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

June 17, 2002

Abstract

One of the most difficult problems in cluster analysis is the identification of the number of groups in a given data set. Most previously suggested approaches to this problem are either somewhat ad hoc or require parametric assumptions and complicated calculations. We suggest an alternative, non-parametric, method based on distortion, a quantity that measures the average distance, per dimension, between each observation and its closest cluster center. Our approach has the following key advantages. We are able to provide a rigorous theoretical justification for the method without relying on parametric assumptions. As a result our approach should be successful in a wide range of problems. We demonstrate its effectiveness, not only for choosing the number of clusters, but also for identifying underlying structure, on a variety of simulated and real world data sets. In addition our methodology is intuitively easy to understand and computationally efficient and straightforward to implement. In the process of developing the theoretical foundation for our methodology we uncover links to the subfield of electrical engineering known as rate distortion theory which allow us to describe the behavior of the distortion in both the presence and absence of clustering. Finally, we note that rate distortion theory and the connections we have developed between it and statistics can potentially be extended to a wide range of model selection problems besides cluster analysis.

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 have been suggested. See Milligan and Cooper (1985) and Hardy (1996) for a detailed list. Examples include Wolfe’s test (Wolfe, 1970), Marriot’s test (Marriot, 1971), Hartigan’s rule (Hartigan, 1975), the upper tail rule (Mojena, 1978) and the moving average quality control rule (Mojena, 1978). More recent work includes a Gaussian model based approach using approximate Bayes factors (Frayley and Raftery, 1998; Schwarz, 1978; Kass and Raftery, 1995) and the gap statistic which compares the change in within-cluster dispersion with that expected under an appropriate null distribution (Tibshirani et al., 2001). 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 require parametric assumptions or to be computation-intensive or both.

In this paper we develop an alternative approach to choosing the number of clusters. Our method has the advantages that it 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. Note that, 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. 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, \( G \). The absence of clus-
tering corresponds to \( G = 1 \). However, empirically the methodology that we develop works well in
more general settings.

Our procedure for choosing the number of clusters is based on “distortion”, which is a measure
of the within cluster dispersion. Formally, let \( \mathbf{X} \) be a \( p \)-dimensional random variable having a
mixture distribution of \( G \) components each with covariance \( \Gamma \), and let \( \mathbf{c}_* \) be the cluster center
closest to \( \mathbf{X} \). Then the distortion associated with \( K \) centers is

\[
    d_K = \frac{1}{p} \min_{\mathbf{c}_1, \ldots, \mathbf{c}_K} E(\mathbf{X} - \mathbf{c}_*)^T \Gamma^{-1} (\mathbf{X} - \mathbf{c}_*)
\]

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

A natural, but overly simplistic approach to choosing the number of clusters, is to plot \( d_K \)
versus \( K \) and to look for the point at which the resulting “distortion curve” levels off. This curve
is always monotone decreasing but 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 clusters. 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 contains two overlapping species and a third well
separated one and could thus be viewed as having either two or three clusters. Figures 1(b) and
(c) give the distortion curves for a mixture of six Gaussian distributions and a single Gaussian
respectively. This 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 curves 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. For
the data sets of Figures 1(b) and (c) our “jump” method correctly chooses \( G = 6 \) and \( G = 1 \)
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 results from the subfield of electrical engineering
known as rate distortion theory and show how they apply 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 theory in turn suggests two
simple procedures for choosing the number of clusters using a transformation of the distortion curve
which we demonstrate on a variety of simulated data sets. In Section 4 we develop a general theory
that shows, that for almost any mixture distribution, the jump method of Section 3 is guaranteed to
produce the correct answer provided the clusters do not overlap too severely. We also demonstrate
the success of this approach on several real world datasets. Hypothesis test and confidence intervals
for the true number of clusters are developed in Section 5. We conclude in Section 6 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. 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 cluster analysis. Section 2.2 contains a more formal development, the main purpose of which is to define the distortion rate function, a close relative of the distortion curve from cluster analysis. Section 2.3 lists some results about the distortion rate function that in turn give 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. 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. This is an analogue of the main problem of rate distortion theory, which is to represent or 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 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 a representation of members of their respective groups, and the squared distance between an observation and its closest center serves as the distortion function. Coding theory and cluster analysis both attempt to minimize the distortion using a fixed number of codewords or equivalently cluster centers. In coding theory one is principally interested in the 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 $K = 2^R$.

