Estimating the number of clusters in a dataset via the Gap statistic

Robert Tibshirani, [*]Guenther Walther [†]and Trevor Hastie [‡]

March 29, 2000

Abstract

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 to that expected under an appropriate reference null distribution. Some theory is developed for the proposal and a simulation study that shows that the Gap statistic usually outperforms other methods that have been proposed in the literature. We also briefly explore application of the same technique to the problem for estimating the number of linear principal components.

1 Introduction

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 estimation of the optimal number of "clusters". Figure 1 (top right) 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

[*]Division of Biostatistics and Department of Statistics, Stanford, University, Stanford CA 94305; tibs@stat.stanford.edu

[†]Department of Statistics, Stanford, University, Stanford CA 94305; walther@stat.stanford.edu

[‡]Department of Statistics Division of Biostatistics, Stanford, University, Stanford CA 94305; trevor@stat.stanford.edu

the number of clusters k increases, but from some k on 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), Sugar et al. (1999). A comprehensive survey of methods for estimating the number of clusters is given in Milligan & Cooper (1985), while 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 $\{x_{ij}\}, i = 1, 2, \ldots n; , j = 1, 2, \ldots p$ consists 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 $\sum_j (x_{ij} - x_{i'j})^2$.

Suppose 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

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

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

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

So if the distance d is the squared Euclidean distance, then W_k is 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.

The idea of our approach is to standardize the graph of $\log(W_k)$ by comparing it to its expectation under an appropriate null reference distribution of the data. 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

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

where E_n^* denotes expectation under a sample of size n from the reference distribution (this expectation generally depends on n due to the data-dependent optimization implicit in the computation of W_k). 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 centers. Then assuming that the centers align themselves in an equally spaced fashion, the expectation of $\log(W_k)$ is approximately

$$\log(pn/12) - (2/p)\log(k) + \textit{constant} \tag{4}$$

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 $k \leq K$. When $k > K$, we are essentially adding an (unnecessary) cluster center 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 & Simons (1971). To develop the Gap statistic into an operational procedure, we need to find an appropriate reference distribution and control the sampling distribution of the Gap statistic.

3 The reference distribution

We will work in the framework that we adopt a null model of a single component, and reject it in favor of a k component model ($k > 1$), if the strongest evidence for any such k warrants it. That is, we wish to simultaneously screen the evidence over all $k > 1$. This approach of guarding against erroneous rejection of the one-component model is similar to Roeder (1994). A component (cluster) of the distribution can be appropriately modeled 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 $\psi(x) = -\frac{1}{2} \parallel x \parallel^2$) and the uniform distribution with convex support. Reasons for modeling the components as log-concave densities instead of the often used unimodal densities are given in Walther (2000). 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. 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:

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

where $\mathrm{MSE}_X(k) = \mathrm{E}\min_{\mu \in A_k} \parallel X - \mu \parallel^2$, with the k-point set $A_k \subset \mathcal{R}^p$ 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 favorable single-component reference distribution on X^* such that $g(k) \leq 0$ for all $X \in \mathcal{S}^p$ and all $k \geq 1$. The next 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 $k \geq 1$*

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

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

Note that 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 & 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 (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 in $\mathcal{S}^p$. This MLE can be shown to exist, as opposed to the MLE of a unimodal distribution. However, at this point in time, no simple algorithms for estimating and simulating from this MLE seem to exist. On the other hand, 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 the columns have mean zero and compute the singular value decomposition $X = UDV^T$. 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 backtransform via $Z = Z'V^T$ 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 along 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 control 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 $s_k = \sqrt{1 + 1/B}\,\mathrm{sd}(k)$. Using this we choose the cluster size $\hat{k}$ to be the smallest k such that $\mathrm{Gap}(k) \geq \mathrm{Gap}(k+1) - s_{k+1}$. This “one standard error ” -style rule is used elsewhere (e.g. Breiman et al. (1984)), and we have found empirically that it works well in the present problem. 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

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

2. Generate B reference datasets, 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:

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

3. Let $\bar{l} = (1/B) \sum_b \log(W_{kb}^*)$, compute the standard deviation $\mathrm{sd}_k = [(1/B) \sum_b (\log(W_{kb}^*) - \bar{l})^2]^{1/2}$, and define $s_k = \mathrm{sd}_k \sqrt{1 + 1/B}$. Finally choose the number of clusters via

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

Figure 1 shows an example using K-means clustering. The data (top left panel) fall in two distinct clusters. The within sum of squares function W_k is displayed in the top right. The functions $\log(W_k)$ and $\hat{E}_n^*(\log(W_k))$ are shown in the bottom left panel, with the gap curve displayed in the bottom right, with ±1 standard error bars.

