Using characteristics of the residuals as an assessment of the effectiveness of processing the results of measurements

Vladimir Marchuk[1,*], *Alexander* Minaev[1], and *Igor* Shrafel[1]

[1]Don State Technical University, 344000 Rostov-on-Don, Russia

Abstract. The article considers an analytical assessment of the use of characteristics of residuals for evaluating the efficiency of processing measurement results. The proved is provided that the method of reproduction of estimates is highly effective in processing the measurement results.

1 Introduction

At present, in practical implementation of the processing of measurement results, in many cases, analysis of residuals is used as a criterion for extracting a useful signal [1–4]. This is due to the fact that any model, however accurate it may be, assumes the presence of residuals - deviations of the theoretical values of a series from empirical data. In the absence of a useful component model, the standard deviation criterion is usually used. The application of this criterion is widely used in practice. In this regard, we will prove analytically using the example that using the method of reproduction of estimates of the useful component [5-12] allows the dispersion of residuals to be approximated to the value of the dispersion of the additive noise component. It should be noted that in terms of efficiency, this criterion is inferior to the criterion for minimizing the root-mean-square error, using the useful component model, since it requires a priori information about the dispersion of the additive noise component and there is no pronounced minimum error.

2 Research

In this paper, we will consider the case of calculating the variance of the estimate of the constant signal by the method of multiplying the estimates with a random division of the implementation of the original process into two parts [5-6]. Define arbitrary positive integers n, l, m such that

$$m \le n - 2l + 1, \tag{1}$$

* Corresponding author: marchuk@sssu.ru

and consider a set of n pairwise uncorrelated random variables $x_1,\ldots,x_n$ with equal mathematical expectations and equal variances: $Mx_k = a$, $Dx_k = D$, $1 \le k \le n$. The value a will be considered unknown, and the random vector $(x_1,\ldots,x_n)$ - the sum of the constant signal a and noise at the moments of time $t_1 < \ldots < t_n$. As a result we get the implementation of a random process $(x_k)_{k=1}^n$. Estimate the value a using the method of reproduction estimates. We randomly divide m times the set $\{1,\ldots,n\}$ into two parts U_j, V_j, so that for all the following conditions are met:

1) $|U_j| \ge l$, $|V_j| \ge l$,(hereinafter $|A|$ is the number of elements of the set A);

2) $\max U_j < \min V_j$;

3) $U_j \ne U_i$, $i = 1,\ldots,j-1$ (with $j = 1$);

4) the probability of choosing a pair U_j, V_j is considered equal for all pairs sets that satisfy conditions 1-3.

From inequality (1) it follows that the described m-parts partition of the set $\{1,\ldots,n\}$ is always possible. Estimation a_{kj} of the signal a in the k-th section $(1 \le k \le n)$ at the multiplication of estimates j $(1 \le j \le m)$, is the average value of the estimates of this signal found by the least squares method (LSM) in the part $(x_i)_{i \in T_j}$ of the original implementation, where through T_j the one of the sets U_j, V_j the number belongs to is denoted. In our case, the LSM are used - estimates of zero degree polynomials:

$$\overline{a}_{kj} = \frac{1}{|T_j|}\sum_{i \in T_j} x_i \text{, then } \overline{a}_k = \frac{1}{m}\sum_{j=1}^{m} a_{kj} \text{ .} \qquad (2)$$

Obviously, $\overline{a}_{kj}$,and therefore $\overline{a}_k$ are unbiased estimates of a. The result of a m-times division of the set $\{1,\ldots,n\}$ into parts U_j, V_j, that satisfy conditions 1–4, can be represented as a random chosen subset A with m elements of the set $\{l,\ldots,n-l\}$, each element p of which symbolizes a certain pair of sets $U_j = \{1,\ldots,p\}$, $V_j = \{p+1,\ldots,n\}$. Because of condition 4, the random choice A has a uniform distribution law. Consequently, the probability P of its results is the inverse of the number of ways in which such a choice can be made $P = \dfrac{1}{C_{n-2l+1}^m}$. The average variance $D_{cp} = \dfrac{1}{n}\sum_{k=1}^{n} D_{a_k}$ is used as the measure of the quality of the assessment multiplication method. It is necessary to calculate D_{cp} and prove that with constant n, l and increasing m quality of processing improves, i.e. D_{cp} decreases. Note that inequality $n \ge 2l$ follows from (1).