2.2 The distortion rate function

In this section we more formally present some of the main concepts in the field of rate distortion theory. Assume that a source produces a sequence $X_1, X_2, \ldots$ i.i.d. from a distribution $q(X)$ on the sample space or source alphabet $\mathcal{X}$. Typically, this alphabet will simply be a Euclidean space, $\mathcal{R}^p$. The representation space (also usually a Euclidean space) from which the codewords are drawn will be denoted by $\mathcal{Y}$.

**Definition 1** A distortion function is a mapping from the set of source-representation pairs to the non-negative real numbers: $d : \mathcal{X} \times \mathcal{Y} \to \mathcal{R}^+$. It gives the cost of approximating $x$ by $\hat{x}$.

In this paper, we will use squared error distortion which takes this cost to be $d(x, \hat{x}) = ||x - \hat{x}||^2$, or more generally the Mahalanobis distance, $||x - \hat{x}||^2_{\text{Cov}(x)^{-1}} = (x - \hat{x})^T \text{Cov}(x)^{-1} (x - \hat{x})$. However the main results of rate distortion theory hold true for a large family of loss functions.

It turns out that one can reduce the expected distortion for a given distribution by coding more than one number or source symbol at a time. This leads to the following set of definitions:

**Definition 2** A code has 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 $\hat{X}^m \in \mathcal{Y}^m$. An encoder is a function, $f_m$, that maps a source sequence $X^m$ to an index number. For a given rate $R$, $f_m : \mathcal{X}^m \to \{1, 2, \ldots, 2^{mR}\}$ since a sequence of $m$ samples will take (approximately) $mR$ base two digits to represent. A decoder, $g_m$, takes the index value and returns a representation sequence from $\mathcal{X}^m$; that is $g_m : \{1, 2, \ldots, 2^{mR}\} \to \mathcal{X}^m$.

The distortion between sequences of $m$ symbols is taken to be the average of the per-symbol distortion, $d(x^m, \hat{x}^m) = \frac{1}{m} \sum d(x_i, \hat{x}_i)$. 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 results of Sections 3 and 4 we make use of the second paradigm, in which case the relationship between the number of clusters and the rate is

$$K = 2^{mR}. \quad (2)$$

**Definition 3** The combination of an encoding and decoding function for an index set corresponding to a fixed rate and distortion function is called a rate distortion code. Specifically, a $(2^{mR}, m)$ rate distortion code encodes and decodes sequences of length $m$ using a set of $2^{mR}$ codewords.
Definition 4 A rate distortion pair, $(R,D)$ is said to be achievable if there exists a sequence of $(2^{mR},m)$ rate distortion codes such that the limiting distortion is less than or equal to $D$. In other words, there exist pairs $(f_m,g_m)$ such that

$$\lim_{m \to \infty} \text{Ed}(X^m, g_m(f_m(X^m))) = \lim_{m \to \infty} \text{Ed}(X^m, X^m) \leq D$$

(3)

The rate distortion region for a source is the closure of the set of achievable rate distortion pairs $(R,D)$.

This definition only makes sense provided the limit in question exists. It is an interesting, and counter-intuitive, result that letting the block length, $m$, increase improves the average per-symbol distortion even if the individual elements of the block are i.i.d. This is essentially a space-filling argument. An orthogonal grid is not necessarily the best way to place $2^m$ points in $m$ dimensions. (See, Cover and Thomas (1991), pages 357-358 for a brief discussion of the Gaussian case, and Na and Neuhoff (1995) for a more general, and more complex, discussion.) Now we are at last ready to define the key idea of a distortion rate function.

Definition 5 The (asymptotic) rate distortion function $R(D)$ is the infimum of rates $R$ such that $(R,D)$ is in the rate distortion region for a given distortion $D$. In other words, $R(D)$ is essentially the minimum bit length (or number of codewords) needed to describe a source with no more than distortion $D$. The distortion rate function, $D(R)$, is the reverse of the rate distortion function. It gives the minimum distortion that can be achieved with a given length of representation. For a fixed $m$, the finite block length distortion rate function, $D_m(R)$, is the infimum of distortions $D$ that can be achieved by codes of rate $R$ and block length $m$.

