Demystifying Information-Theoretic Clustering

Greg Ver Steeg[1,2]   GREGV@ISI.EDU
Aram Galstyan[1,2]   GALSTYAN@ISI.EDU
Fei Sha[2]   FEISHA@USC.EDU
Simon DeDeo[3,4]   SIMON.DEDEO@GMAIL.COM

[1] Information Sciences Institute, 4676 Admiralty Way, Marina del Rey, CA 90292, USA

[2] University of Southern California, Los Angeles, CA 90089, USA

[3] Santa Fe Institute, 1399 Hyde Park Rd., Santa Fe, NM 87501, USA

[4] School of Informatics and Computing, Indiana University, 901 E 10th St., Bloomington, IN 47408, USA

Abstract

We propose a novel method for clustering data which is grounded in information-theoretic principles and requires no parametric assumptions. Previous attempts to use information theory to define clusters in an assumption-free way are based on maximizing mutual information between data and cluster labels. We demonstrate that this intuition suffers from a fundamental conceptual flaw that causes clustering performance to deteriorate as the amount of data increases. Instead, we return to the axiomatic foundations of information theory to define a meaningful clustering measure based on the notion of consistency under coarse-graining for finite data.

1. Introduction

Clustering is a fundamental problem in machine learning (Jain, 2010). As the objective of clustering is typically exploratory in nature, we desire clustering algorithms that make as few assumptions about the data as possible. We would like those algorithms to be flexible enough to reveal complex data patterns that are nonlinear, multi-modal and invariant with respect to changes in data representation. Ideally, we would like to achieve those goals without explicitly defining some notion of similarity between data points (Slonim et al., 2005) or defining "prototypical" clusters (Böhm et al., 2006).

Latent variable models are a common approach to clustering. This is exemplified by the wide adoption of Gaussian mixture models and its simplified version k-means. While

maximizing likelihood of the data under some probabilistic model is a clear and operationally meaningful criteria, such models are not invariant to changes of data representation and require a fully specified generative model.

Ostensibly, information-theoretic criteria satisfy all our desiderata. There has been a growing interest in information-theoretic clustering where we assign cluster labels to data points such that the mutual information between data and labels is maximized (Faivishevsky & Goldberger, 2010; Wang & Sha, 2011; Müller et al., 2012; Sugiyama et al., 2011). Mutual information captures higher-order statistics in the data and is suitable for complex data patterns where linear approaches or Gaussian assumptions are inadequate. Moreover, mutual information is invariant with respect to smooth, invertible transformations of data (Cover & Thomas, 2006). It can also be estimated non-parametrically from the data samples without defining a parametric space of probability distributions (Kraskov et al., 2004). This is especially attractive for clustering high-dimensional data.

The main contribution of this paper is to demonstrate that information-theoretic clustering based on mutual information is fundamentally flawed and to derive a novel alternative grounded in information-theoretic principles. We derive a non-parametric estimator for our clustering objective and a tractable heuristic to optimize it. We demonstrate simple scenarios that cause mutual information clustering to fail dramatically while our method succeeds.

Mutual information is naturally interpreted as a measure of compressibility of data. However, *compression alone does not capture natural cluster structure in the data*. As we illustrate with an exemplar problem, with more information about the underlying probability distribution, mutual information clustering will prefer to split the probability mass into arbitrary but equal-sized masses of probability, completely ignoring the data's intrinsic structure!

How do we reconcile this observation with the many previously reported empirical successes of information-theoretic clustering? We argue that the good clustering performance demonstrated by those methods is not due to the objective but rather reflect transient effects of the estimators used. Thus, counterintuitively, all those methods eventually fail when given *more* data. In the large data limit, the estimators converge to their true values yielding equal-sized, but meaningless, clusters.

We show that a fix cannot be constructed by simply tweaking the information-theoretic objective. Instead, we construct an objective from first principles that preserves information theory axioms even when applied to finite samples from an unknown distribution. We motivate our approach by appeal to the axiom of consistency under coarse-graining that forms the definition of entropy (Shannon, 1948). While the consistency axiom is preserved exactly in the limit of infinite data, empirical estimation will generically lead to its violation. We shall show how larger violations signal a non-robust partition. A lower violation of consistency under coarse-graining is an important property preserved by natural cluster structures, but not by the spurious equal-sized partitions implied by mutual information. We thus propose data be clustered such that consistency is violated minimally. We construct a non-parametric estimator for this quantity and demonstrate an alternate interpretation of consistency violation as *cluster label uncertainty*.

We validate the proposed approach on synthetic data and commonly used datasets for clustering and contrast to existing approaches for information-theoretic clustering. For synthetic data, we show that our measure overcomes the shortcomings of previous information-theoretic estimators, recovering natural clusters even in the limit of large data. We also introduce a heuristic to optimize this objective and we show that it recovers non-convex clusters and achieves competitive clustering results on standard datasets.

The rest of the paper is organized as follows. In Sec. 2, we describe the basic idea of information-theoretic clustering and point out the pitfalls of the status quo which equates compression with clustering. In Sec. 3, we describe the proposed approach, starting by developing the idea of coarse-graining. In Sec. 4, we report empirical studies on applying our approach to synthetic and real-world datasets. Related work, conclusions, and future research directions are discussed in Sec. 5.

