BU-290-M

STRONGLY CONSISTENT ESTIMATES FOR FINITES
MIXTURES OF DISTRIBUTION FUNCTIONS

Keewhan Choi
Cornell University

ABSTRACT

The problems with which we are concerned in this note are those of identifiability and strongly consistent estimates for mixing measures of a finite mixture of distribution functions. We show that the identifiability is a necessary and sufficient condition for the existence of a strongly consistent estimate of the mixing measure. An estimator for the mixing measure of a finite mixture of distribution functions is proposed and its strong consistency is proven.

Paper No. BU-122, Biometrics Unit, and Paper No. 502, Plant Breeding Department, Cornell University

STRONGLY CONSISTENT ESTIMATES FOR FINITE MIXTURES OF DISTRIBUTION FUNCTIONS

Keewhan Choi
Cornell University

Summary

The problems with which we are concerned in this note are those of identifiability and strongly consistent estimates for mixing measures of a finite mixture of distribution functions. We show that the identifiability is a necessary and sufficient condition for the existence of a strongly consistent estimate of the mixing measure. An estimator for the mixing measure of a finite mixture of distribution functions is proposed and its strong consistency is proven.

1. Introduction

Mixtures of distribution functions have received considerable attention recently. The mixtures are of general interest not only for their mathematical aspects but also for the large number of applied problems in which mixtures occur. Estimation of mixing measures of known component distributions is of the most general interest, however, it is necessary to investigate first the identifiability of mixing measures.

In a series of papers, Teicher (1960, 1961, 1963) has investigated extensively the identifiability problem. In particular, Teicher (1963) has

---

Paper No. BU-122, Biometrics Unit, and Paper No. 502, Plant Breeding Department, Cornell University

given a necessary and sufficient condition for the identifiability of finite mixtures. In Section 2, we show that identifiability is equivalent to the existence of a strongly consistent (i.e. converging to the true parameter with probability one) estimate for the mixing measure of finite mixtures of distribution functions.

Robbins (1964) has proposed a strongly consistent estimator of mixing measures of finite mixtures of distribution functions. (It is obvious in Robbins (1964) that his estimate is asymptotically multivariate-normal.) The estimate proposed in this note is easier to compute than that of Robbins (1964). See Blischke (1963) for references on estimation porblems for finite mixtures of parametric families of distributions and applications of mixtures of distributions.

2. Identifiability and strongly consistent estimates of the mixing measure

Let $f = \{P_1, P_2, \ldots, P_m\}$ be a known family of one-demensional distribution functions. Let $G = (g_1, g_2, \ldots, g_m)^t$ denote a probability vector, i.e. $g_i \geq 0$ for $i = 1, 2, \ldots, m$ and $\sum_1^m g_i = 1$. Then the new distribution function

$$P_G(x) = \sum_1^m g_i P_i(x)$$

is called a G-mixture of f, and G the mixing measure.

The problem is to find a strongly consistent estimate of G from n independent observations $x_1, x_2, \ldots, x_n$ with the common distribution function P_G. However, we must first investigate the question of identifiability.

Let $\mathcal{G}$ denote the class of all such probability vectors (mixing measures) and $\mathfrak{F}$ the induced class of mixtures. Then $\mathfrak{F}$ is said to be identifiable in $\mathcal{G}$ (with respect to f) if G and $\bar{G}$ are any two mixing measures such that

$P_G(x) = P_{\bar{G}}(x)$ for all x then $G = \bar{G}$. For brevity let us define, in this section, estimability to mean the existence of a strongly consistent estimate of the unknown G.

Theorem 1. Identifiability (i.e. $\mathcal{G}$ is identifiable) is a necessary and sufficient condition for estimability.

Proof. Necessity is obvious. Sufficiency is proven by showing that Teicher's necessary and sufficient condition for identifiability (I) implies the necessary and sufficient condition for estimability (II) given by Robbins (1964).

I. There exists m real values $x_1, x_2, \ldots, x_m$ for which the determinant $P_i(x_j)$, $1 \leq i,j \leq m$ does not vanish.

