ProbStat Forum, Volume 04, October 2011, Pages 89–97
ISSN 0974-3235

ProbStat Forum is an e-journal.
For details please visit www.probstat.org.in

Nonparametric estimation of distribution function in the presence of additional information based on two unit-parallel system

Patil P.Y.[a], Rattihalli R.N.[b], Moeng S.R.T.[c]

[a] *Department of Statistics, Shivaji University, Kolhapur, INDIA*
[b] *Department of Statistics, Central University of Rajasthan, Kishangarh, INDIA*
[c] *Department of Statistics, University of Botswana, Gaborone, BOTSWANA*

Abstract. Nonparametric inference is becoming more popular because of its wide applicability and computational facilities. Vardi [The Annals of Statistics, S10, 2, 616 – 620] has considered nonparametric estimation of the distribution function in the presence of length biased additional information. In this paper we consider a similar problem when additional information is on parallel system of two identical independent units, each having a common cumulative distribution function F. The nonparametric maximum likelihood estimator ($NPMLE$) of F, not necessarily continuous, is obtained and based on extensive simulations some of its properties are discussed.

1. Introduction

Let $X_1, X_2, \ldots, X_m$ be m independent identically distributed (iid) random variables with a common cumulative distribution function (cdf) F, not necessarily continuous. Consider the problem of estimation of F in the presence of n additional observations $Y_1, Y_2, \ldots, Y_n$. These additional observations need not have the same cdf F, but a cdf G, which is a functional of F. This type of additional information may be available in many situations, similar to the following.

A manufacturer is interested to assess the quality of the units produced, say based on the life length X. For the purpose he may conduct an experiment on m units yielding observations $X_1, X_2, \ldots, X_m$. Suppose these units are used in a system as a subsystem of two components. The service station maintains n records $Y_1, Y_2, \ldots, Y_n$ on the life times of parallel subsystems. Thus the problem of interest is to estimate F based on m iid observations having cdf F and n iid observations having cdf $G = F^2$.

In the following section we obtain the nonparametric maximum likelihood estimator ($NPMLE$) of F in the absence and the presence of ties in the combined data. Illustrative examples are also given for both these cases. $MATLAB$ programs have been developed to obtain the estimator and different norms, these can be obtained on request from the authors. In section 4, the performance of the estimators has been studied based on extensive simulations followed by conclusion section. The simulated results (Table and Graphs) are given in Appendix.

2010 *Mathematics Subject Classification.* Primary 62G10
Keywords. Nonparametric estimation, Generalized MLE, Additional information
Received: 08 May 2011; Revised: 25 June 2011; Accepted: 15 September 2011
Email addresses: pypatil_stats@rediffmail.com (Patil P.Y.), rnr5@rediffmail.com (Rattihalli R.N.), moengsrt@mopipi.ub.bw (Moeng S.R.T.)

2. The Maximum Likelihood Estimators of F

Let $\underline{X} = (X_1, X_2, \ldots, X_m)$ be m *iid* observations with *cdf* F and $\underline{Y} = (Y_1, Y_2, \ldots, Y_n)$ be n *iid* observations independent of X_i's having *cdf* F^2. Let $t_1 < t_2 < \cdots < t_h$ be h ordered observations from combined sample. Let ξ_i and η_i be multiplicity (number) of $X's$ and $Y's$ at t_i respectively, for $i = 1, 2, \ldots, h$. Let $\underline{t} = (t_1, t_2, \ldots, t_h)$, $\underline{\xi} = (\xi_1, \xi_2, \ldots, \xi_h)$ and $\underline{\eta} = (\eta_1, \eta_2, \ldots, \eta_h)$. The combined data $\underline{X} \cup \underline{Y}$ is likelihood equivalent to $(\underline{\xi}, \underline{\eta}, \underline{t})$. Of course, one may suppress either $\underline{\xi}$ or $\underline{\eta}$, but for notational convenience we shall retain both of them. The likelihood function $L(F : \underline{\xi}, \underline{\eta}, \underline{t})$ is given by

$$L(F : \underline{\xi}, \underline{\eta}, \underline{t}) = \prod_{i=1}^{h} \{dF(t_i)\}^{\xi_i} \{dG(t_i)\}^{\eta_i} = \prod_{i=1}^{h} \{dF(t_i)\}^{\xi_i} \{2F(t_i)dF(t_i)\}^{\eta_i}.$$

