INLIER PRONESS IN NORMAL DISTRIBUTION

K. Muralidharan[1] and Arti Khabia [2]

•

Department of Statistics, Faculty of Science
The Maharajah Sayajirao University of Baroda, Vadodara 390 002, India.
[1]Email: lmv_murali@yahoo.com
[2]Email: 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 first formal treatment for inliers is discussed in Muralidharan and Lathika (2004). Some recent studies on inlier model related problems in exponential distribution are by Kale and Muralidharan (2000), Muralidharan and Kale (2007, 2008) and Muralidharan and Arti (2008) and the references contained therein.

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

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2, \quad -\infty<x<+\infty, -\infty<\mu<+\infty, \quad \sigma>0 \tag{1.1}$$

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 $\Im$ = $\{F(x, \theta), x \geq 0, \theta \in \Omega \subset R_m\}$ where the form of the distribution function (df) is known except for labeling parameter, m dimensional θ 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 θ 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 $P(X_i = 0 \mid \theta) = 0$. To accommodate such instantaneous failures, the model $\Im$ is modified to model G $= \{G(x,\theta,p), x \geq 0, \theta \in \Omega, 0 < p < 1\}$, where

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

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 G_1 is given by

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

The corresponding probability density function is given by

$$g_1(x,p,\theta)=\begin{cases}0, & x<\delta\\ 1-p+pF(\delta,\theta), & x=\delta\\ pf(x,\theta), & x>\delta\end{cases} \tag{2.3}$$

The likelihood of this model can be written as

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

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 $f \in \Im$ (or $g_1 \in G_1$) restricted to (δ, ∞). Since r is binomial with probability of success given by $1-p+pF(\delta,\theta)$, the distribution is complete for fixed θ and $p\in[0,1]$. Therefore, the optimal estimating equation for θ ignoring p is the conditional score function given r or $\frac{\partial \ln L_r}{\partial\theta}=0$, where $L_r=\prod_{x_i>\delta}\frac{f(x_i,\theta)}{1-F(\delta,\theta)}$. Maximum likelihood (ML) equations corresponds to two parameter normal models are given as

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

$$\frac{\partial \ln L}{\partial p}=0\Rightarrow\frac{-r\bar{F}(\delta,\theta,\sigma_1)}{1-p\bar{F}(\delta,\theta,\sigma_1)}+\frac{(n-r)}{p}=0 \tag{2.6}$$

$$\frac{\partial \ln L}{\partial \theta}=0\Rightarrow\frac{-rp\frac{\partial}{\partial\theta}\bar{F}(\delta,\theta,\sigma_1)}{1-p\bar{F}(\delta,\theta,\sigma_1)}+\sum_{r+1}^{n}\left(\frac{x_i-\theta}{\sigma_1^2}\right)=0 \tag{2.7}$$

and

$$\frac{\partial \ln L}{\partial \sigma_1}=0\Rightarrow\frac{-rp\frac{\partial}{\partial\sigma_1}\bar{F}(\delta,\theta,\sigma_1)}{1-p\bar{F}(\delta,\theta,\sigma_1)}-\left(\frac{n-r}{\sigma_1}\right)+\sum_{r+1}^{n}\frac{(x_i-\theta)^2}{\sigma_1^{\,3}}=0 \tag{2.8}$$

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

$$f(x) = p\delta_d(x - x_0) + q\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x-\theta}{\sigma_1}\right)^2\right), \quad p+q=1,\ 0<p<1 \qquad (2.9)$$

$$\sigma_1 > 0,\ -\infty < \theta < +\infty$$

Where

$$\delta_d(x - x_0) = \begin{cases} \frac{1}{d}, & x_0 \le x \le x_0 + d \\ 0, & o.w. \end{cases}, \qquad (2.10)$$