The rate distortion function and distortion rate function should not be confused with what we have referred to as the distortion curve in the cluster analysis setting. In fact the latter is a finite block length distortion rate function. Unfortunately, the finite block length function, $D_m(R)$, is difficult to analyze theoretically, but it does converge to $D(R)$ as the block length goes to infinity. Thus, if one is close enough to “asymptopia,” results about distortion rate functions are still informative. It should be noted that in general there is no closed form solution for finding the distortion curve given by (1). Furthermore, in practice, one does not know the underlying population distribution. Thus the distortion curve one actually observes will be an estimate of the finite block length distortion rate function for the empirical distribution of the sample data.

2.3 Asymptotic rate distortion theory results

This section gives some useful results from asymptotic rate distortion theory. Most of the fundamental theory is due to C.E. Shannon. The interested reader should see Cover and Thomas (1991), Chapter 13 for a more complete development including extensive references and proofs.

Theorem 1 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 $R$.

This 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 since finite block length distortion rate functions are less
well understood. Author (1999) gives a proof of convexity of the distortion curve under certain hierarchical restrictions on the clustering methodology.

In general the rate distortion function is difficult to compute. However, for multivariate normal distributions, one has the following result.

**Theorem 2** If $X$ is $p$-dimensional normal with mean vector $\mu$, and covariance structure $\sigma^2 I$, then, under squared-error distortion, the rate distortion function is

$$ R(D) = \frac{p}{2} \log_2 \frac{p\sigma^2}{D} $$

(4)

Versions of this result exist for more complex covariance structures. It is also possible to find simple bounds on the rate distortion function in a wide range of other cases. For instance, for a one-dimensional distribution one gets the following result:

**Theorem 3** If $X$ has variance $\sigma^2$ then the following are bounds on the rate distortion and distortion rate functions of $X$ based on squared error distortion:

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

$$ \frac{2^{-2R_2^{2h(X)}}}{2\pi e} \leq D(R) \leq \sigma^2 2^{-2R} $$

(5)

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

For example, for a uniform distribution, $X \sim U(a, b)$, $h(X) = \log_2 (b - a)$ and $\sigma^2 = (b - a)^2 / 3$ so 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. The form of the bounds is usually mirrored by the shape of the distortion curve in practice. Secondly, for both the normal and the general bounds of Theorem 3 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 us to transformations of the distortion curve that prove extremely valuable for identifying the true number of clusters.

## 3 The distortion curve for Gaussian clusters

In Section 3.1 we show how the results from Section 2.3 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. Based on these results we develop two methods for choosing the true number of clusters and illustrate them on simulated data in Section 3.2.

### 3.1 Asymptotic results for a mixture of Gaussian clusters

In order to properly utilize the distortion function, $d_K$, to choose the correct number of clusters we must first understand its functional forms both when the data set consists of a single cluster
Figure 2: Distortion curves for simulated data sets with (a) a single mixture component and (b) 6 mixture components.

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 by plotting 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 is approximately \( d_K \propto K^{-0.4} \). In fact Theorem 4 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 4** Suppose that \( X \) has an arbitrary \( p \)-dimensional Gaussian distribution. Let \( K = \lfloor k^p \rfloor \) where \( k \) can be any positive number. Then

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

(6)

The proof of Theorem 4 is given in Appendix A. The quantity \( k \) is essentially the \( p \)th root of the number of centers, \( K \). Hence Theorem 4 suggests that for large enough \( p \),

\[
d_K^{-p/2} \approx k^p = K
\]

(7)

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 the linearity can hold for relatively low values of \( p \). In practice we have found that this approximate relationship exists in most situations. However for low values of \( p \) the slope is generally somewhat less than the value of 1 predicted by (7). This is because \( d_K \) is actually a finite block length distortion rate function and thus must be systematically greater than the limiting distortion rate function. After transforming to a negative power this results in a lower slope. For instance the slope in Figure 2(a2) is approximately 0.83. Theorem 4 illustrates a fundamental flaw with the “intuitive” approach which involves examining the raw distortion curve for points where it levels off. Since a single Gaussian will have a distortion curve of the form \( d_K \approx K^{-2/p} \), the distortion will decline rapidly and then level off, 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 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 $K = 6$. 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 clear cut as in this example. Both the linearity for $K \geq G$ and the jump at $K = G$ occur in general. Theorem 5 gives the asymptotic form of the distortion curve for a mixture of $G$ clusters which provides an explanation for this phenomena.