2. Information-theoretic clustering and its pitfalls

Given samples drawn i.i.d. from a *known* distribution, the Shannon entropy of the distribution can be interpreted as the minimum number of bits needed to encode each sam-

ple (on average). For clustering, we are given samples of an *unknown* distribution, and we would like to label (encode) each sample to reflect some natural structure. Even if we knew the Shannon entropy of the distribution, a code that achieves this optimal compression does not necessarily reflect the natural structure of the underlying distribution.

2.1. Basic concepts and entropy estimation

We begin with the generic clustering problem in which we are given some samples $\mathbf{x}^{(i)} \in \mathbb{R}^d$ for $i = 1, \ldots, N$, drawn from some unknown distribution, $p(\mathbf{x})$. The goal is to associate a discrete label, $y^{(i)} \in \{0, 1\}$ (for simplicity we use only binary labels), that somehow reflects a *natural* clustering of the data. Of course, the main difficulty is in defining what qualifies as a natural clustering. In what follows, we consider various information-theoretic criteria.

Entropy is defined in the usual way as $H(X) = \mathbb{E}[-\log p(\mathbf{x})]$. We use base-two logarithms so that information is measured in bits. Using standard notation, capital X denotes a random variable whose instances are denoted in lowercase, and the fact that entropy is a functional of the probability distribution is only explicitly written when clarity demands it. The expectation value may be a sum for discrete variables or an integral in the continuous case. Higher-order entropies can be constructed in various ways from this standard definition. For reference, we provide a few alternate forms of the mutual information, $I(X;Y)$.

$$\begin{aligned} I(X;Y) &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

A function that estimates entropy from some i.i.d. samples drawn from $p(\mathbf{x})$ we denote with $\hat{H}(X)$. An intuitive way to estimate entropy directly from samples in a non-parametric way follows.

$$\begin{aligned} H(X) = \mathbb{E}(\log \frac{1}{p(\mathbf{x})}) &\approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{1}{p(\mathbf{x}^{(i)})} \quad (1) \\ &\approx \frac{1}{N} \sum_{i=1}^{N} \log \frac{\epsilon_{i,k}^d}{k/N} \end{aligned}$$

$$\hat{H}(X) \equiv \log(N/k) + \frac{d}{N} \sum_{i=1}^{N} \log \epsilon_{i,k} + c_{k,N} \quad (2)$$

In the first line, we take the sample mean instead of the expectation value, but we still have the (unknown) density, $p(\mathbf{x}^{(i)})$, in the expression. On the next line, we locally approximate this density by making the smallest box that contains k neighboring points. Then the density is approximated by k/N, the fraction of points in the box,

over the volume, $\epsilon_{i,k}^d$, where $\epsilon_{i,k}$ denotes the distance to the k-th nearest neighbor of point $\mathbf{x}^{(i)}$ according to the max-norm. In our definition in the second line, we add a small constant factor $c_{k,N} = \psi(N) - \psi(k) + \log(2k/N)$ to match with the non-parametric, asymptotically unbiased Kozachenko-Leonenko entropy estimator (Kozachenko & Leonenko, 1987) (or k-nearest neighbor, or kNN, estimator), which requires a more involved derivation as provided by Kraskov et. al. (Kraskov et al., 2004). We write it in this alternate format to increase intuition and to ease later derivations. Note that this estimator depends only on distances between neighboring data points. The estimator has also been empirically shown to have good performance for small amounts of data, with $k = 3$ representing a good choice (Kraskov et al., 2004; Khan et al., 2007). For discrete variables, we use the standard plug-in estimator. DeDeo et. al. discuss more nuanced alternatives in the discrete case (DeDeo et al., 2013).

2.2. Pitfalls

Compression $\neq$ clustering A frequently invoked and plausible sounding intuition is that we should maximize mutual information between data and cluster labels. Consider first the simple case of clustering discretely distributed data, exemplified in Fig. 1(a). Is there a purely information-theoretic criteria that clusters the bins of this discrete probability distribution into the two natural groups separated by the gap? There cannot be, because from a purely information-theoretic perspective the bin labels are arbitrary. That is, the bins can be re-ordered arbitrarily so that, e.g., there is no gap, and this re-ordering does not affect any information-theoretic quantity because they depend only on the values p_i. While no one would attempt to cluster these categorical variables without defining some relationship between the variables, the same issue arises in a more subtle form for continuous distributions.

In the simplest picture of clustering for continuous variables, we have a mixture of two uniform probability distributions in one dimension, shown in Fig. 1(b). If we split the two pieces according to the intuitive clustering, a simple analytic calculation gives $I(X;Y) = H_0(\alpha/(\alpha + \beta))$, where H_0 represents the binary entropy function which is in the range $[0,1]$. If we split the space into two equally sized masses of probability, we maximize the mutual information, $I(X;\bar{Y}) = 1$. Clearly, maximizing mutual information does not have the intended effect. Even slightly unbalanced clusters will not be found. Essentially, the scenario in Fig. 1(b) is the limit of the case in Fig. 1(a) with an infinite number of bins that are infinitely narrow (Cover & Thomas, 2006). These infinitesimal bins can be re-ordered arbitrarily without affecting the value of any information-theoretic (IT) measure.
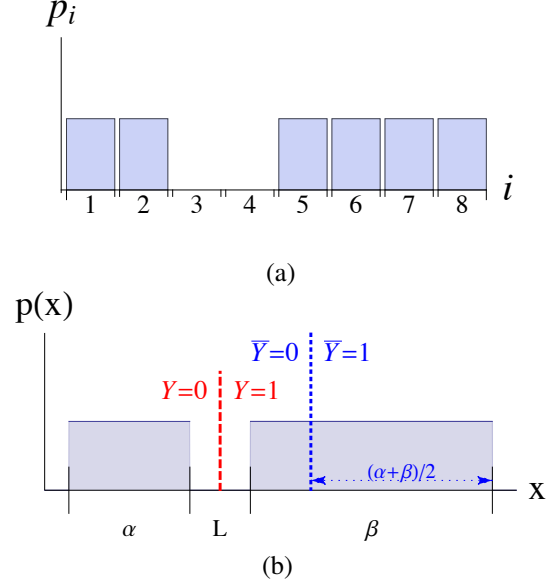

