SD-DP: Sparse Dual of the Density Peaks Algorithm for Cluster Analysis of High-Dimensional Data

November 5, 2018




Dimitris Floros[1] Tiancheng Liu[2] Nikos Pitsianis[12] Xiaobai Sun[2]

[1]Department of Electrical and Computer Engineering, Aristotle University of Thessaloniki
[2]Department of Computer Science, Duke University

The first two authors contributed equally to this work

Outline

1. **Cluster analysis of high-dimensional data**

2. **The Density Peaks (DP) and other influential algorithms**

3. **SD-DP: Sparse Dual of the DP algorithm**

4. **Experimental evidence**
   - **Benchmarks**
   - **Exploratory results**

1. **Cluster analysis of high-dimensional data**

2. The Density Peaks (DP) and other influential algorithms

3. SD-DP: Sparse Dual of the DP algorithm

4. Experimental evidence
    - Benchmarks
    - Exploratory results

Cluster analysis of high-dimensional data

Premise: intrinsic heterogeneous group/cluster structures in real-word data of research interest

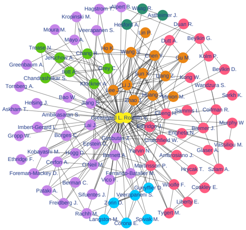
Cluster analysis: uncover cluster structures in data, with noise and uncertainty, with quantified features, governed by certain differentiation criteria

- massive data of many attributes/features
- supervised vs. *un-supervised*

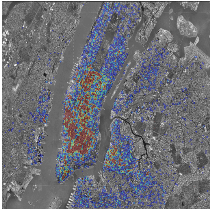
Fundamental to various research studies

Domain-specific analysis	Feature description
Molecular dynamics trajectory patterns [1]	kinetic, spectral measurements
Classification of astronomical events [2]	Gamma ray measurements
Community detection in complex system [3, 4, 5]	link features
Image segmentation/denoising [6, 7]	intensity, patch texture
Content-based image retrieval [8]	semantic content descriptor
Image object recognition [9, 10]	SIFT [11], HOG [12] descriptors
Gene expression pattern analysis [13, 14, 15, 16, 17]	gene-expression matrix
Thematic categorization of documents [18, 19]	word frequency vector
Statistical semantic or sentiment analysis	GloVe [20] word vector
Statistical categorization of musical genres [21]	musical surface features
Consumer profiling/market segmentation [22]	purchase history

Abell 901/902 supercluster [23]



Co-authorship communities [25]



\[-1.5em]

Uber & Taxi demand in NYC [24]

US city lights [26]

DP, other influential algorithms & SD-DP

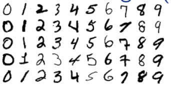
Algorithms / Desirable properties[1]	K-MEANS [27] (1982)	DBSCAN [28] (1996)	OPTICS [29] (1999)	MEAN SHIFT [30] (2002)	GN [3] (2002)	COMBO [5] (2014)	**DP** [31] **(2014)**	**SD-DP** [32] **(2018)**
No prescription of # clusters		✓	✓	✓	✓	✓	✓	✓
No restriction in cluster shape		✓	✓	✓	✓	✓	✓	✓
Free choice of metrics		✓	✓		✓	✓	✓	✓
Agnostic to distribution		✓	✓	✓			✓	✓
Easy or no tuning	✓				✓	✓		✓
Robust in high-dim. space								✓
Accurate in high-dim. space								✓
Low computation cost								✓

Checkmarks are based on limited benchmarking experiments

[1] Additional properties include low program complexity, stability and more

DP vs SD-DP: classification accuracy

60,000 images of handwritten digits (MNIST dataset) [33]



DP (2018) [34]

Estimated Clusters (rows) vs True Classes (columns)

	0	1	2	3	4	5	6	7	8	9	
0	5893 9.8%	4 0.0%	0 0.0%	5 0.0%	2 0.0%	3 0.0%	12 0.0%	0 0.0%	1 0.0%	3 0.0%	99.5% 0.5%
1	0 0.0%	5032 8.4%	1688 2.8%	0 0.0%	6 0.0%	0 0.0%	5 0.0%	5 0.0%	3 0.0%	3 0.0%	74.6% 25.4%
2	42 0.1%	141 0.2%	5699 9.5%	16 0.0%	3 0.0%	2 0.0%	13 0.0%	32 0.1%	4 0.0%	6 0.0%	95.7% 4.3%
3	2 0.0%	8 0.0%	304 0.5%	5690 9.5%	1 0.0%	46 0.1%	5 0.0%	22 0.0%	28 0.0%	25 0.0%	92.8% 7.2%
4	3 0.0%	34 0.1%	15 0.0%	1 0.0%	5405 9.0%	0 0.0%	19 0.0%	7 0.0%	2 0.0%	356 0.6%	92.5% 7.5%
5	5 0.0%	3 0.0%	58 0.1%	104 0.2%	9 0.0%	5089 8.5%	82 0.1%	6 0.0%	8 0.0%	57 0.1%	93.9% 6.1%
6	14 0.0%	21 0.0%	1 0.0%	1 0.0%	3 0.0%	9 0.0%	5867 9.8%	0 0.0%	1 0.0%	1 0.0%	99.1% 0.9%
7	1 0.0%	44 0.1%	1117 1.9%	0 0.0%	21 0.0%	0 0.0%	0 0.0%	5048 8.4%	0 0.0%	34 0.1%	80.6% 19.4%
8	11 0.0%	48 0.1%	36 0.1%	115 0.2%	42 0.1%	91 0.2%	34 0.1%	7 0.0%	5374 9.0%	93 0.2%	91.8% 8.2%
9	10 0.0%	3 0.0%	22 0.0%	54 0.1%	1065 1.8%	12 0.0%	1 0.0%	51 0.1%	14 0.0%	4717 7.9%	79.3% 20.7%
	98.5% 1.5%	94.3% 5.7%	63.7% 36.3%	95.1% 4.9%	82.4% 17.6%	96.9% 3.1%	97.2% 2.8%	97.5% 2.5%	98.9% 1.1%	89.1% 10.9%	89.7% 10.3%

1688 2.8% error

74.6% 25.4% precision

89.1% 10.9% recall

89.7% 10.3% total accuracy

Intensity feature vector ($D = 28 \times 28 = 784$)
Tangent distance
Manual intervention in peak selection and cluster merge