3 Solution

The trivial case $n = 2l$ was considered in [6]; therefore, we consider only the main case $n \geq 2l+1$. The sum $\sum_{j=p}^{q}$ at $p > q$ is considered equal to zero. Fix arbitrary $1 \leq k \leq n$. Now

$$a_k = \frac{1}{m}\sum_{j=1}^{m} a_{kj} = \frac{1}{m}\sum_{i=1}^{n} x_i \left(\sum_{\substack{j=l \\ j\in A}}^{\min(i,k)-1} \frac{1}{n-j} + \sum_{\substack{j=\max(i,k) \\ j\in A}}^{n-l} \frac{1}{j} \right) = \\ = \frac{1}{m}\sum_{i=1}^{m} x_i \left(\sum_{j=l}^{\min(i,k,n-l)-1} \frac{\tau_j}{n-j} + \sum_{j=\max(i,k,l)}^{n-l} \frac{\tau_j}{j} \right), \quad (3)$$

where $\tau_j = \left| A \cap \{j\} \right|$. The first sum in brackets can be converted to $\sum_{j=n+1-\min(i,k,n-l)}^{n-l} \frac{\tau_{n-j}}{j}$.

Consequently, $a_k = \frac{1}{m}\sum_{i=1}^{n} x_i y_{ki}$, where $y_{ki} = \sum_{j=\max(n-i,n-k,l)+1}^{n-l} \frac{\tau_{n-j}}{j} + \sum_{j=\max(i,k,l)}^{n-l} \frac{\tau_j}{j}$.

Then

$$D_{a_k} = M\left[\left(\frac{1}{m}\sum_{i=1}^{m} x_i y_{ki} \right)^2 \right] - \left(M_{a_k} \right)^2 = M\left(\frac{1}{m^2}\sum_{i,p=1}^{n} x_i x_p y_{ki} y_{kp} \right) - a^2. \quad (4)$$

Given the independence of random variables $x_i x_p$ and $y_{ki} y_{kp}$ and pairwise uncorrelated random variables $x_1, \ldots, x_n$, we obtain

$$D_{a_k} = \frac{1}{m^2}\sum_{i,p=1}^{n} M\left(x_i x_p\right) M\left(y_{ki} y_{kp}\right) - a^2 = \\ = \frac{1}{m^2}\left(\sum_{i,p=1}^{n} M\left(x_i^2\right) M\left(y_{ki}^2\right) + \sum_{\substack{i,p=1 \\ i\neq p}}^{n} M_{x_i} M_{x_p} M\left(y_{ki} y_{kp}\right) \right) - a^2 = \\ = \frac{1}{m^2}\left(\sum_{i=1}^{n} \left(D + a^2\right) M\left(y_{ki}^2\right) + a^2 \sum_{\substack{i,p=1 \\ i\neq p}}^{n} M\left(y_{ki} y_{kp}\right) \right) - a^2 = \quad (5) \\ = \frac{1}{m^2}\left(D\sum_{i=1}^{n} M\left(y_{ki}^2\right) + a^2 \sum_{i,p=1}^{n} M\left(y_{ki} y_{kp}\right) \right) - a^2 =$$

$$= \frac{D}{m^2}\sum_{i=1}^{n} M\left(y_{ki}^2\right) + \frac{a^2}{m^2} M\left(\sum_{i,p=1}^{n} M\left(y_{ki}y_{kp}\right)\right) - a^2 =$$

$$= \frac{D}{m^2}\sum_{i=1}^{n} M\left(y_{ki}^2\right) + \frac{a^2}{m^2} M\left(\left(\sum_{i,p=1}^{n} y_{ki}\right)^2\right) - a^2 = \frac{D}{m^2}\sum_{i=1}^{n} M\left(y_{ki}^2\right),$$

