Ann. Inst. Statist. Math.
Vol. 49, No. 2, 231–235 (1997)

A NOTE ON THE BEST INVARIANT ESTIMATOR OF A DISTRIBUTION FUNCTION UNDER THE KOLMOGOROV-SMIRNOV LOSS*

ZHIQIANG CHEN AND ESWAR PHADIA

Department of Mathematics, William Paterson College of New Jersey, Wayne, NJ 07470, U.S.A.

(Received February 23, 1996; revised May 13, 1996)

Abstract. For the invariant decision problem of estimating a continuous distribution function with the Kolmogorov-Smirnov loss within the class of 'proper' distribution functions, it is proved that the sample distribution function is the best invariant estimator only for the sample size $n = 1$ and 2. Further it is shown that the best invariant estimator is minimax. Exact jumps of the best invariant estimator are derived for $n \leq 4$.

Key words and phrases: Best invariant estimator, Kolmogorov-Smirnov loss, minimaxity.

1. Introduction

The best invariant estimators for a continuous cumulative distribution function F defined on R^1 under monotone transformations and the weighted Cramer-von Mises loss function were introduced by Aggarwal (1955). Since then there had been a longstanding conjecture that the best invariant estimator, d_0 is minimax for $n \geq 1$ under the loss

$$L(F,a) = \int |F(t) - a(t)|^k h(F(t))dF(t), \tag{1.1}$$

where k is a positive integer, $h(t)$ is a nonnegative weight function and $a(t)$ is a nondecreasing function from $(-\infty, \infty)$ into $[0,1]$ (see, for example, Ferguson (1967)). This conjecture was proved in Yu (1992) and Yu and Chow (1991).

A parallel problem was to consider the Kolmogorov-Smirnov loss function,

$$L(F,a) = \sup_t\{|F(t) - a(t)|\}, \tag{1.2}$$

* Partially supported by the Center For Research, School of Science and Health, William Paterson College of New Jersey Grant.

which is also invariant under the monotone transformations. This loss function is difficult to handle analytically and therefore not much was accomplished for a long time. Brown (1988) obtained the best invariant estimator under this loss for the sample size $n = 1$ by hand and investigated its admissibility under the assumption that the unknown distribution function is discrete. This was followed up by Friedman *et al.* (1988) who obtained the best invariant estimator for sample sizes $n > 1$ and proved its uniqueness. Again, the obvious question is whether the best invariant estimator under this loss is minimax. This question was answered affirmatively in Yu and Phadia (1992).

For a continuous distribution function, it should be noted that all invariant estimators under monotone transformations and either the von Mises type or Kolmogorov-Smirnov type loss functions are of step function form. Furthermore, except for a particular loss function ($k = 2$ and $h(t) = t^{-1}(1-t)^{-1}$) considered by Aggarwal when the sample distribution function is the best invariant estimator, none of these best invariant estimators are proper distributions. Some natural questions that arise are: If we restrict the action space to 'proper' (to be defined below) distribution functions, does there exist a unique best invariant estimator? Will it be the sample distribution function? Will it be a minimax estimator? We consider these questions in this note and provide affirmative answer to the first and third question and a negative answer to the second question except for the trivial case of $n = 1$ and 2. Our proofs heavily depend upon the two papers cited above, viz. Friedman *et al.* (1988) and Yu and Phadia (1992).

For any ordered sample $\{X_i\}$ of size n from F, all invariant estimators are of the form

$$d(t) = \sum_{i=1}^{n+1} u_i I(x_{i-1} \le t \le x_i) \tag{1.3}$$

where $x_0 = -\infty$, $x_{n+1} = +\infty$, $0 \le u_1 \le u_2 \le \cdots \le u_{n+1} \le 1$ are constants and $I(A)$ is the indicator function of the set A. Under our restrictive setting that an estimator has to be a proper distribution, we have $u_1 = 0$ and $u_{n+1} = 1$ for the invariant class of estimators. Thus it is natural to take the action space of "proper" distributions as

$$A = \{a(t) : a(t) = 0 \text{ for } t < X_1, 1 \text{ for } t \ge X_n\}, \tag{1.4}$$

where $a(t)$ is a measurable function of the order statistics X_i.

