BAYESIAN PARTITIONING ANALYSIS WITH TWO COMPONENT MIXTURES

By

Michael J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 691

July 1970

"Bayesian Partitioning Analysis with Two Component Mixtures"

By Michael J. Symons

Department of Biostatistics
University of North Carolina at Chapel Hill

Summary

The problem of partitioning a sample into disjoint and exhaustive sets is examined from a Bayesian point of view. It is assumed that the sample comes from a mixture of two distributions. A definition of the "best" partition is proposed and, on the basis of this definition, a detailed examination is given when the mixture is of two univariate homogeneous normals and when the mixture is of two univariate heterogeneous normals with specified variance ratio. Relation of these results to the "outlier problem" and a multivariate extension with implications in cluster analysis are considered. An analysis with Darwin's data illustrates the application of these ideas to outliers.

1. Introduction

Consider a mixture of two distributions of a p component random vector $\underline{x}$, i.e.

$$f(\underline{x}|\underline{\theta}) = \alpha f_1(\underline{x}|\underline{\theta}_1) + (1-\alpha)f_2(\underline{x}|\underline{\theta}_2). \tag{1.1}$$

the parameter vector is $\underline{\theta} = (\underline{\theta}_1,\underline{\theta}_2,\alpha)$, where $\underline{\theta}_1$ is the vector of parameters for f_1, correspondingly for $\underline{\theta}_2$, and α is the mixing parameter. Denoting a random sample of size n from $f(\underline{x}|\underline{\theta})$ by $\underline{\bar{X}} = (\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n)$, the joint distribution of $\underline{\bar{X}}$ is

$$f(\underline{\bar{X}}|\underline{\theta}) = \prod_{i=1}^{n} [\alpha f_1(\underline{x}_i|\underline{\theta}_1) + (1-\alpha)f_2(\underline{x}_i|\underline{\theta}_2)]. \tag{1.2}$$

Box and Tiao (1968) have shown that expanding the product of sums in (1.2) results in a sum of 2^n terms. Specifically,

$$f(\underline{\bar{X}}|\underline{\theta}) = \sum_{m=0}^{n} \alpha^m(1-\alpha)^{n-m} \sum_{k=1}^{\binom{n}{m}} [\prod_{I_{1mk}} f_1(\underline{x}_i|\underline{\theta}_1) \prod_{I_{2mk}} f_2(\underline{x}_i|\underline{\theta}_2)], \tag{1.3}$$

where $P_{mk}(N) = (I_{1mk},I_{2mk})$ is a partition of the set of subscripts of $\underline{\bar{X}}$, i.e., of $N = \{1,2,\ldots,n\}$, and hence equivalently a partition $\underline{\bar{X}}$. The set $I_{1mk} = \{i|\underline{x}_i$ is assumed distributed as $f_1\}$ and $I_{2mk} = N\backslash I_{1mk}$. The subscript m denotes the number of $\underline{x}_i$ allocated to f_1 by the partition $P_{mk}(N)$, i.e. m is the number of elements in I_{1mk}; k indexes the $\binom{n}{m}$ allocations of m $\underline{x}_i$ to f_1. The range of the index k is $1,2,\ldots,\binom{n}{m}$, so strictly this index should be k_m or k(m) but for simplicity just k is used. We can write

$$f(\underline{\bar{X}}|\underline{\theta}) = \sum_{m=0}^{n} \sum_{k=1}^{\binom{n}{m}} f(\underline{\bar{X}}|P_{mk}(N),\underline{\theta}) f(P_{mk}(N)|\underline{\theta}) \tag{1.4}$$

and hence by Bayes' Theorem

$$f(P_{mk}(N)|\underline{\bar{X}},\underline{\theta}) = \frac{f(\underline{\bar{X}}|P_{mk}(N),\underline{\theta}) f(P_{mk}(N)|\underline{\theta})}{f(\underline{\bar{X}}|\underline{\theta})}, \tag{1.5}$$

where $f(P_{mk}(N)|\underline{\theta}) = \alpha^{m}(1-\alpha)^{n-m}$ and $f(\underline{\bar{X}}|P_{mk}(N),\underline{\theta})$ is given in the brackets of (1.3).

When the parameter $\underline{\theta}$ is specified, the partition corresponding to the largest $f(P_{mk}(N)|\underline{\bar{X}},\underline{\theta})$ is the partition with largest likelihood (posterior probability). In reality $\underline{\theta}$ is seldom even partially specified. In this case, a Bayesian definition of the optimal partition is the $P_{mk}(N)$ corresponding to the largest $f(P_{mk}(N)|\underline{\bar{X}})$, where

$$f(P_{mk}(N)|\underline{\bar{X}}) \propto \int_{\Theta} f(P_{mk}(N)|\underline{\bar{X}},\underline{\theta}) p(\underline{\theta}) d\underline{\theta}; \tag{1.6}$$

$p(\underline{\theta})$ is the prior distribution on $\underline{\theta}$ with parameter space Θ.

The partition so defined has the largest marginal posterior probability; marginality is achieved by averaging the likelihood and prior over the parameter space. Such treatment of $\underline{\theta}$ qualifies it as a vector of "nuisance parameters" in the sense of Lindley (1965). The most probable partition is of interest; no inference concerning the parameters is desired.

The optimal partition of the sample is denoted by $P_{m^*k^*}(N)$ where this partition corresponds to

$$f(P_{m^*k^*}(N)|\underline{\bar{X}}) = \max_{m,k} [f(P_{mk}(N)|\underline{\bar{X}})] \tag{1.7}$$

for $m=0,1,2,\ldots,n$ and for each value of $m, k=1,2,\ldots,\binom{n}{m}$. The real problem of

finding the optimal partition is to locate efficiently the largest of these 2^n terms, (1.6). With a sample size of any consequence, a count, much less a calculation, of the 2^n terms (1.6) is prohibitive. Nevertheless, with some models all 2^n terms need not be examined to find $P_{m^*k^*}(N)$.

