Finding the number of clusters in a data set : An information theoretic approach

CATHERINE A. SUGAR AND GARETH M. JAMES
Marshall School of Business,
University of Southern California

Abstract

One of the most difficult problems in cluster analysis is the identification of the number of groups in a data set. Most previously suggested approaches to this problem are either somewhat ad hoc or require parametric assumptions and complicated calculations. In this paper we develop a simple yet powerful non-parametric method for choosing the number of clusters based on *distortion*, a quantity that measures the average distance, per dimension, between each observation and its closest cluster center. Our technique is computationally efficient and straightforward to implement. We demonstrate empirically its effectiveness, not only for choosing the number of clusters but also for identifying underlying structure, on a wide range of simulated and real world data sets. In addition, we give a rigorous theoretical justification for the method based on information theoretic ideas. Specifically, results from the subfield of electrical engineering known as *rate distortion theory* allow us to describe the behavior of the distortion in both the presence and absence of clustering. Finally, we note that these ideas potentially can be extended to a wide range of other statistical model selection problems.

1 Introduction

A fundamental, and largely unsolved, problem in cluster analysis is the determination of the "true" number of groups in a data set. Numerous approaches to this problem have been suggested over the years. Milligan and Cooper (1985) and Hardy (1996) provide a detailed set of references. Examples in the statistics literature include Calinski and Harabasz's index (Calinski and Harabasz, 1974), Hartigan's rule (Hartigan, 1975), the Kranowski and Lai test (Krzanowski and Lai, 1985) and the silhouette statistic (Kaufman and Rousseeuw, 1990). Two newer proposals are a Gaussian model-based approach using approximate Bayes factors (Kass and Raftery, 1995; Frayley and Raftery, 1998) and the gap statistic which compares the change in within-cluster dispersion with that expected under an appropriate null distribution (Tibshirani *et al.*, 2001). There have also been several recent papers devoted to this issue in the information theoretic engineering literature where it is known as the *cluster validation problem*. (See, for example, Roberts *et al.* (1998), Frigui and Krishnapuram (1999), Biernacki *et al.* (2000) and references therein.) Unfortunately, many of the approaches that have been suggested for choosing the number of clusters were developed for a specific problem and are somewhat ad hoc. Those methods that are more generally applicable tend either to be model-based, and hence require strong parametric assumptions, or to be computation-intensive, or both.

In this paper we develop an alternative approach to choosing the number of clusters that makes limited parametric assumptions, can be rigorously theoretically motivated using ideas from the field of rate distortion theory, is both simple to understand and compute, and is highly effective on a wide range of problems. The procedure is based on “distortion” which is a measure of within cluster dispersion. Formally, let $\mathbf{X}$ be a p-dimensional random variable having a mixture distribution of G components, each with covariance Γ, let $\mathbf{c}_1, \mathbf{c}_2, \ldots, \mathbf{c}_K$ be a set of candidate cluster centers, and let $\mathbf{c}_\mathbf{x}$ be the one closest to $\mathbf{X}$. Then the minimum achievable distortion associated with fitting K centers to the data is

$$d_K = \frac{1}{p} \min_{\mathbf{c}_1,\cdots,\mathbf{c}_K} E\left[(\mathbf{X}-\mathbf{c}_\mathbf{x})^T \Gamma^{-1} (\mathbf{X}-\mathbf{c}_\mathbf{x})\right] \tag{1}$$

which is simply the average Mahalanobis distance, per dimension, between $\mathbf{X}$ and $\mathbf{c}_\mathbf{x}$. Note that in the case where Γ is the identity matrix distortion is simply mean squared error. In practice one generally estimates d_K using $\widehat{d_K}$, the minimum distortion obtained by applying the k-means clustering algorithm (Hartigan, 1975) to the observed data.

A natural, but overly simplistic approach to choosing the number of clusters, is to plot d_K versus K and look for the point at which the resulting “distortion curve” levels off. This curve is always monotone decreasing. However, intuitively one would expect much smaller drops for K greater than the true number of clusters, G, because past this point adding more centers simply partitions within rather than between groups. Figure 1 shows distortion curves for three different data sets. Since the curves all have similar shapes, the ad hoc method described above would suggest that they have roughly the same number of clusters. This is not the case. Figure 1(a) corresponds to the classic iris data set (Fisher, 1936) which consists of two species whose characteristics overlap and a third well separated one, and could thus be viewed as having either two or three clusters. Figures 1(b) and 1(c) give the distortion curves for a mixture of six Gaussian distributions and a single Gaussian respectively.

The above example clearly illustrates that there are problems with using the raw distortion. None-the-less, all the requisite information for choosing the correct number of clusters is contained in the distortion curve. It is simply necessary to understand more precisely the curve’s functional form in both the presence and absence of clustering. In this paper we show, both theoretically and empirically, that for a large class of distributions the distortion curve, when transformed to an appropriate negative power, will exhibit a sharp jump at the “true” number of clusters. Our basic procedure, which we call the “jump method” has the following simple steps for estimating the true number of clusters:

1. Run the k-means algorithm for different numbers of clusters, K, and calculate the corresponding distortions, $\widehat{d_K}$.
2. Select a transformation power, $Y > 0$. (A typical value is $Y = p/2$.)
3. Calculate the “jumps” in transformed distortion, $J_K = \widehat{d}_K^{-Y} - \widehat{d}_{K-1}^{-Y}$.
4. Estimate the number of clusters in the data set by $K^* = \arg\max_K J_K$, the value of K associated with the largest jump. (Note that we define $d_0^{-Y} \equiv 0$ so the method can select $K^* = 1$ if there is no clustering in the data.)

For the data sets of Figures 1(b) and 1(c) our jump method correctly chooses $K^* = G = 6$ and $K^* = G = 1$ respectively. For the iris data it indicates that either two or three clusters is a reasonable choice.

In Section 2 we introduce some of the key information theoretic results from the subfield of electrical engineering known as rate distortion theory and show how they relate to the cluster analytic distortion curve. These results are used in Section 3 to derive the exact asymptotic form of the distortion curve for both a single Gaussian distribution and a mixture of G Gaussians. This in turn motivates the jump algorithm, which we demonstrate on a variety of simulated data sets. In Section 4 we develop a general theory which shows that, for almost any mixture distribution, this approach is guaranteed to produce the correct answer provided

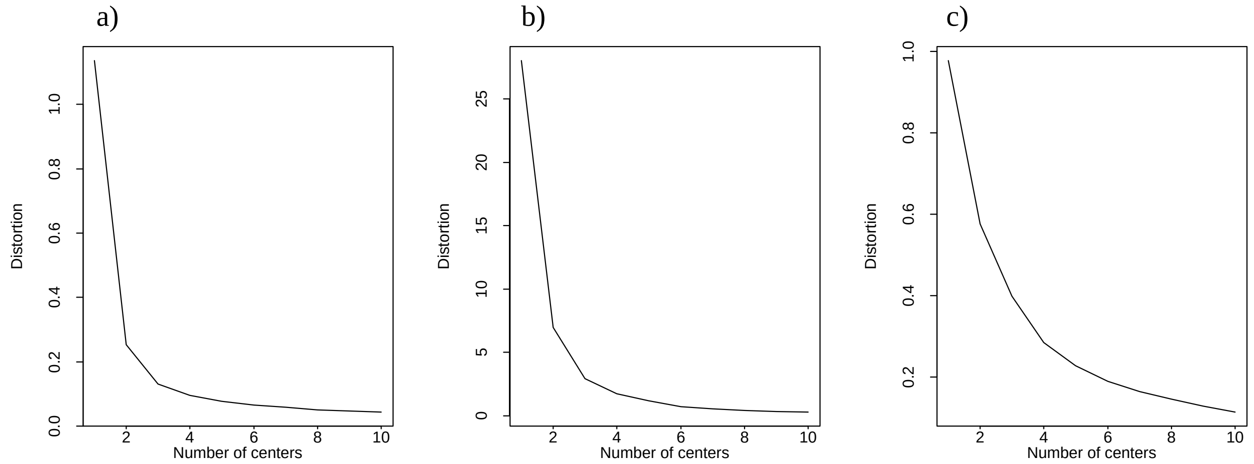

Figure 1: *Distortion curves for (a) the iris data, (b) a simulated data set with* 6 *mixture components and (c) a single Gaussian cluster.*

the clusters do not overlap too severely. We then illustrate the jump method on several real world data sets. Hypothesis tests and confidence intervals for the true number of clusters are developed in Section 5. In Section 6 we present a comparative simulation study to assess the performance of the jump method versus five competing approaches. We conclude in Section 7 by discussing possible extensions of this work. In particular, we believe that the ideas from rate distortion theory which are applied in this paper to cluster analysis may potentially prove useful for a much larger class of statistical model selection problems.

2 Rate distortion theory

Figure 1(c) suggests that the distortion curve is smooth (approximately hyperbolic) when there is little or no clustering in the data. Information theoretic results from the area of electrical engineering known as *rate distortion theory* explain this phenomenon and provide a theoretical underpinning for approaches to estimating and performing tests about the optimal number of clusters. Section 2.1 gives an intuitive introduction to rate distortion theory and explains its relationship to statistics in general and cluster analysis in particular. Section 2.2 presents some results that provide insight about the functional form of the distortion curve.

2.1 Connections between rate distortion theory and cluster analysis

One can characterize cluster analysis as an attempt to find the best possible representation of a population using a fixed number of points. This can be thought of as performing data compression or quantization on i.i.d. draws from a given distribution. In exchange for compressing the information contained in the data one must introduce some imprecision in or “distortion” of the original values in much the same way as with a histogram. In order to minimize the error one uses a finite list of representatives chosen so that, with the exception of regions of low probability, no point will be too far from its representation. This entails placing the representatives in the regions of highest density, in other words where the data are clustered. In this paradigm, each cluster center provides a representation for nearby observations and the distortion, d_K, gives a measure of the best possible level of accuracy that can be obtained using K clusters. The data will be well-summarized when one picks the correct number of centers.

