LETTER Communicated by Joachim Buhmann

How Many Clusters? An Information-Theoretic Perspective

Susanne Still
susanna@princeton.edu
William Bialek
wbialek@princeton.edu
Department of Physics and Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ 08544, U.S.A.

Clustering provides a common means of identifying structure in complex data, and there is renewed interest in clustering as a tool for the analysis of large data sets in many fields. A natural question is how many clusters are appropriate for the description of a given system. Traditional approaches to this problem are based on either a framework in which clusters of a particular shape are assumed as a model of the system or on a two-step procedure in which a clustering criterion determines the optimal assignments for a given number of clusters and a separate criterion measures the goodness of the classification to determine the number of clusters. In a statistical mechanics approach, clustering can be seen as a trade-off between energy- and entropy-like terms, with lower temperature driving the proliferation of clusters to provide a more detailed description of the data. For finite data sets, we expect that there is a limit to the meaningful structure that can be resolved and therefore a minimum temperature beyond which we will capture sampling noise. This suggests that correcting the clustering criterion for the bias that arises due to sampling errors will allow us to find a clustering solution at a temperature that is optimal in the sense that we capture maximal meaningful structure—without having to define an external criterion for the goodness or stability of the clustering. We show that in a general information-theoretic framework, the finite size of a data set determines an optimal temperature, and we introduce a method for finding the maximal number of clusters that can be resolved from the data in the hard clustering limit.

1 Introduction

Much of our intuition about the world around us involves the idea of clustering: many different acoustic waveforms correspond to the same syllable, many different images correspond to the same object, and so on. It is plausible that a mathematically precise notion of clustering in the space of sensory data may approximate the problems solved by our brains. Clustering methods also are used in many different scientific domains as a practical tool

to evaluate structure in complex data. Interest in clustering has increased recently because of new areas of application, such as data mining, image and speech processing, and bioinformatics. In particular, many groups have used clustering methods to analyze the results of genome-wide expression experiments, hoping to discover genes with related functions as members of the same cluster (see, e.g., Eisen, Spellman, Brown, & Botstein, 1998). A central issue in these and other applications of clustering is how many clusters provide an appropriate description of the data. The estimation of the true number of classes has been recognized as "one of the most difficult problems in cluster analysis" by Bock (1996), who gives a review of some methods that address the issue.

The goal of clustering is to group data in a meaningful way. This is achieved by optimization of a so-called clustering criterion (an objective function), and a large variety of intuitively reasonable criteria have been used in the literature (a summary is given in Gordon, 1999). Clustering methods include agglomerative clustering procedures such as described by Ward (1963) and iterative reallocation methods, such as the commonly used K-means algorithm (Lloyd, 1957; MacQueen, 1967), which reduces the sum of squares criterion, the average of the within-cluster squared distances. More recently, algorithms with physically inspired criteria were introduced (Blatt, Wiseman, & Domany 1996; Horn & Gottlieb, 2002). All of these clustering methods have in common that the number of clusters has to be found by another criterion. Often a two-step procedure is performed: the optimal partition is found for a given data set, according to the defined objective function, and then a separate criterion is applied to test the robustness of the results against noise due to finite sample size. Such procedures include the definition of an intuitively reasonable criterion for the goodness of the classification, as in Tibshirani, Walther, and Hastie (2001) or performing cross-validation (Stone, 1974) and related methods in order to estimate the prediction error and to find the number of clusters that minimizes this error (e.g., Smyth, 2000). Roth, Lange, Braun, and Buhmann (2004) quantify the goodness of the clustering with a resampling approach.

It would be attractive if these two steps could be combined in a single principle. In a sense, this is achieved in the probabilistic mixture model approach, but at the cost of assuming that the data can be described by a mixture of N_c multivariate distributions with some parameters that determine their shape. Now the problem of finding the number of clusters is a statistical model selection problem. There is a trade-off between the complexity of the model and goodness of fit. One approach to model selection is to compute the total probability that models with N_c clusters can give rise to the data, and then one finds that phase-space factors associated with the integration over model parameters serve to discriminate against more complex models (Balasubramanian, 1997). This Bayesian approach has been used to determine the number of clusters (Fraley & Raftery, 2002).

From an information-theoretic point of view, clustering is most fundamentally a strategy for lossy data compression: the data are partitioned into groups such that the data could be described in the most efficient way (in terms of bit cost) by appointing a representative to each group. The clustering of the data is achieved by compressing the original data, throwing away information that is not relevant to the analysis. While most classical approaches in statistics give an explicit definition of a similarity measure, in rate distortion theory, we arrive at a notion of similarity through a fidelity criterion implemented by a distortion function (Shannon, 1948). The choice of the distortion function provides an implicit distinction between relevant and irrelevant information in the raw data. The notion of relevance was made explicit by Tishby, Pereira, and Bialek (1999), who defined relevant information as the information that the data provide about an auxiliary variable and performed lossy compression, keeping as much relevant information as possible. This formulation, termed the information bottleneck method (IB), is attractive because the objective function follows only from information-theoretic principles. In particular, this formulation does not require an explicit definition of a measure for similarity or distortion. The trade-off between the complexity of the model, on one hand, and the amount of relevant information it captures, on the other hand, is regulated by a trade-off parameter. For a given problem, the complete range of this trade-off is meaningful, and the structure of the trade-off characterizes the "clusterability" of the data. However, for a finite data set, there should be a maximal value for this trade-off, after which we start to "overfit," and this issue has not yet been addressed in the context of the IB. In related work, Buhmann and Held (2000) derived for a particular class of histogram clustering models a lower bound on the annealing temperature from a bound on the probability of a large deviation between the error made on the training data and the expected error.

In this article, we follow the intuition that if a model—which, in this context, is a (probabilistic) partition of the data set—captures information (or structure) in the data, then we should be able to quantify this structure in a way that corrects automatically for the overfitting of finite data sets. Attempts to capture only this "corrected information" will, by definition, not be sensitive to noise. Put another way, if we would separate at the outset real structure from spurious coincidences due to undersampling, then we could fit only the real structure. In the context of information estimation from finite samples, there is a significant literature on the problem, and we argue here that a finite sample correction to information estimates is (in some limits) sufficient to achieve the "one-step" compression and clustering in the sense described above, leading us naturally to a principled method of finding the best clustering that is consistent with a finite data set.

We should point out that in general, we are not looking for the "true" number of clusters, but rather for the maximum number of clusters that can be resolved from a finite data set. This number equals the true number only

if there exists a true number of classes and if the data set is large enough to allow us to resolve them.

2 Rate Distortion Theory and the Information Bottleneck Method

If data $x \in X$ are chosen from a probability distribution $P(x)$, then a complete description of a single data point requires an average code length equal to the entropy of the distribution, $S(x) = -\sum_x P(x) \log_2[P(x)]$ bits. On the other hand, if we assign points to clusters $c \in \{1, 2, \ldots, N_c\}$, then we need at most $\log_2(N_c)$ bits. For $N_c \ll |X|$, we have $\log_2(N_c) \ll S(x)$, and our intuition is that many problems will allow substantial compression at little cost if we assign each x to a cluster c and approximate x by a representative x_c.

2.1 Rate Distortion Theory. The cost of approximating the signal x by x_c as the expected value of some distortion function, $d(x, x_c)$ (Shannon, 1948). This distortion measure can, but need not be, a metric. Lossy compression is achieved by assigning the data to clusters such that the mutual information

$$I(c; x) = \sum_{xc} P(c|x)P(x) \log_2 \left[\frac{P(c|x)}{P(c)} \right] \tag{2.1}$$

is minimized. The minimization is constrained by fixing the expected distortion,

$$\langle d(x, x_c) \rangle = \sum_{xc} P(c|x)P(x)d(x, x_c). \tag{2.2}$$

This leads to the variational problem:

$$\min_{P(c|x)} [\langle d(x, x_c) \rangle + TI(c; x)]. \tag{2.3}$$

The (formal) solution is a Boltzmann distribution,[1]

$$P(c|x) = \frac{P(c)}{Z(x; T)} \exp \left[-\frac{1}{T'} d(x, x_c) \right], \tag{2.4}$$

with the distortion playing the role of energy, and the normalization,

$$Z(x, T) = \sum_c P(c) \exp \left[-\frac{1}{T'} d(x, x_c) \right], \tag{2.5}$$

[1] $T' = T/\ln(2)$, because the information is measured in bits in equation 2.1. $P(c)$ is calculated as $P(c) = \sum_x P(c|x)P(x)$.

playing the role of a partition function (Rose, Gurewitz, & Fox, 1990). The representatives, x_c, often simply called cluster centers, are determined by the condition that all of the "forces" within each cluster balance for a test point located at the cluster center,[2]

$$\sum_x P(x|c)\frac{\partial}{\partial x_c}d(x, x_c) = 0. \tag{2.6}$$

Recall that if the distortion measure is the squared distance, $d(x, x_c) = (x - x_c)^2$, then equation 2.6 becomes $x_c = \sum_x xP(x|c)$. The cluster center is in fact the center of mass of the points assigned to the cluster.

