

# The cluster structure function

Andrew R. Cohen and Paul M.B. Vitányi

**Abstract**—For each partition of a data set into a given number of parts there is a partition such that every part is as much as possible a good model (an “algorithmic sufficient statistic”) for the data in that part. Since this can be done for every number between one and the number of data, the result is a function, the *cluster structure function*. It maps the number of parts of a partition to values related to the deficiencies of being good models by the parts. Such a function starts with a value at least zero for no partition of the data set and descends to zero for the partition of the data set into singleton parts. The optimal clustering is the one selected by analyzing the cluster structure function. The theory behind the method is expressed in algorithmic information theory (Kolmogorov complexity). In practice the Kolmogorov complexities involved are approximated by a concrete compressor. We give examples using real data sets: the MNIST handwritten digits and the segmentation of real cells as used in stem cell research.

**Index Terms**— Cluster, similarity, classification, Kolmogorov complexity, algorithmic sufficient statistic, pattern recognition, data mining.

## I. INTRODUCTION

The aim of this work is to introduce the cluster structure function and apply it to propose a method for finding the number of clusters in a given dataset that is unsupervised, feasible, justifiable in terms of its theory, and more accurate than previous methods for this task. Clustering is a fundamental task in unsupervised learning, partitioning a set of objects into groups called clusters such that objects in the same cluster are more similar to each other than to those in other groups [26]. Every object in a computer is represented by a finite sequence of 0’s and 1’s: a finite binary string (abbreviated to “string” in the sequel). There are many methods and algorithms for clustering and determining the number of clusters in data as for example surveyed in [2], [14], [16], [26]. We explore a new method for determining the number of clusters based on Kolmogorov’s notion of algorithmic sufficient statistic [24], [8] which is expressed in terms of Kolmogorov complexity [17]. For technical reasons we use *prefix Kolmogorov complexity* [19]. In the sequel we also use  $K$  for the number of clusters in the data, agreeing with customary use. Confusion is avoided by the context.

A brief overview of the needed notions is given here. Details and proofs can be found in the textbook [21]. A prefix Turing machine is a Turing machine (we use a binary alphabet) such that the set of input programs for which the machine halts is a

prefix code (no input program is a proper prefix of another one). The prefix Turing machines can be computationally enumerated  $T_1, T_2, \dots$  and this list has a universal prefix Turing machine  $U$  such that  $U(i, p) = T_i(p)$  for all integers  $i$  and halting programs  $p$  for  $T_i$ . Formally, the *conditional prefix Kolmogorov complexity*  $K(x|y)$  is the length of the shortest input string  $z$  such that the reference universal prefix Turing machine  $U$  on input  $z$  with auxiliary information  $y$  outputs  $x$ . The *unconditional prefix Kolmogorov complexity*  $K(x)$  is defined as  $K(x|\epsilon)$  where  $\epsilon$  is the empty string. The quantity  $K(x)$  is the length of a shortest binary string  $x^*$  from which  $x$  can be effectively reconstructed. If there are more than one candidates for  $x^*$  we use the first one in the enumeration. The string  $x^*$  accounts for every effective regularity in  $x$ . In these definitions both  $x$  and  $y$  can consist of strings into which finite multisets of finite binary strings are encoded.

Informally, a finite set  $A$  of strings containing  $x$  is an *algorithmic sufficient statistic* for  $x$  iff  $K(A) + \log |A| = K(x)$ . That is, the encoding of  $x$  by giving  $A$  (a model) and the index of  $x$  in  $A$  is as short as a shortest computer program for  $x$  (sometimes one adds also a small value). This means that  $A$  is a good model for  $x$  [31]. As we show in Lemma 1 it is impossible that  $A$  is such a good model for all  $y \in A$ . Therefore we have to relax the condition of sufficiency. If the equality above holds up to some additive term then this term is called the *optimality deficiency*. We propose to group the elements from a data set (a multiset) into clusters (submultisets) such that the optimality deficiencies in every cluster are minimal in some sense. This seems to require a specification of the number of clusters. However, the aim is to find the number of clusters. To solve this conundrum the proposed method proceeds as follows. The cluster structure function has the number of clusters as argument and a quantity involving the optimality deficiencies as value. Such a function decreases to 0 when the number of clusters grows to the cardinality of the data set. The optimal number of clusters can then be selected related to the cluster structure function.

We give the definitions and the ideal method of application in Section II. Proofs are deferred to Section II-B. An explanation of the probability relations of members of a cluster is given in Section III. A brief survey of related literature is given in Section IV. Finally, Section V shows examples of real applications including estimating the number of unique digits in a set of MNIST handwritten digits and an ensemble segmentation approach to human stem cell nuclear segmentation.

## II. THEORY OF THE CLUSTER STRUCTURE FUNCTION

The aim is to partition a multiset into submultisets such that each submultiset constitutes a cluster. In probabilistic statistics

Andrew Cohen is with the Department of Electrical and Computer Engineering, Drexel University. Address: A.R. Cohen, 3120–40 Market Street, Suite 313, Philadelphia, PA 19104, USA. Email: andrew.r.cohen@drexel.edu

Paul Vitányi is with the National Research Center for Mathematics and Computer Science in the Netherlands (CWI), and the University of Amsterdam. Address: CWI, Science Park 123, 1098XG Amsterdam, The Netherlands. Email: Paul.Vitanyi@cwi.nl.

the relevant notion is the “sufficient statistic” due to R.J. Fisher [13], [8]. According to Fisher:

“The statistic chosen should summarise the whole of the relevant information supplied by the sample. This may be called the Criterion of Sufficiency . . . In the case of the normal curve of distribution it is evident that the second moment is a sufficient statistic for estimating the standard deviation.”

This type of sufficient statistic pertains to probability distributions. In the problem at hand the data are individual strings. Therefore the probabilistic notion is not appropriate. For individual strings the analogous notion is the “algorithmic sufficient statistic.” For convenience we delete the adjective “algorithmic” in the sequel (probabilistic sufficient statistic doesn’t occur in the sequel). We equate a multiset being a cluster with the multiset being, as close as possible according to a given criterion, an (algorithmic) sufficient statistic for the members of the cluster. The new method partitions  $S$  such that each resulting part is as close as possible to the given criterion a sufficient statistic for all of its members. Therefore they are good models for its members [31]. This is different from existing methods which use some metric which does not say much about this aspect.

**DEFINITION 1.** A multiset  $A$  of strings is an *algorithmic sufficient statistic* abbreviated as *sufficient statistic* for a element  $x \in A$  if  $K(A) + \log |A| = K(x)$ .

Here  $A$  is a model and the  $\log |A|$  term allows us to pinpoint  $x$  in  $A$ . Therefore, every  $y \in A$  satisfies  $K(A) + \log |A| \geq K(y)$ . Reference [11] tells us that if  $A$  is a sufficient statistic for the string  $x$  then  $K(A|x) = O(1)$ . That is,  $A$  is almost completely determined by  $x$ . If  $A$  is a sufficient statistic for  $x$ , then  $K(x|A) = \log |A|$ . Namely,  $K(x) \leq K(A) + K(x|A) \leq K(A) + \log |A| = K(x)$ . We call  $x$  a *typical* member of  $A$ .

This is akin to the minimum description length (MDL) principle in Statistics [12]. To illustrate, if the length of a binary string  $x$  is  $n$  and  $K(x) = n + K(n)$  (the maximum) which means that  $x$  is random then  $A = \{x\}$  is a sufficient statistic of  $x$  (the minimal one) and  $A = \{0, 1\}^n$  is also a sufficient statistic of  $x$  (the maximal one). There is a tradeoff between the cardinality of a sufficient statistic  $A$  of a string  $x$  and the amount of effective regularities in the string  $x$  it represents. The greater the cardinality of  $A$  is the smaller is  $K(A)$  which is the amount of effective regularities it represents. The multiset  $A$  accounts for as many effective regularities in  $x$  as is possible for a set of the cardinality of  $A$ . This means that  $A$  is the model of best fit, which we call the best model, for  $x$  which is possible [31, Section IV-B]. Thus, if  $A$  has the property that for every  $y \in A$  it is as much as possible a sufficient statistic, then all members of  $A$  share as many effective regularities as is possible. All the  $y \in A$  are similar in the sense of [20], [4]. We cluster the data according to this criterium.

If  $A$  contains elements  $y$  such that  $K(A) + \log |A| > K(y)$  (trivially  $<$  is impossible) then  $K(A|y) \neq O(1)$ . Let us look closer at what this implies and consider  $A$  containing only elements of length  $n$ . Then by the symmetry of information [10] we have  $K(A|x) = K(A) + K(x|A) - K(x) + O(\log n)$ .

For example, let  $A$  be the set containing all integers in an interval with complex endpoints and  $x$  an integer in this interval of low complexity. For example  $K(A) = \Omega(n)$  and  $K(x) = o(n/4)$ . Therefore  $K(x|A) = o(n/4)$  and this yields  $K(A|x) = \Omega(n)$ . That is,  $A$  is not at all determined by  $x$ .

**DEFINITION 2.** The *optimality deficiency* of  $A$  as a *sufficient statistic* for  $x \in A$  is

$$\delta(A, x) = K(A) + \log |A| - K(x). \quad (\text{II.1})$$

The *mean* of the optimality deficiencies of a set  $A$  is

$$\mu_A = \frac{1}{|A|} \sum_{x \in A} \delta(A, x).$$

Here  $\delta(A, x) \geq 0$  with equality for a proper sufficient statistic. If  $\mu_A = 0$  then  $\delta(A, x) = 0$  for all  $x \in A$ , that is,  $A$  is a sufficient statistic for all of its elements. But this is not possible for  $|A| \geq 2$  by the following lemma.

**LEMMA 1.** Let  $A$  be a finite multiset of strings of length  $n$ .

(i) Let  $\delta(A, x) = 0$  for some  $x \in A$ . For all  $y \in A$  holds  $K(y) \leq K(x)$  and if  $|A| > 2$  then  $K(y) < K(x)$  for some  $y \in A$ ,  $\delta(A, y) > 0$ , and  $\mu_A > 0$ .

