INLIER PRONESS IN NORMAL DISTRIBUTION

K. Muralidharan¹ and Arti Khabia²

Department of Statistics, Faculty of Science
The Maharajah Sayajirao University of Baroda, Vadodara 390 002, India.
¹Email: lmv_murali@yahoo.com
²Email: artimkhabia@yahoo.com

ABSTRACT

Inliers in a data set are subset of observations not necessarily all zeroes, which appears to be inconsistent with the remaining data set. They are either the resultant of instantaneous or early failures usually encountered in life testing, financial, clinical trial and many other studies. We study the estimation of inliers in Normal distribution. The masking effect problem for correctly identifying the inliers is also discussed. An illustration and a real life example is presented with detailed discussions.

Key Words: inliers; optimal estimating equations; mixture distribution; MLEs; asymptotic distribution; early failures; Schwarz’s Information criterion; modified likelihood test.

1. Introduction

The normal distribution is a very important statistical model occurring in many natural phenomena, such as measurement of height, blood pressure, lengths of objects produced by machines, etc. Usually normal distributions are symmetrical with a single central peak at the mean (average) of the data. But many times we may get normal distribution as mixture of two groups. For example the life time of an electronic item will have two sets of observations, where one set of data may have zero or small life times due to instantaneous or early failures (together called inliers) and the other set contains positive life times called target life times. This may create two symmetrical curved graphs, where the mean of inliers group is much less than the mean of target group. Such failures usually discard the assumption of a unimodal distribution and hence the usual method of modeling and inference procedures may not be accurate in practice. Usually, these situations are handled by modifying commonly used parametric models suitably incorporating inconsistent observations. The modified model is then a non-standard distribution and we call such models as inliers prone models.

Normal mixture distributions are arguably the most important mixture models, and also the most technically challenging. The likelihood function of the normal mixture model is unbounded based on a set of random samples, unless an artificial bound is placed on its component variance parameter. There has been extensive research on finite normal mixture models, but much of it addresses merely consistency of the point estimation or useful practical procedures, and many results require undesirable restrictions on the parameter space.


The object of this paper is to consider the problems associated with the inliers detection in normal distribution as given in (1.1) as the distribution has many potential applications in life testing experiments with instantaneous and early failures. A two parameter normal family has the probability density
\[ f(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right), \quad -\infty < x < +\infty, \quad -\infty < \mu < +\infty, \quad \sigma > 0 \quad (1.1) \]

In section 2, we present various inlier prone models and their estimation belonging to (1.1). Section 3 deals with an illustrative example, where the estimates of the parameters under various models are discussed. The inliers detection using information criterion is presented in Section 4. In section 5 we list down two statistical tests useful to detect whether all observation belong to single normal population or they belong to mixture of two normal populations. The masking effect of the inliers is presented in the last section.

2. Inlier(s) prone models and estimation

2.1 Normal with instantaneous failures

In a parametric model for Failure time distribution (FTD) we start with a family of FTD \( \mathcal{X} = \{F(x, \theta), x \geq 0, \theta \in \Omega \subset \mathbb{R}_m\} \) where the form of the distribution function (df) is known except for labeling parameter, \( m \)-dimensional \( \theta \) and \( F \) is absolutely continuous function with probability density function (pdf), \( f(x, \theta) \) with respect to Lebesgue measure. The basic problem is to infer about unknown \( \theta \) or a suitable functions thereof say \( \psi(\theta) \), on the basis of a random sample of size \( n \) on the observable random variable say \( X_1, X_2,..., X_n \). The occurrence of instantaneous failures when some items put on test giving \( X_i = 0 \) is quite common in electronic component and life testing situations. Note that because of the limited accuracy of measuring failure time it is possible that we record \( X_i = 0 \) for some units although \( P(X_i = 0|\theta) = 0 \). To accommodate such instantaneous failures, the model \( \mathcal{X} \) is modified to model \( \mathcal{G} = \{G(x,\theta, p), x \geq 0, \theta \in \Omega, 0 < p < 1\} \), where

\[
G(x,\theta, p) = \begin{cases} 
1 - p, & x = 0 \\
1 - p + pF(x, \theta), & x > 0 
\end{cases} 
\quad (2.1)
\]

where \( F(x, \theta) \) is according to normal distribution and \( p \) is the mixing proportion. The estimation of parameters in the above model is straight forward and depends on only the positive observations in the model.

