Asymptotics of $k$-mean clustering under non-i.i.d. sampling

Regis J. Serinko*, Gutti Jogesh Babu

Department of Statistics, 219 Pond Laboratory, The Pennsylvania State University, University Park, PA 16802, USA

Received January 1993; revised June 1994

Abstract

The asymptotic theory of $k$-mean clustering is extended to stationary mixing processes, both $\varphi$-mixing and strongly mixing. In addition, a consistency result is obtained for non-identically distributed independent observations.

AMS 1980 Subject Classification: 60F05, 62E20

Keywords: Bahadur's representation; Singular Hessian; Non-identically distributed; Stationary mixing processes; Strong consistency; Weak limits

1. Introduction

Hartigan (1978) and Pollard (1981, 1982a, b) have found regularity conditions which assure consistency and asymptotic normality, with a convergence rate of $n^{1/2}$, of the $k$-mean estimators. One of the regularity conditions is that the Hessian of the between group sum-of-squares is nonsingular. Serinko and Babu (1992) have found a sufficient condition to assure the existence of an asymptotic distribution of the estimators, in the nonregular case of a singular Hessian. Further, all distributions possible under this weaker conditions are characterized. The rate of convergence is slower than $n^{1/2}$.

In this note the asymptotics of $k$-mean clustering under two sets of relaxed sampling assumptions are reported. The first set is that the observations are independent but come from different distributions. For this set of assumptions, substantial modifications of the regularity conditions lead only to a consistency result. The second set of sampling assumptions is that the observations form a stationary mixing process, either strongly mixing or $\varphi$-mixing. Under these assumptions only minor modifications of the regularity conditions are needed in order that the entire asymptotic theory carry over. This case includes autoregressive processes.

The significance and potential usefulness of these results lie in the widespread use of the $k$-mean procedure. For example in astronomy, Murtagh (1992) has employed this procedure to isolate distinct populations in
the galaxy from the the infrared signatures of the sources. The assumption of independent observations in this application is, at best, questionable because of the gravitational interactions. The results of this paper are, in part, a move toward more realistic sampling assumptions for such applications. (Also see Babu and Feigelson, 1994, in this regard.)

Section 2 contains some background material. The independent, non-identically distributed case is presented in Section 3. The results for stationary mixing observations are given in Section 4.

2. Background

The presentation of the material in this section follow Serinko and Babu (1992). It assumes, for simplicity, that each of the observations have a common distribution function. The appropriate modification for the non-identically distributed case is discussed in Section 3. No modification is needed for the stationary mixing case.

To discuss this formulation some notation is needed. For any distribution function $G$, define the quantile function by

$$G^{-1}(p) = \inf\{x: G(x) \geq p\} \quad 0 < p < 1. \tag{1}$$

Let $V_k = \{t \in \mathbb{R}^{k-1}: 0 < t_1 < t_2 < \ldots < t_{k-2} < t_{k-1} < 1\}$, $t_0 = 0$ and $t_k = 1$. Let $F$ be a distribution function with finite first moment and let the components of $\mu(t, k) \in \mathbb{R}^k$ be given by

$$\mu_j(t, k) = (t_j - t_{j-1})^{-1} \int_{t_{j-1}}^{t_j} F^{-1}(u) \, du \quad t \in V_k, \ j = 1, 2, \ldots, k. \tag{2}$$

The function

$$B(t, k) = \sum_{j=1}^{k} (t_j - t_{j-1})^2 \mu(t, k)^2 - \mu^2 \tag{3}$$

defined on $V_k$ is called the split function, where $\mu = \int_0^1 F^{-1}(u) \, du$. For $t$ in the boundary of $V_k$, let $B(t, k) = B(v, m)$, where $(m - 1)$ is the number of distinct components of $t$ which are different from zero or unity and $v \in V_m$ is the $(m - 1)$ dimensional vector of these components. If none of the components differ from zero or unity, then $B(t, k)$ is taken as zero. If $F$ has finite second moment this definition of $B$ on the boundary assures that it is continuous on the closure of $V_k$. Suppose that $B(t, k)$ has a unique maximum at $p \in V_k$. If $F$ has finite second moment, then $p$ and $\mu(p, k)$ are called the split point vector and cluster center vector, respectively. The second moment is needed for this terminology to be consistent with the terminology used in an alternate formulation, which is based on the minimization of the within group sum-of-squares (see Serinko and Babu, 1992). Let $F_n$ be the empirical distribution function of $n$ observations each with distribution function $F$. Define $\mu_n(t, k) \in \mathbb{R}^k$ with components

