Journal of Reliability and Statistical Studies (ISSN: 0974-8024)
Vol. 1, Issue 1(2008) 1-5

INLIERS DETECTION USING SCHWARTZ INFORMATION CRITERION

K. Muralidharan [1] **and B. K. Kale** [2]
1. Department of Statistics, M. S. University of Baroda, Vadodara 390 002, India.
Email: lmv_murali@yahoo.com
2. Department of Statistics,University of Pune, Poona, 400 007 India.
Email: bkkale@stats.unipune.ernet.in

Abstract

In failure time distributions, inliers in a data set are subset of observations sufficiently small relative to the rest of the observations, which appears to be inconsistent with the remaining data set. They are either the resultant of instantaneous failures or early failures, experienced in many life-testing experiments. The model $M_r(E)$ used in outliers, where r observations are outliers is modified by $\partial G / \partial F$ strictly decreasing instead of increasing function of X to represent this situation, where F is the target distribution and G generates inliers. Usually number of inliers is not known and is to be determined. We use the information criterion given by Schwartz (1978) to detect the number of inliers in the model. The method is illustrated with a simulated experiment and a real life data.

Key Words: Exponential Distribution, Instantaneous Failures, Mixture Distribution, Maximum Likelihood, Asymptotic Distribution, Labeled Slippage Model, Early Failures, Inliers.

1. Introduction

The occurrence of instantaneous or early failures in life testing experiment is observed in electronic parts as well as in clinical trials. These occurrences may be due to inferior quality or faulty construction or due to no response to the treatments. The usual method of modeling and inference procedures may not be valid in this case. These situations can be modeled by modifying commonly used parametric models such as exponential, gamma, Weibull and lognormal distribution among others. The modified model is then a non-standard mixture of distribution by mixing a singular distribution at zero to accommodate instantaneous failures.

Consider a model $\Im = \{F(x, \theta), x \geq 0, \theta \in \Omega\}$ where $F(x, \theta)$ is a continuous failure time distribution function (df) with $F(0)=0$. To accommodate a real life situation, where instantaneous failures are observed at the origin, the model $\Im$ is modified to $\mathcal{G} = \{G(x, \theta, \alpha), x \geq 0, \theta \in \Omega, 0<\alpha<1\}$ by using a mixture in the proportion $1-\alpha$ and α respectively of a singular random variable Z at zero and with a random variable X with df $F \in \Im$. Thus the modified failure time distribution is given by

$$G(x,\theta,\alpha) = \begin{cases} 1-\alpha & , x = 0 \\ 1-\alpha+\alpha F(x,\theta), & x > 0 \end{cases}, \tag{1.1}$$

and the corresponding density function is

$$g(x,\theta,\alpha) = \begin{cases} 1-\alpha, & x = 0 \\ \alpha f(x,\theta), & x > 0 \end{cases} \tag{1.2}$$

The Above model has been in use for studying inferences about (α,θ) and has has received considerable attention particularly when X is exponential with mean θ. Some of the references are Aitchison (1955), Kleyle and Dahiya (1975), Jayade and Prasad (1990), Vannman (1991), Muralidharan (1999, 2000), Kale and Muralidharan (2000) and references therein. Vannman (1995) and Muralidharan and Kale (2002) considered the case where F is a two-parameter Gamma distribution with shape parameter β and scale parameter θ.

The family $\Im$ is further modified to accommodate early failures as $\mathcal{G}_1 = \{G_1(x,\theta,\alpha),\ x\geq 0,\ \theta\in\Omega,\ 0<\alpha<1\}$ where the df. corresponding to $G_1\in\mathcal{G}_1$ is given by

$$G_1(x,\alpha,\theta) = (1-\alpha)H(x) + \alpha F(x,\theta), \tag{1.3}$$

where H(x) is a df. with H(δ)=1 for δ sufficiently small and assumed known and specified in advance. We also assume that the early failures are recorded as a class with notional failure time δ so that the modified family G has a pdf w.r.t measure μ which is sum of Lebesgue measure on (δ,∞) and a singular measure at δ. The corresponding pdf is then given by

$$g(x,\alpha,\theta) = \begin{cases} 0, x<\delta \\ 1-\alpha+\alpha F(\delta,\theta), x=\delta \\ \alpha f(x,\theta), x>\delta \end{cases} \tag{1.4}$$

Some of the references, which treat early failure analysis with exponential distributions, are Kale and Muralidharan (2000) and Kale (2003), wherein they treat early failures as inliers using the sample configurations.