(ii) There exist  $A$  and  $x \in A$  such that  $\delta(A, x) < 0$  and for such  $A$  no  $y \in A$  satisfies  $\delta(A, y) = 0$ .

**REMARK 1.** The optimality deficiency should not be confused with the *randomness deficiency* of  $x \in A$  with respect to  $A$ :

$$\delta(x|A) = \log |A| - K(x|A).$$

By the symmetry of information law  $K(A) + K(x|A) = K(x) + K(A|x)$  up to a logarithmic additive term  $O(\log K(A))$ . Therefore  $\delta(x|A) + K(A|x) = \log |A| + K(A) - K(x) + O(\log K(A))$  and hence  $\delta(A, x) = \delta(x|A) + K(A|x) + O(\log K(A))$ .  $\diamond$

For clustering we want ideally the model to be a sufficient statistic for all elements in it. But we have to deal with optimality deficiencies which are greater than 0, and with real data typically they are all greater than 0. There are many ways to combine the optimality deficiencies (or other aspects) to obtain criteria for selection. This is formulated in the criterion function as follows.

Let  $\mathcal{N}$  denote the natural numbers and  $S = \{x_1, \dots, x_n\}$  be a finite nonempty multiset of strings. Consider a partition  $\pi$  of  $S$  into  $k$  nonempty subsets  $S_1, \dots, S_k$  such that  $\bigcup_{i=1}^k S_i = S$  and  $S = \{x_1, \dots, x_n\}$  be a finite nonempty multiset of strings. Consider a partition  $\pi$  of  $S$  into  $k$  nonempty subsets  $S_1, \dots, S_k$  such that  $\bigcup_{i=1}^k S_i = S$  and  $S_i \cap S_j = \emptyset$  for  $i \neq j$ . Denote the set of partitions of  $S$  into  $k$  submultisets by  $\Pi_k$  and the set of all partitions by  $\Pi$ . The *criterion function*  $f : \Pi \rightarrow \mathcal{N}$  takes as argument a partition  $\pi \in \Pi$  of  $S$  and as value a natural number computed from the optimality deficiencies involved in the partition subject to the following: (i) the value of  $f(\pi)$  does not increase if one or more optimality deficiencies are changed to 0; and (ii)  $f(\pi) = 0$  if all optimality deficiencies are 0. (One can use other aspects as well.)

DEFINITION 3. The *Cluster Structure Function* (CSF)<sup>1</sup> for a multiset  $S$  of  $n$  strings is defined by

$$H_S^f(k) = \min_{\pi \in \Pi_k} f(\pi) \quad (\text{II.2})$$

where  $f$  is the criterion function, for each  $k$  ( $1 \leq k \leq n$ ). The graph of this function is called the *CSF curve*. If  $f$  is understood we may write  $H_S$  for the CSF function.

EXAMPLE 1. Let  $\pi \in \Pi_k$ . The *bandwidth* of  $S_i$  is  $b_i = \max_{x \in S_i} \{\delta(S_i, x)\} - \min_{x \in S_i} \{\delta(S_i, x)\}$ . Define  $f(\pi) = \min \sum_{1 \leq i \leq k} b_i$ . For every  $k$  ( $1 \leq k \leq n$ ) the value  $H_S^f(k)$  is based on the partition  $\pi \in \Pi_k$  that minimizes the minimal sum of of the bandwidths of the parts in a  $k$ -partition of  $S$ . If we consider the graph of  $H_S^f$  in a two-dimensional plane with the horizontal axis denoting the number  $k$  of parts of  $S$ , and the vertical axis denoting the value of  $H_S^f$ , then left of the graph of  $H_S^f$  there are no possible  $k$ -partitions while right of the graph of  $H_S^f$  there are redundant  $k$ -partitions. On the graph of  $H_S^f$  occur the witness partitions.  $\diamond$

REMARK 2. By Lemma 1 parts  $A$  with  $|A| > 2$  of a witness partition  $\pi$  of  $S$  can not be a sufficient statistic for all of its elements and therefore  $f(\pi) > 0$ .  $\diamond$

REMARK 3. In clusters the members of a cluster typically share some characteristics but not all characteristics. It turns out that the members of a cluster are probabilistically close, Section III.  $\diamond$

#### A. Properties

It is convenient to ignore possible  $O(1)$  additive terms in the sequel.

LEMMA 2. Let  $S = \{x_1, \dots, x_n\}$  with  $n \geq 1$ . For every  $f$  we have  $H_S^f(n) = 0$  and  $H_S^f$  is monotonic non-increasing with increasing arguments on its domain  $[1, n]$ .

The graph of  $H_S^f$  descends until  $H_S^f(k) = 0$  for the least  $k \leq n$ ,  $H_S^f(k) = \dots = H_S^f(n) = 0$ . We give a lower bound on  $H^f$  for some datasets  $S$ .

LEMMA 3. There exist  $S \subseteq \{0, 1\}^m$  with  $|S| = n$  and  $n \leq m$  such that  $H_S^f(k) = 0$  for all  $1 \leq k \leq n$  up to an additive term of  $O(\log K(S))$ .

The following lemma establishes that there are sets  $S$  of  $n$  elements such that  $H_S^f$  stays at a high value for arguments  $1, \dots, n-1$  and drops suddenly to 0 for argument  $n$ .

LEMMA 4. There exists a set  $S \subseteq \{0, 1\}^m$  and  $|S| = n$  with  $m$  a sufficiently large multiple of  $n$  such that  $H_S^f(k) \geq m/n$  for  $1 \leq k \leq n-1$  and  $H_S^f(n) = 0$ .

In practice we may use the optimality deficiencies within the standard deviation around the mean to determine the criterion function  $f(\pi)$  for a partition  $\pi \in \Pi_k$  of  $S$  into parts  $S_1, \dots, S_k$ . This is a more refined method since it eliminates

the outliers, only counting the central items (68.2% if they are normally distributed) of the optimality deficiencies in each part  $S_i$ . The *mean* of  $S$  is  $\mu_S = 1/|S| \sum_{x \in S} x$ . The *standard deviation* of the  $\delta(S, x)$  of a multiset  $S$  is

$$\sigma_S = \frac{1}{|S|} \sqrt{\sum_{x \in S} (\delta(S, x) - \mu_S)^2}.$$

DEFINITION 4. Let  $S$  be a multiset of strings,  $S_\sigma = \{x \in S : |x - \mu_S| \leq \sigma_S\}$  and  $f_\sigma$  is the criterion function of  $S_\sigma$ .

$$H_{S_\sigma}^{f_\sigma}(k) = \min_{\pi \in \Pi_k} \max_{1 \leq i \leq k} f_\sigma(\pi), \quad (\text{II.3})$$

where  $\pi$  divides  $S_\sigma = \bigcup_{1 \leq i \leq k} S_{\sigma,i}$  into  $k$  parts  $S_{\sigma,1}, \dots, S_{\sigma,k}$

That is,  $H_{S_\sigma}^{f_\sigma}(k)$  is the minimum over all partitions of  $S_\sigma$  into  $k$  parts. It clusters possibly better since  $H_{S_\sigma}^{f_\sigma}(k) \leq H_S^f(k)$  for all  $k$  ( $1 \leq k \leq n$ ) implying by Section III that the conditional probabilities between most members of a part of a witness partition may be larger but never smaller using  $H_{S_\sigma}^{f_\sigma}(k)$  than using  $H_S^f(k)$ .

LEMMA 5. Let  $S = \{x_1, \dots, x_n\}$  with  $n \geq 2$ . Then  $H_{S_\sigma}^{f_\sigma}(1) > 0$ ,  $H_{S_\sigma}^{f_\sigma}(n) = 0$ , and  $H_{S_\sigma}^{f_\sigma}$  is monotonic non-increasing.

LEMMA 6. There exists  $S \subseteq \{0, 1\}^m$  with  $|S| = n$  such that  $H_{S_\sigma}^{f_\sigma}(k) = 0$  for all  $1 \leq k \leq n$  up to an additive term  $O(\log K(S))$ .

LEMMA 7. There exists a multiset  $S \subset \{0, 1\}^m$  and  $|S| = n$  with  $m$  multiple of  $n$  such that  $H_{S_\sigma}^{f_\sigma}(k) \geq m/n$  for  $1 \leq k \leq n-1$  and  $H_{S_\sigma}^{f_\sigma}(n) = 0$ .

#### B. Proofs

*Proof.* of Lemma 1 (i) For all  $y \in A$  we have  $K(y) \leq K(A) + \log |A|$  which implies  $K(y) \leq K(x)$  (since  $K(x) = K(A) + \log |A|$ ) and therefore  $\delta(A, y) \geq 0$  and hence  $\mu_A \geq 0$ . For  $|A| > 2$  there are  $y \in A$  such that  $K(y) < K(x)$  since  $K(y) < K(A) + \log |A| = K(x)$ . For example if  $y$  is the first element of  $A$  and therefore  $K(y) \leq K(A)$ . Hence  $\delta(A, y) > 0$  and  $\mu_A > 0$ .

Ad (ii) There is an  $x \in A$  such that  $\delta(A, x) < 0$ . For example  $A$  is a sufficiently long interval of integers of (represented by  $n$ -bit strings) of length  $O(2^n)$  with end points of  $O(\log n)$  Kolmogorov complexity and  $x \in A$  is a random string in that interval which means  $K(x) = \Omega(n)$ . Then  $\delta(A, x) < 0$  and by Item (i) there are no  $y \in A$  such that  $\delta(A, y) = 0$ .  $\square$

*Proof.* of Lemma 2 The graph of  $H_S^f$  starts with the partition of  $S$  into 1 part (no partition).

$n = 1$ . The optimality deficiency involved is 0 and by Definition 3 we have  $H_S^f(1) = 0$ .

$n > 1$ . Let  $1 \leq k < |S|$ . By Item (i) in the definition of the criterion function, if  $\pi \in \Pi_{k+1}$  and we change one of the optimality deficiencies of the elements to 0 then the criterion function  $f$  does not increase. Hence the minimum of  $f$  for a partition in  $\Pi_k$  is not larger than the minimum of  $f$  for a partition in  $\Pi_{k+1}$ . Therefore  $H_S^f$  is monotonic non-increasing.

