Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 28 (2025), No. 5, pp. 821–837
DOI : 10.47974/JSMS-1206

Estimation of the number of faulty components of outliers in signal processing

Madhusudan Bhandary
Department of Mathematics
Columbus State University
4225 University Avenue
Columbus
GA 31907
U.S.A.

Abstract

Methods of estimation of the number of faulty components in the outliers in signal processing have been suggested. The problem is solved for both mean-slippage and dispersion-slippage outliers by Roy's [8] union-intersection principle and by modified information criterion proposed by the author. The estimators are strongly consistent.

Subject Classification (2020): *Primary 62H15.*

Keywords: *Dispersion-slippage outliers, Mean-slippage outliers, Information criterion, Signal processing, Union-intersection principle.*

1. Introduction

The theory of outliers is quite popular in the literature. The procedure to detect outliers in signal processing is discussed in Bhandary [2]. Suppose somebody is collecting data in the area of signal processing. During this process, suddenly some blizzard or heavy rain or heavy storm or some other natural disturbance occurred so that some of the receivers (out of p receivers, which are placed in p different locations far from one another), where the natural disturbances occurred, received defective signals. After the process of collection of data, everything is put together. In this data naturally some outliers are mixed with the regular data. Using the procedure of outlier detection in signal processing by Bhandary [2] and estimation of number of outliers by Bhandary [3], we can separate between

E-mail: **bhandary_madhusudan@columbusstate.edu**

outliers and non-outliers. Now, among the outliers, our goal is to find out the number of defective components caused by the natural disturbance in some of receivers during the process of data collection. All the observations among the outliers in this case will have equal number of defective components due to the natural disturbance in the same receivers while collecting the data. Without loss of generality we can put these defective components among the outliers into the last components of these observations. The problem is solved first by Roy's [8] union-intersection principle and some drawbacks of the estimate in this case are discussed and then by modified information criterion proposed by the other.

Notation and the model (also outlier model) in signal processing are described in sectin 2.

In section 3, we state the solution by union-intersection principle and the drawbacks of the solution in this case.

The solution by modified information theoretic criteria and the strong consistency of the estimator are discussed in section 4.

In section 5, we do the same as in section 4 for the dispersion-slippage outlier model.

Some simulation results are presented in section 6.

2. Notation and the model

The general model in signal processing is described as follows:

$$\underset{\sim}{x}(t) = A\underset{\sim}{s}(t) + \underset{\sim}{n}(t) \tag{2.1}$$

where $\underset{\sim}{x}(t) = (x_1(t), \ldots, x_p(t))' = p \times 1$ vector of observations at time t.

$\underset{\sim}{s}(t) = (s_1(t), \ldots, s_p(t))' = q \times 1$ random signal vector at time t.

$\underset{\sim}{n}(t) = (n_1(t), \ldots, n_p(t))' = p \times 1$ vector of random noise at time t.

$A = [A(\Phi_1), \ldots, A(\Phi_q)] = p \times q$ matrix of coefficients which are not known.

$A(\Phi_i): p \times 1$ vector of functions of the elements of unknown Φ_i related with i^{th} signal and $q < p$.

In model (2.1) above, $\underset{\sim}{x}(t)$ is assumed to be distributed as normal with zero mean vector and covariance matrix $A\Psi A' + \sigma^2 I_p = \Gamma + \sigma^2 I_p$, where $\Gamma = A\Psi A'$ and is of rank $q(<p)$ and $\Psi =$ covariance matrix $\underset{\sim}{s}(t)$.

The mean-slippage and dispersion-slippage outlier structure of observations in signal processing is discussed in Bhandary [2]. In our problem, the outlier model is described as follows:

$$\underset{\sim}{x}_i \sim N_p(\underset{\sim}{o}, \Gamma + \sigma^2 I_p), \;\; i = 1, \ldots, n_1, \tag{2.2}$$

$$\underset{\sim}{x}_{0i} \sim N_p(\underset{\sim}{r}, \Gamma + \sigma^2 I_p), \;\; i = n_1 + 1, \ldots, n; \;\; \text{where } n = n_1 + n_2$$

$$\text{where,} \qquad \underset{p\times 1}{\underset{\sim}{r}} = \begin{pmatrix} \underset{\sim}{o} \\ \underset{(p-m)\times 1}{\underset{\sim}{a}} \end{pmatrix} \text{ and } \underset{\sim}{a} = \begin{pmatrix} a_{m+1} \\ . \\ . \\ . \\ a_p \end{pmatrix}, \; o \le m \le p-1$$

and $\underset{m\times 1}{\underset{\sim}{o}}$ is an $m \times 1$ vector of zero's and N_p means p-variate normal distribution.

Here n_2 observations $\underset{\sim}{x}_{0i}(i = n_1 + 1, \ldots, n)$ are mean-slippage outliers and *(p-m)* components in outliers are defective. Actually estimation of number of outliers present in a data set is discussed in Bhandary [3]. Using the technique of Bhandary [3] we can separate between outliers and non-outliers. Our problem is to estimate *m*. We have assumed here without loss of generality that the last *(p-m)* components in the outliers are defectives.