$$\mu_{jn}(t, k) = (t_j - t_{j-1})^{-1} \int_{t_{j-1}}^{t_j} F_n^{-1}(u) \, du \quad t \in V_k, \ j = 1, 2, \ldots, k \tag{4}$$

and set $\mu_n = \int_0^1 F_n^{-1}(u) \, du$. The sample split point function is defined by

$$B_n(t, k) = \sum_{j=1}^{k} (t_j - t_{j-1})^2 \mu_n(t, k)^2 - \mu_n^2 \quad t \in V_k, \ n \geq k.$$ 

A point $p_n \in V_k$ which maximizes $B_n$ is taken as an estimator of $p$; it is called the sample split point vector. The sample cluster vector $\mu_n(p_n, k)$ is the estimator of $\mu(p, k)$. 
Whenever $k$ is fixed, which is most often the case, the notation is somewhat simplified by suppressing the dependence of various quantities on the number of clusters.

Both $F_1$ and $F_n$ are left continuous functions, hence the directional derivatives of both $B$ and $B_n$ exist. $B_j^{(\pm)}$ and $B_n^{(\pm)}$ will denote the directional derivatives in the direction of $\pm e_j$ of $B$ and $B_n$, respectively, where $e_j \in \mathbb{R}^{k-1}$ and it has all zero entries save the jth, which is unity. The directional derivatives are given by

$$B_j^{(\pm)}(t) = \left[ \mu_{j+1}(t) - \mu_j(t) \right] \left[ \mu_{j+1}(t) + \mu_j(t) - 2F_1(t_j^{\pm 1}) \right] \quad j = 1, 2, \ldots, k - 1,$$

where $F_1(t_j^{\pm 1}) = \lim_{t_j \to \pm\infty} F_1(t_j)$. An expression for $B_n^{(\pm)}$ is obtained by affixing the subscript $n$ to all terms in (5). The corresponding vectors of directional derivatives are denoted by $\mathbf{B}^{(\pm)}$ and $\mathbf{B}_n^{(\pm)}$. (Hartigan, 1978, uses the notation $dB^{(\pm)} / dp$ for the vector of directional derivatives). If the components of $t$ are continuity points of $F_1$, then $\mathbf{B}^{(\pm)}(t)$, the vector of first partial derivatives at $t$ exists. Further, the Hessian $\mathbf{B}^{(2)}$ of $B$ exists at $t$ whenever $F_1$ has derivatives at the components of $t$.

3. Non-identically distributed observations

When observations are not identically distributed, the split function and other population quantities defined in Section 2 are meaningless. The first order of business is to modify the definitions to take into account the lack of a single distribution function. To this end, suppose the observations, $X_1, X_2, \ldots, X_n$, are independent with $X_i$ from distribution function $G_i$. The role of the population distribution function is played by $\bar{G}_n(t) = 1/n \sum_{i=1}^n G_i(t)$. The definitions of population quantities are modified by replacing $F$ and $F_1$ with $\bar{G}_n$ and $\bar{G}_n^{-1}$, respectively. In order to make their dependence on $n$ explicit, a subscript $n$ will be affixed to population quantities. To avoid confusion with sample quantities, which originally were distinguished by the subscript $n$, hats will be placed on these. For example, $\hat{p}_n$ will, in this section, denote the population split point vector and $\tilde{p}_n$ will denote a sample split point vector. The empirical distribution function will still be denoted by $F_n$.

Define $\delta_n(t) = t - \bar{G}_n^{-1}(t)$ and denote the Lebesgue measure on $(0, 1)$ by $\lambda$. Also, let $\eta^i = \text{Var}(X^2_i)$, where $X_i$ has distribution function $G_i$; $i = 1, 2, \ldots$. The regularity condition for non-identically distributed observations are

(N1) The sequence of distribution functions $\{G_i\}$ satisfies $\sum_{j=1}^{\infty} \left(1/j^2\right) \eta^i_j < \infty$;

(N2) $\bar{G}_n$ is a distribution functions such that $\lambda \left\{ \{t \in (0, 1) : \liminf_{n \to \infty} \delta_n(t) > 0 \} \right\} = 1$;

(N3) $\bar{G}_n$ gives rise to a split function with a unique maximum at $p_n$ in $V_k$ and $\lim\inf_{n \to \infty} (B_n(p_n) - B_n(t)) > 0$ for all $t \in V_k$;

(N4) The derivative $\bar{g}_n$ of $\bar{G}_n$ exists and is continuous in some open neighborhood $I_i$ of $p_i = \bar{G}_n^{-1}(p_n)$ with $\lim\inf_{n \to \infty} \inf_{x \in I_i} \bar{g}_n(x) > 0$; $i = 1, 2, \ldots, k - 1$. Further, the derivative of $\bar{g}_n$ exists and is bounded in $I_i$; $i = 1, 2, \ldots, k - 1$.