Figure 1. (a) A discrete clustering problem. Because the bin numbers are arbitrary, no purely information-theoretic criteria can recover the intuitive clustering that splits the bins based on the gap. (b) The continuous version can be viewed as the limiting case of the discrete picture above with infinitely precise bin widths. The "clustering" which maximizes the mutual information between cluster labels and x is denoted with $\bar{Y}$, leading to $I(X;\bar{Y}) = 1$.

The mystery about the empirical success Although this example makes it clear that a good clustering is not the one that maximizes mutual information, or any solely IT criteria, this leads to a mystery. How have so many papers achieved good clustering performance using this criteria (Wang & Sha, 2011; Faivishevsky & Goldberger, 2010; Müller et al., 2012)? To understand this result, it helps to write mutual information in the form $I(X;Y) = H(Y) - H(Y|X)$. The first term, $H(Y)$ is maximized for equally sized clusters. The second term, $H(Y|X)$, should be 0 for any exact partitioning of the input space (in which case $y = f(\mathbf{x})$).

However, it is easy to see that if we cluster a finite sample of data points, as in Fig. 2(b), that our *estimate*, $\hat{H}(Y|X)$, will not be zero (using, e.g., the non-parametric estimator introduced in Sec. 2.1). In particular, this will be the case near the boundary between the two clusters. The natural clustering will have a smaller value for $\hat{H}(Y|X)$ (due to the gap of width L separating the clusters). On the other hand, $\hat{H}(Y)$ will be larger for clusters of equal size. These two terms compete. For small amounts of data, we can see in Fig. 2(a) that the natural clustering is preferred, i.e. has higher estimated mutual information. However, as the amount of data increases, the uncertainty decreases and eventually equal-sized clusters will be preferred. Ironically, more data leads to a less desirable result. This behavior does not depend on the estimator used: any estimator

that is asymptotically unbiased will converge to the same value in the large N limit. The behavior also persists for arbitrary distributions because the contribution of $\hat{H}(Y|X)$ comes from points near the boundary. For any clusters, the percentage of points near the boundary will decrease as N increases. Tests with previous information-theoretic clustering objectives focused on small, nearly balanced datasets (like many of the UCI datasets considered in Sec. 4), so that these shortcomings went unnoticed.
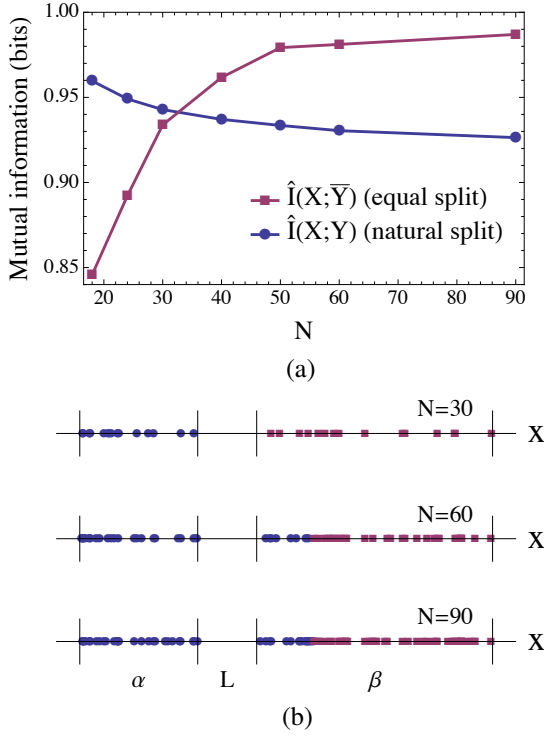

Figure 2. (a) We compare the estimated mutual information of the "natural" clustering with the one that splits the probability mass equally. We used the distribution in Fig. 1(b) with $\alpha = 1, L = 0.5, \beta = 2$. As N, the number of samples, increases, eventually the equal split becomes the preferred one (maximizing estimated mutual information). (b) Clustered samples at various N.

Precision is both the problem and the solution. Increasing N reduces the uncertainty in cluster labels near the cluster boundaries, so that any clustering has approximately the same value of $\hat{H}(Y|X)$. On the other hand, given a sample of N points, we can always resample to include fewer points. Then we can test whether a clustering is more natural in the sense that it reduces the uncertainty in the cluster labels, even for small amounts of data. We make this intuition precise in the next section.

3. Proposed approach

We saw that previous information-theoretic clustering schemes only work accidentally insofar as they mis-

estimate the true uncertainty of cluster labels near the cluster boundaries because of finite data. This behavior is closely linked to the axioms of information theory. We briefly make this connection before deriving a simple expression that explicitly makes use of the uncertainty near cluster boundaries due to limited data.