Figure 2 examines the behavior 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 $\hat{k} = 1$.

Example: Application to hierarchical clustering and DNA microarray data

In this example our data is a 1000×64 matrix of gene expression measurements. Each row represents a gene, and each column a human tumor. The data is taken from Ross et al. (1999). 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 dendogram in Figure 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. (1999).

The results for the Gap statistic are shown in Figure 4. The estimated number of clusters is 2. The corresponding cut of the dendogram is indicated by the horizontal line in Figure 3. However the gap function starts to rise again after 6 clusters, suggesting that there are 2 well separated clusters and more less separated ones.

5 Other approaches

There have been many methods 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 dataset 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 of whether the data should be clustered at all.

Milligan & Cooper (1985) carry out a comprehensive simulation comparison of 30 different procedures. Among the global methods performing the

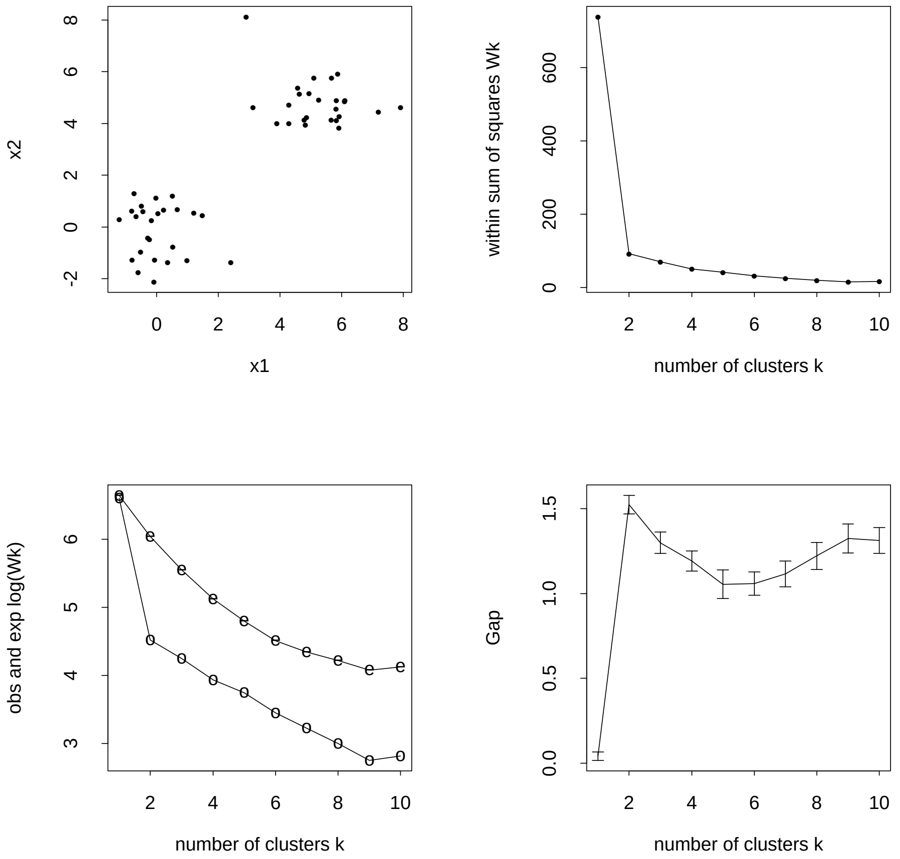

Figure 1: *Results for two cluster example. The data is in top left, and the within sum of squares function W_k is displayed in the top right. The functions* $\log(W_k)$ *and* $\hat{E}_n^*(\log(W_k))$ *are shown in the bottom left panel (plotting symbol "o" and "e" respectively), with the gap curve displayed in the bottom right.*

