Sparse Dual of the Density Peaks Algorithm for Cluster Analysis of High-dimensional Data

Dimitris Floros* Tiancheng Liu† Nikos Pitsianis*† Xiaobai Sun†

*Department of Electrical and Computer Engineering
Aristotle University of Thessaloniki
Thessaloniki 54124, Greece

†Department of Computer Science
Duke University
Durham, NC 27708, USA

***Abstract*—The density peaks (DP) algorithm for cluster analysis, introduced by Rodriguez and Laio in 2014, has proven empirically competitive or superior in multiple aspects to other contemporary clustering algorithms. Yet, it suffers from certain drawbacks and limitations when used for clustering high-dimensional data. We introduce SD-DP, the sparse dual version of DP. While following the DP principle and maintaining its appealing properties, we establish a sparse descriptor of local density as a robust representation. By analyzing and exploiting the consequential properties, we are able to use sparse graph-matrix expressions and operations throughout the clustering process. As a result, SD-DP has provably linear-scaling computation complexity under practical conditions. We show, with experimental results on several real-world high-dimensional datasets, that SD-DP outperforms DP in robustness, accuracy, self-governance, and efficiency.**

***Index Terms*—Clustering algorithms, density peaks clustering, nearest neighbors.**

I. Introduction

The density peaks (DP) algorithm for cluster analysis, introduced by Rodriguez and Laio in 2014 on Science magazine [20], has proven empirically competitive or superior in multiple aspects to other contemporary clustering algorithms. DP received immediate and growing attention from many research communities. Real-world data of research interest have intrinsic, heterogeneous group structures and contain noise and uncertainty. Clustering is to detect such structures in data with intra-group similarity and inter-group dissimilarity, using context-specific feature description and metric, governed by certain differentiation principles or criteria. Cluster analysis has been long employed in scientific studies, such as molecular dynamics trajectory analysis [23], classification of astronomical events [31] and community detection in complex systems [8], [18], [27]. It is increasingly recognized as fundamental, in existing and emerging study domains, to exploratory data analysis, unsupervised learning, knowledge discovery and subsequent higher-level tasks. In computer vision, cluster analysis is embodied in image segmentation [25] and/or denoising [2], content-based image retrieval [26], and image object recognition and tracking [15], [16]. Recent years witness many new applications: gene expression pattern analysis in biology [9], [11], [13], [19], [32], thematic categorization of text documents, author identification, or statistical analysis of word semantics in natural language processing [6], [30], statistical categorization or identification of musical genres [22], and recommendation systems over social or commercial networks [24], to name a few. The data gathered and to be analyzed reside typically in a high dimensional feature space. The feature dimension can be several thousand or much higher, see for example Fig. 1, 2 and 5. High-dimensional data challenge many existing clustering algorithms.

(a) Parthenon image (b) Segmentation result

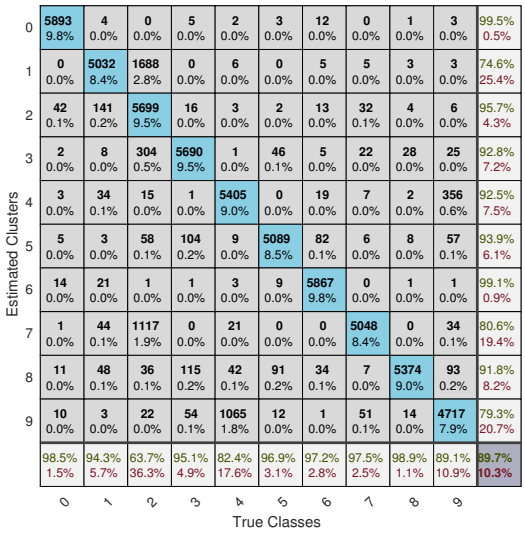
Fig. 1: Image segmentation by SD-DP. **(a)** The Parthenon image from the Berkeley Segmentation Dataset and Benchmark [17] with $N = 481 \times 321 = 154{,}401$ pixels per color channel. **(b)** Segmentation result (3 segments) by SD-DP ($k = 71$). The features are patches of 5×5 pixels per color ($D = 5 \times 5 \times 3 = 75$). The execution time is 3 seconds on MATLAB (excluding kNN construction time).

The DP principle promises a potentially effective and efficient solution to the problem. We assess existing algorithms for cluster analysis against a list of desirable properties, such as capability to estimate the number of cluster, without prescription, admission of non-spherical, non-convex cluster shapes, permission of any similarity or distance metric, high efficiency, and low program complexity. The k-means algorithm [14] requires a prescribed cluster number and favors clusters of spherical shape. Among non-spherical-shape clustering algorithms, the mean-shift algorithm [3] is effective with low-dimensional data, or must be preceded by a dimension reduction [10], [21], [28], [29], which may not preserve local structures. DBSCAN-based algorithms [1], [7] also suffer from the curse of dimensionality, in a subtle way. They require two specified bounds a priori on a so-called signal neighborhood: an upper bound on the neighborhood radius and a lower bound on the population count in the neighborhood. In comparison, the DP algorithm is advantageous in many aspects. However, the DP algorithm suffers from certain drawbacks and limitations, especially when it is used for high dimensional data clustering. Some of the issues are not necessarily attributed to the DP principle.