Condition (N1) is needed so that the strong law of large numbers applies to the second moments of $X_1, X_2, \ldots, X_n$. Condition (N2) is needed to assure that $(\bar{G}_n^{-1}(t) - F_1^{-1}(t))$ goes to zero except on a null set $N \subseteq (0, 1)$ and outside a null subset $A$ of the sampled space. These are used in the proof of Lemma 3. The second point of (N3) assures that the value of $B_n$ at $p_n$ does not come arbitrarily close to the value of $B_n$ at any other point in $V_k$ in the limit of large $n$. This used in the proof of Theorem 3. Condition (N4) is a condition for the strong representation of sample quantiles of non-identically distributed observations (Liu and Singh, 1989; Babu and Singh, 1984) which is used in the proof of Lemmas 2 and 3.

To prove strong consistency, it is first shown that $\sup_{t \in V_k} |B_n(t) - \bar{B}_n(t)| \to 0$ as $n \to \infty$ a.s. Second, it is argued that $p_n$ is the unique maximum of $B_n$ on the closure of $V_n$. Third, it is observed that $B_n$ is defined so as to be continuous on $V_k$ closure when $F$ has a finite second moment. These results are combined to give strong consistency.
Lemma 1. Suppose that $\bar{G}_n$ satisfies (N1)–(N2), then

$$\lim_{n \to \infty} \sup_{t \in V_k} |B_n(t) - \hat{B}_n(t)| = 0 \quad \text{a.s.}$$

Proof. To simplify the notation, let

$$D_n = \sup_{t \in V_k} |B_n(t) - \hat{B}_n(t)|,$$

$$D_{in} = (t_i - t_{i-1}) (\mu_{in}(t) - \hat{\mu}_{in}(t))$$

and $D_{on} = \mu_n^2 - \hat{\mu}_n^2$. In this notation,

$$D_n = \sup_{t \in V_k} \left| \sum_{i=1}^k (D_{in}) - D_{on} \right|.$$

The triangle inequality gives

$$D_n \leq \sum_{i=1}^k \sup_{t \in V_k} |D_{in}| + |D_{on}|.$$

A uniform bound in $i$ is obtained for $|D_{in}|$. One may write

$$D_{in} = (ti - t_{i-1}) (\mu_{in}(t) - \hat{\mu}_{in}(t))^2 + 2(t_i - t_{i-1}) \mu_{in}(t) \hat{\mu}_{in}(t) (\mu_{in}(t) - \hat{\mu}_{in}(t)). \quad (6)$$

Recall that

$$(\mu_{in}(t) - \hat{\mu}_{in}(t)) = (t_i - t_{i-1}) \int_{t_{i-1}}^{t_i} (\bar{G}_n^{-1}(u) - F_n^{-1}(u)) \, du.$$

Jensen's inequality is applied to the first term in (6) and Hölder's to both factors in the second term to yield

$$D_{in} \leq \int_{t_{i-1}}^{t_i} \left( \bar{G}_n^{-1}(u) - F_n^{-1}(u) \right)^2 \, du + 2 \left[ \int_{t_{i-1}}^{t_i} (F_n^{-1}(u))^2 \, du \int_{t_{i-1}}^{t_i} (\bar{G}_n^{-1}(u) - F_n^{-1}(u))^2 \, du \right]^{1/2}.$$

The integrands are all positive, hence

$$D_{in} \leq \int_{0}^{1} \left( \bar{G}_n^{-1}(u) - F_n^{-1}(u) \right)^2 \, du + 2 \left[ \int_{0}^{1} (F_n^{-1}(u))^2 \, du \int_{0}^{1} (\bar{G}_n^{-1}(u) - F_n^{-1}(u))^2 \, du \right]^{1/2} \equiv d_n.$$