To find generalized $NPMLE$ of F, it is enough to find a probability distribution $\underline{p} = (p_1, p_2, \ldots, p_h)$ that maximizes $L(F : \underline{\xi}, \underline{\eta}, \underline{t})$, where $p_i = dF(t_i)$ indicates the jump of the *cdf* at t_i for $i = 1, 2, \ldots, h$. For further details one may refer to Scholz [2]. It is to be noted that $\sum_{j=1}^{h} p_j = 1$. Following Vardi [3], the above likelihood function $L(F : \underline{\xi}, \underline{\eta}, \underline{t})$ becomes

$$L(F : \underline{\xi}, \underline{\eta}, \underline{t}) = \prod_{i=1}^{h} p_i^{\xi_i} \left\{ 2 \left(\sum_{j=1}^{i} p_j \right) p_i \right\}^{\eta_i} e^{-\lambda\left(\sum_{j=1}^{h} p_j - 1\right)} = 2^n \prod_{i=1}^{h} p_i^{(\xi_i + \eta_i)} \left(\sum_{j=1}^{i} p_j \right)^{\eta_i} e^{-\lambda\left(\sum_{j=1}^{h} p_j - 1\right)} \quad (1)$$

where λ is the Lagrange's multiplier.

For convenience in subsection 2.1, we consider the case of no ties in the combined data and as its generalization in subsection 2.2 we consider the possibility of ties.

2.1. Data without ties

In this case $\xi_i + \eta_i = 1$ and $h = m + n$. Let $\eta_i = 1$ for $i = k_1, k_2, \ldots, k_n$. That is k_i's are the positions of $Y - observations$ in the ordered combined data, for $i = 1, 2, \ldots, n$. Hence, from (1) the log-likelihood function l is given by

$$l = n \log 2 + \sum_{i=1}^{h} \log p_i + \sum_{i=1}^{n} \eta_{k_i} \log \left(\sum_{j=1}^{k_i} p_j \right) - \lambda \left(\sum_{j=1}^{h} p_j - 1 \right).$$

Case 2.1.1 ($\eta_h = 0$, that is $k_n < h$): Differentiating l with respect to (w.r.t.) $p_1, p_2, \ldots, p_h, \lambda$ and equating to zero, we have the following $n + 1$ sets of equations:

$$\frac{1}{p_i} + \frac{1}{\sum_{j=1}^{k_1} p_j} + \frac{1}{\sum_{j=1}^{k_2} p_j} + \cdots + \frac{1}{\sum_{j=1}^{k_n} p_j} = \lambda, \text{ for } i = 1, 2, \ldots, k_1, \quad (A_1)$$

$$\frac{1}{p_i} + \frac{1}{\sum_{j=1}^{k_2} p_j} + \frac{1}{\sum_{j=1}^{k_3} p_j} + \cdots + \frac{1}{\sum_{j=1}^{k_n} p_j} = \lambda, \text{ for } i = k_1 + 1, k_1 + 2, \ldots, k_2, \quad (A_2)$$

$$\cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots$$

$$\frac{1}{p_i} + \frac{1}{\sum_{j=1}^{k_n} p_j} = \lambda, \text{ for } i = k_{n-1} + 1, k_{n-1} + 2, \ldots, k_n, \quad (A_n)$$

$$\frac{1}{p_i} = \lambda, \text{ for } i = k_n + 1, k_n + 2, \ldots, h, \quad (A_{n+1})$$

$$\sum_{j=1}^{h} p_j = 1.$$

Note that (A_1) is a set of k_1 equations formed by differentiating l w.r.t. $p_1, p_2, \ldots, p_{k_1}$ and in these equations only the first term changes. Hence by subtracting u^{th} equation from the v^{th} in the set of (A_1) equations, we will have $p_u = p_v$, for $1 \leq u < v \leq k_1$. Let $p_1 = p_2 = \cdots = p_{k_1} = q_1$ (say). In general, we have

$$p_{k_{j-1}+1} = p_{k_{j-1}+2} = \cdots = p_{k_j} = q_j \text{ (say)} \tag{2}$$

for $j = 1, 2, \ldots, n+1$ with the convention that $k_0 = 0$ and $k_{n+1} = h$.

By rewriting the above $(n+1)$ sets of equations in terms of $q_1, q_2, \ldots, q_{n+1}$, from the set of equations (A_1) we will have

$$\frac{1}{q_1} + \frac{1}{k_1 q_1} + \frac{1}{k_1 q_1 + (k_2 - k_1)q_2} + \cdots + \frac{1}{k_1 q_1 + (k_2 - k_1)q_2 + \cdots + (k_n - k_{n-1})q_n} = \lambda. \tag{B_1}$$

Similarly, from the set of equations (A_2), ... (A_{n+1}), we will have

$$\frac{1}{q_2} + \frac{1}{k_1 q_1 + (k_2 - k_1)q_2} + \cdots + \frac{1}{k_1 q_1 + (k_2 - k_1)q_2 + \cdots + (k_n - k_{n-1})q_n} = \lambda, \tag{B_2}$$

... ...