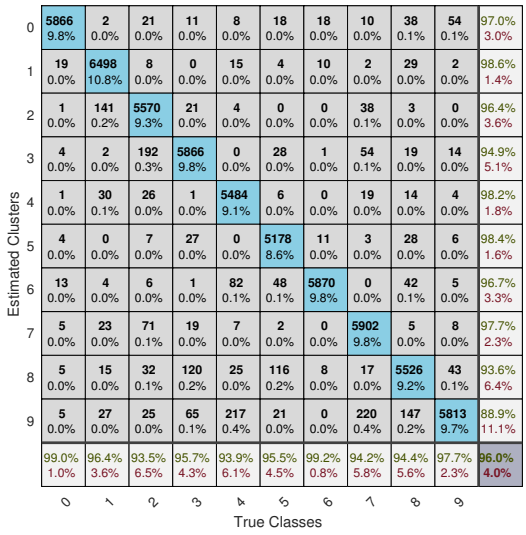
In this paper, we introduce SD-DP, the sparse dual of the DP algorithm. SD-DP follows the DP principle, inherits its appealing properties, and more importantly, surpasses DP in robustness, self-governance, accuracy, computation complexity

The first two authors contributed equally to this work.

TABLE 1: Comparison in accuracy between DP[1] and SD-DP on $N = 60{,}000$ images of MNIST handwritten digits, with known true class labels. The accuracy of each algorithm is summarized in the corresponding confusion/error matrix: **(a)** estimation by DP, using intensity feature vectors ($D = 28 \times 28 = 784$) and tangent distance, **(b)** estimation by SD-DP, using HOG[2] feature descriptors ($D = 144$) and Euclidean distance. The far-right column contains the precision (a.k.a. positive predictive value) and the false discovery rate. The bottom row contains the recall (a.k.a. true positive rate) and the false negative rate. The bottom-right cell is the F_1 score (a.k.a. Sørensen–Dice coefficient) for the overall accuracy.



(a) Confusion matrix with estimation by DP



(b) Confusion matrix with estimation by SD-DP

	0	1	2	3	4	5	6	7	8	9
DP	0.99	0.83	0.77	0.94	0.87	0.95	0.98	0.88	0.95	0.84
SD-DP	0.98	0.98	0.95	0.95	0.96	0.97	0.98	0.96	0.94	0.93

(c) F_1 scores (a.k.a. Sørensen–Dice coefficients)

and potential concurrency for parallel execution. Instead of seeking a trade-off of accuracy for efficiency, we establish a sparse descriptor of local density as a robust representation first and foremost. This leads to many beautiful properties for us to exploit. Particularly, SD-DP has proven linear complexity. We are also able to use sparse graph or matrix expressions and operations throughout the clustering process.

We present experimental results on four sets of real-world high dimensional data. SD-DP outperforms DP in accuracy and efficiency. Table 1 gives the comparison between DP and SD-DP in accuracy on the dataset of MNIST handwritten digits [12]. SD-DP outperforms DP by all three measures: recall, precision, and F_1 scores. We illustrate in section IV a segmentation of a million-pixel image. With data of this size, SD-DP outpaces DP by at least two orders of magnitude.

II. The primal DP algorithm

In [20], Rodriguez and Laio describe their principle assumptions drawn from empirical observation of cluster properties, introduce the density peaks (DP) algorithm and provide testimonial clustering results. They assume *that cluster centers are surrounded by neighbors with lower local density and that they are at a relatively large distance from any points with a higher local density*. Here, we give only a procedural review of the DP algorithm. We make connections to graph expressions and operations. We also comment on potential faults.

[1]Modified DP algorithm by d'Errico et al. [5]
[2]Histogram of oriented gradients [4]

Provided with a set X of N discrete data points in a feature space, a distance function $d(\cdot, \cdot)$ and a specified value of r for neighborhood radius, the DP algorithm proceeds by the following key steps.

