

# Data validation with unknown variance matrix

D. Maquin<sup>†</sup>, S. Narasimhan<sup>‡</sup>, J. Ragot<sup>†</sup>

<sup>†</sup>Centre de Recherche en Automatique de Nancy, Institut National Polytechnique de Lorraine  
2, Avenue de la Forêt de Haye - 54 516 Vandoeuvre - France

<sup>‡</sup>Department of Chemical Engineering, Indian Institute of Technology,  
Chennai, 600 036 - India

## Abstract

The data validation consists in obtaining an estimation of the true values of process variables that respect the balance equations. Generally, the procedure needs the knowledge of the variance of the measurement errors; unfortunately, in most situations, we only have a rough estimation of this variance and therefore the data validation procedure gives results depending on this poor estimation. A pioneer work of Almsy and Mah (1984) presents a solution to this problem based on the analysis of the constraint residuals. Darouach *et al.* (1989) developed a slightly different approach based on a maximum likelihood estimator. Here we present a direct method that simultaneously estimates the variances of the measurement errors and reconciles the data with respect to the balance equations. Some numerical results illustrate the efficiency of the proposed method.

*Keywords:* Data validation, Variance estimation, Balance equations, Analytical redundancy.

## Introduction

Most of the data validation techniques are based on the assumption that the measurement errors are random variables obeying a known statistical distribution. Almost without exception, the techniques use a given variance matrix of the measurement errors. However, in most practical situations, this matrix is unknown. The problem of estimating simultaneously the measurements and their variances has already been investigated. Almsy and Mah (1984) have proposed a method that is based on the sample evaluation of the variance matrix of the residuals. More precisely, they minimize the sum of squares of the off-diagonal elements of the measurement error variance matrix subject to the constraint that links this variance to that of the residuals. This method gives an analytical solution and doesn't make any hypothesis on the statistical distribution of measurement errors. Based on this idea, Darouach *et al.* (1989) have also proposed a method of estimating the variance matrix that is based on the maximum likelihood estimator method and makes use of the model constraints and the statistical properties of the residuals. More recently, Keller and Darouach (1998) have enhanced the previous method and proposed to estimate some off-diagonal elements of the measurement error variance matrix due to correlated measurements. The method proposed in this paper is inspired from the last but one cited work of Darouach. The measurement error variance matrix is estimated by using a maximum likelihood estimator, however this estimation is constrained by the process model only and doesn't require a sample evaluation of the residual variance.

The paper is organized as follows. After this short introduction, the second section states the problem in the case of linear steady-state process operating around a

given point. The proposed solution is described in the third section. In the fourth section, it is extended to the case of several steady-state operating points. Some numerical results, presented in the fifth section, illustrate the efficiency of the proposed method.

## Statement of the problem

Let us consider a process, under steady state conditions, characterized by the vector of state variable  $X^*$ . The measurement equation is taken as:

$$X_i = X^* + e_i \quad i = 1, \dots, N \quad X^* \in \mathbf{R}^{v,1} \quad (1a)$$

where  $N$  is the number of measurements,  $e_i$  the measurement errors considered as independant, centered and with a variance matrix  $V$ .

The model of the process is:

$$AX^* = 0 \quad A \in \mathbf{R}^{p,v} \quad (1b)$$

The probability density function of the measurement errors is chosen as:

$$P(e) = (2\pi)^{-Nv/2} \prod_{i=1}^N |V|^{-1/2} \exp\left(-\frac{1}{2} \left(e_i^T V^{-1} e_i\right)\right) \quad (2)$$

that leads to the likelihood function of the estimations  $\hat{X}$ :

$$L(\hat{X}) = (2\pi)^{-Nv/2} \prod_{i=1}^N |V|^{-1/2} \exp\left(-\frac{1}{2} \|X_i - \hat{X}\|_{V^{-1}}^2\right) \quad (3)$$

A simplification is proposed by defining the moment matrix:

$$M(\hat{X}) = \sum_{i=1}^N e_i e_i^T = \sum_{i=1}^N (X_i - \hat{X})(X_i - \hat{X})^T \quad (4)$$

Thus combining (3) and (4) and using the trace operator (Tr) gives:

$$\mathcal{L}(\hat{X}) = (2\pi)^{-Nv/2} \prod_{i=1}^N |V|^{-1/2} \exp\left(-\frac{1}{2} \text{Tr}(V^{-1}M(\hat{X}))\right) \quad (5)$$

Therefore, the problem of simultaneous estimation of the state variables and the measurement error variance matrix comes down to maximize the likelihood function (5) with respect to  $\hat{X}$  and  $V$  and subject to the model constraint (1b).

