J. R. Statist. Soc. B (2001)
63, *Part* 2, *pp.* 411–423

this is a def

Estimating the number of clusters in a data set via the gap statistic

Robert Tibshirani, Guenther Walther and Trevor Hastie

Stanford University, USA

[Received February 2000. Final revision November 2000]

Summary. We propose a method (the 'gap statistic') for estimating the number of clusters (groups) in a set of data. The technique uses the output of any clustering algorithm (e.g. *K*-means or hierarchical), comparing the change in within-cluster dispersion with that expected under an appropriate reference null distribution. Some theory is developed for the proposal and a simulation study shows that the gap statistic usually outperforms other methods that have been proposed in the literature.

Keywords: Clustering; Groups; Hierarchy; *K*-means; Uniform distribution

1. Introduction

So we always get better, but there are diminishing returns to the number of clusters

Cluster analysis is an important tool for 'unsupervised' learning — the problem of finding groups in data without the help of a response variable. A major challenge in cluster analysis is the estimation of the optimal number of 'clusters'. Fig. 1(b) shows a typical plot of an error measure W_k (the within-cluster dispersion defined below) for a clustering procedure *versus* the number of clusters k employed: the error measure W_k decreases monotonically as the number of clusters k increases, but from some k onwards the decrease flattens markedly. Statistical folklore has it that the location of such an 'elbow' indicates the appropriate number of clusters. The goal of this paper is to provide a statistical procedure to formalize that heuristic.

For recent studies of the elbow phenomenon, see Sugar (1998) and Sugar *et al.* (1999). A comprehensive survey of methods for estimating the number of clusters is given in Milligan and Cooper (1985), whereas Gordon (1999) discusses the best performers. Some of these methods are described in Sections 5 and 6, where they are compared with our method.

In this paper we propose the 'gap' method for estimating the number of clusters. It is designed to be applicable to virtually any clustering method. For simplicity, the theoretical part of our analysis will focus on the widely used K-means clustering procedure.

2. The gap statistic

Our data $\{x_{ij}\}$, $i = 1, 2, \ldots, n$, $j = 1, 2, \ldots, p$, consist of p features measured on n independent observations. Let $d_{ii'}$ denote the distance between observations i and i'. The most common choice for $d_{ii'}$ is the squared Euclidean distance $\Sigma_j (x_{ij} - x_{i'j})^2$.

Suppose that we have clustered the data into k clusters $C_1, C_2, \ldots, C_k$, with C_r denoting the indices of observations in cluster r, and $n_r = |C_r|$. Let

Address for correspondence: Robert Tibshirani, Department of Health Research and Policy and Department of Statistics, Stanford University, Stanford, CA 94305, USA.
E-mail: tibs@stat.stanford.edu

1369–7412/01/63411

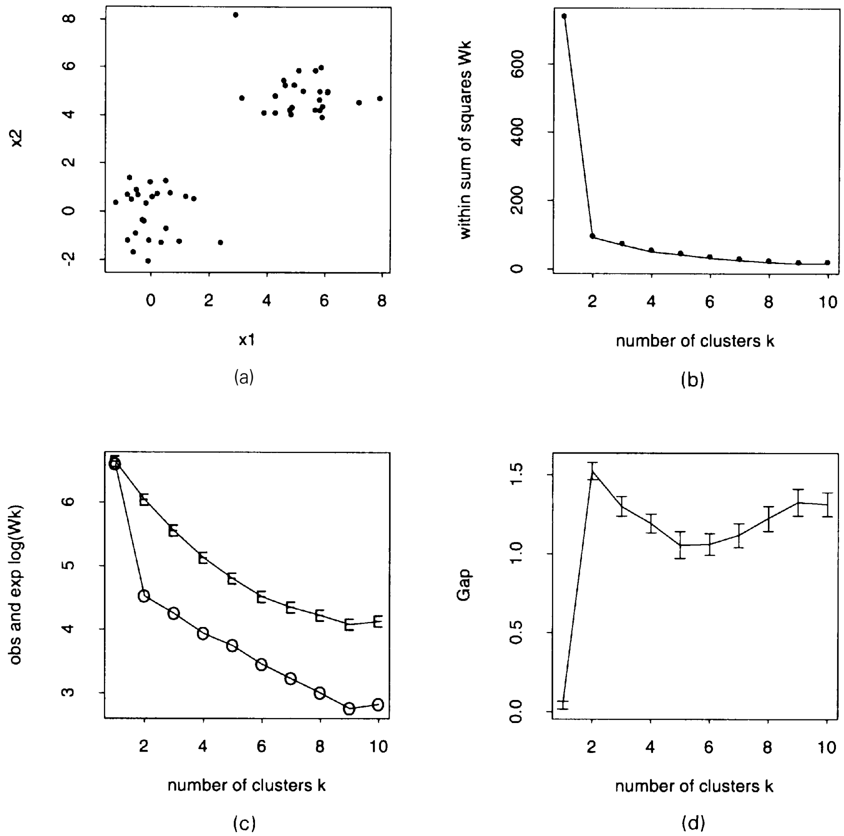

Fig. 1. Results for the two-cluster example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

Sum of distances for points in a single cluster

$$D_r = \sum_{i,i' \in C_r} d_{ii'} \tag{1}$$

be the sum of the pairwise distances for all points in cluster r, and set

Sum over clusters

Paired distance

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r. \tag{2}$$

?!? size of cluster

So, if the distance d is the squared Euclidean distance, then W_k is the pooled within-cluster sum of squares around the cluster means (the factor 2 makes this work exactly). The sample size n is suppressed in this notation.

