MODEL-BASED GAUSSIAN AND NON-GAUSSIAN CLUSTERING

by

Jeffrey D. Banfield
Adrian E. Raftery

TECHNICAL REPORT No. 186

December 1989

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195 USA

Model-based Gaussian and non-Gaussian Clustering

Jeffrey D. Banfield

Department of Mathematical Sciences
Montana State University
Bozeman, Montana 59715.

Adrian E. Raftery

Department of Statistics, GN-22
University of Washington
Seattle, Washington 98195.

ABSTRACT

The classification maximum likelihood approach is sufficiently general to encompass many current clustering algorithms, including those based on the sum of squares criterion and on the criterion of Friedman and Rubin (1967). However, as currently implemented, it does not allow the specification of which features (orientation, size and shape) are to be common to all clusters and which may differ between clusters. Also, it is restricted to Gaussian distributions and it does not allow for noise.

We propose ways of overcoming these limitations. A reparameterization of the covariance matrix allows us to specify that some features, but not all, be the same for all clusters. A practical framework for non-Gaussian clustering is outlined, and a means of incorporating noise in the form of a Poisson process is described. An approximate Bayesian method for choosing the number of clusters is given.

The performance of the proposed methods is studied by simulation, with encouraging results. The methods are applied to the analysis of a data set arising in the study of diabetes, and the results seem better than those of previous analyses.

KEYWORDS: Bayes factors; Classification; Diabetes; Hierarchical agglomeration; Iterative relocation; Mixture models.

Jeffrey D. Banfield is Assistant Professor, Department of Mathematical Sciences, Montana State University, Bozeman, MT 59717. Adrian E. Raftery is Associate Professor, Department of Statistics, GN-22, University of Washington, Seattle, WA 98195. Banfield's research was supported in part by the Office of Naval Research under Contract no. N-00014-89-J-1114. Raftery's research was supported by the Office of Naval Research under Contract no. N-00014-88-K-0265.

1. INTRODUCTION

Cluster analysis has developed mainly through the invention and empirical investigation of *ad-hoc* methods, in isolation from more formal statistical procedures. In recent years it has been found that basing cluster analysis on a probability model can be useful both for understanding when existing methods are likely to be successful, and for suggesting new methods (Symons 1981; McLachlan 1982; McLachlan and Basford 1988).

One such probability model is that the population of interest consists of G different subpopulations, and that the density of a p-dimensional observation $\mathbf{x}$ from the kth subpopulation is $f_k(\mathbf{x};\boldsymbol{\theta})$ for some unknown vector of parameters $\boldsymbol{\theta}$. Given observations $\mathbf{x}=(\mathbf{x}_1,\ldots,\mathbf{x}_n)$, we let $\boldsymbol{\gamma}=(\gamma_1,\ldots,\gamma_n)^T$ denote the identifying labels, where $\gamma_i=k$ if $\mathbf{x}_i$ comes from the kth subpopulation. In the so-called classification maximum likelihood procedure, $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ are chosen so as to maximize the likelihood

$$L(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\gamma})=\prod_{i=1}^{n} f_{\gamma_i}(\mathbf{x}_i;\boldsymbol{\theta}). \tag{1.1}$$

Scott and Symons (1971) have worked out the solution when $f_k(\mathbf{x};\boldsymbol{\theta})$ is multivariate normal with mean vector $\boldsymbol{\mu}_k$ and variance matrix Σ_k, a distribution which we denote by $\mathrm{MVN}(\mu_k,\Sigma_k)$. When $\Sigma_k=\sigma^2 I \quad (k=1,\ldots,G)$ this reduces to the sum of squares criterion (Gordon, 1981), while when $\Sigma_k=\Sigma \quad (k=1,\ldots,G)$ it yields the criterion of Friedman and Rubin (1967). For a more detailed review of these ideas, see Gordon (1981).

However, as currently implemented, the classification maximum likelihood procedure has several limitations:

(1) It considers only the restrictive model where the covariance matrices are constant across all clusters, or the unparsimonious model where they are arbitrary and unequal. The latter is rarely used in practice, probably because of difficulties caused by its very generality and lack of parsimony (Symons 1981). It would seem desirable to have criteria based on intermediate models which allow some of the characteristics of the covariance matrices to differ across clusters. For example, clusters may be elliptical with roughly the same size

and shape, but oriented in different directions.

(2) It allows only for Gaussian distributions, whereas other distributions may be more appropriate in some situations. An example of this arises frequently in unsupervised pattern recognition, where edges may be represented by points clustered uniformly, rather than normally, along a straight line.

(3) It does not, in general, allow for noise, or data points which do not fit the prevailing pattern of clusters. Indeed, when the covariance matrices are unequal, each cluster must contain at least $p+1$ observations (Symons, 1981).

In this article, we present a framework for model-based clustering which is sufficiently general to overcome these limitations. In Section 2, we develop maximum likelihood criteria for Gaussian clustering which allow clusters to have different orientations or sizes, while preserving some common features, such as shape. In Section 3, we present practical criteria for non-Gaussian clustering, and we extend the framework to incorporate Poisson noise. In Section 4, we present a model-based approximate Bayesian approach to choosing the number of clusters. In Section 5 we report the results of a Monte Carlo study of the methods presented, and in Section 6 we study their performance on three data sets, of which two are simulated and one is real.

2. ALLOWING ORIENTATION AND SIZE TO VARY BETWEEN CLUSTERS IN THE GAUSSIAN CASE

When $f_k(\mathbf{x};\boldsymbol{\theta})$ is a MVN $(\boldsymbol{\mu}_k, \Sigma_k)$ density, the likelihood (1.1) has the form

$$L(\mathbf{x};\boldsymbol{\theta},\boldsymbol{\gamma}) = \text{const.} \prod_{k=1}^{G} \prod_{i \in E_k} |\Sigma_k|^{-½} \exp\{-½(\mathbf{x}_i-\boldsymbol{\mu}_k)^T \Sigma_k^{-1} (\mathbf{x}_i-\boldsymbol{\mu}_k)\}, \qquad (2.1)$$

where $E_k = \{i : \gamma_i = k\}$. The maximum likelihood estimator of $\boldsymbol{\mu}_k$ is $\bar{\mathbf{x}}_k = n_k^{-1} \sum_{i \in E_k} \mathbf{x}_i$, where n_k is the number of elements in E_k. Substituting this into (2.1) yields the concentrated log-likelihood

$$l\ (\mathbf{x};\boldsymbol{\theta},\boldsymbol{\gamma}) = \text{const.} - ½ \sum_{k=1}^{G} \{\text{tr}(W_k \Sigma_k^{-1}) + n_k \log|\Sigma_k|\}, \qquad (2.2)$$

where W_k is the sample cross-product matrix for the kth cluster, namely

$$W_k = \sum_{i \in E_k} (\mathbf{x}_i - \bar{\mathbf{x}}_k)(\mathbf{x}_i - \bar{\mathbf{x}}_k)^T .$$

If $\Sigma_k = \sigma^2 I \quad (k = 1, \ldots, G)$, then the log-likelihood (2.2) is maximized by choosing $\boldsymbol{\gamma}$ so as to minimize $\mathrm{tr}(W)$, where $W = \sum_{k=1}^{G} W_k$. This is the sum of squares criterion which underlies, for example, Ward's (1963) agglomerative hierarchical clustering method. If $\Sigma_k = \Sigma \quad (k = 1, \ldots, G)$, then the log-likelihood (2.2) is maximized by choosing $\boldsymbol{\gamma}$ so as to minimize $|W|$, the criterion of Friedman and Rubin (1967). Finally, when the Σ_k are not constrained in any way, the likelihood is maximized by choosing $\boldsymbol{\gamma}$ so as to minimize $\sum_{k=1}^{G} n_k \log \left| \frac{W_k}{n_k} \right|$. This is similar to, but not the same as, equation (14) of Scott and Symons (1971), which we have been unable to reproduce exactly.

Here we develop new criteria which are more general than that of Friedman and Rubin (1967), but based on more parsimonious models than that of Scott and Symons (1971). They allow some but not all of the features of cluster distributions (orientation, size and shape) to vary between clusters, while constraining others to be the same. The key to this is a reparameterization of the covariance matrix Σ_k in terms of its eigenvalue decomposition

$$\Sigma_k = D_k \Lambda_k D_k^T, \tag{2.3}$$

where D_k is the matrix of eigenvectors and Λ_k is a diagonal matrix with the eigenvalues of Σ_k on the diagonal. The orientation of the principal components of Σ_k is determined by D_k, while Λ_k specifies the size and shape of the density contours. We write $\Lambda_k = \lambda_k A_k$, where λ_k is the first eigenvalue of Σ_k, $A_k = \mathrm{diag}\{\alpha_{1k}, \ldots, \alpha_{pk}\}$, and $1 = \alpha_{1k} \geq \alpha_{2k} \geq \cdots \geq \alpha_{pk} > 0$. Thus D_k determines the orientation of the kth cluster, λ_k its size, and A_k its shape. If the α_{jk}'s are of similar magnitude, then the kth cluster will tend to be nearly hyperspherical, while if $\alpha_{2k} \ll 1$, it will be concentrated about a line, and if $\alpha_{2k} \approx 1$ and $\alpha_{3k} \ll 1$ it will be concentrated about a two-dimensional plane in p-space.

This analysis suggests that the sum of squares criterion is likely to be most appropriate when the clusters are all hyperspherical with the same dispersion (Symons, 1981). The criterion of Friedman and Rubin (1967) is likely to work best when the clusters are ellipsoidal with the same orientation, size and shape. The criterion of Scott and Symons (1971) allows clusters of different orientations, sizes and shapes, but its very generality and lack of parsimony may cause problems. For example, Symons (1981) reported that criteria designed for clusters of different shapes may produce clusters of different shapes and sizes even when presented with homogeneous-shaped clusters that are close together. The criterion of Friedman and Rubin (1967) is based on the assumption that D_k, λ_k and A_k are the same for each cluster, while the criterion of Scott and Symons (1971) assumes them all to be different. By allowing some but not all of these quantities to vary between clusters, we obtain criteria that are appropriate for various intermediate situations.