This is an analogue of the main problem of rate distortion theory, which, in engineering terminology, is to *code*, as accurately and efficiently as possible, the output of a *source*. Typically the source output consists of a sequence of realizations of a continuous random variable. Representing or transmitting a real number with perfect accuracy requires storing an infinite number of *bits* (base two digits) which is not feasible. Instead, a finite set of *codewords* is chosen so as to approximate the numbers or *source symbols* as well as possible. One defines a distance function, the *distortion,* between a source symbol and its representation to measure the "goodness" of the code. A typical criterion for a good code is that it should minimize the expected distortion for a draw from the underlying probability distribution of the source. Therefore, the central problem in rate distortion theory is to find the best possible distortion achievable with a given number of codewords. In the statistical setting, the number of clusters, K, is equivalent to the number of codewords, the cluster centers provide canonical representations of members of their respective groups, and the squared (Mahalanobis) distance between an observation and its closest center serves as the distortion function.

In coding theory one is principally interested in the average number of bits that will be required for a representation. This quantity is referred to as the *rate*, R, (per source symbol) of a code. For a simple code, the relationship between the rate and the number of codewords or cluster centers is given by $K = 2^R$. The minimum rate achievable for any given distortion is called the *rate distortion function,* $R(D)$, and, correspondingly, the minimum distortion achievable for any given rate is the *distortion rate function,* $D(R)$. Essentially, $R(D)$ and $D(R)$ provide a way to formalize how many representatives to use and how good a job they can do at data summarization. The distortion rate function, $D(R)$, is intuitively the cluster analytic distortion curve–i.e. the minimum distortion achievable for a given number of representatives–substituting the number of centers in place of the rate. $D(R)$ and d_K are not technically completely equivalent. However, $D(R)$ does provide a lower bound for d_K and empirical evidence suggests that the two curves behave similarly.

The rate distortion and distortion rate functions have an information theoretic interpretation. In fact, the key result of rate distortion theory states that

$$R(D) = \min_{f(\hat{x}|x): E_{X,\hat{X}}[d(X,\hat{X})] \leq D} I(X;\hat{X}) \tag{2}$$

where $d(X,\hat{X})$ is the distortion between the source, X, and its representation, $\hat{X}$, and $I(X;\hat{X})$ is the Shannon mutual information between X and $\hat{X}$. The mutual information is defined as

$$I(X;\hat{X}) = \iint \log \frac{f_{X,\hat{X}}(x,\hat{x})}{f_X(x)\, f_{\hat{X}}(\hat{x})} f_{X,\hat{X}}(x,\hat{x}) dx d\hat{x}$$

where f_X and $f_{\hat{X}}$ are, respectively, the marginal densities of X and $\hat{X}$, and $f_{X,\hat{X}}$ is the joint distribution. $I(X,\hat{X})$ gives the expected information contained in $\hat{X}$ about a draw from the distribution of X and hence provides a measure of the ability to predict one variable given the other. Equation 2 says that the minimum achievable rate, $R(D)$, is equal to the minimum amount of information about the source, X, that is contained in a conditional distribution of a representation, $\hat{X}$, that achieves distortion, D. The mutual information is more familiar to statisticians as the Kullback-Leibler divergence (Kullback and Leibler, 1951) between $f_{X,\hat{X}}$ and $f_X f_{\hat{X}}$. Hence $I(X;\hat{X})$ gives the divergence between the joint distribution of X and $\hat{X}$ and the product of the two marginal distributions, and can be thought of as a measure of the lack of independence between the two random variables. Mutual information and related ideas such as entropy have been widely used in statistics. Examples include hypothesis testing and information sufficiency (Kullback and Leibler, 1951), the construction of multivariate measures of dependence Joe (1989), the selection of reference priors (Bernardo, 1979; Berger and Bernardo, 1989), the Bayesian Information Criterion (Schwarz, 1978), and Bayesian interpretation of experiments (Lindley, 1956). The latter provides one of the most direct translations of coding theory ideas to statistics. Specifically, in a statistical setting one can interpret the source output as a draw from the prior density on a parameter space, and the received signal as data drawn from the posterior distribution. In this formulation, the mutual information gives the expected information the data will have about the parameter and hence measures the amount of information associated with the experiment. Less frequently, the rate distortion function itself has been used in statistics. For example, Yuan and Clarke (1999a,b) use it as a criterion for likelihood selection. A detailed summary of the information theoretic statistics literature and its relationship to the pioneering work of C.E. Shannon is given by Soofi (1994).

2.2 Asymptotic rate distortion theory results

Below we give some well known results from asymptotic rate distortion theory which are used in Sections 3 and 4 to motivate the jump method:

(I) For a given code, the rate distortion function, $R(D)$, is a non-increasing convex function of D. Similarly, the distortion rate function, $D(R)$, is a non-increasing convex function of R.

(II) If X is p-dimensional normal with mean vector μ, and covariance structure $\sigma^2 I$, then, under squared-error distortion, the rate distortion and distortion rate functions are

$$R(D) = \frac{p}{2}\log_2 \frac{p\sigma^2}{D} \quad \text{and} \quad D(R) = p\sigma^2 2^{-\frac{2R}{p}} \tag{3}$$

(III) For a scalar random variable X with variance σ^2 the following are bounds on the rate distortion and distortion rate functions of X based on squared error distortion:

$$H(X) + \frac{1}{2}\log_2 \frac{1}{(2\pi e)D} \leq R(D) \leq \frac{1}{2}\log_2 \frac{\sigma^2}{D}$$

$$\frac{2^{-2R}2^{2H(X)}}{2\pi e} \leq D(R) \leq \sigma^2 2^{-2R} \tag{4}$$

where $H(X) = -\int f(x)\log_2 f(x)dx$ is the entropy of the distribution of X.

The first result suggests that any choice of the number of clusters based on the distortion curve or monotone transformations thereof will be admissible in the sense that no randomized scheme would do better. It has been conjectured that the distortion curve itself is always convex. However this has proven difficult to establish. Sugar (1999) gives a proof of convexity under certain hierarchical restrictions on the clustering methodology. Results (II) and (III) follow from the maximum entropy property of the Gaussian. Versions of (II) exist for more complex covariance structures. However, it is difficult to calculate the distortion rate function for a general distribution. As an application of the third result, consider the uniform distribution, $X \sim U(a,b)$, where $H(X) = \log_2(b-a)$ and $\sigma^2 = (b-a)^2/3$. One gets

$$\frac{(b-a)^2}{2\pi e 2^{2R}} \leq D(R) \leq \frac{(b-a)^2}{3 \times 2^{2R}}.$$

There are several things worth noting about these bounds. First, the functional forms of the upper and lower bounds are the same in terms of R and D. The only difference is in the multiplicative constants. In practice, the shape of the distortion curve usually mirrors the bounds. Second, in the case of both the normal distribution and the more general bounds of (III) we see that there is an inverse relationship between rate and distortion of the form $R \propto -\log_2 D$ or equivalently $D \propto 2^{-2R}$. Empirically this pattern holds in general and will lead

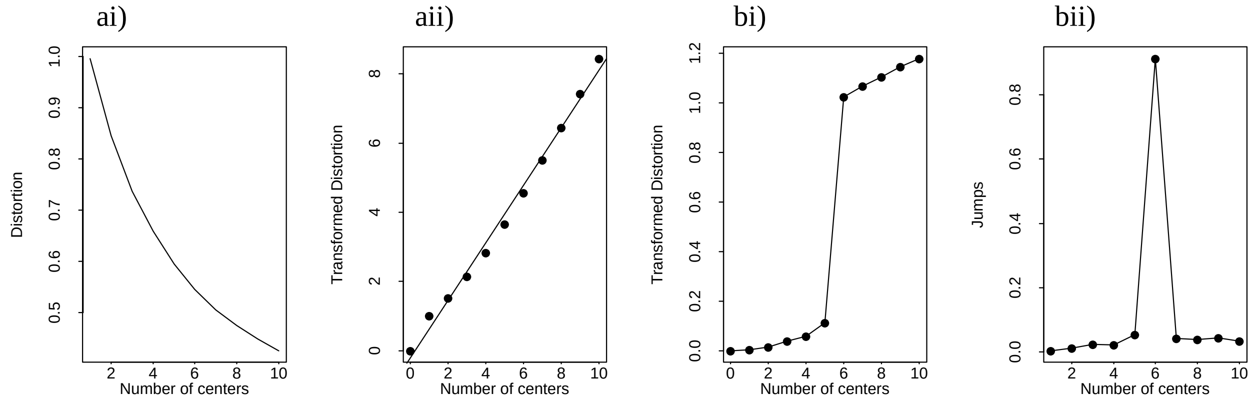

Figure 2: *Distortion curves for simulated data sets with (a) a single mixture component and (b)* 6 *mixture components.*

us to transformations of the distortion curve that prove extremely valuable for identifying the true number of clusters.

Most of the fundamental work in this area is due to C.E. Shannon who pioneered the field of mathematical communication (Shannon, 1948). The notion of a rate distortion function was introduced in Shannon (1959) The interested reader should see Cover and Thomas (1991) for a more complete development including extensive references and proofs presented from a fairly statistical point of view. Other sources include Berger (1971), a classic monograph on rate distortion theory, Gersho and Gray (1992) on vector quantization and signal compression, and the more general information theoretic texts Gallager (1968), McEliece (1977), Csiszar and Korner (1981) and Blahut (1987).