3. Solution by union-intersection principle

We want to test $H_{0k}: \; m = k$ vs. $H_{1k}: \; m < k, \;\; k = 0, 1, \ldots, (p-1)$.
Alternative hypotheses means that m is less than k.
Let

$$H = \left\{ \underset{\sim}{w} : \underset{\sim}{w} = \begin{pmatrix} \underset{\sim}{w}_o \\ \underset{(\text{p-k})\times 1}{\underset{\sim}{o}} \end{pmatrix}, \; \underset{\sim}{w}_o \ne \underset{\sim}{o} \text{ and } \underset{\sim}{w}_o \text{ represents } k \times 1 \text{ constant vector} \right\}$$

Define the transformation $y_i = \underset{\sim}{w}' \underset{\sim}{x}_i, \;\; i = 1, \ldots, n_1,$

and $y_{oi} = \underset{\sim}{w}' \underset{\sim}{x}_{oi}, \;\; i = n_1 + 1, \ldots, n$

where $\underset{\sim}{w} \varepsilon H$.

Then

$$y_i \sim N_1(o, \sigma_o^2), \; i = 1, \ldots, n_1 \tag{3.1}$$

and $$y_{oi} \sim N_1(\underset{\sim}{w}' \underset{\sim}{r}, \sigma_o^2), \; i = n_1 + 1, \ldots, n_1$$

where $$\sigma_o^2 = \underset{\sim}{w}'(\Gamma + \sigma^2 I)\underset{\sim}{w}.$$

From (3.1), it is clear that H_{0k} is equivalent to $H_{0k}^* : \mu_i = 0$ and H_{1k} is equivalent to $H_{1k}^* : \mu_i \neq 0$, $i = 1,2,\ldots,n$, where $\mu_i = E(u_i)$, $i = 1, \ldots, n$ and $u_i = y_i$, for $i=1,\ldots,n_1$, and $u_i = y_{0i}$, for $i = n_1+1,\ldots,n$.

Standard test is available in the literature, if we want to test H_{0k}^* vs. H_{1k}^* based on the observations $y_i (i = 1,\ldots,n_1)$ and $y_{0i}(i = n_1+1,\ldots,n)$.

The rejection region for testing H_{0k}^* vs. H_{1k}^* is

$\left|\frac{\sqrt{n}\bar{y}}{S_y}\right| > c$, where c represents a constant

i.e. $\left|\frac{\sqrt{n}\underset{\sim}{w}'\bar{\underset{\sim}{x}}}{\sqrt{\underset{\sim}{w}' s \underset{\sim}{w}}}\right| > c$ for at least one $\underset{\sim}{w} \varepsilon H$

i.e. $n(\bar{\underset{\sim}{x}}_1{}' s_{11}^{-1} \bar{\underset{\sim}{x}}_1) > c^*$ where c^* is some other constant,

$$s_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

$$\underset{p\times 1}{\bar{\underset{\sim}{x}}} = \frac{1}{n}\left[\sum_{i=1}^{n_1}\underset{\sim}{x}_i + \sum_{i=n_1+1}^{n}\underset{\sim}{x}_{0i}\right] = \begin{pmatrix} \bar{x}_1 \\ \bar{x}_2 \\ (\text{p-k}) \times 1 \end{pmatrix}$$

and $$s = \frac{1}{n}\left[\sum_{i=1}^{n_1}(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_i - \bar{\underset{\sim}{x}})' + \sum_{i=n_1+1}^{n_1}(\underset{\sim}{x}_{0i} - \bar{\underset{\sim}{x}})(\underset{\sim}{x}_{0i} - \bar{\underset{\sim}{x}})'\right]$$

$$= \begin{pmatrix} S_{11} & S_{12} \\ k \times k & \\ S_{21} & S_{22} \end{pmatrix}.$$

$$(p\text{-}k) \times (p\text{-}k)$$

It is clear that under $H_{0k}, \bar{\underset{\sim}{x}}_1{}' s_{11}^{-1} \bar{\underset{\sim}{x}}_1 \sim$ Hotelling's T^2. So, we can sequentially test the null hypothesis for different values of $k = 0, 1, 2, \ldots, (p-1)$ and we get a solution. But the drawback of this method is that we have ignored the structure $\Gamma + \sigma^2 I$ of the dispersion matrix, i.e. the original problem implies the problem reduced by union-intersection principle, but the converse is not true. Hence, rejection of null hypothesis for the reduced problem implies the rejection of null hypothesis for the original problem. But, when null hypothesis is accepted for the reduced problem, we can not say that for the original problem.

Another drawback is that we are testing sequentially for different values of $k = (0, 1, \ldots, p-1)$. If there is rejection of null hypothesis for $k = 0$, we go to test for $k = 1$. In this case, the critical level α will be as follows:

$$\alpha = p\{\text{Reject the null hypothesis for } k = 1 / null$$

$$\text{hypothesis is rejected for } k = 0\} \tag{3.2}$$

Krishnaiah [6] discussed some drawback about this sequential procedure. The probability α given by (3.2) is very complicated to evaluate.

In order to avoid the above difficulties, we can formulate the problem within the framework of model selection procedure. No critical value is needed in this approach.

The procedure involves certain information theoretic criteria. This model selection method has the optimum property that the estimator is strongly consistent. More discussion is given in section 4.

4. Soultion by information theoretic criteria

Akaike's Information criterion (AIC) [1] and Schwartz-Rissanen [9],[10] minimum distance length (MDL) criterion were used in Wax and Kailath [11] in order to estimate the number of signals in the presence of noise. Some modified information criterion used by Zhao, Krishnaiah and Bai [12], [13] gives consistent estimate of integer parameter of a model.

Chen [5], Chen [4], Kundu [7] etc. developed procedures for estimating the number of signals .

