A NOTE ON DETECTION OF GROSS ERRORS IN CHEMICAL ENGINEERING MEASUREMENTS

František MADRON

Chemopetrol, Research Institute of Inorganic Chemistry, 400 60 Ústí n.L.

Received March 15th, 1983

Detection of gross errors in chemical engineering measurements is studied, based on statistical analysis of redundant data. A method of simple evaluation of the gross error detection efficiency is presented. The construction of the power characteristics is illustrated by an example. The study complements the previous authors work on this subject, where the same problem was studied by stochastic simulation.

Several papers dealing with the detection of the gross errors in constrained data sets[1-8] have appeared in recent years. Their review can be found in the article by Crowe, Campos and Hrymak[9].This pressing problem is associated particularly with the more and more extensive use of automatic measuring systems and process computers, whose function is often adversly affected by gross and systematic errors of measurement.

The important question is the efficiency of the gross errors detection procedure, which can be expressed by the power of the test. This problem obtained only limited attention in the literature[5,10,11].

The present communication aims at an characterization of measured variables as regards the possibility of detection of gross errors. The study complements the previous authors paper on this subject[10], where the same problem was studied by stochastic simulation.

Reconciliation of Redundant Data

Let us consider a mathematical model expressed by a system of J independent linear equations in I measured variables (vector $\boldsymbol{x}$) and K unmeasured variables (vector $\boldsymbol{y}$)

$$\boldsymbol{f} + \boldsymbol{A}\boldsymbol{x} + \boldsymbol{B}\boldsymbol{y} = \boldsymbol{0} \tag{1}$$

where the vector $\boldsymbol{f}(J \times 1)$ and matrices $\boldsymbol{A}(J \times I)$ and $\boldsymbol{B}(J \times K)$ are known. The vector of measured variables $\boldsymbol{x}$ is known only approximately on the basis of measurement. It holds that

$$\boldsymbol{x}^+ = \boldsymbol{x} + \boldsymbol{e} \tag{2}$$

where $\mathbf{x}^+$ is the vector of measured values and $\mathbf{e}$ is the vector of measuring errors.

The model (*1*) is quite general. Previously described models[3,5,7] may be obtained by simplification of the model (*1*). It forms a basis for reconciliation in nonlinear implicit models[10].

Let us further suppose that $\mathbf{e}$ are random variables with I-variate normal distribution with zero mean values and known positive-definite covariance matrix. In the most frequent case the covariance matrix $\boldsymbol{F}$ is diagonal with squared standard deviations on the diagonal $F_{ii} = \sigma_i^2$.

Now, let us focus our attention on the case, when at least some measurements are redundant and all nonmeasured variables are identifiable. This assumption is expressed by the following equations and inequalities[2]

$$r(\boldsymbol{A}, \boldsymbol{B}) = J\,; \quad r(\boldsymbol{B}) = K\,; \quad I > J - K > 0 \tag{3}$$

where $r(\cdot)$ is the rank of the matrix.

The procedure for measured data treatment is following. We are trying to obtain statistically adjusted values $\hat{\boldsymbol{x}}$

$$\hat{\boldsymbol{x}} = \boldsymbol{x}^+ + \boldsymbol{v}\,, \tag{4}$$

where $\boldsymbol{v}$ are so-called adjustments. On the one hand the adjusted values must exactly satisfy the model equations (*1*), and on the other hand they have to be the minimum ones in a certain sense. The most frequently is minimized the quadratic form of adjustments (the co-called generalized method of least squares).

$$Q_{\min} = \boldsymbol{v}^{\mathrm{T}}\boldsymbol{F}^{-1}\boldsymbol{v}\,. \tag{5}$$