3 The distortion curve for Gaussian clusters

Given the wide variety of applications of cluster analysis, from partitioning a data space to searching for areas of high density to identifying distinct sub-populations, it is difficult even to define what is meant by the "true" number of clusters in a data set. One common and natural approach, which we adopt for the theoretical development in this paper, is to assume that the data come from a mixture distribution and to equate the number of clusters with the number of mixture components, G. In this paradigm, the absence of clustering corresponds to $G = 1$. In Section 3.1 we show how the results from Section 2.2 can be used to derive the asymptotic form of the distortion curve, d_K, for data generated from a mixture of Gaussian distributions. An extension to non-Gaussian clusters is made in Section 4. These results are used to motivate the jump method for choosing the number of clusters, which we illustrate on simulated data in Section 3.2.

3.1 Asymptotic results for a mixture of Gaussian clusters

In order to utilize the distortion function, d_K, to choose the correct number of clusters we must first understand its functional form both when the data set consists of a single cluster and when it is a mixture of G different clusters. Consider Figure 2(ai) which provides a plot of d_K versus the number of centers, K, for a simulated data set. The data were generated from a single Gaussian distribution with identity covariance, $p = 5$ dimensions and $n = 300$ observations. The relationship appears to be hyperbolic. Figure 2(aii) provides confirmation, giving a plot for the same data after raising d_K to the power of $-p/2 = -2.5$. A strong linear relationship is evident with $R^2 = 99.3\%$. For this data, the functional form of the distortion curve is approximately $d_K \propto K^{-0.4}$. In fact, Theorem 1 suggests that in the limit as p approaches infinity such a relationship between distortion and number of centers will always exist for Gaussian data.

Theorem 1 *Suppose that* $\mathbf{X}$ *has an arbitrary p-dimensional Gaussian distribution. Let* $K = \lfloor k^p \rfloor$ *where k can be any positive number. Then*

$$\lim_{p\to\infty} d_K = k^{-2} \tag{5}$$

The proof of Theorem 1 is given in Appendix A.1. This result derives from the fact that $d_K \to D(\log_2 k)$ as $p \to \infty$. The asymptotic form of d_K for Gaussian data then follows from (II). The quantity, k, is essentially the pth root of the number of centers, K. Hence Theorem 1 suggests that, for large enough p, the following relationship holds approximately

$$d_K^{-p/2} \propto k^p \approx K, \tag{6}$$

which explains the observed linear relationship. Even though the result is asymptotic in the dimension of the space, we see from Figure 2(a) that linearity can hold for relatively low values of p. In practice we have found that this approximate relationship exists in most situations. One might naively imagine that the constant of proportionality in (6) should be 1. However, it turns out that for most values of p the slope is strictly less than 1 and decreases as the dimension increases. For instance, the slope in Figure 2(aii) is approximately 0.83. Theorem 1 illustrates a fundamental flaw with the “intuitive” approach of examining the raw distortion curve for points where it levels off. Since a single Gaussian will have a curve approximately of the form $d_K \propto K^{-2/p}$, the distortion will decline rapidly and then plateau, leaving the impression of clustering even when none exists.

Next we consider the form of the distortion curve when the data consist of a mixture of G Gaussian clusters. Figure 2(bi) provides a plot of the transformed distortion, $d_K^{-5/2}$, versus number of centers, K, generated from a simulated data set consisting of a mixture of $G = 6$ Gaussian distributions. Notice that the plot is approximately linear for $K \geq 6$ clusters and that there is a significant jump between $K = 5$ and $K = 6$. Intuitively this jump occurs because of the sharp increase in performance that results from not having to summarize two disparate groups using the same representative. Adding subsequent cluster centers reduces the within group rather than the between group distortion and thus has a smaller impact. An alternative visualization is provided by Figure 2(bii) which plots the successive jumps in the transformed distortion. This “jump plot” proves particularly useful when the true number of clusters is not as obvious as in this example. Both the linearity for $K \geq G$ and the jump at $K = G$ occur in general. Theorem 2 gives the asymptotic form of the distortion curve for a mixture of G clusters which provides a theoretical explanation for these phenomena.

Theorem 2 *Suppose that the distribution of* $\mathbf{X}$ *is a mixture of G Gaussian clusters with equal priors and common covariance* Γ_p. *Let* $\Delta\sqrt{p}$ *be the minimum Euclidean distance between cluster means after standardizing the space by multiplying by* $\Gamma_p^{-1/2}$. *Then for* $K < G$

$$\lim_{p\to\infty} d_K = \infty$$

provided Δ *is bounded away from zero. Furthermore for* $K = \lfloor k^p \rfloor$

$$\lim_{p\to\infty} d_K = k^{-2}$$

provided $\Delta > 6$.

The proof is given in Appendix A.2. As with Theorem 1, this result derives from the fact that the distortion associated with each individual Gaussian cluster converges to the corresponding distortion rate function so

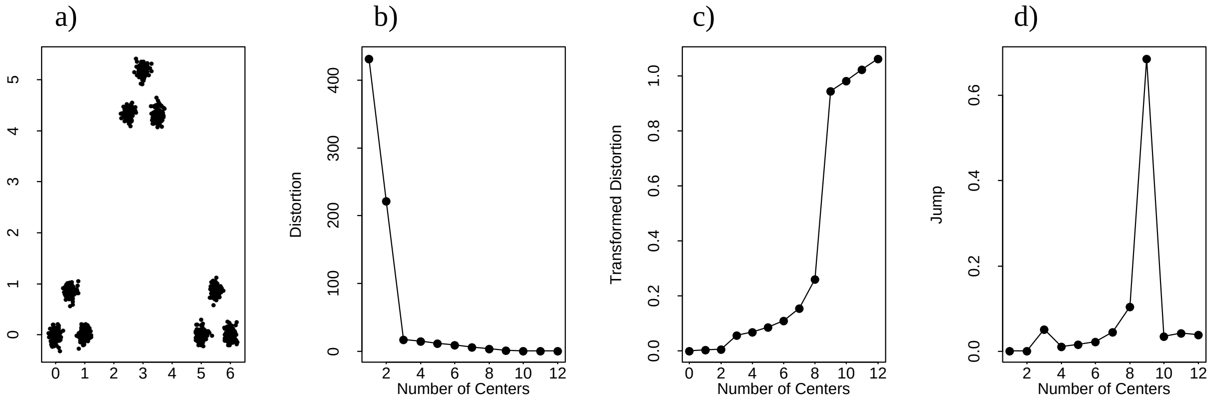

Figure 3: *a) A mixture of nine Gaussian clusters, b) the raw distortion curve which suggests only three clusters, c) the transformed curve which clearly indicates nine clusters and d) the corresponding jump curve which also clearly indicates nine clusters.*

that (II) can be applied. Theorem 2 implies that for large enough p and $K < G$, $d_K^{-p/2} \approx 0$ while for $K > G$, $d_K^{-p/2} \propto k^p \approx K$. In fact the proof of Theorem 2 suggests that the slope is proportional to $1/G$, yielding

$$d_K^{-p/2} \approx \begin{cases} a\frac{K}{G} & ,K \geq G \\ 0 & ,K < G \end{cases} \tag{7}$$

where $0 < a < 1$. This explains both the jump at $K = G$ and the linearity thereafter as seen in Figure 2(bii). As with Theorem 1, even though these results are asymptotic in p, in practice they appear to hold even in low dimensions.

Equation (7) suggests several possible procedures for utilizing the distortion curve to determine G. In particular it provides motivation for the jump method which estimates G using

$$\arg\max_K \left[\widehat{d}_K^{-Y} - \widehat{d}_{K-1}^{-Y}\right],$$

the value of K associated with the largest jump in the transformed distortion. Furthermore it suggests that an appropriate value for Y would be $p/2$. Other approaches are also possible. For example, one could use a "broken line" method by finding the value, K^*, that produces the minimum sum of squared errors when fitting two straight lines to $d_K^{-p/2}$, the first for $K < K^*$ and the second for $K \geq K^*$. This approach is based on the fact that the transformed distortion should be approximately linear for $K < G$ and for $K \geq G$. Empirically the jump method and the broken line method both work extremely well. The broken line method has the advantage of being global rather than local and as a result is potentially more robust. However, its theoretical motivation depends on the Gaussian assumption. In contrast, the jump method is almost wholly non-parametric. In Section 4.1 we show that for a general class of distributions it is guaranteed to choose $K = G$ provided that the separation between cluster means is large enough. Hence we focus primarily on the jump method for the remainder of the paper.

3.2 Simulation results

Equation 7 suggests that the jump and broken line methods will both perform well on high-dimensional Gaussian data. In this section we use empirical simulation studies to show that both methods also perform well on low-dimensional data. Figure 3 provides an example of a data set for which not only do the jump and broken

