Scaling Analysis of Affinity Propagation

Cyril Furtlehner,* Michèle Sebag,† and Xiangliang Zhang‡

(Dated: June 9, 2010)

We analyze and exploit some scaling properties of the *Affinity Propagation* (AP) clustering algorithm proposed by Frey and Dueck (2007). Following a divide and conquer strategy we setup an exact renormalization-based approach to address the question of clustering consistency, in particular, how many cluster are present in a given data set. We first observe that the divide and conquer strategy, used on a large data set hierarchically reduces the complexity $\mathcal{O}(N^2)$ to $\mathcal{O}(N^{(h+2)/(h+1)})$, for a data-set of size N and a depth h of the hierarchical strategy. For a data-set embedded in a d-dimensional space, we show that this is obtained without notably damaging the precision except in dimension $d = 2$. In fact, for d larger than 2 the relative loss in precision scales like $N^{(2-d)/(h+1)d}$. Finally, under some conditions we observe that there is a value s^* of the penalty coefficient, a free parameter used to fix the number of clusters, which separates a fragmentation phase (for $s < s^*$) from a coalescent one (for $s > s^*$) of the underlying hidden cluster structure. At this precise point holds a self-similarity property which can be exploited by the hierarchical strategy to actually locate its position, as a result of an exact decimation procedure. From this observation, a strategy based on AP can be defined to find out how many clusters are present in a given dataset.

I. INTRODUCTION

Clustering techniques are useful data-mining tools in Machine Learning with many applications in biology, astrophysics, pattern recognition, library archiving and more generally for data processing. The question is how to partition an ensemble of objects such that similar ones pertain to the same classes. A precise statement of the problem requires the definition of a similarity measure between objects and of a cost function. As such, it turns out to be an optimization problem, which is generally NP-Hard. Many algorithms have been proposed, ranging from expectation-maximization (EM) types approaches [1] like k-centers and k-means [2] to percolation-like methods for building hierarchies. From the statistical physics viewpoint depending on the form of the cost function, the clustering solution may be reformulated as the ground sate of a q-states Potts model which can be solved by Monte-Carlo based methods [3]. This type of models are suitable for Bethe-Peierls approximations, which algorithmic counterpart is known to be the belief-propagation (BP) algorithm of Pearl [4, 5]. This algorithm was initially introduced in the context of Bayesian inference, but for optimization problems this has a well defined zero temperature limit, the so-called min-sum algorithm [6].

Considering a relaxed version of the cost function were clusters are identified by exemplars, and only the similarity of data to their exemplars are taken into account, Frey and Dueck have recently proposed the *affinity propagation* algorithm [7] as an instance of the min-sum algorithm to solve the clustering problem. Their algorithm turns out to be very efficient compared to other center-based methods like k-centers and k-means by avoiding of getting stuck into some local minimum when the size of the dataset increases. The price to pay for these understandability and stability properties is a quadratic computational complexity, except if the similarity matrix is made sparse with help of a pruning procedure. Nevertheless, a pre-treatment of the data would also be quadratic in the number if item, which is severely hindering the usage of AP on large scale datasets. The basic assumption behind AP, is that cluster are of spherical shape. This limiting assumption has actually been addressed by Leone and co-authors in [8, 9], by softening a hard constraint present in AP, which impose that any exemplar has first to point to itself as oneself exemplar. Another drawback, which is actually common to most clustering techniques, is that there is a free parameter to fix which ultimately determines the number of clusters. Some methods based on EM [10] or on information-theoretic consideration have been proposed [11], but mainly use a precise parametrization of the cluster model. There exists also a different strategy based on similarity statistics [12], that have been already recently combined with AP [13], at the expense of a quadratic price.

In an earlier work [14, 15], a hierarchical approach, based on a divide and conquer strategy was proposed, to decrease the AP complexity and adapt AP to the context of Data Streaming. In this paper we basically analyze in greater details the combination of AP with a divide and conquer strategy from a scaling point of view and, realizing that this can be restated as an exact renormalization procedure by decimation of the dataset, we define a new stability criteria to asses the validity of the clustering solution. The main results of the paper are on one hand, the information loss estimation when AP is combined with this hierarchical procedure, and a simple recipe to identify (almost for free after the hierarchical treatment) the number of cluster present in the dataset,

*INRIA-Saclay, LRI Bât. 490, F-91405 Orsay(France)
†CNRS, LRI Bât. 490, F-91405 Orsay(France)
‡INRIA, Université Paris Sud, LRI Bât. 490, F-91405 Orsay(France)

on the other hand.
The paper is organized as follows. In Section II we start from a brief description of BP and some of its properties. We summarize how AP and its extension, soft constraint affinity propagation (SCAP), originate from it. Then in Section III, we define our hierarchical approach HI-AP and analyze its computational complexity. In Section IV we compute the leading behavior, of the resulting error measured on the distribution of exemplars, which depends on the dimension and on the size of the subsets. Based on these results we enforce the self-similarity of HI-AP in Section V to develop a renormalized version of AP (in the statistical physics sense) and discuss how to fix in a self-consistent way the penalty coefficient, conjugate to the number of clusters, present in AP. Finally Section VI is devoted to experimental tests: we present proof of principle of this method on artificial dataset and analyze its robustness and limits on real-world dataset.

II. INTRODUCTION TO BELIEF-PROPAGATION AND AP

A. Local marginal computation

The belief propagation algorithm is intended to computing marginals of joint-probability measure of the type

$$P(\mathbf{x}) = \prod_a \psi_a(x_a) \prod_i \phi(x_i), \tag{II.1}$$

where $\mathbf{x} = (x_1, \dots, x_N)$ is a set of variables, $x_a = \{x_i, i \in a\}$ a subset of variables involved in the factor ψ_a, while the ϕ_i's are single variable factors. The structure of the joint measure P_a is conveniently represented by a factor graph [6], i.e. a bipartite graph with two set of vertices, $\mathcal{F}$ associated to the factors, and $\mathcal{V}$ associated to the variables, and a set of edges $\mathcal{E}$ connecting the variables to their factors. Computing the single variables marginals scales in general exponentially with the size of the system, except when the underlying factor graph has a tree like structure. In that case all the single site marginals may be computed at once, by solving the following iterative scheme due to J. Pearl [4]:

$$m_{a\to i}(x_i) \longleftarrow \sum_{\substack{x_j \\ j\in a, j\neq i}} \psi_a(x_a) \prod_j n_{j\to a}(x_j)$$

$$n_{i\to a}(x_i) \longleftarrow \phi_i(x_i) \prod_{b\ni i, b\neq a} m_{b\to i}(x_i).$$

$m_{a\to i}(x_i)$ is called the message sent by factor node a to variable node i, while $n_{i\to a}(x_i)$ is the message sent by variable node i to a. These quantities would actually appear as intermediate computations terms, while deconditioning (II.1). On a singly connected factor graph, starting from the leaves, two sweeps are sufficient to obtain the fixed points messages, and the beliefs (the local marginals) are then obtained from these sets of messages using the formulas:

$$b_i(x_i) = \frac{1}{Z_i}\phi_i(x_i) \prod_{a\ni i} m_{a\to i}(x_i)$$

$$b_a(x_a) = \frac{1}{Z_a}\psi_a(x_a) \prod_{i\in a} n_{i\to a}(x_i)$$

with Z_i and Z_a insuring normalization of the beliefs. On a multiply connected graph, this scheme can be used as an approximate procedure to compute the marginals, still reliable on sparse factor graph, while avoiding the exponential complexity of an exact procedure. Many connections with mean field approaches of statistical physics have been recently unravelled, in particular the connection with the TAP equations introduced in the context of spin glasses [16], and with the Bethe approximation of the free energy[5].

B. AP and SCAP as min-sum algorithms

The AP algorithm is a message-passing procedure proposed by Frey and Dueck [7] that performs a classification by identifying exemplars. It solves the following optimization problem

$$\mathbf{c}^* = \operatorname{argmin}\big(E[\mathbf{c}]\big),$$

with

$$E[\mathbf{c}] \stackrel{\text{def}}{=} -\sum_{i=1}^{N} S(i, c_i) - \sum_{\mu=1}^{N} \log \chi_\mu[\mathbf{c}] \tag{II.2}$$

where $\mathbf{c} = (c_1, \dots, c_N)$ is the mapping between data and exemplars, $S(i, c_i)$ is the similarity function between i and its exemplar. For datapoints embedded in an Euclidean space, the common choice for S is the negative squared Euclidean distance. A free positive parameter is given by

$$s \stackrel{\text{def}}{=} -S(i, i), \qquad \forall i,$$

the penalty for being oneself exemplar. $\chi_\mu^{(p)}[\mathbf{c}]$ is a set of constraints. They read

$$\chi_\mu[\mathbf{c}] = \begin{cases} p, & \text{if } c_\mu \neq \mu, \exists i \text{ s.t. } c_i = \mu, \\ 1, & \text{otherwise.} \end{cases}$$

