Journal of Classification 8:31-44 (1991)

Large-Sample Results for Optimization-Based Clustering Methods

Peter G. Bryant

University of Colorado at Denver

Abstract: Many common (nonhierarchical) clustering and classification methods are optimization-based methods, in the sense described by Windham (1987) in this Journal. This paper gives some large sample properties for estimates derived by such methods. Under appropriate conditions, such estimates converge with probability one to a limit, and are asymptotically normally distributed around that limiting value. The conditions are satisfied by most of the common examples of optimization-based methods.

Keywords: Classification; Clustering; Maximum likelihood; Asymptotic properties.

1. Introduction

Windham (1986, 1987) described a class of clustering and classification methods called optimization-based methods, and showed that many common (nonhierarchical) methods fall in this class. While whole classes of hierarchical methods have been studied (see Day and Edelsbrunner 1985), for example), nonhierarchical methods have typically been treated

Author's Address: Peter G. Bryant, Graduate School of Business Administration, University of Colorado at Denver, 1200 Larimer Street, Campus Box 165, Denver, CO, 80217-3364, USA. He thanks Michael Windham and John Williamson for several valuable conversations on this topic. Prepared for the 2nd International Conference, International Federation of Classification Societies, Charlottesville, VA, 1989. Supported in part by summer research funds, Graduate School of Business Administration, University of Colorado at Denver.

individually, and in an ad hoc manner. Windham's characterization of nonhierarchical methods seems to be the first which provides a basis for a coherent theory of such methods.

This paper extends Windham's work by giving some large-sample statistical properties of the common optimization-based methods. Specifically, it shows that under appropriate conditions, estimates of cluster characteristics derived by such methods will tend to limiting values with probability one, and will be approximately normally distributed around those values for large samples. In a random sample of observations from a population, we might say that the sample average "tries" to estimate the population mean, because it tends to the population mean as the sample size increases. The limiting values for an optimization-based method tell us, in the same sense, what the method is "trying" to estimate. As will be seen in Section 4, the results are sometimes surprising.

The paper also extends the work of Marriott (1975) and Bryant and Williamson (1978, 1986) by showing that the large-sample results derived there for particular cases apply to optimization-based methods generally.

Section 2 contains the necessary background and notation. In the interests of making this paper somewhat self-contained, several ideas and examples from Windham (1987) are included there. Section 3 gives the general large-sample results, and discusses when optimization-based methods will satisfy the required conditions. Some general observations on the lack of consistency of the resulting estimators and several numerical examples are given in Section 4.

2. Notation and Background

We consider a d by n matrix $\mathbf{X} = (\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ of n d-dimensional observations $\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n$ to be grouped into k classes characterized in the aggregate by a vector parameter θ taking values in some set Θ. Often, but not always, θ will be of the form $\theta = (\theta_1,\ldots,\theta_k : \sigma)$, where θ_j denotes the parameters (such as location parameters) characterizing class j, and σ denotes the parameters (such as common scale parameters) common to all classes. Let the vector $\mathbf{a} = (\mathbf{a}_1,\ldots,\mathbf{a}_k)$ denote the *degrees of membership* of a generic observation $\mathbf{x}$ in each of the k classes, and let $\mathbf{a}_1 = (\mathbf{a}_{i1},\mathbf{a}_{i2},\ldots,\mathbf{a}_{ik})$ denote them for a particular observation $\mathbf{x}_i$. The entries $\mathbf{a}_{ij}$ are restricted to be nonnegative numbers satisfying $\Sigma_j\, \mathbf{a}_{ij} = 1$. When the $\mathbf{a}_{ij}$ take on only the values 0 or 1, we have a *partition*, and $\mathbf{a}_{ij} = 1$ means that $\mathbf{x}_i$ is assigned to class j. We may interpret values of $\mathbf{a}_{ij}$ between 0 and 1 as probabilities of membership in class j, or (in the language of fuzzy sets) degrees of membership, though neither of these interpretations is required. We call the n by k matrix A whose i-th row is $\mathbf{a}_i$ a *standard classification matrix*, and let A^* denote the set of all such matrices. The degrees of membership play key roles in optimization-based methods, as will be seen.

In what follows, the ranges of the indices i and j in summations are assumed to be 1 to n and 1 to k, respectively. *Optimization-based* clustering methods include:

Classification methods: those which choose the classifications $\mathbf{A} \in A^*$ so as to minimize some function $C(\mathbf{A},\mathbf{X})$;

Estimation (or cluster description) methods: those which choose parameters $\theta \in \Theta$ so as to minimize some function $E(\theta,\mathbf{X})$; and

Combined methods: those which choose $\mathbf{A} \in A^*$ and $\theta \in \Theta$ so as to minimize some function $B(\mathbf{A},\theta,\mathbf{X})$.