Data that is similar to what we have, but w/ no clustering

The idea of our approach is to standardize the graph of $\log(W_k)$ by comparing it with its expectation under an appropriate null reference distribution of the data. (The importance of the choice of an appropriate null model is demonstrated in Gordon (1996).) Our estimate of the optimal number of clusters is then the value of k for which $\log(W_k)$ falls the farthest below this reference curve. Hence we define

Under the reference distribution

$$\mathrm{Gap}_n(k) = E_n^*\{\log(W_k)\} - \log(W_k), \tag{3}$$

n data points

Where is the gap between observed & expected under no clustering the greatest?

where E_n^* denotes expectation under a sample of size n from the reference distribution. Our estimate $\hat{k}$ will be the value maximizing $\mathrm{Gap}_n(k)$ after we take the sampling distribution into account. Note that this estimate is very general, applicable to any clustering method and distance measure $d_{ii'}$.

As a motivation for the gap statistic, consider clustering n uniform data points in p dimensions, with k centres. Then, assuming that the centres align themselves in an equally spaced fashion, the expectation of $\log(W_k)$ is approximately

$$\log(pn/12) - (2/p)\log(k) + \text{constant}. \qquad \text{This needs ? rewriting} \tag{4}$$

If the data actually have K well-separated clusters, we expect $\log(W_k)$ to decrease faster than its expected rate $(2/p)\log(k)$ for $k \leqslant K$. When $k > K$, we are essentially adding an (unnecessary) cluster centre in the middle of an approximately uniform cloud and simple algebra shows that $\log(W_k)$ should decrease *more slowly* than its expected rate. Hence the gap statistic should be largest when $k = K$.

As a further motivation, note that, in the case of a special Gaussian mixture model, $\log(W_k)$ has an interpretation as a log-likelihood; see Scott and Symons (1971). To develop the gap statistic into an operational procedure, we need to find an appropriate reference distribution and to assess the sampling distribution of the gap statistic.

3. The reference distribution

In our framework we assume a null model of a single component, and we reject it in favour of a k-component model ($k > 1$), if the strongest evidence for any such k warrants it, i.e. we wish to screen the evidence over all $k > 1$ simultaneously. This approach of guarding against erroneous rejection of the one-component model is similar to that of Roeder (1994). A component (cluster) of the distribution can be appropriately modelled by a log-concave distribution, i.e. by a density of the form $\exp\{\psi(x)\}$, where ψ is a concave function (unless the distribution is degenerate). Standard examples are of course the normal distribution (with $\psi(x) = -\frac{1}{2}\|x\|^2$) and the uniform distribution with convex support. In Walther (2001) it is shown there that it is impossible to set confidence intervals (even one sided) for the number of modes in a multivariate distribution, a crucial aspect for the goal of this paper. Thus we model the components as log-concave densities instead of the often-used unimodal densities. We denote by $\mathcal{S}^p$ the set of such single-component distributions (or random variables) on $\mathbf{R}^p$.

To see how to find an appropriate reference distribution, consider for a moment the population version corresponding to the gap statistic in the case of K-means clustering:

$$g(k) = \log\left\{\frac{\mathrm{MSE}_{X^*}(k)}{\mathrm{MSE}_{X^*}(1)}\right\} - \log\left\{\frac{\mathrm{MSE}_X(k)}{\mathrm{MSE}_X(1)}\right\},$$

where $\mathrm{MSE}_X(k) = E(\min_{\mu \in A_k} \| X - \mu \|^2)$, with the k-point set $A_k \subset \mathbf{R}^p$ chosen to minimize this quantity, is the population version corresponding to W_k. We subtracted off the logarithms of the variances to make $g(1) = 0$. So we are looking for a least favourable single-component reference distribution on X^* such that $g(k) \leqslant 0$ for all $X \in \mathcal{S}^p$ and all $k \geqslant 1$. The first theorem shows that in the univariate case such a reference distribution is given by the uniform distribution $U = U[0, 1]$.

Theorem 1. Let $p = 1$. Then for all $k \geqslant 1$

$$\inf_{X \in \mathcal{S}^p} \left\{ \frac{\mathrm{MSE}_X(k)}{\mathrm{MSE}_X(1)} \right\} = \frac{\mathrm{MSE}_U(k)}{\mathrm{MSE}_U(1)}. \tag{5}$$

In other words, among all unimodal distributions, the uniform distribution is the most likely to produce spurious clusters by the gap test.

Note that the above problem is invariant under changes in location and scale, thus allowing us to restrict attention to the uniform distribution supported on the unit interval. Calculations show that $\mathrm{MSE}_U(k)/\mathrm{MSE}_U(1) = 1/k^2$. So there is a formal similarity to a proposal by Krzanowski and Lai (1985), following Marriott (1971), who suggested to estimate k by comparing successive differences of $W_k k^{2/p}$. Note, however, that their procedure is not defined for the important single-component case $k = 1$. Even more importantly, such an approach will generally fail in a multivariate situation.

Theorem 2. If $p > 1$ then no distribution $U \in \mathcal{S}^p$ can satisfy equation (5) unless its support is degenerate to a subset of a line.

Note that the assertion of the last theorem is not contingent on our definition $\mathcal{S}^p$ of a single-component model. The same conclusion would apply if we based it on, say, unimodal densities instead. Simple calculations show that employing a reference distribution with degenerate support will result in an ineffectual procedure. Thus the upshot of the theorem is that in a multivariate situation we will not be able to choose a generally applicable and useful reference distribution: the geometry of the particular null distribution matters.