$p = 0$ is the constraint of the model of Frey-Dueck. Note that this strong constraint is well adapted to well-balanced clusters, but probably not to ring-shape ones. For this reason Leone et. al. [8, 9] have introduced the smoothing parameter p. Introducing the inverse temperature β,

$$P[\mathbf{c}] \stackrel{\text{def}}{=} \frac{1}{Z} \exp(-\beta E[\mathbf{c}])$$

represents a probability distribution over clustering assignments c. At finite β the classification problem reads

$$\mathbf{c}^* = \mathrm{argmax}\big(P[\mathbf{c}]\big).$$

The AP or SCAP equations can be obtained from the standard BP equation [7, 8] as an instance of the Max-Product algorithm. For self-containess, let us sketch the derivation here. The BP algorithm provides an approximate procedure to the evaluation of the set of single marginal probabilities $\{P_i(c_i = \mu)\}$ while the min-sum version obtained after taking $\beta \to \infty$ yields the affinity propagation algorithm of Frey and Dueck. The factor-graph involves variable nodes $\{i, i = 1 \dots N\}$ with corresponding variable c_i and factor nodes $\{\mu, \mu = 1 \dots N\}$ corresponding to the energy terms and to the constraints (see Figure II.1). Let $A_{\mu\to i}(c_i)$ the message sent by factor μ to variable i and $B_{i\to\mu}(c_i)$ the message sent by variable i to node μ. The belief propagation fixed point equations read:

$$A_{\mu\to i}(c_i = c) = \frac{1}{Z_{\mu\to i}} \sum_{\{c_j\}} \prod_{j\neq i} B_{j\to\mu}(c_j) \chi_\mu^\beta[\{c_j\}, c] \tag{II.3}$$

$$B_{i\to\mu}(c_i = c) = \frac{1}{Z_{i\to\mu}} \prod_{\nu\neq\mu} A_{\nu\to i}(c) e^{\beta S(i,c)} \tag{II.4}$$

Once this scheme has converged, the fixed points messages provide a consistency relationship between the two sets of beliefs

$$b_\mu[\{c_i\} = \mathbf{c}] = \frac{1}{Z_\mu} \chi_\mu^\beta[\mathbf{c}] \prod_{i=1}^N B_{i\to\mu}(c_i) \tag{II.5}$$

$$b_i(c_i = c) = \frac{1}{Z_i} \prod_{\mu=1}^N A_{\mu\to i}[c] e^{\beta S(i,c)} \tag{II.6}$$

The joint probability measure then rewrites

$$P[\mathbf{c}] = \frac{1}{Z_b} \frac{\prod_{\mu=1}^N b_\mu[\mathbf{c}]}{\prod_{i=1}^N b_i^{N-1}(c_i)}$$

with Z_b the normalization constant associated to this set of beliefs. In (II.3) we observe first that

$$\hat{A}_{\mu\to i} \stackrel{\text{def}}{=} A_{\mu\to i}(c_i = \nu \neq \mu), \tag{II.7}$$

is independent of ν and secondly that $A_{\mu\to i}(c_i = c)$ depends only on $B_{j\to\mu}(c_j = \mu)$ and on $\sum_{\nu\neq\mu} B_{j\to\mu}(c_j = \nu)$. This means that the scheme can be reduced to the propagation of four quantities, by letting

$$A_{\mu\to i} \stackrel{\text{def}}{=} A_{\mu\to i}(c_i = \mu),$$

$$\hat{A}_{\mu\to i} \stackrel{\text{def}}{=} \frac{1 - A_{\mu\to i}}{N - 1}$$

$$B_{i\to\mu} \stackrel{\text{def}}{=} B_{i\to\mu}(c_i = \mu)$$

$$\bar{B}_{i\to\mu} \stackrel{\text{def}}{=} 1 - B_{i\to\mu},$$

which reduce to two types of messages $A_{\mu\to i}$ and $B_{i\to\mu}$. At this point we introduce the log-probability ratios,

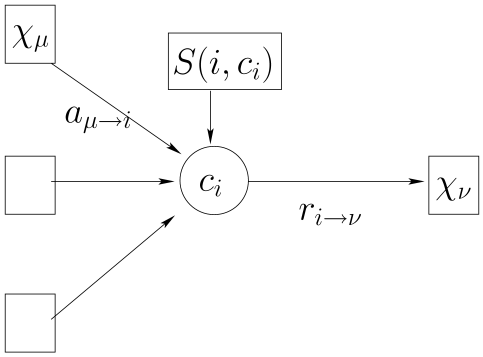

FIG. II.1: Factor graph corresponding to AP. Small squares represents the constraints while large ones are associated to pairwise contributions in $E(\mathbf{c})$.

$$a_{\mu\to i} \stackrel{\text{def}}{=} \frac{1}{\beta} \log\Big(\frac{A_{\mu\to i}}{\hat{A}_{\mu\to i}}\Big),$$

$$r_{i\to\mu} \stackrel{\text{def}}{=} \frac{1}{\beta} \log\Big(\frac{B_{i\to\mu}}{\bar{B}_{i\to\mu}}\Big),$$

corresponding respectively to the "availability" and "responsibility" messages of Frey-Dueck. with $q \stackrel{\text{def}}{=} -\frac{1}{\beta}\log p$. Taking the limit $\beta \to \infty$ at fixed q yields

$$a_{\mu\to i} = \min\Big(0, \max\big(-q, \min(0, r_{\mu\to\mu})\big) + \sum_{j\neq i} \max(0, r_{j\to\mu})\Big),$$

$$\mu \neq i, \tag{II.8}$$

$$a_{i\to i} = \min\Big(q, \sum_{j\neq i} \max(0, r_{j\to i})\Big), \tag{II.9}$$

$$r_{i\to\mu} = S(i,\mu) - \max_{\nu\neq\mu}\big(a_{\nu\to i} + S(i,\nu)\big). \tag{II.10}$$

After reaching a fixed point, exemplars are obtained according to

$$c_i^* = \mathrm{argmax}_\mu\big(S(i,\mu) + a_{\mu\to i}\big) = \mathrm{argmax}_\mu\big(r_{i\to\mu} + a_{\mu\to i}\big). \tag{II.11}$$

Altogether, II.8,II.9,II.10 and II.11 constitute the equations of SCAP which reduce to the equations of AP when q tends to $-\infty$.

III. DECREASING THE COMPLEXITY OF AFFINITY PROPAGATION

As already mentioned the AP computational complexity is expected to scale like $\mathcal{O}(N^2)$; it involves the matrix S

of pair distances, with quadratic complexity in the number N of items, severely hindering its use on large-scale datasets[17]. This AP limitation which is for example not adapted to streaming of data, can be overcome through a Divide-and-Conquer heuristics inspired from [18], which we have proposed in [14, 15]. Let us describe here this approach.

A. Hierarchical affinity propagation

The basic procedure goes as follows: The dataset $\mathcal{E}$ of size N is randomly partitioned into $\sqrt{N}$ subsets. AP is launched on every subset and outputs a set of exemplars, which in turn are clustered to yield the final result. The complexity is then $\sqrt{N} \times (\sqrt{N})^2 = N^{3/2}$ as long as the number of exemplars K produced by each individual subset is $o(N^{1/4})$ because the last clustering step costs $(K\sqrt{N})^2$ in complexity. This Divide-and-Conquer strategy could be actually combined with any other basic clustering algorithm, and this procedure can be easily extended to have more than one hierarchical levels, thus reducing further the computational cost as follows.

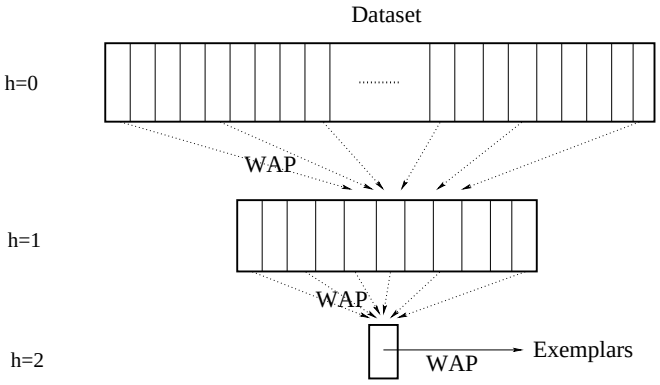

FIG. III.2: Sketch of the HI-AP procedure for 2 hierarchical levels. At each elementary clustering steps, items are weighted in proportion to what they represent as exemplars, i.e. WAP is in use instead of AP.

Let h be the total number of hierarchical levels, starting at $h = 0$ for the basic dataset so that by convention HI-AP with $h = 0$ simply reduces to AP. At each level of the hierarchy, the penalty parameter s is set such that the expected number of exemplars extracted along each clustering step is a constant K. If b is the ratio of subset number between two hierarchical levels, $M = N/b^h$ is the size of each subset to be clustered at level h; at level $h-1$, each clustering problem thus involves $bK = M$ exemplars with corresponding complexity

$$C(0) = K^2\big(\frac{N}{K}\big)^{\frac{2}{h+1}}.$$

The total number N_{cp} of clustering procedures involved is

$$N_{cp} = \sum_{i=0}^{h} b^i = \frac{b^{h+1}-1}{b-1},$$

with overall computational complexity:

$$C(h) = K^2\big(\frac{N}{K}\big)^{\frac{2}{h+1}} \frac{\frac{N}{K}-1}{\big(\frac{N}{K}\big)^{\frac{1}{h+1}}-1} \underset{N\gg K}{\approx} K^2\big(\frac{N}{K}\big)^{\frac{h+2}{h+1}}.$$

It is seen that $C(0) = N^2$, $C(1) \propto N^{3/2}$,..., and $C(h) \propto N$ for $h \gg 1$.
Note that this procedure is naturally implemented in a streaming context; the partition is made automatically by buffering the data as they arrive in a buffer of size M. When it is full, AP is run on this set, and the exemplars are stored in another buffer of identical size M but corresponding to the next hierarchical level. The procedure can be continued indefinitely as long as the data flow is not too large, i.e. the run-time taken by AP to treat one single buffer at lowest hierarchical level should not exceed the time needed for the same buffer to be full again.