**Theorem 5** Suppose that the distribution of $X$ is a mixture of $G$ Gaussian clusters with equal priors and common covariance $\Gamma_p$. 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 $\Delta$ is bounded away from zero. Furthermore for $K = \lceil k^p \rceil$

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

provided $\Delta > 6$.

The proof is given in Appendix B. The result implies that for large enough $p$ and $K < G$, $d_{K}^{-p/2} \approx 0$ while for $K \geq G$, $d_{K}^{-p/2} \propto k^p = K$. In fact the proof of Theorem 5 suggests that the constant of proportionality is $1/G$ which is consistent with (7) where $G = 1$. Thus Theorem 5 suggests

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

which explains both the jump at $K = G$ and the linearity thereafter as seen in Figure 2(bii). As with Theorem 4, even though these results are asymptotic in $p$, in practice they appear to hold even in low dimensions.

Equation (8) suggests two alternative procedures for utilizing the distortion curve to determine $G$. We call the first approach the “jump” method. It estimates $G$ using

$$\arg\max_K \left[ d_{K}^{-p/2} - d_{K-1}^{-p/2} \right],$$

the value of $K$ associated with the largest jump in the transformed distortion. The second approach is called the “broken line” method. It works 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 both methods 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 8 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 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. This ability to detect hierarchical structure in the clustering is an added benefit of this 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 but with differing degrees of separation. The data set of Figure 4(a) contains well separable 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 4. 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 a more general version of the jump method will pick \( K = 1 \) in the absence of clustering.

The theory developed in Section 3.1 assumes Gaussian data and the simulations we have examined so far deal with that situation. However, in practice the functional forms suggested by
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).

Theorems 4 and 5 also seem to be appropriate for non-Gaussian distributions. As an example consider Figure 5 which shows distortion plots for two-dimensional exponential distributions. Figure 5(a) gives the inverse distortion for a single exponential with means of one in each dimension. This curve is very close to linear with an $R^2$ of 98.8%. Figures 5(b) and (c) give the inverse distortion and corresponding jumps for a mixture of four exponential clusters. There is a clear jump at $K = 4$ followed by an approximately linear curve. These results are easily replicated for a wide variety of non-Gaussian distributions such as the uniform and $t$.

4 The distortion curve for non-Gaussian clusters

The theoretical and empirical results of Sections 3.1 and 3.2 show that the distortion curve, appropriately transformed, provides an excellent basis for choosing the correct number of Gaussian mixtures. 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

Figure 5 at the end of Section 3.2 suggests that the jump and broken line methods should provide good results for a wide range of distributions. In particular, Figure 5(b) shows that a jump at $K = G$ exists even for non-Gaussian data. Note however that the transformed distortion curve after $G$ is less strongly linear, suggesting that the broken line method may be less reliable. We
Figure 5: a) The transformed distortion curve for a two dimensional exponential distribution, b) the transformed distortion curve for a mixture of four exponential clusters and c) the corresponding jumps for each value of $K$.

We have found in general that the jump method produces some what superior results for non-Gaussian data.

Results from rate distortion theory can also be applied to non-Gaussian data. In particular Theorem 3 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 6.