Local density. Every point x_i in X is equipped with a local density score ρ_i,

$$\rho_i = \begin{cases} |\mathcal{N}_r(x_i)|, & \text{hard cutoff} \\ \sum_j \exp\left(-d_{ij}^2/r^2\right), & \text{soft cutoff,} \end{cases} \tag{1}$$

where $d_{ij} = d(x_i, x_j)$ and $\mathcal{N}_r(x_i) = \{x_j \mid 0 < d_{ij} \leq r\}$ contains the neighbors of point x_i within the spherical neighborhood with radius r. In the hard-cutoff case, ρ_i is the neighborhood population count. We note that this neighbor-search step results in a graph $G_r(X, E)$, with X as the node set and E representing the neighbor connections: $e_{ij} \in E$ if and only if $x_j \in \mathcal{N}_r(x_i)$. We refer to G_r as the rNN graph. In the soft-cutoff case, the graph becomes complete with weight $\exp\left(-d_{ij}^2/r^2\right)$ on edge e_{ij}. The costs in computation and storage are $\mathcal{O}(N^2)$.

Decision graph and density peak selection. The DP algorithm constructs a decision graph, from which one locates local density peaks. The decision graph may be seen as a 2D embedding of all data points in X, see for example the scatter plot in Fig. 2b. Point x_i is mapped to (ρ_i, δ_i), with δ_i defined as follows,

$$\delta_i = \min_j\{d_{ij} \mid \rho_j > \rho_i\}, \;\; x_p = \arg\min_j\{d_{ij} \mid \rho_j > \rho_i\}. \tag{2}$$

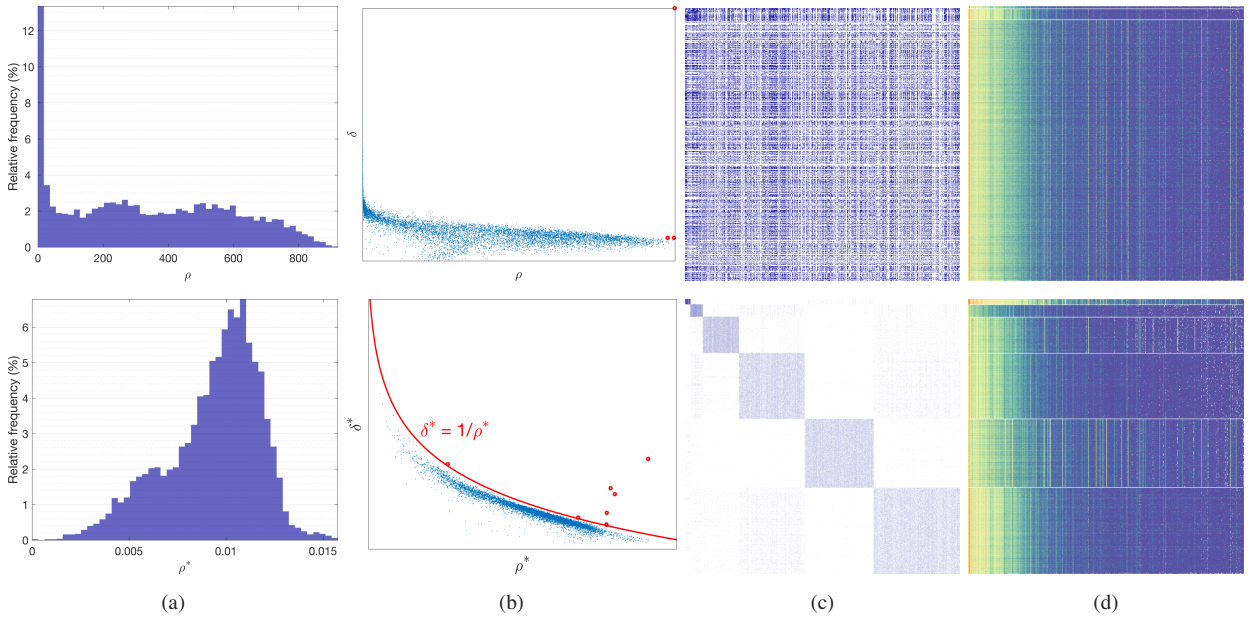


Fig. 2: Differences between DP (top row) and SD-DP (bottom row) in two counterpart components, shown in columns (a) and (b), and in rendered results, shown in columns (c) and (d), on dataset PBMCs-8k [33]. The dataset has $N = 8,000$ cells described by $D = 21,322$ gene expression elements. **Column (a)**: Local density histograms, with 50 equispaced bins for each. The local density ρ with DP is calculated by the soft-cutoff version of (2) with $r = 97.75$, related to $p = 2\%$ via (3). The DP histogram (top) indicates that a large number of data points have near-zero density, i,e., have near-empty neighborhoods. The dual local density ρ^* with SD-DP is calculated by (4) with $k = 35$. The relative frequencies with SD-DP (bottom) are well spread across the bins. **Column (b)**: In each scatter plot, the y-axis for the ascending distance is in log-scale. The decision graph at the top is central to DP for density peak selection. Three DP density peaks (in red) are chosen by the heuristic that peaks have the largest values of $\rho\delta$. The bottom scatter plot in the ρ^*-δ^* plane confirms the proven separation between local maxima and the rest by the red curve $\delta^* = 1/\rho^*$, see Theorem 1. This property gives SD-DP the self-governance and linear complexity in determining the location of local density maxima/peaks. The plot is a by-product for making the comparison, playing no part in decision making. **Column (c)**: The cluster analysis results rendered by DP (top) and SD-DP (bottom) are shown in cell-cell neighbor-interaction matrices. The cells are re-ordered according to the cluster configurations. **Column (d)**: The cluster analysis results are shown in the data array – the rows correspond to the cells; the columns to the gene expression elements. The rows are ordered by the cluster configurations, white horizontal lines are added to highlight cluster boundaries. The columns are ordered in non-descending column sums, only the leading 500 columns are shown. **Conclusion**: The results by SD-DP (cluster revision not shown) suggest 4 large sub-populations of the cells and 2 smaller ones. The data array at the bottom of column (d) offers a visual confirmation of the intra-group similarity and inter-group dissimilarity.