<sup>1</sup>The cluster structure function is named in analogy with the Kolmogorov structure function  $h_x : \mathcal{N} \rightarrow \mathcal{N}$  defined by  $h_x(k) = \min_{S \subseteq \{0, 1\}^*} \{\log |S| : x \in S, K(S) \leq k\}$  associated with a binary finite string  $x$  [31].

For  $k = n$  the multiset  $S$  is partitioned into singleton sets which all have optimality deficiency 0. Hence  $H_S^f(n) = 0$ .  $\square$

*Proof.* of Lemma 3 Choose  $x \in \{0,1\}^m$  and  $S$  with  $|S| = n$  such that  $S = \{y : |y| = m \text{ and } y \text{ equals } x \text{ with the } i\text{th bit flipped } (1 \leq i \leq n)\}$ . Then for each  $y \in S$  we have  $K(S) = K(y) + O(\log n)$ . Therefore  $\delta(S, y) = K(S) + \log n - K(y) = O(\log n)$  for all  $y \in S$ . Hence  $H_S^f(1) = O(\log n) = O(\log K(S))$ . For every  $k$  ( $1 < k \leq n$ ) we describe the partition  $\pi \in \Pi_k$  which witnesses  $H_S^f(k)$  by giving  $S$  in  $K(S)$  bits, the integer  $k$  in  $O(\log n)$  bits and an  $O(1)$  program. This program does the following: given  $k$  and  $S$  it generates all finitely many partitions  $\pi \in \Pi_k$ . A partition  $\pi \in \Pi_k$  of  $S$  divides it into, say,  $S_1, \dots, S_k$ . By the symmetry of information law [10] we have  $K(S) = K(S_i) + K(S|S_i) + O(\log K(S))$  or  $K(S_i) \leq K(S) - O(\log K(S))$ . For every  $y \in S_i$  therefore  $\delta(S_i, y) = K(S_i) + \log |S_i| - K(y) \leq K(S) - O(\log K(S)) + \log |S| - K(y) = \delta(S, y) - O(\log K(S))$ . Since  $1 \leq k \leq n$  this proves the lemma.  $\square$

*Proof.* of Lemma 4 Let  $S = \{x_1, \dots, x_n\}$  with  $K(x_i) = im/n$  for  $1 \leq i \leq n$ . (This is possible since all  $n$  members of  $S$  are strings of length  $m$  and they can have complexity varying continuously between at least  $m$  and close to 0.) Since for each finite multiset  $A$  and  $x \in A$  we have  $\delta(A, x) = K(A) + \log |A| - K(x)$  and therefore

$$\max_{x \in A} \delta(A, x) - \min_{x \in A} \delta(A, x) = \max_{x \in A} \{K(A) + \log |A| - K(x)\} - \min_{x \in A} \{K(A) + \log |A| - K(x)\}.$$

For a  $k$ -partition of  $S$  at least one  $S_i$  in the partition has cardinality at least  $n/k$ . Therefore, if  $n/k > 1$  then by the displayed equality  $H_S^f(k) > m/n$ . This holds for  $k = 1, \dots, n-1$ . For  $k = n$  all parts  $S_i$  in the partition are singleton sets and hence  $H_S^f(k) = 0$ .  $\square$

*Proof.* of Lemma 5. Similar to the proof of Lemma 4.  $\square$

*Proof.* of Lemma 6. Similar to proof in Lemma 3.  $\square$

*Proof.* of Lemma 7. Similar to the proof of Lemma 4.  $\square$

### C. Computing the number of clusters

To determine the number  $K$  of clusters in data  $S$  we compare a cluster structure function used on  $S$  with the same cluster function on *reference set* of  $|S|$  data distributed uniformly. We do this comparison as the logarithm of the ratio. Using the cluster function  $H_S^f$  on the data set  $S$  the number  $K$  of clusters in  $S$  is the  $k$  where the log-ratio  $D^f(k)$  is greatest. Formally

$$D^f(k) = \log H_N^f(k) - \log H_S^f(k) \\ K = \arg \max_k D^f(k),$$

with the reference placement is the uniform distribution of  $|S|$  data samples over the range spanned by  $S$ . For example if  $S$  is a set of numbers then its range is the interval  $I = [\min S, \max S]$ . Note that every set  $S$  is represented in a computer memory as a finite set of finite strings of 0's and 1's and that therefore  $\min S$  and  $\max S$  are well defined. Divide  $I$  in  $n$  equal parts  $I_1, \dots, I_n$  with  $\bigcup_{i=1}^n I_i = I$  and  $I_i \cap I_j = \emptyset$

for  $1 \leq i \neq j \leq n$ . Item  $i \in N$  is positioned in the middle of subinterval  $I_i$  ( $1 \leq i \leq n$ ).

To deal with the incomputability of the function  $K$  we approximate  $K$  from above by a good compressor  $Z$ . If  $x$  is a string then  $Z(x)$  is the length of the by  $Z$  compressed version of  $x$ . The function  $Z$  is by construction a computable function, even a feasibly computable one (for example  $Z$  is bzip2 or some other compressor). Because  $K$  is incomputable there are strings  $x$  such that  $K(x) \ll Z(x)$  and the difference  $Z(x) - K(x)$  is incomputable. However for natural data we assume that they encode no universal computer or problematic mathematical constants like the ratio of the circumference of a circle to its diameter 3.14... We assume that for the natural data we encounter the compression by  $Z$  has a length which is close to its prefix Kolmogorov complexity. The same holds for a multiset  $A$  of strings. We represent  $A = \{x_1, \dots, x_n\}$  as a string  $s(A) = 1^{|x_1|}0x_1 \dots 1^{|x_n|}0x_n$  with  $|s(A)| = |x_1 \dots x_n| + O(\log |x_1| + \dots + \log |x_n|)$ .

For a partition  $\pi \in \Pi_k$  of  $S$  ( $|S| = n$ ) we compute the  $\delta(S_i, x)$ 's by computing  $Z(S_i)$  ( $1 \leq i \leq k$ ) and  $Z(x)$  for all  $x \in S$ . To do so we require at most  $k + n$  compressions. We write "at most" since a member of a multiset  $S$  can occur more than once.

### III. PROBABILITIES AMONG MEMBERS OF CLUSTERS

By Lemma 1 a part  $A$ , with more than two members, of a witness partition of  $S$  can not be a sufficient statistic for all of its elements. In clusters the members of a cluster typically share some characteristics but not all characteristics. It turns out that in an appropriate sense the members of a cluster are nonetheless probabilistically close.

We define a conditional probability of  $n$ -bit strings following [22]. We start with the unconditional probability. Let a finite set  $A$  of  $n$ -bit strings be chosen randomly with probability  $\mathbf{m}(A) = 2^{-K(A)}$ , and subsequently  $x \in A$  is chosen with uniform probability from  $A$ , that is,  $x$  is chosen with probability  $\mathbf{m}(A)/|A|$ . (Since  $K(x)$  is a length of a prefix code we have by Kraft's inequality [8] that  $\sum_x 2^{-K(x)} \leq 1$ . Hence  $\mathbf{m}$  is a semiprobability. A semiprobability is just like a probability but may sum to less than 1. The particular semiprobability  $\mathbf{m}$  is called *universal* since it is the largest lower semicomputable semiprobability [19]. In absence of any information about  $A$  we can assign  $\mathbf{m}(A)$  as its probability. Properties are discussed in the text [21]).

DEFINITION 5. For each  $y \in A$  we define the *conditional probability*  $p(y|x)$  by

$$p(y|x) = \frac{\sum_{A \ni x, y} \mathbf{m}(A)/|A|}{\sum_{A \ni x} \mathbf{m}(A)/|A|}$$

We show below that all pairs of strings in a part of a witness partition of multiset  $S$  of  $n$  strings have an expectation of the conditional  $p$ -probability with respect to each other which is at least  $2^{-H_S^f(k)}$  for some  $k \leq n$ . Hence the smaller  $H_S^f(k)$  is the more all strings in a part of a witness partition of  $H_S^f(k)$  have a large conditional probability with respect to each other: they form a cluster.

**THEOREM 1.** Let  $S \subseteq \{0, 1\}^n$  (consider only  $n$ -length strings) and a witness  $k$ -partition of  $S$  for  $H_S^f(k)$  that divides  $S$  into parts  $S_1, \dots, S_k$ . The expectation taken over a random variable  $p(y|x)$  for pairs  $x, y \in S_i$  for some  $i$  ( $1 \leq i \leq k$ ) is  $\mathbf{E}[p(y|x)] \geq 2^{-H_S^f(k) - O(\log n)}$  and  $\mathbf{E}[p(y|x)]$  becomes at least  $(1/n)^{O(1)}$  for  $k \rightarrow n$ .