This maximization is equivalent to the solving of the following problem:

$$\begin{cases} \min_{\hat{X}, V} \phi = \frac{N}{2} \text{Log}|V| + \frac{1}{2} \text{Tr}(V^{-1}M(\hat{X})) \\ \text{s. t. } A\hat{X} = 0 \end{cases} \quad (6)$$

### General solution of the problem

From (6), let us define the Lagrangian

$$L = \phi + \lambda^T A\hat{X} \quad \lambda \in \mathbf{R}^{p-1} \quad (7)$$

The conditions of stationarity for the Lagrangian (7) are expressed as:

$$\begin{cases} \frac{\partial L}{\partial \hat{X}} = V^{-1} \sum_{i=1}^N (\hat{X} - X_i) + A^T \lambda = 0 \\ \frac{\partial L}{\partial V} = \frac{N}{2} V^{-1} - \frac{1}{2} V^{-1} M(\hat{X}) V^{-1} = 0 \\ \frac{\partial L}{\partial \lambda} = A\hat{X} = 0 \end{cases} \quad (8)$$

The system (8) is non-linear, so no analytical solution may be found. A solution based on a hierarchical calculus using the method of relaxation may be proposed. From (8b), one obtains:

$$V(\hat{X}) = \frac{1}{N} M(\hat{X}) \quad (9)$$

In that expression, the variance matrix depends on the state variable estimations. Equations (8a) and (8c) allow the state variable estimation to be expressed as:

$$\hat{X} = \left( I_v - V(\hat{X})A^T (AV(\hat{X})A^T)^{-1} A \right) \bar{X} \quad (10a)$$

where:

$$\bar{X} = \frac{1}{N} \sum_{i=1}^N X_i \quad (10b)$$

Equation (10a) clearly shows that the state variable estimations depend on the variance matrix.

Then, for solving the non-linear system (8), the following algorithm is proposed:

- 1)  $k = 0$ . Select initial values:  $\hat{X}_k = \bar{X}$
- 2) Compute the moment matrix  $M(\hat{X}_k)$  from (4)
- 3) Deduce the variance matrix  $V(\hat{X}_k)$  from (9)
- 4) Compute the new estimation according
 
$$\hat{X}_{k+1} = \left( I_v - V(\hat{X}_k)A^T (AV(\hat{X}_k)A^T)^{-1} A \right) \bar{X}$$
- 5) According to a convergence test, decide to stop or change  $k$  to  $k+1$  and go to step 2.

### Extension to several steady-state points

In real process operation, the operating points are continuously undergoing changes and steady-state is, in fact, almost never attained. On a practical point of view, steady-state has meaning only within the time interval that is considered. Therefore, it is interesting for state and variance estimation purpose, to consider sequences of measurements corresponding to different operating points even if these sequences are very short. The proposed method may be extended to that case. If we carry out  $p$  series of measurements, each involving  $N_j$  measurements around an operating point  $X_j^*$  representing a steady-state operating point of the process, the set of measurements can then be written as:

$$X_{ij} = X_j^* + e_{ij} \quad i = 1, \dots, N_j \quad j = 1, \dots, p \quad (11a)$$

These measurements are then linked by the model:

$$AX_j^* = 0 \quad j = 1, \dots, p \quad (11b)$$

With  $N$ , the total number of measurements ( $N = \sum_{j=1}^p N_j$ ), the probability density function of the errors is defined by:

$$P(e) = (2\pi)^{-Nv/2} \prod_{j=1}^p \prod_{i=1}^{N_j} |V|^{-1/2} \exp\left(-\frac{1}{2} (e_{ij}^T V^{-1} e_{ij})\right) \quad (12)$$

Defining a moment matrix for each operating point:

$$M(\hat{X}_j) = \sum_{i=1}^{N_j} e_{ij} e_{ij}^T = \sum_{i=1}^{N_j} (X_{ij} - \hat{X}_j)(X_{ij} - \hat{X}_j)^T \quad (13)$$