Now let us consider the following example as a natural occurrence of a physical phenomenon: 0, 0, 0, 0, 0.02, 0.03, 0.09, 1.51, 1.98, 1.20, 1.76, 2.54, 3.91 and 4.11. Here, the first four observations may be treated as instantaneous failures, next three observations may be treated as early failures (by specifying δ=0.09 or 0.1 etc) and other observations may be treated as coming from any failure time distribution F. The observations that are identified as instantaneous and early failures together are called inliers in this paper. Thus inliers in a data set are a subset of observations not necessarily all zeroes, which appears to be inconsistent with the remaining data set.

Kale and Muralidharan (2000) have first introduced the term inliers in connection with the estimation of (α,θ) of early failure model with modified failure time distribution (FTD) being an exponential distribution with mean θ assuming δ known. Using (1.4), the MLE of θ in the modified model is given by likelihood equation based on a random sample of size $(n-n_0)$ from an exponential distribution truncated to (δ,∞) and obtain $\hat{\alpha} = \frac{n-n_0}{n} e^{\delta/\hat{\theta}}$. Further, they showed that the parameter α is not orthogonal to the parameter θ. A recent paper by Muralidharan and Lathika (2004) discusses various methods of inliers model like identified inliers model and labeled slippage inliers model in particular and estimation procedures in these models. The object of this paper is to apply the information criteria suggested by Schwartz (1978) to detect the inliers. In Section 2, we formally describe the problem and in Section 3, we study the detection of inliers using the information criteria. The Section 4 will be devoted for illustration and application. The referee has suggested that the problem can be viewed as a change point problem, but in that case all observations are coming from the target population only. We have therefore, used exchangeable model as has been done by Barnett and Lewis (1984) and also Gather and Kale (1986).

2. The Problem

Let $X_1, X_2, ..., X_n$ be a sequence of independent random variables with some known FTD. As pointed out by Barnett and Lewis (1984) and also by Gather and Kale (1986) the situation can be modeled by assuming that the failure times of n units put on test are such that (n-r) of these are independently and identically distributed (i.i.d) with FTD belonging to $\Im$ characterizing target population and remaining r are i.i.d with FTD in a different class G causing inlier observations where $G\in\mathcal{G}$ and $F\in\Im$ are such that $\partial G/\partial F$ is decreasing in x. Further $X_{i_1}, X_{i_2}, ..., X_{i_r}$ are i.i.d from $G\in\mathcal{G}$ and independent of $X_{i_{r+1}}, X_{i_{r+2}}, ..., X_{i_n}$ which are i.i.d

from $F \in \Im$. The indexing set $\nu = (\nu_1, \nu_2, ..., \nu_r)$ of observations is itself a parameter varying over the parameter space given by $\binom{n}{r}$ choices of r integers out of (1,2,…,n) for fixed r known. The most general setup is defined by allowing r to vary over (n+1) values r=0,1,2,…,n. The identification of r will then decide the number of inliers in the model. We formulate the problem in the following way:

Denoting the expectation of X_i by λ_i, i=1,2,…,n, we consider the following model of no inliers in the model as Model (0) : $\lambda_i = \theta$, i=1,2,…,n (2.1)

And the model with r inliers as

Model(r) : $\lambda_i = \begin{cases} \phi, 1 \le i \le r \\ \theta, r+1 \le i \le n \end{cases}$, (2.2)

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 expectation θ. In the next section, we describe the Schwartz information criterion (SIC) to identify the value of r and then decide those observations as inliers.