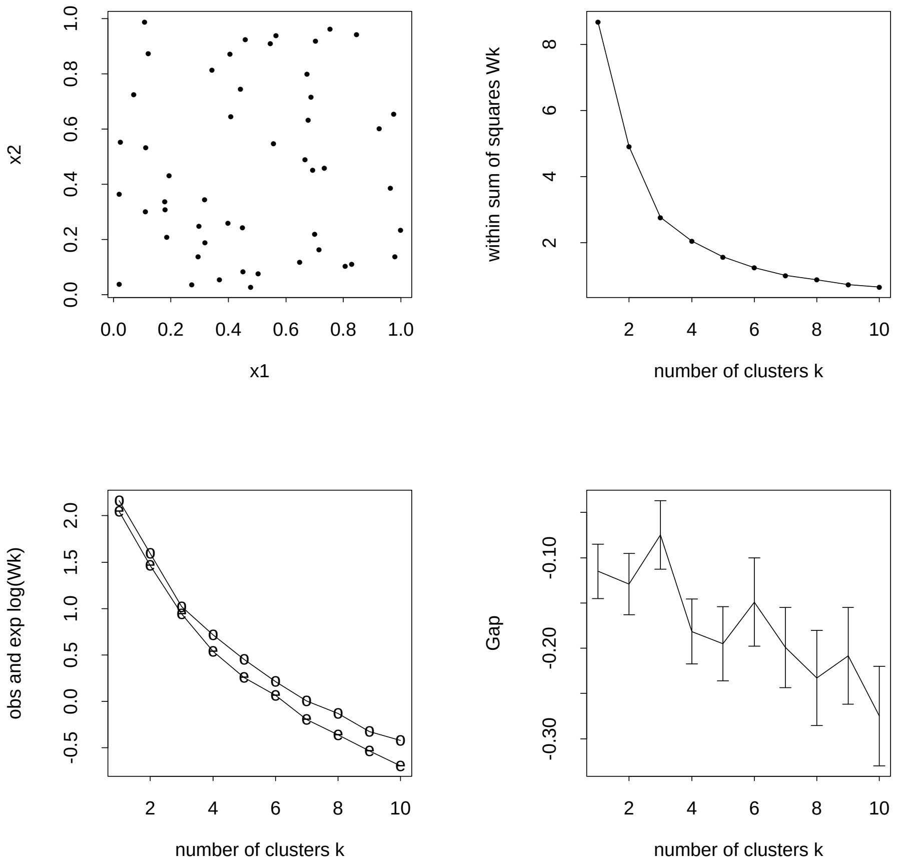

Figure 2: *Results for uniform data example. The data is in top left, and the within sum of squares function W_k is displayed in the top right. The functions* $\log(W_k)$ *and* $\hat{E}_n^*(\log(W_k))$ *are shown in the bottom left panel (plotting symbol "o" and "e" respectively), with the gap curve displayed in the bottom right.*