3.1. Shannon's Axiom of Consistency under Coarse-Graining

Shannon's original derivation of entropy (Shannon, 1948) begins with properties that a measure of uncertainty should be expected to obey and concludes that (Shannon) entropy is the unique measure that satisfies these properties. Consider a discrete probability distribution where outcome j occurs with probability $p_j \equiv p(y = j)$. The first two desired properties are continuity (a small change in p_j does not cause a large change in uncertainty) and that uncertainty should grow with k if there are k equally likely events. The final property we refer to as *consistency under coarse-graining*.

$$h(p_1, p_2, p_3) = h(p_1, p_2 + p_3)+$$
$$(p_2 + p_3)\, h(p_2/(p_2 + p_3), p_3/(p_2 + p_3)).$$

Intuitively, the measure of uncertainty should not change if we lump together (coarse-grain) bins 2 and 3. After combining the uncertainty of the coarse-grained bins with the (weighted) uncertainty within each coarse-grained bin we should recover the original measure of uncertainty. DeDeo et. al. make the further point that any *estimator* of entropy should at least approximately satisfy this property or it risks losing its meaning as an information-theoretic measure altogether (DeDeo et al., 2013).

What is the natural analogue of the coarse-graining criteria for continuous distributions? Suppose we split our space into two disjoint regions $\mathcal{R}_0, \mathcal{R}_1$ so that $\mathcal{R}_1 = \mathbb{R}^d \setminus \mathcal{R}_0$. For some continuous distribution $p(\mathbf{x})$, we can define a joint probability $p(\mathbf{x}, y)$, where $y = j \leftrightarrow \mathbf{x} \in \mathcal{R}_j$, or equivalently, $y = f(\mathbf{x}), f : \mathbb{R}^d \to \{0, 1\}$. Then the probability of drawing a point in region j is just $\int_{\mathcal{R}_j} dx\, p(\mathbf{x}) = \int dx\, p(\mathbf{x}, y = j) = p(y = j)$. Consistency under coarse-graining would demand,

$$H(p(\mathbf{x})) = H(p(y))+ \tag{3}$$
$$p(y = 0)H(p(\mathbf{x}|y = 0)) + p(y = 1)H(p(\mathbf{x}|y = 1)),$$

which is easily shown to be satisfied for differential entropy. This can be written in the more succinct, standard notation as $H(X) = H(Y) + H(X|Y)$[1].

Following the logic for discrete entropy estimators, we would like to check if differential entropy estimators based

---

[1]Note that $H(X|Y) \equiv \sum_i p(y = i)H(p(x|y = i))$, and that this condition should only hold if $y = f(\mathbf{x})$
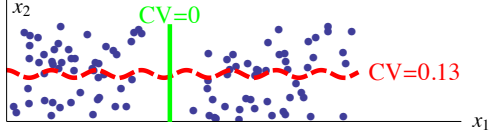
Figure 3. Splitting the space using the dashed line is one (unnatural) way of coarse-graining. The solid line represents a coarse-graining that is more natural as quantified in Eq. 5.

on finite data obey some notion of consistency under coarse-graining. It is easily shown that no estimator can satisfy this condition for *arbitrary* coarse-grainings and remain asymptotically unbiased.[2] Since we cannot construct an asymptotically unbiased entropy estimator that is consistent under arbitrary coarse-graining, instead we start with an unbiased estimator and search for coarse-grainings that lead to consistent entropy estimates. We refer to these coarse-grainings as *natural*. In principle, this argument can be applied to any entropy estimator, but we choose to focus on non-parametric estimators in keeping with our goal to minimize assumptions about the data.

Given samples, $\mathbf{x}^{(i)}$, from a continuous distribution, $p(\mathbf{x})$, a coarse-graining (clustering) can be defined by associating a discrete label $y^{(i)}$ with each sample point. We can then quantify the *consistency violation* (CV) with respect to Eq. 3, where $CV = 0$ if the equality is exactly satisfied.

$$CV = \hat{H}(Y) + \hat{H}(X|Y) - \hat{H}(X) \qquad (4)$$

Here $\hat{H}(X), \hat{H}(X|Y)$ are just defined by a differential entropy estimator like the one in Eq. 1. Fig. 3 gives an example of two different ways of coarse-graining the same data. The *natural* coarse-graining which separates well-defined clusters (solid line) produces a small consistency violation, CV, while the alternate coarse-graining (dashed line) leads to large violations. CV can be viewed as a measure of how well we can estimate the global entropy given the entropy of clusters of data points (i.e. the coarse-grained data). For well-separated clusters, the global entropy is just a weighted average of the cluster entropies. This observation provides some extra theoretical motivation for the clustering objective we propose next.

The right-hand side of Eq. 4 looks familiar as an alternate form for writing $H(Y|X)$. We can simply define an estimator for the uncertainty of the cluster label given a sample point, $\hat{H}(Y|X)$, in terms of our previously defined estima-

tors: $\hat{H}(Y|X) \equiv \hat{H}(Y) + \hat{H}(X|Y) - \hat{H}(X)$. Our estimate of the uncertainty of cluster labels is exactly the same as the violation of the coarse-graining axiom due to limited data. This gives us two interpretations of our clustering objective. We want our coarse-graining to be natural in the sense that we do not violate information-theoretic axioms, even for small amounts of data. Equivalently, we want our estimated uncertainty about cluster labels to be as low as possible, even for small amounts of data.