In a combined method specified by B, let $\hat{\mathbf{A}} = \hat{\mathbf{A}}(\theta,\mathbf{X})$ be the value of $\mathbf{A}$ which minimizes B as a function of $\mathbf{A}$. Then $E(\theta,\mathbf{X}) = B(\hat{\mathbf{A}},\theta,\mathbf{X})$ is an estimation method *generated* by $B(\mathbf{A},\theta,\mathbf{X})$. Similarly, if $\hat{\theta} = \hat{\theta}(\mathbf{A},\mathbf{X})$ minimizes B as a function of θ, $C(\mathbf{A},\mathbf{X}) = B(\mathbf{A},\hat{\theta},\mathbf{X})$ is a classification method generated by $B(\mathbf{A},\theta,\mathbf{X})$. Windham (1987) and Bryant (1988) showed that given a classification method or an estimation method, there exists a combined method which generates it. Without loss of generality, then, we may restrict attention to combined methods.

The following examples illustrate the range of methods included. Consult Windham (1987) for further discussion and references. It is convenient to have the following terms defined:

$$n_j = \Sigma_i a_{ij} \,,$$

$$\bar{\mathbf{x}}_j = [\Sigma_i a_{ij} \mathbf{x}_i] \,/\, n_j \,,$$

$$\mathbf{W}_j = \Sigma_i a_{ij} (\mathbf{x}_i - \bar{\mathbf{x}}_j)(\mathbf{x}_i - \bar{\mathbf{x}}_j)^t \text{ , and}$$

$$\mathbf{W} = \Sigma_j \mathbf{W}_j \,,$$

where $\mathbf{x}^t$ denotes the transpose of $\mathbf{x}$. These are the analogues of the number of observations "assigned" to class j, the mean of class j, and the total scatter matrix.

Example 1. The trace criterion. Let $\theta = (\mu_1, \ldots, \mu_k)$, and let $B(\mathbf{A},\theta,\mathbf{X}) = \Sigma_i \Sigma_j a_{ij} \,|\mathbf{x}_i - \mu_j|^2$, where $|\;\;|$ denotes the Euclidean norm. The optimal value of μ_j is $\hat{\mu}_j = \bar{\mathbf{x}}_j$, and $B(\mathbf{A},\hat{\theta},\mathbf{X}) = \Sigma_j \, tr\,(\mathbf{W}_j)$. This is essentially the k-means criterion.

Example 2. The within-cluster determinant criterion. Let $B(\mathbf{A},\theta,\mathbf{X}) = \Sigma_i \Sigma_j a_{ij} (\mathbf{x}_i - \mu_j)\mathbf{M}_j(\mathbf{x}_i - \mu_j)^t$, where the $\mathbf{M}_j$ are positive definite symmetric d by d matrices with determinant $(1/d)^d$. Here $\theta = (\mu_1, \ldots, \mu_k, \mathbf{M}_1, \ldots, \mathbf{M}_k)$, the optimal values are $\hat{\mu}_j = \bar{\mathbf{x}}_j$, $\hat{\mathbf{M}}_j = d^{-1} \det(\mathbf{W}_j)^{1/d} (\mathbf{W}_j)^{-1}$, and $B(\mathbf{A},\hat{\theta},\mathbf{X}) = \Sigma_j(\det \mathbf{W}_j)^{1/d}$.

Example 3. The determinant criterion. Let $B(\mathbf{A},\theta,\mathbf{X}) = \Sigma_i\Sigma_j a_{ij}(\mathbf{x}_i - \mu_j)\mathbf{M}(\mathbf{x}_i - \mu_j)'$, where $\mathbf{M}$ is a positive definite d by d matrix with determinant $(1/d)^d$. For this case, $\theta = (\mu_1, \ldots, \mu_k, \mathbf{M})$, $\hat{\mu}_j = \bar{\mathbf{x}}_j$, $\hat{\mathbf{M}} = d^{-1} \det(\mathbf{W})^{1/d}(\mathbf{W})^{-1}$, and $B(\mathbf{A},\hat{\theta},\mathbf{X}) = (\det \mathbf{W})^{1/d}$.

Example 4. Mixture analysis. Let $f_j(\mathbf{x},\omega)$ $(j = 1, \ldots, k)$ be probability densities, and let $B(\mathbf{A},\theta,\mathbf{X}) = -\Sigma_i\Sigma_j a_{ij} \log\{p_j f_j(\mathbf{x}_i,\omega)\} + \Sigma_i\Sigma_j a_{ij} \log(a_{ij})$. Here $\theta = (p_1, \ldots, p_k, \omega)$ and $\Sigma_j p_j = 1$. For this case, $\hat{a}_{ij} = [p_j f_j(\mathbf{x}_i,\omega)] / \Sigma_j[p_j f_j(\mathbf{x}_i,\omega)]$, and $B(\hat{\mathbf{A}},\theta,\mathbf{X}) = -\Sigma_i \log\{\Sigma_j p_j f_j(\mathbf{x}_i,\omega)\}$. Thus, B generates the estimation procedure known as mixture maximum likelihood.