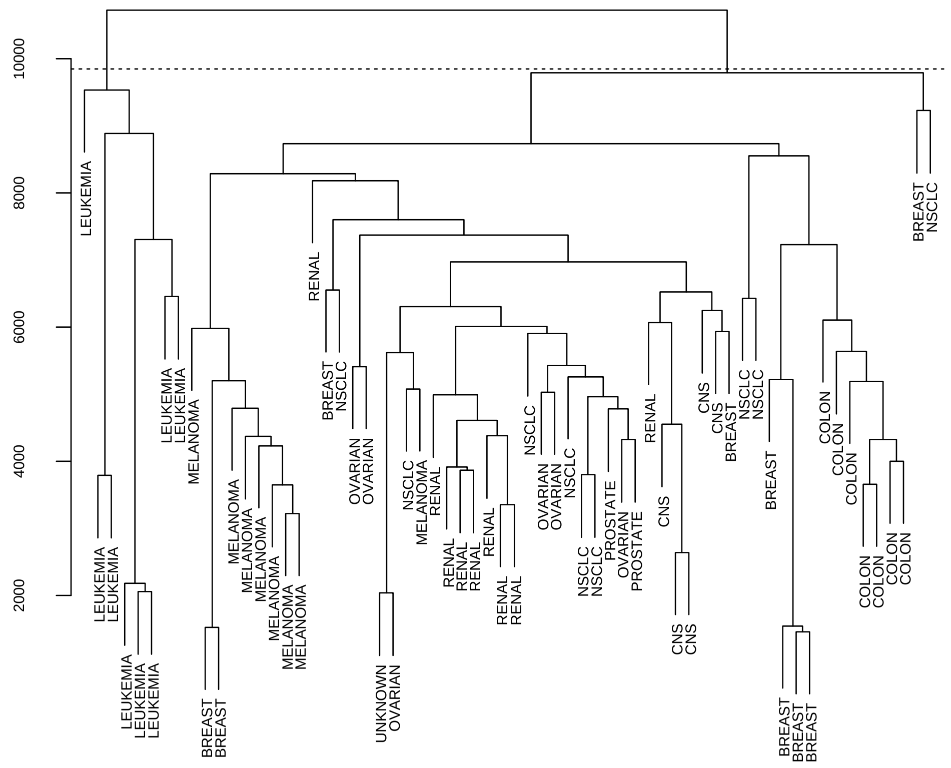

Figure 3: *Dendogram from DNA microarray data. The horizontal line cuts the tree, leaving two clusters as suggested by the Gap statistic*

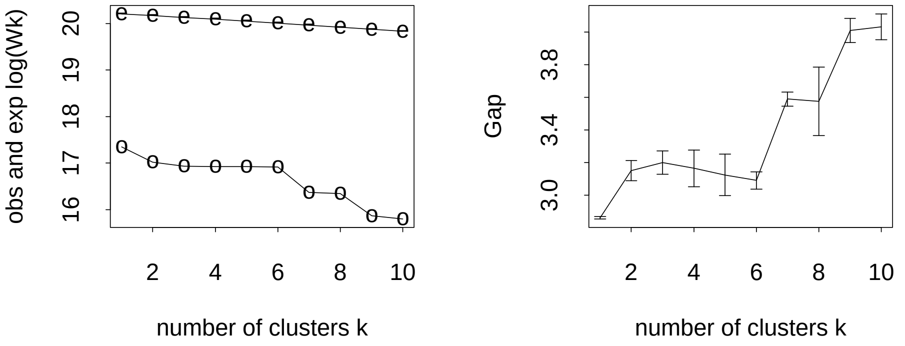

Figure 4: *(Log) observed and expected within sum of squares curves (left) and gap statistic (right) for DNA microarray data*

best was the index due to Calinski & Harabasz (1974):

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

where $B(k)$ and $W(k)$ are the between and within cluster sums of squares, with k clusters. The idea is to maximize $CH(k)$ 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 zero. Since $CH(k) > 0$ for $k > 1$, the maximum would never occur at $k = 1$.

As mentioned earlier, Krzanowski & 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 & 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 the authors argue 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

$$H(k) = [\frac{W(k)}{W(k+1)} - 1](n - k - 1) \tag{9}$$

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

Kaufman & 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. The authors propose to choose the optimal number of clusters $\hat{k}$ as the value maximizing the average $s(i)$ over the dataset. Note that $s(i)$ is not defined for $k = 1$ cluster.

6 Simulations

We generated datasets in five different scenarios:

1. *Null (single cluster) data in ten dimensions:* 200 data points uniformly distributed over the unit square in ten dimensions

2. *3 clusters in 2 dimensions*: the clusters are standard normal variables with (25, 25, 50) observations, centered at (0,0), (0,5), and (5,-3)

3. *4 clusters in 3 dimensions*: each cluster was randomly chosen to have 25 or 50 observations, with centers randomly chosen as $N(0, 5 \cdot I)$. Any simulation with clusters less than 1.0 units apart was discarded.

4. *4 clusters in 10 dimensions*: each cluster was randomly chosen to have 25 or 50 observations, with centers randomly chosen as $N(0, 1.9 \cdot I)$. Any simulation with clusters less than 1.0 units apart was discarded. In this and the previous scenario, the settings are such that about one-half of the random realizations were discarded.