The Lagrange parameter T regulates the trade-off between the detail we keep and the bit cost we are willing to pay. In analogy with statistical mechanics, T often is referred to as the temperature (Rose et al., 1990). T measures the softness of the cluster membership. The deterministic limit ($T \to 0$) is the limit of hard clustering solutions. As we lower T, there are phase transitions among solutions with different numbers of distinct clusters, and if we follow these transitions, we can trace out a curve of $\langle d\rangle$ versus $I(c; x)$, both evaluated at the minimum. This is the rate distortion curve and is analogous to plotting energy versus (negative) entropy, with temperature varying parametrically along the curve. Crucially, there is no optimal temperature that provides the unique best clustering, and thus there is no optimal number of clusters: more clusters always provide a more detailed description of the original data and hence allow us to achieve smaller average values of the distortion while the cost of the encoding increases.

2.2 Information Bottleneck Method. The distortion function implicitly selects the features that are relevant for the compression. However, for many problems, we know explicitly what it is that we want to keep information about while compressing the data, but one cannot always construct the distortion function that selects for these relevant features. In the information bottleneck method (Tishby, Pereira, & Bialek, 1999), the relevant information in the data is defined as information about another variable, $v \in V$. Both x and v are random variables, and we assume that we know the distribution of co-occurrences, $P(x, v)$. We wish to compress x into clusters c, such that the relevant information (the information about v) is maximally preserved. This leads directly to the optimization problem,

$$\max_{P(c|x)}[I(c; v) - TI(c; x)]. \tag{2.7}$$

[2] This condition is not independent of the original variational problem. Optimizing the objective function with respect to x_c, we find $\frac{\partial}{\partial x_c}[\langle d(x, x_c)\rangle + TI(c; x)] = 0 \Leftrightarrow \frac{\partial}{\partial x_c}\langle d(x, x_c)\rangle = p(c)\sum_x P(x|c)\frac{\partial}{\partial x_c}d(x, x_c) = 0;\ \forall c \Rightarrow$ equation 2.6.

One obtains a solution similar to equation 2.4,

$$P(c|x) = \frac{P(c)}{Z(x,T)} \exp\left[-\frac{1}{T} D_{KL}[P(v|x)\|P(v|c)]\right], \tag{2.8}$$

in which the Kullback–Leibler divergence,

$$D_{KL}[P(v|x)\|P(v|c)] = \sum_v P(v|x) \ln\left[\frac{P(v|x)}{P(v|c)}\right], \tag{2.9}$$

emerges in the place of the distortion function, providing a notion of similarity between the distributions $P(v|x)$ and $P(v|c)$, where $P(v|c)$ is given by (Tishby et al., 1999)

$$P(v|c) = \frac{1}{P(c)} \sum_x P(v|x)P(c|x)P(x). \tag{2.10}$$

When we plot $I(c;v)$ as a function of $I(c;x)$, both evaluated at the optimum, we obtain a curve similar to the rate distortion curve, the slope of which is given by the trade-off between compression and preservation of relevant information:

$$\frac{\delta I(c;v)}{\delta I(c;x)} = T. \tag{2.11}$$

3 Finite Sample Effects

The formulation above assumes that we know the probability distribution underlying the data, but in practice, we have access to only a finite number of samples, and so there are errors in the estimation of the distribution. These random errors produce a systematic error in the computation of the cost function. The idea here is to compute the error perturbatively and subtract it from the objective function. Optimization with respect to the assignment rule is now by definition insensitive to noise, and we should (for the IB) find a value for the trade-off parameter, T^*, at which the relevant information is kept maximally.

The compression problem expressed in equation 2.7 gives us the right answer if we evaluate the functional 2.7 at the true distribution $P(x,v)$. But in practice, we do not know $P(x,v)$; instead, we have to use an estimate $\hat{P}(x,v)$ based on a finite data set. We use perturbation theory to compute the systematic error in the cost function that results from the uncertainty in the estimate.

We first consider the case that $P(x)$ is known, and we have to estimate only the distribution $P(v|x)$. This is the case in many practical clustering problems, where x is merely an index to the identity of samples, and hence

$P(x)$ is constant, and the real challenge is to estimate $P(v|x)$. In section 5, we discuss the error that comes from uncertainty in $P(x)$ and also what happens when we apply this approach to rate distortion theory.

Viewed as a functional of $P(c|x)$, $I(c; x)$ can have errors arising only from uncertainty in estimating $P(x)$. Therefore, if $P(x)$ is known, then there is no bias in $I(c; x)$. We assume for simplicity that v is discrete. Let N be the total number of observations of x and v. For a given x, the (average) number of observations of v is then $NP(x)$. We assume that the estimate $\hat{P}(v|x)$ converges to the true distribution in the limit of large data set size $N \to \infty$. However, for finite N, the estimated distribution will differ from the true distribution, and there is a regime in which N is large enough such that we can approximate (compare Treves & Panzeri, 1995),

$$\hat{P}(v|x) = P(v|x) + \delta P(v|x), \tag{3.1}$$

where we assume that $\delta P(v|x)$ is some small perturbation and its average over all possible realizations of the data is zero:

$$\langle \delta P(v|x) \rangle = 0. \tag{3.2}$$

Taylor expansion of $I^{\text{emp}}(c; v) := I(c; v)|_{\hat{P}(v|x)}$ around $P(v|x)$ leads to a systematic error $\Delta I(c; v)$,

$$I(c; v)|_{\hat{P}(v|x)=P(v|x)+\delta P(v|x)} = I(c; v)|_{P(v|x)} + \Delta I(c; v), \tag{3.3}$$

where the error is calculated as an average over realizations of the data

$$\Delta I(c; v) = \sum_{n=1}^{\infty} \frac{1}{n!} \sum_{v} \sum_{x^{(1)}} \cdots \sum_{x^{(n)}} \frac{\delta^n I(c; v)}{\prod_{k=1}^{n} \delta P(v|x^{(k)})} \left\langle \prod_{k=1}^{n} \delta P(v|x^{(k)}) \right\rangle, \tag{3.4}$$

with

$$\frac{\delta^n I(c; v)}{\prod_{k=1}^{n} \delta P(v|x^{(k)})} = \frac{(-1)^n (n-2)!}{\ln 2} \left[\sum_{c} \frac{\prod_{k=1}^{n} P(c, x^{(k)})}{(P(c, v))^{n-1}} - \frac{\prod_{k=1}^{n} P(x^{(k)})}{(P(v))^{n-1}} \right] \tag{3.5}$$

is given by

$$\Delta I(c; v) = \frac{1}{\ln 2} \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \sum_{v} \times \left(\sum_{c} \frac{\langle (\sum_x \delta P(v|x) P(c, x))^n \rangle}{(P(c, v))^{n-1}} - \frac{\langle (\sum_x \delta P(v|x) P(x))^n \rangle}{(P(v))^{n-1}} \right). \tag{3.6}$$

Note that the terms with $n = 1$ vanish because of equation 3.2 and that the second term in the sum is constant with respect to $P(c|x)$.