*Proof.* The parts of a witness to  $H_S^f(k)$  form clusters because intuitively if the conditional probabilities in Definition 5 of different strings in a part of the witness partition are small then the conditional Kolmogorov complexities are small:

CLAIM 1.

$$p(x|y) = \frac{\Theta(\mathbf{m}(x, y))}{\Theta(\mathbf{m}(x))} = 2^{-K(x|y) - O(\log n)}.$$

*Proof.* Start from Definition 5. The first equality holds by the following reasoning: since  $\sum_{A \ni x} \mathbf{m}(A)/|A| = \Theta(\mathbf{m}(x))$  because the lefthand side of the equation is a lower semi-computable function of  $x$  and hence it is  $O(\mathbf{m}(x))$ ; moreover if  $A = \{x\}$  then the lefthand side equals  $\mathbf{m}(x)$ . The same argument can be used for the pair  $\{x, y\}$ . The second equality uses the coding theorem [19] which states  $\mathbf{m}(x) = 2^{K(x) + O(1)}$  and the symmetry of information law [10] which shows both the trivial  $K(x, y) \leq K(x) + K(y|x)$  and  $K(x, y) \geq K(x) + K(y|x) - O(\log K(x, y)) = K(x) + K(y|x) - O(\log n)$ . The  $\Theta$  order of magnitude is an  $O(1)$  term in the exponent and absorbed in the  $O(\log n)$  term.  $\square$

The conditional probabilities of pairs of strings in a part of a  $k$ -partition of  $S$  which is a witness to  $H_S^f(k)$  satisfy the following. By [22, Theorem 5] if  $x, y \in S_i$  for a *particular*  $i$  ( $1 \leq i \leq k$ ) and  $\delta(S_i, x) \leq d$  then  $p(y|x) \geq 2^{-d - O(\log n)}$ , while if  $p(y|x) \geq 2^{-d}$  then  $\delta(S_i, x) \leq d + O(\log n)$ . Hence  $p(y|x) = 2^{-\delta(S_i, x) \pm O(\log n)}$ . The expectation of  $p(y|x)$  over  $S_i$  is given by

$$\begin{aligned} \mathbf{E}[p(y|x)] &= 1/|S_i| \sum_{x \in S_i} 2^{-\delta(S_i, x) \pm O(\log n)} \\ &\geq 2^{-\sum_{x \in S_i} (\delta(S_i, x) \pm O(\log n)) / |S_i|} \\ &= 2^{-\mu_{S_i} \pm O(\log n)}, \end{aligned}$$

using in the second line the inequality of arithmetic and geometric means. This implies that if  $x, y \in S_i$  for *some*  $i$  ( $1 \leq i \leq k$ ) then the expectation of  $p(y|x)$  over all  $S_i$  in a witness partition of  $S$  is given by

$$\begin{aligned} \mathbf{E}[p(y|x)] &= (1/k) \sum_{i=1}^k 2^{-\mu_{S_i} \pm O(\log n)} \\ &\geq 2^{-(1/k) \sum_{i=1}^k (\mu_{S_i} \pm O(\log n))} \\ &\geq 2^{-H_S(k) \pm O(\log n)}, \end{aligned}$$

using in the second line again the inequality of arithmetic and geometric means and in the third line that  $H_S(k) \geq (1/k) \sum_{i=1}^k \mu_{S_i}$  by Definition 3. Since  $H_S(n) = 0$  we have  $\mathbf{E}[p(y|x)]$  is at least  $(1/n)^{O(1)}$  for  $k \rightarrow n$ .  $\square$

**REMARK 4.** Roughly, the smaller  $H_S$  becomes the larger the conditional probabilities of the elements in a part of the witness partition become.  $\diamond$

## IV. RELATED LITERATURE

This paper extends previous work in the field of algorithmic statistics [11], [31], [29]. The applications build particularly on the field of semi-supervised spectral learning [15], [23], [6]. Most previous approaches to estimating the number of clusters in a dataset utilize probabilistic statistical modeling of the data. The Bayesian and Akaike information criteria both formulate the question w.r.t. underlying distributions estimated either parametrically or empirically [26], [37], [36], [38]. Bayesian methods are well suited when the likelihood function and prior probabilities are known. In comparison, the algorithmic statistics approach proposed here works with the particular dataset rather than probabilities across a hypothesized population. Recently, alternative approaches based on characteristics of the specific data set in question, rather than a population-level model, have been considered [40], [39]. These include particularly the widely used Gap statistic [27] that is very similar in spirit to the implementation described here. The connection between the gap statistic and the field of algorithmic statistics was one of the key motivators for this work [5]. The gap statistic compares the spatial characteristics of the data being clustered to that of a randomly generated reference distribution. Our approach is similar in both theory and practice to the gap statistic. Many of the advantages of the two approaches are shared. Both are effective when  $K=1$ , that is there are no meaningful clusters among the data. Both are reasonably efficient to compute. DBScan combines the clustering and  $K$  estimation into a single task, and provides parameters for fine tune control [9]. In theory the cluster structure function might be used in an automated parameter search with such an algorithm.

In the computational biological microscopy image analysis area we build on previous work for optimally partitioning connected components of foreground pixels into elliptical regions [35]. A key advantage of the cluster structure function compared to all other approaches is the very broad and powerful theoretical structure of Kolmogorov complexity and Algorithmic Statistics. The techniques are generally parameter free, beyond the selection of a suitable compression algorithm. In theory it will be possible to automatically identify the optimal compression by considering ensembles of algorithms and choosing the best results among them via the structure function.

## V. EXAMPLE APPLICATIONS

### A. How many different digits are in a MNIST digit set?

Here we apply the optimality deficiency to estimating the number of different digits in a set of digits sampled from the MNIST handwritten digits dataset. Classification of the MNIST digits using supervised learning techniques is well studied but there has been little application of unsupervised learning to this problem. One key challenge is establishing a ground truth number of different classes. Different styles of handwriting were taught at different times in different locations. These differences are likely reflected in the underlying data as distinct categories, even within digits of the same class label. Another challenge is the difficulty in unsupervised

classification of digits even when the correct number of classes is known. While supervised solutions for the MNIST digit classification are extremely accurate, unsupervised clustering of MNIST digits is still a difficult problem. The MNIST data has been normalized to 28x28 8-bit grayscale (0,...,255) images. The MNIST database contains a total of 70,000 handwritten digits consisting of 60,000 training examples and 10,000 test examples. Originally the input looks as Figure 1.

We apply the cluster structure function to the question of estimating how many different MNIST digits are represented in a large set. Configure a sampler to choose a set of 100 digits at a time randomly given a fixed  $K$  value. In each digit set, the cardinality of each digit is given by  $\lceil \frac{100}{K} \rceil$ . For  $K = 3$  there would be 33 of each digit [0,1,2]. For  $K = 10$ , there would be 10 of each digit [0, 9]. Spectral clustering [23] is used to cluster MNIST digit sets [7]. The spectral clustering approach starts with the matrix of pairwise normalized compression distances (NCD) as in [4] among all pairs of digits. We used the free lossless image format (FLIF) compressor [25] for the MNIST digits, and found it to significantly outperform the previously used BZIP and JPEG2000 compressors. After computing the NCD matrix between digit pairs, the classic spectral clustering algorithm [23] is applied. Table I shows the results of spectral clustering for all ten digits. Clustering accuracy for all ten digit types [0..9] was 46% with 95% confidence intervals of [.4622,.4657] obtained via bootstrapping [26].

To compute the CSF function following Section II-C we proceed as follows. For each digit set  $S$ , we generate Cluster Structure Function (CSF) curves. The digit set  $S$  is clustered at different values of  $K$ . Random samples of the data forming subset  $\tilde{S} \subseteq S$  are chosen iteratively. Statistics of the pointwise CSF curves are formed from the random samples  $\tilde{S}$ . The results here were generated using 1000 random samples of each digit set as follows. For each digit set  $S$  compute the pairwise NCD matrix  $D$  between all elements of  $S$  using the FLIF image compression. For each  $K$  on  $[1, \dots, K_{\max}]$  use spectral clustering to partition the elements of  $S$  into  $K$  groups. Each cluster (partition) is labeled  $A_p = \{x_1, x_2, \dots, x_{|A_p|}\}$  and  $|A_p|$  is the number of points in cluster  $A_p$ . After the points have

been clustered for a particular value of  $K$ , pick subsets at random from  $S$  to form  $\tilde{S}$ ,  $|\tilde{S}| = 5 * K$ . Using the cluster assignments for each of the randomly selected points, compute the optimality deficiency for each random sample across each of the  $K$  clusters  $A_p$

$$\delta(A_p, x_i) = \begin{cases} 0 & |A_p| < 2 \\ Z(A_p) - Z(x_i) + \log |A_p| & |A_p| \geq 2, \end{cases} \quad (\text{V.1})$$

where  $Z(A_p)$  is the size in bytes of the FLIF compressed image formed by concatenating all of the digit images in  $\tilde{S}$  belonging to cluster  $A_p$  and  $Z(x_i)$  is the size in bytes of the compressed image corresponding to digit  $x_i$ . We write  $\delta(A_p) = \{\delta(A_p, x_1), \delta(A_p, x_2), \dots, \delta(A_p, x_{|A_p|})\}$  to denote the set of optimality deficiencies for each  $x_i \in A_p$ . After computing  $\delta(A_p)$  for each cluster from the digit set subsample  $\tilde{S}$ , the results of V.1 are combined to compute the related cluster structure function:

$$H_{\tilde{S}}(K) = \frac{\sum_{p=1}^{K_{\max}} \log_2(\max(\delta(A_p)) - \min(\delta(A_p)) + 1)}{K_{\max}}. \quad (\text{V.2})$$

The final cluster structure function (CSF) curve is then generated using the mean and standard deviation of  $H_{\tilde{S}}(K)$  across all random subsamples  $\tilde{S}$  and  $K$  values. The optimal value of  $K$  for such a CSF curve is chosen using the technique proposed in [27], as the first value of  $K$  where the CSF curve decreases more than one standard deviation from the previous value. A robust estimator for standard deviation may be useful in identifying the minimum value of the cluster structure function for some applications. Figure 2 shows two example CSF curves. In the left panel, the correct value is obtained at  $K^{\text{pred}} = K^{\text{true}}$ . In the right pane of Figure 2, the selected value is obtained at  $K^{\text{pred}} = 5$  and does not match the  $K^{\text{true}} = 9$  correct value, although there is a minor decrease at nine for that example.

The minimum of the empirical CSF curve is not in itself significant since the ideal theoretic CSF curve is monotonic non-decreasing and the minimum is always at 0 (Lemma 2). What makes the minimum possibly a little meaningful is when the minimum occurs just after a sharp decrease in the CSF curve. The empirical CSF curves of Figure 2 seem in contradiction with Lemma 2. The curves in the figure roughly follow Lemma 2, but they are the results of several heuristics so they may not be perfectly monotonic non-decreasing. The heuristics are among others: approximation from above of the non-computable Kolmogorov complexity, the spectral heuristic of finding the number of clusters rather than inspecting all the subsets of the data, and repeated random sampling of a subset  $\tilde{S} \subseteq S$  computing the CSF curve of each  $\tilde{S}$  and taking the average. To identify the number of clusters in the data one takes the number following the sharp decrease of the CSF curve. Here the criterion to select the clusters is optimally satisfied.