Assuming that $\Sigma_k = \lambda_k I$ leads to a generalization of the sum of squares criterion. The fact that Σ_k is a multiple of the identity matrix indicates that the underlying densities are spherical. Allowing λ_k to vary between densities allows the sizes of the clusters to differ. The resulting criterion to be minimized is

$$\sum_{k=1}^{G} n_k \log \operatorname{tr}\left(\frac{W_k}{n_k}\right)$$

Our analysis indicates that this criterion will be most appropriate when the clusters are hyperspherical, but of different sizes.

Next, we allow the orientations to vary while keeping size and shape constant, by requiring that $\lambda_k = \lambda, A_k = A \quad (k = 1, \ldots, G)$ where A is known, and by allowing the D_k's to vary between clusters. By noting that $|\Sigma_k| = \lambda^p \prod_{j=1}^{p} \alpha_j$, replacing D_k and λ in (2.2) with their maximum likelihood estimators and writing the eigenvalue decomposition of W_k as

$$W_k = L_k \, \Omega_k \, L_k^T \tag{2.4}$$

where $\Omega_k = \operatorname{diag}\{\omega_{1k}, \ldots, \omega_{pk}\}$ and ω_{jk} is the jth eigenvalue of W_k, we see that the resulting log-likelihood is maximized by choosing $\boldsymbol{\gamma}$ so as to minimize $S = \sum_{k=1}^{G} S_k$, where $S_k = \mathrm{tr}(A^{-1}\Omega_k)$. When $p=2$, this is the criterion that underlies the clustering method of Murtagh and Raftery (1984).

We now allow both size, λ_k, and orientation, D_k, to vary between clusters, while assuming that the shape matrix A is constant across clusters. In this setting, the maximum likelihood estimator of $\boldsymbol{\gamma}$ is obtained by minimizing

$$S^* = \sum_{k=1}^{G} n_k \log(S_k/n_k). \tag{2.5}$$

Table 1 shows the relationship between the different criteria discussed in this section.

Table 1
Constraints imposed on clusters by different criteria

Criterion	Origin	Distribution	Orientation	Size	Shape
$\mathrm{tr}(W)$	Ward (1963)	Spherical	None	Same	Same
$\lvert W \rvert$	Friedman and Rubin (1967)	Ellipsoidal	Same	Same	Same
S	Murtagh and Raftery (1984)	Ellipsoidal	Different	Same	Same
$\sum_{k=1}^{G} n_k \log \mathrm{tr}(W_k/n_k)$	This article	Spherical	None	Different	Same
S^*	This article	Ellipsoidal	Different	Different	Same
$\sum_{k=1}^{G} n_k \log \left\lvert \frac{W_k}{n_k} \right\rvert$	Scott and Symons (1981)	Ellipsoidal	Different	Different	Different

It is usually not feasible to find the global minimum by evaluating the criterion for all possible partitions of the observations. Many algorithms have been devised for finding local minima or good sub-optimal solutions, particularly for the sum of squares criterion. These involve agglomeration, iterative relocation or other methods; for reviews see Gordon (1981, 1987), Murtagh (1985) and Jain and Dubes (1988). Algorithms developed for the sum of squares criterion can be adapted for use with the other criteria in Table 1. For example, Murtagh and Raftery (1984) showed how Ward's (1963) agglomerative hierarchical method based on the sum of squares criterion can be generalized for use with the criterion S.

3. NON-GAUSSIAN CLUSTERING AND NOISE

3.1 Non-Gaussian clustering: The uniform-normal case

The model (1.1) is general enough to encompass clusters with non-Gaussian distributions. To date, attention has been focused on the multivariate normal distribution because it leads to relatively simple criteria. Here we suggest practical criteria for some non-Gaussian situations.

The basic idea is the use of a local parameterization. We assume that there are matrices D_k $(k = 1, \ldots, G)$ such that if $\mathbf{z}_i = D_{\gamma_i}(\mathbf{x}_i - \boldsymbol{\mu}_{\gamma_i})$, then $\mathbf{z}_i$ has the density $g_{\gamma_i}(\mathbf{z}_i ; \boldsymbol{\theta})$; often these densities will be the same, perhaps modulo a scale parameter. In this general framework, criteria can be derived by maximizing the likelihood, as in Section 2. When the distribution of $\mathbf{x}_i$ is $\mathrm{MVN}(\boldsymbol{\mu}_{\gamma_i}, \Sigma_{\gamma_i})$, and D_k is defined by (2.3), then $\mathbf{z}_i$ is the value of $\mathbf{x}_i$ in the local coordinate system with origin at $\boldsymbol{\mu}_{\gamma_i}$ and axes along the principal components of Σ_{γ_i}.

We now carry out a more detailed analysis of one specific, but important, non-Gaussian situation. This is when observations are clustered uniformly along and tightly about a line segment in p-space. Such situations arise in ecology when the data include the geographic locations of plants or animals which may be clustered about roughly linear natural features such as rivers or valleys. They also arise in unsupervised pattern recognition, where observations may be edge elements in an image, or data points in a point pattern with a linear feature.

We let $u_i = z_{i1}$, and $\mathbf{v}_i = (v_{i1}, \ldots, v_{i,p-1})^T = (z_{i2}, \ldots, z_{ip})^T$. We assume that u_i is uniformly distributed between $\phi_{\gamma_i,1}$ and $\phi_{\gamma_i,2}$, and that $\mathbf{v}_i \sim \mathrm{MVN}(\mathbf{0}, \Sigma_k)$. Let $\phi_k = \phi_{k2} - \phi_{k1}$ and $\Sigma_k = \sigma_k^2 \mathrm{I}$; typically σ_k will be small compared to ϕ_k.

We now derive an approximate maximum likelihood estimator for $\boldsymbol{\gamma}$ under this model. We estimate D_k by $\hat{D}_k = L_k$, where L_k is defined by (2.4), and we estimate $\boldsymbol{\mu}_k$ by $\hat{\boldsymbol{\mu}}_k = \bar{\mathbf{x}}_k$. We then define $\hat{\mathbf{z}}_i = \hat{D}_{\gamma_i}(\mathbf{x}_i - \bar{\mathbf{x}}_{\gamma_i})$, with corresponding definitions for $\hat{u}_i$ and $\hat{\mathbf{v}}_i$. Conditionally on these estimated values of D_k and $\boldsymbol{\mu}_k$, the log-likelihood for $\phi = (\phi_1, \ldots, \phi_G)^T$, σ_k^2 and $\boldsymbol{\gamma}$ is

$$l(\mathbf{x}; \phi, \sigma^2, \boldsymbol{\gamma}) = \text{const.} - \sum_{k=1}^{G} \{ n_k \log\phi_k + ½(p-1) n_k \log\sigma_k^2 + \frac{1}{2\sigma_k^2} \sum_{i \in E_k} \hat{\mathbf{v}}_i^T \hat{\mathbf{v}}_i \}. \qquad (3.1)$$

If we assume that $\sigma_k^2 = \sigma^2$ $(k = 1, \ldots, G)$, then the log-likelihood in equation (3.1) is maximized by

$$\hat{\phi}_k = \max_{i \in E_k} \{\hat{u}_i\} - \min_{i \in E_k} \{\hat{u}_i\},$$

$$\hat{\sigma}^2 = \{n(p-1)\}^{-1} \sum_{i=1}^{n} \hat{\mathbf{v}}_i^T \hat{\mathbf{v}}_i .$$

We therefore choose $\boldsymbol{\gamma}$ so as to minimize the criterion

$$½(p-1) n \log\hat{\sigma}^2 + \sum_{k=1}^{G} n_k \log\hat{\phi}_k . \qquad (3.2)$$

In the situation where the σ_k^2's are not constant across clusters we obtain

$$\hat{\sigma}_k^2 = \{n_k(p-1)\}^{-1} \sum_{i \in E_k} \hat{\mathbf{v}}_i^T \hat{\mathbf{v}}_i ,$$

and $\boldsymbol{\gamma}$ is chosen so as to minimize the criterion

$$U = \sum_{k=1}^{G} \{ ½(p-1) n_k \log\hat{\sigma}_k^2 + n_k \log\hat{\phi}_k \}. \qquad (3.3)$$

Many variations on this "uniform-normal" theme are possible, and lead to simple criteria. For example, clusters may be distributed tightly about a two-dimensional planar region in p-space; this can be represented by specifying the distribution of (z_{i1}, z_{i2}) to be concentrated on such a region. Also, the distribution of the scatter about the main line segment or planar region may be more general than assumed above, leading, for example, to a range of values for the covariance matrix of $\mathbf{v}_i$, such as those considered in Section 2.

3.2 Allowing for noise

So far, we have assumed that each observation belongs to a cluster. However, even when a data set is made up mainly of clusters of the prescribed type, there may be other data points that do not follow this pattern. We allow for this possibility by extending our model to include such observations, assumed to arise from a Poisson process with intensity ν. Let $E = \bigcup_{k=1}^{G} E_k$ and $n_0 = n - \sum_{k=1}^{G} n_k$. Then the likelihood (1.1) is modified as follows:

$$L(\mathbf{x}; \boldsymbol{\theta}, \nu, \boldsymbol{\gamma}) = \nu^{n_0} e^{-\nu A} \prod_{i \in E} f_{\gamma_i}(\mathbf{x}_i ; \boldsymbol{\theta}), \tag{3.4}$$

where A is the hypervolume of the region from which the data have been drawn. The clustering criteria developed so far can easily be modified so as to be based on (3.4). Taking account of noise in this way facilitates our proposed method for choosing the number of clusters, described in Section 4.