Example 5. Classification Maximum Likelihood. Let $B(\mathbf{A},\theta,\mathbf{X}) = -\Sigma_i\Sigma_j a_{ij} \log\{f_j(\mathbf{x}_i,\omega)\}$, where f_j are as in Example 4. For this example, $\hat{a}_{ij} = 1$ when $j = \arg\max\{f_j(\mathbf{x}_i,\omega)\}$ and zero otherwise, and $B(\hat{\mathbf{A}},\theta,\mathbf{X}) = -\Sigma_i \max_j[\log\{f_j(\mathbf{x}_i,\omega)\}]$.

Example 6. Penalized Classification Maximum Likelihood. Let $B(\mathbf{A},\theta,\mathbf{X}) = -\Sigma_i\Sigma_j a_{ij} \log\{p_j f_j(\mathbf{x}_i,\omega)\}$, where f_j and p_j are as in Example 4. This example differs from Example 5 by the penalty term $\Sigma_i\Sigma_j a_{ij} \log(p_j)$, and is discussed further in Section 4.

Example 7. The fuzzy k-means criterion. Let $\theta = (\mu_1, \ldots, \mu_k)$, and let $B(\mathbf{A},\theta,\mathbf{X}) = \Sigma_i\Sigma_j(a_{ij})^2 \,|\mathbf{x}_i - \mu_j|^2$. The optimal value of a_{ij} is $\hat{a}_{ij} = |\mathbf{x}_i - \mu_j|^{-2} / \Sigma_j |\mathbf{x}_i - \mu_j|^{-2}$, and $B(\hat{\mathbf{A}},\theta,\mathbf{X}) = \Sigma_i\{\Sigma_j |\mathbf{x}_i - \mu_j|^{-2}\}^{-1}$.

In all of these examples, $B(\mathbf{A},\theta,\mathbf{X})$ is of the form

$$B(\mathbf{A},\theta,\mathbf{X}) = \Sigma_i b(\mathbf{a}_i,\theta,\mathbf{x}_i)\,, \tag{2.1}$$

for some function $b(\mathbf{a},\theta,\mathbf{x})$, and it is this form which allows application of some standard results of mathematical statistics. Let us call such a method a *linear optimization-based (LOB) method* for short.

3. Large Sample Properties for LOB Methods

3.1 Main Results

Let us now suppose that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ are a random sample from some distribution with density $f(\mathbf{x})$ with respect to some underlying measure u, and denote by $E\{\cdot\}$ expectations with respect to this density. Let $\hat{\theta}_n$ be the (random) optimal value of θ determined by the $\mathbf{x}$'s using a given LOB method $B(\mathbf{A},\theta,\mathbf{X}) = \Sigma_i b(\mathbf{a}_i,\theta,\mathbf{x}_i)$ — i.e., that value of θ which minimizes $B(\hat{\mathbf{A}},\theta,\mathbf{X})$ as a function of θ. The large sample properties of $\hat{\theta}_n$ may then be summarized as follows.

Theorem. *If the generalized Wald sufficient conditions are satisfied, then as n becomes large* $\hat{\theta}_n$ *converges to* θ_0 *almost surely, where* θ_0 *is the value of* θ *which minimizes* $Q(\theta) = E\{b[\hat{\mathbf{a}},\theta,\mathbf{x}]\}$ *as a function of* θ. *If, in addition, the normality conditions are satisfied, then the distribution of* $n^{1/2}(\hat{\theta}_n - \theta_0)$ *approaches a normal distribution with mean vector* $\mathbf{0}$ *and variance-covariance matrix* $\mathbf{P}^{-1}\mathbf{H}\mathbf{P}^{-1}$ *where*

$$\mathbf{P} = \{P_{ij}\} = \{(\partial^2 / \partial\theta_i\partial\theta_j)Q(\theta)\},$$

$$\mathbf{H} = \{H_{ij}\} = E\{b_i(\theta,\mathbf{x})b_j(\theta,\mathbf{x})\}, \text{ and}$$

$$b_i(\theta,\mathbf{x}) = (\partial / \partial\theta_i)b(\hat{\mathbf{a}},\theta,\mathbf{x})$$

The conditions referred to in the theorem are:

Generalized Wald Sufficient Conditions (for a.s. convergence):

(R1) Θ is a closed subset of R^m, for some m.

(R2) For each $\theta \in \Theta$, there is a neighborhood of θ and a function $L(\mathbf{x},\theta)$ with $E\{L^3(\mathbf{x},\theta)\}$ finite such that for all θ' and θ'' in the neighborhood,

$$|b(\hat{\mathbf{a}},\mathbf{x},\theta') - b(\hat{\mathbf{a}},\mathbf{x},\theta'')| < L(\mathbf{x},\theta)|\theta' - \theta''|.$$

(R3) $Q(\theta)$ has a unique minimum at $\theta_0 \in \Theta$.

(R4) If Θ is not bounded, then there exists a compact set $C \subset \Theta$ for which $E\{\sup b(\hat{\mathbf{a}},\mathbf{x},\theta)\}$ is finite, where the supremum is over those θ not in C.