$$\frac{1}{q_n} + \frac{1}{k_1 q_1 + (k_2 - k_1)q_2 + \cdots + (k_n - k_{n-1})q_n} = \lambda, \tag{B_n}$$

$$\frac{1}{q_{n+1}} = \lambda. \tag{B_{n+1}}$$

By subtracting (B_2) from (B_1) we will have $\frac{1}{q_1} + \frac{1}{k_1 q_1} = \frac{1}{q_2}$. That is, $k_1 q_1 = (k_1 + 1)q_2$. By subtracting (B_3) from (B_2) we will have $\frac{1}{q_2} + \frac{1}{k_1 q_1 + (k_2 - k_1)q_2} = \frac{1}{q_3}$. That is, $(k_2 + 1)q_2 = (k_2 + 2)q_3$. In general, we have the recursive relation

$$(k_r + (r-1))q_r = (k_r + r)q_{r+1}, \text{ for } r = 1, 2, \ldots, n. \tag{3}$$

By using (3) we have $q_2 = \frac{k_1}{k_1+1} q_1$, $q_3 = \frac{k_1}{k_1+1}\frac{k_2+1}{k_2+2} q_1$, $q_4 = \frac{k_1}{k_1+1}\frac{k_2+1}{k_2+2}\frac{k_3+2}{k_3+3} q_1$ and so on. In general,

$$q_{r+1} = \frac{k_1}{k_1+1}\frac{k_2+1}{k_2+2}\cdots\frac{k_r+r-1}{k_r+r} q_1, \text{ for } r = 1, 2, \ldots, n. \tag{4}$$

However, as $\sum_{j=1}^{h} p_j = 1$, we have $k_1 q_1 + (k_2 - k_1)q_2 + \cdots + (k_n - k_{n-1})q_n + (h - k_n)q_{n+1} = 1$. That is,

$$k_1 q_1 + \sum_{r=2}^{n} (k_r - k_{r-1}) \prod_{j=1}^{r-1} \frac{(k_j + j - 1)}{(k_j + j)}\, q_1 + (h - k_n) \prod_{j=1}^{n} \frac{(k_j + j - 1)}{(k_j + j)} q_1 = 1$$

gives,

$${q_1}^{-1} = k_1 + \sum_{r=2}^{n} (k_r - k_{r-1}) \prod_{j=1}^{r-1} \frac{(k_j + j - 1)}{(k_j + j)} + (h - k_n) \prod_{j=1}^{n} \frac{(k_j + j - 1)}{(k_j + j)}\,. \tag{5}$$

Now by using (5) and (4), one can obtain $q_1, q_2, \ldots, q_{n+1}$. Hence obtain $\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_h$ from (2).

Case 2.1.2 ($\eta_h = 1$, that is $k_n = h$): In this case there shall not be the last term $(\sum_{j=1}^{k_n} p_j)^{-1}$ in the system of equations (A_1) to (A_n) and hence there shall not be q_{n+1}.

2.2. Data with ties

Let r and s be be the number of distinct $X-values$ and $Y-values$, respectively. Let $\delta_i = \xi_i + \eta_i$ for $i = 1, 2, \ldots, h$ and $\underline{\delta} = (\delta_1, \delta_2, \ldots, \delta_h)$. In this case $\delta_i \geq 1$ and $h \leq r + s \leq m + n$. Let $\eta_i \geq 1$ for $i = k_1, k_2, \ldots, k_s$ and $\eta_i = 0$ otherwise. For convention let $k_0 = 0$ and $k_{s+1} = h$. Note that, k_i's are the positions of $Y-observations$ in the ordered combined data, for $i = 1, 2, \ldots, s$. Hence, from (1) the log-likelihood function l is given by

$$l = n \log 2 + \sum_{i=1}^{h} \delta_i \log p_i + \sum_{i=1}^{s} \eta_{k_i} \log \left(\sum_{j=1}^{k_i} p_j \right) - \lambda \left(\sum_{j=1}^{h} p_j - 1 \right).$$

Case 2.2.1 ($\eta_h = 0$, that is $k_s < h$): Differentiating l w.r.t. $p_1, p_2, \ldots, p_h, \lambda$ and equating to zero, we have the following $s + 1$ sets of equations:

$$\frac{\delta_i}{p_i} + \frac{\eta_{k_1}}{\sum_{j=1}^{k_1} p_j} + \frac{\eta_{k_2}}{\sum_{j=1}^{k_2} p_j} + \cdots + \frac{\eta_{k_s}}{\sum_{j=1}^{k_s} p_j} = \lambda, \text{ for } i = 1, 2, \ldots, k_1 \qquad (C_1)$$

$$\frac{\delta_i}{p_i} + \frac{\eta_{k_2}}{\sum_{j=1}^{k_2} p_j} + \frac{\eta_{k_3}}{\sum_{j=1}^{k_3} p_j} + \cdots + \frac{\eta_{k_s}}{\sum_{j=1}^{k_s} p_j} = \lambda, \text{ for } i = k_1 + 1, k_1 + 2, \ldots, k_2, \qquad (C_2)$$

$$\cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots \cdots$$

$$\frac{\delta_i}{p_i} + \frac{\eta_{k_s}}{\sum_{j=1}^{k_s} p_j} = \lambda, \text{ for } i = k_{s-1} + 1, k_{s-1} + 2, \ldots, k_s, \qquad (C_s)$$

$$\frac{\delta_i}{p_i} = \lambda, \text{ for } i = k_s + 1, k_s + 2, \ldots, h, \qquad (C_{s+1})$$

$$\sum_{j=1}^{h} p_j = 1.$$

Note that in the above set (C_1) of k_1 equations is formed by differentiating l w.r.t. $p_1, p_2, \ldots, p_{k_1}$ and in these only the first term changes.