SD-DP

Estimated Clusters (rows) vs True Classes (columns)

	0	1	2	3	4	5	6	7	8	9	
0	5866 9.8%	2 0.0%	21 0.0%	11 0.0%	8 0.0%	18 0.0%	18 0.0%	10 0.0%	38 0.1%	54 0.1%	97.0% 3.0%
1	19 0.0%	6498 10.8%	8 0.0%	0 0.0%	15 0.0%	4 0.0%	10 0.0%	2 0.0%	29 0.0%	2 0.0%	98.6% 1.4%
2	1 0.0%	141 0.2%	5570 9.3%	21 0.0%	4 0.0%	0 0.0%	0 0.0%	38 0.1%	3 0.0%	0 0.0%	96.4% 3.6%
3	4 0.0%	2 0.0%	192 0.3%	5866 9.8%	0 0.0%	28 0.0%	1 0.0%	54 0.1%	19 0.0%	14 0.0%	94.9% 5.1%
4	1 0.0%	30 0.1%	26 0.0%	1 0.0%	5484 9.1%	6 0.0%	0 0.0%	19 0.0%	14 0.0%	4 0.0%	98.2% 1.8%
5	4 0.0%	0 0.0%	7 0.0%	27 0.0%	0 0.0%	5178 8.6%	11 0.0%	3 0.0%	28 0.0%	6 0.0%	98.4% 1.6%
6	13 0.0%	4 0.0%	6 0.0%	1 0.0%	82 0.1%	48 0.1%	5870 9.8%	0 0.0%	42 0.1%	5 0.0%	96.7% 3.3%
7	5 0.0%	23 0.0%	71 0.1%	19 0.0%	7 0.0%	2 0.0%	0 0.0%	5902 9.8%	5 0.0%	8 0.0%	97.7% 2.3%
8	5 0.0%	15 0.0%	32 0.1%	120 0.2%	25 0.0%	116 0.2%	8 0.0%	17 0.0%	5526 9.2%	43 0.1%	93.6% 6.4%
9	5 0.0%	27 0.0%	25 0.0%	65 0.1%	217 0.4%	21 0.0%	0 0.0%	220 0.4%	147 0.2%	5813 9.7%	88.9% 11.1%
	99.0% 1.0%	96.4% 3.6%	93.5% 6.5%	95.7% 4.3%	93.9% 6.1%	95.5% 4.5%	99.2% 0.8%	94.2% 5.8%	94.4% 5.6%	97.7% 2.3%	96.0% 4.0%

8 0.0% error

98.6% 1.4% precision

97.7% 2.3% recall

96.0% 4.0% total accuracy

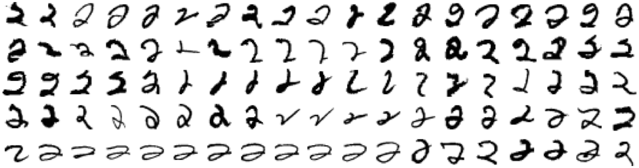
HOG descriptors ($D = 144$)
Euclidean distance
Unsupervised cluster revision

DP vs SD-DP: classification accuracy

Digit	**DP**(2018) semi-supervised	**SD-DP** un-supervised
0	0.99	0.98
1	0.83	0.98
2	0.77	0.95
3	0.94	0.95
4	0.87	0.96
5	0.95	0.97
6	0.98	0.98
7	0.88	0.96
8	0.95	0.94
9	0.84	0.93

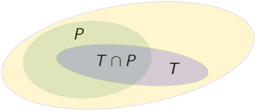
Comparison in Dice similarity coefficients (DSC)
a.k.a. F1 scores and Sørensen-Dice coefficients
60,000 images of handwritten digits (MNIST dataset)

All misclassified digit-*0* images by SD-DP



Subset of misclassified digit-*2* images by SD-DP

$$\text{DSC} = \frac{2TP}{2TP + FP + FN} = \frac{2|T \cap P|}{|T| + |P|}$$



1. Cluster analysis of high-dimensional data

2. The Density Peaks (DP) and other influential algorithms

3. **SD-DP: Sparse Dual of the DP algorithm**

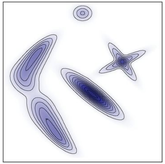
4. Experimental evidence
   - Benchmarks
   - Exploratory results

The Density Peaks principle

[Rodriguez and Laio, Science, 2014]

Principle

> *"Cluster centers are characterized by a higher density than their neighbors and by a relatively large distance from points with higher densities".*

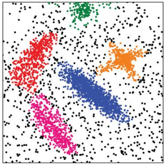


Probability distribution from which point distributions are drawn. The regions with lowest intensity correspond to a background uniform probability of 20%.

Local density description

population in neighborhood of specified radius r

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



Point distribution for samples of 4000 points. Points are colored according to the cluster to which they are assigned. Black points belong to the cluster halos.

Fundamental facts about deep feature space

Deep feature space[1]: $D > 100$

Fact 1:

$$2^D >> N$$

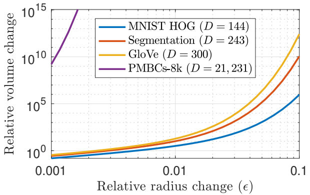
Data are **sparsely**, **non-uniformly** scattered

Fact 2: With D fixed, the hyper-ball volume is **highly sensitive** to radius change

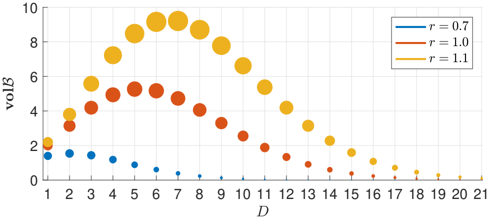
$$\textbf{vol}\mathcal{B}(r(1+\epsilon))/\textbf{vol}\mathcal{B}(r) = (1+\epsilon)^D$$

Fact 3: With radius r fixed, the hyper-ball volume is **vanishing**

$$\textbf{vol}\mathcal{B}(r) \to 0 \text{ as } D \to \infty$$



Fact 2 on specific feature dimensions for 4 particular datasets



Fact 3 on 3 radious values at the low end of dimensions
Each hyper-ball $\mathcal{B}$ is depicted by the disk of area $\textbf{vol}\mathcal{B}(r)$

[1] Largest database (as of 2018): World Data Center for Climate (WDCC) – 6 petabytes (2^{50} bytes) of data