Here, x_p, the parent node of x_i, is the closest point among all with higher local density, and δ_i is the ascending distance from x_i to its parental node. We refer to (2) as the ascending rule. The density peaks are selected from the scatter plot of the decision graph by a heuristic: *only points of high δ and high ρ are the cluster centers* [20]. Additional cues may be used, manually or semi-automatically.

Label propagation. Every density peak is seen as a cluster center, it holds a unique cluster label. Every non-peak point gets a label via its ascending path (2) to one of the density peaks.

Uncertainty assessment. Finally, the DP algorithm grades the points with the same label into two status tiers: core points and halo points. The halo points are de-labeled, for lack of significant affinity. We omit the detail.

A few remarks are in order. (i) The construction of the decision graph is essential to density peak selection. The computational cost for constructing the decision graph is $\mathcal{O}(N^2)$, quadratic with N. (ii) It is not implausible for x_i to face more than one parental candidates. The parental selection of x_i affects the label assignment of x_i and that of its descendants. (iii) There is also uncertainty in numerical calculation of the density and ascending distance. We address, in addition, the problem of determining an appropriate value of radius r for uniform-sized neighborhoods in Section III-A.

III. Sparse Dual of the DP Algorithm

We are mainly concerned with cluster analysis of high dimensional data. We introduce SD-DP, the sparse dual of the DP algorithm, which stems from our understanding of high-dimensional data and the limiting factors of the DP algorithm.

A. Fundamental facts about high-dimensional data

Consider data in a D-dimensional feature space. We assume that every element of the feature descriptor is relevant to discriminative data analysis, in the sense that the element ranges over at least two distinct values. The feature space can thereby house 2^D or more different feature vectors. For convenience, we say the feature space is *deep* if $D > 100$. It is important to respect the fundamental facts about data in a deep feature space. We recognize the following: (i) For dataset X in a deep feature space, the assertion $N = |X| \ll 2^D$ not only holds true today but can also be maintained for many years to come.[3] In other words, the data are sparsely and non-

[3]The largest database as of 2018 is reportedly at the World Data Center for Climate (WDCC), with 220 terabytes of web data and 6 petabytes of extra data.

uniformly scattered in a deep feature space. (ii) With a large and fixed D, the volume of a spherical neighborhood is highly sensitive to a small change in the radius. When the relative change in the radius is ε, the relative change in the volume is $(1+\varepsilon)^D - 1 > D\varepsilon$. (iii) If we fix the radius and let D increase, then the volume of a spherical neighborhood with radius r becomes decreasing when D passes certain value and vanishing as $D \to \infty$.

By the fundamental facts above, one may become aware of the problem with the DP algorithm in selecting the parameter r for uniform-sized neighborhoods. Consider the dataset of single-cell gene expressions for about $8,000$ peripheral blood mononuclear cells (PBMCs-8k) [33]. The feature dimension is high, $D = 21{,}322$, see Fig. 2. The cells are surely sparse and non-uniform in the deep feature space. A small radius will result in empty neighborhoods at many cell points; a large radius may make many neighborhoods equally crowded. Furthermore, the transition from small neighborhoods to large ones, or in between, is difficult to control. If we increase or decrease the DP parameter r by only 1%, then the neighborhood volume expands or shrinks relatively by 2 orders of magnitude. An appropriate value of r is therefore elusive for discriminative data analysis in a deep feature space.

Some attempts in mitigating the problem resort to dimension reduction, which may distort the local densities. Rodriguez and Laio suggested to set the radius at r_p,

$$r_p = \min_r \left\{ r \mid \textstyle\sum_i |\mathcal{N}_r(x_i)| \geq p\, N^2 \right\}, \tag{3}$$

with $p = 1\%$ or 2% so that the average neighbor population is pN. In Fig. 2, the density histogram with the radius set to r_p indicates many near-empty neighborhoods. The heuristic parameter setting (3) is at odds with the fundamental facts about data in a deep feature space. We take a radical departure to resolve the problem.