Let

$$\frac{1}{Q_c} = \lambda - \sum_{l=c}^{s} \left\{ \frac{\eta_{k_l}}{\sum_{j=1}^{k_l} p_j} \right\} \text{ for } c = 1, 2, \ldots, s + 1.$$

The above sets of equations (C_1), (C_2), $\ldots$, (C_{s+1}) can be rewritten as:

$$p_i = \delta_i Q_c \text{ for } i = k_{c-1} + 1, k_{c-1} + 2, \ldots, k_c \text{ and } c = 1, 2, \ldots, s + 1. \qquad (6)$$

Hence, $\sum_{j=1}^{h} p_j = 1$ implies that, $\sum_{j=1}^{k_1} \delta_j Q_1 + \sum_{j=k_1+1}^{k_2} \delta_j Q_2 + \cdots + \sum_{j=k_s+1}^{h} \delta_j Q_{s+1} = 1$. That is,

$$\sum_{r=1}^{s+1} T_r Q_r = 1, \text{ where } T_r = \sum_{j=k_{r-1}+1}^{k_r} \delta_j \text{ for } r = 1, 2, \ldots, s + 1. \qquad (7)$$

Now, from the definition of $Q_i's$, we have $\frac{1}{Q_1} + \frac{\eta_{k_1}}{T_1 Q_1} = \frac{1}{Q_2}$. That is, $Q_2 = \frac{T_1}{T_1 + \eta_{k_1}} Q_1$. Similarly, $\frac{1}{Q_2} + \frac{\eta_{k_2}}{\sum_{j=1}^{2} T_j Q_j} = \frac{1}{Q_3}$. That is, $Q_3 = \sum_{j=1}^{2} T_j Q_j + \eta_{k_1} \left(\sum_{j=1}^{2} T_j Q_j + \sum_{j=1}^{2} \eta_{k_j} \right)^{-1} Q_2$. In general, we have the following recursive relations:

$$Q_j = H_{j-1} Q_{j-1} \text{ for } j = 2, 3, \ldots, s + 1. \qquad (8)$$

where

$$H_{j-1} = \frac{\sum_{i=1}^{j-1} T_j + \sum_{i=1}^{j-2} \eta_{k_j}}{\sum_{i=1}^{j-1} T_j + \sum_{i=1}^{j-1} \eta_{k_j}} \text{ for } j = 2, 3, \ldots, s+1. \tag{9}$$

Hence,

$$Q_j = \prod_{i=1}^{j-1} H_i Q_1 \text{ for } j = 2, 3, \ldots, s+1. \tag{10}$$

From (7) and (10), we have

$$\sum_{j=1}^{s+1} T_j \prod_{i=1}^{j-1} H_i Q_1 = 1$$

with the convention that $\prod_{i=1}^{0} H_i = 1$. Hence,

$$Q_1 = \left\{ \sum_{j=1}^{s+1} T_j \prod_{i=1}^{j-1} H_i \right\}^{-1}. \tag{11}$$

Thus, $Q_j's$ can be obtained from (11) and (10) for $j = 1, 2, 3, \ldots, s+1$. Hence obtain $\hat{p}_i's$ from (6) for $i = 1, 2, 3, \ldots, h$.

Case 2.2.2 ($\eta_h > 0$, that is $k_s = h$): In this case there shall not be the last term $\eta_{k_s} \left(\sum_{j=1}^{k_s} p_j \right)^{-1}$ in the system of equations (C_1) to (C_s) and hence there shall not be Q_{s+1}.

In either of the above cases, the generalized $NPMLE$ of F is given by

$$F_{m,n}^{(M)}(t) = \sum_{t_i \le t} \hat{p}_i, \tag{12}$$

where, for $i = 1, 2, \ldots, h$; $\hat{p}_i's$ are solutions of $p_i's$ obtained as described in the respective cases.

2.3. Illustrations

In this section we illustrate methods of obtaining $NPMLE$ developed in the above sub sections. Examples 1 and 2 are for data without ties and Examples 3 and 4 are for data with ties. In the first two examples we just indicate the order of the observations while for the others we consider the exact values of the observations.