2. Main result

PROPOSITION 2.1. *The best invariant estimator for a continuous distribution function under the Kolmorov-Smirnov loss, monotone transformations and action space A uniquely exists. It is symmetric in the sense that* $u_i = 1 - u_{n+2-i}$ *for all* $i = 1, 2, \ldots, [\frac{n+1}{2}]$ *and satisfies the partial derivative equations*

$$\frac{\partial E[L(F,d)]}{\partial u_i} = 2(Vol(u_i - x_{i-1} = \max_j l_j) - Vol(x_i - u_i = \max_j l_j)) = 0, \tag{2.1}$$

where $l_j = u_j - x_{j-1}$ *or* $x_j - u_j$, $j = 1, 2, \ldots, 2n+2$; $Vol(x_i - u_i = \max_j l_j) = \int I(x_i - u_i = \max_j l_j) I(0 \le x_1, \ldots, x_n \le 1) \prod_1^n dx_i$ *and* x_i *are order statistics from the uniform distribution on* $[0,1]$, $x_0 = 0$, $x_{n+1} = 1$.

PROOF. Notice that the set of all invariant estimators in A is closed under convex combination operation. The proof of this proposition is now the same as in Friedman *et al.* (1988) and will not be repeated here.

When $n = 1$ or $n = 2$, the restriction that the best invariant estimator has to be within A obviously yields the sample distribution function as the best invariant estimator. However, in general it is not so as the following proposition shows.

PROPOSITION 2.2. *For* $n = 1$ *and* 2, *the sample distribution function is the best invariant estimator of* F *under the Kolmogorov-Smirnov loss, monotone transformations and action space* A. *For* $n \ge 3$, *the sample distribution is not the best invariant estimator.*

PROOF. For $n = 1$ and 2, by symmetry, it is easy to check that the best invariant estimator has to be the sample distribution. But when $n \ge 3$, if d is the best invariant estimator, then, $\frac{\partial R(F,d)}{\partial u_i}$ has to be zero for all $i = 1, 2, \ldots, n$. In particular it should be so for $i = 2$, i.e., $\frac{\partial R(F,d)}{\partial u_2} = 0$, or $Vol\{u_2 - X_1 = \max_j l_j\} = Vol\{X_2 - u_2 = \max_j l_j\}$. To show that the sample distribution function is not the best, we only need to show that for the sample distribution function with $u_i = \frac{i-1}{n}$, the above equality is not satisfied. Since $l_i = u_i - x_{i-1}$ or $x_i - u_i$, the $Vol\{u_2 - X_1 = \max_j l_j\}$ in this case can be computed as follows. $1/n - x_1 = \max_j l_j$ implies that for each i, $i = 1, \ldots, n$, $\frac{1}{n} - x_1 \ge \frac{i}{n} - x_i$ and $\frac{1}{n} - x_1 \ge x_i - \frac{i-1}{n}$. Therefore, $0 \le x_1 \le \frac{1}{2n}$, and $\frac{i-1}{n} + x_1 \le x_i \le \frac{i}{n} - x_1$ for $i = 1, 2, \ldots, n$. All these regions for x_i do not overlap, so $Vol\{1/n - x_1 - \max_j l_j\} = \int_0^{1/2n} (\frac{1}{n} - 2x_1)^{n-1} dx_1 = \frac{1}{2n^{n+1}}$.

On the other hand, for $Vol\{X_2 - \frac{1}{n} = \max_j l_j\}$, $X_2 - \frac{1}{n} = \max l_j$ implies that $\frac{i+1}{n} - x_2 \le x_i \le x_2 + \frac{i-2}{n}$ for $i \ne 2$ and $x_2 \ge \frac{3}{2n}$. It is clear that $Vol\{X_2 - \frac{1}{n} = \max_j l_j\}$ is strictly greater than the volume computed under the restriction that $\frac{3}{2n} \le x_2 \le \frac{2}{n}$. In the latter case, all x_i and x_j have no overlapping regions and therefore the volume can be computed easily as $\int_{3/2n}^{2/n} (2x_2 - \frac{3}{n})^{n-1} dx_2 = \frac{1}{2n^{n+1}}$. The proof is completed.