This leads to

$$D_n \leq kd_n + |D_{on}|.$$

Now $\hat{\mu}_n = \bar{X}_n$ and $\mu_n = 1/n \sum_{i=1}^n E[X_1]$, hence the strong law of large numbers (SLLN) (Feller, 1971, p. 239) gives $\lim_{n \to \infty} D_{on} = 0$ a.s. It remains to show that $d_n \to 0$ as $n \to \infty$ a.s. Condition (N1) and the SLLN assures that the first factor in the second term of $d_n$ is bounded. Therefore, it suffices to show $\lim_{n \to \infty} \int_{0}^{1} (\bar{G}_n^{-1}(u) - F_n^{-1}(u))^2 \, du = 0$ a.s. Condition (N2) assures that there exists a single null subset $A$ of the sample space that $\bar{G}_n^{-1}(u) - F_n^{-1}(u) \to 0$ as $n \to \infty$ outside of $A$ except for $u$ in a null set $N \subseteq (0, 1)$. The strong law of large numbers assures that $(\bar{G}_n^{-1}(u) - F_n^{-1}(u))^2$ is uniformly integrable (see Billingsley, 1986, p. 219). Therefore $d_n \to 0$ as $n \to \infty$ a.s. This completes the proof of the Lemma. □

Proposition 1. Suppose that $G_n$ satisfies (N3), then $p_n$ is the unique maximum of $B_n$ on the closure of $V_k$.

Proof. Let $t \in V_r$ and suppose $\bar{G}_n^{-1}$ increases on $[t_{i-1}, t_i)$. For arbitrary $s \in (t_{i-1}, t_i)$, let $s = (t_1, t_2, \ldots, t_{i-1}, s, t_i, \ldots, t_{j-1}) \in V_{r+1}$. The first step is to show that $B_n(s, r + 1) - B_n(t, r) > 0$. By (N3), $p_n$ is the unique maximum of $B_n$ on $V_k$. The second step is to use this fact to show, that for $t \in V_j, j < k$ at least
one interval \([t_{i-1}, t_i)\) exists such that \(\tilde{G}_n^{-1}(t)\) increases on it. These results are combined to complete the proof.

Step one begins with the following.

\[
B_n(s, r + 1) - B_n(t, r) = (t_i - s)^{-1} \left[ \int_s^{t_i} \tilde{G}_n^{-1}(u) \, du \right]^2 + (s - t_i)^{-1} \left[ \int_{t_i}^{t} \tilde{G}_n^{-1}(u) \, du \right]^2
\]

\[\quad - (t_i - t_i)^{-1} \left[ \int_{t_i}^{s} \tilde{G}_n^{-1}(u) \, du + \int_{s}^{t} \tilde{G}_n^{-1}(u) \, du \right]^2 \]

\[\quad = (t_i - t_i)^{-1} \left[ \left( \frac{t_i - s}{s - t_i} \right)^{1/2} \int_{t_i}^{s} \tilde{G}_n^{-1}(u) \, du - \left( \frac{s - t_i}{t_i - s} \right)^{1/2} \int_{s}^{t} \tilde{G}_n^{-1}(u) \, du \right]^2. \]

The right-hand side is positive unless the quantity in the bracket vanishes. However, it only vanishes if \(\tilde{G}_n^{-1}(u)\) is constant on \([t_{i-1}, t_i)\). Hence, if \(\tilde{G}_n^{-1}(u)\) increases on \([t_{i-1}, t_i), B_n(s, r + 1) - B_n(t, r) > 0.\)

By (N3), \(p_*\) is the location of the unique maximum of \(B_n\) on \(V_k\). Suppose that, for some \(t \in V_{k-1}, \tilde{G}_n^{-1}\) is constant on each \([t_{i-1}, t_i), i = 1, 2, \ldots, k - 1.\) Then for any \(s \in (t_{r-1}, t_r)\) the vector \((t_1, t_2, \ldots, t_{r-1}, s, t_r, \ldots, t_{k-1}) \in V_k\) is the location of a maximum. This contradicts (N3). Therefore, for any \(t \in V_{k-1}, \tilde{G}_n^{-1}\) increases on at least one interval \([t_{i-1}, t_i), i = 1, 2, \ldots, k - 1.\) This implies that for any \(t \in V_{j, j < k - 1}\) the resulting partition, which is rougher, must have at least one interval on which \(\tilde{G}_n^{-1}\) increases. Therefore, by the first part, one may increase \(B_n(t, j)\) by splitting this interval. Hence for each \(t \in V_{j, j < k}\) there exists an \(s \in V_k\) such that \(B_n(s, k) - B_n(t, j) > 0.\) This along with the definition of \(B_n(t, k)\) on the boundary of \(V_k\) leads to \(B_n(p_*, k) - B_n(t, k) > 0\) for \(t\) in the boundary of \(V_k\). This completes the proof of the proposition.

