Clustering processes

Daniil Ryabko   DANIIL@RYABKO.NET

INRIA Lille, 40 avenue Halley, 59650 Villeneuve d'Ascq, France

Abstract

The problem of clustering is considered, for the case when each data point is a sample generated by a stationary ergodic process. We propose a very natural asymptotic notion of consistency, and show that simple consistent algorithms exist, under most general non-parametric assumptions. The notion of consistency is as follows: two samples should be put into the same cluster if and only if they were generated by the same distribution. With this notion of consistency, clustering generalizes such classical statistical problems as homogeneity testing and process classification. We show that, for the case of a known number of clusters, consistency can be achieved under the only assumption that the joint distribution of the data is stationary ergodic (no parametric or Markovian assumptions, no assumptions of independence, neither between nor within the samples). If the number of clusters is unknown, consistency can be achieved under appropriate assumptions on the mixing rates of the processes. In both cases we give examples of simple (at most quadratic in each argument) algorithms which are consistent.

1. Introduction

Given a finite set of objects, the problem is to "cluster" similar objects together. This intuitively simple goal is notoriously hard to formalize. Most of the work on clustering is concerned with particular parametric data generating models, or particular algorithms, a given similarity measure, and (very often) a given number of clusters. It is clear that, as in almost learning problems, in clustering finding the right similarity measure is an integral part of the problem. However, even if one assumes the similarity measure known, it is hard to define what a good clustering is (Kleinberg, 2002;

Zadeh & Ben-David, 2009). What is more, even if one assumes the similarity measure to be simply the Euclidean distance (on the plane), and the number of clusters k known, then clustering may still appear intractable for computational reasons. Indeed, in this case finding k centres (points which minimize the cumulative distance from each point in the sample to one of the centres) seems to be a natural goal, but this problem is NP-hard (Mahajan et al., 2009).

In this work we concentrate on a subset of the clustering problem: clustering processes. That is, each data point is itself a sample generated by a certain discrete-time stochastic process. This version of the problem has numerous applications, such as clustering biological data, financial observations, or behavioural patterns, and as such it has gained a tremendous attention in the literature.

The main observation that we make in this work is that, in the case of clustering processes, one can benefit from the notion of *ergodicity* to define what appears to be a very natural notion of consistency. This notion of consistency is shown to be satisfied by simple algorithms that we present, which are polynomial in all arguments. This can be achieved without any modeling assumptions on the data (e.g. Hidden Markov, Gaussian, etc.), without assuming independence of any kind within or between the samples. The only assumption that we make is that the joint distribution of the data is stationary ergodic. The assumption of stationarity means, intuitively, that the time index itself bares no information: it does not matter whether we have started recording observations at time 0 or at time 100. By virtue of the ergodic theorem, any stationary process can be represented as a mixture of stationary ergodic processes. In other words, a stationary process can be thought of as first selecting a stationary ergodic process (according to some prior distribution) and then observing its outcomes. Thus, the assumption that the data is stationary ergodic is both very natural and rather weak. At the same time, ergodicity means that, in asymptotic, the properties of the process can be learned from observation.

This allows us to define the clustering problem as follows. N samples are given: $\mathbf{x}_1 =$

$(x_1^1, \ldots, x_{n_1}^1), \ldots, \mathbf{x}_N = (x_1^N, \ldots, x_{n_N}^N)$. Each sample is drawn by one out of k different stationary ergodic distributions. The samples are *not* assumed to be drawn independently; rather, it is assumed that the joint distribution of the samples is stationary ergodic. The target clustering is as follows: those and only those samples are put into the same cluster that were generated by the same distribution. The number k of target clusters can be either known or unknown (different consistency results can be obtained in these cases). A clustering algorithm is called asymptotically consistent if the probability that it outputs the target clustering converges to 1, as the lengths $(n_1, \ldots, n_N)$ of the samples tend to infinity (a variant of this definition is to require the algorithm to stabilize on the correct answer with probability 1). Note the particular regime of asymptotic: not with respect to the number of samples N, but with respect to the length of the samples $n_1, \ldots, n_N$.

Similar formulations have appeared in the literature before. Perhaps the most close approach is mixture models (Smyth, 1997; Zhong & Ghosh, 2003): it is assumed that there are k different distributions that have a particular known form (such as Gaussian, Hidden Markov models, or graphical models) and each one out of N samples is generated independently according to one of these k distributions (with some fixed probability). Since the model of the data is specified quite well, one can use likelihood-based distances (and then, for example, the k-means algorithm), or Bayesian inference, to cluster the data. Clearly, the main difference from our setting is in that we do not assume any known model of the data; not even between-sample independence is assumed.

The problem of clustering in our formulation generalizes two classical problems of mathematical statistics. The first one is homogeneity testing, or the two-sample problem. Two samples $\mathbf{x}_1 = (x_1^1, \ldots, x_{n_1}^1)$ and $\mathbf{x}_2 = (x_1^2, \ldots, x_{n_2}^2)$ are given, and it is required to test whether they were generated by the same distribution, or by different distributions. This corresponds to clustering just two data points ($N = 2$) with the number k of clusters unknown: either $k = 1$ or $k = 2$. The second problem is process classification, or the three-sample problem. Three samples $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3$ are given, it is known that two of them were generated by the same distribution, while the third one was generated by a different distribution. It is required to find out which two were generated by the same distribution. This corresponds to clustering three data points, with the number of clusters known: $k = 2$. The classical approach is of course to consider Gaussian i.i.d. data, but general non-parametric solutions

exist not only for i.i.d. data (Lehmann, 1986), but also for Markov chains (Gutman, 1989), and under certain mixing rates conditions. What is important for us here, is that the three-sample problem is easier than the two-sample problem; the reason is that k is known in the latter case but not in the former. Indeed, in (Ryabko, 2010b) it is shown that in general, for stationary ergodic (binary-valued) processes, there is no solution to the two-sample problem, even in the weakest asymptotic sense. However, a solution to the three-sample problem, for (real-valued) stationary ergodic processes was given in (Ryabko & Ryabko, 2010).