(R5) If Θ is not bounded, then except for $\mathbf{x}$ in a set of u-measure zero, as $\min_j|\theta_j|$ tends to infinity, so does $b(\hat{\mathbf{a}},\mathbf{t},\mathbf{x})$.

For the case in which Θ is not bounded, (R4) and (R5) will basically allow us to confine our search for $\hat{\theta}$ to the compact set C eventually. Under these conditions, the proof of almost sure convergence is the same as that in Bryant and Williamson (1978), using $b(\hat{a},\theta,\mathbf{x})$ in place of the $ln\,[f_{i(\mathbf{x},\theta)}(\mathbf{x},\theta)]$ used there.

Normality Conditions:

(W1) $Q(\theta) - Q(\theta_0) = (1/2)(\theta - \theta_0)' \, \mathbf{M}'' (\theta - \theta_0) + |\theta - \theta_0|^2 h(\theta - \theta_0)$, where $\mathbf{M}''$ is megative definite, $h(v)$ tends to zero as v does, and in some neighborhood of θ_0, the first partial derivatives of h exist and are continuous except possibly at θ_0, and are bounded by $\Gamma / |\theta - \theta_0|^{\gamma}$, for some constant Γ and some $\gamma < 1$.

(W2) For each $\theta \in \Theta$, the partial derivatives b_j exist almost surely (with respect to f), and as a function of θ_j alone, b_j has only finitely many discontinuities, at least over some bounded open interval containing θ_0 for each $\mathbf{x}$.

(W3) For each compact subset C of Θ, and each $\delta > 0$, there exists a random variable $Z(\mathbf{x},C,\delta)$ with $E\{Z\} = O(\delta)$ and $E\{Z^2\} = O(\delta)$, such that for $i = 1, \ldots, k$,

$$\sup |b_i(\theta',\mathbf{x}) - b_i(\theta'',\mathbf{x})| < Z(\mathbf{x},C,\delta),$$

where the supremum is over the set where $|\theta' - \theta''| < \delta$ and $\theta', \theta'' \in C$.

Under these conditions, Daniels' (1961) proof as augmented by Williamson (1984) can be extended to the case where θ is multidimensional, as is done in Bryant and Williamson (1984).

3.2 Remarks

The theorem in the previous section is one of many generalizations of the corresponding classical results for maximum likelihood estimators. Example 5 with $k = 1$ reduces to ordinary maximum likelihood, for example. The results are known in various forms and under a variety of conditions. The basic result (in reasonable generality) is given by Huber (1967) in the context of M-estimators. His formulation has the advantage of applying to an open parameter set, which would include the usual scale parameters, but his conditions are hard to verify in practice. In Example 5 and similar examples, the derivatives b_j are often discontinuous at a few points, and this situation motivated the normality conditions above, which include such cases and are typically easier to verify in classification problems. Note that the smoothness required for the asymptotic normality is on $Q(\theta)$, not on b. Even when the b_j are discontinuous, the expectations involved will normally have the required smoothness, as the discontinuities typically are isolated. The restriction to a closed set is unappealing mathematically, but is of little practical importance, as we would almost always be willing to restrict a variance (say) to be greater than or equal to *some* number (10^{-9}, say) greater than zero. Recently, Bardwell (1989) has used an ingenious argument to show how this restriction can sometimes be removed.

When $b(\hat{a},\theta,\mathbf{x})$ has continuous third partial derivatives which in some neighborhood of θ_0 are bounded by an f-integrable function which is independent of θ, an analogue of the usual proofs for maximum likelihood for smooth conditions applies here, and the rather technical conditions (W1)-(W3) need not be verified. These results are discussed from other points of view by Pollard (1981, 1982), Foutz and Srivastava (1977), Hartigan (1978), White (1982), Bock (1985), Dupacova and Wets (1988), Windham (1989) and Boente and Fraiman (1988) among others.

3.3 Application to LOB Methods

Whether or not the conditions of the theorem are satisfied depends on the assumptions made about the particular function b and the underlying probability density, of course, but it will often be possible to demonstrate that they are satisfied. As Windham points out, in all the common examples, including those in Section 2, b takes one of the following three forms, except possibly for additive or multiplicative constants:

Form 1: $b(\mathbf{a},\theta,\mathbf{x}) = \Sigma_j a_j D_j(\mathbf{x},\theta)$;
Form 2: $b(\mathbf{a},\theta,\mathbf{x}) = \Sigma_j a_j D_j(\mathbf{x},\theta) + \Sigma_j a_j \log(a_j) + \log(k)$;
or
Form 3: $b(\mathbf{a},\theta,\mathbf{x}) = \Sigma_j k^{\gamma}(a_j)^{1+\gamma} D_j(\mathbf{x},\theta)$, for some $\gamma > 0$,

where $D_j(\mathbf{x},\theta)$ is some measure of the discrepancy between the observation x and class j as described by θ. Typically, $D_j = |\mathbf{x} - \theta_j|^2$, for some norm, or $D_j = -\log\{f_j(\mathbf{x},\theta)\}$, or something similar. When the optimal values of a are used, we obtain:

Form 1: $b(\hat{a},\theta,\mathbf{x}) = \min_j\{D_j(\mathbf{x},\theta)\}$;
Form 2: $b(\hat{a},\theta,\mathbf{x}) = -\log[k^{-1}\,\Sigma_j \exp\{-D_j(\mathbf{x},\theta)\}]$
Form 3: $b(\hat{a},\theta,\mathbf{x}) = [k^{-1}\Sigma_j\{D_j(\mathbf{x},\theta)\}^{-1/\gamma}]^{-\gamma}$.

In each of these cases, b is some sort of "combination" of the discrepancies D_j:

$$b(\mathbf{a},\theta,\mathbf{x}) = m[D_1(\mathbf{x},\theta), \ldots, D_k(\mathbf{x},\theta)]\,,$$

where the function $m(\mathbf{y}) = m(y_1,y_2,\ldots,y_k)$ satisfies the following conditions:

(C1): m is symmetric and continuous in its arguments.

(C2): $\min_j\{y_j\} \le m(y) \le \max_j\{y_j\}$.
(C3): The partial derivatives $m_j = \partial m / \partial y_j$ exist and are continuous except possibly for a set of Lebesgue measure zero.
(C4): There exists a constant K such that:
$0 \le m_j(\mathbf{y}) \le K$
$\Sigma_j m_j(\mathbf{y}) \le K$
$|m(\mathbf{y}) - m(\mathbf{z})| \le K\ |\mathbf{y} - \mathbf{z}|$

These properties allow us in many cases to verify the conditions of the theorem by verifying the corresponding properties for the discrepancies $D_j(\mathbf{x},\theta)$. The problem is thus reduced to the corresponding problem for $k = 1$. For example, it is easy to show that under Conditions C1-C4:

(1) R2 will be satisfied if the D_j satisfy it, that is R2 holds with $b(\hat{\mathbf{a}},\theta,\mathbf{x})$ replaced by $D_j(\mathbf{x},\theta)$. In the important example in which $D_j(\mathbf{x},\theta) = |\mathbf{x} - \theta_j|^2$, $|D_j(\mathbf{x},\theta_1) - D_j(\mathbf{x},\theta_2)| \le 2\{|\theta - \mathbf{x}| + \varepsilon\}\ |\theta_1 - \theta_2|$ when θ_1 and θ_2 are within ε of θ. R2 will then be staisfied if $\mathbf{x}$ has finite third moments.
(2) R4 and R5 will be staisfied if the D_j satisfy them. When $D_j(\mathbf{x},\theta) = |\mathbf{x} - \theta_j|^2$, R5 is automatically satisfied, and R4 will be satisfied for such distributions as the normal.
(3) R3 must generally be verified independently. As some of the examples in the next section show, R3 is not always satisfied.
(4) If the D_j have continuous third derivatives with the appropriate bounds, W1-W3 will be satisfied. This will not be true for b's of Form 1, though, when W1-W3 must be verified independently. Since the discontinuities of m_j typically occur where f assigns zero measure, this is usually demonstrable.

For LOB methods satisfying C1-C4, then, we can usually demonstrate that the conditions of the theorem are satisfied.

4. Consistency Results

4.1 Introduction

When the true density is $f = f(\theta^*,\mathbf{x})$, a member of some parametric family, with true value of $\theta = \theta^*$, it is not necessarily the case that the limiting value $\theta_0 = \theta^*$, and thus $\hat{\theta}_n$ may be an asymptotically biased (inconsistent) estimator of the true value θ^*. This result can also be seen in the following formulation.

By the Law of Large Numbers, $n^{-1}B(\hat{\mathbf{A}},\theta_n,\mathbf{X})$ tends to $Q(\theta_0)$ as n becomes large. For any given $b(\mathbf{a},\mathbf{x},\theta)$, let $I(\theta) = \int \exp\{-b[\hat{\mathbf{a}},\theta,\mathbf{x}]\}d\mathbf{x}$, and let $b^*(\mathbf{x},\theta) = \exp\{-b(\hat{\mathrm{a}},\mathbf{x},\theta)\} / I(\theta)$, so that b^* is a probability density. Note that

$$Q(\theta_0) = K(\theta_0) + E\{-\log f(\mathrm{x})\} - \log I(\theta_0)\,, \tag{4.1}$$

where

$$K(\theta) = E\{-\log\,[b^*(\hat{\mathrm{a}},\theta,\mathrm{x}) / f(\mathrm{x})]\}$$