The maximization of the corresponding likelihood function of the estimations leads to the following problem:

$$\begin{cases} \min_{\hat{X}_j, V} \phi = \frac{N}{2} \text{Log}|V| + \frac{1}{2} \sum_{j=1}^p \text{Tr}(V^{-1}M(\hat{X}_j)) \\ \text{s. t. } A\hat{X}_j = 0 & j = 1, \dots, p \end{cases} \quad (14)$$

The associated Lagrangian may be written as:

$$L = \phi + \sum_{j=1}^p \lambda_j^T A\hat{X}_j \quad (15)$$

Its first order stationarity conditions constitute a non-linear system that defines the searched solution:

$$\begin{cases} \frac{\partial L}{\partial \hat{X}_j} = V^{-1} \sum_{i=1}^{N_j} (\hat{X}_j - X_{ij}) + A^T \lambda_j = 0 & j = 1, \dots, p \\ \frac{\partial L}{\partial V} = \frac{N}{2} V^{-1} - \frac{1}{2} V^{-1} \sum_{j=1}^p M(\hat{X}_j) V^{-1} = 0 \\ \frac{\partial L}{\partial \lambda_j} = A\hat{X}_j = 0 & j = 1, \dots, p \end{cases} \quad (16)$$

As previously, equation (16b), together definition (13), gives the expression of the variance matrix:

$$V(\hat{X}) = \frac{1}{N} \sum_{j=1}^p M(\hat{X}_j) \quad (17)$$

and equations (16a) and (16c) allow, for each operating point  $j$ , the state variable estimation to be expressed as:

$$\hat{X}_j = \left( I_v - V(\hat{X})A^T(AV(\hat{X})A^T)^{-1}A \right) \bar{X}_j \quad (18a)$$

where  $\bar{X}_j = \frac{1}{N_j} \sum_{i=1}^{N_j} X_{ij}$  (18b)

Then, the solution of the non-linear system (16) may be obtained using a hierarchical calculus which scheme is presented figure 1.

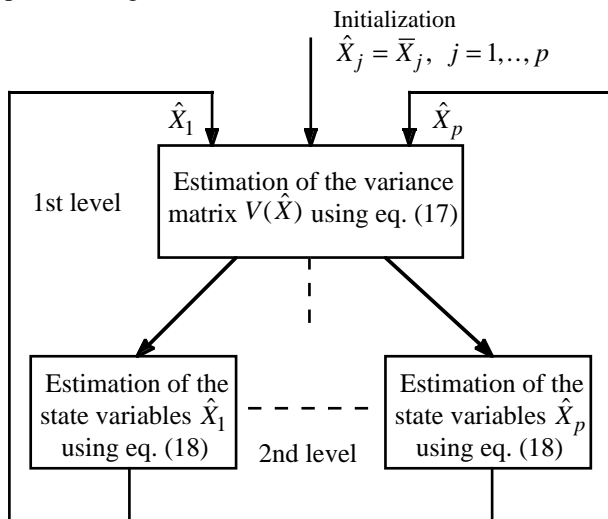


Figure 1 : Hierarchical estimation

### Simulation experiments

Let us consider the carriage network made up of four process units and eight streams represented on figure 2:

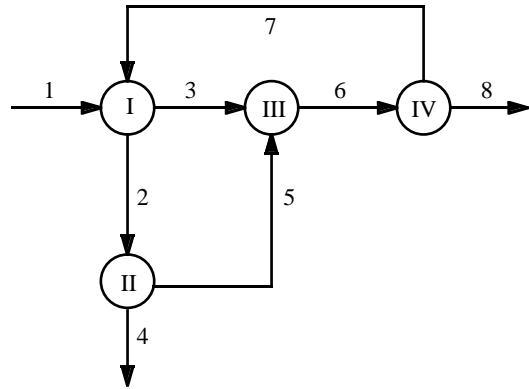


Figure 2 : A carriage network

The incidence matrix of the corresponding graph that describes the model of the process may be written as:

$$A = \begin{pmatrix} 1 & -1 & -1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & -1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & -1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 & -1 \end{pmatrix}$$