Our idea is to subtract this error from the objective function, equation 2.7, and recompute the distribution that maximizes the corrected objective function,

$$\max_{P(c|x)} \left[I^{\rm emp}(c;v) - TI^{\rm emp}(c;x) - \Delta I(c;v) + \mu(x) \sum_c P(c|x) \right]. \tag{3.7}$$

The last constraint ensures normalization, and the optimal assignment rule $P(c|x)$ is now given by

$$P(c|x) = \frac{P(c)}{Z(x,T)} \exp\left[-\frac{1}{T} \left(D_{KL}[P(v|x)\|P(v|c)] - \sum_v \sum_{n=2}^{\infty} (-1)^n \times \left[P(v|x) \frac{\langle (\delta P(v|c))^n \rangle}{n(P(v|c))^n} - \frac{\langle \delta P(v|x)(\delta P(v|c))^{n-1} \rangle}{(n-1)(P(v|c))^{n-1}} \right] \right) \right], \tag{3.8}$$

which has to be solved self-consistently together with equation 2.10 and

$$\delta P(v|c) := \sum_x \delta P(v|x) P(x|c). \tag{3.9}$$

The error $\Delta I(c;v)$ is calculated in equation 3.6 as an asymptotic expansion, and we are assuming that N is large enough to ensure that $\delta P(v|x)$ is small $\forall (x,v)$. Let us thus concentrate on the term of leading order in $\delta P(v|x)$, which is given by (disregarding the term which does not depend on $P(c|x)$)[3]

$$(\Delta I(c;v))^{(2)} \simeq \frac{1}{2\ln(2)N} \sum_{vc} \frac{\sum_x [P(c|x)]^2 P(x|v)}{\sum_x P(c|x)P(x|v)}, \tag{3.10}$$

where we have made use of equation 2.10 and the approximation (for counting statistics)

$$\langle \delta P(v|x) \delta P(v|x') \rangle \simeq \delta_{xx'} \frac{P(v|x)}{NP(x)}. \tag{3.11}$$

[3] We arrive at equation 3.10 by calculating the first leading-order term ($n = 2$) in the sum of equation 3.6:

$$(\Delta I(c;v))^{(2)} \cong \frac{1}{2\ln(2)} \sum_{vc} \frac{\sum_x \sum_{x'} P(c,x)P(c,x') \langle \delta P(v|x) \delta P(v|x') \rangle}{P(c,v)},$$

making use of approximation 3.11 and summing over x', which leads to

$$(\Delta I(c;v))^{(2)} \simeq \frac{1}{2\ln(2)N} \sum_{vc} \frac{\sum_x [P(c|x)]^2 P(v|x)P(x)}{P(c|v)P(v)},$$

and then substituting $P(v|x)P(x)/P(v) = P(x|v)$ and $P(c|v) = \sum_x P(c|x)P(x|v)$ (compare equation 2.10).

Can we say something about the shape of the resulting "corrected" optimal information curve by analyzing the leading-order error term, equation 3.10? This term is bounded from above by the value it assumes in the deterministic limit ($T \to 0$), in which assignments $P(c|x)$ are either 1 or 0 and thus $[P(c|x)]^2 = P(c|x)$,[4]

$$(\Delta I(c;v))^{(2)}_{T\to 0} = \frac{1}{2\ln(2)} \frac{K_v}{N} N_c. \tag{3.12}$$

K_v is the number of bins we have used to obtain our estimate $\hat{P}(v|x)$. Note that if we had adopted a continuous, rather than a discrete, treatment, then the volume of the (finite) V-space would arise instead of K_v.[5] If one does not assume that $P(x)$ is known, and instead calculates the bias by Taylor expansion of $I^{\text{emp}}(c;v)$ around $P(x,v)$ (see section 5), then the upper bound to the leading order term (see equation 5.7) is given by $N_c(K_v - 1)/(2N\ln(2))$. This is the N_c-dependent part of the leading correction to the bias as derived in Treves and Panzeri (1995, equation 2.11, term C_1). Similar to what these authors found when they computed higher-order corrections, we also found in numerical experiments that the leading term is a surprisingly good estimate of the total bias, and we therefore feel confident to approximate the error by equation 3.10, although we cannot guarantee convergence of the series in equation 3.6.[6]

---

[4] Substitution of $[P(c|x)]^2 = P(c|x)$ into equation 3.10 gives $(\Delta I(c;v))^{(2)} = \frac{1}{2\ln(2)N} \sum_{vc} \frac{\sum_x P(c|x)P(x|v)}{\sum_x P(c|x)P(x|v)} = \frac{1}{2\ln(2)N} \sum_{vc} 1 = \frac{1}{2\ln(2)N} K_v N_c$.

[5] For choosing the number of bins, K_v, as a function of the data set size N, we refer to the large body of literature on this problem, for example, Hall and Hannan (1988).

[6] For counting statistics (binomial distribution), we have

$$\langle (\delta P(v|x))^n \rangle = \frac{1}{N^{n-1}} (P(v|x))^n \sum_{k=0}^{n} (-1)^{(n-k)} \frac{n!}{k!(n-k)!} \frac{N^k}{(P(v|x))^k}$$
$$\times \sum_{\{l_1 \cdots l_k\}} \frac{k!N!}{(N-l)!} (1-p)^{(N-l)} \prod_{q=1}^{k} \frac{1}{l_q!} \left(\frac{p}{q!} \right)^{l_q},$$

where $l = \sum_{q=1}^{k} l_q$, the l_q are positive integers, and the sum $\sum_{\{l_1 \cdots l_k\}}$ runs over all partitions of k, that is, values of $l_1, \ldots, l_k$ such that $\sum_{q=1}^{k} q l_q = k$. There is a growing number of contributions to the sum at each order n, some of which can be larger than the smallest terms of the expression at order $n-1$. If there are enough measurements such that $NP(x)$ is large, the binomial distribution approaches a normal distribution with $\langle (\delta P(v|x))^{2n} \rangle = (2n-1)!! \frac{1}{N^n P(x)^n} (P(v|x) - P(v|x)^2)^n$, and $\langle (\delta P(v|x))^{2n-1} \rangle = 0$ $(n = 1, 2, \ldots)$. Substituting this into equation 3.6, and considering only terms with $x^{(1)} = x^{(2)} = \cdots = x^{(n)}$, we get $\frac{1}{\ln 2} \sum_{xvc} P(c,v) \sum_{k=1}^{\infty} \frac{(2k-1)!!}{2k(2k-1)} \left[\frac{1}{N} \frac{P(x)}{P(v)} \frac{(P(c|x))^2}{(P(c|v))^2} (P(v|x) - P(v|x)^2) \right]^k$, which is not guaranteed to converge.

The lower bound of the leading-order error, equation 3.10, is given by[7]

$$\frac{1}{2\ln(2)}\frac{1}{N}2^{I(c;x)}, \tag{3.13}$$

and hence the "corrected information curve," which we define as

$$I^{\rm corr}(c;v) := I^{\rm emp}(c;v) - \Delta I(c;v), \tag{3.14}$$

is (to leading order) bounded from above by

$$I^{\rm corr}_{UB}(c;v) = I^{\rm emp}(c;v) - \frac{1}{2\ln(2)}\frac{1}{N}2^{I(c;x)}. \tag{3.15}$$

The slope of this upper bound is $T - 2^{I(c;x)}/2N$ (using equation 2.11), and there is a maximum at

$$T^*_{UB} = \frac{1}{2N}2^{I(c;x)}. \tag{3.16}$$

If the hard clustering solution assigns equal numbers of data to each cluster, then the upper bound on the error, equation 3.12, can be rewritten as

$$\frac{1}{2\ln(2)}\frac{K_v}{N}2^{I(c;x)}. \tag{3.17}$$

and, therefore the lower bound on the information curve,

$$I^{\rm corr}_{LB}(c;v) = I^{\rm emp}(c;v) - \frac{1}{2\ln(2)}\frac{K_v}{N}2^{I(c;x)}, \tag{3.18}$$

has a maximum at

$$T^*_{LB} = \frac{K_v}{2N}2^{I(c;x)}. \tag{3.19}$$

Since both upper and lower bound coincide at the end point of the curve, where $T \to 0$ (see Figure 1), the actual corrected information curve must

[7] Proof: $\sum_{xvc}\frac{[P(c|x)]^2P(x|v)}{P(c|v)} = \sum_{xc}P(x,c)\frac{P(c|x)}{P(c)}\sum_v\frac{P(v|x)}{P(v|c)} > \sum_{xc}P(x,c)\frac{P(c|x)}{P(c)} = \sum_{xc}P(x,c)2^{\log_2[\frac{P(c|x)}{P(c)}]} \geq 2^{I(c;x)}$.