An obvious solution would be to generate reference data from the maximum likelihood estimate (MLE) in $\mathcal{S}^p$. This is the nonparametric MLE of the density under the restriction of being log-concave. This MLE can be shown to exist, as opposed to the MLE of a unimodal distribution. In one dimension, this MLE can be computed with the help of the iterative convex minorant algorithm (see Walther (2000)). However, we do not know how to compute the MLE in higher dimensions, but the next section shows how the insights gained from theorems 1 and 2 can be used to construct a simple and effective reference distribution.

4. The computational implementation of the gap statistic

The lesson of theorem 2 was that the multivariate variance structure matters. Our idea is to exploit the shape information in the principal components instead of the more complicated structure provided by the MLE.

We consider two choices for the reference distribution:

(a) generate each reference feature uniformly over the range of the observed values for that feature;

(b) generate the reference features from a uniform distribution over a box aligned with the principal components of the data. In detail, if X is our $n \times p$ data matrix, assume that the columns have mean 0 and compute the singular value decomposition $X = UDV^{\mathrm{T}}$. We transform via $X' = XV$ and then draw uniform features Z' over the ranges of the columns of X', as in method (a) above. Finally we back-transform via $Z = Z'V^{\mathrm{T}}$ to give reference data Z.

Method (a) has the advantage of simplicity. Method (b) takes into account the shape of the data distribution and makes the procedure rotationally invariant, as long as the clustering method itself is invariant.

In each case, we estimate $E_n^*\{\log(W_k)\}$ by an average of B copies $\log(W_k^*)$, each of which is computed from a Monte Carlo sample $X_1^*, \ldots, X_n^*$ drawn from our reference distribution. Finally, we need to assess the sampling distribution of the gap statistic. Let $\mathrm{sd}(k)$ denote the standard deviation of the B Monte Carlo replicates $\log(W_k^*)$. Accounting additionally for the simulation error in $E_n^*\{\log(W_k)\}$ results in the quantity

$$s_k = \surd(1 + 1/B)\,\mathrm{sd}(k).$$

Using this we choose the cluster size $\hat{k}$ to be the smallest k such that $\mathrm{Gap}(k) \geqslant \mathrm{Gap}(k+1) - s_{k+1}$. This '1-standard-error' style of rule is used elsewhere (e.g. Breiman *et al.* (1984)). In the simulation studies later in this paper and in other real data examples, we have found empirically that it works well. A more refined approach would employ a multiplier to the s_k for better control of the rejection of the null model.

Computation of the gap statistic proceeds as follows.

Step 1: cluster the observed data, varying the total number of clusters from $k = 1, 2, \ldots, K$, giving within-dispersion measures W_k, $k = 1, 2, \ldots, K$.

Step 2: generate B reference data sets, using the uniform prescription (a) or (b) above, and cluster each one giving within-dispersion measures W_{kb}^*, $b = 1, 2, \ldots, B$, $k = 1, 2, \ldots, K$. Compute the (estimated) gap statistic

$$\mathrm{Gap}(k) = (1/B) \sum_b \log(W_{kb}^*) - \log(W_k).$$

Step 3: let $\bar{l} = (1/B)\, \Sigma_b \log(W_{kb}^*)$, compute the standard deviation

$$\mathrm{sd}_k = [(1/B) \sum_b \{\log(W_{kb}^*) - \bar{l}\}^2]^{1/2}$$

and define $s_k = \mathrm{sd}_k \surd(1 + 1/B)$. Finally choose the number of clusters via

$$\hat{k} = \text{smallest } k \text{ such that } \mathrm{Gap}(k) \geqslant \mathrm{Gap}(k+1) - s_{k+1}.$$

Fig. 1 shows an example using K-means clustering. The data (Fig. 1(a)) fall in two distinct clusters. The within sum of squares function W_k is displayed in Fig. 1(b). The functions $\log(W_k)$ and $\hat{E}_n^*\{\log(W_k)\}$ are shown in Fig. 1(c), with the gap curve displayed in Fig. 1(d), with ±1 standard error bars. The gap curve has a clear maximum at $\hat{k} = 2$.

Fig. 2 examines the behaviour of the gap estimate with unclustered data. The raw data are 100 observations uniformly distributed over the unit square. The observed and expected curves are very close, and the gap estimate is $\hat{k} = 1$.

4.1. *Example: application to hierarchical clustering and DNA microarray data*

In this example our data are a 6834 × 64 matrix of gene expression measurements. Each row represents a gene, and each column a human tumour. The data are taken from Ross *et al.* (2000) and are available at `http://www-genome.stanford.edu/nci60`. The columns have a label (cancer type), but this label was not used in the clustering. We applied hierarchical (agglomerative) clustering to the columns, using squared error and average linkage, and obtained the dendrogram in Fig. 3. Not surprisingly, many cancers of the same type are clustered together. For more on the utility of hierarchical clustering for microarray data, see Ross *et al.* (2000).