**Theorem 1.** Suppose \(\tilde{G}_n\) satisfies (N1)-(N3). Then \(\lim_{n \to \infty} (p_* - \hat{p}_n) = 0\) a.s.

**Proof.** By definition of \(\hat{p}_n\) one has

\[
\tilde{B}_n(\hat{p}_n) - \tilde{B}_n(p_n) \geq 0.
\]

Terms are added and subtracted to give

\[
[\tilde{B}_n(\hat{p}_n) - \tilde{B}_n(p_n)] + [B_n(p_n) - \tilde{B}_n(p_n)] + [B_n(\hat{p}_n) - B_n(p_n)] \geq 0.
\]

The conditions of Lemma 1 are met, hence the first two terms go to zero as \(n \to \infty\) a.s. and one has

\[
0 \leq B_n(p_n) - B_n(\hat{p}_n) \to 0 \quad \text{as } n \to \infty \quad \text{a.s.}
\]

Suppose \(\lim_{n \to \infty} p_n - \hat{p}_n \neq 0\) a.s., then for some \(\delta > 0, |p_n - \hat{p}_n| \geq \delta\) infinitely often on a set of nonzero measure. If \(|p_n - t| \geq \delta\), then there exists an \(\epsilon > 0\) such that \(B_n(p_n) - B_n(t) \geq \epsilon.\) This follows from Proposition 4, the continuity of \(B_n\) on the closure of \(V_k\), and the second part of (N3). These considerations lead to \(B_n(p_n) - B_n(\hat{p}_n) \geq \epsilon\) infinitely often on a set of nonzero measure. This contradicts (7), hence \(\lim_{n \to \infty} (p_n - \hat{p}_n) = 0\) a.s. The proof is now complete.

Consideration is now turned to the weak limits of the estimators. There are analogue to Lemmas 1 and 2 of Serinko and Babu (1992). The proofs of the analogous lemmas follow the proofs of Lemmas 1 and 2 very closely, though there are two differences in the tactics of proof. The first is that the strong representation of sample quantiles for non-identically distributed observations (see Liu and Singh, 1989; Babu and Singh 1984) is invoked. The second is that one must use the Lindeberg central limit theorem (Billingsley, 1986, p. 369) and
the Cramér-Wold device (Billingsley, 1986, p. 397) in place of the standard multivariate CLT. The analogous lemmas, the proofs of which are omitted, are as follows.

**Lemma 2.** Suppose \( G_n \) is a distribution function which satisfies (N1)–(N4), then

\[
(\hat{\mu}_n(p_n) - \mu_n(p_n)) = (\hat{\mu}_n(p_n) - \mu_n(p_n)) + o_p(n^{-1/2})
\]

and further

\[
\sqrt{n}(\Sigma_n^{(0)})^{-1/2}((\hat{\mu}_n(p_n) - \mu_n(p_n)) \Rightarrow \text{MVN}_k(0, I) \text{ as } n \to \infty.
\]

**Lemma 3.** Under the conditions of Lemma 2

\[
B_n^{(1)}(\hat{\mu}_n) = -Z_n + o_p(n^{-1/2}),
\]

where

\[
\sqrt{n} \Sigma_n^{-1/2} Z_n \Rightarrow \text{MVN}_{k-1}(0, I).
\]

The matrices \((\Sigma_n^{(0)})^{-1/2}\) and \(\Sigma_n^{-1/2}\) are the square roots of the inverses of the variance–covariance matrices \(\Sigma_n^{(0)}\) and \(\Sigma_n\), respectively. The form of these latter matrices are obtained by affixing the subscript \(n\) to the equations for \(\Sigma^{(0)}\) and \(\Sigma\) in the appendix.

It is unreasonable to assume the existence of a limit for \(B_n^{(2)}(p_n)\). Hence, the best weak limit one might obtain is

\[
\Sigma_n^{-1/2} B_n^{(2)}(p_n)n^{1/2}(\hat{\mu}_n - p_n) \Rightarrow \text{MVN}_{k-1}(0, I)
\]
as \(n \to \infty\).