because $\frac{1}{m}\sum_{i=1}^{n} y_{ki}$ is the sum of random coefficients of a linear function $a_k = a_k\left(x_1,\ldots,x_n\right)$, which is always equal to 1. Find the expected value of random variables $\tau_i, \tau_j \ \left(l \le i, j \le n-l\right)$ with $i = j \ \ \tau_i\tau_j = \tau_i^2 = \tau_i$

$$M\left(\tau_i^2\right) = \frac{C_{n-2l}^{m-1}}{C_{n-2l+1}^{m}} = \frac{(n-2l)!m!(n-2l+1-m)!}{(m-1)!(n-2l-m+1)!(n-2l+1)!} = \frac{m}{n-2l+1} \tag{6}$$

If $i \ne j$, then

$$M\left(\tau_i\tau_j\right) = \frac{C_{n-2l-1}^{m-2}}{C_{n-2l+1}^{m}} = \frac{(n-2l-1)!m!(n-2l+1-m)!}{(m-2)!(n-2l-m+1)!(n-2l+1)!} = \frac{m(m-1)}{(n-2l+1)(n-2l)} \tag{7}$$

This expression is also true when $m = 1$, because then $\tau_i, \tau_j = 0$ with probability 1. For all $\tau_i, \tau_j = 0$ we have

$$M\left(y_{ki}^2\right) = M\left[\left(\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\frac{\tau_{n-j}}{j}\right)^2 + \left(\sum_{j=\max(i,k,l)}^{n-l}\frac{\tau_j}{j}\right)^2 + 2\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\frac{\tau_{n-j}}{j}\sum_{j=\max(i,k,l)}^{n-l}\frac{\tau_{n-j}}{j}\right] =$$

$$= M\left[\sum_{j,p=\max(n-i,n-k,l)+1}^{n-l}\frac{\tau_{n-j}\tau_{n-p}}{jp} + \sum_{j,p=\max(i,k,l)}^{n-l}\frac{\tau_j\tau_p}{jp} + 2\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\sum_{p=\max(i,k,l)}^{n-l}\frac{\tau_{n-j}\tau_p}{jp}\right] =$$

$$= \sum_{j,p=\max(n-i,n-k,l)+1}^{n-l}\frac{M\left(\tau_{n-j}\tau_{n-p}\right)}{jp} + \sum_{j,p=\max(i,k,l)}^{n-l}\frac{M\left(\tau_j\tau_p\right)}{jp} + 2\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\sum_{p=\max(i,k,l)}^{n-l}\frac{M\left(\tau_{n-j}\tau_p\right)}{jp} =$$

$$= \frac{m}{n-2l+1}\left(\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\frac{1}{j^2} + \sum_{j=\max(i,k,l)}^{n-l}\frac{1}{j^2}\right) + \frac{m(m-1)}{(n-2l+1)(n-2l)}\times$$

$$\times\left(\sum_{\substack{j,p=\max(n-i,n-k,l)+1\\ j\ne p}}^{n-l}\frac{1}{jp} + \sum_{\substack{j,p=\max(i,k,l)\\ j\ne p}}^{n-l}\frac{1}{jp} + 2\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\sum_{p=\max(i,k,l)}^{n-l}\frac{1}{jp}\right) =$$

$$= \frac{m(n-2l-m+1)}{(n-2l+1)(n-2l)}\left(\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\frac{1}{j^2}+\sum_{j=\max(i,k,l)}^{n-l}\frac{1}{j^2}\right)+\frac{m(m-1)}{(n-2l+1)(n-2l)}\times$$

$$\times\left(\sum_{j,p=\max(n-i,n-k,l)+1}^{n-l}\frac{1}{jp}+\sum_{j,p=\max(i,k,l)}^{n-l}\frac{1}{jp}+2\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\sum_{p=\max(i,k,l)}^{n-l}\frac{1}{jp}\right).$$