In this work we demonstrate that, if the number k of clusters is known, then there is an asymptotically consistent clustering algorithm, under the only assumption that the joint distribution of data is stationary ergodic. If k is unknown, then in this general case there is no consistent clustering algorithm (as follows from the mentioned result for the two-sample problem). However, if an upper-bound α_n on the α-mixing rates of the joint distribution of the processes is known, and $\alpha_n \to 0$, then there is a consistent clustering algorithm. Both algorithms are rather simple, and are based on the empirical estimates of the so-called distributional distance. For two processes ρ_1, ρ_2 a distributional distance d is defined as $\sum_{k=1}^{\infty} w_k |\rho_1(B_k) - \rho_2(B_k)|$, where w_k are positive summable real weights, e.g. $w_k = 2^{-k}$, and B_k range over a countable field that generates the sigma-algebra of the underlying probability space. For example, if we are talking about finite-alphabet processes with the binary alphabet $A = \{0, 1\}$, B_k would range over the set $A^* = \cup_{k \in \mathbb{N}} A^k$; that is, over all tuples $0, 1, 00, 01, 10, 11, 000, 001, \ldots$ (of course, we could just as well omit, say, 1 and 11); therefore, the distributional distance in this case is the weighted sum of differences of probabilities of all possible tuples. In this work we consider real-valued processes, so B_k have to range through a suitable sequence of intervals, all pairs of such intervals, triples, etc. (see the formal definitions below). This distance has proved a useful tool for solving various statistical problems concerning ergodic processes (Ryabko & Ryabko, 2010; Ryabko, 2010a).

Although this distance involves infinite summation, we show that its empirical approximations can be easily calculated. For the case of a known number of clusters, the proposed algorithm (which is shown to be consistent) is as follows. (The distance in the algorithms is a suitable empirical estimate of d.) The first sample is assigned to the first cluster. For each $j = 2..k$, find a point that maximizes the minimal distance to those points already assigned to clusters, and assign it to the cluster j. Thus we have one point in each of the k clus-

ters. Next, assign each of the remaining points to the cluster that contains the closest points from those k already assigned. For the case of an unknown number of clusters k, the algorithm simply puts those samples together that are not farther away from each other than a certain threshold level, where the threshold is calculated based on the known bound on the mixing rates. In this case, besides the asymptotic result, finite-time bounds on the probability of outputting an incorrect clustering can be obtained. Each of the algorithms is shown to be at most quadratic in each argument.

Therefore, we show that for the proposed notion of consistency, there are simple algorithms that are consistent under most general assumptions. While these algorithms can be easily implemented, we have left the problem of trying them out on particular applications, as well as optimizing the parameters, for future research. It may also be suggested that the empirical distributional distance can be replaced by other distances, for which similar theoretical results can be obtained. An interesting direction, that could preserve the theoretical generality, would be to use data compressors. These were used in (Ryabko & Astola, 2006) for the related problems of hypotheses testing, leading both to theoretical and practical results. As far as clustering is concerned, compression-based methods were used (without asymptotic consistency analysis) in (Cilibrasi & Vitanyi, 2005), and (in a different way) in (Bagnall et al., 2006). Combining our consistency framework with these compression-based methods is a promising direction for further research.

2. Preliminaries

Let A be an alphabet, and denote A^* the set of tuples $\cup_{i=1}^{\infty} A^i$. In this work we consider the case $A = \mathbb{R}$; extensions to the multidimensional case, as well as to more general spaces, are straightforward. Distributions, or (stochastic) processes, are measures on the space $(A^{\infty}, \mathcal{F}_{A^{\infty}})$, where $\mathcal{F}_{A^{\infty}}$ is the Borel sigma-algebra of A^{∞}. When talking about joint distributions of N samples, we mean distributions on the space $((A^N)^{\infty}, \mathcal{F}_{(A^N)^{\infty}})$.

For each $k, l \in \mathbb{N}$, let $B^{k,l}$ be the partition of the set A^k into k-dimensional cubes with volume $h_l^k = (1/l)^k$ (the cubes start at 0). Moreover, define $B^k = \cup_{l \in \mathbb{N}} B^{k,l}$ and $\mathcal{B} = \cup_{k=1}^{\infty} B^k$. The set $\{B \times A^{\infty} : B \in B^{k,l}, k, l \in \mathbb{N}\}$ generates the Borel σ-algebra on $\mathbb{R}^{\infty} = A^{\infty}$. For a set $B \in \mathcal{B}$ let $|B|$ be the index k of the set B^k that B comes from: $|B| = k : B \in B^k$.

We use the abbreviation $X_{1..k}$ for $X_1, \ldots, X_k$. For a sequence $\mathbf{x} \in A^n$ and a set $B \in \mathcal{B}$ denote $\nu(\mathbf{x}, B)$ the

frequency with which the sequence $\mathbf{x}$ falls in the set B.

$$\nu(\mathbf{x}, B) :=$$
$$\begin{cases} \frac{1}{n-|B|+1} \sum_{i=1}^{n-|B|+1} I_{\{(X_i, \ldots, X_{i+|B|-1}) \in B\}} & \text{if } n \geq |B|, \\ 0 & \text{otherwise.} \end{cases}$$

A process ρ is *stationary* if $\rho(X_{1..|B|} = B) = \rho(X_{t..t+|B|-1} = B)$ for any $B \in A^*$ and $t \in \mathbb{N}$. We further abbreviate $\rho(B) := \rho(X_{1..|B|} = B)$. A stationary process ρ is called *(stationary) ergodic* if the frequency of occurrence of each word B in a sequence $X_1, X_2, \ldots$ generated by ρ tends to its a priori (or limiting) probability a.s.: $\rho(\lim_{n \to \infty} \nu(X_{1..n}, B) = \rho(B)) = 1$. Denote $\mathcal{E}$ the set of all stationary ergodic processes.