2. Mixtures of Two Univariate Normals

Consider a mixture of two univariate normals with variance ratio specified, i.e., (1.1) becomes

$$f(x|\underline{\theta}) = \frac{\alpha}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu_1)^2}{2\sigma^2}\right) + \frac{1-\alpha}{\sqrt{2\pi}\,h\,\sigma} \exp\left(-\frac{(x-\mu_2)^2}{2h^2\sigma^2}\right). \tag{2.1}$$

The means of the components are μ_1 and μ_2; the variances are σ^2 and $h^2\sigma^2$, $h \geq 1$, and α is the mixing parameter. The usual restrictions on these parameters are assumed. When $h>1$, the mixture is heterogeneous, but if $h=1$, (2.1) is a homogeneous mixture of two normals.

The prior distribution assumed on the vector of unknown parameters, $\underline{\theta} = (\mu_1,\mu_2,\sigma^2,\alpha)$, is

$$p(\underline{\theta}) = \left[\frac{1}{2\pi\sigma^2|\underline{\underline{V}}'|^{\frac{1}{2}}} \exp\left(-\frac{1}{2\sigma^2}(\underline{\mu}-\underline{m}')\underline{\underline{N}}'(\underline{\mu}-\underline{m}')^t\right)\right]\left[\frac{1}{\sigma^2}\right]\left[\frac{\Gamma(a_1'+a_2')}{\Gamma(a_1')\Gamma(a_2')}\,\alpha^{a_1'-1}(1-\alpha)^{a_2'-1}\right]. \tag{2.2}$$

This is a bivariate normal prior distribution on the two means $\underline{\mu} = (\mu_1,\mu_2)$ given σ^2 with mean $\underline{m}' = (m_1',m_2')$ and covariance matrix $\sigma^2\underline{\underline{V}}'$, where $\underline{\underline{N}}' = (\underline{\underline{V}}')^{-1}$. A lower case t denotes transpose. When the positive-definite symmetric matrix $\underline{\underline{V}}'$ is diagonal, μ_1 and μ_2 are independent with univariate normal distributions conditional on σ^2. The marginal prior distribution of σ^2 is assumed "flat" on

the logarithm of σ^2. The mixing parameter is considered statistically independent of $\underline{\mu}$ and σ^2 in (2.2); a beta prior is assumed on α with a_1' and a_2' both restricted to be positive.

The choice of prior distributions has been amply discussed by Raiffa and Schlaifer (1961), chapter 3. It should be pointed out that in analyses concerning mixtures, it is not uncommon to assume previous samples from the constituent components, see for example Anderson (1958), Dunsmore (1966), Geisser (1964), and Rao (1952). More than sample information is often assumed on the mixing parameter; α is typically assumed known, e.g., Anderson (1958) or Box and Tiao (1968).

The computation of $f(P_{mk}(N)|\bar{\underline{X}})$ is not difficult once $f(\bar{\underline{X}}|\underline{\theta})$ in (1.3) is written out for (2.1) and combined with $p(\underline{\theta})$ in (2.2). The required integrations for (1.6) are in Raiffa and Schlaifer (1961), chapter 12. The result is

$$f(P_{mk}(N)|\bar{\underline{X}}) = K\Gamma(m+a_1')\Gamma(n-m+a_2')h^{-(n-m)}|\underline{\underline{V}}''_m|^{\frac{1}{2}}(W''_{mk})^{-\frac{1}{2}(n-3)}, \tag{2.3}$$

where K is composed of the constant factors from the prior distribution, constant factors from the likelihood and a normalization factor so that the $f(P_{mk}(N)|\bar{\underline{X}})$ sum over m=0,1,2,...,n and k=1,2,...,$\binom{n}{m}$ to unity. The term W''_{mk} is given as

$$W''_{mk} = \sum_{I_{1mk}} (x_i-\bar{x}_{1mk})^2 + \frac{1}{h^2}\sum_{I_{2mk}} (x_i-\bar{x}_{2mk})^2 + (\bar{\underline{x}}_{mk}-\underline{m}')\underline{\underline{N}}^*_m(\bar{\underline{x}}_{mk}-\underline{m}')^t, \tag{2.4}$$

where $\bar{\underline{x}}_{mk} = (\bar{x}_{1mk},\bar{x}_{2mk})$ and $\bar{x}_{1mk}$ is the average of the observations allocated to f_1, viz., $\bar{x}_{1mk} = \frac{1}{m}\sum_{I_{1mk}} x_i$, and similarly for $\bar{x}_{2mk}$. When m=0 or n some caution is necessary; examining $f(\bar{\underline{X}}|\underline{\theta})$ for these two cases will indicate the appropriate forms. The matrix $\underline{\underline{N}}^*_m = \underline{\underline{N}}_m(\underline{\underline{N}}'')^{-1}\underline{\underline{N}}'$, $\underline{\underline{N}}''_m = \underline{\underline{N}}_m+\underline{\underline{N}}'$ and

$$\underline{\underline{N}}_m = \begin{pmatrix} m & 0 \\ 0 & \frac{n-m}{h} \end{pmatrix}. \tag{2.5}$$

The matrix $\underline{\underline{V}}''_m$ is the inverse of $\underline{\underline{N}}''_m$.

As a special case of the above, with h=1 and a diagonal $\underline{\underline{V}}'$, we can deduce from the previous result a form for the homogeneous mixture with a prior distribution of

$$p(\underline{\theta}) = [\frac{\sqrt{n_1'}}{\sqrt{2\pi}\,\sigma} \exp(-\frac{n_1'(\mu_1-m_1')^2}{2\sigma^2})][\frac{\sqrt{n_2'}}{\sqrt{2\pi}\,\sigma} \exp(-\frac{n_2'(\mu_2-m_2')^2}{2\sigma^2})]$$

$$[\frac{1}{\sigma^2}][\frac{\Gamma(a_1'+a_2')}{\Gamma(a_1')\Gamma(a_2')} \alpha^{a_1'-1} (1-\alpha)^{a_2'-1}]. \tag{2.6}$$

The form for this important special case is

$$f(P_{mk}(N)\,|\,\underline{\bar{X}}) = J \frac{\Gamma(m+a_1')\Gamma(n-m+a_2')}{\sqrt{m+n_1'}\,\sqrt{n-m+n_2'}} (U''_{mk})^{-\frac{1}{2}(n-3)}, \tag{2.7}$$

where J is composed of the constant factors and the term U''_{mk} is given as

$$U''_{mk} = \sum_{I_{1mk}} (x_i-\bar{x}_{1mk})^2 + \sum_{I_{2mk}} (x_i-\bar{x}_{2mk})^2 + \frac{mn_1'}{m+n_1'} (\bar{x}_{1mk}-m_1)^2 + \frac{(n-m)n_2'}{n-m+n_2'} (\bar{x}_{2mk}-m_2')^2. \tag{2.8}$$

The prior covariance matrix is $\underline{\underline{V}}'$, with diagonal elements v_1' and v_2' and zeros off the diagonal. Since $\underline{\underline{N}}' = (\underline{\underline{V}}')^{-1}$ we have $n_1' = \frac{1}{v_1'}$ and $n_2' = \frac{1}{v_2'}$.