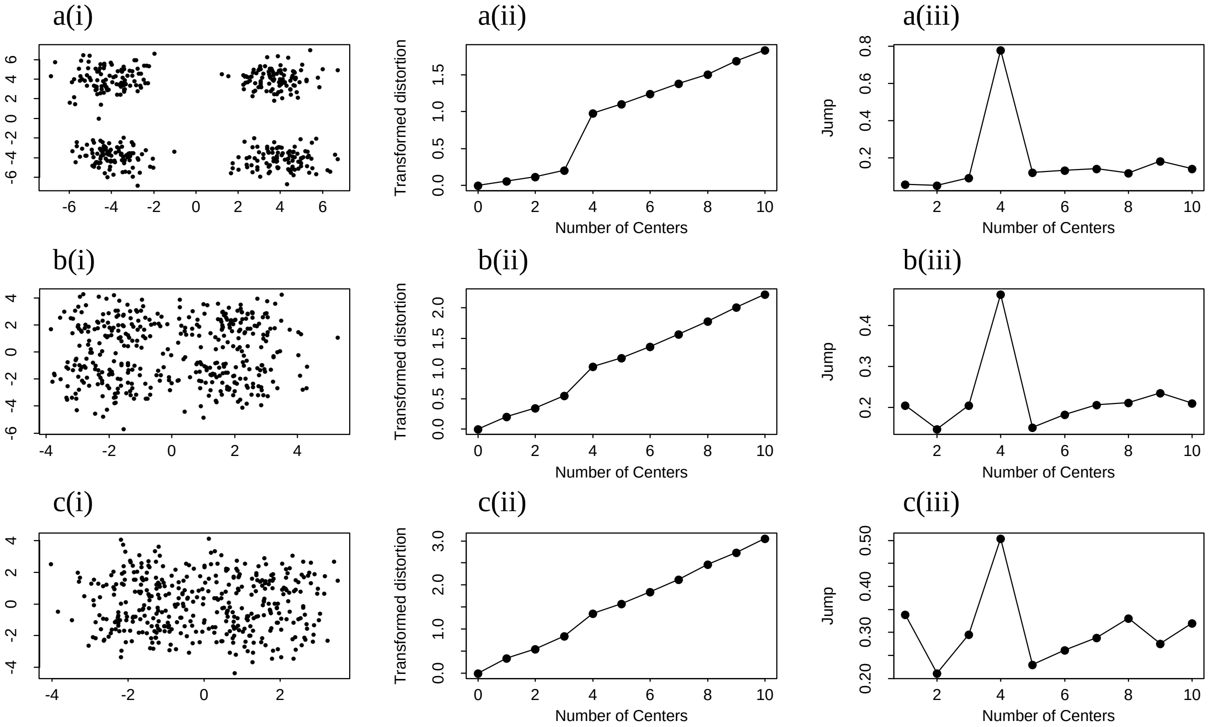

Figure 4: *Three simulated data sets, each with four Gaussian clusters, (i). Transformed distortion curves for each data set (ii), and the jumps associated with each curve (iii).*

line methods work well but using the raw distortion curve fails. Figure 3(a) shows a two-dimensional data set consisting of nine well separated clusters. In Figures 3(b) and (c) we have plotted the raw and transformed distortion curves for this data. Because the nine mixture components are themselves grouped, the raw distortion curve strongly suggests that there are only three clusters. However, after transforming the distortion curve the true number of clusters becomes readily apparent. Both the jump and broken line methods correctly select nine clusters. It is worth noting that the corresponding jump plot in Figure 3(d) exhibits a secondary peak at $K = 3$ corresponding to the three clusters of clusters. The ability to detect hierarchical structure in the clustering is an added benefit of our approach.

Figure 3 illustrates a situation in which the groups are well separated. However, the jump and broken line methods also perform well when the clusters overlap to a large extent. Figure 4 shows three data sets, each a mixture of four Gaussians. The data set of Figure 4(a) contains well separated clusters, that of Figure 4(b) has some overlap and that of Figure 4(c) is almost indistinguishable from a single cluster. The corresponding plots of transformed distortion reflect this decreasing level of separation. Figure 4a(ii) shows a clear jump at $K = 4$. The jump in Figure 4b(ii) is less extreme, while that in Figure 4c(ii) is difficult to detect. However, the corresponding jump plots all clearly indicate four clusters. As the separation between clusters decreases the transformed distortion curve becomes closer to linear as predicted by Theorem 1. However, this example shows that the jump and broken line methods can still produce accurate answers for highly confounded clusters. To estimate the statistical power of these approaches we simulated 100 data sets from the distribution used in Figure 4(c). The broken line method correctly picked $K = 4$ on 92% of the data sets and the jump method on 100%. As an aside, it is interesting to note that in Figure 4 the jump at $K = 1$ steadily increased with the confounding of the groups. In Section 4.1 we present results which show that under appropriate conditions the jump method will pick $K = 1$ in the absence of clustering.

4 The distortion curve for non-Gaussian clusters

The theoretical and empirical results of Section 3 show that the distortion curve, appropriately transformed, provides an excellent basis for choosing the correct number of Gaussian mixture components. In Section 4.1 we extend the theory of Section 3.1 to a large class of non-Gaussian distributions while also relaxing the asymptotic requirement on p. In Section 4.2 we apply the jump method to several real world data sets.

4.1 Theoretical results for mixtures of non-Gaussian clusters

Results from rate distortion theory can also be applied to non-Gaussian data. In particular, (4) provides bounds on the distortion for arbitrary distributions. While it is not possible to use these bounds to derive the exact theoretical form of the distortion curve in the general case, this result does allow us to prove, under suitable conditions, that the largest jump in transformed distortion will be at $K = G$. We summarize our findings in Theorem 3.

Theorem 3 *Suppose that the distribution of* $\mathbf{X}$ *is a mixture of* G *p-dimensional clusters with equal priors. Furthermore, assume that the clusters are identically distributed with covariance* Γ_p *and finite fourth moments in each dimension. Let* $\Delta\sqrt{p}$ *be the minimum Euclidean distance between cluster means after standardizing. Let* $H^*(X)$ *be the minimum entropy, conditional on cluster membership, over each of the p dimensions after standardizing. Finally, let*

$$W = 1 - \frac{6^4 V_{\mathbf{X}}}{(\Delta^2 - 36)^2} \tag{8}$$

where

$$V_{\mathbf{X}} = Var\left(\frac{1}{p}||\mathbf{X} - \mu_j||^2_{\Gamma^{-1}}|\mathbf{X} \text{ in } jth \text{ cluster}\right). \tag{9}$$

Suppose d_K *is computed for* $1 \leq K \leq K_{max}$*. Then as long as* $G \leq K_{max}$*, the jump*

$$\left[d_K^{-Y} - d_{K-1}^{-Y}\right]$$

will be maximized when $K = G$ *provided that* $\Delta > 6$ *and there exists* $Y > 0$ *such that*

$$\left(\frac{p\Delta^2 W}{9G}\right)^{-Y} + \left(W\left[\frac{2^{2H^*(X)}}{K_{max}^2 2\pi e} - \left(\frac{\Delta}{6}\right)^2 (1-W)\right]\right)^{-Y} < 2 \quad and \quad \left(\frac{p\Delta^2 W}{9G}\right)^{-Y} < 1/2 \tag{10}$$

Furthermore, in the limit as $\Delta \to \infty$ *the jump method is guaranteed to produce the correct answer for all p provided that*

$$0 < Y < \left[\log_2(K_{max}^2 2\pi e) - 2H^*(X)\right]^{-1}. \tag{11}$$

Finally, if the dimensions are independent, the bounds on Y provided by (11) apply in the limit as $p \to \infty$ *for all* $\Delta > 6$*.*

The proof is given in Appendix A.3 and has two main parts. First we show that the transformed distortion is bounded above for all values of $K < G$ provided that there is some separation in the clusters. Second, we show that the transformed distortion must be no less than 1 for $K = G$ and that the transformed distortion is also bounded for $K > G$. Provided that both bounds are tight enough, this proves that the maximum jump must be at $K = G$. The final bound is established using (4). The proof provides some intuition as to why there is a large jump at $K = G$. Provided the clusters have reasonable separation the distortion will be large for $K < G$ and hence the transformed distortion will be low. At $K = G$ the distortion will be no more than 1 and hence the transformed distortion will jump to at least 1. Finally, (4) guarantees that the distortion for $K > G$ must be bounded away from zero and hence the transformed distortion can not exhibit any other large jumps.

As a consequence of Theorem 3 we can easily prove that when there is no clustering in the data the maximum jump will be at $K = 1$ for sufficiently low values of Y. We state this result in Corollary 1.

Corollary 1 *Define* $d_0^{-Y} \equiv 0$. *In the absence of clustering (*$G = 1$*) and assuming the distribution of* $\mathbf{X}$ *has a finite fourth moment in each dimension, then for* $1 \leq K \leq K_{max}$ *the jump*

$$\left[d_K^{-Y} - d_{K-1}^{-Y}\right]$$

will be maximized when $K = 1$ *provided*

$$0 < Y < \left[\log_2(K_{max}^2 2\pi e) - 2H^*(X)\right]^{-1}. \qquad (12)$$

The proof is given in Appendix A.4. Note that (12) is not an asymptotic result. It holds for any value of p and any distribution with finite fourth moment. Corollary 1 proves very useful in Section 5 when we develop hypothesis tests for the presence of clustering in a data set.

Theorem 3 and Corollary 1 together guarantee that, provided there is sufficient separation between centers and an appropriate transformation is used, the jump method will produce the correct answer for clusters having any distribution with finite fourth moments. In practice we have found that the constraints given by (10) are overly conservative and in particular that the jump method is effective even for very small values of Δ. Interestingly, it can be shown that for Gaussian mixtures the upper bound in (11) and (12) can be replaced by infinity, but this is not true for any other distribution. This is a consequence of the maximum entropy characterization of the Gaussian and suggests that the further the cluster distributions are from Normal, the smaller the transformation power should be. However, it is not obvious how to choose the optimal value of Y. The constraints in (10) and (12) are useful for proving existence but can not be calculated in real applications. In Section 6 we discuss a promising approach, based on effective dimensions, which we use to guide our choices of Y in the examples of Section 4.2. This is an area of ongoing research.

4.2 Applications

In this section we apply the jump method to three real world data sets. The first is the well known iris data (Fisher, 1936) which contains 150 measurements on four variables for three different species of iris. The second is the Wisconsin breast cancer data set (Wolberg and Mangasarian, 1990) which consists of measurements of nine variables for each of 683 patients. Biopsies for 444 of these patients were benign, while those of the remaining 239 were malignant. Finally we explore the auto data (Quinlan, 1993) which records eight measurements for each of 398 types of cars. Because of high correlations between some variables, the actual clustering on the auto data was performed on a two-dimensional data set formed using principal components. The auto data provide a good example of a situation in which the number of groups is possibly large and not known *a priori*. The breast cancer and auto data sets were both taken from the University of California - Irvine machine learning repository.