Limitations of DP in deep feature space

By the fundamental Facts about data in a deep feature space

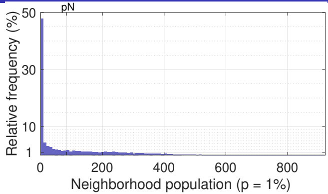
- small radius → many *empty* neighborhoods
- large radius → many *equally crowded* neighborhoods
- adequately discriminative radius values are *elusive*

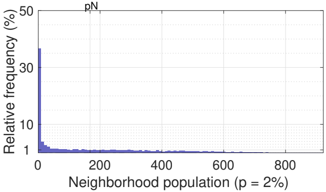
Rodriguez and Laio suggested a heuristic approach: let radius

$$r = \min_d \left\{ d \mid \sum_i |\mathcal{N}_d(x_i)| \geq p\, N^2 \right\}$$

with $p = 1\%$, 2% so that $\mathrm{avg}(\rho) = p\,N$

See the histograms to the right





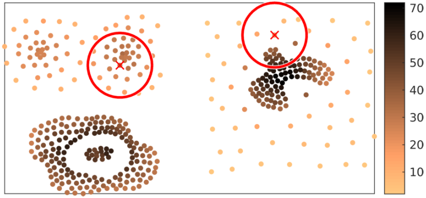
Histograms of neighborhood population, over 50 equispaced bins, with dataset PBMCs-8k of $N = 8{,}000$ cells, $D = 21{,}321$ genes [35]. The neighor radius values are determined by the heurstic described on the left with $p = 1\%, 2\%$ for the top and bottom histograms, respectively. In each case, the local density at a large portion of data points is close to zero

Duality in local density description: neighborhood radius vs population

DP $\rho(\boldsymbol{r})$

#neighbors within distance r

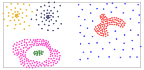
$$\rho_i(r) = \begin{cases} |\mathcal{N}_r(x_i)|, & \text{hard cutoff} \\ \sum_j \exp\left(-d_{ij}^2/r^2\right), & \text{soft cutoff} \end{cases}$$



Local density $\boldsymbol{\rho}$ with $\boldsymbol{r} = 3.1$

Free parameter $\boldsymbol{r}$: real-valued
elusive, volatile in deep space

Labeled Data
Compound[36]

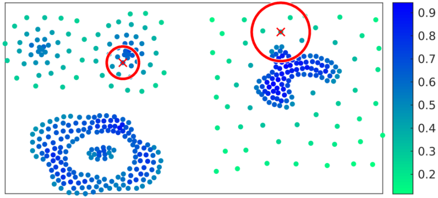


399 points
6 classes

SD-DP $\rho^*(\boldsymbol{k})$

reciprocal distance to the k-th nearest neighbor

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



Dual local density $\boldsymbol{\rho}^*$ with $\boldsymbol{k} = 15$

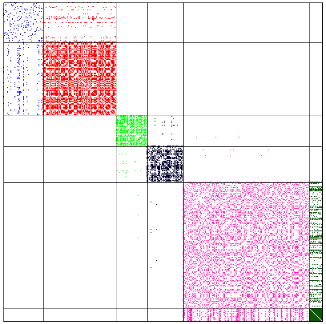
Free parameter $\boldsymbol{k}$: discrete
within grasp, tunable in deep space

Duality in local density description: neighborhood size vs population

DP

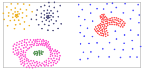
#neighbors within distance r

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



$\mathbf{G}_r$: rNN matrix (Boolean values)
rows/columns ordered by true classes

Labeled Data
Compound

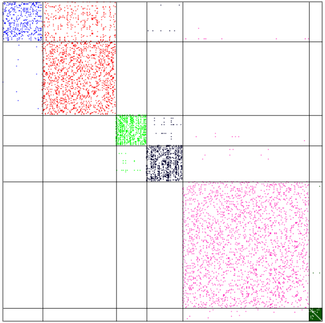


399 points
6 classes

SD-DP

reciprocal distance to the k-th nearest neighbor

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

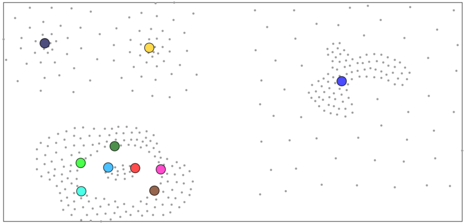


$\mathbf{G}_k$: kNN matrix (Boolean values)
rows/columns ordered by true classes

Density peak location

DP

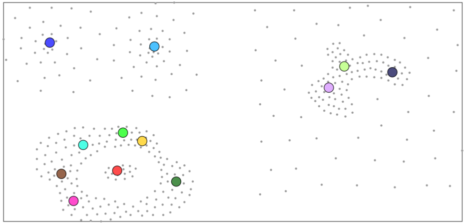
Density peaks are located on ρ-δ **decision graph**
– chosen heuristically or manually
– $O(N^2)$ for ρ-δ graph construction



Compound density peaks (color-coded) with $r = 3.1$

SD-DP

Density peaks are **local maxima** in density
– determined simultaneously, automatically
– $O(N)$, each point makes comparisons with k neighbors



Compound dual density peaks (color-coded) with $k = 15$

Each peak holds a unique label
The rest get labels by **ascending** to the peaks

Ascending rule & ρ-δ graph

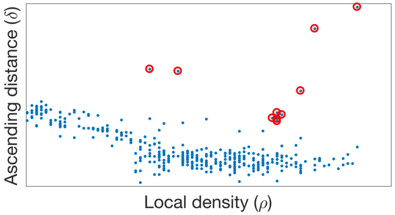
Ascending rule

Every non-peak point i connects to its *nearest* point of *higher* density

$$x_i = \arg\min_j \{d_{ij} \mid \rho_j > \rho_i\}, \quad \text{parental node}$$
$$\delta_i = \min_j \{d_{ij} \mid \rho_j > \rho_i\}, \quad \text{ascending distance}$$

$O(N^2)$ for ρ-δ graph construction

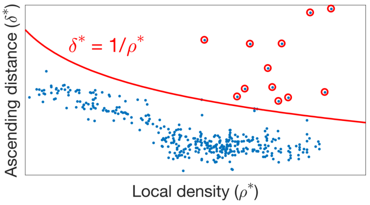
DP decision graph in the ρ-δ plane
Mandatory for peak selection by the heuristic:
"only points of high δ and high ρ are the cluster centers"