Brief discussion about AIC procedure and procedure by Zhao, Krishnaiah and Bai [12], [13] are given as follows:

Suppose k is an unknown integer parameter of a model. Let the range for k be $0 \le k \le m$. For example, k may be the number of variables important for prediction in a regression equation. In our case, k is the number of faulty components in outliers. Using AIC criterion we can say that $\hat{k}$ is an estimate of k where,

$$AIC(\hat{k}) = \min\{AIC(0), AIC(1), \ldots, AIC(m)\},$$

and $$AIC(q) = -2\log 1_q + 2\gamma(q, m)$$

1_q = likelihood of observations under the null hypothesis $H_0 : k = q$

$\gamma(q, m) =$ number of free parameters to be estimated under H_0.

According to Zhao, Krishnaiah and Bai's criterion, $\hat{k}$ is an estimate of k where,

$$I(\overleftarrow{k}, C_n) = \min\{I(0, C_n), I(1, C_n), \ldots, I(m, C_n)\}$$

and
$$I(q, C_n) = -\log L_q + C_n \gamma(q, m)$$

$L_q =$ statistic obtained from likelihood ratio test under H_0

$\gamma(q, m) =$ number of free parameters to be estimated under H_0 and C_n satisfies

(i) $\lim_{n\to\infty} \frac{C_n}{n} = 0$

(ii) $\lim_{n\to\infty} \frac{C_n}{\log\log n} = \infty.$

In our problem, we will modify information criterion as follows:

Estimate k by $\hat{k}$, where $\hat{k}$ is such that

$$I^*(\hat{k}, C_n) = \min_{0\le k\le p-1} I^*(k, C_n) \tag{4.1}$$

and
$$I^*(q, C_n) = -nT_{n,q} + C_n \gamma(q, m)$$

$T_{n,q} =$ some test statistic obtained from testing $H_0 : k = q$ vs. $H_1 :$ not H_0

$\gamma(q, m)$ = defined before and C_n satisfies:

(i) $\lim_{n\to\infty} \frac{C_n}{n} = 0$

(ii) $\lim_{n\to\infty} \frac{C_n}{\sqrt{n\log\log n}} = \infty. \qquad (4.2)$

Let us assume that σ^2 is unknown. Under the setup (2.2), we are interested to test:

$$H_{0k}: \; m = k \text{ vs. } H_{1k}: \; m > k, \; k = 0, 1, \ldots, (p-1).$$

It is easy to show that the logarithm of the modified likelihood ratio test statistic for testing H_{0k} vs. H_{1k} is:

$$\log\ L_k^* = -\frac{n}{2}\left[\sum_{i=1}^{q}\log\delta_{i,k} - \sum_{i=1}^{q}\log\delta_i + (p-q)\log\left(\frac{\sum_{i=q+1}^{p}\delta_{i,k}}{p-q}\right) - (p-q)\log\left(\frac{\sum_{i=q+1}^{p}\delta_i}{p-q}\right)\right] \tag{4.3}$$

where $\delta_{1,k} \geq \ldots \geq \delta_{p,k}$ are the non-zero eigen values of

$$\frac{1}{n}\left\{S_1 + \sum_{i=n_1+1}^{n}\begin{pmatrix}\underset{\sim}{x}_{01i} \\ \underset{\sim}{x}_{02i} - \underset{\sim}{\bar{x}}_{02}\end{pmatrix}\begin{pmatrix}\underset{\sim}{x}_{01i} \\ \underset{\sim}{x}_{02i} - \underset{\sim}{\bar{x}}_{02}\end{pmatrix}'\right\}$$

$$S_1 = \sum_{i=1}^{n_1} \underset{\sim}{x}_i \underset{\sim}{x}_i'$$

$$\underset{p\times 1}{\underset{\sim}{x}_{0i}} = \begin{pmatrix}\underset{\sim}{x}_{01i} \\ \underset{\sim}{x}_{02i}\end{pmatrix}$$

$$(p\text{-}k) \times 1$$

$$\underset{p\times 1}{\underset{\sim}{\bar{x}}_0} = \frac{1}{n_2}\sum_{i=n_2+1}^{n}\underset{\sim}{x}_{0i} = \begin{pmatrix}\underset{\sim}{\bar{x}}_{01} \\ \underset{\sim}{\bar{x}}_{02}\end{pmatrix}$$

$$(p\text{-}k) \times 1$$

and $\delta_1 \geq \ldots \geq \delta_p$ are the ordered eigen values of $S = \frac{1}{n}\sum_{i=1}^{n}\underset{\sim}{x}_i \underset{\sim}{x}_i'$.

The reason we call (4.3) as the logarithm of modified likelihood ratio test statistic is that the modified estimate of $\underset{\sim}{a}$ under a setup (2.2) and H_{0k} is taken as $\underset{\sim}{\bar{x}}_{02}$ and this estimate is not exact maximum likelihood estimate of $\underset{\sim}{a}$. But asymptotically both the estimate (exact and modified) of $\underset{\sim}{a}$ converage stochastically to the same parameter.

In our problem, we take $T_{n,k}$ in formula (4.1) as $\frac{1}{n}\log L_k^*$. The formula $I^*(k, C_n)$ in (4.1) now reduces to

$$I^*(k, C_n) = -\log L_k^* + C_n\left(p - k + \frac{q(2p-q+1)}{2} + 1\right) \tag{4.4}$$

Now, we will show that the estimator $\hat{k}$ of k in (4.1) is strongly consistent estimator of k_0, where k_0 is the true value of k. We need the following lemmas in order to establish strong consistency of $\hat{k}$.