II. If $\sum_1^m c_i P_i(B) = 0$ for every set B, then $c_i = 0$ for all i.

The condition I implies that there exists $(x_1, \ldots, x_m)$ such that the rows of $\left(P_i(x_j)\right)$ $1 \leq i,j \leq m$ are linearly independent, which is equivalent to:

there exists m sets $(-\infty, x_1), (-\infty, x_2), \ldots, (-\infty, x_m)$ such that if $\sum_{i=1}^m c_i[P_i(x_j) - P_i(-\infty)] = 0$ for $j = 1, 2, \ldots, m$ then $c_i = 0$ for all i.

The last statement now implies II.

Let F_n be the emperical distribution function of $(x_1, x_2, \ldots, x_n)$. Then a sequence of non-negative vectors $G^*_{(n)} = \{g^*_{j(n)} \cdot j = 1, 2, \ldots, m\}$ which minimizes

$$\delta(P_G, F_n) \equiv \int [P_G(x) - F_n(x)]^2 dP_G(x)$$

is a strongly consistent estimate of the true mixing measure G. Since we are assuming that every component of the true mixing measure G is positive, we exclude from our consideration those $G^*_{(n)}$ every component of which approaches zero as n increases.

Without loss of generality, we could assume that all the P_i's are absolutely continuous with respect to a σ-finite measure μ and such that their densities $f_i = dP_i/d\mu$ are square integrable (See, for instance, Robbins (1964)). Using the densities f_i, $\delta(P_G, F_n)$ is expressed as:

$$\int [\sum_1^m g_i P_i(x) - F_n(x)]^2 \sum_1^m g_i f_i(x) d\mu(x) .$$

Then, by definition

(1) $\delta(P_{G^*(n)}, F_n) = \inf_G (P_G, F_n)$.

By the Glivenko-Cantelli theorem we have with probability one

(2) $\delta(P_G, F_n) \to 0$.

Hence, with probability one

(3) $\delta(P_{G^*(n)}, F_n) \to 0$.

$$\begin{aligned} &\delta(P_{G^*(n)}, F_n) \\ &= \int [P_{G^*(n)}(x) - P_G(x) + P_G(x) - F_n(x)]^2 dP_{G^*(n)}(x) \\ &= \int \{[P_{G^*(n)}(x) - P_G(x)]^2 + [P_G(x) - F_n(x)]^2 \\ &\quad + 2[P_{G^*(n)}(x) - P_G(x)][P_G(x) - F_n(x)]\} dP_{G^*(n)}(x) . \end{aligned}$$

Since the integrand converges to

$$\lim_{n\to\infty} [P_{G^*(n)}(x) - P_G(x)]^2$$

(which is bound by an integrable function)

application of the Slutsky theorem [p 255 Cramér (1946)] and Lebesgue Dominated Convergence theorem gives us that

(4) $\int [P_{G^*(n)}(x) - P_G(x)]^2 dP_{G^*(n)}(x)$

$= \sum_1^m g^*_{i(n)} \int [\sum_1^m (g^*_{j(n)} - g_j)P_j(x)]^2 dP_i(x)$

converges to 0 with probability one.

This, in turn, implies:

(5) $P_{G^*(n)}(x) = \sum_1^m g^*_{j(n)} P_j(x) \to P_G(x) = \sum_1^m g_j P_j(x)$

with probability one.

To see the validity of the implication (4) ⇒ (5), it is adequate to consider the following case only:

all $g^*_{i(n)}$ converges and at least one $g^*_{j(n)}$ converges to a positive member. Then the convergence of the expression (4) to zero implies that each term must converge to zero, i.e.,

$$g^*_{j(n)} \int [\sum_1^m g^*_{i(n)} P_i(x) - \sum_1^m g_i P_i(x)]^2 dP_j(x) \to 0$$

for all j. This convergence, in turn, implies (5) with probability one.

Let us rewrite (5) as follows:

(5′) $\lim_{n\to\infty} \sum_1^m (g^*_{j(n)} - g_j)P_j(x)$

$= \sum_1^m (\lim_{n\to\infty} g^*_{j(n)} - g_j)P_j(x)$

$= 0$.

Then the assumption of identifiability gives us:

(6) $\lim_n g^*_{j(n)} - g_j = 0$ for all j.

Acknowledgements

The author wishes to thank Professor J. Wolfowitz for several helpful discussions. The proof of strong consistency in Section 2 was inspired by the minimum distance method of Wolfowitz (1957).

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.

Cramér, H. (1946). Mathematical Methods of Statistics. Princeton University Press.

Robbins, H. (1964). The emperical 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.