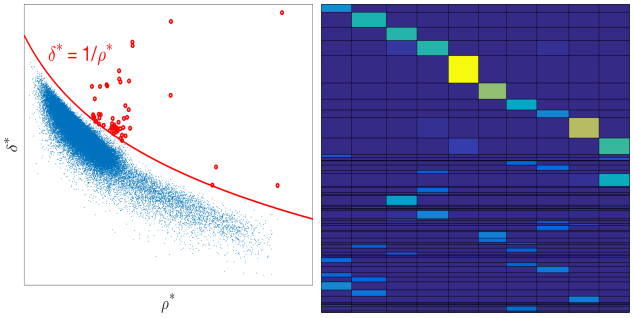
B. Duality in local density description

The local density of point x may be expressed in two alternative ways. One may first specify the radius or range r of a neighborhood, then count the number of points within the neighborhood, as in the DP algorithm. Alternatively, one first specifies the number k of nearest neighbors, then measures the radius of the neighborhood that contains all k neighbors and has the k-th neighbor on the boundary. The local density is higher if the distance to the k-th neighbor is smaller. With high-dimensional data, however, the first choice encounters serious issues as discussed earlier, while the alternative remains effective. The number of nearest neighbors is within grasp, in interpretation as well as in parameter tuning, regardless of the dimension.

Based on this profound difference, we define the *dual local density* at every point x_i as follows,

$$\rho_i^* = 1/\max_j \{ d_{ij} \mid x_j \in \mathcal{N}_k(x_i) \}, \tag{4}$$

where $\mathcal{N}_k(x_i)$ is the set of k nearest neighbors of x_i, and $k > 0$ is a modest constant, not varying with N. The nearest neighbor search results in a kNN graph, which is the sparse counterpart of the rNN graph with DP. There exist efficient algorithms for kNN search.



(a) SD-DP ρ^*-δ^* graph (b) Confusion matrix image

Fig. 3: Local maxima for and confusion matrix with the initial cluster configuration of SD-DP ($k = 48$) on $N = 60{,}000$ images of MNIST handwritten digits with HOG features ($D = 144$). **(a)** ρ^*-δ^* graph, δ^* is in log scale. The red curve $\delta^* = 1/\rho^*$ depicts the proven separation of the 53 local maxima (in red) from the rest of the points, by Theorem 1. **(b)** The confusion matrix between the initial clusters (rows) and the true classes (columns). Color intensity relates to the number of elements per block, brighter blocks contain more elements. The size of each row or column block is the number of data points in the corresponding initial cluster or true class. As observed, each initial cluster is associated with only one true class.

Local density peaks are at *local density maxima*. Point x_i is a local maximum if its local density is higher than that of any other point in its neighborhood. We denote by LocMax$[k]$ the set of local maximum points, specific to k nearest neighbors. The local maximum points, or simply local maxima, can be determined locally, autonomously, and simultaneously at all neighborhoods. There are only k comparisons in local density at each neighborhood. The total computation cost for identifying the local maxima and setting them apart from the rest is $\mathcal{O}(N)$.

C. Label propagation

Every density peak is a local density maximum and holds a unique label. Each non-peak point x_i is connected to a peak via an ascending path and gets the same label the peak holds. The ascending path for each non-peak point is unfolded step by step by the ascending rule:

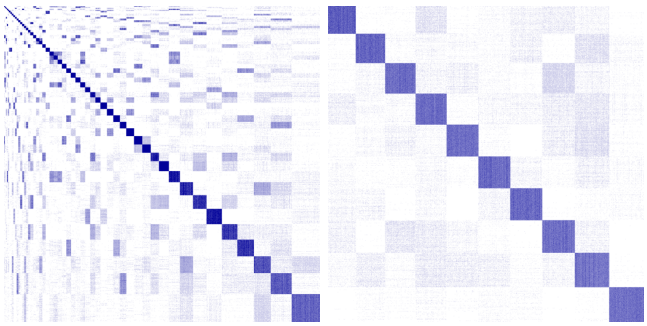
$$\delta_i^* = \min_j \{ d_{ij} \mid \rho_j^* > \rho_i^* \}, x_p = \arg\min_j \{ d_{ij} \mid \rho_j^* > \rho_i^* \}. \tag{5}$$

Here, x_p is the parental point and δ_i^* is the ascending distance from x_i to x_p. We state the following properties:

Theorem 1. *Let $k > 0$ be the specified number of nearest neighbors for each and every point. For any $x_i \notin$ LocMax[k],*

1) *$\rho_i^* \delta_i^* < 1$; and*
2) *its parental node x_p of (5) is in the neighborhood $\mathcal{N}_k(x_i)$.*