Figures 5(a) and (b) show jump plots for the iris data set with $Y = 2/3$ and $Y = 1$ respectively. In the first plot the maximum jump is at $K = 2$ but the jump at $K = 3$ is almost as large. In the second plot the situation is reversed. Thus there is strong evidence for either two or three clusters but it is unclear which of these is

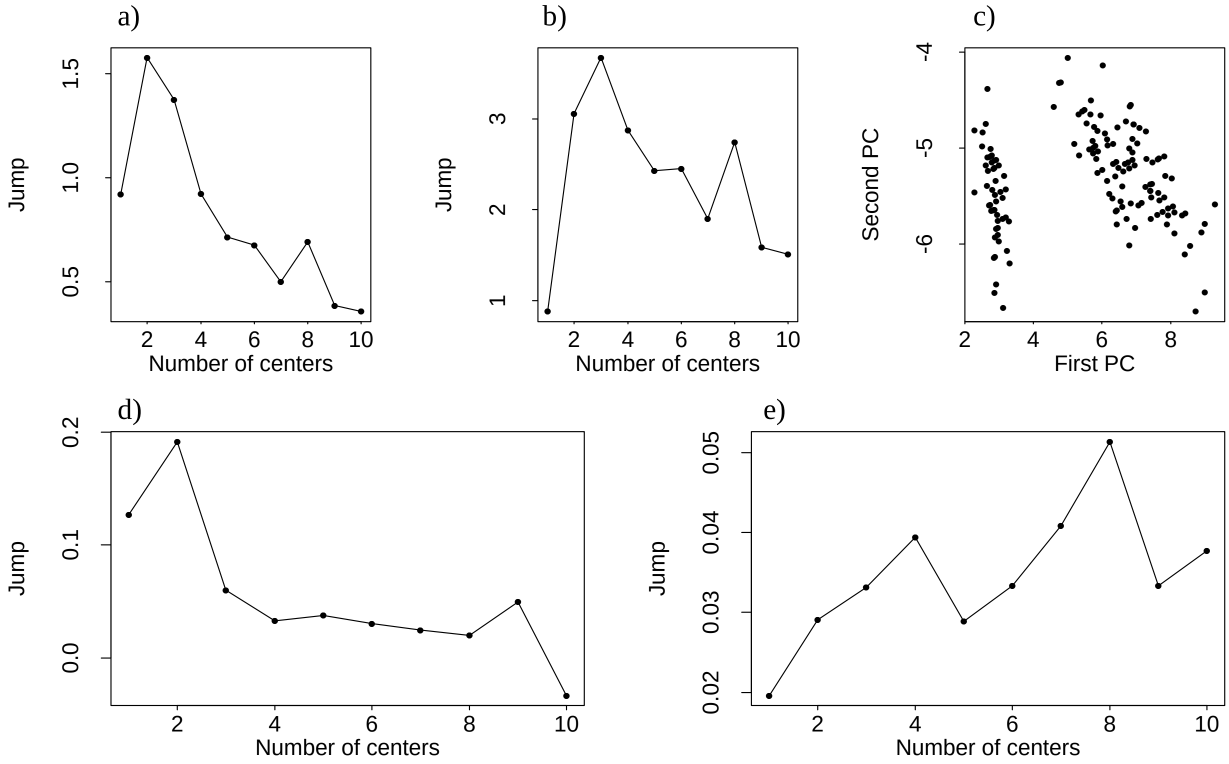

Figure 5: *Jump plots for the iris data using (a)* $Y = 2/3$ *and (b)* $Y = 1$, *(c) a plot of the iris data, and jump plots for (d) the breast cancer data and (e) the auto data.*

the best choice. This is exactly the outcome we should expect. Recall that the iris data set contains three classes. However, Figure 5(c), which plots the first two principal components of the iris data, illustrates that the clusters for two of the species overlap while the third is quite distinct. Thus from a clustering, as opposed to classification, point of view it is not clear whether the data should be treated as one large and one small cluster or as three small clusters. This is another nice example of the way in which the transformed distortion curve can be used to identify fine points of structural detail. Figure 5(d) gives the jump plot for the breast cancer data using $Y = 1$. It shows a sharp peak at $K = 2$. The clustering separates patients almost perfectly based on whether their biopsies were benign or malignant. All numbers of clusters greater than two have significantly smaller jumps, indicating that there is no evidence of sub-clusters within these two groups. The jump plot for the auto data with $Y = 2/3$, Figure 5(e), has a quite different pattern. The largest jump is at $K = 8$ but there are also substantial jumps at a variety of other values. This suggests that there are multiple clusters in the auto data set but it is difficult to tell exactly how many. This will be clarified in the following section where we develop hypothesis tests and confidence intervals for the number of clusters and also discuss the choice of the transformation power Y.

The results of Sections 3 and 4 are based on the expected distortion curve given by (1). In practice one must estimate this function by applying the k-means algorithm to the observed data. Potential sources of error arise from the use of the empirical rather than underlying distribution of the data and from the fact that it is not always possible to obtain the true k-means solution. A third form of uncertainty is introduced because the covariance matrix, Γ, is rarely known in practice. One solution is to estimate Γ as part of the clustering process. Another option is to ignore Γ by using squared error rather than Mahalanobis distance. In our experience, the shape of the distortion curve based on squared error is robust to a wide range of covariances,

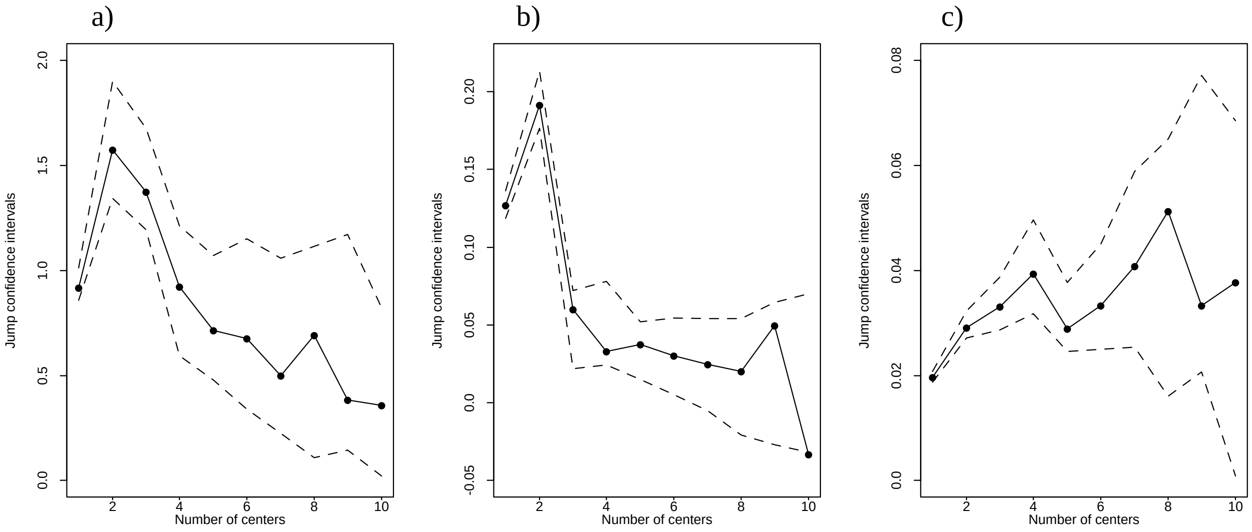

Figure 6: *Approximate* 90% *confidence intervals for the jumps on the a) iris data, b) breast cancer data and c) auto data.*

so we used this approach in our examples.

5 Testing and validation

The results of the previous sections show that the jump method provides accurate estimates of the number of clusters for a wide variety of problems. By examining the relative sizes of the jumps it is also possible to evaluate informally the certainty of these estimates. For example, Figure 5(d) shows that for the breast cancer data the jumps at $K = 1$ and 2 are by far the largest, strongly indicating that there are no more than two clusters in the data. However, for the auto data there appear to be many reasonable choices for the estimate of G. Next we develop some more formal approaches for assessing the certainty in the choice of the number of clusters.

Ideally one wishes to estimate the variability associated with each jump in order to test for statistical significance. A natural approach to this problem is to use the bootstrap (Efron and Tibshirani, 1993). Simply draw with replacement from the given data set to produce a bootstrap sample with the same number of observations as the original and calculate the jumps associated with this new data set. Repeat this process B times. We produced $B = 100$ bootstrap replicates of the jumps at each value of K and used their 5th and 95th percentiles to obtain pointwise 90% confidence intervals for the jump plots of Figures 5(a), (d), and (e). Figure 6 shows the results, with dashed lines denoting the confidence boundaries. Figure 6(a) makes it clear that there are either two or three clusters in the iris data but that it is not possible to distinguish between these two answers. Figure 6(b) provides strong evidence of two clusters in the breast cancer data, while Figure 6(c) gives convincing evidence of the existence of clusters but no indication of the actual number.

A related approach is to calculate, for each value of K, the fraction of bootstrap data sets that have their maximum jump at K. One can then take as a $(1-\alpha)100\%$ confidence interval the smallest collection of K's that account for at least $1-\alpha$ of the total. For example, for the iris data 99% of all bootstrapped data sets had their maximum jump at either $K = 2$ or 3 so a 99% confidence interval would consist of these two numbers. For the breast cancer data the jump method selected $K = 2$ for all 100 bootstrap data sets so any confidence interval for this data would contain just the value two. Interestingly, despite the ambiguity in Figure 6(c), this procedure decisively indicates that there are a large number of clusters in the auto data, with an 87% interval consisting of the values $K = 8$ through 10 and a 97% interval including $K = 7$ through 10.

The above procedure also allows one to perform a simple hypothesis test for the presence of clustering, i.e. the existence of at least two clusters in the data. Corollary 1 indicates that in the absence of clustering the largest jump should be at $K = 1$. Hence if a $(1-\alpha)100\%$ confidence interval does not include $K = 1$ then one can be confident at level α that there is clustering in the data. The 97% confidence intervals for the iris, breast cancer and auto data sets all failed to include $K = 1$ so we can be confident that they each had some form of clustering.