Table II shows the average and standard deviations from subsampling digit sets of varying  $K^{\text{true}}$ . Digits sets with

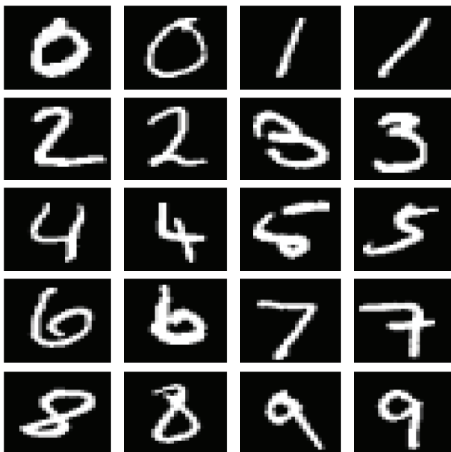


Fig. 1: Example MNIST handwritten digits

|            |   | Predicted Digit |       |       |       |       |       |       |       |       |       |
|------------|---|-----------------|-------|-------|-------|-------|-------|-------|-------|-------|-------|
|            |   | 0               | 1     | 2     | 3     | 4     | 5     | 6     | 7     | 8     | 9     |
| True Digit | 0 | 5,584           | 322   | 924   | 575   | 487   | 587   | 551   | 285   | 504   | 181   |
|            | 1 | 222             | 8,550 | 264   | 170   | 107   | 143   | 198   | 88    | 186   | 72    |
|            | 2 | 1,815           | 666   | 3,615 | 732   | 403   | 447   | 1,248 | 334   | 526   | 214   |
|            | 3 | 1,470           | 480   | 1,020 | 3,557 | 554   | 789   | 362   | 538   | 761   | 469   |
|            | 4 | 1,095           | 533   | 638   | 715   | 2,990 | 628   | 708   | 1,032 | 565   | 1,096 |
|            | 5 | 1,350           | 453   | 679   | 1,221 | 704   | 3,281 | 608   | 568   | 673   | 463   |
|            | 6 | 549             | 469   | 638   | 271   | 269   | 294   | 7,182 | 74    | 157   | 97    |
|            | 7 | 377             | 410   | 292   | 349   | 732   | 426   | 70    | 5,657 | 309   | 1,378 |
|            | 8 | 1,514           | 674   | 815   | 1,119 | 754   | 932   | 255   | 696   | 2,701 | 540   |
|            | 9 | 414             | 338   | 234   | 522   | 1,151 | 479   | 143   | 2,754 | 444   | 3,521 |

TABLE I: Confusion matrix for spectral clustering sets of random digits. Each digit set contains 5 of each digit [0,9]. The digit sets are clustered into 10 clusters. For evaluation, each cluster is labeled with the mode (most common element) of the true digit values in that cluster. This was repeated ten thousand times. Overall accuracy of clustering the ten digit classes is 46%.

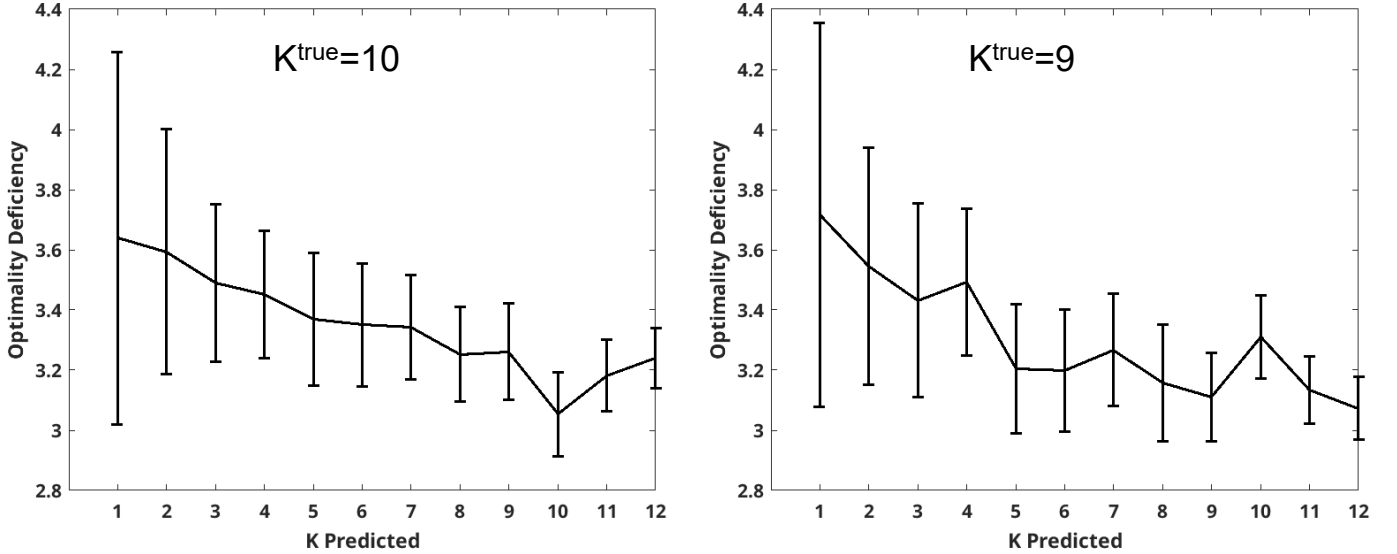


Fig. 2: Curves showing mean and standard deviation of the cluster structure function (CSF) for two different digit sets. Subsets are chosen repeatedly from each digit set, and clustered into  $K$  groups. The value of  $K$  is chosen as the first  $K$  that is one standard deviation smaller than the previous value. The left curve selects the value  $K = K^{true}$ , correctly identifying the value of  $K$  corresponding to the number of different digits in the set. The right curve incorrectly selects  $K = 5$ .

$K^{true} = 1$  have a higher  $K^{pred}$  value, and a much higher standard deviation compared to digits sets with other values of  $K^{true}$ . Omitting digit sets with  $K^{true} = 1$  significantly increases the correlation between the selected point on the CSF curve and  $K^{true}$ . For the CSF, the correlation  $r$  between  $K^{true}$  and  $K^{pred}$  for  $K^{true} > 1$  is  $r = 0.93$ , with a p-value  $p = 3e-4$ . For the Gap Statistic,  $r = -0.84$  ( $p = 5e-3$ ). Based on that observation, a shallow feedforward neural network was used to map the CSF curves to a predicted value  $K^{pred}$ .

The approach is now to use the 20 element vector composed of the mean and standard deviations of the CSF curves evaluated at the numbers of clusters  $K = [1..10]$  as a feature vector to identify the optimal value of  $K$ . We use one thousand

examples each of digit sets from  $K = [1..10]$  as training data (ten thousand total digit sets). Using the MATLAB patternet() classifier with all default parameters, a shallow feed forward neural network with 20 input layer nodes, 10 hidden layer nodes and 10 output layer nodes is trained using ten thousand digit sets, one thousand examples each from  $K \in [1..10]$ . We classify 100 unknown digit sets. When the classification confidence is low, we repeat the sampling, selecting a new  $S$  up to 10 times and average the results to form the prediction. Table III shows the resulting predictions. The vertical axis of the table represents  $K^{true}$ , the horizontal axis represents  $K^{pred}$ . Elements on the diagonal represent correct classifications. Overall accuracy, measured as the percentage of non-zero results that fall on the diagonal of the confusion matrix



| $K^{true}$ | digit set | Cluster Structure Function |                    | Gap Statistic   |                    |
|------------|-----------|----------------------------|--------------------|-----------------|--------------------|
|            |           | $\mu(K^{pred})$            | $\sigma(K^{pred})$ | $\mu(K^{pred})$ | $\sigma(K^{pred})$ |
| 1          | [0]       | 10.09                      | 4.61               | 1.73            | 0.649              |
| 2          | [0,1]     | 2.01                       | 0.10               | 2.52            | 0.559              |
| 3          | [0:2]     | 3.98                       | 3.74               | 1.96            | 0.567              |
| 4          | [0:3]     | 6.12                       | 4.44               | 1.51            | 0.628              |
| 5          | [0:4]     | 6.44                       | 3.92               | 1.21            | 0.518              |
| 6          | [0:5]     | 8.76                       | 4.15               | 1.09            | 0.288              |
| 7          | [0:6]     | 8.7                        | 3.58               | 1.17            | 0.403              |
| 8          | [0:7]     | 9.14                       | 3.29               | 1.05            | 0.219              |
| 9          | [0:8]     | 8.85                       | 3.26               | 1.06            | 0.239              |
| 10         | [0:9]     | 10.01                      | 3.48               | 1.08            | 0.273              |

TABLE II: Unsupervised cluster structure function (CSF) (left) and Gap Statistic (right) estimates of the number of unique digits  $K$  in a MNIST digit set. Both CSF and Gap Statistic predictions  $K^{pred}$  are correlated with  $K^{true}$  except in case  $K = 1$  (where both exhibit much higher standard deviation). Omitting  $K^{true} = 1$ , the CSF correlation is 0.93 ( $p = 3e - 4$ ) and the Gap Statistic correlation is  $-0.84$  ( $p = 5e - 3$ ).

is 86% with a 95% confidence interval [0.84,0.88] established by bootstrapping. We used the same procedure on the mean and standard deviation values obtained from the Gap Statistic (as in Table II) and obtained an accuracy of 54% [0.51,0.57].

### B. Cell Segmentation

Cell segmentation is the identification of individual cells in microscopy images. The identification of cell nuclei in microscopy images is an important question. Human stem cells (HSCs) are particularly challenging to segment as the cells are highly adherent, forming in naturally densely packed colonies. HSC colonies, or groups of touching cells, consist of dividing and differentiating cells that present a wide variety of sizes and shapes. The large morphological variation arises from both the presence of cells in developmental states and the mechanical interaction among adjacent cells deforming their shape, texture, and behavior [32], [33]. Timelapse microscopy of living cells further complicates the problem, requiring reduced imaging energy to lessen phototoxicity, and also introducing temporal variations due to imaging as well as cell and colony appearance variability. It is much easier to segment cells that all have a similar appearance, for example shape and size. Here we present a technique for combining multiple simultaneous segmentations of the same image, each with varying underlying segmentation parameters. We refer to the collective set of segmentation results as an ensemble. The segmentations in the ensemble are combined by using optimality deficiency to select among overlapping segmentations. We use a previously described unsupervised underlying segmentation [32], [34], [33] that takes a single parameter of cell size in  $\mu m$ . The method works as follows. The segmentations are run across a range of expected radius values. The results are combined, with cells that overlap each other placed in common "buckets". The question is then to choose the optimal number of cells  $K$  in each bucket. Every segmentation is given a score based on its appearance and how well it captures the underlying pixels.