4. CHOOSING THE NUMBER OF CLUSTERS: AN APPROXIMATE BAYESIAN APPROACH

Here we suggest an approximate Bayesian approach to the choice of the number of clusters. We first write down an exact Bayesian solution, but this usually cannot be computed in a reasonable amount of time. Arguing heuristically, we obtain an approximation to the Bayesian solution which seems to work well in numerical experiments, some of which are reported in Section 6.

We view the problem of estimating the number of clusters as one of choosing between competing models for the same data. The exact Bayesian solution consists of finding the posterior probability $p(G \mid \mathbf{x})$ of each number of clusters G given the data $\mathbf{x}$. This approach seems to have advantages over the alternative of hypothesis testing in the general context of model comparison, as it avoids the problems of multiple comparisons, comparing non-nested models, and the tendency of hypothesis tests to select unparsimonious models when the sample size is large (Berger and Sellke 1987; Raftery 1986b, 1988b). The details have been worked out for many statistical problems, including the general linear model (Smith and Spiegelhalter 1980), generalized linear models (Raftery 1986a, 1988b), change-points and point processes (Akman and Raftery 1986; Raftery and Akman 1986), and software reliability (Raftery 1987, 1988a).

Technically, it is easiest to start with the Bayes factor, or ratio of posterior to prior odds for $G=r$ against $G=s$, defined by

$$B_{rs} = p(\mathbf{x} \mid G=r) / p(\mathbf{x} \mid G=s). \tag{4.1}$$

In (4.1), $p(\mathbf{x} \mid G=r)$ is the marginal likelihood

$$p(\mathbf{x} \mid G=r) = \sum_{\boldsymbol{\gamma} \in \Gamma_r} \iint L(\mathbf{x}; \boldsymbol{\theta}, \nu, \boldsymbol{\gamma}) p(\boldsymbol{\theta}, \nu, \boldsymbol{\gamma}) \, d\boldsymbol{\theta} \, d\nu,$$

where $\Gamma_r = \{0,1, \ldots, r\}^n$, $L(\mathbf{x}; \boldsymbol{\theta}, \nu, \boldsymbol{\gamma})$ is the generalized likelihood defined by (3.3), and $p(\boldsymbol{\theta}, \nu, \boldsymbol{\gamma})$ is the joint prior density of $\boldsymbol{\theta}$, ν, and $\boldsymbol{\gamma}$. When $\gamma_i=0$, $\mathbf{x}_i$ belongs to the "noise" and appears in the Poisson part of the likelihood (3.3). A different but related approach is described by Rissanen (1988).

Here we concentrate on the approximate calculation of $B_{r,r+1}$ $(r=1, \ldots, n-1)$. This yields posterior probabilities $p(G=r \mid \mathbf{x})$ directly, as follows. Noting that $B_{rs} = \prod_{t=1}^{s-r} B_{r+t-1,r+t}$ $(r<s)$ and $B_{sr} = B_{rs}^{-1}$, we calculate B_{rs_0} for $r=1, \ldots, n-1$ and some fixed s_0. Then

$$p(G=r \mid \mathbf{x}) = B_{rs_0} p(G=r) \Big/ \sum_{t=1}^{n-1} B_{ts_0} p(G=t), \tag{4.2}$$

where $p(G=r)$ is the prior probability that there are r clusters.

We approximate $B_{r,r+1}$ as follows. In an agglomerative hierarchical clustering algorithm, the choice between $G=r+1$ and $G=r$ is a decision whether or not to merge two particular clusters into one. In the p-dimensional multivariate normal case, this is exactly a standard comparison of nested hypotheses in the general linear model, and Smith and Spiegelhalter

(1980) have shown that in that case minus twice the logarithm of the Bayes factor is approximately equal to

$$\lambda_r - \{\tfrac{3}{2}+\log(pn_{r,r+1})\}\,\delta_r, \tag{4.3}$$

where λ_r is the likelihood ratio test statistic, δ_r is the number of degrees of freedom in the asymptotic chi-square distribution of λ_r, and $n_{r,r+1}$ is the number of observations in the merged cluster. However, (4.3) is invalid in the clustering context because the regularity conditions on which it is based do not hold. Wolfe (1971) suggested getting around the problem by doubling the number of degrees of freedom, and Hernandez-Alvi (1979) found that to be a reasonable approximation. Aitkin, Anderson and Hinde (1981) had some reservations about the use of Wolfe's (1971) approximation when δ_r is large, but the simulations of Everitt (1981) showed it to perform well for values of δ_r between 1 and 5, which is the range of primary interest to us. We therefore use the approximation

$$\begin{aligned} -2\log B_{r,r+1} &\approx \lambda_r - \{\tfrac{3}{2}+\log(pn_{r,r+1})\}\,2\delta_r \\ &= E_r, \end{aligned} \tag{4.4}$$

where δ_r is now the decrease in the number of parameters caused by going from $G=r+1$ to $G=r$.

In Table 2, for the case where the data are two-dimensional, we show the values of δ_r and the individual cluster parameters that must be estimated for the clustering criteria from Sections 2 and 3. We write $D = \begin{bmatrix} \cos\psi & -\sin\psi \\ \sin\psi & \cos\psi \end{bmatrix}$, where ψ is the orientation of the cluster. For the criteria in Section 3, ϕ_k can be superefficiently estimated, and so it is not included in the count. The term λ_r in (4.4) involves only the contributions to the likelihood of the clusters involved in the merger. If we define the maximized likelihoods for the two clusters that are merged as $l_{k'}$ and $l_{k''}$ and the maximized likelihood for the cluster resulting from the merger as l_k, we may write

$$\lambda_r = -2(l_{k'} + l_{k''} - l_k) \tag{4.5}$$

The likelihoods for the clusters that are not involved in the merger cancel out in the likelihood ratio.

Table 2
Decrease in the number of parameters caused by reducing the number of clusters by one, for several criteria.

Criterion	δ_r	Parameters
$\mathrm{tr}(W)$	2	μ_x, μ_y
$\lvert W \rvert$	2	μ_x, μ_y
S	3	μ_x, μ_y, ψ
$\sum_{k=1}^{G} n_k \log \mathrm{tr}(W_k/n_k)$	3	μ_x, μ_y, λ_k
S^*	4	$\mu_x, \mu_y, \psi, \lambda_k$
$\sum_{k=1}^{G} n_k \log \lvert \frac{W_k}{n_k} \rvert$	5	$\mu_x, \mu_y, \psi, \lambda_k, \alpha_{2k}$
Equation (3.2)	3	μ_x, μ_y, ψ
U	4	$\mu_x, \mu_y, \psi, \sigma_k^2$

If we assume that the clusters are embedded in a Poisson process, the outcomes of the mergers are slightly more complicated since at each stage in the agglomerative process the number of clusters, G, can increase, decrease or remain the same. The reason for this is that we have two types of data, clusters and noise. If we form a new cluster by merging two singletons that were considered noise, then G will increase. If we merge a singleton with an existing cluster, then G will not change. If two existing clusters are merged, then G will decrease. If two singletons are merged to form cluster k, then $\lambda_r = 2l_k$, and δ_r for the merger is equal to minus the value of δ_r given in table 2. If a singleton is merged with cluster k' to form cluster k then $\lambda_r = -2(l_{k'} - l_k)$ and $\delta_r = 0$ since the parameterization has not changed. When two existing clusters, k' and k'', are merged to form cluster k, λ_r is given by equation (4.4) and δ_r is as given in Table 2.

Having obtained $B_{r,r+1}$ $(r = 1, \ldots, n-1)$ from (4.4), we may calculate $p(G=r \mid \mathbf{x})$ $(r = 1, \ldots, n-1)$ using (4.2). A simple approach is to use as an estimate of the number of clusters the value of r for which $p(G{=}r \mid \mathbf{x})$ is greatest. However, (4.4) provides a rather crude approximation to $p(G{=}r \mid \mathbf{x})$. We therefore prefer to consider several values of the number of clusters, guided by the values of $p(G{=}r \mid \mathbf{x})$, or, equivalently, by $F_r = \sum_{t=1}^{r-1} E_t \approx \text{constant} + 2\log p\,(G{=}r \mid \mathbf{x})$. Following Good (1983), we refer to F_r as the *approximate weight of evidence (AWE)* for the number of clusters being r. In our experience, the change in the approximate weight of evidence, $E_r = F_{r+1} - F_r$ is often large and positive for the first few values of r, $r = 1, \ldots, R$, say, and small or negative thereafter. If that is the case, considerations of parsimony lead us to consider $G = R+1$, as well as the value of r which maximizes the approximate weight of evidence, and any intervening values.

5. SIMULATION RESULTS

To compare the performance of our clustering criteria with that of standard, commonly used criteria, we carried out a Monte Carlo study. The standard criteria used were the single-link method (SL), and Ward's sum of squares criterion, tr(W). These were compared with the three criteria S, introduced for two dimensions by Murtagh and Raftery (1984) and generalized in Section 2, S^* defined by (2.5), and U defined by (3.3).

To compare the criteria we generated 100 random samples from each of three types of data for each of four values of α, giving a total of 1200 samples. The three types of data correspond to the three models for which S, S^* and U are optimal criteria. When generating the data for which U was optimal, ϕ_k was generated from a U(.2, .6) distribution and $\hat{\sigma}_k^2$ was proportional to $\alpha\phi_k^2$. Each sample consisted of three clusters, the orientation of each cluster was randomly chosen from a U(0, π) distribution, and the centers were randomly chosen in the unit square. The number of points in each cluster was generated from a discrete uniform distribution on the integers between 15 and 25.