B. AP clustering of aggregated data points

The exemplars at some level may not represent a fixed number of data points from lower levels, so a slight adjustment may be needed in some cases. The question then is how should we adapt the update rules of AP when the data points are the result of some prior aggregation. Assuming that a subset $\mathcal{S} \subset \mathcal{E}$ of n points, supposed to be at average mutual distance ϵ is aggregated into a single point $c \in \mathcal{S}$, how should we change the update rules so as to keep the result stable when ϵ is small? This is done by rewriting the similarity matrix as follows:

$$S(c,i) \longrightarrow nS(c,i), \qquad \forall i \in \bar{\mathcal{S}} \qquad \text{(III.1)}$$

$$S(i,c) \longrightarrow S(i,c), \qquad \forall i \in \bar{\mathcal{S}} \qquad \text{(III.2)}$$

$$S(c,c) \longrightarrow \sum_{i\in\mathcal{S}} S(i,c), \qquad \text{(III.3)}$$

and all lines and columns with index $i \in \mathcal{S}\backslash\{c\}$ are suppressed from the similarity matrix. The first transformation (III.1) simply states that c as a datapoint accounts now for n former points in $\mathcal{S}$, while (III.2) reflects that c taken as exemplar accounts for himself only in (II.2). In the last transform (III.3) it is implicitly assumed that if c is its own exemplar, then all points in $\mathcal{S}$ would adopt him as their exemplar too. This redefinition of the similarity matrix is not anymore symmetric and yields non-uniform penalty coefficients. In the basic update equations (II.8), (II.9), (II.10) and (II.11), nothing prevents from having different self-similarities because the key property (II.7) for deriving these equations is not affected by this. For

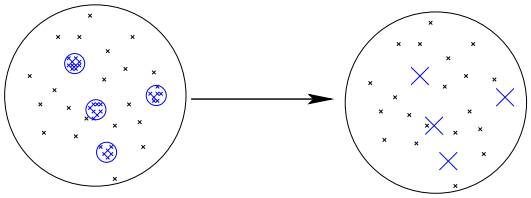

FIG. III.3: Aggregation of data points into single weighted items.

a valid clustering with the hard constraint of AP, the energy cost per data point obtained from (II.2) reads

$$e[\mathbf{c}] = \frac{1}{N}\sum_{c=1}^{n}\big(s + \sum_{i\in c} d^2(i,c)\big) = \frac{n}{N}s + \frac{1}{N}\sum_{i=1}^{N} d^2(i,c_i), \tag{III.4}$$

if n is the total number of cluster found by the solution, and we specify the similarity measure with help of the Euclidean distance

$$d(i,j) = |\mathbf{r}_i - \mathbf{r}_j|, \qquad \forall (i,j) \in \mathcal{E}^2.$$

To insure a basic scale invariance of the result, i.e. that the same solution is recovered, when the number of points in the dataset is rescaled we see that s has to scale like N. Now, if we deal directly with weighted data points in an Euclidean space, the preceding considerations concerning the re-weighting of the similarity matrix suggests that one may start directly from the following cost function:

$$e[\mathbf{c}] \stackrel{\text{def}}{=} ns + \frac{1}{Z}\sum_{c=1}^{n}\sum_{i\in c} w_i d^2(i,c). \tag{III.5}$$

Z being the normalization constant

$$Z \stackrel{\text{def}}{=} \sum_{i\in\mathcal{E}} w_i.$$

The $\{w_i, \forall i \in \mathcal{S}\}$ is a set of weights attached to each datapoint and the self-similarity has been rescaled uniformly

$$s \longrightarrow \sum_{i\in\mathcal{E}} w_i \; s.$$

with respect to the total weight of the dataset. Update rules of AP will be modified accordingly to III.5 and will be referred as to weighted affinity propagation (WAP) in the following. This along with the partitioning mechanism defines in principle completely our hierarchical affinity propagation algorithm (Hi-AP). However, as we shall see in V, the penalty s may be force to scale differently between hierarchies, under some additional constraints related to clustering stability. Beforehand we have to analyse some scaling effects associated to Hi-AP, motivated by the estimation of the error caused by the Divide-and-Conquer strategy.

IV. INFORMATION LOSS OF HI-AP

Assessing the error made by Hi-AP is not an easy task for a general distribution of data points and for an arbitrary setting of the penalty s, because this requires in principle the knowledge, or at least a good estimation of the joint distribution of exemplars found by AP. Nevertheless, we can say something in the following relevant situation, where the underlying dataset presents well separated clusters and by assuming that the penalty s is correctly tuned: this means that AP do not fragment or merge these underlying clusters, it is selecting exactly one single exemplar per existing true cluster. In such a case, clusters can be considered independently. In what follows we consider a single cluster with *finite variance*, with its center of mass *inside* the cluster, and make the assumption that the exemplar selected by AP is the nearest neighbor to the center of mass (see Figure. IV.4). Indeed, by construction, AP aims at finding the cluster exemplar $\mathbf{r_c}$ nearest to the center of mass of the sample points noted $\mathbf{r}_{cm}$:

$$e(\mathbf{c}) = s + \frac{1}{N}\sum_{i=1}^{N} |\mathbf{r}_i \; - \mathbf{r_c}|^2 = |\mathbf{r}_{cm} \; - \mathbf{r_c}|^2 + Cst.$$

A. No loss in dimension $d \neq 2$ for a well-tuned dilute AP clustering

To assess the information loss incurred by Hi-AP it turns out to be more convenient to compare the results in distribution, i.e. the distribution $P_{\mathbf{c}}$ of the cluster exemplars computed by AP, and the distribution $P_{\mathbf{c}(h)}$ of the cluster exemplar computed by Hi-AP with hierarchy-depth h, by considering for example their relative Kullback Leibler distance, or more basically by comparing their variance.

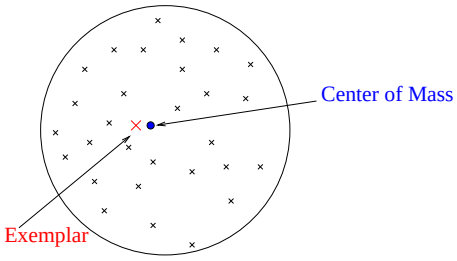

FIG. IV.4: The point minimizing the energy cost for a single cluster distribution

Let

$$\tilde{\mathbf{r}}_{\mathbf{c}} = \mathbf{r_c} - \mathbf{r}_{cm}.$$

denotes the relative position of exemplar $\mathbf{r_c}$ with respect to the center of mass $\mathbf{r}_{cm}$. Assuming a spherical symmetry for the cluster (in fact this has to be true only locally near the center of mass as shown in the next subsection),

the probability distribution of $\tilde{\mathbf{r}}_{\mathbf{C}}$ conditionally to $\mathbf{r}_{cm}$ is cylindrical; the cylinder axis supports the segment $(0, \mathbf{r}_{cm})$, where 0 is the origin of the d-dimensional space. As a result, the probability distribution of $\mathbf{r}_{cm} + \tilde{\mathbf{r}}_{\mathbf{C}}$ is the convolution of a spherical distribution, governed by the central limit theorem, with a cylindrical one governed by extreme value events.

