Chapter 3

Inliers estimation in normal models

3.1 Introduction

A normal distribution is a very important statistical data distribution pattern occurring in many natural phenomena, such as height, blood pressure of person, 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 inlier and target groups. For example life time of a battery follows normal distribution, it is possible in the data set, we may get two sets of observations. The first set of data may have zero or small life time compared to another group with target life time. This may create two symmetrical curved graphs, where the mean of inlier group is much less than the mean of target group. Many authors have worked on mixture of normal distributions.

In this chapter the occurrence of instantaneous or early failures in life testing experiment, which is a phenomenon observed in electronic parts as well as in clinical trials is modeled as mixture of two normal distributions. These occurrences may be due to inferior quality or faulty construction or due to no response of the treatments. The modified model is then a non-standard distribution and we call such models as inlier(s) 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. Moreover, the model is not strongly identifiable so it is hard to differentiate between over dispersion caused by the presence of a mixture and that caused by a large variance, and it has infinite Fisher information with respect to mixing proportions. 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.

In the developments below we consider $N(\theta,\sigma^2)$ as our target population, and the instantaneous and early failures are inlier components. A two parameter Normal (target) family has the probability density function

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp-\frac{1}{2}\left(\frac{x-\theta}{\sigma}\right)^2,\quad -\infty<x<+\infty,-\infty<\theta<+\infty,\quad \sigma>0 \qquad (3.1.1)$$

3.2 Inlier(s) prone models and estimation

Many times in real life data, we observe that data contains inliers. The data is mostly from normal population hence, we fit models which will incorporate mixture distribution of inlier and target observations with normal distributions. The assumption considered in this chapter is that the inlier and target population defer only in their mean values, where as population variances are same.

3.2.1 Normal with instantaneous failures

In a parametric model for FTD we start with a family of FTD $\Im$={F(x, θ), x≥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 function thereof say $\psi(\theta)$, on the basis of a random sample of size n on the observable random variable say, $X_1, X_2, \ldots\ldots X_n$. The occurrence of instantaneous failures when some items are put on test giving $X_i = 0$ is quite common in electronic component and some other 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 $\Im$ is modified to model $\mathcal{G} = \{G(x,\theta,\alpha), x \geq 0, \theta \in \Omega, 0 < \alpha < 1\}$, where

$$G(x;\theta,\alpha) = \begin{cases} 1-\alpha, & x = 0 \\ 1-\alpha+\alpha F(x,\theta), & x > 0 \end{cases} \tag{3.2.1}$$

and $F(x,\theta) = \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{x} \exp -\frac{1}{2\sigma^2}(y_i-\theta)^2 dy$ is df according to Normal distribution and α 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. Thus

$$\hat{\alpha} = \frac{n-r}{n} \tag{3.2.2}$$

$$\hat{\theta} = \frac{\sum_{x_i>0} x_i}{n-r} \quad \text{and} \quad \hat{\sigma}_0^2 = \frac{\sum_{x_i>0} (x_i-\bar{x})^2}{n-r} \tag{3.2.3}$$

are easily obtainable. r denotes number of units that fail instantaneously. As we are considering life times of an object we get non-negative observations.

3.2.2 Normal with early failures

As we have already defined early failures in chapter 2, section (2.3), we can directly write the likelihood of this model as

$$L(x,\alpha,\theta)=\left[1-\alpha+\alpha F(\delta,\theta)\right]^{r}\left(\alpha\left[1-F(\delta,\theta)\right]\right)^{n-r}\prod_{x_i>\delta}\frac{f(x_i,\theta)}{1-F(\delta,\theta)} \tag{3.2.4}$$

where

$$F(\delta,\theta)=\frac{1}{\sqrt{2\pi}\sigma}\int_{-\infty}^{\delta}\exp-\frac{1}{2\sigma^2}(x_i-\theta)^2\,dx$$

that is, the likelihood of the sample under $g_1\in \mathcal{G}_1$ is the product of the likelihoods of r (inliers) and the conditional likelihood of the sample given r which is same as the likelihood of $(n\text{-}r)$ observations coming from the truncated version of $f\in\Im$ (or $g_1\in\mathcal{G}_1$) restricted to (δ, ∞). Now r is binomial with probability of success given by $1-\alpha+\alpha F(\delta,\theta)$. For fixed θ and $\alpha\in[0,1]$ this binomial family is complete. Therefore, the optimal estimating equation for θ ignoring α is the conditional score function given r or $\frac{\partial \ln L_r}{\partial\theta}=0$, where $L_r=\frac{\prod_{x>\delta}f(x_i,\theta)}{1-F(\delta,\theta)}$. Hence optimal estimating equation for θ is given by equation (3.2.7). Thus, it is same as the estimator given by optimal estimating $\hat{\theta}$ equation for θ ignoring α. ML equations correspond to two parameter Normal models are given as

$$\ln L=r\ln\left[1-\alpha\bar{F}(\delta,\theta)\right]+(n-r)[\ln\alpha-\ln\sigma_1]-\frac{1}{2}\sum_{x_i>\delta}\frac{(x_i-\theta)^2}{\sigma_1^2} \tag{3.2.5}$$