The results for the gap statistic are shown in Fig. 4. The estimated number of clusters is 2. The corresponding cut of the dendrogram is indicated by the dotted line in Fig. 3. However,

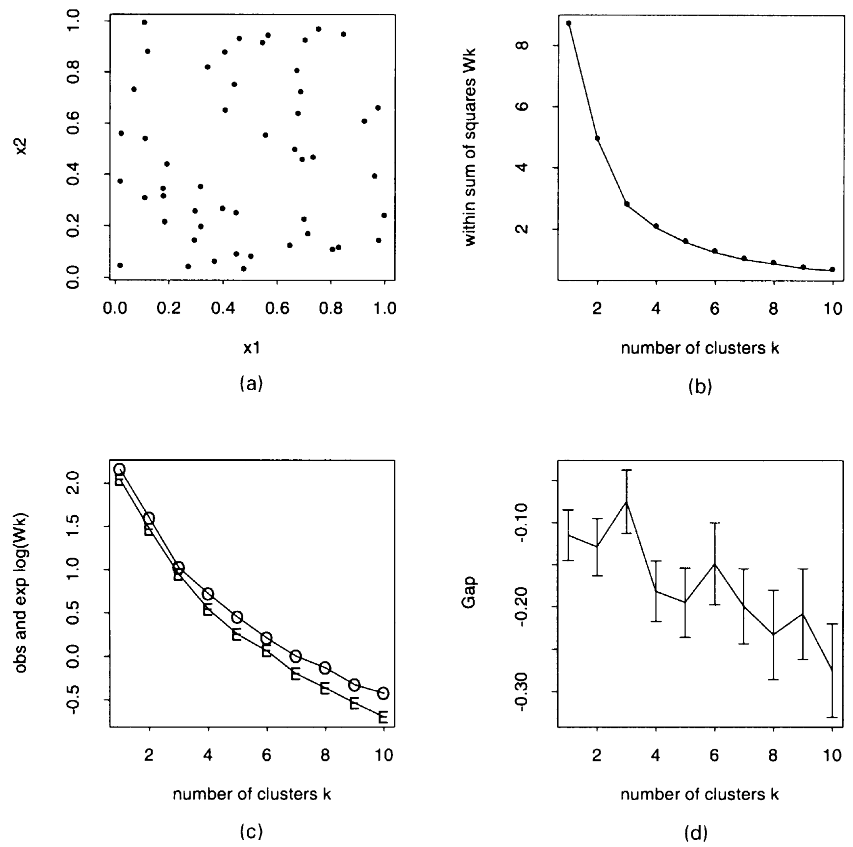

Fig. 2. Results for the uniform data example: (a) data; (b) within sum of squares function W_k; (c) functions $\log(W_k)$ (O) and $\hat{E}_n^*\{\log(W_k)\}$ (E); (d) gap curve

the gap function starts to rise again after six clusters, suggesting that there are two well-separated clusters and more less separated ones. The derivation for the gap test assumes that there are well-separated uniform clusters. In cases where there are smaller subclusters within larger well-separated clusters, it can exhibit non-monotone behaviour. Hence it is important to examine the entire gap curve rather than simply to find the position of its maximum.

5. Other approaches

Many methods have been proposed for estimating the number of clusters: a good summary is given by Gordon (1999). He divides the approaches into global and local methods. The former evaluate some measure over the entire data set and optimize it as a function of the number of clusters. The latter consider individual pairs of clusters and test whether they should be amalgamated. Hence the gap method is a global procedure.

According to Gordon, most global methods have the disadvantage that they are undefined for one cluster and hence offer no indication whether the data should be clustered at all. A very recent proposal is given by Cuevas *et al.* (2000); however, this relies on a high dimensional density estimate, which may suffer from the curse of dimensionality.

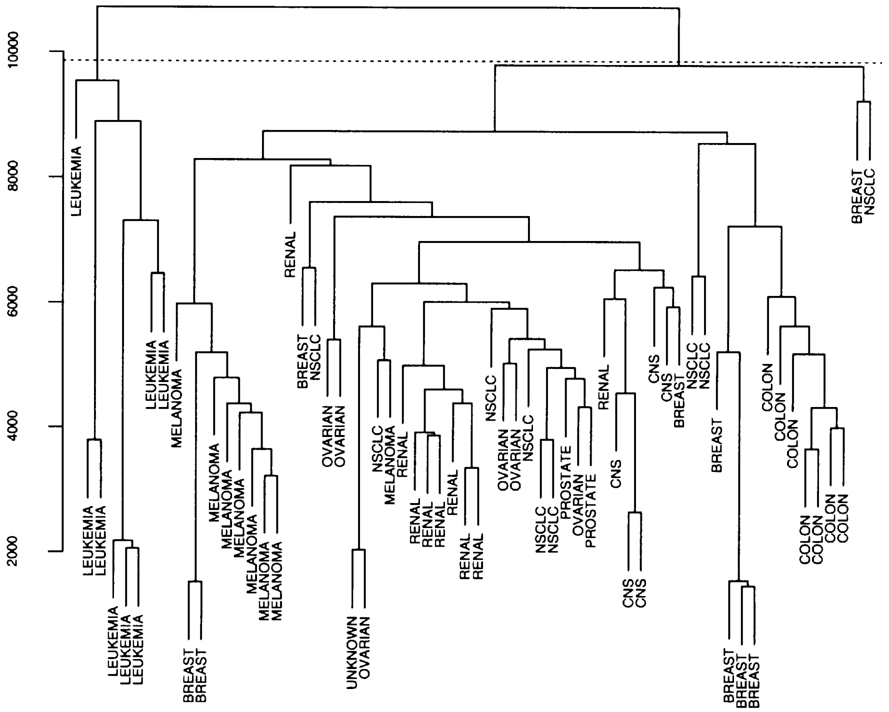

Fig. 3. Dendrogram from the deoxyribonucleic acid (DNA) microarray data: the dotted line cuts the tree, leaving two clusters as suggested by the gap statistic