In the sequel, subscripts sd refers to sample data, ex to the exemplar, and cm to center of mass, $x_{\cdot}$ denotes the corresponding square distances to the origin, $f_{\cdot}$ the corresponding probability densities and $F_{\cdot}$ their cumulative distribution. With these notations, the variance of the bare sample data distribution reads

$$\sigma \stackrel{\text{def}}{=} \mathbb{E}[x_{sd}] = \int_0^\infty x f_{sd}(x) dx, \qquad \text{(IV.1)}$$

and we assume it to be finite, as well as the following quantity,

$$\alpha \stackrel{\text{def}}{=} -\lim_{x \to 0} \frac{\log(F_{sd}(x))}{x^{\frac{d}{2}}}, \qquad \text{(IV.2)}$$

which encodes the short distance behaviour of the sample data distribution. The cumulative distribution of x_{cm} of a sample of size M then satisfies

$$\lim_{M \to \infty} F_{cm}(\frac{x}{M}) = \frac{\Gamma\left(\frac{d}{2}, \frac{dx}{2\sigma}\right)}{\Gamma\left(\frac{d}{2}\right)},$$

where $\Gamma(x, y)$ is the incomplete gamma function, by virtue of the central limit theorem. Meanwhile, $x_{\widetilde{ex}} = |\mathbf{r}_{ex} - \mathbf{r}_{cm}|^2$ has a universal extreme value distribution (up to rescaling, see e.g. [19] for general methods):

$$\lim_{M \to \infty} F_{\widetilde{ex}}(\frac{1}{M^{2/d}} x) = \exp\left(-\alpha x^{\frac{d}{2}}\right). \qquad \text{(IV.3)}$$

To see how the clustering error propagates along with the hierarchical process, we proceeds by induction. At hierarchical level h, one exemplar is selected out of M sample data spherically distributed with variance $\sigma^{(h)}$; it is the closest one to the center of mass and become a sample data at next level. Therefore, at hierarchical level $h+1$, the sample data distribution is the convolution of two spherical distributions, the exemplar and center of mass distributions obtained at level h. Assuming α and σ are given by $\alpha_d^{(h)}$ and $\sigma^{(h)}$ at level h, the following scaling recurrence property holds at level $h+1$ (See appendix A for details):

$$\lim_{M \to \infty} F_{sd}^{(h+1)}(\frac{x}{M^{(h+1)}}) = \begin{cases} \frac{\Gamma(\frac{1}{2}, \frac{x}{2\sigma^{(h+1)}})}{\Gamma(\frac{1}{2})} & d = 1 \\ \exp(-\alpha_2^{(h+1)} x) & d = 2 \end{cases}$$

$$\lim_{M \to \infty} F_{sd}^{(h+1)}(\frac{x}{M^{2(h+1)/d}}) = \exp\left(-\alpha_d^{(h+1)} x^{\frac{d}{2}}\right) \quad d > 2.$$

with $\sigma^{(h+1)} = \sigma^{(h)}$ for $d = 1$, $\alpha_d^{(h+1)} = \alpha_d^{(h)}$ for $d > 2$, while $\alpha_2^{(h+1)} = \alpha_2^{(h)}/2$ in dimension 2. As a result the distortion loss incurred by Hi-AP does not depend on the hierarchy depth h except in dimension $d = 2$. Figure IV.5 shows the radial distribution of exemplars obtained with different hierarchy-depth h and depending on the dimension d of the dataset. The curve for $h = 0$ corresponds to the AP situation, so the comparison with $h > 0$ shows that the information loss due to the hierarchical approach is moderate to negligible in dimension $d \neq 2$ provided that the number of samples per cluster at each clustering level is "sufficient" (say, $M > 30$ for the law of large numbers to hold). In dimension $d > 2$, the distance of the center of mass to the origin is negligible with respect to its distance to the nearest exemplar; the behaviour of the cost is thus governed by the Weibull distribution which is stable by definition (with an increased sensitivity to small sample size M as d approaches 2). In dimension $d = 1$, the distribution is dominated by the variance of the center of mass, yielding the gamma law which is also stable with respect to the hierarchical procedure. In dimension $d = 2$ however, the Weibull and gamma laws do mix at the same scale; the overall effect is that the width of the distribution increases like 2^h, as shown in Fig. IV.5 (top right).

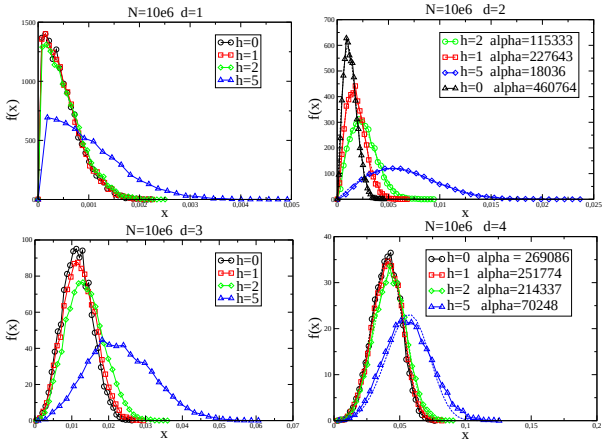

FIG. IV.5: Radial distribution plot of exemplars obtained by clustering of Gaussian distributions of $N = 10^6$ samples in $\mathbb{R}^d$ in one single cluster exemplar, with hierarchical level h ranging in 0,1,2,5, for diverse values of d: $d = 1$ (upper left), $d = 2$ (upper right), $d = 3$ (bottom left) and $d = 4$ (bottom right). Fitting functions are of the form $f(x) = Cx^{d/2-1} \exp(-\alpha x^{d/2})$.

B. Corrections for finite size dataset

In practice the number of data points per cluster might be not so large, hence it would be interesting to have an estimation of the error made by Hi-AP when M is finite. In order to limit the assumption on the shape of the underlying cluster we observe first that the parameter α defined in the preceding section is related to density at

the center of the cluster $p_{sd}(0)$ by

$$\alpha = p_{sd}(0)\frac{\Omega_d}{d}, \tag{IV.4}$$

with $\Omega_d = 2\pi^{d/2}/\Gamma(d/2)$ the d-dimensional solid angle, as long as the distribution is locally spherical around this point. Still, the shape of the cluster has some influence on the final result and we characterize it by defining the following ad hoc shape factor:

$$\omega \stackrel{\text{def}}{=} \frac{\sigma\alpha^{2/d}}{\Gamma\big(1+\frac{2}{d}\big)}. \tag{IV.5}$$

This dimensionless coefficient, relating α i.e. the density at the center of cluster to its variance σ prove to be useful for our purpose. By definition it reduces to $\omega = 1$ for the universal Weibull distribution (IV.3) and some other values depending on the spatial dimension are displayed in Table I. For $d > 2$, assuming $\alpha = \alpha^{(h)}$, $\sigma = \sigma^{(h)}$

	Weibull (IV.3)	Gaussian	Uniform L_1-sphere	Uniform L_2-sphere
ω	1	$\frac{d}{2}\frac{\pi}{\Gamma\big(1+\frac{2}{d}\big)}$	$\frac{\pi}{6}\frac{\big(\frac{d}{2}\big)^{2-2/d}}{\Gamma\big(\frac{2}{d}\big)\big(\Gamma\big(\frac{d}{2}\big)\big)^{2/d}}$	$\frac{d}{d+2}\frac{1}{\Gamma\big(1+\frac{2}{d}\big)}$

TABLE I: Different values of ω for various distributions

and $\omega = \omega^{(h)}$ at level h we find the following recurrence property (see Appendix B for details):

$$\begin{aligned}\sigma^{(h+1)} &= \sigma^{(h)} + o\big(M^{2/d-1}\big),\\ &= \frac{\sigma^{(0)}}{\omega^{(0)}}\Big(1+\frac{1}{M^{1-2/d}}\Big) + o\big(M^{2/d-1}\big)\end{aligned}$$

For a dataset of size N, $M = N^{1/h}$ when there are $h-1$ hierarchical levels, so if we now compare the variance

$$\sigma^{(h)} = \frac{\sigma^{(0)}}{\omega^{(0)}}\Big(1+N^{2/dh-1/h}\Big) + o\big(N^{2/hd-1/h}\big),$$

of the exemplar distribution in that case, to

$$\sigma^{(1)} = \frac{\sigma^{(0)}}{\omega^{(0)}}\Big(1+\frac{\omega^{(0)}}{N^{1-2/d}}\Big) + o\big(N^{2/d-1}\big),$$

obtained directly with AP , we get

$$\frac{\sigma^{(h)}}{\sigma^{(1)}} - 1 = N^{2/dh-1/h} + o\big(N^{2/dh-1/h}\big),$$

when d is larger than 2. This is consistent with the numerical check shown on Figure IV.6.

V. RENORMALIZING AFFINITY PROPAGATION

In Section III we left aside the question concerning the penalty coefficient s, how should it be modified from one

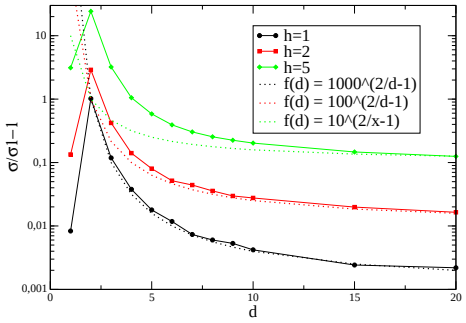

FIG. IV.6: $\sigma^{(h+1)}/\sigma^{(1)} - 1$ for $h = 1, 2, 5$ as a function of the dimension, when finding exemplars of a single cluster of 10^6 points (repeated 10^4 times)

hierarchical level to the next one. We address this question in the present section by applying a simple and exact renormalization principle to AP , based on the results of the preceding section, to yield a way to determine the number of true underlying clusters in a dataset.

By convenience we setup a thermodynamic limit where data point and clusters are distributed in a large spatial volume V and go to infinity independently with a fixed density of underlying clusters. After dividing s by V, the clustering cost per datapoint (III.5) reads for large n and N, $n << N$:

$$e(\rho) = \sigma(\rho) + s\rho, \tag{V.1}$$

with $\rho = n/V$ denoting a fixed density of clusters found by AP;

$$\sigma(\rho) \stackrel{\text{def}}{=} \sum_{c=1}^{\rho V} \nu_c\sigma_c, \tag{V.2}$$

denotes the distortion function, with $\nu_c = N_c/N$ the fraction of points in cluster c and σ_c the corresponding variance of the AP-cluster c.