DP decision graph with dataset **Compound** for peak selection
Red circles annotate density peaks

$O(N)$, parents located locally on the kNN graph

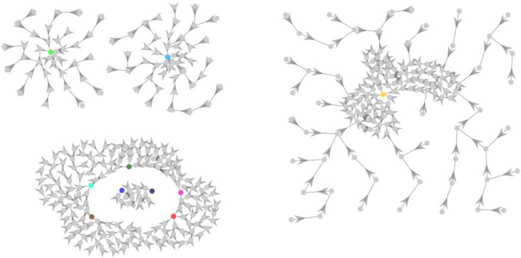
SD-DP (ρ^*-δ^*) **graph**
Visualizing the proven properties of autonomous, linear-cost separation of local maxima from the rest



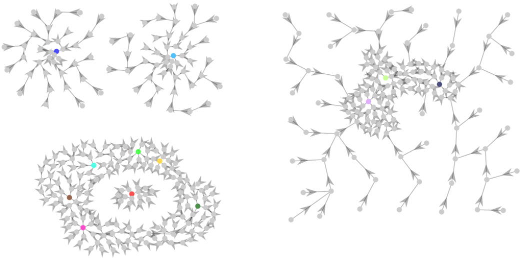
SD-DP visualization graph with dataset **Compound**
Red circles annotate local maxima

Label propagation by ascending rule

DP



SD-DP

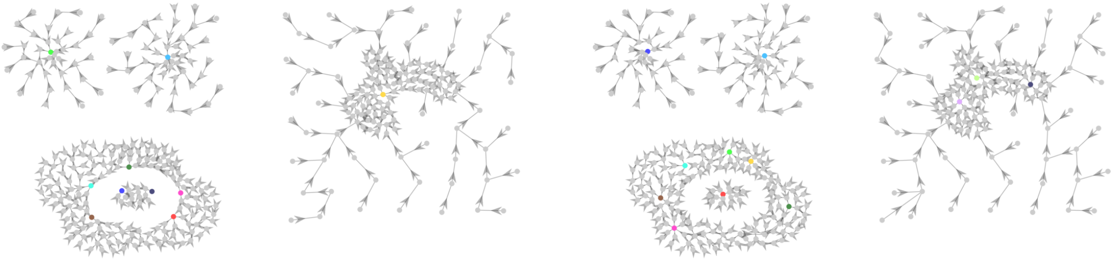


Animation of label propagation with dataset **Compound**

Each peak holds a unique label
The rest get labels by **ascending** to the peaks

Label propagation by ascending rule

DP **SD-DP**

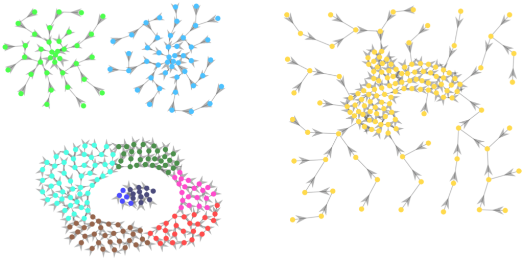


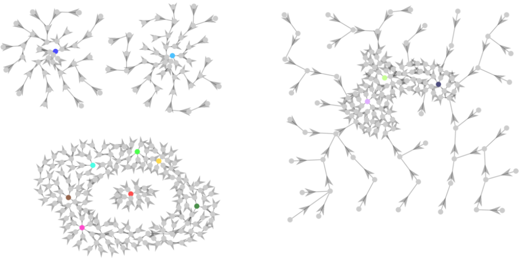
Animation of label propagation with dataset **Compound**

Each peak holds a unique label
The rest get labels by **ascending** to the peaks

Label propagation by ascending rule

DP **SD-DP**





Animation of label propagation with dataset **Compound**

Each peak holds a unique label
The rest get labels by **ascending** to the peaks

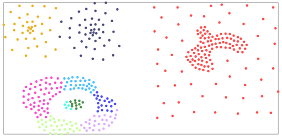
Autonomous revision of cluster configuration

Rationale: multi-source **uncertainty**

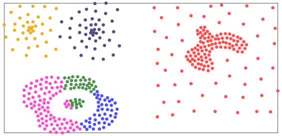
- noise in data
- numerical sensitivity in density calculation
- random tie-breaking in parental node selection

DP

Forward process of peak selection and label propagation without revision



Configuration with 10 clusters

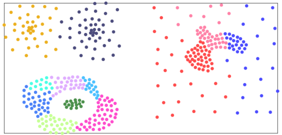


Configuration with #clusters set to 6

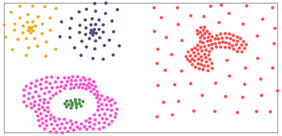
The bottom-left mixture of **Compound** clustered **incorrectly**

SD-DP

Initial configuration of ascending trees at local maxima autonomous revision of cluster configuration



Initial configuration



After revision

The bottom-left mixture of **Compound** clustered **correctly**

Autonomous cluster revision: governing criteria

The weighted kNN matrix

$$\mathbf{G}_k(i,j) = \underbrace{\mathbf{B}_k(i,j)}_{k\text{NN adjacency}} \exp\left(-\left(\underbrace{d_{ij}\,\rho_i^*}_{\text{relative distance}}/\sigma\right)^2\right)$$

is sparse and encodes density-distance information

Initial configuration: L clusters $\{\mathcal{C}_p\}$, $1 \le p \le L$

$\mathbf{G}_k(\{\mathcal{C}_p\})$ is $\mathbf{G}_k$ with columns/rows ordered according to the configuration $\{\mathcal{C}_p\}$

Optimization Objective:

$$\{\mathcal{C}_\ell\} = \arg\min_{\{\mathcal{C}_p\}} f(\{\mathcal{C}_p\}), \qquad f(\{\mathcal{C}_p\}) = \sum_p |\mathcal{C}_p|^2$$

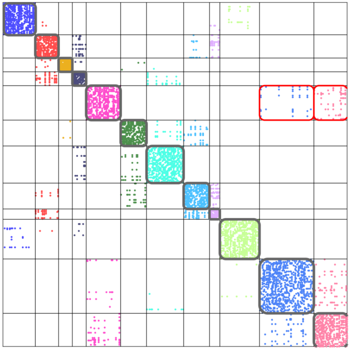
subject to

$$h(\mathbf{G}_k(\mathcal{C}_p, \{\mathcal{C}_q\} - \mathcal{C}_p)) < \tau \cdot h(\mathbf{G}_k(\mathcal{C}_p, \mathcal{C}_p))$$

where

$h(\mathbf{G}_k(\mathcal{C}_p, \mathcal{C}_q))$: aggregated interaction strength of (sub)matrix