Definition 1 (distributional distance). *The distributional distance is defined for a pair of processes ρ_1, ρ_2 as follows (e.g. (Gray, 1988))*

$$d(\rho_1, \rho_2) = \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{m,l}} |\rho_1(B) - \rho_2(B)|,$$

where $w_j = 2^{-j}$.

(The weights in the definition are fixed for the sake of concreteness only; we could take any other summable sequence of positive weights instead.) In words, we are taking a sum over a series of partitions into cubes of decreasing volume (indexed by l) of all sets A^k, $k \in \mathbb{N}$, and count the differences in probabilities of all cubes in all these partitions. These differences in probabilities are weighted: smaller weights are given to larger k and finer partitions. It is easy to see that d is a metric. We refer to (Gray, 1988) for more information on this metric and its properties.

The clustering algorithms presented below are based on *empirical estimates of the distance d*:

$$\hat{d}(X_{1..n_1}^1, X_{1..n_2}^2) =$$
$$\sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{m,l}} |\nu(X_{1..n_1}^1, B) - \nu(X_{1..n_2}^2, B)|, \quad (1)$$

where $n_1, n_2 \in \mathbb{N}$, $\rho \in \mathcal{S}$, $X_{1..n_i}^i \in A^{n_i}$.

Although the expression (1) involves taking three infinite sums, it will be shown below that it can be easily calculated.

Lemma 1 ($\hat{d}$ is consistent). *Let $\rho_1, \rho_2 \in \mathcal{E}$ and let two samples $\mathbf{x}_1 = X_{1..n_1}^1$ and $\mathbf{x}_2 = X_{1..n_2}^2$ be generated by a distribution ρ such that the marginal distribution of $X_{1..n_1}^i$ is ρ_i, $i = 1, 2$, and the joint distribution ρ is stationary ergodic. Then*

$$\lim_{n_1, n_2 \to \infty} \hat{d}(X_{1..n_1}^1, X_{1..n_2}^2) = d(\rho_1, \rho_2) \ \rho\text{-a.s.}$$

Proof. The idea of the proof is simple: for each set $B \in \mathcal{B}$, the frequency with which the sample $\mathbf{x}_1$ falls into B converges to the probability $\rho_1(B)$, and analogously for the second sample. When the sample sizes grow, there will be more and more sets $B \in \mathcal{B}$ whose frequencies have already converged to the probabilities, so that the cumulative weight of those sets whose frequencies have not converged yet, will tend to 0.

For any $\varepsilon > 0$ we can find an index J such that $\sum_{i,j=J}^{\infty} w_i w_j < \varepsilon/3$. Moreover, for each m,l we can find such elements $B_1^{m,l}, \ldots, B_{t_{m,l}}^{m,l}$, for some $t_{m,l} \in \mathbb{N}$, of the partition $B^{m,l}$ that $\rho_i(\cup_{i=1}^{t_{m,l}} B_i^{m,l}) \geq 1 - \varepsilon/6Jw_m w_l$. For each $B_j^{m,l}$, where $m,l \leq J$ and $j \leq t_{m,l}$, we have $\nu((X_1^1, \ldots, X_{n_1}^1), B_j^{m,l}) \to \rho_1(B_j^{m,l})$ a.s., so that

$$|\nu((X_1^1, \ldots, X_{n_1}^1), B_j^{m,l}) - \rho_1(B_j^{m,l})|$$
$$< \rho_1(B_j^{m,l})\varepsilon/(6Jw_j)$$

for all $n_1 \geq u$, for some $u \in \mathbb{N}$; define $U_j^{m,l} := u$. Let $U := \max_{m,l \leq J, j \leq t_{m,l}} U_j^{m,l}$ (U depends on the realization $X_1^1, X_2^1, \ldots$). Define analogously V for the sequence $(X_1^2, X_2^2, \ldots)$. Thus for $n_1 > U$ and $n_2 > V$ we have

$$|\hat{d}(\mathbf{x}_1, \mathbf{x}_2) - d(\rho_1, \rho_2)| =$$
$$\left| \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{k,l}} (|\nu(\mathbf{x}_1, B) - \nu(\mathbf{x}_2, B)| - |\rho_1(B) - \rho_2(B)|) \right|$$
$$\leq \sum_{m,l=1}^{\infty} w_m w_l \sum_{B \in B^{k,l}} w_i(|\nu(\mathbf{x}_1, B) - \rho_1(B)| + |\nu(\mathbf{x}_2, B) - \rho_2(B)|)$$
$$\leq \sum_{m,l=1}^{J} w_m w_l \sum_{i=1}^{t_{k,l}} (|\nu(\mathbf{x}_1, B_i^{m,l}) - \rho_1(B_i^{m,l})|$$
$$+ |\nu(\mathbf{x}_2, B_i^{m,l}) - \rho_2(B_i^{m,l})|) + 2\varepsilon/3$$
$$\leq \sum_{m,l=1}^{J} w_m w_l \sum_{i=1}^{t_{k,l}} (\rho_1(B_i^{m,l})\varepsilon/(6Jw_m w_l)$$
$$+ \rho_2(B_i^{m,l})\varepsilon/(6Jw_m w_l)) + 2\varepsilon/3 \leq \varepsilon,$$