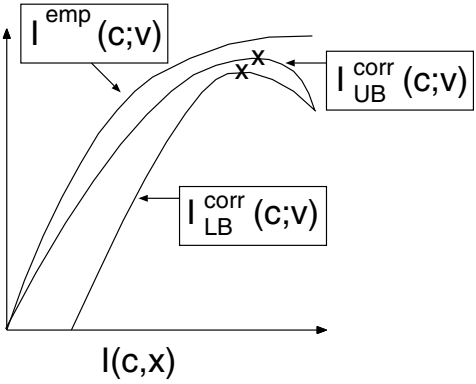

Figure 1: Sketch of the lower and upper bound on the corrected information curve, which both have a maximum under some conditions (see equations 3.16 and 3.19), indicated by x, compared to the empirical information curve, which is monotonically increasing.

have a maximum at

$$T^* = \frac{\gamma}{2N} 2^{I(c;x)}, \tag{3.20}$$

where $1 < \gamma < K_v$.

In general, for deterministic assignments, the information we gain by adding another cluster saturates for large N_c, and it is reasonable to assume that this information grows sublinearly in the number of clusters. That means that the lower bound on $I^{\text{corr}}(c; v)$ has a maximum (or at least a plateau). This ensures us that $I^{\text{corr}}(c; v)$ must have a maximum (or plateau) and, hence, that an optimal temperature exists.

In the context of the IB, asking for the number of clusters that are consistent with the uncertainty in our estimation of $P(v|x)$ makes sense only for deterministic assignments. From the above discussion, we know the leading-order error term in the deterministic limit, and we define the "corrected relevant information" in the limit $T \to 0$ as[8]

$$I^{\text{corr}}_{T\to 0}(c; v) = I^{\text{emp}}_{T\to 0}(c; v) - \frac{K_v}{2\ln(2)N} N_c, \tag{3.21}$$

where $I^{\text{emp}}_{T\to 0}(c; v)$ is calculated by fixing the number of clusters and cooling the temperature to obtain a hard clustering solution. While $I^{\text{emp}}_{T\to 0}(c; v)$ increases monotonically with N_c, we expect $I^{\text{corr}}_{T\to 0}(c; v)$ to have a maximum (or

[8] This quantity is not strictly an information anymore—thus, the quotation marks.

at least a plateau) at N_c^*, as we have argued above. N_c^* is then the optimal number of clusters in the sense that using more clusters, we would not capture more meaningful structure (or, in other words, would overfit the data), and although in principle we could always use fewer clusters, this comes at the cost of keeping less relevant information $I(c; v)$.

4 Numerical Results

4.1 Simple Synthetic Test Data. We test our method for finding N_c^* on data that we understand well and where we know what the answer should be. We thus created synthetic data drawn from normal distributions with zero mean and five different variances (for Figures 2 and 3).[9] We emphasize that we chose an example with gaussian distributions not because any of our analysis makes use of gaussian assumptions, but rather because in the gaussian case, we have a clear intuition about the similarity of different distributions and hence about the difficulty of the clustering task. This will become important later, when we make the discrimination task harder (see Figure 6). We use $K_v = 100$ bins to estimate $\hat{P}(v|x)$. In Figures 2 and 3, we compare how $I^{\rm emp}_{T\to 0}(c; v)$ and $I^{\rm corr}_{T\to 0}(c; v)$ behave as a function of the number of clusters. The number of observations of v, given x, is $N_v = N/N_x$, and the average number of observations per bin is given by N_v/K_v. Figure 3 shows the average $I^{\rm corr}_{T\to 0}(c; v)$, computed from 31 different realizations of the data.[10] All of the 31 individual curves have a maximum at the true number of clusters, $N_c^* = 5$, for $N_v/K_v \geq 2$. They are offset with respect to each other, which is the source of the error bars. When we have too few data ($N_v/K_v = 1$), then we can resolve only four clusters (65% of the individual curves peak at $N_c^* = 4$, the others at $N_c^* = 5$). As N_v/K_v becomes very large, $I^{\rm corr}_{T\to 0}(c; v)$ approaches $I^{\rm emp}_{T\to 0}(c; v)$, as expected.

The curves in Figures 2 and 3 differ in the average number of examples per bin, N_v/K_v. The classification problem becomes harder as we see fewer data. However, it also becomes harder when the true distributions are closer. To separate the two effects, we create synthetic data drawn from gaussian distributions with unit variance and N_A different, equidistant means α, which are $d\alpha$ apart.[11] N_A is the true number of clusters. This problem becomes intrinsically harder as $d\alpha$ becomes smaller. Examples are shown in Figures 4 and 5. The problem becomes easier as we are allowed to look at more data, which corresponds to an increase in N_v/K_v. We are interested

[9] $P(x) = 1/N_x$ and $P(v|x) = \mathcal{N}(0, \alpha(x))$ where $\alpha(x) \in A$, and $|A| = 5$, with $P(\alpha) = 1/5$; and $N_x = 50$. N_x is the number of "objects" we are clustering.

[10] Each time we compute $I^{\rm emp}_{T\to 0}(c; v)$, we start at 100 different, randomly chosen initial conditions to increase the probability of finding the global maximum of the objective functional.

[11] $P(v|x) = \mathcal{N}(\alpha(x), 1)$, $\alpha(x) \in A$, $N_A := |A|$.

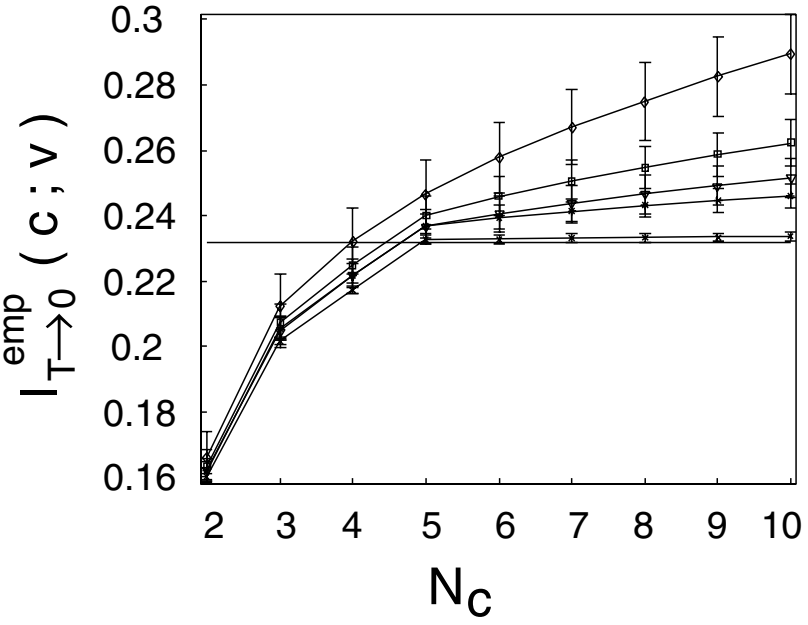

Figure 2: Result of clustering synthetic data with $P(v|x) = \mathcal{N}(0, \alpha(x))$; five possible values for α. Displayed is the relevant information kept in the compression, computed from the empirical distribution, $I^{\text{emp}}_{T\to 0}(c; v)$, which increases monotonically as a function of the number of clusters. Each curve is computed as the mean of 31 different curves, obtained by virtue of creating different realizations of the data. The error bars are $\pm$ 1 standard deviation. N_v/K_v equals 1 (diamonds), 2 (squares), 3 (triangles), 5 (stars), and 50 (X's). $N_x = 50$ and $K_v = 100$ for all curves. The line indicates the value of the information $I(x; v)$, estimated from 10^6 data points.

in the regime in the space spanned by N_v/K_v and $d\alpha$ in which our method retrieves the correct number of clusters.