2.2 Normal with early failures

If early failures are nominally reported as \( X = \delta \) then the distribution function of the modified model \( \mathcal{G}_1 \) is given by

\[
G_1(x, p, \theta) = \begin{cases} 
0, & x < \delta \\
1 - p + pF(\delta, \theta), & x = \delta \\
1 - p + pF(x, \theta), & x > \delta 
\end{cases} 
\quad (2.2)
\]

The corresponding probability density function is given by
K. Muralidharan and Arti Khabia - INLIER PRONNESS IN NORMAL DISTRIBUTION

The likelihood of this model can be written as

\[ L(x, p, \theta) = \left[1 - p + pF(\delta, \theta)\right]^r \left(p[1 - F(\delta, \theta)]\right)^{n-r} \prod_{x_i > \delta} f(x_i, \theta) \] (2.4)

That is, the likelihood of the sample under \( g_1 \in G_1 \) is the product of the likelihoods of \( r \) and the conditional likelihood of the sample given \( r \) which is same as the likelihood of \( (n-r) \) observations coming from the truncated version of \( f \in \mathfrak{F} \) (or \( g_1 \in G_1 \)) restricted to \((\delta, \infty)\). Since \( r \) is binomial with probability of success given by \( 1 - p + pF(\delta, \theta) \), the distribution is complete for fixed \( \theta \) and \( p \in [0,1] \). Therefore, the optimal estimating equation for \( \theta \) ignoring \( p \) is the conditional score function given \( r \) or

\[ \ln L = r \ln \left[1 - pF(\delta, \theta)\right] + (n-r) \left[\ln p - \ln \sigma_1\right] \frac{1}{2} \sum_{x_i > \delta} \frac{(x_i - \theta)^2}{\sigma_i^2} \] (2.5)

\[ \frac{\partial \ln L}{\partial r} = 0 \Rightarrow \frac{-rF(\delta, \theta, \sigma_1)}{1 - pF(\delta, \theta, \sigma_1)} + \frac{(n-r)}{p} = 0 \] (2.6)

\[ \frac{\partial \ln L}{\partial \theta} = 0 \Rightarrow \frac{-rp \frac{\partial}{\partial \theta} F(\delta, \theta, \sigma_1)}{1 - pF(\delta, \theta, \sigma_1)} + \sum_{i=1}^{n-r} \frac{x_i - \theta}{\sigma_i^2} = 0 \] (2.7)

and

\[ \frac{\partial \ln L}{\partial \sigma_i} = 0 \Rightarrow -rp \frac{\partial}{\partial \sigma_i} F(\delta, \theta, \sigma_i) \frac{(n-r)}{\sigma_i} + \sum_{i=1}^{n-r} \frac{(x_i - \theta)^2}{\sigma_i^3} = 0 \] (2.8)

Here equations (2.7) and (2.8) may be solved simultaneously. The above equations give reasonably good estimates of the parameters for \( \delta \) fixed.

2.3 Normal with nearly instantaneous failures

Let \( F(x) \) and \( R(x) = 1 - F(x) \) denote the cumulative distribution function and the survival function of the mixture, respectively. The component distribution functions and their Survival functions are \( F_i(x) \) and \( R_i(x) = 1 - F_i(x) \) respectively, \( i = 1,2 \). The failure rate of a lifetime distribution is defined as \( h(x) = \frac{f(x)}{R(x)} \) provided the density exists. Instead of assuming an instant
or an early failures to occur at a particular point, as in the original model of Lai et al. (2007), we now represent this model as a mixture of the generalized Dirac delta function and the 2-parameter normal as opposed to a mixture of a singular distribution with normal. Thus the resulting modification gives rise to a density function:

\[
f(x) = p\delta_d(x-x_0) + q\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x-\theta}{\sigma_1}\right)^2\right),\quad p + q = 1,\ 0 < p < 1
\]

\[
\sigma_1 > 0,\ -\infty < \theta < +\infty
\]

Where

\[
\delta_d(x-x_0) = \begin{cases} 
\frac{1}{d}, & x_0 \leq x \leq x_0 + d \\
0, & \text{o.w.}
\end{cases}
\]

for sufficiently small \(d\). Here \(p\) is the mixing proportion and \(p > 0\). Also note that

\[
\lim_{d \to 0} \delta_d(x-x_0) = \delta(x-x_0)
\]

where \(\delta(.)\) is the Dirac delta function. We may view the Dirac delta function as approximately normal distribution having a zero mean and standard deviation that tends to 1 (see Strichartz (1994) and Li and Wong (2008) for details). For fixed value of \(d\), (2.10) denotes a uniform distribution over an interval \([x_0, x_0 + d]\) so the modified model is now effectively a mixture of a normal with a uniform distribution. Instead of including a possible instantaneous failure in the model (2.10) allows for a possible “near instantaneous” failure to occur uniformly over a very small time interval. Note that the case \(x_0 = 0\) corresponds to instantaneous failures, whereas \(x_0 \neq 0\) (but small) corresponds to the case with early failures. The survival function and failure rate functions can be obtained as follows: Since \(f(x) = p f_1(x) + q f_2(x)\) and \(F(x) = p F_1(x) + q F_2(x)\). We have,