is the Kullback-Leibler measure of discrepancy between b^* and f. Our limiting minimum value of b, then, is $E\{-\log f(\mathrm{x})\}$, the entropy inherent in the true distribution of the data, modified by two additive terms: $\log I(\theta_0)$, a modification because our function e^{-b} does not necessarily integrate to 1 (is not a probability density), and $K(\theta_0)$, a modification because even if e^{-b} *did* integrate to 1, it would not equal the true density f. It is standard to show that K is positive unless $b^* = f$ almost surely. We may interpret this as a reminder that not knowing the form of f can only hurt us. If we were using $b = -\log f$, we would always get a lower minimum, and the minimum would occur at $\theta = \theta_0$. In general, then, there is no reason to expect that estimates so derived will be consistent. Indeed, they often are not. Further, in the light of (4.1) no simple modification can be expected to correct the problem, unless it is based either on knowledge of the form of the true distribution or else somehow adapts to it for large samples. Several approaches using penalty functions have been tried, though, and the theorem in Section 3 allows us to assess (at least asmptotically) their effects.

4.2 Examples

In the context of clustering methods, the first use of these ideas seems to be Marriott (1975) and Bryant and Williamson (1978), where it is noted that estimates derived by classification maximum likelihood (CML, Example 5) are generally inconsistent. In this section we illustrate such inconsistency and evaluate two attempts to remove the inconsistency by using "penalty" functions. We limit attention to the case when $k = 2$. The results for CML from Bryant and Williamson (1978) are repeated here to facilitate comparison with the other two attempts.

The CML procedure in such a case is a LOB method with

$$b(a,\theta,\mathrm{x}) = -\log\,\{[f_1(\theta,\mathrm{x})]^a\,[f_2(\theta,\mathrm{x})]^{1-a}\}$$

in the variation considered by Scott and Symons (1971), Marriott (1975) and Bryant and Williamson (1978, 1986). Let us call it CML variation 1.

Equation (4.1) suggests that using $b + \log I(\theta)$ instead of b might serve to improve the performance. Unless $b = -\log f$ exactly, of course, we cannot guarantee consistency, but perhaps we can come closer by including some penalty like $\log I(\theta)$ which accounts for the fact that e^{-b} is not a probability density. One approach to this is the penalized CML approach (Example 6) in which

$$b(a,\theta,\mathrm{x}) = -\log\{[pf_1(\theta,\mathrm{x})]^a\,[(1-p)f_2(\theta,\mathrm{x})]^{1-a}\}\ .$$

This has been considered by Symons (1981), Marriott (1982), and Windham (1986). Let us call it CML variation 2. It has the advantage that the penalty term does not depend on the functional form of the f_j, but since in practice, we know the form of the f_j, though not the form of the true f, we could also use $\log\{I(\theta)\}$ itself:

$$\begin{aligned} b(a,\theta,\mathrm{x}) = &-\log\{[f_1(\theta,\mathrm{x})]^a\,[f_2(\theta,\mathrm{x})]^{1-a}\} \\ &+\log\{I(\theta)\}\ . \end{aligned}$$

Let us call this one CML variation 3.

We now consider the univariate case, with f_j equal to the normal density with mean μ_j and standard deviation $\sigma(\mu_1 < \mu_2)$. For any of the three variations of CML, one can show that

$$\hat{a} = 1 \text{ if } x < M$$

and 0 otherwise, where

$$M = (\mu_1 + \mu_2)/2$$

for variation 1 or variation 3, and

$$M = (\mu_1 + \mu_2)/2 + \{\sigma^2 \log[p/(1-p)]/(\mu_2-\mu_1)\}$$

for variation 2. Using these results and the fact that for this case, $I(\theta) = 2\Phi\{\mu_2 - \mu_1)/(2\sigma)\}$, where Φ denotes the cumulative standard normal distribution, direct calculations show that if there is a global minimum in the interior of Θ, the optimal values θ_0 satisfy

$$\begin{aligned} &p_0 = F(M) \\ &F(M)\mu_{10} = \int_{x<M} x\,dF(x) + \sigma_0(\Delta/2) \\ &[1-F(M)]\mu_{20} = \int_{x>M} x\,dF(x) - \sigma_0(\Delta/2) \end{aligned}$$

$$(\sigma_0)^2(1-\delta\Delta) = \int_{x<M}(x-\mu_{10})^2 dF(x) + \int_{x>M}(x-\mu_{20})^2 dF(x)$$

where $\delta = (\mu_{20} - \mu_{10}) / \sigma_0$, $\Delta = \Phi'(\delta/2) / \Phi(\delta/2)$ for variation 3 and $\Delta = 0$ for variations 1 and 2. If the true mean and variance of X are $E\{X\} = \mu_0$ and variance $(X) = V^2$, respectively, then

$$P_0\mu_{10} + (1-p_0)\mu_{20} = \mu_0 \text{ and}$$
$$(\sigma_0)^2 = V^2 - p_0(1-p_0)(\mu_{20}-\mu_{10})^2 \ .$$

Suppose further that the true underlying distribution $F(x)$ is a mixture of two normals differing in mean with a common variance. Conditions R1-R5 and W1-W3 follow easily by direct calculation of the various expectations, except for R3, which must be verified numerically. The equations above may then be solved numerically to yield the limiting values. Such values are given for typical parameters in Tables 1, 2, and 3. Note that for variations 2 and 3, it is sometimes the case that the minimum in the interior is a local minimum only. The minimum obtained there is greater than the minimum obtained by letting the cut-point M tend to either plus or minus infinity (in effect, classifying all the data in one class). Intuitively, this means that while classifying into two populations improves the fit (and gives a lower minimum) compared to a model using a single normal distribution, the improvement is not so much as to compensate for the ''penalty'' term in Q.