$$\frac{\partial \ln L}{\partial\alpha}=0\Rightarrow\frac{-r\alpha\bar{F}(\delta,\theta,\sigma_1)}{1-\alpha\bar{F}(\delta,\theta,\sigma_1)}+\frac{(n-r)}{\alpha}=0 \tag{3.2.6}$$

$$\frac{\partial \ln L}{\partial\theta}=0\Rightarrow\frac{-r\alpha\frac{\partial}{\partial\theta}\bar{F}(\delta,\theta,\sigma_1)}{1-\alpha\bar{F}(\delta,\theta,\sigma_1)}+\sum_{r+1}^{n}\left(\frac{x_i-\theta}{\sigma_1^2}\right)=0 \tag{3.2.7}$$

and

$$\frac{\partial \ln L}{\partial \sigma_1}=0 \Rightarrow \frac{-r\alpha \frac{\partial}{\partial \sigma_1}\bar{F}(\delta,\theta,\sigma_1)}{1-\alpha \bar{F}(\delta,\theta,\sigma_1)}-\left(\frac{n-r}{\sigma_1}\right)+\sum_{r+1}^{n}\frac{(x_i-\theta)^2}{\sigma_1^{\,3}}=0 \qquad (3.2.8)$$

Here equations (3.2.7) and (3.2.8) may be solved simultaneously using Newton Raphson method. The above model gives reasonably good estimates of the parameters for δ fixed. See the example in the section (3.8), at the end of the chapter.

3.3 Normal with nearly instantaneous failures

With reference to equation (2.4.4) in chapter 2, normal with nearly instantaneous failures distribution can be written as

$$f(x)=p\delta_d(x-x_0)+q\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x-\theta}{\sigma_1}\right)^2\right) \quad ,p+q=1,\ 0<p<1 \qquad (3.3.1)$$

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

where

$$\delta_d(x-x_0)=\begin{cases}\frac{1}{d}, & x_0\le x\le x_0+d\\ 0, & otherwise\end{cases}, \qquad (3.3.2)$$

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

$$\delta(x-x_0)=\lim_{d\to 0}\delta_d(x-x_0) \qquad (3.3.3)$$

Since

$$f_1(x)=\delta_d(x-x_0)$$

and

$$f_2(x)=\frac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\frac{1}{2}\left(\frac{x-\theta}{\sigma_1}\right)^2\right),\ \sigma_1>0,\ -\infty<\theta<+\infty$$

where $f(x)$ is given by

$$f(x)=p f_1(x)+q f_2(x) \quad \text{where} \quad p+q=1,\ 0<p<1. \tag{3.3.4}$$

and the corresponding survival function and hazard function of the mixture distribution are

$$R(x)=pR_1(x)+qR_2(x) \tag{3.3.5}$$

and

$$h(x)=\frac{pf_1(x)+qf_2(x)}{pR_1(x)+qR_2(x)} \tag{3.3.6}$$

respectively.

The components of $R(x)$ and $h(x)$ can be obtained as

$$R_1(x)=\begin{cases}1, & 0\le x<x_0\\ \dfrac{d+x_0-x}{d}, & x_0\le x\le x_0+d\\ 0, & x\ge x_0+d\end{cases} \tag{3.3.7}$$

and

$$R_2(x)=1-F_2(x) \quad x>x_0+d \tag{3.3.8}$$

$$h_1(x)=\begin{cases}0, & 0\le x<x_0\\ \dfrac{1}{d+x_0-x}, & x_0\le x\le x_0+d\\ \infty, & x\ge x_0+d\end{cases} \tag{3.3.9}$$

and

$$h_2(x)=\frac{\dfrac{1}{\sqrt{2\pi}\sigma_1}\exp\left(-\dfrac{1}{2}\left(\dfrac{x-\theta}{\sigma_1}\right)^2\right)}{1-F_2(x)} \tag{3.3.10}$$

As a special case of the model, we obtain the Normal with "nearly instantaneous failure" model, when $t_0 = 0$ in equation (3.3.2). Accordingly the simplified expressions of the components in the failure rate and survival functions are

$$h_1(x) = \begin{cases} \dfrac{1}{d-x}, & 0 \le x \le d \\ \infty, & x > d \end{cases} \tag{3.3.11}$$

and its survival rate function in equation (3.3.7) is given as

$$R_1(x) = \begin{cases} \dfrac{d-x}{d}, & 0 \le x \le d \\ 0, & x > d \end{cases} \tag{3.3.12}$$

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

$$R(x) = \begin{cases} \dfrac{p(d-x)}{d} + q\left[1 - F_2(x)\right], & 0 \le x \le d \\ q\left[1 - F_2(x)\right], & x > d \end{cases} \tag{3.3.13}$$

and

$$h(x) = \begin{cases} \dfrac{p}{p(d-x) + dq\left(1 - F_2(x)\right)}\left[1 - \dfrac{dp}{p(d-x) + dq\left(1 - F_2(x)\right)}\right]\dfrac{f_2(x)}{R_2(x)}, & 0 \le x \le d \\ \dfrac{qf_2(x)}{R_2(x)}, & x > d \end{cases} \tag{3.3.14}$$

Nearly instantaneous calculations are performed for the example in section (3.8).

3.3.1 Graphs

In various figures below we provide the graphs for $f(x)$, $R(x)$ and $h(x)$ for left values of mixing proportions and parametric values.