For the first trial, 100 measurements of the eight variables have been simulated around a single operating point. This point is described by:

$$X^* = (15.0 \ 7.5 \ 12.5 \ 3.5 \ 4.0 \ 16.5 \ 5.0 \ 11.5)^T$$

The variance of the measurement errors have been chosen equal to:

$$dV = (3.0 \ 0.9 \ 2.4 \ 0.2 \ 0.5 \ 5.7 \ 0.3 \ 1.6)^T$$

The table 1 shows the estimation results for  $N=100$ . Since the “measurement errors” are precisely known in the simulation experiments, we are able to compute the true variance matrix of measurement errors based on the sample population. Therefore, in order to appreciate the quality of the variance estimation, the sample variances of the  $N$  repeated measurements have been calculated ; they are also given in the last column of table 1. Notice that this calculus is possible in that case as measurements concern a unique steady-state operating point. Of course, this kind of “verification” cannot be done in the general case.

There is a good correspondence between sample standard deviation of the measurements and their estimation. Of course, when the number of measurement is increasing, one obtains a better approximation.

Stream	Estimation	Estimated Std	Sample Std
1	14.989	1.744	1.747
2	7.455	0.926	0.931
3	12.624	1.595	1.603
4	3.469	0.419	0.421
5	3.985	0.796	0.800
6	16.609	2.664	2.678
7	5.089	0.513	0.516
8	11.520	1.335	1.338

Table 1. Estimations of states and variances ( $N=100$ )

For the second trial, measurements issued from four steady-state operating points have been simulated. The first comprises 15 samples, the second 25, the third 40 and the last comprises 20 samples. The "true" flowrates at the different operating points are the following:

$$X^* = \begin{pmatrix} 15.0 & 7.5 & 12.5 & 3.5 & 4.0 & 16.5 & 5.0 & 11.5 \\ 22.5 & 11.5 & 18.5 & 5.5 & 6.0 & 24.5 & 7.5 & 17.0 \\ 17.5 & 8.5 & 14.5 & 4.0 & 4.5 & 19.0 & 5.5 & 13.5 \\ 15.0 & 7.5 & 12.5 & 3.5 & 4.0 & 16.5 & 5.0 & 11.5 \end{pmatrix}^T$$

The variances of the measurement errors have been chosen identical to that of the previous trial. The table 2 presents the obtained estimation results. The four first columns are relative to the four operating points; the last column shows the estimated standard deviation.

Stream	Estimation				Estimated Std
1	15.105	22.622	17.363	14.710	1.636
2	7.499	11.666	8.450	7.501	1.041
3	12.367	18.559	14.390	12.234	1.756
4	3.503	5.605	4.004	3.530	0.475
5	3.997	6.062	4.446	3.971	0.704
6	16.364	24.621	18.836	16.205	2.334
7	4.761	7.603	5.477	5.024	0.569
8	11.603	17.017	13.359	11.336	1.200

Table 2. Estimations of states and variances

To have an idea of the noise level, the time evolution of the two flowrates number 2 and 8 are given at the bottom of that page.

### Conclusion

The proposed method is very useful in data reconciliation. The major advantage is that the weights of the measurement (which correspond to the variances of the measurement errors) are not arbitrarily fixed but estimated using the available data. Simulations have shown a good agreement between theoretical values of the variances and their estimations even if the number of measurement for each operating point is small. This method constitutes an alternative to that proposed recently by Mandel *et al.* (1998) who represent the uncertainties on the measurement error variances by intervals variables.

### References

- Almasy, G.A. and Mah, R.S.H., (1984), Estimation of measurement error variances from process data. *Ind. Eng. Chem. Process. Des. Dev.*, 23, p. 779.
- Darouach, M., Ragot, J., Zasadzinski, M. and Krzakala, G. (1989), Maximum likelihood estimator of measurement error variances in data reconciliation. IFAC Congress Advanced Information Processing in Automatic Control, AIPAC '89, Nancy, France, July 3-7.
- Keller, J.Y. and Darouach, M., (1998), Estimation of measurement errors covariance and fault detection in steady state linear systems. IFAC Workshop on "On line Detection and Supervision in the Chemical Process Industries", Solaize, France, June 4-5.
- Keller, J.Y., Darouach, M. and Krzakala, G., (1994), Analytical estimator of measurement error variances in data reconciliation. *Comp. and Chem. Eng.*, 16, p. 185.
- Mandel, D., Abdollahzadeh, A., Maquin, D. and Ragot J. (1998), Data reconciliation by inequality balance equilibration. A LMI approach. *Int. J. Min. Proc.*, 53, p. 157.