\[
R(x) = 1 - F(x) = p + q - p F_1(x) + q F_2(x) = p R_1(x) + q R_2(x)
\]

and the corresponding failure rate function as

\[
h(x) = \frac{pf_1(x) + qf_2(x)}{pR_1(x) + qR_2(x)}
\]

where

\[
R_1(x) = \begin{cases} 
1, & 0 \leq x < x_0 \\
\frac{d + x_0 - x}{d}, & x_0 \leq x \leq x_0 + d \\
0, & x > x_0 + d
\end{cases}
\]

and

\[
R_2(x) = 1 - F_2(x) = x > x_0 + d
\]

Similarly, the failure rates for each component is given by
Consider the special case of model (2.9) whereby \( x_0 = 0 \). The model may be called the normal with “nearly instantaneous failure” model. In this case, (2.16) can be simplified as

\[
h_1(x) = \begin{cases} 
0, & 0 \leq x < x_0 \\
\frac{1}{d + x_0 - x}, & x_0 \leq x \leq x_0 + d \\
\infty, & x > x_0 + d
\end{cases}
\] (2.18)

and (2.14) simplifies to

\[
R_1(x) = \begin{cases} 
\frac{d - x}{d}, & 0 \leq x \leq d \\
0, & x > d
\end{cases}
\] (2.19)

Thus the normal model with “nearly instantaneous failure” occurring uniformly over \([0, d]\) has

\[
R(x) = \begin{cases} 
p\left(\frac{d - x}{d}\right) + q\left[1 - F_2(x)\right], & 0 \leq x \leq d \\
q\left[1 - F_2(x)\right], & x > d
\end{cases}
\] (2.20)

and

\[
h(x) = \begin{cases} 
\frac{p}{p(d-x)+dq(1-F_2(x))} + \frac{dp}{p(d-x)+dq(1-F_2(x))}f_2(x)R_2(x), & 0 \leq x \leq d \\
\frac{qf_2(x)}{R_2(x)}, & x > d
\end{cases}
\] (2.21)

respectively. The plots for reliability and failure functions are presented in figures 1 to 3 below.
3. An illustrative example

This example is due to Vannman (1991). A batch of wooden boards is dried by a particular chemical process and the object of the experiment is to compare two processes as regards the extent of deformation of boards due to checking. The measure of damage to the board is the checking area \( x \) defined as

\[
x = \frac{\overline{d}}{h l_0} \times 100,
\]

where \( l \) is the length of the check, \( \overline{d} \) is the mean depth of the check, \( h \) is the thickness of the board area and \( l_0 \) is the length of the board. Thus \( x \) is the check area measured as percentage of the board area. The boards are dried at the same time under different schedule and under some climatic conditions. When drying boards not all of them will get the checks and a typical sample of wood contain several observations with \( x_i = 0 \) or \( x_i > 0 \) but relatively small compared to the rest of the checks. These observations will correspond to instantaneous failures or early failures. Note that the larger the number of instantaneous failures better is the process. We
below reproduce the data of Schedule 1 and 2 of Experiment 3. The estimates are presented in Table 1.

E-3, S-1: \( x_i = 0, \ i = 1,2,\ldots,13 \) and the other positive observations arranged in increasing order are 0.08, 0.32, 0.38, 0.46, 0.71, 0.82, 1.15,1.23, 1.40, 3.00, 3.23, 4.03, 4.20, 5.04, 5.36, 6.12, 6.79, 7.90, 8.27, 8.62, 9.50, 10.15, 10.58 and 17.49.

E-3, S-2: \( x_i = 0, \ i = 1,2,\ldots,17 \) and the other 20 positive observations arranged in increasing order are 0.02, 0.02, 0.04, 0.09, 0.23, 0.26, 0.37, 0.93, 0.94, 1.02, 2.23, 2.79, 3.93, 4.47, 5.12, 5.19, 5.39, 6.83 and 8.22.