In Figure 6, points mark those values of $d\alpha$ and N_v/K_v (evaluated on the shown grid) at which we find the true number of clusters. The different shapes of the points summarize results for 2, 5, and 10 clusters. A missing point on the grid indicates a value of $d\alpha$ and N_v/K_v at which we did not find the correct number of clusters. All of these missing points lie in a regime characterized by a strong overlap of the true distributions combined with scarce data. In that regime, our method always tells us that we can resolve fewer clusters than the true number of clusters. For small sample sizes, the correct number of clusters is resolved only if the clusters are well separated, but as we accumulate more data, we can recover the correct number of classes for more and more overlapping clusters. To illustrate the performance of

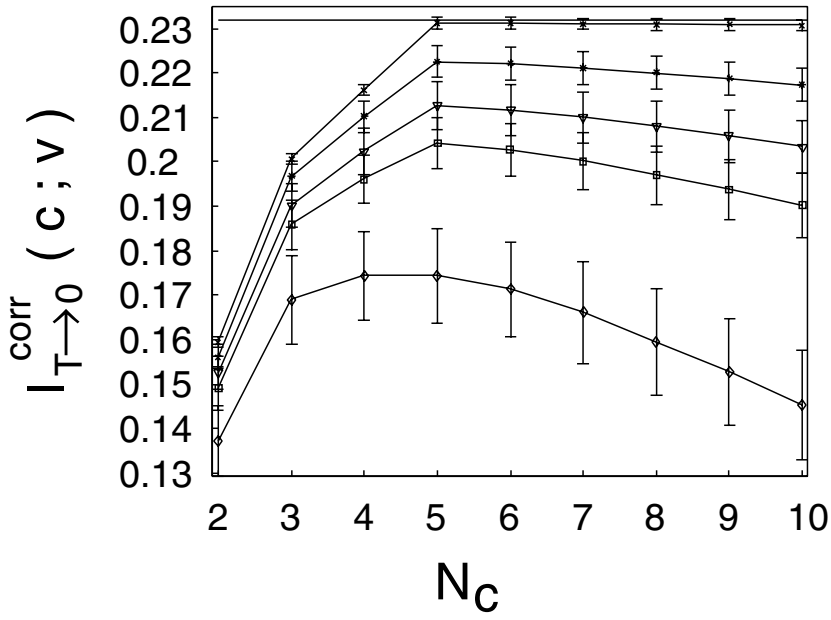

Figure 3: Result of clustering synthetic data with $P(v|x) = \mathcal{N}(0, \alpha(x))$; five possible values for α. Displayed is the "corrected relevant information" in the hard clustering limit, $I^{\rm corr}_{T\to 0}(c; v)$ (see equation 3.21), as a function of the number of clusters. Each curve is computed as the mean of 31 different curves, obtained by virtue of creating different realizations of the data. For $N_v/K_v \geq 2$, all individual 31 curves peak at $N_c^* = 5$, but are offset with respect to each other. The error bars are $\pm$ 1 standard deviation. N_v/K_v equals 1 (diamonds), 2 (squares), 3 (triangles), 5 (stars), and 50 (X's). For $N_v/K_v = 1$, 20 of the 31 curves peak at $N_c^* = 4$, the other 11 at $N_c^* = 5$. The line indicates the value of the information $I(x; v)$, estimated from 10^6 data points.

the method, we show in Figure 5 the distribution $P(x, v)$ in which $\alpha(x)$ has five different values that occur with equal probability, $P(\alpha(x)) = 1/5$ and differ by $d\alpha = 0.2$. For this separation, our method still retrieves five as the optimal number of clusters when we have at least $N_v = 2000$.[12]

Our method detects when only one cluster is present, a case in which many methods fail (Gordon, 1999). We verified this for data drawn from one gaussian distribution and for data drawn from the uniform distribution.

[12] We used $K_v = 100$ bins to estimate $P(v|x)$. $P(x) = 1/N_x$ with $N_x = 20$.

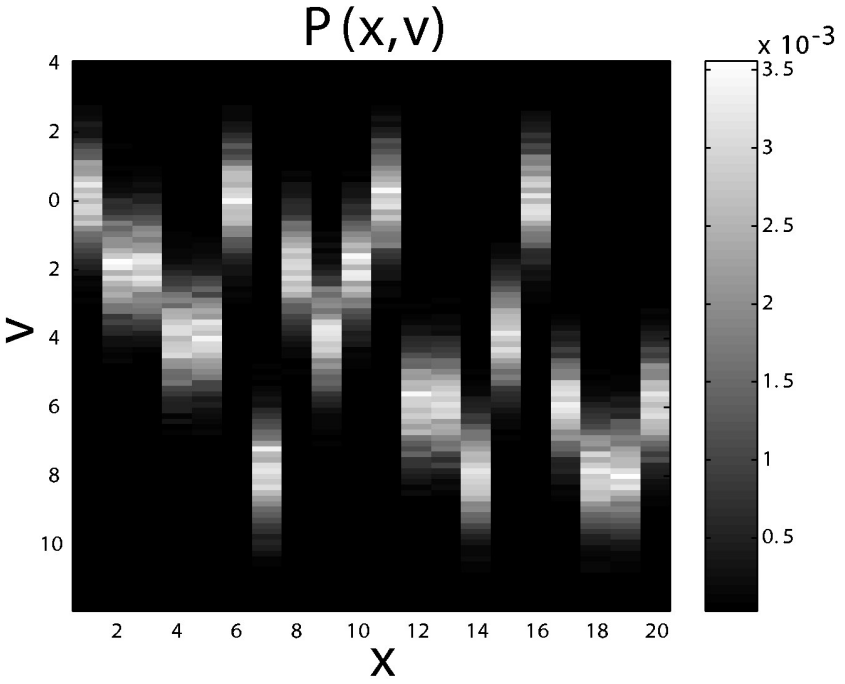

Figure 4: A trivial example of those data sets on which we found the correct number of clusters (results are summarized in Figure 6). Here, $P(v|x) = \mathcal{N}(\alpha(x), 1)$ with five different values for α, spaced $d\alpha = 2$ apart. $K_v = 100$, $N_x = 20$, $N_v/K_v = 20$.

4.2 Synthetic Test Data That Explicitly Violate Mixture Model Assumptions. We consider data drawn from a radial normal distribution, according to $P(r) = \mathcal{N}(1, 0.2)$, with $x = r\cos(\phi)$, $v = r\sin(\phi)$, and $P(\phi) = 1/2\pi$, as shown in Figure 7. The empirical information curves (see Figure 8) and corrected information curves (see Figure 9) are computed as the mean of seven different realizations of the data for different sample sizes.[13] The corrected curves peak at N_c^*, which is shown as a function of N in Figure 10. For fewer than a few thousand samples, the optimal number of clusters goes roughly as $N_c^* \propto N^{2/3}$, but there is a saturation around $N_c^* \approx 25$. This number corresponds to half of the number of x-bins (and therefore half of the number of "objects" we are trying to cluster), which makes sense given the symmetry of the problem.

[13] Each time we compute $I^{\text{emp}}_{T\to 0}(c; v)$, we start at 20 different, randomly chosen initial conditions to increase the probability of finding the global maximum of the objective functional. Increasing the number of initial conditions would decrease the error bars at the cost of computational time.

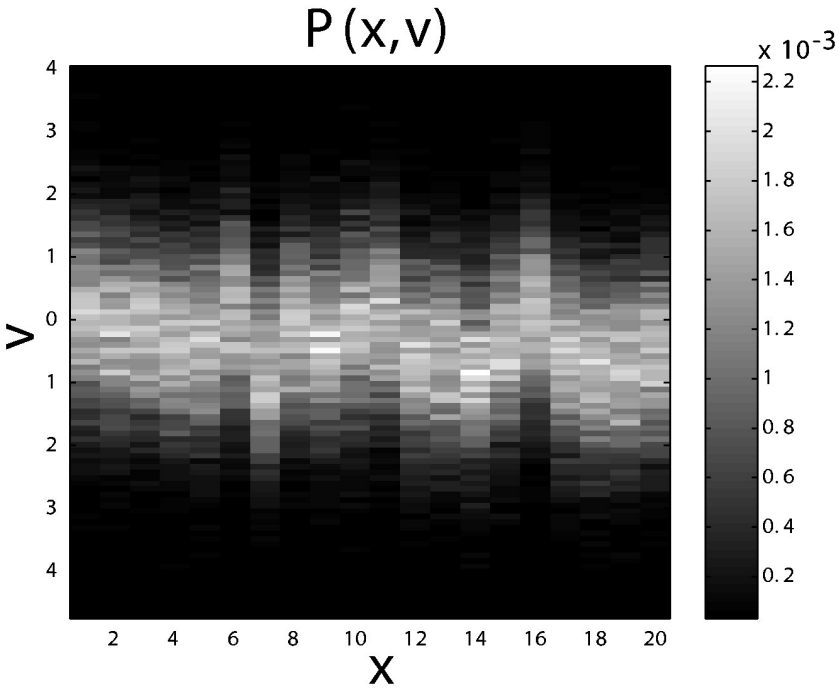

Figure 5: One of the difficult examples of those data sets on which we found the correct number of clusters (results are summarized in Figure 6). Here, $P(v|x) = \mathcal{N}(\alpha(x), 1)$ with five different values for α, spaced $d\alpha = 0.2$ apart. $K_v = 100$, $N_x = 20$, $N_v/K_v = 20$.