Lemma 4.1: *Let* $\frac{n_2}{n} \to \delta(>0)$ *as* $n \to \infty$. *Then for* $k > k_0$, $\delta_{i,k}$, $\hat{\sigma}^2_{(k)}$ *are strongly consistent estimator of* λ_i, σ^2 *respectively for* $i = 1,\ldots,q$, *where* $\hat{\sigma}^2_{(k)} = \sum_{i=q+1}^{p} \frac{\delta_{i,k}}{p-q}$ *and* $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_q > \sigma^2$ *are the ordered eigen values of* $(\Gamma + \sigma^2 I)$.

Lemma 4.2: *Let* $\frac{n_2}{n} \to \delta(>0)$ *as* $n \to \infty$. *Then for* $k > k_0$, $\lim_{n\to\infty}(\delta_{i,k} - \lambda_i) \geq 0$ *a.s. for* $i = 1,\ldots,q$ *and* $\lim_{n\to\infty}(\hat{\sigma}^2_{(k)} - \sigma^2) \geq 0$ *a.s. with inequality strict for one* $k(> k_0)$.

Lemma 4.3: *Let* $0 < \alpha < \frac{n_1}{n} < \beta < 1$ *for all n. Then for* $k \leq k_0$, $\delta_{i,k} - \lambda_i = 0\left(\sqrt{\frac{\log\log n}{n}}\right)$ *a.s. for large n and for* $i = 1,\ldots,q$.

Theorem 4.1: *Let*

(i) $0 < \alpha < \frac{n_1}{n} < \beta < 1$ *for all* n

(ii) $\frac{n_2}{n} \to \delta(>0)$ *as* $n \to \infty$.

Then the estimate $\hat{k}$ is a strongly consistent estimator of k_0, where $\hat{k}$ is obtained from (4.1).

Proof: It is enough to show that the difference

$$I^*(k, C_n) - I^*(k_0, C_n) > 0 \text{ a.s. as } n \to \infty .$$

Now, the difference:

$$I^*(k,C_n) - I^*(k_0,C_n) = \frac{n}{2}\left[\sum_{i=1}^{q}(\log\delta_{i,k} - \log\lambda_i) + (p-q)\left\{\log\left(\frac{\sum_{i=q+1}^{p}\delta_{i,k}}{p-q}\right) - \log\sigma^2\right\}\right.$$

$$\left.-\sum_{i=1}^{q}(\log\delta_{i,k_0} - \log\lambda_i) - (p-q)\left\{\log\left(\frac{\sum_{i=q+1}^{p}\delta_{i,k_0}}{p-q}\right) - \log\sigma^2\right\}\right] + C_n(k_0 - k) \qquad (4.5)$$

Case 1: $k < k_0$

Using lemma 4.3 and equation (4.5) we get

$$I^*(k,C_n) - I^*(k_0,C_n) = 0\left(\sqrt{n\log\log n}\right) + C_n(k_0 - k) \text{ a.s.} \qquad (4.6)$$

Hence, using the condition (ii) of (4.2) on C_n in (4.6), we have

$$I(k,C_n) - I(k_0,C_n) > 0 \textit{ a.s for large n}. \qquad (4.7)$$

Case 2: $k < k_0$

By lemma 4.2, we see that

$$\sum_{i=1}^{q}(\log\delta_{i,k} - \log\lambda_i) + (p-q)\left\{\log\left(\frac{\sum_{i=q+1}^{p}\delta_{i,k}}{p-q}\right) - \log\sigma^2\right\} > 0 \text{ a.s. as } n \to \infty \quad (4.8)$$

Using lemma 4.1, we have

$$\sum_{i=1}^{q}(\log\delta_{i,k_0} - \log\lambda_i) + (p-q)\left\{\log\left(\frac{\sum_{i=q+1}^{p}\delta_{i,k_0}}{p-q}\right) - \log\sigma^2\right\} > 0 \text{ a.s. as } n \to \infty \quad (4.9)$$

By (4.5), (4.8), (4.9) and condition (i) of (4.2) on C_n, we have

$$I^*(k, C_n) - I^*(k_0, C_n) > 0 \text{ a.s. as for large } n. \quad (4.10)$$

From (4.7) and (4.10), we have the result.

Remark: For estimating *k*, in practical purposes, we can choose $C_n = \sqrt{n \log n}$ which satisfies conditions (i) and (ii) of (4.2) on C_n. Without loss of generality, we can assume $\sigma^2 = 1$ when σ^2 is known. In this case, using the approach of Zhao, Krishnaiah and Bai [12], one can find out the estimator $\hat{k}$ and prove strong consistency of $\hat{k}$.

5. Detection of the number of defective components in the dispersion-slippage outliers

Model for dispersion-slippage outlier(s) is discussed in Bhandary [2]. In our case, outlier model is as follows:

$$\underset{\sim}{x}_i \sim N_p(\underset{\sim}{0}, \Gamma + \sigma^2 I), \quad i = 1, \ldots, n_1, \quad (5.1)$$

and $$\underset{\sim}{x}_{0i} \sim N_p\left(0, \Gamma + \begin{pmatrix} \sigma_1^2 & & & & & & & \\ & \cdot & & & & 0 & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & \sigma_m^2 & & & \\ & & & & & \sigma^2 & & \\ & 0 & & & & & \cdot & \\ & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & & \sigma^2 \end{pmatrix}\right), \quad i = n_1 + 1, \ldots, n$$

Where $\sigma_j^2 > \sigma^2$, $j = 1, \ldots, m$ and $m = 1, 2, \ldots, p - q$ and $n = n_1 + n_2$. Here, we assumed without loss of generality that the first m components in the outliers are defectives.

Here, we have assumed that n_2 observations $\underset{\sim}{x}_{0i} (i = n_1 + 1, \ldots, n)$ are outliers having dispersion-slippage structure and the first m components of the outliers are defectives. Our problem is to estimate m.