The solution is obtained by solving the following set of equations[2]

$$\left(\begin{array}{c|c} \boldsymbol{AFA}^{\mathrm{T}} & \boldsymbol{B} \\ \hline \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{0} \end{array}\right)\left(\begin{array}{c} \boldsymbol{k} \\ \hline \hat{\boldsymbol{y}} \end{array}\right) + \left(\begin{array}{c} \boldsymbol{f} + \boldsymbol{Ax}^+ \\ \hline \boldsymbol{0} \end{array}\right) = \boldsymbol{0} \tag{6}$$

$$\boldsymbol{v} = \boldsymbol{FA}^{\mathrm{T}}\boldsymbol{k} \tag{7}$$

$$\hat{\boldsymbol{x}} = \boldsymbol{x}^+ + \boldsymbol{v} \tag{8}$$

where $\boldsymbol{k}(J \times 1)$ is a vector of Lagrange multipliers.

If we denote

$$\left(\begin{array}{c|c} \boldsymbol{AFA}^{\mathrm{T}} & \boldsymbol{B} \\ \hline \boldsymbol{B}^{\mathrm{T}} & \boldsymbol{0} \end{array}\right)^{-1} = \left(\begin{array}{c|c} \boldsymbol{Q}_{11} & \boldsymbol{Q}_{12} \\ \hline \boldsymbol{Q}_{21} & \boldsymbol{Q}_{22} \end{array}\right) \tag{9}$$

the solution has the form

$$\hat{\mathbf{y}} = -\mathbf{Q}_{21}(\mathbf{f} + \mathbf{A}\mathbf{x}^{+}) \tag{10}$$

$$\mathbf{v} = -\mathbf{F}\mathbf{A}^{\mathrm{T}}\mathbf{Q}_{11}(\mathbf{f} + \mathbf{A}\mathbf{x}^{+}) \tag{11}$$

$$\hat{\mathbf{x}} = \mathbf{x}^{+} + \mathbf{v} \tag{12}$$

The matrices (9) have some remarkable properties, one of which is[2]

$$\mathbf{Q}_{11}\mathbf{A}\mathbf{F}\mathbf{A}^{\mathrm{T}}\mathbf{Q}_{11} = \mathbf{Q}_{11} \tag{13}$$

Further holds

$$\mathbf{v} = -\mathbf{F}\mathbf{A}^{\mathrm{T}}\mathbf{Q}_{11}\mathbf{A}\mathbf{e} \tag{14}$$

The vector of adjustments $\mathbf{v}$ has zero mean value and the covariance matrix $\mathbf{F}_{\mathrm{v}}$

$$\mathbf{F}_{\mathrm{v}} = \mathbf{F}\mathbf{A}^{\mathrm{T}}\mathbf{Q}_{11}\mathbf{A}\mathbf{F} \tag{15}$$

$Q_{\min}$ defined by equation (5) has central chi-square distribution with $\nu = J - K$ degrees of freedom $\chi^2(\nu)$.

The Presence of Gross Error

Let us suppose now, that the j-th measurement is corrupted by the gross error d_j, which will be expressed as a multiple of the standard deviation of the random error σ_j.

$$E(e_j) = d_j = q_j\sigma_j \tag{16}$$

$$E(e_i) = 0 \quad \text{for} \quad i \neq j \tag{17}$$

where $q_j = d_j/\sigma_j$ represents dimensionless gross error. The mean value of adjustments is no more zero

$$E(\mathbf{v}) = -\mathbf{F}\mathbf{A}^{\mathrm{T}}\mathbf{Q}_{11}\mathbf{A}E(\mathbf{e}) \tag{18}$$

$Q_{\min}$ has then the noncentral chi-square distribution with $\nu = J - K$ degrees of freedom with the parameter of noncentrality $\delta - \chi^2(\nu, \delta)$ (see literature[6,8]). Readers who are not familiar with this distribution may refer to the cited book[8]. The parameter δ is obtained by substitution of $E(\mathbf{v})$ in the equation (5) instead of $\mathbf{v}$.

$$\delta = [E(\mathbf{v}^{\mathrm{T}})\,\mathbf{F}^{-1}E(\mathbf{v})]^{1/2} \tag{19}$$