5 Uncertainty in $P(x)$

In the most general case, x can be a continuous variable drawn from an unknown distribution $P(x)$. We then have to estimate the full distribution $P(x, v)$, and if we want to follow the same treatment as above, we have to assume that our estimate approximates the true distribution

$$\hat{P}(v, x) = P(v, x) + \delta P(v, x), \tag{5.1}$$

where $\delta P(v, x)$ is some small perturbation and its average over all possible realizations of the data is zero:

$$\langle \delta P(v, x) \rangle = 0. \tag{5.2}$$

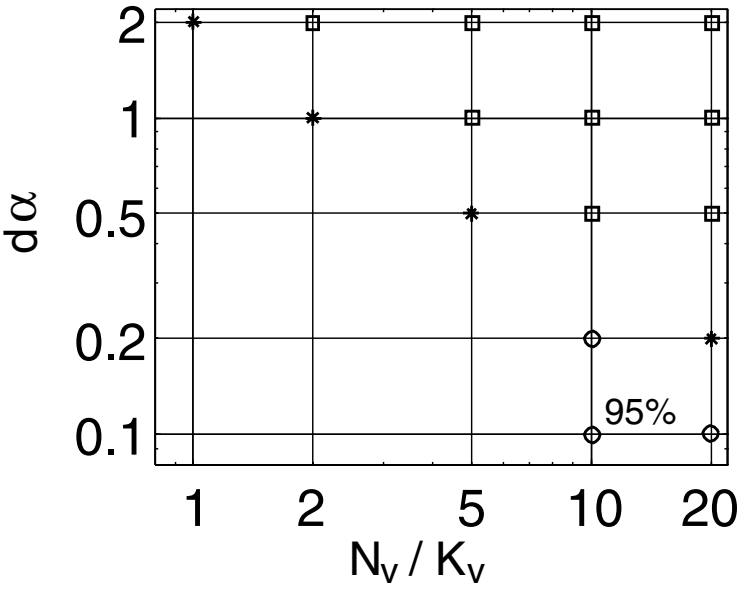

Figure 6: Result of finding the correct number of clusters with our method for a synthetic data set of size $N = N_x N_v$, ($N_x = 20$) with $P(v|x) = \mathcal{N}(\alpha(x), 1)$ and with either 2, 5, or 10 possible values for α, spaced $d\alpha$ apart. We indicate values of $d\alpha$ and the "resolution" N_v/K_v ($K_v = 100$) at which the correct number of clusters is found: for 2, 5, and 10 clusters (squares); for only 2 and 5 clusters (stars); for only 2 clusters (circles). The classification error (on the training data) is 0 for all points except for the one that is labeled with 95% correct.

Now, this estimate induces an error not only in $I^{\text{emp}}(c; v)$ but also in $I^{\text{emp}}(c; x)$. Taylor expansion of these two terms gives

$$\Delta I(c; v) = \frac{1}{\ln(2)} \sum_{vc} \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \left(\frac{1}{(P(v,c))^{n-1}} - \frac{1}{(P(c))^{n-1}} \right) \times \left\langle \left(\sum_x P(c|x) \delta P(x, v) \right)^n \right\rangle - \Lambda(P(v)) \tag{5.3}$$

$$\Lambda(P(v)) = \frac{1}{\ln(2)} \sum_{v} \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \frac{1}{(P(v))^{n-1}} \left\langle \left(\sum_x \delta P(x, v) \right)^n \right\rangle \tag{5.4}$$

$$\Delta I(c; x) = -\frac{1}{\ln(2)} \sum_{vc} \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \frac{1}{(P(c))^{n-1}} \times \left\langle \left(\sum_x P(c|x) \delta P(x, v) \right)^n \right\rangle. \tag{5.5}$$

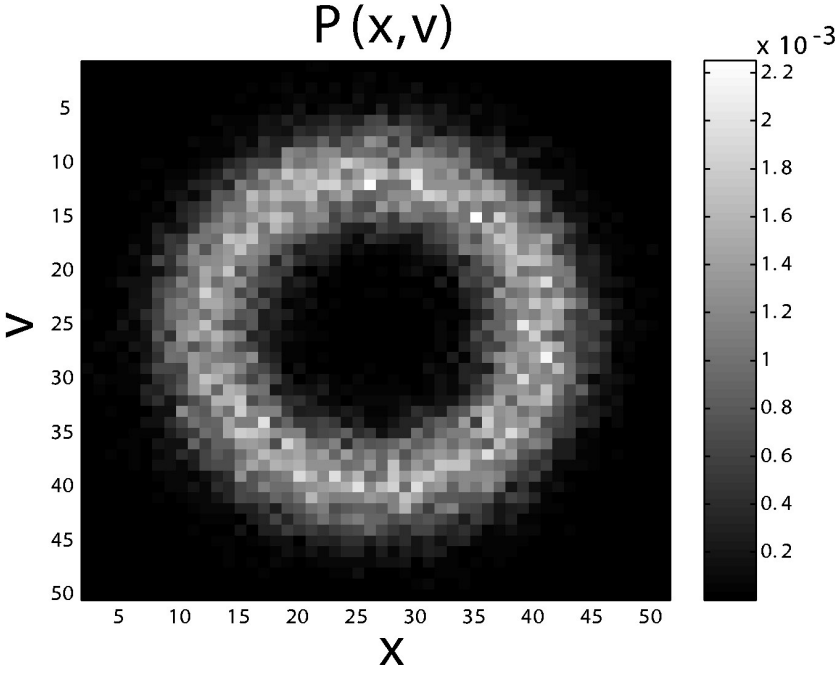

Figure 7: Twenty thousand data points drawn from a radial distribution, according to $P(r) = \mathcal{N}(1, 0.2)$, with $x = r\cos(\phi)$, $v = r\sin(\phi)$, and $P(\phi) = 1/2\pi$. Displayed is the estimated probability distribution (normalized histogram with 50 bins along each axis).

This results in a correction to the objective function ($F^{\rm corrected} = F^{\rm emp} - \Delta F$), given by:

$$\Delta F = \frac{1}{\ln(2)} \sum_{vc} \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \frac{1}{(P(c))^{n-1}} \left(\frac{1}{(P(v|c))^{n-1}} + T - 1 \right) \times \left\langle \left(\sum_x P(c|x) \delta P(x,v) \right)^n \right\rangle - \Lambda(P(v)), \tag{5.6}$$

where $\Lambda(P(v))$ is constant in $P(c|x)$ and therefore not important. At critical temperature $T = 1$, the error due to uncertainty in $P(x)$ made in calculating $I^{\rm emp}(c; v)$ cancels that made in computing $I^{\rm emp}(c; x)$. For small T, the largest contribution to the error is given by the first term in the sum of equation 5.6, since $1/(P(v|c))^n \geq 1$, $\forall\{n, v, c\}$. Therefore, the procedure that we have suggested for finding the optimal number of clusters in the deterministic limit ($T \to 0$) remains unchanged, even if $P(x)$ is unknown. To illustrate this