Table 2 also lists for each value of π^* the minimum separation required between the means for the interior minimum to be a global minimum. In those cases where the convergence of CML variation 2 is assured, the limiting values are perhaps marginally better than those for variation 1, but this happens only for well-separated components and the biases in such cases are small in any case.

For variation 3, things are a bit different. As with variation 2, the optimum does not always occur in the interior, but for each given distance between the means, there is a range of values of the mixing proportion π^* for which the optimum occurs in the interior, roughly .22 to .78. Exact values are given in Table 3. There seems to be little reason to prefer the limiting values for variation 3 to those of either variation 1 or variation 2.

In short, the large sample results of Section 3 show that there is some evidence that penalty functions can reduce the asymptotic bias in CML estimates, but not in a uniformly useful way. When the components are well separated and the mixing proportions are not extreme, the bias is reduced, but

TABLE 1.

LIMITING VALUES OF CML VARIATION 1 ESTIMATES
FOR TWO NORMALS DIFFERING IN MEAN

μ_2^*	π^*	μ_{10}	μ_{20}	σ^2_0	p_0
1	.1	.048	1.713	.398	.488
	.2	-.087	1.632	.422	.484
	.3	-.204	1.555	.437	.486
	.4	-.306	1.477	.445	.492
	.5	-.396	1.396	.448	.500
2	.1	.735	2.593	.516	.427
	.2	.348	2.440	.589	.401
	.3	.098	2.336	.621	.418
	.4	-.059	2.253	.635	.455
	.5	-.167	2.167	.639	.500
3	.1	1.122	3.391	.720	.305
	.2	.345	3.200	.796	.280
	.3	.108	3.143	.813	.344
	.4	.003	3.103	.819	.420
	.5	-.059	3.059	.821	.500
4	.1	.560	4.102	.913	.142
	.2	.139	4.060	.927	.219
	.3	.048	4.046	.930	.312
	.4	.007	4.033	.931	.406
	.5	-.016	4.017	.932	.500

Notes:
(1) Values tabulated are limiting values of estimates of the group means, common variance and proportion in the lower class, derived as described in the text.

(2) True distribution is the mixture $\pi^*N(0,1)+(1-\pi^*)N(\mu_2^*,1)$.

TABLE 2.

LIMITING VALUES OF CML VARIATION 2 ESTIMATES
FOR TWO NORMALS DIFFERING IN MEAN

π^*	$\mu_{2,min}$	μ_2^*	μ_{10}	μ_{20}	σ^2_0	p_0
.1	2.597	3	-.441	2.959	.998	.076
		4	-.102	3.995	.975	.096
.2	2.822	3	-.219	2.987	.901	.183
		4	-.056	4.000	.953	.197
.3	2.936	3	-.131	3.014	.851	.291
		4	-.036	4.005	.941	.298
.4	3.010	4	-.025	4.011	.934	.399
.5	3.036	4	-.017	4.017	.931	.500

Notes:
(1) Values tabulated are limiting values of estimates of the group means, common variance and proportion in the lower class, derived as described in the text.

(2) True distribution is the mixture $\pi^*N(0,1)+(1-\pi^*)N(\mu_2^*,1)$.

(3) Values of $\mu_{2,min}$ are the minimum values of μ_2^* for which Variation 2 of CML has an interior minimum, as described in the text.

TABLE 3.

LIMITING VALUES OF CML VARIATION 3 ESTIMATES
FOR TWO NORMALS DIFFERING IN MEAN

μ_2^*	π_{min}	π^*	μ_{10}	μ_{20}	σ^2_0	p_0
1	.214	.25	.736	.763	1.175	.494
		.30	.666	.733	1.181	.495
		.40	.536	.663	1.183	.497
		.50	.425	.575	1.183	.500
2	.220	.25	1.353	1.626	1.605	.462
		.30	1.045	1.703	1.478	.461
		.40	.620	1.722	1.327	.474
		.50	.387	1.613	1.285	.500
3	.225	.25	1.382	2.782	1.765	.380
		.30	.706	2.927	1.346	.372
		.40	.325	2.907	1.205	.429
		.50	.173	2.827	1.182	.500
4	.227	.25	.293	4.001	1.116	.270
		.30	.201	3.993	1.099	.315
		.40	.107	3.972	1.087	.407
		.50	.058	3.942	1.084	.500

Notes:

(1) Values tabulated are limiting values of estimates of the group means, common variance and proportion in the lower class, derived as described in the text.