Let us consider a one level Hi-AP where the N-size dataset is randomly partitioned into $M = 1/\lambda$ subsets of λN points each and where the reduced penalty s is fixed to some value such that each clustering procedure yields n exemplars on average. Considering the n/λ-size set of exemplars, the question is to adjust the value $s^{(\lambda)}$ for clustering this new dataset, in order to recover the same result as obtained by clustering the initial dataset with penalty s. Let us make some assumptions on the dataset:

- (**i**): the initial dataset samples n^* non-overlapping distributions, with common *shape factor* ω.
- (**ii**): there exists a value s^* of s for which AP yields the n^* true underlying clusters when N tends to infinity.
- (**iii**): $\sigma(\rho)$, the mean square distance of the sample data to their exemplars in the thermodynamic limit, is assumed to be a smooth decreasing convex function of the density $\rho = n/V$ of exemplars (obtained by AP) with possibly a cusp at $\rho = \rho^*$.

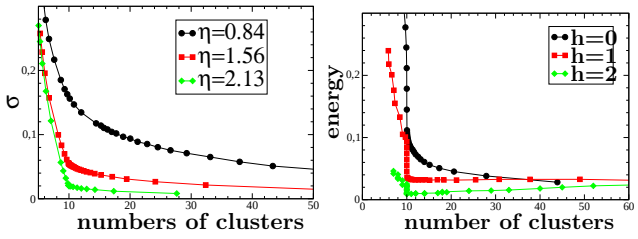

FIG. V.7: The distortion function for various values of η (left panel).The energy (i.e. the distortion plus the penalties) as a function of the number of clusters obtained at each hierarchical level of a single $h = 2$ HI-AP procedure with $\eta = 0.85$ (right panel). In both plots $d = 5$ $\lambda N = 300$ and $n^* = 10$

Assumption (**i**) can be approximately measured through parameter η, where d_{min} is the minimal distance between cluster centers and R_{max} is the maximal value of cluster radius:

$$\eta \stackrel{\text{def}}{=} \frac{d_{min}}{2R_{max}}, \tag{V.3}$$

Indeed clusters are expected to be separable for $\eta > 1$. In practice, gradually increasing s decreases the number of clusters, one by one, either merging two clusters fragments of a true cluster (de-fragmenting phase) or merging two (truly distinct) clusters (coalescent phase). Assumption (**ii**) implies that merging two truly distinct clusters entails a higher cost than fragmenting a true cluster. It follows that, by gradually increasing s, one first observes the de-fragmenting phase – until some threshold value s^* is reached. At that point the de-fragmentation phase ends and is replaced by the coalescent one.
Performing the clustering using one or two hierarchical levels should yield the same result. This basic requirement indicates how s should be renormalized. It is obtained by reinterpreting the Divide-and-Conquer as a decimation procedure by enforcing the self-consistency of HI-AP as illustrated in Figure V.8. Let n_1 [resp. n_2] be the number of clusters obtained after the first [resp. second] clustering stage. Depending on s the proper rescaling may vary, but for $s \simeq s^*$ this is supposed to behave in a universal way, because in that case, the clusters are preserved while their variance, as shown in the preceding section is simply multiplied by $\left(N\lambda/n_1\right)^{-2/d}/\omega = \lambda^{2/d}/\omega$ in dimension $d > 2$. Therefore we choose to rescale s as

$$s^{(\lambda)} = \frac{\lambda^{2/d}}{\omega}\, s. \tag{V.4}$$

When $\lambda^{2/d}/\omega \ll 1$, i.e. when there is a sufficient amount of data points per cluster, we expect the following property of HI-AP to hold:

$$\text{if} \quad \begin{cases} s < s^* & \text{then} \quad n_2 \geq n_1 \geq n^* \\ s = s^* & \text{then} \quad n_2 = n_1 = n^*. \\ s > s^* & \text{then} \quad n_2 = n_1 \leq n^* \end{cases} \tag{V.5}$$

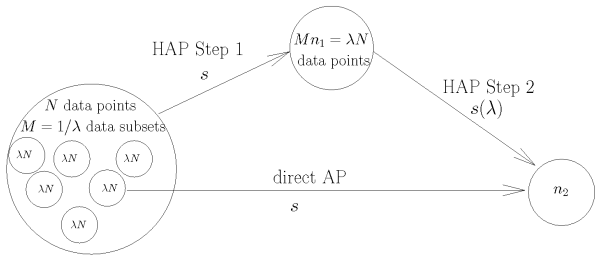

FIG. V.8: Divide-and-Conquer strategy translated in a Kadanoff decimation procedure.

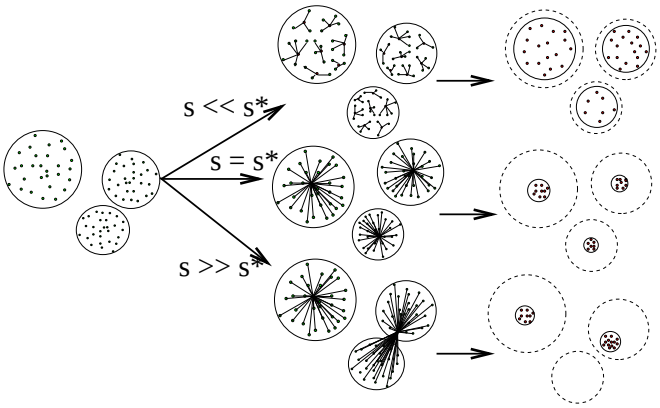

FIG. V.9: Transformation of the clusters after the first HI-AP step depending on s. $s^{(\lambda)}$ is defined to insure clustering stability when $s \simeq s^*$.

The reason is the following. In the thermodynamic limit the value n_1 for n, which minimizes the energy is obtained for $\rho_1 = n_1/V$ as the minimum of (V.1):

$$s + \sigma'(\rho_1) = 0.$$

At the second stage one has to minimize with respect to ρ,

$$e^{(\lambda)}(\rho) = \frac{\lambda^{2/d}}{\omega}\Big[\frac{\omega}{\lambda^{2/d}}\sigma^{(\lambda)}(\rho, \rho_1) + \rho s\Big],$$

where $\sigma^{(\lambda)}(\rho, \rho_1)$ denotes the distortion function of the second clustering stage when the first one yields a density ρ_1 of clusters. This amounts to find $\rho = \rho_2$ such that

$$s + \frac{\omega}{\lambda^{2/d}}\frac{\partial \sigma^{(\lambda)}}{\partial \rho}(\rho, \rho_1) = 0, \tag{V.6}$$

We need now to see how, depending on ρ_1,

$$\tilde{\sigma}^{(\lambda)}_{\rho_1}(\rho) \stackrel{\text{def}}{=} \lambda^{-2/d}\sigma^{(\lambda)}(\rho, \rho_1)$$

compares with $\sigma(\rho)$. This is depicted on Figure V.10.a. The qualitative justification of this plot is given in Appendix C. The point which is selected then graphically corresponds to the one for which the slope of $\sigma(\rho)$ is $-s$,

which results in the behaviour depicted on Figure V.10.b. from which this property (V.5) holds. The true number of clusters $n^* = \rho^* V$ is then easily identified as the junction point of the two curves in this figure.

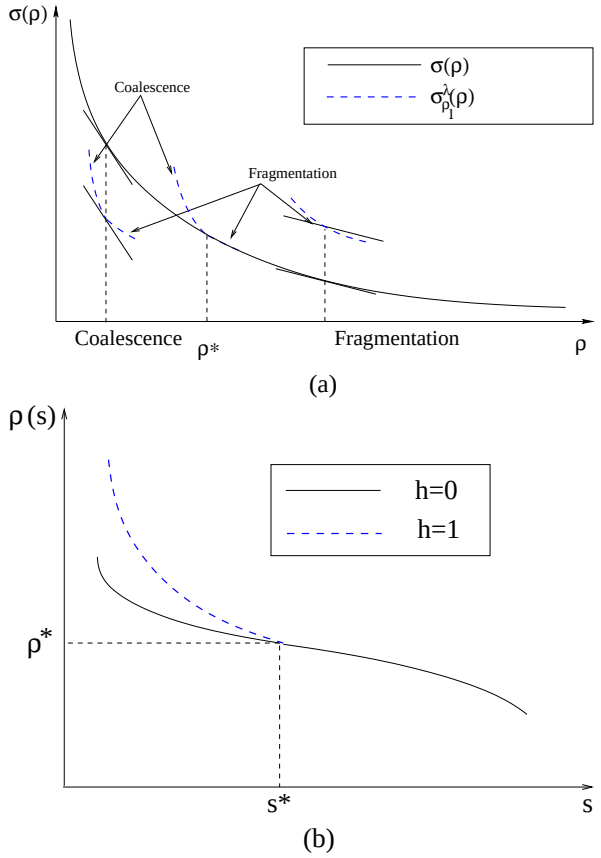

FIG. V.10: Sketch of the rescaling property. Comparison of the distortion function between two stages of Hi-AP (a). Corresponding result in terms of the number of clusters as a function of s (b).