5. *Two elongated clusters in 3 dimensions.* Each cluster is generated as follows: set $x_1 = x_2 = x_3 = t$ with t taking on 100 equally spaced valued from -0.5 to 0.5 and then Gaussian noise with standard deviation .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.

Fifty realizations were generated from each setting. In the non-null settings, the clusters have no overlap, so that 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 (6), (9) and (10). *Gap/unif* is the gap method with 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 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.

Table 1: *Results of simulation study. Numbers are counts out of 50 trials. "*" indicates column corresponding to correct number of clusters.*

	Estimate of number of clusters $\hat{k}$									
Method	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
Three 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 dims.										
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 dims.										
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
Two 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

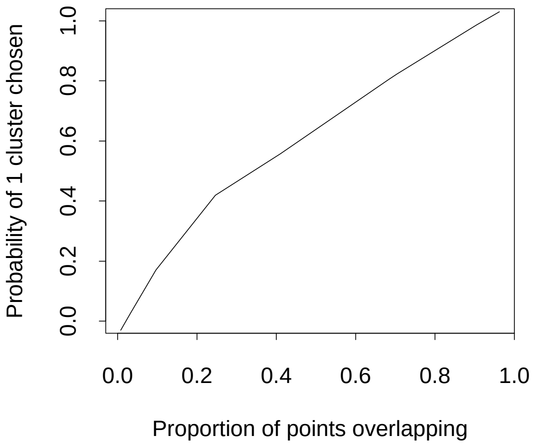

Figure 5: *Gap method for overlapping data: proportion of times that the method chose one cluster, as a function of the proportion of points in the overlap region between the two subpopulations.*

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 $(\Delta, 0)$, and identity covariance. For each sample we computed the Gap estimate of the number of clusters, and also recorded the proportion of data points in the overlap region between the two centroids. This done for 10 values of Δ running from 0 up to 5. The results are shown in Figure 5. Roughly speaking, if the overlap proportion is p, then the probability of selecting one cluster is also about p.

8 Estimating the number of principal components

Consider the different (but related) problem of estimating the number of linear principal components needed for approximating a data cloud. As before we have data x_{ij} on $j = 1, 2, \ldots p$ features for each of $i = 1, 2, \ldots n$ observa-

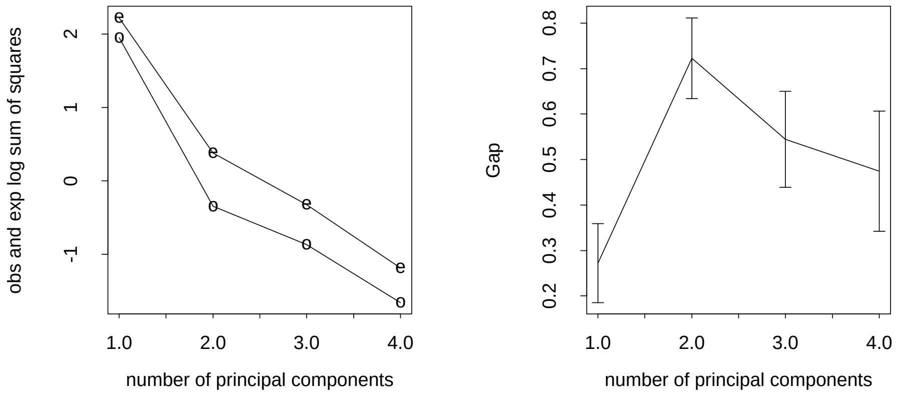

Figure 6: *Gap method applied to the problem of choosing the optimal order for linear principal components. Left panel shows the observed ("o") and expected ("e") value of the (log) reconstruction error* $\log(W_k')$. *Right panel shows the Gap curve with its maximum at the correct value of two principal components.*

tions. Let W_k' be the reconstruction error in using k principal components to approximate the data. In detail, let $\{\hat{x}_{ij}^k\}, j = 1, 2, \ldots p$ be the projection of the *ith* data point onto the rank k principal component approximation. Then

$$W_k' = \sum_i \sum_j (x_{ij} - \hat{x}_{ij}^k)^2 \tag{11}$$

In order to estimate the best value of k, we apply the Gap method using W_k' in place of W_k.