3.2. Using conditional entropy for clustering

Nonparametric estimation of conditional entropy We first derive a more compelling form for Eq. 4 in Sec. A of the supplementary material. For each sample point, $\mathbf{x}^{(i)}$, we associate a discrete cluster label, $y^{(i)}$, then,

$$\hat{H}_k(Y|X) = \frac{d}{N} \sum_{i=1}^{N} \log \frac{\bar{\epsilon}_{i,k}}{\epsilon_{i,k}}, \qquad (5)$$

where $\epsilon_{i,k}$ represents the distance to the k-th nearest neighbor while $\bar{\epsilon}_{i,k}$ is the distance to the k-nn restricted to points in the same cluster as sample i.[3] As long as each point's k nearest neighbors lie within the same cluster, the consistency violation, or estimated uncertainty about the cluster label, will be zero, as in Fig. 3 where we used $k = 3$.

Total cluster label uncertainty For any fixed partitioning of the space, CV approaches zero for large N. We want our cluster label uncertainty to be small even under arbitrary resampling with limited data, so we now imagine that each point is kept with random, independent probability $1 - \alpha$. We can estimate the expected value of Eq. 5 for any α, $\hat{H}_{\alpha,k}(Y|X) \equiv \mathbb{E}_\alpha[\hat{H}_k(Y|X)]$, where the expectation is over all possible resamplings. Toward this end, we note that a point's k-th nearest neighbor in the resampled dataset corresponds to its m-th nearest neighbor ($m \geq k$) in the original dataset with probability $q(m|k) = C_{m-1}^{k-1}(1 - \alpha)^k \alpha^{m-k}$, yielding,

$$\hat{H}_{\alpha,k}(Y|X) = \qquad (6)$$
$$\frac{d}{N} \sum_{i=1}^{N} \sum_{m=k}^{N-1} C_{m-1}^{k-1}(1 - \alpha)^k \alpha^{m-k} \log \frac{\bar{\epsilon}_{i,m}}{\epsilon_{i,m}}.$$

If we want $\hat{H}_{\alpha,k}(Y|X)$ to be small for all α, we can just consider its value integrated over all α. We refer to this quantity as the *total cluster label uncertainty* or *total consistency violation*. After performing an elementary integra-

---

[2]Imagine we have N samples from an unknown distribution, $p(\mathbf{x})$. We can randomly partition the N samples into two equal size sets and define regions $\mathcal{R}_0, \mathcal{R}_1$ accordingly to contain both sets. Call the entropy estimates for all N points and for a random subset of $N/2$ points $\hat{H}_N(X), \hat{H}_{N/2}(X)$, respectively. In the large N limit, we expect both of these quantities to converge to the true entropy for an asymptotically unbiased estimator. On the other hand, consistency (Eq. 3) would require $\hat{H}_N(X) = 1 + \hat{H}_{N/2}(X)$, a clear conflict.

[3]An ambiguity in this definition arises if there are k or fewer points in a cluster. Let n_y represent the number of points with the cluster label y. Then if $k > n_{y^{(i)}} + 1$ we define $\bar{\epsilon}_{i,k} = \epsilon_{i,N-1}$. This definition reflects the fact that our uncertainty is maximal if we are not given sufficient data. See Appendix A for more details.

tion and changing the summation variable we obtain

$$\hat{H}_{T,k}(Y|X) = \int_0^1 d\alpha \; \hat{H}_{\alpha,k}(Y|X) = \tag{7}$$

$$\frac{d}{N} \sum_{i=1}^{N} \sum_{l=1}^{N-k} \frac{k}{(l+k)(l+k-1)} \log \frac{\bar{\epsilon}_{i,l+k-1}}{\epsilon_{i,l+k-1}}.$$

For entropy estimators, $k = 1$ leads to the lowest bias but higher values of k are often chosen to reduce variance (Khan et al., 2007). Because averaging over resamplings already reduces variance, we choose $k = 1$, simplifying the expression further, and we will refer to this quantity succinctly as $\hat{H}_T(Y|X)$.[4]

Although this expression looks simple, $\epsilon_{i,l}$ is a function of all the $\mathbf{x}^{(j)}$ and $\bar{\epsilon}_{i,l}$ is a function of all the $\mathbf{x}^{(j)}, y^{(j)}$. In principle, this quantity should vary between 0 (a perfect clustering) and $H(Y)$ (a completely random clustering). We can search for partitions that minimize the ratio of $\hat{H}_T(Y|X)/\hat{H}(Y)$ so that we can objectively compare clusterings on a scale of zero to one. We will see in the examples that natural partitions have low CV while most partitions have a CV orders of magnitude larger. Unlike mutual information estimators, this quantity distinguishes natural cluster structure even in the large N limit.

Numerical procedure   While the focus of this work is on deriving a principled approach to information-theoretic clustering, we should briefly mention some practical concerns. Optimizing Eq. 7 over all possible partitions is a difficult problem. We consider a heuristic approach which involves solving a tractable semidefinite program in Appendix B of the supplementary material. We note that other information-theoretic approaches also require heuristic approaches to optimize (Wang & Sha, 2011). We leave a detailed exploration of the best heuristics for this optimization, along with extensive comparisons to other clustering objectives, for future work. In the next section, we consider simple clustering scenarios where we can compare all partitions to develop intuition about the meaning of this objective. We briefly compare results from our heuristic solver to other clustering methods for some standard datasets.