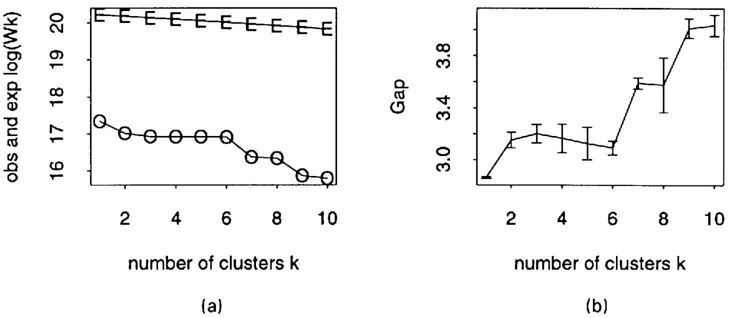

Fig. 4. (a) Logarithmic observed (O) and expected (E) within sum of squares curves and (b) the gap statistic for the DNA microarray data

Milligan and Cooper (1985) carried out a comprehensive simulation comparison of 30 different procedures. Among the global methods performing the best was the index due to Calinski and Harabasz (1974):

$$\mathrm{CH}(k)=\frac{B(k)/(k-1)}{W(k)/(n-k)} \tag{6}$$

where $B(k)$ and $W(k)$ are the between- and within-cluster sums of squares, with k clusters. The idea is to maximize $\mathrm{CH}(k)$ over the number of clusters k. CH(1) is not defined; even if it were modified by replacing $k-1$ with k, its value at 1 would be 0. Since $\mathrm{CH}(k) > 0$ for $k > 1$, the maximum would never occur at $k = 1$.

As mentioned earlier, Krzanowski and Lai (1985) proposed the quantity $W_k k^{2/p}$ as a criterion for choosing the number of clusters. This followed a proposal by Marriott (1971), who used the determinant, rather than the trace, of the within sum of squares matrix. The actual proposal of Krzanowski and Lai (1985) defined

$$\mathrm{DIFF}(k) = (k-1)^{2/p} W_{k-1} - k^{2/p} W_k \tag{7}$$

and chose k to maximize the quantity

$$\mathrm{KL}(k) = \left| \frac{\mathrm{DIFF}(k)}{\mathrm{DIFF}(k+1)} \right|. \tag{8}$$

This is similar to maximizing $W_k k^{2/p}$, but Krzanowski and Lai (1985) argued that it may have better properties. Note that $\mathrm{KL}(k)$ is not defined for $k = 1$ and hence cannot be used for testing one cluster *versus* more than one.

Hartigan (1975) proposed the statistic

$$H(k) = \left\{ \frac{W(k)}{W(k+1)} - 1 \right\} \Big/ (n-k-1). \tag{9}$$

The idea is to start with $k = 1$ and to add a cluster as long as $H(k)$ is sufficiently large. One can use an approximate F-distribution cut-off; instead Hartigan suggested that a cluster be added if $H(k) > 10$. Hence the estimated number of clusters is the smallest $k \geqslant 1$ such that $H(k) \leqslant 10$. This estimate is defined for $k = 1$ and can potentially discriminate between one *versus* more than one cluster.

Kaufman and Rousseeuw (1990) proposed the *silhouette* statistic, for assessing clusters and estimating the optimal number. For observation i, let $a(i)$ be the average distance to other points in its cluster, and $b(i)$ the average distance to points in the nearest cluster besides its own nearest is defined by the cluster minimizing this average distance. Then the silhouette statistic is defined by

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}}. \tag{10}$$

A point is well clustered if $s(i)$ is large. Kaufman and Rousseeuw (1990) proposed to choose the optimal number of clusters $\hat{k}$ as the value maximizing the average $s(i)$ over the data set. Note that $s(i)$ is not defined for the $k = 1$ cluster.

6. Simulations

We generated data sets in five different scenarios:

(a) *null (single-cluster) data in 10 dimensions*—200 data points uniformly distributed over the unit square in 10 dimensions;

(b) *three clusters in two dimensions*—the clusters are standard normal variables with (25, 25, 50) observations, centred at (0, 0), (0, 5) and (5, −3);

(c) *four clusters in three dimensions*—each cluster was randomly chosen to have 25 or 50 standard normal observations, with centres randomly chosen as $N(0, 5I)$ (any simulation with clusters having a minimum distance less than 1.0 units between them was discarded;

(d) *four clusters in 10 dimensions*—each cluster was randomly chosen to have 25 or 50 standard normal observations, with centres randomly chosen as $N(0, 1.9I)$ (any simulation with clusters having a minimum distance less than 1.0 units between them was discarded; in this and the previous scenario, the settings are such that about half of the random realizations were discarded);

(e) *two elongated clusters in three dimensions*—each cluster is generated as follows. Set $x_1 = x_2 = x_3 = t$ with t taking 100 equally spaced values from -0.5 to 0.5 and then Gaussian noise with standard deviation 0.1 is added to each feature. Cluster 2 is generated in the same way, except that the value 10 is added to each feature at the end. The result is two elongated clusters, stretching out along the main diagonal of a three-dimensional cube.

50 realizations were generated from each setting. In the non-null settings, the clusters have no overlap, so there is no confusion over the definition of the 'true' number of clusters. We applied six different methods for estimating the number of clusters: *CH*, *KL*, *Hartigan* and *Silhouette* are given by equations (6), (8), (9) and (10) respectively. *Gap/unif* is the gap method with a uniform reference distribution over the range of each observed feature; *Gap/pc* uses the uniform reference in the principal component orientation. The results are given in Table 1.

The gap estimate using the uniform reference does well except in the last problem, where the oblong shape of the data adversely affects it. The Gap/pc method, using a uniform reference in the principal components orientation, is the clear winner overall.