**Theorem 6** Suppose that the distribution of $X$ is a mixture of $G$ $p$-dimensional clusters with equal priors. Furthermore, assume that the clusters are identically distributed with covariance $\Gamma_p$ 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

$$W = 1 - \frac{6^4 V_X}{(\Delta^2 - 36)^2}$$

where

$$V_X = \text{Var} \left( \frac{1}{p} ||X - \mu_j||^2 \big| X \text{ in } j \text{th cluster} \right).$$

Then for $1 \leq K \leq K_m$, with $G \leq K_m$, 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

$$\left( \frac{p \Delta^2 W}{9G} \right)^{-Y} + \left( W \left[ \frac{2^{2h^*(X)}}{K^2 \pi^2 e} - \left( \frac{\Delta}{6} \right)^2 (1 - W) \right] \right)^{-Y} < 2 \quad \text{and} \quad \left( \frac{p \Delta^2 W}{9G} \right)^{-Y} < 1/2$$

Furthermore, in the limit as $\Delta \to \infty$, the jump method is guaranteed to produce the correct answer
for all
\[ 0 < Y < \left[ \log_2 (K_m^2 2\pi e) - 2h^* (X) \right]^{-1}. \]  
(12)

Finally, if the dimensions are independent and \( \Delta > 6 \), the bounds on \( Y \) provided by (12) apply in the limit as \( p \to \infty \).

The proof is given in Appendix C. Theorem 6 is a powerful result because it guarantees that, provided that there is enough separation between cluster 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 (11) are overly conservative and in particular that the jump method is effective even for very small values of \( \Delta \). Interestingly, it can be shown that for Gaussian mixtures the upper bound in (12) can be replaced by infinity but this is not true for any other distribution. This is a consequence of the fact that the Gaussian distribution has maximum entropy for a given variance. This suggests that the further the cluster distribution is from Gaussian, as measured by entropy, the smaller the transformation power should be. However, it is not obvious how to choose the optimal value of \( Y \). We discuss this issue further in Sections 5 and 6. We have found that values of \( Y \) near one work well in most circumstances.

As a consequence of Theorem 6 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 \( d_0^Y = 0 \). In the absence of clustering \( (G = 1) \) and assuming the distribution of \( X \) has a finite fourth moment in each dimension, then for \( 1 \leq K \leq K_m \) the jump
\[ \left[ d_K^Y - d_{K-1}^Y \right] \]
will be maximized when \( K = 1 \) provided
\[ 0 < Y < \left[ \log_2 (K_m^2 2\pi e) - 2h^* (X) \right]^{-1}. \]  
(13)

The proof is given in Appendix D. Note that (13) is not an asymptotic result. It holds for any value of \( p \) and any distribution with finite fourth moment. Furthermore, for a single Gaussian cluster, the upper bound in (13) can be replaced by infinity. Corollary 1 proves very useful in Section 5 when we develop hypothesis tests for the presence of clustering in a data set.

### 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. However, 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 6(a) and (b) show jump plots with Y = 2/3 and Y = 1 respectively. In the first plot the maximum jump is at K = 2 with the jump at K = 3 being 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 the best choice. This is exactly the outcome we should expect. Recall that the iris data has three classes. However, Figure 6(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 unclear 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 finer levels of structure. Figure 6(d) gives the jump plot for the breast cancer data using Y = 1. It shows a sharp peak at two clusters. 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. Finally, the jump plot for the auto data with Y = 2/3, Figure 6(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.
Figure 7: Approximate 90% confidence intervals for the jumps on the a) iris data, b) breast cancer data and c) auto data.

5 Testing and validation

The results of the previous section 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 informally evaluate the certainty of these estimates. For example, Figure 6(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 possible choices for $G$. 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. Then calculate the jumps associated with this new data set and repeat the process $B$ times. We produced $B = 100$ bootstrap replicates of the jumps at each value of $K$ and used there 5th and 95th percentiles to obtain pointwise 90% confidence intervals. Figure 7 illustrates this technique for the jump plots of Figures 6(a), (d) and (e) with dashed lines denoting the confidence boundaries. Figure 7(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 7(b) provides strong evidence of two clusters in the breast cancer data. While Figure 7(c) provides convincing evidence for 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. Interesting, despite the ambiguity in Figure 7(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 $\alpha$ 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 the choice of 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 (13) 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 harming the significance level will depend on how close the cluster distributions are to Gaussian. For approximately Gaussian 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. For the examples in this section involving the iris and auto data we used $Y = 2/3$ because we were clustering on a small number of dimensions. For the breast cancer data which involved many more variables we used $Y = 1$. A rigorous approach for choosing $Y$ is a subject for continuing research. However, we have found that, in general, values near $Y = 1$ work well with a small adjustment up or down depending on the number of variables used in the clustering.