τ: a small threshold



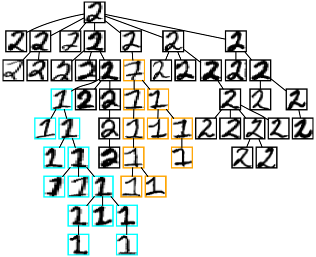
$\mathbf{G}_k$: kNN matrix with rows/columns ordered by initial configuration on **Compound**

Total area of diagonal blocks: $f(\{\mathcal{C}_p\})$
Aggregated interaction strength: $h(\mathbf{G}_k(\mathcal{C}_p, \mathcal{C}_q))$

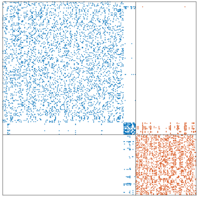
Autonomous cluster revision: split-and-merge

$\mathbf{G}_k$ is sparse, encodes density-distance information
$\mathbf{G}_k(\{\mathcal{C}_p\})$ encodes inter-/intra-cluster interaction strength in addition

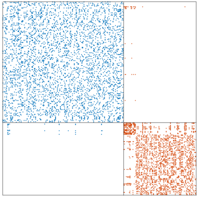
A sub-cluster with weak intra-cluster interaction and stronger interaction with another cluster is **split** from its parent and **merged** to the other
$\Longrightarrow$ Inter-cluster interaction strength h decreases



Subtrees of digit-*1* images, initially attached to the parental tree of digit-*2* images by local density and the ascending rule, are automatically differentiated from the rest and split from the parental tree



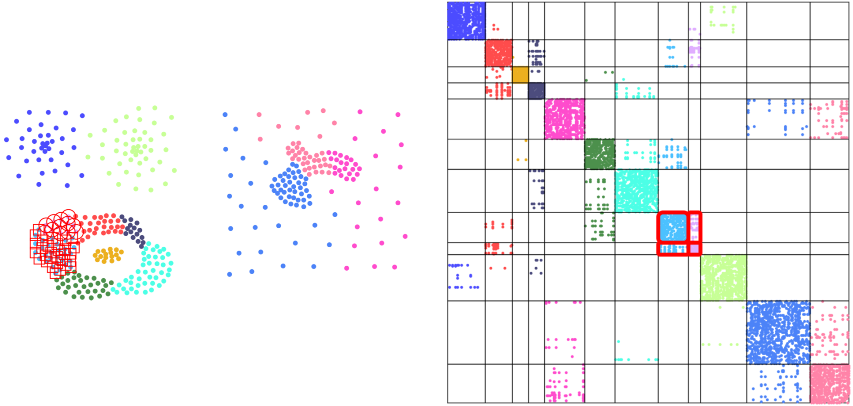
Before split-and-merge



After split-and-merge

Matrix view of split and merge (synthetic construction)

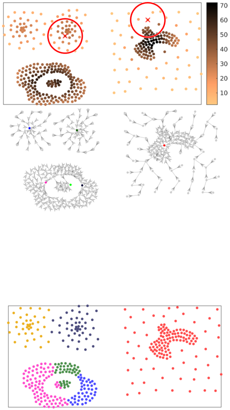
Autonomous cluster revision



Animation of autonomous cluster revision with dataset **Compound**

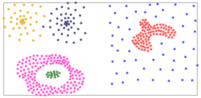
DP vs SD-DP: clustering process & results

DP



The bottom-left mixture clustered **incorrectly**

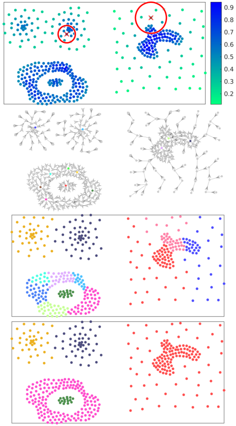
Compound



399 points
6 classes

The right mixture was not separated;
it does not adhere to the DP principle

SD-DP



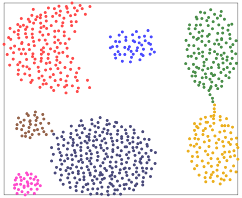
The bottom-left mixture clustered **correctly**

1. Cluster analysis of high-dimensional data

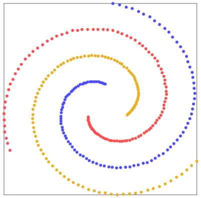
2. The Density Peaks (DP) and other influential algorithms

3. SD-DP: Sparse Dual of the DP algorithm

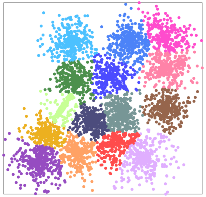
4. **Experimental evidence**
    - **Benchmarks**
    - **Exploratory results**



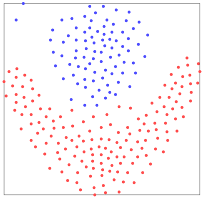
Aggregation
788 points; **7** classes



Spiral
312 points; **3** classes



S3
5,000 points; **15** classes



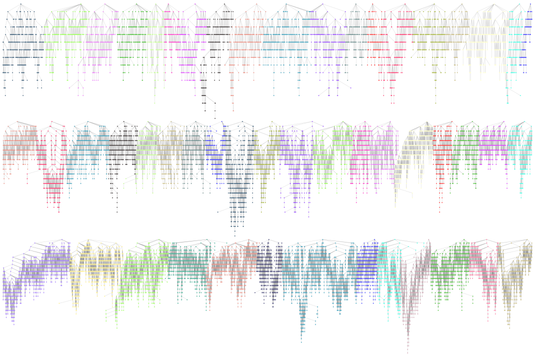
Flame
240 points; **2** classes

SD-DP correctly recovers the numbers and the shapes of the true classes [37]