3. Optimal Partition with Mixtures of Two Univariate Normals

One can schematically rewrite (2.3) and (2.7) as follows:

$$f(P_{mk}(N)\,|\,\underline{\bar{X}}) = CN_mA_{mk}, \tag{3.1}$$

where C is composed of those constant factors which do not depend upon m or k, the factor N_m depends only on m, and the factor A_{mk} depends on m and k. The search for the optimal partition can now be viewed as a two stage operation. First, find the allocation of m of the x_i to f_1 which maximizes A_{mk}, i.e., find $P_{mk*}(N)$ such that

$$A_{mk*} = \max_k [A_{mk}], \tag{3.2}$$

for $k=1,2,\ldots,\binom{n}{m}$. Second, the optimal partition, $P_{m*k*}(N)$, corresponds to $N_{m*}A_{m*k*}$, where

$$N_{m*}A_{m*k*} = \max_m [N_mA_{mk*}], \tag{3.3}$$

for m=0,1,2,...,n. It is a relatively easy task to examine all of the n+1 values of N_mA_{mk*}. Hence, the problem lies in finding A_{mk*}, the assignment of the m x_i which maximizes A_{mk}, for $k=1,2,\ldots,\binom{n}{m}$.

The A_{mk} factor for the heterogeneous mixture of two normals (2.1) is $(W''_{mk})^{-\frac{1}{2}(n-3)}$ where W''_{mk} is given in (2.4). Minimizing W''_{mk} is equivalent to maximizing $(W''_{mk})^{-\frac{1}{2}(n-3)}$. By considering all the parameters in (2.1) as specified, it is not difficult to convince oneself that the optimal partition of the order statistics, $x_{(1)},x_{(2)},\ldots,x_{(n)}$, where $x_{(n)}$ denotes the largest x_i, should be of the following form,

$$P_{mk*}(N) = (\{i+1,\ldots,i+m\},\{1,\ldots,i\}\cup\{i+m+1,\ldots,n\}) \tag{3.4}$$

or

$$P_{mk*}(N) = (\{1,\ldots,i\}U\{i+n-m+1,\ldots,n\},\{i+1,\ldots,i+n-m\}), \qquad (3.5)$$

where i assumes values to allow in (3.4) any allocation of m consecutive $x_{(i)}$ to f_1 and in (3.5) any allocation n-m consecutive $x_{(i)}$ to f_2. When the parameters are unspecified one also expects a form like (3.4) or (3.5) to minimize W''_{mk}.

The Appendix establishes that it is sufficient to examine the W''_{mk} value corresponding to the n partitions for each value of m=2,3,...,n-2. In addition the W''_{mk} values for the two partitions corresponding to m=0 and n and the 2n W''_{mk} values for the 2n partitions with m=1 and n-1 also need to be inspected. Hence the optimal partition, $P_{m*k*}(N)$, can be determined by computing a specific n^2-n+2 W''_{mk} values corresponding to the like number of "candidate" partitions. Granted this is a large number of partitions, but far fewer than 2^n.

When the mixture of two normals is homogeneous, i.e., h=1, the search for the optimal partition is considerably reduced. As would be expected the optimal partition for fixed values of m=1,2,...,n-1 corresponds to an assignment of the smallest or largest m $x_{(i)}$ to the first component of the mixture of normals. It is therefore sufficient to examine 2n of the 2^n partitions and their corresponding $N_m A_{mk}$ values to determine $P_{m*k*}(N)$.

This result corresponding to the homogeneous mixture can be established by using the general tact taken in the Appendix. Using the ideas introduced there, one can show that for any partition composed of three or more "runs" an "exchange of allocation" exists which increases the value of A_{mk} (h=1). One of two exchanges results in an increase in A_{mk}, an exchange either at the third and second "interface" or at the second and first interface. The details of the proof are not included but they are less involved than those in the Appendix. It can also be shown without added complication that these same results for the homogeneous and heterogeneous mixture of normals hold if σ^2 has a prior distribution

of the inverted gamma form, Raiffa and Schlaifer (1961).

When the prior distribution on (μ_1,μ_2) tends toward a flat prior or "prior of ignorance", the limiting form of the W''_{mk} function for the heterogeneous mixture is interesting. The limiting form of W'_{mk}, as n'_1 and n'_2 tend toward zero, has the limit

$$\sum_{I_{1mk}} (x_i-\bar{x}_{1mk})^2 + \frac{1}{h^2} \sum_{I_{2mk}} (x_i-x_{2mk})^2. \tag{3.6}$$

This is a "weighted" within-partition sum of squares; the between-partition and prior sum of squares portion of W''_{mk} becomes insignificant in the limit (n'_1 and $n'_2 \to 0$).

In the homogeneous mixture (h=1), the corresponding limiting form is more familiar, viz.,

$$\sum_{I_{1mk}} (x_i-\bar{x}_{1mk})^2 + \sum_{I_{2mk}} (x_i-\bar{x}_{2mk})^2. \tag{3.7}$$

This is the within partition sum of squares; the between partition and prior sum of squares is zero in the limit as n'_1 and n'_2 tend to zero.

Two comments on these "limiting" forms are appropriate. The first is to note that in this limit the $N_m A_{mk}$ terms are zero when m=0 or n. This is due to the fact that the limiting prior on (μ_1,μ_2) is everywhere zero. As a consequence, the search for the optimal partition should be restricted so that $1 \leq m \leq n-1$. In effect then the optimal partition is restricted to be a partition with at least one observation assigned to each of the constituent components of the mixture.

The second comment is to point out that Fisher (1958) established that the optimal partition corresponding to (3.7) consists of at most two "runs". This same result has also been established by W. A. Ericson by other methods in an Appendix to Sonquist and Morgan (1964).