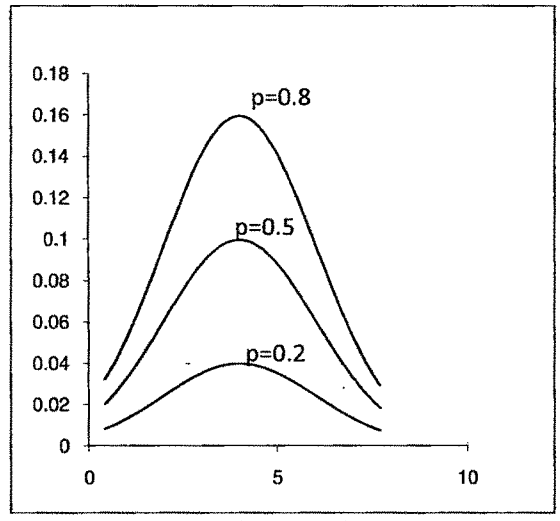

Fig. 3.3.1. Density function for $\mu = 4$ and $\sigma = 2$

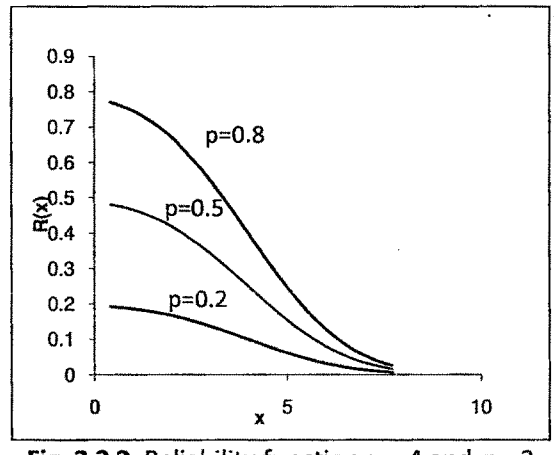

Fig. 3.3 2. Reliability function $\mu = 4$ and $\sigma = 2$

Graph (3.3.4) and (3.3.5) are plotted on the basis of random sample generated from mixture of two normal distributions. From the graph (3.3.4) we can clearly identify two symmetrical curves, where first curve has inlier distribution with mean 4 remarkably less than second curve which can be considered as target distribution with mean 20. Graph (3.3.5) is known as normal quantile-quantile (Q-Q) plot. A sample from single normal distribution should produce a linear plot on this graph, which is not in our case. Hence both the graph clearly represents the presence of two groups.

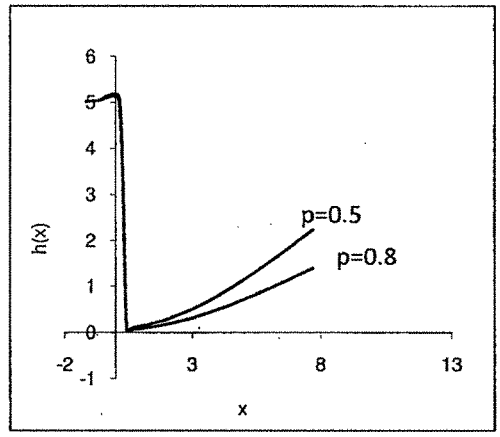

Fig. 3.3.3. Failure distribution for $\mu = 4$ and $\sigma = 2$

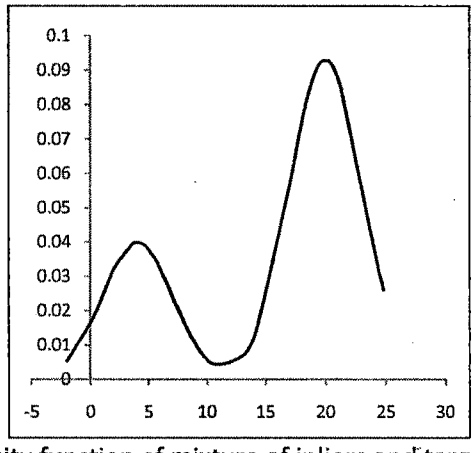

Fig 3.3.4. Density function of mixture of inliers and target distributions

3.4 Inlier detection methods

Here we obtain number of inliers for different data set by various methods, viz identified inlier model, labeled slippage methods and information criteria.

3.4.1 Identified inlier model (M_k)

Referring to equation (2.5.14) of section (2.5.2) from chapter 2 the identified inliers model with $g(x)$ as inliers and $f(x)$ as target distribution is written as

$$L(x|\phi,\theta,v,r)=\prod_{i=1}^{r} g(x_i)\prod_{i=r+1}^{n} f(x_i) \qquad (3.4.1)$$

$$= \prod_{i=1}^{r} \frac{1}{\sqrt{2\pi}\sigma_0} \exp -\frac{1}{2}\left(\frac{x_i - \phi}{\sigma_0}\right)^2 \prod_{i=r+1}^{n} \frac{1}{\sqrt{2\pi}\sigma_1} \exp -\frac{1}{2}\left(\frac{x_i - \theta}{\sigma_1}\right)^2 \quad (3.4.2)$$

The likelihood function in (3.4.2) assumes that between the experiments when units are placed on test we do not know which of the units fail instantaneously. Equivalently $X_{i_1} = 0, X_{i_2} = 0, \ldots X_{i_r} = 0$ which fail early i.e. those units whose failure time distribution is $g(x_{(i)}, \phi)$ with failure rate much larger than that of the failure time distribution of the target population whose failure rate is considerably smaller. The identification is done as follows: evaluate for each fixed *r* where *r* = 0,1,2,...*n*-1 the maximum likelihood equation $\hat{L}_r$, and then consider $\hat{r}$ being that value of *r* for which likelihood is maximum. The computation for example of detection of inliers is done in section (3.5) and (3.8).