which proves the statement. $\square$

3. Main results

The clustering problem can be defined as follows. We are given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$, where each sample $\mathbf{x}_i$ is a string of length n_i of symbols from A: $\mathbf{x}_i = X_{1..n_i}^i$. Each sample is generated by one out of k different *unknown* stationary ergodic distributions $\rho_1, \ldots, \rho_k \in \mathcal{E}$. Thus, there is a partitioning $I = \{I_1, \ldots, I_k\}$ of the set $\{1..N\}$ into k *disjoint* subsets $I_j, j = 1..k$

$$\{1..N\} = \cup_{j=1}^{k} I_j,$$

such that $\mathbf{x}_j$, $1 \leq j \leq N$ is generated by ρ_j if and only if $j \in I_j$. The partitioning I is called the *target clustering* and the sets $I_i, 1 \leq i \leq k$, are called the *target clusters*. Given samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and a target clustering I, let $I(\mathbf{x})$ denote the cluster that contains $\mathbf{x}$.

A *clustering function* F takes a finite number of samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ and an optional parameter k (the target number of clusters) and outputs a partition $F(\mathbf{x}_1, \ldots, \mathbf{x}_N, (k)) = \{T_1, \ldots, T_k\}$ of the set $\{1..N\}$.

Definition 2 (asymptotic consistency). *Let a finite number N of samples be given, and let the target clustering partition be I. Define $n = \min\{n_1, \ldots, n_N\}$. A clustering function F is strongly asymptotically consistent if $F(\mathbf{x}_1, \ldots, \mathbf{x}_N, (k)) = I$ from some n on with probability 1. A clustering function is weakly asymptotically consistent if $P(F(\mathbf{x}_1, \ldots, \mathbf{x}_N, (k)) = I) \to 1$.*

Note that the consistency is asymptotic with respect to *the minimal length of the sample*, and not with respect to the *number of samples*.

3.1. Known number of clusters

Algorithm 1 is a simple clustering algorithm, which, given the number k of clusters, will be shown to be consistent under most general assumptions. It works as follows. The point $\mathbf{x}_1$ is assigned to the first cluster. Next, find the point that is farthest away from $\mathbf{x}_1$ in the empirical distributional distance $\hat{d}$, and assign this point to the second cluster. For each $j = 3..k$, find a point that maximizes the minimal distance to those points already assigned to clusters, and assign it to the cluster j. Thus we have one point in each of the k clusters. Next simply assign each of the remaining points to the cluster that contains the closest points from those k already assigned. One can notice that Algorithm 1 is just one iteration of the k-means algorithm, with so-called farthest-point initialization (Katsavounidis et al., 1994), and a specially designed distance.

Algorithm 1 The case of known number of clusters k

INPUT: The number of clusters k, samples $\mathbf{x}_1, \ldots, x_N$.

Initialize: $j := 1$, $c_1 := 1$, $T_1 := \{x_{c_1}\}$.

for $j := 2$ to k **do**

$\quad c_j := \text{argmax}\{i = 1, \ldots, N : \min_{t=1}^{j-1} \hat{d}(\mathbf{x}_i, \mathbf{x}_{c_t})\}$

$\quad T_j := \{x_{c_j}\}$

end for

for $i = 1$ to N **do**

$\quad$ Put $\mathbf{x}_i$ into the set $T_{\text{argmin}_{j=1}^{k} \hat{d}(\mathbf{x}_i, \mathbf{x}_{c_j})}$

end for

OUTPUT: the sets T_j, $j = 1..k$.

Proposition 1 (calculating $\hat{d}(\mathbf{x}_1, \mathbf{x}_2)$)**.** *For two samples* $\mathbf{x}_1 = X^1_{1..n_1}$ *and* $\mathbf{x}_2 = X^2_{1..n_2}$ *the computational complexity (time and space) of calculating the empirical distributional distance* $\hat{d}(\mathbf{x}_1, \mathbf{x}_2)$ *(1) is* $O(n^2 \log s_{\min}^{-1})$, *where* $n = \max(n_1, n_2)$ *and*

$$s_{\min} = \min_{i=1..n_1, j=1..n_2, X^1_i \neq X^2_j} |X^1_i - X^2_j|.$$

Proof. First, observe that for fixed m and l, the sum

$$T^{m,l} := \sum_{B \in B^{m,l}} |\nu(X^1_{1..n_1}, B) - \nu(X^2_{1..n_2}, B)| \qquad (2)$$

has not more than $n_1 + n_2 - 2m + 2$ non-zero terms (assuming $m \leq n_1, n_2$; the other case is obvious). Indeed, for each $i = 0, 1$, in the sample $\mathbf{x}_i$ there are $n_i - m + 1$ tuples of size k: $X^i_{1..m}, X^i_{2..m+1}, \ldots, X^i_{n_1-m+1..n_1}$. Therefore, the complexity of calculating $T^{m,l}$ is $O(n_1 + n_2 - 2m + 2) = O(n)$. Furthermore, observe that for each m, for all $l > \log s_{\min}^{-1}$ the term $T^{m,l}$ is constant. Therefore, it is enough to calculate $T^{m,1}, \ldots, T^{m, \log s_{\min}^{-1}}$, since for fixed m

$$\sum_{l=1}^{\infty} w_m w_l T^{m,l} = w_m w_{\log s_{\min}^{-1}} T^{m, \log s_{\min}^{-1}} + \sum_{l=1}^{\log s_{\min}^{-1}} w_m w_l T^{m,l}$$