4. Application to Outliers with an Analysis of Darwin's Data

Ferguson (1961) credited Dixon with providing model substance to the outlier problem. Dixon (1950) assumed that the "good" observations are distributed normally with mean μ and variance σ^2, i.e., they are distributed $N(\mu,\sigma^2)$, and that the "outliers" are $N(\mu+\lambda\sigma,\sigma^2)$ or $N(\mu,\lambda^2\sigma^2)$ observations. That is, the maverick observation "slipped" by an amount $\lambda\sigma$ or were more variable by a factor $\lambda^2(\lambda>1)$, respectively. Stated another way, sampling is desired from a $N(\mu_1,\sigma_1^2)$ population, but an occasional "outlier" might be observed from a $N(\mu_2,\sigma_2^2)$ population.

The proportion of outliers is usually thought to be small. Regardless of the amount of the mix of "outliers" with "good observations", a mixture of two normals is the appropriate model for data so generated.

De Finetti (1961) wrote a general paper on the rejection of outliers. He was concerned with the effect of outliers on the posterior distribution of the parameter(s) of interest. Presented in the paper are some very specific points of a Bayesian position on the rejection of outliers: (1) There exist no observations to be rejected; (2) The posterior distribution of the parameter(s) of interest is to be determined on the basis of all the observations taken; (3) The Bayesian method leads to an exact result where the influence of the outliers on the posterior distribution is weak or practically negligible. Quoting from his conclusion,

> "...This paper provides no conclusions in the form of formulae or direct and general application. Rather it purports to show that no conclusions of this nature are possible. The whole problem in fact depends on the particular form of the distribution of errors..."

Box and Tiao (1968) wrote another Bayesian paper concerned with outliers and a mixture of two normals as the model. Each of the normal components had the same mean but the contaminating normal component had a larger variance. This is an example of a "particular form of the distribution of errors" as described by de Finetti. The emphasis of Box and Tiao was consistent with the attitude of de Finetti, i.e., they considered the effect of outliers on the posterior distribution of the common mean of their constituent normals.

In this section the identification of the observation(s) which are "most likely" outliers is considered. That is, assuming a mixture of two distributions as the model, those observations in the optimal partition which are allocated to the contaminating component are dubbed as "outliers". With the mixture of two normals in (2.1), the optimal partition of a sample from such a mixture could be determined by the methods discussed in the third section. In the partition with largest posterior probability, those observations allocated to the contaminating component would be "most likely" the outliers. Implicit in this statement is a definition of the phrase, the "most likely" outliers.

A modified analysis of the example in the Box and Tiao paper is used as an illustration. Darwin conducted a very careful experiment to assess the difference in height associated with cross-fertilizing and self-fertilizing plants. The resulting data were the differences in height of fifteen paired, cross-with self-fertilized plants.

Table 1: Darwin's Data

$x_{(1)}$	-67	$x_{(6)}$	16	$x_{(11)}$	41
$x_{(2)}$	-48	$x_{(7)}$	23	$x_{(12)}$	49
$x_{(3)}$	6	$x_{(8)}$	24	$x_{(13)}$	56
$x_{(4)}$	8	$x_{(9)}$	28	$x_{(14)}$	60
$x_{(5)}$	14	$x_{(10)}$	29	$x_{(15)}$	75

The units were in eighths of an inch; more details of this particular data set and the experimental design are given by Fisher (1960).

As noted by Box and Tiao the two negative observations appear rather discrepant as compared with those remaining. Box and Tiao modeled these data as a mixture of normals with common mean, the contiminating normal having a five-fold larger standard deviation than that of the "good" component. An alternative "slippage" model is proposed.

In view of the fact that Darwin's data were the product of experiments carried out over eleven years, it seems quite possible that there could have been a simple mistaken interchange or erroneous recording of labels on two of the pairs. That is, a cross-fertilized seed or plant may have been paired with a self-fertilized one, but the labels were confused or interchanged. In effect then we are suggesting that the signs of $x_{(1)}$ and $x_{(2)}$ in Table 1 should be positive. Such an underlying error process would generate observations distributed as

$$(1-\alpha)N(\mu_1,\sigma^2) + \alpha N(\mu_2,\sigma^2) \qquad (4.1)$$

where we are considering $\mu_2 = -\mu_1$; σ^2 is the common variance and α is the probability that a paired difference is an outlier.

The prior distribution assumed is given (2.6). The idea that μ_2 may be the negative of μ_1 is reflected in the prior by taking $m_2' = -m_1'$. Table 2 exhibits some results for various prior parameter values.

Table 2: Three "Most Likely" Partitions with Various Prior Parameters

Prior No.	m_1'	m_2'	$n_1'=n_2'$	$a_1'/(a_1'+a_2')$	2211...1	2111...1	1111...1
1	16.0	-16.0	1.0	9.5/10.0	1.000	0.195	0.524
2	48.0	-48.0	1.0	9.5/10.0	1.000	0.086	0.111
3	48.0	-48.0	5.0	9.5/10.0	1.000	0.079	0.083
4	8.0	-8.0	1.0	9.5/10.0	0.909	0.269	1.000
5	48.0	-48.0	1.0	1.0/2.0	1.000	0.040	0.016

The notation 2211...1 denotes the partition allocating $x_{(1)}$ and $x_{(2)}$ to the outlier component and $x_{(3)}$ through $x_{(15)}$ as "good data". Similarly for the partition 2111...1; 111...1 denotes the partition allocating all fifteen observations to the $N(\mu_1,\sigma^2)$ component. The entries in the columns for each of the three partitions is the posterior probability of that partition, expressed as a fraction of the "most likely" partition.

The overall impression of the table for various combinations of prior parameters identifies $x_{(1)}$ and $x_{(2)}$ as "most likely" the outliers. Notice that with the very weak prior in the fourth row the optimal partition is 1111...1, but the posterior probability of the partition 2211...1 is 90.9% of that for the "all good observations" partition. This prior assumes only one ($n_1'=1.0$) "good observation" at a one inch ($m_1'=8.0$ eights of an inch) increase in height of cross-fertilized over the self-fertilized plants. Such a slight difference probably would not have been noticed by Darwin. Larger differences probably

occurred in Darwin's experience to have motivated him to design such a careful experiment. It is an example of good design in Fisher's Design of Experiments.