Tables 3, 4 and 5 show the proportion of points misclassified by each of the five criteria considered. The single-link method performed poorly, while Ward's sum of squares did only slightly better. The three criteria S, S^* and U all performed much better. Of these three, S^* did marginally better than the others, but the differences between them were small. As one might expect, each of the three criteria S, S^* and U performed best on the type of data for which it was designed, but it also performed well on the other kinds of data.

The clear superiority of S, S^* and U to the single-link and Ward's method held for each combination of the three kinds of data with the four values of α. The results for the three kinds of data were quite similar. As α increased, the proportion of points misclassified by S, S^* and U increased. This reflects the fact that as α increases, the data generating mechanism more closely approximates that for which tr(W) is the best criterion, and so the superiority of S, S^* and U becomes less marked. Averaged over the 1,200 random samples generated, the proportion of points misclassified was 16% for S, 14% for S^*, 15% for U, 47% for the single-link method and 43% for Ward's sum of squares.

It is assumed that some prior information about α is available. This can come from a training sample or knowledge of the mechanism generating the data, for example the resolution of the edge detector used in processing a digital image. Our numerical work, including the analysis of Example 3 described in Section 6.3, indicates that our criteria are not sensitive to errors in the estimation of α. In the simulation study the correct value of α was used in S, S^* and U. This provides information on the best performance that can be expected.

6. EXAMPLES

6.1 Example 1: Simulated clusters

Figure 1(a) shows three clusters generated from bivariate normal distributions with the same shape but different sizes and orientations. It is typical of the 400 random samples for which the results are summarized in Table 4.

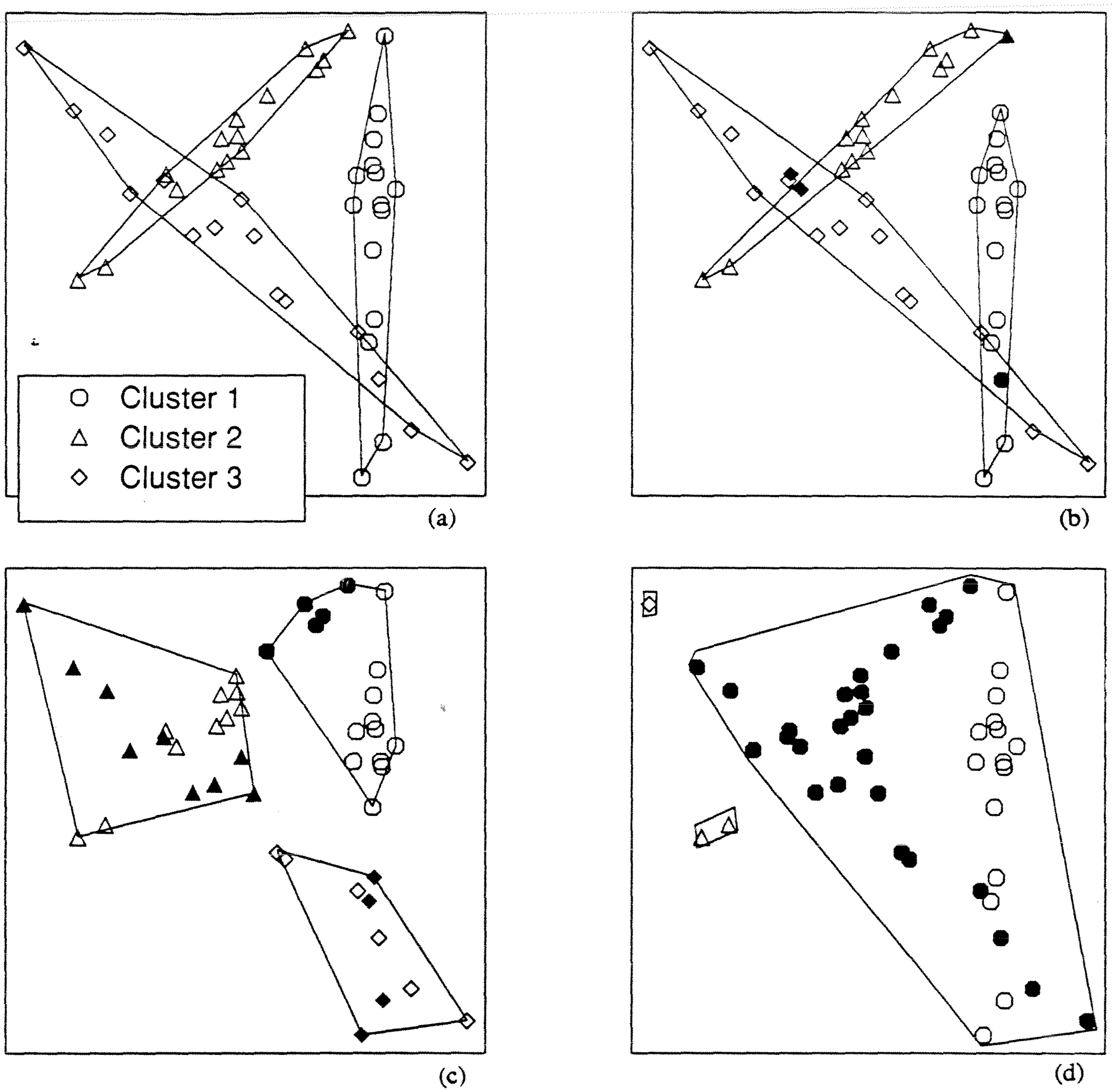

Figure 1. (a) Three clusters generated from bivariate normal distributions with the same shape but different sizes and orientations. The solid lines are the convex hulls of the groups. (b) The clusters formed by the S^* criterion. The filled-in symbols represent misclassified points. For example, the filled-in triangle at the top right-hand corner was classified as a diamond, but in fact is a circle. (c) The clusters found by Ward's sum-of-squares criterion, tr(W). (d) The clusters found by the single-link method.

Table 3

Bivariate normal clusters with the same size and shape but different orientations. S is the optimal criterion. 100 random samples were generated for each value of α. The entries in the table are the percentages of points misclassified.

Criterion	α			
	.001	.005	.01	.025
S	4	13	16	25
S^*	4	16	18	24
U	7	19	26	30
SL	51	51	52	52
tr(W)	40	41	41	40

Table 4

Bivariate normal clusters with the same shape but different sizes and orientations. S^* is the optimal criterion. 100 random samples were generated for each value of α. The entries in the table are the percentages of points misclassified.

Criterion	α			
	.001	.005	.01	.025
S	14	17	24	31
S^*	7	12	17	22
U	7	12	18	26
SL	46	47	50	48
tr(W)	48	48	46	46

Figures 1(b,c,d) show the results of grouping the data into three clusters using the criteria S^*, tr(W) and SL respectively. The S^* criterion performed well. Three of the four misclassified points are within or close to the intersections of the clusters. This is inevitable, since even the human eye, with its remarkable pattern recognition and classification abilities, finds it hard to classify points at the intersection of clusters. Ward's criterion, tr(W), misclassified 18 of the 45 points and did not reproduce the general shape of the clusters. As can be seen from Figure 1(c),

Table 5

Bivariate uniform-normal clusters. The observations are clustered uniformly along and tightly about a line segment in two-dimensional space, as described in Section 3.1. U is the optimal criterion. 100 random samples were generated for each value of α. The entries in the table are the percentages of points misclassified.

Criterion	α			
	.001	.005	.01	.025
S	4	11	13	18
S^*	5	9	12	19
U	3	7	9	14
SL	38	41	45	43
tr(W)	43	41	44	43

it tends, instead, to find "circular" clusters. The single-link method has been suggested for finding long clusters such as those in Figure 1(a). However, as can be seen from Figure 1(d) and Tables 3, 4 and 5, it does not perform well when the clusters intersect.

Clusters that are physically separate, in whatever metric is being used, are easy to distinguish with most clustering criteria. The clusters we have been working with are distinguished from each other by their structure. A point within one cluster may be closer, in Euclidean distance, to points in other clusters than to any point in the cluster to which it belongs, yet we are able to classify it correctly due to the structure of the clusters. For example, consider Figure 1(b). Note the two points on the left that have been correctly classified as belonging to cluster 2 (triangles) yet they are closer to points in cluster 3 (diamonds) than to any point in cluster 2. Criteria based strictly on distance measures, such as the single link method, are unable to handle clusters that are defined by their structure.

6.2 Example 2: Simulated clusters with noise

Figure 2 shows three clusters with added noise. The clusters were generated from bivariate normal distributions with the same shape but different sizes and orientations while the noise was generated by a Poisson process. This example differs from Example 1 in that noise has been added and that we do not assume the numbers of clusters to be known in advance.

After clustering the data in Figure 2 using S^* in a hierarchical agglomeration procedure, the approximate weight of evidence (AWE) was calculated at each iteration, as shown in Figure 3. The AWE is maximized at iteration 47 and falls off sharply after that, indicating that the clustering algorithm should be stopped at the 47th iteration. Figure 4 shows the results at iteration 47 after using an iterative relocation algorithm to improve upon the original agglomerative results. The three main clusters are well-defined with one misclassification, and only one of the noise points has been misclassified.

6.3 Example 3: Diabetes data

Reaven and Miller (1979) described and analyzed data consisting of the area under a plasma glucose curve (Glucose Area), the area under a plasma insulin curve (Insulin Area) and steady state plasma glucose response (SSPG) for 145 subjects. The subjects were clinically classified into three groups, chemical diabetes, overt diabetes and normal (non-diabetic). Symons (1981) reanalyzed the data using seven different clustering criteria. Taking the clinical classification to be correct, we evaluate one of our criteria and compare it with those considered by Symons (1981), using the data as published in Andrews and Herzberg (1985).