There is an interesting tradeoff in picking the transformation power Y. As we saw with the iris data, this choice can have some effect on the estimated value of G. In general, the closer Y is to zero the more concave the transformed distortion curve will be and hence the more likely it is that the maximum jump will occur at $K = 1$, even in the presence of clustering. Therefore, lower values of Y decrease the power of the above hypothesis test. However, we see from (12) in Corollary 1 that if Y is too large we are no longer guaranteed that the biggest jump will occur at $K = 1$ even if there is no clustering. Thus, if Y is too large, the significance level of the test may be overstated. In general, the largest value that Y can take on without misspecifying the significance level will depend on how close the cluster distributions are to Gaussian. For approximately normal data one may use a large value of Y, but for very non-Gaussian data the transformation power needs to be considerably lower. In some situations it may be possible to estimate the cluster distributions and hence the optimal value of Y. If this is not practical, then we recommend using a relatively low value to guarantee correct significance levels.

6 A comparative simulation study

In this section we present results from a comprehensive simulation study to compare the performance of the jump procedure with five standard approaches. These methods make use of the following statistics.

$$\begin{aligned}
CH(K) &= \frac{B(K)/(K-1)}{W(K)/(n-K)} && (13)\\
KL(K) &= \left|\frac{\text{DIFF}(K)}{\text{DIFF}(K+1)}\right|, \quad \text{DIFF}(K) = (K-1)^{2/p}W(K-1) - K^{2/p}W(K) && (14)\\
H(K) &= (n-K-1)\left[\frac{W(K)}{W(K+1)} - 1\right] && (15)\\
s(i) &= \frac{b(i)-a(i)}{\max[a(i),b(i)]} && (16)\\
\text{Gap}(K) &= \frac{1}{B}\sum_b \log(W_b^*(K)) - \log(W(K)) && (17)
\end{aligned}$$

The first method, suggested in Calinski and Harabasz (1974), chooses the number of clusters as the argument maximizing (13) where $B(K)$ and $W(K)$ are, respectively, the between and within cluster sum of squares with K clusters. $CH(K)$ has the form of an ANOVA F-statistic for testing for the presence of distinct groups. The approach of Krzanowski and Lai (1985) maximizes $KL(K)$ as given in (14). This statistic attempts to measure rate of change in distortion, adjusting for the dimension of the space, p. Hartigan (1975) proposes choosing the smallest value of K such that $H(K)$ in (15) is less than or equal to 10. $H(K)$ is effectively a partial F-statistic for testing whether it is worth adding a $K+1$st cluster to the model. The silhouette statistic, proposed by Kaufman and Rousseeuw (1990) and shown in (16), is a measure of how well the ith point is clustered. The term $a(i)$ is the average distance between the ith point and all other observations in its cluster and $b(i)$ is the average distance to points in the nearest cluster, where nearest is defined as the cluster minimizing $b(i)$. Large values of $s(i)$ indicate strong clustering. Kaufman and Rousseeuw (1990) suggests choosing the number of clusters that maximizes the average value of $s(i)$. Finally, a more recent approach developed in Tibshirani *et al.* (2001) uses the Gap statistic, (17). With this method B different uniform data sets, each with the same range as the original data, are produced and the within cluster sum of squares is calculated for different numbers of clusters. $W_b^*(K)$ is the within cluster sum of squares for the bth uniform data set. One approach would be to maximize $\text{Gap}(K)$. However, to avoid adding unnecessary clusters an estimate of the standard deviation of $\log(W_b^*(K))$, s_K, is produced and the smallest value of K such that

$$\text{Gap}(K) \geq \text{Gap}(K+1) - s_{K+1}$$

is chosen as the number of clusters.

We compared these methods with the jump approach in five different simulations. The first simulated data set was generated from a basic two dimensional mixture of five Gaussian clusters, each with identity covariance. The second simulation, which was designed to test the effectiveness of the methods on highly multivariate data, also used a Gaussian mixture with five components, but in ten dimensions. The third simulation examined performance when there was dependence among the dimensions. It used a distribution with four Gaussian clusters each with two by two covariance matrix with correlation 0.7. We tested the effect of differing covariances in simulation four by producing four Gaussian clusters in two dimensions with correlations of -0.7, -0.3, 0.3 and 0.7 respectively. Finally, in simulation five we produced four non-Gaussian clusters arranged in a two-dimensional square using an exponential distribution with mean one independently in each dimension. All the simulated data sets contained 100 observations equally divided among the clusters. For each of the five scenarios we produced 100 data sets, ran k-means with 20 random restarts on each, and then applied the six procedures to the resulting fits. The results are shown in Table 1. All simulations report results for the jump method with $Y = p/2$ but we have also included outcomes for some other values of Y.

The jump method appears to be extremely robust. It performed well using the transformation power $Y = p/2$ in all the scenarios, while each of the other approaches did poorly in at least two. Although this simulation study is not exhaustive, it does suggest conditions under which the jump method will be effective. In particular, the jump method strongly outperformed the other approaches in simulations four and five in which the cluster distributions either had differing covariances or were non-Gaussian. In some of the simulations the jump approach occasionally incorrectly chose a very large number of clusters. It appears that the method can be somewhat sensitive to a non-optimal fit of the k-means algorithm. Originally a handful of the data sets produced this effect. We reran our procedure on these data sets with 100 random restarts of k-means rather than 20 and produced slightly improved results.

An important practical issue with the jump method is the choice of the transformation power, Y. The theory of Section 3 would suggest setting $Y = p/2$. However, these results are based on the Mahalanobis distortion which is equivalent to assuming the data have been standardized so as to be uncorrelated. When squared error distortion is used and strong correlations exist between dimensions, values of Y somewhat less that $p/2$ may produce superior results. This was the case for simulations three through five. Empirically, a promising approach involves estimating the "effective" number of dimensions in the data and transforming accordingly. For example, the iris data is four-dimensional which suggests using $Y = 2$. However, several of the variables are highly correlated. As a result, the effective dimension of this data set is closer to 2, implying that a transformation power near $Y = 1$ may be more appropriate. This is an area of ongoing research.

7 Discussion

We have shown that the jump method is highly successful at selecting the correct number of clusters on a wide range of practical problems. Moreover, our empirical results illustrate that the transformed distortion curve

Simulation	Method	Cluster estimates									
		1	2	3	4	5	6	7	8	9	10
One	CH	0	0	0	0	98	0	1	1	0	0
(Five	KL	0	0	26	0	34	9	10	16	5	0
clusters,	Hartigan	0	0	0	0	0	0	1	5	18	76
two	Silouette	0	51	21	4	24	0	0	0	0	0
dimensions)	Gap	0	0	77	0	23	0	0	0	0	0
	Jump (Y=1)	0	0	3	4	92	0	0	0	1	0
Two	CH	0	96	4	0	0	0	0	0	0	0
(Five	KL	0	0	0	0	98	0	1	1	0	0
clusters,	Hartigan	0	0	0	0	100	0	0	0	0	0
ten	Silhouette	0	100	0	0	0	0	0	0	0	0
dimensions)	Gap	0	0	0	0	100	0	0	0	0	0
	Jump (Y=4)	0	0	0	0	100	0	0	0	0	0
	Jump (Y=5)	0	0	0	0	97	0	0	0	1	2
Three	CH	0	0	0	26	2	1	4	15	22	30
(Four	KL	0	0	0	87	2	1	2	6	2	0
clusters,	Hartigan	0	0	0	0	0	0	6	18	30	46
common	Silhouette	0	0	0	100	0	0	0	0	0	0
non-identity	Gap	0	1	0	91	8	0	0	0	0	0
covariance)	Jump (Y=0.7)	0	0	0	100	0	0	0	0	0	0
	Jump (Y=1)	0	0	0	97	0	0	0	1	1	1
Four	CH	0	0	0	83	5	5	3	0	1	1
(Four	KL	0	0	0	76	7	2	3	8	4	0
clusters,	Hartigan	0	0	0	0	0	0	8	20	23	49
differing	Silhouette	0	34	0	65	1	0	0	0	0	0
covariances)	Gap	0	20	0	78	2	0	0	0	0	0
	Jump (Y=0.7)	0	0	0	100	0	0	0	0	0	0
	Jump (Y=1)	0	0	0	98	0	0	0	1	1	0
Five	CH	0	0	0	22	11	19	10	6	15	17
(Four	KL	0	0	0	71	17	4	3	0	5	0
exponential	Hartigan	0	0	0	0	0	2	7	9	17	65
clusters)	Silhouette	0	0	0	60	30	8	1	1	0	0
	Gap	85	9	0	6	2	0	0	0	0	0
	Jump (Y=0.7)	0	0	0	99	1	0	0	0	0	0
	Jump (Y=1)	0	0	0	87	4	0	1	1	4	3

Table 1: Simulation results. Simulation 1 had cluster means of $(0,0),(2.5,2.5),(5,5),(-2.5,2.5)$ and $(-5,-5)$. Simulations 2 to 4 had clusters evenly spaced on a line with separations of $1.6,5$ and 3.5 respectively in each dimension. The clusters in simulation 5 were arranged on a square with sides of length 4. All simulations had standard deviations of 1 in each dimension.

and corresponding jump plots are just as valuable as exploratory tools. For example, they can be used to detect underlying hierarchical structures in clustering, as seen with the iris and triangular nine cluster data sets. Additionally, the theory of Sections 3 and 4 potentially can be extended in several directions. First, empirical evidence suggests that the linearity of the transformed distortion holds even for non-Gaussian distributions and low values of p. Recent advances in an area known as Bennett theory, which deals with non-asymptotic rate distortion functions, may prove useful for formalizing this observation(Na and Neuhoff, 1995). Second, in practice, the requirements in (10) from Theorem 3 to guarantee the success of the jump method are overly conservative and can probably be relaxed. Related to this is the question of how best to select the transformation power, Y. Third, results from rate distortion theory can be applied to many distortion measures besides squared error. For example, codes based on Hamming distance, the number of matching coordinates, have been widely studied and their properties could be very useful when clustering categorical data such as genetic sequences.