It should be emphasized that the model assumed for the data of Table 1 is given in (4.1). Notice that μ_2 is not restricted to be $-\mu_1$, but the prior information ($m_2'=-m_1'$) was used to reflect this feeling. By using a negative covariance (v_{12}') in the matrix $\underline{\underline{V}}'$ of the prior (2.2), this could be strengthened. The prior correlation between μ_1 and μ_2 is $\rho_{12}' = v_{12}'/\sqrt{v_1'v_2'}$ and the closer ρ_{12}' is taken to -1, the underlying thought that $\mu_2 = -\mu_1$ will be represented ever more strongly. Despite the subjective nature of these suggestions, the alternative of assuming $\mu_2 = -\mu_1$ is less appealing.

5. A Multivariate Extension and Relation to Cluster Analysis

Consider a mixture of two multivariate normal distributions of a p component random vector $\underline{x}$, i.e.,

$$\alpha N(\underline{\mu}_1,\sigma^2\underline{\underline{H}}_1) + (1-\alpha)N(\underline{\mu}_2,\sigma^2\underline{\underline{H}}_2). \tag{5.1}$$

where $\underline{\mu}_1$ and $\underline{\mu}_2$ are p component mean vectors and the p by p variance-covariance matrices are specified within a common variance parameter σ^2. The prior for the parameters of (5.1) and corresponding to (2.6) with $\underline{\mu}_1$ and $\underline{\mu}_2$ independent given σ^2 is

$$N(\underline{m}_1', \frac{\sigma^2}{n_1'}\underline{\underline{H}}_1)N(\underline{m}_2', \frac{\sigma^2}{n_2'}\underline{\underline{H}}_2)(\frac{1}{\sigma^2})\frac{\Gamma(a_1'+a_2')}{\Gamma(a_1')\Gamma(a_2')}\alpha^{a_1'-1}(1-\alpha)^{a_2'-1}. \tag{5.2}$$

The search for the optimal partition can be reduced to finding, for fixed m, the partition of a sample $\underline{x}_1,\underline{x}_2,\ldots,\underline{x}_n$ from (5.1) which minimizes

$$\sum_{I_{1mk}} (\underline{x}_i - \bar{\underline{x}}_{1mk}) \underline{\underline{H}}_1^{-1} (\underline{x}_i - \bar{\underline{x}}_{1mk})^t + \sum_{I_{2mk}} (\underline{x}_i - \bar{\underline{x}}_{2mk}) \underline{\underline{H}}_2^{-1} (\underline{x}_i - \bar{\underline{x}}_{2mk})^t \tag{5.3}$$

$$+ \frac{mn_1'}{m+n_1'} (\bar{\underline{x}}_{1mk} - \underline{m}_1') \underline{\underline{H}}_1^{-1} (\bar{\underline{x}}_{1mk} - \underline{m}_1')^t + \frac{(n-m)n_2'}{n-m+n_2'} (\bar{\underline{x}}_{2mk} - \underline{m}_2') \underline{\underline{H}}_2^{-1} (\bar{\underline{x}}_{2mk} - \underline{m}_2')^t.$$

If $\underline{\underline{H}}_1 = \underline{\underline{H}}_2 = \underline{\underline{I}}$, the p by p identity matrix, and limiting vague priors are assumed on $\underline{\mu}_1$ and $\underline{\mu}_2$ (n_1' and n_2' tending to zero), the minimization reduces to the multivariate analogue to (3.7), i.e.,

$$\sum_{I_{1mk}} (\underline{x}_i - \bar{\underline{x}}_{1mk})^t (\underline{x}_i - \bar{\underline{x}}_{1mk}) + \sum_{I_{2mk}} (\underline{x}_i - \bar{\underline{x}}_{2mk})^t (\underline{x}_i - \bar{\underline{x}}_{2mk}) \tag{5.4}$$

As discussed with (3.7), the comparison of values of (5.3) is reasonable only for the values of m, $1 \leq m \leq n-1$. Since the proof in the Appendix is dependent on a scalar ordering of the sample observations, other methods are necessary to limit the search for $P_{m^*k^*}(N)$ with multivariate data.

Notice that (5.4) is the objective function of the Edwards and Cavalli-Sforza (1965) approach to cluster analysis. That is, their method divides the n observations into two sets such that the sum of squares of distances between the two sets is maximal, or equivalently, minimizes the within partition sum of squares, (5.4). Their search is over $2^{n-1}-1$ partitions; at least one observation is allocated to each set (2^n-2) and no identity is associated with either set ($(2^n-2)/2$).

Ward (1963) considered hierarchical grouping on the basis of optimizing an objective function, but offered no comment on the selection of the objective function measuring homogeneity or similarity. Besides the "within sum of squares" criterion, other authors have proposed various criterion and based

a cluster analysis on such, e.g., Sokal and Michener's "weighted mean pair" similarity criterion, (1958). A more recent description has been given by Sokal and Sneath (1963). A comparison of clustering techniques is given by Gower (1967) with mathematical considerations being given primary emphasis.

The remainder of the section will be spent offering a relation between cluster analysis and mixture models. Although cluster analysis is typically given a non-parametric context, Cox (1966) mentioned that mixtures and "internal classification" schemes are related. If the data can be modeled by a mixture of meaningful sub-populations, it seems quite natural to try to "estimate" which observations "most likely" came from which constituent of the mixture. The optimal partition of the sample is a candidate for determining which observations "most likely" came from which component.

Finding the optimal partition for a sample from a specific mixture model (only two components here) was viewed as equivalent to optimizing an objective function, specifically, maximizing the marginal posterior probability of a partition of the sample. (Recall that the marginality was achieved by integrating over the parameter space; no inference is desired of the parameters.) As presented in the third section this function, $N_m A_{mk}$, is composed of two gross factors. The N_m factor depends only on how many x_i are allocated to the "first" component. The A_{mk} factor depends on which m of n x_i are allocated to the "first" component. The point must be made that the mixture model assumed should offer some information on the form of the function which is to be optimized. Of course, different mixtures are expected to be associated with correspondingly different measures of interset distance.