VI. NUMERICAL EXPERIMENTS

We have tested this renormalized procedure both on artificial and real-world datasets, for proofs of principle and to discuss the robustness and limits of the approach.

A. Artificial datasets

The study conducted on artificial datasets investigates the impact of the cluster shapes, their overlapping, the dimensionality and the size of the dataset. The typical observed behaviour is the one shown on Figure VI.11.a VI.11.c and VI.11.d . The self-similar point is clearly identified when plotting the number of clusters against the bare penalty, when η is not to small. As expected from the scaling (V.4), the effect is less sensible when the dimension increases, but remains perfectly visible and exploitable at least up to $d = 30$. The absence of information loss of the hierarchical procedure can be seen on the mean-error plots on Figure VI.11.b , in the region of s around the critical value s^*. The results are stable, when we take into account at the first stage of the hierarchical procedure the influence of the shape of the clusters. This is done by fixing the value of the factor form ω to the correct value. In that case, at subsequent levels of the hierarchy the default value $\omega = 1$ is the correct one to give consistent results. Nevertheless if the factor form is unknown and set to false default value, the results are spoiled at subsequent levels, and the underlying number of clusters turns out to be more difficult to identify, depending on the discrepancy of ω with respect to its default value. We have observed also that the identification of the transition point is still possible when the number of datapoints per cluster get smaller, down to 6 in these tests. To obtain these curves e.g. with two hierarchical levels the total number of clustered items vary in the range $10^4 - 10^6$ which is out of the range of a single AP run on a complete graph.

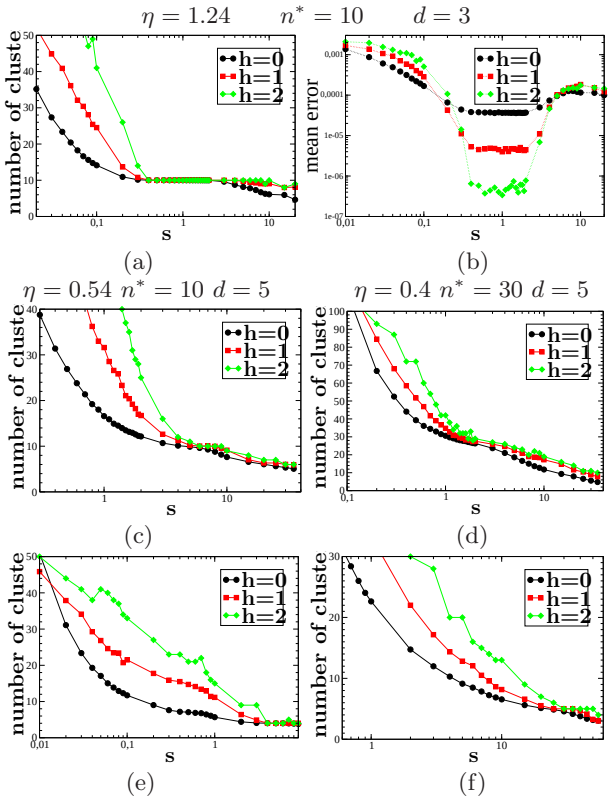

FIG. VI.11: Number of clusters obtained at each hierarchical level as a function of s, with fixed size of individual partition $\lambda N = 300$, for various spatial dimension, separability indexes and number of underlying clusters (a), (c) and (d), for the EGEE dataset (e) and of a jpeg image (f) of 1.5 10^5 pixels size. Error distance of the exemplars from the true underlying centers (b) corresponding to clustering (a).

B. Real-world datasets

On real data the situation is different since we have no direct way to know whether the conditions of validity of the approach are satisfied. To be of interest a real dataset has to be very large, typically $10^5 - 10^6$ items, to let several hierarchical comparisons. In addition the data have to be embedded in $\mathbb{R}^d$.

The EGEE dataset, publicly available from [15] and comprising 5 million datapoints, has been used. Each datapoint, originally describing a job submitted on the EGEE grid, is described by 6 continuous variables and 6 boolean ones. For the sake of the study, boolean values have been replaced by continuous values in $[0, 1]$, with addition of a small amount of noise uniformly distributed in $[0, 0.1]$. Finally all components are rescaled such as to fit in a window of identical unit size and the standard Euclidean distance is used. The output of HI-AP is shown on Figure VI.11.e. It is seen that the curves join near $s = 4[20]$, yielding then basically $n = 4$ different clusters. When looking more carefully at the exemplars, we see that the clusters correspond to different combinations of labels (the initially binary components), while in the continuous subspace HI-AP does not detect any structure; all exemplars found at $s = 4$ share the same continuous components. Looking at distributions of the whole dataset along the axes shows instead well defined structures; unfortunately these clusters are very unbalanced by a factor of $\simeq 100 - 1000$, which certainly prevents the condition (ii) from being satisfied. By contrast the structures on the (initially) discrete features are perfectly identified, although clusters seem also unbalanced by a factor of $\simeq 10$ in this subspace.

The strategy to circumvent this limitation of the algorithm is certainly related to redefining the distance, accounting for the spatial variation of densities [21].

A second large image dataset has been considered, where the datapoints actually reflect the pixels in the images. Each datapoint lies in a (almost) continuous $5-$dimensional space, the 2 first components corresponding to the pixel spatial position while the 3 other corresponding to RGB encoding of the colors. We rescale as before each variable to fit in a window of size one, and consider as well the Euclidean distance. On the example seen on Figure VI.11.f we again see a point of convergence of the three curves indicating a number of cluster equal to 5. We observe this despite the little offset between $h = 1, 2$ and $h = 2, 3$, showing that these clusters are far from being spherical. Looking then again at the distribution along some axes (color axes) reveal on one hand that we should probably identify more detailed clusters, but also how far we are from the working assumptions made and supporting the renormalization approach on the other hand, as these clusters do overlap quite significantly. While a similar situation is observed on most pictures we have been testing, HI-AP is found to yield a rather relevant selection of clusters though.

VII. DISCUSSION AND PERSPECTIVES

The present analysis of the scaling properties of AP, within a divide-and-conquer setting gives us a simple way to identify a self-similar property of the special point s^*, for which the exact structure of the clusters is recovered. Our main contribution hence is a principled approach for identifying the true cluster structure when using AP. While earlier work has been intensively examining the stability of k-means or PCA approaches (see e.g. [22]), to our best knowledge, the use of a renormalization approach is original in this context. This property can be actually exploited, when the dimension is not too large and when the clusters are sufficiently far apart and sufficiently populated. The separability property is actually controlled by the parameter η introduced in V.3, and in the vicinity of s^*, the absence of information loss, deduced from the single cluster analysis is effective. The approach can be turned into a simple line-search algorithm and this perspective will be investigated further on real data sets to obtain an on-line self-tuning of s, i.e. during the hierarchical treatment itself.

From the theoretical viewpoint, this renormalization approach to the self-tuning of algorithm parameter could be applied in other context, where self-similarity is a key property at large scale. First it is not yet clear how we could adapt our method to the SCAP context. The principal component analysis and associated spectral clustering provide other examples, where the fixing of the number of selected components is usually not obtained by some self-consistent procedure and where a similar approach to the one presently proposed could be used.

ACKNOWLEDGMENTS

This work was supported by the French National Research Agency (ANR) grant No ANR-08-SYSC-017.

[1] A.P. Dempster, Laird N.M., and D.B. Rubin. Maximum likelihood for incomplete data via the EM algorithm. *J. Royal Stat. Soc. B*, 39(1):1–38, 1977.

[2] in these algorithms, k is the number of centers to be obtained by alternatively assigning datapoints to candidate centers (expectation step) and then taking the mean of each newly defined cluster as new candidate centers (optimization step).

[3] S. Wiseman, M. Blatt, and E. Domany. Superparamagnetic clustering of data. *Phys. Rev. E*, 57:3767–3787, 1998.