The technical results in this paper depend heavily on information theory. Other recent work in information theoretic clustering includes Roberts *et al.* (1998), Frayley and Raftery (1998) and Biernacki *et al.* (2000) which develop Bayesian methods for choosing the number components in a Gaussian mixture distribution. These approaches differ from ours both in that they are model-based and that they make no explicit use of the distortion curve. However, it may be possible to use the results of Section 3 to establish a theoretical link with this work. Frigui and Krishnapuram (1999) suggests a more non-parametric clustering method based on an objective function involving a distortion type measure which is optimized over both cluster assignments and number of groups. However, their procedure for choosing the number of clusters can not easily be used with other methods since it is integrated into the overall clustering algorithm. In contrast, the jump approach can be applied with many clustering techniques besides k-means. For instance, James and Sugar (2003) integrates the jump method into a more model-based procedure for clustering funtional data. Of the recently suggested clustering algorithms, perhaps the one making the most use of information theoretic ideas is that of Gokcay and Principe (2002). While this algorithm does use a measure of the divergence between clusters, it does not provide any approach for choosing the number of clusters.

This paper has focused on identifying the number of groups in a data set. In addressing this problem we have drawn links between the fields of rate distortion theory and cluster analysis. We believe that these ideas can be applied to numerous other model selection problems in statistics. In such situations a common approach is to plot a goodness of fit measure versus the statistic of interest and to use the resulting curve to select the model parameter. Examples include using the sum of squared errors to choose the number of predictors in a standard regression setting, or the penalty term in a ridge regression. Similarly, a plot of cumulative explained variability is frequently used to select the optimal number of dimensions in a principal components analysis. These are special cases of a more general paradigm in which likelihood curves are used to choose modeling parameters. Often the resulting “distortion” curve is monotone so choosing the global optimum fails to produce a sensible result. Cross-validation may alleviate this problem but is computationally expensive and potentially unstable. Instead, one often attempts to find a point at which the curve levels off, indicating that there will be little improvement in goodness of fit associated with further increasing the number of parameters. This leads to the same difficulties as using the raw distortion curve to choose the number of clusters. Transformations similar to those used in the clustering context may also lead to better model selection procedures in the wide range of statistical problems that use goodness of fit measures akin to distortion.

Acknowledgments

We would like to thank Bob Gray and Richard Olshen for introducing us to the field of rate distortion theory and the associate editor and referees for numerous useful suggestions and references.

A Proofs of the theorems

Here we briefly define the notation used in the proofs. Let $X_1, X_2, \ldots$ be an *i.i.d.* sequence on the sample space or *source alphabet* $\mathcal{X}$. Typically, this alphabet will simply be a Euclidean space, $\mathcal{R}^p$. The representation space from which the codewords are drawn (also usually a Euclidean space) will be denoted by $\hat{\mathcal{X}}$. A code is said to have block length m if each codeword represents not a single source symbol but m source symbols at once. Mathematically, $X^m = (X_1, \ldots, X_m) \in \mathcal{X}^m$ is represented by $\hat{X}^m \in \hat{\mathcal{X}}^m$. Note that, regardless of the block length, each single source symbol effectively is assigned a representation symbol. For a block length 1 code, the representation symbol associated with a particular source value will always be the same. However, this need not be the case for a block length m code. Using this set of definitions, clustering can be visualized in two different ways. It can be thought of as a coding problem with a block length of $m = 1$ and p-dimensional source and representation spaces, or, alternatively, as a coding problem with a block length of $m = p$ and 1-dimensional source and representation spaces. In the proofs of Theorems 1 through 3 we make use of the second paradigm, in which case the relationship between the number of clusters and the rate is

$$K = 2^{pR}. \tag{18}$$

Let $R(D)$ be the (asymptotic) rate distortion function and $D(R)$ be the distortion rate function. Finally, we denote the finite block length distortion rate function by $D_m(R)$. This represents the lowest distortion that can be achieved with rate R and block length m.

A.1 Proof of Theorem 1

First we prove a lemma:

Lemma 1

Let $D_p(R_p)$ be the distortion rate function with finite block length p and rate R_p and suppose that $\lim_{p\to\infty} R_p = R$. Then

$$\lim_{p\to\infty} D_p(R_p) = D(R)$$

Proof
We need to show that for every $\varepsilon > 0$ there exists N_ε s.t. for all $p > N_\varepsilon$, $|D_p(R_p) - D(R)| < \varepsilon$.
First note that since $D(R)$ is continuous there exists a δ such that for all $|y - R| \leq \delta$, $|D(y) - D(R)| < \varepsilon/2$. Let $x = R - \delta$. Since $R_p \to R$ we can choose an N_1 s.t. for all $p > N_1$, $|R_p - R| < \delta$ which also implies $|D(R_p) - D(R)| < \varepsilon/2$. Therefore, since $D_p(\cdot) \geq D(\cdot)$, we have already shown that for large enough p, $D_p(R_p) - D(R) > -\varepsilon$. Now choose N_2 s.t. for all $p > N_2$, $|D_p(x) - D(x)| < \varepsilon/2$. Then for all $p > \max(N_1, N_2)$

$$D_p(R_p) - D(R) \leq [D_p(x) - D(x)] + [D(x) - D(R)] < \varepsilon/2 + \varepsilon/2 = \varepsilon.$$

Hence $|D_p(R_p) - D(R)| < \varepsilon$.
Proof of Theorem 1
First note that we may assume without loss of generality that $\Gamma = I$ so that d_K is calculated in terms of squared error. If not, one can produce an identity covariance by multiplying $\mathbf{X}$ by $\Gamma^{-1/2}$. Hence $\mathbf{X}$ can be viewed as a p-dimensional Gaussian with identity covariance or as p i.i.d. normals with variance one. Therefore, using the second formulation, block length and dimension are equivalent and sending p or block length to infinity is the same thing.

Since $K = \lfloor k^p \rfloor$ this implies $k^p - 1 \leq K \leq k^p$. Hence our distortion function d_K is simply $D_p(R_p)$ where

$\frac{1}{p}\log_2(k^p-1) \leq R_p \leq \log_2 k$. Therefore $\lim_{p\to\infty} R_p = \log_2 k$ and by Lemma 1

$$\lim_{p\to\infty} d_K = D(\log_2 k). \tag{19}$$

By (3), for a one-dimensional normal with variance one

$$R(D) = -\frac{1}{2}\log_2 \quad \Rightarrow \quad D(R) = 2^{-2R}. \tag{20}$$

Combining (19) and (20) gives

$$\lim_{p\to\infty} d_K = 2^{-2\log_2 k} = k^{-2}$$

A.2 Proof of Theorem 2

First we prove a lemma:

Lemma 2

Suppose that $\mathbf{X}$ comes from a mixture distribution of G identically distributed p-dimensional clusters with equal priors and covariance Γ. Let d_{K_j} be the average distortion per observation when allocating K_j clusters to the jth mixture component. Then, provided that $\Delta > 6$

$$W\left[\min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G} - \left(\frac{\Delta}{6}\right)^2 (1-W)\right] \leq d_K \leq \min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G}$$

where $W = 1 - \frac{6^4 V_{\mathbf{X}}}{(\Delta^2-36)^2}$ and $V_{\mathbf{X}} = Var\left(\frac{1}{p}||\mathbf{X}-\mu_j||^2_{\Gamma^{-1}}|\mathbf{X} \text{ in } j\text{th cluster}\right)$.

Proof
First note that, as with Theorem 1, we can assume that $\Gamma = I$ because if not this can be achieved by transforming to $\Gamma^{-1/2}\mathbf{X}$. Clearly $d_K \leq \min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G}$ because the right hand side is a restricted version of the left hand side. Now suppose we produce truncated distributions by constructing spheres of radius $\sqrt{p}\Delta/6$ around each cluster mean and only considering observations that fall inside a sphere, i.e.

$$||\mathbf{X}_j - \mu_j||^2 = \sum_{l=1}^{p} (X_{jl} - \mu_{jl})^2 \leq p\Delta^2/36, \quad j = 1, \ldots, G$$

where $\mathbf{X}_j$ are observations from cluster j. Let $d^*_{K_j}$ be the equivalent of d_{K_j} and d^*_K the equivalent of d_K but for the truncated data. Then it is clear that

$$d^*_K = \min_{\sum_j K_j = K}\left[\frac{\sum_j d^*_{K_j}}{G}\right]$$

because the spheres are separated by at least twice their width so that every center will be uniquely associated with the observations from only one sphere. Furthermore

$$\begin{aligned} d_K &= P(\text{Inside sphere}) \times \text{Avg dist inside sphere} + P(\text{Outside sphere}) \times \text{Avg dist outside sphere} \\ &\geq P(\text{Inside sphere}) \times d^*_K \end{aligned}$$

where $P(\text{Inside sphere}) = 1 - P(\text{Outside sphere})$ and

$$\begin{aligned} P(\text{Outside sphere}) &= P\left(\frac{1}{p}\sum_{l=1}^{p}(X_{jl}-\mu_{jl})^2 > \frac{\Delta^2}{36}\right) \\ &\leq \frac{6^4 V_{\mathbf{X}}}{(\Delta^2-36)^2} = 1-W \quad \text{(by Chebychev provided } \Delta > 6) \end{aligned}$$

Finally note that for all j

$$\begin{aligned} d_{K_j} &= P(\text{Inside sphere}) \times E(d_{K_j}|\text{Inside sphere}) + P(\text{Outside sphere}) \times E(d_{K_j}|\text{Outside sphere}) \\ &\leq d^*_{K_j} + P(\text{Outside sphere}) \times E(d_1|\text{Outside sphere}) \\ &= d^*_{K_j} + \int_{\frac{\Delta^2}{36}}^{\infty} d_1 f(d_1) d(d_1) \\ &\leq d^*_{K_j} + \left(\frac{\Delta}{6}\right)^2 (1-W) \end{aligned}$$

