STRONGLY CONSISTENT ESTIMATOR FOR MIXTURES OF DISTRIBUTION FUNCTIONS

Keewhan Choi
Cornell University

ABSTRACT

The problem with which we are dealing in this paper is that of estimating mixing measures of mixtures of known distributions. An estimator is proposed and its strong consistency is proven. Asymptotic normality of the estimator for finite mixtures is proven. The estimator is in the spirit of Wolfowitz's minimum distance method.

STRONGLY CONSISTENT ESTIMATOR FOR MIXTURES OF DISTRIBUTION FUNCTIONS

Keewhan Choi
Cornell University

1. Introduction

The problem with which we are concerned in this paper is that of estimating mixing measures of mixtures of known distributions. Development of the strongly consistent estimator (i.e., an estimator which converges to true values with probability one) discussed in this paper results from the consideration of the minimum distance method of Wolfowitz (1957). Roughly speaking, the method of this paper consists of choosing that mixing measure which is "closest" to the empirical distribution function.

Since identifiability (i.e., one-to-one correspondence between mixing measures and the induced distributions) is a necessary condition for estimation, it follows that, in many problems, identifiability implies estimability by the method in this paper. For mixtures of finite distributions, it is obvious from the result of Robbins (1964) that identifiability implies estimability. In a series of papers, Teicher (1960, 1961, 1963) has investigated the identifiability problem of mixtures of distribution functions.

Robbins (1964) has proposed a strongly consistent estimator of mixing measures for finite mixtures of distribution functions. (His estimator is asymptotically multivariate-normal.) The estimator proposed in this paper is easier to compute than that of Robbins (1964) since our method only requires finding the minimum of a quadratic form. See Blischke (1963) for references on estimation problems for finite mixtures of parametric families of distribution functions.

An estimator of the mixing measures is proposed and its strong consistency is proven in Section 2. The asymptotic normality of the estimator for finite mixture is proven in Section 3.

2. Estimator and its strong consistency.

Let $\{F(x,\theta)\ ;\ \theta \in \text{Ⓗ}\}$ be a known family of distribution functions and G an (unknown) probability measure on Ⓗ. We shall assume that $F(x,\theta)$ is continuous and strictly increasing in x for each θ and continuous in θ for each x. The identifiability is assumed since it is a necessary condition for estimation.

The problem is to estimate the mixing measure G from n independent observations $\underset{\sim}{x}_n = (x_1, x_2, \ldots, x_n)$ from the common induced distribution (mixture) $P_G(x) = \int_{\text{Ⓗ}} F(x,\theta)dG(\theta)$. Denote the empirical distribution function of $\underset{\sim}{x}_n$ by $F_n(x)$. Then $G^*_{(n)}$ which minimizes $\int \left[P_{G_{(n)}}(x) - F_n(x)\right]^2 dF_n(x)$ is a strongly consistent estimator of G, i.e. $G^*_{(n)}$ conveyes weakly to G with probability one.

If Ⓗ consists of a finite number of points, then $P_G(x) = \sum_{j=1}^{m} g_j F(x,\theta_j)$ where $\sum_{j=1}^{m} g_j = 1$, $g_j > 0$ for $j = 1, 2, \ldots, m$. Then $G^*_{(n)} = (g^*_{1_{(n)}}, g^*_{2_{(n)}}, \ldots, g^*_{m_{(n)}})$ will be an m-dimensional probability vector. For all other cases, the support of $G^*_{(n)}$ consists of at most n + 1 points in Ⓗ. To see the validity of this assertion, consider the set C (in the n-dimensional Euclidean space) whose generic element is

$$\left(\int F(x_{(1_n)}, \theta)dG(\theta),\ \int F(x_{(2_n)}, \theta)dG(\theta),\ \ldots,\ \int F(x_{(n_n)}, \theta)dG(\theta)\right) \equiv \left(\int F(x_{(i)}, \theta)dG(\theta)\right)$$

where $x_{(i_n)}$ is the i^{th} order statistic of $\underset{\sim}{x}_n$ and G is any probability measure whose support is contained in Ⓗ. Then C is convex and compact. (Convexity is obvious. Take any sequence in C, $\left\{\left(\int F(x_{(i)}, \theta)dG_k(\theta)\right) ; k = 1, 2, \cdots\right\} \equiv S$. Since $\{G_k, k \geq 1\}$ is a sequence of distribution functions and $F(x,\theta)$ is continuous in θ there is a subsequence $\left\{G_{a_i}, i \geq 1\right\}$ such that $\int F(x_{(i)}, \theta)dG_{a_i}(\theta)$ converges to $\int F(x_{(i)}, \theta)dG(\theta)$. This implies that S has a convergent subsequence.) Because $\int\left[P_{G_{(n)}}(x) - F_n(x)\right]^2 dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left[\int F(x_{(i_n)}, \theta)dG(\theta) - \frac{i_n}{n}\right]^2$ is a continuous function on the compact set C, it achieves its minimum on C. Since C is convex, the point at which the minimum is achieved is a mixture of at most n + 1 extreme points of C.