<table>
<thead>
<tr>
<th>Schedule</th>
<th>Instantaneous</th>
<th>Early failures</th>
<th>Nearly instantaneous</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 (( \delta = 1.5 ))</td>
<td>( \phi )</td>
<td>4.867917</td>
<td>7.352</td>
</tr>
<tr>
<td></td>
<td>( \delta_1 )</td>
<td>4.398309</td>
<td>3.745867</td>
</tr>
<tr>
<td>2 (( \delta = 0.9 ))</td>
<td>( \phi )</td>
<td>2.439000</td>
<td>3.919167</td>
</tr>
<tr>
<td></td>
<td>( \delta_1 )</td>
<td>2.606334</td>
<td>2.390099</td>
</tr>
</tbody>
</table>

### Table 1: Estimation for instantaneous failure, early failures and nearly instantaneous

4. Inliers detection using Information criterion

Denoting the parameter of \( X \) by \( \alpha_i = \theta, \ i = 1,2,\ldots,n \). We consider the following model of no inliers in the Model as

\[
\text{Model}(0): \alpha_i = \theta, \quad i = 1,2,\ldots,n \tag{4.1}
\]

and the model with \( r \) inliers as

\[
\text{Model}(r): \alpha_i = \begin{cases} 
\phi, & 1 \leq i < r \\
\theta, & r + 1 \leq i < n
\end{cases} \tag{4.2}
\]

where \( r, 1 \leq r \leq n-1 \), is the unknown index of the inliers. Model(0) may also be interpreted as having all observations from the target distribution \( F \) with common parameter \( \theta \).

Suppose that the life times of \( X_1, X_2, \ldots, X_n \) is sequence of independent random variables with normal distribution having unknown mean \( \theta \). According to the procedure, the model(0) is selected with no inliers if \( SIC(0) < \min \limits_{1 \leq r \leq n-1} SIC(r) \). And the model(r) is selected if \( SIC(0) > \min \limits_{1 \leq r \leq n-1} SIC(r) \). Here \( SIC \) is the Schwartz Information criterion. Thus we have

\[
SIC(0) = 2n \log \sigma_1 + \sum_{i=1}^{n} \left( \frac{x(i) - \theta}{\sigma_1} \right)^2 + p \log n \tag{4.3}
\]

and
\[ SIC(r) = 2r \log \sigma_0 + 2(n-r) \log \sigma_1 + \sum_{i=1}^{r} \left( \frac{x_{(i)} - \phi}{\sigma_0} \right)^2 + \sum_{i=r+1}^{n} \left( \frac{x_{(i)} - \theta}{\sigma_1} \right)^2 + p \log n \]  \hspace{1cm} (4.4)

The estimate of inliers say \( r \) is such that \( SIC(r) = \min_{1 \leq k \leq n} SIC(k) \). The above procedure is implemented through other information criteria’s like the Bayesian Information criterion:

\[ BIC = -\ln L(\Theta) + \frac{0.5 \ln(n)}{n} \]

and the Hannan-Quinn criterion given by:

\[ HQ = -\ln L(\Theta) + p \ln[\ln(n)] \]

where \( L(\Theta) \) the maximum likelihood function and \( p \) is the number of free parameters that need to be estimated under the model. The method is illustrated through numerical examples in the later sections.

5. Testing of hypothesis

Here we are interested to test the hypothesis that, whether sample observations belong to inliers population from \( N(\phi, \sigma_0^2) \) against the hypothesis that it belongs to target population from \( N(\theta, \sigma_1^2) \), assuming \( \sigma = \sigma_0 = \sigma_1 \). Equivalently, the hypothesis can be written as \( H_0: \mu = \phi \) versus \( H_1: \mu = \theta \). Below we discuss two computationally simple test procedures to detect inliers in a model.

5.1 Modified likelihood ratio test

The study of the modified likelihood approach to finite normal mixture models with a common and unknown variance in the mixing components and a test of the hypothesis of a homogeneous model versus a mixture on two or more components were done by Chen and Kalbfleisch (2005).

We define \( M_1 : \{ F(x) / x \sim N(\theta, \sigma^2) \} \), That is, all observations come from target population and \( M_2 : \{ F(x) = (1-p)F_1(x) + pF_2(x) \} \), That is, the observations comes from a mixture of two normal distributions, with \( F_1(x) \) and \( F_2(x) \) are distribution functions of inliers and target populations respectively, as defined in previous sections.

We want to test null hypothesis \( H_0 : p = 1 \) against \( H_0 : p < 1 \) or in other words a test of the hypothesis \( X \in M_1 \) versus \( X \in M_2 \) then ordinary LRT statistics is given by

\[ \ln \lambda = 2 \left[ \sup_{\theta, X \in M_2} \ln(\phi, \theta, X) - \sup_{\phi, X \in M_1} \ln(\theta, X) \right] \]  \hspace{1cm} (5.1)

Due to non-regularity of the finite mixture models \( \ln \lambda \) does not have usual chi-squared distribution. Therefore, we modify the likelihood as

\[ m \ln(\phi, \theta, X) = \ln(\phi, \theta, X) + C \ln\{ 4p(1-p) \} \]  \hspace{1cm} (5.2)
where \( C \) is a positive constant. The purpose of the penalty term \( C \ln \{4p(1-p)\} \) is to restore regularity to the problem by avoiding estimates of \( p \) on or near the boundary. Let \( \ln(\theta, X) \) maximizes \( m \ln(\theta, X) \) for \( X \in M_1 \) and \( \ln(\phi, \theta, X) \) maximizes \( m \ln(\phi, \theta, X) \) for \( X \in M_2 \). Then the modified likelihood ratio statistic is

\[
\ln \lambda = 2 \left[ \ln(\phi, \theta, X) - \ln(\theta, X) \right] \tag{5.3}
\]

The null hypothesis is rejected for large values of \( \ln \mathcal{E} \), where \( \ln \mathcal{E} \) follows \( \chi^2(2) \) distribution.