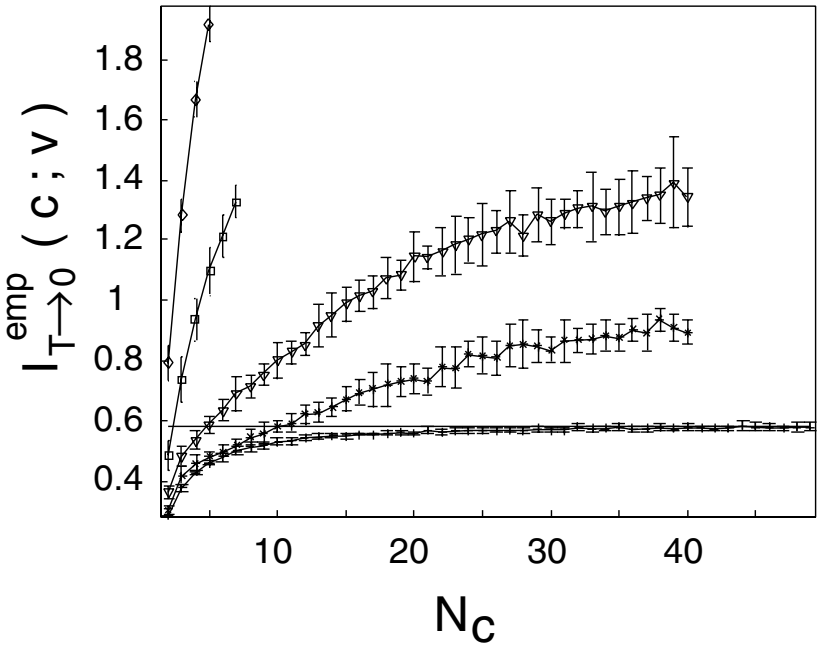

Figure 8: $I^{\text{emp}}_{T\to 0}(c; v)$ as a function of the number of clusters, averaged over seven different realizations of the data. Error bars are $\pm$ 1 standard deviation. The information $I(x; v)$, calculated from 100,000 data points, equals 0.58 bits (line). Data set size N equals 100 (diamonds), 300 (squares), 1000 (triangles), 3000 (stars), and 100,000 (crosses).

point, let us consider, as before, the leading-order term of the error (using the approximation in equation 3.11),

$$(\Delta F)^{(2)} \simeq \frac{1}{2N\ln(2)} \sum_{cv} \frac{1}{p(c)} \left(\frac{1}{p(v|c)} + T - 1 \right) \sum_x (P(c|x))^2 P(x, v). \qquad (5.7)$$

In the $T \to 0$ limit, this term becomes $N_c(K_v - 1)/2N\ln(2)$, and we find

$$I^{\text{corr}}_{T\to 0}(c; v) = I^{\text{emp}}_{T\to 0}(c; v) - \frac{K_v - 1}{2\ln(2)N} N_c, \qquad (5.8)$$

which is insignificantly different from equation 3.21 in the regime $K_v \gg 1$.

Only for very large temperatures $T \gg 1$ (i.e., at the onset of the annealing process) could the error that results from uncertainty in $P(x)$ make a significant difference.

The corrected objective function is now given by

$$F^{\text{corr}} = I^{\text{emp}}(c; v) - TI^{\text{emp}}(c; x) - \Delta F + \mu(x) \sum_c P(c|x), \qquad (5.9)$$

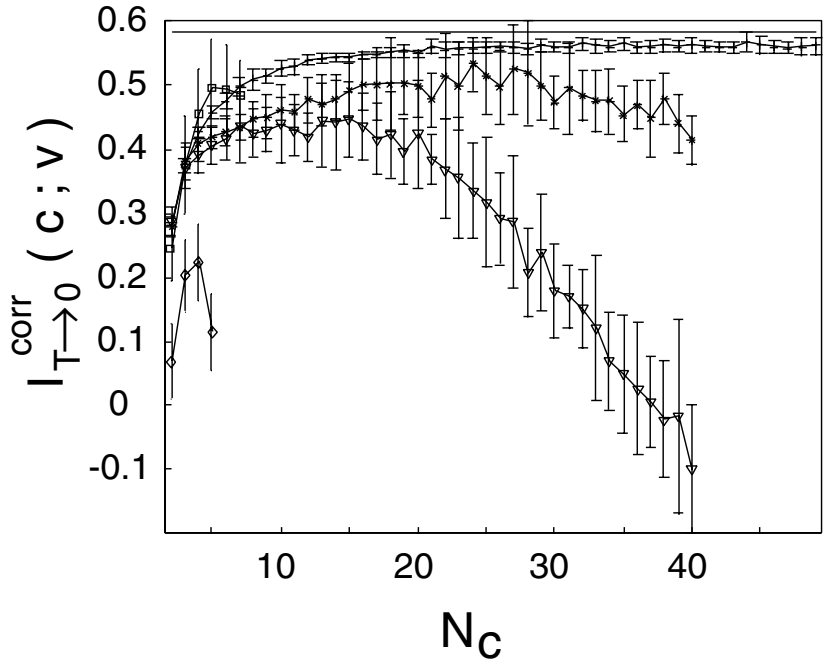

Figure 9: $I^{\mathrm{corr}}_{T\to 0}(c;v)$ as a function of the number of clusters, averaged over seven different realizations of the data. Error bars are $\pm$ 1 standard deviation. The information $I(x;v)$ calculated from 100,000 data points equals 0.58 bits (line). Data set size N equals: 100 (diamonds), 300 (squares), 1000 (triangles), 3000 (stars), and 100,000 (crosses).

and the optimal assignment rule is given by

$$P(c|x) = \frac{P(c)}{Z(x,T)} \exp\Bigg[-\frac{1}{T}\Bigg(D_{KL}\left[P(v|x)\|P(v|c)\right] - \frac{1}{P(x)} \sum_{v} \sum_{n=2}^{\infty} (-1)^n \Bigg[\frac{P(x)}{n(P(c))^n} \left(\frac{P(v|x)}{(P(v|c))^n} + T - 1 \right) \times \left\langle (\delta P(v,c))^n \right\rangle - \frac{1}{(n-1)(P(c))^{n-1}} \left(\frac{1}{(P(v|c))^{n-1}} + T - 1 \right) \times \left\langle \delta P(x,v) (\delta P(v,c))^{n-1} \right\rangle \Bigg] \Bigg) \Bigg], \tag{5.10}$$

which has to be solved self-consistently together with equation 2.10 and

$$\delta P(v,c) := \sum_{x} \delta P(v,x) P(c|x). \tag{5.11}$$

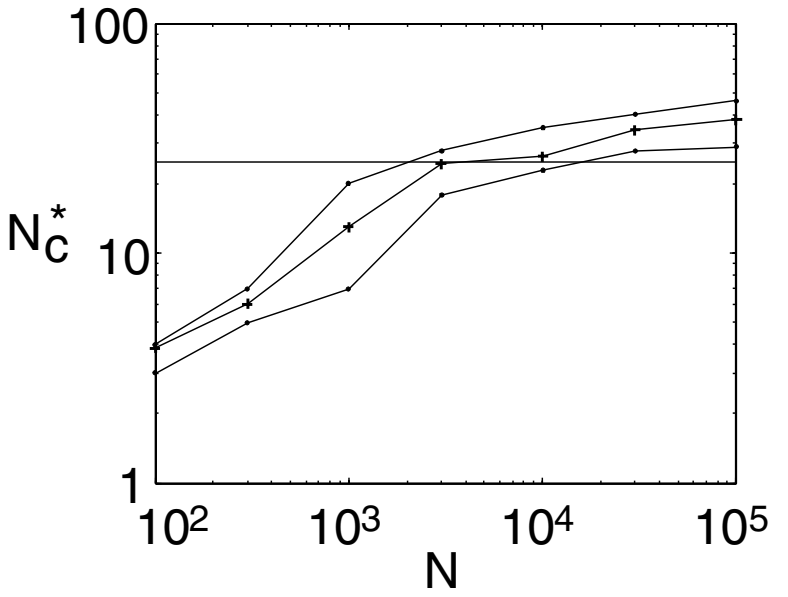

Figure 10: Optimal number of clusters, N_c^*, as found by the suggested method, as a function of the data set size N. The middle curve (crosses) represents the average over seven different realizations of the data, points on the upper/lower curve are maximum/minimum values, respectively. Line at 25.