Reaven and Miller (1979, Figures 1-4) showed four two-dimensional projections of the data. The data have the 3-dimensional shape of a boomerang with two wings and a fat middle. One of the wings corresponds to patients with overt diabetes, the other wing is composed primarily of patients with chemical diabetes and the "fat middle" is composed of normal patients. By viewing the data using a rotating 3-dimensional scatterplot, such as the ones provided in MacSpin (Donoho, Donoho and Gasko, 1988) or XLISP-STAT (Tierney, 1988), this

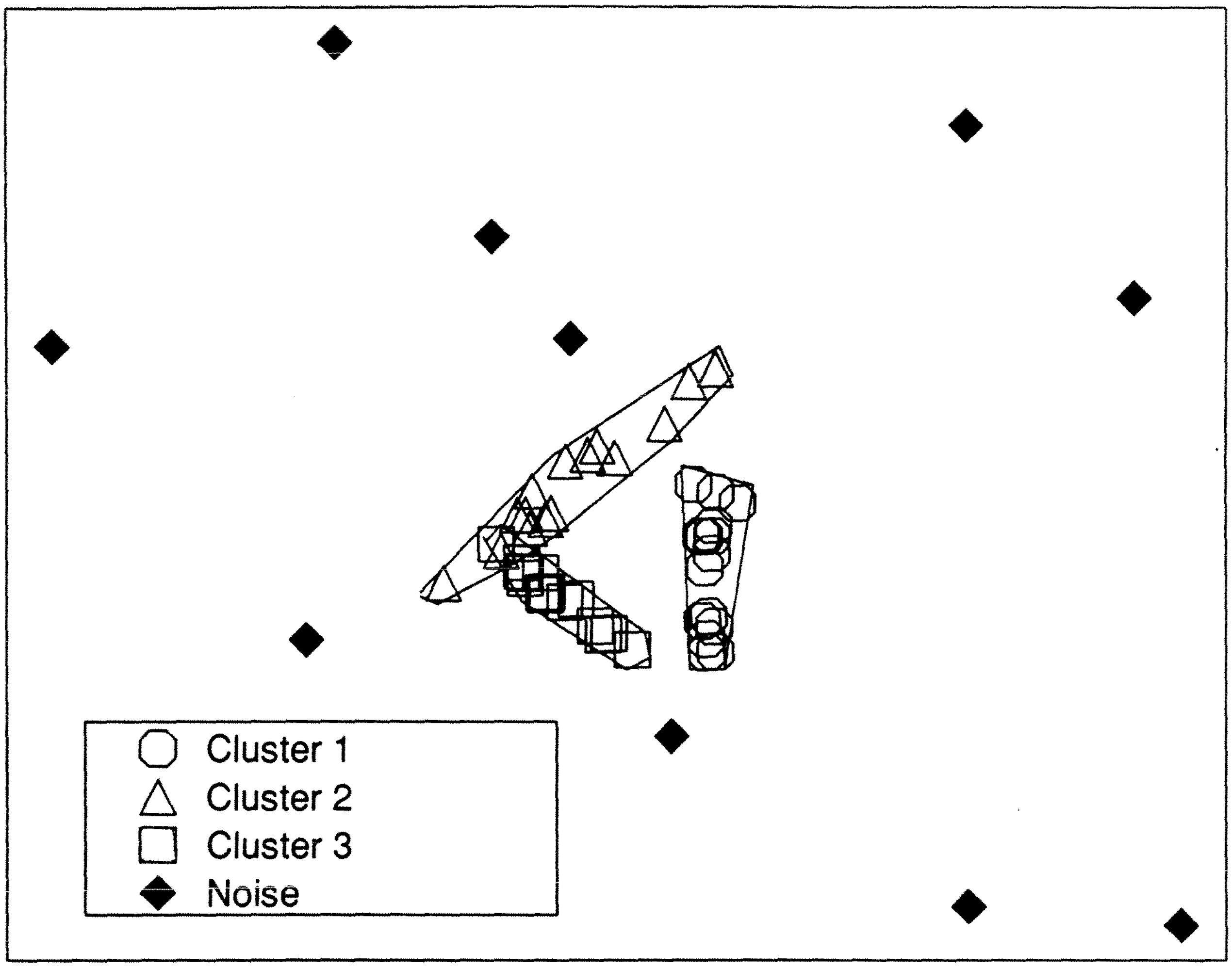

Figure 2. Three clusters with noise. The clusters were generated from bivariate normal distributions with the same shape but different sizes and orientations. The noise was generated from a Poisson process.

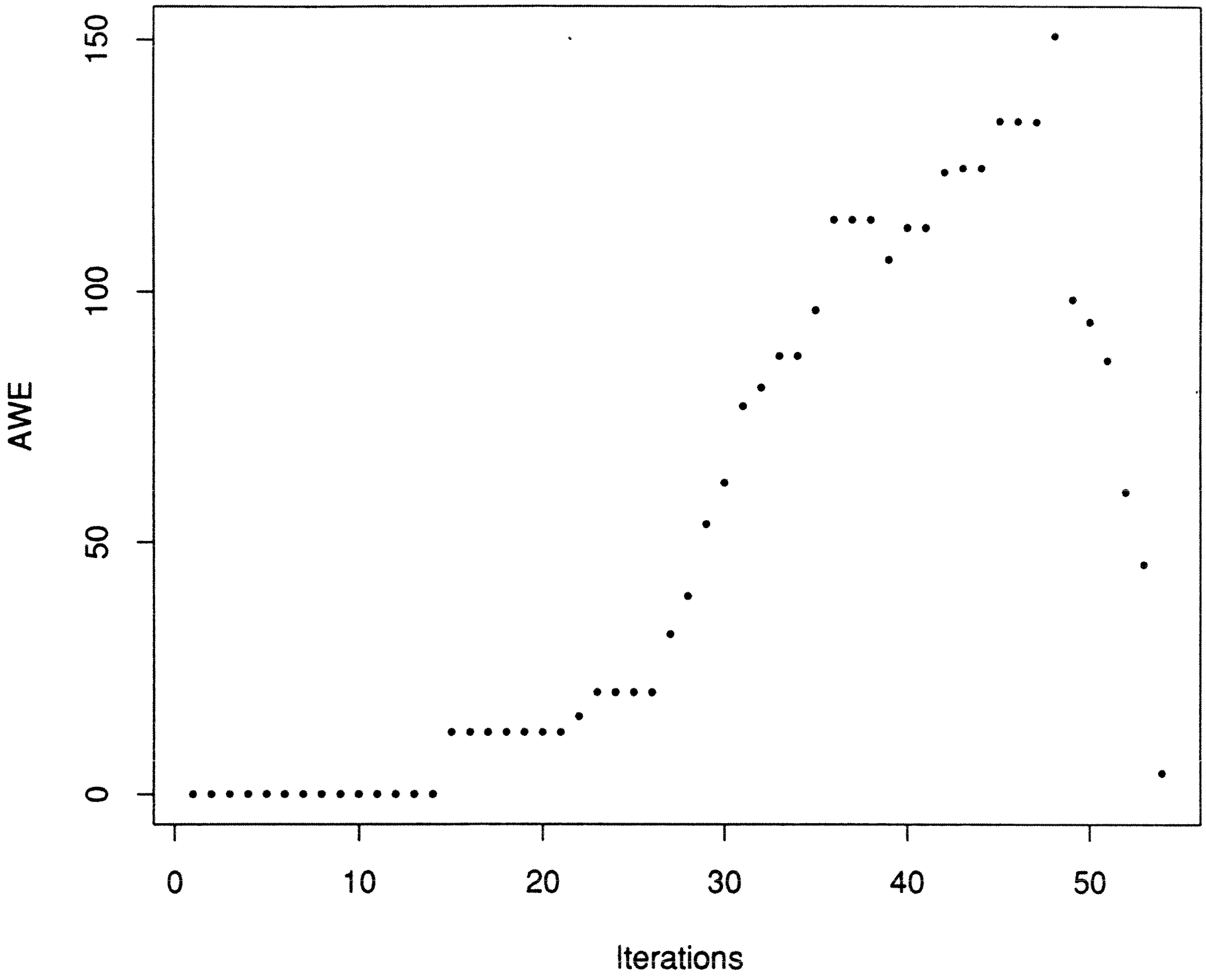

Figure 3. Approximate weight of evidence (AWE) for the number of clusters in Figure 2 using the criterion S^*. The maximum occurs at iteration 47 and leads to the clusters shown in Figure 4.