(that is, we double the weight of the last non-zero term). Thus, the complexity of calculating $\sum_{l=1}^{\infty} w_m w_l T^{m,l}$ is $O(n \log s_{\min}^{-1})$. Finally, for all $m > n$ we have $T^{m,l} = 0$. Since $\hat{d}(\mathbf{x}_1, \mathbf{x}_2) = \sum_{m,l=1}^{\infty} w_m, w_l T^{m,l}$, the statement is proven. □

Theorem 1. *Let* $N \in \mathbb{N}$ *and suppose that the samples* $\mathbf{x}_1, \ldots, \mathbf{x}_N$ *are generated in such a way that the joint distribution is stationary ergodic. If the correct number of clusters k is known, then Algorithm 1 is strongly asymptotically consistent. Algorithm 1 makes* $O(kN)$ *calculations of* $\hat{d}(\cdot, \cdot)$, *so that its computational complexity is* $O(kN n_{\max}^2 \log s_{\min}^{-1})$, *where* $n_{\max} = \max_{i=1}^k n_i$ *and*

$$s_{\min} = \min_{u,v=1..N, u \neq v, i=1..n_u, j=1..n_v, X^u_i \neq X^v_j} |X^u_i - X^v_j|.$$

Observe that the samples are not required to be generated independently. The only requirement on the distribution of samples is that the joint distribution is stationary ergodic. This is perhaps one of the mildest possible probabilistic assumptions.

Proof. By Lemma 1, $\hat{d}(\mathbf{x}_i, \mathbf{x}_j)$, $i, j \in \{1..N\}$ converges to 0 if and only if $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same cluster. Since there are only finitely many samples $\mathbf{x}_i$, there exists some $\delta > 0$ such that, from some n on, we will have $\hat{d}(\mathbf{x}_i, \mathbf{x}_j) < \delta$ if $\mathbf{x}_i, \mathbf{x}_j$ belong to the same

target cluster ($I(\mathbf{x}_i) = I(\mathbf{x}_j)$), and $\hat{d}(\mathbf{x}_i, \mathbf{x}_j) > \delta$ otherwise ($I(\mathbf{x}_i) \neq I(\mathbf{x}_j)$). Therefore, from some n on, for every $j \leq k$ we will have $\max\{i = 1, \ldots, N : \min_{t=1}^{j-1} \hat{d}(\mathbf{x}_i, \mathbf{x}_{c_t})\} > \delta$ and the sample $\mathbf{x}_{c_j}$, where $c_j = \operatorname{argmax}\{i = 1, \ldots, N : \min_{t=1}^{j-1} \hat{d}(\mathbf{x}_i, \mathbf{x}_{c_t})\}$, will be selected from a target cluster that does not contain any $\mathbf{x}_{c_i}$, $i < j$. The consistency statement follows.

Next, let us find how many pairwise distance estimates $\hat{d}(\mathbf{x}_i, \mathbf{x}_j)$ the algorithm has to make. On the first iteration of the loop, it has to calculate $\hat{d}(\mathbf{x}_i, \mathbf{x}_{c_1})$ for all $i = 1..N$. On the second iteration, it needs again $\hat{d}(\mathbf{x}_i, \mathbf{x}_{c_1})$ for all $i = 1..N$, which are already calculated, and also $\hat{d}(\mathbf{x}_i, \mathbf{x}_{c_2})$ for all $i = 1..N$, and so on: on jth iteration of the loop we need to calculate $d(\mathbf{x}_i, \mathbf{x}_{c_j})$, $i = 1..N$, which gives at most kN pairwise distance calculations in total. The statement about computational complexity follows from this and Proposition 1: indeed, apart from the calculation of $\hat{d}$, the rest of the computations is of order $O(kN)$. □

Complexity–precision trade–off. The bound on the computational complexity of Algorithm 1, given in Theorem 1, is given for the case of *precisely* calculated distance estimates $\hat{d}(\cdot, \cdot)$. However, precise estimates are not needed if we only want to have an asymptotically consistent algorithm. Indeed, following the proof of Lemma 1, it is easy to check that if we replace in (1) the infinite sums with sums over any number of terms m_n, l_n that grows to infinity with $n = \min(n_1, n_2)$, and if we replace partitions $B^{m,l}$ by their (finite) subsets $B^{m,l,n}$ which increase to $B^{m,l}$, then we still have a consistent estimate of $d(\cdot, \cdot)$.

Definition 3 ($\check{d}$)**.** *Let* m_n, l_n *be some sequences of numbers,* $B^{m,l,n} \subset B^{m,l}$ *for all* $m, l, n \in \mathbb{N}$, *and denote* $n := \min\{n_1, n_2\}$. *Define*

$$\check{d}(X^1_{1..n_1}, X^2_{1..n_2}) := \sum_{m=1}^{m_n} \sum_{l=1}^{l_n} w_m w_l$$
$$\sum_{B \in B^{m,l,n}} |\nu(X^1_{1..n_1}, B) - \nu(X^2_{1..n_2}, B)|. \quad (3)$$

Lemma 2 ($\check{d}$ *is consistent*)**.** *Assume the conditions of Lemma 1. Let* l_n *and* m_n *be any sequences of integers that go to infinity with n, and let, for each* $m, l \in \mathbb{N}$, *the sets* $B^{m,l,n}$, $n \in \mathbb{N}$ *be an increasing sequence of subsets of* $B^{m,l}$, *such that* $\cup_{n \in \mathbb{N}} B^{m,l,n} = B^{m,l}$. *Then*

$$\lim_{n_1, n_2 \to \infty} \check{d}(X^1_{1..n_1}, X^2_{1..n_2}) = d(\rho_1, \rho_2) \ \rho\text{-}a.s..$$

The proof is analogous to that of Lemma 1.