3. The SIC Scheme

Suppose that the life times $X_1, X_2, ..., X_n$ is a sequence of independent random variables with exponential distribution having unknown mean θ. Our aim is to detect those information's (inliers) from the n models given by (2.2). In that case the mean life of inliers say ϕ and the mean life of the target distribution θ will be such that ϕ<<θ. The SIC given by Schwartz (1978) is defined as

SIC = $-2 \log L(\hat{\Theta}) + p \log n$,

where $L(\hat{\Theta})$ the maximum of likelihood function and p is is the number of free parameters that need to be estimated under the model. Schwartz (1978) used the above criterion for estimating the dimension of a given model. This criterion is slightly different from Akaike (1974), in which Schwartz's procedure leans more than Akaike's towards lower-dimensional models. For large number of observations the procedures differ markedly from each other.

According to the procedure, the model(0) is selected with no inliers if SIC(0) $\le \min_{1 \le r \le n-1} SIC(r)$

and the model(r) is selected if SIC(0) > $\min_{1 \le r \le n-1} SIC(r)$.

For exponential distribution, we have

SIC(0) = $2n \log \hat{\theta} + 2 \sum_{i=1}^{n} x_i / \theta + p \log n$ (3.1)

And

SIC(r) = $2r \log \hat{\phi} + 2 \sum_{i=1}^{r} x_i / \hat{\phi} + 2(n-r) \log \hat{\theta} + 2 \sum_{i=r+1}^{n} x_i / \hat{\theta} + p \log n$, (3.2)

where $\hat{\phi} = \sum_{i=1}^{r} x_i / r$ and $\hat{\theta} = \sum_{i=r+1}^{n} x_i / (n-r)$. The estimate of inliers say $\hat{r}$ is such that $SIC(\hat{r}) = \min_{1 \le r \le n-1} SIC(r)$. As discussed in section 2, this estimate of $\hat{r}$ is the number of observations coming from the inliers distribution G and the remaining observations will be from the target distribution F. Balakrishanan and Chen (2004) have used the above criterion for

detecting a change point in a sequence of extreme value observations. We now illustrate the idea on a simulated data and a real life data set.

4. Illustration

(a). Simulated data set: We have generated 15 independent random observations, where five of them are coming from exponential with mean life, ϕ=0.04 and the remaining ten observations from exponential distribution with mean life, θ=5. The observations are 0.01339, 0.02679, 0.03442, 0.05519, 0.09459, 0.32254, 0.64367, 1.19427, 3.00276, 3.14612, 3.15643, 3.94635, 5.17659, 9.79405 and 12.52736. The SIC(0) is calculated using (3.1) and is equal to 64.3963 and SIC(r) for r=1,2,…,n-1 are calculated based on (3.2) and they are as follows: 58.2884, 50.9455, 43.9152, 37.9862, 33.5080, 34.8518, 37.1259, 40.7128, 48.2497, 51.5893, 53.2503, 55.0619, 57.3022, 62.3729. Clearly

SIC(0)=64.3963>SIC(5)= $\min_{1\le r\le n-1} SIC(r) = 33.5080$. Therefore, $\hat{r}$=5 and the corresponding estimates for ϕ and θ are 0.044876 and 4.2910 respectively. Which clearly shows that this method works. Encouraged by this, we carried out an experiment with 1000 samples each of size 15 and number of inliers as 3,4, 5 and 6 each with ϕ = 0.5 and θ = 2,3,4,5,6 and 7. The following table entitled power of SIC procedure presents the number of times the SIC procedure correctly identified the number of inliers as a proportion to total number of samples.

Table 1
Power of SIC procedure

r	θ/ϕ= 4	θ/ϕ= 6	θ/ϕ= 8	θ/ϕ =10	θ/ϕ= 12
3	0.509	0.785	0.896	0.946	0.982
4	0.514	0.822	0.934	0.968	0.996
5	0.555	0.874	0.957	0.999	1.00
6	0.586	0.887	0.988	1.00	1.00

(b). Example: The example is based on Vanmann's (1995) data on drying of woods under different experiments and schedules. We reproduce Vanmann's data on Experiment 2 on one batch of 37 boards by using two different schedules.