Example 1. X-observation is the largest in a combined data: Let $m = 4$, $n = 3$ ($h = 7$), x and y be the observations on X and Y respectively. Suppose the observations in the combined sample have the order $x\,y\,x\,x\,y\,y\,x$. Here $k_1 = 2$, $k_2 = 5$ and $k_3 = 6$. Then from (2) we get $p_1 = p_2 = q_1$; $p_3 = p_4 = p_5 = q_2$; $p_6 = q_3$ and $p_7 = q_4$. From (4) we have $q_2 = \frac{k_1}{k_1+1} q_1 = \frac{2}{3} q_1$; $q_3 = \frac{k_1}{k_1+1} \frac{k_2+1}{k_2+2} q_1 = \frac{4}{7} q_1$; $q_4 = \frac{k_1}{k_1+1} \frac{k_2+1}{k_2+2} \frac{k_3+2}{k_3+3} q_1 = \frac{32}{63} q_1$. From (5) we have $q_1 = \frac{63}{320}$. Thus $\hat{p}_1 = \hat{p}_2 = \frac{63}{320}$; $\hat{p}_3 = \hat{p}_4 = \hat{p}_5 = \frac{42}{320}$; $\hat{p}_6 = \frac{36}{320}$; $\hat{p}_6 = \frac{32}{320}$. Now for the specified values of x and y , the generalized $NPMLE$ of F can be obtained from (12).

Example 2. Y-observation is the largest in a combined data : Let $m = 4$, $n = 3$ ($h = 7$) and the observations in the combined sample have the order $x\ y\ x\ y\ x\ x\ y$. Hence $k_1 = 2$, $k_2 = 4$ and $k_3 = 7$. Then from (2) $p_1 = p_2 = q_1$; $p_3 = p_4 = q_2$; $p_5 = p_6 = p_7 = q_3$. From (4) we have $q_2 = \frac{k_1}{k_1+1} q_1 = \frac{2}{3} q_1$; $q_3 = \frac{k_1}{k_1+1} \frac{k_2+1}{k_2+2} q_1 = \frac{5}{9} q_1$. From (5) we have $q_1 = \frac{1}{5}$. Thus $\hat{p}_1 = \hat{p}_2 = \frac{1}{5}$; $\hat{p}_3 = \hat{p}_4 = \frac{2}{15}$; $\hat{p}_5 = \hat{p}_6 = \hat{p}_7 = \frac{1}{9}$. For the specified values of x and y, the generalized $NPMLE$ of F is given by (12).

Example 3. Y-observation is not the largest in a combined data: Let $\underline{X} = (0.3, 0.5, 0.7, 1.3, 1.5, 1.5, 2.5, 3.5, 3.5, 4.0, 6.0)$, $\underline{Y} = (0.3, 0.35, 0.5, 0.5,\ 0.9, 1.9, 5.0, 5.0)$. Distinct observations in a combined sample are

$\underline{t} = (0.3, 0.35, 0.5, 0.7, 0.9, 1.3, 1.5, 1.9, 2.5, 3.5, 4.0, 5.0, 6.0)$. Observe that Y-observation is not the largest in a combined data that is, $\eta_h = 0$.

Here, $m = 11, n = 8, r = 9, s = 6, h = 13$, $\underline{\xi} = (1, 0, 1, 1, 0, 1, 2, 0, 1, 2, 1, 0, 1)$, $\underline{\eta} = (1, 1, 2, 0, 1, 0, 0, 1, 0, 0, 0, 2, 0)$, and $\underline{\delta} = (2, 1, 3, 1, 1, 1, 2, 1, 1, 2, 1, 2, 1)$, $\eta_i > 0$ for $i = 1, 2, 3, 5, 8, 12$ and $\eta_i = 0$ otherwise. Hence $k_1 = 1, k_2 = 2$, $k_3 = 3, k_4 = 5, k_5 = 8, k_6 = 12$. From (7), $T_1 = 2, T_2 = 1, T_3 = 3, T_4 = 2, T_5 = 4, T_6 = 6, T_7 = 1$. From (9), $H_1 = 2/3$, $H_2 = 4/5$, $H_3 = 4/5$, $H_4 = 12/13$, $H_5 = 17/18$ and $H_6 = 12/13$. Hence, from (11) $Q_1 = 4225/39168$ and from (8), $Q_2 = 4225/58752$, $Q_3 = 845/1468$, $Q_4 = 169/3672$, $Q_5 = 13/306$, $Q_6 = 221/5508$, $Q_7 = 1/27$. From (6), we have, $\hat{p}_1 = 4225/19584, \hat{p}_2 = 4225/58752, \hat{p}_3 = 845/4896, \hat{p}_4 = 169/3672, \hat{p}_5 = 169/3672, \hat{p}_6 = 13/306, \hat{p}_7 = 13/153, \hat{p}_8 = 13/306, \hat{p}_9 = 221/5508, \hat{p}_{10} = 221/2754, \hat{p}_{11} = 221/5508, \hat{p}_{12} = 221/2754$, and $\hat{p}_{13} = 1/27$. Using these $\hat{p}_i's$ the generalized $NPMLE$ of F can be obtained from (12).