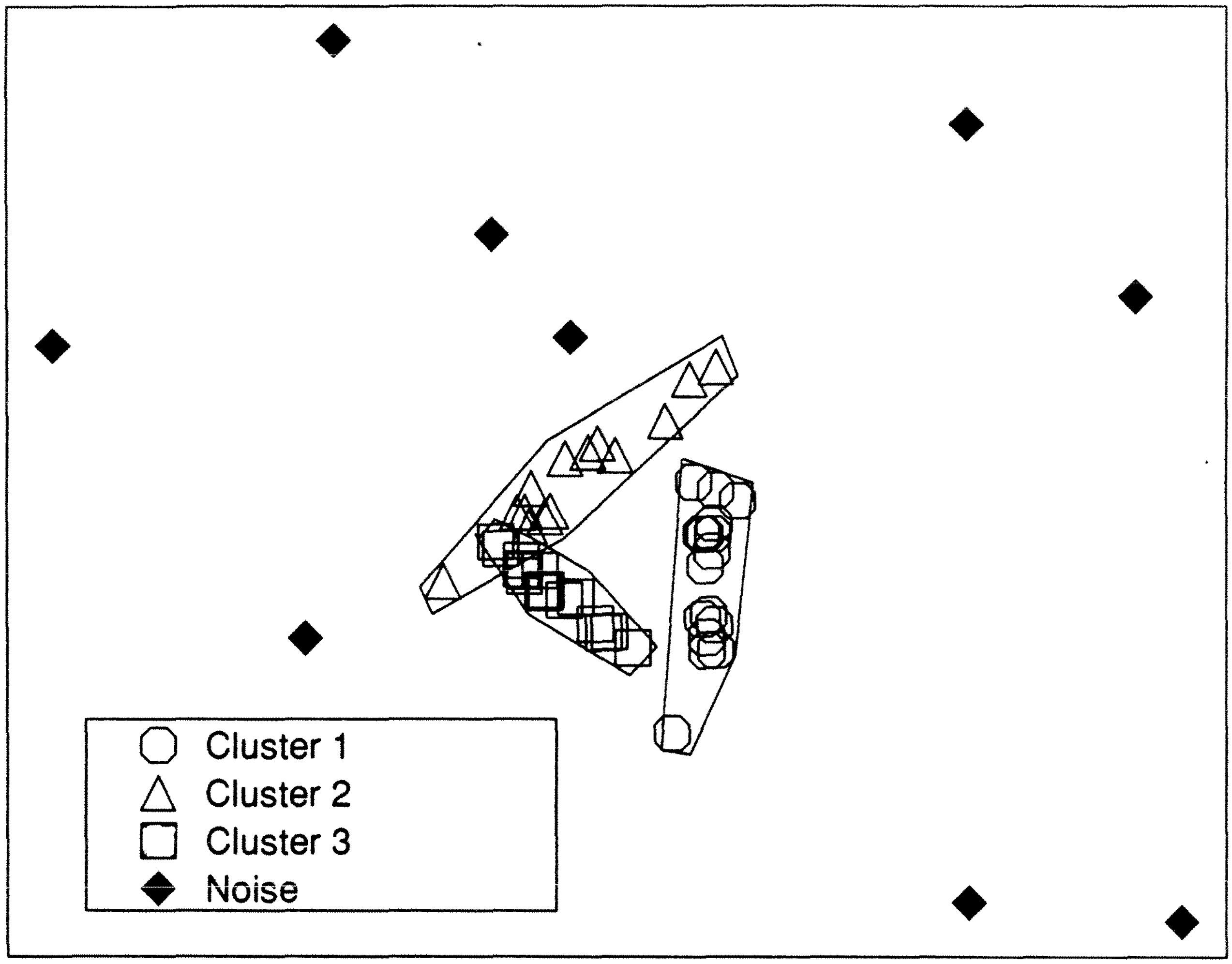

Figure 4. The clusters resulting from the data in Figure 2 using the criterion S^* and stopping at the 47th iteration as indicated by the AWE in Figure 3.

structure is obvious and several other features become apparent. One of the "wings" is almost planar, the other is linear with some curvature, and the "fat middle" has a shape similar to an American football. Four two-dimensional projections of the data are shown in Figure 5.

Based on this information, we could use the approach developed in Sections 2 and 3 to design a purpose-built clustering criterion for this application. However, we prefer to use a very general criterion of the form S^*, where $A_k = \mathrm{diag}\{1, \alpha, \alpha\}$. This criterion is optimal for trivariate normal clusters with different sizes and orientations but the same "tubular" shape, clustered circularly about a line in $\mathbf{R}^3$. The estimated values of α for the three clinically identified groups are .09, .19 and .34. The results were relatively insensitive to changes in α so long as it remained in that broad range. The results we report are for $\alpha = .2$.

Starting from the correct clinical classification and using a single point iterative relocation algorithm with the criterion S^*, the optimal classification, as given in Table 6, resulted in only 10% of the points being misclassified. This compares favorably with the results given by all seven clustering criteria used by Symons (1981) for this data set. We also used a hierarchical agglomerative clustering algorithm followed by iterative relocation. Once again, the results compare favorably with those of Symons (1981). The clusters found are shown in Figure 6.

The AWE for the hierarchical agglomerative clustering algorithm increased steadily until the final 5 iterations. Figure 7 shows the number of clusters versus the AWE over the last 20 iterations. From this it can be seen that the AWE increases sharply as one goes from one cluster to two, and again from two to three. It increases slightly as the number of clusters goes up to four and five, and decreases thereafter. If we did not know the true number of clusters this would lead us to focus attention on the groupings into three, four and five clusters, and to perform a more detailed analysis on these sets of clusters.

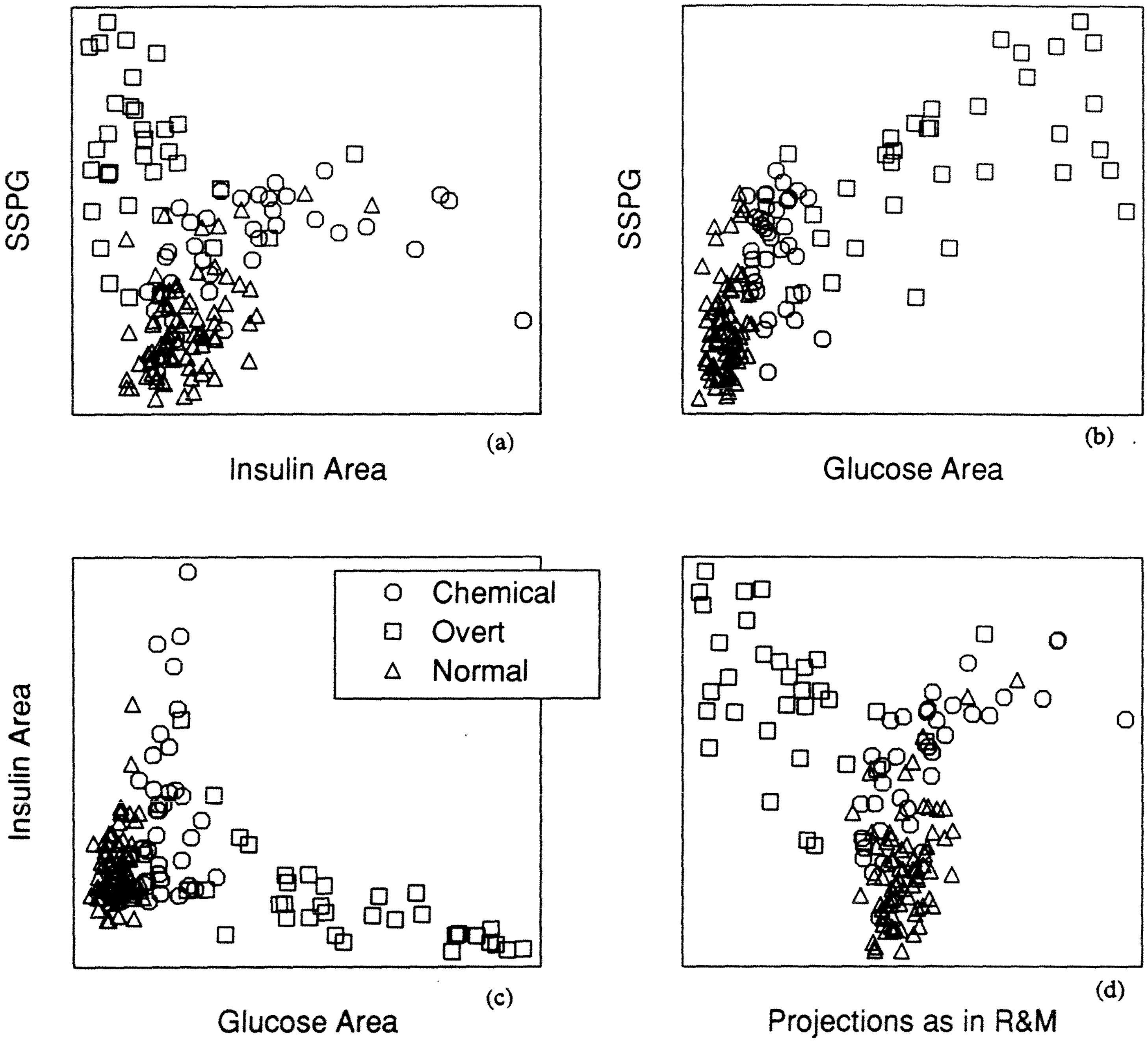

Figure 5. Four two-dimensional projections of the three-dimensional diabetes data of Reaven and Miller (1979). The symbols indicate the clinical classification of subjects as having chemical diabetes, overt diabetes or being normal. (d) Shows the approximate projections represented by the artist's sketch in Reaven and Miller (1979) and reproduced in Symons (1981).