<u>Proof of the strong consistency of $G^*_{(n)}$</u>

(i) $\int\left[P_G(x) - F_n(x)\right]^2 dF_n(x) \to 0$ w.p. 1,

where G is the (unknown) true mixing measure.

Since $F_n \to P_G$ uniformly w.p. 1, the convergence follows from the following fact:

$$\left.\begin{array}{l} f_n \geq 0 \text{ and measurable for all } n \\ f_n \to f_0 \text{ uniformly, and } f_0 \text{ bounded} \\ H_n \to H \text{ in distribution} \end{array}\right\} \Rightarrow \int f_n(x)dH_n(x) \to \int f_0(x)dH(x).$$

(ii) $\int\left[P_{G^*_{(n)}}(x) - F_n(x)\right]^2 dF_n(x) \to 0$ w.p. 1, and this implies

$P_{G^*_{(n)}}(x) \to P_G(x)$ w.p. 1.

The first assertion follows from (i), because (by definition)

$$\frac{1}{n}\sum_{i=1}^{n}\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 = \int\left[P_{G^*_{(n)}}(x) - F_n(x)\right]^2 dF_n$$

$$\leq \int\left[P_G(x) - F_n(x)\right]^2 dF_n(x) = \frac{1}{n}\sum_{i=1}^{n}\left[P_G(x_{(i_n)}) - \frac{i_n}{n}\right]^2.$$

Consider a sequence $\left\{\int\left[P_{G^*_{(n)}}(x) - F_n(x)\right]^2 dF_n(x),\ n \geq 1\right\}$. For given $\epsilon > 0$,

we have for all sufficiently large n,

$$\frac{1}{n}\sum_{i=1}^{n}\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \leq \frac{1}{n}\sum_{i=1}^{n}\left[P_G(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \leq \epsilon^2 .$$

Suppose $\left[P_{G^*_{(n)}}(x_{(i_0)}) - \frac{i_0}{n}\right]^2 \geq \epsilon^2$ for some i_0.

(Case 1) $\quad P_{G^*_{(n)}}(x_{(i_0)}) - \frac{i_0}{n} \geq \epsilon$

For all i_n such that $i_0 \leq i_n \leq i_0 + \frac{\epsilon}{2}n$

$$\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \frac{\epsilon^2}{4} ,$$

because

$$P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n} \geq P_{G^*_{(n)}}(x_{(i_0)}) - \frac{i_n}{n} \geq P_{G^*_{(n)}}(x_{(i_0)}) - \frac{1}{n}(i_0 + \frac{\epsilon}{2}n) \geq \frac{\epsilon}{2}.$$

Then

$$\frac{1}{n}\sum_{i=1}^{n}\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \frac{1}{n}\sum_{i=i_0}^{i_0+\frac{\epsilon}{2}n}\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \frac{\epsilon^3}{8} .$$

(Case 2) $\quad P_{G^*_{(n)}}(x_{(i_0)} - \frac{i_0}{n}) < - \epsilon$

Take i_n such that $i_0 - \frac{\epsilon}{2}n \leq i_n \leq i_0$.

Then

$$\frac{1}{n}\sum_{i=1}^{n}\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \frac{\epsilon^3}{8}$$

exactly as in (Case 1).

We have shown that $\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \epsilon^2$ for some i_n implies that $\int\left[P_{G^*_{(n)}}(x) - F_n(x)\right]^2 dF_n(x) \geq \frac{\epsilon^3}{8}$. Then it follows that the probability of $\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \epsilon^2$ for infinitely many n is less than or equal to the probability of $\int\left[P_{G^*_{(n)}}(x) - F_n(x)\right]^2 dF_n(x) \geq \frac{\epsilon^3}{8}$ for infinitely many n. Since the latter probability is zero by (i), the former probability is zero. To wit we have $\Pr\left\{\left[P_{G^*_{(n)}}(x_{(i_n)}) - \frac{i_n}{n}\right]^2 \geq \epsilon^2 \text{ for infinitely many } n\right\}=0.$

Now, take any sequence $\{i_n\}$ such that $\frac{i_n}{n} \to p$ and x such that $p = P_G(x)$. (Since $F(x,\theta)$ is strictly increasing in x, so is $P_G(x)$. Hence x exists uniquely.)