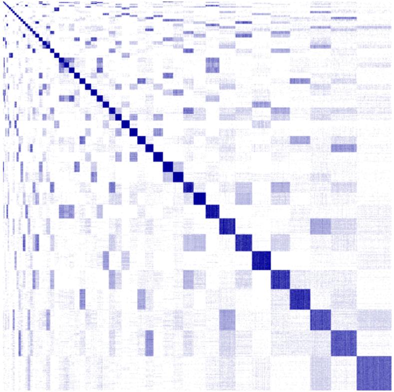
http://cs.uef.fi/sipu/datasets/

Benchmark experiments: handwritten digit recognition

Peaks
(local maxima)



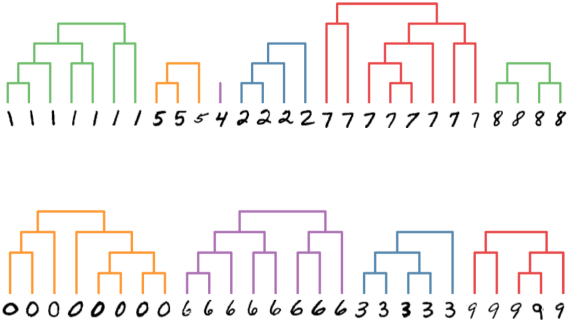
Ascending trees rooted at 53 local maxima; unique color for each tree



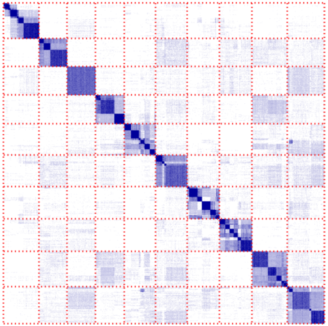
$\mathbf{G}_k$: kNN matrix with rows/columns ordered by clusters.
Clusters are arranged in order of size, $k = 48$

60,000 images of handwritten digits (MNIST dataset)
HOG descriptor (144 dimensions) for each digit image

Benchmark experiments: unsupervised revision



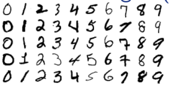
Unsupervised cluster merging, unique color for each merged cluster
Splits took place at a finer level (not shown)



$\mathbf{G}_k$: rows/columns ordered according to two cluster levels – the initial one and the merged one

DP vs SD-DP: classification accuracy

60,000 images of handwritten digits (MNIST dataset) [33]



DP (2018) [34]

Estimated Clusters (rows) × True Classes (columns)

	0	1	2	3	4	5	6	7	8	9	
0	5893 9.8%	4 0.0%	0 0.0%	5 0.0%	2 0.0%	3 0.0%	12 0.0%	0 0.0%	1 0.0%	3 0.0%	99.5% 0.5%
1	0 0.0%	5032 8.4%	1688 2.8%	0 0.0%	6 0.0%	0 0.0%	5 0.0%	5 0.0%	3 0.0%	3 0.0%	74.6% 25.4%
2	42 0.1%	141 0.2%	5699 9.5%	16 0.0%	3 0.0%	2 0.0%	13 0.0%	32 0.1%	4 0.0%	6 0.0%	95.7% 4.3%
3	2 0.0%	8 0.0%	304 0.5%	5690 9.5%	1 0.0%	46 0.1%	5 0.0%	22 0.0%	28 0.0%	25 0.0%	92.8% 7.2%
4	3 0.0%	34 0.1%	15 0.0%	1 0.0%	5405 9.0%	0 0.0%	19 0.0%	7 0.0%	2 0.0%	356 0.6%	92.5% 7.5%
5	5 0.0%	3 0.0%	58 0.1%	104 0.2%	9 0.0%	5089 8.5%	82 0.1%	6 0.0%	8 0.0%	57 0.1%	93.9% 6.1%
6	14 0.0%	21 0.0%	1 0.0%	1 0.0%	3 0.0%	9 0.0%	5867 9.8%	0 0.0%	1 0.0%	1 0.0%	99.1% 0.9%
7	1 0.0%	44 0.1%	1117 1.9%	0 0.0%	21 0.0%	0 0.0%	0 0.0%	5048 8.4%	0 0.0%	34 0.1%	80.6% 19.4%
8	11 0.0%	48 0.1%	36 0.1%	115 0.2%	42 0.1%	91 0.2%	34 0.1%	7 0.0%	5374 9.0%	93 0.2%	91.8% 8.2%
9	10 0.0%	3 0.0%	22 0.0%	54 0.1%	1065 1.8%	12 0.0%	1 0.0%	51 0.1%	14 0.0%	4717 7.9%	79.3% 20.7%
	98.5% 1.5%	94.3% 5.7%	63.7% 36.3%	95.1% 4.9%	82.4% 17.6%	96.9% 3.1%	97.2% 2.8%	97.5% 2.5%	98.9% 1.1%	89.1% 10.9%	89.7% 10.3%

1688
2.8%
error

74.6%
25.4%
precision

89.1%
10.9%
recall

89.7%
10.3%
total accuracy

Intensity feature vector ($D = 28 \times 28 = 784$)
Tangent distance
Manual intervention in peak selection and cluster merge

SD-DP

Estimated Clusters (rows) × True Classes (columns)

	0	1	2	3	4	5	6	7	8	9	
0	5866 9.8%	2 0.0%	21 0.0%	11 0.0%	8 0.0%	18 0.0%	18 0.0%	10 0.0%	38 0.1%	54 0.1%	97.0% 3.0%
1	19 0.0%	6498 10.8%	8 0.0%	0 0.0%	15 0.0%	4 0.0%	10 0.0%	2 0.0%	29 0.0%	2 0.0%	98.6% 1.4%
2	1 0.0%	141 0.2%	5570 9.3%	21 0.0%	4 0.0%	0 0.0%	0 0.0%	38 0.1%	3 0.0%	0 0.0%	96.4% 3.6%
3	4 0.0%	2 0.0%	192 0.3%	5866 9.8%	0 0.0%	28 0.0%	1 0.0%	54 0.1%	19 0.0%	14 0.0%	94.9% 5.1%
4	1 0.0%	30 0.1%	26 0.0%	1 0.0%	5484 9.1%	6 0.0%	0 0.0%	19 0.0%	14 0.0%	4 0.0%	98.2% 1.8%
5	4 0.0%	0 0.0%	7 0.0%	27 0.0%	0 0.0%	5178 8.6%	11 0.0%	3 0.0%	28 0.0%	6 0.0%	98.4% 1.6%
6	13 0.0%	4 0.0%	6 0.0%	1 0.0%	82 0.1%	48 0.1%	5870 9.8%	0 0.0%	42 0.1%	5 0.0%	96.7% 3.3%
7	5 0.0%	23 0.0%	71 0.1%	19 0.0%	7 0.0%	2 0.0%	0 0.0%	5902 9.8%	5 0.0%	8 0.0%	97.7% 2.3%
8	5 0.0%	15 0.0%	32 0.1%	120 0.2%	25 0.0%	116 0.2%	8 0.0%	17 0.0%	5526 9.2%	43 0.1%	93.6% 6.4%
9	5 0.0%	27 0.0%	25 0.0%	65 0.1%	217 0.4%	21 0.0%	0 0.0%	220 0.4%	147 0.2%	5813 9.7%	88.9% 11.1%
	99.0% 1.0%	96.4% 3.6%	93.5% 6.5%	95.7% 4.3%	93.9% 6.1%	95.5% 4.5%	99.2% 0.8%	94.2% 5.8%	94.4% 5.6%	97.7% 2.3%	96.0% 4.0%