If we use the estimate $\check{d}(\cdot,\cdot)$ in Algorithm 1 (instead of $\hat{d}(\cdot,\cdot)$), then we still get an asymptotically consistent clustering function. Thus the following statement holds true.

Proposition 2. *Assume the conditions of Theorem 1. For all sequences m_n, l_n of numbers that increase to infinity with n, there is a strongly asymptotically consistent clustering algorithm, whose computational complexity is $O(kNn_{\max}m_{n_{\max}}l_{n_{\max}})$.*

On the one hand, Proposition 2 can be thought of as an artifact of the asymptotic definition of consistency; on the other hand, in practice precise calculation of $\check{d}(\cdot,\cdot)$ is hardly necessary. What we get from Proposition 2 is the possibility to select the appropriate trade–off between the computational burden, and the precision of clustering before asymptotic.

Note that the bound in Proposition 2 does not involve the sizes of the sets $B^{m,l,n}$; in particular, one can take $B^{m,l,n} = B^{m,l}$ for all n. This is because, for every two samples $X_{1..n}^1$ and $X_{1..n}^2$, this sum has no more than $2n$ non-zero terms, whatever are m, l. However, in the following section, where we are after clustering with an unknown number of clusters k, and thus after controlled rates of convergence, the sizes of the sets $B^{m,l,n}$ will appear in the bounds.

3.2. Unknown number of clusters

So far we have shown that when the number of clusters is known in advance, consistent clustering is possible under the only assumption that the joint distribution of the samples is stationary ergodic. However, under this assumption, in general, consistent clustering with unknown number of clusters is impossible. Indeed, as was shown in (Ryabko, 2010b), when we have only two *binary-valued* samples, generated *independently* by two stationary ergodic distributions, it is impossible to decide whether they have been generated by the same or by different distributions, even in the sense of weak asymptotic consistency (this holds even if the distributions come from a smaller class: the set of all B-processes). Therefore, if the number of clusters is unknown, we have to settle for less, which means that we have to make stronger assumptions on the data. What we need is known rates of convergence of frequencies to their expectations. Such rates are provided by assumptions on the mixing rates of the distribution generating the data. Here we will show that under rather mild assumptions on the mixing rates (and, again, without any modeling assumptions or assumptions of independence), consistent clustering is possible when the number of clusters is unknown.

In this section we assume that all the samples are $[0,1]$-valued (that is, $X_i^j \in [0,1]$); extension to arbitrary bounded (multidimensional) ranges is straightforward. Next we introduce *mixing coefficients*, mainly following (Bosq, 1996) in formulations. Informally, mixing coefficients of a stochastic process measure how fast the process forgets about its past. Any one-way infinite stationary process $X_1, X_2, \ldots$ can be extended backwards to make a two-way infinite process $\ldots, X_{-1}, X_0, X_1, \ldots$ with the same distribution. In the definition below we assume such an extension. Define the α mixing coefficients as

$$\alpha(n) = \sup_{A \in \sigma(\ldots, X_{-1}, X_0), B \in \sigma(X_n, X_{n+1}, \ldots))} |P(A \cap B) - P(A)P(B)|, \quad (4)$$

where $\sigma(..)$ stays for the sigma-algebra generated by random variables in brackets. These coefficients are non-increasing. A process is called strongly α-*mixing* if $\alpha(n) \to 0$. Many important classes of processes satisfy the mixing conditions. For example, if a process is a stationary irreducible aperiodic Hidden Markov process, then it is α-mixing. If the underlying Markov chain is finite-state, then the coefficients decrease exponentially fast. Other probabilistic assumptions can be used to obtain bounds on the mixing coefficients, see e.g. (Bradley, 2005) and references therein.

Algorithm 2 is very simple. Its inputs are: samples $\mathbf{x}_1, \ldots, x_N$; the threshold level $\delta \in (0,1)$, the parameters $m, l \in \mathbb{N}$, $B^{m,l,n}$. The algorithm assigns to the same cluster all samples which are at most δ-far from each other, as measured by $\check{d}(\cdot,\cdot)$. The estimate $\check{d}(\cdot,\cdot)$ can be calculated in the same way as $\hat{d}(\cdot,\cdot)$ (see Proposition 1 and its proof). We do not give a pseudo code implementation of this algorithm, since it's rather obvious.

The idea is that the threshold level δ is selected according to the minimal length of a sample and the (known bounds on) mixing rates of the process ρ generating the samples (see Theorem 2).

The next theorem shows that, if the joint distribution of the samples satisfies $\alpha(n) \le \alpha_n \to 0$, where α_n are known, then one can select (based on α_n only) the parameters of Algorithm 2 in such a way that it is weakly asymptotically consistent. Moreover, a bound on the probability of error before asymptotic is provided.

Theorem 2 (Algorithm 2 is consistent, unknown k)**.** *Fix sequences $\alpha_n \in (0,1)$, $m_n, l_n, b_n \in \mathbb{N}$, and let $B^{m,l,n} \subset B^{m,l}$ be an increasing sequence of finite sets, for each $m, l \in \mathbb{N}$. Set $b_n := \max_{l \le l_n, m \le m_n} |B^{m,l,n}|$. Let also $\delta_n \in (0,1)$. Let $N \in \mathbb{N}$ and suppose that the samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$ are generated in such a way*

that the (unknown) joint distribution ρ is stationary ergodic, and satisfies $\alpha_n(\rho) \leq \alpha_n$, for all $n \in \mathbb{N}$. Then for every sequence $q_n \in [0..n/2]$, Algorithm 2, with the above parameters, satisfies

$$\rho(T \neq I) \leq 2N(N+1)(m_n l_n b_n \gamma_n(\delta_n) + \gamma_n(\varepsilon_\rho)) \quad (5)$$

where

$$\gamma(\delta) = (2e^{-q_n\delta^2/32} + 11(1 + 4/\delta)^{1/2} q_n \alpha_{(n-2m_n)/2q_n}),$$

T is the partition output by the algorithm, I is the target clustering, ε_ρ is a constant that depends only on ρ, and $n = \min_{i=1..N} n_i$.