Here we apply the approach to the question of identifying elliptical cells or nuclei. Rather than using compression-based similarity, the score is built on an appearance model.

The segmentation model expects cells that are convex, brighter in the interior compared to the exterior, and to contain a well defined boundary between a bright interior and dark exterior. Given a particular cell segmentation  $C$ , the score is a combined measure of convex efficiency, background efficiency and boundary efficiency. The term efficiency describes a normalized measure capturing how close to the model the data achieves. The convex efficiency is defined as

$$e_{convex}(C) = \frac{|C|}{|C_{convex}|},$$

where  $|C|$  is the area (volume) of segmentation  $C$  and  $|C_{convex}|$  is the area of the convex hull of  $C$ . The boundary efficiency is computed from the normalized ([0,1]) image pixel values, defined as

$$e_{boundary}(C) = 1 - \text{mean}(R(\beta(C)) - T(\beta(C))),$$

where  $R(\beta(C))$  is the maximal intensity in the region surrounding the boundary voxels  $\beta(C)$ , and  $T(\beta(C))$  is the mean adaptive threshold value for voxels along the boundary. The background efficiency is defined as

$$e_{background}(C) = \frac{\text{mean}(I(C) - T(C))}{\text{mean}(I(\hat{C}) - T(\hat{C}))},$$

where  $I(C)$  is the source image,  $T(C)$  is the adaptive threshold image of segmentation  $C$ , and  $\hat{C}$  represents the image background. The final segmentation score is the sum of the three scores,

$$e_C = e_{convex}(C) + e_{boundary}(C) + e_{background}(C). \quad (\text{V.3})$$

After each cell has been scored, the goal is to select the set of non-overlapping segmentations from the ensemble that maximize the sum of the individual segmentation scores. This



|            |       | Cluster Structure Function |     |     |    |    |    |    |    |    |    | Gap Statistic   |     |    |    |   |   |    |   |    |    |
|------------|-------|----------------------------|-----|-----|----|----|----|----|----|----|----|-----------------|-----|----|----|---|---|----|---|----|----|
|            |       | $K^{predicted}$            |     |     |    |    |    |    |    |    |    | $K^{predicted}$ |     |    |    |   |   |    |   |    |    |
| digit set  |       | 1                          | 2   | 3   | 4  | 5  | 6  | 7  | 8  | 9  | 10 | 1               | 2   | 3  | 4  | 5 | 6 | 7  | 8 | 9  | 10 |
| $K^{true}$ | [0]   | 1                          | 100 | 0   | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 100             | 0   | 0  | 0  | 0 | 0 | 0  | 0 | 0  | 0  |
|            | [0,1] | 2                          | 0   | 100 | 0  | 0  | 0  | 0  | 0  | 0  | 0  | 0               | 100 | 0  | 0  | 0 | 0 | 0  | 0 | 0  | 0  |
|            | [0:2] | 3                          | 0   | 0   | 99 | 1  | 0  | 0  | 0  | 0  | 0  | 0               | 1   | 99 | 0  | 0 | 0 | 0  | 0 | 0  | 0  |
|            | [0:3] | 4                          | 0   | 0   | 1  | 99 | 0  | 0  | 0  | 0  | 0  | 15              | 0   | 27 | 57 | 1 | 0 | 0  | 0 | 0  | 0  |
|            | [0:4] | 5                          | 0   | 0   | 0  | 0  | 99 | 1  | 0  | 0  | 0  | 1               | 0   | 0  | 37 | 9 | 0 | 44 | 4 | 2  | 3  |
|            | [0:5] | 6                          | 0   | 0   | 0  | 0  | 23 | 70 | 7  | 0  | 0  | 0               | 0   | 0  | 2  | 2 | 1 | 39 | 3 | 9  | 44 |
|            | [0:6] | 7                          | 0   | 0   | 0  | 0  | 0  | 3  | 94 | 1  | 1  | 2               | 0   | 0  | 4  | 2 | 1 | 50 | 3 | 10 | 28 |
|            | [0:7] | 8                          | 0   | 0   | 0  | 0  | 0  | 0  | 1  | 58 | 20 | 0               | 0   | 0  | 0  | 2 | 0 | 12 | 6 | 30 | 50 |
|            | [0:8] | 9                          | 0   | 0   | 0  | 0  | 0  | 0  | 1  | 13 | 49 | 0               | 0   | 0  | 0  | 0 | 0 | 2  | 5 | 28 | 65 |
|            | [0:9] | 10                         | 0   | 0   | 0  | 0  | 0  | 0  | 1  | 1  | 5  | 0               | 0   | 0  | 0  | 0 | 0 | 1  | 2 | 8  | 89 |

TABLE III: Supervised cluster structure function (CSF) (left) and Gap Statistic (right) estimates of the number of unique digits  $K$  in a NIST digit set. Each digit set contains 100 digits, split equally among the  $K$  digit classes. The algorithm is given a digit set sampler that can pull repeatedly from the same distribution ( $K$  value) with the goal of estimating  $K$ . The results here were generated by classifying one hundred each of digit sets with  $K^{true} \in [1..10]$ . A 20-element vector consisting of mean and standard deviations of the CSF and the Gap Statistic was the input to a shallow feed-forward neural network. Overall accuracy for the CSF was 86% [0.84,0.88] and 54% for the Gap Statistic [0.51,0.57].

is equivalent to selecting the  $H_S(k)$  from Equation II.2 where the  $\delta(A, x)$  in Equation II.1 are approximated by the individual cell segmentation scores. Figure 3 demonstrates the ensemble segmentation for a colony of HSCs imaged using a fluorescent nuclear marker (H2B).

Quantitative validation for the ensemble segmentation approach was done using ground truth data from the cell tracking challenge [28] reference datasets. Twelve time-lapse datasets in 2-D and 3-D of live cells were processed using the ensemble segmentation with an empirically selected range of radius parameters. Ground truth scores were obtained for each radius parameter setting run separately and also for the ensemble segmentation. We consider the detection (DET) score here, as our concern is not primarily the accuracy of pixel assigned to each segmentation, but rather that we detect the correct number of cells in each frame. We use the training movies for validation because our method is unsupervised and training is not required. Our results are competitive on these movies with the supervised algorithms evaluated on the testing challenge datasets. In each of the 12 movies, the ensemble segmentation outperformed the best result selected from segmentations run separately. The results for the optimality deficiency based ensemble segmentation were statistically significantly better compared to the best score obtained from the single radius segmentation data for both the detection (DET) ( $p = 5e - 4$ , Wilcoxon paired sign-rank test) and tracking (TRA) scores ( $p = 2e - 3$ ). This is significant because the best radius result varied even within pairs of movies from the same application type, showing the value of the ensemble segmentation approach. Table IV shows the results for the ensemble classification as well as the best and worst performing individual segmentation for each of the datasets processed here.

### C. Synthetic Dataset

We evaluate the performance of the cluster structure function using synthetic data generated as random points from  $K = 3$  different 2-D standard normal distributions, each with covariance  $\Sigma = [1, 0; 0, 1]$ . Position the  $K = 3$  clusters along the x-axis at  $x = [0, r, 2 * r]$  with cluster spacing  $r = [0.5 : 0.25 : 1.5]$ . In each of the 100 trials, generate 1e4 points from each of the  $K = 3$  distributions. Supplementary Figure 1 shows a histogram of an example synthetic dataset with cluster spacing = 1.0. To evaluate the cluster structure function, approximate  $K(A) - K(x)$ , as in eqn. II.1 using the Euclidean distance between point  $x$  and the centroid of cluster  $A$ . As in the examples above, we include only the points that fall within one standard deviation of the centroid for each cluster and then average this result across each cluster. We estimate the value of  $K$  using the cluster structure function and compare to results from the Gap statistic, the Akaike Information Criteria (AIC) and the Bayesian Information Criteria (BIC) [26]. The CSF performed significantly better compared to all three alternatives, with the AIC the next closest. The AIC was the only alternative that was competitive with the CSF for this application. Fig. 4 shows results for the CSF and AIC. The good performance of the cluster structure function here follows from the optimality of Euclidean distance used to estimate  $K(A) - K(x)$  as in eqn. II.1.

## VI. SOURCE CODE AVAILABILITY

All of the source code used to generate results in this paper is available open source from <https://git-bioimage.coe.drexel.edu/opensource/ncd>. This includes MATLAB implementations of the NCD and clustering algorithms. There is also limited support for a Python implementation, with ongoing development on that task. The ensemble segmentation algorithms are available at <https://leverjs.net/git>.