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

$$\delta(x - x_0) = \lim_{d\to 0} \delta_d(x - x_0) \qquad (2.11)$$

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,

$$R(x) = 1 - F(x) = p + q - pF_1(x) + q\,F_2(x) = pR_1(x) + qR_2(x) \qquad (2.12)$$

and the corresponding failure rate function as

$$h(x) = \frac{pf_1(x) + qf_2(x)}{pR_1(x) + qR_2(x)} \qquad (2.13)$$

where

$$R_1(x) = \begin{cases} 1, & 0 \le x < x_0 \\ \frac{d + x_0 - x}{d}, & x_0 \le x \le x_0 + d \\ 0, & x > x_0 + d \end{cases} \qquad (2.14)$$

and $$R_2(x) = 1 - F_2(x) \quad x > x_0 + d \qquad (2.15)$$

Similarly, the failure rates for each component is given by

$$h_1(x) = \begin{cases} 0, & 0 \le x < x_0 \\ \dfrac{1}{d + x_0 - x}, & x_0 \le x \le x_0 + d \\ \infty, & x > x_0 + d \end{cases} \tag{2.16}$$

and

$$h_2(x) = \frac{\dfrac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\theta}{\sigma_1}\right)^2\right)}{1-F_2(x)} \tag{2.17}$$

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

$$h_1(x) = \begin{cases} \dfrac{1}{d-x}, & 0 \le x \le d \\ \infty, & x > d \end{cases} \tag{2.18}$$

and (2.14) simplifies to

$$R_1(x) = \begin{cases} \dfrac{d-x}{d}, & 0 \le x \le d \\ 0, & x > d \end{cases} \tag{2.19}$$

Thus the normal model with "nearly instantaneous failure" occurring uniformly over [0, d] has

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

and

$$h(x) = \begin{cases} \dfrac{p}{p(d-x)+dq\left(1-F_2(x)\right)} + \left[1 - \dfrac{dp}{p(d-x)+dq\left(1-F_2(x)\right)}\right]\dfrac{f_2(x)}{R_2(x)}, & o \le x \le d \\ \dfrac{qf_2(x)}{R_2(x)}, & x > d \end{cases} \tag{2.21}$$