6 Discussion

Distortion curves provide a useful way of understanding clustering, especially in high dimensional settings where the data are not easily visualized. In this paper we use them to attack one of the most difficult problems in clustering, namely choosing the number of clusters. Our method involves transforming the raw distortion curve to a negative power and searching for the maximum jump. We are able to provide a strong theoretical motivation for our approach using ideas from rate distortion theory. In particular we show that, under suitable conditions, the jump method will choose the “correct” number of clusters for virtually any distribution and hence is non-parametric in nature. Despite the complexity of the underlying theory this approach is both conceptually and computationally extremely simple to implement and can be automated in a straightforward manner. Furthermore, because highly efficient algorithms exist for finding the solution to the k-means problem, our method has a distinct advantage over more computationally expensive procedures. As we showed in Sections 3 and 4 the jump method is highly successful at selecting the correct number of clusters on a wide range of practical problems. Associated tools such as confidence intervals and hypothesis tests for the presence of clustering are easy to construct and automate. However, we hope that the empirical results of this paper illustrate that the transformed distortion curve and corresponding jump plots are just as useful as exploratory tools. For example, they can be used to detect underlying hierarchical structures in clustering, as we saw with the iris and triangular nine cluster data sets.

The results presented in Sections 3 and 4 give the asymptotic functional form of the distortion
curve for mixtures of Gaussian data as well as conditions under which the jump method will be optimal for a wide range of distributions. This theory can potentially be extended in several directions. First, empirical evidence suggests that the linearity of the transformed distortion also holds even for non-Gaussian distributions and low values of $p$. The latter case corresponds to finite block length distortion curves which are less well understood than their asymptotic counterparts. However, recent advances in variable rate coding and Bennett theory (Na and Neuhoff, 1995) may prove useful for attacking this problem. Second, in practice, the requirements in (11) from Theorem 6 to guarantee the success of the jump method are overly conservative, and can probably be relaxed. Third, the theory we have developed focuses on squared error or Mahalanobis distance but the results from rate distortion theory can also be applied to many other distortion measures. For example, codes based on Hamming distance, the number of matching coordinates, have been widely studied and their properties could potentially be very useful when clustering categorical data such as genetic sequences. Finally, it would be useful to understand the connections between our approach and some of the other recently proposed methods for choosing the number of clusters, especially the gap test of Tibshirani et al. (2001) which also utilizes distortion and the Gaussian model based clustering paradigm of Frayley and Raftery (1998). In particular it should be possible to use the results of Section 3 which give the exact form of the distortion curve for Gaussian mixtures to establish a theoretical link with the latter work.

The theoretical results of this paper 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, $\Gamma$, is rarely known in practice. One solution is to estimate $\Gamma$ as part of the clustering process. Another option is to ignore $\Gamma$ 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 so we used this approach in our examples. Another important practical issue 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 curve which effectively is equivalent to assuming the data have been standardized to have zero correlation in each dimension. When squared error distortion is used and strong correlations exist between dimensions, $Y = p/2$ is too large. 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 $Y = 4/2 = 2$. 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.

This paper has concentrated on identifying the best number of groups to use in a clustering situation. In addressing this problem we have drawn links between the fields of rate distortion theory and cluster analysis. However, we believe that similar ideas can be used to connect rate distortion theory to numerous other model selection problems in statistics. In such situations a common statistical 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 mean squared error to choose the number of predictors in a standard regression setting or the penalty term in a ridge regression. Similarly a plot of cumulative variability explained is frequently used to select the optimal number of dimensions in a principal components analysis. These are special cases, more generally likelihood curves are used to choose parameters in a wide variety of modeling situations. Often the resulting “distortion” curve is monotone so choosing the global optimum fails to produce
a sensible result. Cross validation sometimes alleviates this problem but is usually computationally expensive and potentially unstable. Instead one often attempts to find a point where the curve levels off, indicating that there will be little improvement in goodness of fit associated with further increasing the parameter. This leads to the same difficulties, illustrated in Figure 1, as using the raw distortion curve to choose the number of clusters. We believe that transformations similar to those that proved so successful in the clustering context may also allow for more accurate model selection in the wide range of statistical problems that use goodness of fit measures akin to distortion.