8
0.0%
error

98.6%
1.4%
precision

97.7%
2.3%
recall

96.0%
4.0%
total accuracy

HOG descriptors ($D = 144$)
Euclidean distance
Unsupervised cluster revision

1. Cluster analysis of high-dimensional data

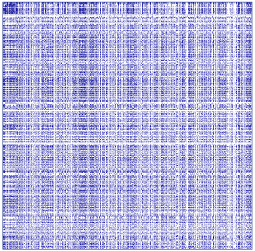
2. The Density Peaks (DP) and other influential algorithms

3. SD-DP: Sparse Dual of the DP algorithm

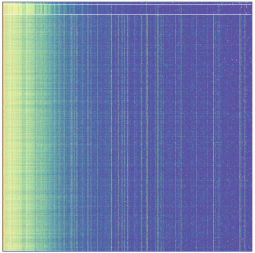
4. **Experimental evidence**
   - Benchmarks
   - **Exploratory results**

DP vs SD-DP: clustering of high-dimensional data

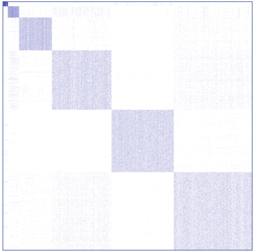
DP | **SD-DP**



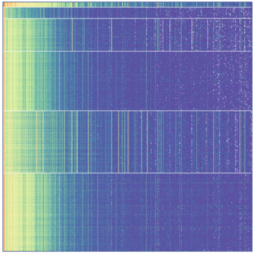
$\mathbf{G}_r$: rNN matrix with rows/columns ordered by rendered clusters



Data matrix
cells (rows) vs genes (columns)



$\mathbf{G}_k$: kNN matrix with rows/columns ordered by rendered clusters



Data matrix
cells (rows) vs genes (columns)

DP with $r = 97.75$ ($p = 2\%$)
Rendered 2 small and 1 large cluster

SD-DP with $k = 35$
Rendered 2 small and 4 large clusters

Dataset PBMCs-8k [35]: $N = 8{,}000$ cells, $D = 21{,}321$ genes

Exploratory experiments: fast image segmentation

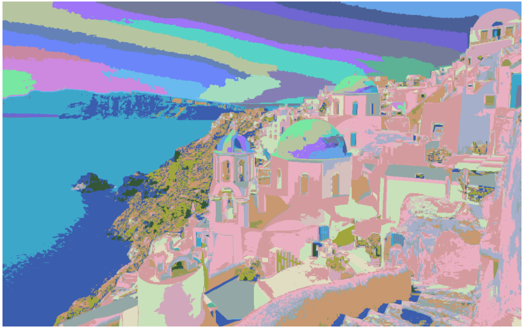
Parthenon image [38] (481 $\times$ 321, N = 154,401)

Segmentation result (3 segments)
5 $\times$ 5 patch feature per color; $D = 5 \times 5 \times 3 = 75$

Segmentation time: 3 seconds in MATLAB (excluding kNN construction)
SD-DP outpaces DP by two orders of magnitude

Exploratory experiments: fast high-definition image segmentation

Santorini image[1] (1280 × 800, $N = 1{,}024{,}000$)



Illustrative segmentation result (30 segments)
9×9 patch feature per color; $D = 9 \times 9 \times 3 = 243$

Segmentation time: 15 seconds in MATLAB (excluding kNN construction)
SD-DP outpaces DP by at least two orders of magnitude

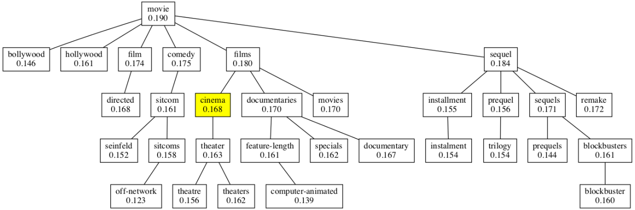
[1] https://blog.ryanair.com/wp-content/uploads/2015/08/santorini123.jpg

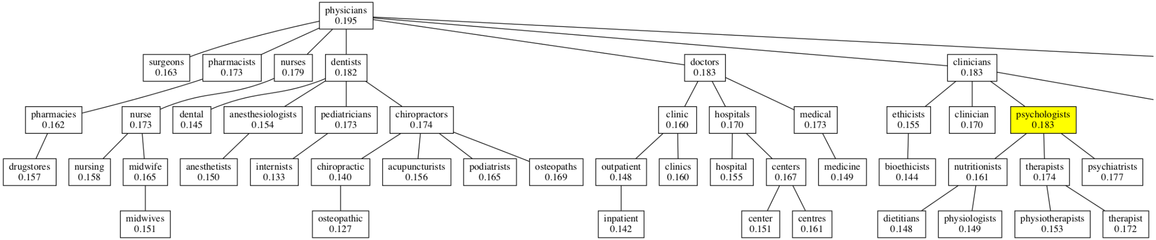
Exploratory experiments: statistical hierarchy of word semantics

$N = 400{,}000$ `GloVe` [20] word vectors[2] ($D = 300$)
Semantically related words, based on word co-occurrence from text content, are closer in the `GloVe` space

SD-DP ($k = 5$) produces a statistical hierarchy of word semantics
A word with higher density has more general meaning
A word with lower density has more specific meaning