Then,

$$\lim_n P_{G^*_{(n)}}(x_{(i_n)}) = p = \lim_n P_G(x_{(i_n)}).$$

Because $F(x,\theta)$ is continuous in x, the Dominated Convergence Theorem shows that $P_G(x)$ is continuous.

Since $x_{(i_n)} \to x$ w.p. 1,

$$\lim_n P_G(x_{(i_n)}) = P_G(x) \qquad \text{w.p. } 1.$$

Hence

$$\lim_n P_{G^*_{(n)}}(x) = P_G(x) \qquad \text{w.p. } 1.$$

(iii) $\left[P_{G^*_{(n)}}(x) \to P_G(x) \text{ for all } x\right] \Rightarrow \left[G^*_{(n)} \xrightarrow{W} G\right]$,

where $\xrightarrow{W}$ denotes weak convergence.

Let $\mathcal{G} = \left\{G \mid \int_{(H)} dG(\theta) \le 1\right\}$, and

$$H = \left\{P_G(\cdot) \mid P_G(x) = \int_{(H)} F(x,\theta)\, dG(\theta),\ G \in \mathcal{G},\ x \in X\right\} .$$

Let h be the function from $\mathcal{G}$ onto H defined by $h(G) = P_G(\cdot)$.

Then $\mathcal{G}$ is a compact space with the usual weak topology. Since H is a topological space with its topology given by pointwise convergence of the functions $P_G(\cdot)$, it is a Hausdorff space. (If H is not a Hausdorff space, there is a sequence $\left\{P_{G_n}(\cdot),\ n \ge 1\right\}$ such that it converges to at least two different functions $f_G(\cdot)$, and $\tilde{f}_G(\cdot)$. Then there exists $x_0 \in X$ such that $f_G(x_0) \ne \tilde{f}_G(x_0)$. However, by the Helly-Bray theorem

$$P_{G_n}(x_0) \to f_G(x_0) \text{ and } P_{G_n}(x_0) \to \tilde{f}_G(x_0).$$

Therefore, $f_G(x_0) = \tilde{f}_G(x_0)$, which is a contradiction.)

Since h is a continuous (by the Helly-Bray theorem) and one-to-one (by the identifiability assumption) function of a compact space onto a Hausdorff space, h has a continuous inverse.

3. Asymptotic normality of $G^*_{(n)}$ when $\Theta = \{\theta_1, \theta_2, \cdots, \theta_m\}$.

Denote $F(x,\theta_j)$ by $F_j(x)$. Then $G^*_{(n)}$ minimizes among all $G = (g_1, \cdots, g_m)$,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{i_n}{n} - \sum_{j=1}^{m} g_j F_j(x_{(i_n)})\right)^2 .$$

Let

$$\overline{T}_n(G) = \frac{1}{n}\begin{bmatrix} \sum_{i=1}^{n} F_1(x_{(i_n)})\left(\frac{i_n}{n} - \sum_{j=1}^{m} g_j F_j(x_{(i_n)})\right) \\ \sum_{i=1}^{n} F_2(x_{(i_n)})\left(\frac{i_n}{n} - \sum_{j=1}^{m} g_j F_j(x_{(i_n)})\right) \\ \vdots \\ \sum_{i=1}^{n} F_m(x_{(i_n)})\left(\frac{i_n}{n} - \sum_{j=1}^{m} g_j F_j(x_{(i_n)})\right) \end{bmatrix}$$

Then $\overline{T}_n(G^*_{(n)}) = 0$.

By the Taylor series expansion

$$\overline{T}_n(G) = \overline{T}_n(G_0) + \dot{\overline{T}}_n(G_0)(G - G_0)$$

where G_0 is the true mixing measure and $\dot{\overline{T}}_n(G)$ is the $m \times m$ matrix of the partial derivatives of $\overline{T}_n(G)$ with respect to g_j. Under the assumptions that F_i are continuous and

$$|xF_i(x)| \le K\left[F_i(x)\left(1 - F_i(x)\right)\right]^{-1+\delta}$$

for some $\delta > 0$, constant K and $j = 1, 2, \cdots, m$, it follows from the results of Govindarajulu (1965) that $\Sigma_n^{-\frac{1}{2}}\,\overline{T}_n(G_0)$ converges in distribution to the normal variable with mean 0 and covariance matrix $I(\Sigma_n$ given below). Since