A Proof of Theorem 4

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 \( \epsilon > 0 \) there exists \( N_\epsilon \) s.t. for all \( p > N_\epsilon \), \( |D_p(R_p) - D(R)| < \epsilon \).

First note that since \( D(R) \) is continuous there exists a \( \delta \) such that for all \( |y - R| \leq \delta \), \( |D(y) - D(R)| < \epsilon /2 \). 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 \( |D(R_p) - D(R)| < \epsilon /2 \). Therefore, since \( D_p(\cdot) \geq D(\cdot) \), we have already shown that for large enough \( p \), \( D_p(R_p) - D(R) > -\epsilon \). Now choose \( N_2 \) s.t. for all \( p > N_2 \), \( |D_p(x) - D(x)| < \epsilon /2 \). Then for all \( p > \max(N_1, N_2) \)

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

Hence \( |D_p(R_p) - D(R)| < \epsilon \).

**Proof of Theorem 4**

First note that we may assume without loss of generality that \( \Gamma = I \) so that \( d_K \) is calculated in terms of squared error because if not one can produce an identity covariance by multiplying \( X \) by \( \Gamma^{-1/2} \). Hence \( X \) 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 \( K = \left\lceil k^p \right\rceil \) 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

\[
\lim_{p \to \infty} d_K = D(\log_2 k). \tag{14}
\]

By Theorem 2, for a one dimensional normal with variance one

\[
R(D) = \frac{1}{2} \log_2 \Rightarrow D(R) = 2^{-2R}. \tag{15}
\]

Combining (14) and (15) gives

\[
\lim_{p \to \infty} d_K = 2^{-2\log_2 k} = k^{-2}
\]
B Proof of Theorem 5

First we prove a lemma.

Lemma 2

Suppose that $X$ comes from a mixture distribution of $G$ identically distributed $p$-dimensional clusters with equal priors and covariance $\Gamma$. Let $d_{K_j}$ be the average distortion per observation when allocating $K_j$ clusters to the $j$th 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 $W = 1 - \frac{6^4 V_X}{(\Delta^2 - 36)}$ and $V_X = Var \left( \frac{1}{p} ||X - \mu_j||^2 \right) \ |X \in j$th cluster).

Proof

First note that, as with Theorem 4 we can assume that $\Gamma = I$ because if not this can be achieved by taking $\Gamma^{-1/2} X$. 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.

$$ ||X_j - \mu_j||^2 = \sum_{i=1}^{p} (X_{ji} - \mu_{ji})^2 \leq p\Delta^2 / 36, \quad j = 1, \ldots, G $$

where $X_j$ 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

$$ d_K^* = \min_{\sum_j K_j = K} \left[ \sum_j d_{K_j}^* \right] $$

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

$$ d_K = P(\text{Inside sphere}) \times \text{Avg dist inside sphere} + P(\text{Outside sphere}) \times \text{Avg dist outside sphere} $$

$$ \geq P(\text{Inside sphere}) \times d_K^* $$

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

$$ P(\text{Outside sphere}) = P \left( \frac{1}{p} \sum_{i=1}^{p} (X_{ji} - \mu_{ji})^2 > \frac{\Delta^2}{36} \right) $$