### 5.3. Most powerful test

The most powerful test for testing \( H_0 : \mu = \phi \) against \( H_0 : \mu = \theta \) where \( \mu \) is the mean of normal population and \( p \) known is given by

\[
\psi(x) = \begin{cases} 
1, & \frac{P_1(x)}{P_0(x)} > C_\alpha \\
0, & \frac{P_1(x)}{P_0(x)} < C_\alpha 
\end{cases} \tag{5.4}
\]

where \( P_1(x) \) and \( P_0(x) \) are likelihood functions under distribution of target population \( \mathcal{Z} \) and inlier population \( \mathcal{G} \) respectively, and \( C_\alpha \) is such that \( P_{H_0}[\psi(x)] = \alpha \), where \( \alpha \) is the level of significance.

We reject \( H_0 \) for large values of the ratio \( \frac{P_1(x)}{P_0(x)} \). Also, the value of \( C_\alpha \) is obtained as

\( C_\alpha = \phi + \sigma z_\alpha \), after some numerical computation.

### 6. Simulation Study

To illustrate the method of identifying inliers model we have generated 15 independent random samples, where 5 of them are from normal distribution with mean \( \phi = 4 \) and \( \sigma^2 = 2 \), and remaining ten observations from normal distribution with parameter mean \( \theta = 20 \) and \( \sigma^2 = 3 \). The observations are 1.44852, 3.667636, 3.949972, 5.548854, 6.017887, 17.61194, 19.26654, 20.09814, 20.23482, 20.36071, 20.64048, 21.08915, 21.26954, 22.53701 and 24.23439.

The identification is done as follows: Evaluate for each fixed \( r \) the maximum likelihood equation \( \mathcal{E}_r \), and then consider \( \mathcal{E} \) being that value of \( r \) for which likelihood is maximum. The estimates are presented in table 2. It is interesting to note that the likelihood is maximum
corresponds to \( r=5 \), which is expected. The corresponding estimates of the parameters are \( \hat{\phi} = 4.126574 \), \( \sigma_0 = 1.80372 \) and \( \hat{\theta} = 20.73427 \), \( \sigma_1 = 1.783219 \).

### Table 2: The Likelihood and Information criterions

<table>
<thead>
<tr>
<th>( r )</th>
<th>( L )</th>
<th>( SIC )</th>
<th>( BIC )</th>
<th>( HQ )</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td>-38.1951</td>
<td>69.8294</td>
<td>-3.4621</td>
<td>-2.6464</td>
</tr>
<tr>
<td>3</td>
<td>-34.5019</td>
<td>62.4430</td>
<td>-3.3604</td>
<td>-2.5447</td>
</tr>
<tr>
<td>4</td>
<td>-31.2064</td>
<td>55.8519</td>
<td>-3.2600</td>
<td>-2.4443</td>
</tr>
<tr>
<td>5</td>
<td>-20.7104</td>
<td>34.8599</td>
<td>-2.8501</td>
<td>-2.0344</td>
</tr>
<tr>
<td>6</td>
<td>-26.0540</td>
<td>45.5470</td>
<td>-3.0796</td>
<td>-2.2639</td>
</tr>
<tr>
<td>7</td>
<td>-28.5460</td>
<td>50.5312</td>
<td>-3.1709</td>
<td>-2.3552</td>
</tr>
<tr>
<td>8</td>
<td>-30.9970</td>
<td>55.4332</td>
<td>-3.2533</td>
<td>-2.4376</td>
</tr>
<tr>
<td>9</td>
<td>-33.0941</td>
<td>59.6274</td>
<td>-3.3188</td>
<td>-2.5031</td>
</tr>
<tr>
<td>10</td>
<td>-34.9391</td>
<td>63.3174</td>
<td>-3.3730</td>
<td>-2.5573</td>
</tr>
<tr>
<td>11</td>
<td>-36.6837</td>
<td>66.8065</td>
<td>-3.4218</td>
<td>-2.6061</td>
</tr>
<tr>
<td>12</td>
<td>-38.4748</td>
<td>70.3887</td>
<td>-3.4694</td>
<td>-2.6537</td>
</tr>
<tr>
<td>13</td>
<td>-39.6796</td>
<td>72.7984</td>
<td>-3.5003</td>
<td>-2.6846</td>
</tr>
</tbody>
</table>