4. Stationary mixing observations

The formulation of the \(k\)-mean procedure for stationary mixing observations is the same as for i.i.d. observation. In fact the proofs parallel those already presented, with only small changes. For example, when Bahadur’s representation of sample quantiles is invoked in the i.i.d. proof, one invokes its Babu and Singh (1978) version for mixing observations. Likewise, mixing versions of the strong law of large numbers and the CLT are used in place of their i.i.d. counterparts.

A sequence of random variables \(\{X_n\}\) with common distribution function \(F\) is called a stationary \(\varphi\)-mixing process, if there exists a sequence \(\{\alpha(n)\}\) such that \(1 \geq \varphi(n)10\) as \(n \to \infty\) and

\[
\sup_{k \geq 1} \sup_{B \in M_n^i} \sup_{A \in M_n^j} |P(A \cap B) - P(A)P(B)| \leq \varphi(n),
\]

where \(M_n^i\) denotes the \(\sigma\)-field generated by \(X_i a \leq i \leq b\). The process is called strong mixing if there exists a sequence \(\{\alpha(n)\}\) such that \(1 \geq \alpha(n)10\) as \(n \to \infty\) and

\[
\sup_{k \geq 1} \sup_{B \in M_n^i} \sup_{A \in M_n^j} |P(A \cap B) - P(A)P(B)| \leq \alpha(n).
\]

It is worth noting that strong mixing sequences arise in the study of time series, since many autoregressive processes are strong mixing (see Gordtski, 1977; Chandra, 1976).
The variance-covariance matrices $\Sigma_M^{(0)}$ and $\Sigma_M$, which appear below, are defined in the appendix. The regularity conditions for stationary mixing observations are:

(M1) $F$ is a distribution function with finite fourth moment;

(M2) $F$ is a distribution function which gives rise to a split function with a unique maximum at $p$ in $V_k$;

(M3) The derivative $f$ of $F$ exists and is continuous in some open neighborhood of $\tilde{\mu}_i = F^{-1}(p_i)$ and $f(\tilde{\mu}_i) > 0$ $i = 1, 2, \ldots, k - 1$;

(M4a) The mixing coefficients satisfy $\sum_{i=1}^{\infty} \varphi^{1/2}(i) < \infty$;

(M4b) The mixing coefficients satisfy $\varphi(n) = O(e^{-\theta n})$ for some $\theta > 0$;

(M5) The matrices $\Sigma_M^{(0)}$ and $\Sigma_M$ are nonsingular;

(M6) The Hessian $B^{(2)}(p)$ is nonsingular.

The fourth moment is needed for the strong law of large numbers to apply to the second moments of the observations. The conditions (M4) are needed for the strong representation of sample quantiles (Babu and Singh, 1978) and the central limit theorem, both of which are invoked in the proofs of Lemmas 4 and 5. A reference to (M4) should be understood to be either condition (M4a) or (M4b) according to whether the observations are $\phi$-mixing or strong mixing, respectively. (M5) is needed for the central limit theorem. The proofs in this section are omitted. The reader is referred to analogous results in Section 3 or Serinko and Babu (1992) for the strategy of proof.

The consistency result for mixing observations is

**Theorem 2.** Suppose $F$ satisfies (M1) and (M2), then $\lim_{n \to \infty} p_n = p$ a.s.

The proof of this theorem follows the proof of Theorem 1. The lemmas which correspond to Lemmas 1 and 2 of Serinko and Babu (1992) and Lemmas 2 and 3 in the non-identically distributed case are the following.

**Lemma 4.** Suppose $F$ satisfies (M1)–(M3) and the observations satisfy (M4) and (M5), then

$$(\mu_n(p_n) - \mu(p_n)) = (\mu_n(p) - \mu(p)) + o_p(n^{-1/2})$$

and further

$$\sqrt{n} (\mu_n(p) - \mu(p)) \Rightarrow MVN_k(0, \Sigma_M^{(0)}) \text{ as } n \to \infty.$$  

**Lemma 5.** Under the condition of Lemma 4

$$B^{(1)}(p_n) = -Z_n + o_p(n^{-1/2}),$$

where

$$\sqrt{n} Z_n \Rightarrow MVN_{k-1}(0, \Sigma_M).$$

If (M6) is satisfied the mean-value theorem and Lemma 5 yield a central limit theorem for $\sqrt{n}(p_n - p)$.