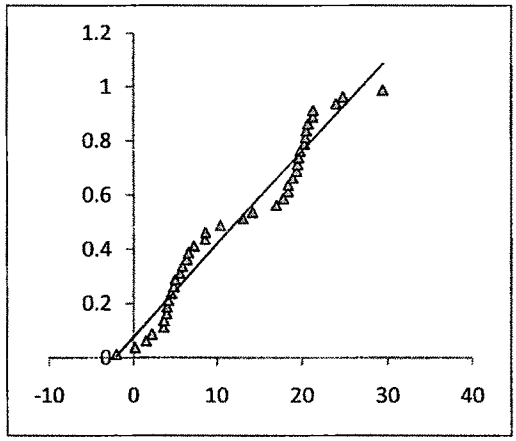

Fig. 3.3.5. Normal Probability Plot for mixture of two distributions

3.4.2 Inlier detection in Labeled slippage model (L_k)

With $g(x)$ and $f(x)$ as described in section (3.4.1), the likelihood under labeled slippage model referring to section (2.5) and substituting in equation (2.5.1), gives

$$\ln L = r_0 \ln(1-p) + (n-r_0)\ln p - \ln \varphi_{r_1}(\phi,\theta) + n\ln\sigma - \frac{\sum_{i=1}^{r_1}\left(x_{(i)}-\phi\right)^2}{2\sigma^2} - \frac{\sum_{i=r_1+1}^{n}\left(x_{(i)}-\theta\right)^2}{2\sigma^2} \tag{3.4.3}$$

and the corresponding likelihood equations are

$$\frac{\partial \ln L}{\partial p} = \frac{-r_0}{(1-p)} + \frac{(n-r_0)}{p} = 0 \tag{3.4.4}$$

$$\frac{\partial \ln L}{\partial \phi} = -\frac{\partial}{\partial \phi}\ln \varphi_{r_1}(\phi,\theta) + \frac{\sum_{i=1}^{r_1} x_{(i)}}{r_1} \tag{3.4.5}$$

$$\frac{\partial \ln L}{\partial \theta} = -\frac{\partial}{\partial \theta}\ln \varphi_{r_1}(\phi,\theta) + \frac{\sum_{i=r_1+1}^{n} x_{(i)}}{(n-r_0-r_1)} \tag{3.4.6}$$

and

$$\frac{\partial \ln L}{\partial \sigma} = 0 \Rightarrow \hat{\sigma} = \frac{\sum_{i=1}^{r_1}\left(x_{(i)}-\phi\right)^2 + \sum_{i=r_1+1}^{n}\left(x_{(i)}-\theta\right)^2}{n} \tag{3.4.7}$$

Here (3.4.4) can be solved to get the estimate of p as $\hat{p} = (n-r_0)/n$. The equations (3.4.5) and (3.4.6) contains gamma and digamma functions. The function

$$\varphi_{r_1}(\phi,\theta) = \frac{(n-r_0-r_1)}{\sqrt{2\pi}\sigma}\int_0^{\infty}\left\{G(x)\right\}^{r_1}\left[\bar{F}(x)\right]^{n-r_0-r_1} e^{-\frac{1}{2\sigma^2}(x-\theta)^2} dx$$

where $G(x)$ and $F(x)$ are cumulative distribution functions of inlier and target population. The function $\varphi_{r_1}(\phi,\theta)$ is difficult to evaluate and can only be evaluated using some numerical method.

3.4.3 Information criterion for detection of inliers

As defined in chapter 2, section (2.6) here for Normal distribution, we have *SIC* for model with no inliers as

$$SIC(0) = 2n\log\sigma_1 + \sum_{i=1}^{n}\left(\frac{x_i-\theta}{\sigma_1}\right)^2 + p\log n \qquad (3.4.8)$$

and model with *r* inliers is defined as

$$SIC(r) = 2r\log\sigma_0 + 2(n-r)\log\sigma_1 + \sum_{i=1}^{r}\left(\frac{x_i-\phi}{\sigma_0}\right)^2 + \sum_{i=r+1}^{n}\left(\frac{x_i-\theta}{\sigma_1}\right)^2 + p\log n \qquad (3.4.9)$$

The estimate of inliers say *r* is such that $SIC(r) = \min_{1\le r\le n} SIC(r)$.

Here we use three information criteria such as *SIC, BIC* and *HQ* already defined in chapter 2. Hence *SIC* = $-2\ln L(\Theta) + p \ln n$, *BIC* = $-\ln L(\Theta) + \frac{0.5p\ln(n)}{n}$ *and* *HQ* = $-\ln L(\Theta) + p\ln[\ln(n)]$ can be used to detect the inliers, where $L(\Theta)$ the maximum likelihood function and *p* is the number of free parameters that need to be estimated under the model. We now illustrate this method using the simulated example discussed in the next section. Table (3.5.2) also presents the parameter estimates and the information criterion values.