4. Results

4.1. Synthetic datasets

Our goal is to find clusterings, determined by $y^{(i)}$, that optimize the objective in Eq. 7. Recall the simple example in Sec. 2 which mutual information failed to cluster correctly for large N. For data taken from the simple one-

dimensional distribution in Fig. 2, we can just measure the ratio $\hat{H}_T(Y|X)/\hat{H}(Y)$ (CVR for short from here on) for all possible ways of splitting the x axis, shown in Fig. 4. We discard the trivial coarse-graining where all points are in the same group which has an undefined CVR of $0/0$, and then the best partition exactly corresponds to our intuition of the most natural clustering. We also see from Fig. 4 that more data leads to a better separation between good and bad clusterings, as desired.
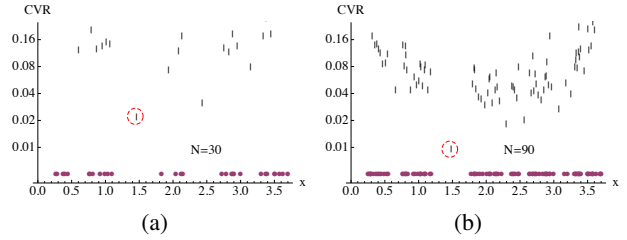

Figure 4. The vertical bars measure the consistency violation ratio (CVR) for partitioning the data at each value of x. The samples, $x^{(i)}$, are plotted near the x axis. In this case, $\alpha = 1, \beta = 2, L = 0.5$ (see Fig. 2(b)) and (a)$N = 30$, (b)$N = 90$. Unlike the mutual information objective in Fig. 2, adding more data does not make the clusters harder to distinguish.

For a less trivial example that highlights the shortcomings of other information-theoretic clustering criteria, we consider points drawn from the distribution in Fig. 5. The clusters in this case are non-convex and unbalanced in size. Let Y_r denote a cluster label that distinguishes whether a point is at radius greater than r or not. We evaluate the quality of this partitioning as a function of r and N using CVR, the mutual information (MI) using the estimator in Eq. 1, and the mutual information inspired objective NIC (Faivishevsky & Goldberger, 2010). For each objective, we label $r^* = \arg \mathrm{opt}_r \, \mathrm{Objective}(Y_r, X)$. The results are shown in Fig. 6. Note that the mutual information estimator we use is asymptotically unbiased (Kraskov et al., 2004), and so we expect a similar result for any mutual information estimator. We see that mutual information converges to an incorrect solution that equally splits the probability mass. Also note that while NIC is meant to approximate mutual information, it actually converges to a different incorrect value because it is not an unbiased estimator (see note in Sec. 5). We see that CVR performs best for all values of N and converges quickly to the correct solution.

Fig. 7 shows that these results are robust to parameter changes, though NIC, unlike MI, converges to correct solutions if the cluster size imbalance is small. CVR is the only objective to prefer the correct partition over a wide range of parameter values. While in this simple example, we only considered partitions defined by Y_r, we show in Sec. B that our heuristic optimizer is able to search over all possible partitions to recover the correct, non-convex clusters.
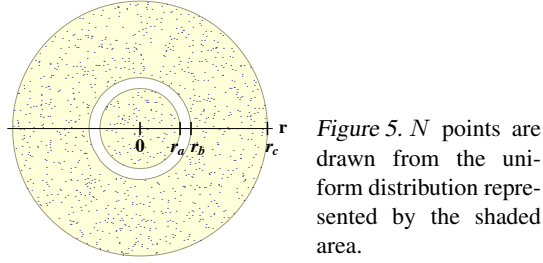
---

[4]For $k = 1$, the log-distance of the l-th nearest neighbor is weighted by a factor of $1/(l(l + 1))$. For comparison, the NIC objective (Faivishevsky & Goldberger, 2010) weights all the log-distances equally. See Sec. 5 for a more detailed comparison.

Figure 5. N points are drawn from the uniform distribution represented by the shaded area.
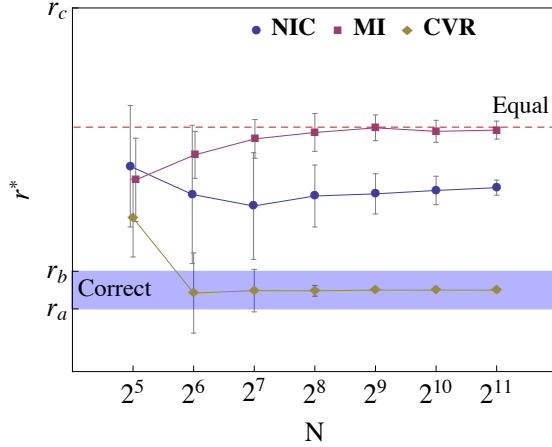

Figure 6. The ideal radius to partition the data is plotted for several information-theoretic criteria as a function of N. For each point, we calculate r^* from 100 random datasets and we plot the sample mean and standard deviation. In the plotted example, $r_a, r_b, r_c = 1.1, 1.4, 3.5$.

4.2. Real-world datasets

Editor activity on Wikipedia In Fig. 8, we consider the behavior of users who frequently edit the Wikipedia page for George W. Bush (DeDeo, 2012). On Wikipedia, users may contest any article's point of view by directly editing the text. When a change is made by some user, any other user can choose to reject the change by "reverting" to a prior version. Certain types of users are likely to get into turf wars or, conversely, to engage in pro-social vandalism repair, so that a large fraction of their activity consists of "reverts". We look at the fraction of a user's activities that does not consist of reverting to previous versions, f_{nr}. For the 50 most active users on the George W. Bush Wikipedia page, we plot f_{nr} in Fig. 8. Both CVR and Bayesian latent variable models like k-means discover the same natural cluster structure when used to determine a binary partition.