The entire theory of Serinko and Babu (1992) carries over to mixing observations. That theory is built on three propositions, a lemma and a corollary. Propositions 1 and 2 of Serinko and Babu (1992) are proven under conditions (M1)–(M3), there called (H1)–(H3). On the otherhand, the proofs of Lemma 3 and Corollary 1 of that paper, in addition to these assumptions, uses the assumption of i.i.d. observations, through Lemmas 1 and 2 of that paper. Since the conclusions of those lemmas, here called Lemmas 4 and 5, hold under the mixing conditions and the assumptions (M4) and (M5), the conclusions of Lemma 3 and Corollary 1 will also hold with the i.i.d. assumption replaced by the mixing conditions and assumptions (M4) and (M5). Proposition 3 of Serinko and Babu (1992) is proven without reference to the nature of the observations, but it does require an additional assumption. This assumption is stated below for completeness. First, some new notation must be introduced.
Under conditions (M1)-(M3), whenever $\text{Det} \mathbf{B}^{(2)}(\mathbf{p}) = 0$, the linear span of the set

$$
\mathbf{W}^{(s)}(t) = \{ \mathbf{x} \in \mathbb{R}^{k-1} : \mathbf{x}^T \mathbf{B}^{(2)}(t) = \mathbf{x}^T \lambda_1(t) \},
$$

where $\lambda_1(t)$ is the largest eigenvalue of $\mathbf{B}^{(2)}(t)$, is one-dimensional for $t$ in a neighborhood $\mathbf{W}^{(s)}$ of $\mathbf{p}$. Let $\mathbf{e}^{(s)}(t)$ denote the unit eigenvector associated with $\lambda_1(t)$. Then one may define a function $g$ on $\mathbf{W}^{(s)}$ by

$$
g(x\mathbf{e}^{(s)}(t)) = -\mathbf{e}^{(s)}(p)^T \mathbf{B}^{(1)}(p + x\mathbf{e}^{(s)}(t)),
$$

where $x$ is a sufficiently small real.

The following condition on $g$ is sufficient to assure the existence of normalizing constants $\{a_n\}$ such that $a_n(\mathbf{p}_n - \mathbf{p})$ will have a weak limit.

(M7) $F$ is a distribution function, with $\text{Det} \mathbf{B}^{(2)}(\mathbf{p}) = 0$, for which there exists a increasing, real-valued function $r$, defined on a neighborhood $V$ of the origin of $\mathbb{R}$, with the following properties

(a) $\lim_{x \to 0, t \to p, r = o(x)} g(x\mathbf{e}^{(s)}(t))/r(x) = 1$;
(b) $\frac{r(x)}{r(ax)} = \lim_{x \to o^+} \frac{r(ax)}{r(x)}$ exists for $a$ in a dense set $A \subseteq [0, \infty)$.

For a random variable $Y$, let $Y^+ = \max \{0, + Y\}$. The second and final major result can now be stated.

**Theorem 3.** Suppose that the mixing process satisfies (M1)-(M5) and (M7). Then there exists a real sequence $\{a_n\}$ with $\lim_{n \to \infty} a_n = \infty$ and a random variable $X$ such that

$$
a_n(\mathbf{p}_n - \mathbf{p}) \Rightarrow X \mathbf{e}^{(s)} \quad \text{as } n \to \infty.
$$

Further, $X$ is given by one of the following:

$$
X = (Z^+)^\beta - c(Z^-)^\beta, \quad X = (Z^+)^\beta, \quad X = -(Z^-)^\beta,
$$

where $0 < \beta \leq 1$, $0 < c < \infty$, and $Z \sim N(0, \sigma^2)$ with $\sigma^2 = \mathbf{e}^{(s)}(\mathbf{p})^T \Sigma \mathbf{e}^{(s)}(\mathbf{p})$.

**Remark 1.** $g(x) = o(x)$ implies that $a_n = o(n^{1/2})$. That is, the rate of convergence is slower for distributions with singular Hessians than for those with nonsingular Hessians.

The weak limit of the cluster center vector is given in the following corollary.