In particular, if $\alpha_n = o(1)$, then, selecting the parameters in such a way that $\delta_n = o(1)$, $q_n, m_n, l_n, b_n = o(n)$, $q_n, m_n, l_n \to \infty$, $\cup_{k \in \mathbb{N}} B^{m,l,k} = B^{m,l}$, $b_n^{m,l} \to \infty$, for all $m, l \in \mathbb{N}$, and, finally,

$$m_n l_n b_n (e^{-q_n\delta_n^2} + \delta_n^{-1/2} q_n \alpha_{(n-2m_n)/2q_n}) = o(1),$$

as is always possible, Algorithm 2 is weakly asymptotically consistent *(with the number of clusters k unknown). The computational complexity of Algorithm 2 is $O(N^2 m_{n_{\max}} l_{n_{\max}} b_{n_{\max}})$, and is bounded by $O(N^2 n_{\max}^2 \log s_{\min}^{-1})$, where $n_{\max}$ and $\log s_{\min}^{-1}$ are defined as in Theorem 1.*

Proof. We use the following bound from (Bosq, 1996): for any zero-mean random process $Y_1, Y_2, \ldots,$ every $n \in \mathbb{N}$ and every $q \in [1..n/2]$ we have

$$P\left(|\sum_{i=1}^{n} Y_i| > n\varepsilon\right)$$
$$\leq 4\exp(-q\varepsilon^2/8) + 22(1 + 4/\varepsilon)^{1/2} q\alpha(n/2q).$$

For every $j = 1..N$, every $m < n$, $l \in \mathbb{N}$, and $B \in B^{m,l}$, define the processes $Y_1^j, Y_2^j, \ldots,$ where

$$Y_t^j := \mathbb{I}_{(X_t^j, \ldots, X_{t+m-1}^j) \in B} - \rho(X_{1..m}^j \in B).$$

It is easy to see that α-mixing coefficients for this process satisfy $\alpha(n) \leq \alpha_{n-2m}$. Thus,

$$\rho(|\nu(X_{1..n_j}^j, B) - \rho(X_{1..m}^j \in B)| > \varepsilon/2) \leq \gamma_n(\varepsilon) \quad (6)$$

Then for every $i, j \in [1..N]$ such that $I(\mathbf{x}_i) = I(\mathbf{x}_j)$ (that is, $\mathbf{x}_i$ and $\mathbf{x}_j$ are in the same cluster) we have

$$\rho(|\nu(X_{1..n_i}^i, B) - \nu(X_{1..n_j}^j, B)| > \varepsilon) \leq 2\gamma_n(\varepsilon).$$

Using the union bound, summing over m, l, and B, we obtain

$$\rho(\check{d}(\mathbf{x}_i, \mathbf{x}_j) > \varepsilon) \leq 2m_n l_n b_n \gamma_n(\varepsilon). \quad (7)$$

Next, let i, j be such that $I(\mathbf{x}_i) \neq I(\mathbf{x}_j)$. Then, for some $m_{i,j}, l_{i,j} \in \mathbb{N}$ there is $B_{i,j} \in B^{m_{i,j}, l_{i,j}}$ such that $|\rho(X_{1..|B_{i,j}|}^i \in B_{i,j}) - \rho(X_{1..|B_{i,j}|}^j \in B_{i,j})| > 2\tau_{i,j}$ for some $\tau_{i,j} > 0$. Then for every $\varepsilon < \tau_{i,j}/2$ we have

$$\rho(|\nu(X_{1..n_i}^i, B_{i,j}) - \nu(X_{1..n_j}^j, B_{i,j})| < \varepsilon) \leq$$
$$\rho(|\nu(X_{1..n_i}^i, B_{i,j}) - \rho(X_{1..|B|}^i \in B_{i,j})| > \tau_{i,j})$$
$$+ \rho(|\nu(X_{1..n_j}^j, B_{i,j}) - \rho(X_{1..|B_{i,j}|}^j \in B_{i,j})| > \tau_{i,j})$$
$$\leq 2\gamma_n(\tau_{i,j}). \quad (8)$$

Moreover, for $\varepsilon < w_{m_{i,j}} w_{l_{i,j}} \tau_{i,j}/2$

$$\rho(\check{d}(\mathbf{x}_i, \mathbf{x}_j) > \varepsilon) \leq 2\gamma_n(w_{m_{i,j}} w_{l_{i,j}} \tau_{i,j}). \quad (9)$$

Define $\varepsilon_\rho := \min_{i,j=1..N:I(\mathbf{x}_i) \neq I(\mathbf{x}_j)} w_{m_{i,j}} w_{l_{i,j}} \tau_{i,j}/2$. Clearly, from this and (8), for every $\varepsilon < 2\varepsilon_\rho$ we obtain

$$\rho(\check{d}(\mathbf{x}_i, \mathbf{x}_j) > \varepsilon) \leq 2\gamma_n(\varepsilon_\rho). \quad (10)$$