After substitution from equation (*14*) we have

$$\delta = [E(\mathbf{e}^T)\,\mathbf{A}^T\mathbf{Q}_{11}\mathbf{A}\,E(\mathbf{e})]^{1/2} = (d_j^2 G_{jj})^{1/2} = q_j\sigma_j(G_{jj})^{1/2} \qquad (20)$$

where $\mathbf{G} = \mathbf{A}^T\mathbf{Q}_{11}\mathbf{A}$. After the comparison of $\mathbf{G}$ with the equation (*15*), for diagonal $\mathbf{F}$ the parameter δ can be expressed as

$$\delta = q_j F_{v_{jj}}^{1/2}/\sigma_j = q_j\sigma_{v_j}/\sigma_j \qquad (21)$$

Detection of Gross Error

Let us consider the hypothesis, that the gross error is not present

$$H_0: \quad E(\mathbf{e}) = \mathbf{0} \qquad (22)$$

The hypothesis will be rejected, if it holds

$$Q_{\min} > \chi^2_{1-\alpha}(v)\,. \qquad (23)$$

The hypothesis is rejected with the probability α when it is true (the error of the first kind).

When H_0 is not correct $[E(e_j) = d_j \neq 0]$ and is not rejected, the error of the second kind has occurred. It we denote the probability of the error of the second kind by γ, the value $\beta = 1 - \gamma$ is the power of the test. β is the function of the parameters v and δ and can be found in standard graphs[8].

The important values of δ are defined by $\beta = 0{\cdot}5$ and $\beta = 0{\cdot}9$ (the power of the test is 50 and 90%). These values of δ will be further denoted by δ_{50} and δ_{90}. As

$$\beta = P[\chi^2_{1-\alpha}(v) < \chi^2(v, \delta)] \qquad (24)$$

the value δ_{50} is defined by

$$0{\cdot}5 = P[\chi^2_{1-\alpha}(v) < \chi^2(v, \delta_{50})] \qquad (25)$$

and analogously for δ_{90}. The values δ_{50} and δ_{90} for $\alpha = 0{\cdot}05$ were computed from graphs[8] and are in Table I.

The table can serve for construction of power characteristics of the tests (curves of β as the function of magnitude of gross error). In the coordinates $|q_j| - \beta$ the power characteristics has three important points (Fig. 1).
For

$$q_j = d_j/\sigma_j = 0 \quad \text{obviously} \quad \beta = \alpha \qquad (26)$$

$$|q_j| = |d_j/\sigma_j| = \delta_{50}\sigma_j/\sigma_{v_j}\,; \quad \beta = 0{\cdot}5 \tag{27}$$

$$|q_j| = |d_j/\sigma_j| = \delta_{90}\sigma_j/\sigma_{v_j}\,; \quad \beta = 0{\cdot}9 \tag{28}$$

Equations (*27*, *28*) result from Eq. (*21*). Values q_j defined by equations (*27*) and (*28*) will further be denoted by $q_{j,50}$ and $q_{j,90}$. For example, the value $q_{j,50}$ means that the absolute value of error d_j must be at least $q_{j,50}\sigma_j$ in order to detect it with the probability at least 50%.

In real cases, the number of measured variables is high and the construction of power characteristics may be inconvenient in practice. The values $q_{j,50}$ and $q_{j,90}$ then represent simple characteristics of measured variables from the point of view of detection of their gross errors.

The measured variables, which are not redundant, have adjustments identically zero and also $\sigma_{v_i} = 0$. The detection of gross errors of these variables is not possible ($v_i = 0$ regardless the magnitude of their errors).