UCI datasets We also consider three standard clustering datasets from the UCI Machine Learning database (Bache & Lichman, 2013): *glass*, *iris*, and *wine*. Although we cannot optimize the objective in Eq. 7 exactly, we describe details of a heuristic approach in Appendix B. The results are summarized in Table 1. We achieve competitive results to
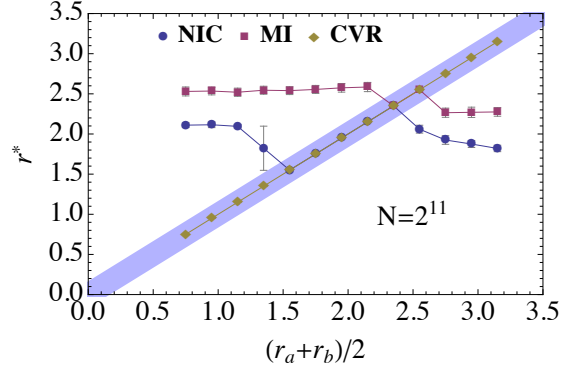

Figure 7. We compare the accuracy of various information-theoretic clustering objectives over a range of parameter values using a fixed number of samples, $N = 2^{11}$. For the distribution in Fig. 5, we set $r_c = 3.5, r_b = r_a + 0.3$ and we vary r_a. For each point, we calculate r^* from 20 random datasets and we plot the sample mean and standard deviation. The shaded area indicates the correct range for r^*.
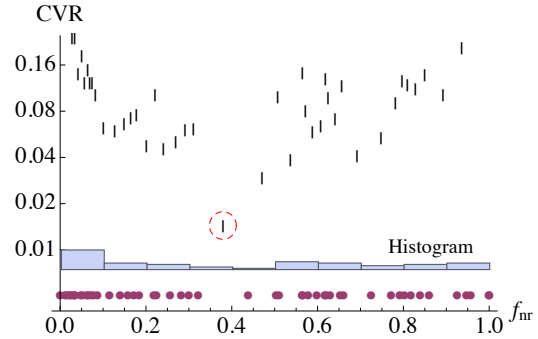

Figure 8. We estimate CVR for editor behavior on Wikipedia.

k-means and the best of IT methods that have been previously compared (Wang & Sha, 2011). Note that the cluster sizes for *iris* and *glass* are exactly and approximately balanced respectively which benefits mutual information since it is biased towards balanced solutions.

We also compare the CVR of the ground truth clustering to the best clustering found according to our heuristic. The CVR of the clusters found for *iris* and *wine* is comparable to that of the ground truth and much smaller than the maximal possible CVR of 1. On the other hand, the CVR for the ground truth clustering for *glass* is actually larger than 1 (possible due to noisy estimators). This implies cluster labels of neighboring sample points are nearly random. Note that the Rand index (Rand, 1971) of the clustering in which each point is its own cluster achieves a score of 0.74, comparable to all the results listed in Table 1.

	Stats			Rand index				
Dataset	N	#clusters	dim	k-means	ITC	CVR	CVR_{bf}	CVR_{gt}
iris	150	3	4	0.868	0.94	0.925	0.08	0.09
wine	178	3	13	0.927	0.92	0.936	0.18	0.25
glass	214	6	9	0.678	0.75	0.671	0.33	1.16

Table 1. Summary of results from clustering with the UCI datasets (Bache & Lichman, 2013). The middle columns report cluster quality by calculating the Rand index (Rand, 1971) of a candidate clustering with respect to the ground truth. For information-theoretic clustering (ITC), we report recent state-of-the-art results for optimizing mutual information using an SDP (Wang & Sha, 2011). The right columns report CVR for the ground truth clustering, Y_{gt} and the best (lowest) CVR found by our heuristic, Y_{bf}.

5. Discussion

Related Work Many clustering approaches explicitly combine information-theoretic methods with a similarity measure (Slonim et al., 2005; Böhm et al., 2006; Gokcay & Principe, 2002). That is they define some function $f(x^{(i)}, x^{(j)})$ that characterizes the similarity of two points and then maximize intra-cluster similarity while simultaneously optimizing some information-theoretic quantity. While these methods side-step our critique of purely information-theoretic clustering methods, they also lose generality with their insistence on defining some *ad hoc* notions of similarity. A line of recent work has attempted to maximize the mutual information between data and cluster labels using non-parametric entropy estimators. Faivishevsky and Goldberger (Faivishevsky & Goldberger, 2010) attempt to use a modified version of a non-parametric entropy estimator (Kraskov et al., 2004) to create a more tractable objective for optimization.[5] Wang and Sha demonstrate a semidefinite optimization based on this criteria (Wang & Sha, 2011). Other attempts to define tractable optimization problems invoke a squared loss variant of mutual information (Sugiyama et al., 2011) and estimators based on constructing minimum spanning trees (Müller et al., 2012). While these methods are general and have shown some success, we showed in Sec. 2 and Sec. 4 that these approaches are ultimately flawed and will fail in the large data limit.

We showed that our approach, unlike MI-based methods, is not biased towards balanced clusters. This is in keeping with our desire to minimize the number of assumptions. On the other hand, there may be scenarios in which a bias

towards balanced clusters is desirable. One line of work attempts to define intuitive properties of clustering methods (like sensitivity to cluster balance) and then categorize them with respect to these properties; see (Ackerman et al., 2012) and references therein.