Example 4. Y-observation is the largest in a combined data: Let $\underline{X} = (0.5, 1.3, 1.3, 2.0, 2.5, 2.5, 2.5, 3.0, 3.0)$, $\underline{Y} = (0.7, 1.5, 1.5, 2.5, 2.5, 3.5, 3.5, 3.5)$. Distinct observations in a combined sample are $\underline{t} = (0.5, 0.7, 1.3, 1.5, 2.0, 2.5, 3.0, 3.5)$. Observe that $Y - observation$ is largest in a combined data, that is $\eta_h > 0$.

Here, $m = 9, n = 8, r = 5, s = 4, h = 8$, $\underline{\xi} = (1, 0, 2, 0, 1, 3, 2, 0)$, $\underline{\eta} = (0, 1, 0, 2, 0, 2, 0, 3)$ and $\underline{\delta} = (1, 1, 2, 2, 1, 5, 2, 3)$, $\eta_i > 0$ for $i = 2, 4, 6, 8$ and $\eta_i = 0$ otherwise. Hence $k_1 = 2, k_2 = 4, k_3 = 6, k_4 = 8$. From (7), $T_1 = 2, T_2 = 4, T_3 = 6, T_4 = 5$. From (9), $H_1 = 2/3$, $H_2 = 7/9$, $H_3 = 15/17$, $H_4 = 22/25$. Hence, from (11) $Q_1 = 153/1540$ and from (8), $Q_2 = 51/770$, $Q_3 = 119/2310$, $Q_4 = 1/22$. From (6), we have, $\hat{p}_1 = 153/1540$, $\hat{p}_2 = 153/1540$, $\hat{p}_3 = 51/385$, $\hat{p}_4 = 51/385$, $\hat{p}_5 = 119/2310$, $\hat{p}_6 = 119/462$, $\hat{p}_7 = 1/11$, $\hat{p}_8 = 3/22$. Using these $\hat{p}_i's$ the generalized $NPMLE$ of F can be obtained from (12).

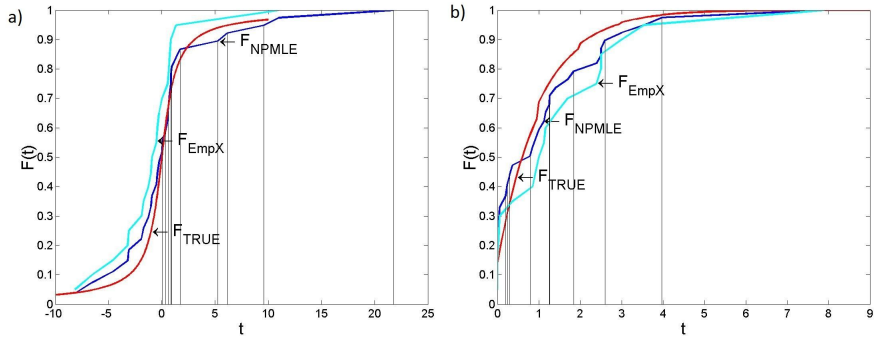

Figure 1: (a) When true F is $C(0, 1)$. (b) When true F is given by (13). F_{TRUE} : True *cdf*; F_{NPMLE} : $NPMLE$ given by (12) and F_{EmpX} : Empirical *cdf* based on *X-observations* only.

3. Simulation Study

In this section we carry out simulations to study the role of the additional information and the convergence behavior of the estimator. Data without ties case is simulated by considering F as standard Cauchy distribution ($C(0, 1)$) functions. Data with ties case is simulated by considering following F :

$$F(x) = 0.8\left(1 - e^{-x}\right) + 0.2\left(0.7\sum_{y=0}^{[x]} 0.3^y\right), \qquad \text{if} \qquad 0 \le x < \infty, \tag{13}$$

a mixture of standard exponential and geometric distribution with $p = 0.7$. We evaluate the performance of the proposed estimator given by (12), for different values of m and n. For each (m, n), 1000 simulated samples are considered. In Appendix I the values of the simulated norms between true F and its $NPMLE$ are tabulated together with the corresponding three dimensional graphs and their contours.

Figure 1(a) and Figure 1(b) are based on random samples of sizes $m = 20$ and $n = 10$. The proposed estimator given by (12) is closer to the true *cdf* as compared to the empirical based only on $X-$observations which is given by $F_m(t) = \sum_{i=1}^{m} I_{[X_i \leq t]}$, where I_A is an indicator function.

From Table 1 and Figures 2 and 3, it is evident that the $NPMLE$ tends to the true distribution function as (m, n) increases. Though convergence of $NPMLE$ in m is faster than the convergence in n, by using the additional information from the n observations we get better estimator.