Example. Four mass flows connected with the chemical reactor are measured[7]. There are three elemental balances

$$\begin{aligned} 0{\cdot}1x_1 + 0{\cdot}6x_2 - 0{\cdot}2x_3 - 0{\cdot}7x_4 &= 0 \\ 0{\cdot}8x_1 + 0{\cdot}1x_2 - 0{\cdot}2x_3 - 0{\cdot}1x_4 &= 0 \\ 0{\cdot}1x_1 + 0{\cdot}3x_2 - 0{\cdot}6x_3 - 0{\cdot}2x_4 &= 0 \end{aligned} \tag{29}$$

The vector of standard deviations $\boldsymbol{\sigma}^T = 0{\cdot}017, 0{\cdot}05, 0{\cdot}024, 0{\cdot}2)$. Eq. (*29*) is the model (*1*) without $\mathbf{B}, \mathbf{y}$ and $\mathbf{f}$. $I = 4$, $J = 3$ and $K = 0$. After the reconciliation we have the covariance matrix of adjustments $\mathbf{F}_v$. The standard deviations of adjustments are square roots of the diagonal elements

TABLE I

Values of δ_{50} and δ_{90} for $\alpha = 0{\cdot}05$

ν	δ_{50}	δ_{90}	ν	δ_{50}	δ_{90}
1	1·96	3·24	11	3·09	4·60
2	2·23	3·56	12	3·15	4·67
3	2·40	3·76	13	3·20	4·74
4	2·53	3·93	14	3·25	4·80
5	2·64	4·06	15	3·30	4·86
6	2·74	4·17	16	3·34	4·91
7	2·82	4·28	17	3·38	4·96
8	2·90	4·37	18	3·42	5·02
9	2·97	4·45	19	3·46	5·06
10	3·03	4·53	20	3·50	5·11

of $\mathbf{F}_v - \boldsymbol{\sigma}_v^T = (0{\cdot}0169, 0{\cdot}0241, 0{\cdot}0215, 0{\cdot}197)$. The number of degrees of freedom $v = 3$ and from Table I we have $\delta_{50} = 2{\cdot}40$ and $\delta_{90} = 3{\cdot}76$. The values $q_{j,50}$ and $q_{j,90}$ can be calculated from Eq (*27*, *28*). For example $q_{1,50} = 2{\cdot}40 \times 0{\cdot}017/0{\cdot}0169 = 2{\cdot}41$. The gross error of x_1 must be at least $2{\cdot}41\sigma_1 = 0{\cdot}041$ in order to detect it with the probability at least 50%. The results are summarized in Table II where are also the values obtained by the method described by Mah and Tamhane[5] (the values are denoted by apostrophes). The individual power characteristics are on Fig. 2.

The described method enables simple detection of the presence of gross errors and at the same place the characterization of individual measured variables from the point of view of detection of their gross errors.

The method differs somewhat from the alternative method described by Mah and Tamhane[5] (further denoted as the second method). The chi-square test is mostly slightly more powerful than the test used by the second method (Table II). The difference is more significant in real cases, where the number of measured variables

TABLE II

Values of $q_{j,50}$ and $q_{j,90}$

j	$q_{j,50}$	$q_{j,90}$	$q'_{j,50}$	$q'_{j,90}$
1	2·41	3·78	2·52	3·79
2	4·98	7·82	5·21	7·85
3	2·67	4·20	2·79	4·21
4	2·44	3·83	2·55	3·83

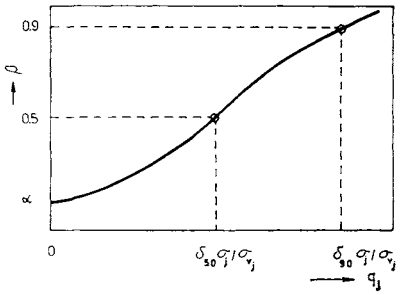

FIG. 1
Construction of the power characteristics

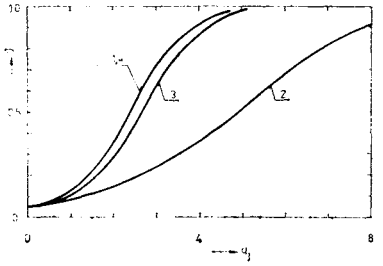

FIG. 2
The power characteristics

is high (for example 50 or more). This follows from the fact, that the test used by the second method is conservative (the exact value of the probability of the IInd type error is not known).