Sum up by i,k from 1 to n each of the five sums in brackets.

$$\sum_{i,k=1}^{n}\sum_{j=\max(n-i,n-k,l)+1}^{n-l}\frac{1}{j^2}=\sum_{j=l+1}^{n-l}\frac{1}{j^2}\sum_{i,k=n-j+1}^{n}1=\sum_{j=l+1}^{n-l}1=n-2l;$$

$$\sum_{i,k=1}^{n}\sum_{j=\max(i,k,l)+1}^{n-l}\frac{1}{j^2}=\sum_{j=l}^{n-l}\frac{1}{j^2}\sum_{i,k=1}^{j}1=\sum_{j=l}^{n-l}1=n-2l+1;$$

$$\sum_{i,k=1}^{n}\sum_{j,p=\max(n-i,n-k,l)+1}^{n-l}\frac{1}{jp}=\sum_{j,p=l+1}^{n-l}\frac{1}{jp}\sum_{i,k=\max(n-j,n-p)+1}^{n}1=$$

$$=\sum_{j,p=l+1}^{n-l}\frac{\left(\min(j,p)\right)^2}{jp}=\sum_{j,p=l+1}^{n-l}\frac{\min(j,p)}{\max(j,p)}=\sum_{j=l+1}^{n-l}\frac{\min(j,j)}{\max(j,j)}+\sum_{\substack{j,p=l+1\\ j>p}}^{n-l}\frac{p}{j}+\sum_{\substack{j,p=l+1\\ j<p}}^{n-l}\frac{j}{p}=$$

$$=\sum_{j=l+1}^{n-l}1+2\sum_{\substack{j,p=l+1\\ j>p}}^{n-l}\frac{p}{j}=n-2l+2\sum_{j=l+2}^{n-l}\frac{1}{j}\sum_{p=l+1}^{j-1}p=n-2l+\sum_{j=l+2}^{n-l}\frac{(j+l)(j-l-1)}{j}=$$

$$=n-2l+\sum_{j=l+2}^{n-l}(j-1)-l(l+1)\sum_{j=l+2}^{n-l}\frac{1}{j}=n-2l+\frac{n(n-2l-1)}{2}-l(l+1)\sigma_{l+2}=$$

(hereafter the designation $\sigma_j=\sum_{i=j}^{n-l}\frac{1}{i}$ is used; note that when $j>n-l$ $\sigma_j=0$)

$$=\frac{n^2+n-2l\cdot n-4l}{2}-l(l+1)\sigma_{l+2};$$

$$\sum_{i,k=1}^{n}\sum_{j,p=\max(i,k,l)}^{n-l}\frac{1}{jp}=\sum_{j,p=l}^{n-l}\frac{1}{jp}\sum_{i,k=1}^{\min(j,p)}1=\sum_{j,p=l}^{n-l}\frac{\left(\min(j,p)\right)^2}{jp}=$$

$$=\sum_{j,p=l+1}^{n-l}\frac{\left(\min(j,p)\right)^2}{jp}+\frac{\left(\min(l,l)\right)^2}{l^2}+\sum_{p=l+1}^{n-l}\frac{l^2}{lp}+\sum_{j=l+1}^{n-l}\frac{l^2}{jl}=$$

$$= \frac{n^2 + n - 2l \cdot n - 4l}{2} - l(l+1)\sigma_{l+2} + 1 + 2l\sigma_{l+1};$$

$$\sum_{i,k=1}^{n} \sum_{j=\max(n-i,n-k,l)+1}^{n-k} \frac{1}{j} \sum_{h=\max(i,k,l)}^{n-l} \frac{1}{p} = \sum_{j=l+1}^{n-l} \frac{1}{j} \sum_{p=l}^{n-l} \frac{1}{p} \sum_{i,k=n+1-j}^{p} 1 =$$

$$= \sum_{j=l+1}^{n-l} \frac{1}{j} \sum_{p=\max(l,n+1-j)}^{n-l} \frac{p-n+j}{p} = \sum_{j=l+1}^{n-l} \frac{1}{j} \sum_{p=n+1-j}^{n-l} 1 + \sum_{j=l+1}^{n-l} \sum_{p=n+1-j}^{n-l} \frac{1}{p} - n\sigma =$$