$$\bar{T}_n(G) = -\frac{1}{n}\left(\sum_{k=1}^{n} F_i(x_k)\, F_j(x_k)\right),$$

by the strong law of large numbers, $\dot{\bar{T}}_n(G)$ converges to $\left(-E\left[F_i(X)\, F_j(X)\right]\right)$ with probability one. Hence, by Pratt (1959),

$$\sqrt{n}\,(G^*_{(n)} - G_0) - \sqrt{n}\left(-E\left[F_i(X)\, F_j(X)\right]\right)^{-1}\left(-\bar{T}_n(G_0)\right) \to 0 \qquad \text{w.p. } 1,$$

provided $\left(E\left[F_i(X)\, F_j(X)\right]\right)$ is not singular. Therefore

$$\Sigma_n^{-\frac{1}{2}}\,(G^*_{(n)} - G_0) \to N(0, I)$$

in distribution where

$$\Sigma_n = \left(E\left[F_i(X)\, F_j(X)\right]\right)^{-1} \Sigma_n^* \left(E\left[F_i(X)\, F_j(X)\right]\right)^{-1}.$$

It remains only to show $\Sigma_n^* = (\sigma_{ij_{(n)}})$, where $\sigma_{ij_{(n)}}$ is the covariance between the two random variables

$$\frac{1}{n}\left[\sum_{k=1}^{n} F_i(X_{(k)})\,\frac{k}{n} - \sum_{\ell=1}^{m} g_{0\ell} \sum_{k=1}^{n} F_i(X_{(k)})\, F_\ell(X_{(k)})\right]$$

and

$$\frac{1}{n}\left[\sum_{k=1}^{n} F_j(X_{(k)})\,\frac{k}{n} - \sum_{\ell=1}^{m} g_{0\ell} \sum_{k=1}^{n} F_j(X_{(k)})\, F_\ell(X_{(k)})\right].$$

After some computation, $\sigma_{ij_{(n)}}$ is given as follows:

$$\sigma_{ij_{(n)}} = \text{cov}(a_i,c_j) - \text{cov}(a_i,d_j) - \text{cov}(b_i,c_j) + \text{cov}(b_i,d_j)$$

where

$$\text{cov}(a_i,c_j) = \frac{2}{n} \int\int_{-\infty<x<y<\infty} P_{G_0}(x)\left[1 - P_{G_0}(x)\right] P_{G_0}(x)\, P_{G_0}(y)\, dF_i(x)\, dF_j(x)$$

$$\operatorname{cov}(a_i, d_j) = \frac{2}{n} \int\int_{-\infty<x<y<\infty} P_{G_0}(x) \left[1 - P_{G_0}(y)\right] P_{G_0}(x)\, dF_i(x)\, d(F_j P_{G_0})(y)$$

$$\operatorname{cov}(b_i, c_j) = \frac{2}{n} \int\int_{-\infty<x<y<\infty} P_{G_0}(y) \left[1 - P_{G_0}(x)\right] P_{G_0}(y)\, dF_j(y)\, d(F_i P_{G_0})(x)$$

$$\operatorname{cov}(b_i, d_j) = \frac{2}{n} \int\int_{-\infty<x<y<\infty} P_{G_0}(x) \left[1 - P_{G_0}(y)\right] d(F_i P_{G_0})(x)\, d(F_j P_{G_0})(y).$$

Acknowledgements

The author wishes to thank Professor R. H. Farrell for several helpful discussions and proofs in Section 2.

References

Blischke, W. R. (1963). Mixtures of discrete distributions. Paper presented at the International Symposium of the Classical Discrete Distributions. McGill University, Montreal, August 15-20, 1963.

Govindarajulu, Z. (1965). Asymptotic normality of linear functions of order statistics in one and multi-samples. Tech. Rept. under U. S. Air Force Grant No. AF-AFOSR-741-65, 1-26.

Pratt, J. W. (1959). On a general concept of "in probability". Ann. Math. Stat. 30, 549-558.

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. Ann. Math. Stat. 35, 1-20.

Teicher, H. (1960). On the mixture of distributions. Ann. Math. Stat. 19, 360-369.

Teicher, H. (1961). Identifiability of mixtures. Ann. Math. Stat. 32, 244-248.

Teicher, H. (1963). Identifiability of finite mixtures. Ann. Math. Stat. 34, 1265-1269.

Wolfowitz, J. (1957). The minimum distance method. Ann. Math. Stat. 38, 75-88.