As an example, we simulated 100 data points according to the model

$$\begin{aligned} X_1 &= U_1 \\ X_2 &= U_2 \\ X_3 &\sim .05 \cdot Z_3;\ X_4 \sim .05 \cdot Z_4\ ; X_5 \sim .05 \cdot Z_5 \end{aligned} \tag{12}$$

with $U_j \sim U(0,1)$, $Z_j \sim N(0,1)$. Hence the data lie close to a two-dimensional manifold in five dimensional space. The results of applying the Gap test are shown in Figure 6. The Gap curve has a clear maximum at two principal components, as it should. Of course this example is just an illustration, and this problem requires further study.

9 Discussion

The problem of estimating the number of clusters in a dataset is a difficult one, underlined by the fact that there is no clear definition of a "cluster". Hence in data that is not clearly separated into groups, different people might have different opinions about the number of distinct clusters. In this paper, we have focussed on well-separated clusters, and proposed the Gap statistic for estimating the number of groups. When used with a uniform reference 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 are concentrated on 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 the Gap curve can have many local maxima, and these themselves can be informative.

There are many avenues for further research. One is consideration of other possibilities for the reference distribution: for example, one 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 likely produce better results.

It would be especially useful to develop methods for efficient simulation of reference data from the log-concave maximum likelihood estimate. The use of this distribution in the Gap method could then be compared to the uniform reference distribution.

10 Proofs

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

$$\sum_{i=1}^{k} P(X \in I_i)\ \mathrm{Var}_{I_i} X \geq \frac{1}{k^2}\mathrm{Var}\, X \tag{13}$$

for every partition $I_1, \ldots, I_k$ of the support of X. Here we write $\mathrm{Var}_I X = \int_I (x - \int_I x dP_X / P(X \in I))^2 dP_X / P(X \in I)$ 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 Thm. 1.10 in Dharmadhikari & Joag-dev (1988), it is enough to consider a nondegenerate cdf 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 $\frac{d}{dt} f(F^{-1}(t)) = f'(F^{-1}(t))/f(F^{-1}(t)) = \frac{d}{dx} \log f(x)|_{x=F^{-1}(t)}$. But $\frac{d}{dx} \log f(x)$ is nonincreasing as f is logarithmically concave. Together with the fact that $F^{-1}(t)$ is nondecreasing, it follows that $f(F^{-1}(\cdot))$ has a nonincreasing derivative, and hence is concave on $[0,1]$.

Next, write

$$\begin{aligned} \operatorname{Var} X &= 1/2 \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y-x)^2 f(x) f(y) dx dy \\ &= 1/2 \int_0^1 \int_0^1 (F^{-1}(v) - F^{-1}(u)))^2 du dv \\ &= \int_0^1 \int_u^1 \Big(\int_u^v \frac{1}{f(F^{-1}(t))} dt \Big)^2 du dv \end{aligned}$$

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

$$\operatorname{Var} X = \int_{z=0}^1 \int_{u=0}^{1-z} \Big(\int_u^{u+z} \frac{1}{f(F^{-1}(t))} dt \Big)^2 du dz. \tag{14}$$

Proceeding likewise with $\operatorname{Var}_{I_i} X$ one obtains

$$\sum_{i=1}^k F(I_i) \operatorname{Var}_{I_i} X = \sum_{i=1}^k \int_{z=0}^1 z^2 \int_{u=0}^{1-z} F^3(I_i) \Big(\frac{1}{F(I_i) z} \int_{s_{i-1}+F(I_i)u}^{s_{i-1}+F(I_i)(u+z)} \frac{1}{f(F^{-1}(t))} dt \Big)^2 du dz,$$

where we set $s_i := \sum_{j \le i} F(I_j)$, $i = 0, \dots, k$.