3.5 Simulation study

To Illustrate the method of identifying inliers we have generated 15 independent random samples, where 5 of them are coming from normal distribution with parameter mean $\phi = 4$ and variance $\sigma_0^2 = 2$ and remaining ten observations from Normal distribution with parameter mean $\theta = 20$ and variance $\sigma_1^2 = 2$. The sample values 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. We note that $SIC(0) = 58.4562 > SIC(5) = \min_{1\le r\le n} SIC(r) = 34.85999$.

Table 3.5.1. The Likelihood and Information criterions

r	L	SIC	BIC	HQ
2	-38.1951	69.82944	-3.46217	-2.64648
3	-34.5019	62.44302	-3.36048	-2.54479
4	-31.2064	55.85195	-3.26009	-2.44439
5	**-20.7104**	**34.85999**	**-2.8501**	**-2.03441**
6	-26.054	45.54709	-3.07963	-2.26394
7	-28.546	50.53121	-3.17098	-2.35529
8	-30.997	55.43326	-3.25336	-2.43766
9	-33.0941	59.62746	-3.31882	-2.50313
10	-34.9391	63.31742	-3.37307	-2.55738
11	-36.6837	66.80655	-3.4218	-2.6061
12	-38.4748	70.38878	-3.46947	-2.65377
13	-39.6796	72.79842	-3.5003	-2.68461

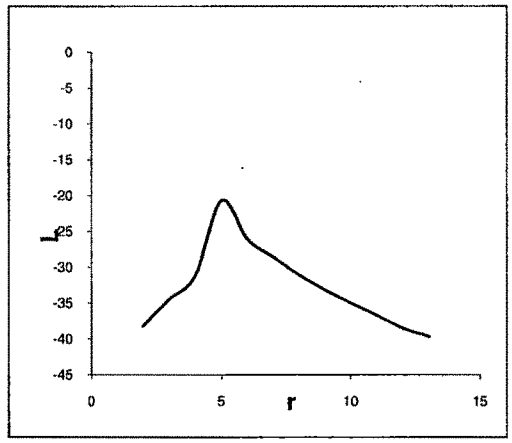

Fig. 3.5.1. Likelihood plot

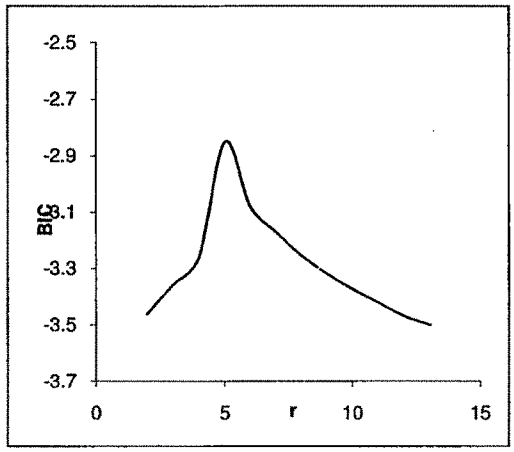

Fig. 3.5.2. *BIC* plot

A similar conclusion can be drawn in the case of other information criterions BIC and HQ also. Hence $r=5$ and the estimates are $\hat{\phi}=4.126574$, $\hat{\sigma}_0=1.803727$ $\hat{\theta}=20.73427$, and $\hat{\sigma}_1=1.783219$ respectively. The graphical representations of the likelihood and *BIC* plots are given in figure (3.5.1) and (3.5.2).

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 $\theta=6, 9, 12$ and 15. The table (3.5.2) entitled power of SIC procedure presents the number of times the SIC procedure correctly identified the number of inliers in proportion to total number of samples. The values clearly indicate the effectiveness of the method in detecting the inliers. One of the important problem 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.

Table 3.5.2. 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

3.6 Testing of hypothesis for inliers

After detection of number of inliers, it is necessary to test whether the methods used for detection are valid or not. Hence different tests are applied to test whether data truly represents our model of mixture of inliers and target population.

3.6.1 Sequential Probability Ratio Test (SPRT) to detect number of inliers

We want to test the hypothesis whether sample observations belong to inliers population from $N(\phi,\sigma_0^2)$ against hypothesis that it belongs to target population from $N(\theta,\sigma_1^2)$, assuming $\sigma=\sigma_0=\sigma_1$.

H_0 : sample observations are taken from normal population with mean ϕ

H_1 : sample observations are taken from normal population with mean θ

We use SPRT test given as follows:

The likelihood ratio λ_m is given by $\lambda_m = \frac{L_{1m}}{L_{0m}}$ or equivalently

$$\ln\lambda_m = \sum_{i=1}^{m} \ln\frac{f\left(x_{(i)},\theta\right)}{g\left(x_{(i)},\phi\right)}$$

$$= \frac{m\left(\phi^2-\theta^2\right)+2\left(\theta-\phi\right)\sum_{i=1}^{m} x_{(i)}}{2\sigma^2}, \quad m=1,2,....n \qquad (3.6.1)$$

For deciding number of inliers *r*, first arrange the observations in ascending order and then we continue to take likelihood ratio for m= 1, 2.... ,*n* by including observations one by one till we reject H_0. That is

If $\sum_{i=1}^{m} z_{(i)} \leq \ln B$ then accept H_0 and take the next observation.

and

If $\sum_{i=1}^{m} z_{(i)} \geq \ln A$ reject H_0 and stop.