The other methods do quite well, except in the null setting where the gap estimate is the only one to show a reasonable performance. Of course it might be possible to modify any of the methods to handle the null (single-cluster) case: one possibility would be to simulate their null distribution under uniform data, in a manner similar to the gap estimate.

7. Overlapping classes

The simulation studies suggest that the gap estimate is good at identifying well-separated clusters. When data are not well separated, the notion of a cluster is not any more well defined in the literature.

In this section, we did a small experiment to assess how the gap method responds to non-separated data. Each simulated data set consists of 50 observations from each of two bivariate normal populations, with means (0, 0) and (Δ, 0), and identity covariance. For each sample we computed the gap estimate of the number of clusters and also recorded the proportion of data points from the first population that were closer to the second population mean, or vice versa. We call this the amount of 'overlap'. This was done for 10 values of Δ running from 0 to 5, with 10 simulations done for each value of Δ. The results are shown in Fig. 5. Roughly speaking, if the overlap proportion is p, then the probability of selecting one cluster is also about p.

8. Discussion

The problem of estimating the number of clusters in a data set is difficult, underlined by the fact that there is no clear definition of a 'cluster'. Hence, in data that are not clearly separated into groups, different people might have different opinions about the number of distinct clusters. In this paper, we have focused on well-separated clusters and have proposed the gap statistic for estimating the number of groups. When used with a uniform reference

Table 1. Results of the simulation study†

Method	*Estimates of the following numbers of clusters $\hat{k}$:*									
	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Null model in 10 dimensions										
CH	0‡	50	0	0	0	0	0	0	0	0
KL	0‡	29	5	3	3	2	2	0	0	0
Hartigan	0‡	0	1	20	21	6	0	0	0	0
Silhouette	0‡	49	1	0	0	0	0	0	0	0
Gap/unif	49‡	1	0	0	0	0	0	0	0	0
Gap/pc	50‡	0	0	0	0	0	0	0	0	0
3-cluster model										
CH	0	0	50‡	0	0	0	0	0	0	0
KL	0	0	39‡	0	5	1	1	2	0	0
Hartigan	0	0	1‡	8	19	13	3	3	2	1
Silhouette	0	0	50‡	0	0	0	0	0	0	0
Gap/unif	1	0	49‡	0	0	0	0	0	0	0
Gap/pc	2	0	48‡	0	0	0	0	0	0	0
Random 4-cluster model in 3 dimensions										
CH	0	0	0	42‡	8	0	0	0	0	0
KL	0	0	0	35‡	5	3	3	3	0	0
Hartigan	0	1	7	3‡	9	12	8	2	3	5
Silhouette	0	20	15	15‡	0	0	0	0	0	0
Gap/unif	0	1	2	47‡	0	0	0	0	0	0
Gap/pc	2	2	4	42‡	0	0	0	0	0	0
Random 4-cluster model in 10 dimensions										
CH	0	1	4	44‡	1	0	0	0	0	0
KL	0	0	0	45‡	3	1	1	0	0	0
Hartigan	0	0	2	48‡	0	0	0	0	0	0
Silhouette	0	13	20	16‡	5	0	0	0	0	0
Gap/unif	0	0	0	50‡	1	0	0	0	0	0
Gap/pc	0	0	4	46‡	0	0	0	0	0	0
2 elongated clusters										
CH	0	0‡	0	0	0	0	0	7	16	27
KL	0	50‡	0	0	0	0	0	0	0	0
Hartigan	0	0‡	0	1	0	2	1	5	6	35
Gap/unif	0	0‡	17	16	2	14	1	0	0	0
Gap/pc	0	50‡	0	0	0	0	0	0	0	0

†Numbers are counts out of 50 trials. Some rows do not add up to 50 because the number of clusters chosen was greater than 10.
‡Column corresponding to the correct number of clusters.

distribution in the principal component orientation, it outperforms other proposed methods from the literature in our simulations. The simpler uniform reference (over the range of the data) works well except when the data lie near a subspace.

The DNA microarray example shows the importance of graphing the gap statistic, rather than simply extracting the estimated maximum. With real data the gap curve can have many local maxima, and these themselves can be informative.

There are many avenues for further research. One is a consideration of other possibilities for the reference distribution: for example, we could proceed sequentially. Having found k clusters, we could generate reference data from k separate uniform distributions, over the support of each of the k estimated data clusters. As before, a principal component orientation would probably produce better results. The gap method can also be used with adaptive

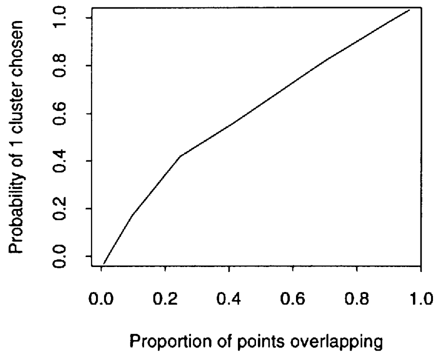

Fig. 5. Gap method for overlapping data: the proportion of times that the method chose one cluster, as a function of the proportion of points in the overlap region between the two subpopulations

versions of K-means clustering (see for example Diday and Govaert (1977)), which may be better at finding elongated clusters than the standard version. Similarly, it may be applicable to model-based clustering (Fraley and Raftery, 1998).

A referee raised the interesting question of how to carry out the gap test when the dimension p of the data is unknown and only pairwise dissimilarities are available. One possibility would be to use multidimensional scaling to map the data into a low dimensional space while preserving the dissimilarities, and then to proceed in this space as described in the paper. However, a more direct method would be preferable and we leave this as an open problem.

It would be especially useful to develop methods for an efficient simulation of reference data from the log-concave MLE. The use of this distribution in the gap method could then be compared with the uniform reference distribution.