As a conclusion, notice that if the mixture of two multivariate normals in (5.1) with common variance-covariance matrices, $\sigma^2 \underline{\underline{I}}$, model the data, the prior

information on $\underline{\mu}_1$ and $\underline{\mu}_2$ is of the limiting vague form discussed earlier, and the N_m factor is ignored, then the optimal partition corresponds to minimizing the within partition sum of squares (5.4).

Acknowledgements

This work was supported primarily by funds from the National Institutes of Health administered by the Department of Biostatistics, The University of Michigan. I would like to express my appreciation for the guidance of Dr. William A. Ericson throughout the investigation.

References

Anderson, T. W. (1958). An Introduction to Multivariate Statistical Analysis. Wiley, New York.

Box, G. E. P. and Tiao, G. C. (1968). A Bayesian approach to some outlier problems. Biometrika, 55:119-129.

Cox, G. R. (1966). Notes on the analysis of mixed frequency distributions. British Journal of Mathematical and Statistical Psychology, 19:39-47.

De Finetti, B. (1961). The Bayesian approach to rejection of outliers. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1:199-210.

Dixon, W. J. (1950). Analysis of extreme values. Annals of Mathematical Statistics, 21:488-506.

Dunsmore, I. R. (1966). A Bayesian approach to classification. Journal of the Royal Statistical Society, Series B, 28:568-577.

Edwards, A. W. F. and Cavalli-Sforza, L. L. (1965). A method for cluster analysis. Biometrics, 21:362-375.

Ferguson, T. S. (1961). On the rejection of outliers. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. 1:253-287.

Fisher, R. A. (1960). Design of Experiments. (7th edition). Haefner, New York.

Fisher, W. D. (1958). On Grouping for maximum homogeneity. Journal of the American Statisticsl Association, 53:789-798.

Geisser, S. (1964). Posterior odds for multivariate normal classifications. Journal of the Royal Statistical Society, Series B. 26:69-76.

Gower, J. C. (1967). A comparison of some methods of cluster analysis. Biometrics, 23:623-637.

Lindley, D. V. (1965). Introduction to Probability and Statistics from a Bayesian Viewpoint, Part 2: Inference, Cambridge University Press, New York.

Raiffa, H. and Schlaifer, R. (1961). Applied Statistical Decision Theory. Harvard Business School, Boston.

Rao, C. R. (1952). Advanced Statistical Methods in Biometric Research. Wiley, New York.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. University of Kansas Science Bulletin, 38: 1409-1438.

Sokal, R. R. and Sneath, P. H. (1963). Principles of Numerical Taxonomy. Freeman, San Francisco.

Sonquist, J. A. and Morgan, J. N. (1964). The detection of interaction effects. Monograph No. 35: Survey Research Center. Institute for Social Research, The University of Michigan.

Symons, M. J. (1969). A Bayesian Test of Normality with a Mixture of Two Normals as the Alternative and Applications to 'Cluster' Analysis. Unpublished Doctoral Dissertation. The University of Michigan.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association, 58:236-244.

Appendix

This appendix establishes that the optimal partition in a mixture of two univariate normals with specified variance ratio is of the form

$$P_{mk*}(N) = (\{i+1,\ldots,i+m\},\{1,\ldots,i\}U\{i+m+1,\ldots,n\}) \qquad (I.1)$$

or

$$P_{mk*}(N) = (\{1,\ldots,i\}U\{i+n-m+1,\ldots,n\},\{i+1,\ldots,i+n-m\}). \qquad (I.2)$$

The value of i is to allow any allocation of m consecutive order statistics to f_1 in (2.1) in (I.1) and any allocation of n-m consecutive order statistics to the second component in (I.2).

Without loss of generality, h is taken as greater than or equal to one, i.e., the larger variance component is labeled as the second constituent of the mixture. As remarked in the third section, minimizing W'_{mk} is equivalent to maximizing A_{mk}.

There are some preliminary ideas to be discussed. First a partition of the order statistics can be thought of as a sequence of n 1's and 2's indicating whether each $x_{(i)}$ is allocated to the first or second component of the mixture, e.g., with n=5 and m=2, 21122 represents a partition denoting $x_{(2)}$ and $x_{(3)}$ are allocated to the first component and $x_{(1)}$, $x_{(4)}$, $x_{(5)}$ are allocated to the second. Defining a "run" as a sequence of consecutive 1's or 2's, the partition 21122 has three runs. An "exchange of allocation", or exchange, is to interchange the component assignments of two $x_{(i)}$'s in a partition. For example, consider the partition 21122. An exchange of allocation of $x_{(3)}$ and $x_{(4)}$ results in the partition 21212.

With these ideas, the optimal partition is therefore composed of three or fewer runs and when m=0, 1, n-1, or n all the partitions are of this form;

hence one can assume $2 \leq m \leq n-2$. This requires n to be greater than or equal to four.

Before proceeding with the details, the idea behind the proof is sketched. Let n and m be fixed. Consider any partition with four or more runs, e.g., 2112122. It will be shown that there exists an exchange of allocation such that W''_{mk} has a distinctly smaller value when evaluated for the resulting partition than for the former partition with four or more runs. The resulting partition may also have four or more runs; if so, the exchange procedure is repeated. Eventually the process will lead to a partition with three or fewer runs. Hence, any partition with four or more runs is dominated by (has a larger W''_{mk} value than) some partition with three or fewer runs. No "infinite cycle" is possible since each exchange produces a positive decrease in W''_{mk} and there are only $\binom{n}{m}$ possible partitions. To insure a positive decrease in W''_{mk}, the n order statistics are assumed distinct, i.e., $x_{(1)} < x_{(2)} < \ldots < x_{(n)}$.