The corresponding value of m represents the first observation from target population and number of inliers $\hat{r}=m-1$. A and B are given as

$$B=\frac{\beta}{1-\alpha}, \quad A=\frac{1-\beta}{\alpha} \qquad (3.6.2)$$

where α represents probability of type I error and β represents probability of type II error.

Test criteria for rejection of H_0 is

$$\ln\lambda_m > \ln A \Rightarrow \sum_{i=1}^{m} x_{(i)} > \frac{\sigma^2}{(\theta-\phi)}\ln A + \frac{m}{2}(\phi+\theta) \qquad (3.6.3)$$

Corresponding value of m for which H_0 was accepted last becomes number of inliers r. The criteria is applied in example in section (3.8).

3.6.2 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 was done by Chen and Kalbfleisch (2005). Here we use it to study the test for hypothesis

H_0 : sample observations are taken from single target normal population with mean θ

H_1 : sample observations are taken from mixture of inliers with mean ϕ and target distribution with mean θ.

We define $M_1 = \{F(x) : x \sim N(\theta, \sigma^2)\}$ i.e. all observations come from target population. $M_2 = \{F(x) = (1-p)F_1(x) + pF_2(x)\}$ *i.e.* X comes from mixture of two Normal distribution where $F_1(x)$ and $F_2(x)$ are distribution functions of inliers and target population, respectively, as defined in previous section.

Then the null hypothesis proceeds with testing $H_0 : p = 1$ against $H_1 : p < 1$ or in other words a test of the hypothesis $X \in M_1$ versus $X \in M_2$. The usual likelihood (LRT) statistics is given by

$$\ln\lambda = 2\left[\sup_{\theta, X \in M_1} \ln(\theta, X) - \sup_{\phi, \theta, X \in M_2} \ln(\phi, \theta, X)\right] \tag{3.6.4}$$

Due to non-regularity of the finite mixture models ln λ does not have usual chi-squared distribution. Therefore we proceed with a modified likelihood approach where the quantity $\ln(\phi, \theta, X)$ is replaced as

$$m\ln(\phi, \theta, X) = \ln(\phi, \theta, X) + c\ln\{4p(1-p)\} \tag{3.6.5}$$

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 estimate of p on or near the boundary. Let $\ln(\hat{\theta},\hat{X}_1)$ maximizes $m\ln(\theta,X)$ for $X\in M_1$ and $\ln(\hat{\phi},\hat{\theta},\hat{X}_2)$ maximizes $m\ln(\phi,\theta,X)$ for $X\in M_2$. Thus modified likelihood ratio statistic is

$$\ln\hat{\lambda}=2\left[\ln(\hat{\theta},\hat{X}_1)-\ln(\hat{\phi},\hat{\theta},\hat{X}_2)\right] \tag{3.6.6}$$

The null hypothesis is rejected for values of $\ln\hat{\lambda}$ that are sufficiently large. Here $\ln\hat{\lambda}$ follows $\chi^2_{(2)}$ distribution.

3.6.3 Most powerful test for detection of inliers

The most powerful test for testing the hypothesis as given in (1.6.1) whether the sample is from single population, we frame the hypothesis with common parameter μ

$H_0:\mu=\phi$ i.e sample observations are from inliers normal population

$H_1:\mu=\theta$ i.e sample observations are from target normal population

where μ is the mean of normal population and $\theta>\phi$.

Then the most powerful test is as given below

$$\psi(x)=\begin{cases}1, & \dfrac{P_1(x)}{P_0(x)}>C_\alpha\\ 0, & \dfrac{P_1(x)}{P_0(x)}<C_\alpha\end{cases} \tag{3.6.7}$$

which can be simplified as

$$\psi(x)=\begin{cases}1, & \sum_{i=1}^{n}x_i>\dfrac{C_\alpha\sigma^2}{(\theta-\phi)}+\dfrac{n(\theta+\phi)}{2}\\ 0, & o.w\end{cases} \tag{3.6.8}$$

where C_α is such that the test attains level of the test when H_0 is true. Thus we reject H_0 for large values of the $\sum_{i=1}^{n} x_i$ with $C_\alpha = \phi + \sigma z_\alpha$.

3.6.4 F- test to test whether data contains inlier observations

To test whether the data is taken from single normal population or from mixture of inlier and target (both normal) distributions, we proceed with the F-test as follows

$H_0 : x_1, x_2, \ldots x_n$ are independent and follows $N(\theta, \sigma^2)$

$H_1 : x_{(1)}, x_{(2)}, \ldots x_{(r)}$ follows $N(\phi, \sigma_0^2)$ and $x_{(r+1)}, x_{(r+2)}, \ldots x_{(n)}$ follows $N(\theta, \sigma_1^2)$

where $\phi < \theta$.

Then test statistic obtained by Titterrington(1985) gives the maximum ratio of between sum of squares to within sum of squares as

$$F_{\max} = \frac{\max n_1 n_2 (\bar{x}_1 - \bar{x}_2)^2}{\left[(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2\right](n_1 + n_2)} \qquad (3.6.9)$$

where the maximum is over all partitioning of data set into two groups..