Acknowledgements

Tibshirani was partially supported by National Institutes of Health grant 2 R01 CA72028 and National Science Foundation grant DMS-9971405. Hastie was partially supported by grant DMS-9803645 from the National Science Foundation and grant ROI-CA-72028-01 from the National Institutes of Health.

Appendix A: Proofs

A.1. Proof of theorem 1

Setting $\mu_j := (j-\frac{1}{2})/k$ for $1 \leqslant j \leqslant k$ shows that $\mathrm{MSE}_U(k) \leqslant E\{\min_{\mu_j}(U-\mu_j)^2\} = 1/12k^2$, whence $\mathrm{MSE}_U(k)/\mathrm{MSE}_U(1) \leqslant 1/k^2$. Thus it is enough to prove

$$\sum_{i=1}^{k} P(X \in I_i)\,\mathrm{var}_{I_i}(X) \geqslant \frac{1}{k^2}\,\mathrm{var}(X) \tag{11}$$

for every partition $I_1, \ldots, I_k$ of the support of X. Here we write

$$\mathrm{var}_I(X) = \frac{\displaystyle\int_I \left\{x - \int_I x\,\mathrm{d}P_X/P(X \in I)\right\}^2 \mathrm{d}P_X}{P(X \in I)}$$

for the conditional variance of X given $X \in I$.

By standard arguments (e.g. convolution with a Gaussian kernel and using Ibragimov's convolution result; see theorem 1.10 in Dharmadhikari and Joag-dev (1988)), it is enough to consider a non-degenerate cumulative density function F of X that has a density f which is logarithmically concave and differentiable in the interior of its support and so does not vanish there. Hence

$$\frac{\mathrm{d}}{\mathrm{d}t} f\{F^{-1}(t)\} = \frac{f'\{F^{-1}(t)\}}{f\{F^{-1}(t)\}} = \frac{\mathrm{d}}{\mathrm{d}x} \log\{f(x)\}|_{x=F^{-1}(t)}.$$

But $\mathrm{d}[\log\{f(x)\}]/\mathrm{d}x$ is non-increasing as f is logarithmically concave. Together with the fact that $F^{-1}(t)$ is non-decreasing, it follows that $f\{F^{-1}(\cdot)\}$ has a non-increasing derivative and hence is concave on [0, 1].

Next, write

$$\begin{aligned}\mathrm{var}(X) &= \frac{1}{2} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y-x)^2 f(x) f(y) \,\mathrm{d}x \,\mathrm{d}y \\ &= \frac{1}{2} \int_0^1 \int_0^1 \{F^{-1}(v) - F^{-1}(u)\}^2 \,\mathrm{d}u \,\mathrm{d}v \\ &= \int_0^1 \int_u^1 \left\{ \int_u^v \frac{1}{f\{F^{-1}(t)\}} \,\mathrm{d}t \right\}^2 \mathrm{d}u \,\mathrm{d}v\end{aligned}$$

by symmetry and the fundamental theorem of calculus. The change of variable $z = v - u$ gives

$$\mathrm{var}(X) = \int_{z=0}^1 \int_{u=0}^{1-z} \left[\int_u^{u+z} \frac{1}{f\{F^{-1}(t)\}} \,\mathrm{d}t \right]^2 \mathrm{d}u \,\mathrm{d}z. \tag{12}$$

Proceeding likewise with $\mathrm{var}_{I_i}(X)$ we obtain

$$\sum_{i=1}^k F(I_i) \,\mathrm{var}_{I_i}(X) = \sum_{i=1}^k \int_{z=0}^1 z^2 \int_{u=0}^{1-z} F^3(I_i) \left[\frac{1}{F(I_i) z} \int\limits_{s_{i-1}+F(I_i)u}^{s_{i-1}+F(I_i)(u+z)} \frac{1}{f\{F^{-1}(t)\}} \,\mathrm{d}t \right]^2 \mathrm{d}u \,\mathrm{d}z,$$

where we set $s_i := \Sigma_{j \leqslant i} F(I_j)$, $i = 0, \ldots, k$.

Using the concavity of $f\{F^{-1}(\cdot)\}$ and Holder's inequality it can be shown that the above expression is not smaller than

$$\begin{aligned}&\sum_{i=1}^k \int_{z=0}^1 \frac{z^2}{k^2} \int_{u=0}^{1-z} \left[\frac{1}{z} \int\limits_{s_{i-1}(1-z)+F(I_i)u}^{s_{i-1}(1-z)+F(I_i)u+z} \frac{1}{f\{F^{-1}(t)\}} \,\mathrm{d}t \right]^2 F(I_i) \,\mathrm{d}u \,\mathrm{d}z \\ &\qquad = \frac{1}{k^2} \int_{z=0}^1 \sum_{i=1}^k \int\limits_{v=s_{i-1}(1-z)}^{s_{i-1}(1-z)+F(I_i)(1-z)} \left[\int_{t=v}^{v+z} \frac{1}{f\{F^{-1}(t)\}} \,\mathrm{d}t \right]^2 \mathrm{d}v \,\mathrm{d}z \\ &\qquad = \frac{1}{k^2} \mathrm{var}(X) \qquad\qquad \text{by equation (12),}\end{aligned}$$

proving inequality (11).