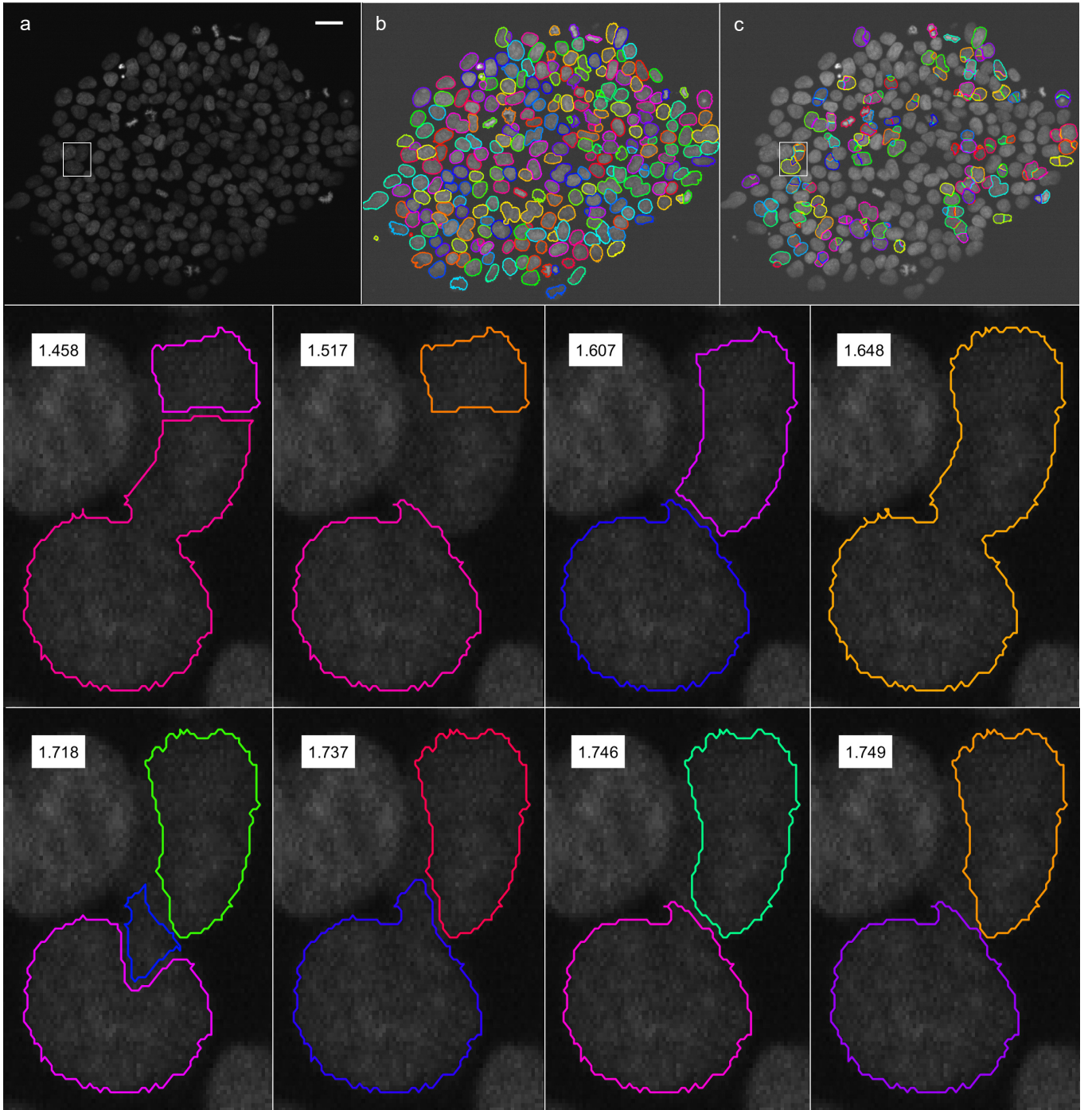


Fig. 3: The ensemble segmentation combines results from different segmentation algorithms using the optimality deficiency to select the best results for overlapping segmentations. Frame segmentations are run at each of a range of different parameter values. The resulting segmentations are each treated as a possible clustering of the underlying pixels into objects. An example is shown here for a single image frame taken from a 1200 frame movie showing the development of live human stem cells (HSCs). The top row shows a raw image (a), the final segmentation results (b) and the overlapping ensemble regions (c). The bottom two rows show different possible combinations of segmentation results from the region shown in the rectangle in (a) and (c). The segmentation results are scored from worst (lowest score) to best (highest score). The optimal set of segmentation results are selected using a greedy optimization to maximize the scores in each overlapping region. Segmentation scores are generated from the convexity, boundary and background efficiencies.

| dataset name       | DET*  | SEG   | TRA   | radius         |
|--------------------|-------|-------|-------|----------------|
| BF-C2DL-HSC_01     | 0.911 | 0.681 | 0.903 | [2.5:0.25:4]   |
| BF-C2DL-HSC_01     | 0.902 | 0.668 | 0.892 | 2.75           |
| BF-C2DL-HSC_01     | 0.817 | 0.593 | 0.801 | 2.5            |
| BF-C2DL-HSC_02     | 0.549 | 0.425 | 0.540 | [2.5:0.25:4]   |
| BF-C2DL-HSC_02     | 0.542 | 0.417 | 0.531 | 2.75           |
| BF-C2DL-HSC_02     | 0.476 | 0.354 | 0.463 | 2.5            |
| Fluo-N2DH-GOWT1_01 | 0.996 | 0.843 | 0.996 | [2.5:0.25:4]   |
| Fluo-N2DH-GOWT1_01 | 0.992 | 0.834 | 0.992 | 2.75           |
| Fluo-N2DH-GOWT1_01 | 0.989 | 0.837 | 0.988 | 4              |
| Fluo-N2DH-GOWT1_02 | 0.923 | 0.859 | 0.923 | [2.5:0.25:4]   |
| Fluo-N2DH-GOWT1_02 | 0.913 | 0.846 | 0.913 | 2.75           |
| Fluo-N2DH-GOWT1_02 | 0.892 | 0.876 | 0.892 | 4              |
| Fluo-N2DH-SIM+_01  | 0.986 | 0.848 | 0.985 | [1.5:0.25:3.5] |
| Fluo-N2DH-SIM+_01  | 0.983 | 0.845 | 0.981 | 2              |
| Fluo-N2DH-SIM+_01  | 0.919 | 0.782 | 0.907 | 3.5            |
| Fluo-N2DH-SIM+_02  | 0.807 | 0.539 | 0.799 | [1.5:0.25:3.5] |
| Fluo-N2DH-SIM+_02  | 0.802 | 0.533 | 0.793 | 2              |
| Fluo-N2DH-SIM+_02  | 0.320 | 0.220 | 0.311 | 3.5            |
| Fluo-N2DL-HeLa_01  | 0.954 | 0.700 | 0.952 | [2.5:0.25:5]   |
| Fluo-N2DL-HeLa_01  | 0.953 | 0.700 | 0.950 | 3.5            |
| Fluo-N2DL-HeLa_01  | 0.944 | 0.670 | 0.938 | 5              |
| Fluo-N2DL-HeLa_02  | 0.912 | 0.786 | 0.909 | [2.5:0.25:5]   |
| Fluo-N2DL-HeLa_02  | 0.910 | 0.782 | 0.905 | 3.25           |
| Fluo-N2DL-HeLa_02  | 0.901 | 0.756 | 0.894 | 5              |

TABLE IV: Ensemble segmentation combines results from segmentation algorithms run at different parameter settings on 2-D and 3-D image data. Optimality deficiency estimates the number of cells  $K$  in each region of overlapping segmentations. The approach here is optimizing the detection (DET) metric for the cell tracking challenge datasets. The first row in each group shows the ensemble results and radius parameter settings, the subsequent two rows show the best and worst performing single segmentations. The ensemble segmentation significantly outperforms the best individual segmentations ( $p = 5e - 4$ ).

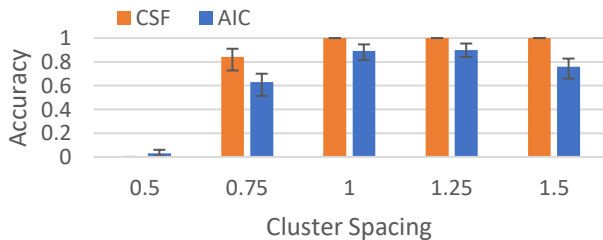


Fig. 4: Estimating the number of clusters in data generated from  $K = 3$  normal distributions, all with  $\Sigma = [1, 0; 0, 1]$ . The distributions are located along the X axis at multiples  $[0, 1, 2]$ . \* Cluster Spacing. The cluster structure function (CSF) significantly outperforms the Akaike Information Criteria (AIC). Error bars show 95% confidence intervals from bootstrapping.

## VII. ACKNOWLEDGEMENTS

Portions of this work were supported by NIH NIA (R01AG041861) and by the Human Frontiers Science Program (RGP0043/2019-203). The authors wish to thank Prof. Rafael Carazo Salas from the Univ. of Bristol UK and his group for providing sample HSC image data.

## REFERENCES

- [1] P. Adriaans and P.M.B. Vitányi, Approximation of the two-part MDL code, *IEEE Trans. Inform. Theory*, 55:1(2009), 444–457.
- [2] M.R. Anderberg, *Cluster Analysis for Applications*, Academic Press, 2014.
- [3] D. Bradley, G. Roth, Adapting Thresholding Using the Integral Image, *Journal of Graphics Tools*, 12:2(2007), 13–21.
- [4] R.L. Cilibrasi, P.M.B. Vitányi, Clustering by compression, *IEEE Trans. Inform. Theory*, 51:4(2005), pp. 1523–1545.
- [5] A.R. Cohen, C. Björnsson, S. Temple, G. Banker and B. Roysam, Automatic summarization of changes in biological image sequences using algorithmic information theory, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31:8(2009), 1386–1403.
- [6] A.R. Cohen, F. Gomes, B. Roysam, M. Cayouette, Computational prediction of neural progenitor cell fates, *Nature Methods*, 7:3(2010), 213–218.