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

$$\begin{aligned} &\sum_{i=1}^k \int_{z=0}^1 z^2/k^2 \int_{u=0}^{1-z} \Big(\frac{1}{z} \int_{s_{i-1}(1-z)+F(I_i)u}^{s_{i-1}(1-z)+F(I_i)u+z} \frac{1}{f(F^{-1}(t))} dt \Big)^2 F(I_i) du dz \\ &= \frac{1}{k^2} \int_{z=0}^1 \sum_{i=1}^k \int_{v=s_{i-1}(1-z)}^{s_{i-1}(1-z)+F(I_i)(1-z)} \Big(\int_{t=v}^{v+z} \frac{1}{f(F^{-1}(t))} dt \Big)^2 dv dz \\ &= \frac{1}{k^2} \operatorname{Var} X \quad \text{by (14),} \end{aligned}$$

proving (13). □

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 $\mu_j = (j - 1/2, \epsilon/2, \ldots, \epsilon/2)$, $1 \leq j \leq k$, shows $\mathrm{MSE}_X(k) \leq E \min_{\mu_j} \parallel X - \mu_j \parallel^2 = (1 + (p-1)\epsilon^2)/12$. So $\inf_{X \in \mathcal{S}^p} \mathrm{MSE}_X(k)/\mathrm{MSE}_X(1) \leq 1/k^2$, even if we were to consider only $X \in \mathcal{S}^p$ with nondegenerate support.

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

$$\mathrm{MSE}_{U_i}(1) \leq k^2 \mathrm{MSE}_{U_i}(k) \quad \text{for all } i \text{ by Theorem 1,} \tag{15}$$

and clearly

$$\sum_{i=1}^{p} \mathrm{MSE}_{U_i}(k) \leq \mathrm{MSE}_U(k) \quad \text{for all } k > 1. \tag{16}$$

So $\mathrm{MSE}_U(1) = \sum_{i=1}^{p} \mathrm{MSE}_{U_i}(1) \leq k^2 \sum_{i=1}^{p} \mathrm{MSE}_{U_i}(k) \leq \mathrm{MSE}_U(k)$, and hence $\mathrm{MSE}_U(k)/\mathrm{MSE}_U(1) = 1/k^2$ can only hold if we have equality in (15) and (16).

To avoid technicalities we will only give the main arguments for the remainder of the proof. Proceeding similarly as in the proof of Theorem 1 one concludes from equality in (15) that the U_i must have a uniform distribution, with the optimal centers $\gamma_i(j)$, $1 \leq j \leq k$, equally spaced. Let l_i be the length of the support of U_i. One then checks that (16) can hold with equality only if with probability one the center $\gamma_i(j)$ 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 $k \prod_{i=1}^{p} l_i/k \to 0$ as $k \to \infty$. Hence by Prekopa's theorem (theorem 2.8 in [Dharmadhikari and Joag-Dev]), 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. □

Acknowledgements: Tibshirani and Hastie were supported by grants from NSF and NIH.

References

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

Calinski, R. B. & Harabasz, J. (1974), 'A dendrite method for cluster analysis', *Communications in statistics* **3**, 1–27.

Dharmadhikari, S. & Joag-dev, K. (1988), *Unimodality, convexity, and applications*, Academic Press.

Gordon, A. (1999), *Classification (2nd edition)*, Chapman and Hall/CRC press, London.

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

Kaufman, L. & Rousseeuw, P. (1990), *Finding groups in data: an introduction to cluster analysis*, New York; Wiley.

Krzanowski, W. J. & 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. & 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', *Journal of the American Statistical Association* **89**, 487–495.

Ross, D., Scherf, U., Eisen, M., Perou, C., Spellman, P., Iyer1, V., Rees, C., Jeffery, S., Van de Rijn, M., Waltham, M., Alexander, P., Lee, J., Lashkari, D., Shalon, D., Myers, T., Weinstein, J., Botstein, B. & Brown, P. (1999), Systematic variation in gene expression patterns in human cancer cell lines, Technical report, Stanford University.

Scott, A. & Simons, 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, Technical report, Stanford University. Ph.D. dissertation in Statistics, R. Olshen supervisor.

Sugar, C., Lenert, L. & 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.

Walther, G. (2000), On the nonparametric analysis of a mixture distribution, Technical report. In preparation.