The Annals of Statistics
1980, Vol. 8, No. 1, 226–229

A NOTE ON EMPIRICAL BAYES ESTIMATION OF A DISTRIBUTION FUNCTION BASED ON CENSORED DATA[1]

BY E. G. PHADIA

University of California, Davis and The William Paterson College of New Jersey

Susarla and Van Ryzin exhibited an empirical Bayes estimator of a distribution function F based on randomly right-censored observations. In a later paper they obtained a different estimator which alleviates the weaknesses of their earlier estimator and showed that it is asymptotically optimal with rate of convergence n^{-1}. The purpose of this note is to present a slightly different estimator which is simpler and is also asymptotically optimal with the same rate of convergence. Their numerical example is reworked to show that the estimator is a proper distribution function.

1. Introduction. In their recent paper, Susarla and Van Ryzin [2] considered the empirical Bayes approach and obtained estimators of a distribution function (df) based on randomly right-censored data. Unfortunately, their estimators have certain weaknesses, foremost among them being that they are not monotonic (see their remark at the end of their example in the Appendix). In a later paper [4], while avoiding these weaknesses, they obtained an estimator which is asymptotically optimal with rate of convergence n^{-1} when the censoring random variables are independent but not identically distributed. About the same time, the author obtained an estimator which is slightly different from theirs, looks simpler, and is also free from the weaknesses their earlier estimators suffered. The purpose of this note is to report this estimator, show that it is also asymptotically optimal with the same rate of convergence n^{-1}, and rework their numerical example.

In Section 2 we first introduce the notation and define our estimator. Since this note is directly related to [2], we shall mostly adopt their notation. For the sake of brevity many of the details will be omitted and the reader is referred to their papers for further details. The asymptotic optimality is discussed in Section 3. In Section 4, we rework their example to show that the monotonicity of the estimator need not be compromised.

2. Notation and formulation of the problem. Let (F_i, X_i), $i = 1, 2, \cdots$ be a sequence of independent random variables. The F_i's are distributed according to the Dirichlet process with parameter α, to be denoted by $\mathcal{D}(\alpha)$. (Here α is a finite nonnull measure defined on the Borel σ-field $\mathcal{B}$ of R the real line). Given $F_i = F$,

Received July 1978; revised August 1979.

[1]Research supported partially by a grant from the National Science Foundation MCS 77-01653.

AMS 1970 *subject classifications.* Primary 62C99; secondary 62G05.

Key words and phrases. Empirical Bayes estimation, right-censored observations, Dirichlet process priors, nonparametric estimation of a distribution function, asymptotic optimality.

X_i is distributed according to F. Further, let Y_i, $i = 1, 2, \cdots$ be a sequence of independent random variables (censoring variables), independent of (F_i, X_i) and each distributed according to a continuous df H (censoring distribution). (Here we slightly differ from Susarla and Van Ryzin since they do not consider the Y_i's to be identically distributed. However, we feel that in practice the assumption of a common censoring distribution is reasonable. Moreover, it is possible to extend this result to the nonidentical case with a little more complexity). We take both F and H to be right-sided df's, i.e., $H(u) = P(Y_1 > u)$ and $F(u) = P(X_i > u | F_i = F)$.

With this set up, the problem is to estimate the survival fnction $F(u)$, based on the observable random variables δ_i and Z_i where $\delta_i = I[X_i \leqslant Y_i]$ and $Z_i = X_i \wedge Y_i$, $i = 1, 2, \cdots$. Here $I[A]$ denotes the indicator function of set A. We note that $Z_1, Z_2, \cdots$ are independent and that $P(Z_i > u) = F(u)H(u)$ for all i's. The usual integrated squared error loss function $L(F, \hat{F})$ given by

$$L(F, \hat{F}) = \int (F(u) - \hat{F}(u))^2 \, dW(u) \tag{2.1}$$

is also considered here, where W is some known weight function on R.

The following empirical Bayes estimator (at the $(n+1)$th stage) is proposed:

$$\begin{aligned}[1 + \alpha(R)]S_n(u) &= [1 + \alpha(R)]\hat{F}_{n+1,\hat{\alpha}}(u) \\ &= I[Z_{n+1} > u] + \hat{\alpha}(u, \infty) + I[\delta_{n+1} = 0, Z_{n+1} \leqslant u]\frac{\hat{\alpha}(u,\infty)}{\hat{\alpha}[Z_{n+1},\infty)}\end{aligned} \tag{2.2}$$

where

$$\frac{\hat{\alpha}(u,\infty)}{\alpha(R)} = \frac{N^+(u)}{n}\prod_{i=1}^{n}\left(\frac{N^+(Z_i) + 1 + c}{N^+(Z_i) + c}\right)^{I[\delta_i = 0,\, Z_i < u]} \tag{2.3}$$

and $N^+(u)$ = number of Z_i's $> u$, $i = 1, 2, \cdots, n$, and c a positive constant. We shall hereafter denote $\hat{\alpha}(u,\infty) = \hat{\alpha}(u)$ and $\alpha(u,\infty) = \alpha(u)$. The estimator differs from the Susarla and Van Ryzin estimators in the estimate of α.