Now

$$W'_{mk} = \sum_{I_{1mk}} (x_i - \bar{x}_{1mk})^2 + \frac{1}{h^2} \sum_{I_{2mk}} (x_i - \bar{x}_{2mk})^2 + (\underline{\bar{x}}_{mk} - \underline{m}')\underline{\underline{N}}^*_m (\underline{\bar{x}}_{mk} - \underline{m}')^t, \qquad (I.3)$$

where

$$\underline{\underline{N}}^*_m = \underline{\underline{N}}_m (\underline{\underline{N}}''_m)^{-1} \underline{\underline{N}}', \quad \underline{\underline{N}}''_m = \underline{\underline{N}}_m + \underline{\underline{N}}',$$

$$\underline{\underline{N}}_m = \begin{pmatrix} m & 0 \\ 0 & (n-m)/h^2 \end{pmatrix}, \qquad (I.4)$$

and

$$\underline{\underline{N}}' = (\underline{\underline{V}}')^{-1} = \begin{pmatrix} n_1' & n_{12}' \\ n_{12}' & n_2' \end{pmatrix} . \tag{I.5}$$

Algebraically, $\underline{\underline{N}}_m^*$ can be written

$$\underline{\underline{N}}_m^* = \frac{1}{d} \begin{pmatrix} a & b \\ c & d \end{pmatrix}, \tag{I.6}$$

with

$$a = m[n_1'(n-m+h^2 n_2') - h^2 (n_{12}')^2], \tag{I.7}$$

$$b = m(n-m)n_{12}', \tag{I.8}$$

$$c = (n-m)[n_2'(m+n_1') - (n_{12}')^2], \tag{I.9}$$

and

$$d = (m+n_1')(n-m+h^2 n_2') - h^2 (n_{12}')^2. \tag{I.10}$$

Hence, W_{mk}' can be expanded as

$$\begin{aligned} W_{mk}' = {} & SS_{1mk} - \frac{1}{m} S_{1mk} + \frac{1}{h} [SS_{2mk} - \frac{1}{n-m} S_{2mk}] \\ & + \frac{a}{d} [\frac{1}{m^2} S_{1mk}^2 - \frac{2}{m} m_1' S_{1mk} + (m_1')^2] \\ & + \frac{2b}{d} [\frac{1}{m(n-m)} S_{1mk} S_{2mk} - \frac{1}{m} m_2' S_{1mk} - \frac{1}{n-m} m_1' S_{2mk} + m_1' m_2'] \\ & + \frac{c}{d} [\frac{1}{(n-m)^2} S_{2mk}^2 - \frac{2}{n-m} m_2' S_{2mk} + (m_2')^2], \end{aligned} \tag{I.11}$$

where

$$SS_{1mk} = \sum_{I_{1mk}} x_i^2 \ , \ S_{1mk} = \sum_{I_{1mk}} x_i,$$

and similarly for SS_{2mk} and S_{2mk}.

Collecting and simplifying the coefficients of the various terms, the coefficient of $-S_{1mk}^2$ reduces to

$$A = \frac{1}{d}\,(n-m+h^2n_2'). \tag{I.12}$$

The coefficient of $-2S_{1mk}$ becomes

$$B = \frac{1}{d}\,[m_1'(n_1'(n-m+h^2n_2') - h^2(n_{12}')^2) + m_2'(n-m)n_{12}']. \tag{I.13}$$

Similarly, the coefficients of $-\frac{1}{h^2}\,S_{2mk}^2$ and $-\frac{2}{h^2}\,S_{2mk}$ are

$$C = \frac{1}{d}\,(m+n_1') \tag{I.14}$$

and

$$D = \frac{h^2}{d}\,[m_2'(n_2'(m+n_1') - (n_{12}')^2) + n_1'm_1n_{12}'], \tag{I.15}$$

respectively. The coefficient of $2S_{1mk}S_{2mk}$ reduces to

$$E = \frac{n_{12}'}{d}\,. \tag{I.16}$$

The constant terms remaining are denoted by K,

$$K = \frac{a}{d}\,(m_1')^2 + \frac{2b}{d}\,m_1'm_2' + \frac{c}{d}\,(m_2')^2. \tag{I.17}$$

Therefore

$$W_{mk}' = SS_{1mk} - AS_{1mk}^2 - 2BS_{1mk} + \frac{1}{h^2}\,[SS_{2mk} - CS_{2mk}^2 - 2DS_{2mk}] + 2ES_{1mk}S_{2mk} + K. \tag{I.18}$$

Consider a partition with four or more runs. Let y denote an $x_{(i)}$ currently allocated to the first component normal but to be exchanged with an $x_{(i)}$ assigned to the other normal constituent. Let z denote this $x_{(i)}$ assumed to come from the second component. Define

$$SS_{1mky} = SS_{1mk} - y^2 \tag{I.19}$$

$$S_{1mky} = S_{1mk} - y \tag{I.20}$$

$$SS_{2mkz} = SS_{2mk} - z^2 \tag{I.21}$$

and

$$S_{2mkz} = S_{2mk} - z. \tag{I.22}$$

As a function of y and z,

$$W''_{mk}(y,z) = y^2(1-A) - 2y(AS_{1mky} + B - ES_{2mkz}) + \frac{1}{h^2}\left[z^2(1-C) - 2z(CS_{2mkz} + D - h^2 ES_{1mky})\right] + 2Ezy + K', \tag{I.23}$$

where K' absorbed K and the other constant terms. The value of W''_{mk} after the exchange is

$$W''_{mk}(z,y) = z^2(1-A) - 2z(AS_{1mky} + B - ES_{2mkz}) + \frac{1}{h^2}\left[y^2(1-C) - 2y(CS_{2mkz} + D - h^2 ES_{1mky}) + 2Eyz + K'\right]. \tag{I.24}$$

Note that $W''_{mk}(y,z) = W''_{mk}$ and $W''_{mk}(z,y) = W''_{mk'}$, where k' is the subscript of another partition for the same value m.

If the difference, $W''_{mk}(z,y) - W''_{mk}(y,z)$, is greater than or equal to zero, then the exchange did not produce a partition with a smaller W''_{mk} value. A

negative difference means a "successful" exchange. Explicitly the difference can be written as

$$(z-y)[(z+y)F+G], \tag{I.25}$$

where

$$F = 1-A - \frac{1}{h^2}(1-C), \tag{I.26}$$

and

$$G = -2(A+E)S_{1mky} + \frac{2}{h^2}(C+h^2E)S_{2mky} - 2B + \frac{2}{h^2}D. \tag{I.27}$$

It remains to be shown that for all partitions with four or more runs, there exists an exchange which makes (I.25) negative. The set of all partitions with four or more runs can be divided into two disjoint and exhaustive classes. The first class is composed of all those partitions with four or more runs and the right most run is of 2's, e.g., 2212221112. The complementary class is all those with four or more runs and the right most run of 1's, e.g., 2212221121. Note that m and n are assumed fixed, $2 \leq m \leq n-2$, $n \geq 4$.