Clearly \( SIC(0) = 58.4562 > SIC(5) = \min_{1 \leq r \leq 6} SIC(r) = 34.85999 \). A similar conclusion can be drawn in the case of \( BIC \) and \( HQ \). Next, we carried out an experiment with 1000 samples each of size 15 and number of inliers as 3, 4, 5 and 6 each with \( \phi = 3 \) and \( \theta = 6, 9, 12, 15 \). The table 3 entitled power of \( SIC \) procedure presents the number of times the \( SIC \) procedure correctly identified the number of inliers as proportion to total number of samples. The values clearly indicate the effectiveness of the method in detecting the inliers.

### Table 3. Power of SIC procedure

<table>
<thead>
<tr>
<th>0 / ( \phi )</th>
<th>( r )</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \theta / \phi )</td>
<td>3</td>
<td>0.570</td>
<td>0.720</td>
<td>0.700</td>
<td>0.550</td>
</tr>
<tr>
<td></td>
<td>4</td>
<td>0.460</td>
<td>0.480</td>
<td>0.490</td>
<td>0.440</td>
</tr>
<tr>
<td></td>
<td>5</td>
<td>0.460</td>
<td>0.460</td>
<td>0.460</td>
<td>0.462</td>
</tr>
<tr>
<td></td>
<td>6</td>
<td>0.410</td>
<td>0.420</td>
<td>0.430</td>
<td>0.410</td>
</tr>
</tbody>
</table>

### 6.1. Numerical Example

We recall the Vannman (1991) data example discussed in section 3 to illustrate the identification of inliers using information criterions. The computed value \( SIC(0) = 99.45467 \) and below in Table 4, the value of likelihood, \( SIC(r) \) and modified likelihood ratio for different values of \( r \) are given for different information criterions.

Clearly, \( SIC(0) = 99.45467 > SIC(9) = \min_{1 \leq r \leq 6} SIC(r) = 53.87482 \). Also the likelihood is maximum for \( r = 9 \). The corresponding estimates of the parameter are \( \hat{\phi} = 0.727778 \), \( \sigma_0 = \)
0.456858 and $\hat{\theta} = 7.352$, $\sigma_1 = 3.745867$. For modified likelihood ratio test also the maximum $\ln \hat{\xi}$ is attained at $r = 9$.

One of the important problems while detecting the inliers is the masking effect, where masking effect is defined as the loss of power due to wrong detection of more than one inliers. This is discussed in the next section.

### Table 4. Estimates of parameters for various values of $r$.  

<table>
<thead>
<tr>
<th>$r$</th>
<th>Likelihood</th>
<th>SIC</th>
<th>BIC</th>
<th>HQ</th>
<th>$\ln \hat{\xi}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>-36.1743</td>
<td>78.70478</td>
<td>-3.45593</td>
<td>-2.43208</td>
<td>20.74989</td>
</tr>
<tr>
<td>4</td>
<td>-32.7564</td>
<td>71.86888</td>
<td>-3.35668</td>
<td>-2.32323</td>
<td>27.5858</td>
</tr>
<tr>
<td>5</td>
<td>-30.8971</td>
<td>68.15026</td>
<td>-3.29824</td>
<td>-2.27439</td>
<td>31.30442</td>
</tr>
<tr>
<td>6</td>
<td>-28.6342</td>
<td>63.62454</td>
<td>-3.22218</td>
<td>-2.19833</td>
<td>35.83014</td>
</tr>
<tr>
<td>7</td>
<td>-27.5317</td>
<td>61.41942</td>
<td>-3.18292</td>
<td>-2.15907</td>
<td>38.03526</td>
</tr>
<tr>
<td>8</td>
<td>-25.6430</td>
<td>57.64209</td>
<td>-3.11185</td>
<td>-2.08800</td>
<td>41.81259</td>
</tr>
<tr>
<td>9</td>
<td>-23.7594</td>
<td>53.87482</td>
<td>-3.03556</td>
<td>-2.01171</td>
<td>45.57985</td>
</tr>
<tr>
<td>10</td>
<td>-27.4743</td>
<td>61.30473</td>
<td>-3.18083</td>
<td>-2.15698</td>
<td>38.14995</td>
</tr>
<tr>
<td>12</td>
<td>-29.3104</td>
<td>64.97688</td>
<td>-3.24552</td>
<td>-2.22167</td>
<td>34.47799</td>
</tr>
<tr>
<td>13</td>
<td>-29.6057</td>
<td>65.56758</td>
<td>-3.25555</td>
<td>-2.23170</td>
<td>33.88709</td>
</tr>
<tr>
<td>14</td>
<td>-30.5163</td>
<td>67.38864</td>
<td>-3.28584</td>
<td>-2.26199</td>
<td>32.06603</td>
</tr>
<tr>
<td>15</td>
<td>-31.1017</td>
<td>68.55955</td>
<td>-3.30484</td>
<td>-2.28099</td>
<td>30.89513</td>
</tr>
<tr>
<td>16</td>
<td>-32.0722</td>
<td>70.50050</td>
<td>-3.33557</td>
<td>-2.31172</td>
<td>28.95417</td>
</tr>
<tr>
<td>17</td>
<td>-33.2247</td>
<td>72.80552</td>
<td>-3.37087</td>
<td>-2.34702</td>
<td>26.64915</td>
</tr>
<tr>
<td>18</td>
<td>-35.0261</td>
<td>76.40824</td>
<td>-3.42367</td>
<td>-2.39982</td>
<td>23.04643</td>
</tr>
<tr>
<td>19</td>
<td>-36.5309</td>
<td>79.41796</td>
<td>-3.46574</td>
<td>-2.44189</td>
<td>20.03672</td>
</tr>
<tr>
<td>20</td>
<td>-37.8073</td>
<td>81.97070</td>
<td>-3.50008</td>
<td>-2.47623</td>
<td>17.48397</td>
</tr>
<tr>
<td>22</td>
<td>-40.8648</td>
<td>88.08568</td>
<td>-3.57785</td>
<td>-2.55400</td>
<td>11.369</td>
</tr>
</tbody>
</table>