Likelihood ratio test for this problem is very difficult to derive and therefore we cannot use Zhao, Krishnaiah and Bai's [12], [13] method. We will use modified information criterion proposed in (4.1).

In our problem, using Roy's Union-intersection principle, we can find our test statistic $T_{n,k}$ as follows:

Let

$G = \{ \underset{p \times 1}{\underset{\sim}{u}} : \underset{\sim}{u}' \underset{\sim}{u} = 1$ and the first $k(< p)$ components of $\underset{\sim}{u}$ are 0 and

the remaining (p-k) components are arbitrary}.

Define the transformation $y_i = \underset{\sim}{u}' \underset{\sim}{x}_i$, $i = 1, \ldots, n_1$ and

$$\underset{\sim}{y}_{0i} = \underset{\sim}{u}' \underset{\sim}{x}_{0i}, \ \ i = n_1 + 1, \ldots, n \text{ where } \underset{\sim}{u} \, \varepsilon \, G.$$

Then $y_i \sim N_1(0, \sigma_1^{*2})$, $i = 1, \ldots, n_1$

and $y_{0i} \sim N_1(0, \sigma_2^{*2})$, $i = n_1 + 1, \ldots, n$

where $\sigma_1^{*2} = \underset{\sim}{u}' (\Gamma + \sigma^2 I) \underset{\sim}{u}$

and
$$\sigma_2^{*2} = \underset{\sim}{u}' \Gamma \underset{\sim}{u} + \sum_{i=1}^{m} u_i^2 \sigma_i^2 + \sigma^2 \sum_{i=m+1}^{p} u_i^2 \tag{5.2}$$

$$\underset{\sim}{u} = \begin{pmatrix} u_1 \\ u_2 \\ \cdot \\ \cdot \\ \cdot \\ u_p \end{pmatrix}.$$

From (5.2), it is obvious that H_{0k} in Section 4 is equivalent to

H_{0k}^*: $\sigma_1^{*2} = \sigma_2^{*2}$ and H_{1k} in Section 4 is equivalent to

H_{1k}^*: $\sigma_1^{*2} > \sigma_2^{*2}$.

Test for H_{0k}^* vs. H_{1k}^* based on the observations $y_i (i = 1, \ldots, n_1)$ and $y_{0i} (i = n_1 + 1, \ldots, n)$ is a standard test in the literature. The rejection region for testing H_{0k}^* vs. H_{1k}^* is

$$\frac{\frac{1}{n_1}\sum_{i=1}^{n_1} y_i^2}{\frac{1}{n_2}\sum_{i=n_1+1}^{n} y_{0i}^2} > c, \text{ where c is some constant}$$

i.e. $\frac{\underset{\sim}{u}' s_1 \underset{\sim}{u}}{\underset{\sim}{u}' s_2 \underset{\sim}{u}} > c$ for at least one $\underset{\sim}{u} \varepsilon G$

i.e. $\hat{\lambda}_{\max}(s_{1(22)} s_{2(22)}^{-1}) > c$, provided $s_{2(22)}$ is non-singular

$$\text{where } s_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} \underset{\sim}{x}_i \underset{\sim}{x}_i' = \begin{pmatrix} s_{1(11)} & s_{1(12)} \\ k \times k & \\ s_{1(21)} & s_{1(22)} \end{pmatrix}$$

$$(p-k)\times(p-k)$$

$$s_2 = \frac{1}{n_2}\sum_{i=n_1+1}^{n_1} \underset{\sim}{x}_{0i} \underset{\sim}{x}_{0i}' = \begin{pmatrix} s_{2(11)} & s_{2(12)} \\ k \times k & \\ s_{2(21)} & s_{2(22)} \end{pmatrix}$$

$$(p-k)\times(p-k)$$

and $\hat{\lambda}_{\max}(\Lambda)$ = largest eigen value of Λ.

One important thing here is that we have ignored the structure $\Gamma + \sigma^2 I$ of the dispersion matrix. So, the reduced problem can be obtained from the original problem by union-intersection principle, but the converse is not true. Hence, we can use the test statistic $\hat{\lambda}_{\max}$ $(s_{1(22)}\ s_{2(22)}^{-1})$ for rejection purpose only and not for acceptance purpose.

If we take care of the structure $\Gamma + \sigma^2 I$ of the dispersion matrix, where $\underset{p \times p}{\Gamma}$ is non-negative definite of rank $q(< p)$ then we can propose some intuitive test statistic as $\hat{\lambda}_{q+1}$ $(s_{1(22)}\ s_{2(22)}^{-1})$ where $\hat{\lambda}_1 \geq ... \geq \hat{\lambda}_q \geq \hat{\lambda}_{q+1} \geq ... \geq \hat{\lambda}_{p-k}$ are the ordered eigen values of $(s_{1(22)}\ s_{2(22)}^{-1})$. So, in this problem, we take $T_{n,k}$ in formula (4.1) as $\hat{\lambda}_{q+1}$ $(s_{1(22)}\ s_{2(22)}^{-1})$, where $k < p-q$.

Thus, the formula $I^*(k, C_n)$ in (4.1) now reduces to

$$I^*(k, C_n) = -n\hat{\lambda}_{q+1}(s_{1(22)} s_{2(22)}^{-1}) + C_n\left(k + \frac{q(2p-q+1)}{2} + 1\right) \tag{5.3}$$

Now, we will show that the estimator $\hat{k}$ of k obtained by using (5.3) in (4.1) is strongly consistent estimator of k_0, where k_0 is the true value of k. We need the following lemmas in order to establish the strong consistency of $\hat{k}$:

Lemma 5.1: *Let* $\hat{\lambda}_1 \geq \ldots \geq \hat{\lambda}_{p-k}$ *be the ordered eigen values of* $(s_{1(22)} s_{2(22)}^{-1})$. *Then for* $k \geq k_0$, $\hat{\lambda}_1 \to 1$ *a.s. as* $n \to \infty$, $i = 1, 2, \ldots, p-k$.