[4] J. Pearl. *Probabilistic Reasoning in Intelligent Systems:*

Network of Plausible Inference. Morgan Kaufmann, 1988.
[5] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. *Advances in Neural Information Processing Systems*, pages 689–695, 2001.
[6] F. R. Kschischang, B. J. Frey, and H. A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47:498–519, 2001.
[7] B. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315:972–976, 2007.
[8] M. Leone, Sumedha, and M. Weigt. Clustering by soft-constraint affinity propagation: Applications to gene-expression data. *Bioinformatics*, 23:2708, 2007.
[9] M. Leone, Sumedha, and M. Weigt. Unsupervised and semi-supervised clustering by message passing: Soft-constraint affinity propagation. *Eur. Phys. J. B*, pages 125–135, 2008.
[10] C. Fraley and A. Raftery. How many clusters?which clustering method? answer via model-based clustering. *The Computer Journal*, 41(8), 1998.
[11] S. Still and W. Bialek. How many clusters?:an information-theoretic perspective. *Neural Computation*, 16:2483–2506, 2004.
[12] S. Dudoit and J. Fridlyand. A prediction-based resampling method for estimating the number of clusters in a dataset. *Genome Biology*, 2(7):0036.1–0036.21, 2002.
[13] K. Wang, J. Zhang, D. Li, X. Zhang, and T. Guo. Adaptive affinity propagation clustering. *Acta Automatica Sinica*, 33(12):1242–1246, 2007.
[14] X. Zhang, C. Furtlehner, and M. Sebag. Data streaming with affinity propagation. In *ECML/PKDD*, pages 628–643, 2008.
[15] X. Zhang, C. Furtlehner, J. Perez, C. Germain-Renaud, and M. Sebag. Toward autonomic grids: Analyzing the job flow with affinity streaming. In *KDD*, pages 628–643, 2009.
[16] Y. Kabashima. Propagating beliefs in spin-glass models. *J. Phys. Soc. Jpn.*, 72:1645–1649, 2003.
[17] Except if the similarity matrix is sparse, in which case the complexity reduces to $Nklog(N)$ with k the average connectivity of the similarity matrix [7].
[18] S. Guha, A. Meyerson, N. Mishra, R. Motwani, and L. O'Callaghan. Clustering data streams: Theory and practice. In *TKDE*, volume 15, pages 515–528, 2003.
[19] L. de Haan and A. Ferreira. *Extreme Value Theory*. Operations Research and Financial Engineering. Springer, 2006.
[20] the curves coincide also at small s in the region where the number of data point is not sufficient according to condition $\lambda^{2/d} \ll \omega$ in V.4.
[21] Lihi Zelnik-manor and Pietro Perona. Self-tuning spectral clustering. In *Advances in Neural Information Processing Systems 17*, pages 1601–1608. MIT Press, 2004.
[22] M. Meila. The uniqueness of a good optimum for k-means. In *ICML*, pages 625–632. ACM, 2006.
[23] Fluctuations are neglected in this argument. In practice the exemplars which emerge from the coalescence of two clusters might originate from both clusters, when considering different subsets, if the number of data is not sufficiently large.

Appendix A: Proof of loss estimate

The influence between the center of mass and extreme value statistics distribution corresponds to corrections which vanish when M tends to infinity (see Appendix B. Neglecting these corrections, enables us to use a spherical kernel instead of cylindrical kernel and to making no distinction between ex and $\tilde{e}x$, to write the recurrence. Between level h and $h+1$, one has:

$$f_{sd}^{(h+1)}(x) = \int_0^\infty K^{(h,M)}(x,y) f_{ex}^{(h,M)}(y) dy \qquad \text{(A.1)}$$

with

$$\lim_{M\to\infty} M^{-1} K^{(h,M)}(\frac{x}{M}, \frac{y}{M}) = \frac{d}{\sigma^{(h)}} K(\frac{dx}{\sigma^{(h)}}, \frac{dy}{\sigma^{(h)}}) \quad \text{(A.2)}$$

where $K(x,y)$ is the d-dimensional radial diffusion kernel,

$$K(x,y) \stackrel{\text{def}}{=} \frac{1}{2} x^{\frac{d-2}{4}} y^{\frac{2-d}{4}} I_{\frac{d}{2}-2}\big(\sqrt{xy}\big) e^{-\frac{x+y}{2}}.$$

with $I_{\frac{d}{2}-1}$ the modified Bessel function of index $d/2-1$. The selection mechanism of the exemplar yields at level h,

$$F_{ex}^{(h,M)}(x) = \big(F_{sd}^{(h)}(x)\big)^M,$$

and with a by part integration, (A.1) rewrites as:

$$f_{sd}^{(h+1)}(x) = K^{(h,M)}(x,0) + \int_0^\infty \big(F_{sd}^{(h)}(y)\big)^M \frac{\partial K^{(h,M)}}{\partial y}(x,y) dy,$$

with

$$\lim_{M\to\infty} M^{-1} K^{(h,M)}(\frac{x}{M}, 0) = \frac{d}{2\Gamma(\frac{d}{2})\sigma^{(h)}} \big(\frac{dx}{2\sigma^{(h)}}\big)^{\frac{d}{2}-1} \exp\big(-\frac{dx}{2\sigma^{(h)}}\big).$$

At this point the recursive hierarchical clustering is described as a closed form equation. The result of Section IV A is then based on (A.2) and on the following scaling behaviors,

$$\lim_{M\to\infty} F_{ex}^{(h,M)}\big(\frac{x}{M^{\frac{2}{d}}}\big) = \exp\big(-\alpha^{(h)} x^{\frac{d}{2}}\big),$$

so that

$$\lim_{M\to\infty} F_{sd}^{(h+1)}(\frac{x}{M^\gamma}) = \lim_{M\to\infty} M^{1-\gamma} \int_0^\infty dy \int_{\frac{x}{\sigma^{(h)}}}^\infty du f_{ex}^{(h,M)}(\frac{y}{M^{\frac{2}{d}}}) K(M^{1-\gamma}u, \frac{M^{1-\frac{d}{2}} y}{\sigma^{(h)}}).$$

Basic asymptotic properties of $I_{d/2-1}$ yield with a proper choice of γ, the non degenerate limits of the scaling result. In the particular case $d=2$, taking $\gamma=1$, it comes:

$$\lim_{M\to\infty} F_{sd}^{(h+1)}(\frac{x}{M}) = \int_0^\infty dy \int_{\frac{x}{\sigma^{(h)}}}^\infty du f_{ex}^{(h)}(\sigma^{(h)} y) K(u,y)$$

$$= -\int_0^\infty dy \int_{\frac{x}{\sigma^{(h)}}}^\infty du \frac{de^{-\alpha^{(h)}\sigma^{(h)}x}}{dy} I_0(2\sqrt{uy})e^{-(u+y)}$$

$$= \exp\big(-\frac{\alpha^{(h)}}{1+\alpha^{(h)}\sigma^{(h)}}x\big),$$

with help of the identity

$$\int_0^\infty dx x^\nu e^{-\alpha x} I_{2\nu}(2\beta\sqrt{x}) = \frac{1}{\alpha}\big(\frac{\beta}{\alpha}\big)^{2\nu} e^{\frac{\beta}{\alpha}}.$$

Again in the particular case $d = 2$, by virtue of the exponential law one further has $\alpha^{(h)} = 1/\sigma^{(h)}$, finally yielding:

$$\beta^{(h+1)} = \frac{1}{2}\beta^{(h)}. \tag{A.3}$$

Appendix B: Finite size corrections

We consider a given hierarchical level h, $\mathbf{r}$ denotes sample points, $\mathbf{r}_{cm}$ their corresponding center of mass, and $\mathbf{r_C}$ the exemplar, which in turn becomes a sample point at level $h+1$. We have

$$p_{sd}^{(h+1)}(\mathbf{r})d^d\mathbf{r} \;=\; P(\mathbf{r_C} \in d^d\mathbf{r}) \;=\; d^d\mathbf{r}\int d^d\mathbf{r}_{cm}$$
$$p_{sd,cm}^{(h)}(\mathbf{r},\mathbf{r}_{cm})P\big(|\mathbf{r}_{sd}-\mathbf{r}_{cm}| \geq |\mathbf{r}-\mathbf{r}_{cm}|\big|\mathbf{r}_{cm}\big)^{M-1}.$$

We analyse this equation with the help of a generating function:

$$\phi_\centerdot(\Lambda) = \int d^d\mathbf{r} p_\centerdot(\mathbf{r}) e^{-\Lambda\mathbf{r}}.$$

where $\centerdot$ may be indifferently sd, c or cm and $\Lambda\mathbf{r}$ is the ordinary scalar product between two d-dimensional vectors. Let $\lambda = |\Lambda|$, by rotational invariance, $p_\centerdot$ depends only on r and $\phi_\centerdot$ depends solely on λ, so we have

$$g_\centerdot(\lambda) \stackrel{\text{def}}{=} \log(\phi_\centerdot(\Lambda))$$
$$= \log\Big(2\pi^{d/2}\int_0^\infty dr r^{d-1} p_\centerdot(r)\big(\frac{\lambda r}{2}\big)^{1-d/2} I_{d/2-1}(\lambda r)\Big).$$

The joint distribution between $\mathbf{r}_{sd}$ and $\mathbf{r}_{cm}$ takes the following form

$$p_{sd,cm}(\mathbf{r},\mathbf{r}_{cm}) = p_{sd}(r)p_{cm|sd}(|\mathbf{r}_{cm}-\frac{\mathbf{r}}{M}|)$$

where by definition $p_{cm|sd}$ is the conditional density of $\mathbf{r}_{cm}$ to $\mathbf{r}_{sd}$, with

$$g_{cm|sd}(\lambda) = (M-1)g_{sd}\big(\frac{\lambda}{M}\big), \tag{B.1}$$

while

$$g_{cm}(\lambda) = Mg_{sd}\big(\frac{\lambda}{M}\big), \tag{B.2}$$

where g_{sd} is assumed to have a non zero radius Taylor expansion of the form

$$g_{sd}(\lambda) = \frac{\sigma^{(h)}}{2d}\lambda^2 + \sum_{n=2}^\infty \frac{g^{(2n)}(0)}{2n!}\lambda^{2n}, \tag{B.3}$$

since by rotational symmetry all odd powers of λ vanish and where $\sigma^{(h)}$ represents the variance at level h of the sample data distribution. In addition the conditional probability density of $\mathbf{r}_{sd}$ to $\mathbf{r}_{cm}$ reads

$$p_{sd|cm}(\mathbf{r},\mathbf{r}_{cm}) = \frac{p_{sd}(r)}{p_{cm}(r_{cm})}p_{cm|sd}(|\mathbf{r}_{cm}-\frac{\mathbf{r}}{M}|)$$
$$\stackrel{\text{def}}{=} p_{sd|cm}(u,\theta,r_{cm})$$

where $\mathbf{u} = \mathbf{r}-\mathbf{r}_{cm}$ and θ is the angle between $\mathbf{u}$ and $\mathbf{r}_{cm}$. Let

$$f(u,r_{cm}) \stackrel{\text{def}}{=} P(|\mathbf{r}_{sd}-\mathbf{r}_{cm}| \geq u\big|\mathbf{r}_{cm}).$$