Observe that this estimator is a proper right-sided distribution function, and is independent of the censoring distribution. The estimator of α involves a product and very much looks like the Bayes estimator or the Kaplan and Meier maximum likelihood estimator. The role of the constant c in this estimate may be viewed as follows. When made arbitrarily large, it will tend to minimize the influence of censored observations and $\hat{\alpha}(u)$ will behave more like the sample distribution function of Z_i for $i = 1, 2, \cdots, n$. On the other hand, if c is made arbitrarily small $\hat{\alpha}(u)$ will tend to be close to the product-limit estimator of Kaplan and Meier [1]. By a suitable choice of c, a desired degree of smoothness in the empirical Bayes estimator S_n may be obtained.

3. Asymptotic optimality of $S_n(u)$. Under the same assumptions as in [2] and [3], namely, $\alpha(R)$ is known and for any fixed but arbitrary u, $\alpha(u) > 0$, $H(u) > 0$, it

can readily be verified that

$$E|\hat{\alpha}(u) - \alpha(u)|^2 = 0(n^{-1})$$

by following their steps (Theorem 2.2 [3]).

The asymptotic optimality with rate of convergence $0(n^{-1})$ can now easily be verified by following Susarla and Van Ryzin's approach (Theorem 3 [2] or Theorem 5 [4]). The details of the proof are omitted here, but are available with the author.

4. Example. As an example, Susarla and Van Ryzin [2] considered the data of survival times (in weeks) of 81 patients from a melonoma study conducted by the Central Oncology Group of the University of Wisconsin, Madison. They used $n = 80$, $z_{81} = 16^+$ where the plus sign indicates that the 81st observation was censored. Further they used a negative exponential distribution with parameter β as the censoring distribution, where β was estimated from the data, and obtained the empirical Bayes estimator of the survival function by using their formula for the case of known censoring distribution. They did not work out their example for the case of unknown censoring distribution. The problem does not arise in our case since our estimator does not depend upon the censoring distribution.

The estimator (2.2) for this example with $\alpha(R) = 1$ reduces to

$$(4.1) \qquad S_n(u) = \hat{F}_{n+1,\hat{\alpha}}(u) = \tfrac{1}{2}(1 + \hat{\alpha}(u, \infty)) \qquad \text{for } u < z_{81} = 16,$$

$$= \frac{\hat{\alpha}(u,\infty)}{2}\left(1 + \frac{1}{\hat{\alpha}[z_{81},\infty)}\right) \qquad u \geqslant z_{81} = 16,$$

where $\hat{\alpha}(u, \infty)$ is computed using the formula (2.3) with $c = 1$ and $\hat{\alpha}[z_{81}, \infty) = \frac{78}{80}$. In the table we give the values of $S_n(u)$ evaluated at several values of u along with Susarla and Van Ryzin's (S-V) estimator.

u	<13	$14 \leqslant u < 16$	16	20	40	60	80	100	120	140	160	180	200	220	233	>234
S-V estimator	1	1	1	.980	.835	.745	.681	.615	.529	.326	.180	.198	.093	.034	.036	0
Our estimator	1	.988	.988	.962	.744	.650	.593	.552	.479	.380	.302	.302	.259	.173	.173	0

Several comments on this estimate $S_n(u)$ are in order. First, note that our estimator is nonincreasing, unlike theirs. Second, unlike their estimator, our estimator remains constant between two values of u (for e.g., 160–180 and 220–233) in the absence of any observations in-between. Third, their estimator when u is around 220 is close to zero which is enforced by taking a negative exponential censoring distribution, whereas this estimator still gives higher values for the estimator. This may be explained by the fact that 34 of 80 observations (42.5%) are

censored and six out of seven beyond $u = 180$! However, if further smoothening is desired toward the tail end, this can be achieved by taking the value of c somewhat larger. The choice of c in this example was arbitrary and was set to be equal to one for simplicity. It appears in the estimator S_n via the estimator $\hat{\alpha}$ of the parameter α of the prior distribution for F. Any modest amount of variation in c will not drastically affect the estimator S_n, especially if we have more than one observation at the $(n + 1)$th stage. But it seems difficult to specify the exact value. A value of $\frac{1}{2}$ to 5 seems reasonable. However, we feel that this should be left to the discretion of the user who might combine his own intuition and past experience in selecting a suitable value of c. The choice of c may also depend on some optimality criterion.

REFERENCES

[1] KAPLAN, E. L., AND MEIER, P. (1958). Nonparametric estimation from incomplete observations, *J. Amer. Statist. Assoc.* **53** 457–481.

[2] SUSARLA, V. AND VAN RYZIN, J. (1978a). Empirical Bayes estimation of a distribution (survival) function from right censored observations. *Ann. Statist.* **6** 740–754.

[3] SUSARLA, V. AND VAN RYZIN, J. (1978b). Large sample theory for a Bayesian nonparametric survival curve estimator based on censored samples. *Ann. Statist.* **6** 755–768.

[4] SUSARLA, V. AND VAN RYZIN, J. (1979). Large sample theory for survival curve estimation under variable censoring. In *Optimization Theory in Statistics*. (J. S. Rustagi, ed.), pages 475–508. Academic Press, New York.

DEPARTMENT OF MATHEMATICS
WILLIAM PATERSON COLLEGE OF NEW JERSEY
300 POMPTON ROAD
WAYNE, NEW JERSEY 07470