(2) True distribution is the mixture $\pi^*N(0,1)+(1-\pi^*)N(\mu_2^*,1)$.

(3) Values of π_{min} are the minimum values of π* for which Variation 3 of CML has an interior minimum, as described in the text.

it was small then to begin with. For ill-separated components or extreme values of the mixing proportions, the penalized CML approaches don't classify at all, in effect. This may be a useful result when we wish to determine if classes are really present, but it's not really helpful for purposes of reducing bias in the estimates of the cluster characteristics in the case in which we can assume the clusters really are present.

References

BARDWELL, R. A. (1989), "Asymptotic Behavior of Certain Estimators Under Mild Regularity Conditions," Ph.D. dissertation, Department of Mathematics, University of Colorado at Boulder.

BOCK, H.-H. (1985), "On Some Significance Tests in Cluster Analysis," *Journal of Classification, 2,* 77-108.

BOENTE, G., and FRAIMAN, R. (1988), "On the Asymptotic Behaviour of General Maximum Likelihood Estimates for the Nonregular Case Under Nonstandard Conditions," *Biometrika, 75,* 45-56.

BRYANT, P. G. (1988), "On Characterizing Optimization-Based Clustering Criteria," *Journal of Classification, 5,* 81-84.

BRYANT, P. G., and WILLIAMSON, J. A. (1978), "Asymptotic Behaviour of Classification Maximum Likelihood Estimates," *Biometrika, 65*, 273-281.

BRYANT, P. G., and WILLIAMSON, J. A. (1984), "The Asymptotic Distribution of Statistics Derived by Maximizing Sums," Faculty Working Paper Series number UCD-CBA 1984-3, College of Business and Administration, University of Colorado at Denver.

BRYANT, P. G., and WILLIAMSON, J. A. (1986), "Maximum Likelihood and Classification: A Comparison of Three Approaches," in *Classification as a Tool of Research*, Eds. W. Gaul and M. Schader, Amsterdam: North-Holland, 35-45.

DANIELS, H. E. (1961), "The Asymptotic Efficiency of a Maximum Likelihood Estimator," in *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1*, Ed. J. Neyman, Berkeley and Los Angeles: University of California Press, 151-163.

DAY, W. H. E., and EDELSBRUNNER, H. (1985), "Investigation of Proportional Link Linkage Clustering Methods," *Journal of Classification, 2*, 239-254.

DUPACOVA, J., and Wets, R. (1988), "Asymptotic Behavior of Statistical Estimators and of Optimal Solutions of Stochastic Optimization Problems," *Annals of Statistics, 16, 4*, 1517-1549.

FOUTZ, R. V., and SRIVASTAVA, R. C. (1977), "The Performance of the Likelihood Ratio Test When the Model is Incorrect," *Annals of Statistics, 5*, 1183-1194.

HARTIGAN, J. A. (1978), "Asymptotic Distributions for Clustering Criteria," *Annals of Statistics, 6*, 117-131.

HUBER, P. J. (1967), "The Behavior of Maximum Likelihood Estimates under Non-standard Conditions," in *Proceedings, Fifth Berkeley Symposium on Mathematical Statistics and Probability, 1*, Eds. L. M. Le Cam and J. Neyman, Berkeley and Los Angeles: University of California Press, 221-233.

MARRIOTT, F. H. C. (1975), "Separating Mixtures of Normal Dsitributions," *Biometrics, 31*, 767-769.

MARRIOTT, F. H. C. (1982), "Optimization Methods of Cluster Analysis," *Biometrika, 69*, 417-421.

POLLARD, D. (1981), "Strong Consistency of k-means Clustering," *Annals of Statistics, 9*, 135-140.

POLLARD, D. (1982), "A Central Limit Theorem for k-means Clustering," *Annals of Probability, 10*, 919-926.

SCOTT, A. J., and SYMONS, M. J. (1971), "Clustering Methods Based on Likelihood Ratio Criteria," *Biometrics, 27*, 387-397.

SYMONS, M. J. (1981), "Clustering Criteria and Multivariate Normal Mixtures," *Biometrics, 37*, 35-43.

WHITE, H. (1982), "Maximum Likelihood Estimation of Misspecified Models," *Econometrica, 50*, 1-25.

WILLIAMSON, J. A. (1984), "A Note on the Proof by H. E. Daniels of the Asymptotic Efficiency of a Maximum Likelihood Estimator," *Biometrika, 71*, 651-653.

WINDHAM, M. P. (1986), "A Unification of Optimization-Based Numerical Classification Algorithms," in *Classification as a Tool of Research*, Eds. W. Gaul and M. Schader, Amsterdam: North-Holland, 447-452.

WINDHAM, M. P. (1987), "Parameter Modification for Clustering Criteria," *Journal of Classification, 4*, 191-214.

WINDHAM, M. P. (1989), "Statistical Models in Cluster Analysis," Utah State University, Department of Mathematics and Statistics Research Report May/1989/45.