For part 1 of Theorem 1, it is straightforward to verify that $\rho_i^* \delta_i^* \geq 1$, or equivalently, $\delta_i^* \geq 1/\rho_i^*$, if and only if x_i is a local maximum. The condition sets local maxima apart from the rest, without resorting to the ρ^*-δ^* graph for decision making. The graph is a by-product. We use it for numerical confirmation and graphical illustration of the simple



(a) (b)

Fig. 4: Matrix view of autonomous cluster revision (see Section III-D) by SD-DP on $N = 60{,}000$ images of MNIST handwritten digits with HOG features ($D = 144$). **(a)** kNN matrix $\mathbf{G}_k$ ($k = 48$) of (6) in block partition according to the initial cluster configuration (53 local maxima). **(b)** The same matrix in block partition according to the revised cluster configuration (10 clusters).

characterization, see the bottom scatter plot in column (b) of Fig. 2 and Fig. 3a, where the local maxima are indeed above the red curve, $\delta^* = 1/\rho^*$; the rest, below.

By part 2 of Theorem 1, the search for all parental nodes is local, autonomous and can be carried out simultaneously at all neighborhoods. The total complexity for parental linking is linear, $\mathcal{O}(N)$. By recursion argument, the ascending path from a non-peak point to a local density maximum is entirely on the kNN graph.

Corollary 2. *The ascending paths from all non-local-maximum points constitute a forest of non-overlapping trees, each ascending tree is rooted at a local maximum. The ascending trees partition and span the kNN graph.*

We can use the ascending forest as our *initial cluster configuration*. The points on the same ascending tree belong to the same cluster, with the same label as that of the local maximum. The total computation cost for constructing the initial cluster configuration is $\mathcal{O}(N)$.

We also introduce the DP counterpart of Theorem 1. We define the local maximum set LocMax[r] in a similar way.

Theorem 3. *Let r be the radius of the neighborhood of each point. For any point $x_i \notin$ LocMax[r],*

1) $\delta_i < r$, and
2) the parental node is within $\mathcal{N}_r(x_i)$.

In theory, by Theorem 3, we may convert the peak selection in DP into autonomous with the horizontal separation line $\delta = r$. In practice, however, there may be a great number of degenerate local maxima with empty neighborhoods, even with r chosen by (3), see Fig. 2a for example.

D. Revision of cluster configuration

The next original contribution we made in SD-DP is on cluster configuration revision. We start with the initial cluster configuration of the ascending trees rooted at the local maxima, see Section III-C. The initial cluster configuration is susceptible to uncertainty in multiple sources such as noise in data, numerical sensitivity in density, distance calculation, and random tie-breaking in parental node selection, which affects the ascending paths of the descendants. It is imperative to assess the current cluster configuration and make revision when necessary and plausible. We describe briefly three key components of our revision scheme: quantitative evaluation of a cluster configuration, governing criteria for revision and revision algorithm.

In evaluating a cluster configuration, we make use of the weighted kNN matrix $\mathbf{G}_k$ defined as follows:

$$\mathbf{G}_k(i,j) = \mathrm{B}_k(i,j)\exp\big(-(d_{ij}\rho_i/\sigma)^2\big), \tag{6}$$

where B_k is the binary-valued adjacency matrix for the kNN graph, $d_{ij}\rho_i$ is the relative distance between x_i and its neighbor x_j against the distance to the k-th nearest neighbor of x_i, and $\sigma > 0$ is a chosen scaling unit. We coalesce the pairwise interactions in $\mathbf{G}_k$ according to any particular cluster configuration.

Assume that the current configuration $\{\mathcal{C}_p\}$, $1 \le p \le L$, has L clusters. We denote by $\mathbf{G}_k(\{\mathcal{C}_p\})$ the matrix permuted and blocked according to the configuration. The diagonal block $\mathbf{G}_k(\mathcal{C}_p, \mathcal{C}_p)$ represents interactions within cluster $\mathcal{C}_p$. The off-diagonal block $\mathbf{G}_k(\mathcal{C}_p, \mathcal{C}_q)$ contains inter-cluster interactions between two clusters $\mathcal{C}_p$ and $\mathcal{C}_q$. Within each diagonal block, any sub-cluster structure can be maintained in the same fashion, recursively.

To govern a revision process, we formulate the clustering problem as an optimization over all feasible cluster configurations,

$$\begin{aligned} &\{\mathcal{C}_\ell\} = \arg\min_{\{\mathcal{C}_p\}} f(\{\mathcal{C}_p\}) = \sum_p |\mathcal{C}_p|^2 \\ &\text{subject to} \\ &h(\mathbf{G}_k(\mathcal{C}_p, \{\mathcal{C}_q\} - \mathcal{C}_p)) < \tau \cdot h(\mathbf{G}_k(\mathcal{C}_p, \mathcal{C}_p)). \end{aligned} \tag{7}$$

The objective function f measures the total area of the diagonal blocks. The function h, in the inequality constraints, aggregates the interaction strength over a given (sub)matrix. The optimization promotes a non-overlapping cluster configuration with smaller and denser diagonal blocks among other feasible configurations. By the feasible condition, the inter-cluster interactions are relatively weaker than the intra-cluster interactions, $\tau > 0$ is a small threshold on the relative ratio.