Conclusion We have demonstrated why existing information-theoretic clustering methods are conceptually flawed and presented a principled solution without introducing any model-based assumptions[6]. *Consistency under coarse-graining* is essential to the definition of entropy. Formally incorporating this notion in a finite-data setting provides a basic and foundational motivation for defining clusters. Alternately, our objective can be viewed as a way of minimizing the estimated uncertainty of cluster labels. Preliminary results indicate the feasibility of optimizing our criteria and also demonstrate competitive accuracy on standard benchmarks.

Information-theoretic learning methods are attractive because they make no assumptions about the underlying data while maintaining a clear, operational meaning. Real-world learning problems consist of finite samples of data from unknown distributions. To construct an information-theoretic foundation for unsupervised learning, we need to carefully refine our measures so that they make sense in finite-data regimes. We hope that further development of these ideas will contribute to that goal.

---

[5]It should be noted that the entropy estimator in (Faivishevsky & Goldberger, 2008) is not asymptotically correct. They argue "The MeanNN estimator exploits the fact that the kNN estimation is valid for every k and therefore averaging estimators for all possible values of k leads itself to a new estimator of the differential entropy". The k-NN estimator is guaranteed to be consistent only when k grows sufficiently slowly (sublinearly) with N, the number of samples; see (Wang et al., 2009) and references therein. Averaging all values of k up to N violates this condition. A simple example where their estimator can be seen to fail is the distribution in Fig. 1(b). Regardless of N, their estimator is proportional to $\log L$, while the true entropy is independent of L.

[6]While we achieved our goal of avoiding reliance on a *global* similarity measure, non-parametric entropy estimators require a notion of adjacency to find data points that are nearest neighbors according to some metric. A major advantage of taking the information-theoretic approach is that $H(Y|X)$ is invariant under smooth, invertible transformations of the data (X). Therefore, any asymptotically unbiased estimator should converge to the same value regardless of the metric used for finding nearest neighbors.

References

Ackerman, Margareta, Ben-David, Shai, Branzei, Simina, and Loker, David. Weighted clustering. In *AAAI*, pp. 858–863, 2012.

Bache, K. and Lichman, M. UCI machine learning repository, 2013. URL http://archive.ics.uci.edu/ml.

Böhm, Christian, Faloutsos, Christos, Pan, Jia-Yu, and Plant, Claudia. Robust information-theoretic clustering. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 65–75. ACM, 2006.

Cover, Thomas M and Thomas, Joy A. *Elements of information theory*. Wiley-Interscience, 2006.

DeDeo, Simon. Collective phenomena and non-finite state computation in a human social system. *arXiv:1212.0018*, 2012. doi: 10.1371/journal.pone.0075818. To appear, PLoS ONE.

DeDeo, Simon, Hawkins, Robert X. D., Klingenstein, Sara, and Hitchcock, Tim. Bootstrap methods for the empirical study of decision-making and information flows in social systems. *Entropy*, 15(6):2246–2276, 2013. ISSN 1099-4300. doi: 10.3390/e15062246. URL http://www.mdpi.com/1099-4300/15/6/2246.

Dhillon, Inderjit S and Guan, Yuqiang. Information theoretic clustering of sparse cooccurrence data. In *Data Mining, 2003. ICDM 2003. Third IEEE International Conference on*, pp. 517–520. IEEE, 2003.

Faivishevsky, Lev and Goldberger, Jacob. Ica based on a smooth estimation of the differential entropy. In *NIPS*, 2008.

Faivishevsky, Lev and Goldberger, Jacob. A nonparametric information theoretic clustering algorithm. In *Proceedings of ICML*, 2010.

Goemans, M.X. and Williamson, D.P. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. Assoc. Comput. Mach.*, 42:1115–1145, 1995.

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

Jain, Anil K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010.

Khan, Shiraj, Bandyopadhyay, Sharba, Ganguly, Auroop R, Saigal, Sunil, Erickson III, David J, Protopopescu, Vladimir, and Ostrouchov, George. Relative performance of mutual information estimation methods for quantifying the dependence among short and noisy data. *Physical Review E*, 76(2):026209, 2007.

Kozachenko, L. F. and Leonenko, N. N. Sample estimate of the entropy of a random vector. *Probl. Peredachi Inf.*, 23(2):95–101, November 1987.

Kraskov, Alexander, Stögbauer, Harald, and Grassberger, Peter. Estimating mutual information. *Phys. Rev. E*, 69:066138, Jun 2004. doi: 10.1103/PhysRevE.69.066138. URL http://link.aps.org/doi/10.1103/PhysRevE.69.066138.

Müller, Andreas, Nowozin, Sebastian, and Lampert, Christoph. Information theoretic clustering using minimum spanning trees. *Pattern Recognition*, pp. 205–215, 2012.

Rand, William M. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Shannon, C.E. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379423, 1948.

Slonim, Noam, Atwal, Gurinder Singh, Tkačik, Gašper, and Bialek, William. Information-based clustering. *Proceedings of the National Academy of Sciences of the United States of America*, 102(51):18297–18302, 2005.

Sugiyama, Masashi, Yamada, Makoto, Kimura, Manabu, and Hachiya, Hirotaka. Information-maximization clustering based on squared-loss mutual information. *arXiv preprint arXiv:1112.0611*, 2011.

Wang, Meihong and Sha, Fei. Information theoretical clustering via semidefinite programming. AISTATS, 2011.

Wang, Qing, Kulkarni, Sanjeev R., and Verdú, Sergio. Universal estimation of information measures for analog sources. *Found. Trends Commun. Inf. Theory*, 5 (3):265–353, March 2009. ISSN 1567-2190. doi: 10.1561/0100000021. URL http://dx.doi.org/10.1561/0100000021.