$$ \leq \frac{6^4 V_X}{(\Delta^2 - 36)^2} = 1 - W \quad (\text{by Chebychev provided } \Delta > 6) $$
Finally note that for all \( j \)
\[
d_{K_j} = P(\text{Inside sphere}) \times E(d_{K_j}|\text{Inside sphere}) + P(\text{Outside sphere}) \times E(d_{K_j}|\text{Outside sphere})
\leq d_{K_j} + P(\text{Outside sphere}) \times E(d_1|\text{Outside sphere})
= d_{K_j} + \int_{\frac{\Delta^2}{36}}^{\infty} d_1 f(d_1) d(d_1)
\leq d_{K_j} + \left(\frac{\Delta}{6}\right)^2 (1 - W)
\]
The last line comes from the fact that
\[
\int_{\frac{\Delta^2}{36}}^{\infty} d_1 f(d_1) d(d_1) = \int_{\frac{\Delta^2}{36}}^{\infty} (d_1 - 1) f(d_1) d(d_1) + P(d_1 > \Delta^2/36)
\leq \frac{V_X}{\Delta^2 - 1} + \frac{V_X}{\left(\frac{\Delta^2}{36} - 1\right)^2} (by \ Chebychev)
\leq \frac{V_X}{\Delta^2 - 1} + \frac{V_X}{\left(\frac{\Delta^2}{36} - 1\right)^2} = \left(\frac{\Delta}{6}\right)^2 (1 - W)
\]
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 5**

First we consider \( K = \lceil kp \rceil \). Note that for Gaussian data \( V_X \propto 1/p \) 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
\[
\lim_{p \to \infty} \min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G} = k^{-2}. \tag{16}
\]
First we show that
\[
\lim_{p \to \infty} \min_{\sum_j K_j = K} \frac{\sum_j d_{K_j}}{G} \leq k^{-2}. \tag{17}
\]
Note that by setting \( K_j = \lceil kp/G \rceil \), \( d_{K_j} \) is a finite block length distortion rate function with rate \( R_p \to \log_2 k \). Hence by Lemma 1 and Theorem 4 \( \lim_{p \to \infty} d_{K_j} = k^{-2} \). Since this result applies for all \( j = 1, \ldots, G \) we have proven (17). 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 \( K_j = \lceil kp \rceil = K \), which is the largest \( K_j \) can be, it is still the case that \( \lim_{p \to \infty} d_{K_j} = k^{-2} \). Hence (16) is proved.

Now we consider \( K < G \). Since we are only fitting \( K < G \) centers to \( G \) clusters and the
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 clusters mass must lie within $\sqrt{p}\Delta/6$ of its mean,

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

## C Proof of Theorem 6

First note that, as with Theorem 4 we can assume that $\Gamma = I$ because if not this can be achieved by taking $\Gamma^{-1/2}X$. Consider $d_{G-1}$. By exactly the same argument as given above for Theorem 5 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

$$[d_G^{-Y} - d_{G-1}^{-Y}] \geq 1 - \left(\frac{p\Delta^2W}{9G}\right)^{-Y} \quad \text{and} \quad [d_K^{-Y} - d_{K-1}^{-Y}] \leq \left(\frac{p\Delta^2W}{9G}\right)^{-Y}, \quad K < G \quad (18)$$

Consider $d_{K_j}$, the distortion associated with the $j$th 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 (5) 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 $K \leq K_m$

$$\frac{2^{2h^*(X)}}{K_m^2 2\pi e} \leq d_{K_j} \quad (19)$$

Therefore equation (19) along with Lemma 2 imply that

$$d_K^{-Y} \leq \left(W \left[\frac{2^{2h^*(X)}}{K_m^2 2\pi e} - \left(\frac{\Delta}{6}\right)^2 (1 - W)\right]\right)^{-Y} \quad (20)$$

so from (18) and (20) the jump is maximized at $K = G$ provided (11) holds. Notice that for large enough $\Delta$ there is guaranteed to be a $Y$ that fulfills (11). Furthermore, if the dimensions of $X$ are independent from each other, for $\Delta > 6$ and large enough $p$ there is also guaranteed to be a $Y$ that fulfills (11). In fact in the limit as $\Delta$ or $p$ approach infinity (11) becomes

$$\left(\frac{K_m^2 2\pi e}{2^{2h^*(X)}}\right)^Y < 2$$

which is fulfilled provided

$$0 < Y < \left[\log_2 (K_m^2 2\pi e) - 2h^*(X)\right]^{-1}.$$
D Proof of Corollary 1

Clearly \( d_1 = 1 \) so \( d_1^Y - d_0^Y = 1 \). In this case \( \Delta = \infty \) so from (11) the jump is maximized provided

\[
0 < Y < \left[ \log_2 \left( K^2 \frac{2\pi e}{m} \right) - 2h^* (X) \right]^{-1}.
\]

References