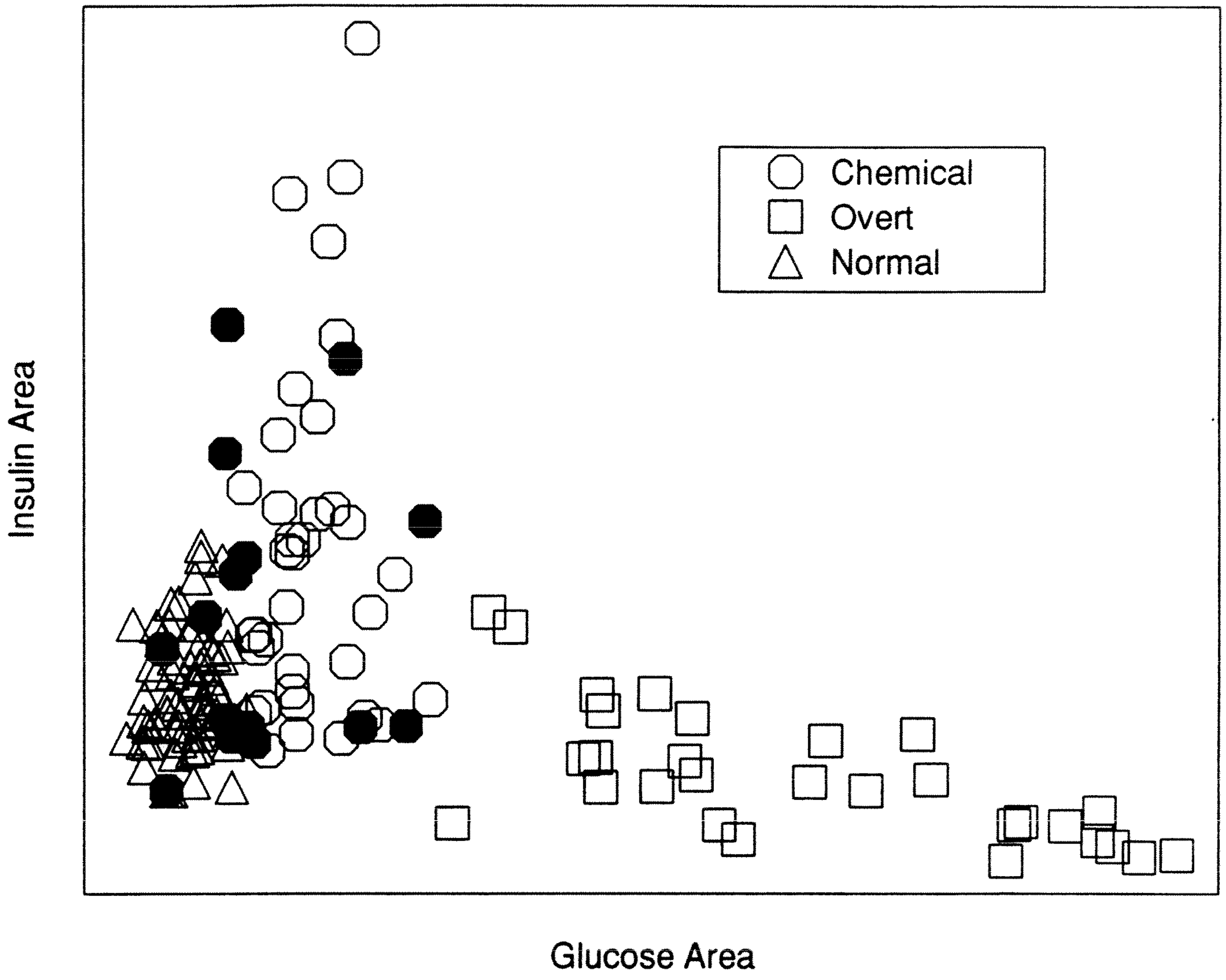

Figure 6. The three clusters in the diabetes data found by hierarchical agglomeration followed by iterative relocation using the criterion S^* with $A_k = \text{diag}\{1, .2, .2\}$. The two-dimensional projection shown is that of Figure 5(c). The symbols indicate the classification of the subjects based on the clustering algorithm. The filled-in symbols represent subjects whose clustering classification differs from the clinical classification.

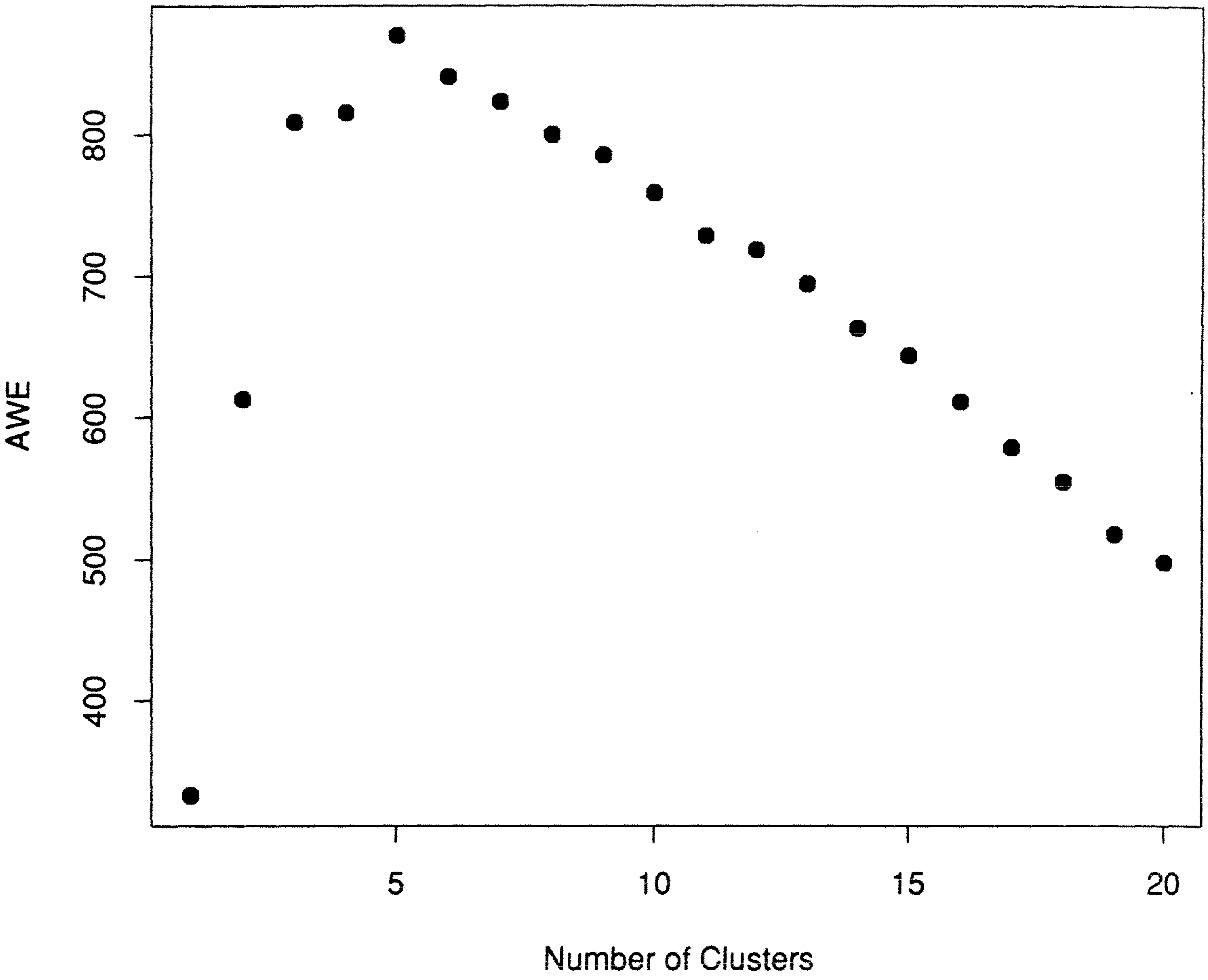

Figure 7. Approximate weight of evidence (AWE) for the number of clusters in the diabetes data over the last 20 iterations of the clustering algorithm. The AWE increases sharply up to three clusters, with further slight increases up to five clusters, and decreases thereafter. This would lead us to focus on the groupings into three, four and five clusters.

Table 6

Results of clustering the diabetes data. The first row shows the result of single point iterative relocation using the criterion S^* with $A_k = \mathrm{diag}\{1, .2, .2\}$, starting with the clinical classification. The second row shows the result of hierarchical agglomeration followed by iterative relocation with the same criterion. The remaining rows show the results of seven other clustering procedures, starting at the clinical classification, as reported by Symons (1981). Criterion (13) of Symons (1981) is due to Maronna and Jacovkis (1974). The error rate % is the percentage of the subjects who were not classified in the same way by the clustering method as by the clinical diagnosis.

Method	Error	Clinical classification		
	rate %	Normal (76,0,0)	Chemical (0,36,0)	Overt (0,0,33)
S^* from clinical	10	(65,0,0)	(11,36,4)	(0,0,29)
S^* agglomerative	10	(65,0,0)	(11,36,4)	(0,0,29)
$\mid W \mid$	19	(73,17,3)	(3,19,4)	(0,0,26)
Reaven and Miller (1979) variant of $\mid W \mid$	14	(73,10,1)	(3,26,6)	(0,0,26)
(8) in Symons (1981)	26	(75,30,6)	(1,6,1)	(0,0,26)
(10) in Symons (1981)	26	(75,30,6)	(1,6,1)	(0,0,26)
(13) in Symons (1981)	13	(73,10,0)	(3,26,7)	(0,0,26)
(11) in Symons (1981)	14	(63,0,0)	(13,30,2)	(0,6,31)
(12) in Symons (1981)	13	(73,9,0)	(3,27,7)	(0,0,26)

7. DISCUSSION

We have proposed ways of overcoming some of the limitations of the classification maximum likelihood procedure for cluster analysis, as currently implemented. These are (1) the inability to specify some but not all features (orientation, size, shape) to be constant across clusters; (2) the restriction to normal distributions; and (3) the failure to account for "noise". We have also proposed an approximate Bayesian solution to the problem of choosing the number of clusters, which seems to avoid some of the difficulties associated with solutions to this problem based on significance testing.

In the context of Gaussian clustering, we reparameterize the covariance matrices in terms of their eigenvalue decompositions. Each group of parameters then corresponds clearly to a particular feature of the cluster (orientation, size or shape), and criteria appropriate for a range of different situations result by constraining none, some or all of these features to be constant across clusters. This leads to a range of criteria which are more general than that of Friedman and Rubin (1967) and more parsimonious than that of Scott and Symons (1971) for the unequal covariance case. The reparameterization of covariance matrices in terms of the eigenvalue decomposition has also been considered by Flury (1988) although he did not view it in the context of cluster analysis and he assumed the eigenvector matrices, D_k, to be the same across all groups.