(Hereafter $\sigma = \sum_{j=l+1}^{n-l} \frac{1}{j} \sum_{p=n+1-j}^{n-l} \frac{1}{p}$)

$$= \sum_{j=l+1}^{n-l} \frac{j-l}{j} + \sum_{p=l+1}^{n-l} \frac{1}{p} \sum_{j=n+1-p}^{n-l} 1 - n\sigma = 2\sum_{j=l+1}^{n-l} 1 - 2l\sigma_{l+1} - n\sigma = 2n - 4l - 2l\sigma_{l+1} - n\sigma .$$

Thereby,

$$D_{cp} = \frac{1}{n}\sum_{k=1}^{n} D_{a_k} = \frac{D}{nm^2}\sum_{i,k=1}^{n} M\left(y_{ki}^2\right) = \frac{D}{nm(n-2l+1)(n-2l)} \cdot \left[(n-2l-m+1)\cdot\right.$$

$$\cdot(n-2l+n-2l+1) + (m-1)\left(2\left(\frac{n^2+n-2l\cdot n-4l}{2} - l(l+1)\sigma_{l+2}\right) + 1 + 2l\sigma_{l+1} + \right.$$

$$\left. + 2(2n-4l-2l\sigma_{l+1}-n\sigma)\right)\Big] = \frac{D}{nm(n-2l+1)(n-2l)}\Big[(n-2l)(2n-4l+1) + (m-1)\cdot$$

$$\cdot\left(-2n+4l-1+n^2+n-2l\cdot n-4l+4n-8l-2l(l+1)\sigma_{l+2}-2l\sigma_{l+1}-2n\sigma+1\right)\Big] =$$

$$= \frac{D}{nm(n-2l+1)(n-2l)}\Big[(n-2l)(2n-4l+1) + (m-1)\Big(n^2-2l\cdot n+3n-8l-2l(l+1)\sigma_{l+2} -$$

$$\left. -2l\left(\sigma_{l+2}+\frac{1}{l+1}\right) - 2n\sigma\right)\Big] = \frac{D}{nm(n-2l+1)(n-2l)}\Big[(n-2l)(2n-4l+1) + (m-1)\times$$

$$\times\left(n^2-2l\cdot n+3n-8l-\frac{2l}{l+1}-2l(l+2)\sigma_{l+2}-2n\sigma\right)\Big].$$

Let us prove that with an increase in the number of reproduction estimates m, the mean variance decreases. The resulting ratio is

$$D_{cp} = \frac{\alpha}{m}(\beta + (m-1)\gamma), \tag{8}$$

where α, β and γ do not depend on m. Now it is required to prove inequality

$$\frac{\alpha}{m}\left(\beta+(m-1)\gamma\right)>\frac{\alpha}{m+1}\left(\beta+m\gamma\right)\Leftrightarrow(m+1)\left(\beta+(m-1)\gamma\right)>m\left(\beta+m\gamma\right)\Leftrightarrow\beta>\gamma,$$

i.e. inequality

$$(n-2l)(2n-4l+1)>n^2-2l\cdot n+3n-8l-\frac{2l}{l+1}-2l(l+2)\sigma_{l+2}-2n\sigma\Leftrightarrow$$

$$\Leftrightarrow 2n^2-4l\cdot n+n-4l\cdot n+8l^2-2l-n^2+2l\cdot n-3n+8l+\frac{2l}{l+1}+2l(l+2)\sigma_{l+2}+2n\sigma>$$

$$>0\Leftrightarrow n^2-6l\cdot n+8l^2-2n+6l+\frac{2l}{l+1}+2l(l+2)\sigma_{l+2}+2n\sigma>0. \qquad (9)$$

