under the null hypothesis. Each strategy was tested under the same scenarios of gross errors introduced. The size of gross error is selected as four times of its corresponding flowrate standard deviation when there is only one gross error, five and three times for two gross errors, five, four and three times for three gross errors. If a leak is introduced, then the minimum standard deviation of flowrates connected to the corresponding process is chosen for the selection of the leak's size.

Table 1 to 4 show the comparisons between MUBET and PCNT-MUBET, MSEGE and PCNT-MSEGE, SICC and PCMT-SICC, and SEM and PCMT-SEM. The results indicate that the two methods for each pair are comparable.

Table 1: Performance Comparison between MUBET and PCNT-MUBET

Gross Error Introduced			MUBET				PCNT-MUBET			
No.	Location	Size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S5	0.8	0.2290	0.9431	0.8898	0.9984	0.1731	0.9995	0.8645	0.9991
	L3	0.8	0.1769	0.8337	0.7835	0.8756	1.7290	0.2983	0.2752	0.9678
2	S6,S7	1.0,0.6	2.7848	0.2515	0.1285	0.6382	1.7704	0.5000	0.0000	0.7061
	S4,L2	2.0,0.6	0.2287	0.8095	0.5876	0.6397	2.5140	0.5000	0.0000	0.7045
3	S6,S7,S9	1.0,0.8,0.36	4.8456	0.0013	0.0000	0.8728	3.8903	0.3335	0.0000	0.9150
	S7,S8,L3	1.0,0.32,0.6	2.3781	0.3641	0.0007	0.6074	1.8516	0.6072	0.0000	0.6866

Note: S_n means a bias in stream S_n and L_n a leak in unit

Table 2: Performance Comparison between MSEGE and PCNT-MSEGE

Gross Error Introduced			MSEGE				PCNT-MSEGE			
No.	Location	Size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S5	0.8	0.0434	0.9998	0.9577	0.9999	0.1006	0.9999	0.9025	1.0000
	L3	0.8	0.0840	0.9555	0.9326	0.9682	0.1328	0.9548	0.8943	0.9712
2	S6,S7	1.0,0.6	0.1335	0.9337	0.8481	0.8821	0.1992	0.9342	0.8219	0.9003
	S4,L2	2.0,0.6	0.3767	0.8321	0.6520	0.6822	0.1791	0.9451	0.8507	0.9134
3	S6,S7,S9	1.0,0.8,0.36	1.0947	0.6449	0.0000	0.9572	1.1736	0.6352	0.0000	0.9659
	S7,S8,L3	1.0,0.32,0.6	0.8449	0.7014	0.4061	0.6501	0.6601	0.7782	0.5466	0.9093

Table 3: Performance Comparison between SICC and PCMT-SICC

Gross Error Introduced			SICC				PCMT-SICC			
No.	Location	Size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S5	0.8	0.0731	1.0000	0.9274	1.0000	0.0631	1.0000	0.9372	1.0000
	L3	0.8	2.0355	0.0000	0.0000	0.9787	2.5566	0.0000	0.0000	0.9403
2	S6,S7	1.0,0.6	0.1380	0.9438	0.8592	0.8983	1.3892	0.5091	0.0212	0.4978
	S4,L2	2.0,0.6	1.9792	0.5000	0.0000	0.8790	1.8590	0.5000	0.0000	0.7987
3	S6,S7,S9	1.0,0.8,0.36	1.0685	0.6455	0.0000	0.9467	1.7177	0.6796	0.0460	0.8146
	S7,S8,L3	1.0,0.32,0.6	2.0572	0.3963	0.0000	0.4052	3.1281	0.1215	0.0000	0.6391

Table 4: Performance Comparison between SEM and PCMT-SEM

Gross Error Introduced			SEM				PCMT-SEM			
No.	Location	Size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S5	0.8	0.0939	1.0000	0.9077	0.9077	0.0917	1.0000	0.9122	0.9122
	L3	0.8	2.0658	0.0000	-	0.9157	3.0131	0.0000	-	0.8712
2	S6,S7	1.0,0.6	0.1580	0.9444	0.8437	0.8461	3.0680	0.0000	0.0000	0.8929
	S4,L2	2.0,0.6	2.0057	0.5000	-	0.8551	2.0316	0.5000	-	0.9133
3	S6,S7,S9	1.0,0.8,0.36	1.1136	0.6454	0.0000	0.8994	4.9877	0.0010	0.0000	0.9937
	S7,S8,L3	1.0,0.32,0.6	2.2657	0.4309	-	0.3070	4.3079	0.0000	-	0.3980