For detection of inliers, we find F_{max} for all possible values of r =1, 2,...n-1. The number of inliers r will be detected for which corresponding value of F_{max} is maximum.

3.7 Masking effect on tests for inliers

Let $X_1, X_2 \ldots X_n$ 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 λ. This parameter is called discordancy parameter, measuring the degree of discordancy. Under H_1 it is assumed that one of the observation 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 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 $X_{(1)}, X_{(2)}, \ldots X_{(n)}$ is same as that under the exchangeable model introduced by Kale (1975) where it is assumed that any set $X_{i_1}, X_{i_2}, \ldots X_{i_k}$ has priori equal probability of being independent and identically distributed as G_λ and the remaining (*n-k*) observations are distributed as *F*, the distribution function of target population.

In exchangeable model $X_{(1)}, X_{(2)}, \ldots X_{(n)}$ 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 by Bendre and Kale (1985) can be studied by assuming $X_{(1)}, X_{(2)}, \ldots X_{(k)}$ correspond to observation coming from $N(\mu - \lambda\sigma, \sigma^2)$ and then taking limit as $\lambda \to \infty$.

$$h\left(x_{(1)}, x_{(2)} \ldots x_{(n)}\right) = \frac{k!(n-k)!}{\varphi_\lambda(1,2,3 \ldots\ldots k)} \prod_{i=1}^{k} g_\lambda(x_i) \prod_{i=k+1}^{n} f(x_i) \quad , \qquad (3.7.1)$$

$$-\infty < X_{(1)} < X_{(2)} \ldots < X_{(n)} < \infty$$

Also f and g_λ 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 $A(n,\alpha)$, for Labelled slippage model L_{sk} for $k \geq 1$, is obtained as

$$\lim_{\lambda\to\infty} P\left[T(x)\in A(n,\alpha)/L_{sk}\right]=\lim_{\lambda\to\infty}\int_{A(n,\alpha)} h\left(x_{(1)},x_{(2)},\ldots\ldots x_{(n)}\right)dx_{(1)}\ldots\ldots dx_{(n)} \qquad (3.7.2)$$

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

3.7.1. Limiting masking effect

For single inlier in left tail, that is to test whether $X_{(1)}$ is an inliers, Grubbs proposed a test proposed by Bendre and Kale (1987).

$$G=\frac{\sum_{i=2}^{n}\left(x_{(i)}-\bar{x}_n\right)^2}{\sum_{i=1}^{n}\left(x_{(i)}-\bar{x}\right)^2}, \qquad (3.7.3)$$

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

The maximum studentized residual T is given by

$$T=\frac{\dfrac{(n-1)}{n}}{\left[\sum\dfrac{\left(x_{(i)}-\bar{x}_{(n)}\right)^2}{\left(x_{(n)}-\bar{x}_{(1)}\right)^2}+\dfrac{(n-1)}{n}\right]^{\frac{1}{2}}} \qquad (3.7.4)$$

where the sum is over $i = 2,3.....n$. Since under L_{s1} corresponds to the one inlier observation coming from $N(\mu-\lambda\sigma,\sigma^2)$ 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 $\lim P_2^G(\lambda)=\lim P\left[T<t_{n,\alpha}\,|\,L_{sk}\right]$ as $\lambda\to\infty$ we write

$$T=\frac{Y_{(1)}-\frac{k}{n}}{\left[\sum Y_{(1)}^2-2k\frac{\sum Y_{(i)}}{n}+\frac{k^2}{n}\right]^{\frac{1}{2}}} \tag{3.7.5}$$

sums are over i =1, 2,...n and where

$$Y_{(i)}=\frac{\left(x_{(i)}-\bar{x}_{(k)}\right)}{\left(\bar{x}'_{(n-k+1)}-\bar{x}_{(k)}\right)} \qquad i=1,2,.....n \tag{3.7.6}$$

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 r.v., 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

$$\lim P_2^G(\lambda)=\begin{cases}1, & \left[\frac{(n-k)}{nk}\right]^{\frac{1}{2}} < t_{n,\alpha} \\ 0 & o.w.\end{cases} \qquad (3.7.7)$$

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 (3.7.1), presents the maximum number of inliers in a sample of size n upto which Grubb's test is free from the limiting masking effect.

Table. 3.7.1 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

3.8 Illustrations

3.8.1 Vannman's data

This example is based on a wood drying experiment. The data of Schedule 1 and 2 of Experiment 3 conducted by Vannman (1991). In both the case n=37. For data refer appendix.

Table (3.8.1) presents the estimates of the parameters of target distribution under instantaneous failure, early failures and nearly instantaneous models.

Table 3.8.1 Estimation for instantaneous failure, early failures and nearly instantaneous failures

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	$\hat{\theta}$	2.439	3.919167	3.0425
	$\hat{\sigma}_1$	2.606334	2.390099	2.581076

3.8.2 Rainfall data

The data, collected by Amutha and Porchelvan (2009), represents average monthly rainfall (in mm) during year 2004 and 2006 for the estimation of surface runoff in Malattar Sub-watershed which is a major tributary of Palar river. The watershed experiences tropical monsoon climate with normal temperature, humidity and evaporation throughout the year. The data was published in Journal of the Indian Society of Remote Sensing. For our illustration's purpose we reproduce two sets of data from the above paper.