We have

$$f(u,r_{cm}) \;=$$
$$1-\Omega_{d-1}\int_0^u dx x^{d-1}\int_0^\pi d\theta \sin\theta^{d-2} p_{sd|cm}(x,\theta,r_{cm}).$$

with

$$\Omega_d = \frac{2\pi^{d/2}}{\Gamma\big(\frac{d}{2}\big)},$$

the d-dimensional solid angle. Let

$$h(u,r_{cm}) \stackrel{\text{def}}{=} \log(f(u,r_{cm})).$$

We have

$$p_{sd}^{(h+1)}(r) = \; p_{sd}^{(h)}(r)\int d^d\mathbf{r}_{cm}$$
$$p_{cm|sd}(|\mathbf{r}_{cm}-\frac{\mathbf{r}}{M}|)\exp\big((M-1)h(|\mathbf{r}-\mathbf{r}_{cm}|,r_{cm})\big).$$

From the expansion (B.3) we see that corrections in g_{cm} and $g_{cm|sd}$ to the Gaussian distribution are of order $1/M^3$, $\sigma_{cm} = \sigma/M$ as expected from the central limit theorem and $\sigma_{cm|sd} = (M-1)\sigma/M^2$. Letting $y = \mathbf{r}_{cm}-\mathbf{r}$ we have

$$p_{sd}^{(h+1)}(r) \;=$$
$$p_{sd}^{(h)}(0)\Big(\frac{dM}{2\pi\sigma^{(h)}}\Big)^{d/2}\int d^d\mathbf{y}\;\exp\Big(-M\psi^{(M)}(\mathbf{r},\mathbf{y})\Big),$$

with

$$\psi^{(M)}(\mathbf{r},\mathbf{y}) \stackrel{\text{def}}{=} -\frac{d}{2}\log\frac{M}{M-1} - \frac{dr^2}{2\sigma^{(h)}} + \log\frac{p_{sd}^{(h)}(r)}{p_{sd}^{(h)}(0)}$$
$$+\frac{dM}{2(M-1)\sigma^{(h)}}|\mathbf{y}+\mathbf{r}|^2 + (M-1)h\big(y,|\mathbf{y}+\mathbf{r}|\big).$$

As observed previously $p_{sd}^{(h+1)}(r/M^{1/d})$ converges to a Weibull distribution when M goes to infinity, and the corrections to this are obtained with help of the following approximation:

$$\psi^{(M)}(\frac{\mathbf{r}}{M^{1/d}},\mathbf{y}) = \frac{d}{2\sigma^{(h)}}|\mathbf{y}+\frac{\mathbf{r}}{M^{1/d}}|^2 + \alpha^{(h)}y^d + O(\frac{1}{M}),$$

with

$$\alpha^{(h)} = p_{sd}^{(h)}(0)\frac{\Omega_d}{d}.$$

As a result, computing the normalization constant $p_{sd}^{(h+1)}(0)$ and the corresponding variance $\sigma^{(h+1)}$, yields the following recurrence relations:

$$\begin{cases} \alpha^{(h+1)} = & \alpha^{(h)} + O\Big(\frac{1}{M}\Big). \\ \sigma^{(h+1)} = & \Gamma\big(1+\frac{2}{d}\big){\alpha^{(h)}}^{-2/d}\Big(1+\frac{\sigma^{(h)}{\alpha^{(h)}}^{2/d}}{\Gamma\big(1+\frac{2}{d}\big)}\frac{1}{M^{1-2/d}}\Big) \\ & +o\big(M^{2/d-1}\big). \end{cases}$$

Letting

$$\omega^{(h)} \stackrel{\text{def}}{=} \frac{\sigma^{(h)}{\alpha^{(h)}}^{2/d}}{\Gamma\big(1+\frac{2}{d}\big)},$$

we get

$$\omega^{(h+1)} = 1 + \frac{\omega^{(h)}}{M^{1-2/d}} + o\big(M^{2/d-1}\big).$$

Consequently, for $h=0$, we have

$$\sigma^{(1)} = \frac{\sigma^{(0)}}{\omega^{(0)}}\Big(1+\frac{\omega^{(0)}}{M^{1-2/d}}\Big) + o\big(M^{2/d-1}\big),$$

while for $h>1$ we get

$$\sigma^{(h+1)} = \sigma^{(h)}\Big(1+\frac{\omega^{(h)}-\omega^{(h-1)}}{M^{1-2/d}}\Big) + o\big(M^{2/d-1}\big).$$

For $h=1$ this reads

$$\sigma^{(2)} = \sigma^{(1)}\Big(1+\frac{1-\omega^{(0)}}{M^{1-2/d}}\Big) + o\big(M^{2/d-1}\big),$$

and thereby

$$\sigma^{(h+1)} = \sigma^{(h)} + o\big(M^{2/d-1}\big), \qquad \text{for } h>1.$$

Appendix C: Clustering stability in HI-AP

Assume first that $\rho_1 = \rho^*$, which is obtained if we set $s=s^*$ in the first clustering stage. This means that each cluster which is obtained at this stage is among the exact clusters with a reduced variance, resulting from the extreme value distribution properties (IV.3) combined with definition (IV.5) of the shape factor ω:

$$\sigma_c^{(\lambda)} = \frac{1}{\omega}\Big(\frac{\lambda N}{n_1}\Big)^{-2/d}\sigma_c = \frac{\lambda^{2/d}}{\omega}\sigma_c. \qquad \text{(C.1)}$$

Note at this point that

$$\frac{\omega}{\lambda^{2/d}} \gg 1,$$

is required to be in the conditions of getting a cluster shaped by the extreme value distribution. For $\rho > \rho^*$, the new distortion involves only the inner cluster distribution of exemplars which is simply rescaled by this $(\rho_1/\lambda)^{2/d}$ factor, so from (V.2) we conclude that

$$\tilde{\sigma}_{\rho^*}^{(\lambda)}(\rho) = \sigma(\rho), \qquad \text{for} \qquad \rho \geq \rho^*.$$

Instead, for $\rho < \rho^*$, the new distortion involves the merging of clusters, which inter distances, contrary to their inner distances, are not rescaled and are the same as in the original data set. This implies that

$$\frac{d\tilde{\sigma}_{\rho^*}^{(\lambda)}}{d\rho}(\rho) \leq \sigma'(\rho), \qquad \text{for} \qquad \rho < \rho^*.$$

As a result the optimal number of clusters is unchanged, $\rho_1 = \rho^*$.

For $\rho_1 < \rho^*$, which is obtained when $s > s^*$, the new distribution of data points, formed of exemplars, is also governed by the extreme value distribution, and all cluster at this level are intrinsically true clusters, with a shape following the Weibull distribution. We are then necessary at the transition point at this stage: $\rho^* = \rho_1$[23]. In addition, the cost of merging two clusters, i.e. when ρ is slightly below ρ_1, is actually greater now after rescaling,

$$\frac{d\tilde{\sigma}_{\rho_1}^{(\lambda)}}{d\rho}(\rho) \leq \sigma'(\rho), \qquad \text{for} \qquad \rho = (\rho_1)_-,$$

because mutual cluster distances appear comparatively larger. Instead, for ρ slightly above ρ_1, the gain in distortion when ρ increases is smaller, because it is due to the fragmentation of Weibull shaped cluster, as compared to the gain of separating clusters in the coalescence phase at former level,

$$\frac{d\tilde{\sigma}_{\rho_1}^{(\lambda)}}{d\rho}(\rho) \geq \sigma'(\rho), \qquad \text{for} \qquad \rho = (\rho_1)_+.$$

As a result, from the convexity property of $\sigma^{(\lambda)}(\rho)$, we then expect again that the solution of (V.6) remains unchanged $\rho_2 = \rho_1$ in the second step with respect to the first one.

Finally, for $\rho_1 > \rho^*$, the new distribution of data points is not shaped by the extreme value statistics when the number of fragmented clusters increases, because in that case the fragments are distributed in the entire volume of the fragmented cluster. In particular,

$$\tilde{\sigma}_{\rho_1}^{(\lambda)}(\rho) \simeq \frac{\omega}{\lambda^{2/d}}\sigma(\rho), \qquad \text{when} \qquad \rho_1 >> \rho^*.$$

The rescaling effect vanishes progressively when we get away from the transition point, so we conclude that the optimal density of clusters ρ_2 is displaced toward larger values in this region.