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

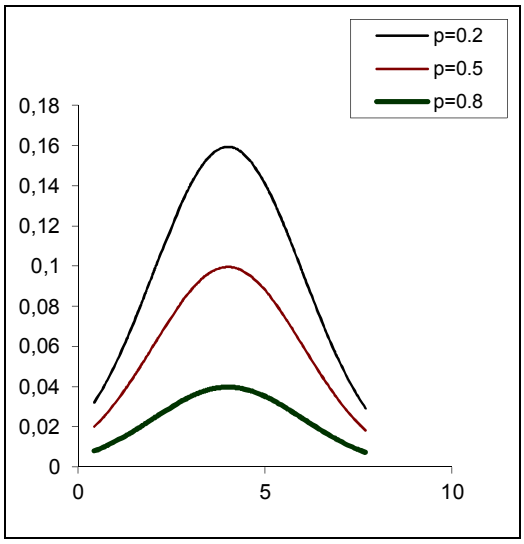

Figure 1. Density function

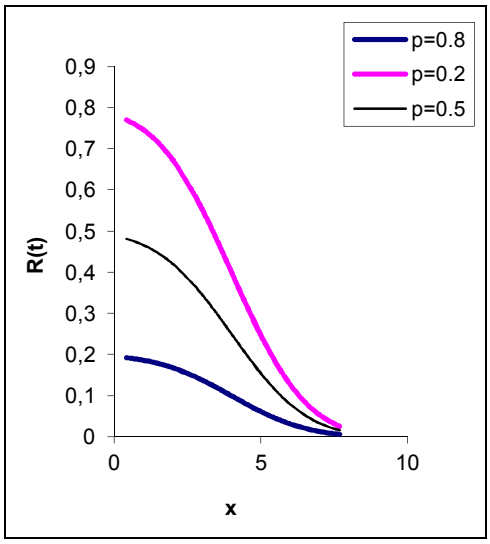

Figure 2. Reliability function

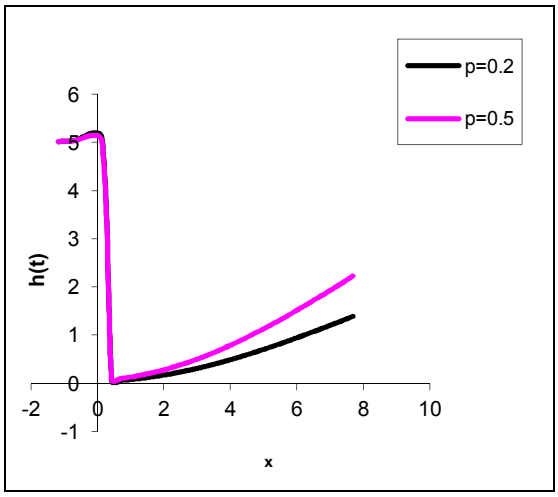

Figure 3. Failure distribution for $\mu = 4$ and $\sigma = 2$

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 $x = \frac{l\bar{d}}{hl_0}100$, where l is the length of the check, $\bar{d}$ 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,…,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,…,17 and the other 20 positive observations arranged in increasing are 0.02, 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 1: Estimation for instantaneous failure, early failures and nearly instantaneous

Schedule		Instantaneous	Early failures	Nearly instantaneous
1 (δ=1.5)	$\hat{\theta}$	4.867917	7.352	5.076087
	$\hat{\sigma}_1$	4.398309	3.745867	4.374601
2 (δ=0.9)	$\hat{\theta}$	2.43900	3.919167	3.042500
	$\hat{\sigma}_1$	2.606334	2.390099	2.581076

4. Inliers detection using Information criterion

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

$$\text{Model(0)}:\ \alpha_i = \theta,\ \ i = 1,2,.....n \tag{4.1}$$

and the model with r inliers as

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

where r, $1 \le r \le$ 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 θ.

Suppose that the life times of $X_{(1)}, X_{(2)},..., X_{(n)}$ is sequence of independent random variables with normal distribution having unknown mean θ. According to the procedure, the model(0) is selected with no inliers if $SIC(0) < \min_{1 \le r \le n-1} SIC(r)$. And the model(r) is selected if $SIC(0) > \min_{1 \le r \le n-1} SIC(r)$. 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

$$SIC(r) = 2r\log\sigma_0 + 2(n-r)\log\sigma_1 + \sum_{i=1}^{r}\left(\frac{x_{(i)}-\phi}{\sigma_0}\right)^2 + \sum_{i=r+1}^{n}\left(\frac{x_{(i)}-\theta}{\sigma_1}\right)^2 + p\log n \qquad (4.4)$$

The estimate of inliers say r is such that $SIC(r) = \min_{1\le k\le n} SIC(k)$. The above procedure is implemented through other information criteria's like the Bayesian Information criterion: $BIC = -\ln L(\Theta) + \frac{0.5p\ln(n)}{n}$ 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

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

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

$$m\ln(\phi,\theta,X) = \ln(\phi,\theta,X) + C\ln\{4p(1-p)\} \qquad (5.2)$$

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 $\ln\left(\hat{\theta}, X\right)$ maximizes $m\ln(\theta, X)$ for $X \in M_1$ and $\ln\left(\hat{\phi}, \hat{\theta}, X\right)$ maximizes $m\ln(\phi, \theta, X)$ for $X \in M_2$. Then the modified likelihood ratio statistic is

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

The null hypothesis is rejected for large values of $\ln \hat{\lambda}$, where $\ln \hat{\lambda}$ 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 μ is the mean of normal population and *p* known is given by

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

where $P_1(x)$ and $P_0(x)$ are likelihood functions under distribution of target population $\Im$ and inlier population G respectively, and C_α is such that $P_{H_0}[\psi(x)] = \alpha$, where α is the level of significance. We reject H_0 for large values of the ratio $\dfrac{P_1(x)}{P_0(x)}$. Also, the value of C_α 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 $\sigma_0^{\,2} = 2$, and remaining ten observations from normal distribution with parameter mean $\theta = 20$ and $\sigma_1^{\,2} = 3$. 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 $\hat{E}_r$, and then consider $\hat{r}$ 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, σ_0=1.80372 and $\hat{\theta}$ =20.73427, σ_1= 1.783219.