Governed by (7), a revision algorithm can be carried out with two basic operations: split and merge. The split operation assumes, and builds upon, sub-cluster structure. In the initial configuration, a cluster is an ascending tree, its sub-clusters are the sub-trees. The hierarchical intra-structure can be maintained in revision. In Fig. 6 we illustrate the revision operations via evaluation of the kNN matrix with two clusters, $\mathcal{C}_p$ and $\mathcal{C}_q$, in the current configuration. The sub-cluster $\mathcal{C}_u$ in $\mathcal{C}_p$ has weak intra-interactions with the rest, see submatrices $\mathbf{G}_k(\mathcal{C}_u, \mathcal{C}_p - \mathcal{C}_u)$ and $\mathbf{G}_k(\mathcal{C}_p - \mathcal{C}_u, \mathcal{C}_u)$. Splitting $\mathcal{C}_u$ from $\mathcal{C}_p$ will decrease the value of the objective function f while maintaining the feasibility conditions. When cluster merges are necessary to reduce inter-cluster interactions, we tailor

(a) Santorini HD image[4] (b) Segmentation result

Fig. 5: Illustrative segmentation of a high-definition (HD) image by SD-DP. **(a)** An HD color image of Santorini with $N = 1280 \times 800 = 1{,}024{,}000$ pixels per color channel. **(b)** Segmentation result (30 segments) by SD-DP ($k = 71$). The features are patches of 9×9 pixels per color ($D = 9 \times 9 \times 3 = 243$). The execution time is 15 seconds on MATLAB (excluding kNN construction time).

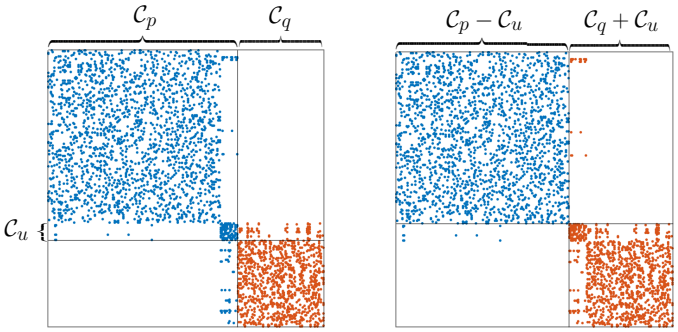


Fig. 6: Illustration of split and merge on a matrix $\mathbf{G}_k$ with two clusters, $\mathcal{C}_p$ and $\mathcal{C}_q$. Interactions ending in $\mathcal{C}_p$ and $\mathcal{C}_q$ are colored blue and brown, respectively. **Left**: Initial cluster configuration. **Right**: Cluster configuration after split and merge. Sub-cluster $\mathcal{C}_u$ of $\mathcal{C}_p$ is automatically detected, split from $\mathcal{C}_p$ and merged into $\mathcal{C}_q$.

and modify the current configuration with splits for measured merges. In Fig. 6, sub-cluster $\mathcal{C}_u$ has a strong interaction with $\mathcal{C}_q$, by evaluation of $\mathbf{G}_k(\mathcal{C}_u, \mathcal{C}_q)$ and $\mathbf{G}_k(\mathcal{C}_q, \mathcal{C}_u)$. If oblivious to the substructure, one would merge the entire $\mathcal{C}_p$ with $\mathcal{C}_q$, at the expense of increasing f. Instead, we split $\mathcal{C}_u$ from $\mathcal{C}_p$ and merge it with $\mathcal{C}_q$. This membership change of $\mathcal{C}_u$ not only decreases the h value but also makes the f value lower. A merge tailored by split renders a much better result than the inordinate merge. One may locate splits at various substructure levels and organize merges in different fashions. For the experiments reported here, we used a simple two-phase procedure for autonomous revision. Phase one uses the substructures at a coarse level for splits and merges, phase two makes further revision at a finer level. Supplementary material and additional information are available at http://sddp.cs.duke.edu.

IV. Concluding Remarks

With SD-DP, the sparse dual version of the DP algorithm, we have made significant advances in non-parametric, unsupervised classification analysis of big and high-dimensional data. SD-DP embodies several intellectual merits. The robustness is the first and foremost. It stems from the dual local density, which respects the fundamental facts about high-dimensional data. The initial configuration, following the DP principle, is entirely unsupervised, proven theoretically efficient with linear complexity and shown empirically faster than DP by orders of magnitude in execution time. We make cluster revision at multiple substructure levels, using merges tailored by splits. The revision is governed by the optimization model (7), which allows us to leverage collective information on relative strength between intra- and inter-cluster interactions. All SD-DP operations can be cast as familiar sparse graph/matrix operations.