7. Masking effect on tests for inlier(s)

Suppose $X_1, X_2, \ldots, X_n$ be sequence of $n$ independent random variables with some known FTD. Under the null hypothesis $H_0$ these random variables are identically distributed with df $F$ whereas under alternative hypothesis $H_1$, discordant observations (inliers) arise from population df $G$. The df of $G$ is assumed to be of same form as that of $F$ with a change in location or scale parameter by an unknown quantity $\lambda$. This parameter is called discordancy parameter, measuring the degree of discordancy. Under $H_1$ it is assumed one of the observations follows df $G$. Let $T(x)$ be a test statistics to detect a single discordant observation with critical region $A(n, \alpha)$. Due to lack of information about the number of discordant observations present in the sample, however, the true situation may not be specified by $H_1$ and more than one discordant observation may be present in the sample. In such cases a test statistics $T(x)$ suggested for detection of a single discordant, may fail to detect a single inlier as discordant even when additional discordant observations are present in the sample. Such a phenomenon is called masking effect.
All tests for detecting a single inlier, $H_0$ against $H_1$ are based on symmetric functions of observations or on functions of order statistics. In the k-inlier model, the joint distribution of order statistics $X_{(1)}, X_{(2)}, ..., X_{(n)}$ is same as that under the exchangeable model introduced by Kale (1975) where it is assumed that any set $X_i, X_j, ..., X_k$ has priori equal probability of being independent and identically distributed as $G_\lambda$ and the remaining $(n-k)$ observation are distributed as $F$, the distribution function of target population.

In exchangeable model $X_{(1)}, X_{(2)}, ..., X_{(n)}$ has minimum posterior probability of coming from $G_\lambda$ such that $G_\lambda F_\lambda$ is the decreasing function in $X$. The limiting masking effect (Bendre and Kale 1987) can be studied by assuming $X_{(1)}, X_{(2)}, ..., X_{(k)}$ correspond to observation coming from $N(\mu - \lambda \sigma, \sigma^2)$ and then taking limit as $\lambda \to \infty$. In the above condition, the joint probability is defined as

$$h(x_{(1)}, x_{(2)}, ..., x_{(n)}) = \frac{k!(n-k)!}{\varphi_\lambda(1,2,3, ..., k)} \prod_{i=1}^{k} g_\lambda(x_i) \prod_{i=k+1}^{n} f(x_i),$$

where $-\infty < x_{(1)} < x_{(2)} < ... < x_{(n)} < \infty$

Also $f$ and $g_\lambda$ are probability density functions of $N(\mu, \sigma^2)$ and $N(\mu - \lambda \sigma, \sigma^2)$ respectively. Thus masking effect on any test statistics $T(x)$ with critical region $A(n, \alpha)$, we have

$$\lim_{\lambda \to \infty} P[T(x) \in A(n, \alpha) / L_{sk}] = \lim_{\lambda \to \infty} \int h(x_{(1)}, x_{(2)}, ..., x_{(n)})dx_{(1)} ... dx_{(n)}$$

(7.2)

Thus under the labeled slippage model, $L_{sk}$ as $\lambda \to \infty$, $x_{(k+1)}, x_{(k+2)}, ..., x_{(n)}$ behave as order statistics of a sample of size $(n-k)$ from $N(\mu, \sigma^2)$ and $x_{(1)}, x_{(2)}, ..., x_{(k)}$ diverge to zero. However if $T(x_{(1)}, x_{(2)}, ..., x_{(n)})$ is a function whose distribution does not depend on $\lambda$ then $T$ converges in distribution to a proper random variable as $\lambda \to \infty$.