Schedule 1. x_i =0, i=1,2,…,13 and the other positive observations arranged in increasing order of their magnitude 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. The computed value of SIC(0)= 127.1460 and the corresponding SIC(r)'s are
124.0335, 121.2333, 118.0726, 115.1593, 113.3191, 111.5499, 110.6135, 109.4866, 108.4856, 110.4540, 111.7338, 113.3368, 114.4581, 115.8601, 117.0348, 118.3385, 119.6651, 121.2188, 122.5813, 123.7872, 125.0636, 126.2976, 127.3799.

Clearly, SIC(0)=127.1460 > SIC(9)= $\min_{1\le r\le n-1} SIC(r)$ = 108.4856. Hence $\hat{r}$=9. i.e. the number of inliers including 13 instantaneous failures are 22. The remaining 15 observations are from the target populations. The corresponding estimates of the parameters are $\hat{\phi}$ = 0.7278 and $\hat{\theta}$= 7.3520.

Acknowledgements

Both the authors thank the referee and Editor for their valuable comments. The first also author thanks the Department of Science and Technology, New Delhi for the financial support in the form of granting a project (NO/DST/MS/143/2K). The second author thanks Professor M. B. Rajarshi, Head, Department of Statistics, University of Pune for providing necessary research facilities.

References

1. Abramovitz, M. and Stegun, I. A. (1965). Handbook of Mathematical Functions, General publishing Company Ltd, Canada.
2. Aitchison, J. (1955). On the distribution of a positive random variable having a discrete probability Mass at the origin, J. Amer. Stat. Assn., Vol. 50, p.901-908.
3. Akaike, H. (1974). A new look at the Statistical identification model, IEEE Trans. Auto. Control, 19, p.716-723.
4. Balakrishanan, N. and Chen, J. (2004). Detecting a change point in a sequence of extreme value observations, J. Prob. Statist. Science, 2(1), p.55-64.
5. Barnett, V. and Lewis, T. (1984). Outliers in Statistical data, John Wiley & Sons, New York.
6. Gather, U. and Kale, B.K (1986). Outlier generating models: A review, Tech. Report No. 117, Dept. of statistics and Acturial Mathematics, University of Iowa, Iowa city, Iowa.
7. Jayade, V.P. and Prasad, M.S. (1990). Estimation of parameters of mixed failure time distribution. Comm, Statist. – Theory and Methods, 19(12), p.4667-4677.
8. Kale, B.K. (2003). Modified failure time distributions to accommodate instantaneous and early failures, Industrial Mathematics and Statistics, Ed. J. C. Misra, p.623-648, Narosa Publishers, New Delhi.
9. Kale, B.K. and Muralidharan, K. (2000). Optimal estimating equations in mixture distributions accommodating instantaneous or early failures. J.Indian . Statist. Assoc., 38, p.317-329.
10. Kleyle , R.M. and Dahiya, R.L. (1975). Estimation of parameters of mixed failure time distribution from censored data, Comm. Statist. – Theory and Methods, 4(9), p.873-882.
11. Muralidharan, K. (1999). Tests for the mixing proportion in the mixture of a degene-rate and exponential distribution, J. Indian Stat. Assn., Vol. 37, issue 2, p.105-119.
12. Muralidharan, K. (2000). The UMVUE and Bayes estimate of reliability of mixedfailure time distribution., Comm. Statist- Simulations & Computations, 29(2), p. 603-619.
13. Muralidharan, K. and Kale, B.K. (2002). Modified Gamma distribution with Singularity at zero. Comm. Statist- Simulations & Computations, 31(1), p.143-158.
14. Muralidharan and Lathika (2004). The Concept of inliers. Proceedings of first Sino-International Symposium on Probability, Statistics and Quantitative Management., Taiwan, October, p. 77-92.
15. Schwartz (1978). Estimating the dimensionality of a model. The Annals of Statistics, Vol. 6(2), p. 461-464.
16. 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, Swedon.