Lemma 5.2: *Let* $0 < \alpha < \frac{n_1}{n} < \beta < 1$ *for all n. Then for* $k \geq k_0$, $(\hat{\lambda}_1 - 1) = 0\left(\sqrt{\frac{\log\log n}{n}}\right)$ *a.s. for large n,* $i = 1, 2, \ldots, p-k$.

Lemma 5.3: *For* $k < k_0$, $\lim_{n\to\infty} (\hat{\lambda}_i - 1) < 0$ *a.s. for* $i = 1, \ldots, p-k$.

Theorem 5.1: *Let* $0 < \alpha < \frac{n_1}{n} < \beta < 1$ *for all n. Then* $\hat{k}$ *obtained by using (5.3) in (4.1) is a strongly consistent estimator of* k_0.

Proof: The idea of proof of Theorem 5.1 will be same as the proof of Theorem 4.1.

The difference $I^*(k, C_n) - I^*(k_0, C_n)$

$$= n\hat{\lambda}_{q+1} \underset{(p-k_0)\times(p-k_0)}{(s_{1(22)} s_{2(22)}^{-1})} - n\hat{\lambda}_{q+1} \underset{(p-k)\times(p-k)}{(s_{1(22)} s_{2(22)}^{-1})} + C_n(k - k_0)$$

$$= n\left(\hat{\lambda}_{q+1}(s_{1(22)} s_{2(22)}^{-1} - 1)\right) - n\left(\hat{\lambda}_{q+1}(s_{1(22)} s_{2(22)}^{-1} - 1)\right) - C_n(k - k_0) \tag{5.4}$$

Case 1: $k > k_0$

Using lemma 5.2 and the condition (ii) of (4.2) on C_n, we get

$$I^*(k, C_n) - I^*(k_0, C_n) > 0 \text{ a.s. an } n \to \infty. \tag{5.5}$$

Case 2: $k \leq k_0$

Using lemma 5.1 we have

$$\hat{\lambda}_{q+1} \underset{(p-k_0)\times(p-k_0)}{(s_{1(22)} s_{2(22)}^{-1})} - 1 \to 0 \text{ a.s. as } n \to \infty \tag{5.6}$$

By lemma 5.3, (5.6) and the condition (i) of (4.2) on C_n, we get

$$I^*(k, C_n) - I^*(k_0, C_n) > 0 \text{ a.s. an } n \to \infty \tag{5.7}$$

From (5.5) and (5.7) we have the result.

6. Simulation results

$$p = 3,\ q = 3,\ \sigma^2 = 1.0,\ n_1 = 40,\ n_2 = 30$$

We generate observations from $N_3\left(\begin{pmatrix}0\\0\\0\end{pmatrix}, \Gamma+\sigma^2 I_3\right)$ and from $N_3\left(\begin{pmatrix}0\\3.183\\4.274\end{pmatrix}, \Gamma+\sigma^2 I_3\right)$ by IMSL subroutine and calculate S_1 and S_2 and eventually calculate $I^*(k, C_n)$ for $C_n = \sqrt{n\log n}$ by using formula (4.3) and (4.4) respectively and then present result in the following table:

Table 1

k	$I^*(k,C_n)$
0	70.0648
1	61.8046
2	62.1879

In table 1 above, since $\min\limits_{0\le k\le 2} I^*(k, C_n) = 61.8046$ and the minimum occurs for $k = 1$, we get $\hat{k} = 1$ and the population true value of $k = k_0 = 1$, because we have generated observations from $N_3\left(\begin{pmatrix}0\\3.183\\4.274\end{pmatrix}, \Gamma+\sigma^2 I_3\right)$.

APPENDIX

Proof of Lemma 4.1: Define the transformation

$$\underset{p\times 1}{\underset{\sim}{y}_i} = \begin{pmatrix} \underset{\sim}{x}_{01i} \\ \underset{(p-k)\times 1}{\underset{\sim}{x}_{02i}} - \underset{\sim}{a} \end{pmatrix}, i = n_1 + 1, \ldots., n. \tag{*}$$

Then $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_{n_1}, \underset{\sim}{y}_{n_1+1}, \ldots, \underset{\sim}{y}_n$ are *i.i.d.* $\sim N_p(\underset{\sim}{0}, \Gamma + \sigma^2 I)$. Under the transformation (*), we can write the following:

$$\frac{1}{n}\left(S_1 + \sum_{i=n_1+1}^{n} \begin{pmatrix} \underset{\sim}{x}_{01i} \\ \underset{\sim}{x}_{02\mathrm{i}} - \underset{\sim}{\overline{x}}_{02} \end{pmatrix}\begin{pmatrix} \underset{\sim}{x}_{01i} \\ \underset{\sim}{x}_{02\mathrm{i}} - \underset{\sim}{\overline{x}}_{02} \end{pmatrix}'\right)$$

$$= \frac{1}{n}\left(S_1 + \sum_{i=n_1+1}^{n} \underset{\sim}{y}_i \underset{\sim}{y}_i' - n_2 \underset{\sim}{\overline{y}}\,\underset{\sim}{\overline{y}}' + n_2 \begin{pmatrix} \underset{\sim}{\overline{x}}_{01} \\ \underset{\sim}{0} \end{pmatrix}\begin{pmatrix} \underset{\sim}{\overline{x}}_{01} \\ \underset{\sim}{0} \end{pmatrix}'\right) \tag{**}$$