The test statistics in the second method are in essence the normalized adjustments v_i/σ_{v_i} (for diagonal F). The nonredundant measured variables (for which $\sigma_{v_i} = 0$) must be therefore excluded from tests. Problems however sustain when variable is almost nonredundant ($\sigma_{v_i} \to 0$). Experiences from real cases show, that due to computational errors in these cases are often detected nonexisting gross errors. The chi-square test does not have this disadvantage. This fact represents the most important advantage of the chi-square test. On the other hand, the chi-square test is not convenient for identification of the source of gross error. For this purpose we recomend other methods, described for example in the paper[9].

Simulation by Monte Carlo method showed, that the described method can be efficiently used even if the mathematical model is non linear[3]. The equations of the mathematical model are linearized to the form of Eq. (*1*) and the problem is solved as earlier described[10].

LIST OF SYMBOLS

d_j	gross error of j-th measured variable
$\mathbf{e}$	vector ($I \times 1$) of random errors e_i
$E(.)$	expectation of
$\mathbf{F}, \mathbf{F}_v$	covariance matrices of vectors $\mathbf{e}$ and $\mathbf{v}$
I	number of measured variables
J	number of equations
K	number of nonmeasured variables
$P.$	probability of
q_j	relative gross error d_j/σ_j
$q_{j,50}, q_{j,90}$	values of q_j for $\beta = 0{\cdot}5$ and $0{\cdot}9$
$Q_{\min}$	quadratic form of adjustments [Eq. (*5*)]
$\mathbf{v}$	vector ($I \times 1$) of adjustments v_i
$\mathbf{x}, \mathbf{x}^+, \hat{\mathbf{x}}$	vectors ($I \times 1$) of correct, measured, and adjusted values of measured variables
$\mathbf{y}, \hat{\mathbf{y}}$	vectors ($K \times 1$) of correct and estimated values of nonmeasured variables
α	level of significance
β	power of the test
δ	parameter of noncentrality of χ^2 distribution defined by equation (*19*)
δ_{50}, δ_{90}	δ for $\beta = 0{\cdot}5$ and $0{\cdot}9$
$\chi^2(\nu)$	central chi-square random variable with ν degrees of freedom
$\chi^2(\nu, \delta)$	noncentral chi-square random variable with ν degrees of freedom and the parameter of noncentrality δ
$\chi^2_{1-\alpha}(\nu)$	$(1 - \alpha)$ th quantile of distribution $\chi^2(\nu)$
σ, σ_v	vectors ($I \times 1$) of standard deviations of vectors $\mathbf{e}$ and $\mathbf{v}$

REFERENCES

1. Knepper J. C., Gorman J. W.: AIChE J. *26*, 260 (1980).
2. Kubáček L.: Matemat. čas. *19*, 270 (1969).
3. Madron F.: *Measurement in Chemical Industry. Errors, Data Treatment and Planning of Experiments* (in Czech). To be published by SNTL, Prague 1985.
4. Madron F., Veverka V., Vaněček V.: AIChE J. *23*, 482 (1977).
5. Mah R. S. H., Tamhane A. C.: AIChE J. *28*, 828 (1982).
6. Rao R. C.: *Linear Statistical Inference and its Applications.* Wiley, New York 1973.
7. Ripps D. L.: Chem. Eng. Progr., Symp. Ser. 55, *61*, 8 (1965).
8. Scheffé H.: *The Analysis of Variance.* Wiley, New York 1959.
9. Crowe C. M., Garcia Campos Y. A., Hrymak A.: AIChE J. *29*, 88 (1983).
10. Madron F.: This Journal *49*, 2614 (1983).
11. Nogita S.: Ind. Eng. Chem. Process Des. Develop. *11*, 197 (1972).