Table 2: The Likelihood and Information criterions

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

Clearly *SIC*(0) = 58.4562 > $SIC(5) = \min_{1\le r\le n} SIC(r) = 34.85999$. 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

θ / ϕ r	2	3	4	5
3	0.570	0.720	0.700	0.550
4	0.460	0.480	0.490	0.440
5	0.460	0.460	0.460	0.462
6	0.410	0.420	0.430	0.410

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 *SIC*(*r*) = 53.87482. Also the likelihood is maximum for *r* = *9*. The corresponding estimates of the parameter are $\hat{\phi}$ = 0.727778, σ_0 =

0.456858 and $\hat{\theta}$ =7.352, σ_1= 3.745867. For modified likelihood ratio test also the maximum $\ln \hat{\lambda}$ 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*.

r	Likelihood	*SIC*	*BIC*	*HQ*	$\ln \hat{\lambda}$
2	-39.7964	85.94886	-3.55136	-2.52751	13.50582
3	-36.1743	78.70478	-3.45593	-2.43208	20.74989
4	-32.7564	71.86888	-3.35668	-2.33283	27.5858
5	-30.8971	68.15026	-3.29824	-2.27439	31.30442
6	-28.6342	63.62454	-3.22218	-2.19833	35.83014
7	-27.5317	61.41942	-3.18292	-2.15907	38.03526
8	-25.6430	57.64209	-3.11185	-2.08800	41.81259
9	**-23.7594**	**53.87482**	**-3.03556**	**-2.01171**	**45.57985**
10	-27.4743	61.30473	-3.18083	-2.15698	38.14995
11	-28.1648	62.68569	-3.20565	-2.18180	36.76899
12	-29.3104	64.97688	-3.24552	-2.22167	34.47779
13	-29.6057	65.56758	-3.25555	-2.23170	33.88709
14	-30.5163	67.38864	-3.28584	-2.26199	32.06603
15	-31.1017	68.55955	-3.30484	-2.28099	30.89513
16	-32.0722	70.50050	-3.33557	-2.31172	28.95417
17	-33.2247	72.80552	-3.37087	-2.34702	26.64915
18	-35.0261	76.40824	-3.42367	-2.39982	23.04643
19	-36.5309	79.41796	-3.46574	-2.44189	20.03672
20	-37.8073	81.97070	-3.50008	-2.47623	17.48397
21	-39.3469	85.04991	-3.54000	-2.51615	14.40476
22	-40.8648	88.08568	-3.57785	-2.55400	11.369

7. Masking effect on tests for inlier(s)

Suppose $X_1, X_2 .... X_n$ 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 λ. 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,α)*. 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_o 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 $X_{i_1}, X_{i_2}, ...X_{i_k}$ has priori equal probability of being independent and identically distributed as G_λ 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_λ such that $\frac{\partial G_\lambda}{\partial F}$ 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 $N\left(\mu - \lambda\sigma, \sigma^2\right)$ and then taking limit as $\lambda \to \infty$. In the above condition, the joint probability is defined as

$$h\left(x_{(1)}, x_{(2)}....x_{(n)}\right) = \frac{k!(n-k)!}{\varphi_\lambda\left(1,2,3.....k\right)} \prod_{i=1}^{k} g_\lambda\left(x_i\right) \prod_{i=k+1}^{n} f\left(x_i\right) \quad , \qquad -\infty < x_{(1)} < x_{(2)}.... < x_{(n)} < \infty \tag{7.1}$$

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

$$\lim_{\lambda\to\infty} P\left[T\left(x\right) \in A\left(n,\alpha\right) / L_{sk}\right] = \lim_{\lambda\to\infty} \int_{A\left(n,\alpha\right)} h\left(x_{(1)}, x_{(2)}, ...x_{(n)}\right) dx_{(1)}.....dx_{(n)} \tag{7.2}$$

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

$$G = \frac{\sum_{i=2}^{n}\left(x_{(i)} - \bar{x}_n\right)^2}{\sum_{i=1}^{n}\left(x_{(i)} - \bar{x}\right)^2} \quad where \ \bar{x}_n = \frac{\sum_{i=2}^{n} x_{(i)}}{n-1} \quad and \ \bar{x} = \frac{\sum_{i=1}^{n} x_{(i)}}{n} \tag{7.3}$$

and the maximum studentized residual *T* as

$$T=\frac{\frac{(n-1)}{n}}{\left[\sum\frac{\left(x_{(i)}-\bar{x}_{(n)}\right)^2}{\left(x_{(n)}-\bar{x}_{(1)}\right)^2}+\frac{(n-1)}{n}\right]^{\frac{1}{2}}} \tag{7.4}$$