**Corollary 1.** Under the conditions of Theorem 3

$$
a_n(\mu_n(\mathbf{p}_n) - \mu(\mathbf{p})) \Rightarrow X \mathbf{M} \mathbf{e}^{(s)},
$$

where $\mathbf{M}$ is a $k \times (k - 1)$ matrix, with elements

$$(\mathbf{M})_{jj} = \frac{\mu_{j+1}(\mathbf{p}) - \mu_j(\mathbf{p})}{2(p_j - p_{j-1})} ; \quad j = 1, 2, \ldots, k - 1;$$

$$(\mathbf{M})_{jj-1} = \frac{\mu_j(\mathbf{p}) - \mu_{j-1}(\mathbf{p})}{2(p_j - p_{j-1})} ; \quad j = 2, 3, \ldots, k;$$

and all the other elements vanish.
Appendix

The variance–covariance matrices which appear first in Lemmas 2–5 are given here. Some notation is introduced first. Let
\[ \sigma_j^2 = (p_j - p_{j-1})^{-1} \int_{p_{j-1}}^{p_j} (F_n^{-1}(u) - F^{-1}(u))^2 \, du \quad \text{and} \quad \mu_j = \mu_j(p) \quad j = 1, 2, \ldots, k. \]

Define a matrix with components
\[ [\Sigma^{(0)}]_{jj} = \sigma_j^2 + \mu_j \left[ \frac{1 - (p_j - p_{j-1})}{(p_j - p_{j-1})} \right] \]
and
\[ [\Sigma^{(0)}]_{jl} = -\frac{\mu_j \mu_l}{(p_j - p_{j-1})(p_l - p_{l-1})}, \quad l \neq j; \quad j, l = 1, 2, \ldots, k. \]

A second matrix is given by
\[ [\Sigma]_{jj} = [\Sigma^{(0)}]_{j+1,j+1} + [\Sigma^{(0)}]_{jj} + \left( \frac{2}{f(\bar{\mu}_j)} \right)^2 p_j(1 - p_j) + 2[\Sigma^{(0)}]_{j+1} \]
\[ + 2 \left( \frac{2}{f(\bar{\mu}_j)} \right) \mu_{j+1} p_j + 2 \left( \frac{2}{f(\bar{\mu}_j)} \right) \mu_j(1 - p_j); \quad j = 1, 2, \ldots, k-1 \]
and
\[ [\Sigma]_{jl} = [\Sigma^{(0)}]_{j+1,l+1} + [\Sigma^{(0)}]_{jl} + [\Sigma^{(0)}]_{j+1,l} + [\Sigma^{(0)}]_{j+1,l+1} - \left( \frac{2}{f(\bar{\mu}_j)} \right) p_l(\mu_{j+1} + \mu_j) - \left( \frac{2}{f(\bar{\mu}_j)} \right)(1 - p_j)(\mu_{l+1} + \mu_l), \quad 1 < j; \quad j, l = 1, 2, \ldots, k-1. \]

The matrices \(\Sigma_{n}^{(0)}\) and \(\Sigma_n\) in Lemmas 2 and 3 are obtained from \(\Sigma^{(0)}\) and \(\Sigma\), respectively, by affixing the subscript \(n\) to the terms in the above expression. The matrix in Lemma 4 is given by
\[ (\Sigma_{M}^{(0)})_{ij} = (\Sigma^{(0)})_{ij} + \sum_{k=1}^{\infty} \{ E[\xi_{i0}\xi_{jk}] + E[\xi_{j0}\xi_{ik}] \}, \]
where the mixing random variables \(\xi_{jr}\) are given by
\[ \xi_{jr} = X_r \cdot I \{ F^{-1}(p_{j-1}) < X_r \leq F^{-1}(p_j) \} - \mu_j(p)(p_j - p_{j-1}); \quad j = 1, 2, \ldots, k; \quad r = 0, 1, 2, \ldots \]

The matrix in Lemma 8 has components
\[ (\Sigma_{M})_{ij} = (\Sigma)_{ij} + \sum_{k=1}^{\infty} \{ E[\eta_{i0}\eta_{jk}] + E[\eta_{j0}\eta_{ik}] \}, \]
where the mixing random variables \(\eta_{jr}\) are given by
\[ \eta_{jr} = \{ X_r \cdot I \{ (p_{j+1} - p_j)^{-1} I \{ F^{-1}(p_j) < X_r \leq F^{-1}(p_{j+1}) \} + (p_j - p_{j-1})^{-1} I \{ F^{-1}(p_{j-1}) < X_r \leq F^{-1}(p_j) \} \}
\[ - 2(f(\bar{\mu}_j))^{-1} I \{ X_r \leq F^{-1}(p_j) \} - [\mu_{j+1}(p) + \mu_j(p) - 2(f(\bar{\mu}_j))^{-1} p_j] \times [\mu_{j+1}(p) - \mu_j(p)] \}
\]
References

Liu, R.Y. and K. Singh (1989), Interpreting i.i.d. inferences under some non-i.i.d. models, preprint.