Table 5: Performance Comparison between SEM and PCMT-SEM for a large plant

Gross Error Introduced			SEM				PCMT-SEM			
No	Location	Size	AVTI	OP	OPF	OPFE	AVTI	OP	OPF	OPFE
1	S15	500000	0.115	1.000	0.889	0.889	0.088	1.000	0.916	0.916
2	S21	60000	0.114	0.992	0.891	0.891	0.090	1.000	0.915	0.915
3	S15, S21	500000, 60000	0.120	0.996	0.885	0.885	0.086	1.000	0.918	0.918
4	S14, S15, S23	-500000, -500000,50000	0.114	0.999	0.889	0.889	2.245	0.667	0.000	0.001
5	S4, S15, S37	7000000, 500000, 200000	0.115	0.999	0.890	0.890	0.083	1.000	0.922	0.922
6	S4,S15,S21,S37	700000,500000,60000, 200000	0.187	0.994	0.827	0.827	0.081	1.000	0.924	0.924
7	S14,S15,S23,S56	-500000,500000 50000, 100000	0.329	0.892	0.552	0.565	4.235	0.500	0.000	0.000

The second example is a Large Plant example. It consists of 93 streams (3 of them are unmeasured: S46, S49 and S50), 11 processes, 14 tanks and 9 mixing and splitting nodes. Real data was used to perform a data reconciliation and then the reconciled data were used as true values in our experiments. Considering the time expense for this large system, in this experiment each result is based on 1000 (rather than 10000) simulation trials where the random errors are changed and the magnitudes of gross errors are fixed.

Table 5 indicates that in 5 out of the total 7 cases PCMT-SEM was successful and got a slightly higher performance than SEM. However, there are two cases that PCMT-SEM completely failed while SEM was still successful.

The failure can be explained in light of the PCA strategy. The assumption that the variable with larger contribution to the larger principal component has larger probability of having a gross error is not always true.

Conclusion

The principal component tests have been added to different collective compensation techniques and serial elimination strategy. The performance has then been compared to the regular techniques using known tests. The simulation results show that the use of PC tests does not necessarily improve the power of serial identification strategies. In fact, it sometimes performs better and sometimes worse.

References

Bagajewicz, M. and Q. Jiang, 1998, Gross Error Modeling and Detection in Plant Linear Dynamic Reconciliation. To appear, *Comp. & Chem. Eng.*

Bagajewicz M., Q. Jiang and M. Sánchez, 1998, Removing Singularities and Assessing Uncertainties in Two Efficient Gross Error Collective Compensation Methods. Submitted to *Chem. Engng. Comm.*

Iordache C., R. Mah and A. Tamhane, 1985, Performance Studies of the Measurement Test for Detection of Gross Errors in Process Data. *AIChE J.*, 31, 1187-1201.

Mah R. S. H., 1990, Chemical Process Structures and Information Flows. Butterworths.

Mah R.S.H., G. Stanley and D. Downing, 1976, Reconciliation and Rectification of Process Flow and Inventory Data. *Ind. Engng. Chem. Process Des. Dev.*, **15**, 175-183.

Madron F., 1992, Process Plant Performance. Measurement and Data Processing for Optimization and Retrofits. Ellis Horwood Ltd, Chichester, England.

Rollins D. and J. Davis, 1992,Unbiased Estimation of Gross Errors in Process Measurements. *AIChE J.*, **38**, 563-572.

Sánchez M. and J. Romagnoli, 1994, Monitoreo de Procesos Continuos: Análisis Comparativo de Técnicas de Identificación y Cálculo de Bias en los Sensores, *AADECA 94 - XIV Simposio Nacional de Control Automático*, Argentina.

Sánchez M., J. Romagnoli, Q. Jiang and M. Bagajewicz, 1998, Simultaneous Estimations of Biases and Leaks in Process Plants. Submitted to *Comp. Chem. Engng.*

Serth R. and W Heenan, 1986, Gross Error Detection and Data Reconciliation in Steam Metering Systems. *AIChE J.*, 32, 733-742.

Tong H. and C. Crowe, 1995, Detection of Gross Errors in Data Reconciliation by Principal Component Analysis, *AIChE J.*, **41**, 7, 1712-1722.

Tong H. and C. Crowe, 1996, Detecting Persistent Gross Errors by Sequential Analysis of Principal Components. *Comp. Chem. Engng.*, S20, S733-S738.

Tong H. and D. Bluck, 1998, An Industrial Application of Principal Component Test to Fault Detection and Identification. Workshop on On-Line Fault Detection and Supervision in the Chemical Process Industries. IFAC, Lyon.