- [7] A.R. Cohen and P.M.B. Vitányi, Normalized compression distance of multisets with applications, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 37:8(2015), 1602–1614.
- [8] T.M. Cover and J.A. Thomas, *Elements of Information Theory*, Wiley, New York, 1991.
- [9] M. Ester, H.P. Kriegel, J. Sander, and X. Xu, A density-based algorithm for discovering clusters in large spatial databases with noise, Proc. 2nd Int. Conf. Knowledge Discovery and Data Mining, 1996.
- [10] P. Gács, On the symmetry of algorithmic information, *Soviet Math. Dokl.*, 15(1974), 1477–1480. Correction: *Ibid.*, 15 (1974) 1480.
- [11] P. Gács, J.T. Tromp, P.M.B. Vitányi, Algorithmic statistics, *IEEE Trans. Inform. Theory*, 47:6(2001), 2443–2463.
- [12] P.D. Grünwald, *The Minimum Description Length Principle*, MIT Press, 2007.
- [13] R.A. Fisher, On the mathematical foundations of theoretical statistics, *Phil. Trans. Royal Soc. London, Ser. A*, 222(1922), 309–368.
- [14] K. Jain, R.C. Dubes, *Algorithms for clustering data*, Prentice-Hall, NJ, USA, 1988.
- [15] S.D. Kamvar, D. Klein, C. D. Manning, Spectral Learning, *Proc. 18th Intl Joint Conference on Artificial Intelligence*, pp. 561–566, 2003.
- [16] L. Kaufman, P.J. Rousseeuw, *Finding Groups in Data: an Introduction to Cluster Analysis*, Wiley-Interscience, 2009.
- [17] A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission*, 1:1(1965), 1–7.
- [18] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, Gradient-based learning applied to document recognition. *Proc. IEEE*, 86:11(1998), 2278–2324.
- [19] L.A. Levin, Laws of information conservation (non-growth) and aspects of the foundation of probability theory, *Problems Inform. Transmission*, 10(1974), 206–210.
- [20] M. Li, X. Chen, X. Li, B. Ma, P.M.B. Vitányi, The similarity metric, *IEEE Trans. Inform. Th.*, 50:12(2004), 3250–3264.
- [21] M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and Its Applications*, 4th Ed., Springer, New York, 2019.
- [22] A. Milovanov, Algorithmic statistics, prediction and machine learning, Proc. 33rd Symp. Theoret. Aspects Comput. Sci., (STACS) LIPICs 47(2016), 54:1–54:13.
- [23] A. Ng, M.I. Jordan, Y. Weiss, On spectral clustering: analysis and an algorithm, NIPS’01: Proc. 14th Int. Conf. Neural Information Processing Systems: Natural and Synthetic, 2001, 849–856.
- [24] A.Kh. Shen, The concept of  $(\alpha, \beta)$ -stochasticity in the Kolmogorov sense, and its properties, *Soviet Math. Dokl.*, 28:1(1983), 295–299.
- [25] J. Sneyers and P. Wuille, FLIF: Free lossless image format based on MANIAC compression, Proc. IEEE Int. Conf. Image Processing (ICIP), 2016, 66–70.
- [26] S. Theodoridis, K. Koutroubas, *Pattern Recognition*, 4th Ed., Academic Press, 2008.
- [27] R. Tibshirani, G. Walther and T. Hastie, Estimating the number of clusters in a dataset via the gap statistic, *J. Royal Stat. Soc.*, 63(2001), 411–423.
- [28] V. Ulman, M. Maška, K.E.G. Magnusson, O. Ronneberger, et al., An objective comparison of cell-tracking algorithms, *Nature Methods*, 14:12(2017), 1141–1152.
- [29] P.M.B. Vitányi, Meaningful information, *IEEE Trans. Inform. Theory*, 52:10(2006), 4617–4626.
- [30] P.M.B. Vitányi, Information distance in multiples, *IEEE Trans. Inform. Theory*, 57:4(2011), 2451–2456.
- [31] N.K. Vereshchagin and P.M.B. Vitányi, Kolmogorov’s Structure functions and model selection, *IEEE Trans. Inform. Theory*, 50:12(2004), 3265–3290.
- [32] M. Winter, M. Liu, D. Monteleone, et al., Computational image analysis reveals intrinsic multigenerational differences between anterior and posterior cerebral cortex neural progenitor cells, *Stem Cell Reports*, 5(2015), 609–620.
- [33] M. Winter, W. Mankowski, E. Wait, et al., Separating touching cells using pixel replicated elliptical shape models, *IEEE Trans. Medical Imaging*, 38:4(2008), 883–893.
- [34] M. Winter, W. Mankowski, E. Wait, et al., LEVER: software tools for segmentation, tracking and lineaging of proliferating cells, *Bioinformatics*, 2016.
- [35] M. Winter, W. Mankowski, E. Wait, E.C.D.L. Hoz, A. Aguinaldo, A.R. Cohen, Separating Touching Cells using Pixel Replicated Elliptical Shape Models, *IEEE Trans Medical Imaging*, 38:4(2018), 883–893.
- [36] S. Wade S. Z. Ghahramani, Bayesian cluster analysis: Point estimation and credible balls (with discussion) *Bayesian Analysis*, 2018, 13(2): 559–626.
- [37] F. K. Teklehaymanot, M. Muma, A.M. Zoubir, Bayesian cluster enumeration criterion for unsupervised learning *IEEE Transactions on Signal Processing*, 2018, 66(20): 5392–5406.
- [38] D. Valle, Y. Jameel, B. Betancourt, E.T. Azeria, N. Attias, J. Cullen, Automatic selection of the number of clusters using Bayesian clustering and sparsity-inducing priors. *Ecol Appl.* 2022 Apr;32(3):e2524. doi: 10.1002/eap.2524. Epub 2022 Feb 22. PMID: 34918421.
- [39] W. Fu, P.O. Perry, Estimating the number of clusters using cross-validation, *Journal of Computational and Graphical Statistics*, 2020, 29(1): 162–173.
- [40] M. Rahman, M. Masud, B. Mazumder, "Estimation of the Number of Clusters based on Simplicial Depth." *2020 2nd International Conference on Sustainable Technologies for Industry*, IEEE, 2020.
- [41] P. Bloem, F. Mota, S. de Rooij, L. Antunes, P. Adriaans, (2014). A Safe Approximation for Kolmogorov Complexity. *Algorithmic Learning Theory. ALT 2014. Lecture Notes in Computer Science*, vol 8776.

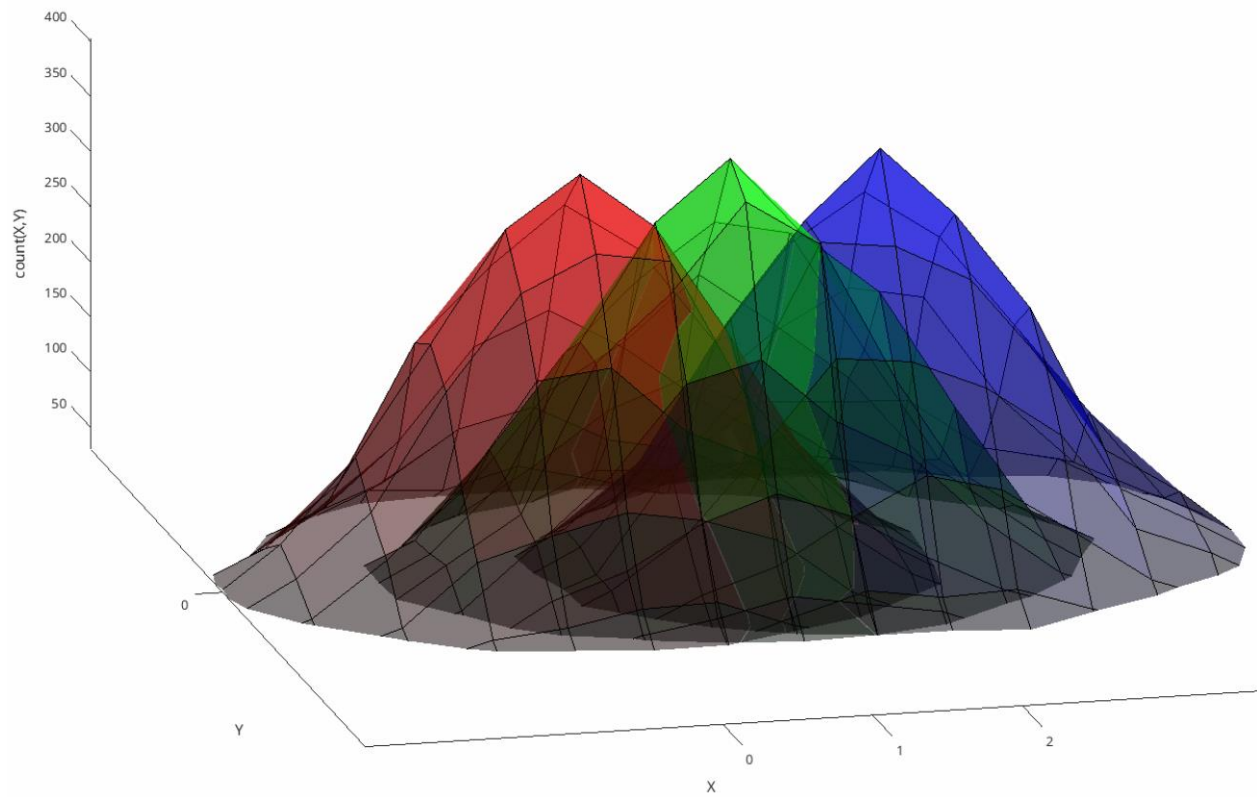


**Andrew R. Cohen** received his Ph.D. from the Rensselaer Polytechnic Institute in May 2008. He is currently an associate professor in the department of Electrical & Computer Engineering at Drexel University. Prior to joining Drexel, he was an assistant professor in the department of Electrical Engineering and Computer Science at the University of Wisconsin, Milwaukee. He has worked as a software design engineer at Microsoft Corp. on the Windows and DirectX teams and as a CPU Product Engineer at Intel Corp. His research interests include 5-D image sequence analysis for applications in biological microscopy, algorithmic information theory, spectral methods, data visualization, and supercomputer applications. He is a senior member of the IEEE.



**Paul M.B. Vitányi** received his Ph.D. from the Free University of Amsterdam (1978). He is a CWI Fellow at the national research institute for mathematics and computer science in the Netherlands, CWI, and Professor of Computer Science at the University of Amsterdam. He served on the editorial boards of Distributed Computing, Information Processing Letters, Theory of Computing Systems, Parallel Processing Letters, International journal of Foundations of Computer Science, Entropy, Information, Journal of Computer and Systems Sciences (guest editor), and elsewhere. He has worked on cellular automata, computational complexity, distributed and parallel computing, machine learning and prediction, physics of computation, Kolmogorov complexity, information theory, quantum computing, publishing more than 200 research papers and some books. He received a Knighthood (Ridder in de Orde van de Nederlandse Leeuw) and is member of the Academia Europaea. Together with Ming Li they pioneered applications of Kolmogorov complexity and co-authored “An Introduction to Kolmogorov Complexity and its Applications,” Springer-Verlag, New York, 1993 (3rd Edition 2008), parts of which have been translated into Chinese, Russian and Japanese.





Supplementary Figure 1: Histogram for synthetic dataset containing 3 clusters. Here cluster spacing equals 1.0. Each cluster is shown in red, green, blue. The data was generated from a mixture of 3 normal distribution with means  $\mu = [0, 1, 2]$  and identical covariance  $\Sigma = [1, 0; 0, 1]$ .