For each $l\geq 1$, we prove the increase of the sequence $(\phi_n)_{n=2l+1}^{\infty}$, where ϕ_n is the left side of inequality (2), reduced by $2n\sigma$. With any $n\geq 2l+1$ we get

$$\phi_{n+1}=\phi_n=(n+1)^2-6l(n+1)+8l^2-2(n+1)+6l+\frac{2l}{l+1}+2l(l+2)\sum_{j=l+2}^{n+1-l}\frac{1}{j}-$$

$$-n^2+6l\cdot n-8l^2+2n-6l-\frac{2l}{l+1}-2l(l+2)\sum_{j=l+2}^{n+l}\frac{1}{j}=2n+1-6l-2+\frac{2l(l+2)}{n+1-l}=$$

$$=\frac{(2n-6l-1)(n+1-l)+2l^2+4l}{n+1-l}=\frac{2n^2-8l\cdot n+8l^2+n-l-1}{n+1-l}=$$

$$=\frac{2(n-2l)^2+n-l-1}{n+1-l}>\frac{0+2l+1-l-1}{n+1-l}>0\cdot$$

Thus, the required thesis is proved. Now we find

$$\phi_{2l+1}=(2l+1)^2-6l(2l+1)+8l^2-2(2l+1)+6l+\frac{2l}{l+1}+2l(l+2)\sum_{j=l+2}^{l+1}\frac{1}{j}=$$

$$=4l^2+4l+1-12l^2-6l+8l^2-4l-2+6l+\frac{2l}{l+1}=\frac{l-1}{l+1}\geq 0.$$

From all this it follows that $\phi_n\geq 0$ for all $n\geq 2l+1$. Note also that with the same values n $\sigma=\sum_{j=l+1}^{n-l}\frac{1}{j}\sum_{p=n+1-j}^{n-l}\frac{1}{p}>\frac{1}{(n-l)^2}>0$. Therefore $\phi_n+2n\sigma>0$, i.e. inequality (9) holds for all $n\geq 2l+1$. Thus, with increasing m value D_{cp} decreases.

In practice, in many cases, the residual analysis serves as a criterion for the quality of models, which allows to determine whether a series of residuals is Gaussian noise or if there are any interdependencies in its structure.

Conclusion

1. It is analytically proved that when using the method of reproduction of estimates for extracting a signal under conditions of a priori uncertainty, the value D_{cp} decreases with m increasing.

2. The obtained ratios make it possible to use comparative analysis of residual characteristics (correlation function, differential distribution density, etc.) as quality criteria for processing measurement results.

References

1. S.M. Perevertkin, A.V. Kantor, N.F. Borodin, T.S. Shcherbakova, *Onboard telemetry equipment of spacecraft* (Mashinostroenie, Moscow, 1977)
2. T. Anderson, *Statistical analysis of time series* (Mir, Moscow, 1976)
3. J. Bendat, A. Pirsol, *Applied analysis of random data* (Mir, Moscow, 1989)
4. M. Kendall, J. A. Stuart, *Statistical findings and communication* (Science, Moscow, 1973)
5. V.I. Marchuk, *Primary processing of measurement results with a limited amount of a priori information: Monograph* (Publishing house TRGU , Taganrog, 2003)
6. V.I. Marchuk, I.S. Shrayfel, *Methods for isolating the useful component with a priori uncertainty and a limited amount of measurement results*. (Publishing House SSSU Shakhty, 2007)
7. V.I. Marchuk, Measuring equipment, **12**, 3 (2003)
8. V.I. Marchuk, K.E. Rumyantsev, Aerospace Instrument Engineering, **4**, 48 (2004)
9. V.I. Marchuk, G.R. Saakyan, Review of Applied and Industrial Mathematics, **9**, 1, 126 (2002)
10. V.I. Marchuk, O.V. Saakyan, G.R. Saakyan, News of universities. North Caucasus region. Technical science, 3, 67 (1999)
11. M. A. Budin, IEEE Transactions on systems, MAN, and cybernetics, 4, 371 (1973)
12. V.I. Marchuk, Serbian journal of electrical engineering, **15**, 3, 365 (2018)