The three dimensional 3D plots of the simulated Sup norm, L_1 norm and L_2 norm and their respective contour plots for the Cauchy distributions are shown in Figure 2. Similar plots for mixture of distributions (13) are shown by Figure 3.

4. Conclusion

In this paper we have obtained generalized maximum likelihood estimator of the distribution function F based on m random variables having *cdf* F and n additional observations from F^2. The estimator being maximum likelihood, it possesses desirable statistical properties like consistency, asymptotic unbiasedness, etc. By considering the Sup, L_1 and L_2 norms and using extensive simulations we have studied the impact of sample sizes on the estimator in both the cases - data without ties and data with ties.

Appendix I

Table 1: Simulated Norms between $NPMLE$ and true *cdf* for different values of sample sizes (m, n). Simulation size is equal to 1000.

Sample Size from F	**Sample Size from G**	**True Distribution**					
		$C(0,1)$			Mixture given by (13)		
m	n	$Sup-Norm$	L_1-Norm	L_2-Norm	$Sup-Norm$	L_1-Norm	L_2-Norm
10	10	0.182328	0.992403	0.082706	0.177112	0.233927	0.021778
10	20	0.151484	0.841741	0.058263	0.146982	0.188603	0.014011
10	30	0.134779	0.754840	0.047609	0.127621	0.161099	0.010047
10	40	0.122947	0.694246	0.040366	0.114543	0.143258	0.007952
10	50	0.113558	0.648138	0.035350	0.105508	0.130604	0.006491
20	10	0.147843	0.792732	0.054235	0.140721	0.196792	0.014580
20	20	0.128372	0.698536	0.041763	0.122859	0.167500	0.010554
20	30	0.116469	0.638206	0.035094	0.110252	0.147704	0.008108
20	40	0.107144	0.592818	0.030263	0.100927	0.133055	0.006590
20	50	0.100804	0.559910	0.027079	0.094099	0.123064	0.005564
30	10	0.128707	0.692245	0.041738	0.121585	0.174449	0.011267
30	20	0.114224	0.624466	0.033219	0.108590	0.152220	0.008580
30	30	0.105217	0.575528	0.028486	0.099427	0.136382	0.006838
30	40	0.097722	0.537413	0.024812	0.092106	0.124525	0.005716
30	50	0.092556	0.510504	0.022390	0.086390	0.116148	0.004912
40	10	0.114683	0.621877	0.033556	0.109148	0.158130	0.009057
40	20	0.103789	0.572591	0.027922	0.099309	0.140770	0.007232
40	30	0.096578	0.532785	0.024364	0.092123	0.128186	0.005979
40	40	0.090734	0.500413	0.021509	0.085987	0.118401	0.005128
40	50	0.086244	0.477652	0.019600	0.081419	0.111175	0.004483
50	10	0.104742	0.568222	0.027834	0.099012	0.145317	0.007575
50	20	0.096387	0.528001	0.023879	0.090947	0.131731	0.006277
50	30	0.090497	0.497184	0.021337	0.085635	0.120959	0.005274
50	40	0.085467	0.469176	0.019022	0.080614	0.112413	0.004592
50	50	0.081445	0.449827	0.017488	0.076709	0.106284	0.004071