Can be used for search in depth and breadth simultaneously
The local density is annotated on each word
Query words are highlighted





[2] Pre-trained word vectors (`Wikipedia 2014 + Gigaword 5`)

Recap: Sparse Dual of Density Peaks

Contributions

Dual local density description
ground for robustness, by recognizing, respecting the fundamental facts of high dimensional data

Initial cluster formation
clusters by ascending trees rooted at local maxima
proven local, parallel, of linear complexity

Autonomous cluster revision
coherent revision criteria at multiple cluster levels

Sparse matrix/graph operations

Experimental findings

Unsupervised classification of handwritten digits
96% overall accuracy reached

Gene clustering
4 large clusters found in 8,000 cells in expression of 21,321 genes

Statistical hierarchy of word semantics
among $400,000$ words in the `GloVe` space ($D = 300$)

HD image segmentation
faster than DP by two orders of magnitude or more

Desirable properties \ Algorithms	K-MEANS (1982)	DBSCAN (1996)	OPTICS (1999)	MEAN SHIFT (2002)	GN (2002)	COMBO (2014)	**DP (2014)**	**SD-DP (2018)**
No prescription of # clusters		✓	✓	✓	✓	✓	✓	✓
No restriction in cluster shape		✓	✓	✓	✓	✓	✓	✓
Free choice of metrics		✓	✓		✓	✓	✓	✓
Agnostic to distribution		✓	✓	✓			✓	✓
Easy or no tuning	✓				✓	✓		✓
Robust in high-dim. space								✓
Accurate in high-dim. space								✓
Low computation cost								✓

Additional information available at `http://sddp.cs.duke.edu`

Acknowledgements

Anonymous reviewers for valuable comments

George Bisbas for assistance in experiments

Alexandros-Stavros Iliopoulos for multiple suggestions

Hellenic General Secretariat of Research and Technology and the ERA.NET RUS Plus program for partial support

References I

[1] J. Shao, S. W. Tanner, N. Thompson, and T. E. Cheatham, "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.

[2] I. Zehavi, M. R. Blanton, J. A. Frieman, D. H. Weinberg, H. J. Mo, M. A. Strauss, S. F. Anderson, J. Annis, N. A. Bahcall, M. Bernardi, J. W. Briggs, J. Brinkmann, S. Burles, L. Carey, F. J. Castander, A. J. Connolly, I. Csabai, J. J. Dalcanton, S. Dodelson, M. Doi *et al.*, "Galaxy clustering in early Sloan digital sky survey redshift data," *The Astrophysical Journal*, vol. 571, no. 1, pp. 172–190, 2002.

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

[4] 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.

[5] S. Sobolevsky, R. Campari, A. Belyi, and C. Ratti, "General optimization technique for high-quality community detection in complex networks," *Physical Review E*, vol. 90, no. 1, 2014.

[6] 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.

[7] 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.

[8] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 22, no. 12, pp. 1349–1380, 2000.

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

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

[11] ——, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

References II

[12] 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.

[13] A. P. Patel, I. Tirosh, J. J. Trombetta, A. K. Shalek, S. M. Gillespie, H. Wakimoto, D. P. Cahill, B. V. Nahed, W. T. Curry, R. L. Martuza, D. N. Louis, O. Rozenblatt-Rosen, M. L. Suvà, A. Regev, and B. E. Bernstein, "Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma," *Science*, vol. 344, no. 6190, pp. 1396–1401, 2014.

[14] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, C. Rolny, G. Castelo-Branco, J. Hjerling-Leffler, and S. Linnarsson, "Cell types in the mouse cortex and hippocampus revealed by single-cell RNA-seq," *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.

[15] D. Grün, A. Lyubimova, L. Kester, K. Wiebrands, O. Basak, N. Sasaki, H. Clevers, and A. van Oudenaarden, "Single-cell messenger RNA sequencing reveals rare intestinal cell types," *Nature*, vol. 525, pp. 251–255, 2015.

[16] A. M. Klein, L. Mazutis, I. Akartuna, N. Tallapragada, A. Veres, V. Li, L. Peshkin, D. A. Weitz, and M. W. Kirschner, "Droplet barcoding for single cell transcriptomics applied to embryonic stem cells," *Cell*, vol. 161, no. 5, pp. 1187–1201, 2015.

[17] 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.

[18] 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.

[19] 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.

[20] J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Empirical Methods in Natural Language Processing*, 2014, pp. 1532–1543.

[21] 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.

References III

[22] A. Shepitsen, J. Gemmell, B. Mobasher, and R. Burke, "Personalized recommendation in social tagging systems using hierarchical clustering," in *Proceedings of the 2008 ACM Conference on Recommender Systems*, 2008, pp. 259–266.

[23] ESO, "An intergalactic heavyweight," 2013, https://www.eso.org/public/images/potw1304a/.

[24] G. Bisbas, "Forecast demand using extended discrete Fourier transform. (in Greek)," Diploma thesis, Aristotle University of Thessaloniki, Greece, 2017.

[25] K. Mylonakis, N. Pitsianis, and X. Sun, "The fast multipole method in three decades," Poster presented at Epstein Greengard Modern Advances in Computational and Applied Mathematics workshop at Yale University, 2017.

[26] NASA, "City Lights of the United States," 2012, https://earthobservatory.nasa.gov/images/79800/city-lights-of-the-united-states-2012.

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

[28] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "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.

[29] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander, "OPTICS: ordering points to identify the clustering structure," *ACM Sigmod Record*, vol. 28, no. 2, pp. 49–60, 1999.

[30] 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.

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

[32] D. Floros, T. Liu, N. Pitsianis, and X. Sun, "Sparse dual of the density peaks algorithm for cluster analysis of high-dimensional data," in *IEEE High Performance Extreme Computing Conference*, 2018.

[33] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

References IV

[34] M. d'Errico, E. Facco, A. Laio, and A. Rodriguez, "Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering," 2018, arXiv:1802.10549.

[35] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu *et al.*, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, 2017.

[36] C. T. Zahn, "Graph-theoretical methods for detecting and describing gestalt clusters," *IEEE Transactions on Computers*, vol. C-20, no. 1, pp. 68–86, 1971.

[37] P. Fränti and S. Sieranoja, "K-means properties on six clustering benchmark datasets," *Applied Intelligence*, vol. 48, 2018, http://cs.uef.fi/sipu/datasets/.

[38] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "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.