For $n = 3$ and 4, in view of the symmetry, we need to determine only one coefficient for the best invariant estimator. The computation of this coefficient can briefly describe as follows. For $n = 3$, 4 we need to find a u_2 such that $Vol\{u_2 - X_1 = \max_j l_j\} = Vol\{X_2 - u_2 = \max_j l_j\}$. The proof of Proposition 2.2 shows that $Vol\{X_2 - u_2 = \max_j l_j\} > Vol\{u_2 - x_1 = \max_j l_j\}$ for $u_2 = 1/n$, $n \ge 3$. As we increase u_2, the first volume decreases whereas the second increases. The two volumes should be equal in order to achieve the best invariant estimator. This suggests that for $n \ge 3$, $u_2 > 1/n$. Under this constrain, for $n = 3$, routine but tedious computation leads to $Vol\{u_2 - X_1 = \max_j l_j\} = 1/12 - 3u_2/4 + 2u_2^2 - 4u_2^3/3$ and $Vol\{X_2 - u_2 = \max_j l_j\} = 1/12 - 3u_2^2/4 + 5u_2^3/6$. Equating these two volumes and solving, we get $u_2 = (33 - 3\sqrt{17})/52 = 0.396744$, which is the answer because of the uniqueness of the best invariant estimator. So the coefficients u_i in (1.3) for $n = 3$ are 0, 0.39674, 0.60326 and 1 (compared to 0.2441, 0.4013, 0.5987, 0.7559 in unrestricted case in Friedman *et al.* (1988)). Similarly, for $n = 4$, the suggestion for u_2 is that it is greater than 1/4, and we get two 4th order polynomials for the corresponding volumes: $u^4/8 - u^3/2 + 9u^2/16 - 17u/96 + 13/768$ and $-301u^4/96 + 107u^3/24 - 37u^2/16 + 47u/96 - 23/768$. An admissible numerical answer (using Mathematica) is $u_2 = 0.324424$. Therefore, for $n = 4$, the coefficients in (1.3) are 0, 0.324424, 0.5, 0.675576 and 1 (compared to 0.2072, 0.3366, 0.5, 0.6634, 0.7928 in unrestricted case (Friedman *et al.* (1988))). The weights assigned to each ordered observation for the best invariant estimator when restricted to a proper distribution are, for $n = 3$ and $n = 4$ respectively, 0.39674, 0.20652, 0.39674; and 0.324424, 0.175576, 0.175576, 0.324424. In both cases, outer observations receiving heavier weights than the inner observations.

As in the unrestricted case in Friedman *et al.* (1988), we have been unable to get an iterative formula to compute the coefficients of the best invariant estimator. However, the same way of computing coefficients as in Friedman *et al.* (1988) must work for $n \geq 5$, but it will not be pursued here.

PROPOSITION 2.3. *The best invariant estimator is a minimax estimator among all estimators in the action space* A.

PROOF. Yu and Chow (1991) showed that for any $a(X,t) \in A$ and any positive ϵ, δ, there is a distribution function F and an (unrestricted) invariant estimator d_1 such that

$$(dF)^{n+1}(\{X_1, \ldots, X_n, t) : |a(X,t) - d_1(X,t)| \geq \epsilon\}) \leq \delta$$

where dF denotes the probability measure induced by the distribution function F. If we simply change the first and last coefficients in above d_1 to 0 and 1 correspondently, and call the resulting invariant estimator d_2, the above property is still true by the virtue of the definition of $a(X,t)$. Therefore, for any $a(X,t) \in A$, and for any $\epsilon > 0$, there is a distribution F and an (restrictive) invariant estimator d_2 such that

$$(dF)^{n+1}(\{X_1, \ldots, X_n, t) : |a(X,t) - d_2(X,t)| \geq \epsilon\}) \leq \epsilon.$$

Now the minimaxity can be concluded as in Yu and Phadia (1992).

REFERENCES

Aggarwal, O. P. (1955). Some minimax invariant procedures of estimating a cumulative distribution function, *Ann. Math. Statist.*, **26**, 450–462.

Brown, L. D. (1988). Admissibility in discrete and continuous invariant nonparametric problems and in their multivariate analogs, *Ann. Statist.*, **16**, 1567–1593.

Ferguson, T. S. (1967). *Mathematical Statistics, a Decision Theoretic Approach*, p. 197, Academic Press, New York.

Friedman, D., Gelman, A. and Phadia, E. (1988). Best invariant estimator of a distribution function under the Kolmogorov-Smirnov loss function, *Ann. Statist.*, **16**, 1254–1261.

Yu, Q. (1992). Minimax estimator in the classical invariant estimators of a distribution function, *Ann. Inst. Statist. Math.*, **44** (4), 728–735.

Yu, Q. and Chow, M. S. (1991). Minimaxity of the empirical distribution function in invariant problem, *Ann. Statist.*, **19** (2), 935–951.

Yu, Q. and Phadia, E. (1992). Minimaxity of the best invariant estimator of a distribution function under the Kolmogorov-Smirnov loss, *Ann. Statist.*, **20** (4), 2192–2195.