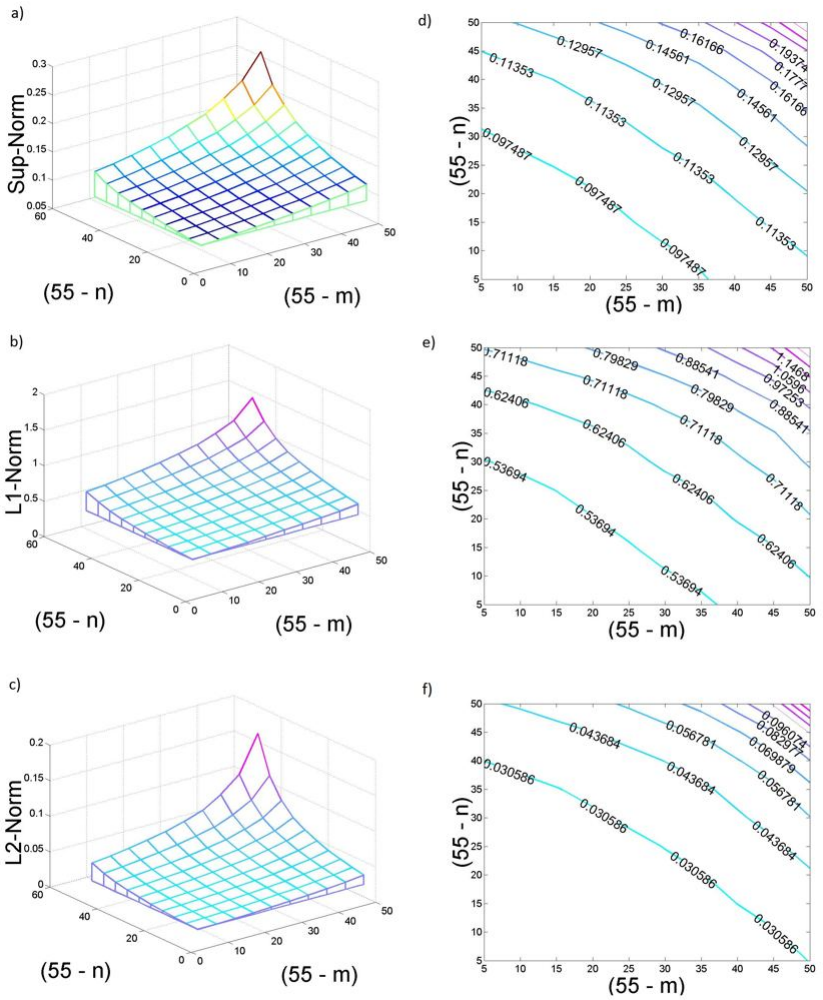

Figure 2: (a), (b), (c) 3-D Plots of [(50-m), (50-n), norms]. Simulations are from $C(0,1)$. (d), (e), (f) Contour Plot of [(50-m), (50-n)], norms]. Simulations are from $C(0,1)$.

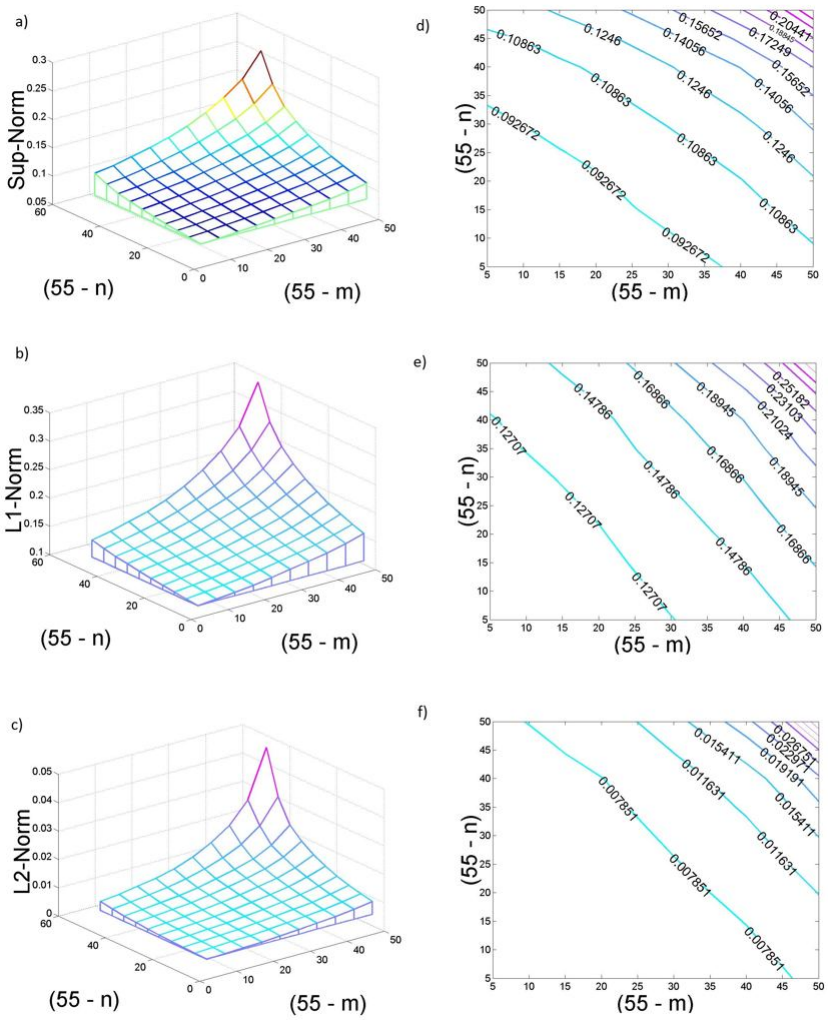

Figure 3: (a), (b), (c) 3-D Plots of [(50-m), (50-n), norms]. Simulations are from (13). (d), (e), (f) Contour Plot of [(50-m), (50-n)], norms]. Simulations are from (13).

References

[1] J. D. Gibbons, S. Chakraborti (2003). Nonparametric Statistical Inference, Marcel Dekker, Inc.
[2] F. W. Scholz (1980). Towards a unified definitition of maximum likelihood, Canadian Journal of Statistics, 8, 193–203.
[3] Y. Vardi (1982). Nonparametric Estimation in the presence of length bias, The Annals of Statistics, S10, 2, 616 – 620.