Consider the class of all partitions with the right most run (RMR) of 2's and having four or more runs. Let z_0 denote the largest order statistic in the fourth RMR, z_1 the smallest order statistic in the third RMR, z_2 the largest order statistic in the third RMR, and z_3 the smallest order statistic in the second RMR. The claim is that one of two exchanges results in a negative value of (I.25), i.e., a partition which dominates the original. The two exchanges are z_0 with z_1 and z_2 with z_3. Since W'_{mk} is translation invariant, let $z_0=0$, $z_1=D_1$, $z_2=D_1+D_2$, and $z_3 = D_1+D_2+D_3$. Now D_1 and D_3 are strictly positive since the order statistics are assumed distinct. It may be that $D_2=0$, if so, then $z_2=z_1$, i.e., the third RMR has only one element.

Let

$$S_{10} = \sum_{I_{1mk}} x_i - z_0, \tag{I.28}$$

$$S_{13} = S_{10} + z_0 - z_3 = S_{10} - D_1 - D_2 - D_3, \tag{I.29}$$

$$S_{21} = \sum_{I_{2mk}} x_i - z_1, \tag{I.30}$$

and

$$S_{22} = S_{21} + z_1 - z_2 = S_{21} - D_2. \tag{I.31}$$

The first exchange identifies z_0 with y and z_1 with z, hence S_{10} as S_{1mky} and S_{21} as S_{2mkz} in (I.25) and (I.27).

Equation (I.25) becomes

$$D_1(D_1F+G). \tag{I.32}$$

The second exchange identifies z_3 with y and z_2 with z, hence S_{13} as S_{1mky} and S_{22} as S_{2mkz} in (I.25) and (I.27). Using (I.29) and (I.31), equation (I.25) becomes

$$-D_3[(2D_1+2D_2+D_3)F+G+2(D_1+D_2+D_3)(A+E) - \frac{2}{h^2} D_2(C+h^2E)]. \tag{I.33}$$

Since D_1 and D_3 are positive, dividing (I.32) and (I.33) by them, respectively, does not alter the sign of the equation. The result is

$$D_1F + G \tag{I.34}$$

and

$$-(D_1F+G) - [(D_1+2D_2+D_3)F + 2(D_1+D_2+D_3)(A+E) - \frac{2}{h^2} D_2(C+h^2E)]. \tag{I.35}$$

Suppose (I.34) is greater than or equal to zero, i.e., the first exchange was not an improvement. Therefore, the first part of (I.35), $-(D_1F+G)$, is

non-positive. The proof is complete if the second part of (I.35) is negative. The coefficients of D_1+D_3 and $2D_2$ in the latter piece reduce to

$$1 - \frac{1}{h^2} + A + 2E + \frac{C}{h^2} \tag{I.36}$$

and

$$1 - \frac{1}{h^2}\ , \tag{I.37}$$

respectively. Since D_2 is greater than or equal to zero and $1-\frac{1}{h^2}$ is non-negative ($h \geq 1$), the result is established if (I.36) is positive. Recall D_1+D_3 is positive since D_1 and D_3 are both positive.

The coefficients of D_1+D_3, (I.36), is positive if $A+2E+C/h^2$ is positive since $1-1/h^2$ is non-negative. From (I.12), (I.14), and (I.16), $A+2E+C/h^2$ reduces to

$$\frac{1}{d}\ [n-m+h^2n_2'+2n_{12}'+m/h^2+n_1'/h^2]. \tag{I.38}$$

Using the facts that $\underline{\underline{N}}'$ is a symmetric, positive-definite matrix and $h \geq 1$, one can verify that d, (I.10), is positive. The result will be established if

$$h^2n_2' + n_1'/h^2 + 2n_{12}' > 0. \tag{I.39}$$

Since $\underline{\underline{N}}'$ is a symmetric, positive-definite matrix,

$$\sqrt{n_1'n_2'} > |n_{12}'|\ . \tag{I.40}$$

Now

$$(h^2n_2'-n_1'/h^2)^2 \geq 0, \tag{I.41}$$

therefore, expanding the left side of (I.41) and adding $4n_1'n_2'$ to both sides of the inequality yields

$$h^4(n_2')^2 + 2n_1'n_2' + (n_1')^2/h^4 \geq 4n_1'n_2', \tag{I.42}$$

or

$$(h^2n_2' + n_1'/h^2)^2 \geq 4n_1'n_2', \tag{I.43}$$

and hence

$$h^2n_2' + n_1'/h^2 \geq 2\sqrt{n_1'n_2'} > 2|n_{12}'| \tag{I.44}$$

from (I.40). The inequality in (I.39) is now established.

The conclusion is that for all partitions with four or more runs and the right most run of 2's, there exists an exchange of allocation which results in a smaller value of W_{mk}''.

To complete the proof, it must be shown that there exists an exchange which results in a smaller value of W_{mk}'' for the class of all partitions with four or more runs and the right most run being of 1's. In the previous case, the improving exchange was between the interface elements of the fourth and third RMR or of the third and second RMR. In this case either the exchange at the third and second interface or at the second and first interface will result in a negative value of (I.25). The proof is not given since it is literally identical to the above proof.

The following result has been established. The partition which maximizes $A_{mk} = (W_{mk}'')^{-\frac{1}{2}(n-3)}$ with W_{mk}'' given in (I.3) for m=2,3,...,n-2 is one which has all m $x_{(i)}$'s allocated to the normal component (with mean μ_1 and variance σ^2) adjacent to one another, or is of the form (n-m) consecutive $x_{(i)}$'s allocated to the normal component with mean μ_2 and variance $h^2\sigma^2$. Stated another way, the optimal partition is of the form (I.1) or (I.2). Note that there is no loss in generality by assuming $h \geq 1$ since the labelling of the components is arbitrary. It was elected to label the larger variance component as "2".

Only n of the $\binom{n}{m}$ partitions need be examined to find A_{mk*}, (3.2). The optimal partition is found by searching the $N_m A_{mk*}$ terms over $0 \leq m \leq n$ as indicated in (3.3).

The basic approach in the proof of this appendix was suggested by Dr. W. A. Ericson, The University of Michigan.