A.2. Proof of theorem 2

If X is uniformly distributed on $U([0, k] \times [0, \epsilon]^{p-1})$, then $\mathrm{MSE}_X(1) = \{k^2 + (p-1)\epsilon^2\}/12$, and taking $\mu_j = (j - 1/2, \epsilon/2, \ldots, \epsilon/2)$, $1 \leqslant j \leqslant k$, shows that $\mathrm{MSE}_X(k) \leqslant E(\min_{\mu_j} \|X - \mu_j\|^2) = \{1 + (p-1)\epsilon^2\}/12$. So

$$\inf_{X \in \mathcal{S}^p} \{\mathrm{MSE}_X(k)/\mathrm{MSE}_X(1)\} \leqslant 1/k^2,$$

even if we were to consider only $X \in \mathcal{S}^p$ with non-degenerate support.

However, suppose that $U \in \mathcal{S}^p$ satisfies $\mathrm{MSE}_U(k)/\mathrm{MSE}_U(1) = 1/k^2$. Each of the marginals U_i of U, $1 \leqslant i \leqslant p$, must be in $\mathcal{S}^1$ by theorem 2.16 in Dharmadhikari and Joag-dev (1988). Hence

$$\mathrm{MSE}_{U_i}(1) \leqslant k^2\, \mathrm{MSE}_{U_i}(k) \qquad \text{for all } i \text{ by theorem 1,} \tag{13}$$

and clearly

$$\sum_{i=1}^{p} \mathrm{MSE}_{U_i}(k) \leqslant \mathrm{MSE}_U(k) \qquad \text{for all } k > 1. \tag{14}$$

So

$$\mathrm{MSE}_U(1) = \sum_{i=1}^{p} \mathrm{MSE}_{U_i}(1) \leqslant k^2 \sum_{i=1}^{p} \mathrm{MSE}_{U_i}(k) \leqslant \mathrm{MSE}_U(k),$$

and hence $\mathrm{MSE}_U(k)/\mathrm{MSE}_U(1) = 1/k^2$ can only hold if we have equality in expressions (13) and (14).

To avoid technicalities we shall only give the main arguments for the remainder of the proof. Proceeding similarly as in the proof of theorem 1 we conclude from equality in expression (13) that the U_i must have a uniform distribution, with the optimal centres $\gamma_i(j)$, $1 \leqslant j \leqslant k$, equally spaced. Let l_i be the length of the support of U_i. We then check that expression (14) can hold with equality only if with probability 1 the centre $\gamma_i(j)$ closest to U_i has the same index j for all marginals i. But the set of $u \in \mathbf{R}^p$ for which the latter statement holds has Lebesgue measure $k \prod_{i=1}^{p} l_i/k \rightarrow 0$ as $k \rightarrow \infty$. Hence, by Prekopa's theorem (theorem 2.8 in Dharmadhikari and Joag-Dev (1988)), the support of U must be degenerate and contained in a linear subspace of $\mathbf{R}^p$. Repeating this argument at most $p-1$ times proves the theorem.

References

Breiman, L., Friedman, J., Olshen, R. and Stone, C. (1984) *Classification and Regression Trees*. Belmont: Wadsworth.

Calinski, R. B. and Harabasz, J. (1974) A dendrite method for cluster analysis. *Communs Statist.*, **3**, 1–27.

Cuevas, A., Febrero, M. and Fraiman, R. (2000) Estimating the number of clusters. *Can. J. Statist.*, **28**, 367–382.

Dharmadhikari, S. and Joag-dev, K. (1988) *Unimodality, Convexity, and Applications*. New York: Academic Press.

Diday, E. and Govaert, G. (1977) Classification automatique avec distances adaptives. *RAIRO Informatique/Computer Sciences*, pp. 329–349.

Fraley, C. and Raftery, A. (1998) How many clusters?; which clustering method?—answers via model-based cluster analysis. *Comput. J.*, **41**, 578–588.

Gordon, A. (1996) Null models in cluster validation. In *From Data to Knowledge* (eds W. Gaul and D. Pfeifer), pp. 32–44. New York: Springer.

——(1999) *Classification*, 2nd edn. London: Chapman and Hall–CRC.

Hartigan, J. (1975) *Clustering Algorithms*. New York: Wiley.

Kaufman, L. and Rousseeuw, P. (1990) *Finding Groups in Data: an Introduction to Cluster Analysis*. New York: Wiley.

Krzanowski, W. J. and Lai, Y. T. (1985) A criterion for determining the number of groups in a data set using sum of squares clustering. *Biometrics*, **44**, 23–34.

Marriott, F. H. C. (1971) Practical problems in a method of cluster analysis. *Biometrics*, **27**, 501–514.

Milligan, G. W. and Cooper, M. C. (1985) An examination of procedures for determining the number of clusters in a data set. *Psychometrika*, **50**, 159–179.

Roeder, K. (1994) A graphical technique for determining the number of components in a mixture of normals. *J. Am. Statist. Ass.*, **89**, 487–495.

Ross, D., Scherf, U., Eisen, M., Perou, C., Spellman, P., Iyerl, V., Rees, C., Jeffery, S., Van de Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J., Lashkari, D., Shalon, D., Myers, T., Weinstein, J., Botstein, D. and Brown, P. (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genet.*, **24**, 227–234.

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

Sugar, C. (1998) Techniques for clustering and classification with applications to medical problems. *PhD Dissertation*. Stanford University, Stanford.

Sugar, C., Lenert, L. and Olshen, R. (1999) An application of cluster analysis to health services research: empirically defined health states for depression from the sf-12. *Technical Report*. Stanford University, Stanford.

Walther, G. (2000) Detecting the presence of mixing with multiscale maximum likelihood. *Technical Report*. Stanford University, Stanford.

——(2001) On the nonparametric analysis of a mixture distribution. *Technical Report*. Stanford University, Stanford.