5.1 Rate Distortion Theory. Let us assume that we estimate the distribution $P(x)$ by $\hat{P}(x) = P(x) + \delta P(x)$, with $\langle \delta P(x) \rangle = 0$, as before. While there is no systematic error in the computation of $\langle d \rangle$, this uncertainty in $P(x)$ does produce a systematic underestimation of the information cost $I(c; x)$:

$$\Delta I(c; x) = -\frac{1}{\ln(2)} \sum_{n=2}^{\infty} \frac{(-1)^n}{n(n-1)} \sum_c \frac{\langle (\sum_x P(c|x)\delta P(x))^n \rangle}{(P(c))^{n-1}}. \tag{5.12}$$

When we correct the cost functional for this error (with $\lambda = 1/T$),

$$F^{\text{corr}} := I(c; x) + \lambda \langle d(x, x_c) \rangle - \Delta I(c; x) + \mu(x) \sum_c P(c|x), \tag{5.13}$$

we obtain for the optimal assignment rule (with $\lambda' = \lambda \ln(2)$),

$$P(c|x) = \frac{P(c)}{Z(x, \lambda)} \exp \left[-\lambda' d(x, x_c) + \sum_c \sum_{n=2}^{\infty} (-1)^n \left(\frac{\langle (\sum_x P(c|x)\delta P(x))^n \rangle}{n(P(c))^n} \right.\right.$$
$$\left.\left. - \frac{1}{P(x)} \frac{\langle \delta P(x) (\sum_x P(c|x)\delta P(x))^{n-1} \rangle}{(n-1)(P(c))^{n-1}} \right) \right]. \tag{5.14}$$

Let us consider the leading-order term of the error made in calculating the information cost,

$$(\Delta I(c;x))^{(2)} = -\frac{1}{2\ln(2)} \sum_c \frac{\langle \sum_x (P(c|x)\delta P(x))^2 \rangle}{P(c)}. \tag{5.15}$$

For counting statistics, we can approximate, as before,

$$(\Delta I(c;x))^{(2)} \simeq -\frac{1}{2\ln(2)N} \sum_c \frac{\sum_x (P(c|x))^2 P(x)}{P(c)}. \tag{5.16}$$

The information cost is therefore underestimated by at least $2^{I(c;x)}/(2\ln(2)N)$ bits.[14] The corrected rate distortion curve with

$$I^{\rm corr}(c;x) := I(c;x) - \Delta I(c;x) \tag{5.17}$$

is then bounded from below by

$$I^{\rm corr}_{LB}(c;x) = I(c;x) + \frac{1}{2\ln(2)N} 2^{I(c;x)}, \tag{5.18}$$

and this bound has a rescaled slope given by

$$\tilde{\lambda} = \lambda' \left(1 - \frac{1}{2N} 2^{I(c;x)}\right), \tag{5.19}$$

but no extremum. Since there is no optimal trade-off, it is not possible to use the same arguments as we have used before to determine an optimal number of clusters in the hard clustering limit. To do this, we would have to carry the results obtained from the treatment of the finite sample size effects in the IB over to rate distortion theory, using insights we have gained in Still, Bialek, and Bottou (2004) about how to use the IB for data that are given with some measure of distance (or distortion).

6 Summary

Clustering, as a form of lossy data compression, is a trade-off between the quality and complexity of representations. In general, a data set (or clustering problem) is characterized by the whole structure of this trade-off—the rate distortion curve or the information curve in the IB method—which quantifies our intuition that some data are more clusterable than others. In

[14] Using $\sum_{xc} P(x,c) \frac{P(c|x)}{P(c)} = \sum_{xc} P(x,c) 2^{\log_2[\frac{P(c|x)}{P(c)}]} \geq 2^{I(c;x)}$.

this sense, there is never a single "best" clustering of the data, just a family of solutions evolving as a function of temperature.

As we solve the clustering problem at lower temperatures, we find solutions that reveal more and more detailed structure and hence have more distinct clusters. If we have only finite data sets, however, we expect that there is an end to the meaningful structure that can be resolved—at some point, separating clusters into smaller groups just corresponds to fitting the sampling noise. The traditional approach to this issue is to solve the clustering problem in full and then to test for significance or validity of the results by some additional statistical criteria. What we have presented in this work is, we believe, a new approach. Because clustering is formulated as an optimization problem, we can try to take account of the sampling errors and biases directly in the objective functional. In particular, for the IB method, all terms in the objective functional are mutual information, and there is a large literature on the systematic biases in information estimation. There is a perturbative regime in which these biases have a universal form and can be corrected. Applying these corrections, we find that at fixed sample size, the trade-off between complexity and quality really does have an end point beyond which lowering the temperature or increasing the number of clusters does not resolve more relevant information. We have seen numerically that in model problems, this strategy is sufficient to set the maximum number of resolvable clusters at the correct value.

Acknowledgments

We thank N. Tishby for helpful comments on an earlier draft. S. S. thanks M. Berciu and L. Bottou for helpful discussions and acknowledges support from the German Research Foundation (DFG), grant Sti197.

References

Balasubramanian, V. (1997). Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions. *Neural Computation, 9,* 349–368.

Blatt, M., Wiseman S., & Domany, E. (1996). Superparamagnetic clustering of data. *Phys. Rev. Lett., 76,* 3251–3254.

Bock, H.-H. (1996). Probability models and hypotheses testing in partitioning cluster analysis. In P. Arabie, L. J. Hubert, & G. De Soete (Eds.), *Clustering and classification* (pp. 378–453). Singapore: World Scientific.

Buhmann, J. M., & Held, M. (2000). Model selection in clustering by uniform convergence bounds. In S. A. Solla, T. K. Leen, & K.-R. Müller (Eds.), *Advances in neural information processing systems, 12*. Cambridge, MA: MIT Press.

Eisen, M., Spellman, P. T., Brown P. O., & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Nat. Acad. Sci. (USA), 95,* 14863.

Fraley, C., & Raftery, A. (2002). Model-based clustering, discriminant analysis, and density estimation. *J. Am. Stat. Assoc., 97*, 611.

Gordon, A. D. (1999). *Classification*. London: Chapman and Hall/CRC Press.

Hall, P., & Hannan, E. J. (1988). On stochastic complexity and nonparametric density estimation. *Biometrika, 75*, 705–714.

Horn D., & Gottlieb, A. (2002). Algorithm for data clustering in pattern recognition problems based on quantum mechanics. *Phys. Rev. Lett., 88*, 018702.

Lloyd, S. (1957). *Least squares quantization in PCM* (Tech. Rep.). Murray Hill, NJ: Bell Laboratories.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. M. L. Cam & J. Neyman (Eds.), *Proc. 5th Berkeley Symp. Math. Statistics and Probability* (Vol. 1, pp. 281–297). Berkeley: University of California Press.

Rose, K., Gurewitz, E., & Fox, G. C. (1990). Statistical mechanics and phase transitions in clustering. *Phys. Rev. Lett., 65*, 945.

Roth, V., Lange, T., Braun M. L., & Buhmann, J. M. (2004). Stability-based validation of clustering solutions. *Neural Computation, 16*, 1299–1323.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Tech. J., 27*, 379–423, 623–656.

Smyth, P. (2000). Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and Computing, 10*, 63.

Still, S., Bialek W., & Bottou, L. (2004). Geometric clustering using the information bottleneck method. In S. Thrun, L. Saul, & B. Schölkopf (Eds.), *Advances in neural information processing systems, 16*. Cambridge, MA: MIT Press.

Stone, M. (1974). Cross-validatory choice and assessment of statistical predictions. *J. R. Stat. Soc., 36*, 111.

Tibshirani, R., Walther, G., & Hastie, T. (2001). Estimating the number of clusters in a dataset via the Gap statistic. *J. R. Stat. Soc. B, 63*, 411.

Tishby, N., Pereira, F., & Bialek, W. (1999). The information bottleneck method. In B. Hajek & R. S. Sreenivas (Eds.), *Proc. 37th Annual Allerton Conf.* Urbana: University of Illinois.

Treves, A., & Panzeri, S. (1995). The upward bias in measures of information derived from limited data samples. *Neural Computation, 7*, 399.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. *J. Am. Stat. Assoc., 58*, 236.

Received June 17, 2003; accepted May 28, 2004.