A general and practical approach to non-Gaussian clustering is introduced. It is developed in detail for the important special case where points are distributed uniformly along and tightly about a line segment in p-space. "Noise" is allowed for by permitting isolated observations to be distributed over the data region according to a Poisson process. We propose an approximate Bayesian method for choosing the number of clusters. We also write down the exact Bayesian solution, which is optimal given the model, but is usually not computable; our approximation seems to perform well in numerical examples.

An alternative specification of the model (1.1), which leads to the so-called mixture maximum likelihood approach, has been considered by Wolfe (1970), Symons (1981), McLachlan (1982) and McLachlan and Basford (1988). This assumes that $\mathbf{x}$ is a random sample from a mixture of the G densities $f_k(\mathbf{x};\theta)$ $(k=1,\ldots,G)$ in the proportions $\varepsilon=(\varepsilon_1,\ldots,\varepsilon_G)^T$. Then θ and ε are estimated, and conditional probabilities $p(\gamma_i=k \mid \mathbf{x},\hat{\theta},\hat{\varepsilon})$ are calculated. Marriott (1975) and Bryant and Williamson (1978) showed that when, unlike here, estimation of θ is of primary interest, then the classification maximum likelihood method is inconsistent. However, when the covariance matrices are unequal, the mixture maximum likelihood approach appears to break down in practice (Day 1969). McLachlan and Basford (1988, Section 2.1) discuss some theoretical results which suggest that it may be possible to apply the mixture maximum likelihood approach when the covariance matrices are unequal, but this does not seem to have been done yet. If it could be done, it seems likely that the methods proposed in this paper could also be extended to the mixture maximum likelihood approach using the EM algorithm (McLachlan and Basford, 1988, Section 1.6).

The classification and mixture maximum likelihood approaches are in conflict only when the primary aim is to estimate $\boldsymbol{\theta}$; the conflict is resolved when, as here, the aim is to estimate $\boldsymbol{\gamma}$, and $\boldsymbol{\theta}$ is a nuisance parameter. This is easiest to see in a Bayesian framework, where the full solution is the posterior distribution $p(\boldsymbol{\gamma} \mid \mathbf{x})$. It follows from equation (2.2) of Binder (1978) that this is the same under the two models when the prior for $\boldsymbol{\gamma}$ in (1.1) is hierarchical and compatible with the prior for ε in the mixture model. Thus the classification maximum likelihood solution $\hat{\boldsymbol{\gamma}}$ may be viewed as a approximation to the posterior mode of $\boldsymbol{\gamma}$ under both models.

References

Aitkin, M., Anderson, D. and Hinde, J. (1981). Statistical modelling of data on teaching styles (with discussion). *Journal of the Royal Statistical Society, Series A* **144,** 419-461.

Akman, V.E. and Raftery, A.E. (1986). Bayes factors for non-homogeneous Poisson processes with vague prior information. *Journal of the Royal Statistical Society, series B* **48,** 322-329.

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for Students and Research Workers.* New York: Springer-Verlag 215-220.

Berger, J.O. and Sellke, T. (1987). Testing a point null hypothesis: the irreconcilability of *P* values and evidence. *Journal of the American Statistical Association* **82,** 112-122.

Binder, D.A. (1978). Bayesian cluster analysis. *Biometrika* **65,** 31-38.

Bryant, P. and Williamson, J.A. (1978). Asymptotic behaviour of classification maximum likelihood estimates. *Biometrika* **65,** 273-278.

Day, N. (1969). Estimating the components of a mixture of normal distributions. *Biometrika* **56,** 463-474.

Donoho, A.W., Donoho, D.L. and Gasko, M. (1988). MACSPIN: Dynamic graphics on a desktop computer. In *Dynamic Graphics for Statistics,* Cleveland, W.S. and McGill, M.E. (eds.), 331-352. Belmont, Ca.: Wadsworth & Brooks/Cole.

Everitt, B.S. (1981). Contribution to the discussion of paper by M. Aitkin, D. Anderson and J. Hinde. *Journal of the Royal Statistical Society, Series A* **144,** 457-458.

Flury, B. (1988). *Common Principal Components and Related Multivariate Models.* New York: Wiley

Friedman, H. P. and Rubin, J. (1967). On some invariant criteria for grouping data. *Journal of the American Statistical Association* **62,** 1159-78.

Good, I.J. (1983). *Good Thinking: The Foundation of Probability and Its Applications.* Minneapolis: University of Minnesota Press.

Gordon, A.D. (1981). *Classification: methods for the exploratory analysis of multivariate data.* New York: Chapman and Hall.

Gordon, A.D. (1987). A review of hierarchical classification. *Journal of the Royal Statistical Society, Series A* **150,** 119-137.

Hernandez-Alvi, A. (1979). Problems in cluster analysis. Unpublished Ph.D. thesis, University of Oxford.

Jain, A.K. and Dubes, R.C. (1988). *Algorithms for Clustering Data.* Englewood Cliffs, N.J.: Prentice Hall.

Maronna, R. and Jacovkis, P. M. (1974). Multivariate clustering procedures with variable metrics. *Biometrics* **30,** 499-505.

Marriott, F. (1975). Separating mixtures of normal distributions. *Biometrics* **31,** 767-769.

McLachlan, G. (1982). The classification and mixture maximum likelihood approaches to cluster analysis. In *Handbook of Statistics* (Vol. 2), Krishnaiah, P.R. and Kanal, L.N. (eds.), 199-208. Amsterdam: North-Holland,

McLachlan, G. and Basford K. (1988). *Mixture Models: Inference and Applications to Clustering*. New York: Dekker.

Murtagh, F. (1985). *Multidimensional Clustering Algorithms.* CompStat Lectures, **4.** Heidelberg: Physica-Verlag.

Murtagh, F. and Raftery, A. E. (1984). Fitting straight lines to point patterns. *Pattern Recognition* **17,** 479-483.

Raftery, A.E. (1986a). A note on Bayes factors for log-linear contingency table models with vague prior information. *Journal of the Royal Statistical Society, series B* **48,** 249-250.

Raftery, A.E. (1986b). Choosing models for cross-classifications. *American Sociological Review* **51,** 145-146.

Raftery, A.E. (1987). Inference and prediction for a general order statistic model with unknown population size. *Journal of the American Statistical Association* **82,** 1163-1168.

Raftery, A.E. (1988a). Analysis of a simple debugging model. *Applied Statistics* **37,** 12-22.

Raftery, A.E. (1988b). Bayes factors for generalized linear models. Technical report no. 121, Department of Statistics, University of Washington.

Raftery, A.E. and Akman, V.E. (1986). Bayesian analysis of a change-point Poisson process. *Biometrika* **73,** 85-89.

Reaven, G.M. and Miller, R.G. (1979). An attempt to define the nature of chemical diabetes using a multidimensional analysis. *Diabetologia* **16,** 17-24.

Rissanen, J. (1988). On optimal number of features in classification. Unpublished manuscript.

Scott, A.J. and Symons, M.J. (1971). Clustering methods based on likelihood ratio criteria. *Biometrics* **27,** 387-397.

Smith, A. and Spiegelhalter, D. (1980). Bayes factors and choice criteria for linear models. *Journal of the Royal Statistical Society, Series A* **42,** 213-220.

Symons, M. (1981). Clustering criteria and multivariate normal mixtures. *Biometrics* **37,** 35-43.

Tierney, L. (1988). XLISP-STAT: A Statistical Environment Based on the XLISP Language. Technical Report Number 528, School of Statistics, University of Minnesota.

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

Wolfe, J.H. (1970). Pattern Clustering by Multivariate Mixture Analysis. *Multivariate Behavioral Research* **5,** 329-350.

Wolfe, J.H. (1971). A Monte Carlo study of the sampling distribution of the likelihood ratio for mixtures of multinormal distributions. *Technical Bulletin STB 72-2.* San Diego: U.S. Naval Personnel and Training Research Laboratory.

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
186

2. GOVT ACCESSION NO.

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Model-based Gaussian and non-Gaussian Clustering

5. TYPE OF REPORT & PERIOD COVERED
Technical Report
5/1/88-9/30/90

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)
Jeffrey D. Banfield
Adrian E. Raftery

8. CONTRACT OR GRANT NUMBER(s)
N-00014-88-K-0265

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Department of Statistics, GN-22
University of Washington
Seattle, WA 98105

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
NR-661-003

11. CONTROLLING OFFICE NAME AND ADDRESS
ONR Code N63374
1107 NE 45th Street
Seattle, WA 98105

12. REPORT DATE
December 1989

13. NUMBER OF PAGES
31

14. MONITORING AGENCY NAME & ADDRESS(if different from Controlling Office)

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

APPROVED FOR PUBLIC RELEASE: DISTRIBUTION UNLIMITED.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

Bayes factors; Classification; Diabetes; Hierachical agglomeration; Iterative relocation; Mixture models.

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

The classification maximum likelihood approach is sufficiently general to encompass many current clustering algorithms, including those based on the sum of squares criterion and on the criterion of Friedman and Rubin (1967). However, as currently implemented it does not allow the specification of which features (orientation, size and shape) are to be common to all clusters and which may differ between clusters. Also, it is restricted

(continued on reverse)

to Gaussian distributions and it does not allow for noise.

We propose ways of overcoming these limitations. A reparameterization of the covariance matrix allows us to specify that some features, but not all, be the same for all clusters. A practical framework for non-Gaussian clustering is outlined, and a means of incorporationg noise in the form of a Poisson process is described. An approximate Bayesian method for choosing the number of clusters is given.

The performance of the proposed methods is studied by simulation, with encouraging results. The methods are applied to the analysis of a data set arising in the study of diabetes, and the results seem better than those of previous analyses.