### 7.1 Limiting masking effect

In line with Grubb’s test, for a single inlier, we propose the test

$$G = \frac{\sum_{i=2}^{n} (x_{(i)} - \bar{x}_n)^2}{\sum_{i=1}^{n} (x_{(i)} - \bar{x})^2} \quad \text{where} \quad \bar{x}_n = \frac{\sum_{i=2}^{n} x_{(i)}}{n-1} \quad \text{and} \quad \bar{x} = \frac{\sum_{i=1}^{n} x_{(i)}}{n}$$

(7.3)

and the maximum studentized residual $T$ as
(7.4)

Since under $L_{s1}$ corresponds to the inlier observation coming from $N(\mu-\lambda \sigma, \sigma^2)$ and

\[
\left( \frac{x_i - \bar{x}_{(n)}}{\bar{x}_{(i)} - \bar{x}_{(n)}} \right)^2 \rightarrow 0 \quad \text{in probability as } \lambda \rightarrow \infty \quad \text{for } i=2,3,4\ldots n \quad \text{and therefore } T \rightarrow \left[ \frac{n-1}{n} \right]^z
\]

probability as $\lambda \rightarrow \infty$. Hence as $\lambda \rightarrow \infty$, $\lim P^G_1(\lambda) = 1$, where $P^G_1(\lambda)$ is the power function of Grubb’s test. To study $\lim P^G_2(\lambda) = \lim P[T < t_{n,\alpha} | L_{sk}]$ as $\lambda \rightarrow \infty$, we write

\[
T = \frac{Y_i - \overline{Y}_i}{\sqrt{\left( \sum Y_i^2 - 2k \sum \frac{Y_i^2}{n} + k^2 \right)}}, \quad i = 1,2,\ldots n
\]

where

\[
Y_i = \frac{\left( \frac{x_i - \bar{x}}{\bar{x}_i - \bar{x}} \right)}{\bar{x}_i - \bar{x}_{(n-k+1)}}
\]

With $\bar{x}_i$ the mean of $x_{(k+1)}, x_{(k+2)}, \ldots, x_{(n)}$ and $\bar{x}_i$ the mean of $x_{(1)}, x_{(2)}, \ldots, x_{(k)}$. Therefore

$Y_i \rightarrow 0$ in probability for $i=1,2,\ldots,k$ because the numerator of $Y_i$ is a proper random variable, while denominator diverges to infinity. For $i=1,2,\ldots,k$, we observe that $Y_i = \frac{\left( x_i - \bar{x}_{(n-k+1)} \right)}{\bar{x}_i - \bar{x}_{(n-k+1)}}$ is such that the numerator has a distribution independent of $\lambda$ and therefore converges to a proper random variable, but denominator diverges to infinity and hence $Y_i \rightarrow 1$ in probability as $\lambda \rightarrow \infty$.

Therefore under $L_{sk}$ as $\lambda \rightarrow \infty$, $T \rightarrow \left[ \frac{(n-k)}{nk} \right]^z$ and

\[
\lim P^G_2(\lambda) = \begin{cases} 
1, & \left[ \frac{(n-k)}{nk} \right]^z < t_{n,\alpha} \\
0, & \text{o.w.}
\end{cases}
\]

(7.7)

Thus Grubb’s test is free from the limiting masking effect for $\left[ \frac{(n-k)}{nk} \right]^z \geq t_{n,\alpha}$ and the performance of the test depends on the sample size $n$ and the number of inliers. In general $t_{n,\alpha}$ is a decreasing function of the sample size and hence for large $n$ with moderate $k$ the test is free from the limiting masking effect. Table 5, presents the maximum number of inliers in a sample of size $n$ up to which Grubb’s test is free from the limiting masking effect.
Table 5 Maximum inliers accommodated by Grubb’s test

<table>
<thead>
<tr>
<th>α</th>
<th>n = 10</th>
<th>n = 15</th>
<th>n = 20</th>
<th>n = 25</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.01</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>0.05</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>0.10</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>

From the table, it is observed that for large sample size more number of inliers may be accommodated.

References