SD-DP promises to surmount multiple serious barriers at once in permissible data type, size and dimension, algorithmic robustness, estimation accuracy, and computational efficiency. Still in its infancy, SD-DP is to be applied to, and assessed by, more and diverse datasets, and likely get improved upon further investigation.

Acknowledgements We thank the anonymous reviewers for their valuable comments, thank George Bisbas for assistance in experiments and thank Alexandros-Stavros Iliopoulos for multiple suggestions on the manuscript revision. This work was partially supported by the Hellenic General Secretariat of Research and Technology and the ERA.NET RUS Plus program.

[4]https://blog.ryanair.com/wp-content/uploads/2015/08/santorini123.jpg

REFERENCES

[1] M. Ankerst *et al.*, "OPTICS: ordering points to identify the clustering structure," *ACM Sigmod Record*, pp. 49–60, 1999.

[2] A. Buades, B. Coll, and J.-M. Morel, "A non-local algorithm for image denoising," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 2, 2005, pp. 60–65.

[3] D. Comaniciu and P. Meer, "Mean shift: a robust approach toward feature space analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603–619, 2002.

[4] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2005, pp. 886–893.

[5] M. d'Errico *et al.*, "Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering," 2018, arXiv:1802.10549.

[6] I. S. Dhillon, "Co-clustering documents and words using bipartite spectral graph partitioning," in *Proceedings of 7th International Conference on Knowledge Discovery and Data Mining*, 2001, pp. 269–274.

[7] M. Ester *et al.*, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining*, 1996, pp. 226–231.

[8] M. Girvan and M. E. Newman, "Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, vol. 99, no. 12, pp. 7821–7826, 2002.

[9] D. Grün *et al.*, "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature*, vol. 525, pp. 251–255, 2015.

[10] G. E. Hinton and S. T. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems*, 2003, pp. 857–864.

[11] A. M. Klein *et al.*, "Droplet barcoding for single cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[12] Y. Lecun *et al.*, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] J. M. Lee and E. L. Sonnhammer, "Genomic gene clustering analysis of pathways in eukaryotes," *Genome research*, vol. 13, no. 5, pp. 875–882, 2003.

[14] S. P. Lloyd, "Least squares quantization in PCM," *IEEE Transactions on Information Theory*, vol. 28, no. 2, pp. 129–137, 1982.

[15] D. G. Lowe, "Local feature view clustering for 3d object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, 2001.

[16] ——, "Object recognition from local scale-invariant features," in *IEEE International Conference on Computer Vision*, vol. 2, 1999, pp. 1150–1157.

[17] D. Martin *et al.*, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 416–423.

[18] M. E. J. Newman, "Modularity and community structure in networks," *Proceedings of the National Academy of Sciences*, vol. 103, no. 23, pp. 8577–8582, 2006.

[19] A. P. Patel *et al.*, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.

[20] A. Rodriguez and A. Laio, "Clustering by fast search and find of density peaks," *Science*, vol. 344, no. 6191, pp. 1492–1496, 2014.

[21] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[22] N. Scaringella, G. Zoia, and D. Mlynek, "Automatic genre classification of music content: a survey," *IEEE Signal Processing Magazine*, vol. 23, no. 2, pp. 133–141, 2006.

[23] J. Shao *et al.*, "Clustering molecular dynamics trajectories: 1. characterizing the performance of different clustering algorithms," *Journal of chemical theory and computation*, vol. 3, no. 6, pp. 2312–2334, 2007.

[24] A. Shepitsen *et al.*, "Personalized recommendation in social tagging systems using hierarchical clustering," in *Proceedings of the 2008 ACM conference on Recommender systems*, 2008, pp. 259–266.

[25] J. Shi and J. Malik, "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.

[26] A. W. Smeulders *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 12, pp. 1349–1380, 2000.

[27] S. Sobolevsky *et al.*, "General optimization technique for high-quality community detection in complex networks," *Physical Review E*, vol. 90, no. 1, 2014.

[28] J. B. Tenenbaum, V. d. Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

[29] L. van der Maaten and G. Hinton, "Visualizing data using t-SNE," *Journal of Machine Learning Research*, vol. 9, pp. 2579–2605, 2008.

[30] W. Xu, X. Liu, and Y. Gong, "Document clustering based on non-negative matrix factorization," in *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, 2003, pp. 267–273.

[31] I. Zehavi *et al.*, "Galaxy clustering in early sloan digital sky survey redshift data," *The Astrophysical Journal*, vol. 571, no. 1, p. 172, 2002.

[32] A. Zeisel *et al.*, "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.

[33] G. X. Zheng *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, 2017.