Since under L_{s1} corresponds to the inlier observation coming from $N\left(\mu-\lambda\sigma,\sigma^2\right)$ and $\frac{\left(x_{(i)}-\bar{x}_{(n)}\right)^2}{\left(x_{(1)}-\bar{x}_{(n)}\right)^2}\to 0$ in probability as $\lambda\to\infty$ for i =2,3,4…….n and therefore $T\to\left[\frac{n-1}{n}\right]^{\frac{1}{2}}$ in probability as $\lambda\to\infty$. Hence as $\lambda\to\infty$, $\lim P_1^G(\lambda)=1$, where $P_1^G(\lambda)$ is the power function of Grubb's test. To study $\lim P_2^G(\lambda)=\lim P\left[T<t_{n,\alpha}\mid L_{sk}\right]$ *as* $\lambda\to\infty$ we write

$$T=\frac{Y_{(1)}-\frac{k}{n}}{\left[\sum Y_{(1)}^2-2k\frac{\sum Y_{(i)}}{n}+\frac{k^2}{n}\right]^{\frac{1}{2}}} \tag{7.5}$$

where

$$Y_{(i)}=\frac{\left(x_{(i)}-\bar{x}_{(k)}\right)}{\left(\bar{x}'_{(n-k+1)}-\bar{x}_{(k)}\right)}\qquad i=1,2,.....n \tag{7.6}$$

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

Therefore under L_{sk} as $\lambda\to\infty$, $T\to\left[\frac{(n-k)}{nk}\right]^{\frac{1}{2}}$ and

$$\lim P_2^G(\lambda)=\begin{cases}1, & \left[\frac{(n-k)}{nk}\right]^{\frac{1}{2}}<t_{n,\alpha}\\ 0 & o.w.\end{cases} \tag{7.7}$$

Thus Grubb's test is free from the limiting masking effect for $\left[\frac{(n-k)}{nk}\right]^{\frac{1}{2}}\geq t_{n,\alpha}$ 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

α	n =10	n = 15	n = 20	n = 25
0.01	1	1	1	2
0.05	1	2	2	2
0.10	1	2	2	3

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

References

Akaike, H. (1974). A new look at the Statistical identification model. *IEEE Trans. Auto. Control,* 19, 716-723.

Bendre S.M. and Kale B.K. (1987). Masking effect on test for outliers in normal samples, *Biometrica,*74(4), 891-896.

Chen, J., Kalbfleisch J.D.(2005). Modified likelihood ratio test in finite mixture models with a structural parameter, *Journal of Statistical Planning and Inference* 129, 93-107.

Lai, C. D., Khoo, B. C., Muralidharan, K. and Xie, M. (2007). Weibull model allowing nearly instantaneous failures. *J. Applied Mathematics and Decision Sciences* Article ID 90842, 11 pages.

Kale**,** B. K. (1975)**.** Trimmed means and the method of maximum likelihood. *Applied Statistical Publishing House, Amsterdam,* 177-185.

Kale**,** B. K. and Muralidharan, K. (2007)**.** Masking effect of inliers. *J. Indian Statistical Association,* 45(1), 33-49.

Kale, B. K. and Muralidharan, K. (2008). Maximum Likelihood estimation in presence of inliers. *Journal of Indian Society for Probability and Statistics*, 10, 65-80.

Li, Y. T., and Wong, R. (2008). Integral and series representations of the Dirac delta function, *Commu. Pure Appl. Analysis*. 7 (2): 229–247.

Muralidharan, K and Lathika, P. (2004). The concept of inliers. Proceedings of the First Sino-International Symposium on Probability, Statistics and Quantitative Management., Taiwan, October, 77-92.

Muralidharan K. and Arti M. (2008). Analysis of instantaneous and early failures in Pareto distribution, *Journal of statistical theory and Applications,* Vol 7, 187-204.

Strichartz, R. (1994). *A Guide to Distribution Theory and Fourier Transforms*, CRC Press, ISBN 0-8493-8273-4.

Titterington, D.M., Smith, A., Makov, U.E.(1985). *Statistical Analysis of finite mixture distribution.* John Wiley and Sons, New York.

Vannman. K. (1991). Comparing samples from nonstandard mixtures of distributions with Applications to quality comparison of wood. Research report 1991:2 submitted to Division of Quality Technology, Lulea University, Lulea, Sweden.