If, for every pair i, j of samples, $\check{d}(\mathbf{x}_i, \mathbf{x}_j) < \delta_n$ if and only if $I(\mathbf{x}_i) = I(\mathbf{x}_j)$, then Algorithm 2 gives a correct answer. Therefore, taking the bounds (7) and (10) together for each of the $N(N+1)/2$ pairs of samples, we obtain (5). The complexity statement can be established analogously to that in Theorem 1. $\square$

While Theorem 2 shows that α-mixing with a known bound on the coefficients is sufficient to achieve asymptotic consistency, the bound (5) on the probability of error includes as multiplicative terms all the parameters m_n, l_n and b_n of the algorithm, which can make it large for practically useful choices of the parameters. The multiplicative factors are due to the fact that we take a bound on the divergence of each individual frequency of each cell of each partition from its expectation, and then take a union bound over all of these. To obtain a more realistic performance guarantee, we would like to have a bound on the divergence of *all* the frequencies of all cells of a given partition from their expectations. Such uniform divergence estimates are possible under stronger assumptions; namely, they can be established under some assumptions on β-mixing coefficients, which are defined as follows

$$\beta(n) = \mathbf{E} \sup_{B \in \sigma(X_n, \ldots))} |P(B) - P(B|\sigma(\ldots, X_0))|.$$

These coefficients satisfy $2\alpha(n) \leq \beta(n)$ (see e.g. (Bosq, 1996)), so assumptions on the speed of decrease of β-coefficients are stronger. Using the uniform bounds given in (Karandikara & Vidyasagar, 2002), one can obtain a statement similar to that in Theorem 2, with α-mixing replaced by β-mixing, and without the multiplicative factor b_n.

4. Conclusion

We have proposed a framework for defining consistency of clustering algorithms, when the data comes as a set of samples drawn from stationary processes. The main advantage of this framework is its generality: no assumptions have to be made on the distribution of the data, beyond stationarity and ergodicity. The proposed notion of consistency is so simple and natural, that it may be suggested to be used as a basic sanity-check for all clustering algorithms that are used on sequence-like data. For example, it is easy to see that the k-means algorithm will be consistent with some initializations (e.g. with the one used in Algorithm 1) but not with others (e.g. not with the random one).

While the algorithms that we presented to demonstrate the existence of consistent clustering methods are computationally efficient and easy to implement, the main value of the established results is theoretical. As it was mentioned in the introduction, it can be suggested that for practical applications empirical estimates of the distributional distance can be replaced with distances based on data compression, in the spirit of (Ryabko & Astola, 2006; Cilibrasi & Vitanyi, 2005; Ryabko, 2009).

Another direction for future research concerns optimal bounds on the speed of convergence: while we show that such bounds can be obtained (of course, only in the case of known mixing rates), finding practical and tight bounds, for different notions of mixing rates, remains open.

Finally, here we have only considered the setting in which the number N of samples is fixed, while the asymptotic is with respect to the lengths of the samples. For on-line clustering problems, it would be interesting to consider the formulation where both N and the lengths of the samples grow.

References

A. Bagnall, C. Ratanamahatana, E. Keogh, S. Lonardi, G. Janacek. A bit level representation for time series data mining with shape based similarity. *Data Mining and Knowledge Discovery*, 13(1): 11–40, 2006.

Bosq, D. *Nonparametric Statistics for Stochastic Processes*. Estimation and Prediction. Springer, 1996.

Bradley, R.C. Basic properties of strong mixing conditions. A survey and some open questions. *Probability Surveys*, 2:107–144, 2005.

Cilibrasi, R. and Vitanyi, P.M.B. Clustering by compression. *IEEE Trans. Inf. Th.*, 51:1523–1545, 2005.

Gray, R. *Probability, Random Processes, and Ergodic Properties*. Springer Verlag, 1988.

Gutman, M. Asymptotically optimal classification for multiple tests with empirically observed statistics. *IEEE Trans. Inf. Th.*, 35(2):402–408, 1989.

Karandikara, R.L. and Vidyasagar, M. Rates of uniform convergence of empirical means with mixing processes. *Stat.&Prob. Lett.*, 58:297–307, 2002.

Katsavounidis, I., Kuo, C.-C. Jay, and Zhang, Zhen. A new initialization technique for generalized Lloyd iteration. *IEEE Signal Proc. Let.*, 1:144–146, 1994.

Kleinberg, J. An impossibility theorem for clustering. In *NIPS*:446–453, 2002.

Lehmann, E. *Testing Statistical Hypotheses, 2nd ed.*. Wiley, New York, 1986.

Mahajan, M., Nimbhorkar, P., and Varadarajan, K. The planar k-means problem is NP-hard. In *WALCOM*: 274–285, 2009.

Ryabko, B. Compression-based methods for nonparametric prediction and estimation of some characteristics of time series. *IEEE Trans. Inf. Th.*, 55:4309–4315, 2009.

Ryabko, B. and Astola, J. Universal codes as a basis for time series testing. *Stat. Methodology*, 3:375–397, 2006.

Ryabko, D. Testing composite hypotheses about discrete-valued stationary processes. In *ITW*: 291–295, 2010a.

Ryabko, D. Discrimination between B-processes is impossible. *J. Theor. Prob.*, 23(2):565–575, 2010b.

Ryabko, D. and Ryabko, B. Nonparametric statistical inference for ergodic processes. *IEEE Trans. Inf. Th.*, 56(3):1430–1435, 2010.

Smyth, P. Clustering sequences with hidden Markov models. In *NIPS*: 648–654. 1997.

Zadeh, R. and Ben-David, S. A uniqueness theorem for clustering. In *UAI*, 2009.

Zhong, S. and Ghosh, J. A unified framework for model-based clustering. *JMLR*, 4:1001–1037, 2003.