By strong law of large numbers, we have

$$=\frac{1}{n}\left(S_1+\sum_{i=n_1+1}^{n}\underset{\sim}{y}_i\,\underset{\sim}{y}_i{}'\right)\to(\Gamma+\sigma^2 I)\text{ a.s.}$$

$$\underset{\underset{p\times 1}{\sim}}{\overline{y}}\to\underset{\sim}{0}\text{ a.s.}$$

and
$$\begin{pmatrix}\underset{\sim}{\overline{x}}_{01}\\ \text{k}\times 1\\ \underset{\sim}{0}\end{pmatrix}\to\underset{\underset{p\times 1}{\sim}}{0}\text{ a.s. for } k\le k_0. \qquad (***)$$

Hence, from (**) and (***), we have

$$\frac{1}{n}\left(S_1+\sum_{i=n_1+1}^{n}\begin{pmatrix}\underset{\sim}{x}_{01i}\\ \underset{\sim}{x}_{02\text{i}}-\underset{\sim}{\overline{x}}_{02}\end{pmatrix}\begin{pmatrix}\underset{\sim}{x}_{01i}\\ \underset{\sim}{x}_{02\text{i}}-\underset{\sim}{\overline{x}}_{02}\end{pmatrix}'\right)\to(\Gamma+\sigma^2 I)\text{ a.s. for } k\le k_0.$$

Hence, using lemma 3.2 of Zhao, Krishnaiah and Bai [12], we get lemma 4.1.

Proof of Lemma 4.2: Using the assumption in lemma 4.2 and strong law of large numbers, we get

$$\frac{n_2}{n}\begin{pmatrix}\underset{\underset{k\times 1}{\sim}}{\overline{x}_{01}}\\ \underset{\sim}{0}\end{pmatrix}\begin{pmatrix}\underset{\sim}{\overline{x}}_{01}\\ \underset{\sim}{0}\end{pmatrix}'\to\begin{pmatrix}\underset{\underset{k\times 1}{\sim}}{0}\\ \underset{\underset{(k-k_0)\times 1}{\sim}}{a^*}\\ \underset{\underset{(p-k)\times 1}{\sim}}{0}\end{pmatrix}\begin{pmatrix}\underset{\sim}{0}\\ \underset{\sim}{a}^*\\ \underset{\sim}{0}\end{pmatrix}'\text{ a.s. for } k\le k_0$$

and $\underset{\sim}{a}^*(\ne 0)\varepsilon\,\Re^{(k-k_0)}$.

Using the same idea as in lemma 4.1, we get

$$\frac{1}{n}\left(S_1+\sum_{i=n_1+1}^{n}\begin{pmatrix}\underset{\sim}{x}_{01i}\\ \underset{\sim}{x}_{02\text{i}}-\underset{\sim}{\overline{x}}_{02}\end{pmatrix}\begin{pmatrix}\underset{\sim}{x}_{01i}\\ \underset{\sim}{x}_{02\text{i}}-\underset{\sim}{\overline{x}}_{02}\end{pmatrix}'\right)\to(\Gamma+\sigma^2 I)+\delta\begin{pmatrix}\underset{\sim}{0}\\ \underset{\sim}{a}^*\\ \underset{\sim}{0}\end{pmatrix}\begin{pmatrix}\underset{\sim}{0}\\ \underset{\sim}{a}^*\\ \underset{\sim}{0}\end{pmatrix}' \qquad (*)$$

The matrix $\delta \begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{a^*} \\ \underset{\sim}{0} \end{pmatrix} \begin{pmatrix} \underset{\sim}{0} \\ \underset{\sim}{a^*} \\ \underset{\sim}{0} \end{pmatrix}'$ has one eigen value $\delta \underset{\sim}{a}^{*'} \underset{\sim}{a}^*$, which is striclty positive since $\delta > 0$ and the other eigen values are all zeros. (**)

Hence, by lemma 3.2 of Zhao, Krishnaiah and Bai [12], and (*) and (**), we get lemma 4.2.

Proof of Lemma 4.3: By law of iterated logarithm, and by the assumption in lemma 4.3, we have

$$\frac{1}{n}\left(\sum_{i=1}^{n_1} \underset{\sim}{x_i}\underset{\sim}{x_i}' + \sum_{i=n_1+1}^{n} \underset{\sim}{y_i}\underset{\sim}{y_i}'\right) - (\Gamma + \sigma^2 I) = 0\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.}$$

$$\underset{\sim}{\overline{y}} = 0\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s.} \tag{*}$$

and
$$\frac{n_2}{n}\begin{pmatrix} \underset{\substack{\sim\\k\times 1}}{\overline{x}_{01}} \\ \underset{\sim}{0} \end{pmatrix}\begin{pmatrix} \underset{\sim}{\overline{x}_{01}} \\ \underset{\sim}{0} \end{pmatrix}' = 0\left(\frac{\log\log n}{n}\right) \text{ a.s. for } k \le k_0.$$

Using the same idea as in lemma 4.1 and (*), we get

$$\frac{1}{n}\left(S_1 + \sum_{i=n_1+1}^{n} \begin{pmatrix} \underset{\sim}{x_{01i}} \\ \underset{\sim}{x_{02i}} - \underset{\sim}{\overline{x}_{02}} \end{pmatrix}\begin{pmatrix} \underset{\sim}{x_{01i}} \\ \underset{\sim}{x_{02i}} - \underset{\sim}{\overline{x}_{02}} \end{pmatrix}'\right) - (\Gamma + \sigma^2 I)$$

$$= 0\left(\sqrt{\frac{\log\log n}{n}}\right) \text{ a.s. for } k \le k_0.$$

Hence, applying lemma 3.2 of Zhao, Krishnaiah and Bai [12], we get lemma 4.4.