The last line comes from the fact that

$$\begin{aligned} \int_{\frac{\Delta^2}{36}}^{\infty} d_1 f(d_1) d(d_1) &= \int_{\frac{\Delta^2}{36}}^{\infty} (d_1-1) f(d_1) d(d_1) + P(d_1 > \Delta^2/36) \\ &\leq \frac{\int_{\frac{\Delta^2}{36}}^{\infty} (d_1-1)^2 f(d_1) d(d_1)}{\frac{\Delta^2}{36}-1} + \frac{V_{\mathbf{X}}}{\left(\frac{\Delta^2}{36}-1\right)^2} \quad \text{(by Chebychev)} \\ &\leq \frac{V_{\mathbf{X}}}{\frac{\Delta^2}{36}-1} + \frac{V_{\mathbf{X}}}{\left(\frac{\Delta^2}{36}-1\right)^2} = \left(\frac{\Delta}{6}\right)^2 (1-W) \end{aligned}$$

Therefore

$$d_K \geq W\left[\min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G} - \left(\frac{\Delta}{6}\right)^2 (1-W)\right]$$

Proof of Theorem 2

First we consider $K = \lfloor k^p \rfloor$. Note that for Gaussian data $V_{\mathbf{X}} \propto 1/p$ and so converges to 0 as $p \to \infty$. Hence by Lemma 2 we see that the lower bound on d_K converges to $\min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G}$ as $p \to \infty$ so we need only show that

$$\lim_{p\to\infty} \min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G} = k^{-2}. \tag{21}$$

First we show that

$$\lim_{p\to\infty} \min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G} \leq k^{-2}. \tag{22}$$

Note that by setting $K_j = \lfloor k^p/G \rfloor$, d_{K_j} is a finite block length distortion rate function with rate $R_p \to \log_2 k$. Hence by Lemma 1 and Theorem 1 $\lim_{p\to\infty} d_{K_j} = k^{-2}$. Since this result applies for all $j = 1, \ldots, G$ we have proven (22). However, it must also be the case that

$$\lim_{p\to\infty}\min_{\sum_j K_j=K}\frac{\sum_j d_{K_j}}{G}\geq k^{-2}$$

because even when we set $K_j=\lfloor k^p\rfloor=K$, which is the largest K_j can be, it is still the case that $\lim_{p\to\infty}d_{K_j}=k^{-2}$. Hence (21) is proved.

Now we consider $K<G$. Since we are only fitting $K<G$ centers to G clusters and the minimum distance between clusters is at least $\sqrt{p}\Delta$ it must be the case that one cluster has no centers within $\sqrt{p}\Delta/2$ of its mean. Furthermore, since at least W of this cluster's mass must lie within $\sqrt{p}\Delta/6$ of its mean,

$$d_K\geq\frac{p\Delta^2}{9G}W\to\infty\quad\text{as}\quad p\to\infty$$

A.3 Proof of Theorem 3

First note that, as with Theorem 1, we can assume that $\Gamma=I$ because if not this can be achieved by transforming to $\Gamma^{-1/2}\mathbf{X}$. Consider d_{G-1}. By exactly the same argument as given above for Theorem 2 it must be the case that

$$d_{G-1}\geq\frac{p\Delta^2}{9G}W$$

It is also clear that with G centers a distortion of at most 1 is achieved with one cluster placed at the mean of each mixture so that $d_G\leq 1$. Hence

$$[d_G^{-Y}-d_{G-1}^{-Y}]\geq 1-\left(\frac{p\Delta^2W}{9G}\right)^{-Y}\quad\text{and}\quad[d_K^{-Y}-d_{K-1}^{-Y}]\leq\left(\frac{p\Delta^2W}{9G}\right)^{-Y},\quad K<G \tag{23}$$

Consider d_{K_j}, the distortion associated with the jth cluster using K_j centers. d_{K_j} is the average distortion over the p dimensions when fitting K_j clusters so as to minimize overall distortion. Furthermore, each of these coordinate-wise distortions must be no less than the distortion achieved by fitting K_j clusters to each dimension individually. However, from (4) we see that each of these latter coordinate-wise distortions must be greater than or equal to,

$$\frac{2^{-2R_j}2^{2H^*(X)}}{2\pi e}$$

where $K_j=2^{R_j}$. But since $K_j\leq K$ for all j and we are only considering $K\leq K_{max}$

$$\frac{2^{2H^*(X)}}{K_{max}^2 2\pi e}\leq d_{K_j}. \tag{24}$$

Therefore equation (24), together with Lemma 2, implies that

$$d_K^{-Y}\leq\left(W\left[\frac{2^{2H^*(X)}}{K_{max}^2 2\pi e}-\left(\frac{\Delta}{6}\right)^2(1-W)\right]\right)^{-Y} \tag{25}$$

so from (23) and (25) the jump is maximized at $K=G$ provided (10) holds. Notice that for large enough Δ there is guaranteed to be a Y that fulfills (10). Furthermore, if the dimensions of $\mathbf{X}$ are independent from each other, for $\Delta>6$ and large enough p there is also guaranteed to be a Y that fulfills (10). In fact in the limit as

Δ or p approach infinity (10) becomes

$$\left(\frac{K_{max}^2 2\pi e}{2^{2H^*(X)}}\right)^Y < 2$$

which is fulfilled provided

$$0 < Y < \left[\log_2(K_{max}^2 2\pi e) - 2H^*(X)\right]^{-1}.$$

A.4 Proof of Corollary 1

Clearly $d_1 = 1$ so $d_1^{-Y} - d_0^{-Y} = 1$. In this case $\Delta = \infty$ so from (10) the jump is maximized provided

$$0 < Y < \left[\log_2(K_{max}^2 2\pi e) - 2H^*(X)\right]^{-1}.$$

References

Berger, J. O. and Bernardo, J. M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association* **84**, 200–207.

Berger, T. (1971). *Rate distortion theory; a mathematical basis for data compression*. Englewood Cliffs, N.J., Prentice-Hall.

Bernardo, J. M. (1979). Expected information as expected utility. *The Annals of Statistics* **7**, 686–690.

Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22**, 7, 719–725.

Blahut, R. E. (1987). *Principles and practice of information theory*. Reading, Mass. : Addison-Wesley.

Calinski, R. B. and Harabasz, J. (1974). A denrite method for cluster analysis. *Communications in Statistics* **3**, 1–27.

Cover, T. M. and Thomas, J. A. (1991). *Elements of Information Theory*. Wiley, 2nd edn.

Csiszar, I. and Korner, J. (1981). *Information theory : coding theorems for discrete memoryless systems*. New York : Academic Press, 2nd edn.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. London : Chapman and Hall.

Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics* **7**, 179–188.

Frayley, C. and Raftery, A. (1998). How many clusters? which clustering method? - answers via model-based cluster analysis. *Technical Report no. 329, Department of Statistics, University of Washington* .

Frigui, H. and Krishnapuram, R. (1999). A robust competitive clustering algorithm with applications in computer vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **21**, 5, 450–465.

Gallager, R. G. (1968). *Information theory and reliable communication*. New York, Wiley.

Gersho, A. and Gray, R. (1992). *Vector Quantization and Signal Compression*. Boston: Kluwer Academic Publishers.

Gokcay, E. and Principe, J. C. (2002). Information theoretic clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**, 2, 158–171.

Hardy, A. (1996). On the number of clusters. *Computational Statistics and Data Analysis* **23**, 83–96.

Hartigan, J. A. (1975). *Clustering Algorithms*. Wiley.

James, G. M. and Sugar, C. A. (2003). Clustering for sparsely sampled functional data. *Journal of the American Statistical Association* (To appear).

Joe, H. (1989). Relative entropy measures of multivariate dependence. *Journal of the American Statistical Association* **84**, 157–164.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association* **90**, 773–795.

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 clusters in a data set. *Biometrics* **44**, 23–34.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* **22**, 79–86.

Lindley, D. V. (1956). On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics* **27**, 986–1005.

McEliece, R. J. (1977). *The theory of information and coding : a mathematical framework for communication*. Reading, Mass. : Addison-Wesley Pub. Co.

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.

Na, S. and Neuhoff, D. (1995). Bennett's integral for vector quantizers. *IEEE Transactions on Information Theory* **41**, 886–900.

Quinlan, R. (1993). *Combining Instance-Based and Model-Based Learning : In Proceedings on the Tenth International Conference of Machine Learning, 236-243, University of Massachusetts, Amherst.* Morgan Kaufmann.

Roberts, S. J., Husmeier, D., Rezek, I., and Penny, W. (1998). Bayesian approaches to gaussian mixture modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20**, 11, 1133–1142.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical Journal* **27**, 379–423.

Shannon, C. E. (1959). Coding theorems for a discrete source with a fidelity criterion. *IRE National Convention Record* **7**, 4, 142–163.

Soofi, E. S. (1994). Capturing the intangible concept of information. *Journal of the American Statistical Association* **89**, 1243–1254.

Sugar, C. A. (1999). An application of cluster analysis to health services research: Empirically defined health states for depression from the sf-12. *Stanford University, Department of Statistcs Technical Report* .

Tibshirani, R., Walther, G., and Hastie, T. (2001). Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society, Series B* **63**, 411–423.

Wolberg, W. H. and Mangasarian, O. L. (1990). Multisurface method of pattern separation for medical diagnosis applied to breast cytology. *Proceedings of the National Academy of Sciences, U.S.A.* **87**, 9193–9196.

Yuan, A. and Clarke, B. (1999a). An information criterion for likelihood selection. *IEEE Transactions on Information Theory* **45**, 562–571.

Yuan, A. and Clarke, B. (1999b). A minimally informative likelihood for decision analysis: illustration and robustness. *Canadian Journal of Statistics* **27**, 649–666.