Set 1 (2004) : 3.40, 0.00, 0.00, 15.80, 232.80, 8.80, 123.20, 47.00, 154.00, 103.20, 89.80 and 12.20.

Set 2 (2006) : 0.00, 0.00, 21.40, 60.20, 53.86, 93.20, 27.80, 45.40, 205.40, 101.20, 128.20 and 0.00.

We have combined the two sets together and arranged in ascending order to obtain inlier detection discussed in section (3.3), (3.4) and (3.6). Table (3.8.2), represents the value of inlier numbers r, likelihood, *SIC(r), BIC(r) HQ(r)* and modified test statistics for different values of *r*.

Table 3.8.2. Detection of number of inliers

r	Likelihood	SIC	BIC	HQ	$\ln\hat{\lambda}$
2	-39.796	85.9489	-3.5514	-2.5275	7.58094
3	-36.174	78.7048	-3.4559	-2.4321	14.8250
4	-32.756	71.8689	-3.3567	-2.3328	21.6609
5	-30.897	68.1503	-3.2982	-2.2744	25.3795
6	-28.634	63.6245	-3.2222	-2.1983	29.9053
7	-27.532	61.4194	-3.1829	-2.1591	32.1104
8	-25.643	57.6421	-3.1119	-2.0880	35.8877
9	**-23.759**	**53.8748**	**-3.0356**	**-2.0117**	**39.6550**
10	-27.474	61.3047	-3.1808	-2.1570	32.2251
11	-28.165	62.6857	-3.2057	-2.1818	30.8441
12	-29.31	64.9769	-3.2455	-2.2217	28.5529
13	-29.606	65.5676	-3.2555	-2.2317	27.9622
14	-30.516	67.3886	-3.2858	-2.2620	26.1412
15	-31.102	68.5595	-3.3048	-2.2810	24.9702
16	-32.072	70.5005	-3.3356	-2.3117	23.0293
17	-33.225	72.8055	-3.3709	-2.3470	20.7243
18	-35.026	76.4082	-3.4237	-2.3998	17.1216
19	-36.531	79.4180	-3.4657	-2.4419	14.1118
20	-37.807	81.9707	-3.5001	-2.4762	11.5591
21	-39.347	85.0499	-3.5400	-2.5161	8.47988
22	-40.865	88.0857	-3.5778	-2.5540	5.44412

Table (3.8.2) gives us *SIC(0)* = 99.45467 > *SIC(9)* = min *SIC(r)* = -23.759. The likelihood is maximum for $r = 9$. The corresponding estimates of the parameter are $\hat{\phi}$ = 0.72778, σ_0 = 0.45686 and $\hat{\theta}$ = 7.352 , σ_1 = 3.74587. For modified likelihood ratio test also maximum $\ln\hat{\lambda}$ corresponds to $r = 9$. We observe that, *SIC(0)* = 183.2181 > *SIC(6)* = min *SIC(r)* = 173.5757 . Also the likelihood is maximum for $r = 6$. The corresponding estimates of the parameter are $\hat{\phi}$ = 14.9, σ_0 = 8.78886 and $\hat{\theta}$ = 111.182, σ_1 = 58.0748. For modified likelihood ratio test also maximum $\ln\hat{\lambda}$ corresponds to $r = 6$. For SPRT, we test H_0: $\phi = 15$ against H_0: $\phi > 15$. For which we considered (α, β) =(0.02, 0.05). Then $\ln A$ = -2.5647 and $\ln B$ = -2.9755, and the computed statistics value is $\frac{\sigma^2}{(\theta-\phi)}\ln A+\frac{m}{2}(\phi+\theta)$ = 101.2454 .

Table 3.8.3 Estimates of parameters and detection of r

r	Likelihood	*SIC*	*BIC*	*HQ*	$\ln\hat{\lambda}$	$\sum_{i=1}^{m} x_{(i)}$
2	-91.4964	188.8817	91.57392	91.49643	-5.663600	12.200
3	-88.8329	183.5547	88.91039	88.83290	-0.336540	24.400
4	-86.5127	178.9142	86.59016	86.51268	4.303914	40.200
5	-84.9457	175.7802	85.02315	84.94566	7.437946	61.600
6	**-83.8434**	**173.5757**	**83.92090**	**83.84341**	**9.642444**	89.400
7	-85.0201	175.9291	85.09758	85.02010	7.289072	134.80
8	-84.4475	174.7838	84.52495	84.44746	8.434343	181.80
9	-83.9214	173.7317	83.99890	83.92141	9.486446	235.66
10	-84.2590	174.4069	84.33647	84.25899	8.811291	303.66
11	-85.9395	177.7679	86.01701	85.93953	5.450209	393.56
12	-86.7336	179.3560	86.81104	86.73355	3.862165	486.76
13	-87.4507	180.7903	87.52822	87.45073	2.427797	587.96
14	-87.6541	181.1971	87.73158	87.65410	2.021071	691.16
15	-88.6952	183.2794	88.77273	88.69525	-0.061230	814.36
16	-89.1081	184.1052	89.18562	89.10814	-0.887010	942.56
17	-89.4825	184.8538	89.55995	89.48247	-1.635670	1096.5

Hence we reject H_0 for first time when inlier r is 7 and conclude that number of inliers in the above data set, see table (3.8.3) is $\hat{r}=6$.