Proof of Lemma 5.1: Lemma 5.1 follows easily by strong law of large numbers and a partition of

$$\Gamma + \sigma^2 I_p = \begin{pmatrix} \Gamma_{11} + \sigma^2 I_k & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} + \sigma^2 I_{p-k} \end{pmatrix}$$

$$\text{and } \Gamma + \begin{pmatrix} \sigma_1^2 & & & & & & & \\ & \cdot & & & 0 & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & \sigma_k^2 & & & \\ & & & & & \sigma^2 & & \\ & 0 & & & & & \cdot & \\ & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & & \sigma^2 \end{pmatrix} = \begin{pmatrix} \Gamma_{11} + \begin{pmatrix} \sigma_1^2 & & & \\ & \cdot & & \\ & & \cdot & \\ & & & \cdot \\ & & & & \sigma_k^2 \end{pmatrix} & \Gamma_{12} \\ \Gamma_{21} & \Gamma_{22} + \sigma^2 I_{p-k} \end{pmatrix}$$

Proof of Lemma 5.2: Lemma 5.2 follows by law of iterated logarithm and a partition as in lemma 5.1 and lemma 3.2 of Zhao, Krishnaiah and Bai [12].

Proof of Lemma 5.3: By strong law of large numbers, we have

$$s_{1(22)} s_{2(22)}^{-1} \to (\Gamma_{22} + \sigma^2 I_{p-k}) \left[\Gamma_{22} + \begin{pmatrix} \sigma_1^2 & & & & & & & \\ & \cdot & & & 0 & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & \sigma_k^2 & & & \\ & & & & & \sigma^2 & & \\ & 0 & & & & & \cdot & \\ & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & & \sigma^2 \end{pmatrix} \right]^{-1} \text{ a.s. for } k < k_0. \quad (*)$$

It is easy to show that

$$(\Gamma_{22} + \sigma^2 I_{p-k}) \left[\Gamma_{22} + \begin{pmatrix} \sigma_1^2 & & & & & & & \\ & \cdot & & & 0 & & & \\ & & \cdot & & & & & \\ & & & \cdot & & & & \\ & & & & \sigma_k^2 & & & \\ & & & & & \sigma^2 & & \\ & 0 & & & & & \cdot & \\ & & & & & & & \cdot \\ & & & & & & & & \cdot \\ & & & & & & & & & \sigma^2 \end{pmatrix} \right]^{-1} < I_{p-k}. \qquad (**)$$

where "<" is in the sense of negative definite.

The above result is true because $\sigma_j^2 > \sigma^2, \ j = 1, \ldots, k_0$.

Hence, from (*) and (**), we get

$\lim\limits_{n\to\infty} (\hat{\lambda}_i - 1) < 0$ a.s. for $k < k_0$ and $i = 1, \ldots, p-k$.

References

[1] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in Proc. Second Int. Symp. Information Theory, *Suppl. to Problems of Control and Information Theory,* pp. 267–281 (1972).

[2] M. Bhandary, "Detection of outliers in signal processing and detection on number of signals in presence of outliers," *Commun. Statist.—Theory Methods,* vol. 18, no. 6, pp. 2233–2261 (1989).

[3] M. Bhandary, "Detection of the number of outliers present in a data set using an information theoretic criterion," *Commun. Statist.—Theory Methods,* vol. 21, no. 11, pp. 3263–3274 (1992).

[4] P. Chen, "A selection procedure for estimating the number of signal components," *J. Stat. Plann. Inference,* vol. 105, pp. 299–301 (2002).

[5] P. Chen, M. C. Wicks, and R. S. Adve, "Development of a statistical procedure for detecting the number of signals in a radar measurement," *Proc. IEE Radar, Sonar Navig.,* vol. 148, no. 4, pp. 219–226 (2001).

[6] P. R. Krishnaiah, "Selection of variables under univariate regression models," in Handbook of Statistics 2, P. R. Krishnaiah, Ed. North-Holland Publishing Company, pp. 805–820 (1982).

[7] D. Kundu, "Estimating the number of signals in the presence of white noise," *J. Stat. Plann. Inference,* vol. 90, pp. 57–68 (2000).

[8] S. N. Roy, "On a heuristic method of test construction and its use in multivariate analysis," *Ann. Math. Statist.,* vol. 24, pp. 220–238 (1953).

[9] J. Rissanen, "Modeling by shortest data description," *Automatica,* vol. 14, pp. 465–471 (1978).

[10] G. Schwartz, "Estimating the dimension of a model," *Ann. Statist.,* vol. 6, pp. 461–464 (1978).

[11] M. Wax and T. Kailath, "Detection of the number of signals by information theoretic criteria," *IEEE Trans. Acoust., Speech, Signal Process.,* vol. ASSP-33, pp. 387–392 (1984).

[12] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of number of signals in presence of white noise," *J. Multivar. Anal.,* vol. 20, pp. 1–25 (1986).

[13] L. C. Zhao, P. R. Krishnaiah, and Z. D. Bai, "On detection of number of signals when the noise covariance matrix is arbitrary," *J. Multivar. Anal.*, vol. 20, pp. 26–49 (1986).

Received April, 2023